

PHIÊN BẢN THỨ 2 ĐÃ
CẬP NHẬT & MỞ RỘNG

DATA DRIVEN CONSTRUCTION

ĐIỀU HƯỚNG KỶ NGUYÊN DỮ LIỆU TRONG NGÀNH XÂY DỰNG

VỚI CÁC TRƯỜNG HỢP SỬ
DỤNG AI & LLM

Artem Boiko


DATA-DRIVE
CONSTRUCTIO
NAVIGATING THE DATA AGE
IN THE CONSTRUCTION INDUS
100+
KEY DATA-RELATED TOPICS


DC
GUIDE


DC
GUIDE


50+
PRACTICAL PROBLEMS SOLVED


300+
UNIQUE ILLUSTRATIONS AND GRAPHS

DATA-DRIVEN CONTRUCTION

Điều hướng trong kỷ nguyên dữ liệu trong ngành xây dựng

Ấn bản thứ hai, đã sửa đổi và bổ sung

ARTEM BOIKO

Boiko là James Carville của IT - trong câu nói nổi tiếng của ông ấy "Nó là kinh tế, ngu ngốc", chỉ cần thay một từ cho cuốn sách nổi tiếng này. "Nó là dữ liệu, ngu ngốc." (không phải phần mềm) Và để tìm đường trong vũ trụ dữ liệu, một câu nói của người La Mã cổ đại từ thời Hy Lạp vẫn còn phù hợp ngày nay: "Navigare necesse est". Tác giả dẫn dắt người đọc xuyên suốt các độ sâu và cạn của đại dương dữ liệu với bàn tay vững vàng và kim chỉ nam không lay chuyển, không kể đến cách tiếp cận lịch sử toàn diện và, cuối cùng nhưng không kém phần quan trọng, đồ họa độc đáo và khiếu hài hước tốt mà không chỉ hiện ra ở cái nhìn thứ hai. Phản ứng quốc tế đối với cuốn sách của Boiko dao động từ sự tán thưởng cuồng nhiệt đến sự hoài nghi khá châm biếm, điều này đã làm cho ấn bản thứ hai tiếng Đức của cuốn sách được hưởng lợi. Boiko là một nhà tư tưởng dữ liệu độc đáo và không giáo điều. Ông cung cấp cho người đọc những hiểu biết thú vị và những luận điểm dũng cảm, thậm chí có phần khiêu khích, làm nảy sinh suy nghĩ thêm. Một liều thuốc xuất sắc cho căn bệnh đồng thuận tiềm tàng của người Đức. Nhân tiện, câu châm ngôn La Mã trên có một bổ sung: "vivere non est necesse." Nó không áp dụng cho cách tiếp cận của Boiko đối với thế giới dữ liệu - dữ liệu sống và sự sống của nó là cần thiết, không cần phải nói, là thiết yếu.

- Dr. Burkhard Talebitari, biên tập viên tự do - bao gồm cho Tạp chí: BIM, phát hành hàng năm bởi Ernst & Sohn từ năm 2013.

Cuốn sách của Artem Boiko là một cột mốc cho việc dân chủ hóa số hóa trong ngành xây dựng - và là một sự thay đổi lớn thực sự cho các doanh nghiệp nhỏ và vừa (SMEs). Điều đặc biệt mang tính đột phá: bằng cách sử dụng các công cụ mã nguồn mở hiện đại với mã số thấp và không mã, các công ty đã có thể tích hợp dữ liệu một cách hiệu quả vào quy trình kinh doanh của họ và phân tích có lợi - mà không cần kiến thức lập trình chuyên sâu. Điều này làm cho việc sử dụng tốn kém các gói phần mềm thương mại rườm rà trở nên dư thừa. Cuốn sách này là một lời kêu gọi hành động! Nó là một hướng dẫn quý giá cho bất kỳ ai không chỉ muốn hiểu sự chuyển đổi số trong ngành xây dựng mà còn muốn chủ động định hình nó - một cách thực tiễn, hiệu quả và hướng tới tương lai. Bây giờ là lúc hợp tác để chia sẻ kiến thức này và tăng cường bền vững năng suất của ngành xây dựng.

- Dr. Michael Max Buehler, Giáo sư Quản lý Xây dựng tại HTWG Konstanz, Đồng sở hữu tại GemeinWerk Ventures, và Giám đốc độc lập tại DevvStream.

Cuốn sách DataDrivenConstruction là một trong những bước đầu tiên vượt qua ranh giới của thế giới nhà thầu thông thường, với các hệ thống thiết kế và quản lý phức tạp của họ, khi mà, dường như, độ phức tạp và bão hòa dữ liệu thậm chí không cho cơ hội nào cho sự đơn giản hóa và tăng cường sự minh bạch trong việc làm việc với dữ liệu xây dựng. Trong cuốn sách của mình, Artem cho thấy bằng ngôn ngữ đơn giản những cơ hội mà công nghệ hiện đại về việc làm việc với dữ liệu mở ra cho chúng ta, và thực sự cung cấp các bước cụ thể mà bạn có thể áp dụng ngay lập tức trong công việc của mình. Tôi kêu gọi mọi người muốn hiểu tìm hiểu hệ thống tự động sẽ tiến triển ra sao trong ngành xây dựng hãy nghiên cứu cuốn sách này cẩn thận để nhận ra rằng cuộc cách mạng dữ liệu trong xây dựng đã gõ cửa chúng ta. Hiện tại chỉ có những người đam mê quan tâm, nhưng trong vài năm nữa, giống như BIM, những phương pháp và phần mềm như vậy sẽ trở nên phổ biến!

- Ihor Rogachew, Trưởng Trung tâm Năng lực IMT, BIM & Chuyển đổi số tại RGD, và Người sáng lập InfraBIM.Pro.

Tôi rất khuyến nghị cuốn sách DataDrivenConstruction đề cập, như tiêu đề đã nói, đến một cách tiếp cận quản lý thông tin dựa trên dữ liệu cho AECO. Tôi hiện đang sử dụng nó để giúp khởi xướng một số cuộc thảo luận với các nhóm khác nhau. Tôi thấy nó là một tài liệu tham khảo rất dễ tiếp cận. Ngoài việc đưa ra cái nhìn tổng quan đầy đủ về bối cảnh lịch sử của các công cụ trong AECO, dữ liệu và giới thiệu một số công nghệ quan trọng, cuốn sách chứa đựng một số sơ đồ rất hữu ích, phác thảo phạm vi các nguồn dữ liệu và các sản phẩm của người sử dụng cuối với các quy trình mẫu. Tôi thấy rằng đây là những sơ đồ kiểu mà chúng ta cần nhiều hơn khi phát triển và theo dõi các chiến lược thông tin và góp phần vào BEP - định nghĩa mô hình dữ liệu toàn doanh nghiệp mà trên đó ranh giới cho một PIM và AIM có thể được phủ lên.

- Paul Ransley, Tư vấn viên chính tại Acmena, và Kỹ sư Tích hợp Hệ thống tại Transport for London.

Nếu "dữ liệu là dầu mới", chúng ta cần học cách định nghĩa, tìm kiếm, khai thác, tinh chế nó, để làm cho nó có giá trị. Tôi thấy cuốn sách DataDrivenConstruction rất hữu ích và sâu sắc. Cuốn sách cung cấp một bối cảnh lịch sử hữu ích và giải thích việc làm việc với dữ liệu bằng ngôn ngữ đơn giản. Đối với những ai quan tâm đến chuyển đổi số, nó giúp cung cấp một hiểu biết tốt về dữ liệu - cách thức hoạt động, cách mà nó được cấu trúc và nó có thể được sử dụng như thế nào.

- Ralph Montague, Giám đốc tại ArcDox, Giám đốc của Hội nghị các Điều phối viên BIM, và Chủ tịch Ủy ban Gương Quốc gia BIM tại Cơ quan Tiêu chuẩn Quốc gia Ireland.

Như đã nhấn mạnh trong cuốn sách, thông tin là một tài sản quan trọng đối với lĩnh vực xây dựng và có nó ở định dạng dễ tiếp cận sẽ tạo điều kiện thuận lợi cho việc ra quyết định chính xác và rút ngắn thời gian thực hiện dự án. Cuốn sách cung cấp một cách tiếp cận trung lập và hiệu quả để truy cập và tận dụng nguồn tài nguyên này trong quá trình ra quyết định. Phương pháp được trình bày trong cuốn sách khai thác cách tiếp cận hiện đại kết hợp giữa lập trình dựa trên trí tuệ nhân tạo và các công cụ mã nguồn mở dễ tiếp cận. Bằng cách tận dụng sức mạnh của trí tuệ nhân tạo và sử dụng phần mềm nguồn mở, phương pháp này nhằm nâng cao tự động hóa, tối ưu hóa quy trình và thúc đẩy tính khả dụng và hợp tác trong lĩnh vực này. Ngôn ngữ của cuốn sách rõ ràng và dễ theo dõi.

- Dr. Salih Ofluoğlu, Dean của Khoa Mỹ thuật và Kiến trúc tại Đại học Antalya Bilim, và Tổ chức Diễn đàn BIM Eurasia.

Tất cả những gì tôi có thể nói là, WOW! Cách bạn kết hợp lịch sử, LLM, đồ họa và sự dễ dàng trong việc hiểu các điểm của bạn thật sự đáng kinh ngạc. Sự mạch lạc của cuốn sách thật tuyệt vời. Có rất nhiều khía cạnh xuất sắc trong cuốn sách này; nó thực sự là một bước ngoặt. Đây là nguồn thông tin tuyệt vời, và tôi khen ngợi bạn vì sự nỗ lực và đam mê mà bạn đã bỏ vào đó. Xin chúc mừng bạn vì đã tạo ra một tác phẩm đáng chú ý như vậy. Tôi có thể tiếp tục, nhưng chỉ cần nói là tôi vô cùng ấn tượng!

— Natasha Prinsloo, Trưởng phòng Thực hành Kỹ thuật số tại energylab_.

Đối với bất kỳ ai trong ngành xây dựng, từ những người mới bắt đầu đến những chuyên gia dày dạn, cuốn sách này là một bước ngoặt! Nó không phải là một cuốn sách khô khan như thường lệ — nó tràn đầy thông tin, chiến lược và một chút hài hước để giữ bạn luôn hứng thú. Từ các phương pháp ghi dữ liệu cổ xưa đến công nghệ số tiên tiến, nó bao quát sự tiến hóa của việc sử dụng dữ liệu trong xây dựng. Nó giống như một chiếc máy thời gian đưa bạn qua sự phát triển của dữ liệu trong xây dựng. Dù bạn là kiến trúc sư, kỹ sư, quản lý dự án hay nhà phân tích dữ liệu, cuốn hướng dẫn toàn diện này sẽ cách mạng hóa cách bạn tiếp cận các dự án. Hãy sẵn sàng tối ưu hóa quy trình, nâng cao quyết định, và quản lý dự án như chưa bao giờ có trước đây!

— Pierpaolo Vergati, Giảng viên tại Đại học Sapienza Roma, và Quản lý Dự án Xây dựng cấp cao tại Fintecna.

Tôi đọc cuốn sách trong một hơi, chưa đầy 6 giờ. Chất lượng sản xuất của cuốn sách thật xuất sắc, với giấy bóng dày, màu sắc hài hòa, phông chữ dễ chịu. Số lượng ví dụ thực tiễn phong phú về cách làm việc với LLM liên quan đến ngành xây dựng sẽ tiết kiệm cho bạn hàng tháng, nếu không muốn nói là hàng năm, tự học. Các ví dụ làm việc rất đa dạng, từ đơn giản đến phức tạp, mà không yêu cầu bạn phải mua phần mềm phức tạp và đắt tiền. Cuốn sách sẽ cho phép chủ sở hữu bất kỳ doanh nghiệp nào trong ngành xây dựng có cái nhìn mới về chiến lược kinh doanh, số hóa và triển vọng phát triển. Và đối với các công ty nhỏ hơn, tăng cường hiệu quả với các công cụ giá cả phải chăng và miễn phí.

— Mikhail Kosarev, Giảng viên và Tư vấn về Chuyển đổi Số trong Ngành Xây dựng tại TIM-ASG.

Cuốn sách "DATA DRIVEN CONSTRUCTION" là một bước ngoặt cho bất cứ ai tò mò về hướng đi của ngành xây dựng trong thời đại dữ liệu. Artem không chỉ chạm đến bề mặt; anh ấy đi sâu vào những phát triển hiện tại, thách thức và cơ hội đầy hứa hẹn trong xây dựng. Điều làm nên sự khác biệt của cuốn sách này là tính khả dụng — Artem giải thích các ý tưởng phức tạp bằng các phép ẩn dụ dễ hiểu, giúp nội dung dễ tiếp cận. Tôi thấy cuốn sách thực sự bổ ích nhưng cũng rất thú vị. Tóm lại, Artem đã tạo ra một tài nguyên quý giá không chỉ thông tin mà còn truyền cảm hứng. Dù bạn là một chuyên gia dày dạn hay một người mới gia nhập ngành xây dựng, cuốn sách này sẽ mở rộng quan điểm của bạn và làm sâu sắc thêm hiểu biết của bạn về hướng đi của ngành. Rất đáng được giới thiệu!

— Moayad Saleh, Kiến trúc sư và Quản lý Thực hiện BIM tại TMM GROUP Gesamtplanungs GmbH.

Tôi phải nói rằng Data-Driven Construction xứng đáng được giảng dạy như một sách giáo khoa ở các trường đại học và là cuốn sách sẽ có những đóng góp quý giá cho lĩnh vực BIM đang phát triển. Cuốn sách Data-Driven Construction bao gồm một từ điển kỹ thuật giải thích các khái niệm rất tốt. Các chủ đề rất khó giải thích được trình bày một cách đơn giản và dễ hiểu với ngôn ngữ hình ảnh rất đẹp. Tôi nghĩ rằng những gì được dự kiến mô tả trong hình ảnh nên được truyền đạt cho người đọc, dù chỉ ngắn gọn. Sự dễ hiểu của một số hình ảnh, nói cách khác, việc đọc hình ảnh yêu cầu thông tin bổ sung. Tôi cũng muốn nói rằng tôi rất vui được giới thiệu tác phẩm quý giá của Artem Boiko trong các bài giảng và hội thảo của tôi tại các trường đại học.

— Dr. Ediz Yazicioglu, Chủ sở hữu tại ArchCube và Giảng viên quản lý dự án xây dựng tại Khoa Kiến trúc của Đại học Kỹ thuật Istanbul và Đại học Medipol.

"Data Driven Construction" truyền đạt sinh động những nguyên tắc cơ bản của công việc dựa trên thông tin với dữ liệu xây dựng. Một cuốn sách xử lý các luồng thông tin và các khái niệm kinh tế căn bản, do đó tách biệt nó với các sách BIM khác, vì nó không chỉ đại diện cho góc nhìn của một nhà sản xuất phần mềm, mà còn cố gắng truyền đạt các khái niệm cơ bản. Đây là một cuốn sách đáng để đọc và xem.

— Jakob Hirn, CEO và Đồng sáng lập Build Informed GmbH, và Người khởi xướng Diễn đàn Đổi mới "On Top With BIM".

"Dữ liệu là dầu mỏ mới", như họ thường nói, vì vậy những người khai thác hoặc thợ mỏ nên có những công cụ và tư duy đúng để trích xuất giá trị từ nguồn tài nguyên thế kỷ 21 này. Ngành xây dựng đã quá lâu mắc kẹt trên con dốc trơn trượt của các quy trình dựa trên thông tin "3D", trong đó việc giao hàng dự án dựa vào thông tin đã được xử lý của người khác (ví dụ: họ đã vạch ra biểu đồ hình tròn hoặc thanh) trong khi "dữ liệu" cơ bản (ví dụ: bảng tính thô) có khả năng cung cấp nhiều hơn thế, đặc biệt là vì sự kết hợp dữ liệu đa dạng và AI mang lại tiềm năng vô hạn. Nếu bạn đang cung cấp (hoặc giảng dạy/nghiên cứu) xây dựng, cuốn sách này là tài nguyên tốt nhất - và cho đến nay là duy nhất - cho việc điều hướng thế giới dựa trên dữ liệu mà chúng ta đang đối mặt.

— Dr Zulfikar Adamu, Phó Giáo sư Công nghệ thông tin chiến lược trong xây dựng tại LSBU, Vương quốc Anh.

"Data-Driven Construction" của Artem Boiko là một tác phẩm ấn tượng cung cấp nền tảng vững chắc cho ngành xây dựng trong thời gian công nghệ và khả năng thông tin liên tục phát triển. Boiko quản lý để trình bày các chủ đề phức tạp một cách dễ hiểu đồng thời giới thiệu những ý tưởng tiên phong. Cuốn sách là một tuyển tập được suy nghĩ kỹ lưỡng không chỉ làm nổi bật các phát triển hiện tại mà còn cung cấp cái nhìn về những đổi mới trong tương lai. Nó được khuyến nghị cao cho bất cứ ai muốn nắm bắt kế hoạch và thực hiện xây dựng dựa trên dữ liệu.

- Markus Eiberger, Giảng viên tại Đại học Khoa học ứng dụng Stuttgart, Quản lý Dự án cấp cao và Phó Giám đốc Chi nhánh tại Konstruktionsgruppe Bauen, Thành viên Hội đồng của Hiệp hội BIM Cluster Baden-Württemberg.

Ấn phẩm thứ hai, tháng 3 năm 2025. © 2025 | Artem Boiko | Karlsruhe

ISBN 978-3-912002-10-2



Artem Boiko Bản quyền

boikoartem@gmail.com
info@datadrivenconstruction.io

Không phần nào của cuốn sách này có thể được sao chép hoặc truyền đạt dưới bất kỳ hình thức nào và bằng bất kỳ phương tiện nào, điện tử hoặc cơ học, bao gồm sao chụp, ghi âm hoặc bất kỳ hệ thống lưu trữ và tìm kiếm thông tin nào, mà không có sự cho phép bằng văn bản của tác giả - ngoại trừ việc phân phối phi thương mại trong tình trạng không thay đổi. Cuốn sách được phát hành miễn phí và có thể được chuyển giao tự do cho những người dùng khác với mục đích cá nhân, giáo dục hoặc nghiên cứu, với điều kiện duy trì quyền tác giả và liên kết đến bản gốc. Tác giả giữ tất cả các quyền không tài sản đối với văn bản và không cung cấp bất kỳ bảo đảm rõ ràng hoặc ngụ ý nào. Các công ty, sản phẩm và tên gọi được đề cập trong cuốn sách có thể là hư cấu hoặc được sử dụng trong các ví dụ. Tác giả không chịu trách nhiệm cho bất kỳ hậu quả nào liên quan đến việc sử dụng thông tin được cung cấp. Thông tin trong cuốn sách được trình bày "như có" mà không có bảo đảm về độ đầy đủ hoặc tính cập nhật. Tác giả không chịu trách nhiệm cho bất kỳ thiệt hại ngẫu nhiên hoặc gián tiếp nào liên quan đến việc sử dụng thông tin, mã hoặc chương trình có trong cuốn sách này. Các ví dụ mã trong cuốn sách chỉ nhằm mục đích giáo dục. Người đọc sử dụng chúng với rủi ro của riêng mình. Tác giả khuyến nghị người đọc kiểm tra tất cả các giải pháp phần mềm trước khi áp dụng vào môi trường sản xuất. Tất cả các thương hiệu và tên sản phẩm được đề cập trong văn bản là thương hiệu đã đăng ký hoặc dấu hiệu dịch vụ của các công ty tương ứng và là tài sản của những người sở hữu tương ứng. Việc sử dụng các tên này trong cuốn sách không có nghĩa là có bất kỳ quan hệ nào với các chủ sở hữu của chúng hoặc sự chấp thuận từ các chủ sở hữu đó. Việc đề cập đến sản phẩm hoặc dịch vụ của bên thứ ba không phải là một khuyến nghị và không ngụ ý sự hỗ trợ từ chúng. Tên công ty và sản phẩm được sử dụng trong ví dụ có thể là thương hiệu của các chủ sở hữu. Các liên kết đến các trang web của bên thứ ba được cung cấp vì mục đích thuận lợi và không có nghĩa là tác giả chứng thực thông tin được cung cấp trên những trang web đó. Tất cả thống kê, trích dẫn và nghiên cứu được đưa ra đều còn tại thời điểm viết sách. Dữ liệu có thể thay đổi theo thời gian.

Cuốn sách này được phát hành theo giấy phép Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Quốc tế (CC BY-NC-ND 4.0). Bạn có thể sao chép và phân phối nó cho các mục đích phi thương mại, với điều kiện giữ nguyên quyền tác giả và không thay đổi.

CC BY-NC-ND 4.0

© 2024 Artem Boiko. Ấn bản đầu tiên. © 2025 Artem Boiko. Ấn bản thứ hai, chỉnh sửa và bổ sung. Tất cả quyền được bảo lưu.

LỜI MỞ ĐẦU CHO ẤN BẢN THỨ HAI

Cuốn sách này là kết quả của một cuộc đối thoại sống động với cộng đồng chuyên nghiệp. Nó dựa trên nhiều thảo luận chuyên nghiệp về công việc với dữ liệu trong ngành xây dựng, diễn ra trên các nền tảng chuyên nghiệp và mạng xã hội khác nhau. Những cuộc thảo luận này đã trở thành cơ sở cho các bài viết, xuất bản phẩm và tài liệu trực quan tạo ra phản hồi mạnh mẽ trong cộng đồng nghề nghiệp. Tài liệu của tác giả hàng năm thu hút hàng triệu lượt xem trên các nền tảng và ngôn ngữ khác nhau, kết nối các chuyên gia trong lĩnh vực số hóa xây dựng.

Trong vòng một năm kể từ khi phát hành phiên bản đầu tiên, cuốn sách đã được đặt hàng bởi các chuyên gia từ hơn 50 quốc gia — từ Brazil và Peru đến Mauritius và Nhật Bản. Phiên bản thứ hai của cuốn sách, mà bạn đang cầm trên tay, đã được chỉnh sửa và bổ sung dựa trên phản hồi của các chuyên gia, các nhận xét phê bình về phiên bản đầu tiên và các cuộc thảo luận trong giới chuyên môn. Nhờ có những phản hồi, phiên bản thứ hai đã được mở rộng đáng kể: bổ sung các chương mới về công nghệ CAD (BIM) và xây dựng quy trình ETL hiệu quả. Số lượng ví dụ thực tiễn và các trường hợp nghiên cứu cũng được tăng cường đáng kể. Đặc biệt giá trị là phản hồi từ các nhà lãnh đạo trong ngành xây dựng, các công ty tư vấn và các công ty công nghệ thông tin lớn nhất, những người đã liên hệ với tác giả với các câu hỏi về số hóa và khả năng tương tác cả trước và sau khi phát hành phiên bản đầu tiên của cuốn sách. Nhiều người trong số họ đã áp dụng các phương pháp mô tả trong cuốn sách hoặc có kế hoạch thực hiện điều đó trong thời gian gần.

Bạn đang cầm trên tay cuốn sách được tạo ra nhờ vào các cuộc thảo luận và trao đổi ý kiến tích cực. Tiến bộ được sinh ra trong đối thoại, từ sự va chạm của các quan điểm và tính cởi mở với các phương pháp mới. Cảm ơn bạn đã trở thành một phần của cuộc đối thoại này. Phê bình mang tính xây dựng của bạn là cơ sở cho những cải tiến trong tương lai. Nếu bạn phát hiện ra bất kỳ lỗi nào trong văn bản hoặc có mong muốn chia sẻ ý tưởng và đề xuất, mọi phản hồi đều được hoan nghênh. Thông tin liên hệ để liên lạc được cung cấp ở cuối cuốn sách.

TẠI SAO CUỐN SÁCH MIỄN PHÍ?

Cuốn sách này được thiết kế như một tài nguyên giáo dục mở, nhằm mục đích truyền bá các phương pháp hiện đại trong quản lý dữ liệu trong ngành xây dựng. Phiên bản đầu tiên của cuốn sách đã làm cơ sở để thu thập ý kiến và đề xuất từ cộng đồng chuyên gia, cho phép cải thiện cấu trúc và nội dung của tài liệu. Tất cả các ý kiến, đề xuất và ý tưởng đã được phân tích cẩn thận và đưa vào phiên bản đã chỉnh sửa này. Mục tiêu của cuốn sách là giúp các chuyên gia trong ngành xây dựng hiểu được tầm quan trọng của việc làm việc với dữ liệu: một cách hệ thống, có ý thức và với cái nhìn dài hạn về giá trị thông tin. Tác giả đã thu thập các ví dụ, minh họa và quan sát thực tiễn trong hơn 10 năm làm việc trong lĩnh vực số hóa xây dựng. Phần lớn các tài liệu này được hình thành trong quá trình thực hiện các dự án thực tế, các cuộc thảo luận với kỹ sư và nhà phát triển, tham gia vào các sáng kiến quốc tế và tổ chức các buổi hội thảo đào tạo. Cuốn sách này là nỗ lực nhằm hệ thống hóa kinh nghiệm đã tích lũy và chia sẻ nó theo một hình thức dễ tiếp cận. Nếu bạn muốn hỗ trợ việc mở rộng hơn nữa các ý tưởng của cuốn sách và nhận được một định dạng tiện lợi để đọc, làm việc với các ví dụ và tài liệu hình ảnh — bạn có thể mua phiên bản in.

QUYỀN SỬ DỤNG TÀI LIỆU

Tất cả các tài liệu, minh họa và đoạn trích trong cuốn sách này có thể được sao chép, trích dẫn hoặc sử dụng dưới bất kỳ hình thức nào và trên bất kỳ phương tiện nào với điều kiện phải ghi nguồn: tác giả Artem Boiko và tiêu đề cuốn sách “Data-Driven Construction”. Cảm ơn bạn đã tôn trọng công sức và việc phổ biến kiến thức.

Với lòng biết ơn chân thành, tôi xin dành cuốn sách này cho gia đình mình, những người từ những năm đầu đời đã truyền cho tôi tình yêu sâu sắc với việc xây dựng, cho thành phố khai thác than nơi tôi sinh ra — vì những bài học về sức sống mạnh mẽ và cho vợ tôi, một kỹ sư địa chất, người đã luôn ủng hộ tôi và là nguồn cảm hứng liên tục trong cuộc sống.

DÀNH CHO AI CUỐN SÁCH NÀY

Được viết bằng ngôn ngữ dễ tiếp cận, cuốn sách này dành cho một lượng lớn độc giả trong ngành xây dựng - từ sinh viên và người mới bắt đầu muốn tìm hiểu những nguyên tắc cơ bản về quy trình xây dựng hiện đại, đến các chuyên gia cần phương pháp quản lý dữ liệu tiên tiến trong xây dựng. Cho dù bạn là kiến trúc sư, kỹ sư, quản đốc, quản lý xây dựng hay nhà phân tích dữ liệu, cuốn sách toàn diện này với nhiều minh họa và đồ thị độc đáo cung cấp những thông tin quý giá về cách sử dụng dữ liệu trong doanh nghiệp nhằm tối ưu hóa và tự động hóa quy trình, cải thiện quyết định và quản lý dự án xây dựng ở các cấp độ khác nhau với sự trợ giúp của những công cụ hiện đại.

Cuốn sách này là một hướng dẫn toàn diện, kết hợp giữa các nguyên tắc lý thuyết và những khuyến nghị thực tiễn về việc tích hợp phương pháp quản lý dữ liệu vào quy trình xây dựng. Cuốn sách tập trung vào việc sử dụng thông tin một cách chiến lược để tối ưu hóa hoạt động kinh doanh, tự động hóa quy trình, cải thiện quyết định và quản lý dự án hiệu quả bằng các công cụ kỹ thuật số hiện đại.

Trong các trang sách này, các khía cạnh lý thuyết và thực tiễn về làm việc với thông tin trong ngành xây dựng được khám phá. Qua các ví dụ chi tiết, phương pháp định hình yêu cầu, thu thập yêu cầu, xử lý dữ liệu không cấu trúc và đa định dạng được nghiên cứu để biến chúng thành những giải pháp hiệu quả cho các công ty xây dựng.

Độc giả sẽ tuần tự trải qua hành trình từ việc xác định yêu cầu và xây dựng các mô hình dữ liệu cơ bản đến các quy trình phức tạp hơn trong việc tích hợp các nguồn thông tin khác nhau, tạo ra các quy trình ETL, xây dựng Pipe thông tin và các mô hình học máy. Cách tiếp cận tuần tự này giúp trình bày rõ ràng các cơ chế tổ chức và tự động hóa quy trình kinh doanh cũng như hệ thống hỗ trợ quyết định trong lĩnh vực xây dựng. Mỗi phần của cuốn sách sẽ kết thúc bằng một chương thực hành, chứa các hướng dẫn từng bước cho phép áp dụng ngay kiến thức đã học vào các dự án thực tế.

TÓM TẮT CÁC PHẦN CỦA CUỐN SÁCH

Cuốn sách này được xây dựng xung quanh khái niệm chuyển đổi dữ liệu trong chuỗi tạo giá trị: từ việc thu thập và đảm bảo chất lượng đến xử lý phân tích và rút ra những giải pháp thực tiễn quý giá bằng cách sử dụng các công cụ và phương pháp hiện đại.

Phần 1: Sự tiến hóa kỹ thuật số trong xây dựng - theo dõi sự chuyển mình lịch sử của quản lý dữ liệu từ bảng đất sét đến các hệ thống kỹ thuật số hiện đại, phân tích sự xuất hiện của các hệ thống mô-đun và sự tăng trưởng tầm quan trọng của việc số hóa thông tin trong bối cảnh các cuộc cách mạng công nghiệp.

Phần 2: Những thách thức thông tin trong ngành xây dựng - nghiên cứu các vấn đề phân mảnh dữ liệu, "kho thông tin", ảnh hưởng của phương pháp HiPPO đến quyết định và hạn chế của các định dạng độc quyền, đồng thời đề nghị xem xét việc chuyển đổi sang các hệ sinh thái AI và LLM.

Phần 3: Hệ thống hóa dữ liệu trong xây dựng - hình thành loại hình dữ liệu xây dựng, mô tả các phương pháp tổ chức chúng, tích hợp với các hệ thống doanh nghiệp và thảo luận về việc tạo ra các trung tâm năng lực để chuẩn hóa các quy trình thông tin.

Phần 4: Đảm bảo chất lượng dữ liệu - tiết lộ các phương pháp chuyển đổi thông tin rời rạc thành dữ liệu chất lượng cao, có cấu trúc, bao gồm việc trích xuất dữ liệu từ các nguồn khác nhau, xác thực và mô hình hóa bằng cách sử dụng LLM.

Phần 5: Tính toán chi phí và thời gian - dành cho việc số hóa các tính toán chi phí và lập kế hoạch, tự động hóa việc lấy khối lượng từ mô hình CAD- (BIM-), công nghệ mô hình 4D-8D và tính toán các chỉ số ESG của các dự án xây dựng.

Phần 6: CAD và BIM - phân tích một cách có hệ thống sự tiến hóa của công nghệ thiết kế, các vấn đề tương thích giữa các hệ thống, xu hướng chuyển đổi sang các định dạng dữ liệu mở và triển vọng ứng dụng trí tuệ nhân tạo trong thiết kế.

Phần 7: Phân tích dữ liệu và tự động hóa - xem xét các nguyên tắc trực quan hóa thông tin, các chỉ số hiệu suất chính, quy trình ETL, công cụ điều phối quy trình làm việc và ứng dụng các mô hình ngôn ngữ cho việc tự động hóa các nhiệm vụ lặp đi lặp lại.

Phần 8: Lưu trữ và quản lý dữ liệu - nghiên cứu các định dạng lưu trữ dữ liệu, các khái niệm kho dữ liệu và hồ dữ liệu, các nguyên tắc quản lý dữ liệu và các phương pháp tiếp cận mới, bao gồm cả cơ sở dữ liệu vector và các phương pháp DataOps và VectorOps.

Phần 9: Dữ liệu lớn và học máy - dành cho việc chuyển sang phân tích khách quan dựa trên dữ liệu lịch sử, Internet vạn vật trên các công trường xây dựng và áp dụng các thuật toán học máy để dự đoán chi phí và thời gian của các dự án.

Phần 10: Ngành xây dựng trong thời đại dữ liệu số - trình bày cái nhìn về tương lai của ngành xây dựng, phân tích sự chuyển đổi từ phân tích nguyên nhân - hệ quả sang làm việc với các mối tương quan, khái niệm "uber hóa" ngành xây dựng và các chiến lược chuyển đổi số.

What is meant by **data-driven construction**?

Led by CEO and project managers

A data-driven approach to construction is the **process** of implementing **connected and automated systems** that enable companies to **continuously improve** **project delivery** and operational efficiency while **reducing costs and risks**, ultimately creating a **sustainable competitive advantage.**

Focuses on real-time analytics

It's a mindset shift!

Streamline operations & enhance collaboration

The ultimate goal



GIỚI THIỆU

Công ty của bạn có thể duy trì tính cạnh tranh trong thế giới mà công nghệ phát triển nhanh chóng và mọi khía cạnh của kinh doanh, từ tính toán thời hạn và chi phí đến phân tích rủi ro, đều được tự động hóa bằng các mô hình học máy bao lâu nữa?

Ngành xây dựng, tồn tại như con người từ thuở ban đầu, đang đứng trước những thay đổi cách mạng, hứa hẹn sẽ thay đổi hoàn toàn những quan niệm truyền thống về xây dựng. Ngay bây giờ, ở các lĩnh vực khác của nền kinh tế, sự số hóa không chỉ thay đổi những quy tắc quen thuộc mà còn một cách không khoan nhượng loại bỏ các công ty không thể thích ứng với những điều kiện xử lý dữ liệu mới và không thể nâng cao tốc độ ra quyết định (Hình. 1).



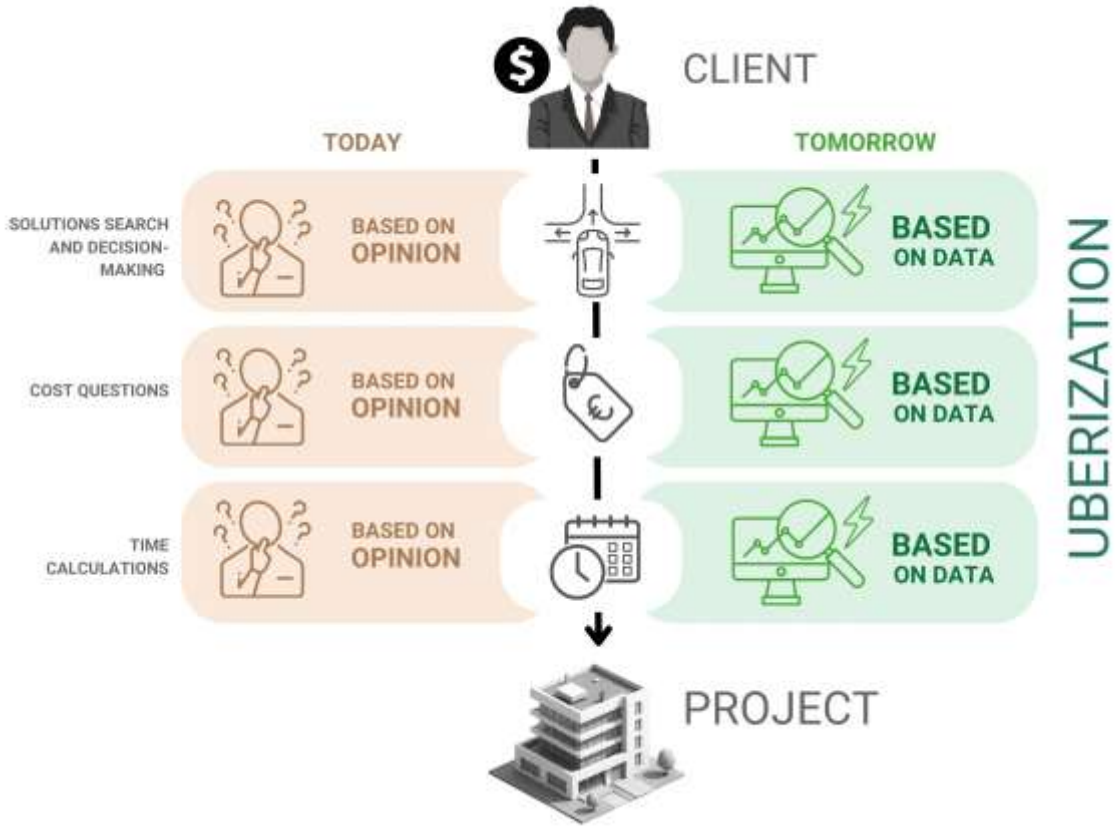
Hình. 1 Tốc độ ra quyết định trong ngành xây dựng thường phụ thuộc vào yếu tố con người nhiều hơn so với các lĩnh vực khác.

Ngành ngân hàng, bán lẻ, logistics và nông nghiệp đang chuyển động nhanh chóng tới sự số hóa hoàn toàn, nơi không còn chỗ cho các lỗi và ý kiến chủ quan. Các thuật toán hiện đại có khả năng phân tích các khối lượng dữ liệu khổng lồ và cung cấp cho khách hàng các dự đoán chính xác - bất kể đó là xác suất hoàn trả khoản vay, tuyến đường giao hàng tối ưu hay dự đoán rủi ro.

Xây dựng là một trong những lĩnh vực cuối cùng sẽ phải thực hiện sự chuyển đổi không thể tránh khỏi từ các quyết định dựa trên ý kiến của các chuyên gia được trả lương cao sang các quyết định dựa trên dữ liệu. Sự chuyển đổi này không chỉ được thúc đẩy bởi các khả năng công nghệ mới, mà còn bởi nhu cầu ngày càng tăng của thị trường và khách hàng về tính minh bạch, độ chính xác và tốc độ.

Tự động hóa, robot hóa quy trình, dữ liệu mở và dự đoán dựa trên chúng - tất cả không còn chỉ là những khả năng, mà đã trở thành sự cần thiết. Hầu hết các công ty trong ngành xây dựng, những công ty trước đây còn chịu trách nhiệm trước khách hàng về tính toán khối lượng, chi phí, thời gian dự án và kiểm soát chất lượng, giờ đây có nguy cơ trở thành những người thực hiện đơn thuần mà không đưa ra các quyết định then chốt.

Với sự phát triển của sức mạnh tính toán, các thuật toán học máy và việc dân chủ hóa quyền truy cập dữ liệu, việc tự động hợp nhất dữ liệu từ nhiều nguồn khác nhau trở nên khả thi, cho phép phân tích sâu hơn về các quá trình, dự đoán rủi ro và tối ưu hóa chi phí ngay từ giai đoạn thảo luận dự án xây dựng. Những công nghệ này tạo ra tiềm năng cho việc nâng cao hiệu quả và giảm chi phí một cách triệt để trong toàn bộ lĩnh vực.



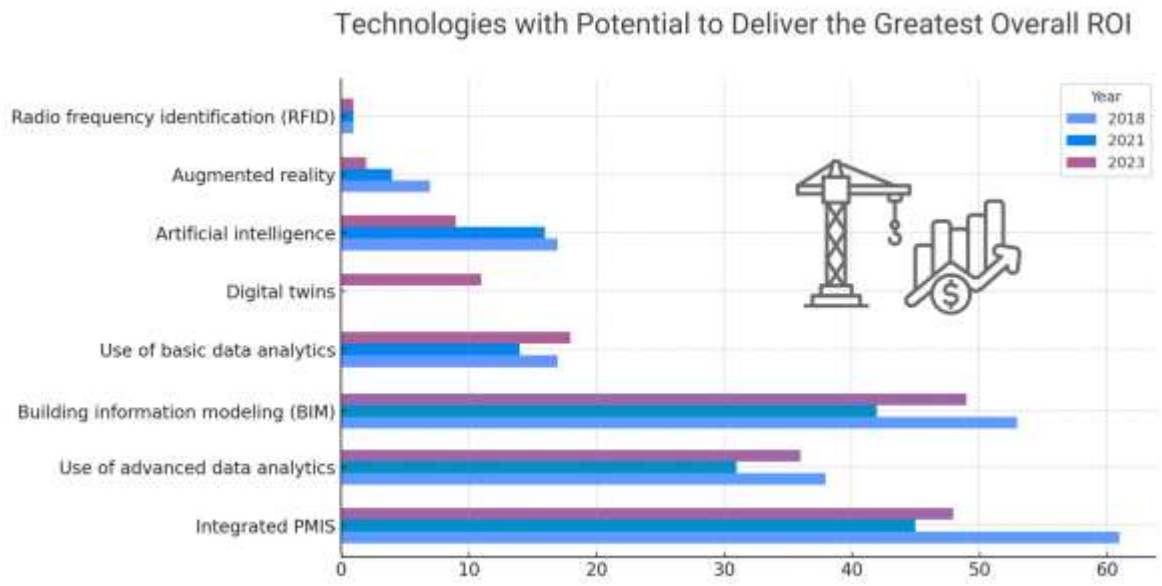
Khách hàng không quan tâm đến yếu tố con người dư thừa trong lộ trình thực hiện dự án của họ.

Mặc dù có nhiều lợi ích từ các công cụ và khái niệm mới, ngành xây dựng vẫn tụt hậu khá nhiều so với các lĩnh vực khác trong nền kinh tế về việc áp dụng công nghệ mới.

> Theo báo cáo "IT Metrics Key Data 2017", ngành xây dựng đứng cuối cùng về chi tiêu cho công nghệ thông tin trong số 19 lĩnh vực kinh tế khác.

Sự gia tăng nhanh chóng về khối lượng dữ liệu và độ phức tạp của các quy trình trở thành nỗi đau đầu cho ban lãnh đạo các công ty, và vấn đề chính trong việc sử dụng công nghệ mới là dữ liệu, mặc dù phong phú, vẫn còn phân tán, không có cấu trúc và thường không tương thích giữa các hệ thống và sản phẩm phần mềm khác nhau. Do đó, nhiều công ty trong ngành xây dựng hôm nay chủ yếu quan tâm đến các vấn đề chất lượng dữ liệu, mà chỉ có thể được giải quyết thông qua việc áp dụng các hệ thống quản lý và phân tích tự động hóa hiệu quả.

> Theo một cuộc khảo sát do KPMG® thực hiện trong số các nhà quản lý công ty xây dựng vào năm 2023, các hệ thống thông tin quản lý dự án (PMIS), phân tích dữ liệu tiên tiến và cơ bản, cùng với mô hình thông tin xây dựng (BIM) là những công nghệ có tiềm năng cao nhất để cải thiện lợi nhuận đầu tư trong các dự án.



Khảo sát giữa các nhà lãnh đạo công ty xây dựng: những công nghệ nào sẽ mang lại tỷ lệ hoàn vốn đầu tư cao nhất trong các dự án vốn?

Giải quyết các vấn đề liên quan đến tích hợp dữ liệu vào quy trình kinh doanh nằm ở việc đảm bảo chất lượng thông tin cao, sử dụng các định dạng dữ liệu phù hợp và áp dụng các phương pháp hiệu quả trong việc tạo ra, lưu trữ, phân tích và xử lý dữ liệu.

Sự nhận thức về giá trị của dữ liệu buộc các ngành khác nhau từ bỏ các ứng dụng rời rạc và các cấu trúc quản lý hành chính phức tạp. Thay vào đó, trọng tâm chuyển sang việc tạo ra các phương pháp tiếp cận mới đối với kiến trúc thông tin, biến các công ty thành những doanh nghiệp hiện đại, được điều hành bởi dữ liệu. Sớm hay muộn, lĩnh vực xây dựng cũng sẽ thực hiện bước chuyển này, từ sự tiến hóa số dần dần đến một cuộc cách mạng số thực sự, ảnh hưởng đến tất cả các công ty.

> Việc chuyển đổi sang các quy trình kinh doanh dựa trên dữ liệu sẽ không dễ dàng. Nhiều công ty sẽ gặp khó khăn, vì các nhà lãnh đạo không phải lúc nào cũng hiểu cách sử dụng các khối dữ liệu hỗn độn để nâng cao hiệu quả và sự phát triển của doanh nghiệp.

Trong cuốn sách này, chúng ta sẽ đắm chìm trong thế giới dữ liệu, nơi thông tin trở thành tài nguyên chiến lược chủ chốt, xác định hiệu quả và sự bền vững của các quy trình kinh doanh. Trong bối cảnh khối lượng thông tin tăng nhanh, các công ty gặp phải những thách thức mới. Chuyển đổi số không còn là một thuật ngữ thời thượng — nó trở thành một nhu cầu thiết yếu.



Hình 4 Dữ liệu và quy trình là nền tảng của xây dựng.

Hiểu được sự chuyển mình có nghĩa là có thể giải thích điều phức tạp bằng những từ đơn giản. Đó chính là lý do cuốn sách này được viết bằng ngôn ngữ dễ tiếp cận và kèm theo các hình minh họa do tác giả tạo ra, được thiết kế đặc biệt để giải thích trực quan các khái niệm chính. Những sơ đồ, đồ thị và hình ảnh trực quan này được nhằm xóa bỏ các rào cản nhận thức và làm cho tài liệu dễ hiểu ngay cả đối với những người trước đây cho rằng các chủ đề này quá phức tạp. Tất cả các hình minh họa, sơ đồ và đồ thị trong cuốn sách này đều được tác giả tạo ra và thiết kế đặc biệt để trực quan hóa các khái niệm chính được mô tả trong văn bản.

> *Một bức tranh đáng giá ngàn lời.*
>
> *– Fred R. Barnard, họa sĩ minh họa người Anh, 1927*

Để liên kết lý thuyết với thực tiễn, chúng ta sẽ sử dụng các công cụ trí tuệ nhân tạo (cụ thể là các mô hình ngôn ngữ), cho phép phát triển giải pháp mà không cần kiến thức sâu về lập trình. Nếu bạn chú trọng vào tài liệu thực tiễn và quan tâm đến công việc thực tiễn với dữ liệu hơn, bạn có thể bỏ qua phần giới thiệu đầu tiên và chuyển ngay đến phần thứ hai của cuốn sách, nơi bắt đầu mô tả các ví dụ và trường hợp cụ thể.

Tuy nhiên, không nên đặt kỳ vọng quá mức vào trí tuệ nhân tạo (AI), học máy và các công cụ mô hình ngôn ngữ lớn (LLM) nói chung. Nếu không có dữ liệu đầu vào chất lượng và hiểu biết sâu sắc về lĩnh vực cụ thể, ngay cả các thuật toán tiên tiến nhất cũng không thể cung cấp kết quả đáng tin cậy và có ý nghĩa.

> Giám đốc điều hành của Microsoft, Satya Nadella, vào đầu năm 2025 đã cảnh báo về nguy cơ xuất hiện bong bóng trong lĩnh vực trí tuệ nhân tạo, so sánh cơn sốt hiện tại với "bong bóng dotcom". Ông nhấn mạnh rằng các tuyên bố về việc đạt được các giai đoạn trí tuệ nhân tạo tổng quát mà không có cơ sở tương ứng là "sự thao túng chỉ số vô nghĩa". Nadella cho rằng thành công thực sự của AI phải được đo lường qua sự đóng góp của nó vào tăng trưởng GDP toàn cầu, thay vì sự chú ý quá mức vào những tuyên bố tai tiếng.

Dưới tất cả những lời nói rầm rộ về công nghệ mới và các khái niệm, là một công việc phức tạp và tỉ mỉ để

đảm bảo chất lượng dữ liệu, chuẩn hóa quy trình kinh doanh và điều chỉnh các công cụ theo các nhiệm vụ thực tế.

> Phương pháp dựa trên dữ liệu không phải là một sản phẩm có thể tải xuống hoặc mua được. Đây là một chiến lược cần được xây dựng. Nó bắt đầu bằng một cái nhìn mới về các quy trình và vấn đề hiện có, sau đó yêu cầu phải có sự kiên trì trong việc tiến về phía đã chọn.

Các nhà phát triển phần mềm hàng đầu và các nhà cung cấp ứng dụng sẽ không trở thành động lực cho những thay đổi trong ngành xây dựng, vì đối với nhiều người trong số họ, phương pháp dựa trên dữ liệu là một mối đe dọa đối với mô hình kinh doanh đã hình thành.

> *Các ngành khác [khác với ngành xây dựng], chẳng hạn như ngành ô tô, đã trải qua những thay đổi triệt để và tàn phá, và quá trình chuyển đổi kỹ thuật số của họ đang diễn ra một cách mạnh mẽ. Các công ty xây dựng cần phải hành động nhanh chóng và quyết đoán: những công ty linh hoạt sẽ thu được lợi ích lớn, trong khi những công ty chần chừ sẽ phải đối mặt với những rủi ro nghiêm trọng. Hãy nhớ đến sự chấn động mà nhiếp ảnh kỹ thuật số đã gây ra trong ngành công nghiệp này.Báo cáo của Diễn đàn Kinh tế Thế giới "Hình thành tương lai ngành xây dựng", 2016*

Các công ty nhận thức kịp thời về cơ hội và lợi ích của cách tiếp cận mới sẽ có thể đạt được lợi thế cạnh tranh bền vững và có thể phát triển, mở rộng mà không phụ thuộc vào quyết định của các nhà cung cấp lớn.

Đây là cơ hội của bạn không chỉ để vượt qua cơn bão số hóa thông tin sắp tới, mà còn để kiểm soát nó. Trong cuốn sách, bạn sẽ tìm thấy không chỉ là phân tích tình trạng hiện tại của ngành, mà còn là các khuyến nghị cụ thể về việc tái tư duy và tái cấu trúc quy trình cũng như doanh nghiệp của bạn để trở thành người dẫn đầu trong kỷ nguyên xây dựng mới và nâng cao kỹ năng chuyên môn của bạn.

Tương lai số của xây dựng không chỉ là việc sử dụng công nghệ và phần mềm mới, mà còn là một sự tái đánh giá cơ bản cách thức làm việc với dữ liệu và các mô hình kinh doanh.

Công ty của bạn có sẵn sàng cho những thay đổi chiến lược này không?

MỤC LỤC

GIỚI THIỆU **1**

MỤC LỤC ..... **I**

I PHẦN TỪ BẢNG ĐẤT SÉT ĐẾN CÁCH MẠNG SỐ: CÁCH THÔNG TIN ĐÃ TIẾN HÓA TRONG NGÀNH XÂY DỰNG. **2**

CHƯƠNG 1.1. SỰ TIẾN HÓA CỦA VIỆC SỬ DỤNG DỮ LIỆU TRONG NGÀNH XÂY DỰNG 3

Sự ra đời của kỷ nguyên dữ liệu trong xây dựng 3

Từ đất sét và giấy cói đến công nghệ số 4

Quy trình như một công cụ của trải nghiệm được điều khiển bởi dữ liệu. 5

Số hóa thông tin trong quy trình xây dựng 7

CHƯƠNG 1.2. CÔNG NGHỆ VÀ HỆ THỐNG QUẢN LÝ TRONG XÂY DỰNG HIỆN ĐẠI 11

Cách mạng số và sự xuất hiện của các hệ thống MRP/ERP mô-đun 11

Hệ thống quản lý dữ liệu: từ khai thác dữ liệu đến các bài toán kinh doanh. 13

Mẫu nấm doanh nghiệp: cách mà dữ liệu kết nối trong quy trình kinh doanh. 16

CHƯƠNG 1.3. CÁCH MẠNG SỐ VÀ BÙNG NỔ DỮ LIỆU 20

Bắt đầu cuộc cách mạng khối lượng dữ liệu như một sóng tiến hóa 20

Khối lượng dữ liệu được tạo ra trong một công ty hiện đại 22

Chi phí lưu trữ dữ liệu: khía cạnh kinh tế 23

Ranh giới của việc tích lũy dữ liệu: từ khối lượng đến ý nghĩa 25

Các bước tiếp theo: từ lý thuyết dữ liệu đến thay đổi thực tiễn 27

II PHẦN NGÀNH XÂY DỰNG VÀ SỰ LÀM CÔN KHI DỮ LIỆU MÃNG **28**

CHƯƠNG 2.1. PHÂN MẢNG VÀ SILO DỮ LIỆU 29

Càng nhiều công cụ, doanh nghiệp càng hiệu quả? 29

Silos dữ liệu và ảnh hưởng của chúng đến hiệu quả công ty 31

Trùng lặp và thiếu chất lượng dữ liệu như hệ quả của sự phân mảnh 34

HiPPO hay nguy cơ ý kiến trong việc ra quyết định 36

Sự gia tăng liên tục về độ phức tạp và tính động của quy trình kinh doanh 39

Cách mạng công nghiệp lần thứ tư (Ngành 4.0) và lần thứ năm (Ngành 5.0) trong xây dựng 42

CHƯƠNG 2.2. BIẾN HỖN LOẠN THÀNH TRẬT TỰ VÀ GIẢM ĐỘ PHỨC TẠP 45

Mã thừa và hệ thống khép kín như rào cản nâng cao năng suất 45

Từ silos đến kho dữ liệu thống nhất 47

Hệ thống lưu trữ tích hợp cho phép chuyển sang sử dụng các tác nhân AI 48
Từ việc thu thập dữ liệu đến ra quyết định: con đường đến tự động hóa 51
Các bước tiếp theo: chuyển đổi hỗn độn thành hệ thống có thể quản lý 53
III PHẦN KHUNG DỮ LIỆU TRONG CÁC QUY TRÌNH KINH DOANH XÂY DỰNG 55
CHƯƠNG 3.1. CÁC LOẠI DỮ LIỆU TRONG NGÀNH XÂY DỰNG 56
Các loại dữ liệu quan trọng nhất trong ngành xây dựng 56
Dữ liệu cấu trúc 60
Cơ sở dữ liệu quan hệ RDBMS và ngôn ngữ truy vấn SQL 61
Các truy vấn SQL trong cơ sở dữ liệu và xu hướng mới 64
Dữ liệu phi cấu trúc 66
Dữ liệu văn bản: giữa hỗn loạn phi cấu trúc và cấu trúc 67
Dữ liệu bán cấu trúc và dữ liệu yếu cấu trúc 68
Dữ liệu hình học và ứng dụng của chúng 69
Dữ liệu CAD: từ thiết kế đến lưu trữ dữ liệu 72
Sự xuất hiện của khái niệm BIM (BOM) và việc sử dụng CAD trong các quy trình 75
CHƯƠNG 3.2. ĐỒNG NHẤT VÀ CẤU TRÚC DỮ LIỆU 81
Điền dữ liệu vào hệ thống trong ngành xây dựng 81
Chuyển đổi dữ liệu: nền tảng quan trọng của phân tích kinh doanh hiện đại 83
Mô hình dữ liệu: mối quan hệ trong dữ liệu và các kết nối giữa các yếu tố 86
Các định dạng sở hữu và ảnh hưởng của chúng đến các quá trình số hóa 91
Các định dạng mở thay đổi cách tiếp cận đến số hóa 94
Thay đổi mô hình: Mã nguồn mở như là sự kết thúc của kỷ nguyên thống trị của các nhà cung cấp phần mềm. 96
Dữ liệu mở được cấu trúc: nền tảng của chuyển đổi số 98
CHƯƠNG 3.3. MÔ HÌNH NGÔN NGỮ LỚN (LLM) VÀ VAI TRÒ CỦA CHÚNG TRONG XỬ LÝ DỮ LIỆU VÀ QUY TRÌNH KINH DOANH 101
Các cuộc trò chuyện LLM: ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok cho tự động hóa quy trình xử lý dữ liệu 101
Các mô hình ngôn ngữ lớn LLM: cách hoạt động của chúng 102
Sử dụng LLM cục bộ cho dữ liệu nhạy cảm của công ty 105
Kiểm soát hoàn toàn về AI trong công ty và cách triển khai LLM riêng 107
RAG: Trợ lý LLM thông minh với quyền truy cập vào dữ liệu doanh nghiệp 109
CHƯƠNG 3.4. IDE HỖ TRỢ LLM VÀ NHỮNG THAY ĐỔI TRONG LẬP TRÌNH TRONG TƯƠNG LAI 111
Lựa chọn IDE: từ các thử nghiệm LLM đến các giải pháp kinh doanh 111

IDE hỗ trợ LLM và các thay đổi trong lập trình trong tương lai 113
Python Pandas: công cụ thiết yếu cho công việc với dữ liệu 114
DataFrame: định dạng dữ liệu bảng phổ quát 118
Các bước tiếp theo: tạo ra khung dữ liệu bền vững 121
IV PHẦN CHẤT LƯỢNG DỮ LIỆU: TỔ CHỨC, CẤU TRÚC, MÔ HÌNH HÓA 123
CHƯƠNG 4.1. CHUYỂN ĐỔI DỮ LIỆU THÀNH HÌNH THỨC CẤU TRÚC 124
Học cách chuyển đổi tài liệu, PDF, hình ảnh và văn bản thành các định dạng cấu trúc 124
Ví dụ về việc chuyển đổi tài liệu PDF thành bảng 125
Chuyển đổi hình ảnh JPEG, PNG thành hình thức cấu trúc 129
Chuyển đổi dữ liệu văn bản thành hình thức cấu trúc 131
Chuyển đổi dữ liệu CAD (BIM) thành hình thức cấu trúc 134
Các nhà cung cấp giải pháp CAD chuyển sang dữ liệu có cấu trúc 139
CHƯƠNG 4.2. PHÂN LOẠI VÀ TÍCH HỢP: NGÔN NGỮ THỐNG NHẤT CHO DỮ LIỆU XÂY DỰNG 142
Tốc độ ra quyết định phụ thuộc vào chất lượng dữ liệu 142
Chuẩn hóa và tích hợp dữ liệu 143
Tính tương thích số bắt đầu từ yêu cầu 145
Ngôn ngữ thống nhất xây dựng: vai trò của các phân loại trong chuyển đổi số 148
Masterformat, OmniClass, Uniclass và CoClass: sự tiến hóa của hệ thống phân loại 150
CHƯƠNG 4.3. MÔ HÌNH HÓA DỮ LIỆU VÀ TRUNG TÂM CỦA KINH NGHIỆM HÀNG ĐẦU 155
Mô hình hóa dữ liệu: mô hình khái niệm, mô hình logic và mô hình vật lý 155
Mô hình hóa dữ liệu thực tế trong bối cảnh xây dựng 158
Tạo cơ sở dữ liệu bằng LLM 161
Trung tâm kinh nghiệm xuất sắc (CoE) về mô hình hóa dữ liệu 163
CHƯƠNG 4.4. HỆ THỐNG CÁC YÊU CẦU VÀ XÁC MINH THÔNG TIN 166
Thu thập và phân tích yêu cầu: chuyển đổi giao tiếp thành dữ liệu có cấu trúc 166
Sơ đồ quy trình và hiệu quả của các sơ đồ khái niệm 170
Yêu cầu có cấu trúc và biểu thức chính quy RegEx 172
Thu thập dữ liệu cho quy trình kiểm tra 177
Kiểm tra dữ liệu và kết quả kiểm tra 179
Hình dung kết quả kiểm tra 184
So sánh kiểm tra chất lượng dữ liệu với nhu cầu thực tế của con người 186
Các bước tiếp theo: biến dữ liệu thành tính toán và kế hoạch chính xác 188
V PHẦN TÍNH TOÁN CHI PHÍ VÀ THỜI GIAN: TRIỂN KHAI DỮ LIỆU TRONG QUY TRÌNH XÂY DỰNG 190

CHƯƠNG 5.1. TÍNH TOÁN CHI PHÍ VÀ DỰ TOÁN DỰ ÁN XÂY DỰNG 191
Cơ bản về xây dựng: đánh giá số lượng, chi phí và thời gian 191
Phương pháp tính toán dự toán chi phí dự án 192
Phương pháp tài nguyên trong lập dự toán và tính toán trong xây dựng 193
Cơ sở dữ liệu tài nguyên xây dựng: danh mục vật liệu và công việc xây dựng 193
Lập dự toán và tính toán chi phí công việc dựa trên cơ sở dữ liệu tài nguyên 195
Tính toán tổng chi phí dự án: từ dự toán đến ngân sách 200
CHƯƠNG 5.2. QUANTITY TAKE-OFF VÀ TỰ ĐỘNG HÓA LẬP DỰ TOÁN VÀ KẾ HOẠCH THỜI GIAN 204
Chuyển từ 3D sang 4D và 5D: sử dụng các tham số khối lượng và số lượng 204
Các thuộc tính 5D và thu thập khối lượng thuộc tính từ CAD 204
QTO Quantity Take-Off: nhóm dữ liệu dự án theo thuộc tính 208
Tự động hóa QTO bằng cách sử dụng LLM và dữ liệu có cấu trúc 213
Tính toán QTO cho toàn bộ dự án bằng cách sử dụng quy tắc từ bảng Excel 216
CHƯƠNG 5.3. 4D, 6D-8D VÀ TÍNH TOÁN KHÍ THẢI CO_2 222
Mô hình 4D: tích hợp thời gian vào dự toán xây dựng 222
Lịch trình xây dựng và tự động hóa dựa trên dữ liệu tính toán 223
Các lớp thuộc tính mở rộng 6D-8D: từ hiệu quả năng lượng đến đảm bảo an toàn 225
Đánh giá CO_2 và tính toán khí thải carbon dioxide trong các dự án xây dựng 228
CHƯƠNG 5.4. HỆ THỐNG ERP VÀ PMIS TRONG XÂY DỰNG 232
Hệ thống ERP xây dựng dựa trên các tính toán và dự toán 232
PMIS: Liên kết giữa ERP và công trường xây dựng 237
Spéculation, lợi nhuận, tính kín đáo và thiếu tính minh bạch trong ERP và PMIS 238
Kết thúc kỷ nguyên ERP/PMIS khép kín: ngành xây dựng cần những cách tiếp cận mới 240
Các bước tiếp theo: sử dụng hiệu quả dữ liệu dự án 242
VI PHẦN CAD VÀ BIM: TIẾP THỊ, THỰC TẾ VÀ TƯƠI LAI CỦA DỮ LIỆU DỰ ÁN TRONG XÂY DỰNG ... 245
CHƯƠNG 6.1. SỰ XA XUẤT CỦA CÁC KHÁI NIỆM BIM TRONG NGÀNH XÂY DỰNG 246
Lịch sử xuất hiện của BIM và open BIM như là các khái niệm tiếp thị từ các nhà cung cấp CAD 246
Thực tế của BIM: thay vì cơ sở dữ liệu tích hợp - các hệ thống mô-đun kín 249
Sự xuất hiện của định dạng mở IFC trong ngành xây dựng 251
Vấn đề của định dạng IFC phụ thuộc vào lõi hình học 253
Sự xuất hiện của chủ đề ngữ nghĩa và siêu hình trong xây dựng 255
Tại sao các công nghệ ngữ nghĩa không đáp ứng được mong đợi trong xây dựng 257
CHƯƠNG 6.2. ĐỊNH DẠNG KHÓA CỦA CÁC DỰ ÁN VÀ VẤN ĐỀ TÍNH TƯƠI LIÊN TỤC 261

Dữ liệu đóng và năng suất giảm sút: ngõ cụt của ngành CAD (BIM) 261
Huyền thoại về khả năng tương tác giữa các hệ thống CAD 263
Chuyển sang USD và dữ liệu phân cấp 267
CHƯƠNG 6.3. HÌNH HỌC TRONG XÂY DỰNG: TỪ ĐƯỜNG THẲNG ĐẾN KHỐI LƯỢNG 271
Khi các đường thẳng trở thành tiền hoặc tại sao hình học lại quan trọng đối với các nhà thầu 271
Từ đường thẳng đến thể tích: cách mà diện tích và thể tích trở thành dữ liệu 271
Chuyển sang MESH, USD và đa giác: sử dụng phân mảnh cho hình học 273
LOD, LOI, LOMD - phân loại chi tiết độc đáo trong CAD (BIM) 275
Các tiêu chuẩn mới trong CAD (BIM) - AIA, BEP, IDS, LOD, COBie 278
CHƯƠNG 6.4. TÍNH TOÁN CỦA THIẾT KẾ VÀ VIỆC SỬ DỤNG LLM ĐỂ LÀM VIỆC VỚI CAD 282
Ảo tưởng về tính độc đáo của dữ liệu CAD (BIM): con đường đến phân tích và định dạng mở 282
Thiết kế thông qua tham số: tương lai của CAD và BIM 285
Sự xuất hiện của LLM trong các quy trình xử lý dữ liệu CAD dự án 288
Phân tích tự động hóa các tệp DWG với LLM và Pandas 291
Các bước tiếp theo: chuyển từ định dạng kín sang dữ liệu mở 297
VII PHẦN QUYẾT ĐỊNH DỰA TRÊN DỮ LIỆU, PHÂN TÍCH, TỰ ĐỘNG HÓA VÀ HỌC MÁY 299
CHƯƠNG 7.1. PHÂN TÍCH DỮ LIỆU VÀ QUYẾT ĐỊNH DỰA TRÊN DỮ LIỆU 300
Dữ liệu như một nguồn lực trong quyết định 300
Trực quan hóa dữ liệu: chìa khóa để hiểu và ra quyết định 303
Các chỉ số hiệu suất KPI và ROI 305
Bảng thông tin và bảng điều khiển: trực quan hóa các chỉ số để quản lý hiệu quả 307
Phân tích dữ liệu và nghệ thuật đặt câu hỏi 309
CHƯƠNG 7.2. DÒNG DỮ LIỆU KHÔNG CẦN NỖ LỰC THỦ CÔNG: TẠI SAO CẦN ETL 312
Tự động hóa ETL: giảm chi phí và tăng tốc độ làm việc với dữ liệu 312
ETL Extract: thu thập dữ liệu 315
ETL Transform: áp dụng quy tắc kiểm tra và chuyển đổi 318
ETL Load: Hình ảnh hóa kết quả dưới dạng biểu đồ và đồ thị 321
ETL Load: Tạo tài liệu PDF tự động 326
ETL Load: Tự động tạo tài liệu bằng FPDF 327
ETL Load: Lập báo cáo và tải lên các hệ thống khác 331
ETL với sự hỗ trợ của LLM: Hình ảnh hóa dữ liệu từ tài liệu PDF 332
CHƯƠNG 7.3. DÂY CHUYỀN ETL TỰ ĐỘNG (PIPELINE) 337
Pipeline: Dây chuyền ETL tự động dữ liệu 337

Quá trình kiểm tra dữ liệu của Pipeline-ETL với LLM 341
Pipeline-ETL: kiểm tra dữ liệu và thông tin các phần tử dự án trong CAD (BIM) 343
CHƯƠNG 7.4. ĐIỀU PHỐI ETL VÀ QUY TRÌNH LÀM VIỆC: GIẢI PHÁP THỰC TẾ 349
DAG và Apache Airflow: tự động hóa và điều phối quy trình làm việc 349
Apache Airflow: ứng dụng thực tiễn trong tự động hóa ETL 350
Apache NiFi cho việc định tuyến và chuyển đổi dữ liệu 354
n8n Low-Code, No-Code cho việc điều phối quy trình 355
Các bước tiếp theo: chuyển từ các thao tác thủ công sang giải pháp dựa trên phân tích 358
VIII PHẦN LƯU TRỮ VÀ QUẢN LÝ DỮ LIỆU TRONG XÂY DỰNG 360
CHƯƠNG 8.1. HẠ TẦNG DỮ LIỆU: TỪ ĐỊNH DẠNG LƯU TRỮ ĐẾN KHO LƯU TRỮ SỐ 361
Nguyên tử dữ liệu: nền tảng của quản lý thông tin hiệu quả 361
Kho lưu trữ thông tin: tệp hoặc dữ liệu 362
Lưu trữ dữ liệu lớn: phân tích các định dạng phổ biến và hiệu quả của chúng 364
Tối ưu hóa lưu trữ dữ liệu với Apache Parquet 367
DWH: Kho dữ liệu Data Warehouse 369
Data Lake - sự tiến hóa của ETL trong ELT: từ làm sạch truyền thống đến xử lý linh hoạt 370
Kiến trúc Data Lakehouse: sự hợp lực giữa kho lưu trữ và hồ dữ liệu 372
CDE, PMIS, ERP hoặc DWH và Data Lake 375
CHƯƠNG 8.2. QUẢN LÝ KHO DỮ LIỆU VÀ NGĂN NGỪA HỖN LOẠN 378
Cơ sở dữ liệu vector và Bounding Box 378
Quản lý dữ liệu (Data Governance), tối giản dữ liệu (Data Minimalism) và đầm lầy dữ liệu (Data Swamp) 381
DataOps và VectorOps: tiêu chuẩn mới trong làm việc với dữ liệu 384
Các bước tiếp theo: từ lưu trữ hỗn loạn đến các kho lưu trữ có cấu trúc 385
IX PHẦN DỮ LIỆU LỚN, HỌC MÁY VÀ DỰ ĐOÁN 387
CHƯƠNG 9.1. DỮ LIỆU LỚN VÀ PHÂN TÍCH CHÚNG 388
Dữ liệu lớn trong xây dựng: từ trực giác đến khả năng dự đoán. 388
Vấn đề về tính hợp lý của dữ liệu lớn: tương quan, thống kê và mẫu dữ liệu. 389
Dữ liệu lớn: phân tích dữ liệu bộ dữ liệu một triệu giấy phép xây dựng ở San Francisco 392
Ví dụ về dữ liệu lớn dựa trên dữ liệu CAD (BIM) 397
IoT Internet of Things và hợp đồng thông minh. 401
CHƯƠNG 9.2. HỌC MÁY VÀ DỰ ĐOÁN 405
Học máy và trí tuệ nhân tạo sẽ thay đổi cách chúng ta xây dựng. 405

Từ đánh giá chủ quan đến dự đoán thống kê 407
Titanic dataset: Xin chào thế giới trong lĩnh vực phân tích dữ liệu và dữ liệu lớn 409
Machine learning in action: from Titanic passengers to project management. 414
Dự đoán và dự báo dựa trên dữ liệu lịch sử 418
Các khái niệm chính trong học máy 420
CHƯƠNG 9.3. DỰ ĐOÁN CHI PHÍ VÀ THỜI GIAN BẰNG HỌC MÁY. 423
Ví dụ về việc sử dụng học máy để xác định chi phí và thời gian của dự án. 423
Dự đoán chi phí và thời gian của dự án bằng cách sử dụng hồi quy tuyến tính. 425
Dự đoán chi phí và thời gian dự án bằng thuật toán K-nearest neighbor (k-NN) 428
Các bước tiếp theo: từ lưu trữ đến phân tích và dự đoán. 432
X PHẦN NGÀNH XÂY DỰNG TRONG THỜI ĐẠI DỮ LIỆU SỐ. CƠ HỘI VÀ THÁCH THỨC. 435
CHƯƠNG 10.1. CHIẾN LƯỢC SINH TỒN: HÌNH THÀNH LỢI THẾ CẠNH TRANH 436
Sự tương quan thay vì tính toán: tương lai của phân tích xây dựng. 436
Cách tiếp cận dựa trên dữ liệu trong xây dựng: hạ tầng cấp độ mới 439
Văn phòng kỹ thuật số thế hệ tiếp theo: cách AI thay đổi không gian làm việc 441
Dữ liệu mở và sự uber hóa là mối đe dọa đối với ngành xây dựng hiện tại. 443
Các vấn đề chưa được giải quyết của mô hình Uber hóa như là cơ hội cuối cùng để tận dụng thời gian cho việc chuyển đổi. 446
CHƯƠNG 10.2. HƯỚNG DẪN THỰC HÀNH VỀ VIỆC TRIỂN KHAI CÁCH TIẾP CẬN DỮ LIỆU HƯỚNG DẪN 451
Từ lý thuyết đến thực tiễn: bản đồ đường đi của chuyển đổi số trong xây dựng. 451
Xây dựng nền tảng số: 1-5 bước đến sự trưởng thành số. 453
Khám phá tiềm năng dữ liệu: 5-10 bước đến sự trưởng thành số. 457
Bản đồ đường đi của quá trình chuyển đổi: từ hỗn loạn đến công ty dựa trên dữ liệu. 463
Xây dựng trong ngành công nghiệp 5.0: cách kiếm tiền khi không thể giấu diếm nữa. 466
KẾ LUẬN 468
VỀ TÁC GIẢ 471
PHẢN HỒI 472
NHẬN XÉT VỀ BẢN DỊCH 472
KỸ NĂNG VÀ KHÁI NIỆM KHÁC 473
TỪ ĐIỂN 477
DANH MỤC TÀI LIỆU VÀ TÀI NGUYÊN TRỰC TUYẾN 483
MỤC LỤC THEO CHỦ ĐỀ 499

TỐI ĐA TIỆN ÍCH VỚI BẢN IN

Bạn đang cầm trên tay phiên bản kỹ thuật số miễn phí của Data-Driven Construction. Để làm việc thuận tiện hơn và truy cập nhanh chóng vào tài liệu, chúng tôi khuyên bạn nên chú ý đến ấn phẩm in:



- Luôn sẵn sàng: cuốn sách ở định dạng in sẽ trở thành công cụ làm việc đáng tin cậy, cho phép bạn nhanh chóng tìm kiếm và sử dụng các hình ảnh và biểu đồ cần thiết trong bất kỳ tình huống làm việc nào.
- Chất lượng hình ảnh cao: tất cả hình ảnh và đồ họa trong ấn phẩm in được trình bày với chất lượng tốt nhất.
- Truy cập nhanh chóng vào thông tin: điều hướng thuận tiện, khả năng ghi chú, đánh dấu và làm việc với cuốn sách ở bất kỳ đâu.

Khi bạn mua phiên bản in đầy đủ của cuốn sách, bạn nhận được công cụ tiện lợi cho việc làm việc thoải mái và hiệu quả với thông tin: khả năng sử dụng nhanh chóng các tài liệu hình ảnh trong công việc hàng ngày, nhanh chóng tìm thấy các sơ đồ cần thiết và ghi chú. Ngoài ra, việc mua hàng của bạn hỗ trợ việc phổ biến kiến thức mở.

Bạn có thể đặt hàng phiên bản in của cuốn sách tại: datadrivenconstruction.io/books



I PHẦN
TỪ BẢNG ĐẤT SÉT ĐẾN CÁCH MẠNG SỐ: CÁCH THÔNG TIN ĐÃ TIẾN HÓA TRONG NGÀNH XÂY DỰNG.

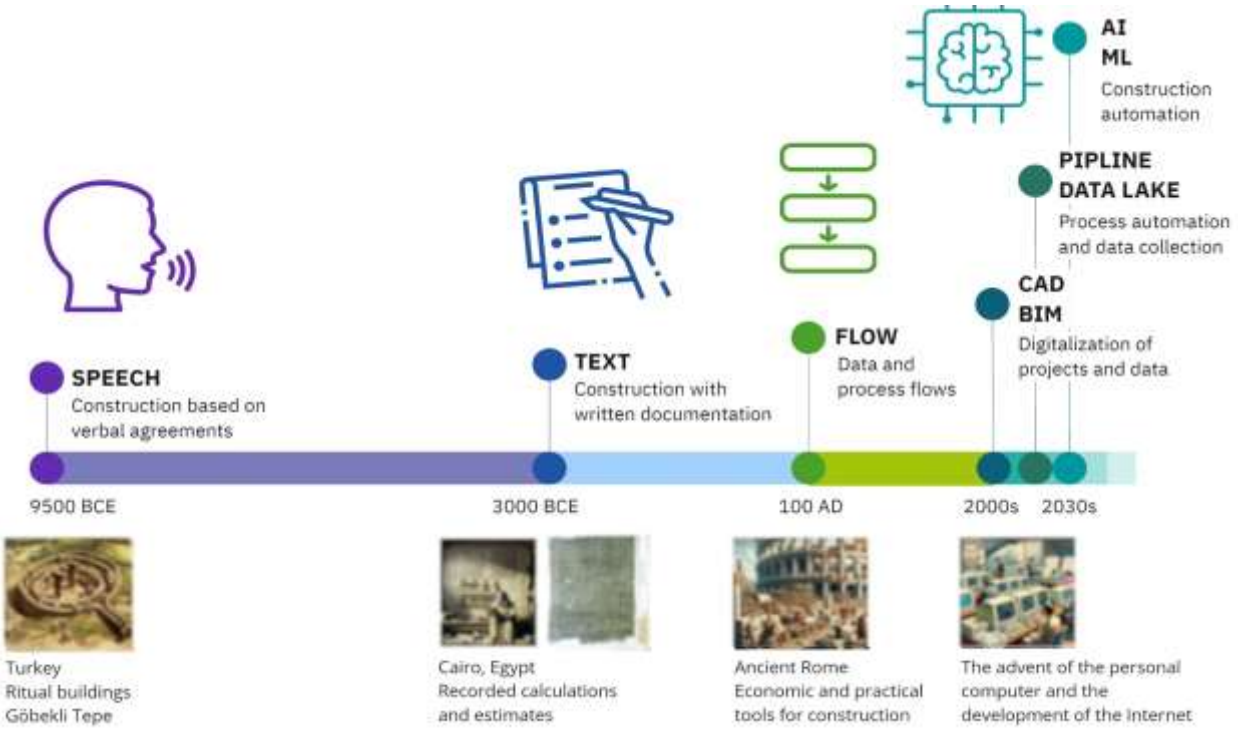
Phần đầu tiên của cuốn sách xem xét sự tiến hóa lịch sử của quản lý dữ liệu trong ngành xây dựng — từ những ghi chép thô sơ trên các phương tiện vật lý đến những hệ sinh thái kỹ thuật số hiện đại. Phân tích sự chuyển đổi công nghệ trong quản lý thông tin, sự xuất hiện của các hệ thống ERP và ảnh hưởng của việc phân mảnh dữ liệu đến hiệu quả của các quy trình kinh doanh. Đặc biệt chú trọng vào quá trình số hóa thông tin và tầm quan trọng ngày càng tăng của phân tích khách quan thay cho các đánh giá chủ quan của chuyên gia. Xem xét chi tiết sự gia tăng theo cấp số nhân của khối lượng thông tin mà ngành xây dựng hiện đại phải đối mặt và các thách thức liên quan đến các hệ thống doanh nghiệp. Nghiên cứu định vị ngành xây dựng trong bối cảnh của cuộc cách mạng công nghiệp lần thứ tư và lần thứ năm, cũng như tiềm năng sử dụng trí tuệ nhân tạo và các phương pháp tập trung vào dữ liệu để tạo ra lợi thế cạnh tranh bền vững.

CHƯƠNG 1.1.
SỰ TIẾN HÓA CỦA VIỆC SỬ DỤNG DỮ LIỆU TRONG NGÀNH XÂY DỰNG

Sự ra đời của kỷ nguyên dữ liệu trong xây dựng

Khoảng 10.000 năm trước, trong thời kỳ đồ đá mới, nhân loại đã thực hiện một bước chuyển cách mạng trong sự phát triển của mình, từ bỏ lối sống du mục để theo đuổi sự ổn định, điều này đã dẫn đến sự xuất hiện của những công trình thô sơ đầu tiên được xây dựng bằng đất sét, gỗ và đá. Từ thời điểm này trở đi, lịch sử của ngành xây dựng bắt đầu.

Khi các nền văn minh phát triển, kiến trúc trở nên ngày càng phức tạp, dẫn đến sự xuất hiện của các ngôi đền và tòa nhà công cộng đầu tiên. Việc phức tạp hóa các dự án kiến trúc yêu cầu các kỹ sư và người quản lý thời xưa phải tạo ra những ghi chép và tính toán đầu tiên. Những ghi chép đầu tiên trên các bảng đất sét và giấy cói thường bao gồm mô tả logic của việc tính toán khối lượng vật liệu xây dựng cần thiết, chi phí của chúng và tính toán tiền công cho công việc đã hoàn thành. Như vậy, thời kỳ sử dụng dữ liệu trong xây dựng đã bắt đầu — từ trước khi xuất hiện công nghệ số hiện đại.

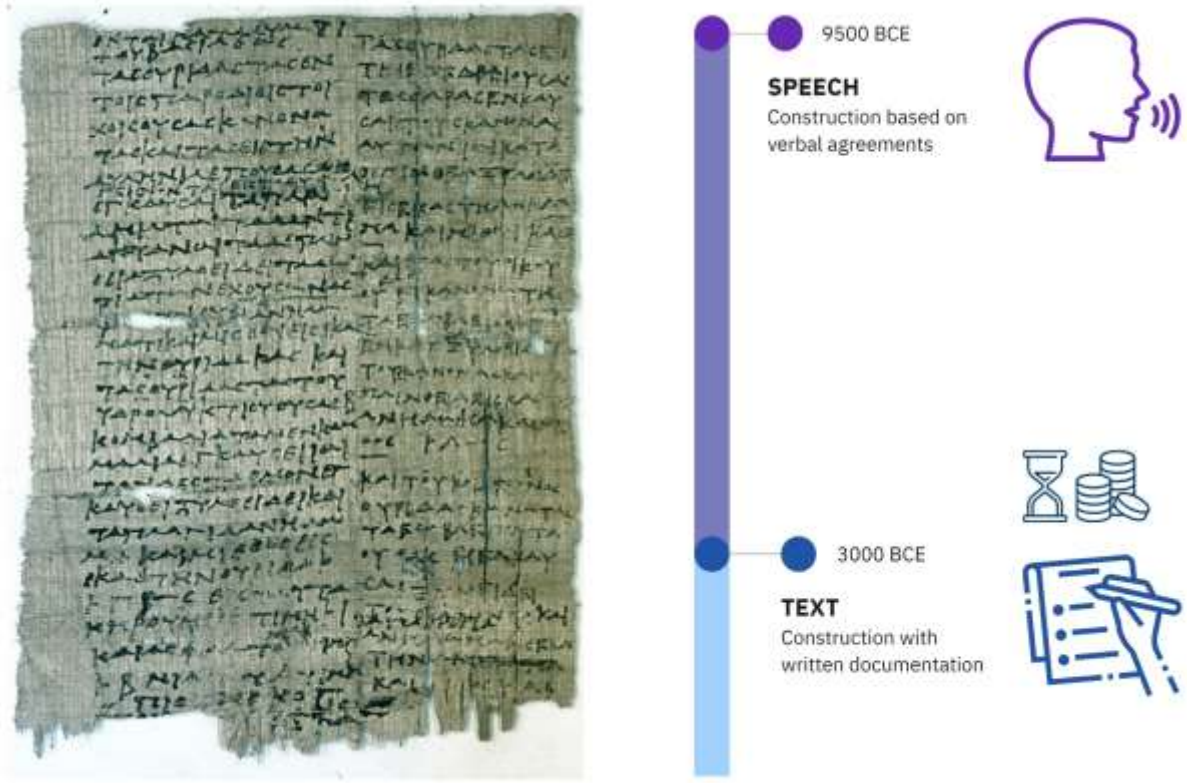


Hình 1.11: Lịch sử phát triển công nghệ thông tin trong xây dựng: từ thông tin bằng lời đến trí tuệ nhân tạo.

Từ đất sét và giấy cói đến công nghệ số

Những tài liệu đầu tiên trong xây dựng có từ thời kỳ xây dựng kim tự tháp, khoảng 3000–4000 năm trước Công nguyên. Kể từ đó, việc lưu giữ hồ sơ viết đã giúp đỡ và đồng hành với sự tiến bộ trong ngành xây dựng, cho phép tích lũy và hệ thống hóa kiến thức, mà trong 10.000 năm tiếp theo đã dẫn đến những đổi mới đáng kể trong phương pháp xây dựng và kiến trúc.

Việc sử dụng những phương tiện vật lý đầu tiên trong xây dựng, chẳng hạn như bảng đất sét, giấy papyrus hàng nghìn năm tuổi hoặc giấy khổ "A0" vào những năm 1980, để ghi chép dữ liệu ban đầu không có ý định sử dụng thông tin này trong các dự án mới. Mục đích chính của những ghi chép như vậy là mô tả chi tiết tình trạng hiện tại của dự án, bao gồm các tính toán về vật liệu cần thiết và chi phí thực hiện. Tương tự, trong thế giới hiện đại, việc sở hữu dữ liệu và mô hình thiết kế số không phải lúc nào cũng đảm bảo việc áp dụng chúng trong các dự án tương lai và thường chỉ đóng vai trò là thông tin cho các tính toán hiện tại về vật liệu cần thiết và chi phí xây dựng.



Hình 1.12: Giấy papyrus thế kỷ III trước Công nguyên, mô tả chi phí vẽ trang trí các loại cửa sổ khác nhau trong cung điện hoàng gia bằng kỹ thuật khắc sáp.

> Nhân loại đã mất khoảng 5.000 năm để chuyển từ các cuộc trò chuyện bằng miệng sang văn bản viết trong việc quản lý các dự án xây dựng, và cũng mất tương tự khoảng thời gian để chuyển từ giấy tờ sang dữ liệu số như nguồn tài nguyên chính cho việc lập kế hoạch và kiểm soát.

Tương tự như việc phát triển các quan hệ thương mại và tiền tệ đã thúc đẩy sự xuất hiện của chữ viết và những luật sư đầu tiên giải quyết các vấn đề tranh chấp, những ghi chép đầu tiên về chi phí vật liệu và khối lượng công việc trong xây dựng đã dẫn đến sự xuất hiện của những quản lý đầu tiên trong ngành xây dựng, người có nhiệm vụ ghi chép, theo dõi và chịu trách nhiệm về thông tin quan trọng liên quan đến thời hạn và chi phí của dự án.

Ngày nay, dữ liệu đóng vai trò quan trọng hơn rất nhiều: chúng không chỉ ghi lại các quyết định đã được đưa ra, mà còn trở thành công cụ dự đoán và mô phỏng tương lai. Trên cơ sở đó, phương pháp quản lý dự án hiện đại được xây dựng - chuyển đổi kinh nghiệm tích lũy thành hệ thống ra quyết định dựa trên dữ liệu có cấu trúc và có thể kiểm tra.

Quy trình như một công cụ của trải nghiệm được điều khiển bởi dữ liệu.

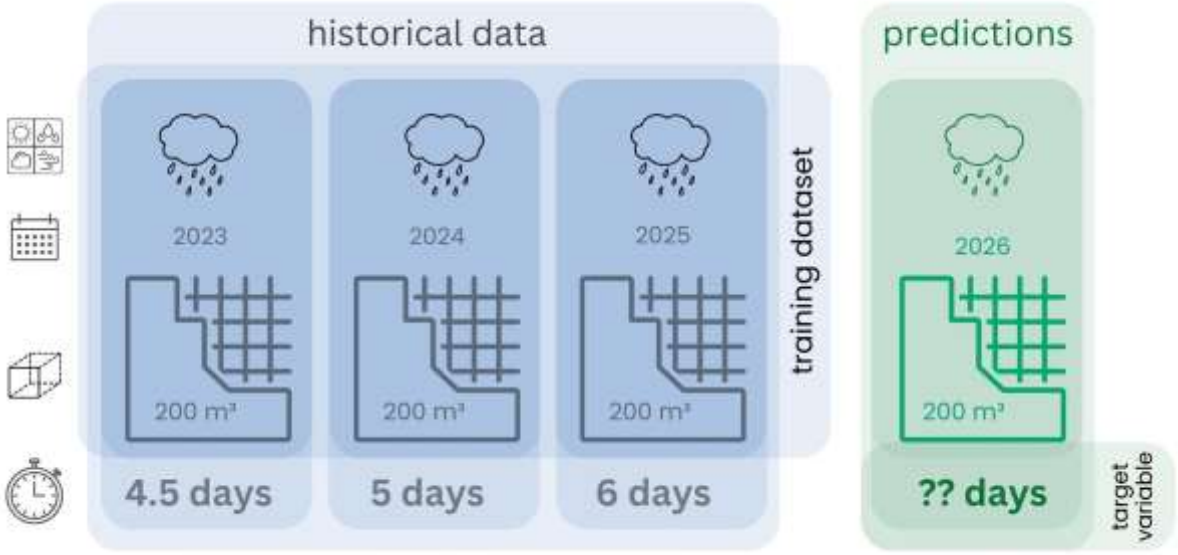
Bất kỳ quá trình nào cũng dựa trên việc chuyển đổi kinh nghiệm trong quá khứ thành công cụ lập kế hoạch cho tương lai. Kinh nghiệm trong nghĩa hiện đại đại diện cho một tập hợp dữ liệu có cấu trúc, việc phân tích nó cho phép đưa ra những dự đoán hợp lý.

> Chính dữ liệu lịch sử là nền tảng cho việc dự đoán, vì chúng thể hiện rõ ràng kết quả công việc đã hoàn thành và cung cấp cái nhìn về các yếu tố ảnh hưởng đến những kết quả này.

Xem xét một ví dụ cụ thể từ xây dựng monolithic: thường thì khi lập kế hoạch thời gian công việc, người ta sẽ xem xét khối lượng bê tông, độ phức tạp của cấu trúc và điều kiện thời tiết. Giả sử rằng một nhà thầu cụ thể trên công trường hoặc dữ liệu lịch sử của công ty trong ba năm qua (2023–2025) cho thấy rằng việc đổ bê tông cho một cấu trúc monolithic có diện tích 200 m² trong thời tiết mưa cần từ 4,5 đến 6 ngày (Hình 1.13). Chính những thống kê tích lũy này trở thành cơ sở cho việc dự đoán thời gian hoàn thành và tính toán tài nguyên khi lập kế hoạch cho các công việc tương tự trong các dự án tương lai. Dựa trên những dữ liệu lịch sử này, nhà thầu hoặc người lập dự toán có thể đưa ra dự đoán hợp lý, dựa trên kinh nghiệm đã thu được, về thời gian cần thiết để thực hiện các công việc tương tự trong năm 2026 trong các điều kiện tương tự.

Trong trường hợp này, việc đánh giá thời gian - quá trình phân tích hoạt động như một cơ chế chuyển đổi dữ liệu rời rạc thành kinh nghiệm có cấu trúc, và sau đó - thành một công cụ lập kế hoạch chính xác. Dữ liệu và quy trình là một hệ sinh thái thống nhất, nơi mà một cái không thể tồn tại mà không có cái kia.

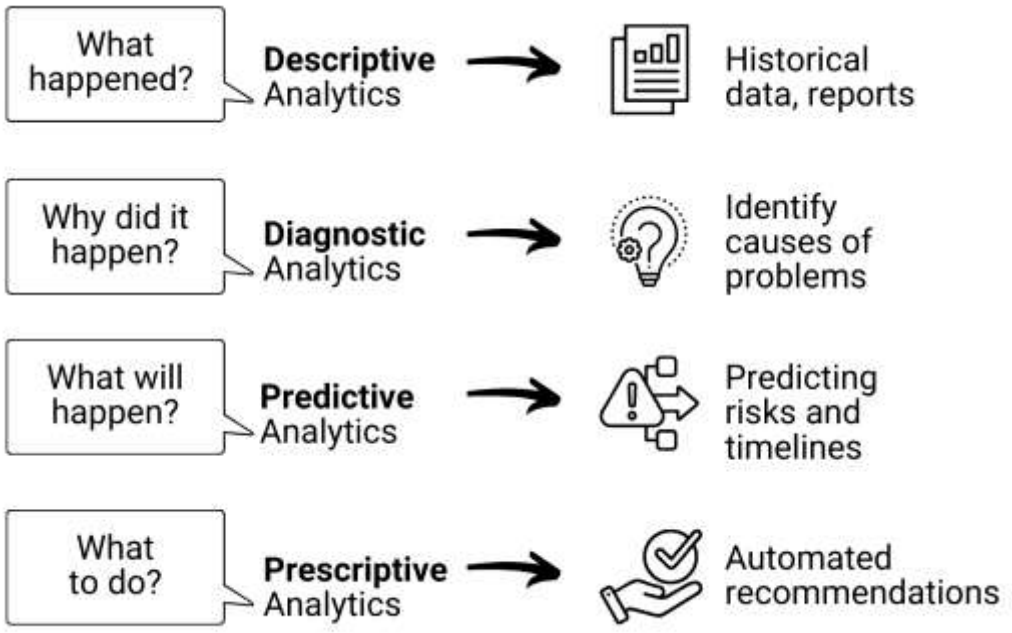
> *Hãy coi những gì có thể đếm được là có thể đếm, đo lường những gì có thể đo lường, và biến những gì không thể đo lường thành có thể đo lường.*
>
> *Galileo Galilei*



Hình 1.13 Dữ liệu lịch sử đóng vai trò như một tập dữ liệu huấn luyện để dự đoán một trong các đại lượng trong tương lai.

Trong bối cảnh kinh doanh hiện đại, phân tích dữ liệu trở thành một thành phần quan trọng trong việc quản lý dự án hiệu quả, tối ưu hóa quy trình và ra quyết định chiến lược. Ngành xây dựng đang dần tiếp cận bốn cấp độ phân tích chính, mỗi cấp độ đều trả lời một câu hỏi cụ thể và cung cấp những lợi ích độc đáo.-

- Phân tích mô tả - trả lời câu hỏi "điều gì đã xảy ra?" và cung cấp dữ liệu lịch sử cũng như báo cáo về các sự kiện và kết quả trong quá khứ: trong ba năm qua (2023–2025), việc đổ bê tông cho cấu trúc monolithic có diện tích 200 m² trong thời tiết mưa cần từ 4,5 đến 6 ngày.
- Phân tích chẩn đoán - trả lời câu hỏi "tại sao điều này xảy ra?", xác định nguyên nhân gây ra vấn đề: phân tích cho thấy thời gian đổ bê tông cấu trúc monolithic tăng lên do thời tiết mưa, làm chậm quá trình đông cứng của bê tông.
- Phân tích dự đoán - hướng tới tương lai, dự đoán các rủi ro có thể xảy ra và thời gian thực hiện công việc, trả lời câu hỏi "điều gì sẽ xảy ra?": dựa trên dữ liệu lịch sử, dự đoán rằng việc đổ bê tông cho một cấu trúc monolithic tương tự có diện tích 200 m² trong thời tiết mưa vào năm 2026 sẽ mất khoảng 5,5 ngày, với sự xem xét tất cả các yếu tố và xu hướng đã biết.
- Phân tích prescriptive cung cấp các khuyến nghị tự động và trả lời câu hỏi "cần làm gì?", cho phép các công ty lựa chọn các hành động tối ưu: Để tối ưu hóa công việc, ví dụ, nên sử dụng các phụ gia đặc biệt để tăng tốc độ đông cứng của bê tông trong điều kiện độ ẩm cao; lập kế hoạch đổ bê tông vào những khoảng thời gian có khả năng mưa thấp nhất; tổ chức các che chắn tạm thời cho công trình, điều này sẽ giúp rút ngắn thời gian thi công xuống còn 4-4,5 ngày ngay cả trong điều kiện thời tiết không thuận lợi.



Hình 1.14 Các loại phân tích chính: từ mô tả quá khứ đến quy trình ra quyết định tự động hóa.

Chuyển đổi số toàn diện, yêu cầu chuyển sang phân tích hệ thống và quản lý dựa trên dữ liệu, không chỉ cần thu hút các nhà thầu bên ngoài mà còn cần hình thành một đội ngũ nội bộ có năng lực. Các thành viên chủ chốt của đội ngũ này nên là các quản lý sản phẩm, kỹ sư dữ liệu, nhà phân tích và lập trình viên, những người sẽ làm việc chặt chẽ với các bộ phận kinh doanh (Hình 4.39). Sự hợp tác này là cần thiết để đặt ra các câu hỏi phân tích hợp lý và định hình hiệu quả các nhiệm vụ kinh doanh trong việc ra quyết định. Trong bối cảnh xã hội thông tin, dữ liệu không chỉ là công cụ hỗ trợ mà còn là nền tảng cho việc dự đoán và tối ưu hóa.-

Trong xây dựng, chuyển đổi số đang thay đổi căn bản các phương pháp thiết kế, quản lý và vận hành các công trình. Quá trình này được gọi là số hóa thông tin - khi tất cả các khía cạnh của quy trình xây dựng được chuyển đổi sang dạng số, phù hợp cho việc phân tích.

Số hóa thông tin trong quy trình xây dựng

Trong suốt hàng ngàn năm, khối lượng thông tin được ghi lại trong xây dựng gần như không thay đổi, nhưng trong vài thập kỷ qua, nó đã tăng vọt (Hình 1.15).

> Theo nghiên cứu của PwC® "Quản lý dữ liệu. Những gì sinh viên cần để thành công trong thế giới kinh doanh đang thay đổi nhanh chóng" (2015) [9], 90% tất cả dữ liệu trên thế giới đã được tạo ra trong hai năm qua (tính đến năm 2015). Tuy nhiên, hầu hết các công ty không sử dụng dữ liệu này một cách đầy đủ, vì chúng hoặc nằm trong các hệ thống rời rạc, hoặc chỉ được lưu trữ mà không có phân tích thực tế.

Trong những năm gần đây, sự gia tăng khối lượng dữ liệu chỉ tăng tốc, gấp đôi từ 15 zettabyte vào năm 2015 lên 181 zettabyte vào năm 2025 [10]. Hàng ngày, các máy chủ của các công ty xây dựng và thiết kế được lấp đầy bằng tài liệu dự án, lịch trình công việc, tính toán và báo cáo tài chính. Đối với bản vẽ 2D/3D,

các định dạng được sử dụng là DWG, DXF và DGN, trong khi các mô hình 3D sử dụng RVT, NWC, PLN và IFC™. Các tài liệu văn bản, bảng tính và bài thuyết trình được lưu trữ dưới định dạng DOC, XLSX và PPT. Thêm vào đó, video và hình ảnh từ công trường được lưu trữ dưới định dạng MPG và JPEG, dữ liệu từ các thành phần IoT, thẻ RFID® (nhận dạng và theo dõi) và hệ thống quản lý tòa nhà BMS (giám sát và kiểm soát) được gửi đến theo thời gian thực.



Hình 1.15 Sự gia tăng theo cấp số nhân của khối lượng dữ liệu 2010–2025 (theo tài liệu [10]).

Trong bối cảnh khối lượng thông tin tăng nhanh, ngành xây dựng đang đối mặt với nhu cầu không chỉ thu thập và lưu trữ dữ liệu mà còn đảm bảo kiểm tra, xác minh, đo lường và xử lý phân tích dữ liệu. Ngày nay, ngành công nghiệp đang trải qua giai đoạn số hóa thông tin tích cực - chuyển đổi hệ thống tất cả các khía cạnh của hoạt động xây dựng sang dạng số, phù hợp cho việc phân tích, diễn giải và tự động hóa.

> Số hóa thông tin có nghĩa là thu thập thông tin về tất cả các thực thể và yếu tố của dự án xây dựng và quy trình xây dựng - bao gồm cả những thứ mà chúng ta trước đây không coi là thông tin - và chuyển đổi nó thành định dạng dữ liệu, để làm cho thông tin có thể đo lường định lượng và thuận tiện cho việc phân tích.

Trong bối cảnh xây dựng, điều này có nghĩa là ghi lại và biểu thị bằng con số thông tin về tất cả các yếu tố của dự án và tất cả các quy trình - từ việc di chuyển thiết bị và con người trên công trường đến các điều kiện thời tiết và khí hậu tại công trường, giá cả vật liệu hiện tại và lãi suất của ngân hàng trung ương - với mục tiêu hình thành các mô hình phân tích.

> *Nếu bạn có thể đo lường những gì bạn đang nói và biểu thị nó bằng con số - có nghĩa là bạn biết điều gì đó về chủ đề đó. Nhưng nếu bạn không thể biểu thị điều đó một cách định lượng, kiến thức của bạn rất hạn chế và không thỏa đáng. Có thể đó là giai đoạn khởi đầu, nhưng đó không phải là mức độ kiến thức khoa học thực sự.*
>
> *- U. Thomson (Lord Kelvin), 1824–1907, nhà khoa học người Anh*

Số hóa thông tin vượt xa cách tiếp cận truyền thống trong việc thu thập thông tin, khi chỉ ghi lại các chỉ số cơ bản - chẳng hạn như giờ công hoặc chi phí thực tế cho vật liệu. Ngày nay, hầu như bất kỳ sự kiện nào cũng có thể được chuyển đổi thành dòng dữ liệu, phù hợp cho phân tích sâu với việc sử dụng các công cụ phân tích tiên tiến và các phương pháp học máy. Ngành xây dựng đã có một sự chuyển mình cơ bản: từ bản vẽ giấy, bảng tính Excel và chỉ dẫn bằng miệng - sang các hệ thống kỹ thuật số, trong đó mỗi yếu tố của đối tượng trở thành nguồn dữ liệu. Ngay cả nhân viên - từ kỹ sư đến công nhân trên công trường - giờ đây cũng được coi là một tập hợp các biến số kỹ thuật số và tập dữ liệu.

> Theo dữ liệu từ KPMG "Vấn đề quen thuộc - cách tiếp cận mới: Tổng quan xây dựng toàn cầu năm 2023", các bản sao kỹ thuật số, AI (Trí tuệ nhân tạo) và Big Data đang trở thành những động lực chính trong việc nâng cao tính sinh lợi của các dự án.

Công nghệ hiện đại không chỉ đơn giản hóa việc thu thập thông tin, làm cho nó phần lớn tự động, mà còn giảm đáng kể chi phí lưu trữ dữ liệu. Kết quả là các công ty từ bỏ cách tiếp cận chọn lọc và thích lưu trữ toàn bộ khối lượng thông tin để phân tích sau này, điều này mở ra những cơ hội tiềm năng cho việc tối ưu hóa quy trình trong tương lai.

> Số hóa thông tin và kỹ thuật số hóa cho phép phát hiện giá trị tiềm ẩn, trước đây chưa được sử dụng của thông tin. Khi được tổ chức một cách hợp lý, dữ liệu có thể có một cuộc sống thứ hai: chúng có thể được tái sử dụng, suy nghĩ lại và tích hợp vào các dịch vụ và giải pháp mới.

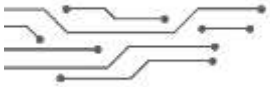
Trong tương lai, số hóa thông tin có thể dẫn đến việc tự động hóa hoàn toàn quy trình tài liệu, triển khai các quy trình xây dựng tự quản và sự xuất hiện của các nghề mới - nhà phân tích dữ liệu xây dựng, chuyên gia quản lý dự án AI và kỹ sư kỹ thuật số. Các công trình xây dựng sẽ trở thành các nguồn thông tin động, và việc ra quyết định sẽ dựa trên các sự kiện kỹ thuật số đáng tin cậy và có thể tái tạo, thay vì dựa vào trực giác hoặc kinh nghiệm chủ quan.

> *Thông tin là dầu mỏ của thế kỷ 21, và phân tích là động cơ đốt trong.*
> *- Peter Sondergaard, Phó Chủ tịch cấp cao Gartner®*

> Theo dữ liệu của IoT Analytics năm 2024, dự kiến rằng chi tiêu toàn cầu cho quản lý dữ liệu và phân tích sẽ tăng mạnh từ 185,5 tỷ đô la vào năm 2023 lên 513,3 tỷ đô la vào năm 2030, với tỷ lệ tăng trưởng hàng năm trung bình là 16%. Tuy nhiên, không phải tất cả các thành phần đều tăng trưởng với tốc độ như nhau: phân tích đang phát triển nhanh chóng, trong khi sự tăng trưởng của các hệ thống lưu trữ dữ liệu đang chậm lại. Phân tích sẽ mang lại sự tăng trưởng nhanh nhất trong hệ sinh thái quản lý dữ liệu: theo dự đoán, quy mô của nó sẽ tăng từ 60,6 tỷ đô la vào năm 2023 lên 227,9 tỷ đô la vào năm 2030, tương ứng với tỷ lệ tăng trưởng hàng năm trung bình là 27%.

Với sự gia tăng nhanh chóng của quá trình số hóa thông tin và sự phát triển mạnh mẽ của khối lượng thông tin, ban lãnh đạo các dự án xây dựng và công ty đang phải đối mặt với nhu cầu lưu trữ, phân tích và xử lý hệ thống các dữ liệu đa dạng, thường là không đồng nhất. Để đáp ứng thách thức này, bắt đầu từ giữa những năm 1990, ngành công nghiệp đã bắt đầu chuyển đổi hàng loạt sang việc tạo ra, lưu trữ và quản lý tài liệu điện tử — từ bảng tính và tính toán dự án đến bản vẽ và hợp đồng.

Các tài liệu giấy truyền thống, yêu cầu chữ ký, lưu trữ vật lý, xem xét định kỳ và lưu trữ trong tủ, đang dần bị thay thế bởi các hệ thống kỹ thuật số, trong đó dữ liệu được lưu trữ dưới dạng có cấu trúc — trong các cơ sở dữ liệu của các ứng dụng chuyên biệt.



CHƯƠNG 1.2.
CÔNG NGHỆ VÀ HỆ THỐNG QUẢN LÝ TRONG XÂY DỰNG HIỆN ĐẠI

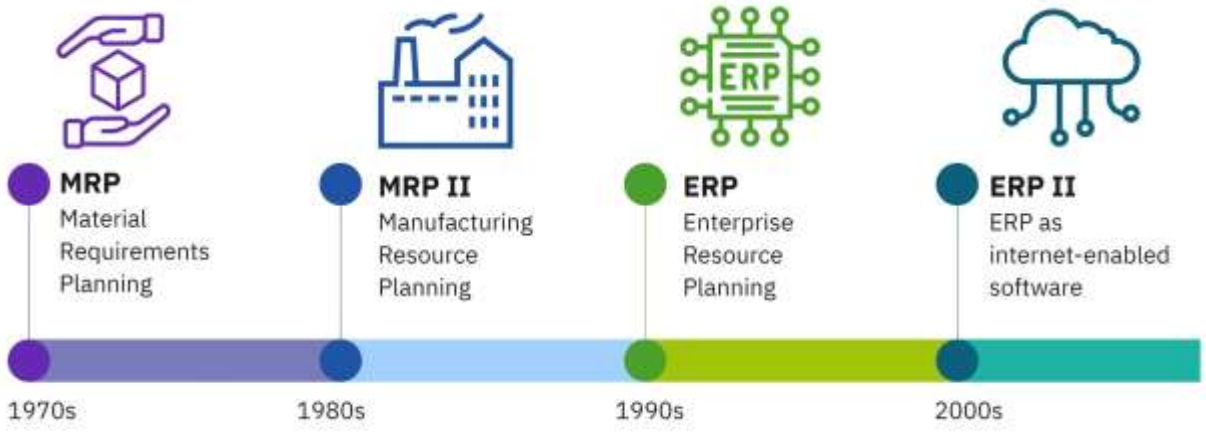
Cách mạng số và sự xuất hiện của các hệ thống MRP/ERP mô-đun

Kỷ nguyên lưu trữ và xử lý dữ liệu hiện đại dưới dạng kỹ thuật số bắt đầu với sự xuất hiện của băng từ vào những năm 1950, mở ra khả năng lưu trữ và sử dụng khối lượng thông tin lớn. Bước đột phá tiếp theo là sự xuất hiện của ổ đĩa, đã thay đổi hoàn toàn cách tiếp cận quản lý dữ liệu trong ngành xây dựng.

Với sự phát triển của các kho dữ liệu, thị trường giải pháp đã xuất hiện nhiều công ty bắt đầu phát triển phần mềm mô-đun để tạo ra, lưu trữ, xử lý dữ liệu và tự động hóa các nhiệm vụ thường xuyên.

> Sự gia tăng theo cấp số nhân của khối lượng thông tin và công cụ đã dẫn đến nhu cầu phát triển các giải pháp mô-đun tích hợp, không chỉ làm việc với các tệp riêng lẻ mà còn giúp quản lý và kiểm soát dòng dữ liệu trong các quy trình và dự án khác nhau.

Các công cụ tích hợp đầu tiên của nền tảng không chỉ cần lưu trữ tài liệu mà còn phải ghi lại tất cả các yêu cầu thay đổi và hoạt động trong các quy trình: ai đã khởi xướng, quy mô yêu cầu là gì và cuối cùng đã được ghi lại dưới dạng giá trị hoặc thuộc tính. Để thực hiện những mục tiêu này, cần có một hệ thống có thể theo dõi các tính toán chính xác và các quyết định đã được đưa ra. Những nền tảng như vậy đã trở thành các hệ thống MRP (Lập kế hoạch yêu cầu vật liệu) và ERP (Lập kế hoạch nguồn lực doanh nghiệp), đã trở nên phổ biến từ đầu những năm 1990.

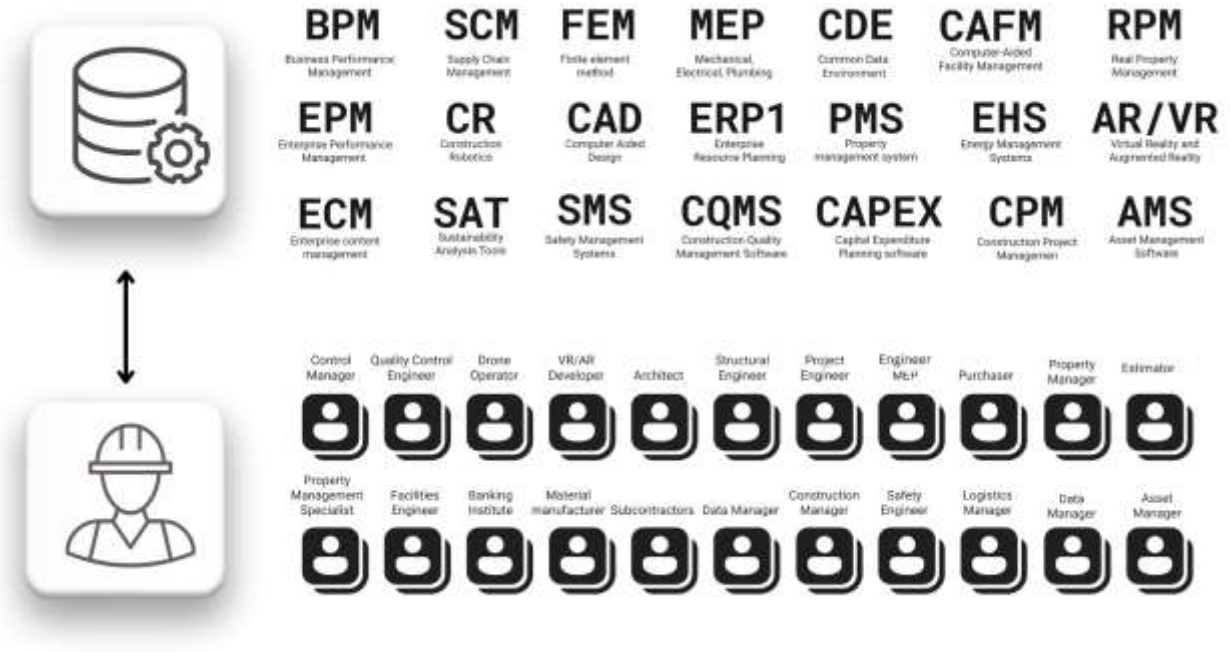


Các thành tựu trong lĩnh vực công nghệ lưu trữ dữ liệu đã dẫn đến sự xuất hiện của các hệ thống ERP vào những năm 1980.

Các hệ thống MRP và ERP đầu tiên đã đặt nền tảng cho kỷ nguyên số hóa trong quản lý quy trình kinh doanh và các dự án xây dựng. Các hệ thống mô-đun, ban đầu được thiết kế để tự động hóa các quy trình kinh doanh chính, đã dần dần tích hợp với các giải pháp phần mềm bổ sung, linh hoạt và thích ứng hơn.

Các giải pháp bổ sung này được thiết kế để xử lý dữ liệu và quản lý nội dung của các dự án, chúng hoặc thay thế các mô-đun nhất định của các hệ thống lớn, hoặc hiệu quả bổ sung cho chúng, mở rộng chức năng của

toàn bộ hệ thống.



Hình 1.22 Các giải pháp phần mềm mới đã thu hút một đội ngũ lớn các nhà quản lý vào doanh nghiệp để quản lý dòng dữ liệu.

Trong vài thập kỷ qua, các công ty đã đầu tư một số tiền đáng kể vào các hệ thống mô-đun, coi chúng như những giải pháp tích hợp lâu dài.

> Theo báo cáo của Software Path năm 2022, ngân sách trung bình cho mỗi người dùng hệ thống ERP là 9.000 đô la Mỹ. Trung bình khoảng 26% nhân viên trong công ty sử dụng các hệ thống này. Do đó, đối với một tổ chức có 100 người dùng, tổng chi phí triển khai ERP đạt khoảng 900.000 đô la.

Đầu tư vào các giải pháp mô-đun đóng kín, độc quyền ngày càng trở nên kém hợp lý trong bối cảnh sự phát triển nhanh chóng của các công nghệ hiện đại, linh hoạt và mở. Nếu những khoản đầu tư như vậy đã được thực hiện, điều quan trọng là phải đánh giá lại một cách khách quan vai trò của các hệ thống hiện có: liệu chúng có thực sự cần thiết trong dài hạn hay không, hoặc các chức năng của chúng có thể được xem xét lại và thực hiện một cách hiệu quả và minh bạch hơn.

Một trong những vấn đề chính của các nền tảng mô-đun hiện đại trong việc xử lý dữ liệu là chúng tập trung hóa việc quản lý dữ liệu trong các ứng dụng đóng kín. Kết quả là, dữ liệu - tài sản chính của công ty - trở nên phụ thuộc vào các giải pháp phần mềm cụ thể, chứ không phải ngược lại. Điều này hạn chế khả năng tái sử dụng thông tin, làm phức tạp quá trình di chuyển và giảm tính linh hoạt của doanh nghiệp trong bối cảnh môi trường số đang thay đổi nhanh chóng.

Nếu có khả năng giảm bớt tầm quan trọng hoặc nhu cầu về kiến trúc mô-đun đóng kín trong tương lai, thì có lý do để công nhận ngay hôm nay rằng các khoản chi phí đã bỏ ra là không thể thu hồi và tập trung vào việc chuyển đổi chiến lược sang một hệ sinh thái số mở, có thể mở rộng và thích ứng hơn.

> Phần mềm độc quyền được đặc trưng bởi sự kiểm soát tuyệt đối của công ty phát triển đối với mã nguồn và dữ liệu người dùng được tạo ra trong quá trình sử dụng các giải pháp như vậy. Khác với phần mềm mã nguồn mở, người dùng không có quyền truy cập vào cấu trúc nội bộ của ứng dụng và không thể tự mình xem xét, thay đổi hoặc điều chỉnh nó theo nhu cầu của mình. Thay vào đó, họ buộc phải mua giấy phép, cho phép quyền sử dụng phần mềm trong các giới hạn do nhà cung cấp quy định.

Cách tiếp cận hiện đại, tập trung vào dữ liệu, đề xuất một mô hình khác: dữ liệu nên được coi là tài sản chiến lược chính - độc lập, bền vững và tách biệt khỏi các giải pháp phần mềm cụ thể. Các ứng dụng, ngược lại, chỉ trở thành công cụ để làm việc với dữ liệu, có thể được thay thế một cách tự do mà không có nguy cơ mất thông tin quan trọng.

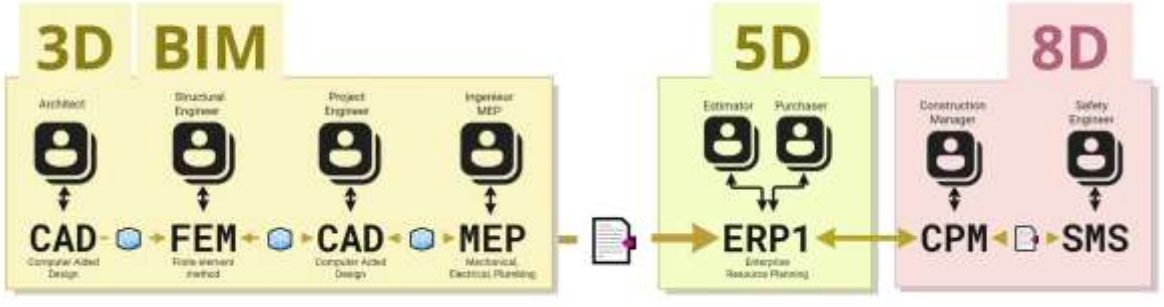
Sự phát triển của các hệ thống ERP và MRP trong những năm 1990 đã cung cấp cho doanh nghiệp những công cụ mạnh mẽ để quản lý quy trình, tuy nhiên cũng đi kèm với một hệ quả bất ngờ - sự gia tăng đáng kể số lượng nhân viên tham gia vào việc quản lý dòng thông tin. Thay vì tự động hóa và đơn giản hóa các nhiệm vụ vận hành, các hệ thống này thường tạo ra các cấp độ phức tạp, quan liêu và phụ thuộc vào các nguồn lực CNTT nội bộ.

Hệ thống quản lý dữ liệu: từ khai thác dữ liệu đến các bài toán kinh doanh.

Các công ty hiện đại đang phải đối mặt với nhu cầu tích hợp nhiều hệ thống quản lý dữ liệu. Việc lựa chọn các hệ thống quản lý dữ liệu, quản lý khéo léo các hệ thống này và tích hợp các nguồn dữ liệu rời rạc trở thành nhiệm vụ quan trọng đối với hiệu quả của doanh nghiệp.

Vào giữa những năm 2020, trong các công ty xây dựng vừa và nhỏ có thể tìm thấy hàng trăm (và trong các công ty lớn - hàng nghìn) hệ thống khác nhau (Hình 1.23), những hệ thống này cần hoạt động hài hòa để tất cả các khía cạnh của quy trình xây dựng diễn ra suôn sẻ và đồng bộ.

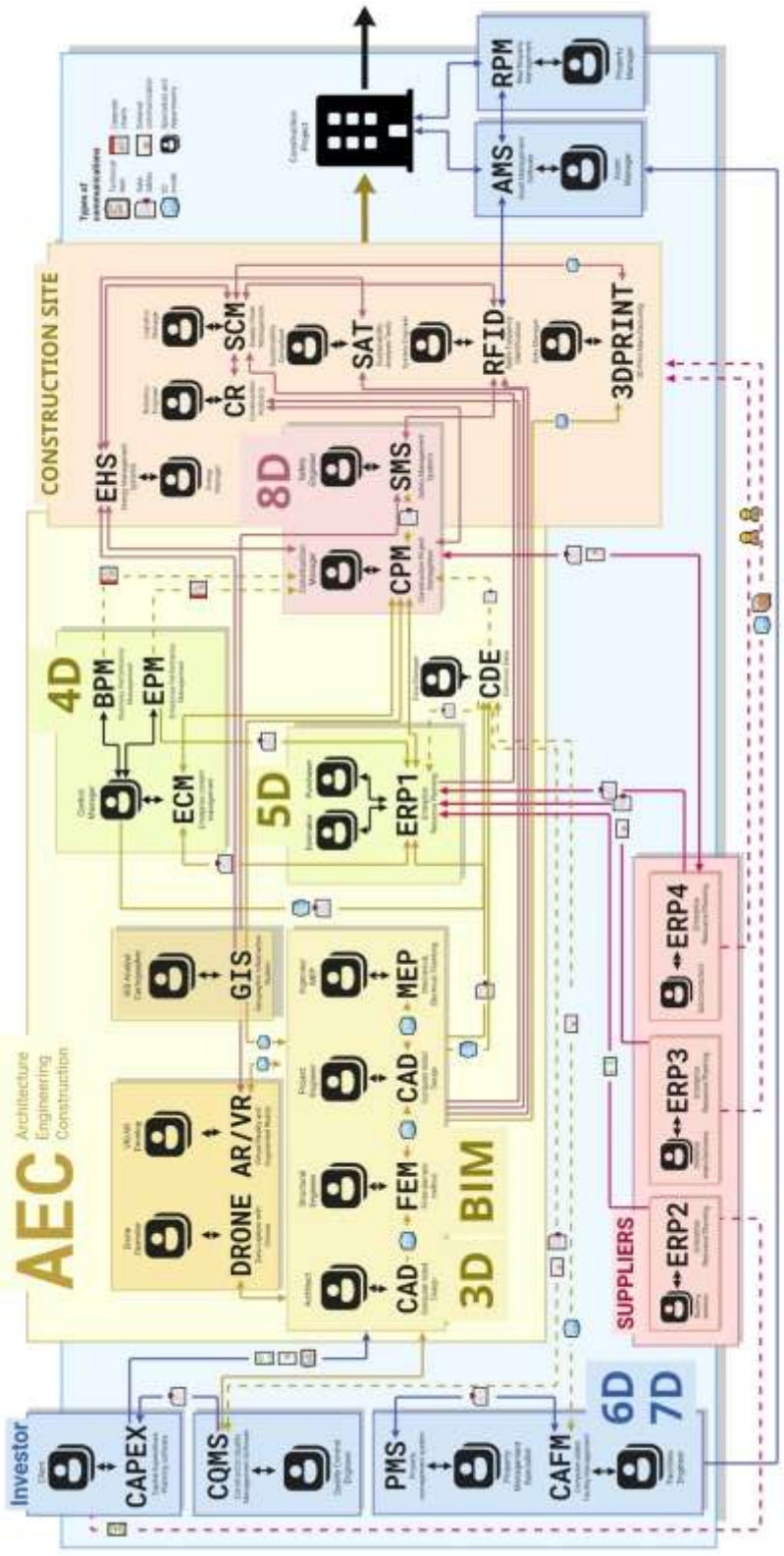
> Theo nghiên cứu của Deloitte® "Quản lý dựa trên dữ liệu trong các dự án vốn kỹ thuật số" năm 2016 - một chuyên gia xây dựng trung bình sử dụng 3,3 ứng dụng phần mềm mỗi ngày, tuy nhiên chỉ có 1,7 trong số đó được tích hợp với nhau.



Hình 1.23 Mỗi hệ thống kinh doanh yêu cầu một đội ngũ chuyên nghiệp và một quản lý có trách nhiệm để quản lý dữ liệu một cách chất lượng.

Dưới đây là danh sách các hệ thống phổ biến cho các công ty vừa và lớn trong ngành xây dựng, được sử dụng trong quản lý hiệu quả các dự án xây dựng:

- ERP (Enterprise Resource Planning) - đảm bảo tích hợp các quy trình kinh doanh, bao gồm kế toán, mua sắm và quản lý dự án.
- CAPEX (Phần mềm Lập kế hoạch Chi tiêu Vốn) - được sử dụng để lập ngân sách và quản lý các khoản đầu tư tài chính trong các dự án xây dựng, giúp xác định chi phí của tài sản cố định và đầu tư vào tài sản dài hạn.
- CAD (Thiết kế Hỗ trợ Máy tính) và BIM (Mô hình Thông tin Xây dựng) - được sử dụng để tạo ra các bản vẽ kỹ thuật chi tiết và chính xác cũng như các mô hình 3D của các dự án. Các hệ thống này tập trung vào việc làm việc với thông tin hình học.
- MEP (Cơ điện và Ống) - các hệ thống kỹ thuật bao gồm các thành phần cơ khí, điện và ống dẫn, chi tiết hóa hệ thống "mạch máu" bên trong của dự án.
- GIS (Hệ thống Thông tin Địa lý) - được sử dụng để phân tích và lập kế hoạch địa hình, bao gồm lập bản đồ và phân tích không gian.
- CQMS (Phần mềm Quản lý Chất lượng Xây dựng) - đảm bảo rằng các quy trình xây dựng tuân thủ các tiêu chuẩn và quy định đã thiết lập, giúp loại bỏ các khuyết điểm.
- CPM (Quản lý Dự án Xây dựng) - bao gồm lập kế hoạch, phối hợp và kiểm soát các quy trình xây dựng.
- CAFM (Quản lý Cơ sở Hỗ trợ Máy tính) - các hệ thống quản lý và bảo trì các tòa nhà.
- SCM (Quản lý Chuỗi Cung ứng) cần thiết để tối ưu hóa dòng chảy vật liệu và thông tin giữa các nhà cung cấp và công trường xây dựng.
- EPM (Quản lý Hiệu suất Doanh nghiệp) - nhằm cải thiện các quy trình kinh doanh và hiệu suất.
- AMS (Phần mềm Quản lý Tài sản) - được sử dụng để tối ưu hóa việc sử dụng, quản lý và bảo trì thiết bị và cơ sở hạ tầng trong suốt vòng đời của tài sản.
- RPM (Quản lý Tài sản Thực) - bao gồm các nhiệm vụ và quy trình liên quan đến quản lý và vận hành các tòa nhà và khu đất, cũng như các tài nguyên và tài sản liên quan.



Hình 1.24 Mối quan hệ giữa các hệ thống, kết nối các quy trình của công ty với dòng thông tin giữa các phòng ban khác nhau.

- CAE (Kỹ thuật Hỗ trợ Máy tính) - kỹ thuật máy tính, bao gồm các hệ thống tính toán và mô phỏng, chẳng hạn như phân tích phần tử hữu hạn (FEA) và động lực học chất lỏng tính toán (CFD).
- CFD (Động lực học Chất lỏng Tính toán) - mô phỏng các dòng chảy của chất lỏng và khí. Là một phân nhánh của CAE.
- CAPP (Lập kế hoạch quy trình hỗ trợ bằng máy tính) – lập kế hoạch quy trình công nghệ. Được sử dụng để tạo ra các bản đồ lộ trình và công nghệ.
- CAM (Sản xuất hỗ trợ bằng máy tính) – sản xuất tự động, tạo ra các chương trình điều khiển cho máy CNC.
- PDM (Quản lý dữ liệu sản phẩm) – quản lý dữ liệu về sản phẩm, hệ thống lưu trữ và quản lý tài liệu kỹ thuật.
- MES (Hệ thống thực thi sản xuất) – hệ thống quản lý quy trình sản xuất trong thời gian thực.
- PLM (Quản lý vòng đời sản phẩm) – quản lý vòng đời của một phần dự án, kết hợp PDM, CAPP, CAM và các hệ thống khác để kiểm soát toàn diện sản phẩm từ phát triển đến tái chế.

Những hệ thống này và nhiều hệ thống khác, bao gồm các giải pháp phần mềm đa dạng, đã trở thành một phần không thể thiếu trong ngành xây dựng hiện đại. Về bản chất, những hệ thống này là các cơ sở dữ liệu chuyên biệt với giao diện thân thiện, đảm bảo việc nhập, xử lý và phân tích thông tin hiệu quả ở tất cả các giai đoạn thiết kế và xây dựng. Việc tích hợp các công cụ kỹ thuật số với nhau không chỉ giúp tối ưu hóa quy trình làm việc mà còn nâng cao đáng kể độ chính xác của các quyết định, điều này có tác động tích cực đến thời gian và chất lượng thực hiện các dự án.

Tuy nhiên, việc tích hợp không diễn ra trong một nửa số trường hợp. Theo thống kê, chỉ có mỗi ứng dụng hoặc hệ thống thứ hai được tích hợp với các giải pháp khác. Điều này chỉ ra sự phân mảnh vẫn tồn tại trong môi trường kỹ thuật số và nhấn mạnh sự cần thiết phải phát triển các tiêu chuẩn mở và giao diện thống nhất để đảm bảo việc trao đổi thông tin xuyên suốt trong khuôn khổ dự án xây dựng.

Một trong những thách thức chính trong việc tích hợp cho các công ty hiện đại vẫn là độ phức tạp cao của các hệ thống kỹ thuật số và yêu cầu về năng lực người dùng cần thiết để tìm kiếm và diễn giải thông tin một cách hiệu quả. Để hỗ trợ mỗi hệ thống được triển khai trong doanh nghiệp, một đội ngũ chuyên gia được hình thành, do một quản lý chính dẫn dắt.

> Quản lý hệ thống chính đóng vai trò quyết định trong việc định hướng dòng dữ liệu và chịu trách nhiệm về chất lượng thông tin cuối cùng, giống như các quản lý đầu tiên hàng ngàn năm trước đã chịu trách nhiệm về các con số được ghi lại trên giấy cói hoặc bảng đất sét.

Để biến các dòng thông tin rời rạc thành công cụ quản lý, cần có khả năng tích hợp hệ thống và quản lý dữ liệu. Trong kiến trúc này, các quản lý phải hoạt động như các phần tử của một mạng lưới thống nhất – giống như mycelium, kết nối các phần riêng lẻ của công ty thành một cơ thể sống toàn diện, có khả năng thích ứng và phát triển.

Mẫu nấm doanh nghiệp: cách mà dữ liệu kết nối trong quy trình kinh doanh.

Quá trình tích hợp dữ liệu vào các ứng dụng và cơ sở dữ liệu dựa trên việc tổng hợp thông tin thu được từ nhiều nguồn khác nhau, bao gồm các phòng ban và chuyên gia khác nhau. Các chuyên gia tìm kiếm dữ liệu cần thiết, xử lý chúng và chuyển giao vào các hệ thống và ứng dụng của họ để sử dụng sau này.

Mỗi hệ thống của công ty, bao gồm một tập hợp các công cụ, công nghệ và cơ sở dữ liệu – là một cây tri

thức, có rễ sâu vào đất dữ liệu lịch sử và phát triển để mang lại những trái ngọt mới dưới dạng các giải pháp hoàn chỉnh: tài liệu, tính toán, bảng biểu, đồ thị và bảng điều khiển. Các hệ thống trong công ty, giống như những cây trong một khu rừng nhất định, tương tác và giao tiếp với nhau, tạo thành một hệ thống phức tạp và được cấu trúc tốt, được duy trì và quản lý bởi các chuyên gia quản lý.

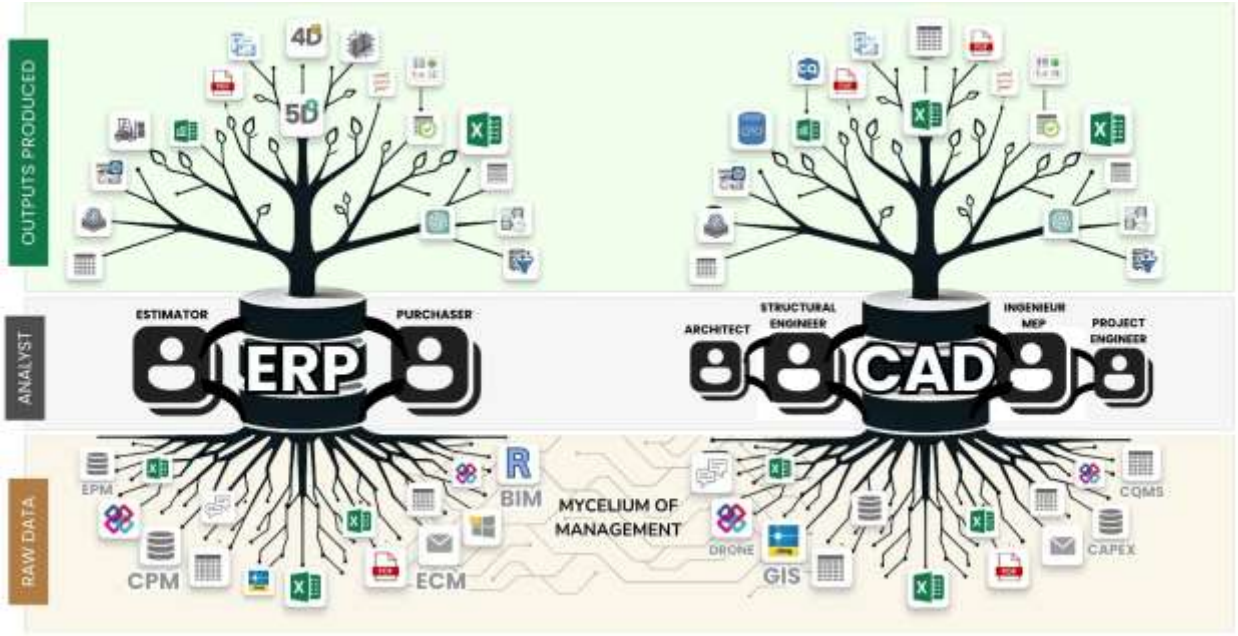
> Hệ thống tìm kiếm và truyền tải thông tin trong công ty hoạt động như một mạng lưới rừng phức tạp, bao gồm các cây (hệ thống) và mycelium của nấm (các nhà quản lý), những người đóng vai trò là người dẫn đường và xử lý, đảm bảo việc truyền tải thông tin và đưa nó vào các hệ thống cần thiết. Điều này giúp duy trì dòng chảy và phân phối dữ liệu khỏe mạnh và hiệu quả trong công ty.

Các chuyên gia, giống như rễ cây, hấp thụ dữ liệu thô ở giai đoạn đầu của dự án, biến chúng thành chất dinh dưỡng cho hệ sinh thái doanh nghiệp. Các hệ thống quản lý dữ liệu và nội dung (Hình 1.24 - ERP, CPM, BIM, v.v.) đóng vai trò là các tuyến thông tin mạnh mẽ, nơi những kiến thức này lưu thông qua tất cả các cấp độ của công ty.

Giống như trong tự nhiên, nơi mỗi yếu tố của hệ sinh thái đóng vai trò của mình, trong bối cảnh kinh doanh của công ty, mỗi người tham gia vào quá trình - từ kỹ sư đến nhà phân tích - đều đóng góp vào sự phát triển và màu mỡ của môi trường thông tin. Những "cây dữ liệu" hệ thống (Hình 1.25) không chỉ là cơ chế thu thập thông tin, mà còn là lợi thế cạnh tranh, đảm bảo sự phát triển bền vững của công ty.-

Các hệ sinh thái rừng phản ánh một cách đáng ngạc nhiên các nguyên tắc tổ chức của các cấu trúc doanh nghiệp kỹ thuật số. Tương tự như cấu trúc nhiều tầng của rừng - từ tầng dưới đến ngọn cây - quản lý doanh nghiệp phân phối nhiệm vụ theo các cấp độ trách nhiệm và các phòng ban chức năng.

> Rễ sâu và phân nhánh của cây cung cấp sự ổn định và truy cập vào chất dinh dưỡng. Tương tự, cấu trúc tổ chức vững chắc và các quy trình làm việc ổn định với dữ liệu chất lượng cao hỗ trợ toàn bộ hệ sinh thái thông tin của công ty, thúc đẩy sự phát triển bền vững ngay cả trong những thời kỳ (gió mạnh) bất ổn thị trường và khủng hoảng.



Hình 1.25 Sự tích hợp dữ liệu qua các hệ thống khác nhau giống như mycelium, kết nối các nhà quản lý và chuyên gia thành một mạng lưới thông tin thống nhất.

Hiểu biết hiện đại về quy mô trong kinh doanh đã tiến hóa. Ngày nay, giá trị của công ty không chỉ được xác định bởi phần nhìn thấy được - "tán cây" dưới dạng các tài liệu và báo cáo cuối cùng - mà còn bởi độ sâu của "hệ thống rễ" từ dữ liệu được thu thập và xử lý một cách có hệ thống. Càng nhiều thông tin được thu thập và xử lý, giá trị của doanh nghiệp càng cao. Các công ty có khả năng tích lũy "phân compost" từ dữ liệu đã được xử lý và biết cách khai thác những thông tin hữu ích từ đó sẽ có lợi thế chiến lược.

> Thông tin lịch sử trở thành một loại vốn mới, đảm bảo sự tăng trưởng, tối ưu hóa quy trình và lợi thế cạnh tranh. Trong một thế giới hướng tới dữ liệu, những người chiến thắng không phải là những ai có nhiều hơn, mà là những ai biết nhiều hơn.

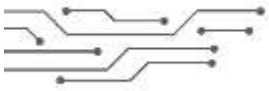
Đối với ngành xây dựng, điều này có nghĩa là chuyển sang quản lý dự án theo thời gian thực, nơi tất cả các quy trình - từ thiết kế và mua sắm đến phối hợp nhà thầu - sẽ dựa trên dữ liệu được cập nhật hàng ngày. Sự tích hợp thông tin từ các nguồn khác nhau (hệ thống ERP, mô hình CAD, cảm biến IoT trên công trường, RFID) sẽ cho phép xây dựng các dự đoán chính xác hơn, phản ứng kịp thời với những thay đổi và tránh được các sự chậm trễ do thiếu dữ liệu cập nhật.

> Theo nghiên cứu "Doanh nghiệp năm 2025, được điều khiển bởi dữ liệu" (McKinsey & Company®, 2022 [18]), các công ty thành công trong tương lai sẽ dựa vào dữ liệu trong tất cả các khía cạnh chính của hoạt động của họ — từ quyết định chiến lược đến tương tác vận hành.

Dữ liệu sẽ không còn chỉ là công cụ phân tích mà sẽ trở thành một phần không thể thiếu trong tất cả các quy trình kinh doanh, đảm bảo tính minh bạch, kiểm soát và tự động hóa quản lý. Cách tiếp cận dựa trên dữ liệu sẽ cho phép các tổ chức giảm thiểu ảnh hưởng của yếu tố con người, giảm rủi ro vận hành và nâng cao tính minh bạch cũng như hiệu quả trong việc ra quyết định.

Thế kỷ XXI đang lật ngược mô hình kinh tế: nếu như trước đây dầu mỏ được gọi là "vàng đen" vì khả năng vận hành các cơ chế và phương tiện, thì ngày nay, dưới áp lực của thời gian, dữ liệu lịch sử trở thành nguồn

tài nguyên chiến lược mới, nuôi dưỡng không phải máy móc mà là các thuật toán ra quyết định, sẽ thúc đẩy sự phát triển của doanh nghiệp.



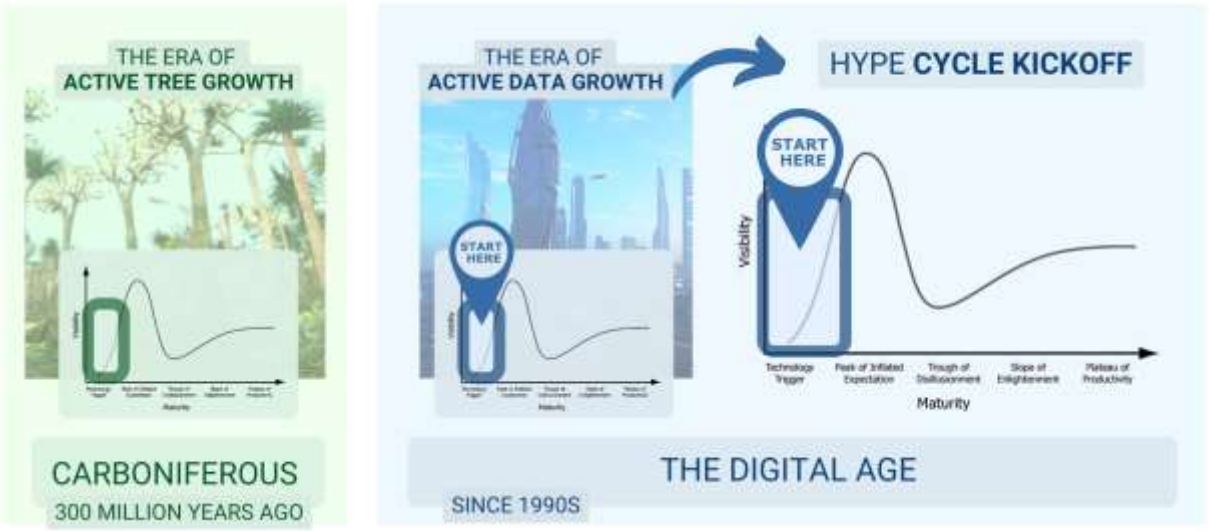
CHƯƠNG 1.3.
CÁCH MẠNG SỐ VÀ BÙNG NỔ DỮ LIỆU

Bắt đầu cuộc cách mạng khối lượng dữ liệu như một sóng tiến hóa

Ngành xây dựng đang trải qua một cuộc bùng nổ thông tin chưa từng có. Nếu tưởng tượng doanh nghiệp như một cái cây tri thức (Hình 1.25), được nuôi dưỡng bằng dữ liệu, thì giai đoạn số hóa hiện tại có thể so sánh với sự phát triển mạnh mẽ của thực vật trong kỷ than đá — thời kỳ mà sinh quyển Trái Đất đã biến đổi nhờ sự tích lũy sinh khối nhanh chóng (Hình 1.31).-

Trong bối cảnh phát triển số toàn cầu, khối lượng thông tin trong lĩnh vực xây dựng tăng gấp đôi hàng năm. Công nghệ hiện đại cho phép thu thập dữ liệu trong nền, phân tích chúng theo thời gian thực và sử dụng ở quy mô mà trước đây còn tưởng chừng không thể.

> Theo định luật Moore, được Gordon Moore (người đồng sáng lập Intel®) phát biểu, mật độ và độ phức tạp của các mạch tích hợp, cũng như khối lượng dữ liệu được xử lý và lưu trữ, tăng gấp đôi khoảng mỗi hai năm [19].



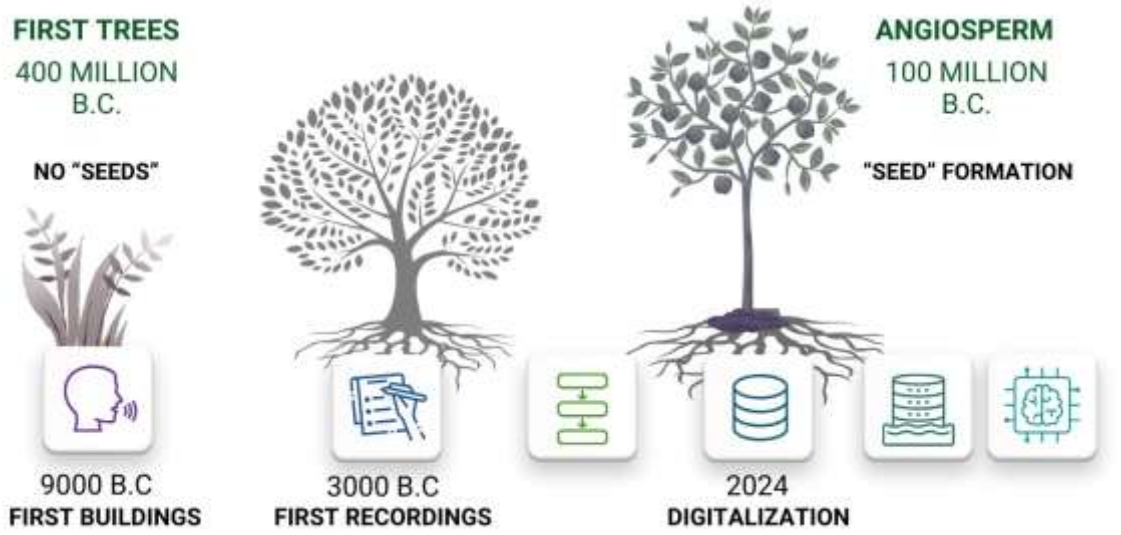
Hình 1.31 Sự khởi đầu của số hóa đã dẫn đến sự gia tăng dữ liệu theo cấp số nhân, giống như sự bùng nổ thực vật trong kỷ than đá.

Nếu như các công trình megalith cổ đại, như Göbekli Tepe (Thổ Nhĩ Kỳ), không để lại kiến thức tài liệu nào có thể tái sử dụng, thì ngày nay công nghệ số cho phép tích lũy và tái sử dụng thông tin. Điều này có thể so sánh với sự chuyển mình tiến hóa từ thực vật bào tử sang thực vật hạt (angiospem): sự xuất hiện của hạt đã thúc đẩy sự phát triển rộng rãi của sự sống trên hành tinh. (Hình 1.32).

Tương tự, dữ liệu từ các dự án trước đây trở thành "hạt giống số" — những người mang kiến thức DNA, có thể mở rộng và sử dụng trong các dự án và sản phẩm mới. Sự xuất hiện của các công cụ trí tuệ nhân tạo hiện đại — học máy và các mô hình ngôn ngữ lớn (LLM), như ChatGPT, LlaMa, Mistral, Claude, DeepSeek,

QWEN, Grok — cho phép tự động trích xuất, diễn giải và áp dụng dữ liệu trong các bối cảnh mới.

> Giống như cách mà hạt giống đã cách mạng hóa sự phát tán sự sống trên một hành tinh ban đầu không có sự sống, "hạt giống dữ liệu" đang trở thành nền tảng cho sự xuất hiện tự động của các cấu trúc thông tin và kiến thức mới, cho phép các hệ sinh thái số phát triển độc lập và thích ứng với các yêu cầu thay đổi của người dùng.



Hình 1.32 "Hạt giống dữ liệu" số đóng vai trò tiến hóa tương tự như angiosperms - thực vật có hoa, đã biến đổi hệ sinh thái của Trái Đất.

Chúng ta đang đứng trước ngưỡng cửa của một kỷ nguyên mới trong xây dựng, nơi bùng nổ dữ liệu và sự lan tỏa tích cực của "hạt giống dữ liệu" - thông tin có cấu trúc từ các dự án trong quá khứ và hiện tại - đang hình thành nền tảng cho tương lai số của ngành. Việc "thụ phấn" chúng bằng các mô hình ngôn ngữ lớn (LLM) không chỉ cho phép quan sát những thay đổi số mà còn tham gia tích cực vào việc tạo ra các hệ sinh thái tự học, thích ứng. Đây không phải là sự tiến hóa - mà là một cuộc cách mạng số, trong đó dữ liệu trở thành vật liệu xây dựng chính cho thực tại mới.

> *Khối lượng dữ liệu trong ngành xây dựng đang gia tăng mạnh mẽ nhờ thông tin từ nhiều lĩnh vực trong suốt vòng đời của các dự án xây dựng. Sự tích lũy dữ liệu khổng lồ này đã thúc đẩy ngành xây dựng bước vào kỷ nguyên Dữ liệu Lớn.*
>
> *- Giáo sư Hằng Dương, Khoa Xây dựng Dân dụng và Kiến trúc, Đại học Công nghệ Vũ Hán, Vũ Hán, Trung Quốc.*

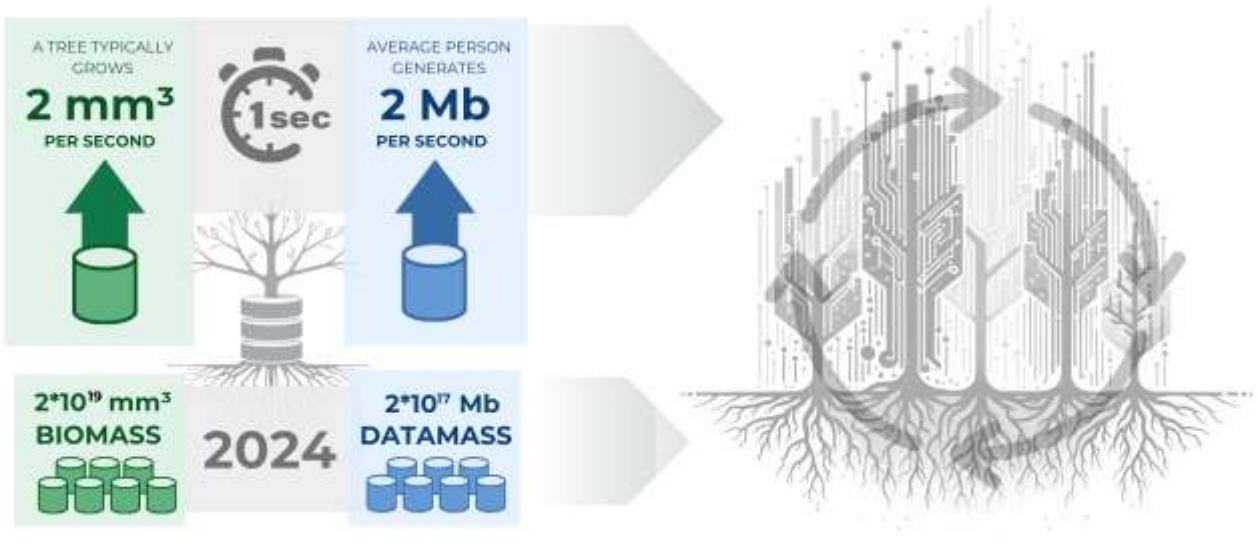
Sự gia tăng khối lượng dữ liệu trong kỷ nguyên thông tin gợi nhớ đến các quá trình tiến hóa trong tự nhiên: giống như sự phát triển của rừng đã thay đổi cảnh quan cổ xưa của hành tinh, thì sự bùng nổ thông tin hiện nay cũng đang thay đổi toàn bộ cảnh quan của ngành xây dựng.

Khối lượng dữ liệu được tạo ra trong một công ty hiện đại

Trong hai năm qua, 90% tất cả dữ liệu tồn tại trên thế giới đã được tạo ra. Tính đến năm 2023, mỗi người, bao gồm cả các chuyên gia trong ngành xây dựng, tạo ra khoảng 1,7 megabyte dữ liệu mỗi giây, và tổng khối lượng dữ liệu trên thế giới đã đạt 64 zettabyte vào năm 2023 và dự kiến sẽ vượt qua 180 zettabyte, hay 180*10^15 megabyte, vào năm 2025.

> Cuộc bùng nổ thông tin này có một tiền lệ lịch sử - sự phát minh của máy in bởi Johannes Gutenberg vào thế kỷ XV. Chỉ sau năm mươi năm kể từ khi nó ra đời, số lượng sách ở châu Âu đã tăng gấp đôi: trong vài thập kỷ, số lượng sách được in ra đã bằng số sách được tạo ra bằng tay trong 1200 năm trước đó. Hôm nay, chúng ta đang chứng kiến sự gia tăng thậm chí còn nhanh hơn: khối lượng dữ liệu trên thế giới tăng gấp đôi mỗi ba năm.

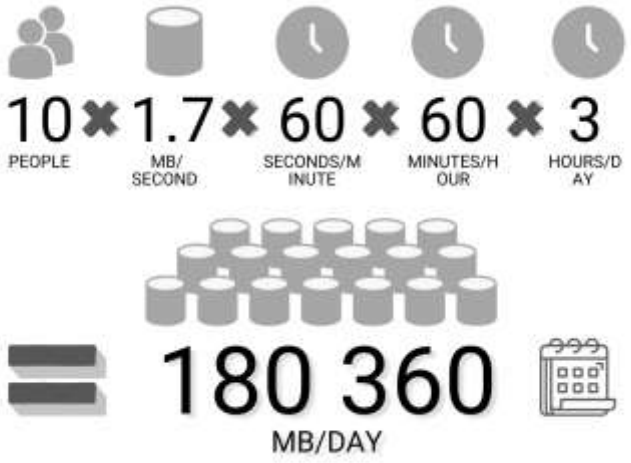
Với tốc độ tăng trưởng dữ liệu hiện tại, ngành xây dựng có thể tạo ra một khối lượng thông tin tương tự như toàn bộ lịch sử của nó trong vài thập kỷ tới.



Hình 1.33 Việc mỗi nhân viên lưu trữ dữ liệu hàng ngày trên các máy chủ của công ty góp phần vào sự gia tăng liên tục khối lượng dữ liệu.

Trong thế giới kinh doanh xây dựng hiện đại, ngay cả những công ty nhỏ cũng tạo ra một lượng lớn thông tin đa dạng hàng ngày, và dấu vết số của một công ty xây dựng nhỏ có thể đạt hàng chục gigabyte mỗi ngày - từ mô hình và bản vẽ đến ghi hình và cảm biến trên công trình. Nếu giả định rằng mỗi chuyên gia trung bình tạo ra khoảng 1,7 MB dữ liệu mỗi giây, thì điều này tương đương với khoảng 146 GB mỗi ngày, hoặc 53 TB mỗi năm.

Với một đội ngũ 10 người làm việc tích cực trong 3 giờ mỗi ngày, tổng khối lượng thông tin được tạo ra trong một ngày đạt 180 gigabyte.



Hình 1.34 Một công ty 10 người tạo ra khoảng 50–200 gigabyte dữ liệu mỗi ngày.

Nếu giả định rằng 30% dữ liệu công việc là mới (phần còn lại được ghi đè hoặc xóa), thì một công ty có 10 người có thể tạo ra khoảng vài trăm gigabyte dữ liệu mới mỗi tháng (các chỉ số thực tế phụ thuộc vào loại hình hoạt động của công ty).

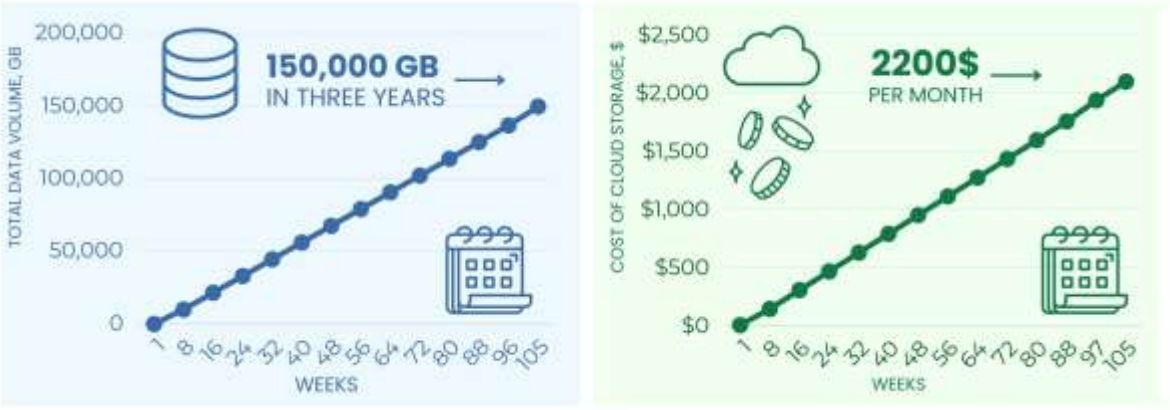
Do đó, trở nên rõ ràng: chúng ta không chỉ tạo ra ngày càng nhiều dữ liệu - mà còn phải đối mặt với nhu cầu ngày càng tăng về việc quản lý, lưu trữ và đảm bảo khả năng truy cập lâu dài cho chúng. Và nếu trước đây dữ liệu có thể "nằm" trên các máy chủ cục bộ mà không tốn kém nhiều, thì trong bối cảnh chuyển đổi số, ngày càng nhiều công ty bắt đầu sử dụng các giải pháp đám mây như nền tảng cho cơ sở hạ tầng thông tin của họ.

Chi phí lưu trữ dữ liệu: khía cạnh kinh tế

Trong những năm gần đây, ngày càng nhiều công ty chuyển giao việc lưu trữ dữ liệu cho các dịch vụ đám mây. Ví dụ, nếu một công ty lưu trữ một nửa dữ liệu của mình trên đám mây, với mức giá trung bình 0,015 đô la cho mỗi gigabyte mỗi tháng, chi phí lưu trữ của họ có thể tăng từ 10 đến 50 đô la mỗi tháng.

Đối với một công ty nhỏ với các mô hình tạo dữ liệu điển hình, chi phí lưu trữ đám mây có thể dao động từ hàng trăm đến có thể hơn một nghìn đô la mỗi tháng (Hình 1.35) sau vài năm, tạo ra gánh nặng tài chính tiềm tàng đáng kể.

> Theo nghiên cứu của Forrester "Các doanh nghiệp chuyển giao việc lưu trữ dữ liệu cho bên thứ ba khi độ phức tạp gia tăng", trong đó có sự tham gia của 214 nhà lãnh đạo quyết định trong lĩnh vực hạ tầng công nghệ, hơn một phần ba tổ chức chuyển giao hệ thống lưu trữ dữ liệu cho bên thứ ba để đối phó với khối lượng và độ phức tạp ngày càng tăng của các hoạt động dữ liệu, trong khi gần hai phần ba doanh nghiệp ưa chuộng mô hình dựa trên đăng ký.



Hình 1.35 Việc chuyển dữ liệu lên đám mây có thể làm tăng chi phí lưu trữ hàng tháng lên đến 2.000 đô la ngay cả đối với một công ty chỉ có 10 nhân viên.

Tình hình trở nên phức tạp hơn với sự chuyển đổi nhanh chóng sang các công nghệ đám mây như CAD (BIM), CAFM, PMIS và hệ thống ERP, điều này làm tăng thêm chi phí lưu trữ và xử lý dữ liệu. Cuối cùng, các công ty buộc phải tìm kiếm các cách để tối ưu hóa chi phí và giảm sự phụ thuộc vào các nhà cung cấp đám mây.

Kể từ năm 2023, với sự phát triển mạnh mẽ của các mô hình ngôn ngữ lớn (LLM), các phương pháp lưu trữ dữ liệu đã bắt đầu thay đổi. Ngày càng nhiều công ty đang suy nghĩ về việc lấy lại quyền kiểm soát dữ liệu, vì việc xử lý thông tin trên các máy chủ riêng trở nên an toàn và có lợi hơn.

Trong bối cảnh này, xu hướng từ bỏ các hệ thống lưu trữ và xử lý dữ liệu đám mây chỉ cho những dữ liệu cần thiết để chuyển sang triển khai cục bộ các LLM và giải pháp AI doanh nghiệp đang nổi lên. Như Giám đốc điều hành của Microsoft đã lưu ý trong một cuộc phỏng vấn của mình, thay vì phụ thuộc vào một số ứng dụng riêng lẻ hoặc giải pháp đám mây SaaS để thực hiện các nhiệm vụ khác nhau, các tác nhân AI sẽ quản lý các quy trình trong cơ sở dữ liệu, tự động hóa các chức năng của các hệ thống khác nhau.

> *Cách tiếp cận cũ đối với vấn đề này [xử lý dữ liệu] như sau: nếu nhớ lại cách mà các ứng dụng kinh doanh khác nhau xử lý việc tích hợp, thì họ đã sử dụng các kết nối. Các công ty đã bán giấy phép cho những kết nối này, và xung quanh đó hình thành một mô hình kinh doanh. SAP [ERP] là một trong những ví dụ cổ điển: việc truy cập dữ liệu SAP chỉ có thể thực hiện được khi có kết nối tương ứng. Vì vậy, tôi nghĩ rằng một điều tương tự sẽ xuất hiện trong trường hợp tương tác [AI] của các tác nhân [..]. Cách tiếp cận, ít nhất là cách mà chúng tôi chấp nhận, là: tôi nghĩ rằng khái niệm tồn tại của các ứng dụng kinh doanh có lẽ sẽ sụp đổ trong kỷ nguyên [AI] của các tác nhân. Bởi vì, nếu suy nghĩ kỹ, chúng về cơ bản là các cơ sở dữ liệu với rất nhiều logic kinh doanh. — Satya Nadella, Giám đốc điều hành Microsoft, phỏng vấn kênh BG2, năm 2024 [28]*

Trong mô hình này, cách tiếp cận dựa trên dữ liệu với việc sử dụng LLM vượt ra ngoài các hệ thống cổ điển. Trí tuệ nhân tạo trở thành trung gian giữa người dùng và dữ liệu (Hình. 2.23, Hình. 2.24), loại bỏ sự cần thiết của nhiều giao diện trung gian và nâng cao hiệu quả của các quy trình kinh doanh. Chúng ta sẽ nói chi tiết

hơn về cách tiếp cận này trong chương "Biến hỗn loạn thành trật tự và giảm độ phức tạp".-

Trong khi kiến trúc tương lai chỉ mới hình thành, các công ty đã phải đối mặt với hậu quả của những quyết định trong quá khứ. Sự số hóa hàng loạt trong vài thập kỷ qua, đi kèm với việc triển khai các hệ thống rời rạc và sự tích lũy dữ liệu không kiểm soát, đã dẫn đến một vấn đề mới — quá tải thông tin.

Ranh giới của việc tích lũy dữ liệu: từ khối lượng đến ý nghĩa

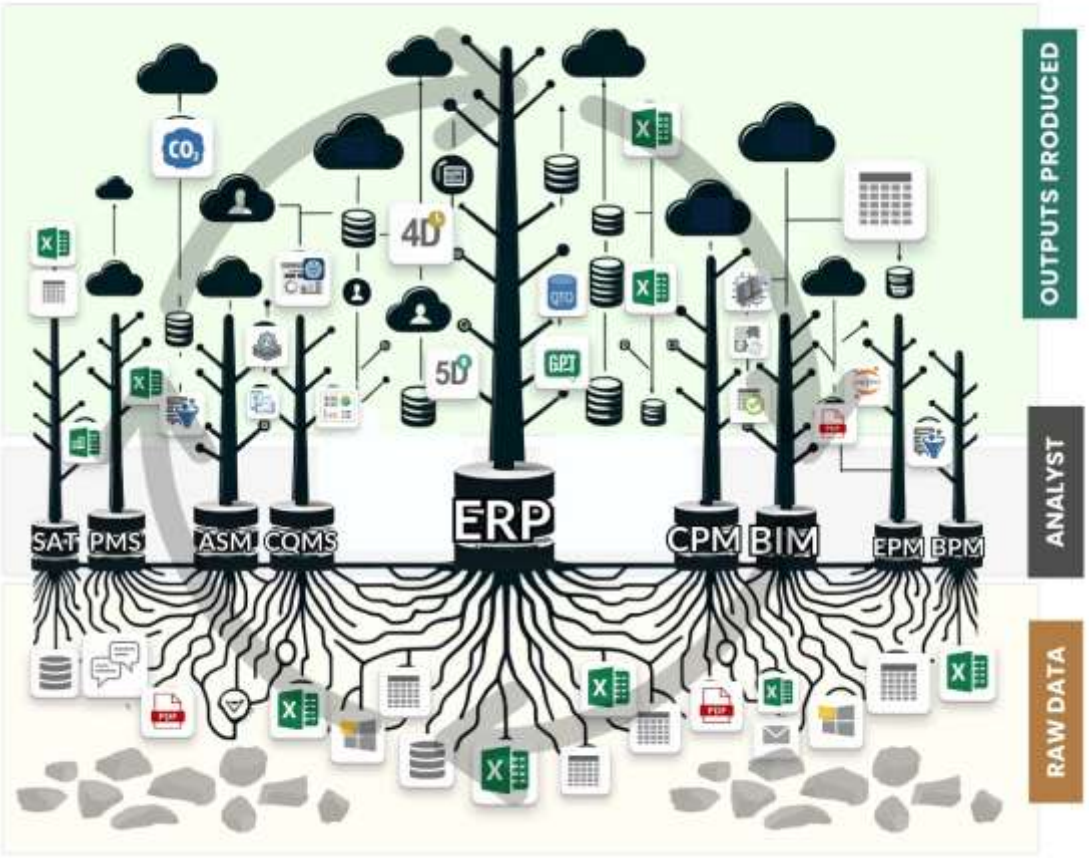
Các hệ thống hiện đại của các công ty phát triển và hoạt động thành công khi có sự tăng trưởng được quản lý, khi khối lượng dữ liệu và số lượng ứng dụng ở trong trạng thái cân bằng với khả năng của các bộ phận CNTT và quản lý. Tuy nhiên, trong vài thập kỷ qua, sự số hóa đã dẫn đến sự gia tăng không kiểm soát về khối lượng và độ phức tạp của dữ liệu, gây ra hiệu ứng bão hòa trong hệ sinh thái thông tin của các công ty.

Ngày nay, các máy chủ và kho lưu trữ đang phải đối mặt với một lượng thông tin thô và đa dạng chưa từng có, mà không kịp biến thành phân bón và nhanh chóng mất tính thời sự. Các nguồn lực hạn chế của công ty không thể xử lý được dòng chảy này, và dữ liệu tích lũy trong các kho lưu trữ bị cô lập (còn gọi là "silo"), đòi hỏi phải xử lý thủ công để trích xuất thông tin hữu ích.

> Kết quả là, giống như một khu rừng bị dây thường xuân bao phủ và bị mốc, các hệ thống quản lý hiện đại của các công ty thường bị ảnh hưởng bởi quá tải thông tin. Thay vì hình thành một lớp thông tin dinh dưỡng, hệ sinh thái doanh nghiệp lại tạo ra những khu vực cô lập của dữ liệu đa dạng, điều này không thể tránh khỏi dẫn đến giảm hiệu quả tổng thể của các quy trình kinh doanh.

Thời gian dài của sự gia tăng theo cấp số nhân về khối lượng dữ liệu, được quan sát trong 40 năm qua, sẽ không thể tránh khỏi chuyển sang giai đoạn bão hòa và làm mát tiếp theo. Khi các kho lưu trữ đạt đến giới hạn, sẽ xảy ra một sự chuyển biến chất lượng: dữ liệu sẽ không còn chỉ là đối tượng lưu trữ, mà sẽ trở thành một nguồn lực chiến lược.

Với sự phát triển của trí tuệ nhân tạo và học máy, các công ty có cơ hội giảm chi phí xử lý thông tin và chuyển từ tăng trưởng định lượng sang sử dụng dữ liệu một cách chất lượng hơn. Trong thập kỷ tới, ngành xây dựng sẽ phải chuyển trọng tâm - từ việc tạo ra những khối lượng dữ liệu mới sang việc đảm bảo tính cấu trúc, toàn vẹn và giá trị phân tích của chúng.



Hình 1.36 Các nguồn dữ liệu tách biệt cản trở việc trao đổi thông tin giữa các hệ thống dữ liệu.

> Giá trị chính không còn nằm ở khối lượng thông tin, mà ở khả năng tự động diễn giải và biến nó thành kiến thức ứng dụng, hữu ích cho việc ra quyết định quản lý. Để dữ liệu trở nên thực sự hữu ích, cần phải quản lý chúng một cách hợp lý: thu thập, kiểm tra, cấu trúc, lưu trữ và phân tích trong bối cảnh các nhiệm vụ kinh doanh cụ thể.

Quy trình phân tích dữ liệu trong công ty giống như chu kỳ sống và phân hủy của cây cối trong rừng và sự xuất hiện của những cây non khỏe mạnh mới: những cây trưởng thành chết đi, phân hủy và trở thành môi trường dinh dưỡng cho những mầm non mới. Các quy trình hoàn thành và kết thúc sau khi sử dụng sẽ rơi vào hệ sinh thái thông tin của công ty, cuối cùng trở thành phân bón thông tin, nuôi dưỡng sự phát triển của các hệ thống và dữ liệu mới trong tương lai.

Tuy nhiên, trong thực tế, chu kỳ này thường bị phá vỡ. Thay vì sự đổi mới hữu cơ, một hỗn loạn nhiều lớp hình thành - giống như các lớp địa chất, trong đó các hệ thống mới chồng lên các hệ thống cũ mà không có sự tích hợp và cấu trúc sâu sắc. Kết quả là xuất hiện các "silo" thông tin rời rạc, cản trở việc lưu thông kiến thức và gây khó khăn cho việc quản lý dữ liệu.

Các bước tiếp theo: từ lý thuyết dữ liệu đến thay đổi thực tiễn

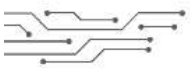
Sự tiến hóa của dữ liệu trong xây dựng là con đường từ các bảng đất sét đến các nền tảng mô-đun hiện đại. Ngày nay, thách thức không nằm ở việc thu thập thông tin, mà là tạo ra một cấu trúc biến các dữ liệu rời rạc và đa dạng thành một nguồn tài nguyên chiến lược. Bất kể vai trò của bạn - giám đốc công ty hay kỹ sư bình thường - hiểu giá trị của dữ liệu và khả năng làm việc với chúng trong tương lai sẽ trở thành một kỹ năng nghề nghiệp then chốt.

Tóm lại phần này, cần nêu rõ các bước thực tiễn chính sẽ giúp áp dụng các phương pháp đã thảo luận vào các nhiệm vụ hàng ngày của bạn:

- Thực hiện kiểm toán cá nhân về các luồng thông tin.
    - Lập danh sách tất cả các hệ thống và ứng dụng mà bạn làm việc hàng ngày.
    - Đánh dấu nơi bạn tốn nhiều thời gian nhất để tìm kiếm hoặc kiểm tra lại dữ liệu.
    - Xác định các nguồn thông tin chính của bạn.
    - Phân tích cảnh quan ứng dụng hiện tại của bạn về sự dư thừa và trùng lặp chức năng.
- Cố gắng tiến bộ trong các quy trình theo các cấp độ trưởng thành phân tích.
    - Bắt đầu công việc của bạn với các nhiệm vụ từ phân tích mô tả (điều gì đã xảy ra?).
    - Dần dần triển khai phân tích chẩn đoán (tại sao điều này xảy ra?).
    - Suy nghĩ về cách bạn có thể chuyển sang phân tích dự đoán (điều gì sẽ xảy ra?) và phân tích quy định (phải làm gì?).
- Bắt đầu cấu trúc dữ liệu công việc của bạn.
    - Triển khai một hệ thống đặt tên tệp và thư mục thống nhất mà bạn thường sử dụng trong công việc của mình.
    - Tạo mẫu cho các tài liệu và báo cáo thường xuyên sử dụng.
    - Thường xuyên lưu trữ các dự án đã hoàn thành với cấu trúc rõ ràng

Ngay cả khi bạn không thể thay đổi toàn bộ hạ tầng thông tin trong nhóm hoặc công ty của mình, hãy bắt đầu từ các quy trình của riêng bạn và những cải tiến nhỏ trong công việc hàng ngày. Hãy nhớ rằng giá trị thực sự của dữ liệu không nằm ở khối lượng của nó, mà ở khả năng khai thác lợi ích thực tiễn từ chúng. Ngay cả những tập hợp thông tin nhỏ nhưng được cấu trúc và phân tích đúng cách cũng có thể mang lại hiệu quả đáng kể nếu chúng được tích hợp vào quy trình ra quyết định.

Trong các phần tiếp theo của cuốn sách, chúng ta sẽ chuyển sang các phương pháp và công cụ cụ thể để làm việc với dữ liệu, xem xét các cách chuyển đổi thông tin không cấu trúc thành các tập hợp có cấu trúc, nghiên cứu công nghệ tự động hóa phân tích và phân tích chi tiết cách xây dựng một hệ sinh thái phân tích hiệu quả trong công ty xây dựng.





II PHẦN **NGÀNH XÂY DỰNG VÀ SỰ LÀM CÔN KHI DỮ LIỆU MÃNG**

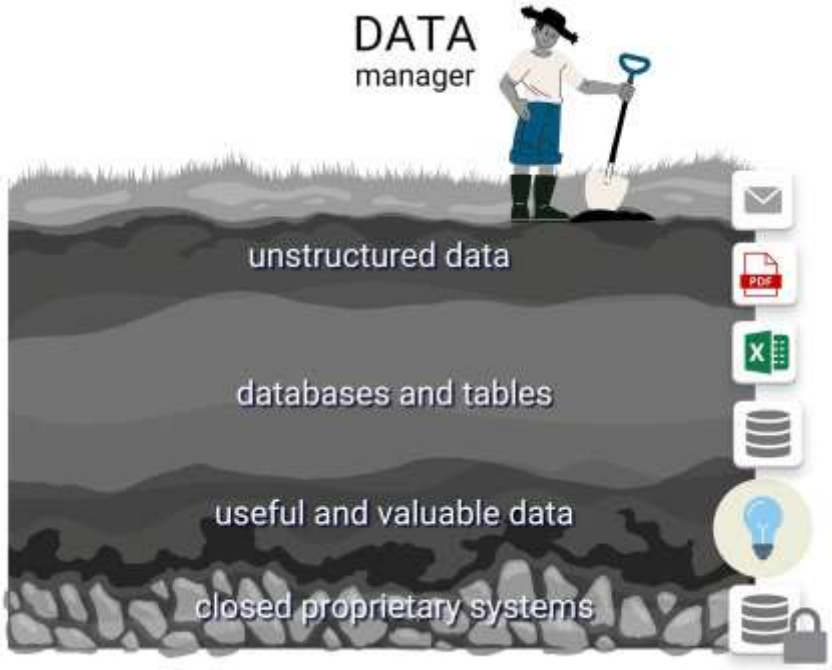
Phần thứ hai tập trung vào phân tích phê phán các vấn đề mà các công ty xây dựng phải đối mặt khi làm việc với khối lượng dữ liệu ngày càng tăng. Các hậu quả của việc phân mảnh thông tin và hiện tượng "dữ liệu trong silo" cản trở việc ra quyết định hiệu quả được xem xét chi tiết. Vấn đề HiPPO (Ý kiến của Người được trả lương cao nhất) và ảnh hưởng của nó đến chất lượng quyết định quản lý trong các dự án xây dựng cũng được nghiên cứu. Tác động của tính động của quy trình kinh doanh và sự phức tạp ngày càng tăng của chúng đối với dòng thông tin và hiệu quả hoạt động được đánh giá. Các ví dụ cụ thể về cách mà sự phức tạp dư thừa của hệ thống làm tăng chi phí và giảm tính linh hoạt của tổ chức được đưa ra. Sự chú ý đặc biệt được dành cho các hạn chế do các định dạng độc quyền tạo ra và triển vọng sử dụng các tiêu chuẩn mở trong ngành xây dựng. Khái niệm chuyển đổi sang các hệ sinh thái phần mềm dựa trên AI và LLM, giúp giảm thiểu sự phức tạp dư thừa và rào cản kỹ thuật, được trình bày.

CHƯƠNG 2.1. **PHÂN MẢNG VÀ SILO DỮ LIỆU**

Càng nhiều công cụ, doanh nghiệp càng hiệu quả?

Thoạt nhìn, có thể có vẻ như việc tăng số lượng công cụ kỹ thuật số dẫn đến hiệu quả cao hơn. Tuy nhiên, trên thực tế, mọi thứ lại không như vậy. Với mỗi giải pháp mới, cho dù là dịch vụ đám mây, hệ thống lỗi thời hay báo cáo Excel mới, công ty lại thêm một lớp vào cảnh quan kỹ thuật số của mình — một lớp mà thường không tích hợp với các lớp khác.

Dữ liệu có thể được so sánh với than hoặc dầu mỏ: chúng hình thành qua nhiều năm, "nén" dưới các lớp hỗn loạn, lỗi lầm, quy trình không cấu trúc và các định dạng bị lãng quên. Để khai thác thông tin thực sự hữu ích từ chúng, các công ty phải thực sự vượt qua những đống giải pháp lỗi thời và tiếng ồn kỹ thuật số.



Hình 2.11 Dữ liệu đa định dạng tạo thành các lớp tách biệt — ngay cả những "thông tin vàng" cũng bị mất trong các lớp địa chất của sự phức tạp hệ thống.

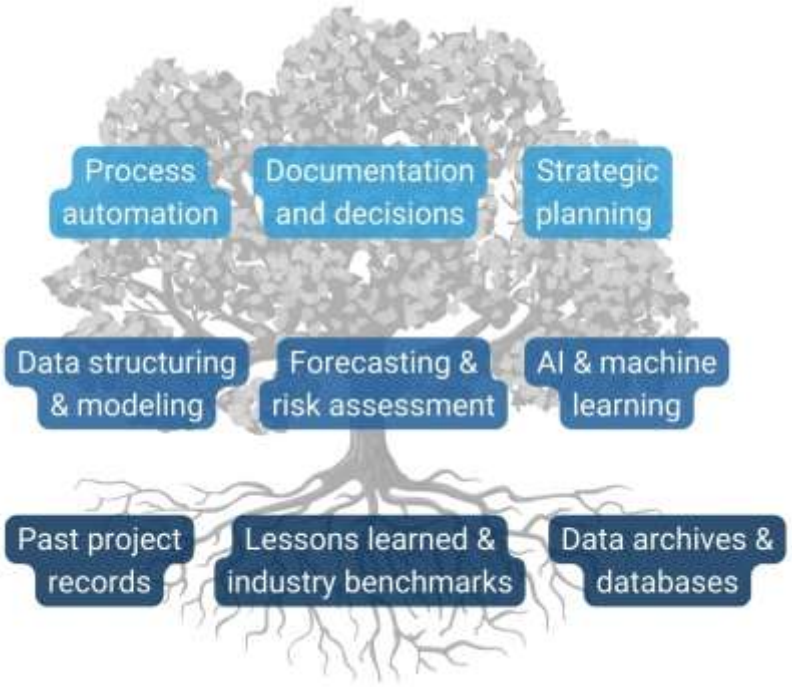
Mỗi ứng dụng mới để lại một dấu vết: tệp, bảng hoặc một "silo" hoàn toàn tách biệt trên máy chủ. Một lớp là đất sét (dữ liệu lỗi thời và bị lãng quên), lớp khác là cát (các bảng và báo cáo rời rạc), lớp thứ ba là granit (các định dạng độc quyền không thể tích hợp). Theo thời gian, môi trường kỹ thuật số của công ty ngày càng giống như một kho lưu trữ lớp với sự tích lũy thông tin không kiểm soát, nơi giá trị bị mất trong sâu thẳm của các máy chủ công ty.

Với mỗi dự án mới và mỗi hệ thống mới, không chỉ hạ tầng trở nên phức tạp hơn mà con đường đến với dữ liệu chất lượng hữu ích cũng vậy. Để có thể khai thác được "mỏ quý", cần phải thực hiện việc làm sạch sâu, cấu trúc thông tin, "phân mảnh", nhóm lại thành các đoạn có ý nghĩa và trích xuất những thông tin chiến lược quan trọng thông qua phân tích và mô hình hóa dữ liệu.

> *Dữ liệu là một tài sản quý giá, và chúng sẽ tồn tại lâu hơn chính các hệ thống [xử lý dữ liệu].*
>
> *— Tim Berners-Lee, cha đẻ của World Wide Web và người sáng tạo ra trang web đầu tiên*

Trước khi dữ liệu có thể trở thành "tài sản quý giá" và nền tảng đáng tin cậy cho việc ra quyết định, chúng cần phải trải qua quá trình chuẩn bị kỹ lưỡng. Chính việc tiền xử lý hợp lý biến những thông tin rời rạc thành kinh nghiệm có cấu trúc, là lớp thông tin hữu ích, từ đó trở thành công cụ dự đoán và tối ưu hóa.

> Có một hiểu lầm rằng để bắt đầu phân tích, cần phải có dữ liệu hoàn toàn sạch, tuy nhiên trên thực tế, khả năng làm việc với dữ liệu "bẩn" là một phần không thể thiếu của quá trình.



Hình 2.12 Dữ liệu là hệ thống gốc và nền tảng của doanh nghiệp, mà từ đó dựa vào các quy trình ra quyết định.

Khi công nghệ không ngừng phát triển, doanh nghiệp của bạn cũng phải tiến lên và học cách tạo ra giá trị từ dữ liệu. Giống như các công ty dầu mỏ và than đá xây dựng hạ tầng để khai thác khoáng sản, doanh nghiệp cũng cần học cách tự quản lý một cách hợp lý các dòng thông tin mới trên các máy chủ của mình và trích xuất thông tin quý giá từ dữ liệu chưa được sử dụng, đa định dạng và lỗi thời, biến chúng thành tài nguyên chiến lược.

Việc tạo ra các mỏ (kho dữ liệu) là bước đầu tiên. Ngay cả những công cụ mạnh mẽ nhất cũng không giải quyết được vấn đề cô lập dữ liệu và dữ liệu đa định dạng, nếu các công ty vẫn tiếp tục hoạt động trong các hệ thống rời rạc. Khi dữ liệu tồn tại tách biệt với nhau, không giao thoa và không trao đổi thông tin, doanh

nghiệp sẽ gặp phải hiệu ứng "silô dữ liệu". Thay vì có một hạ tầng thống nhất và đồng bộ, các công ty buộc phải tiêu tốn tài nguyên để hợp nhất và đồng bộ hóa dữ liệu.

Silos dữ liệu và ảnh hưởng của chúng đến hiệu quả công ty

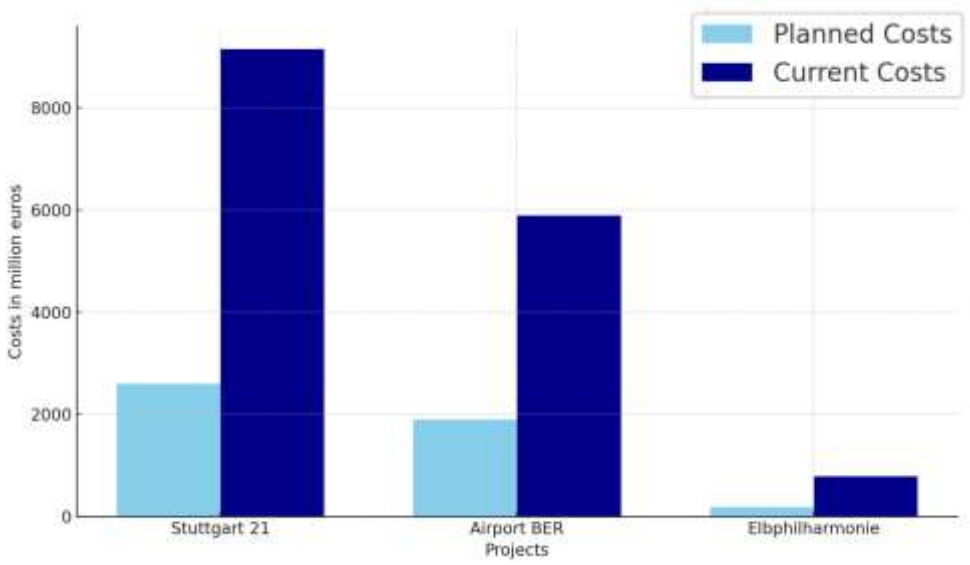
Hãy tưởng tượng rằng bạn đang xây dựng một khu chung cư, nhưng mỗi đội thi công đều có dự án riêng của mình. Một số xây dựng tường, một số lắp đặt hệ thống điện nước, một số khác làm đường, mà không tham khảo ý kiến lẫn nhau. Kết quả là ống nước không khớp với các lỗ trên tường, thang máy không phù hợp với số tầng, và đường phải được tháo dỡ và làm lại.

Tình huống như vậy không chỉ là kịch bản giả thuyết, mà là thực tế của nhiều dự án xây dựng hiện đại. Do số lượng lớn các nhà thầu chính và phụ làm việc với các hệ thống khác nhau và không có một trung tâm điều phối thống nhất, quá trình trở thành một chuỗi các cuộc thỏa thuận, sửa đổi và xung đột vô tận. Tất cả những điều này dẫn đến sự chậm trễ đáng kể và làm tăng chi phí dự án nhiều lần.

Tình huống điển hình khi có sự chậm trễ trên công trường xây dựng: khuôn đã sẵn sàng, nhưng việc cung cấp thép không đến đúng hạn. Khi kiểm tra thông tin trong các hệ thống khác nhau, giao tiếp diễn ra theo cách sau:

- Người giám sát trên công trường vào ngày 20 viết cho quản lý dự án: "Chúng tôi đã hoàn thành việc lắp đặt khuôn, còn thép đâu?"
- Quản lý dự án (PMIS) bộ phận cung ứng: — "Khuôn đã sẵn sàng. Trong hệ thống của tôi [PMIS] ghi rằng thép sẽ đến vào ngày 18. Thép đâu rồi?"
- Chuyên viên cung ứng (ERP): — "Trong ERP của chúng tôi ghi rằng lô hàng sẽ đến vào ngày 25."
- Kỹ sư dữ liệu hoặc bộ phận CNTT (chịu trách nhiệm về tích hợp): — Trong PMIS ghi ngày 18, trong ERP là ngày 25. Không có liên kết theo OrderID giữa ERP và PMIS, vì vậy dữ liệu không được đồng bộ hóa. Đây là một ví dụ điển hình về khoảng cách thông tin.
- Quản lý dự án với giám đốc điều hành — "Việc cung cấp thép bị trì hoãn, công trường đứng yên, và ai chịu trách nhiệm — không rõ."

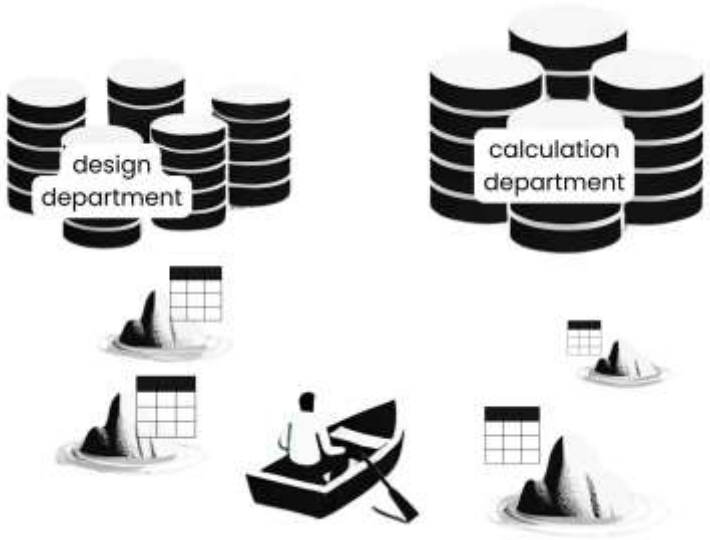
Nguyên nhân của sự cố là do sự cô lập dữ liệu trong các hệ thống rời rạc. Tích hợp và thống nhất các nguồn dữ liệu, tạo ra một kho thông tin duy nhất, cũng như tự động hóa thông qua các công cụ ETL (Apache NiFi, Airflow hoặc n8n) cho phép loại bỏ sự phân tán giữa các hệ thống. Những phương pháp và công cụ này sẽ được xem xét chi tiết trong các phần tiếp theo của cuốn sách.



Hình 2.13 So sánh chi phí dự kiến và thực tế cho các dự án hạ tầng lớn ở Đức.

Điều tương tự xảy ra với các hệ thống doanh nghiệp: trước tiên, các giải pháp cô lập được tạo ra, sau đó phải chi tiêu ngân sách khổng lồ cho việc tích hợp và đồng bộ hóa chúng. Nếu ngay từ đầu đã được suy nghĩ về các mô hình dữ liệu và mối quan hệ, nhu cầu tích hợp sẽ không phát sinh. Dữ liệu phân tán tạo ra sự hỗn loạn trong thế giới số, giống như một quá trình xây dựng không đồng bộ.

> Theo nghiên cứu của KPMG "Cue construction 4.0: Đã đến lúc làm hay phá" năm 2023, chỉ có 36% công ty trao đổi dữ liệu hiệu quả giữa các bộ phận, trong khi 61% gặp phải vấn đề nghiêm trọng do các "silô" dữ liệu cô lập [30].



Hình 2.14 Dữ liệu khó khai thác được thu thập trong nhiều năm tích tụ trong các kho lưu trữ "silô" cô lập, có nguy cơ không bao giờ được sử dụng.

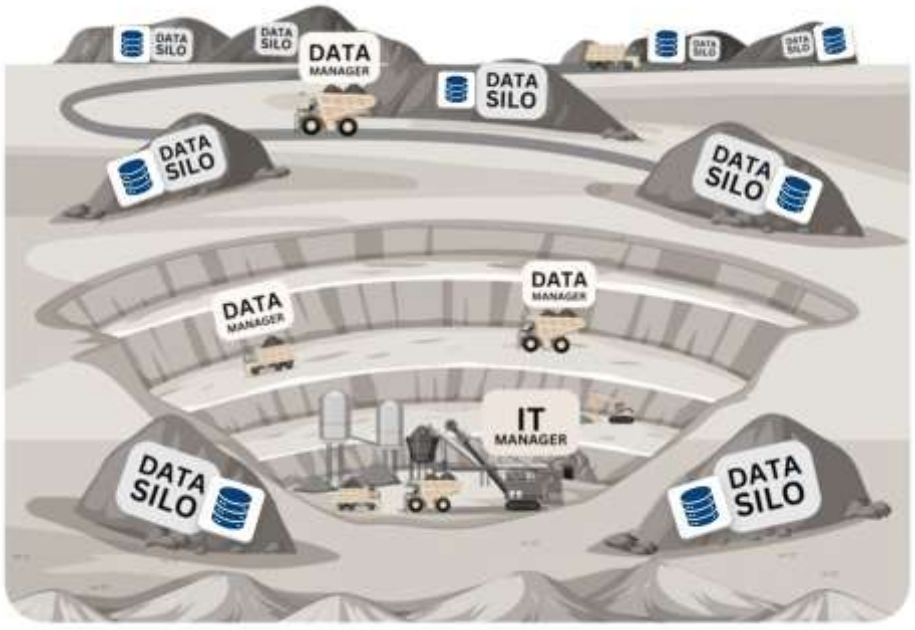
Dữ liệu của công ty được lưu trữ trong các hệ thống cô lập, giống như những cây riêng lẻ, rải rác trên cảnh quan. Mỗi cây chứa thông tin quý giá, nhưng sự thiếu kết nối giữa chúng cản trở việc tạo ra một hệ sinh thái duy nhất, liên kết. Sự phân tán này cản trở dòng chảy dữ liệu và hạn chế khả năng của tổ chức trong việc

nhìn thấy bức tranh toàn cảnh. Kết nối những silô này là một quá trình cực kỳ dài và phức tạp, giống như việc nuôi trồng nấm mốc ở cấp quản lý, nơi mà họ sẽ học cách truyền đạt các mảnh thông tin riêng lẻ giữa các hệ thống.

Theo nghiên cứu của WEF năm 2016, một trong những rào cản chính cho sự chuyển đổi số là sự thiếu hụt các tiêu chuẩn dữ liệu thống nhất và sự phân mảnh.

> *Ngành xây dựng là một trong những ngành phân mảnh nhất trên thế giới và phụ thuộc vào sự tương tác nhịp nhàng của tất cả các bên tham gia trong chuỗi giá trị [5].*
>
> *— Diễn đàn Kinh tế Thế giới 2016: Định hình Tương lai của Ngành Xây dựng.*

Các nhà thiết kế, quản lý, điều phối viên và nhà phát triển thường thích làm việc độc lập, tránh những phức tạp trong việc phối hợp. Xu hướng tự nhiên này dẫn đến việc tạo ra các "silô" thông tin, trong đó dữ liệu bị cô lập trong các hệ thống riêng biệt. Càng nhiều hệ thống bị cô lập như vậy, việc thiết lập sự tương tác giữa chúng càng trở nên khó khăn. Theo thời gian, mỗi hệ thống có cơ sở dữ liệu riêng và bộ phận hỗ trợ chuyên biệt từ các nhà quản lý, điều này càng làm phức tạp thêm việc tích hợp.



Mỗi hệ thống đều cố gắng tạo ra silô dữ liệu độc đáo của riêng mình, mà cần phải xử lý bằng các công cụ phù hợp.

Vòng luẩn quẩn trong các hệ thống doanh nghiệp diễn ra như sau: các công ty đầu tư vào các giải pháp phức tạp và cô lập, sau đó phải đối mặt với chi phí cao để tích hợp chúng, và các nhà phát triển, nhận thấy sự phức tạp trong việc kết hợp các hệ thống, thường thích làm việc trong các hệ sinh thái khép kín của riêng họ. Tất cả những điều này làm tăng sự phân tán trong cảnh quan CNTT và làm phức tạp việc chuyển đổi sang các giải pháp mới. Cuối cùng, các nhà quản lý phê bình sự phân tán của dữ liệu, nhưng hiếm khi phân tích nguyên nhân và cách ngăn chặn nó. Các lãnh đạo phàn nàn về các hệ thống CNTT lỗi thời, nhưng việc thay thế chúng đòi hỏi đầu tư đáng kể và hiếm khi mang lại kết quả như mong đợi. Kết quả là, ngay cả những nỗ lực để giải quyết vấn đề này cũng thường chỉ làm trầm trọng thêm tình hình.

Nguyên nhân chính của sự phân tán là ưu tiên ứng dụng hơn dữ liệu. Các công ty thường phát triển các hệ thống riêng lẻ hoặc mua các giải pháp có sẵn từ các nhà cung cấp, sau đó cố gắng kết hợp chúng, tạo ra các kho lưu trữ và cơ sở dữ liệu trùng lặp và không tương thích với nhau.

Để vượt qua vấn đề phân mảnh, cần một cách tiếp cận hoàn toàn mới - ưu tiên dữ liệu hơn ứng dụng. Các công ty nên phát triển các chiến lược quản lý dữ liệu và mô hình dữ liệu trước tiên, sau đó tạo ra các hệ thống hoặc mua các giải pháp hoạt động với một tập hợp thông tin thống nhất, thay vì tạo ra các rào cản mới.

> *Chúng ta đang bước vào một thế giới mới, nơi dữ liệu có thể quan trọng hơn phần mềm.*
>
> *- Tim O'Reilly, Giám đốc điều hành O'Reilly Media, Inc.*

> Nghiên cứu của Viện Toàn cầu McKinsey "Tái định hình xây dựng: con đường nâng cao năng suất" (2016) cho thấy ngành xây dựng đang tụt lại so với các lĩnh vực khác trong chuyển đổi số. Theo báo cáo, việc triển khai quản lý dữ liệu tự động và các nền tảng kỹ thuật số có thể nâng cao đáng kể năng suất và giảm thiểu tổn thất liên quan đến sự không nhất quán trong quy trình. Nhu cầu chuyển đổi số này cũng được báo cáo của Egan (Vương quốc Anh, 1998) nhấn mạnh, trong đó nêu bật vai trò quan trọng của các quy trình tích hợp và cách tiếp cận hợp tác trong xây dựng.

> Cuối cùng, nếu trong 10.000 năm qua, vấn đề chính của các nhà quản lý dữ liệu là thiếu dữ liệu, thì với sự gia tăng nhanh chóng về số lượng dữ liệu và hệ thống quản lý dữ liệu, người dùng và các nhà quản lý đã phải đối mặt với vấn đề - thừa dữ liệu, gây khó khăn trong việc tìm kiếm thông tin hợp pháp và chất lượng.

Sự phân tán của các silô dữ liệu không thể tránh khỏi dẫn đến một vấn đề nghiêm trọng là giảm chất lượng dữ liệu. Khi có nhiều hệ thống độc lập, cùng một dữ liệu có thể tồn tại ở nhiều phiên bản khác nhau, thường với các giá trị mâu thuẫn, điều này tạo ra thêm khó khăn cho người dùng, những người cần xác định thông tin nào là hiện tại và đáng tin cậy.

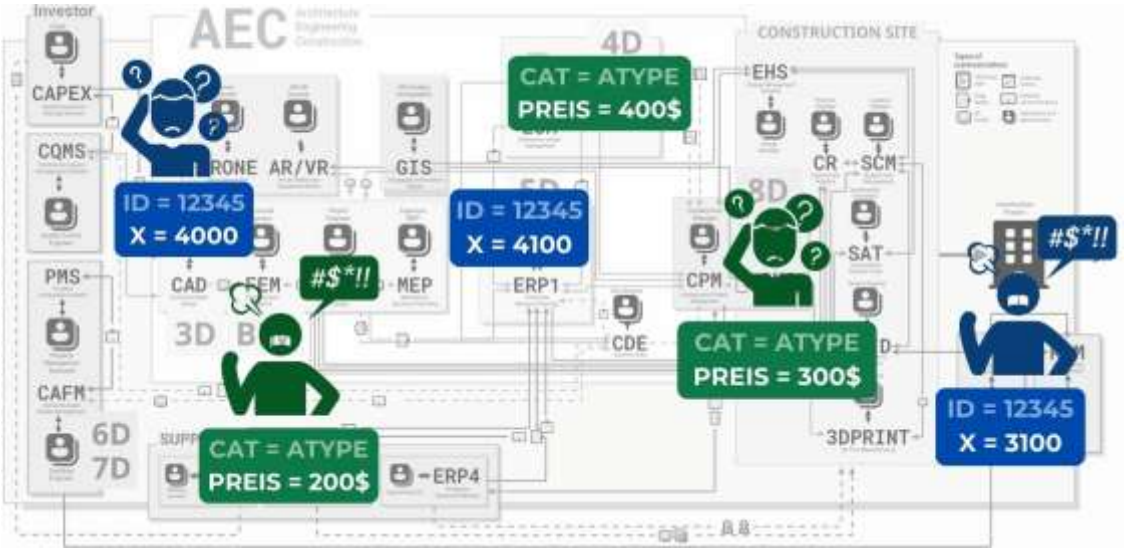
Trùng lặp và thiếu chất lượng dữ liệu như hệ quả của sự phân mảnh

Do vấn đề silo dữ liệu, các nhà quản lý buộc phải dành một khoảng thời gian đáng kể để tìm kiếm và xác minh dữ liệu. Để bảo vệ khỏi các vấn đề về chất lượng, các công ty tạo ra các cấu trúc quản lý thông tin phức tạp, trong đó các cấp quản lý chịu trách nhiệm tìm kiếm, xác thực và đồng bộ hóa dữ liệu. Tuy nhiên, cách tiếp cận này chỉ làm tăng thêm tính quan liêu và làm chậm quá trình ra quyết định. Càng nhiều dữ liệu, việc phân tích và diễn giải càng trở nên khó khăn, đặc biệt khi không có tiêu chuẩn thống nhất về lưu trữ và xử lý chúng.

Với sự xuất hiện của nhiều ứng dụng phần mềm và hệ thống, mọc lên như nấm sau mưa trong thập kỷ qua, vấn đề silo và chất lượng dữ liệu không phù hợp ngày càng trở nên cấp bách đối với người dùng cuối. Cùng một dữ liệu, nhưng với các giá trị khác nhau, giờ đây có thể tìm thấy trong các hệ thống và ứng dụng khác nhau. Điều này dẫn đến khó khăn cho người dùng cuối khi họ cố gắng xác định phiên bản dữ liệu nào là chính xác và hợp lệ trong số nhiều tùy chọn có sẵn. Điều này dẫn đến sai sót trong phân tích và cuối cùng

là trong quá trình ra quyết định.-

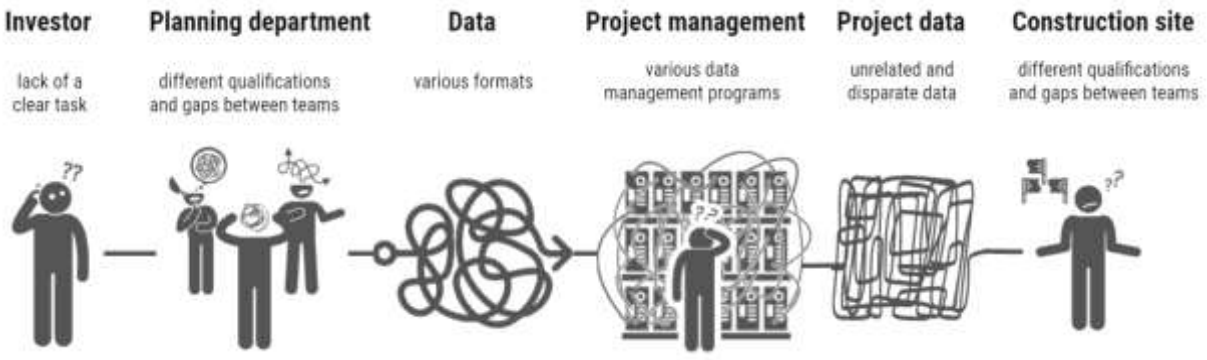
Để bảo vệ khỏi các vấn đề tìm kiếm dữ liệu cần thiết, các nhà lãnh đạo công ty tạo ra một hệ thống quan liêu nhiều cấp từ các nhà quản lý kiểm tra. Nhiệm vụ của họ là nhanh chóng tìm kiếm, xác minh và chuyển tiếp dữ liệu cần thiết dưới dạng bảng và báo cáo, điều hướng trong mê cung của các hệ thống rời rạc.



Khi cố gắng tìm dữ liệu cần thiết, các nhà quản lý phải đảm bảo chất lượng và tính hợp pháp của dữ liệu giữa các hệ thống khác nhau.

Tuy nhiên, trên thực tế, mô hình này tạo ra những phức tạp mới. Khi việc quản lý dữ liệu được thực hiện thủ công và thông tin bị phân tán qua nhiều giải pháp không liên kết với nhau, mỗi nỗ lực để có được thông tin chính xác và cập nhật thông qua một chuỗi các cá nhân có trách nhiệm trở thành một điểm nghẽn - tốn thời gian và dễ mắc lỗi.

Tình hình trở nên tồi tệ hơn với sự gia tăng nhanh chóng số lượng giải pháp kỹ thuật số. Thị trường phần mềm tiếp tục được bổ sung bằng các công cụ mới, có vẻ hứa hẹn. Nhưng nếu không có một chiến lược rõ ràng về quản lý dữ liệu, những giải pháp này sẽ không được tích hợp vào một hệ thống thống nhất, mà ngược lại - tạo ra thêm các lớp phức tạp và trùng lặp. Kết quả là, thay vì đơn giản hóa các quy trình, các công ty lại rơi vào một môi trường thông tin còn phân mảnh và hỗn loạn hơn.



Sự phức tạp của các hệ thống và sự đa dạng của các định dạng dữ liệu dẫn đến mất tính nhất quán trong quá trình xây dựng.

Tất cả các vấn đề đã nêu liên quan đến việc quản lý nhiều giải pháp rời rạc, sớm hay muộn sẽ dẫn đến sự nhận thức quan trọng của ban lãnh đạo công ty: vấn đề không nằm ở khối lượng dữ liệu và không phải là việc tìm kiếm một công cụ "đại chúng" mới để xử lý chúng. Nguyên nhân thực sự nằm ở chất lượng dữ liệu và cách mà tổ chức tạo ra, nhận được, lưu trữ và sử dụng chúng.

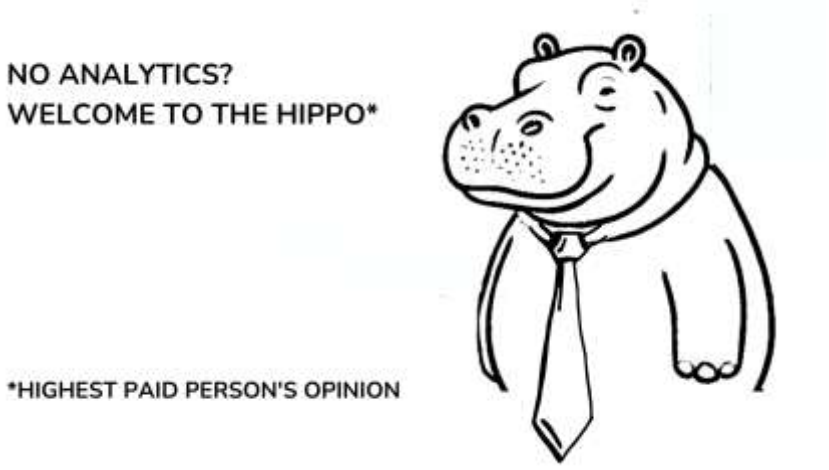
> Chìa khóa cho thành công bền vững không phải là việc theo đuổi các ứng dụng "thần kỳ" mới, mà là việc hình thành một văn hóa làm việc với dữ liệu bên trong công ty. Điều này có nghĩa là dữ liệu được coi là một tài sản chiến lược, và các vấn đề về chất lượng, tính toàn vẹn và tính cập nhật của chúng trở thành ưu tiên ở tất cả các cấp độ của tổ chức.

Giải quyết vấn đề "chất lượng so với số lượng" nằm ở việc tạo ra một cấu trúc dữ liệu thống nhất, loại bỏ sự trùng lặp, xóa bỏ mâu thuẫn và kết hợp các luồng thông tin. Kiến trúc như vậy cho phép hình thành một nguồn dữ liệu duy nhất, đáng tin cậy, trên cơ sở đó đưa ra các quyết định hợp lý, chính xác và kịp thời.

Ngược lại, như vẫn thường xảy ra, các công ty tiếp tục dựa vào ý kiến chủ quan và đánh giá trực giác của các chuyên gia HiPPO, thay vì dựa vào các sự kiện đáng tin cậy. Trong ngành xây dựng, nơi mà kinh nghiệm chuyên gia thường đóng vai trò quan trọng, điều này càng rõ ràng hơn.

HiPPO hay nguy cơ ý kiến trong việc ra quyết định

Truyền thống trong ngành xây dựng, các quyết định chính thường được đưa ra dựa trên kinh nghiệm và đánh giá chủ quan. Thiếu dữ liệu kịp thời và đáng tin cậy, các nhà lãnh đạo công ty phải hành động trong bóng tối, dựa vào trực giác của những nhân viên có mức lương cao nhất (HiPPO - Ý kiến của Người có thu nhập cao nhất), thay vì dựa vào các sự kiện khách quan.-



Hình 2.18 Trong sự thiếu hụt phân tích, doanh nghiệp phụ thuộc vào ý kiến chủ quan của các chuyên gia có kinh nghiệm.

Cách tiếp cận này có thể được biện minh trong điều kiện ổn định và thay đổi chậm, nhưng trong thời đại chuyển đổi số, nó trở thành một rủi ro nghiêm trọng. Các quyết định dựa trên trực giác và phỏng đoán dễ bị sai lệch, thường được xây dựng trên các giả thuyết chưa được xác nhận và không xem xét bức tranh tổng thể được phản ánh trong dữ liệu.

> Những gì được coi là các cuộc tranh luận hợp lý trong công ty thường không dựa trên bất cứ điều gì cụ thể. Sự thành công của công ty không nên phụ thuộc vào uy tín và mức lương của các chuyên gia, mà nên được xác định bởi khả năng làm việc hiệu quả với dữ liệu, phát hiện các quy luật và đưa ra các quyết định cân nhắc.

Quan trọng là từ bỏ khái niệm mà theo đó uy tín hoặc kinh nghiệm tự động có nghĩa là quyết định đúng đắn. Cách tiếp cận dựa trên dữ liệu thay đổi các quy tắc trò chơi: giờ đây, dữ liệu và phân tích trở thành nền tảng cho việc đưa ra quyết định, chứ không phải chức vụ và mức lương. Dữ liệu lớn, học máy và phân tích trực quan cho phép phát hiện các quy luật và dựa vào sự thật, chứ không phải phỏng đoán.-

> *Không có dữ liệu, bạn chỉ là một người có ý kiến.*
>
> *- W. Edwards Deming, nhà khoa học và tư vấn quản lý*

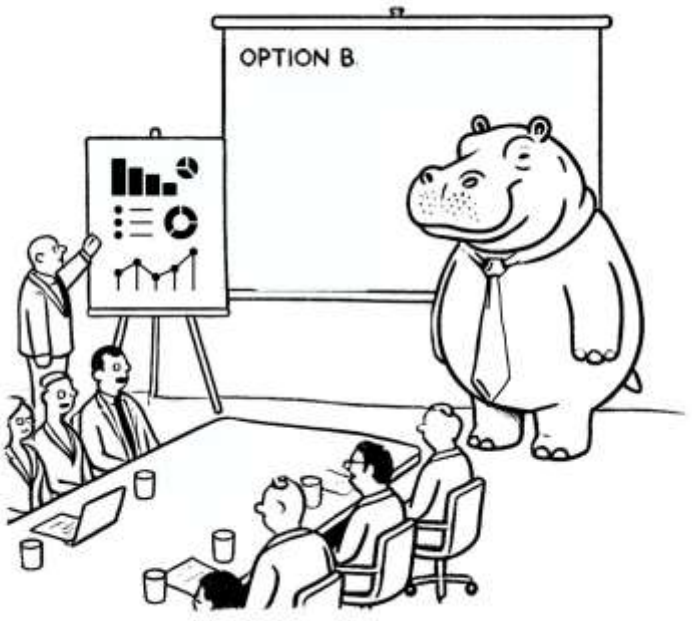
Các phương pháp quản lý dữ liệu hiện đại cũng đảm bảo tính kế thừa của kiến thức trong công ty. Các quy trình được mô tả rõ ràng, tự động hóa và cách tiếp cận hệ thống cho phép chuyển giao ngay cả các vai trò quan trọng mà không làm mất hiệu quả.

Tuy nhiên, sự tin tưởng mù quáng vào dữ liệu cũng có thể dẫn đến những sai lầm nghiêm trọng. Dữ liệu tự nó chỉ là một tập hợp các con số. Nếu không có phân tích hợp lý, bối cảnh và khả năng phát hiện các quy luật, chúng không có giá trị và không thể quản lý các quy trình. Chìa khóa cho sự thành công không nằm ở việc lựa chọn giữa trực giác HiPPO và phân tích, mà là xây dựng các công cụ thông minh, biến thông tin rời rạc thành các quyết định có thể quản lý và có cơ sở.

> Trong điều kiện xây dựng số, các yếu tố quyết định thành công không phải là thâm niên làm việc và vị trí trong hệ thống, mà là tốc độ phản ứng, độ chính xác của các quyết định và hiệu quả sử dụng tài nguyên.

Dữ liệu là công cụ, chứ không phải chân lý tuyệt đối. Chúng nên bổ sung cho tư duy con người, chứ không thay thế nó. Mặc dù có nhiều lợi ích từ phân tích, dữ liệu không thể hoàn toàn thay thế trực giác và kinh nghiệm của con người. Vai trò của chúng là giúp đưa ra quyết định chính xác và có ý thức hơn.

Lợi thế cạnh tranh sẽ không chỉ đạt được bằng cách tuân thủ các tiêu chuẩn, mà còn bằng khả năng vượt qua đối thủ trong việc sử dụng hiệu quả các nguồn lực giống nhau. Trong tương lai, kỹ năng làm việc với dữ liệu sẽ trở nên quan trọng như từng có chữ viết hoặc khả năng toán học. Các chuyên gia biết phân tích và diễn giải dữ liệu sẽ có thể đưa ra quyết định chính xác hơn, thay thế những người chỉ dựa vào kinh nghiệm cá nhân.



Quyết định nên dựa trên phân tích khách quan, chứ không phải ý kiến của nhân viên có mức lương cao nhất.

Các nhà quản lý, chuyên gia và kỹ sư sẽ đóng vai trò như các nhà phân tích dữ liệu, nghiên cứu cấu trúc, động lực và các chỉ số chính của dự án. Tài nguyên con người sẽ trở thành các yếu tố trong hệ thống, yêu cầu điều chỉnh linh hoạt dựa trên dữ liệu để đạt được hiệu quả tối đa.

> *Sai sót khi sử dụng dữ liệu không phù hợp ít hơn nhiều so với việc không có dữ liệu.– Charles Babbage, nhà phát minh máy tính phân tích đầu tiên.*

Sự xuất hiện của dữ liệu lớn và việc triển khai các Mô hình Ngôn ngữ Lớn (LLM) đã thay đổi một cách triệt để không chỉ cách phân tích mà còn cả bản chất của việc ra quyết định. Nếu trước đây, trọng tâm là nguyên nhân (tại sao điều gì đó xảy ra - phân tích chẩn đoán), thì ngày nay, khả năng dự đoán tương lai (phân tích dự đoán) trở thành ưu tiên hàng đầu, và trong tương lai, phân tích chỉ định, nơi mà máy học và AI gợi ý lựa chọn tốt nhất trong quá trình ra quyết định.-

> Theo một nghiên cứu mới của SAP™ "Nghiên cứu mới cho thấy gần một nửa các nhà lãnh đạo tin tưởng vào trí tuệ nhân tạo hơn cả bản thân họ" vào năm 2025, 44% các nhà lãnh đạo cấp cao sẵn sàng thay đổi quyết định trước đó dựa trên khuyến nghị của AI, và 38% sẽ giao cho AI đưa ra quyết định kinh doanh thay mặt họ. Trong khi đó, 74% các nhà lãnh đạo cho biết họ tin tưởng vào lời khuyên của AI hơn cả bạn bè và gia đình của mình, và 55% làm việc trong các công ty mà những hiểu biết thu được từ AI thay thế hoặc thường xuyên vượt qua các phương pháp ra quyết định truyền thống - đặc biệt là trong các tổ chức có doanh thu hàng năm trên 5 tỷ đô la. Hơn nữa, 48% người được khảo sát sử dụng các công cụ AI sinh tạo hàng ngày, trong đó 15% sử dụng nhiều lần trong ngày.

Với sự phát triển của LLM và các hệ thống quản lý dữ liệu tự động, một vấn đề mới nổi lên: làm thế nào để sử dụng thông tin một cách hiệu quả mà không làm mất giá trị của nó trong sự hỗn loạn của các định dạng không tương thích và các nguồn khác nhau, điều này càng được bổ sung bởi sự phức tạp và động lực ngày càng tăng của các quy trình kinh doanh.

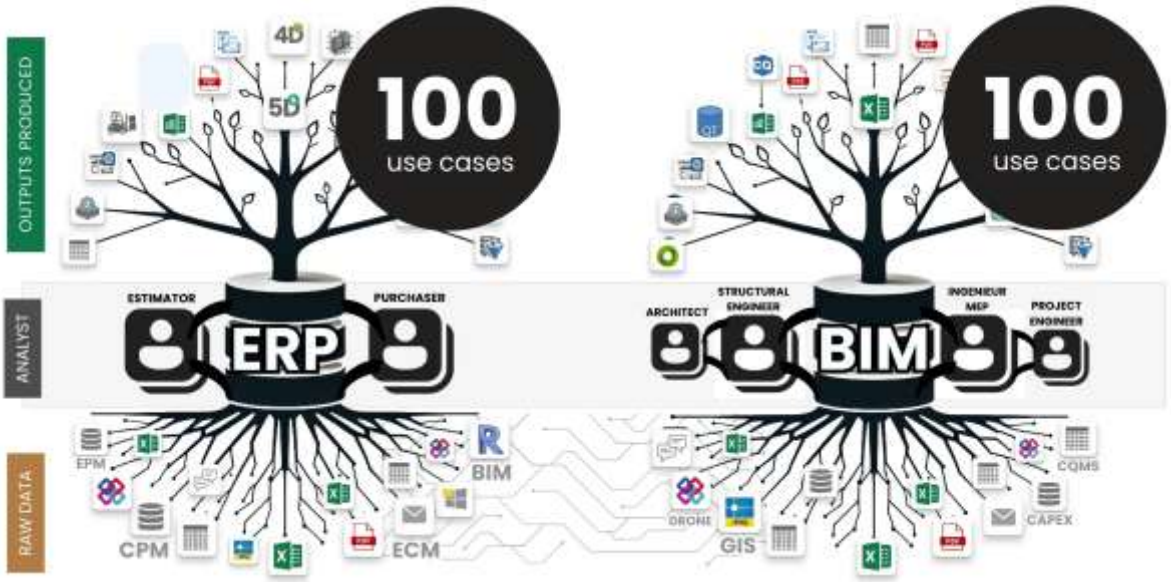
Sự gia tăng liên tục về độ phức tạp và tính động của quy trình kinh doanh

Ngành xây dựng hiện nay đang đối mặt với những thách thức nghiêm trọng trong việc quản lý dữ liệu và quy trình. Những khó khăn chính bao gồm sự phân tán của các hệ thống thông tin, sự quan liêu quá mức và thiếu tích hợp giữa các công cụ số. Những vấn đề này càng trở nên nghiêm trọng khi các quy trình kinh doanh ngày càng phức tạp hơn - dưới tác động của công nghệ, yêu cầu thay đổi từ khách hàng và các quy định mới.

Sự độc đáo của các dự án xây dựng không chỉ nằm ở những đặc điểm kỹ thuật mà còn ở sự khác biệt trong các tiêu chuẩn quốc gia và yêu cầu quy định của các quốc gia khác nhau. Điều này đòi hỏi một cách tiếp cận linh hoạt, cá nhân hóa cho từng dự án, điều mà rất khó thực hiện trong khuôn khổ các hệ thống quản lý mô-đun truyền thống. Do sự phức tạp của các quy trình và khối lượng dữ liệu lớn, nhiều công ty đã tìm đến các nhà cung cấp, những người cung cấp các giải pháp chuyên biệt. Tuy nhiên, thị trường đang bị quá tải - nhiều startup cung cấp các sản phẩm tương tự, tập trung vào các nhiệm vụ hẹp. Kết quả là, thường mất đi cách tiếp cận toàn diện trong việc quản lý dữ liệu.-

> Sự thích ứng với dòng chảy liên tục của các công nghệ mới và yêu cầu của thị trường trở thành yếu tố quan trọng trong khả năng cạnh tranh. Tuy nhiên, các ứng dụng sở hữu và hệ thống mô-đun hiện có có tính thích ứng thấp - bất kỳ thay đổi nào cũng thường yêu cầu những điều chỉnh kéo dài và tốn kém từ phía các nhà phát triển, những người không phải lúc nào cũng hiểu rõ đặc thù của các quy trình xây dựng.

Các công ty trở thành nạn nhân của sự lạc hậu công nghệ, chờ đợi các bản cập nhật mới thay vì nhanh chóng triển khai các phương pháp tích hợp đổi mới. Cuối cùng, cấu trúc nội bộ của các tổ chức xây dựng thường là một hệ sinh thái phức tạp của các hệ thống liên kết với nhau, thường là khép kín, mà sự phối hợp giữa chúng được thực hiện thông qua một mạng lưới quản lý nhiều cấp.

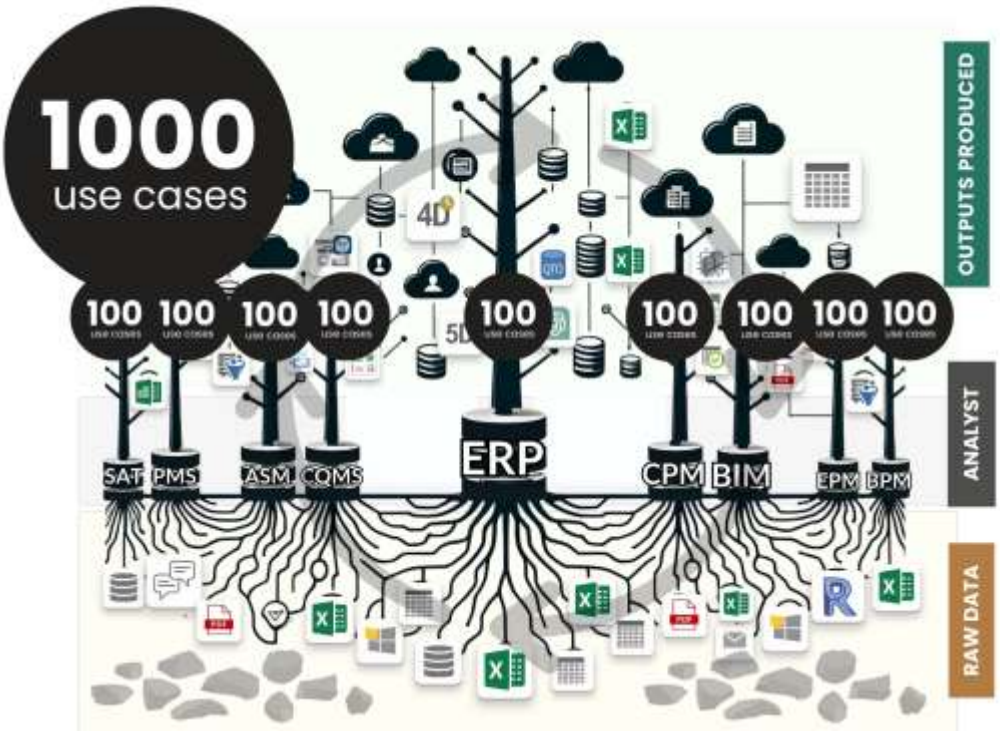


Các công ty bao gồm các hệ thống liên kết, sự kết hợp của chúng tạo ra các quy trình cần được tự động hóa.

> Theo một nghiên cứu được thực hiện bởi Hiệp hội Xây dựng Canada và công ty KPMG tại Canada vào năm 2021, chỉ có 25% các công ty cho rằng họ đang ở vị trí đáng kể hoặc xuất sắc so với các đối thủ trong việc triển khai công nghệ hoặc giải pháp số. Chỉ có 23% người tham gia khảo sát cho biết các giải pháp của họ chủ yếu hoặc rất lớn dựa trên dữ liệu. Trong khi đó, phần lớn người tham gia khảo sát mô tả việc sử dụng một số công nghệ khác là hoàn toàn thử nghiệm hoặc thừa nhận rằng họ không áp dụng chúng.

Sự không muốn tham gia vào các thử nghiệm công nghệ đặc biệt rõ ràng trong các dự án hạ tầng lớn, nơi mà sai sót có thể tốn hàng triệu đô la. Ngay cả những công nghệ tiên tiến nhất - như mô hình số, phân tích dự đoán - thường gặp phải sự kháng cự không phải vì hiệu quả của chúng, mà do thiếu độ tin cậy đã được chứng minh trong các dự án thực tế.

> Theo báo cáo của Diễn đàn Kinh tế Thế giới (WEF) "Hình thành tương lai của ngành xây dựng", việc áp dụng công nghệ mới trong xây dựng không chỉ gặp phải những khó khăn kỹ thuật mà còn cả rào cản tâm lý từ phía khách hàng. Nhiều khách hàng lo ngại rằng việc sử dụng các giải pháp tiên tiến sẽ biến dự án của họ thành một sân chơi thử nghiệm và khiến họ trở thành "chuột thí nghiệm", trong khi những hậu quả không lường trước có thể dẫn đến chi phí và rủi ro bổ sung.



Hình 2.111 Đối với mỗi trường hợp sử dụng dữ liệu, thị trường giải pháp cung cấp các ứng dụng để tối ưu hóa và tự động hóa quy trình.

> Ngành xây dựng rất đa dạng: các dự án khác nhau có yêu cầu khác nhau, đặc điểm khu vực, quy định pháp lý về phân loại, tiêu chuẩn tính toán, v.v. Do đó, gần như không thể tạo ra một ứng dụng hoặc hệ thống độc quyền phổ quát nào hoàn toàn phù hợp với tất cả những yêu cầu và đặc điểm của các dự án.--

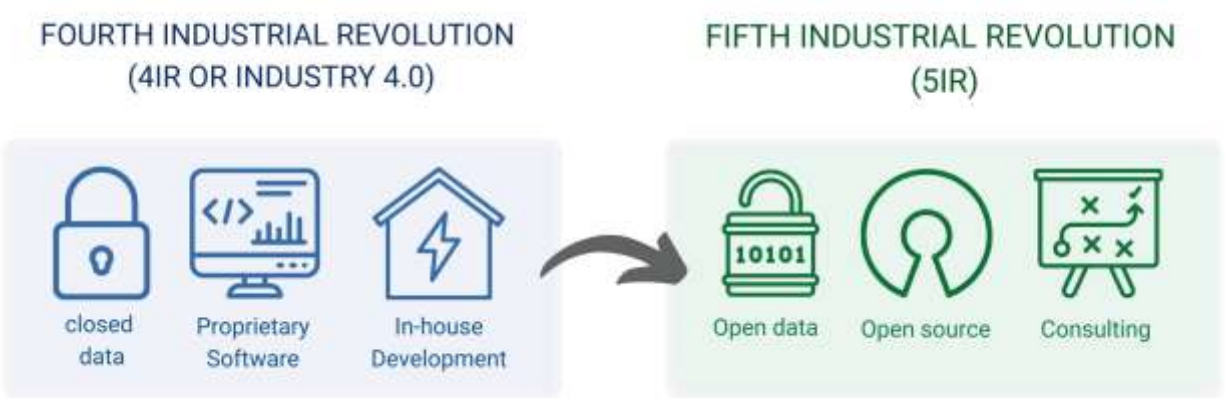
Để đối phó với sự phức tạp ngày càng tăng của các hệ thống và sự phụ thuộc vào nhà cung cấp phần mềm, ngày càng nhiều người nhận ra rằng chìa khóa để quản lý dữ liệu hiệu quả không chỉ là sự minh bạch và tiêu chuẩn hóa, mà còn là việc đơn giản hóa chính kiến trúc quy trình. Sự phức tạp và tính năng động ngày càng tăng của các quy trình kinh doanh đòi hỏi những cách tiếp cận mới, trong đó ưu tiên chuyển từ việc tích lũy dữ liệu sang việc cấu trúc và sắp xếp chúng. Chính sự chuyển dịch này sẽ trở thành bước tiếp theo trong sự phát triển của ngành xây dựng, đánh dấu sự kết thúc của kỷ nguyên thống trị của các nhà cung cấp phần mềm và sự khởi đầu của kỷ nguyên tổ chức thông tin có ý nghĩa.

Nhận thức về sự hạn chế của các giải pháp phổ quát và sự dễ bị tổn thương trước sự gia tăng phức tạp dẫn đến sự thay đổi ưu tiên: từ các nền tảng đóng và tích lũy dữ liệu sang sự minh bạch, khả năng thích ứng và làm việc có cấu trúc với thông tin. Sự chuyển dịch trong tư duy này phản ánh những thay đổi rộng lớn hơn trong nền kinh tế và công nghệ toàn cầu, được mô tả qua lăng kính của các "cuộc cách mạng công nghiệp". Để hiểu ngành xây dựng đang đi về đâu và những định hướng tương lai của nó là gì, cần xem xét vị trí của ngành trong bối cảnh của Cuộc cách mạng công nghiệp lần thứ Tư và lần thứ Năm - từ tự động hóa và số hóa đến cá nhân hóa, tiêu chuẩn mở và mô hình dịch vụ dữ liệu.

Cách mạng công nghiệp lần thứ tư (Ngành 4.0) và lần thứ năm (Ngành 5.0) trong xây dựng

Các cấu trúc công nghệ và kinh tế là những khái niệm lý thuyết được sử dụng để mô tả và phân tích sự tiến hóa của xã hội và kinh tế qua các giai đoạn phát triển khác nhau. Tuy nhiên, các nhà nghiên cứu và chuyên gia khác nhau có thể diễn giải chúng theo những cách khác nhau.

- Cuộc cách mạng công nghiệp lần thứ tư (4IR hoặc Industry 4.0) liên quan đến công nghệ thông tin, tự động hóa, số hóa và toàn cầu hóa. Một trong những yếu tố chính của nó là việc tạo ra các giải pháp phần mềm độc quyền, tức là các sản phẩm kỹ thuật số chuyên biệt được phát triển cho các nhiệm vụ và công ty cụ thể. Những giải pháp này thường trở thành một phần quan trọng của hạ tầng CNTT, nhưng lại khó mở rộng mà không cần sửa đổi thêm.
- Cách mạng công nghiệp lần thứ năm (5IR) hiện đang ở giai đoạn khái niệm và phát triển sớm hơn so với 4IR. Các nguyên tắc chính của nó bao gồm việc nâng cao mức độ cá nhân hóa sản phẩm và dịch vụ. 5IR là một phong trào hướng tới hoạt động kinh tế linh hoạt, thích ứng và cá nhân hóa hơn, với trọng tâm là cá nhân hóa, tư vấn và các mô hình dịch vụ. Một khía cạnh quan trọng của nền kinh tế thứ năm là việc sử dụng dữ liệu để ra quyết định, điều này gần như không thể thực hiện được nếu không có việc áp dụng dữ liệu mở và các công cụ mở (Hình 2.112).-



Hình 2.112 Nền kinh tế thứ tư tập trung vào các giải pháp, trong khi nền kinh tế thứ năm tập trung vào cá nhân hóa và dữ liệu.

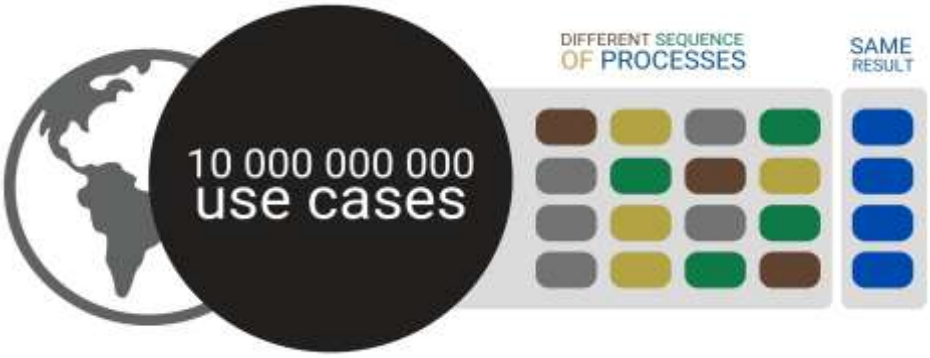
> Việc tạo ra một ứng dụng cho các công ty trong ngành xây dựng, nhằm sử dụng trong mười hoặc một trăm tổ chức, không đảm bảo rằng nó sẽ thành công trong việc mở rộng sang các công ty, khu vực hoặc quốc gia khác mà không có những sửa đổi và điều chỉnh đáng kể. Xác suất thành công trong việc mở rộng các giải pháp như vậy vẫn thấp, vì mỗi tổ chức có quy trình, yêu cầu và điều kiện độc đáo, có thể cần những điều chỉnh cá nhân hóa.

Quan trọng là phải hiểu rằng ngay cả bây giờ, việc tích hợp thành công các giải pháp công nghệ đòi hỏi một cách tiếp cận cá nhân hóa sâu sắc cho từng quy trình, dự án và công ty. Điều này có nghĩa là ngay cả sau khi phát triển một cấu trúc, công cụ hoặc chương trình chung, sẽ cần phải điều chỉnh và cấu hình chi tiết theo các yêu cầu và điều kiện độc đáo của từng công ty và dự án cụ thể.

> Theo báo cáo của PwC "Giải mã cách mạng công nghiệp lần thứ năm" [38], khoảng 50% các giám đốc điều hành cấp cao trong các ngành khác nhau trong năm nay đang đặt cược vào việc tích hợp công nghệ tiên tiến và kinh nghiệm con người. Cách tiếp cận này cho phép nhanh chóng thích ứng với những thay đổi trong thiết kế sản phẩm hoặc yêu cầu của khách hàng, tạo ra sản xuất cá nhân hóa.

Đối với mỗi quy trình, cần phát triển một chức năng hoặc ứng dụng độc đáo, điều này, với quy mô của ngành xây dựng toàn cầu và sự đa dạng của các dự án, dẫn đến sự tồn tại của một lượng lớn các trường hợp kinh doanh, mỗi trường hợp đều có logic Pipeline độc đáo (Hình 2.113). Mỗi trường hợp như vậy có những đặc điểm riêng và yêu cầu một cách tiếp cận cá nhân hóa. Chúng tôi sẽ xem xét chi tiết sự đa dạng của các giải pháp có thể cho cùng một nhiệm vụ phân tích trong bối cảnh các phương pháp khác nhau trong chương dành riêng cho học máy và phân tích tập dữ liệu "Titanic" (Hình 9.29). --

> Pipeline trong bối cảnh các quy trình kỹ thuật số là một chuỗi các hành động, quy trình và công cụ đảm bảo dòng dữ liệu và công việc tự động hóa hoặc có cấu trúc ở các giai đoạn khác nhau của vòng đời dự án.



Hình 2.113 Tính cá nhân và sự biến đổi của các trường hợp kinh doanh khiến cho việc tạo ra các nền tảng và công cụ đóng kín có thể mở rộng trở nên không khả thi.

Cuộc sống của chúng ta đã thay đổi rất nhiều dưới ảnh hưởng của chuyển đổi số, và hôm nay có thể nói về sự xuất hiện của một giai đoạn mới trong sự phát triển kinh tế của ngành xây dựng. Trong "nền kinh tế mới" này, sự cạnh tranh sẽ được tổ chức theo những quy tắc khác: những ai có khả năng chuyển đổi hiệu quả kiến thức công khai và dữ liệu mở thành các sản phẩm và dịch vụ được yêu cầu sẽ có lợi thế then chốt trong bối cảnh cách mạng công nghiệp lần thứ năm.

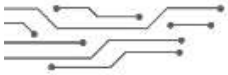
> Như nhà kinh tế Kate Maskus đã chỉ ra trong cuốn sách "Quyền sở hữu tư nhân và vấn đề công cộng: Kinh tế toàn cầu của sở hữu trí tuệ trong thế kỷ XXI" năm 2012, "chúng ta đang sống trong một nền kinh tế tri thức toàn cầu, và tương lai thuộc về những ai biết cách biến các phát hiện khoa học thành hàng hóa."

Sự chuyển mình sang mô hình kinh tế thứ năm đòi hỏi sự chuyển đổi từ các giải pháp CNTT khép kín sang các tiêu chuẩn và nền tảng mở. Các công ty sẽ bắt đầu từ bỏ các sản phẩm phần mềm truyền thống để chuyển sang các mô hình dịch vụ, trong đó tài sản chính sẽ là dữ liệu, chứ không phải công nghệ độc quyền.

> Nghiên cứu của Trường Kinh doanh Harvard năm 2024 cho thấy giá trị kinh tế khổng lồ của phần mềm mã nguồn mở (Open Source Software, OSS). Theo nghiên cứu, OSS có mặt trong 96% tất cả các mã phần mềm, và một số phần mềm thương mại có đến 99,9% là các thành phần OSS. Nếu không có OSS, các công ty sẽ chi tiêu gấp 3,5 lần cho phần mềm.

> Các hệ sinh thái của các công ty xây dựng, theo xu hướng toàn cầu, sẽ dần chuyển sang mô hình kinh tế thứ năm, nơi các dịch vụ phân tích và tư vấn dựa trên dữ liệu sẽ trở nên ưu tiên hơn so với các giải pháp khép kín, tách biệt với các kịch bản sử dụng cứng nhắc.

Thời đại số hóa sẽ thay đổi cán cân quyền lực trong ngành: thay vì phụ thuộc vào các giải pháp của nhà cung cấp, các công ty sẽ xây dựng khả năng cạnh tranh của mình dựa trên khả năng sử dụng hiệu quả dữ liệu. Kết quả là ngành xây dựng sẽ chuyển từ các hệ thống cứng nhắc lỗi thời sang các hệ sinh thái linh hoạt, thích ứng, nơi các tiêu chuẩn mở và các công cụ tương thích sẽ trở thành nền tảng cho việc quản lý dự án. Sự kết thúc của thời đại thống trị của các nhà cung cấp ứng dụng sẽ tạo ra những điều kiện mới, trong đó giá trị sẽ được xác định không phải bởi quyền sở hữu mã nguồn khép kín và các kết nối đặc biệt, mà bởi khả năng biến dữ liệu thành lợi thế chiến lược.



CHƯƠNG 2.2.
BIẾN HỖN LOẠN THÀNH TRẬT TỰ VÀ GIẢM ĐỘ PHỨC TẠP

Mã thừa và hệ thống khép kín như rào cản nâng cao năng suất

Trong vài thập kỷ qua, những thay đổi công nghệ trong lĩnh vực CNTT chủ yếu được xác định bởi các nhà cung cấp phần mềm. Chính họ đã định hướng phát triển, xác định công nghệ nào các công ty nên triển khai và công nghệ nào nên bị bỏ qua. Trong thời kỳ chuyển đổi từ các giải pháp rời rạc sang các cơ sở dữ liệu tập trung và các hệ thống tích hợp, các nhà cung cấp đã thúc đẩy các sản phẩm có giấy phép, đảm bảo kiểm soát truy cập và khả năng mở rộng. Sau đó, với sự xuất hiện của công nghệ đám mây và các mô hình Phần mềm như một Dịch vụ (SaaS), sự kiểm soát này đã chuyển thành mô hình đăng ký, giữ người dùng trong vai trò khách hàng thường xuyên của các dịch vụ kỹ thuật số.

Cách tiếp cận này đã tạo ra một nghịch lý: mặc dù có khối lượng mã phần mềm được tạo ra chưa từng có, nhưng thực tế chỉ có một phần nhỏ trong số đó được sử dụng. Có thể mã đã được viết nhiều lần, gấp hàng trăm hoặc hàng ngàn lần so với nhu cầu thực tế, vì cùng một quy trình kinh doanh được mô tả và sao chép theo nhiều cách khác nhau trong hàng chục hoặc hàng trăm chương trình — ngay cả trong cùng một công ty. Trong khi đó, chi phí cho việc phát triển đã được thanh toán và những chi phí này là không thể hoàn lại. Tuy nhiên, ngành công nghiệp vẫn tiếp tục tái sản xuất chu kỳ này, tạo ra các sản phẩm mới với giá trị gia tăng tối thiểu cho người dùng cuối, thường dưới áp lực của kỳ vọng thị trường hơn là nhu cầu thực tế.

> Theo Hướng dẫn đánh giá chi phí phát triển phần mềm do Đại học Mua sắm Quốc phòng (DAU) biên soạn, chi phí phát triển phần mềm có thể thay đổi đáng kể tùy thuộc vào một số yếu tố, bao gồm độ phức tạp của hệ thống và công nghệ được chọn. Lịch sử cho thấy, chi phí phát triển vào năm 2008 khoảng 100 đô la cho mỗi dòng mã nguồn (SLOC), trong khi chi phí bảo trì có thể tăng lên tới 4.000 đô la cho mỗi SLOC.

Chỉ một trong những thành phần của ứng dụng CAD - lõi hình học - có thể có hàng chục triệu dòng mã. Tình hình tương tự cũng xảy ra trong các hệ thống ERP, mà chúng ta sẽ trở lại thảo luận về độ phức tạp của chúng trong phần năm của cuốn sách. Tuy nhiên, khi xem xét kỹ lưỡng, rõ ràng là một phần đáng kể của mã này không tạo ra giá trị gia tăng, mà chỉ thực hiện chức năng "người đưa thư" - cơ học di chuyển dữ liệu giữa cơ sở dữ liệu, API, giao diện người dùng và các bảng khác của hệ thống. Bất chấp huyền thoại phổ biến về tầm quan trọng của cái gọi là logic kinh doanh, thực tế khắc nghiệt hơn nhiều: các cơ sở mã hiện đại tràn ngập các khối mẫu lỗi thời (mã di sản), mục đích duy nhất của chúng là đảm bảo việc truyền dữ liệu giữa các bảng và các thành phần mà không ảnh hưởng đến việc ra quyết định hoặc tăng hiệu quả kinh doanh.-

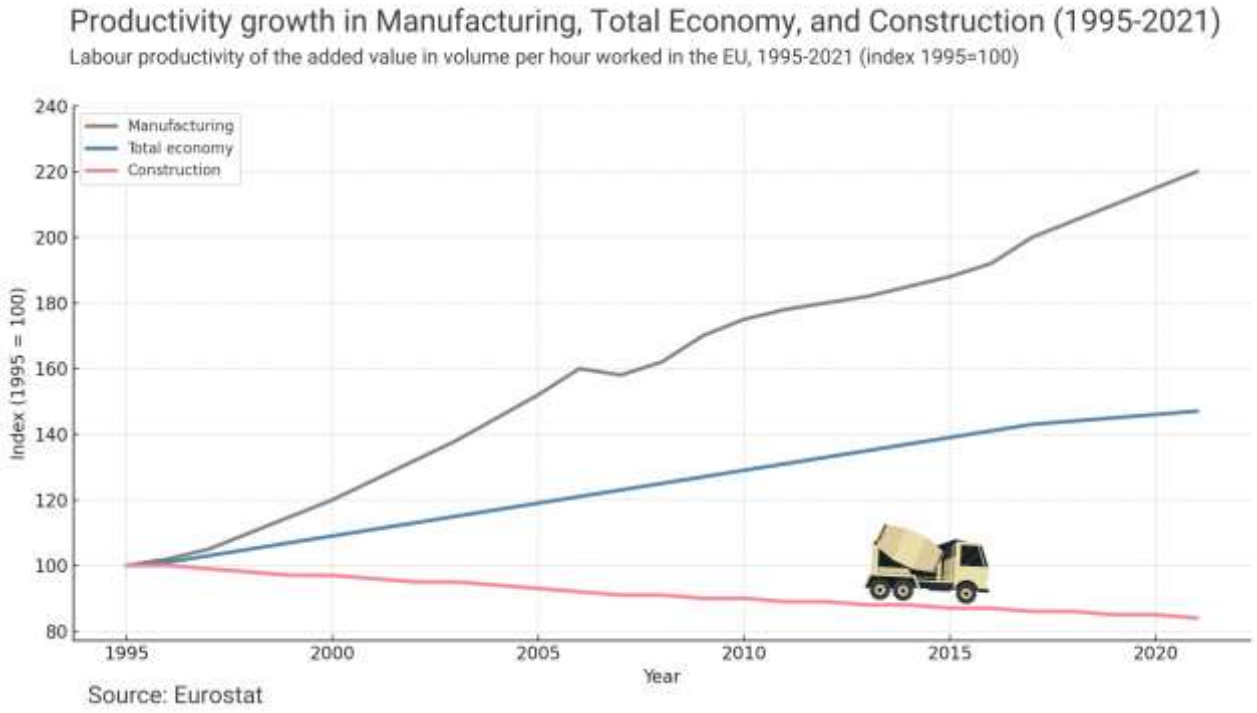
Cuối cùng, các giải pháp đóng, xử lý dữ liệu từ nhiều nguồn khác nhau, không thể tránh khỏi trở thành những "hệ sinh thái mì spaghetti" phức tạp. Chỉ một đội ngũ quản lý khổng lồ, làm việc trong chế độ bán thủ công mới có thể xử lý được những hệ thống phức tạp, rối rắm này. Cách tổ chức quản lý dữ liệu như vậy không chỉ không hiệu quả về mặt tài nguyên mà còn tạo ra những điểm yếu nghiêm trọng trong quy trình kinh doanh, khiến công ty phụ thuộc vào một nhóm chuyên gia hẹp, những người hiểu cách hoạt động của mê cung công nghệ này.

Sự gia tăng liên tục về khối lượng mã, số lượng ứng dụng và sự phức tạp của các khái niệm do các nhà cung cấp đưa ra đã dẫn đến một kết quả tất yếu - sự gia tăng độ phức tạp của hệ sinh thái CNTT trong ngành xây dựng. Điều này đã làm cho việc thực hiện số hóa thông qua việc tăng số lượng ứng dụng trong

ngành trở nên kém hiệu quả. Các sản phẩm phần mềm được phát triển mà không chú ý đúng mức đến nhu cầu của người dùng thường yêu cầu nguồn lực đáng kể cho việc triển khai và hỗ trợ, nhưng không mang lại lợi ích như mong đợi.

> Theo nghiên cứu của McKinsey về "Tăng cường năng suất xây dựng", trong hai thập kỷ qua, mức tăng trưởng năng suất lao động toàn cầu trong ngành xây dựng chỉ đạt trung bình 1% mỗi năm, so với mức tăng 2,8% của nền kinh tế toàn cầu và 3,6% của ngành chế biến. Tại Hoa Kỳ, năng suất lao động trong ngành xây dựng tính trên mỗi công nhân đã giảm một nửa kể từ những năm 1960.

Sự gia tăng độ phức tạp của hệ thống, sự cô lập và tính đóng kín của dữ liệu đã làm xấu đi sự giao tiếp giữa các chuyên gia, khiến ngành xây dựng trở thành một trong những ngành kém hiệu quả nhất.



Sự đóng kín và phức tạp của dữ liệu, và do đó là sự giao tiếp kém giữa các chuyên gia, đã đưa ngành xây dựng trở thành một trong những lĩnh vực kém hiệu quả nhất trong nền kinh tế.

> Như được nhấn mạnh trong nghiên cứu của McKinsey (2024), "Việc đảm bảo năng suất xây dựng không còn là điều tùy chọn", trong bối cảnh nguồn lực ngày càng khan hiếm và ngành công nghiệp có mục tiêu tăng gấp đôi tốc độ tăng trưởng, xây dựng không thể tiếp tục duy trì ở mức năng suất hiện tại. Dự báo cho thấy chi phí xây dựng toàn cầu sẽ tăng từ 13 nghìn tỷ đô la vào năm 2023 lên các giá trị cao hơn đáng kể vào cuối thập kỷ, khiến vấn đề hiệu quả không chỉ trở nên cấp thiết mà còn mang tính chất then chốt.

Một trong những cách quan trọng để nâng cao hiệu quả sẽ là sự thống nhất và đơn giản hóa cấu trúc của các ứng dụng và kiến trúc hệ sinh thái xử lý dữ liệu. Cách tiếp cận này giúp loại bỏ các lớp trừu tượng thừa thãi và độ phức tạp không cần thiết mà đã tích lũy qua nhiều năm trong các hệ thống doanh nghiệp.

Từ silos đến kho dữ liệu thống nhất

Càng nhiều dữ liệu mà tổ chức tích lũy, việc khai thác giá trị thực sự từ chúng càng trở nên khó khăn hơn. Do tính phân mảnh trong việc lưu trữ thông tin trong các silo cách ly, các công ty hiện đại trong quy trình kinh doanh giống như những người xây dựng đang cố gắng xây dựng một tòa nhà chọc trời từ những vật liệu được lưu trữ tại hàng nghìn kho khác nhau. Sự dư thừa thông tin không chỉ làm khó khăn trong việc truy cập dữ liệu có giá trị pháp lý mà còn làm chậm quá trình ra quyết định: mỗi bước đều cần phải được kiểm tra và xác nhận nhiều lần.

> Mỗi nhiệm vụ hoặc quy trình đều bị gắn chặt với một bảng hoặc cơ sở dữ liệu riêng biệt, và việc trao đổi dữ liệu giữa các hệ thống yêu cầu các tích hợp phức tạp. Lỗi và sự không nhất quán trong một hệ thống có thể gây ra sự cố dây chuyền ở những hệ thống khác. Các giá trị không chính xác, cập nhật muộn và thông tin trùng lặp buộc nhân viên phải mất thời gian đáng kể để kiểm tra và xác nhận dữ liệu bằng tay. Cuối cùng, tổ chức tiêu tốn nhiều thời gian hơn để khắc phục hậu quả của sự phân mảnh hơn là để phát triển và tối ưu hóa quy trình.

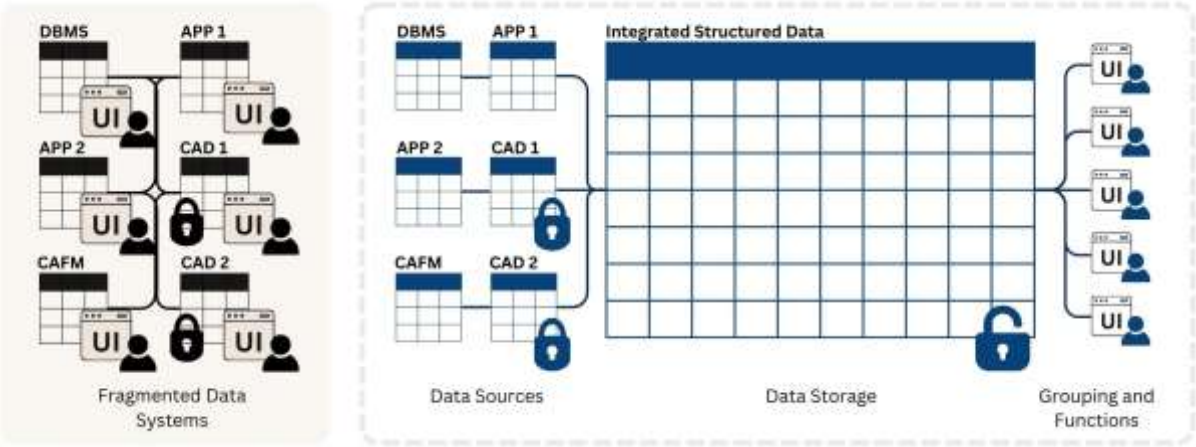
Vấn đề này là phổ quát: một số công ty vẫn tiếp tục vật lộn với sự hỗn loạn, trong khi những công ty khác tìm thấy giải pháp trong tích hợp – chuyển dòng thông tin về một hệ thống lưu trữ tập trung. Hãy tưởng tượng điều này như một bảng lớn duy nhất, nơi có thể lưu trữ bất kỳ thực thể nào liên quan đến nhiệm vụ, dự án và đối tượng. Thay vì hàng chục bảng và định dạng rời rạc, một kho dữ liệu liên kết duy nhất xuất hiện, cho phép: -

- tối thiểu hóa mất mát dữ liệu;
- loại bỏ sự cần thiết phải thường xuyên xác nhận thông tin;
- cải thiện khả năng truy cập và chất lượng của dữ liệu;
- đơn giản hóa việc xử lý phân tích và máy học.

Đưa dữ liệu về một tiêu chuẩn thống nhất có nghĩa là bất kể nguồn gốc, thông tin sẽ được chuyển đổi thành định dạng đồng nhất và có thể đọc được bởi máy. Cách tổ chức dữ liệu như vậy cho phép kiểm tra tính toàn vẹn của chúng, phân tích theo thời gian thực và sử dụng kịp thời cho việc ra quyết định quản lý.

Chúng ta sẽ nói chi tiết về khái niệm hệ thống lưu trữ tích hợp và ứng dụng của chúng trong phân tích và máy học trong chương "Lưu trữ dữ liệu lớn và máy học". Các chủ đề về mô hình hóa và cấu trúc dữ liệu sẽ được làm rõ trong các chương "Chuyển dữ liệu thành dạng cấu trúc" và "Cách các tiêu chuẩn thay đổi trò

chơi: từ tệp ngẫu nhiên đến mô hình dữ liệu được suy nghĩ kỹ".



Hình. 2.22 Tích hợp dữ liệu loại bỏ sự cô lập, cải thiện khả năng truy cập thông tin và tối ưu hóa quy trình kinh doanh.

Sau khi cấu trúc và hợp nhất dữ liệu, bước logic tiếp theo là kiểm tra chúng. Khi có một kho dữ liệu tích hợp duy nhất, quá trình này trở nên đơn giản hơn đáng kể: không còn nhiều sơ đồ không thống nhất, cấu trúc trùng lặp và những mối quan hệ phức tạp giữa các bảng nữa. Tất cả thông tin được đưa về một mô hình dữ liệu thống nhất, điều này loại bỏ các mâu thuẫn nội tại và tăng tốc quá trình xác thực. Kiểm tra và đảm bảo chất lượng dữ liệu là những khía cạnh thiết yếu trong tất cả các quy trình kinh doanh, và chúng ta sẽ xem xét chi tiết về chúng trong các chương tương ứng của cuốn sách.

Tại giai đoạn cuối cùng, dữ liệu được nhóm, lọc và phân tích. Chúng được áp dụng các chức năng khác nhau: tổng hợp (cộng, nhân), tính toán giữa các bảng, cột hoặc hàng (Hình. 2.24). Làm việc với dữ liệu trở thành một chuỗi các bước: thu thập, cấu trúc, kiểm tra, chuyển đổi, xử lý phân tích và xuất ra các ứng dụng cuối cùng, nơi thông tin được sử dụng để giải quyết các nhiệm vụ thực tiễn. Chi tiết về việc xây dựng các kịch bản như vậy, tự động hóa các bước và xây dựng quy trình xử lý sẽ được thảo luận trong các chương liên quan đến các quy trình ETL và phương pháp data pipeline.

Như vậy, chuyển đổi số không chỉ là việc đơn giản hóa cách làm việc với thông tin. Đó là việc từ chối sự phức tạp dư thừa trong quản lý dữ liệu, từ hỗn loạn đến có thể dự đoán, từ nhiều hệ thống đến quy trình có thể kiểm soát. Độ phức tạp của kiến trúc càng thấp, mã cần thiết để hỗ trợ cũng giảm. Và trong tương lai, mã có thể biến mất hoàn toàn, nhường chỗ cho các tác nhân thông minh tự động phân tích, hệ thống hóa và chuyển đổi dữ liệu.

Hệ thống lưu trữ tích hợp cho phép chuyển sang sử dụng các tác nhân AI

Càng ít sự phức tạp của dữ liệu và hệ thống, càng ít mã cần viết và hỗ trợ. Và cách đơn giản nhất để tiết kiệm thời gian phát triển là hoàn toàn loại bỏ mã, thay thế nó bằng dữ liệu. Khi việc phát triển mã ứng dụng chuyển từ mã sang các mô hình dữ liệu, một sự chuyển dịch không thể tránh khỏi hướng về cách tiếp cận dựa trên dữ liệu (data-driven) xảy ra, vì những khái niệm này đứng sau một tư duy hoàn toàn khác.

> Khi con người chọn con đường làm việc với dữ liệu làm trung tâm, họ bắt đầu nhìn nhận vai trò của chúng theo cách khác. Dữ liệu không còn chỉ là "nguyên liệu" cho các ứng dụng - giờ đây chúng là nền tảng xung quanh đó kiến trúc, logic và sự tương tác được xây dựng.

Tuy nhiên, cách tiếp cận truyền thống đối với quản lý dữ liệu thường bắt đầu ở cấp độ ứng dụng và trong xây dựng giống như một hệ thống quan liêu cồng kềnh: nhiều mức phê duyệt, kiểm tra thủ công, vô số phiên bản tài liệu qua các sản phẩm phần mềm tương ứng. Với sự phát triển của công nghệ số, ngày càng nhiều công ty sẽ buộc phải chuyển sang nguyên tắc tối giản - chỉ lưu trữ và sử dụng những gì thực sự cần thiết và sẽ được sử dụng.

Các nhà cung cấp đã nắm bắt được logic tối giản. Để đơn giản hóa quy trình lưu trữ và xử lý dữ liệu, công việc của người dùng được chuyển từ chức năng của các ứng dụng và công cụ ngoại tuyến sang các dịch vụ đám mây và các giải pháp SaaS.

> Khái niệm SaaS (Phần mềm như một dịch vụ) là một trong những hướng đi chính trong các hạ tầng CNTT hiện đại, cho phép người dùng truy cập vào các ứng dụng qua internet mà không cần phải cài đặt và bảo trì phần mềm trên máy tính của riêng mình.

Một mặt, SaaS đã đơn giản hóa việc mở rộng quy mô, quản lý phiên bản và giảm chi phí hỗ trợ và bảo trì, nhưng mặt khác, ngoài việc phụ thuộc vào logic của ứng dụng cụ thể, nó cũng khiến người dùng hoàn toàn phụ thuộc vào hạ tầng đám mây của nhà cung cấp. Nếu dịch vụ gặp sự cố, quyền truy cập vào dữ liệu và quy trình kinh doanh có thể bị chặn tạm thời hoặc thậm chí lâu dài. Hơn nữa, tất cả dữ liệu của người dùng khi làm việc với các ứng dụng SaaS đều được lưu trữ trên máy chủ của nhà cung cấp, điều này tạo ra rủi ro về mặt an ninh và tuân thủ các yêu cầu quy định. Việc thay đổi mức giá hoặc điều kiện sử dụng cũng có thể dẫn đến việc tăng chi phí hoặc cần phải di chuyển khẩn cấp.

Sự phát triển của AI, các tác nhân LLM và phương pháp tiếp cận dựa trên dữ liệu đã đặt ra câu hỏi về tương lai của các ứng dụng dưới hình thức truyền thống và SaaS. Nếu trước đây các ứng dụng và dịch vụ là cần thiết để quản lý logic kinh doanh và xử lý dữ liệu, thì với sự xuất hiện của các tác nhân AI, những chức năng này có thể chuyển giao cho các hệ thống thông minh, hoạt động trực tiếp với dữ liệu.

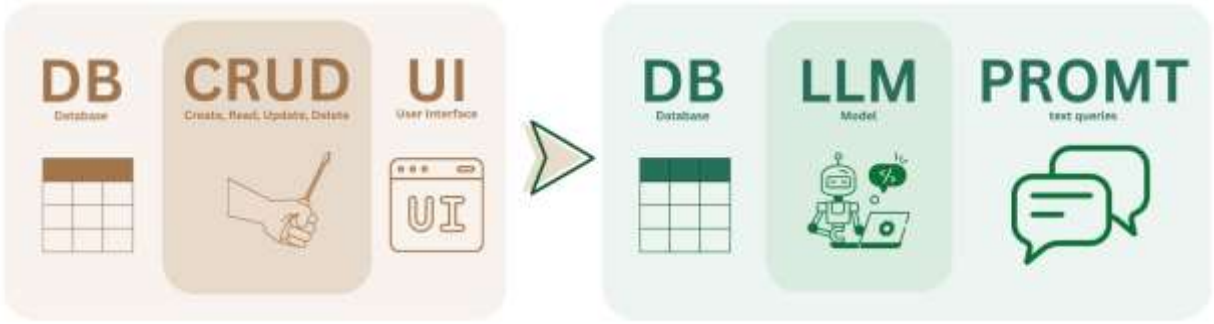
Chính vì lý do đó, ngày càng nhiều trong các phòng IT và ở cấp quản lý thảo luận về kiến trúc hybrid, nơi các tác nhân AI và giải pháp cục bộ bổ sung cho các dịch vụ đám mây, giảm thiểu sự phụ thuộc vào các nền tảng SaaS.

> *Cách tiếp cận mà chúng tôi theo đuổi công nhận rằng các ứng dụng kinh doanh truyền thống hoặc ứng dụng SaaS có thể thay đổi hoàn toàn trong kỷ nguyên của các tác nhân. Những ứng dụng này về cơ bản là CRUD [tạo, đọc, cập nhật và xóa] cơ sở dữ liệu với logic kinh doanh. Nhưng trong tương lai, logic này sẽ chuyển giao cho các tác nhân AI.*
>
> *— Satya Nadella, Giám đốc điều hành Microsoft, năm 2024.*

Phương pháp tiếp cận dựa trên dữ liệu và việc sử dụng các tác nhân AI/LLM cho phép giảm bớt số lượng

quy trình thừa thãi, do đó giảm tải cho nhân viên. Khi dữ liệu được tổ chức một cách hợp lý, việc phân tích, hình dung và ứng dụng để ra quyết định trở nên dễ dàng hơn. Thay vì những báo cáo và kiểm tra vô tận, các chuyên gia có thể truy cập thông tin cập nhật chỉ với vài cú nhấp chuột hoặc thông qua các tác nhân LLM tự động dưới dạng tài liệu và bảng điều khiển đã hoàn chỉnh.

Trong công việc với dữ liệu, các công cụ trí tuệ nhân tạo (AI) và chat LLM sẽ hỗ trợ chúng ta. Trong những năm gần đây, có một xu hướng chuyển đổi từ các thao tác CRUD truyền thống (tạo, đọc, cập nhật, xóa) sang việc sử dụng các mô hình ngôn ngữ lớn (LLM) để quản lý dữ liệu. LLM có khả năng diễn dịch ngôn ngữ tự nhiên và tự động tạo ra các truy vấn phù hợp đến cơ sở dữ liệu, điều này làm đơn giản hóa việc tương tác với các hệ thống quản lý dữ liệu.-



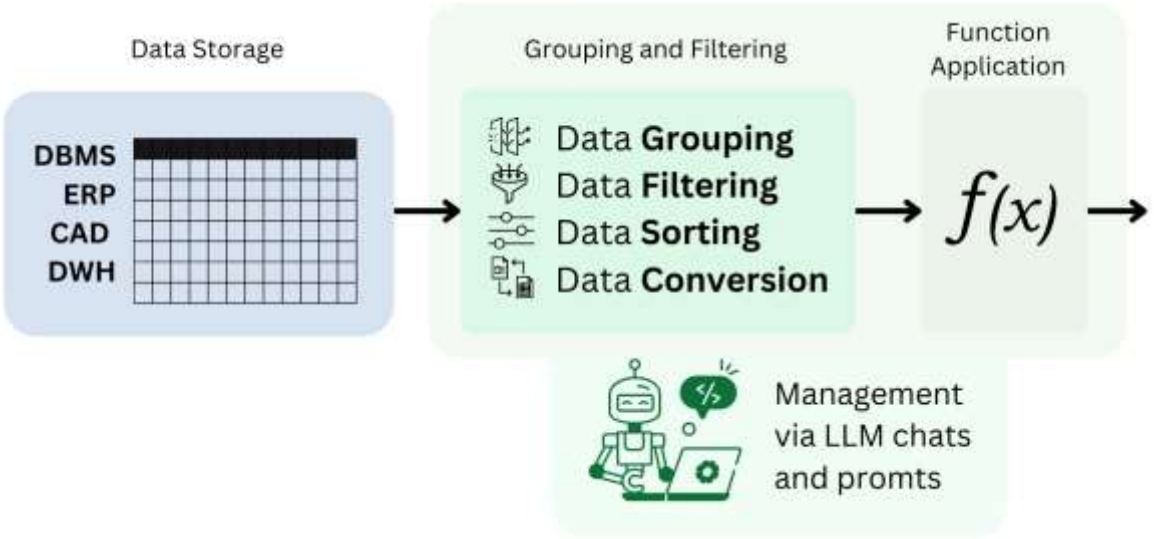
Hình 2.23 AI sẽ thay thế các quyết định liên quan đến quản lý kho dữ liệu và cơ sở dữ liệu cũng như tích hợp chúng, dần dần loại bỏ các ứng dụng truyền thống và các hoạt động CRUD.

> *Trong vòng 3–6 tháng tới, trí tuệ nhân tạo sẽ viết 90% mã, và sau 12 tháng, hầu như toàn bộ mã có thể được tạo ra bởi trí tuệ nhân tạo.*
>
> *— Dario Amodei, Giám đốc điều hành công ty LLM Anthropic, tháng 3 năm 2025.*

Mặc dù công cụ phát triển AI (chẳng hạn như GitHub Copilot) đang phát triển nhanh chóng, vào năm 2025, các lập trình viên vẫn đóng vai trò chủ chốt trong quá trình này. Các tác nhân AI ngày càng trở thành trợ lý hữu ích: chúng tự động diễn giải các yêu cầu của người dùng, tạo ra các truy vấn SQL và Pandas (thêm chi tiết về điều này sẽ có ở các chương sau) hoặc viết mã cho phân tích dữ liệu. Như vậy, trí tuệ nhân tạo đang dần thay thế các giao diện người dùng truyền thống của ứng dụng.

Sự phổ biến của các mô hình trí tuệ nhân tạo, chẳng hạn như các mô hình ngôn ngữ, sẽ thúc đẩy sự phát triển của các kiến trúc lai. Thay vì từ bỏ hoàn toàn các giải pháp đám mây và sản phẩm SaaS, chúng ta có thể thấy sự tích hợp giữa các dịch vụ đám mây với các hệ thống quản lý dữ liệu cục bộ. Chẳng hạn, học liên hợp (federated learning) cho phép sử dụng các mô hình AI mạnh mẽ mà không cần chuyển dữ liệu nhạy cảm lên đám mây. Như vậy, các công ty có thể giữ quyền kiểm soát dữ liệu của mình, đồng thời tiếp cận

được công nghệ tiên tiến.



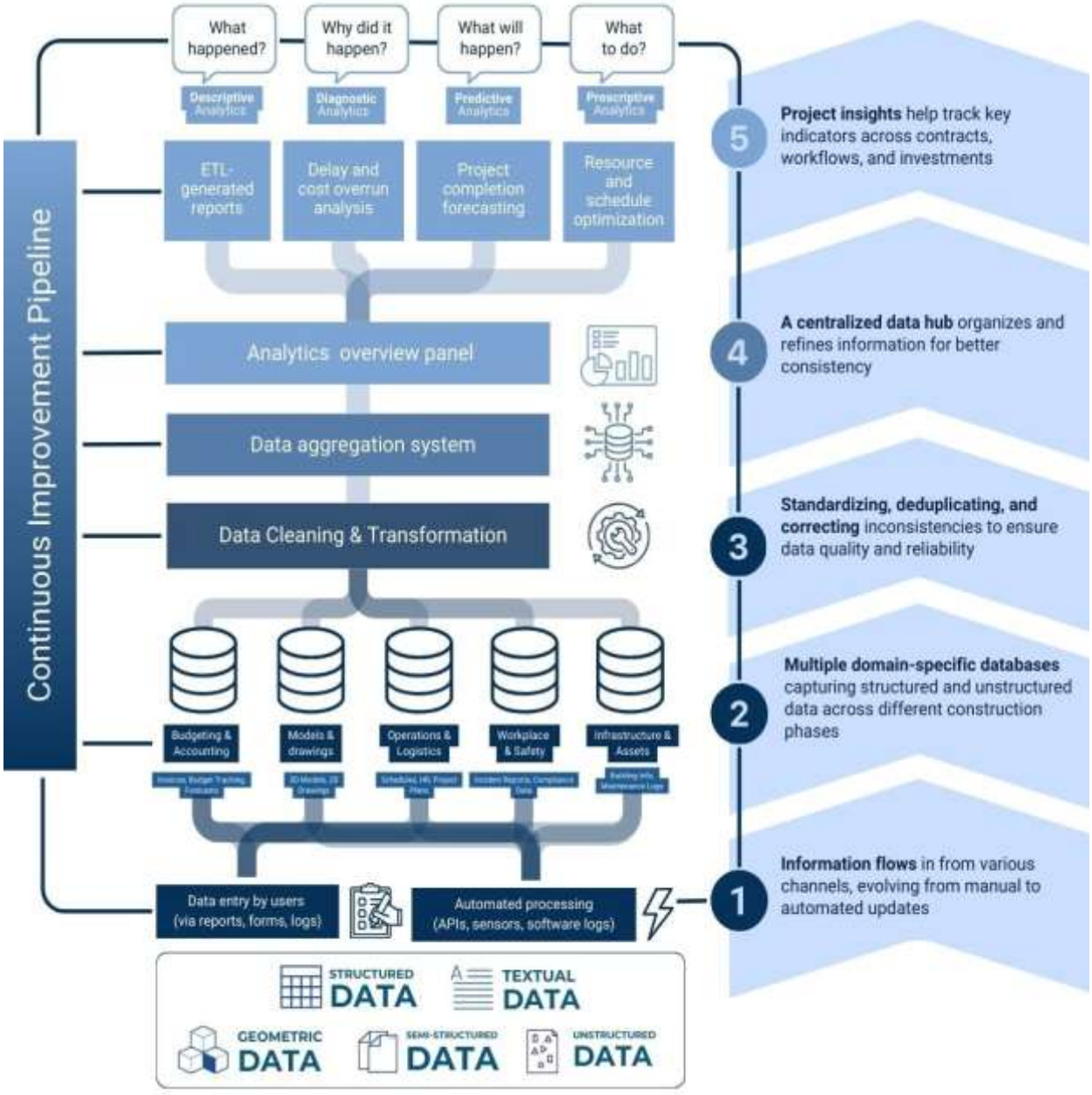
Rís. 2.24 Các thao tác chính về nhóm, lọc và sắp xếp sẽ được thực hiện bởi các cuộc trò chuyện LLM.

Tương lai của ngành xây dựng sẽ dựa trên sự kết hợp giữa các giải pháp cục bộ, công suất đám mây và các mô hình thông minh hoạt động cùng nhau để tạo ra hệ thống quản lý dữ liệu hiệu quả và an toàn. LLM sẽ cho phép người dùng không có kiến thức kỹ thuật sâu sắc tương tác với các cơ sở dữ liệu và kho dữ liệu bằng cách diễn đạt các truy vấn của họ bằng ngôn ngữ tự nhiên. Thêm chi tiết về LLM và các tác nhân AI cũng như cách chúng hoạt động sẽ được bàn luận trong chương "Các tác nhân LLM và định dạng dữ liệu có cấu trúc".

Dữ liệu được tổ chức đúng cách cùng với các công cụ phân tích đơn giản, thân thiện hỗ trợ bởi LLM sẽ không chỉ đơn giản hóa công việc với thông tin mà còn giúp giảm thiểu sai sót, nâng cao hiệu quả và tự động hóa các quy trình.

Từ việc thu thập dữ liệu đến ra quyết định: con đường đến tự động hóa

Trong các phần tiếp theo của cuốn sách, chúng tôi sẽ xem xét chi tiết cách các chuyên gia tương tác với nhau và cách dữ liệu trở thành cơ sở cho việc ra quyết định, tự động hóa và nâng cao hiệu quả công việc. Rís. 2.25 trình bày một ví dụ về sơ đồ, cho thấy quy trình các giai đoạn xử lý dữ liệu trong tiếp cận dựa trên dữ liệu. Sơ đồ này minh họa đường nét của quy trình cải tiến liên tục (Continuous Improvement Pipeline), các phần trong quy trình này sẽ được xem xét chi tiết hơn trong cuốn sách.



Rís. 2.25 Ví dụ về quy trình cải tiến dữ liệu liên tục: quy trình xử lý và phân tích dữ liệu trong các dự án xây dựng.

Hệ thống mô tả các quy trình kinh doanh của một công ty vừa và nhỏ được xây dựng trên nguyên tắc đa lớp. Hệ thống bao gồm: thu thập dữ liệu, làm sạch, tổng hợp, xử lý phân tích và ra quyết định dựa trên kết quả thu được. Tất cả các giai đoạn này sẽ được chúng tôi nghiên cứu trong cuốn sách tiếp theo — cả trong bối cảnh lý thuyết và thông qua các ví dụ thực tiễn:

- Ở cấp độ đầu tiên, quá trình nhập dữ liệu diễn ra (Rís. 3.11). Thông tin được nhập vào theo cả hai cách: theo chế độ thủ công (thông qua báo cáo, biểu mẫu, nhật ký) và dưới dạng tự động (qua API, cảm biến, hệ thống phần mềm). Dữ liệu có thể có nhiều cấu trúc khác nhau: hình học, văn bản, không cấu trúc. Ở giai đoạn này, cần thiết phải có tiêu chuẩn hóa, cấu trúc và đồng hóa các dòng thông

tin.

- Cấp độ tiếp theo là xử lý và chuyển đổi dữ liệu. Bao gồm các quy trình làm sạch, xóa trùng lặp, sửa chữa lỗi và chuẩn bị thông tin cho phân tích tiếp theo. Giai đoạn này cực kỳ quan trọng, vì chất lượng phân tích phụ thuộc trực tiếp vào độ sạch và độ chính xác của dữ liệu.
- Sau đó, dữ liệu được đưa vào các bảng chuyên dụng, khung dữ liệu hoặc cơ sở dữ liệu, phân chia theo các hướng chức năng: lập ngân sách và kế toán, mô hình và bản vẽ, logistics, an ninh và cơ sở hạ tầng. Sự phân chia này cho phép tổ chức quyền truy cập thuận tiện và đảm bảo khả năng phân tích chéo thông tin.
- Sau đó, dữ liệu được tổng hợp và hiển thị trên bảng phân tích (vitrina). Tại đây, các phương pháp phân tích mô tả, chẩn đoán, dự đoán và chỉ dẫn được áp dụng. Điều này cho phép trả lời những câu hỏi chính: đã xảy ra điều gì, tại sao nó xảy ra, điều gì sẽ xảy ra trong tương lai và nên thực hiện những hành động nào. Ví dụ, hệ thống có thể phát hiện sự chậm trễ, dự đoán hoàn thành các dự án hoặc tối ưu hóa nguồn lực. -
- Cuối cùng, ở cấp độ cuối cùng, các kết luận phân tích và các chỉ số chính được hình thành, giúp theo dõi việc thực hiện các hợp đồng, quản lý đầu tư và cải thiện các quy trình kinh doanh. Thông tin này trở thành cơ sở cho việc ra quyết định và chiến lược phát triển của công ty.

Tương tự, dữ liệu trải qua quá trình từ thu thập đến sử dụng trong quản lý chiến lược. Trong các phần tiếp theo của cuốn sách, chúng ta sẽ xem xét từng giai đoạn một cách chi tiết, đặc biệt chú trọng đến các loại dữ liệu, phương pháp xử lý dữ liệu, công cụ phân tích và các trường hợp thực tế sử dụng những cách tiếp cận này trong ngành xây dựng.

Các bước tiếp theo: chuyển đổi hỗn độn thành hệ thống có thể quản lý

Trong phần này, chúng ta đã khảo sát vấn đề của các silo thông tin và xem xét ảnh hưởng của sự phức tạp dư thừa của hệ thống đối với hiệu quả kinh doanh, phân tích sự chuyển đổi từ cuộc cách mạng công nghiệp lần thứ tư sang lần thứ năm, nơi dữ liệu chứ không phải ứng dụng đóng vai trò trung tâm. Chúng ta đã thấy cách các hệ thống thông tin rời rạc tạo ra rào cản cho việc chia sẻ kiến thức, và sự phức tạp ngày càng gia tăng của bối cảnh CNTT làm giảm năng suất và cản trở sự đổi mới trong ngành xây dựng.

Tóm lại phần này, cần nêu rõ các bước thực tiễn chính sẽ giúp áp dụng các phương pháp đã thảo luận vào các nhiệm vụ hàng ngày của bạn:

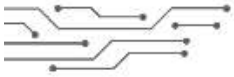
- Hình dung bối cảnh thông tin của bạn
    - [ ] Tạo một bản đồ trực quan về các nguồn dữ liệu (Miro, Figma, Canva) mà bạn thường xuyên làm việc
    - [ ] Thêm vào bản đồ này các hệ thống và ứng dụng mà bạn sử dụng trong công việc
    - [ ] Xác định các chức năng có khả năng bị trùng lặp và các giải pháp dư thừa
    - [ ] Xác định các điểm quan trọng, nơi có thể xảy ra mất mát hoặc sai lệch dữ liệu trong quá trình truyền giữa các hệ thống
- Thực hiện các thực tiễn quản lý dữ liệu cá nhân
    - [ ] Di chuyển trọng tâm từ ứng dụng sang dữ liệu như một tài sản chính trong các quy trình
    - [ ] Tài liệu lại các nguồn dữ liệu và phương pháp xử lý để đảm bảo tính minh bạch
    - [ ] Phát triển các cơ chế đánh giá và nâng cao chất lượng dữ liệu

- Hướng tới việc dữ liệu được nhập một lần và sử dụng nhiều lần - đây là nền tảng của tổ chức quy trình hiệu quả

- Thúc đẩy cách tiếp cận dựa trên dữ liệu trong nhóm của bạn
  - Đề nghị sử dụng các định dạng trao đổi dữ liệu chuẩn hóa và thống nhất giữa các đồng nghiệp.
  - Thường xuyên nêu ra các vấn đề liên quan đến chất lượng và sự sẵn có của dữ liệu trong các cuộc họp của nhóm.
  - Làm quen với các giải pháp mã nguồn mở thay thế cho các công cụ bạn đang sử dụng để giải quyết vấn đề của mình.

Bắt đầu từ những điều nhỏ - chọn một quy trình hoặc tập dữ liệu cụ thể rất quan trọng cho công việc của bạn và áp dụng phương pháp tập trung vào dữ liệu, di chuyển trọng tâm từ công cụ sang dữ liệu. Bằng cách đạt được thành công trong một thử nghiệm cụ thể, bạn sẽ không chỉ có được kinh nghiệm thực tiễn mà còn có một minh chứng rõ ràng cho lợi ích của phương pháp mới đối với nhóm của bạn. Trong quá trình thực hiện hầu hết các bước này, nếu có thắc mắc, bạn có thể tìm đến bất kỳ LLM hiện đại nào để được giải thích và hỗ trợ.

Trong các phần tiếp theo của cuốn sách, chúng ta sẽ chuyển sang xem xét chi tiết hơn về các phương pháp cấu trúc và thống nhất dữ liệu, và nghiên cứu các cách tiếp cận thực tiễn để tích hợp thông tin đa dạng. Chúng ta sẽ đặc biệt chú ý đến việc chuyển đổi từ các kho dữ liệu phân tán sang các hệ sinh thái dữ liệu thống nhất, mà đóng vai trò chủ chốt trong quá trình chuyển đổi số của ngành xây dựng.





III PHẦN **KHUNG DỮ LIỆU TRONG CÁC QUY TRÌNH KINH DOANH XÂY DỰNG**

Phần ba sẽ hình thành một cái nhìn toàn diện về phân loại dữ liệu trong xây dựng và các phương pháp để tổ chức chúng hiệu quả. Các đặc điểm và tính chất làm việc với dữ liệu có cấu trúc, không có cấu trúc, bán cấu trúc, dữ liệu văn bản và dữ liệu hình học sẽ được phân tích trong bối cảnh các dự án xây dựng. Chúng ta sẽ thảo luận về các định dạng lưu trữ hiện đại và các giao thức trao đổi thông tin giữa các hệ thống khác nhau được sử dụng trong ngành. Các công cụ và phương pháp thực tiễn để chuyển đổi dữ liệu đa định dạng thành một môi trường có cấu trúc duy nhất, bao gồm các cách tích hợp dữ liệu CAD (BIM), cũng sẽ được mô tả. Các phương pháp đảm bảo chất lượng dữ liệu thông qua chuẩn hóa và xác thực sẽ được đề xuất, điều này rất quan trọng để đảm bảo độ chính xác của các phép toán trong xây dựng. Các khía cạnh thực tiễn của việc sử dụng các công nghệ hiện đại (Python Pandas, mô hình LLM) sẽ được phân tích chi tiết với các ví dụ mã để giải quyết các vấn đề điển hình trong ngành xây dựng. Giá trị của việc tạo ra một trung tâm năng lực (CoE) như một cấu trúc tổ chức để điều phối và chuẩn hóa các phương pháp quản lý thông tin sẽ được lý giải.

CHƯƠNG 3.1.
CÁC LOẠI DỮ LIỆU TRONG NGÀNH XÂY DỰNG

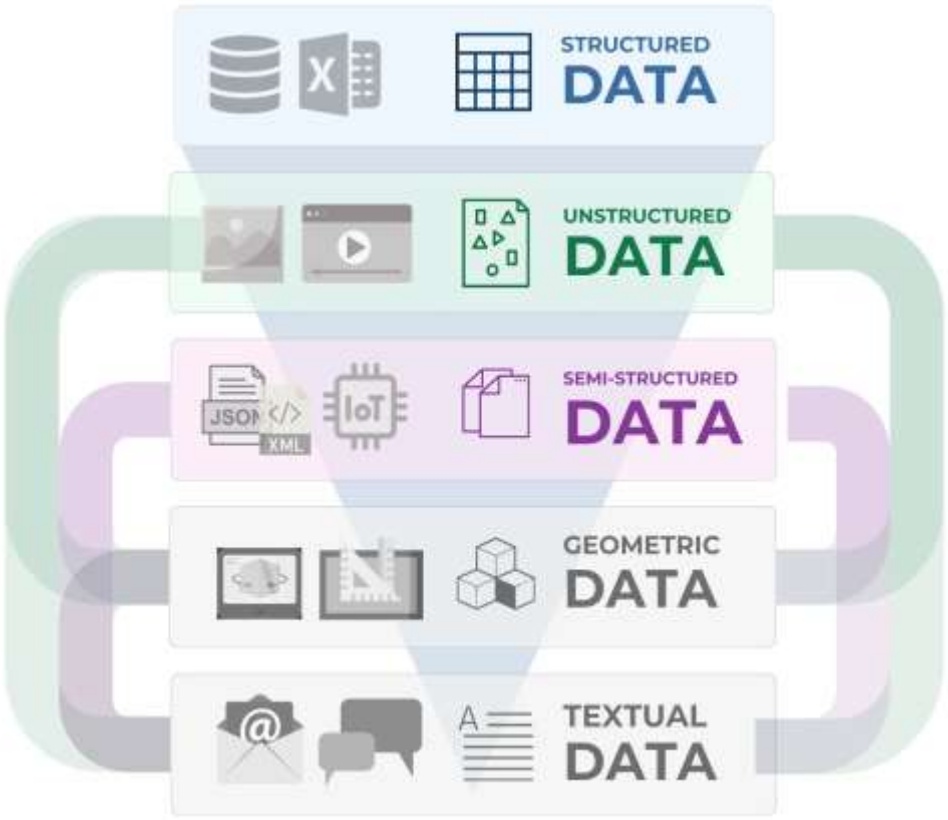
Các loại dữ liệu quan trọng nhất trong ngành xây dựng

Trong ngành xây dựng hiện đại, các hệ thống, ứng dụng và kho dữ liệu của các công ty đang được tích lũy thông tin và dữ liệu từ các loại và định dạng khác nhau (Hình. 3.11). Chúng ta sẽ xem xét kỹ lưỡng các loại dữ liệu chính, tạo nên bức tranh thông tin của một công ty hiện đại hoạt động trong ngành xây dựng:-

- Dữ liệu có cấu trúc: dữ liệu này có cấu trúc tổ chức rõ ràng, chẳng hạn như bảng điện tử Excel và cơ sở dữ liệu quan hệ.
- Dữ liệu không có cấu trúc: đây là thông tin không được tổ chức theo các quy tắc nghiêm ngặt. Ví dụ của dữ liệu này bao gồm văn bản, video, hình ảnh và bản ghi âm.
- Dữ liệu bán cấu trúc: dữ liệu này nằm ở vị trí trung gian giữa dữ liệu có cấu trúc và không có cấu trúc. Chúng chứa các yếu tố cấu trúc, nhưng cấu trúc này không phải lúc nào cũng rõ ràng hoặc thường được mô tả thông qua các sơ đồ khác nhau. Ví dụ về dữ liệu bán cấu trúc trong xây dựng bao gồm: thông số kỹ thuật, tài liệu dự án hoặc báo cáo về công việc đã thực hiện.
- Dữ liệu văn bản: bao gồm tất cả những gì được tạo ra từ các giao tiếp bằng miệng và viết, chẳng hạn như email, biên bản cuộc họp và cuộc gặp.
- Dữ liệu hình học: Dữ liệu này được cung cấp từ các phần mềm CAD, nơi các chuyên gia tạo ra dữ liệu hình học của các thành phần dự án để trực quan hóa, xác nhận giá trị khối lượng hoặc kiểm tra va chạm.

Quan trọng cần lưu ý rằng dữ liệu hình học và dữ liệu văn bản (chữ và số) không phải là các thể loại riêng biệt, mà có thể xuất hiện trong tất cả ba loại dữ liệu. Dữ liệu hình học, ví dụ, có thể là một phần của dữ liệu có cấu trúc (định dạng CAD tham số) hoặc dữ liệu không có cấu trúc (bản vẽ quét). Dữ liệu văn bản cũng tương tự, có thể được tổ chức trong cơ sở dữ liệu (dữ liệu có cấu trúc) hoặc tồn tại dưới dạng tài liệu mà không có cấu trúc rõ ràng.

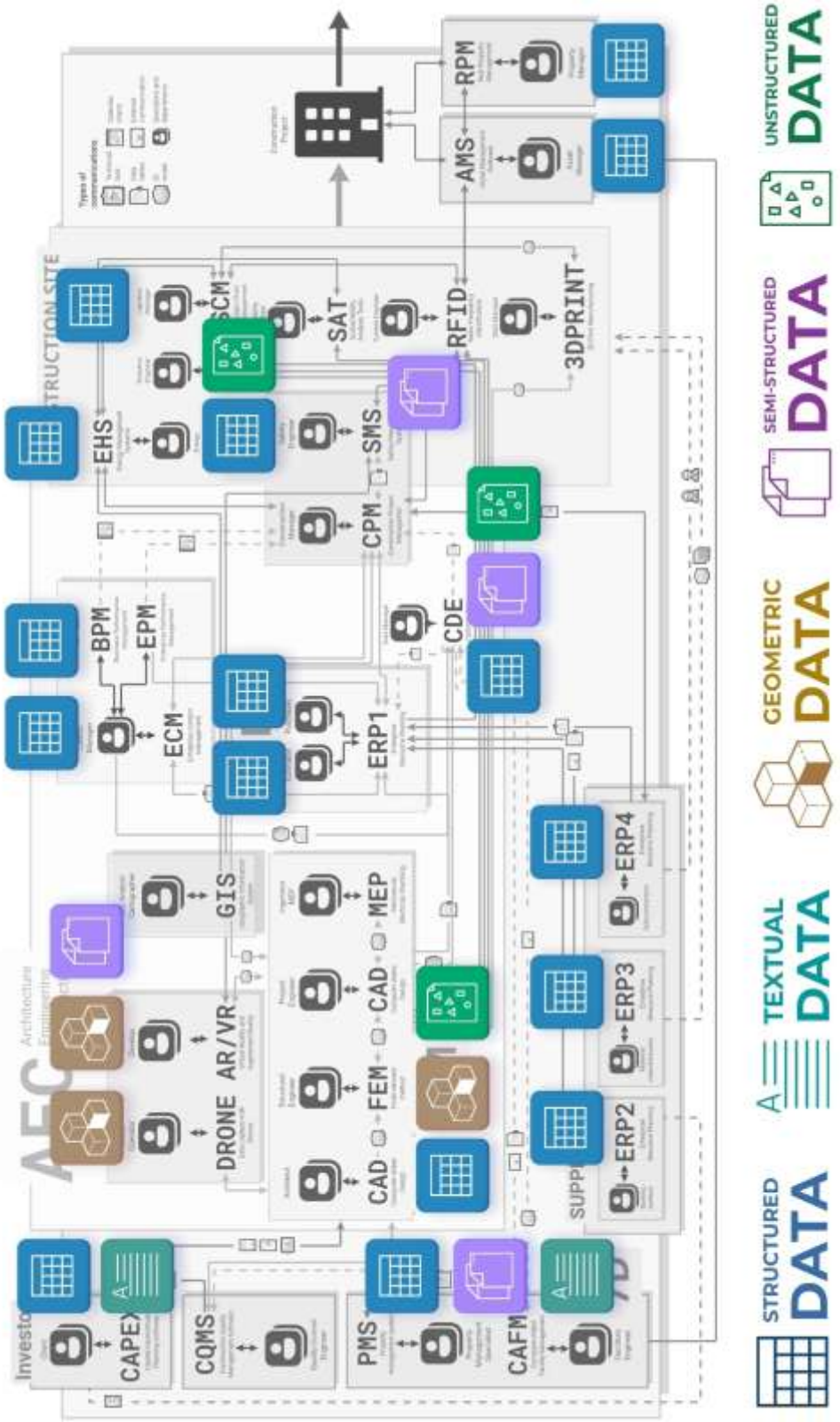
Mỗi loại dữ liệu trong công ty xây dựng đều là một phần độc đáo trong bức tranh thông tin của công ty. Từ dữ liệu không có cấu trúc, chẳng hạn như hình ảnh từ các công trường xây dựng và bản ghi âm cuộc họp, đến các hồ sơ có cấu trúc, bao gồm bảng và cơ sở dữ liệu - mỗi yếu tố đều đóng một vai trò quan trọng trong việc hình thành cảnh quan thông tin của công ty.



Hình 3.11 Các kỹ sư và quản lý dữ liệu phải học cách làm việc với tất cả các loại dữ liệu được sử dụng trong ngành xây dựng.

Đây là một ví dụ về danh sách chỉ một số hệ thống và các loại dữ liệu liên quan (Hình 3.12) được sử dụng trong xây dựng: -

- ERP (Quản lý tài nguyên doanh nghiệp) - thường xử lý dữ liệu có cấu trúc, giúp quản lý tài nguyên của doanh nghiệp và tích hợp các quy trình kinh doanh khác nhau.
- CAD (Thiết kế hỗ trợ máy tính) kết hợp với BIM (Mô hình thông tin xây dựng) - sử dụng dữ liệu hình học và dữ liệu bán cấu trúc để thiết kế và mô hình hóa các dự án xây dựng, đảm bảo độ chính xác và tính nhất quán của thông tin trong giai đoạn thiết kế.
- GIS (Hệ thống thông tin địa lý) - làm việc với dữ liệu hình học và dữ liệu có cấu trúc để tạo ra và phân tích dữ liệu bản đồ và mối quan hệ không gian.
- RFID (Nhận dạng tần số vô tuyến) - sử dụng dữ liệu bán cấu trúc để theo dõi hiệu quả vật liệu và thiết bị tại công trường xây dựng thông qua nhận dạng tần số vô tuyến.
- ECM (Quản lý nội dung kỹ thuật) - là hệ thống quản lý dữ liệu kỹ thuật và tài liệu, bao gồm dữ liệu bán cấu trúc và không có cấu trúc, chẳng hạn như bản vẽ kỹ thuật và tài liệu dự án.

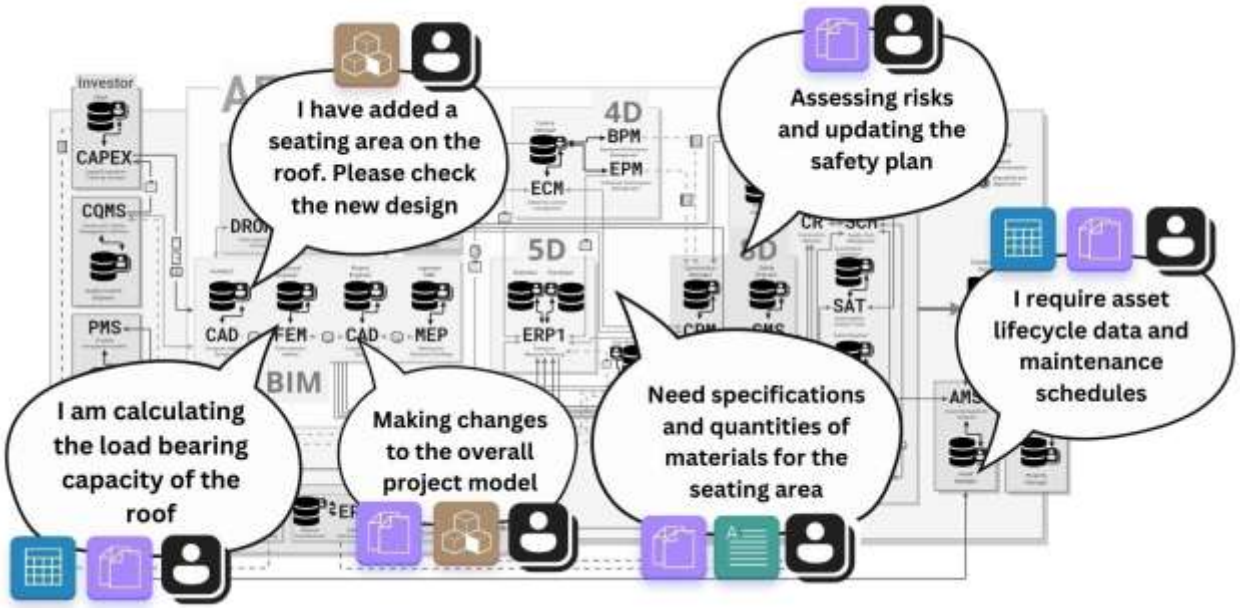


Hình 3.12 Các định dạng và dữ liệu khác nhau lấp đầy các hệ thống khác nhau, yêu cầu chuyển đổi sang dạng có thể tích hợp toàn diện.

Những hệ thống này và nhiều hệ thống khác quản lý một loạt dữ liệu, từ dữ liệu bảng có cấu trúc đến các mô hình hình học phức tạp, đảm bảo sự tương tác tích hợp trong các quy trình thiết kế, lập kế hoạch và quản lý xây dựng.

Trong ví dụ về một cuộc đối thoại đơn giản (Hình 3.13) giữa các chuyên gia dự án xây dựng, có sự trao đổi về nhiều loại dữ liệu khác nhau: -

- Kiến trúc sư: "Dựa trên những mong muốn của khách hàng, tôi đã thêm một khu vực nghỉ ngơi trên mái. Xin vui lòng xem thiết kế mới" (dữ liệu hình học - mô hình).
- Kỹ sư kết cấu: "Dự án đã được nhận. Tôi đang tính toán khả năng chịu lực của mái cho khu vực nghỉ ngơi mới" (dữ liệu có cấu trúc và dữ liệu bán cấu trúc - bảng tính toán).
- Quản lý mua sắm: "Cần có thông số kỹ thuật và số lượng vật liệu cho khu vực nghỉ ngơi, để tổ chức việc mua sắm" (dữ liệu văn bản và dữ liệu bán cấu trúc - danh sách và thông số kỹ thuật).
- Kỹ sư về an toàn lao động và kỹ thuật an toàn: "Nhận dữ liệu về khu vực mới. Đánh giá rủi ro và cập nhật kế hoạch an toàn" (dữ liệu bán cấu trúc - tài liệu và kế hoạch).
- Chuyên gia về mô hình BIM: "Thực hiện thay đổi trong mô hình tổng thể của dự án để điều chỉnh tài liệu làm việc" (dữ liệu hình học và dữ liệu bán cấu trúc).
- Trưởng dự án: "Tôi đưa khu vực nghỉ ngơi mới vào lịch làm việc. Tôi cập nhật lịch trình và tài nguyên trong hệ thống quản lý dự án" (dữ liệu cấu trúc và bán cấu trúc - lịch trình và kế hoạch).
- Chuyên gia bảo trì cơ sở vật chất (FM): "Tôi chuẩn bị dữ liệu cho việc bảo trì khu vực nghỉ ngơi trong tương lai và nhập chúng vào hệ thống quản lý tài sản" (dữ liệu cấu trúc và bán cấu trúc - hướng dẫn và kế hoạch bảo trì kỹ thuật).



Hình 3.13 Giao tiếp giữa các chuyên gia diễn ra cả ở cấp độ văn bản lẫn cấp độ dữ liệu.

Mỗi chuyên gia làm việc với các loại dữ liệu khác nhau, đảm bảo tương tác hiệu quả trong nhóm và hoàn thành dự án thành công. Hiểu rõ sự khác biệt giữa dữ liệu cấu trúc, bán cấu trúc và không cấu trúc giúp nhận thức về vai trò độc nhất của mỗi loại trong các quy trình kinh doanh kỹ thuật số. Điều quan trọng không chỉ là biết có nhiều hình thức dữ liệu khác nhau mà còn là hiểu cách, nơi và lý do chúng được áp dụng.

> Gần đây, ý tưởng về việc hợp nhất các loại dữ liệu khác nhau dường như là điều tham vọng nhưng khó thực hiện. Hôm nay - điều này đã trở thành một phần của thực tiễn hàng ngày. Việc tích hợp dữ liệu của các sơ đồ và cấu trúc khác nhau đã trở thành một phần không thể thiếu trong kiến trúc hệ thống thông tin hiện đại.

Ở các chương tiếp theo, chúng tôi sẽ xem xét chi tiết các tiêu chuẩn và phương pháp chính cho phép hợp nhất dữ liệu cấu trúc, bán cấu trúc và không cấu trúc thành một biểu diễn đồng nhất. Đặc biệt chú ý sẽ được dành cho dữ liệu cấu trúc và cơ sở dữ liệu quan hệ - như là các cơ chế chính để lưu trữ, xử lý và phân tích thông tin trong ngành xây dựng.

Dữ liệu cấu trúc

Trong ngành xây dựng, thông tin đến từ nhiều nguồn - bản vẽ, đặc điểm kỹ thuật, lịch trình và báo cáo. Để quản lý dòng chảy này một cách hiệu quả, cần có sự cấu trúc thông tin. Dữ liệu cấu trúc cho phép tổ chức thông tin theo cách thuận tiện, dễ đọc và dễ tiếp cận.

> Theo báo cáo thường niên lần thứ 5 về công nghệ xây dựng của JB Knowledge [17], 67% chuyên gia quản lý dự án xây dựng theo dõi và đánh giá hiệu suất công việc bằng cách thủ công hoặc với sự trợ giúp của bảng điện tử.

Một trong các định dạng dữ liệu cấu trúc phổ biến nhất là XLSX và CSV. Chúng được sử dụng rộng rãi để lưu trữ, xử lý và phân tích thông tin trong các bảng điện tử. Trong các bảng này, dữ liệu được trình bày dưới dạng hàng và cột, điều này làm cho chúng trở nên thuận tiện cho việc đọc, chỉnh sửa và phân tích.

Định dạng XLSX, được phát triển bởi Microsoft, dựa trên việc sử dụng cấu trúc XML và được nén bằng thuật toán ZIP. Các đặc điểm chính của định dạng:

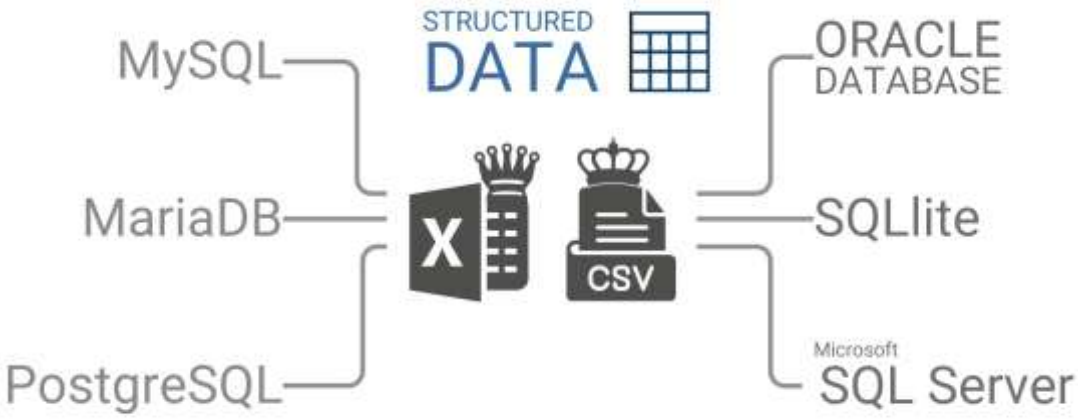
- Hỗ trợ các công thức phức tạp, biểu đồ và macro.
- Khả năng lưu trữ dữ liệu trên nhiều trang, cũng như định dạng thông tin.
- Tối ưu hóa cho việc làm việc trong môi trường Microsoft Excel, nhưng cũng tương thích với các gói văn phòng khác.

Định dạng CSV là một tệp văn bản đơn giản, trong đó các giá trị được phân tách bằng dấu phẩy, dấu chấm phẩy hoặc các ký tự phân tách khác. Các lợi ích chính:

- Tính tương thích phổ quát với các chương trình và hệ điều hành khác nhau.
- Tiện lợi trong việc nhập/xuất dữ liệu vào các cơ sở dữ liệu và hệ thống phân tích.
- Dễ dàng xử lý ngay cả trong các trình soạn thảo văn bản.

Tuy nhiên, CSV không hỗ trợ các công thức và định dạng, vì vậy ứng dụng chính của nó là trao đổi dữ liệu giữa các hệ thống và cập nhật thông tin hàng loạt. Nhờ vào tính phổ quát và độc lập với nền tảng, CSV đã trở thành công cụ chuyển dữ liệu phổ biến trong các môi trường CNTT đa dạng.

Cả hai định dạng XLSX và CSV đều hoạt động như một cầu nối giữa các hệ thống khác nhau làm việc với dữ liệu có cấu trúc (Hình 3.14). Chúng đặc biệt hữu ích trong những nhiệm vụ mà độ đọc được, chỉnh sửa thủ công và khả năng tương thích cơ bản là quan trọng.-



Hình 3.14 Các định dạng XLSX và CSV tạo thành cầu nối giữa các hệ thống khác nhau làm việc với dữ liệu có cấu trúc.

> Tính độc lập với nền tảng làm cho CSV trở thành định dạng phổ biến nhất cho việc truyền tải dữ liệu trong các môi trường và hệ thống CNTT đa dạng.

Tuy nhiên, XLSX và CSV không được thiết kế cho các tính toán hiệu suất cao hoặc lưu trữ lâu dài lượng dữ liệu lớn. Để phục vụ cho các mục đích này, người ta thường sử dụng các định dạng cấu trúc hiện đại hơn, như Apache Parquet, Apache ORC, Feather, HDF5. Những định dạng này sẽ được chúng tôi xem xét chi tiết trong chương "Lưu trữ dữ liệu lớn: phân tích các định dạng phổ biến và hiệu quả của chúng" trong phần chín của cuốn sách.

Trên thực tế, Excel với định dạng XLSX thường được sử dụng cho các tác vụ nhỏ và tự động hóa quy trình hàng ngày. Các kịch bản phức tạp hơn yêu cầu sử dụng các hệ thống quản lý dữ liệu, như ERP, PMIS CAFM, CPM, SCM và những hệ thống khác (Hình 3.21). Chính trong những hệ thống này, dữ liệu có cấu trúc được lưu trữ, do đó cơ sở nền tảng cho tổ chức và quản lý thông tin của công ty.-

Các hệ thống thông tin quản lý dữ liệu hiện đại, được ứng dụng trong ngành xây dựng, dựa vào dữ liệu có cấu trúc được tổ chức dưới dạng bảng. Để quản lý đáng tin cậy, mở rộng và toàn diện các khối lượng thông tin lớn, các nhà phát triển ứng dụng và hệ thống thường sử dụng các hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS).

Cơ sở dữ liệu quan hệ RDBMS và ngôn ngữ truy vấn SQL

Để lưu trữ, xử lý và phân tích dữ liệu một cách hiệu quả, người ta sử dụng cơ sở dữ liệu quan hệ (RDBMS) — các hệ thống lưu trữ thông tin mà tổ chức dữ liệu thành các bảng có mối quan hệ xác định giữa chúng.

> Dữ liệu được tổ chức trong cơ sở dữ liệu (RDBMS) không chỉ đơn thuần là thông tin số; mà chúng còn là nền tảng cho các giao dịch và tương tác giữa các hệ thống khác nhau.

Dưới đây là một số hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS) phổ biến nhất (Hình 3.15):-

- MySQL (Mã nguồn mở) – một trong những RDBMS phổ biến nhất, là một phần của stack LAMP (Linux, Apache, MySQL, PHP/Perl/Python). Được sử dụng rộng rãi trong phát triển web nhờ vào tính đơn giản và hiệu suất cao.
- PostgreSQL (Mã nguồn mở) – một hệ thống đối tượng quan hệ mạnh mẽ, nổi tiếng với độ tin cậy và khả năng mở rộng. Phù hợp cho các giải pháp doanh nghiệp phức tạp.
- Microsoft SQL Server – hệ thống thương mại từ Microsoft, được sử dụng rộng rãi trong môi trường doanh nghiệp nhờ vào việc tích hợp với các sản phẩm khác của công ty và mức độ bảo mật cao.
- Oracle Database – một trong những hệ quản trị cơ sở dữ liệu (CSDL) mạnh mẽ và đáng tin cậy nhất, được áp dụng trong các doanh nghiệp lớn và các ứng dụng quan trọng.
- IBM DB2 – hướng tới các tập đoàn lớn, cung cấp hiệu suất cao và khả năng chịu lỗi.
- SQLite (Mã nguồn mở) – cơ sở dữ liệu nhúng nhẹ, hoàn hảo cho các ứng dụng di động và hệ thống độc lập, chẳng hạn như các chương trình thiết kế CAD (BIM).

Các hệ thống quản lý cơ sở dữ liệu phổ biến trong ngành xây dựng - MySQL, PostgreSQL, Microsoft SQL Server, Oracle® Database, IBM® DB2 và SQLite - làm việc với các dữ liệu có cấu trúc. Tất cả các CSDL này đều là những giải pháp mạnh mẽ và linh hoạt để quản lý một loạt các quy trình kinh doanh và ứng dụng, từ các trang web nhỏ đến các hệ thống doanh nghiệp quy mô lớn (Hình. 3.21). -

> Theo dữ liệu từ Statista [48], vào năm 2022, các hệ thống quản trị cơ sở dữ liệu quan hệ (RDBMS) chiếm khoảng 72% tổng số CSDL đang được sử dụng.



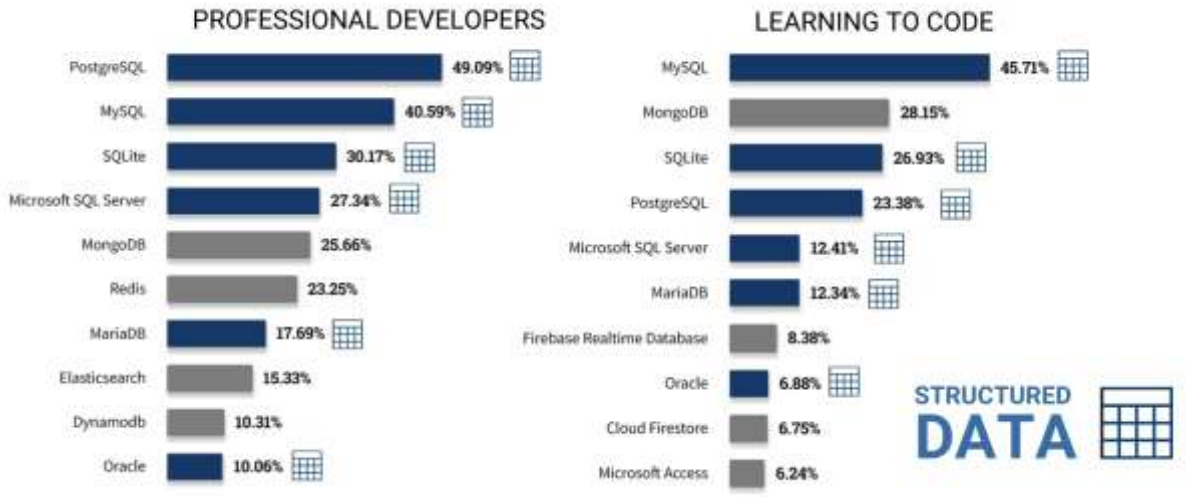
Rank			DBMS	Database Model	Open Source vs Commercial
Mar2025	Feb2025	Mar2024			
1.	1.	1.	**Oracle®**	Relational, Multi-model	Commercial
2.	2.	2.	**MySQL**	Relational, Multi-model	**Open Source**
3.	3.	3.	**Microsoft® SQL Server**	Relational, Multi-model	Commercial
4.	4.	4.	**PostgreSQL**	Relational, Multi-model	**Open Source**
5.	5.	5.	MongoDB	Document, Multi-model	**Open Source**
6.	7.	9.	**Snowflake®**	Relational	Commercial
7.	6.	6.	Redis®	Key-value, Multi-model	**Open Source**
8.	8.	7.	Elasticsearch®	Multi-model	**Open Source**
9.	9.	8.	**IBM Db2**	Relational, Multi-model	Commercial
10.	10.	10.	**SQLite**	Relational	**Open Source**
11.	11.	12.	Apache Cassandra®	Multi-model	**Open Source**
12.	12.	11.	**Microsoft Access®**	Relational	**Open Source**
13.	13.	17.	Databricks®	Multi-model	Commercial
14.	14.	13.	**MariaDB**	Relational, Multi-model	**Open Source**
15.	15.	14.	Splunk	Search engine	Commercial
16.	16.	16.	Amazon DynamoDB	Multi-model	Commercial
17.	17.	15.	**Microsoft Azure SQL**	Relational, Multi-model	Commercial

Hình. 3.15 Sự phổ biến của việc sử dụng cơ sở dữ liệu có cấu trúc (được đánh dấu bằng màu xanh) trong bảng xếp hạng CSDL (theo tài liệu [49]).

Việc cài đặt các cơ sở dữ liệu mã nguồn mở có thể thực hiện khá đơn giản - thậm chí không cần kiến thức kỹ thuật sâu. Các hệ thống mã nguồn mở như PostgreSQL, MySQL hay SQLite đều miễn phí và hoạt động trên hầu hết các hệ điều hành: Windows, macOS và Linux. Tất cả những gì bạn cần làm là truy cập vào trang chính thức của dự án, tải xuống bộ cài đặt và làm theo hướng dẫn. Trong hầu hết các trường hợp, việc cài đặt mất không quá 10-15 phút. Một trong những cơ sở dữ liệu này sẽ được mô phỏng và tạo ra trong phần bốn của cuốn sách (Hình. 4.38).-

Nếu công ty của bạn sử dụng các dịch vụ đám mây (ví dụ như Amazon Web Services, Google Cloud hoặc Microsoft Azure), thì việc triển khai cơ sở dữ liệu có thể thực hiện chỉ bằng vài cú nhấp chuột - nền tảng sẽ cung cấp cho bạn các mẫu cài đặt sẵn. Nhờ vào mã nguồn mở, các cơ sở dữ liệu này dễ dàng được tùy chỉnh theo nhu cầu của mình, và cộng đồng người dùng to lớn luôn sẵn sàng giúp tìm ra giải pháp cho bất kỳ vấn đề nào.

RDBMS vẫn là nền tảng cho nhiều ứng dụng doanh nghiệp và nền tảng phân tích (Hình. 3.16), cho phép các công ty lưu trữ, xử lý và phân tích dữ liệu hiệu quả - có nghĩa là thực hiện quyết định dựa trên thông tin chính xác và kịp thời.-



Hình. 3.16 Khảo sát của các lập trình viên trên StackOverFlow (diễn đàn CNTT lớn nhất) về các cơ sở dữ liệu họ đã sử dụng trong năm ngoái và những gì họ muốn sử dụng trong năm tới (RDBMS được đánh dấu bằng màu xanh) (theo tài liệu [50]).

RDBMS đảm bảo độ tin cậy, tính toàn vẹn dữ liệu, hỗ trợ giao dịch và sử dụng ngôn ngữ truy vấn mạnh mẽ - SQL (Structured Query Language), thường được sử dụng trong phân tích và cho phép dễ dàng lấy, sửa đổi và phân tích thông tin được lưu trữ trong cơ sở dữ liệu. Chính SQL là công cụ chính để làm việc với dữ liệu trong các hệ thống quan hệ.

Các truy vấn SQL trong cơ sở dữ liệu và xu hướng mới

Lợi thế chính của ngôn ngữ SQL, thường được sử dụng trong các cơ sở dữ liệu quan hệ, so với các phương pháp quản lý thông tin khác (chẳng hạn như bảng điện tử cổ điển Excel) là khả năng hỗ trợ các khối lượng cơ sở dữ liệu rất lớn với tốc độ xử lý truy vấn cao.

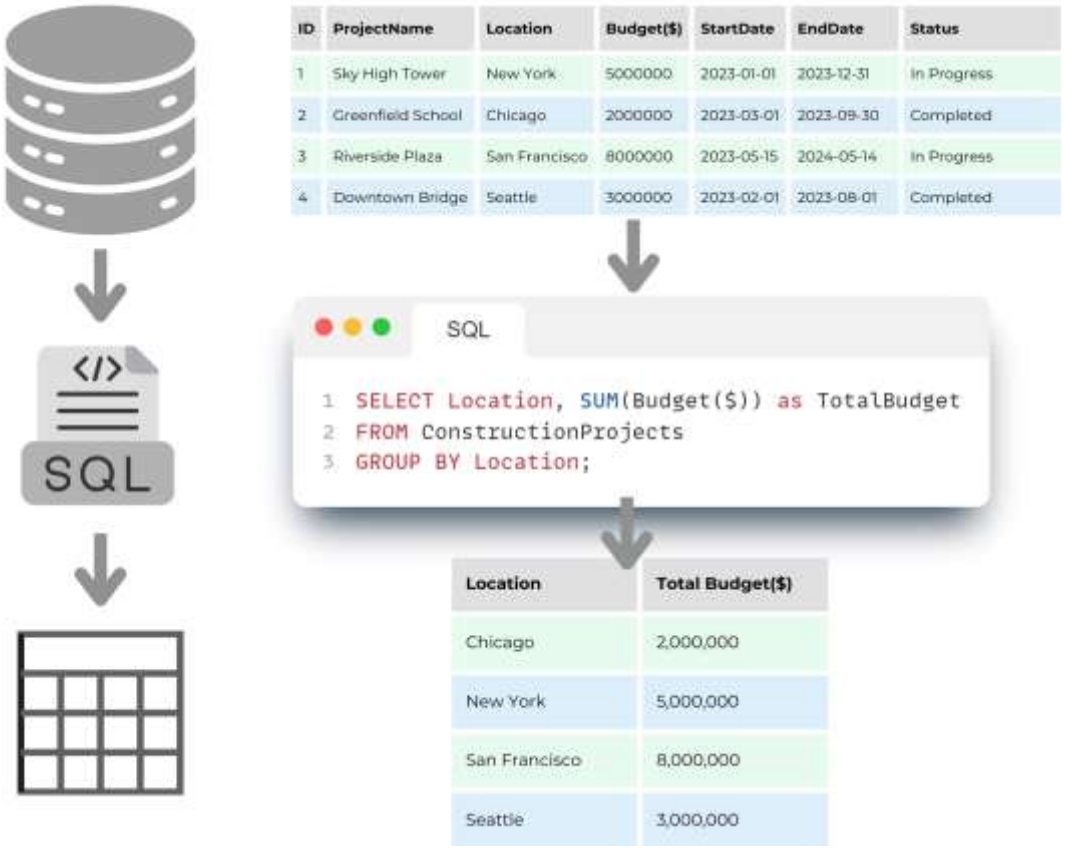
> Ngôn ngữ truy vấn có cấu trúc (SQL) – là một ngôn ngữ lập trình chuyên dụng, nhằm mục đích lưu trữ, xử lý và phân tích thông tin trong các cơ sở dữ liệu quan hệ. SQL được sử dụng để tạo ra, quản lý và truy cập vào dữ liệu, cho phép tìm kiếm, lọc, kết hợp và tổng hợp thông tin một cách hiệu quả. Nó phục vụ như một công cụ chính để truy cập dữ liệu, cung cấp phương pháp tương tác thuận tiện và có tính hình thức với các kho lưu trữ thông tin.

Sự phát triển của các hệ thống SEQUEL-SQL trải qua nhiều sản phẩm và công ty nổi bật như Oracle, IBM DB2, Microsoft SQL Server, SAP, PostgreSQL và MySQL, kết thúc bằng sự xuất hiện của SQLite và MariaDB. SQL cung cấp khả năng làm việc với các bảng mà Excel không có, làm cho việc làm việc với dữ liệu trở nên quy mô hơn, an toàn hơn và dễ dàng hơn cho tự động hóa:

- Tạo dựng và quản lý cấu trúc dữ liệu (DDL): trong SQL, bạn có thể tạo, thay đổi và xóa các bảng trong cơ sở dữ liệu, thiết lập các mối quan hệ giữa chúng và xác định cấu trúc lưu trữ dữ liệu. Trong khi đó, trong Excel, công việc diễn ra với các trang và ô cố định, không có các mối liên kết rõ ràng giữa các trang và tập dữ liệu.
- Các thao tác với dữ liệu (DML): SQL cho phép thêm, thay đổi, xóa và truy xuất dữ liệu hàng loạt với

tốc độ cao, thực hiện các truy vấn phức tạp với lọc, sắp xếp và kết hợp các bảng. Trong Excel, việc xử lý khối lượng lớn thông tin đòi hỏi thao tác thủ công hoặc các macro đặc biệt, điều này làm chậm quá trình và tăng khả năng xảy ra lỗi.-

- Kiểm soát quyền truy cập (DCL): SQL cho phép phân chia quyền truy cập vào dữ liệu cho các người dùng khác nhau, hạn chế khả năng chỉnh sửa hoặc xem thông tin. Trong khi đó, trong Excel, quyền truy cập thường là chung (khi truyền file), hoặc yêu cầu các thiết lập phức tạp với việc chia quyền qua các dịch vụ đám mây.



Ví dụ DML trong SQL: xử lý nhanh chóng, nhóm và tổng hợp thông qua một vài dòng mã để tự động xử lý dữ liệu.

Excel giúp việc làm việc với dữ liệu trở nên dễ dàng nhờ vào cấu trúc trực quan và dễ hiểu của nó. Tuy nhiên, khi khối lượng dữ liệu tăng lên, hiệu suất của Excel giảm sút. Excel cũng gặp phải những giới hạn về khối lượng dữ liệu lưu trữ – tối đa một triệu dòng, và hiệu suất sẽ giảm sút trước khi đạt đến giới hạn đó. Do đó, mặc dù Excel có vẻ ưu việt hơn cho việc hình ảnh hóa và thao tác với các khối lượng dữ liệu nhỏ, nhưng để làm việc với các tập dữ liệu lớn, SQL là sự lựa chọn tốt hơn.

Bước tiến tiếp theo trong sự phát triển của dữ liệu có cấu trúc là sự xuất hiện của các cơ sở dữ liệu theo cột (Columnar Databases), đây là một sự thay thế cho các cơ sở dữ liệu quan hệ truyền thống, đặc biệt khi nói đến các khối lượng dữ liệu lớn hơn nhiều và các phép toán phân tích. Khác với các hệ quản trị cơ sở dữ liệu theo hàng, nơi dữ liệu được lưu trữ theo hàng, trong các cơ sở dữ liệu theo cột, thông tin được ghi theo cột. So với các cơ sở dữ liệu cổ điển, điều này cho phép:

- Giảm thiểu khối lượng lưu trữ nhờ vào việc nén hiệu quả các dữ liệu đồng loại trong các cột.

- Tăng tốc độ các truy vấn phân tích, vì chỉ có các cột cần thiết được đọc chứ không phải toàn bộ bảng.
- Tối ưu hóa việc làm việc với Big Data và các kho dữ liệu, chẳng hạn như trong kiến trúc Data Lakehouse.

Chúng ta sẽ nói chi tiết hơn về các cơ sở dữ liệu theo cột, Pandas DataFrame, Apache Parquet, HDF5, cũng như việc tạo ra các kho dữ liệu Big Data dựa trên chúng để phục vụ cho việc phân tích và xử lý dữ liệu trong các chương tiếp theo của cuốn sách này – "DataFrame: định dạng phổ quát cho dữ liệu bảng" và "Các định dạng lưu trữ dữ liệu và làm việc với Apache Parquet: kho dữ liệu DWH và kiến trúc Data Lakehouse".

Dữ liệu phi cấu trúc

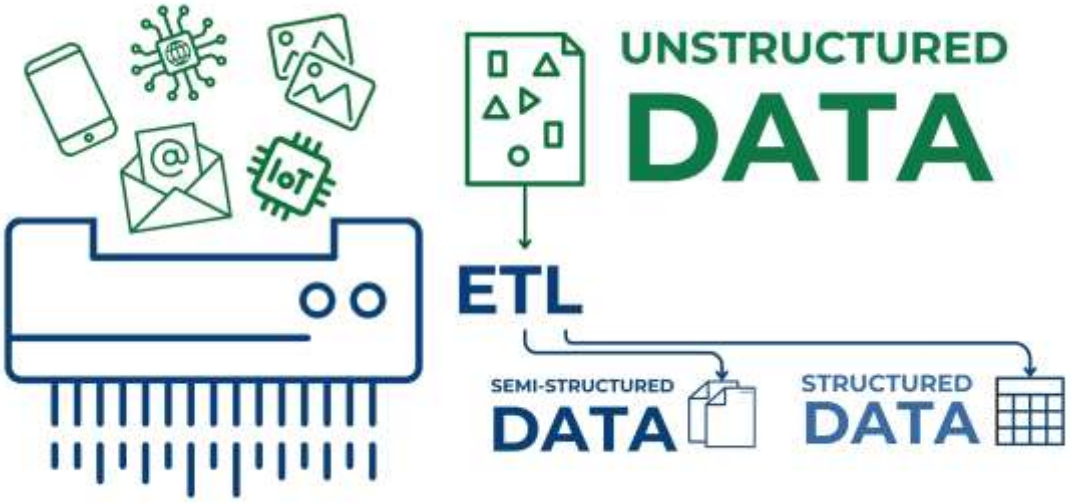
Mặc dù hầu hết dữ liệu được sử dụng trong các ứng dụng và hệ thống thông tin có dạng có cấu trúc, phần lớn thông tin được tạo ra trong ngành xây dựng lại được trình bày dưới dạng dữ liệu không có cấu trúc — hình ảnh, video, tài liệu văn bản, ghi âm và các hình thức nội dung khác. Điều này đặc biệt đúng ở giai đoạn xây dựng, vận hành và giám sát kỹ thuật, nơi mà thông tin hình ảnh và văn bản chiếm ưu thế.

> Dữ liệu không có cấu trúc là thông tin không có mô hình hoặc cấu trúc đã định sẵn, không được tổ chức thành những dòng và cột truyền thống như trong cơ sở dữ liệu hoặc bảng.

Dưới dạng tổng quát, dữ liệu không có cấu trúc có thể được phân loại thành hai loại:

- Dữ liệu không có cấu trúc do con người tạo ra, bao gồm nhiều loại nội dung do con người sáng tạo: tài liệu văn bản, email, hình ảnh, video và các loại tương tự.
- Dữ liệu không có cấu trúc do máy tạo ra được tạo ra từ các thiết bị và cảm biến: đây là các tệp nhật ký, dữ liệu GPS, kết quả hoạt động của Internet of Things (IoT) và, chẳng hạn, các thông tin viễn thông khác từ công trường xây dựng.

Khác với dữ liệu có cấu trúc, được tổ chức một cách thuận tiện trong các bảng và cơ sở dữ liệu, dữ liệu không có cấu trúc yêu cầu các bước xử lý bổ sung trước khi được tích hợp vào các hệ thống thông tin. Việc sử dụng các công nghệ thu thập, phân tích và chuyển đổi tự động cho những dữ liệu này mở ra nhiều cơ hội mới để nâng cao hiệu quả xây dựng, giảm thiểu sai sót và tối thiểu hóa ảnh hưởng của yếu tố con người.



Xử lý dữ liệu không có cấu trúc bắt đầu bằng việc chuyển đổi chúng thành dữ liệu bán cấu trúc và có cấu trúc.

Dữ liệu không có cấu trúc chiếm đến 80% toàn bộ thông tin mà các chuyên gia trong các công ty gặp phải, do đó, chúng tôi sẽ xem xét chi tiết với các ví dụ về các loại và cách xử lý chúng trong các chương tiếp theo của cuốn sách.

Để thuận tiện cho cuộc thảo luận, chúng tôi sẽ tách dữ liệu văn bản thành một loại riêng. Mặc dù chúng là một loại dữ liệu không có cấu trúc, nhưng tầm quan trọng và sự phổ biến của chúng trong ngành xây dựng yêu cầu phải được chú ý đặc biệt.

Dữ liệu văn bản: giữa hỗn loạn phi cấu trúc và cấu trúc

Dữ liệu văn bản trong ngành xây dựng bao phủ một loạt các định dạng và loại thông tin, từ tài liệu giấy cho đến các phương thức giao tiếp không chính thức như thư từ, cuộc trò chuyện, thư trao đổi công việc và các cuộc họp trực tiếp trên công trường. Tất cả những dữ liệu văn bản này đều có thông tin quan trọng cho việc quản lý các dự án xây dựng - từ chi tiết các quyết định thiết kế và thay đổi kế hoạch cho đến thảo luận về các vấn đề an toàn và thương lượng với các nhà thầu và khách hàng.-

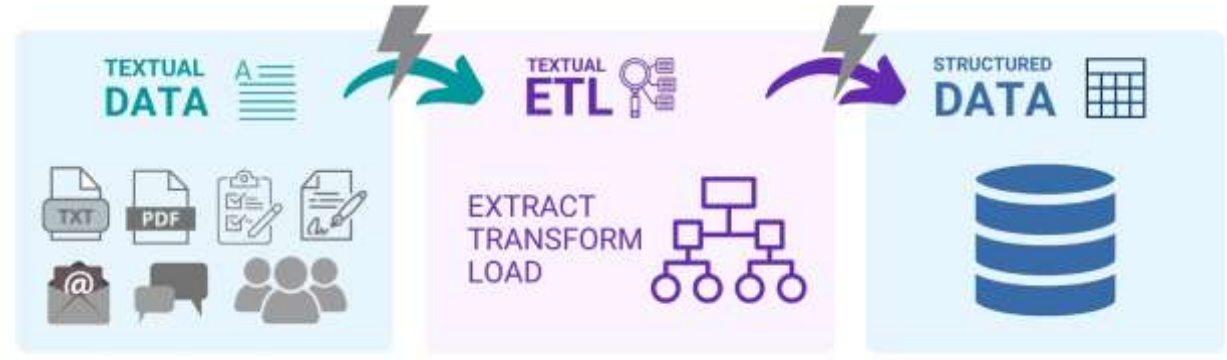


Dữ liệu văn bản là một trong những loại thông tin phổ biến nhất được sử dụng trong giao tiếp giữa các thành viên trong dự án.

Thông tin văn bản có thể là cả dạng chính thức và không chính thức. Các dữ liệu chính thức bao gồm tài liệu định dạng Word (.doc,.docx), PDF, cũng như các tệp văn bản biên bản cuộc họp (.txt). Dữ liệu không chính thức bao gồm trao đổi qua các ứng dụng nhắn tin và email, biên bản các cuộc họp (Teams, Zoom, Google Meet), cũng như ghi âm các cuộc thảo luận (.mp3,.wav), mà cần được chuyển đổi thành văn bản.

Nhưng nếu các tài liệu bằng văn bản, như yêu cầu chính thức, điều kiện hợp đồng và thông điệp điện tử, thường đã có một cấu trúc nhất định, thì các thông điệp miệng và thư từ công việc thường vẫn còn không có cấu trúc, điều này làm cho việc phân tích và tích hợp chúng vào hệ thống quản lý dự án trở nên khó khăn.

Chìa khóa để quản lý dữ liệu văn bản một cách hiệu quả là chuyển đổi chúng thành định dạng có cấu trúc. Điều này cho phép tự động tích hợp thông tin đã xử lý vào các hệ thống hiện có, vốn đã làm việc với dữ liệu có cấu trúc.



Hình 3.110 Chuyển đổi nội dung văn bản thành dữ liệu có cấu trúc.

Để sử dụng thông tin văn bản một cách hiệu quả, cần tự động chuyển đổi nó thành dạng có cấu trúc (Hình 3.110). Quy trình này thường bao gồm một vài bước:-

- Nhận diện văn bản (OCR) — chuyển đổi hình ảnh của tài liệu và bản vẽ thành định dạng có thể đọc bằng máy.
- Phân tích văn bản (NLP) — tự động xác định các tham số chính (ngày tháng, số tiền và số liệu liên quan đến dự án).
- Phân loại dữ liệu — phân bổ thông tin vào các danh mục (tài chính, logistics, quản lý rủi ro).

Sau khi nhận diện và phân loại, dữ liệu đã có cấu trúc có thể được tích hợp vào cơ sở dữ liệu và sử dụng trong các hệ thống báo cáo và quản lý tự động.

Dữ liệu bán cấu trúc và dữ liệu yếu cấu trúc

Dữ liệu bán cấu trúc chứa một mức độ tổ chức nhất định, nhưng không có sơ đồ hoặc cấu trúc nghiêm ngặt. Mặc dù thông tin như vậy bao gồm các yếu tố có cấu trúc (ví dụ: ngày tháng, tên nhân viên và danh sách nhiệm vụ đã hoàn thành), định dạng trình bày có thể khác nhau đáng kể trong các dự án hoặc thậm chí đối với từng nhân viên. Các ví dụ về dữ liệu như vậy bao gồm nhật ký ghi chép thời gian làm việc, báo cáo về công việc đã hoàn thành và biểu đồ, có thể được trình bày dưới nhiều định dạng khác nhau.

Dữ liệu bán cấu trúc dễ phân tích hơn dữ liệu không cấu trúc, tuy nhiên chúng yêu cầu xử lý thêm để tích hợp vào các hệ thống quản lý dự án tiêu chuẩn hóa.

> Việc làm việc với dữ liệu bán cấu trúc, đặc trưng bởi sự thay đổi cấu trúc liên tục, gây ra rất nhiều khó khăn. Điều này liên quan đến việc tính biến đổi của cấu trúc dữ liệu đòi hỏi phải có những cách tiếp cận riêng biệt để xử lý và phân tích mỗi nguồn dữ liệu bán cấu trúc.

Nhưng nếu việc làm việc với dữ liệu không cấu trúc yêu cầu nhiều nỗ lực, thì việc xử lý dữ liệu bán cấu trúc có thể được thực hiện bằng các phương pháp và công cụ tương đối đơn giản.

> Dữ liệu yếu cấu trúc — là một thuật ngữ chung mô tả dữ liệu có cấu trúc tối thiểu hoặc không đầy đủ. Thường là các tài liệu văn bản, trò chuyện, email, nơi có một số siêu dữ liệu (ví dụ: ngày tháng, người gửi), nhưng phần lớn thông tin được trình bày một cách lộn xộn.

Trong xây dựng, dữ liệu yếu cấu trúc xuất hiện trong các quy trình khác nhau. Ví dụ, chúng có thể bao gồm:

- Tính toán chi phí và đề xuất thương mại – bảng dữ liệu về nguyên vật liệu, khối lượng và chi phí, nhưng không có định dạng thống nhất.
- Bản vẽ và sơ đồ kỹ thuật – các tệp PDF hoặc DWG chứa chú thích văn bản và siêu dữ liệu, nhưng không có cấu trúc cố định.
- Biểu đồ thi công – dữ liệu từ MS Project, Primavera P6 hoặc các hệ thống khác, có thể có cấu trúc xuất khác nhau.
- CAD (mô hình BIM) – chứa các yếu tố cấu trúc, nhưng cách trình bày dữ liệu phụ thuộc vào phần mềm và tiêu chuẩn dự án.

Dữ liệu hình học được tạo ra bởi các hệ thống CAD có thể được phân loại giống như dữ liệu bán cấu trúc. Tuy nhiên, chúng ta sẽ tách riêng dữ liệu CAD (BIM) hình học thành một loại riêng, bởi vì chúng, giống như dữ liệu văn bản, có thể thường được xem như là một loại dữ liệu riêng biệt trong các quy trình của công ty.

Dữ liệu hình học và ứng dụng của chúng

Nếu siêu dữ liệu về các yếu tố của dự án gần như luôn được lưu trữ dưới dạng bảng, định dạng có cấu trúc hoặc bán cấu trúc, thì dữ liệu hình học của các yếu tố dự án trong hầu hết các trường hợp được tạo ra bằng cách sử dụng các công cụ CAD chuyên dụng, cho phép hình dung chi tiết các yếu tố của dự án dưới dạng một tập hợp các đường (2D) hoặc hình khối (3D).-

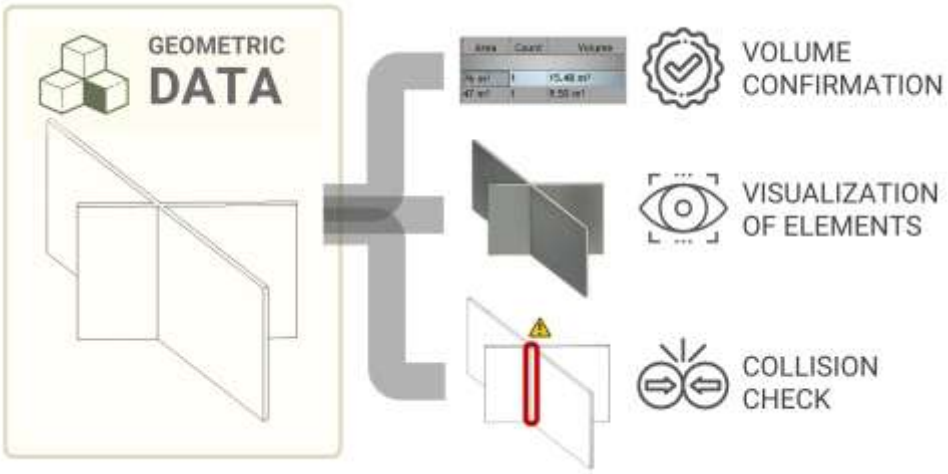


Các công cụ CAD đã giúp chuyển đổi thông tin hình học từ các phương tiện vật lý sang một dạng cơ sở dữ liệu.

Khi làm việc với dữ liệu hình học trong xây dựng và kiến trúc, có thể phân biệt ba lĩnh vực chính trong việc ứng dụng dữ liệu hình học:-

- Xác nhận thể tích: dữ liệu hình học, được tạo ra bên trong các phần mềm CAD (BIM) qua các lõi hình học chuyên dụng, cần thiết để xác định thể tích và kích thước của các yếu tố dự án một cách tự động và chính xác. Dữ liệu này bao gồm diện tích, thể tích, chiều dài và các thuộc tính quan trọng khác được tính toán tự động, cần thiết cho lập kế hoạch, dự toán ngân sách và đặt hàng tài nguyên và vật liệu.
- Hình dung dự án: trong trường hợp có bất kỳ thay đổi nào trong dự án, việc hình dung các yếu tố cho phép tự động tạo ra các bản vẽ cập nhật ở các mặt phẳng khác nhau. Hình dung dự án ở giai đoạn đầu cho phép tăng tốc sự hiểu biết giữa tất cả các bên tham gia, nhằm tiết kiệm thời gian và tài nguyên trong quá trình xây dựng.
- Kiểm tra va chạm: trong các dự án xây dựng và kỹ thuật phức tạp, nơi mà sự tương tác của nhiều loại yếu tố (ví dụ, ống dẫn và tường) mà không có "xung đột hình học" là vô cùng quan trọng, việc kiểm tra va chạm đóng vai trò then chốt. Sử dụng phần mềm phát hiện va chạm cho phép phát hiện trước những xung đột hình học tiềm năng giữa các yếu tố của dự án, ngăn ngừa những sai sót tốn kém trong quá trình xây dựng.

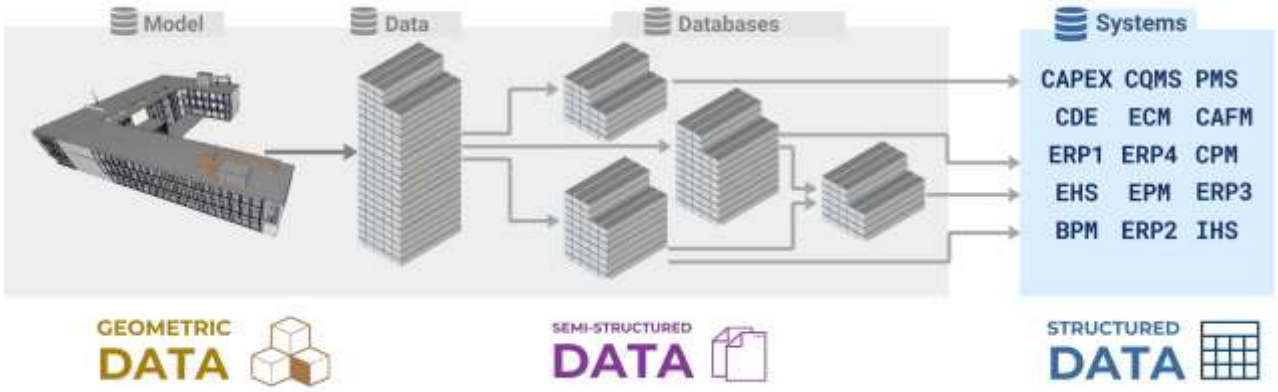
Kể từ khi bắt đầu xuất hiện các văn phòng kỹ thuật thiết kế, từ thời điểm xây dựng các công trình phức tạp đầu tiên, các kỹ sư thiết kế đã cung cấp thông tin hình học dưới dạng bản vẽ, đường và các yếu tố hình học phẳng (trên giấy papyrus, giấy vẽ “A0” hoặc trong các định dạng DWG, PDF, PLT), dựa trên đó các kỹ sư thi công và lập dự toán, trong suốt hàng ngàn năm, đã sử dụng thước và thước đo để thu thập các thể tích thuộc tính hoặc số lượng các yếu tố và nhóm yếu tố. -



Hình học là cơ sở để thu nhận các tham số thể tích của các yếu tố, mà sau đó được sử dụng để tính toán chi phí và thời gian của dự án.

Ngày nay, nhiệm vụ thủ công và tốn công này đã được giải quyết thông qua việc tự động hóa hoàn toàn nhờ vào sự xuất hiện của mô hình thể tích trong các công cụ CAD (BIM) hiện đại, cho phép tự động nhận được các thuộc tính thể tích của bất kỳ yếu tố nào mà không cần tính toán các tham số thể tích một cách thủ công.

Các công cụ CAD hiện đại cũng cho phép phân loại và phân nhóm các phần tử của dự án, để có thể xuất các bảng đặc tả từ cơ sở dữ liệu của dự án để sử dụng trong các hệ thống khác nhau, chẳng hạn như đánh giá chi phí, lập lịch trình hoặc tính toán CO2 (Hình 3.113). Chúng ta sẽ nói về việc nhận được các đặc tả, bảng QTO và khối lượng, cũng như các ví dụ thực tế trong chương "Nhận khối lượng và tính toán định lượng".-



Hình 3.113 Các công cụ CAD (BIM) lưu trữ dữ liệu trong các cơ sở dữ liệu, được thiết kế để tích hợp và tương tác với các hệ thống khác.

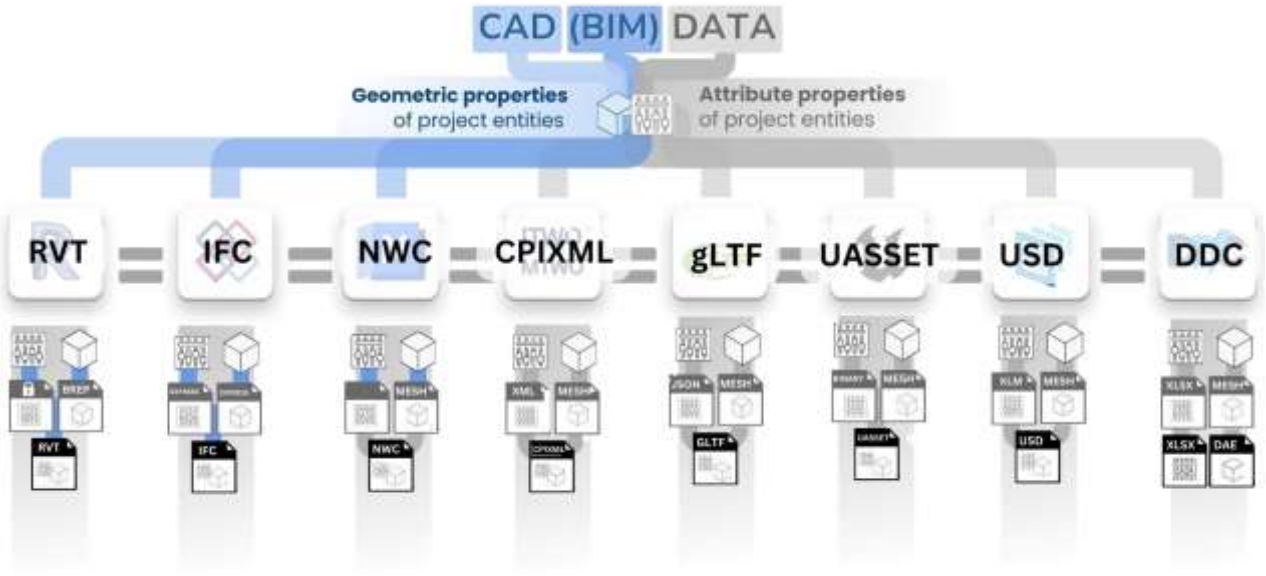
Do tính đóng kín của các cơ sở dữ liệu và định dạng sử dụng trong môi trường CAD, dữ liệu hình học được tạo ra trong các giải pháp CAD thực sự đã hình thành thành một loại thông tin riêng biệt. Nó kết hợp cả hình học của các phần tử và siêu thông tin (được cấu trúc hoặc bán cấu trúc), được lưu trữ trong các tệp và định dạng chuyên dụng.

Dữ liệu CAD: từ thiết kế đến lưu trữ dữ liệu

Trong các hệ thống CAD và BIM hiện đại, dữ liệu được lưu trữ trong các định dạng độc quyền riêng, thường là định dạng proprietary: DWG, DXF, RVT, DGN, PLN và các định dạng khác. Những định dạng này hỗ trợ cả bản vẽ 2D và 3D, đồng thời lưu giữ không chỉ hình học mà còn các thuộc tính liên quan đến các đối tượng. Dưới đây là một số phổ biến nhất:

- DWG là định dạng tệp nhị phân, được sử dụng để lưu trữ dữ liệu thiết kế 2D (và hiếm khi 3D) và siêu dữ liệu.
- DXF là định dạng văn bản để trao đổi bản vẽ 2D và 3D giữa các hệ thống CAD. Nó chứa hình học, lớp và dữ liệu thuộc tính, hỗ trợ cả định dạng ASCII và nhị phân.
- RVT là định dạng lưu trữ nhị phân cho các mô hình CAD, bao gồm hình học 3D, thuộc tính của các phần tử, liên kết và tham số của dự án.
- IFC là định dạng văn bản mở để trao đổi dữ liệu xây dựng giữa các hệ thống CAD (BIM). Nó bao gồm hình học, thuộc tính của các đối tượng và thông tin về mối liên hệ của chúng.

Ngoài các định dạng đó, còn có các định dạng khác được sử dụng: PLN, DB1, SVF, NWC, CPIXML, BLEND, BX3, USD, XLSX, DAE. Mặc dù chúng khác nhau về mục đích và mức độ mở (Hình 3.114), tất cả đều có thể đại diện cho cùng một mô hình thông tin của dự án dưới các hình thức khác nhau. Trong các dự án phức tạp, các định dạng này thường được sử dụng song song — từ thiết kế đến phối hợp các mô hình của dự án.



Hình 3.114 Các định dạng phổ biến để lưu trữ thông tin từ CAD mô tả hình học thông qua các tham số BREP hoặc MESH, bổ sung cho chúng bằng dữ liệu thuộc tính.

Tất cả các định dạng được liệt kê trên cho phép lưu trữ dữ liệu về từng phần tử của dự án xây dựng và trong tất cả các định dạng đã đề cập đều có hai loại dữ liệu chính:

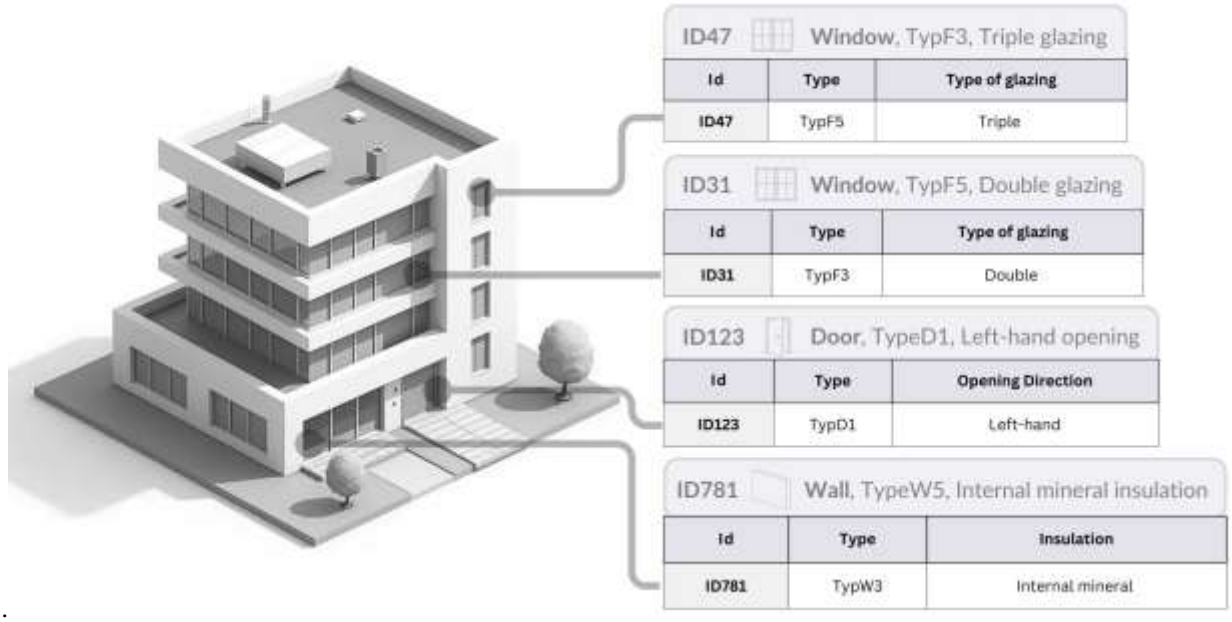
- Các tham số hình học — mô tả hình dạng, vị trí và kích thước của đối tượng. Chi tiết về hình học và việc sử dụng sẽ được thảo luận trong phần sáu của sách, dành riêng cho các giải pháp CAD (BIM);
- Các thuộc tính thuộc tính — chứa thông tin khác nhau: về vật liệu, loại phần tử, thông số kỹ thuật, các định danh duy nhất và các thuộc tính khác mà các phần tử của dự án có thể có.

Trong các dự án hiện đại, dữ liệu thuộc tính trở nên đặc biệt quan trọng, vì chúng xác định các đặc điểm vận hành của các đối tượng, cho phép thực hiện các tính toán kỹ thuật, tài chính và đảm bảo tương tác liên tục giữa các bên tham gia thiết kế, xây dựng và vận hành. Ví dụ:

- Đối với cửa sổ và cửa ra vào, cần chỉ định: loại kết cấu, kiểu kính, hướng mở (Hình 3.21).
- Đối với tường, thông tin về vật liệu, cách nhiệt và đặc tính âm học được ghi lại.
- Đối với các hệ thống kỹ thuật, các thông số của ống dẫn, ống gió, đường cáp và các mối nối của chúng được lưu trữ.

Những thông số này có thể được lưu trữ cả bên trong các tệp CAD (BIM) và trong các cơ sở dữ liệu bên ngoài - thông qua việc xuất khẩu, chuyển đổi hoặc truy cập trực tiếp vào các cấu trúc bên trong CAD thông qua các công cụ đảo ngược kỹ thuật. Cách tiếp cận này giúp dễ dàng tích hợp thông tin dự án với các hệ thống và nền tảng doanh nghiệp khác.

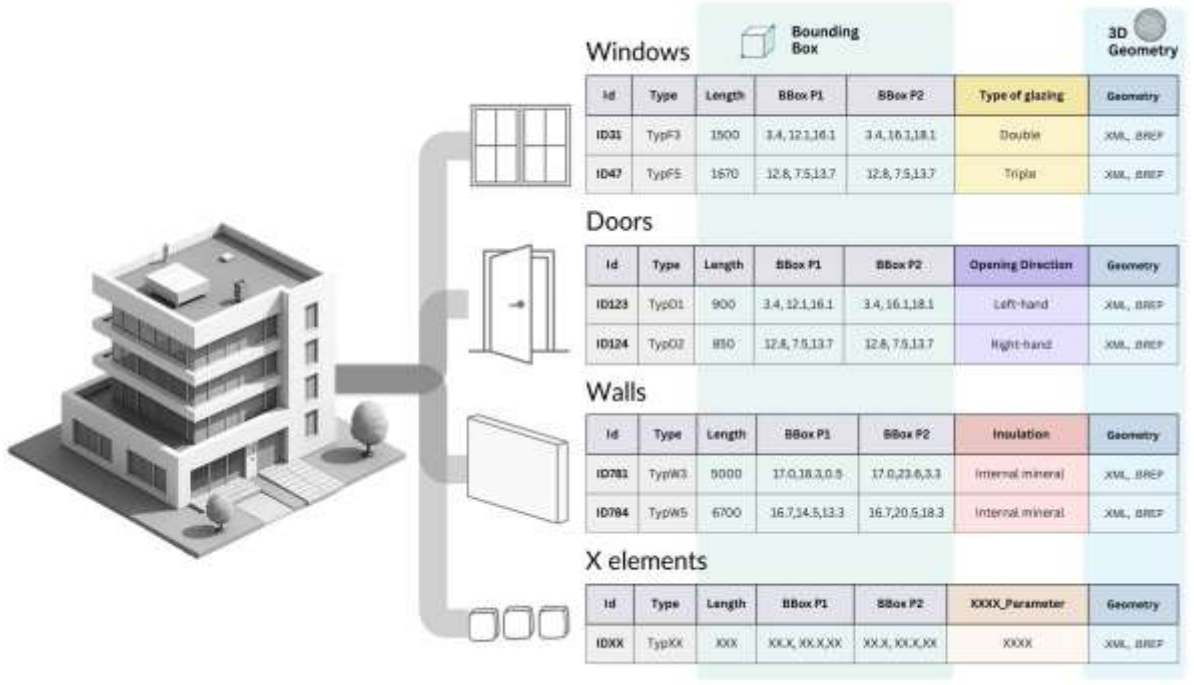
> Kỹ thuật đảo ngược trong bối cảnh CAD (BIM) là quá trình trích xuất và phân tích cấu trúc bên trong của mô hình số nhằm tái tạo logic, cấu trúc dữ liệu và các phụ thuộc của nó mà không cần truy cập vào các thuật toán hay tài liệu gốc.



Hình 3.115 Một phần tử của dự án, ngoài mô tả hình học tham số hoặc đa giác, còn chứa thông tin về các tham số và thuộc tính của các phần tử.

Cuối cùng, xung quanh mỗi phần tử hình thành một tập hợp độc đáo các tham số và thuộc tính, bao gồm cả những đặc điểm độc nhất của từng đối tượng (ví dụ: mã nhận dạng và kích thước) cũng như các thuộc tính chung cho nhóm các phần tử. Điều này không chỉ cho phép phân tích các phần tử riêng lẻ của dự án mà còn kết hợp chúng thành các nhóm logic, mà các chuyên gia khác có thể sử dụng cho các nhiệm vụ và tính toán của mình trong các hệ thống và cơ sở dữ liệu.

Thực thể (tiếng Anh: entity) là một đối tượng cụ thể hoặc trừu tượng của thế giới thực, có thể được phân biệt rõ ràng, mô tả và thể hiện dưới dạng dữ liệu.

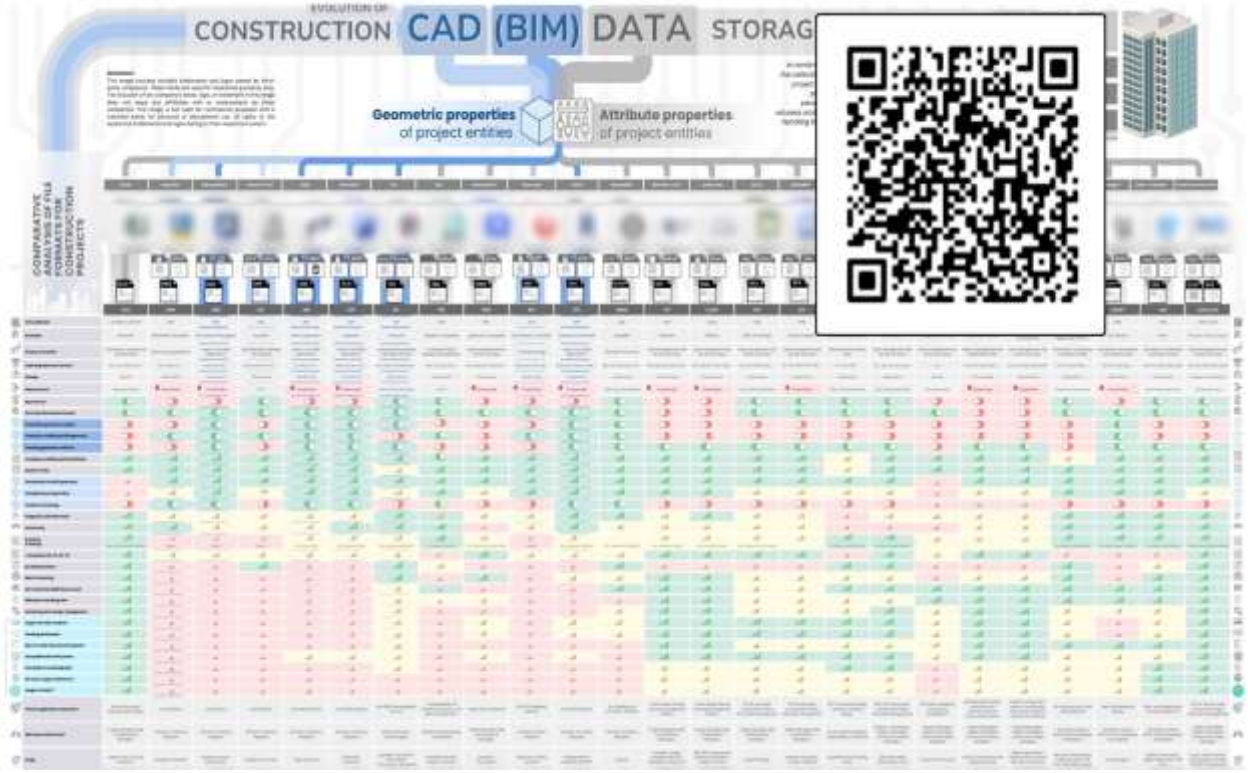


Hình 3.116 Mỗi phần tử của dự án chứa các thuộc tính, có thể được ghi vào bởi người thiết kế hoặc được tính toán bên trong chương trình CAD.

Trong vài thập kỷ qua, ngành xây dựng đã xuất hiện nhiều định dạng CAD (BIM) mới, đơn giản hóa việc tạo, lưu trữ và truyền tải dữ liệu. Các định dạng này có thể là đóng hoặc mở, dạng bảng, tham số hoặc đồ thị. Tuy nhiên, sự đa dạng và phân mảnh của chúng gây khó khăn đáng kể cho việc quản lý dữ liệu trong tất cả các giai đoạn của vòng đời dự án. Bảng so sánh các định dạng chính được sử dụng để trao đổi thông tin trong xây dựng được trình bày trên Hình 3.117 (bản đầy đủ có sẵn qua mã QR).-

Để giải quyết các vấn đề về khả năng tương tác và truy cập dữ liệu CAD, các quản lý (BIM) và điều phối viên được đưa vào công việc, nhiệm vụ của họ là kiểm soát việc xuất khẩu, kiểm tra chất lượng dữ liệu và tích hợp các phần dữ liệu CAD (BIM) vào các hệ thống khác.

Tuy nhiên, do tính đóng kín và phức tạp của các định dạng, việc tự động hóa quy trình này gặp khó khăn, buộc các chuyên gia phải thực hiện nhiều thao tác thủ công mà không có khả năng thiết lập quy trình xử lý dữ liệu hoàn chỉnh (pipeline).



Hình 3.117 Bảng so sánh các định dạng dữ liệu chính, trong đó lưu trữ thông tin về các phần tử của dự án [53].

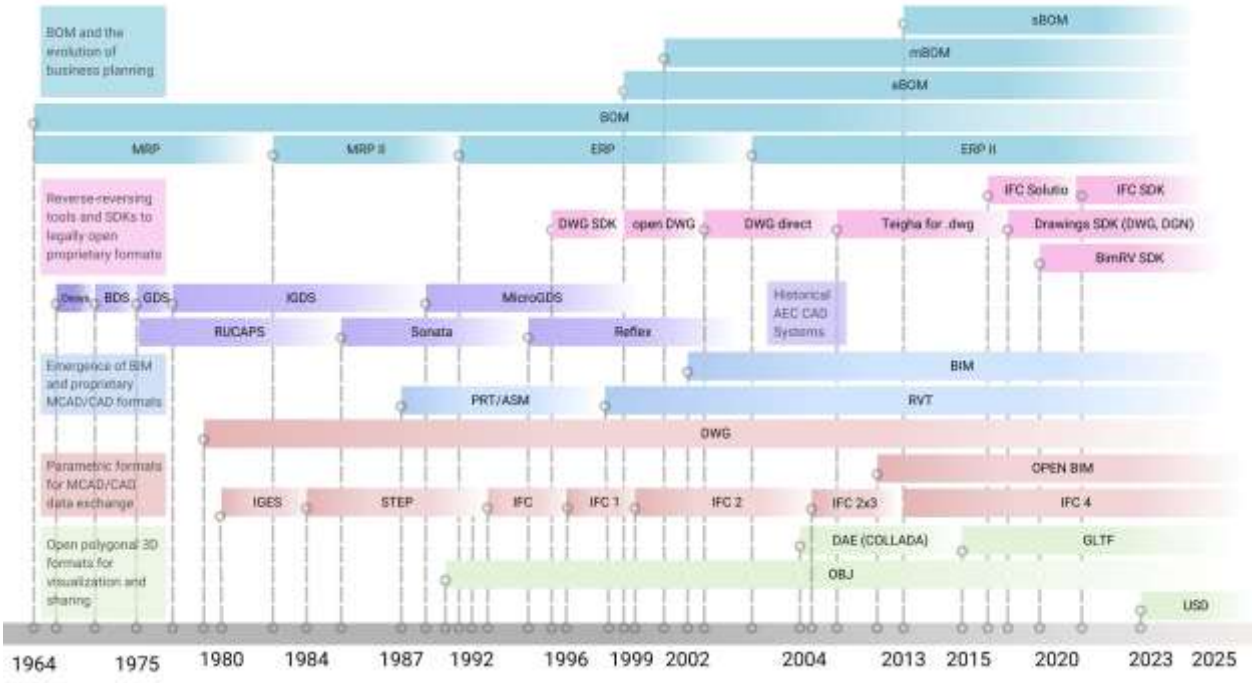
Để tìm hiểu tại sao lại có nhiều định dạng dữ liệu khác nhau và tại sao hầu hết trong số đó là đóng, quan trọng là phải đi sâu vào các quy trình diễn ra bên trong các chương trình CAD (BIM), điều này sẽ được thảo luận chi tiết trong phần sáu của cuốn sách.

Một lớp thông tin bổ sung được thêm vào hình học đã được các nhà phát triển hệ thống CAD giới thiệu dưới dạng khái niệm BIM (Mô hình Thông tin Xây dựng) - một thuật ngữ tiếp thị được thúc đẩy mạnh mẽ trong ngành xây dựng kể từ năm 2002 [54].

Sự xuất hiện của khái niệm BIM (BOM) và việc sử dụng CAD trong các quy trình

Khái niệm mô hình thông tin xây dựng (BIM), lần đầu tiên được trình bày trong tài liệu Whitepaper BIM năm 2002, đã xuất hiện nhờ các sáng kiến tiếp thị của các nhà sản xuất phần mềm CAD. Nó phát sinh từ các sáng kiến tiếp thị của các nhà phát triển phần mềm CAD và trở thành một nỗ lực để điều chỉnh các nguyên tắc đã được chứng minh hiệu quả trong ngành cơ khí cho nhu cầu của ngành xây dựng.

Nguồn cảm hứng cho BIM là khái niệm BOM (Bill of Materials) - thông số thành phần của sản phẩm, đã được áp dụng tích cực trong công nghiệp từ cuối những năm 1980. Trong ngành cơ khí, BOM cho phép liên kết dữ liệu từ các hệ thống CAD với các hệ thống PDM (Quản lý Dữ liệu Sản phẩm), PLM (Quản lý Vòng đời Sản phẩm) và ERP, cung cấp quản lý thông tin kỹ thuật toàn diện trong suốt vòng đời sản phẩm.-



Hình 3.118 Sự tiến hóa của các thông số (BOM), mô hình thông tin (BIM) và các định dạng số trong ngành xây dựng kỹ thuật.

Sự phát triển hiện đại của khái niệm BOM đã dẫn đến sự xuất hiện của cấu trúc mở rộng - XBOM (Extended BOM), không chỉ bao gồm thành phần sản phẩm mà còn có các kịch bản hành vi, yêu cầu về vận hành, thông số bền vững và dữ liệu cho phân tích dự đoán. XBOM về bản chất thực hiện vai trò giống như BIM trong xây dựng: cả hai phương pháp đều nhằm biến mô hình kỹ thuật số thành một nguồn thông tin đáng tin cậy duy nhất (Single Source of Truth) cho tất cả các bên tham gia dự án trong suốt vòng đời của đối tượng.

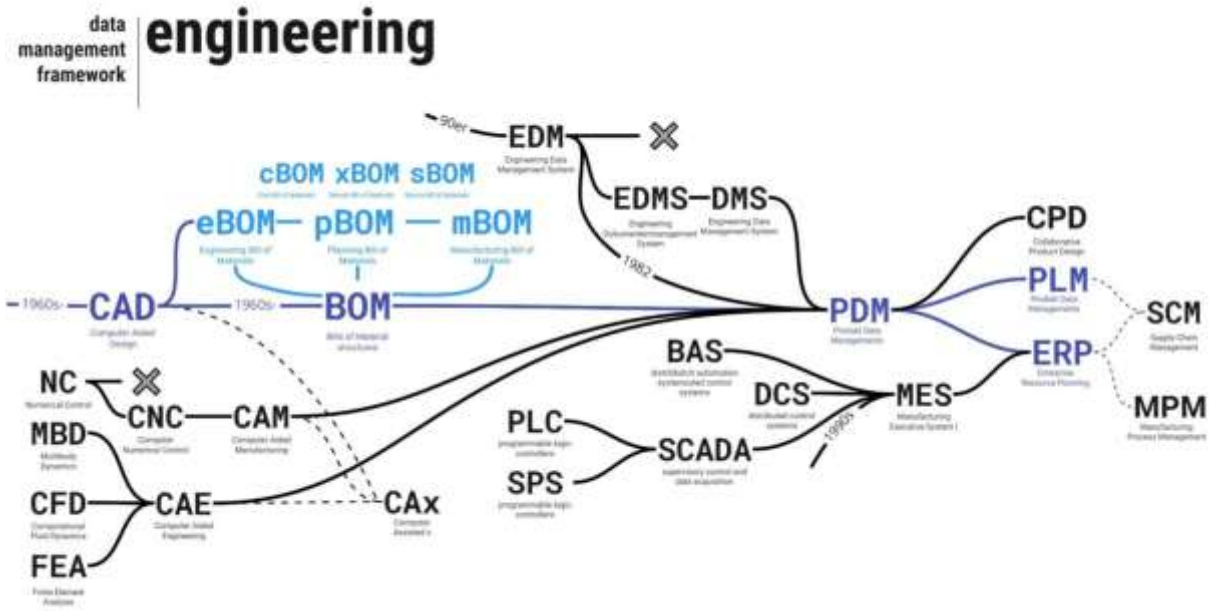
Một trong những bước quan trọng trong việc áp dụng BOM trong xây dựng là sự ra đời của CAD tham số đầu tiên vào năm 2002 (MCAD), được thiết kế đặc biệt cho ngành xây dựng. Nó được phát triển bởi nhóm đã tạo ra Pro-E® - hệ thống MCAD cách mạng cho ngành cơ khí, xuất hiện từ cuối những năm 1980 và trở thành tiêu chuẩn ngành.

Vào cuối những năm 1980, mục tiêu là loại bỏ các hạn chế tồn tại trong các chương trình CAD thời đó. Nhiệm vụ chính là giảm chi phí lao động cho việc thay đổi các tham số của các yếu tố thiết kế và đảm bảo khả năng cập nhật mô hình dựa trên dữ liệu bên ngoài các chương trình CAD thông qua cơ sở dữ liệu. Vai trò quan trọng trong điều này sẽ thuộc về việc tham số hóa: tự động lấy thông số từ cơ sở dữ liệu và sử dụng chúng để cập nhật mô hình bên trong hệ thống CAD.

> Pro-E và khái niệm mô hình tham số theo phần với BOM làm nền tảng đã có ảnh hưởng đáng kể đến sự phát triển của thị trường CAD và MCAD. Trong suốt 25 năm qua, mô hình này trong ngành và nhiều hệ thống hiện đại đã trở thành những người thừa kế khái niệm của nó.

> *Mục tiêu là tạo ra một hệ thống đủ linh hoạt để khuyến khích kỹ sư dễ dàng xem xét các kết cấu khác nhau. Và chi phí cho việc thay đổi thiết kế phải gần như bằng không. Phần mềm CAD/CAM truyền thống bị hạn chế thực sự đối với việc thực hiện thay đổi giá rẻ chỉ trong giai đoạn đầu của quy trình thiết kế.- Samuel Geisenberg, người sáng lập Công ty Parametric Technology Corporation®, nhà phát triển sản phẩm MCAD Pro-E và là giáo viên của người sáng tạo sản phẩm CAD sử dụng định dạng RVT.*

Trong ngành chế tạo máy, các nền tảng chính đã trở thành hệ thống PDM, PLM, MRP và ERP. Chúng đóng vai trò trung tâm trong việc quản lý dữ liệu và quy trình, thu thập thông tin từ các hệ thống CAx (CAD, CAM, CAE) và tổ chức hoạt động dự án dựa trên cấu trúc sản phẩm (BOM: eBOM, pBOM, mBOM) (Hình 3.118). Sự tích hợp này cho phép giảm thiểu sai sót, tránh việc trùng lặp dữ liệu và đảm bảo theo dõi xuyên suốt ở tất cả các giai đoạn - từ thiết kế đến sản xuất.



Hình 3.119 BOM đã ra đời từ những năm 1960 như một cách để cấu trúc dữ liệu từ các hệ thống CAx và truyền tải chúng vào các hệ thống quản lý.

Việc một trong những nhà cung cấp hàng đầu mua một giải pháp CAD được phát triển bởi đội ngũ cũ của Pro-E, dựa trên cách tiếp cận BOM, đánh dấu việc công bố gần như ngay lập tức một loạt Whitepaper về BIM (2002-2003) [60][61]. Từ giữa những năm 2000, khái niệm BIM bắt đầu được thúc đẩy mạnh mẽ trong ngành xây dựng, làm tăng rõ rệt sự quan tâm đến phần mềm tham số. Sự phổ biến tăng nhanh đến mức, nhánh xây dựng từ Pro-E của ngành chế tạo máy - CAD tham số do nhà cung cấp này phát triển - gần như đã loại bỏ các đối thủ cạnh tranh trong lĩnh vực thiết kế kiến trúc và xây dựng (Hình 3.120). Đến đầu năm 2020, nó đã củng cố sự thống trị toàn cầu về thị trường BIM (CAD) [62]. -



Hình 3.120 Mức độ phổ biến của truy vấn tìm kiếm trên Google (RVT so với IFC): CAD tham số, được tạo ra bởi đội ngũ cũ của Pro-E với hỗ trợ BOM-BIM, đã đạt được sự phổ biến gần như ở hầu hết các quốc gia trên thế giới.

Trong 20 năm qua, thuật ngữ BIM đã tích lũy rất nhiều ý nghĩa, mà sự đa nghĩa này bắt nguồn từ các khái niệm tiếp thị ban đầu được đưa ra vào đầu những năm 2000. Tiêu chuẩn ISO 19650, đã đóng vai trò quan trọng trong việc phổ biến thuật ngữ này, thực chất đã xác lập BIM là một phương pháp "dựa trên thông tin khoa học" trong quản lý thông tin. Tuy nhiên, trong chính văn bản của tiêu chuẩn liên quan đến quản lý dữ liệu trong suốt vòng đời của các đối tượng sử dụng BIM, mặc dù thuật ngữ BIM được đề cập, nhưng chưa bao giờ nhận được định nghĩa rõ ràng.

Trên trang web gốc của nhà cung cấp, người đã công bố một loạt Whitepaper về BIM vào năm 2002 [60] và 2003 [61], thực tế đã tái hiện các tài liệu tiếp thị về các khái niệm BOM (Bills of Materials) và PLM (Quản lý vòng đời sản phẩm), đã được áp dụng trong phần mềm chế tạo máy Pro-E từ những năm 1990 [63].

> *Mô hình thông tin xây dựng - một phương pháp tiếp cận đổi mới cho thiết kế, xây dựng và quản lý các tòa nhà, được giới thiệu bởi công ty... [tên công ty CAD nhà cung cấp] vào năm 2002, đã thay đổi cách nhìn nhận của các chuyên gia trong ngành trên toàn cầu về cách công nghệ có thể được áp dụng trong thiết kế, xây dựng và quản lý các tòa nhà. Whitepaper BIM, 2003 [61]*

Trong những tài liệu sớm này, BIM được liên kết trực tiếp với khái niệm cơ sở dữ liệu tích hợp trung tâm. Như đã được chỉ ra trong Whitepaper năm 2003, BIM là quản lý thông tin về tòa nhà, trong đó tất cả các cập nhật xảy ra trong một kho dữ liệu duy nhất, đảm bảo đồng bộ hóa tất cả các bản vẽ, mặt cắt và thông số kỹ thuật (BOM - Bills of Materials).

> *BIM được mô tả như là quản lý thông tin về tòa nhà, nơi tất cả các cập nhật và thay đổi đều diễn ra trong cơ sở dữ liệu. Do đó, cho dù bạn đang làm việc với sơ đồ, cắt lớp hay bản vẽ trên giấy, mọi thứ luôn được phối hợp, thống nhất và cập nhật.— trang web của nhà cung cấp CAD với Whitepaper BIM, năm 2003 [54]*

Ý tưởng quản lý thiết kế qua một cơ sở dữ liệu tích hợp duy nhất đã được thảo luận rộng rãi trong các nghiên cứu từ những năm 1980. Ví dụ, khái niệm BDS của Charles Eastman [57] bao gồm 43 lần đề cập đến thuật ngữ "cơ sở dữ liệu" (Hình. 6.12). Đến năm 2004, trong tài liệu về BIM, con số này đã giảm gần một nửa — xuống còn 23 trong Whitepaper năm 2002 [64]. Và đến giữa những năm 2000, chủ đề cơ sở dữ liệu gần như biến mất khỏi tài liệu tiếp thị của các nhà cung cấp và nói chung khỏi chương trình nghị sự số hóa.

Mặc dù chính cơ sở dữ liệu và quyền truy cập vào nó ban đầu được thiết kế như là lõi của hệ thống BIM, nhưng theo thời gian, trọng tâm đã chuyển sang hình học, hình ảnh hóa và 3D. Tuy nhiên, chính đơn vị đăng ký tiêu chuẩn IFC vào năm 1994, đã xuất bản Whitepaper BIM vào năm 2002 — cũng là nhà cung cấp đó — trong Whitepaper đầu những năm 2000 đã chỉ ra rõ ràng sự hạn chế của các định dạng trung lập như IGES, STEP và IFC và nhu cầu về việc truy cập trực tiếp vào các cơ sở dữ liệu CAD:

> *Các ứng dụng khác nhau có thể không tương thích, và dữ liệu được nhập lại có thể không chính xác […]. Kết quả của thiết kế tự động hóa truyền thống [CAD]: tăng chi phí, kéo dài thời gian ra mắt sản phẩm và giảm chất lượng sản phẩm. Ngày nay, tất cả các ứng dụng chính sử dụng các giao diện chuẩn ngành để trao đổi dữ liệu ở cấp độ thấp. Bằng cách sử dụng các tiêu chuẩn cũ IGES hoặc mới STEP [IFC thực tế và về mặt pháp lý là bản sao của định dạng STEP/IGES] để trao đổi dữ liệu giữa các ứng dụng của các nhà cung cấp khác nhau, người dùng có thể đạt được một mức độ tương thích dữ liệu nhất định giữa các sản phẩm tốt nhất trong ngành. Nhưng IGES và STEP chỉ hoạt động ở các cấp thấp, và chúng không thể trao đổi được dữ liệu phong phú như thông tin được tạo ra bởi các ứng dụng hàng đầu hiện nay […]. Và mặc dù các tiêu chuẩn này và tiêu chuẩn khác đang được cải tiến gần như hàng ngày, chúng sẽ luôn bị tụt lại phía sau các sản phẩm của các nhà sản xuất hiện đại về độ phong phú của dữ liệu. […] các chương trình trong khung ứng dụng phải có khả năng trao đổi dữ liệu và duy trì độ phong phú của chúng mà không cần đến các bộ chuyển đổi trung lập như IGES, STEP [IFC] hoặc PATRAN. Thay vào đó, các ứng dụng của khung cần phải có khả năng truy cập trực tiếp vào cơ sở dữ liệu chính của CAD, để không mất đi sự chi tiết và chính xác của thông tin.— Whitepaper của nhà cung cấp CAD (IFC, BIM) «Thiết kế và sản xuất tích hợp: Lợi ích và lý do», năm 2000 [65]*

Do đó, ngay từ những năm 1980 và đầu những năm 2000, yếu tố then chốt của thiết kế số trong môi trường CAD được coi là cơ sở dữ liệu, chứ không phải định dạng tệp hoặc định dạng trung lập IFC. Đã đề xuất từ bỏ các bộ chuyển đổi và đảm bảo truy cập trực tiếp cho các ứng dụng đến dữ liệu. Tuy nhiên, trong thực tế, đến giữa những năm 2020, khái niệm BIM đã trở nên giống như chiến lược "chia để trị", nơi lợi ích của các nhà cung cấp phần mềm sử dụng các lõi hình học đóng vẫn được ưu tiên hơn là phát triển việc trao đổi thông tin mở.

Hôm nay, BIM được coi là một phần không thể thiếu của ngành xây dựng. Nhưng trong hai thập kỷ qua, những lời hứa về sự tương tác đơn giản và tích hợp dữ liệu vẫn chủ yếu chưa được thực hiện. Hầu hết các giải pháp vẫn phụ thuộc vào các định dạng khép kín hoặc định dạng trung lập, cũng như các công cụ chuyên biệt. Chúng ta sẽ xem xét chi tiết các vấn đề về lịch sử hình thành BIM, open BIM và IFC, cũng như các vấn đề về khả năng tương tác và các lõi hình học trong phần thứ sáu của cuốn sách "CAD và BIM: marketing,

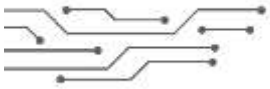
thực tế và tương lai của dữ liệu thiết kế trong xây dựng".

> Hiện nay, ngành công nghiệp đang đối mặt với một thách thức lớn — chuyển từ cách hiểu truyền thống về CAD (BIM) như một công cụ mô hình hóa sang việc sử dụng nó như một cơ sở dữ liệu hoàn chỉnh. Điều này đòi hỏi những cách tiếp cận mới trong việc làm việc với thông tin, từ bỏ sự phụ thuộc vào các hệ sinh thái khép kín và áp dụng các giải pháp mở.

Với sự phát triển của các công cụ kỹ thuật đảo ngược, cho phép truy cập vào các cơ sở dữ liệu CAD, cũng như nhờ vào sự lan rộng của mã nguồn mở và công nghệ LLM, người dùng và các nhà phát triển trong ngành xây dựng ngày càng dần rời xa các thuật ngữ mơ hồ từ các nhà cung cấp phần mềm. Thay vào đó, sự chú ý được chuyển vào những gì thực sự quan trọng: dữ liệu (cơ sở dữ liệu) và quy trình.

> Ẩn sau những chữ viết tắt và hình ảnh thời thượng là các thực tiễn quản lý dữ liệu tiêu chuẩn: lưu trữ, truyền tải và biến đổi — tức là quá trình ETL cổ điển (Extract, Transform, Load). Giống như trong các ngành khác, số hóa trong xây dựng không chỉ yêu cầu các tiêu chuẩn trao đổi, mà còn cần một cách làm việc rõ ràng với thông tin đa dạng.

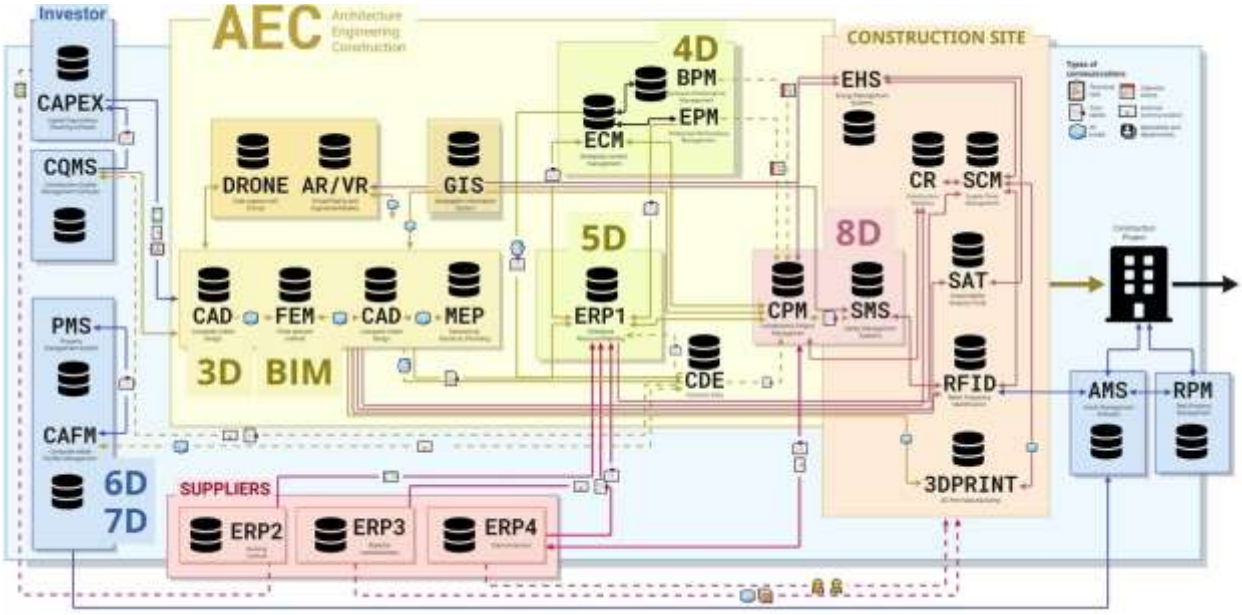
Để tận dụng tối đa tiềm năng của dữ liệu CAD (BIM), các công ty cần phải xem xét lại cách tiếp cận của mình trong việc quản lý thông tin. Điều này không thể tránh khỏi dẫn đến một yếu tố chính của quá trình chuyển đổi số — thống nhất, tiêu chuẩn hóa và cấu trúc có ý nghĩa của dữ liệu mà các chuyên gia trong ngành xây dựng hàng ngày làm việc.



CHƯƠNG 3.2. **ĐỒNG NHẤT VÀ CẤU TRÚC DỮ LIỆU**

Điền dữ liệu vào hệ thống trong ngành xây dựng

Dù là các tập đoàn lớn hay các công ty vừa, các chuyên gia hàng ngày đều tham gia vào việc điền vào các hệ thống phần mềm và cơ sở dữ liệu với nhiều giao diện thông tin định dạng khác nhau (Hình 3.21), mà thông qua các quản lý cần đảm bảo chúng tương tác với nhau một cách nhịp nhàng. Chính tập hợp các hệ thống và quy trình tương tác này cuối cùng tạo ra doanh thu và lợi nhuận cho công ty.

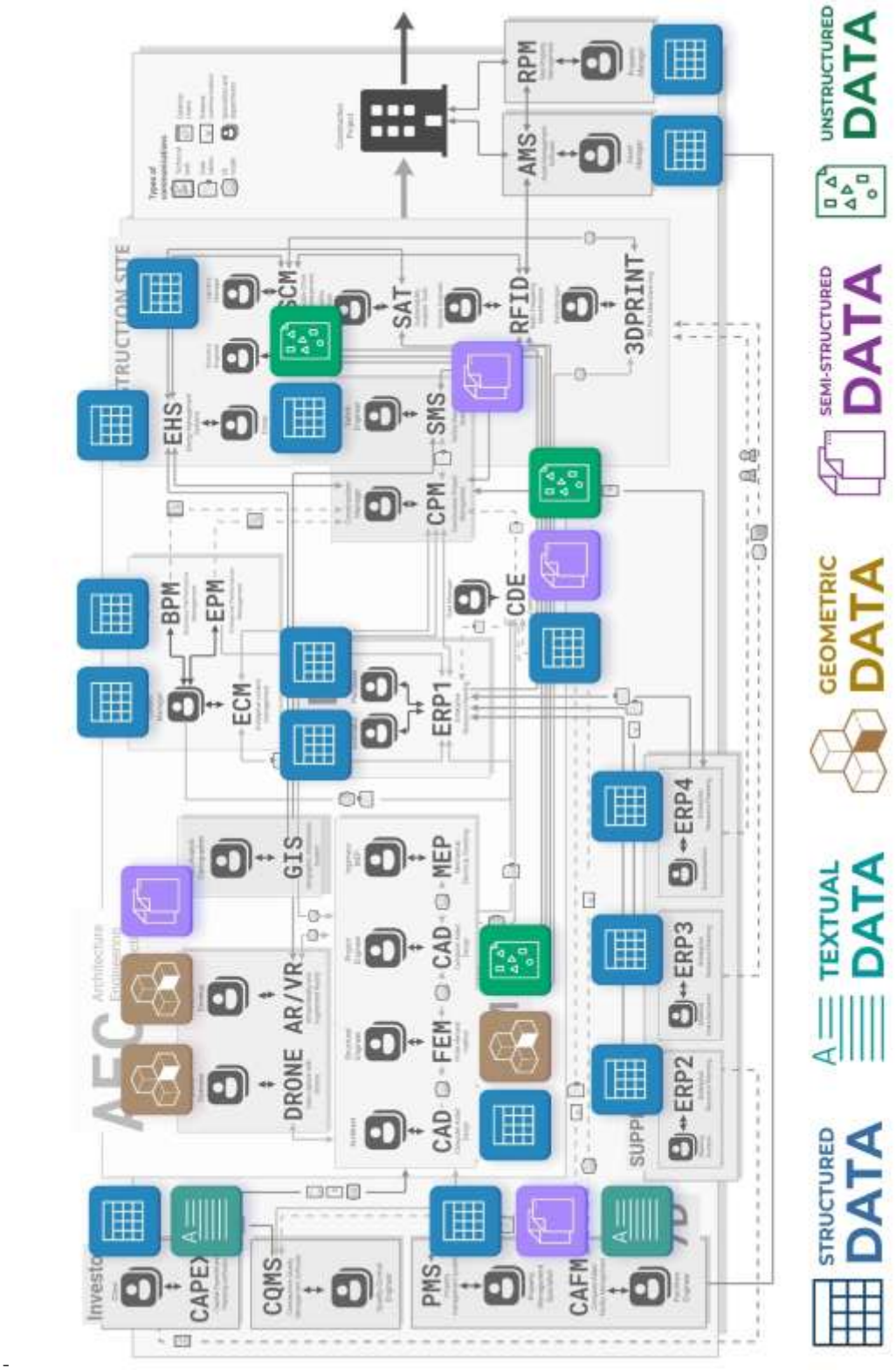


Hình 3.21 Hầu như mỗi hệ thống hoặc ứng dụng được sử dụng trong ngành xây dựng đều có một trong những cơ sở dữ liệu RDBMS phổ biến làm nền tảng.

Mỗi loại hệ thống đã đề cập trước đó và áp dụng trong ngành xây dựng hoạt động với những loại dữ liệu của riêng mình, tương ứng với vai trò chức năng của các hệ thống đó. Để chuyển từ mức độ trừu tượng đến độ cụ thể, chúng ta chuyển từ loại dữ liệu đến biểu diễn của chúng dưới dạng định dạng và tài liệu.

Đối với danh sách hệ thống đã được trình bày trước đó (Hình 1.24), chúng ta sẽ thêm các loại định dạng và tài liệu cụ thể mà chúng thường làm việc: -

- Nhà đầu tư (CAPEX)
  - Dữ liệu tài chính: ngân sách, dự đoán chi phí (dữ liệu có cấu trúc).
  - Dữ liệu về xu hướng thị trường: phân tích thị trường (dữ liệu có cấu trúc và không có cấu trúc).
  - Dữ liệu pháp lý và hợp đồng: hợp đồng (dữ liệu văn bản).



Hình 3.22 Trong ngành xây dựng, có rất nhiều hệ thống với các giao diện khác nhau hoạt động với các loại dữ liệu khác nhau.

- Hệ thống quản lý (PMS, CAFM, CQMS)

- Dữ liệu dự án: biểu đồ, nhiệm vụ (dữ liệu có cấu trúc).
- Dữ liệu bảo trì đối tượng: kế hoạch bảo trì (dữ liệu văn bản và dữ liệu bán cấu trúc).
- Dữ liệu kiểm soát chất lượng: tiêu chuẩn, báo cáo kiểm tra (dữ liệu văn bản và dữ liệu không cấu trúc).

- CAD, FEM và BIM
  - Bản vẽ kỹ thuật: kế hoạch kiến trúc, kết cấu (dữ liệu hình học, dữ liệu không cấu trúc).
  - Mô hình tòa nhà: mô hình 3D, dữ liệu vật liệu (dữ liệu hình học và dữ liệu bán cấu trúc).
  - Tính toán kỹ thuật: phân tích tải (dữ liệu có cấu trúc).
- Hệ thống quản lý công trường (EHS, SCM)
  - Dữ liệu an toàn và sức khỏe: biên bản an toàn (dữ liệu văn bản và dữ liệu có cấu trúc).
  - Dữ liệu chuỗi cung ứng: tồn kho, đơn đặt hàng (dữ liệu có cấu trúc).
  - Báo cáo hàng ngày: thời gian làm việc, năng suất (dữ liệu có cấu trúc).
- Drone, AR/VR, GIS, in 3D
  - Dữ liệu địa lý: bản đồ địa hình (dữ liệu hình học và dữ liệu có cấu trúc).
  - Dữ liệu thời gian thực: video và hình ảnh (dữ liệu không cấu trúc).
  - Mô hình cho in 3D: bản vẽ kỹ thuật số (dữ liệu hình học).
- Hệ thống quản lý bổ sung (4D BPM, 5D ERP1)
  - Dữ liệu về thời gian và chi phí: biểu đồ, ước lượng (dữ liệu có cấu trúc).
  - Quản lý thay đổi: bản ghi thay đổi trong dự án (dữ liệu văn bản và dữ liệu có cấu trúc).
  - Báo cáo về kết quả hoạt động: chỉ tiêu thành công (dữ liệu có cấu trúc).
- Tích hợp dữ liệu và kết nối (CDE, RFID, AMS, RPM)
  - Trao đổi dữ liệu: trao đổi tài liệu, mô hình dữ liệu (dữ liệu có cấu trúc và dữ liệu văn bản).
  - RFID và dữ liệu theo dõi: logistics, quản lý tài sản (dữ liệu có cấu trúc).
  - Giám sát và kiểm soát: cảm biến tại hiện trường (dữ liệu có cấu trúc và dữ liệu không cấu trúc).

Do đó, mỗi hệ thống trong ngành xây dựng — từ hệ thống quản lý công trường đến cơ sở dữ liệu vận hành — hoạt động với loại thông tin riêng: có cấu trúc, văn bản, hình học, và nhiều hơn nữa. "Cảnh quan dữ liệu", mà các chuyên gia hàng ngày phải làm việc rất đa dạng. Tuy nhiên, việc chỉ liệt kê các định dạng không tiết lộ toàn bộ độ phức tạp trong công việc thực tế với thông tin.

Trên thực tế, các công ty gặp phải vấn đề là dữ liệu, mặc dù được thu thập từ các hệ thống, không sẵn sàng để sử dụng "như là". Đặc biệt là các thông tin văn bản, hình ảnh, tài liệu PDF, tệp CAD và các định dạng khác, rất khó để phân tích bằng các công cụ tiêu chuẩn. Đó là lý do tại sao bước then chốt tiếp theo là chuyển đổi dữ liệu — quá trình mà không có nó, không thể tự động hóa hiệu quả việc xử lý, phân tích, trực quan hóa và ra quyết định.

Chuyển đổi dữ liệu: nền tảng quan trọng của phân tích kinh doanh hiện đại

Hôm nay, hầu hết các công ty đối mặt với một nghịch lý: khoảng 80% quy trình hàng ngày của họ vẫn dựa trên dữ liệu cấu trúc cổ điển — các bảng Excel quen thuộc và cơ sở dữ liệu quan hệ (RDBMS). Tuy nhiên, 80% thông tin mới mà các công ty nhận được trong hệ sinh thái số của họ có tính chất không cấu trúc hoặc yếu cấu trúc. Đây là văn bản, đồ họa, hình học, hình ảnh, mô hình CAD, tài liệu PDF, bản ghi âm và video, thư điện tử và nhiều thứ khác.-

Hơn nữa, khối lượng dữ liệu không cấu trúc tiếp tục tăng trưởng nhanh chóng — tỷ lệ tăng hàng năm được

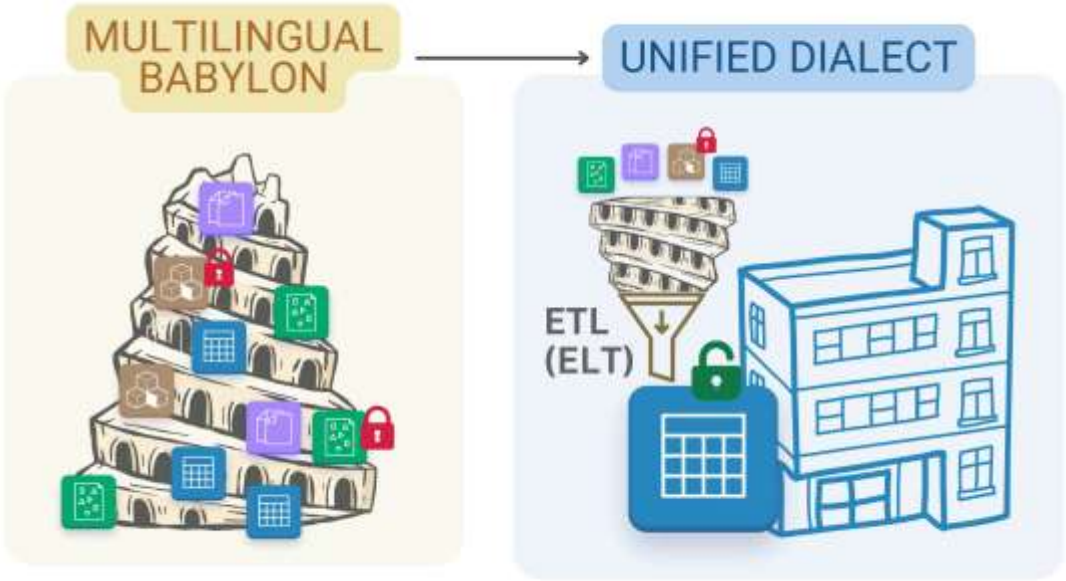
ước tính từ 55 đến 65%. Động thái này tạo ra những khó khăn nghiêm trọng trong việc tích hợp thông tin mới vào quy trình kinh doanh hiện có. Việc bỏ qua dòng dữ liệu đa dạng này dẫn đến việc hình thành các khoảng trống thông tin và giảm khả năng quản lý toàn bộ môi trường số của công ty.



Sự gia tăng hàng năm của khối lượng dữ liệu không cấu trúc tạo ra những vấn đề trong việc tích hợp thông tin liên tục vào các quy trình kinh doanh.

> Việc bỏ qua các dữ liệu không cấu trúc phức tạp và dữ liệu yếu cấu trúc trong các quy trình tự động hóa có thể dẫn đến những khoảng trống đáng kể trong các bức tranh thông tin của công ty. Trong thế giới hiện đại với sự di chuyển thông tin không kiểm soát và bùng nổ, các công ty cần áp dụng một phương pháp tiếp cận hỗn hợp để quản lý dữ liệu, bao gồm các phương pháp hiệu quả cho tất cả các loại dữ liệu.

Chìa khóa để quản lý dữ liệu hiệu quả nằm ở việc tổ chức, cấu trúc và phân loại các loại dữ liệu "Babylon" khác nhau (bao gồm dữ liệu không cấu trúc, văn bản và định dạng hình học, cũng như dữ liệu cấu trúc hoặc yếu cấu trúc). Quá trình này chuyển đổi đống dữ liệu hỗn loạn thành các cấu trúc được tổ chức để tích hợp vào các hệ thống, từ đó cho phép ra quyết định dựa trên chúng.-



Nhiệm vụ chính của các phòng quản lý dữ liệu là chuyển đổi "Babylon" của các dữ liệu đa dạng và đa định dạng thành một hệ thống có cấu trúc và phân loại.

Một trong những trở ngại chính trong việc đạt được sự thống nhất này vẫn là mức độ tương thích thấp giữa các nền tảng số khác nhau - các "silo", như chúng ta đã đề cập trong các chương trước.

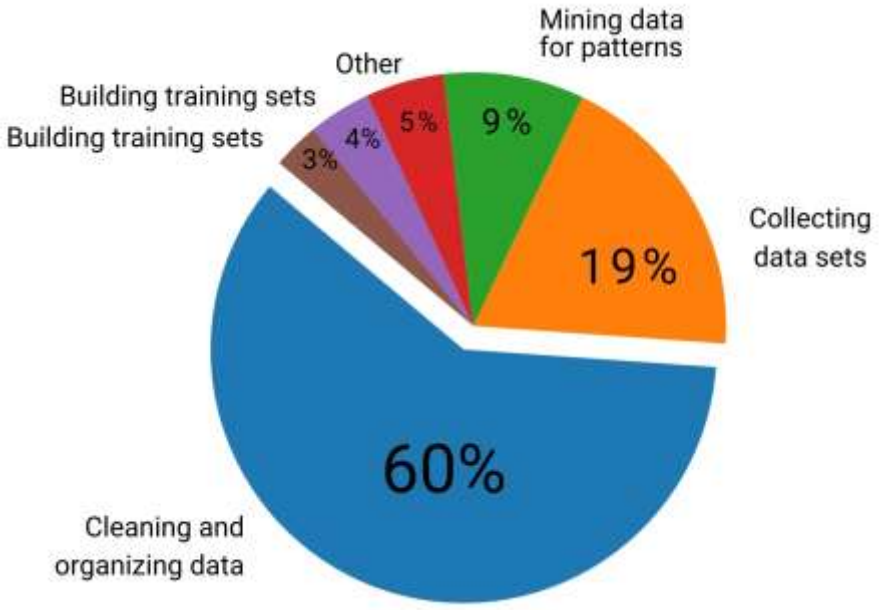
> Theo báo cáo của Viện Tiêu chuẩn và Công nghệ Quốc gia (NIST, Hoa Kỳ), sự tương thích kém giữa các nền tảng xây dựng khác nhau dẫn đến mất mát thông tin và chi phí bổ sung đáng kể. Chỉ vào năm 2002, các vấn đề về tương thích phần mềm đã gây ra tổn thất trong xây dựng ở Mỹ lên tới 15,8 tỷ đô la mỗi năm, trong đó hai phần ba số tổn thất này thuộc về các chủ sở hữu và nhà điều hành tòa nhà, đặc biệt trong quá trình vận hành và bảo trì. Nghiên cứu cũng chỉ ra rằng việc chuẩn hóa các định dạng dữ liệu có thể giảm thiểu những mất mát này và nâng cao hiệu quả làm việc ở tất cả các giai đoạn của vòng đời dự án.

Theo một nghiên cứu của CrowdFlower vào năm 2016, với 16.000 chuyên gia làm việc với dữ liệu trên toàn cầu, vấn đề chính vẫn là dữ liệu "bẩn" và đa định dạng. Theo nghiên cứu này, nguồn tài nguyên quý giá nhất không phải là các cơ sở dữ liệu cuối cùng hay các mô hình học máy, mà là thời gian của các chuyên gia được dành cho việc chuẩn bị thông tin.

Việc làm sạch, định dạng và tổ chức dữ liệu chiếm tới 60 phần trăm thời gian làm việc của nhà phân tích và quản lý dữ liệu. Gần một phần năm thời gian được dành cho việc tìm kiếm và thu thập các tập dữ liệu cần thiết, thường bị ẩn trong các kho dữ liệu khép kín ("silô") và không khả dụng cho phân tích. Chỉ khoảng 9 phần trăm thời gian được dành riêng cho mô hình hóa, phân tích, xây dựng dự đoán và kiểm tra giả thuyết. Tất cả những việc khác bao gồm giao tiếp, trực quan hóa, báo cáo và nghiên cứu các nguồn thông tin phụ trợ.

Trung bình công việc của quản lý dữ liệu được phân bổ như sau (Hình. 3.25):-

- Làm sạch và tổ chức dữ liệu (60%): sự có mặt của các dữ liệu sạch và có cấu trúc có thể giảm đáng kể thời gian làm việc của nhà phân tích và tăng tốc quá trình hoàn thành nhiệm vụ.
- Thu thập dữ liệu (19%): thách thức chính đối với các chuyên gia trong lĩnh vực khoa học dữ liệu là việc tìm kiếm các tập dữ liệu liên quan. Thông thường, dữ liệu của các công ty được lưu trữ trong các "silô" không có tổ chức, gây khó khăn trong việc truy cập thông tin cần thiết.
- Mô hình hóa/học máy (9%): thường xuyên bị khó khăn bởi sự không rõ ràng trong các mục tiêu kinh doanh từ khách hàng. Thiếu sự đặt ra vấn đề rõ ràng có thể làm giảm tiềm năng của ngay cả mô hình chất lượng nhất.
- Các nhiệm vụ khác (5%): bên cạnh việc xử lý dữ liệu, các nhà phân tích cũng phải nghiên cứu, xem xét dữ liệu từ nhiều góc độ khác nhau, giao tiếp kết quả thông qua các trực quan hóa và báo cáo, cũng như đưa ra các khuyến nghị về tối ưu hóa quy trình và chiến lược.



Hình. 3.25 Những gì các quản lý dữ liệu chi tiêu nhiều thời gian nhất (theo tài liệu [70]).

> Những ước tính này cũng được xác nhận bởi các nghiên cứu khác. Theo nghiên cứu của Xplenty, được công bố trong BizReport vào năm 2015 [71], từ 50% đến 90% thời gian của các chuyên gia phân tích kinh doanh (BI) được sử dụng cho việc chuẩn bị dữ liệu cho phân tích.

Làm sạch, kiểm tra và tổ chức dữ liệu là nền tảng quan trọng cho tất cả các quy trình xử lý và phân tích dữ liệu sau này, chiếm tới 90% thời gian của các chuyên gia làm việc với dữ liệu.

Công việc tỉ mỉ này, không thể nhìn thấy bởi người dùng cuối, là điều kiện thiết yếu. Các lỗi trong dữ liệu đầu vào sẽ làm sai lệch kết quả phân tích, gây nhầm lẫn và có thể dẫn đến những sai lầm quản lý tốn kém. Đó là lý do tại sao các quy trình làm sạch và chuẩn hóa dữ liệu - từ việc loại bỏ các bản sao và điền vào các khoảng trống đến việc điều chỉnh các đơn vị đo lường và đưa về cùng một mô hình - trở thành nền tảng của chiến lược số hiện đại.

Do đó, việc chuyển đổi, làm sạch và chuẩn hóa dữ liệu không chỉ chiếm phần lớn thời gian của các chuyên gia (đến 80% công việc với dữ liệu), mà còn xác định khả năng sử dụng hiệu quả của chúng trong các quy trình kinh doanh hiện đại. Tuy nhiên, tự bản thân việc tổ chức và làm sạch dữ liệu không đủ để đáp ứng bài toán tối ưu hóa quản lý luồng thông tin của công ty. Trong giai đoạn tổ chức và cấu trúc, việc lựa chọn mô hình dữ liệu phù hợp trở nên quan trọng, ảnh hưởng trực tiếp đến sự thuận tiện và hiệu quả trong việc làm việc với thông tin trong các giai đoạn xử lý tiếp theo.

Vì dữ liệu và mục tiêu kinh doanh khác nhau, việc hiểu rõ các mô hình dữ liệu và khả năng lựa chọn hoặc tạo cấu trúc phù hợp là rất quan trọng. Tùy thuộc vào mức độ cấu trúc và cách mô tả mối quan hệ giữa các yếu tố, có ba mô hình chính: mô hình có cấu trúc, mô hình kém cấu trúc và mô hình đồ thị. Mỗi mô hình thích hợp cho các nhiệm vụ khác nhau và có những điểm mạnh và điểm yếu riêng.

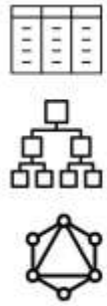
Mô hình dữ liệu: mối quan hệ trong dữ liệu và các kết nối giữa các yếu tố

Dữ liệu trong các hệ thống thông tin được tổ chức theo nhiều cách khác nhau – tùy thuộc vào các nhiệm vụ

và yêu cầu về lưu trữ, xử lý và truyền tải thông tin. Sự khác biệt chính giữa các loại mô hình dữ liệu, dưới dạng của thông tin được lưu trữ, nằm ở mức độ cấu trúc và cách mô tả mối quan hệ giữa các yếu tố.

> Dữ liệu có cấu trúc có một sơ đồ rõ ràng và lặp lại: chúng được tổ chức dưới dạng bảng với các cột cố định. Định dạng này đảm bảo tính dự đoán, đơn giản trong xử lý và hiệu quả khi thực hiện các truy vấn SQL, lọc và tổng hợp. Ví dụ: cơ sở dữ liệu (RDBMS), Excel, CSV.

Dữ liệu kém cấu trúc cho phép một cấu trúc linh hoạt: các yếu tố khác nhau có thể chứa một tập hợp thuộc tính khác nhau và được lưu trữ dưới dạng cấu trúc phân cấp. Ví dụ: JSON, XML hoặc các định dạng tài liệu khác. Những dữ liệu này tiện lợi khi cần mô hình hóa các đối tượng và mối quan hệ lồng nhau, nhưng mặt khác, nó làm phức tạp việc phân tích và chuẩn hóa dữ liệu (Hình 3.26).-



Data Model	Storage Format	Example
Relational	CSV, SQL	A table of doors in Excel
Hierarchical	JSON, XML	Nested door objects inside a room
Graph-based	RDF, GraphDB	Relationships between building elements

Hình 3.26 Mô hình dữ liệu - là cấu trúc logic mô tả cách thức tổ chức, lưu trữ và xử lý dữ liệu trong hệ thống.

Việc chọn định dạng phù hợp phụ thuộc vào các nhiệm vụ:

- Nếu tốc độ lọc và phân tích quan trọng - các bảng quan hệ (SQL, CSV, RDBMS, cơ sở dữ liệu cột) sẽ phù hợp.
- Nếu cần cấu trúc linh hoạt - nên sử dụng JSON hoặc XML.
- Nếu dữ liệu có các mối quan hệ phức tạp - các cơ sở dữ liệu đồ thị cung cấp tính trực quan và khả năng mở rộng.

Trong các cơ sở dữ liệu quan hệ cổ điển (RDBMS), mỗi thực thể (chẳng hạn như cửa) được đại diện bởi một hàng, và thuộc tính của nó - là các cột trong bảng. Ví dụ, bảng các hạng mục thuộc danh mục "Cửa" có thể chứa các trường ID, Chiều cao, Chiều rộng, Chống cháy và ID Phòng, chỉ định vị trí (Hình 3.27).

Trong các cơ sở dữ liệu quan hệ cổ điển (RDBMS), các mối quan hệ được hình thành dưới dạng bảng, trong đó mỗi bản ghi đại diện cho một đối tượng, và các cột - là các tham số của nó. Dưới dạng bảng, thông tin về cửa trong dự án xuất hiện như vậy, trong đó mỗi hàng đại diện cho một yếu tố riêng - cửa với mã xác định duy nhất và các thuộc tính của nó, trong khi mối quan hệ với phòng được thực hiện thông qua tham số "ID Phòng".

Door ID	Room ID	Height (mm)	Width (mm)	Fireproof
ID1001	101	2000	900	Yes
ID1002	101	2100	800	No
ID1003	102	2000	850	Yes

Hình 3.27 Thông tin về ba yếu tố thuộc danh mục "Cửa" của dự án dưới dạng bảng có cấu trúc.

Trong các định dạng kém cấu trúc, chẳng hạn như JSON hoặc XML, dữ liệu được lưu trữ dưới dạng phân cấp hoặc lồng nhau, trong đó các yếu tố có thể chứa các đối tượng khác bên trong, và cấu trúc của chúng có thể thay đổi. Điều này cho phép mô hình hóa các mối quan hệ phức tạp giữa các yếu tố. Thông tin tương tự về cửa trong dự án, được ghi dưới dạng cấu trúc (Hình 3.27), trong định dạng kém cấu trúc (JSON) sẽ được thể hiện như sau (Hình 3.28), trong đó chúng trở thành các đối tượng lồng nhau bên trong các phòng (Rooms – ID), phản ánh hợp lý cấu trúc phân cấp.--

doors.json

```
{
  "Rooms": [
    {
      "ID": 101,
      "Doors": [
        {"ID": 1, "Height": 2000, "Width": 900, "Fireproof": "Yes"},
        {"ID": 2, "Height": 2100, "Width": 800, "Fireproof": "No"}
      ]
    },
    {
      "ID": 102,
      "Doors": [
        {"ID": 3, "Height": 2000, "Width": 850, "Fireproof": "Yes"}
      ]
    }
  ]
}
```

Hình 3.28 Thông tin về các yếu tố trong danh mục "Cửa" của dự án dưới định dạng JSON.

Trong mô hình đồ thị, dữ liệu được biểu diễn dưới dạng các nút (đỉnh) và các kết nối (cạnh) giữa chúng. Điều này cho phép hiển thị trực quan các mối quan hệ phức tạp giữa các đối tượng và thuộc tính của chúng. Đối với dữ liệu về cửa và phòng trong dự án, hình ảnh đồ thị trông như sau:

- Các nút (đỉnh) đại diện cho các thực thể chính: các phòng (Phòng 101, Phòng 102) và các cửa (ID1001, ID1002, ID1003)
- Các cạnh (kết nối) cho thấy mối quan hệ giữa các thực thể này, ví dụ, sự thuộc về của một cửa đối với một phòng cụ thể.
- Các thuộc tính được gắn liền với các nút và chứa các đặc điểm của các thực thể (chiều cao, chiều rộng, khả năng chống cháy cho cửa).

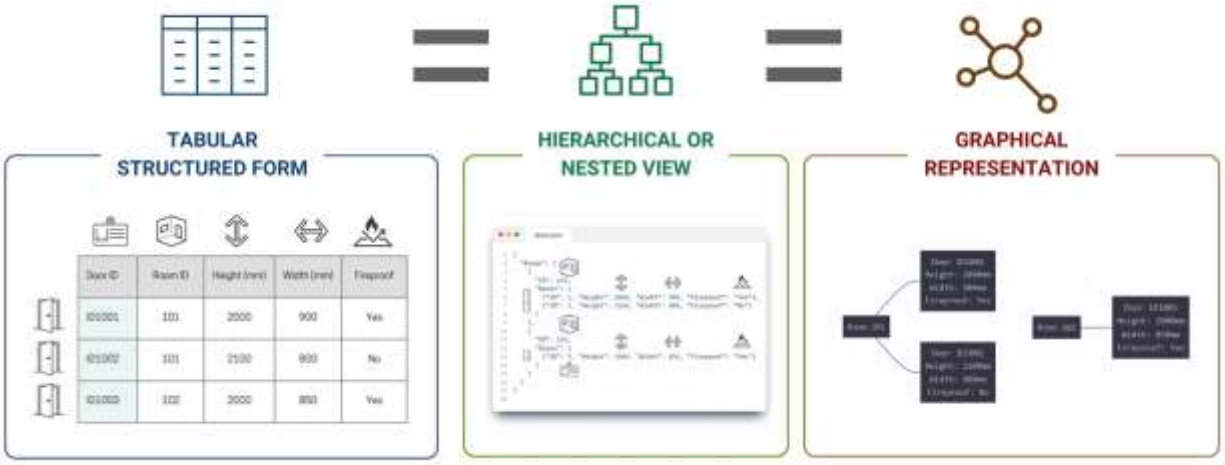


Hình 3.29 Thông tin về các thực thể cửa trong dự án được biểu diễn bằng đồ thị.

Trong mô hình đồ thị dữ liệu, mỗi phòng và mỗi cửa là các nút riêng biệt. Các cửa được liên kết với các phòng thông qua các cạnh, chỉ ra sự thuộc về của cửa đối với một phòng cụ thể. Trong khi đó, các thuộc tính của cửa (chiều cao, chiều rộng, khả năng chống cháy) được lưu trữ như các thuộc tính của các nút tương ứng. Chúng ta sẽ bàn thêm về các định dạng đồ thị và cách mà ngữ nghĩa đồ thị xuất hiện trong ngành xây dựng trong chương "Sự xuất hiện của ngữ nghĩa và ontology trong xây dựng".

Các cơ sở dữ liệu đồ thị hiệu quả trong những trường hợp mà không chỉ dữ liệu mà các mối quan hệ giữa chúng là quan trọng, chẳng hạn như trong các hệ thống gợi ý, hệ thống định tuyến hoặc mô hình hóa các mối quan hệ phức tạp trong các dự án quản lý tài sản. Định dạng đồ thị đơn giản hóa việc tạo ra các mối quan hệ mới, cho phép thêm các loại dữ liệu mới vào đồ thị mà không cần thay đổi cấu trúc kho lưu trữ. Tuy nhiên, so với các bảng quan hệ và các định dạng có cấu trúc, thì trong đồ thị không tồn tại sự liên kết bổ sung giữa các dữ liệu — việc chuyển đổi dữ liệu từ cơ sở dữ liệu hai chiều sang đồ thị không tăng số lượng mối quan hệ và không cho phép thu thập thêm thông tin mới.

> Hình thức và sơ đồ dữ liệu cần phải phù hợp với trường hợp sử dụng cụ thể và các nhiệm vụ đang được giải quyết. Để hoạt động hiệu quả trong các quy trình kinh doanh, việc sử dụng các công cụ và các mô hình dữ liệu giúp tối ưu hóa việc thu thập kết quả một cách nhanh chóng và đơn giản là rất quan trọng.



Hình 3.210 Thông tin tương tự về các phần tử trong dự án có thể được lưu trữ dưới nhiều định dạng khác nhau thông qua các mô hình dữ liệu khác nhau.

> Ngày nay, hầu hết các công ty lớn đang phải đối mặt với vấn đề phức tạp thừa thãi của dữ liệu. Mỗi ứng dụng trong số hàng trăm hoặc hàng nghìn ứng dụng sử dụng một mô hình dữ liệu riêng, điều này tạo ra sự phức tạp dư thừa — một mô hình riêng lẻ thường phức tạp hơn nhiều lần so với yêu cầu cần thiết, và tổng hợp tất cả các mô hình có thể phức tạp gấp nghìn lần. Sự phức tạp thừa thãi này gây ra nhiều khó khăn cho cả các nhà phát triển và người dùng cuối.

Sự phức tạp này đặt ra những hạn chế nghiêm trọng đối với sự phát triển và bảo trì của hệ thống trong công ty. Mỗi yếu tố mới trong mô hình yêu cầu thêm mã, triển khai các logic mới, kiểm tra cẩn thận và thích ứng với các giải pháp đã tồn tại. Tất cả những điều này làm tăng chi phí và làm chậm quá trình làm việc của nhóm tự động hóa trong công ty, biến ngay cả những nhiệm vụ đơn giản thành những quá trình tốn kém và tốn thời gian.

Sự phức tạp ảnh hưởng đến tất cả các cấp độ kiến trúc dữ liệu. Trong các cơ sở dữ liệu quan hệ, điều này thể hiện ở việc gia tăng số lượng bảng và cột, thường là thừa thãi. Trong các hệ thống hướng đối tượng, độ phức tạp gia tăng do sự tồn tại của nhiều lớp và các thuộc tính liên quan lẫn nhau. Trong các định dạng như XML hoặc JSON, sự cồng kềnh thể hiện qua các cấu trúc lồng ghép rắc rối, các khóa duy nhất và các sơ đồ không nhất quán.

> Sự phức tạp thừa thãi của các mô hình dữ liệu không chỉ làm giảm hiệu quả của hệ thống mà còn khiến cho người dùng cuối và trong tương lai, các mô hình ngôn ngữ lớn (LLM) gặp khó khăn trong việc tiếp cận. Vấn đề hiểu biết và độ phức tạp của các mô hình dữ liệu và việc xử lý chúng đặt ra câu hỏi: làm thế nào để dữ liệu có thể được sử dụng đủ đơn giản để thực sự mang lại lợi ích một cách nhanh chóng.

Ngay cả khi mô hình dữ liệu được lựa chọn một cách khéo léo, lợi ích thực tiễn của chúng giảm sút mạnh mẽ nếu quyền truy cập vào dữ liệu bị hạn chế. Các định dạng sở hữu và các nền tảng đóng góp phần cản trở tích hợp, làm phức tạp hóa tự động hóa và lấy đi quyền kiểm soát của công ty đối với thông tin của chính mình, tạo ra không chỉ là silo dữ liệu mới mà còn là silo bị khóa, chìa khóa chỉ có thể nhận được thông qua sự cho phép của nhà cung cấp. Để hiểu quy mô của vấn đề, quan trọng là xem xét cách mà sự đóng kín của

các hệ thống ảnh hưởng đến các quy trình kỹ thuật số trong lĩnh vực xây dựng.

Các định dạng sở hữu và ảnh hưởng của chúng đến các quá trình số hóa

Một trong những vấn đề chính mà các công ty xây dựng gặp phải trong quá trình số hóa là quyền truy cập hạn chế vào dữ liệu. Điều này làm khó khăn cho việc tích hợp các hệ thống, giảm chất lượng thông tin và làm phức tạp hóa tổ chức các quy trình hiệu quả. Nguồn gốc của những khó khăn này thường nằm ở việc sử dụng các định dạng sở hữu và các giải pháp phần mềm đóng.

Thật không may, cho đến nay nhiều chương trình được sử dụng trong ngành xây dựng cho phép người dùng chỉ lưu trữ dữ liệu trong các định dạng riêng của họ hoặc trong các kho đám mây, mà việc truy cập chỉ có thể thực hiện thông qua các giao diện bị hạn chế nghiêm ngặt. Đồng thời, thường xuyên các giải pháp như vậy được xây dựng dựa trên các hệ thống đóng kín hơn của các nhà cung cấp lớn hơn. Kết quả là ngay cả những nhà phát triển muốn cung cấp các kiến trúc mở hơn cũng buộc phải tuân theo các quy tắc do các nhà cung cấp lớn thiết lập.

> Trong khi các hệ thống quản lý dữ liệu xây dựng hiện đại ngày càng hỗ trợ các định dạng và tiêu chuẩn mở (Hình 3.15), cơ sở dữ liệu CAD (BIM), cùng với các hệ thống ERP và CAFM liên quan vẫn còn là những "hòn đảo" sở hữu độc lập trong bối cảnh số của ngành (Hình 3.211).-

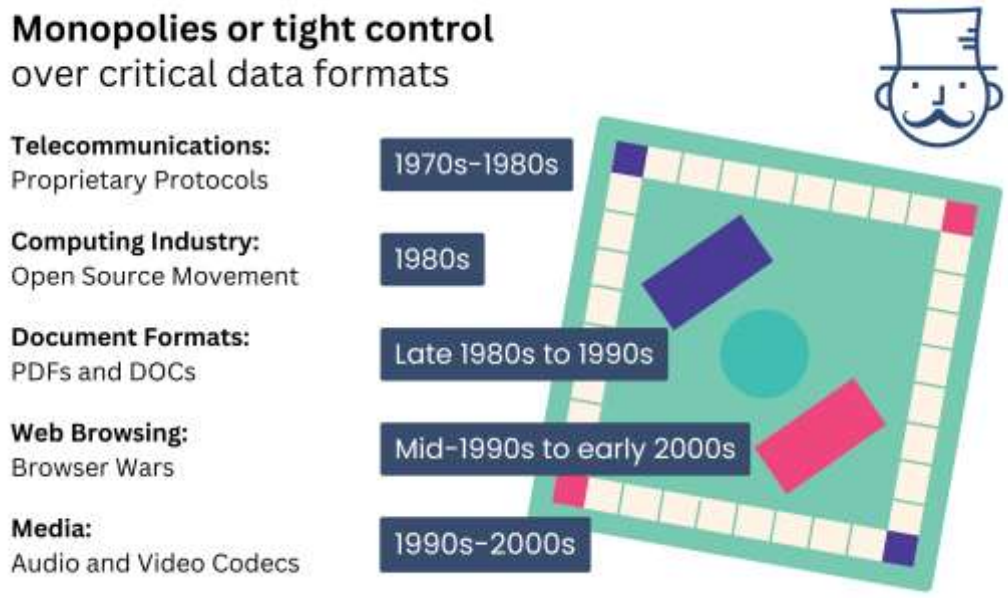


Hình 3.211 Tính chất đóng kín và sở hữu của dữ liệu tạo ra rào cản đối với việc tích hợp và truy cập dữ liệu.

Sự đóng kín và độc quyền về định dạng và giao thức không chỉ là vấn đề của ngành xây dựng. Trong nhiều lĩnh vực kinh tế, cuộc chiến chống lại các tiêu chuẩn đóng và quyền truy cập hạn chế vào dữ liệu đã bắt đầu với việc làm chậm đổi mới (Hình 3.212), sự tồn tại của các rào cản nhân tạo đối với việc gia nhập của các diễn viên mới và làm sâu sắc thêm sự phụ thuộc vào các nhà cung cấp lớn. Trong bối cảnh sự gia tăng

nhanh chóng của tầm quan trọng của dữ liệu, các cơ quan chống độc quyền đơn giản là không kịp phản ứng với các thách thức liên quan đến các thị trường số mới và cuối cùng, các định dạng đóng và quyền truy cập đóng vào dữ liệu về cơ bản trở thành những "biên giới" số, kìm hãm sự di chuyển của thông tin và sự phát triển.

> *Nếu máy móc sản xuất mọi thứ mà chúng ta cần, thì vị trí của chúng ta sẽ phụ thuộc vào cách phân phối những tài nguyên này. Mọi người sẽ có thể tận hưởng cuộc sống trong sự đủ đầy chỉ khi tài sản do máy móc sản xuất trở thành tài sản chung. Ngược lại, đa số người dân có thể sẽ sống trong cảnh nghèo đói khủng khiếp nếu những người sở hữu máy móc có thể vận động thành công chống lại việc phân phối lại tài sản. Cho đến nay, có vẻ như tình hình đang đi theo phương án thứ hai, công nghệ dẫn đến sự gia tăng khoảng cách bất bình đẳng.*
>
> *— Stephen Hawking, nhà vật lý thiên văn, 2015*



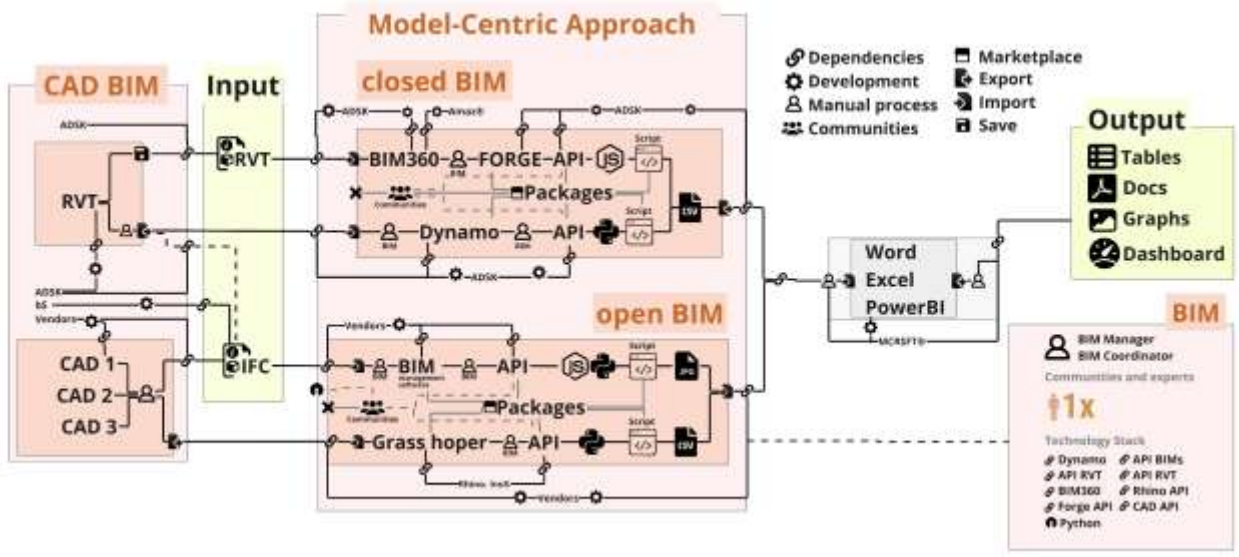
Hình 3.212 Quyền sở hữu độc quyền đối với các định dạng và giao thức dữ liệu chính không phải là một vấn đề riêng biệt của ngành xây dựng.

Cuối cùng, do việc truy cập hạn chế vào các cơ sở dữ liệu của phần mềm, các nhà quản lý dữ liệu, nhà phân tích, chuyên gia CNTT và lập trình viên, những người phát triển ứng dụng truy cập và xử lý dữ liệu cũng như tự động hóa trong ngành xây dựng, hiện đang phải đối mặt với nhiều phụ thuộc vào nhà cung cấp phần mềm (Hình 3.213). Những phụ thuộc này dưới dạng các lớp truy cập bổ sung yêu cầu tạo ra các giải pháp với các kết nối API chuyên biệt và các công cụ, phần mềm đặc thù.-

> API (Giao diện Lập trình Ứng dụng) — là một giao diện được chuẩn hóa, cho phép một chương trình tương tác với một chương trình khác bằng cách trao đổi dữ liệu và chức năng mà không cần truy cập vào mã nguồn. API mô tả các yêu cầu mà hệ thống bên ngoài có thể thực hiện, định dạng mà chúng cần có và những phản hồi mà nó nhận được. Đây là một "hợp đồng" tiêu chuẩn giữa các mô-đun phần mềm.

Số lượng lớn các phụ thuộc vào các giải pháp đóng khiến toàn bộ kiến trúc mã và logic quy trình kinh doanh trong công ty trở thành "kiến trúc mì spaghetti" từ các công cụ phụ thuộc vào chính sách của nhà cung cấp phần mềm về việc cung cấp quyền truy cập dữ liệu chất lượng.

Sự phụ thuộc vào các giải pháp và nền tảng khép kín không chỉ dẫn đến mất tính linh hoạt mà còn tạo ra những rủi ro kinh doanh thực sự. Sự thay đổi điều kiện cấp phép, việc đóng cửa truy cập vào dữ liệu, thay đổi định dạng hoặc cấu trúc API - tất cả những điều này có thể chặn đứng các quy trình quan trọng. Đột nhiên, việc cập nhật một bảng dữ liệu yêu cầu phải làm lại cả một khối tích hợp và kết nối (Hình 3.213), trong khi bất kỳ bản cập nhật quy mô lớn nào của phần mềm hoặc API của nhà cung cấp đều trở thành mối đe dọa tiềm tàng đối với sự ổn định của toàn bộ hệ thống công ty.-



Hình 3.213 Ví dụ về nhiều phụ thuộc trong việc xử lý dữ liệu CAD tạo ra rào cản cho việc tích hợp dữ liệu trong hệ sinh thái của các công ty xây dựng.

> Các nhà phát triển và kiến trúc sư hệ thống trong những điều kiện như vậy buộc phải làm việc không phải để đi trước, mà để tồn tại. Thay vì triển khai các giải pháp mới, họ thích nghi. Thay vì phát triển, họ cố gắng duy trì tính tương thích. Thay vì tự động hóa và tăng tốc các quy trình, họ dành thời gian để nghiên cứu các giao diện đóng, tài liệu API và tái cấu trúc mã lặp đi lặp lại.

Làm việc với các định dạng và hệ thống đóng không chỉ là một vấn đề kỹ thuật mà còn là một hạn chế chiến lược. Mặc dù có những cơ hội rõ ràng mà công nghệ tự động hóa hiện đại, trí tuệ nhân tạo, mô hình ngôn ngữ lớn (LLM) và phân tích dự đoán mang lại, nhiều công ty vẫn không thể khai thác toàn bộ tiềm năng của chúng. Những rào cản do định dạng sở hữu đặt ra đã tước đoạt quyền truy cập của doanh nghiệp vào dữ liệu của chính mình. Đây có lẽ là sự châm biếm lớn nhất trong quá trình chuyển đổi số trong ngành xây dựng.

-

> Sự minh bạch về dữ liệu và tính mở của hệ thống không phải là sự xa xỉ mà là điều kiện cần thiết cho tốc độ và hiệu quả. Trong bối cảnh thiếu tính minh bạch, các quy trình kinh doanh thường bị lấp đầy bởi quan liêu không cần thiết, chuỗi phê duyệt nhiều lớp và sự phụ thuộc ngày càng tăng vào nguyên tắc HiPPO - quyết định dựa trên ý kiến của người có thu nhập cao nhất.

Tuy nhiên, một sự chuyển mình đang hình thành. Mặc dù các giải pháp sở hữu đang thống trị, ngày càng nhiều công ty nhận ra những hạn chế của kiến trúc được xây dựng theo mô hình Cách mạng Công nghiệp lần thứ tư. Hiện tại, hướng đi đang chuyển sang các nguyên tắc của Cách mạng Công nghiệp lần thứ năm, nơi dữ liệu được coi là tài sản chiến lược, các giao diện mở (API) và tính tương tác thực sự giữa các hệ thống giữ vai trò trung tâm.

Sự chuyển mình này đánh dấu sự thoát ra khỏi các hệ sinh thái khép kín sang kiến trúc số linh hoạt, mô-đun, nơi các định dạng, tiêu chuẩn mở và sự minh bạch trong việc trao đổi dữ liệu giữ vai trò then chốt.

Các định dạng mở thay đổi cách tiếp cận đến số hóa

Ngành xây dựng là một trong những lĩnh vực cuối cùng tiếp cận vấn đề đóng kín và sở hữu dữ liệu. Khác với các lĩnh vực khác trong nền kinh tế, tiến trình số hóa tại đây diễn ra chậm chạp. Nguyên nhân bao gồm tính bảo thủ truyền thống của ngành, sự thống trị của các giải pháp rời rạc và sự ăn sâu của quy trình quản lý tài liệu giấy. Trong suốt nhiều thập kỷ, các quy trình chính trong xây dựng dựa vào bản vẽ vật lý, cuộc gọi điện thoại và cơ sở dữ liệu không đồng bộ. Trong bối cảnh này, các định dạng đóng đã được coi là tiêu chuẩn, chứ không phải là rào cản.

Kinh nghiệm từ các ngành khác cho thấy: việc loại bỏ các rào cản liên quan đến dữ liệu đóng dẫn đến sự bùng nổ của đổi mới, tăng tốc phát triển và gia tăng cạnh tranh. Trong khoa học, việc trao đổi dữ liệu mở giúp tăng tốc độ khám phá và phát triển hợp tác quốc tế. Trong y tế, việc này cải thiện hiệu quả chẩn đoán và điều trị. Trong kỹ thuật phần mềm, nó tạo ra các hệ sinh thái hợp tác sáng tạo và nhanh chóng hoàn thiện sản phẩm.

> Theo báo cáo của McKinsey "Dữ liệu mở: Khai phóng đổi mới và năng suất thông qua dòng thông tin" năm 2013, dữ liệu mở có khả năng giải phóng từ 3 đến 5 nghìn tỷ đô la hàng năm trong bảy ngành kinh tế chính, bao gồm xây dựng, giao thông, y tế và năng lượng. Theo cùng một nghiên cứu, các hệ sinh thái dữ liệu phi tập trung cho phép các công ty xây dựng lớn và nhà thầu giảm chi phí phát triển và duy trì phần mềm, từ đó tăng tốc việc áp dụng công nghệ số.

Sự chuyển đổi sang kiến trúc mở, đã bắt đầu từ lâu ở các lĩnh vực khác của nền kinh tế, dần dần cũng ảnh hưởng tới ngành xây dựng. Các công ty lớn và các nhà thầu nhà nước, đặc biệt là các tổ chức tài chính kiểm soát đầu tư vào các dự án xây dựng, ngày càng đưa ra yêu cầu về việc sử dụng dữ liệu mở và đảm bảo quyền truy cập vào mã nguồn của các tính toán, bảng định giá và ứng dụng. Các nhà phát triển không chỉ được yêu cầu tạo ra các giải pháp kỹ thuật số và trình bày các số liệu cuối cùng của dự án — mà còn được kỳ vọng về tính minh bạch, khả năng tái tạo và độc lập khỏi các nhà cung cấp ứng dụng bên ngoài.

> Việc sử dụng các giải pháp mã nguồn mở giúp khách hàng yên tâm rằng ngay cả khi các nhà phát triển bên ngoài ngừng hợp tác hoặc rời bỏ dự án, điều đó cũng sẽ không ảnh hưởng đến khả năng phát triển tiếp theo của các công cụ và hệ thống. Một trong những lợi thế chính của dữ liệu mở là khả năng xóa bỏ sự phụ thuộc của các nhà phát triển ứng dụng vào các nền tảng cụ thể để truy cập vào dữ liệu.

Nếu công ty không thể hoàn toàn từ bỏ các giải pháp độc quyền, một lựa chọn hợp lý có thể là sử dụng các phương pháp kỹ thuật đảo ngược. Những phương pháp hợp pháp và có cơ sở kỹ thuật này cho phép chuyển đổi các định dạng đóng thành các định dạng có thể truy cập hơn, có cấu trúc và dễ dàng tích hợp. Điều này đặc biệt quan trọng trong những trường hợp cần kết nối với các hệ thống lỗi thời hoặc di chuyển thông tin từ một môi trường phần mềm sang môi trường khác.

> Một trong những ví dụ điển hình trong lịch sử chuyển đổi sang định dạng mở và việc áp dụng kỹ thuật đảo ngược (hack hợp pháp các hệ thống độc quyền) trong xây dựng là cuộc chiến giành việc mở định dạng DWG, được sử dụng rộng rãi trong các hệ thống thiết kế tự động (CAD). Năm 1998, để đối phó với sự độc quyền của một trong những nhà cung cấp phần mềm, 15 nhà cung cấp CAD khác đã thành lập một liên minh mang tên "Open DWG", nhằm cung cấp cho các nhà phát triển các công cụ miễn phí và độc lập để làm việc với định dạng DWG (chuẩn de facto cho việc trao đổi bản vẽ) mà không cần sử dụng phần mềm độc quyền hoặc API đóng. Sự kiện này đã trở thành một bước ngoặt, cho phép hàng chục nghìn công ty tiếp cận miễn phí với định dạng đóng của giải pháp CAD phổ biến từ cuối những năm 1980 cho đến ngày nay và tạo ra các giải pháp tương thích, góp phần phát triển cạnh tranh trên thị trường CAD. Hiện nay, SDK "Open DWG", lần đầu tiên được tạo ra vào năm 1996, được sử dụng trong hầu hết các giải pháp mà có thể nhập, chỉnh sửa và xuất định dạng DWG, ngoài ứng dụng chính thức của nhà phát triển định dạng DWG.

Các cuộc chuyển đổi tương tự cũng đang diễn ra bắt buộc ở các tập đoàn công nghệ khác. Microsoft, từng là biểu tượng của cách tiếp cận độc quyền, đã mở mã nguồn của.NET Framework, bắt đầu sử dụng Linux trong cơ sở hạ tầng dịch vụ đám mây Azure và mua lại GitHub để củng cố vị thế trong cộng đồng mã nguồn mở. Công ty Meta (trước đây là Facebook) đã phát hành các mô hình AI mã nguồn mở, chẳng hạn như dòng sản phẩm Llama, để thúc đẩy đổi mới và hợp tác trong lĩnh vực phát triển AI. Giám đốc điều hành Mark Zuckerberg cho rằng các nền tảng mã nguồn mở sẽ dẫn dắt tiến bộ công nghệ trong thập kỷ tới.

> Mã nguồn mở là một mô hình phát triển và phân phối phần mềm, trong đó mã nguồn được mở để sử dụng tự do, nghiên cứu, sửa đổi và phân phối.

Dữ liệu mở và các giải pháp mã nguồn mở không chỉ đơn thuần là một xu hướng, mà còn trở thành nền tảng cho sự bền vững kỹ thuật số. Chúng mang lại cho các công ty sự linh hoạt, khả năng thích ứng với sự thay đổi, kiểm soát các giải pháp của chính mình và khả năng mở rộng các quy trình kỹ thuật số mà không phải phụ thuộc vào chính sách của nhà cung cấp. Và, điều quan trọng không kém, chúng trả lại cho doanh nghiệp quyền kiểm soát tài nguyên quý giá nhất của thế kỷ XXI — dữ liệu của chính mình.

Thay đổi mô hình: Mã nguồn mở như là sự kết thúc của kỷ nguyên thống trị của các nhà cung cấp phần mềm.

Ngành công nghiệp xây dựng sẽ trải qua một sự chuyển đổi mà không thể kiếm tiền theo cách thông thường. Khái niệm chuyển sang sử dụng dữ liệu, cách tiếp cận theo trung tâm dữ liệu và việc sử dụng các công cụ mã nguồn mở dẫn tới việc xem xét lại các quy tắc mà các gã khổng lồ phần mềm trên thị trường dựa vào.

> Khác với những chuyển biến công nghệ trước, sự chuyển đổi này sẽ không được các nhà cung cấp chủ động thúc đẩy. Sự thay đổi mô hình đe dọa đến các mô hình kinh doanh truyền thống của họ, dựa trên cấp phép, đăng ký và tư vấn. Thực tế mới không yêu cầu một sản phẩm "trong hộp" hay đăng ký trả phí — nó đòi hỏi phải tái cấu trúc quy trình và tư duy.

Để quản lý và phát triển các giải pháp trung tâm dữ liệu dựa trên công nghệ mở, các công ty sẽ cần xem xét lại các quy trình nội bộ. Các chuyên gia từ các phòng ban khác nhau sẽ không chỉ ở mức độ tương tác, mà còn cần phải xem xét lại các phương pháp làm việc cùng nhau.

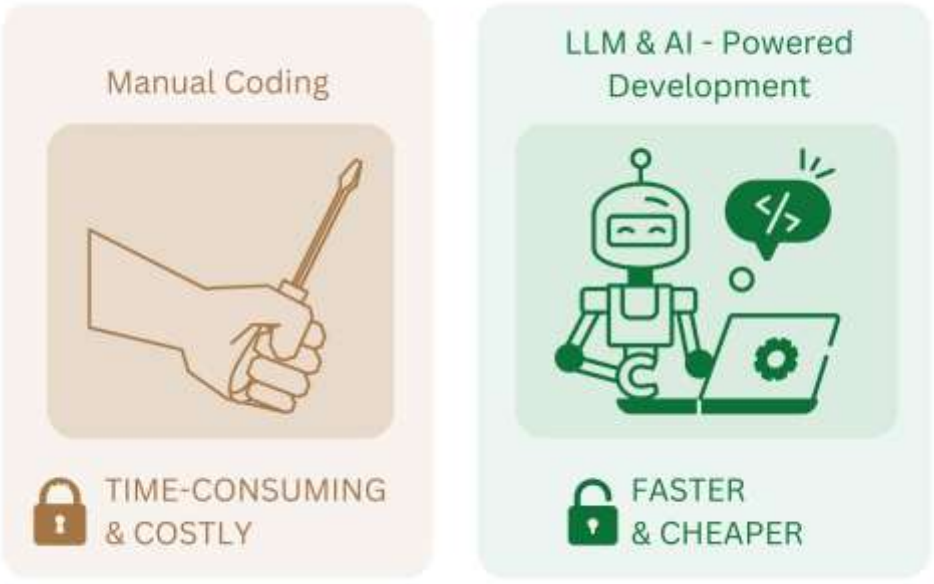
Mô hình mới này ngụ ý việc sử dụng dữ liệu mở và các giải pháp mã nguồn mở, trong đó vai trò quan trọng trong việc tạo ra mã phần mềm sẽ không còn chỉ thuộc về các lập trình viên, mà còn bởi các công cụ dựa trên trí tuệ nhân tạo và các mô hình ngôn ngữ lớn (LLM). Đến giữa năm 2024, hơn 25% mã mới trong Google sẽ được tạo ra bằng sự hỗ trợ của AI. Trong tương lai, việc lập trình với LLM sẽ hoàn thành 80% công việc chỉ trong 20% thời gian.

> Theo một nghiên cứu của McKinsey năm 2020, trong lĩnh vực phân tích, GPU ngày càng chiếm ưu thế hơn CPU - nhờ hiệu suất cao của chúng và được hỗ trợ bởi các công cụ mã nguồn mở hiện đại. Điều này cho phép các công ty tăng tốc độ xử lý dữ liệu mà không cần đầu tư đáng kể vào phần mềm đắt tiền hay thuê các chuyên gia khan hiếm.

Các công ty tư vấn hàng đầu như McKinsey, PwC và Deloitte nhấn mạnh tầm quan trọng ngày càng tăng của các tiêu chuẩn mở, các ứng dụng mã nguồn mở trong nhiều ngành công nghiệp khác nhau.

> Theo báo cáo PwC Open Source Monitor năm 2019, 69% các công ty có từ 100 nhân viên trở lên đã sử dụng các giải pháp mã nguồn mở một cách có ý thức. Đặc biệt, OSS được sử dụng mạnh mẽ trong các công ty lớn: 71% trong số các công ty có 200–499 nhân viên, 78% trong danh mục 500–1999 nhân viên và lên tới 86% trong số các công ty có hơn 2000 nhân viên. Theo báo cáo Synopsys OSSRA năm 2023, 96% các cơ sở mã được phân tích chứa các thành phần mã nguồn mở.

Tương lai của vai trò lập trình viên không phải là viết mã bằng tay, mà là thiết kế các mô hình dữ liệu, kiến trúc luồng và quản lý các tác nhân AI tạo ra các tính toán cần thiết theo yêu cầu. Giao diện người dùng sẽ trở nên tối giản, trong khi tương tác sẽ mang tính đối thoại. Lập trình cổ điển sẽ nhường chỗ cho thiết kế cấp cao và điều phối các giải pháp kỹ thuật số (Hình. 3.214). Các xu hướng hiện đại — như các nền tảng low-code (Hình. 7.46) và hệ sinh thái hỗ trợ LLM (Hình. 7.44) — sẽ giúp giảm đáng kể chi phí cho việc phát triển và bảo trì các hệ thống CNTT.--



Hình. 3.214 Nếu như ngày nay các ứng dụng được tạo ra thủ công bởi lập trình viên, thì trong tương lai một phần lớn mã sẽ được tạo ra bởi các giải pháp dựa trên AI và LLM.

Sự chuyển mình này sẽ không giống như những lần trước đây và có khả năng các nhà cung cấp phần mềm lớn sẽ không trở thành chất xúc tác cho nó.

> Nghiên cứu của Trường Kinh doanh Harvard "Giá trị của phần mềm mã nguồn mở" năm 2024 [40], tổng giá trị của phần mềm mã nguồn mở được đánh giá từ hai góc độ. Một mặt, nếu tính toán xem sẽ mất bao nhiêu chi phí để tạo ra tất cả các giải pháp mã nguồn mở hiện có từ đầu, thì số tiền này vào khoảng 4,15 tỷ đô la. Mặt khác, nếu tưởng tượng rằng mỗi công ty phát triển các giải pháp tương tự mã nguồn mở của riêng mình (điều này đang diễn ra phổ biến), mà không có quyền truy cập vào các công cụ đã có, thì tổng chi phí của doanh nghiệp cho điều này sẽ đạt tới con số khổng lồ 8,8 nghìn tỷ đô la — đó là chi phí của nhu cầu.

> Không khó để đoán rằng không một nhà cung cấp phần mềm lớn nào muốn thu hẹp thị trường phần mềm với giá trị tiềm năng 8,8 nghìn tỷ đô la xuống còn 4,15 tỷ. Điều này có nghĩa là sự giảm sút nhu cầu sẽ lớn hơn 2.000 lần. Sự chuyển mình như vậy đơn giản là không có lợi cho các nhà cung cấp, những người có mô hình kinh doanh dựa vào việc duy trì sự phụ thuộc của khách hàng vào các giải pháp đóng trong nhiều năm. Do đó, các công ty mong đợi rằng ai đó sẽ cung cấp cho họ một giải pháp tiện lợi và mở "trọn gói" có thể sẽ bị thất vọng — những người bán như vậy đơn giản là sẽ không xuất hiện.

Sự chuyển đổi sang kiến trúc kỹ thuật số mở không có nghĩa là giảm số lượng việc làm hay thu nhập. Ngược lại, nó tạo ra điều kiện cho các mô hình kinh doanh linh hoạt và thích ứng, có thể dần dần thay thế thị trường truyền thống của giấy phép và phần mềm hộp.

Thay vì bán giấy phép — là các dịch vụ, thay vì các định dạng đóng — là các nền tảng mở, thay vì phụ thuộc vào nhà cung cấp — là sự độc lập và khả năng xây dựng các giải pháp đáp ứng nhu cầu thực tế. Những người trước đây chỉ sử dụng các công cụ sẽ có thể trở thành đồng tác giả của chúng. Còn những ai biết làm việc với dữ liệu, mô hình, kịch bản và logic — sẽ ở trung tâm của nền kinh tế kỹ thuật số mới trong

ngành. Chúng ta sẽ nói chi tiết hơn về những thay đổi này và những vai trò mới, mô hình kinh doanh và hình thức hợp tác đang hình thành xung quanh dữ liệu mở trong phần cuối cùng, phần mười của cuốn sách.

> Các giải pháp dựa trên dữ liệu mở và mã nguồn mở sẽ cho phép các công ty tập trung không vào việc chiến đấu với các API lỗi thời và tích hợp các hệ thống đóng mà vào hiệu quả của quy trình kinh doanh. Sự chuyển đổi có ý thức sang kiến trúc mở mang lại khả năng cải thiện năng suất một cách đáng kể và giảm sự phụ thuộc vào nhà cung cấp.

Sự chuyển mình sang thực tại mới không chỉ là sự thay đổi trong cách tiếp cận phát triển phần mềm, mà còn là việc tái định nghĩa nguyên tắc làm việc với dữ liệu. Trong tâm điểm của sự chuyển đổi này không phải là mã nguồn, mà là thông tin: cấu trúc, khả năng truy cập và khả năng diễn giải của nó. Và chính ở đây, dữ liệu mở và có cấu trúc trở thành trọng tâm, trở thành một phần không thể thiếu của kiến trúc số mới.

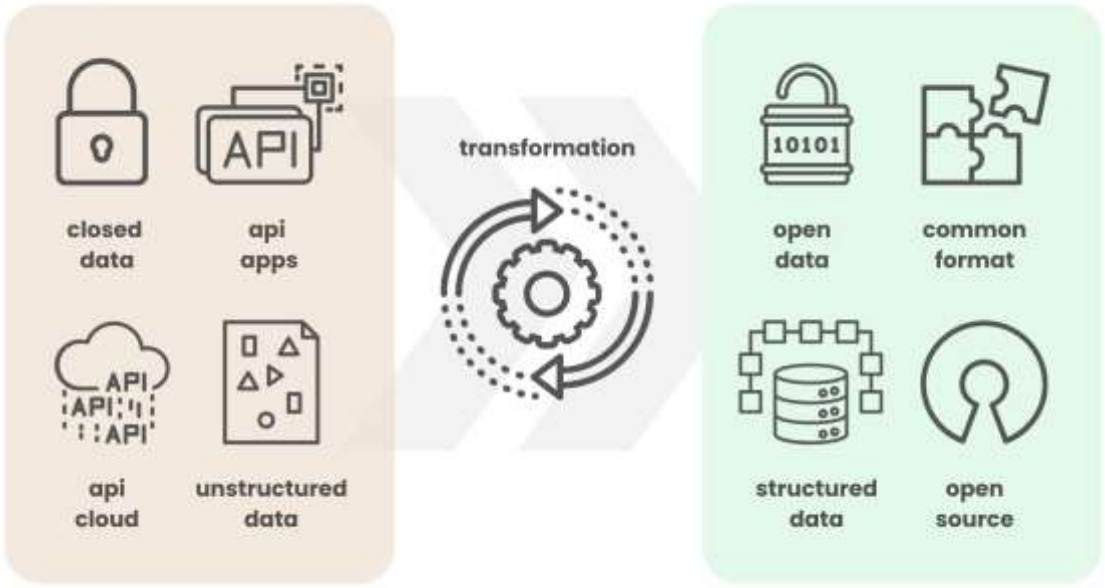
Dữ liệu mở được cấu trúc: nền tảng của chuyển đổi số

Nếu trong những thập kỷ trước, sự ổn định của doanh nghiệp phần lớn được xác định bởi sự lựa chọn các giải pháp phần mềm và sự phụ thuộc vào các nhà cung cấp cụ thể, thì trong điều kiện kinh tế số hiện nay, yếu tố then chốt trở thành chất lượng dữ liệu và khả năng làm việc hiệu quả với chúng. Mã nguồn mở là một phần quan trọng của công nghệ mới, nhưng tiềm năng của nó chỉ thực sự được khai thác khi có dữ liệu có tổ chức, dễ hiểu và có thể đọc máy. Giữa tất cả các loại mô hình dữ liệu, dữ liệu mở có cấu trúc trở thành nền tảng cho quá trình chuyển đổi số bền vững.

Lợi thế chính của dữ liệu mở có cấu trúc là khả năng diễn giải một cách rõ ràng và khả năng xử lý tự động. Điều này cho phép nâng cao hiệu quả một cách đáng kể, cả ở cấp độ cá nhân và trên quy mô toàn tổ chức.

> Theo báo cáo của Deloitte "Quá trình chuyển giao dữ liệu trong các chuyển đổi doanh nghiệp", công việc với IT để quản lý chuyển giao dữ liệu có cấu trúc là rất quan trọng. Theo báo cáo của chính phủ Vương quốc Anh "Phân tích dữ liệu và AI trong việc thực hiện dự án của chính phủ" (2024), việc loại bỏ các rào cản trong việc chia sẻ dữ liệu giữa các dự án và tổ chức khác nhau là một yếu tố then chốt để nâng cao hiệu quả trong quản lý dự án. Tài liệu nhấn mạnh rằng tiêu chuẩn hóa các định dạng dữ liệu và triển khai các nguyên tắc của dữ liệu mở cho phép tránh việc trùng lặp thông tin, giảm thiểu thời gian lãng phí và nâng cao độ chính xác của dự đoán.

Đối với ngành xây dựng, nơi mà traditionally có mức độ phân mảnh cao và sự đa dạng về định dạng, quá trình cấu trúc-đồng nhất và dữ liệu mở có cấu trúc đóng vai trò quyết định trong việc hình thành các quy trình nhất quán và có thể quản lý. Chúng cho phép các bên tham gia dự án tập trung vào việc tăng năng suất, thay vì phải giải quyết các vấn đề kỹ thuật liên quan đến sự không tương thích của các nền tảng đóng, mô hình dữ liệu và định dạng.



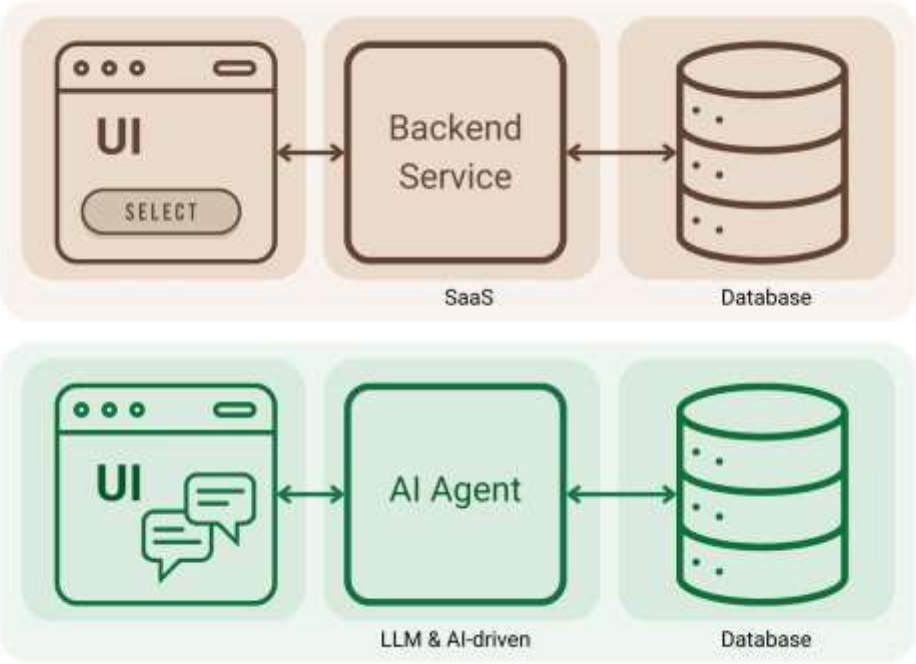
Dữ liệu mở có cấu trúc làm giảm sự phụ thuộc vào các giải pháp phần mềm và nền tảng và tăng tốc độ đổi mới.

Các công nghệ hiện đại và công cụ mà chúng ta sẽ xem xét chi tiết trong cuốn sách này không chỉ cho phép thu thập thông tin, mà còn tự động làm sạch nó: loại bỏ trùng lặp, sửa chữa lỗi, chuẩn hóa các giá trị. Điều này có nghĩa là các nhà phân tích và kỹ sư không làm việc với các tài liệu rời rạc, mà với một cơ sở tri thức có tổ chức, phù hợp để phân tích, tự động hóa và ra quyết định.

> *Làm cho nó đơn giản nhất có thể, nhưng không đơn giản hơn.*
>
> *– Albert Einstein, nhà vật lý lý thuyết (nguồn gốc của câu nói này còn gây tranh cãi).*

Ngày nay, hầu hết các giao diện người dùng để làm việc với dữ liệu có thể được tạo tự động - mà không cần viết mã thủ công cho từng trường hợp kinh doanh. Để làm điều này, cần có một lớp hạ tầng có thể hiểu cấu trúc dữ liệu, mô hình và logic mà không cần hướng dẫn bổ sung (Hình. 4.115). Chính những dữ liệu có cấu trúc đã làm cho phương pháp này trở nên khả thi: các biểu mẫu, bảng, bộ lọc và các chế độ xem có thể được tạo tự động với chi phí lập trình tối thiểu.

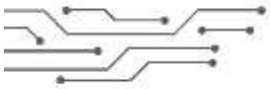
Các giao diện quan trọng nhất, quyết định cho người dùng, vẫn có thể yêu cầu chỉnh sửa thủ công. Nhưng trong hầu hết các trường hợp - từ 50 đến 90 phần trăm của các kịch bản làm việc - việc tự động tạo ra ứng dụng và tính toán là đủ mà không cần sử dụng các ứng dụng chuyên dụng (Hình. 3.216), điều này giảm đáng kể chi phí phát triển và duy trì, giảm số lượng lỗi và tăng tốc thực hiện các giải pháp số.



Hình. 3.216 Các mô hình kiến trúc làm việc với dữ liệu: kiến trúc ứng dụng truyền thống và mô hình hướng tới AI với LLM.

Việc chuyển từ các kiến trúc được xây dựng dựa trên các ứng dụng riêng lẻ sang các hệ thống được quản lý thông minh dựa trên các mô hình ngôn ngữ (LLM) - là bước tiếp theo trong sự tiến hóa số. Trong kiến trúc này, dữ liệu có cấu trúc không chỉ là đối tượng lưu trữ, mà còn là nền tảng cho tương tác với các công cụ AI, có khả năng phân tích, diễn giải và đề xuất hành động dựa trên ngữ cảnh.

Trong các chương tiếp theo, chúng tôi sẽ xem xét các ví dụ thực tế về việc triển khai kiến trúc dựa trên dữ liệu có cấu trúc mở, cũng như cho thấy cách mà các mô hình ngôn ngữ được áp dụng trong việc diễn giải, xác thực và xử lý dữ liệu tự động. Những trường hợp thực tiễn này sẽ giúp hiểu rõ hơn cách mà logic số mới hoạt động trên thực tế - và những lợi ích mà nó mang lại cho các công ty sẵn sàng chuyển đổi.



CHƯƠNG 3.3. **Mô hình ngôn ngữ lớn (LLM) và vai trò của chúng trong xử lý dữ liệu và quy trình kinh doanh**

Các cuộc trò chuyện LLM: ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok cho tự động hóa quy trình xử lý dữ liệu

Sự xuất hiện của các mô hình ngôn ngữ lớn (LLM) đã trở thành sự tiếp nối tự nhiên của việc hướng tới dữ liệu mở có cấu trúc và triết lý mã nguồn mở. Khi dữ liệu trở nên có tổ chức, có sẵn và có thể đọc được bởi máy, bước tiếp theo là công cụ có khả năng tương tác với thông tin này mà không cần phải viết mã phức tạp hoặc có kiến thức kỹ thuật chuyên biệt.

> LLM là sản phẩm trực tiếp của sự cởi mở: các bộ dữ liệu lớn mở, các ấn phẩm và phong trào mã nguồn mở. Nếu không có các bài viết khoa học mở, dữ liệu văn bản công khai và văn hóa phát triển hợp tác, sẽ không có ChatGPT hay bất kỳ LLM nào khác. LLM, theo một cách nào đó, là "mật rượu" của kiến thức số mà nhân loại đã tích lũy, được thu thập và huấn luyện dựa trên các nguyên tắc cởi mở.

Các mô hình ngôn ngữ lớn hiện đại (LLM - Large Language Models), chẳng hạn như ChatGPT® (OpenAI), LlaMa™ (Meta AI), Mistral DeepSeek™, Grok™ (xAI), Claude™ (Anthropic), QWEN™ cung cấp cho người dùng khả năng đăng câu hỏi dữ liệu bằng ngôn ngữ tự nhiên. Điều này làm cho việc làm việc với thông tin trở nên dễ tiếp cận không chỉ cho các nhà phát triển, mà còn cho các nhà phân tích, kỹ sư, nhà thiết kế, quản lý và các chuyên gia khác, những người trước đây xa lạ với lập trình.

> LLM (Large Language Model) là một trí tuệ nhân tạo được huấn luyện để hiểu và tạo văn bản dựa trên một lượng lớn dữ liệu được thu thập từ toàn bộ internet. Nó có khả năng phân tích ngữ cảnh, trả lời câu hỏi, tham gia đối thoại, viết văn bản và tạo mã lập trình.

Nếu trước đây việc trực quan hóa, xử lý hoặc phân tích dữ liệu yêu cầu kiến thức về ngôn ngữ lập trình chuyên biệt như Python, SQL, R hoặc Scala, cũng như khả năng làm việc với các thư viện như Pandas, Polars hoặc DuckDB và nhiều thư viện khác, thì bắt đầu từ năm 2023, tình hình đã thay đổi một cách đáng kể. Bây giờ, người dùng chỉ cần mô tả những gì họ muốn nhận được — và mô hình sẽ tự động sinh mã, thực thi nó, hiển thị bảng hoặc biểu đồ và giải thích kết quả. Lần đầu tiên sau hàng thập kỷ, sự phát triển công nghệ không đi theo con đường phức tạp hóa, mà theo hướng đơn giản hóa và làm cho thông tin trở nên dễ tiếp cận hơn.

Nguyên tắc này — "xử lý dữ liệu bằng từ ngữ (prompt)" — đã đánh dấu một giai đoạn mới trong tiến hóa làm việc với thông tin, thực chất nâng cao việc tạo ra giải pháp lên một cấp độ tr trừ cao hơn. Tương tự như việc trước đây người dùng không cần phải hiểu các nguyên tắc kỹ thuật của Internet để chạy cửa hàng trực tuyến hoặc tạo trang web bằng WordPress, Joomla và các hệ thống mã nguồn mở khác (tác giả cuốn sách đã làm việc với các hệ thống như vậy từ năm 2005, bao gồm trong lĩnh vực nền tảng trực tuyến giáo dục và kỹ thuật) — điều này đã dẫn đến sự phát triển mạnh mẽ của nội dung số và kinh doanh trực tuyến — ngày nay, kỹ sư, nhà phân tích và quản lý có thể tự động hóa quy trình làm việc mà không cần biết các ngôn ngữ lập trình. Điều này được hỗ trợ bởi các mô hình ngôn ngữ lớn mạnh mẽ — cả miễn phí và mã nguồn mở, như LLaMA, Mistral, Qwen, DeepSeek và nhiều mô hình khác — làm cho các công nghệ tiên tiến trở nên dễ dàng tiếp cận

với đông đảo công chúng.

Các mô hình ngôn ngữ lớn LLM: cách hoạt động của chúng

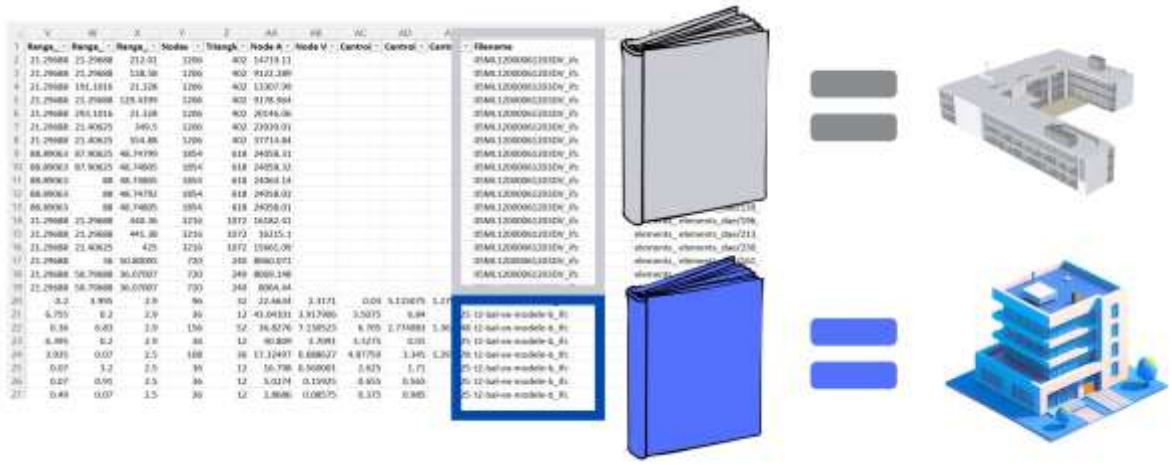
Các mô hình ngôn ngữ lớn (ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok) là các mạng nơ-ron được huấn luyện trên một khối lượng dữ liệu văn bản khổng lồ từ Internet, sách, bài báo và các nguồn khác. Nhiệm vụ chính của chúng là hiểu bối cảnh của ngôn ngữ con người và sinh ra các phản hồi có ý nghĩa.

Trong cơ sở của các LLM hiện đại là kiến trúc transformer, được đề xuất bởi các nhà nghiên cứu Google vào năm 2017. Thành phần chính của kiến trúc này là cơ chế attention, cho phép mô hình xem xét các mối liên hệ giữa các từ mà không phụ thuộc vào vị trí của chúng trong văn bản.

> Quá trình huấn luyện LLM giống như cách mà con người học ngôn ngữ — chỉ là với quy mô lớn gấp triệu lần. Mô hình phân tích hàng tỷ ví dụ về cách sử dụng từ và câu, tìm ra các quy luật trong cấu trúc ngôn ngữ và trong logic của các chuyển tiếp ý nghĩa. Trong quá trình này, toàn bộ văn bản được chia thành các token — đơn vị cảm nghĩa tối thiểu (các từ hoặc phần của chúng), sau đó được chuyển đổi thành các vector trong không gian nhiều chiều. Những biểu diễn vector này cho phép máy "hiểu" các mối liên hệ tiềm ẩn giữa các khái niệm, thay vì chỉ đơn thuần xử lý văn bản như một chuỗi ký tự.-

Các mô hình ngôn ngữ lớn không chỉ đơn thuần là công cụ để sinh văn bản. Chúng có khả năng nhận diện ý nghĩa, tìm kiếm mối liên hệ giữa các khái niệm và làm việc với dữ liệu, ngay cả khi chúng được trình bày dưới nhiều định dạng khác nhau. Điều quan trọng là thông tin phải được chia thành các mô hình dễ hiểu và được đại diện dưới dạng token mà LLM có thể làm việc.

Cách tiếp cận tương tự có thể áp dụng cho các dự án xây dựng. Nếu tưởng tượng dự án như một văn bản độc đáo, nơi mỗi tòa nhà, yếu tố hoặc cấu trúc là một mã token, chúng ta có thể bắt đầu xử lý thông tin này theo cách tương tự. Các dự án xây dựng có thể được so sánh với những cuốn sách được chia thành các thể loại, chương và nhóm đoạn, với các mã token tối thiểu - các yếu tố của dự án xây dựng. Bằng cách chuyển đổi các mô hình dữ liệu sang định dạng có cấu trúc, chúng ta cũng có thể chuyển dữ liệu có cấu trúc thành các cơ sở dữ liệu véc tơ, đây là nguồn lý tưởng cho học máy và các công nghệ như LLM.-



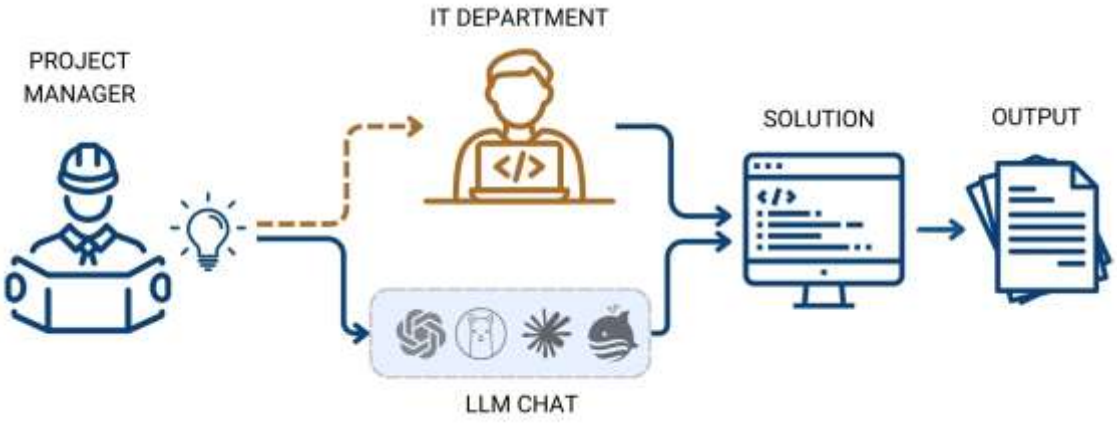
Yếu tố của dự án xây dựng giống như mã token trong văn bản: đơn vị tối thiểu từ đó hình thành các nhóm (đoạn) và các phần (thể loại) của toàn bộ dự án.

Nếu dự án xây dựng được số hóa và các yếu tố của nó được đại diện dưới dạng mã token hoặc véc tơ, thì có thể truy cập chúng không thông qua các truy vấn cứng nhắc, mà bằng ngôn ngữ tự nhiên. Đây chính là một trong những lợi thế chính của LLM - khả năng hiểu ý nghĩa của truy vấn và liên kết nó với các dữ liệu tương ứng.

Kỹ sư không còn cần phải viết truy vấn SQL hoặc mã Python để lấy dữ liệu cần thiết - họ chỉ cần, hiểu được cách hoạt động của LLM và cấu trúc dữ liệu, phác thảo nhiệm vụ theo cách quen thuộc: "Tìm tất cả các cấu trúc bê tông cốt thép có lớp bê tông lớn hơn B30 và tính tổng thể tích của chúng." Mô hình sẽ nhận ra ý nghĩa của truy vấn, chuyển đổi nó thành định dạng có thể đọc bởi máy móc, tìm dữ liệu (nhóm và chuyển đổi) và trả về kết quả cuối cùng.

Tài liệu, bảng, mô hình dự án được chuyển đổi thành các đại diện véc tơ (embedding) và lưu trữ trong cơ sở dữ liệu. Khi người dùng đặt câu hỏi, truy vấn cũng được chuyển đổi thành véc tơ, và hệ thống tìm kiếm các dữ liệu gần nhất về mặt ý nghĩa. Điều này cho phép LLM dựa vào không chỉ kiến thức đã được đào tạo của nó, mà còn vào dữ liệu doanh nghiệp hiện tại, thậm chí nếu chúng đã xuất hiện sau khi kết thúc quá trình đào tạo mô hình.

> Một trong những lợi thế quan trọng nhất của LLM trong xây dựng là khả năng sinh mã lập trình. Thay vì truyền giao nhiệm vụ kỹ thuật cho lập trình viên, các chuyên gia có thể mô tả nhiệm vụ bằng ngôn ngữ tự nhiên, và mô hình sẽ tạo ra mã cần thiết, mà có thể được sử dụng (sao chép từ cuộc trò chuyện) trong việc tạo ra mã tự động hóa quy trình. Các mô hình LLM cho phép các chuyên gia không có kiến thức sâu về lập trình đóng góp vào việc tự động hóa và cải tiến quy trình kinh doanh của công ty.



Các mô hình LLM cung cấp khả năng cho người dùng viết mã và nhận kết quả mà không cần phải có kỹ năng lập trình.

> Theo một nghiên cứu được thực hiện bởi Wakefield Research và tài trợ bởi SAP vào năm 2024, trong đó có sự tham gia của 300 giám đốc điều hành cấp cao tại các công ty có doanh thu hàng năm từ ít nhất 1 tỷ đô la Mỹ: 52% các giám đốc điều hành cấp cao tin tưởng vào AI trong các vấn đề phân tích dữ liệu và cung cấp khuyến nghị cho quyết định. Còn 48% sử dụng AI để xác định các rủi ro chưa được ghi nhận trước đó, và 47% - để đề xuất các kế hoạch thay thế. Hơn nữa, 40% áp dụng AI trong phát triển sản phẩm mới, lập kế hoạch ngân sách và thực hiện nghiên cứu thị trường. Nghiên cứu cũng cho thấy tác động tích cực của AI đối với đời sống cá nhân: 39% người tham gia cho rằng cải thiện sự cân bằng giữa công việc và cuộc sống cá nhân, 38% báo cáo về cải thiện sức khỏe tâm lý, trong khi 31% cảm thấy giảm căng thẳng.

Tuy nhiên, với tất cả sức mạnh của mình, LLM vẫn chỉ là công cụ mà chúng ta cần sử dụng một cách có ý thức. Giống như bất kỳ công nghệ nào, chúng đều có những hạn chế. Một trong những vấn đề nổi tiếng nhất là hiện tượng "ảo giác" - những trường hợp mà mô hình tự tin đưa ra câu trả lời có vẻ hợp lý nhưng thực tế lại sai. Do đó, điều quan trọng là phải hiểu cách thức hoạt động của mô hình: dữ liệu và mô hình dữ liệu nào mà nó có thể diễn giải mà không gặp lỗi, cách nó diễn giải các truy vấn và thông tin mà nó lấy từ đâu. Cũng cần nhớ rằng kiến thức của LLM bị giới hạn bởi ngày tháng của quá trình đào tạo, và không có kết nối với dữ liệu bên ngoài, mô hình có thể không xem xét các quy định, tiêu chuẩn, giá cả hoặc công nghệ hiện hành.

Giải pháp cho những vấn đề này là cập nhật thường xuyên các cơ sở dữ liệu vector, kết nối với các nguồn dữ liệu hiện hành và phát triển các AI-agent tự trị, không chỉ trả lời câu hỏi mà còn chủ động sử dụng dữ liệu để học hỏi, quản lý nhiệm vụ, phát hiện rủi ro, đề xuất các phương án tối ưu hóa và kiểm soát tiến độ dự án.

> Việc chuyển sang giao diện LLM trong xây dựng không chỉ là một công nghệ mới. Đây là sự chuyển mình trong tư duy, xóa bỏ rào cản giữa con người và dữ liệu. Đây là cơ hội làm việc với thông tin dễ dàng như chúng ta trò chuyện với nhau - và vẫn nhận được những kết quả chính xác, được kiểm chứng và sẵn sàng hành động.

Những công ty bắt đầu sử dụng những công cụ như vậy sớm hơn những công ty khác sẽ có được lợi thế cạnh tranh đáng kể. Đây là việc tăng tốc công việc, giảm chi phí, và nâng cao chất lượng quyết định dự án nhờ vào việc truy cập nhanh vào phân tích dữ liệu và khả năng tìm kiếm đáp án cho những câu hỏi phức tạp. Nhưng cũng cần xem xét các vấn đề về an ninh. Việc sử dụng dịch vụ LLM đám mây có thể liên quan đến rủi ro rò rỉ dữ liệu. Do đó, ngày càng nhiều tổ chức tìm kiếm các giải pháp thay thế cho phép triển khai các

công cụ LLM trong cơ sở hạ tầng của riêng họ - cục bộ, với sự bảo vệ và kiểm soát hoàn toàn về thông tin.

Sử dụng LLM cục bộ cho dữ liệu nhạy cảm của công ty

Sự xuất hiện của các chatbot LLM đầu tiên vào năm 2022 đánh dấu một giai đoạn mới trong sự phát triển của trí tuệ nhân tạo. Tuy nhiên, ngay sau khi những mô hình này trở nên phổ biến, một câu hỏi tự nhiên đã xuất hiện: mức độ an toàn khi truyền tải dữ liệu và yêu cầu liên quan đến công ty lên đám mây là bao nhiêu? Hầu hết các mô hình ngôn ngữ đám mây đều lưu trữ lịch sử giao tiếp và các tài liệu đã tải lên trên máy chủ của họ, và với những công ty làm việc với thông tin nhạy cảm, đây đã trở thành một rào cản lớn trong việc áp dụng AI.

> Một trong những giải pháp bền vững và hợp lý nhất cho vấn đề này là triển khai Open Source LLM nội bộ, trong cơ sở hạ tầng CNTT của doanh nghiệp. Khác với các dịch vụ đám mây, các mô hình cục bộ hoạt động mà không cần kết nối Internet, không truyền dữ liệu lên các máy chủ bên ngoài và cho phép các công ty kiểm soát hoàn toàn thông tin.

> *Mô hình mã nguồn mở tốt nhất hiện nay về hiệu suất tương đương với các mô hình đóng [như ChatGPT, Claude], nhưng còn tụt lại khoảng một năm.*
>
> *- Ben Cottier, nhà nghiên cứu hàng đầu của tổ chức nghiên cứu phi lợi nhuận Epoch AI, 2024.*

Các công ty công nghệ lớn đã bắt đầu cung cấp các mô hình ngôn ngữ lớn (LLM) của họ cho việc sử dụng nội bộ. Dòng mô hình LLaMA mở từ Meta và dự án DeepSeek đang phát triển nhanh chóng từ Trung Quốc đã trở thành ví dụ cho sự chuyển đổi sang kiến trúc mở. Bên cạnh đó, Mistral và Falcon cũng đã phát hành các mô hình mạnh mẽ, không bị ràng buộc bởi các nền tảng sở hữu. Những sáng kiến này không chỉ tăng tốc sự phát triển của trí tuệ nhân tạo toàn cầu mà còn mang đến cho các công ty, mà sự riêng tư là vấn đề quan trọng, những lựa chọn thực sự về độc lập, linh hoạt và sự tuân thủ các tiêu chuẩn an ninh.

Trong môi trường doanh nghiệp, đặc biệt là trong ngành xây dựng, bảo vệ dữ liệu không chỉ là vấn đề tiện lợi mà còn là tuân thủ quy định. Làm việc với tài liệu thầu, dự toán, bản vẽ và thư từ bí mật yêu cầu phải kiểm soát chặt chẽ. Tại đây, các LLM nội bộ cung cấp sự chắc chắn cần thiết rằng dữ liệu sẽ ở lại trong bức tường bao quanh công ty.

	Cloud LLMs (OpenAI, Claude)	Local LLMs (DeepSeek, LLaMA)
Data Control	Data is transmitted to third parties	Data remains within the company's network
License	Proprietary, paid	Open-source (Apache 2.0, MIT)
Infrastructure	Requires internet	Operates in an isolated environment
Customization	Limited	Full adaptation to company needs
Cost	Pay-per-token/request	One-time hardware investment + maintenance costs
Scalability	Easily scalable with cloud resources	Scaling requires additional local hardware
Security & Compliance	Risk of data leaks, may not meet strict regulations (GDPR, HIPAA)	Full compliance with internal security policies
Performance & Latency	Faster inference due to cloud infrastructure	Dependent on local hardware, may have higher latency
Integration	API-based integration, requires internet access	Can be tightly integrated with on-premise systems
Updates & Maintenance	Automatically updated by provider	Requires manual updates and model retraining
Energy Consumption	Energy cost is covered by provider	High power consumption for inference and training
Offline Availability	Not available without an internet connection	Works completely offline
Inference Cost	Pay-per-use model (cost scales with usage)	Fixed cost after initial investment

Rò 3.33 Các mô hình nội bộ đảm bảo kiểm soát và an ninh hoàn toàn, trong khi các giải pháp đám mây cung cấp tích hợp thuận tiện và các bản cập nhật tự động.

Những lợi ích chính của LLM mã nguồn mở nội bộ:

- Kiểm soát hoàn toàn đối với dữ liệu. Tất cả thông tin sẽ ở lại bên trong công ty, giúp loại trừ việc truy cập trái phép và rò rỉ dữ liệu.
- Hoạt động độc lập. Loại bỏ sự phụ thuộc vào kết nối internet, điều này đặc biệt quan trọng cho công việc trong các hạ tầng CNTT bị cô lập. Điều này cũng đảm bảo hoạt động liên tục trong bối cảnh các lệnh trừng phạt hoặc việc chặn các dịch vụ đám mây.
- Linh hoạt trong ứng dụng. Mô hình có thể được sử dụng để tạo nội dung, phân tích dữ liệu, viết mã, hỗ trợ thiết kế và quản lý quy trình kinh doanh.
- Được điều chỉnh theo các nhiệm vụ của doanh nghiệp. LLM có thể được đào tạo trên các tài liệu nội bộ, cho phép xem xét các đặc thù công việc và các đặc điểm ngành nghề của công ty. LLM nội bộ có thể kết nối với các nền tảng CRM, ERP hoặc BI, giúp tự động hóa phân tích yêu cầu khách

hàng, tạo báo cáo hoặc thậm chí dự đoán xu hướng.

> Việc triển khai mô hình miễn phí và mã nguồn mở DeepSeek-R1-7B trên máy chủ, để cả đội ngũ người dùng có thể truy cập, với chi phí 1.000 USD mỗi tháng, có thể tốn kém ít hơn so với các khoản thanh toán hàng năm cho các API đám mây như ChatGPT hoặc Claude và giúp các công ty hoàn toàn kiểm soát dữ liệu, loại trừ việc truyền dữ liệu ra internet và giúp tuân thủ các yêu cầu quy định như GDPR.

Trong các ngành khác, các LLM nội bộ đã thay đổi cách tiếp cận tự động hóa. Trong các dịch vụ hỗ trợ, chúng trả lời các câu hỏi thường gặp từ khách hàng, làm giảm gánh nặng cho các nhân viên. Trong bộ phận nhân sự, chúng phân tích hồ sơ xin việc và tìm kiếm các ứng viên phù hợp. Trong thương mại điện tử, chúng tạo ra các đề xuất cá nhân hóa mà không tiết lộ dữ liệu người dùng.

Trong lĩnh vực xây dựng, hiệu ứng tương tự được kỳ vọng. Nhờ vào việc tích hợp LLM với dữ liệu thiết kế và các quy định, có thể tăng tốc độ chuẩn bị tài liệu, tự động hóa việc lập dự toán và phân tích chi phí dự đoán. Một hướng đi đầy triển vọng đặc biệt là việc sử dụng LLM kết hợp với các bảng và dataframe có cấu trúc.

Kiểm soát hoàn toàn về AI trong công ty và cách triển khai LLM riêng

Các công cụ hiện đại cho phép các công ty triển khai mô hình ngôn ngữ lớn (LLM) tại chỗ chỉ trong vài giờ. Điều này cung cấp quyền kiểm soát hoàn toàn đối với dữ liệu và cơ sở hạ tầng, loại bỏ sự phụ thuộc vào các dịch vụ đám mây bên ngoài và giảm thiểu rủi ro rò rỉ thông tin. Giải pháp này đặc biệt quan trọng đối với các tổ chức làm việc với tài liệu dự án nhạy cảm hoặc dữ liệu thương mại bí mật.

Tùy thuộc vào nhiệm vụ và tài nguyên, có nhiều kịch bản triển khai khác nhau — từ các giải pháp có sẵn "trong hộp" đến các kiến trúc linh hoạt và có thể mở rộng hơn. Một trong những công cụ đơn giản nhất là Ollama, cho phép bạn chạy các mô hình ngôn ngữ chỉ với một cú nhấp chuột, mà không cần kiến thức kỹ thuật sâu. Bắt đầu nhanh chóng với Ollama:

1. Tải xuống bản phân phối cho hệ điều hành của bạn (Windows / Linux / macOS) từ trang web chính thức: ollama.com
2. Cài đặt mô hình qua dòng lệnh. Ví dụ, để chạy mô hình Mistral:

```
ollama run mistral
```

3. Sau khi khởi động, mô hình đã sẵn sàng để làm việc — bạn có thể gửi các yêu cầu văn bản qua terminal hoặc tích hợp nó vào các công cụ khác. Khởi động mô hình và thực hiện yêu cầu:

```
ollama run mistral “Làm thế nào để tạo một phép tính với tất cả các tài nguyên cho
công việc lắp đặt một vách ngăn thạch cao rộng 100mm?”
```

Đối với những ai thích làm việc trong môi trường trực quan quen thuộc, có LM Studio — ứng dụng miễn phí với giao diện giống như ChatGPT:

- Cài đặt LM Studio bằng cách tải xuống bản phân phối từ trang web chính thức - lmstudio.ai
- Thông qua danh mục tích hợp sẵn, chọn mô hình (ví dụ, Falcon hoặc GPT-Neo-X) và tải nó về
- Làm việc với mô hình qua giao diện trực quan, giống như ChatGPT, nhưng hoàn toàn tại chỗ

	Developer	Parameters	GPU Requirements (GB)	Features	Best For
Mistral 7B	Mistral AI	7	8 (FP16)	Fast, supports multimodal tasks (text + images), fully open-source code	Lightweight tasks, mobile devices, laptops
LLaMA 2	Meta	7–70	16–48 (FP16)	High text generation accuracy, adaptable for technical tasks, CC-BY-SA license	Complex analytical and technical tasks
Baichuan 7B/13B	Baichuan Intelligence	7–13	8–16 (FP16)	Fast and efficient, great for large data processing, fully open-source code	Data processing, automating routine tasks
Falcon 7B/40B	Technology Innovation Institute (TII)	7–40	8–32 (FP16)	Open-source, high performance, optimized for fast work	Workloads with limited computational resources
DeepSeek-V3	DeepSeek	671	1543 (FP16) / 386 (4-bit)	Multilingual, 128K token context window, balanced speed and accuracy	Large enterprises, SaaS platforms, multitasking scenarios
DeepSeek-R1-7B	DeepSeek	7	18 (FP16) / 4.5 (4-bit)	Retains 92% of R1 capabilities in MATH-500, local deployment support	Budget solutions, IoT devices, edge computing

Hình. 3.34 So sánh các mô hình LLM mã nguồn mở phổ biến tại chỗ.

Việc chọn mô hình phụ thuộc vào yêu cầu về tốc độ, độ chính xác và khả năng phần cứng có sẵn (Hình. 3.34). Các mô hình nhỏ, như Mistral 7B và Baichuan 7B, phù hợp với các nhiệm vụ nhẹ và thiết bị di động, trong khi các mô hình mạnh mẽ, như DeepSeek-V3, yêu cầu tài nguyên tính toán đáng kể nhưng cung cấp hiệu suất cao và hỗ trợ nhiều ngôn ngữ. Trong những năm tới, thị trường LLM sẽ phát triển nhanh chóng — chúng ta sẽ thấy ngày càng nhiều mô hình nhẹ và chuyên biệt hơn. Thay vì các LLM đa năng bao trùm toàn bộ nội dung nhân loại, sẽ xuất hiện những mô hình được đào tạo dựa trên chuyên môn hẹp. Ví dụ, có thể mong đợi sự xuất hiện của các mô hình chỉ dành riêng cho việc xử lý các phép tính kỹ thuật, báo giá xây dựng hoặc dữ liệu ở định dạng CAD. Những mô hình chuyên biệt này sẽ nhanh hơn, chính xác hơn và an toàn hơn khi sử dụng — đặc biệt trong môi trường chuyên nghiệp, nơi độ tin cậy cao và độ sâu kiến thức là cực kỳ quan trọng.

Sau khi khởi động LLM tại chỗ, nó có thể được điều chỉnh cho các nhiệm vụ cụ thể của công ty. Để làm điều đó, kỹ thuật tinh chỉnh (fine-tuning) được áp dụng, trong đó mô hình trải qua quá trình huấn luyện bổ sung trên các tài liệu nội bộ, hướng dẫn kỹ thuật, mẫu hợp đồng hoặc tài liệu dự án.

RAG: Trợ lý LLM thông minh với quyền truy cập vào dữ liệu doanh nghiệp

Bước tiến tiếp theo trong tiến hóa ứng dụng LLM trong kinh doanh là việc tích hợp các mô hình với dữ liệu doanh nghiệp hiện tại theo thời gian thực. Cách tiếp cận này được gọi là RAG (Retrieval-Augmented Generation) - tạo ra với hỗ trợ truy xuất. Trong kiến trúc này, mô hình ngôn ngữ không chỉ đơn thuần là một giao diện trò chuyện, mà còn là một trợ lý thông minh toàn diện, có khả năng định hướng trong tài liệu, bản vẽ, cơ sở dữ liệu và cung cấp câu trả lời chính xác, theo ngữ cảnh.

> Lợi thế chính của RAG là khả năng sử dụng dữ liệu nội bộ của công ty mà không cần phải đào tạo lại mô hình, đồng thời vẫn giữ được độ chính xác cao và tính linh hoạt trong việc làm việc với thông tin.

Công nghệ RAG kết hợp hai thành phần chính:

- Truy xuất thông tin (Retrieval): mô hình kết nối với các kho dữ liệu - tài liệu, bảng tính, tệp PDF, bản vẽ - và trích xuất thông tin liên quan theo yêu cầu của người dùng.
- Sinh câu trả lời (Augmented Generation): dựa trên dữ liệu đã trích xuất, mô hình tạo ra một câu trả lời chính xác, có cơ sở, xem xét ngữ cảnh và đặc thù của yêu cầu.

Để khởi động LLM với hỗ trợ RAG, cần thực hiện một vài bước:

- Chuẩn bị dữ liệu: thu thập các tài liệu cần thiết, bản vẽ, thông số kỹ thuật, bảng tính. Chúng có thể ở nhiều định dạng và cấu trúc khác nhau, từ PDF đến Excel.
- Lập chỉ mục và vector hóa: thông qua các công cụ như LlamaIndex hoặc LangChain, dữ liệu được chuyển đổi thành các đại diện vector, cho phép tìm thấy các mối liên hệ ý nghĩa giữa các đoạn văn bản (về cơ sở dữ liệu vector và việc chuyển đổi các khối lượng lớn thành đại diện vector, bao gồm cả dự án CAD, sẽ được bàn luận chi tiết trong phần 8).
- Yêu cầu trợ lý: sau khi tải dữ liệu, có thể đặt câu hỏi cho mô hình và nó sẽ tìm kiếm câu trả lời trong cơ sở dữ liệu doanh nghiệp, chứ không phải trong kiến thức tổng quát được thu thập từ internet.

Giả sử trong công ty có một thư mục constructionsite_docs, nơi lưu trữ hợp đồng, hướng dẫn, bảng dự toán và bảng tính. Bằng cách sử dụng một đoạn mã Python (Hình. 3.35), có thể quét thư mục này và xây dựng chỉ mục vector: mỗi tài liệu sẽ được chuyển đổi thành một tập hợp các vector phản ánh nội dung ý nghĩa của văn bản. Điều này biến các tài liệu thành một "bản đồ ý nghĩa", mà qua đó mô hình có thể định hướng một cách hiệu quả và tìm kiếm các mối liên kết giữa các thuật ngữ và cụm từ.-

Ví dụ, mô hình "nhớ" rằng từ "trả lại" và "khiếu nại" thường xuất hiện trong phần của hợp đồng liên quan đến việc giao hàng vật liệu đến công trường xây dựng. Sau đó, nếu đặt câu hỏi - ví dụ, "Thời gian trả hàng của chúng ta là gì?" (Hình. 3.35 - 11 dòng mã) - LLM sẽ phân tích các tài liệu nội bộ và tìm thấy thông tin chính xác, hành động như một trợ lý thông minh, có khả năng đọc và hiểu nội dung của tất cả các tập tin

doanh nghiệp.

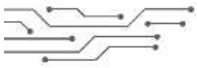


Hình. 3.35 LLM đọc thư mục các tập tin - tương tự như cách mà con người mở nó ra và tìm kiếm tài liệu cần thiết.

Đoạn mã có thể được chạy trên bất kỳ máy tính nào có cài đặt Python. Thêm chi tiết về việc sử dụng Python và IDE để chạy mã sẽ được bàn luận trong chương tiếp theo.

Triển khai LLM cục bộ không chỉ là một xu hướng, mà là một quyết định chiến lược cho các công ty coi trọng sự an toàn và linh hoạt. Tuy nhiên, triển khai LLM, cho dù trên máy tính cục bộ của công ty hay sử dụng các giải pháp trực tuyến, chỉ là bước đầu tiên. Để ứng dụng khả năng của LLM vào các nhiệm vụ thực tế, các công ty cần sử dụng những công cụ không chỉ cung cấp câu trả lời trong các cuộc trò chuyện mà còn lưu giữ logic đã tạo dưới dạng mã có thể thực thi ngoài bối cảnh sử dụng LLM. Điều này rất quan trọng cho việc mở rộng các giải pháp - các quy trình được tổ chức đúng cách cho phép áp dụng những phát kiến của AI ngay lập tức vào nhiều dự án hoặc thậm chí trên toàn bộ công ty.

Trong bối cảnh này, việc lựa chọn môi trường phát triển phù hợp (IDE) đóng một vai trò quan trọng. Các công cụ lập trình hiện đại không chỉ cho phép phát triển các giải pháp dựa trên LLM mà còn tích hợp chúng vào các quy trình kinh doanh hiện có, biến chúng thành các ETL Pipeline tự động.



CHƯƠNG 3.4. **IDE hỗ trợ LLM và những thay đổi trong lập trình trong tương lai**

Lựa chọn IDE: từ các thử nghiệm LLM đến các giải pháp kinh doanh

Khi đắm chìm trong thế giới tự động hóa, phân tích dữ liệu và trí tuệ nhân tạo - đặc biệt là khi làm việc với các mô hình ngôn ngữ lớn (LLM) - việc lựa chọn một môi trường phát triển tích hợp (IDE) phù hợp cực kỳ quan trọng. Chính nó sẽ trở thành công cụ làm việc chính của bạn: nơi mà mã được LLM tạo ra sẽ được chạy, cả trên máy tính cục bộ lẫn trong mạng lưới doanh nghiệp. Việc lựa chọn IDE không chỉ ảnh hưởng đến sự thoải mái khi làm việc, mà còn quyết định tốc độ bạn có thể chuyển từ các yêu cầu thử nghiệm trong LLM sang các giải pháp đầy đủ, được tích hợp vào quy trình kinh doanh thực tế.

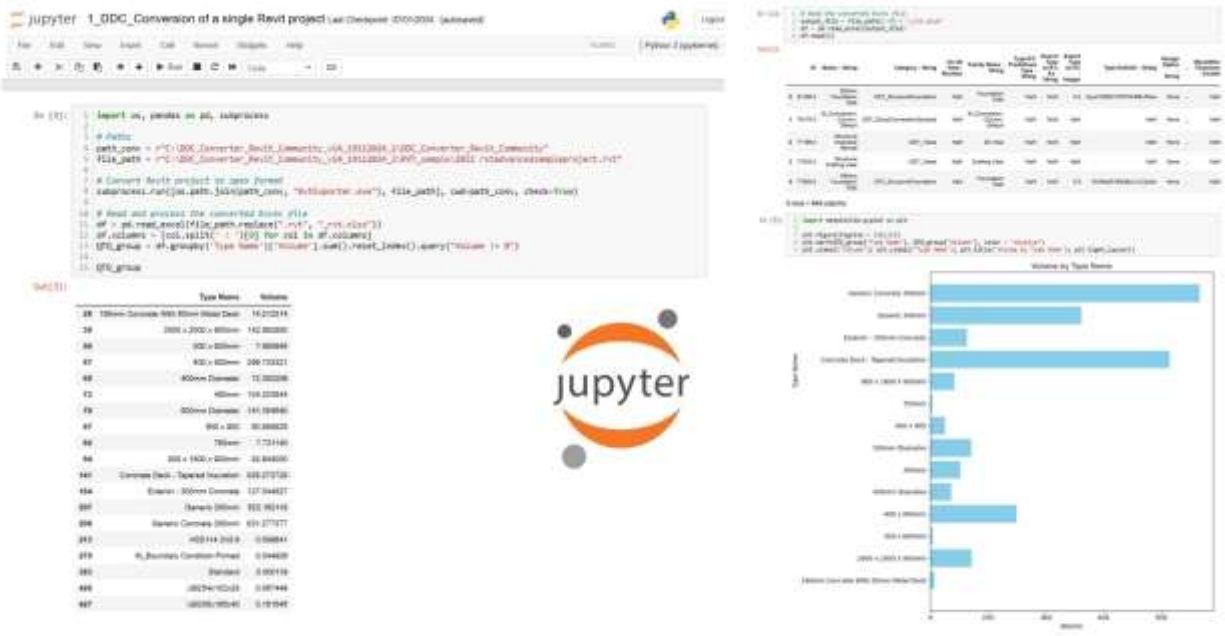
> IDE (Môi Trường Phát Triển Tích Hợp) là một công cụ đa chức năng trên máy tính của bạn để tự động hóa quy trình và xử lý dữ liệu. Thay vì phải lưu trữ riêng lẻ cưa, búa, khoan và các công cụ khác, bạn có một thiết bị duy nhất có thể làm mọi thứ - cắt, ghép, khoan và thậm chí kiểm tra chất lượng vật liệu. IDE cho lập trình viên là một không gian duy nhất, nơi bạn có thể viết mã (trong sự tương tự của xây dựng - tạo ra các bản vẽ), thử nghiệm công việc của nó (xây dựng mô hình tòa nhà), tìm kiếm lỗi (như kiểm soát độ bền của cấu trúc trong xây dựng) và chạy dự án hoàn chỉnh (bàn giao nhà đi vào sử dụng).

Tổng quan về các IDE phổ biến:

- PyCharm® (JetBrains) là một IDE chuyên nghiệp mạnh mẽ cho Python. Nó rất phù hợp cho các dự án nghiêm túc nhờ vào số lượng lớn các chức năng tích hợp. Tuy nhiên, hỗ trợ cơ bản cho các tệp Jupyter tương tác (IPYNB) chỉ có trong phiên bản trả phí, và giao diện có thể gây rối cho những người mới bắt đầu.

> Tệp có phần mở rộng IPYNB (Interactive Python Notebook) là một định dạng của các sổ tay tương tác Jupyter® Notebook (Hình 3.41), nơi mã, hình ảnh và giải thích được kết hợp trong cùng một tài liệu. Định dạng này lý tưởng cho việc xây dựng báo cáo, phân tích và kịch bản huấn luyện.-

- VS Code® (Microsoft) là một công cụ nhanh chóng, linh hoạt và có thể tùy chỉnh với miễn phí hỗ trợ IPYNB và nhiều plugin. Nó phù hợp cho cả người mới và chuyên gia. Cho phép tích hợp GitHub Copilot và các plugin làm việc với các mô hình ngôn ngữ, điều này khiến nó trở thành lựa chọn xuất sắc cho các dự án trong lĩnh vực AI và khoa học dữ liệu.
- Jupyter Notebook — Lựa chọn cổ điển và phổ biến cho việc thử nghiệm và học tập. Cho phép viết mã, thêm chú thích, hình dung các kết quả trong cùng một giao diện (Hình 3.41). Hoàn hảo cho việc kiểm tra giả thuyết nhanh, làm việc với LLM và tạo ra các bước phân tích dữ liệu có thể tái tạo. Để quản lý các phụ thuộc và thư viện, nên sử dụng Anaconda Navigator — giao diện trực quan để quản lý môi trường Python.



Hình 3.41 Jupyter Notebook là một trong những công cụ tiện lợi và phổ biến nhất để tạo ra quy trình Pipeline.

- Google Collab™ (và nền tảng Kaggle (Hình 9.25)) — một lựa chọn trên đám mây thay thế Jupyter, cung cấp quyền truy cập miễn phí vào GPU/TPU. Đây là giải pháp tuyệt vời để bắt đầu — không cần cài đặt phần mềm cục bộ và có thể làm việc ngay từ trình duyệt. Hỗ trợ tích hợp với Google Drive và gần đây là với Gemini (LLM của Google).-

	PyCharm	VS Code	Jupyter Notebook	Google Colab
Complexity	High	Medium	Low	Low
.ipynb support	Paid	Free	Built-in	Built-in
Copilots	Yes	Yes	Yes	Yes
Computing resources	Local	Local	Local	Cloud
For whom	Professionals	Universal	Beginners	Experimenters

Hình 3.42 So sánh các IDE: Jupyter Notebook là một trong những công cụ tiện lợi và đơn giản nhất để tạo ra quy trình Pipeline.

Lựa chọn IDE phụ thuộc vào nhiệm vụ của bạn. Nếu bạn muốn bắt đầu nhanh với AI, hãy thử Jupyter Notebook hoặc Google Collab. Đối với các dự án nghiêm túc, tốt hơn là sử dụng PyCharm hoặc VS Code. Điều quan trọng là bắt đầu. Các công cụ hiện đại cho phép nhanh chóng biến các thử nghiệm thành các giải pháp khả thi.

Tất cả các IDE đã mô tả cho phép tạo ra các quy trình xử lý dữ liệu — tức là chuỗi các mô-đun của khối mã (có thể được sinh ra bởi LLM), mỗi khối chịu trách nhiệm cho từng bước, chẳng hạn như:

- kịch bản phân tích,
- chuỗi trích xuất thông tin từ tài liệu,
- phản hồi tự động dựa trên RAG,
- tạo báo cáo và hình dung.

Nhờ vào cấu trúc mô-đun, mỗi bước có thể được trình bày như một khối riêng biệt: tải dữ liệu → lọc → phân tích → hình dung → xuất kết quả. Những khối này có thể được tái sử dụng, - thích nghi và lắp ghép thành các chuỗi mới, như một bộ xây dựng, chỉ dành cho dữ liệu.

Đối với các kỹ sư, lãnh đạo và nhà phân tích, điều này mở ra khả năng tài liệu hóa logic ra quyết định dưới dạng mã, có thể được sinh ra với sự trợ giúp của LLM. Cách tiếp cận này giúp tăng tốc các tác vụ tẻ nhạt, tự động hóa các hoạt động điển hình và hình thành các quy trình có thể tái tạo, trong đó mỗi bước được ghi chép rõ ràng và minh bạch cho tất cả các thành viên trong nhóm.

Thông tin chi tiết về các Pipeline ETL tự động (Hình 7.23), các công cụ Apache Airflow (Hình 7.44), Apache NiFi (Hình 7.45) và n8n (Hình 7.46) để xây dựng các khối logic khi tự động hóa quy trình sẽ được thảo luận trong phần 7 và 8 của cuốn sách. ----

IDE hỗ trợ LLM và các thay đổi trong lập trình trong tương lai

Sự tích hợp của trí tuệ nhân tạo vào quy trình phát triển đang thay đổi bối cảnh lập trình. Các môi trường hiện đại không chỉ là những trình chỉnh sửa văn bản với đánh dấu cú pháp — mà còn trở thành những trợ lý thông minh, có khả năng hiểu logic của dự án, hoàn thiện mã và thậm chí giải thích cách thức hoạt động của một đoạn mã nào đó. Trên thị trường, các sản phẩm ngày càng xuất hiện, bằng cách sử dụng AI mở rộng ranh giới của việc phát triển truyền thống:

- GitHub Copilot (tích hợp vào VS Code, PyCharm): Trợ lý AI, tạo ra mã dựa trên các chú thích hoặc mô tả một phần, biến các gợi ý văn bản thành các giải pháp hoàn chỉnh.
- Cursor (nhánh của VS Code với lõi AI): cho phép không chỉ hoàn thiện mã mà còn đặt câu hỏi cho dự án, tìm kiếm sự phụ thuộc và học hỏi từ cơ sở mã.
- JetBrains AI Assistant: plugin cho IDE JetBrains (bao gồm PyCharm) với chức năng giải thích mã phức tạp, tối ưu hóa và tạo thử nghiệm.
- Amazon CodeWhisperer: tương tự như Copilot với trọng tâm vào an ninh và hỗ trợ các dịch vụ AWS từ Amazon.

Lập trình trong những năm tới sẽ trải qua những thay đổi căn bản. Trọng tâm chính sẽ chuyển từ việc viết mã rập khuôn sang thiết kế mô hình và kiến trúc dữ liệu - các nhà phát triển sẽ tập trung nhiều hơn vào thiết kế hệ thống, trong khi AI sẽ đảm nhận những nhiệm vụ định hình: tạo mã, thử nghiệm, tài liệu và các chức năng cơ bản. Tương lai của lập trình là sự hợp tác giữa con người và AI, trong đó máy móc đảm nhận sự tẻ nhạt về mặt kỹ thuật, còn con người tập trung vào sự sáng tạo.

Lập trình bằng ngôn ngữ tự nhiên sẽ trở thành điều thường nhật. Cá nhân hóa IDE sẽ lên một tầm cao mới - các môi trường phát triển sẽ biết cách thích ứng theo phong cách làm việc của người dùng, và các công ty sẽ dự đoán các mẫu, đề xuất giải pháp theo ngữ cảnh và học hỏi từ các dự án trước đó.

> Điều này không hủy bỏ vai trò của nhà phát triển, nhưng biến đổi nó một cách mạnh mẽ: từ việc viết mã sang quản lý tri thức, chất lượng và quy trình. Sự tiến hóa này cũng sẽ ảnh hưởng đến lĩnh vực phân tích kinh doanh, nơi việc tạo báo cáo, trực quan hóa và ứng dụng hỗ trợ quyết định ngày càng diễn ra thông qua việc tạo mã và logic bằng AI và LLM, chat và các tác nhân giao diện.

Sau khi công ty đã thiết lập các cuộc trò chuyện LLM và chọn môi trường phát triển phù hợp, bước quan trọng tiếp theo là tổ chức dữ liệu. Quá trình này bao gồm việc trích xuất thông tin từ các nguồn rời rạc, làm sạch, chuyển đổi thành hình thức có cấu trúc và tích hợp vào các hệ thống doanh nghiệp.

Trong phương pháp quản lý dữ liệu Data-Centric hiện đại, mục tiêu chính là đưa dữ liệu về một hình thức thống nhất, linh hoạt có thể tương thích với nhiều công cụ và ứng dụng. Để làm việc với các quy trình cấu trúc và dữ liệu có cấu trúc, cần có các thư viện chuyên dụng. Một trong những thư viện mạnh mẽ, linh hoạt và phổ biến nhất là thư viện Pandas cho Python. Nó cho phép xử lý dữ liệu bảng một cách thuận tiện: lọc, nhóm, làm sạch, bổ sung, thực hiện tổng hợp và xây dựng báo cáo.

Python Pandas: công cụ thiết yếu cho công việc với dữ liệu

Trong thế giới phân tích và tự động hóa dữ liệu, Pandas chiếm một vị trí đặc biệt. Đây là một trong những thư viện phổ biến nhất và được sử dụng rộng rãi trong ngôn ngữ lập trình Python, được thiết kế để làm việc với dữ liệu có cấu trúc.

Thư viện giống như một bộ công cụ hoàn chỉnh: các chức năng, mô-đun, lớp. Như trên công trường xây dựng không cần phát minh lại búa hay thước hàng, thì trong lập trình, các thư viện cho phép giải quyết nhanh chóng các vấn đề mà không cần phát minh lại các chức năng và giải pháp cơ bản.

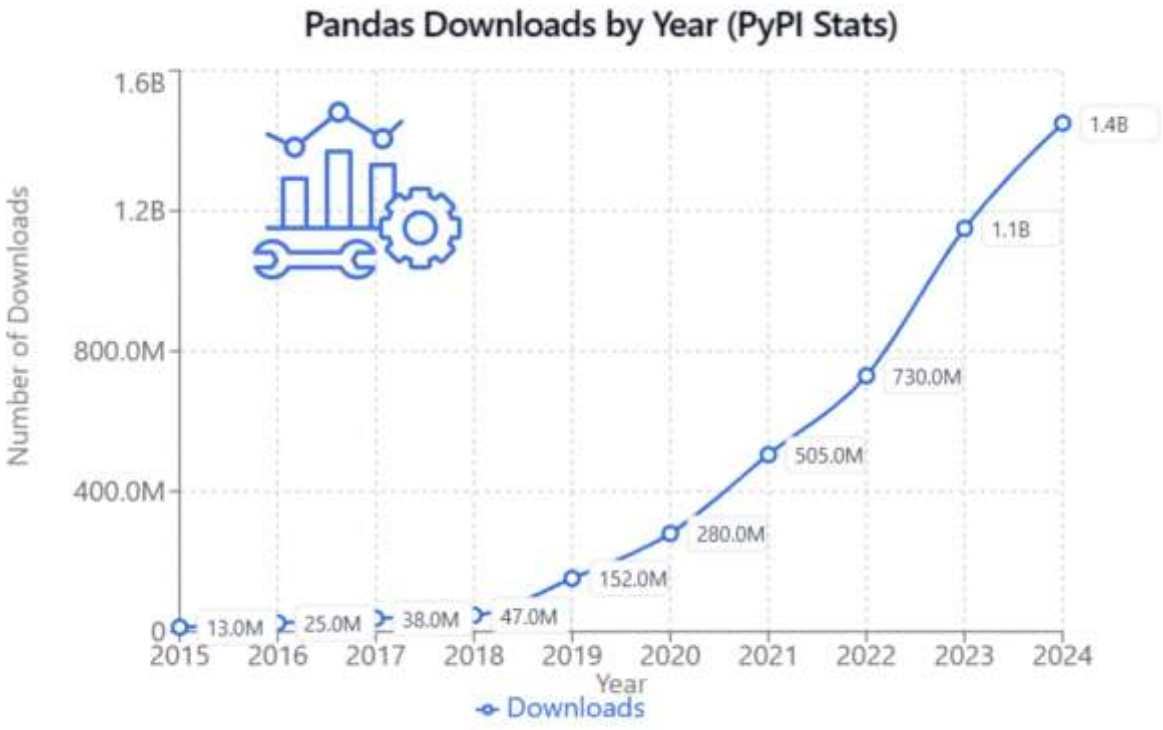
> Pandas là thư viện Python mã nguồn mở, cung cấp các cấu trúc dữ liệu hiệu suất cao và dễ hiểu, đặc biệt là DataFrame - định dạng chính để làm việc với bảng. Pandas là con dao đa năng cho các nhà phân tích, kỹ sư và nhà phát triển làm việc với dữ liệu.

Python — là một ngôn ngữ lập trình bậc cao với cú pháp đơn giản, được áp dụng rộng rãi trong phân tích dữ liệu, tự động hóa, học máy và phát triển web. Sự phổ biến của nó được giải thích bởi tính đọc dễ của mã, khả năng tương thích đa nền tảng và hệ sinh thái thư viện phong phú. Đến nay, đã có hơn 137.000 gói mã nguồn mở được tạo ra cho Python [87], và con số này tiếp tục tăng lên gần như hàng ngày. Mỗi thư viện như vậy là một kho chứa các hàm sẵn có: từ các phép toán toán học cơ bản cho đến các công cụ phức tạp để xử lý hình ảnh, phân tích dữ liệu lớn, làm việc với mạng nơ-ron và tích hợp với các dịch vụ bên ngoài.

> Nói cách khác, hãy tưởng tượng rằng bạn có quyền truy cập miễn phí và mở vào hàng trăm nghìn giải pháp phần mềm sẵn có — các thư viện và công cụ mà bạn có thể tích hợp trực tiếp vào quy trình kinh doanh của mình. Đây giống như một danh mục ứng dụng khổng lồ dành cho tự động hóa, phân tích, trực quan hóa, tích hợp và nhiều thứ khác — và tất cả đều có sẵn ngay sau khi cài đặt Python.

Pandas — là một trong những gói phổ biến nhất trong hệ sinh thái Python. Vào năm 2022, số lượng tải xuống trung bình của thư viện Pandas đạt 4 triệu mỗi ngày (Hình 3.43), trong khi đến đầu năm 2025, con số này

đã tăng lên 12 triệu lượt tải xuống mỗi ngày, phản ánh sự phổ biến ngày càng tăng và ứng dụng rộng rãi của nó trong phân tích dữ liệu và trò chuyện LLM [86]. -

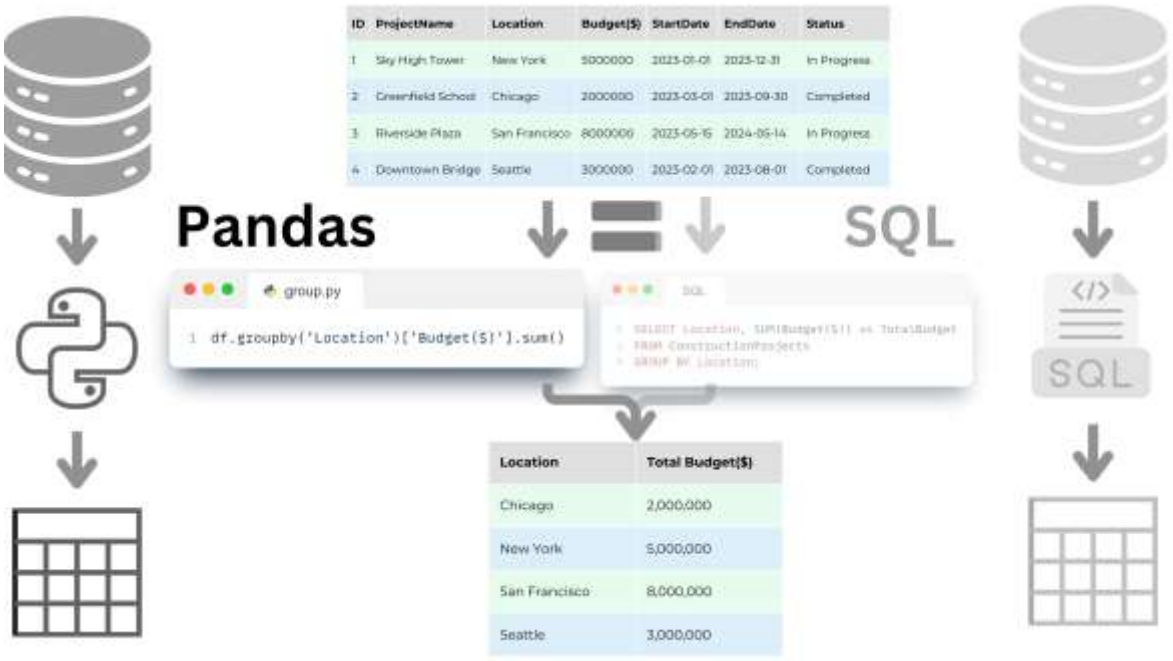


Hình 3.43 Pandas – là một trong những thư viện được tải xuống nhiều nhất. Vào năm 2024, số lượt tải xuống hàng năm của nó đã vượt qua 1,4 tỷ.

Ngôn ngữ truy vấn trong thư viện Pandas về cơ bản tương tự như ngôn ngữ truy vấn SQL mà chúng ta đã thảo luận trong chương "Cơ sở dữ liệu quan hệ và ngôn ngữ truy vấn SQL".

> Trong thế giới phân tích và quản lý dữ liệu có cấu trúc, Pandas nổi bật với tính đơn giản, tốc độ và sức mạnh, cung cấp cho người dùng một loạt công cụ phong phú để phân tích và xử lý thông tin hiệu quả.

Cả hai công cụ — SQL và Pandas — đều cung cấp những khả năng mạnh mẽ để làm việc với dữ liệu, đặc biệt là so với Excel truyền thống. Chúng hỗ trợ các thao tác như lựa chọn, lọc (Hình 3.44), chỉ khác là SQL được tối ưu hóa để làm việc với các cơ sở dữ liệu quan hệ, trong khi Pandas xử lý dữ liệu trong bộ nhớ (RAM), cho phép nó chạy trên bất kỳ máy tính nào mà không cần tạo ra các cơ sở dữ liệu hoặc triển khai hạ tầng riêng biệt.-

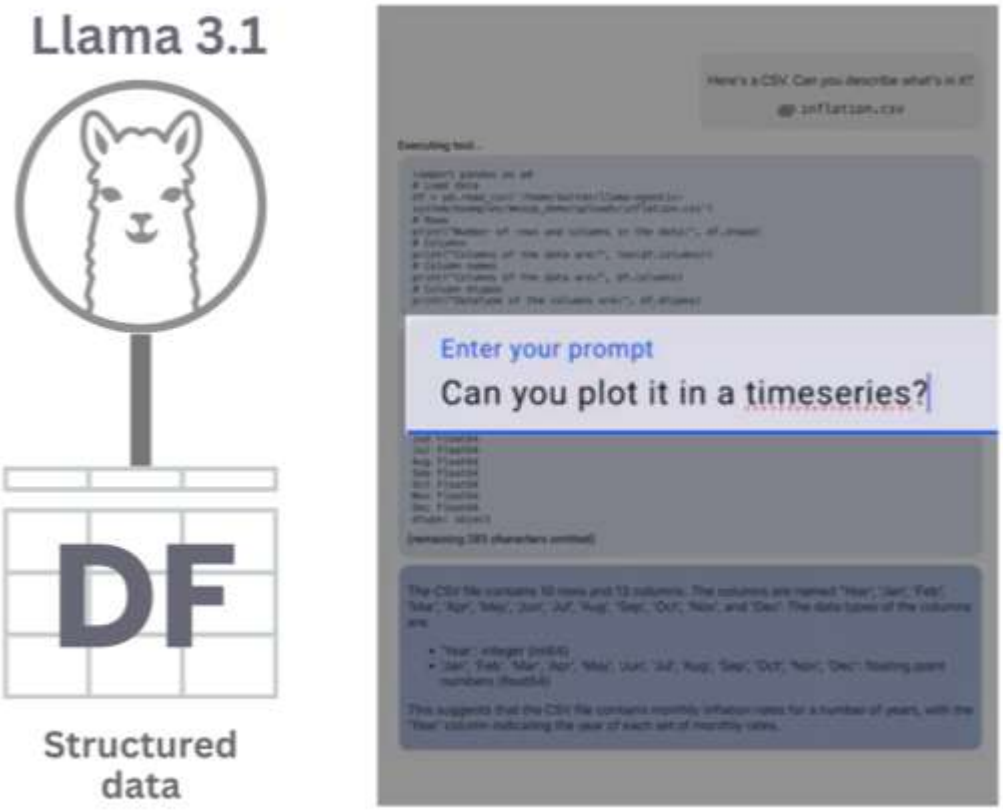


Hình 3.44 Pandas, khác với SQL, có tính linh hoạt khi làm việc với các định dạng dữ liệu khác nhau, không bị giới hạn trong các cơ sở dữ liệu.

Thư viện Pandas của ngôn ngữ lập trình Python thường được sử dụng trong nghiên cứu khoa học, tự động hóa quy trình, tạo đường ống (bao gồm cả ETL) và thao tác dữ liệu trên Python, trong khi SQL là tiêu chuẩn quản lý cơ sở dữ liệu và thường được sử dụng trong môi trường doanh nghiệp để làm việc với những khối lượng dữ liệu lớn.

> Thư viện Pandas của ngôn ngữ lập trình Python cho phép thực hiện không chỉ những thao tác cơ bản, chẳng hạn như đọc và ghi bảng, mà còn thực hiện những nhiệm vụ phức tạp hơn, bao gồm hợp nhất dữ liệu, nhóm dữ liệu và thực hiện các phép toán phân tích phức tạp.

Hôm nay, thư viện Pandas không chỉ được sử dụng trong nghiên cứu học thuật và phân tích kinh doanh, mà còn kết hợp với các mô hình LLM. Ví dụ, bộ phận Meta® (Facebook™) khi phát hành mô hình mã nguồn mở LlaMa 3.1 vào năm 2024 đã đặc biệt chú trọng vào việc làm việc với dữ liệu có cấu trúc, biến xử lý các DataFrame có cấu trúc (Hình 3.45) ở định dạng CSV và tích hợp với thư viện Pandas trực tiếp trong chat trở thành một trong những trường hợp chính trong bản phát hành của họ.-



Hình 3.45 Một trong những trường hợp đầu tiên và quan trọng nhất mà đội ngũ Meta đã trình bày trong LlaMa 3.1 vào năm 2024 là xây dựng các ứng dụng sử dụng Pandas.

> *Pandas là công cụ cần thiết cho hàng triệu nhà khoa học dữ liệu, những người xử lý và chuẩn bị dữ liệu cho trí tuệ nhân tạo sinh tạo. Việc tăng tốc Pandas mà không cần thay đổi mã sẽ là một bước tiến lớn. Các nhà khoa học dữ liệu sẽ có thể xử lý dữ liệu trong vài phút thay vì hàng giờ, và thu thập được nhiều dữ liệu hơn để huấn luyện các mô hình trí tuệ nhân tạo sinh tạo [88].— Jensen Huang, người sáng lập và Giám đốc Điều hành của NVIDIA*

Sử dụng Pandas, người dùng có thể quản lý và phân tích các tập dữ liệu lớn hơn rất nhiều so với khả năng của Excel. Trong khi Excel thường chỉ có thể xử lý tối đa 1 triệu dòng dữ liệu, thì Pandas có thể dễ dàng làm việc với các tập dữ liệu (Hình 9.12, Hình 9.110) chứa hàng chục triệu dòng [89]. Khả năng này cho phép người dùng thực hiện phân tích dữ liệu phức tạp và trực quan hóa trên các tập dữ liệu lớn, cung cấp cái nhìn sâu sắc và hỗ trợ việc đưa ra quyết định dựa trên dữ liệu. Hơn nữa, Pandas nhận được sự hỗ trợ mạnh mẽ từ cộng đồng [90]: hàng trăm triệu nhà phát triển và nhà phân tích trên toàn thế giới (Kaggle.com, Google Collab, Microsoft® Azure™ Notebooks, Amazon SageMaker) sử dụng nó hàng ngày, trực tuyến hoặc ngoại tuyến, cung cấp một lượng lớn giải pháp đã có sẵn cho bất kỳ bài toán kinh doanh nào.--

Cơ sở của hầu hết các quy trình phân tích trên Python là một dạng dữ liệu có cấu trúc gọi là DataFrame, được cung cấp bởi thư viện Pandas. Đây là một công cụ mạnh mẽ và linh hoạt để tổ chức, phân tích và trực

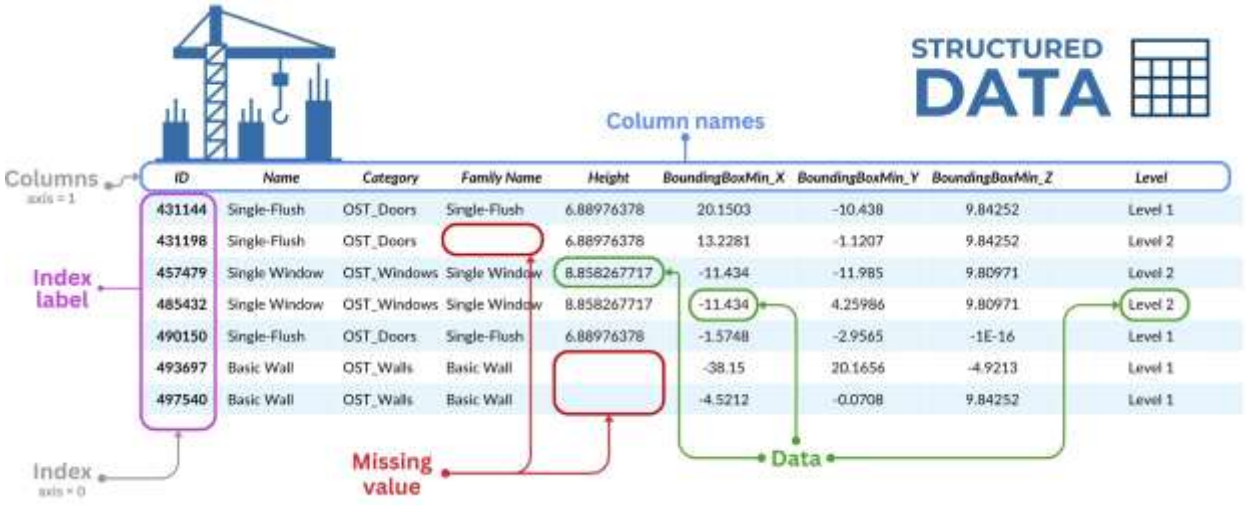
quan hóa dữ liệu dạng bảng.

DataFrame: định dạng dữ liệu bảng phổ quát

DataFrame là cấu trúc trung tâm trong thư viện Pandas, đại diện cho một bảng hai chiều (Hình 3.46), trong đó các hàng tương ứng với các đối tượng hoặc bản ghi riêng lẻ, còn các cột là các thuộc tính, tham số hoặc danh mục của chúng. Cấu trúc này tương tự về mặt hình ảnh với các bảng trong Excel, nhưng vượt xa chúng về tính linh hoạt, khả năng mở rộng và chức năng.

> DataFrame là một cách đại diện và xử lý dữ liệu dạng bảng, được lưu trữ trong bộ nhớ tạm của máy tính.

DataFrame là một cách đại diện và xử lý dữ liệu dạng bảng, được lưu trữ trong bộ nhớ tạm của máy tính. Trong bảng, các hàng có thể phản ánh, chẳng hạn, các yếu tố trong một dự án xây dựng, trong khi các cột phản ánh các thuộc tính của chúng: danh mục, kích thước, tọa độ, chi phí, thời gian, v.v. Hơn nữa, trong bảng này có thể chứa thông tin về một dự án (Hình 4.113), cũng như dữ liệu về hàng triệu đối tượng từ hàng nghìn dự án khác nhau (Hình 9.110). Nhờ vào các phép toán vector hóa của Pandas, người dùng có thể dễ dàng lọc, nhóm và tổng hợp khối lượng thông tin lớn này với tốc độ cao.-



Hình 3.46 Dự án xây dựng dưới dạng DataFrame - là một bảng hai chiều với các yếu tố ở các hàng và thuộc tính ở các cột.

Theo đánh giá của Nvidia, hiện nay đã có tới 30% tất cả các tài nguyên máy tính được sử dụng để xử lý dữ liệu có cấu trúc - các DataFrame, và tỷ lệ này vẫn tiếp tục tăng.

> *Xử lý dữ liệu là lĩnh vực mà có lẽ một phần ba tất cả các phép tính trên toàn thế giới đang thực hiện tại mỗi công ty. Xử lý dữ liệu và dữ liệu của hầu hết các công ty nằm trong DataFrame, ở định dạng bảng.*
>
> *- Jensen Huang, Giám đốc điều hành của Nvidia*

Hãy liệt kê một số đặc điểm chính của DataFrame trong Pandas:

- Cột: trong DataFrame, dữ liệu được tổ chức thành các cột, mỗi cột có một tên duy nhất. Các cột thuộc tính có thể chứa dữ liệu của nhiều loại khác nhau, giống như các cột trong cơ sở dữ liệu hoặc các cột trong bảng.
- Pandas Series là một cấu trúc dữ liệu một chiều trong Pandas, tương tự như danh sách hoặc cột trong bảng, nơi mỗi giá trị đi kèm với một chỉ số riêng.

> Trong Pandas Series có hơn 400 thuộc tính và phương thức, điều này làm cho việc làm việc với dữ liệu trở nên vô cùng linh hoạt. Bạn có thể áp dụng trực tiếp một trong bốn trăm chức năng có sẵn vào cột, thực hiện các phép toán toán học, lọc dữ liệu, thay thế giá trị, làm việc với ngày tháng, chuỗi và nhiều thứ khác. Hơn nữa, Series hỗ trợ các phép toán vector hóa, điều này giúp tăng tốc độ xử lý các tập dữ liệu lớn so với các phép tính theo chu kỳ. Ví dụ, bạn có thể dễ dàng nhân tất cả các giá trị với một số, thay thế dữ liệu thiếu hoặc áp dụng các biến đổi phức tạp mà không cần viết những vòng lặp phức tạp.

- Dòng: trong DataFrame có thể được lập chỉ mục bằng các giá trị duy nhất. Chỉ mục này cho phép thay đổi và điều chỉnh dữ liệu trong các dòng nhất định một cách nhanh chóng.
- Chỉ số: theo mặc định khi tạo DataFrame, Pandas gán chỉ số cho mỗi dòng từ 0 đến N-1 (trong đó N là số lượng tất cả các dòng trong DataFrame). Tuy nhiên, chỉ số có thể được thay đổi để bao gồm các ký hiệu đặc biệt, như ngày tháng hoặc các đặc điểm duy nhất.
- Lập chỉ mục dòng trong DataFrame có nghĩa là mỗi dòng được gán một tên hoặc nhãn duy nhất, được gọi là chỉ số của DataFrame.
- Các loại dữ liệu: DataFrame hỗ trợ nhiều loại dữ liệu khác nhau, bao gồm: `int`, `float`, `bool`, `datetime64` và `object` cho dữ liệu văn bản. Mỗi cột trong DataFrame có định dạng loại dữ liệu riêng, điều này xác định những phép toán nào có thể thực hiện trên nội dung của nó.
- Các phép toán trên dữ liệu: DataFrame hỗ trợ một loạt các phép toán để xử lý dữ liệu, bao gồm tổng hợp (`groupby`), hợp nhất (`merge` và `join`), nối (`concat`), chia-tai-đề và nhiều phương pháp chuyển đổi dữ liệu khác.
- Điều chỉnh kích thước: DataFrame cho phép bạn thêm và xoá các cột và dòng, điều này làm cho nó trở thành một cấu trúc động có thể được thay đổi theo nhu cầu phân tích dữ liệu.
- Trực quan hóa dữ liệu: sử dụng các phương thức trực quan hóa tích hợp sẵn hoặc tương tác với các thư viện trực quan hóa dữ liệu phổ biến như Matplotlib hoặc Seaborn, DataFrame có thể dễ dàng chuyển đổi thành các biểu đồ và đồ thị để đại diện cho dữ liệu dưới dạng đồ họa.
- Nhập và xuất dữ liệu: Pandas cung cấp các chức năng để đọc, nhập và xuất dữ liệu sang nhiều định dạng tệp khác nhau, như CSV, Excel, JSON, HTML và SQL, điều này khiến DataFrame trở thành một trung tâm quan trọng cho việc thu thập và phân phối dữ liệu.

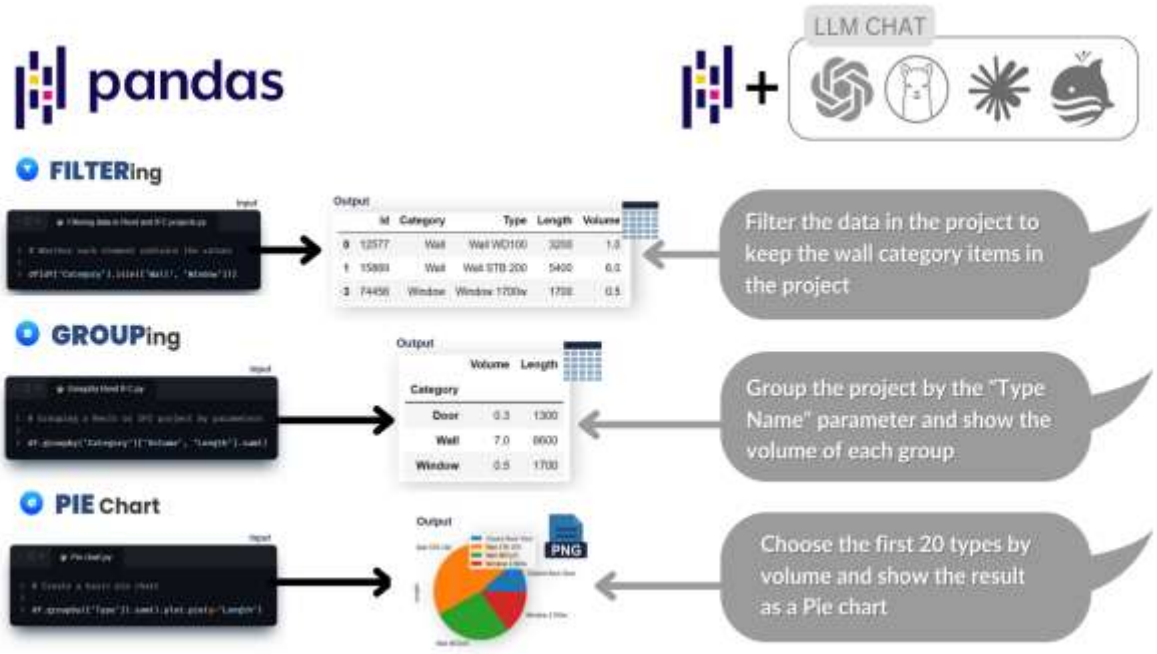
Khác với các định dạng CSV và XLSX, Pandas DataFrame cung cấp độ linh hoạt và hiệu suất cao hơn khi làm việc với dữ liệu: nó cho phép xử lý khối lượng thông tin lớn trong bộ nhớ, hỗ trợ các loại dữ liệu nâng cao (bao gồm ngày tháng, giá trị logic và chuỗi thời gian), và cung cấp nhiều khả năng lọc, tổng hợp, hợp nhất và trực quan hóa dữ liệu. Trong khi CSV không lưu trữ thông tin về loại dữ liệu và cấu trúc, XLSX thường bị quá tải với định dạng và có khả năng mở rộng kém, thì DataFrame vẫn là lựa chọn tối ưu cho phân tích nhanh, tự động hóa quy trình và tích hợp với các mô hình AI. Trong các chương tiếp theo, chúng tôi sẽ xem xét chi tiết từng khía cạnh của dữ liệu này, và trong phần 8 của cuốn sách sẽ xem xét chi tiết các định dạng tương tự như Parquet, Apache Orc, JSON, Feather, HDF5 và kho dữ liệu.--



	XLSX	CSV	Pandas DataFrame
Storage	Tabular	Tabular	Tabular
Usage	Office tasks, data presentation	Simple data exchange	Data analysis, manipulation
Compression	Built-in	None	None (in-memory)
Performance	Low	Medium	High (memory dependent)
Complexity	High (formatting, styles)	Low	Low
Data Type Support	Limited	Very limited	Extended
Scalability	Low	Low	Medium (memory limited)

DataFrame - lựa chọn tối ưu để thao tác với dữ liệu với hiệu suất cao và hỗ trợ loại dữ liệu nâng cao.

Nhờ vào độ linh hoạt, sức mạnh và sự dễ sử dụng của nó, thư viện Pandas và định dạng DataFrame đã trở thành tiêu chuẩn de facto trong lĩnh vực phân tích dữ liệu trên Python. Chúng hoàn toàn phù hợp cho việc tạo ra các báo cáo đơn giản cũng như xây dựng các quy trình phân tích phức tạp, đặc biệt là khi kết hợp với các mô hình LLM.



Các mô hình LLM đơn giản hóa việc tương tác với Pandas: thay vì mã, chỉ cần một truy vấn văn bản.

Ngày nay, Pandas được sử dụng rộng rãi trong các cuộc trò chuyện trên nền tảng LLM - chẳng hạn như ChatGPT, LlaMa, DeepSeek, QWEN và nhiều cái khác. Trong nhiều trường hợp, khi mô hình nhận được truy vấn liên quan đến xử lý bảng, kiểm tra dữ liệu hoặc phân tích, nó sẽ tự động sinh mã bằng cách sử dụng thư viện Pandas. Điều này làm cho DataFrame trở thành "ngôn ngữ" tự nhiên để đại diện cho dữ liệu trong các cuộc hội thoại với AI.-

Các công nghệ xử lý dữ liệu hiện đại như Pandas đã làm đơn giản hóa đáng kể việc phân tích, tự động hóa và tích hợp dữ liệu vào quy trình kinh doanh. Chúng cho phép nhanh chóng nhận kết quả, giảm tải cho các chuyên gia và đảm bảo tính tái lập của các thao tác.

Các bước tiếp theo: tạo ra khung dữ liệu bền vững

Trong phần này, chúng ta đã xem xét những loại dữ liệu chính được sử dụng trong ngành xây dựng, làm quen với các định dạng lưu trữ khác nhau và phân tích vai trò của các công cụ hiện đại, bao gồm LLM và IDE, trong việc xử lý thông tin. Chúng tôi đã nhận ra rằng quản lý dữ liệu hiệu quả là nền tảng cho việc đưa ra quyết định dựa trên dữ liệu và tự động hóa các quy trình kinh doanh. Các tổ chức có khả năng cấu trúc và hệ thống hóa dữ liệu của họ sẽ đạt được lợi thế cạnh tranh đáng kể trong các giai đoạn xử lý và chuyển đổi dữ liệu.

Tóm lại phần này, cần nêu rõ các bước thực tiễn chính sẽ giúp áp dụng các phương pháp đã thảo luận vào các nhiệm vụ hàng ngày của bạn:

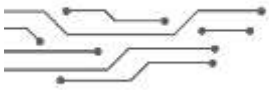
- Tiến hành kiểm toán dữ liệu trong các quy trình của bạn
  - [ ] Lập danh sách tất cả các loại dữ liệu mà bạn sử dụng trong các dự án

- [ ] Xác định loại và mô hình dữ liệu nào là quan trọng nhất cho các quy trình kinh doanh của bạn
- [ ] Phát hiện các vấn đề nơi thông tin thường không được cấu trúc, cấu trúc kém hoặc không thể truy cập

- Bắt đầu xây dựng chiến lược quản lý dữ liệu
  - [ ] Đặt ra các câu hỏi về chính sách và tiêu chuẩn để làm việc với các loại dữ liệu khác nhau
  - [ ] Phân tích xem những quy trình làm việc nào của bạn có thể được cải thiện bằng cách chuyển đổi dữ liệu không được cấu trúc thành dữ liệu có cấu trúc
  - [ ] Tạo quy định về lưu trữ và truy cập dữ liệu, xem xét an ninh và quyền riêng tư.
- Thiết lập và làm quen với các công cụ cơ bản để làm việc với dữ liệu
  - [ ] Chọn IDE phù hợp với nhiệm vụ của bạn (ví dụ, cài đặt VS Code hoặc Jupyter Notebook)
  - [ ] Hãy thử cài đặt LLM cục bộ để xử lý bảo mật dữ liệu cá nhân của bạn
  - [ ] Bắt đầu thử nghiệm với thư viện Pandas để xử lý dữ liệu bảng XLSX
  - [ ] Mô tả cho LLM các nhiệm vụ điển hình mà bạn xử lý trong các công cụ bảng hoặc cơ sở dữ liệu và yêu cầu LLM tự động hóa công việc thông qua Pandas

Việc áp dụng các bước tương tự sẽ giúp bạn dần thay đổi cách tiếp cận trong làm việc với dữ liệu, chuyển từ các mảng thông tin rời rạc, không có cấu trúc sang một hệ sinh thái thống nhất, nơi dữ liệu trở thành tài sản có sẵn và dễ hiểu. Bắt đầu từ những điều nhỏ – tạo khung dữ liệu đầu tiên trong Pandas, khởi chạy LLM cục bộ, tự động hóa nhiệm vụ định kỳ đầu tiên bằng Python (ví dụ, làm việc với bảng trong Excel).

Phần thứ tư của cuốn sách sẽ tập trung vào các vấn đề về chất lượng dữ liệu, tổ chức của chúng, cấu trúc và mô hình hóa. Chúng ta sẽ tập trung vào các phương pháp cho phép chuyển đổi các nguồn thông tin rời rạc — từ PDF và văn bản đến hình ảnh và mô hình CAD — thành các mảng có cấu trúc, phù hợp cho phân tích và tự động hóa. Chúng ta cũng sẽ tìm hiểu cách xác định yêu cầu về dữ liệu, cách xây dựng các mô hình khái niệm và logic trong các dự án xây dựng, và cách mà các mô hình ngôn ngữ hiện đại (LLM) có thể hỗ trợ trong quá trình này.





IV PHẦN
CHẤT LƯỢNG DỮ LIỆU: TỔ CHỨC, CẤU TRÚC, MÔ HÌNH HÓA

Phần thứ tư tập trung vào các phương pháp và công nghệ đảm bảo việc chuyển đổi thông tin rời rạc thành các mảng dữ liệu có cấu trúc chất lượng cao. Các quy trình hình thành và tài liệu hóa yêu cầu về dữ liệu được xem xét chi tiết như là cơ sở của kiến trúc thông tin hiệu quả trong các dự án xây dựng. Các phương pháp thực tiễn để trích xuất thông tin có cấu trúc từ các nguồn khác nhau (tài liệu PDF, hình ảnh, tập tin văn bản, mô hình CAD) được trình bày với các ví dụ cụ thể. Việc áp dụng các biểu thức chính quy (RegEx) và các công cụ khác để xác thực và xác nhận dữ liệu tự động được phân tích. Quy trình mô hình hóa dữ liệu theo từng bước ở cấp độ khái niệm, logic và vật lý được mô tả với sự xem xét đặc thù của ngành xây dựng. Các ví dụ cụ thể về việc sử dụng các mô hình ngôn ngữ (LLM) để tự động hóa các quy trình cấu trúc và kiểm tra thông tin được trình bày. Những phương pháp hiệu quả để trực quan hóa kết quả phân tích, nâng cao tính khả dụng của thông tin phân tích cho tất cả các cấp quản lý trong các dự án xây dựng được đề xuất.

CHƯƠNG 4.1.
CHUYỂN ĐỔI DỮ LIỆU THÀNH HÌNH THỨC CẤU TRÚC

Trong kỷ nguyên kinh tế dựa trên dữ liệu, dữ liệu không còn là rào cản mà trở thành nền tảng cho việc ra quyết định. Thay vì liên tục điều chỉnh thông tin cho mỗi hệ thống mới và định dạng của nó, các công ty ngày càng hướng đến việc hình thành một mô hình dữ liệu có cấu trúc thống nhất, phục vụ như một nguồn chân lý phổ quát cho tất cả các quy trình. Các hệ thống thông tin hiện đại không được thiết kế xoay quanh các định dạng và giao diện, mà xoay quanh ý nghĩa của dữ liệu — bởi vì cấu trúc có thể thay đổi, trong khi ý nghĩa của thông tin vẫn không thay đổi lâu hơn.

> Chìa khóa để làm việc hiệu quả với dữ liệu không nằm ở việc chuyển đổi và biến đổi của nó vô tận, mà là ở việc tổ chức đúng ngay từ đầu: tạo ra một cấu trúc đa năng, có thể đảm bảo tính minh bạch, tự động hóa và tích hợp ở tất cả các giai đoạn của vòng đời dự án.

Phương pháp truyền thống buộc phải điều chỉnh thủ công khi triển khai mỗi nền tảng mới: chuyển dữ liệu, thay đổi tên thuộc tính, điều chỉnh định dạng. Những bước này không cải thiện chất lượng của chính dữ liệu, mà chỉ che giấu các vấn đề, tạo ra chu trình khép kín của các chuyển đổi vô tận. Hậu quả là các công ty trở nên phụ thuộc vào các giải pháp phần mềm cụ thể, và sự chuyển đổi số bị chậm lại.

Trong các chương tiếp theo, chúng ta sẽ xem xét cách cấu trúc dữ liệu một cách hợp lý, sau đó là cách tạo ra các mô hình đa năng, giảm thiểu sự phụ thuộc vào nền tảng và tập trung vào điều chính - dữ liệu như một nguồn lực chiến lược, xoay quanh đó hình thành các quy trình bền vững.

Học cách chuyển đổi tài liệu, PDF, hình ảnh và văn bản thành các định dạng cấu trúc

Trong các dự án xây dựng, phần lớn thông tin tồn tại dưới dạng không cấu trúc: đó là các tài liệu kỹ thuật, biên bản công việc, bản vẽ, thông số kỹ thuật, biểu đồ, biên bản. Sự đa dạng của chúng - cả về định dạng lẫn nội dung - làm phức tạp việc tích hợp và tự động hóa.

> Quá trình chuyển đổi thành các định dạng cấu trúc hoặc bán cấu trúc có thể thay đổi tùy thuộc vào loại dữ liệu đầu vào và kết quả mong muốn trong quá trình xử lý.

Chuyển đổi dữ liệu từ dạng không cấu trúc sang dạng cấu trúc là cả nghệ thuật lẫn khoa học. Quá trình này thay đổi tùy theo loại dữ liệu đầu vào và mục tiêu phân tích và thường chiếm một phần đáng kể công việc của kỹ sư (Hình 3.25) trong việc xử lý và phân tích dữ liệu, với mục tiêu nhận được tập dữ liệu sạch, có tổ chức.



Hình 4.11 Chuyển đổi tài liệu quét không cấu trúc thành định dạng bảng cấu trúc.

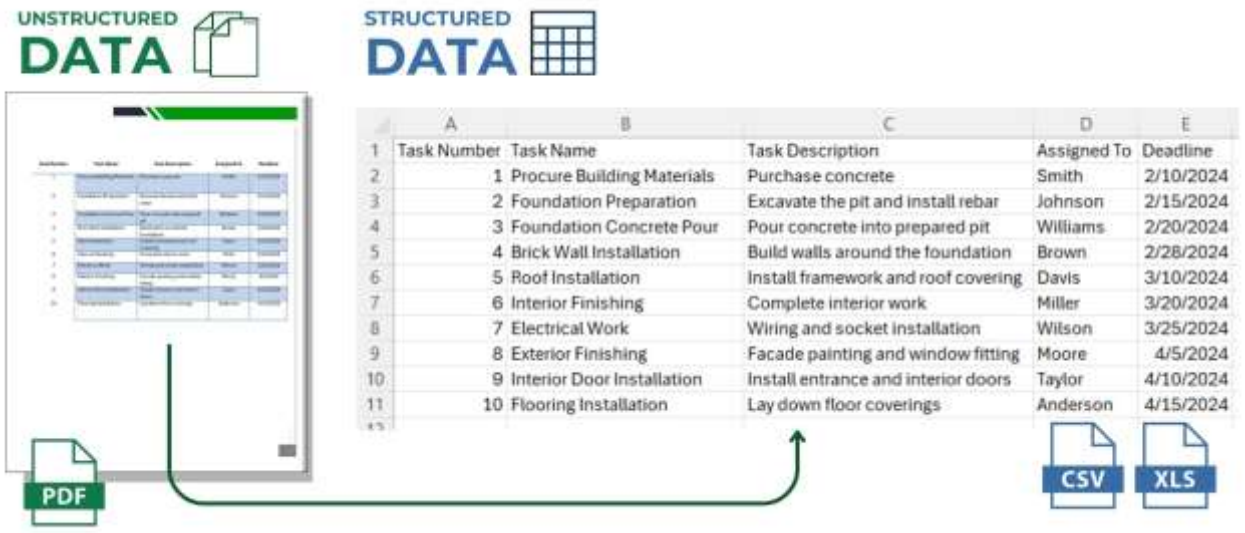
Việc chuyển đổi tài liệu, PDF, hình ảnh và văn bản thành định dạng cấu trúc (Hình 4.11) là một quá trình từng bước, bao gồm các giai đoạn sau:-

- Trích xuất dữ liệu (Extract): ở giai đoạn này, tài liệu hoặc hình ảnh gốc chứa dữ liệu không cấu trúc được tải lên. Điều này có thể là một tài liệu PDF, bức ảnh, bản vẽ hoặc sơ đồ.
- Chuyển đổi dữ liệu (Transform): tiếp theo là giai đoạn chuyển đổi dữ liệu không cấu trúc thành định dạng có cấu trúc. Ví dụ, điều này có thể bao gồm việc nhận diện và diễn giải văn bản từ hình ảnh bằng cách sử dụng nhận dạng ký tự quang học (OCR) hoặc các phương pháp xử lý khác.
- Tải lên và lưu trữ dữ liệu (Load): giai đoạn cuối cùng liên quan đến việc lưu trữ dữ liệu đã xử lý ở các định dạng khác nhau như CSV, XLSX, XML, JSON, để phục vụ cho công việc sau này, trong đó việc lựa chọn định dạng phụ thuộc vào các yêu cầu và sở thích cụ thể.

Quá trình này, được gọi là ETL (Extract, Transform, Load), đóng vai trò then chốt trong việc xử lý dữ liệu tự động, mà chúng ta sẽ bàn chi tiết hơn trong chương "ETL và Pipeline: Extract, Transform, Load". Tiếp theo, chúng ta sẽ cùng xem xét các ví dụ về cách tài liệu ở nhiều định dạng khác nhau được chuyển đổi thành dữ liệu có cấu trúc.

Ví dụ về việc chuyển đổi tài liệu PDF thành bảng

Một trong những nhiệm vụ thường gặp nhất trong các dự án xây dựng là xử lý các yêu cầu kỹ thuật ở định dạng PDF. Để minh họa cho quá trình chuyển đổi từ dữ liệu không cấu trúc sang định dạng có cấu trúc, hãy xem xét một ví dụ thực tiễn: trích xuất bảng từ tài liệu PDF và chuyển đổi nó thành định dạng CSV hoặc Excel (Hình 4.12). -



	A	B	C	D	E
1	Task Number	Task Name	Task Description	Assigned To	Deadline
2	1	Procure Building Materials	Purchase concrete	Smith	2/10/2024
3	2	Foundation Preparation	Excavate the pit and install rebar	Johnson	2/15/2024
4	3	Foundation Concrete Pour	Pour concrete into prepared pit	Williams	2/20/2024
5	4	Brick Wall Installation	Build walls around the foundation	Brown	2/28/2024
6	5	Roof Installation	Install framework and roof covering	Davis	3/10/2024
7	6	Interior Finishing	Complete interior work	Miller	3/20/2024
8	7	Electrical Work	Wiring and socket installation	Wilson	3/25/2024
9	8	Exterior Finishing	Facade painting and window fitting	Moore	4/5/2024
10	9	Interior Door Installation	Install entrance and interior doors	Taylor	4/10/2024
11	10	Flooring Installation	Lay down floor coverings	Anderson	4/15/2024

Hình 4.12 Khác với PDF, các định dạng CSV và XLSX rất phổ biến và dễ dàng tích hợp vào nhiều hệ thống quản lý dữ liệu khác nhau.

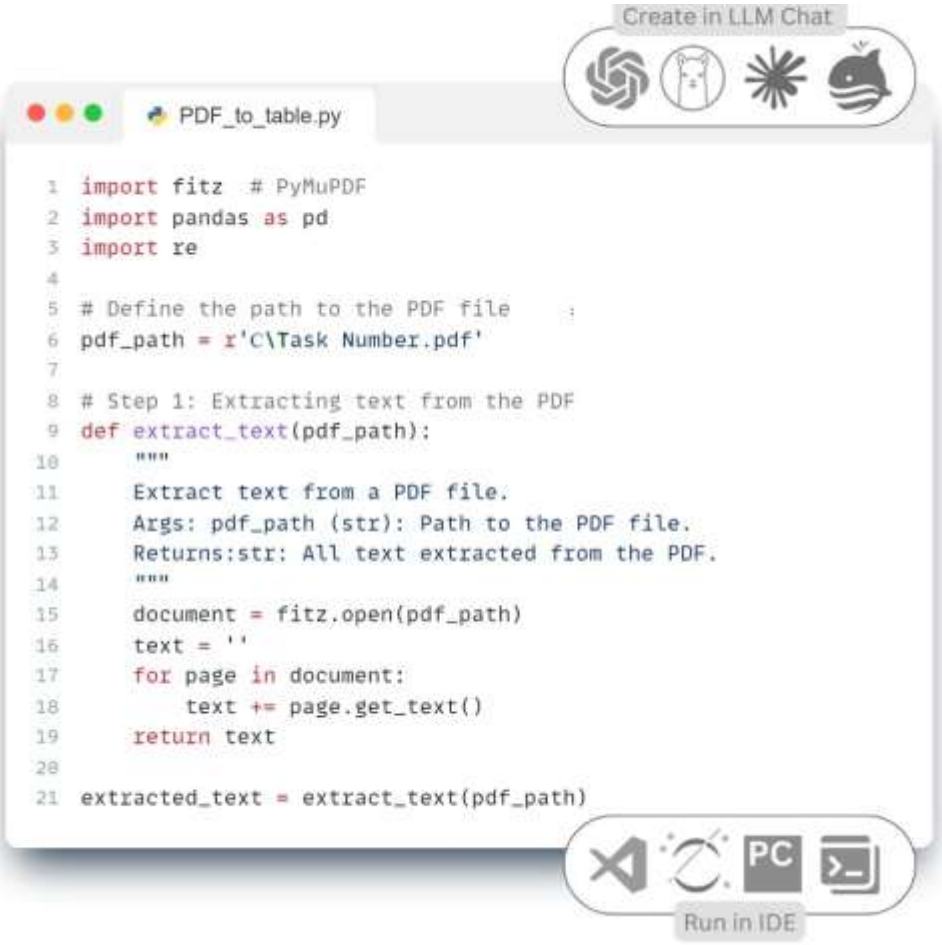
Các mô hình ngôn ngữ LLM, chẳng hạn như ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN, làm đơn giản hóa đáng kể công việc của các chuyên gia với dữ liệu, giảm thiểu nhu cầu phải nghiên cứu sâu về các ngôn ngữ lập trình và cho phép giải quyết nhiều vấn đề bằng các truy vấn văn bản.

Do đó, thay vì dành thời gian để tìm kiếm giải pháp trên internet (thông thường đó là trang web StackOverFlow hoặc các diễn đàn và trò chuyện chuyên ngành) hoặc hỏi ý kiến từ các chuyên gia về xử lý dữ liệu, chúng ta có thể tận dụng các khả năng của các LLM hiện đại trực tuyến hoặc cục bộ. Chỉ cần đặt một truy vấn, và mô hình sẽ cung cấp mã sẵn sàng để chuyển đổi tài liệu PDF thành định dạng bảng.

- Gửi truy vấn văn bản sau đây đến bất kỳ mô hình LLM nào (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ mô hình nào khác):

> Vui lòng viết mã để trích xuất văn bản từ tệp PDF có chứa bảng. Mã phải nhận đường dẫn đến tệp làm tham số và trả về bảng đã trích xuất dưới dạng DataFrame.

- Phản hồi của mô hình LLM trong hầu hết các trường hợp sẽ được trình bày dưới dạng mã Python, vì ngôn ngữ này được sử dụng rộng rãi cho xử lý dữ liệu, tự động hóa và làm việc với các định dạng tệp khác nhau:



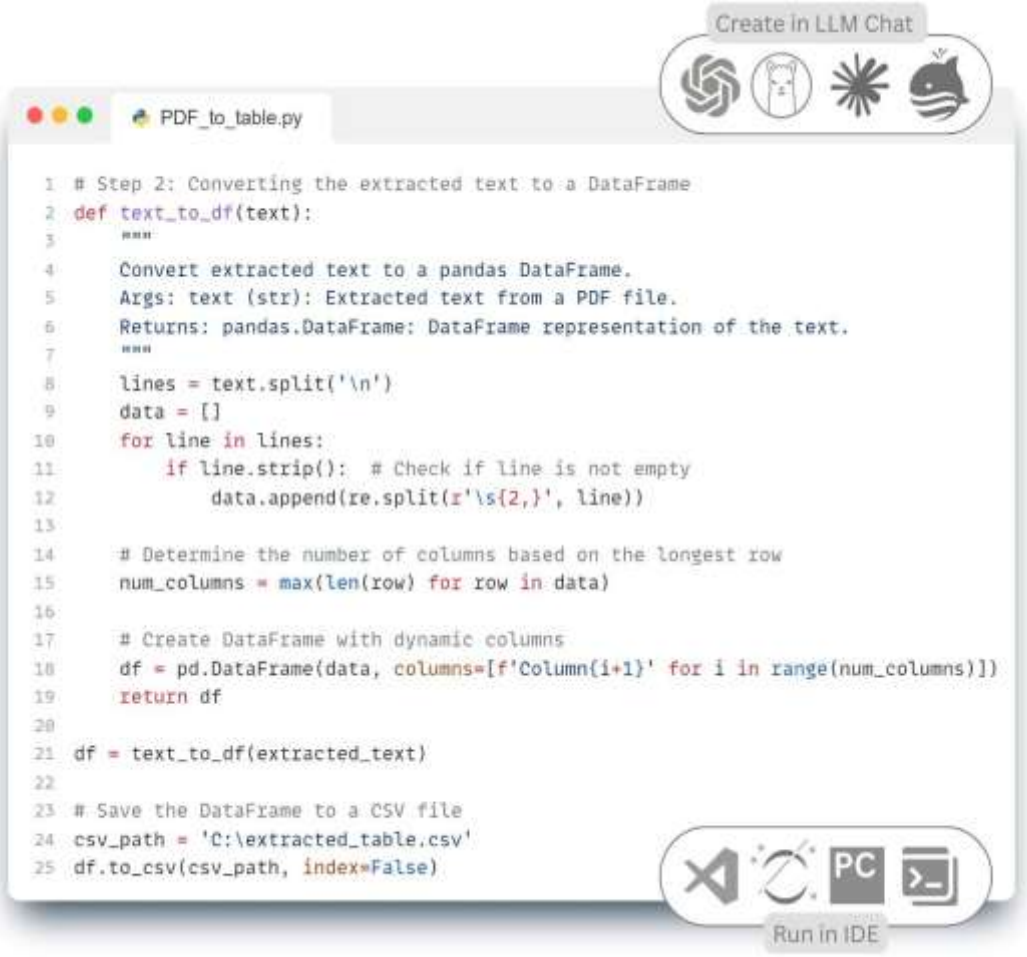
Hình. 4.13 Phản hồi của LLM dưới dạng mã Python và các thư viện cùng gói (Pandas, Fitz) trích xuất văn bản từ tệp PDF.

Mã này (Hình. 4.13) có thể được chạy trong một trong các IDE phổ biến mà chúng ta đã đề cập ở trên, ở chế độ ngoại tuyến: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse với plugin PyDev, Thonny, Wing IDE, IntelliJ IDEA với plugin Python, JupyterLab hoặc các công cụ trực tuyến phổ biến: Kaggle.com, Google Colab, Microsoft Azure Notebooks, Amazon SageMaker.

- Ở giai đoạn "Chuyển đổi", chúng tôi sử dụng thư viện Pandas phổ biến (mà chúng tôi đã thảo luận chi tiết trong chương "Python Pandas: công cụ không thể thiếu cho làm việc với dữ liệu") để đọc văn bản đã trích xuất vào DataFrame và lưu DataFrame vào tệp bảng CSV hoặc XLXS.

> Tôi cần mã sẽ chuyển đổi bảng kết quả từ tệp PDF thành DataFrame. Vui lòng thêm mã để lưu DataFrame vào tệp CSV.

Phản hồi từ LLM:



Hình. 4.14 Chuyển đổi bảng đã trích xuất từ PDF sang DataFrame và lưu bảng vào tệp CSV.

Nếu gặp lỗi khi thực hiện mã (Hình. 4.13, Hình. 4.14) — chẳng hạn như do thiếu thư viện hoặc đường dẫn đến tệp không đúng — bạn chỉ cần sao chép thông báo lỗi cùng với mã nguồn và gửi lại cho mô hình LLM. Mô hình sẽ phân tích thông điệp lỗi, giải thích vấn đề và đề xuất sửa chữa hoặc các bước bổ sung.-

> Do đó, việc tương tác với AI LLM trở thành một chu trình hoàn chỉnh: truy vấn → phản hồi → kiểm tra → phản hồi → điều chỉnh — mà không cần có kiến thức kỹ thuật sâu sắc.

Sử dụng truy vấn văn bản thông thường trong chat LLM và một vài dòng Python mà chúng ta có thể chạy

cục bộ trong bất kỳ IDE nào, chúng ta đã chuyển đổi tài liệu PDF thành định dạng bảng CSV, mà không giống như tài liệu PDF, dễ đọc cho máy và dễ dàng tích hợp vào bất kỳ hệ thống quản lý dữ liệu nào.

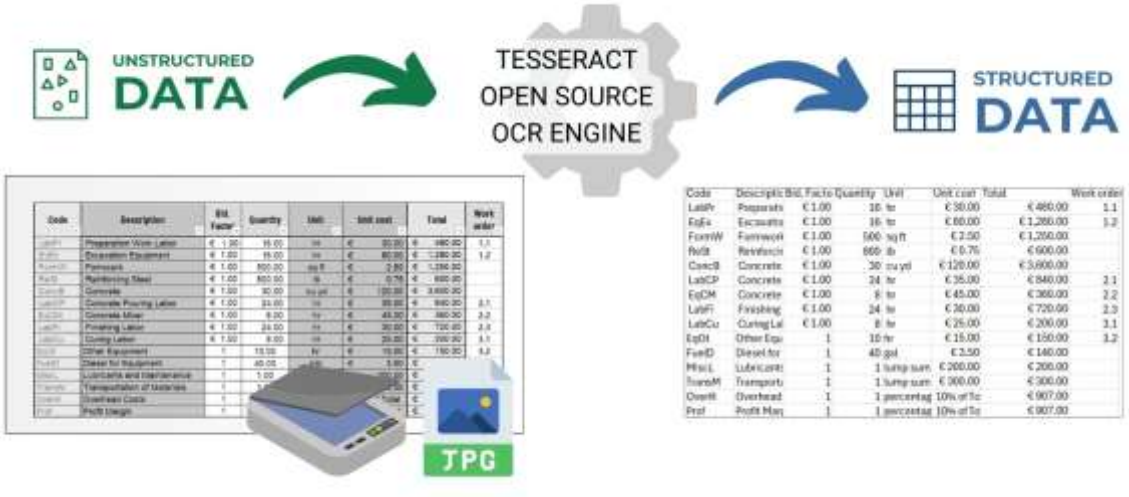
Chúng ta có thể áp dụng mã này (Hình. 4.13, Hình. 4.14), sao chép nó từ bất kỳ cuộc trò chuyện LLM nào, cho hàng chục và hàng nghìn tài liệu PDF mới trên máy chủ, từ đó tự động hóa quá trình chuyển đổi dòng tài liệu không có cấu trúc thành định dạng bảng có cấu trúc CSV.-

Tuy nhiên, tài liệu PDF không phải lúc nào cũng chứa văn bản; thường thì đây là các tài liệu đã scan, cần được xử lý như hình ảnh. Mặc dù hình ảnh về bản chất là không có cấu trúc, việc phát triển và áp dụng các thư viện nhận dạng cho phép trích xuất, xử lý và phân tích nội dung của chúng, từ đó giúp chúng ta có thể tận dụng những dữ liệu này trong các quy trình kinh doanh một cách đầy đủ.

Chuyển đổi hình ảnh JPEG, PNG thành hình thức cấu trúc

Hình ảnh là một trong những dạng dữ liệu không có cấu trúc phổ biến nhất. Trong ngành xây dựng và nhiều lĩnh vực khác, một lượng lớn thông tin được lưu trữ dưới dạng tài liệu đã scan, sơ đồ, hình ảnh và bản vẽ. Các dữ liệu này chứa rất nhiều thông tin quý giá, nhưng không thể được xử lý trực tiếp như một bảng Excel hoặc cơ sở dữ liệu. Hình ảnh mang lại nhiều thông tin phức tạp bởi vì nội dung, màu sắc, và kết cấu của chúng rất đa dạng, và việc trích xuất thông tin hữu ích yêu cầu một quy trình xử lý đặc biệt.

Sự phức tạp trong việc sử dụng hình ảnh như một nguồn dữ liệu nằm ở việc thiếu cấu trúc. Hình ảnh không truyền đạt ý nghĩa theo cách đơn giản, dễ dàng định lượng mà máy tính có thể ngay lập tức hiểu hoặc xử lý, như một bảng tính Excel hoặc bảng dữ liệu. Để chuyển đổi dữ liệu không có cấu trúc từ hình ảnh sang dạng có cấu trúc, cần phải sử dụng các thư viện đặc biệt, có khả năng diễn giải thông tin hình ảnh có trong đó.-



Hình. 4.15 Việc chuyển đổi các tài liệu đã scan và hình ảnh thành các định dạng có cấu trúc là khả thi thông qua các công cụ OCR đặc biệt.

Để trích xuất văn bản từ hình ảnh, công nghệ OCR (Optical Character Recognition) – nhận dạng ký tự quang học được áp dụng. Nó cho phép nhận biết các chữ cái và con số trong các bản scan của tài liệu, hình ảnh và tệp PDF, biến chúng thành văn bản có thể chỉnh sửa và máy tính có thể đọc được. Công nghệ OCR đã được sử dụng lâu dài trong tự động hóa quy trình xử lý tài liệu và ngày nay dễ dàng tích hợp vào bất kỳ quy trình kinh doanh hoặc ứng dụng Python nào. Một trong những công cụ OCR phổ biến nhất là Tesseract, mã

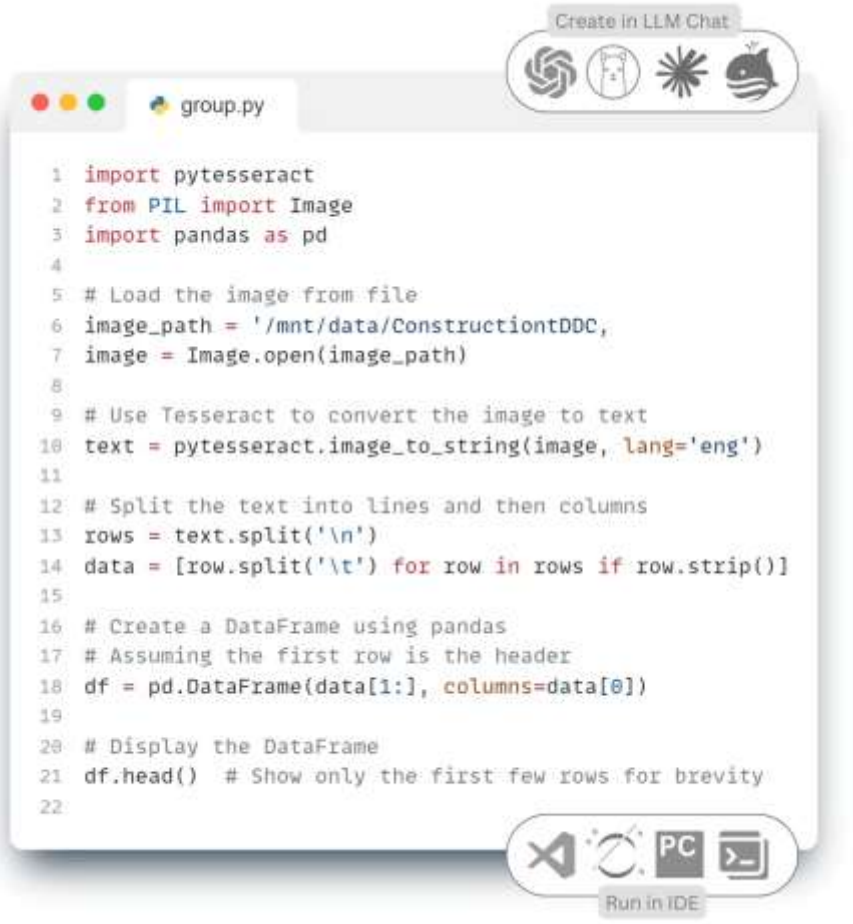
nguồn mở, ban đầu được phát triển bởi HP™ và hiện nay được hỗ trợ bởi Google™. Nó hỗ trợ hơn 100 ngôn ngữ và nổi bật về độ chính xác trong việc nhận diện.

Hãy yêu cầu LLM chat viết một ví dụ mã để lấy dữ liệu từ bảng đã scan hoặc chụp ảnh về dạng có cấu trúc.

- Gửi yêu cầu văn bản đến LLM chat (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ công cụ nào khác):

> Viết mã cho phép chuyển đổi hình ảnh JPEG chứa một bảng thành DataFrame ⏎

- Phản hồi từ LLM trong hầu hết các trường hợp sẽ đề xuất sử dụng thư viện Pytesseract để nhận diện văn bản trong hình ảnh.



Hình. 4.16 Chuyển đổi văn bản được trích xuất từ bảng của hình ảnh hoặc ảnh chụp thành định dạng bảng có cấu trúc.

Trong ví dụ này – mã (Hình. 4.16) thu được từ LLM, sử dụng thư viện pytesseract (Tesseract cho Python) để chuyển đổi hình ảnh thành văn bản với công nghệ OCR và thư viện Pandas để chuyển đổi văn bản này thành dạng có cấu trúc, tức là DataFrame.

> Quy trình chuyển đổi thường bao gồm việc xử lý trước để cải thiện chất lượng hình ảnh, sau đó áp dụng các thuật toán khác nhau để phát hiện mẫu, trích xuất đặc trưng hoặc nhận diện đối tượng. Kết quả là thông tin hình ảnh không có cấu trúc được chuyển đổi thành dữ liệu có cấu trúc.

Mặc dù PDF và hình ảnh là những nguồn thông tin không có cấu trúc chủ yếu, nhưng thực sự nhà vô địch về khối lượng lại là văn bản được tạo ra trong email, chat, họp và các ứng dụng nhắn tin. Những dữ liệu này không chỉ phong phú — chúng còn phân tán, không chính thức và cực kỳ kém cấu trúc.

Chuyển đổi dữ liệu văn bản thành hình thức cấu trúc

Ngoài các tài liệu PDF chứa bảng (Hình 4.12) và phiên bản quét của các biểu mẫu bảng (Hình 4.15), một phần đáng kể thông tin trong tài liệu dự án được trình bày dưới dạng văn bản. Đây có thể là các câu liên kết trong các tài liệu văn bản hoặc các ghi chú rời rạc được phân tán trên các bản vẽ và sơ đồ. Trong điều kiện xử lý dữ liệu hiện đại, một trong những nhiệm vụ thường gặp là chuyển đổi văn bản như vậy thành định dạng có cấu trúc mà có thể phân tích, trực quan hóa và đưa ra quyết định.-

Yếu tố trung tâm của quá trình này là phân loại — một hệ thống phân loại cho phép tổ chức thông tin theo các danh mục và tiểu danh mục dựa trên các đặc điểm chung.

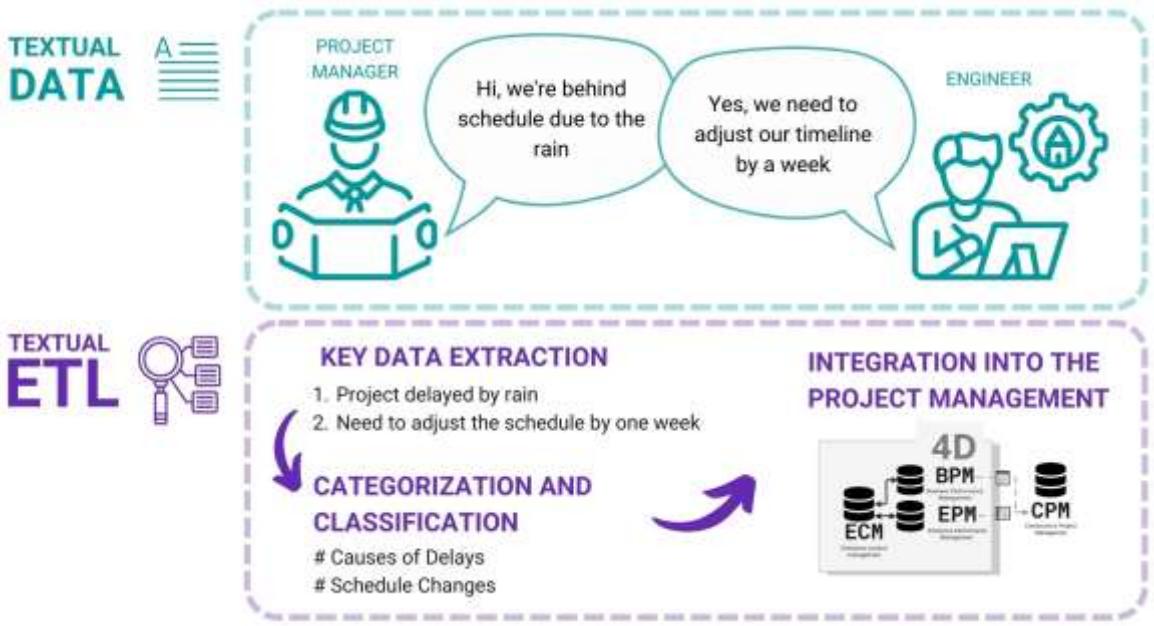
> Phân loại là một cấu trúc phân loại theo hệ thống phân cấp, được sử dụng để nhóm và tổ chức các đối tượng. Trong ngữ cảnh xử lý văn bản, nó phục vụ như một cơ sở cho việc phân phối có hệ thống các yếu tố theo các danh mục có nghĩa, từ đó đơn giản hóa phân tích và nâng cao chất lượng xử lý dữ liệu.

Việc tạo ra phân loại kèm theo các giai đoạn trích xuất thực thể, phân loại chúng và kết nối với ngữ cảnh. Để mô hình hóa quy trình trích xuất thông tin từ dữ liệu văn bản, cần thực hiện các bước tương tự như những gì chúng tôi đã áp dụng để cấu trúc dữ liệu từ các tài liệu PDF:

- Trích xuất dữ liệu (Extract): cần phân tích dữ liệu văn bản để trích xuất thông tin về sự trì hoãn và các thay đổi trong lịch trình dự án.
- Phân loại và phân loại (Transform): phân bổ thông tin thu được vào các danh mục, chẳng hạn như nguyên nhân trì hoãn và thay đổi trong lịch trình.
- Tích hợp (Load): cuối cùng, chuẩn bị dữ liệu có cấu trúc để tích hợp vào các hệ thống quản lý dữ liệu bên ngoài.

Hãy xét một tình huống: chúng ta có một cuộc đối thoại giữa quản lý dự án và kỹ sư, trong đó thảo luận về các vấn đề với lịch trình công việc. Mục tiêu của chúng ta là trích xuất các yếu tố chính (nguyên nhân trì hoãn, điều chỉnh thời gian) và trình bày chúng dưới dạng có cấu trúc (Hình 4.17).

Chúng ta sẽ thực hiện trích xuất dựa trên các từ khóa chính dự kiến, tạo một DataFrame để mô phỏng việc trích xuất dữ liệu và sau đó, sau khi biến đổi, một DataFrame mới sẽ có các cột cho ngày, sự kiện (ví dụ, nguyên nhân trì hoãn) và hành động (ví dụ, thay đổi lịch trình).



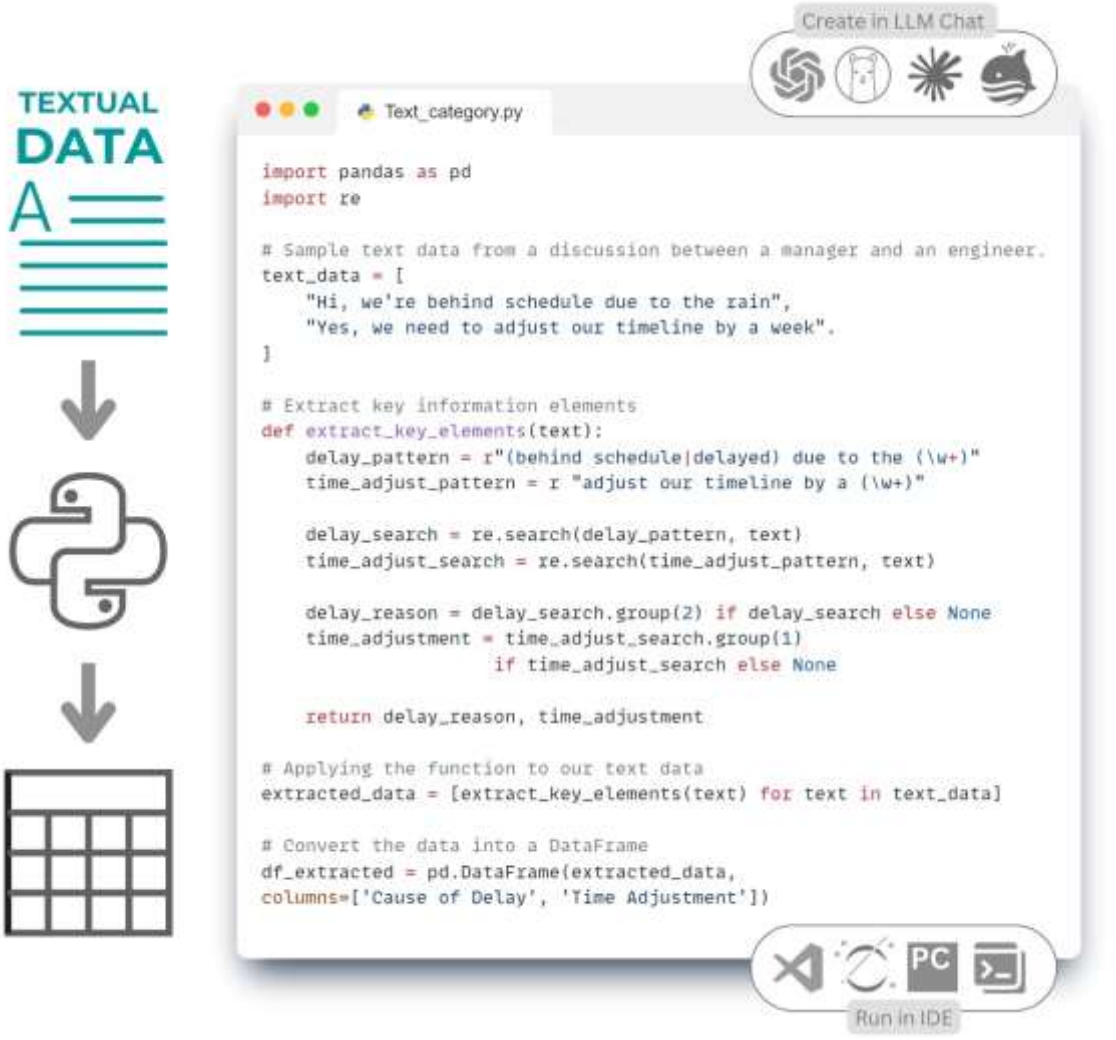
Hình 4.17 Trích xuất thông tin chính từ văn bản liên quan đến việc cần điều chỉnh lịch trình và tích hợp các thay đổi vào hệ thống quản lý dự án.

Chúng ta sẽ cung cấp mã cho nhiệm vụ này sử dụng truy vấn văn bản trong một trong các mô hình ngôn ngữ, như đã trình bày trong các ví dụ trước.

- Gửi truy vấn văn bản đến bất kỳ trò chuyện LLM nào:

> Tôi có một cuộc trò chuyện giữa người quản lý: "Xin chào, chúng tôi đang chậm tiến độ do mưa" và kỹ sư: "Vâng, chúng ta cần điều chỉnh thời gian thêm một tuần". Tôi cần một kịch bản sẽ phân tích các cuộc đối thoại văn bản tương tự trong tương lai, trích xuất từ chúng các nguyên nhân trì hoãn và những điều chỉnh thời gian cần thiết, sau đó tạo một DataFrame từ những dữ liệu này. Sau đó, DataFrame cần được lưu vào file CSV.

- Phản hồi từ LLM thường sẽ bao gồm mã Python sử dụng biểu thức chính quy (re - Regex) và thư viện Pandas (pd):

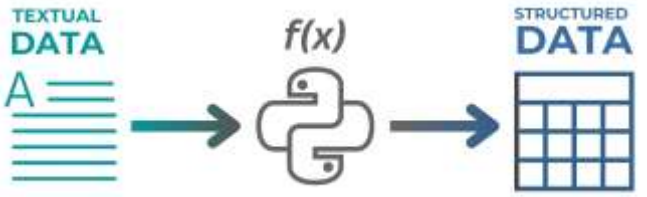


Hình 4.18 Tách thông tin chính từ văn bản về sự cần thiết điều chỉnh thời gian dưới dạng bảng.

Trong ví dụ này (Hình 4.17), dữ liệu văn bản chứa thư tín giữa quản lý dự án và kỹ sư được phân tích để nhận diện và trích xuất thông tin cụ thể có thể ảnh hưởng đến quản lý các dự án tương lai với các cuộc đối thoại tương tự. Sử dụng biểu thức chính quy (chúng ta sẽ nói thêm về biểu thức chính quy trong chương "Yêu cầu có cấu trúc và biểu thức chính quy RegEx"), xác định qua các mẫu nguyên nhân gây trì hoãn dự án và các điều chỉnh thời gian cần thiết. Hàm được viết trong ví dụ này trích xuất từ chuỗi hoặc nguyên nhân trì hoãn hoặc điều chỉnh thời gian, dựa trên các mẫu: xác định từ ngay sau "do", như là nguyên nhân trì hoãn hoặc từ ngay sau "về", như là điều chỉnh thời gian.

Nếu trong chuỗi đề cập đến trì hoãn do thời tiết, thì nguyên nhân được xác định là "mưa"; nếu trong chuỗi đề cập đến việc điều chỉnh thời gian trong một khoảng thời gian cụ thể, thì khoảng thời gian đó được trích

xuất như là điều chỉnh thời gian (Hình 4.19). Việc thiếu bất kỳ từ nào trong số này trong chuỗi sẽ dẫn đến giá trị "Không" cho thuộc tính-cột tương ứng.



Text string	Cause of Delay	Time Adjustment
0	rain	None
1	None	week

Hình 4.19 Bảng tổng hợp, nhận được dưới dạng DataFrame, sau khi thực hiện mã chứa thông tin về sự tồn tại của trì hoãn và các điều chỉnh thời gian cần thiết.

Việc cấu trúc và tham số hóa các điều kiện từ văn bản (đối thoại, thư tín, tài liệu) cho phép nhanh chóng giải quyết các trì hoãn trong xây dựng: chẳng hạn, thiếu nhân công có thể ảnh hưởng đến tốc độ làm việc trong thời tiết xấu, vì vậy các công ty, biết các tham số trì hoãn từ các cuộc đối thoại (Hình 4.19) giữa chỉ huy công trình tại hiện trường và quản lý dự án - có thể tăng cường đội ngũ khi có dự báo không thuận lợi.

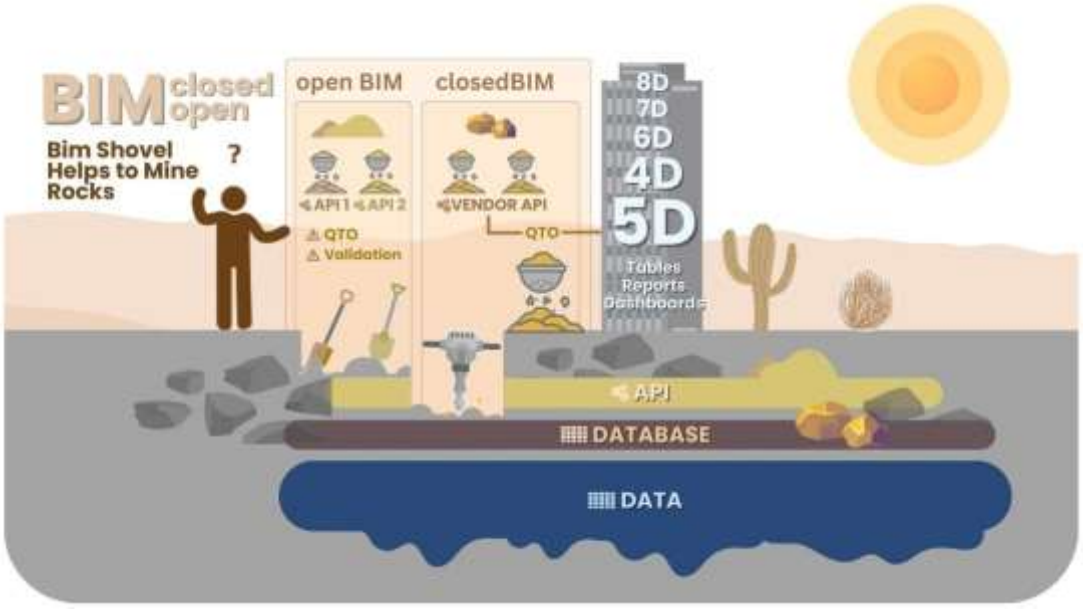
Việc chuyển đổi tài liệu và hình ảnh vào định dạng có cấu trúc có thể đạt được thông qua các công cụ tương đối đơn giản, mở và miễn phí, dựa trên phân loại.

Phân loại các yếu tố cũng là một phần quan trọng trong việc làm việc với dữ liệu dự án, đặc biệt là trong bối cảnh sử dụng phần mềm CAD (BIM).

Chuyển đổi dữ liệu CAD (BIM) thành hình thức cấu trúc

Việc cấu trúc và phân loại dữ liệu CAD (BIM) là một nhiệm vụ phức tạp hơn, vì dữ liệu được lưu trữ từ cơ sở dữ liệu CAD (BIM) gần như luôn được trình bày dưới dạng các định dạng bảo mật hoặc khó khăn về tham số, thường kết hợp đồng thời các yếu tố dữ liệu hình học (có cấu trúc bán phần) và các yếu tố siêu thông tin (dữ liệu có hoặc không có cấu trúc).

Các định dạng dữ liệu nguyên bản trong các hệ thống CAD (BIM) thường được bảo vệ và không thể sử dụng trực tiếp, trừ khi áp dụng phần mềm chuyên dụng hoặc giao diện API của chính nhà phát triển (Hình 4.110). Sự cách ly dữ liệu như vậy tạo ra các kho lưu trữ khép kín - được gọi là "silô", hạn chế việc trao đổi thông tin tự do và cản trở việc tạo ra các quy trình số liên kết trong công ty.-



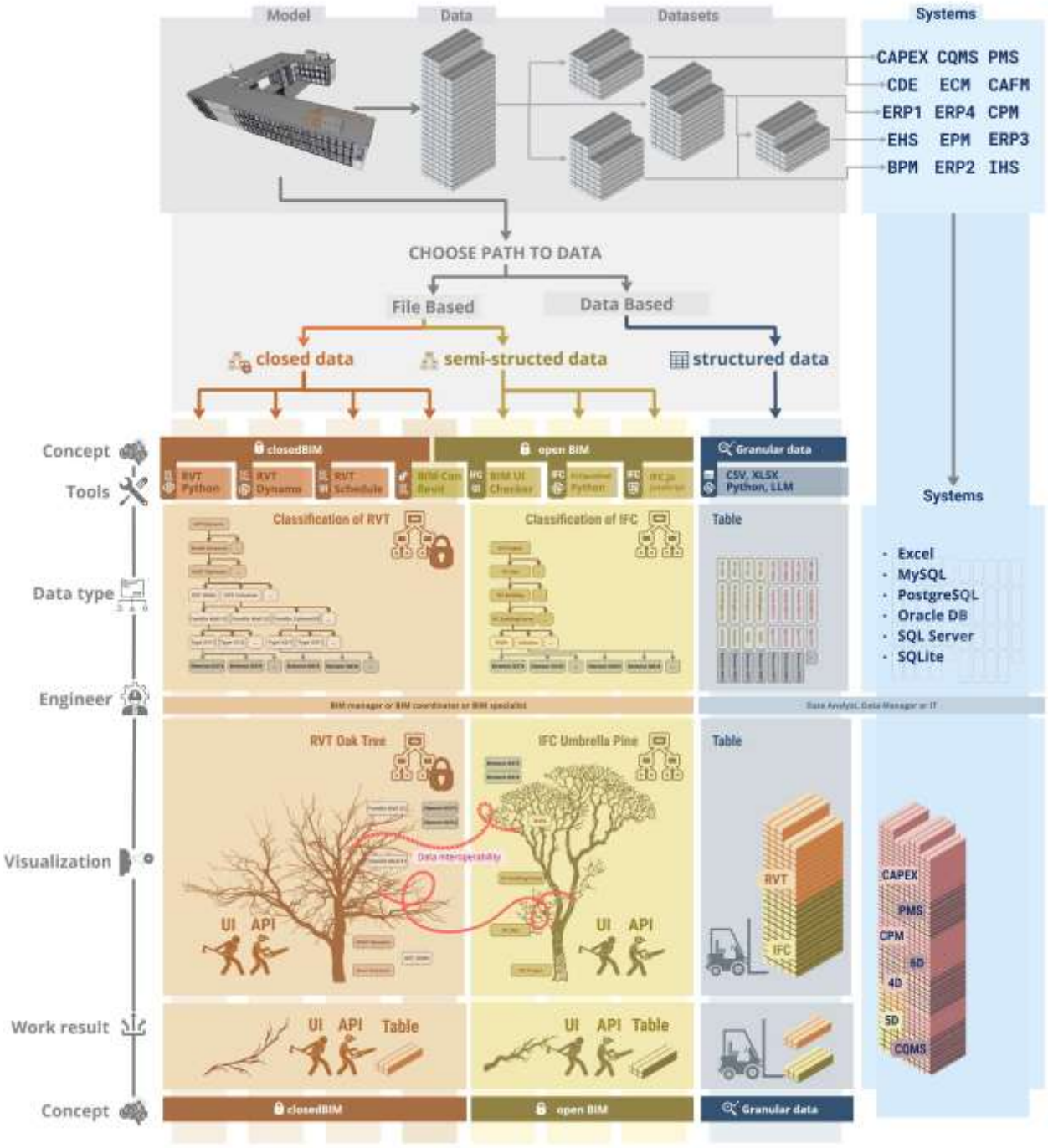
Hình 4.110 Các chuyên gia CAD (BIM) có thể truy cập dữ liệu nguyên bản thông qua các kết nối API hoặc công cụ của nhà cung cấp.

Trong các định dạng CAD (BIM) đặc biệt, thông tin về các đặc điểm và thuộc tính của các yếu tố dự án được thu thập trong một hệ thống phân loại theo thứ bậc, nơi các thực thể với các thuộc tính tương ứng được sắp xếp, giống như các trái cây của một cái cây, ở các nút cuối cùng của các nhánh phân loại dữ liệu (Hình 4.111).-

Việc trích xuất dữ liệu từ các hệ thống phân cấp như vậy có thể thực hiện theo hai cách: hoặc là thủ công, nhấp vào từng nút như thể đang xử lý một cái cây, chặt đi những nhánh được chọn của các danh mục và loại. Cách thay thế là sử dụng các giao diện lập trình ứng dụng (API) - cho phép một phương pháp tự động hóa hiệu quả hơn trong việc thu thập và nhóm dữ liệu, cuối cùng biến nó thành bảng có cấu trúc để sử dụng trong các hệ thống khác.

Để trích xuất bảng dữ liệu có cấu trúc từ các dự án CAD (BIM), có thể sử dụng nhiều công cụ khác nhau, chẳng hạn như Dynamo, pyRvt, Pandamo (Pandas + Dynamo), ACC, hoặc các giải pháp nguồn mở như IfcOpSh hoặc IFCjs cho định dạng IFC.

Các công cụ xuất và chuyển đổi dữ liệu hiện đại cho phép đơn giản hóa việc xử lý và chuẩn bị dữ liệu bằng cách tách nội dung của các mô hình CAD thành hai thành phần chính: thông tin hình học và dữ liệu thuộc tính - siêu thông tin mô tả các thuộc tính của các phần tử cấu trúc. Hai lớp dữ liệu này vẫn được liên kết với nhau thông qua các định danh duy nhất, nhờ đó có thể khớp chính xác từng phần tử với mô tả hình học (thông qua các tham số hoặc đa giác) với các thuộc tính của nó: tên, vật liệu, giai đoạn thực hiện, chi phí và nhiều hơn nữa. Cách tiếp cận này đảm bảo tính toàn vẹn của mô hình và cho phép sử dụng dữ liệu một cách linh hoạt cho cả việc trực quan hóa (dữ liệu hình học của mô hình) cũng như cho các tác vụ phân tích hoặc quản lý (có cấu trúc hoặc kém cấu trúc), làm việc với hai loại dữ liệu một cách riêng biệt hoặc song song.-



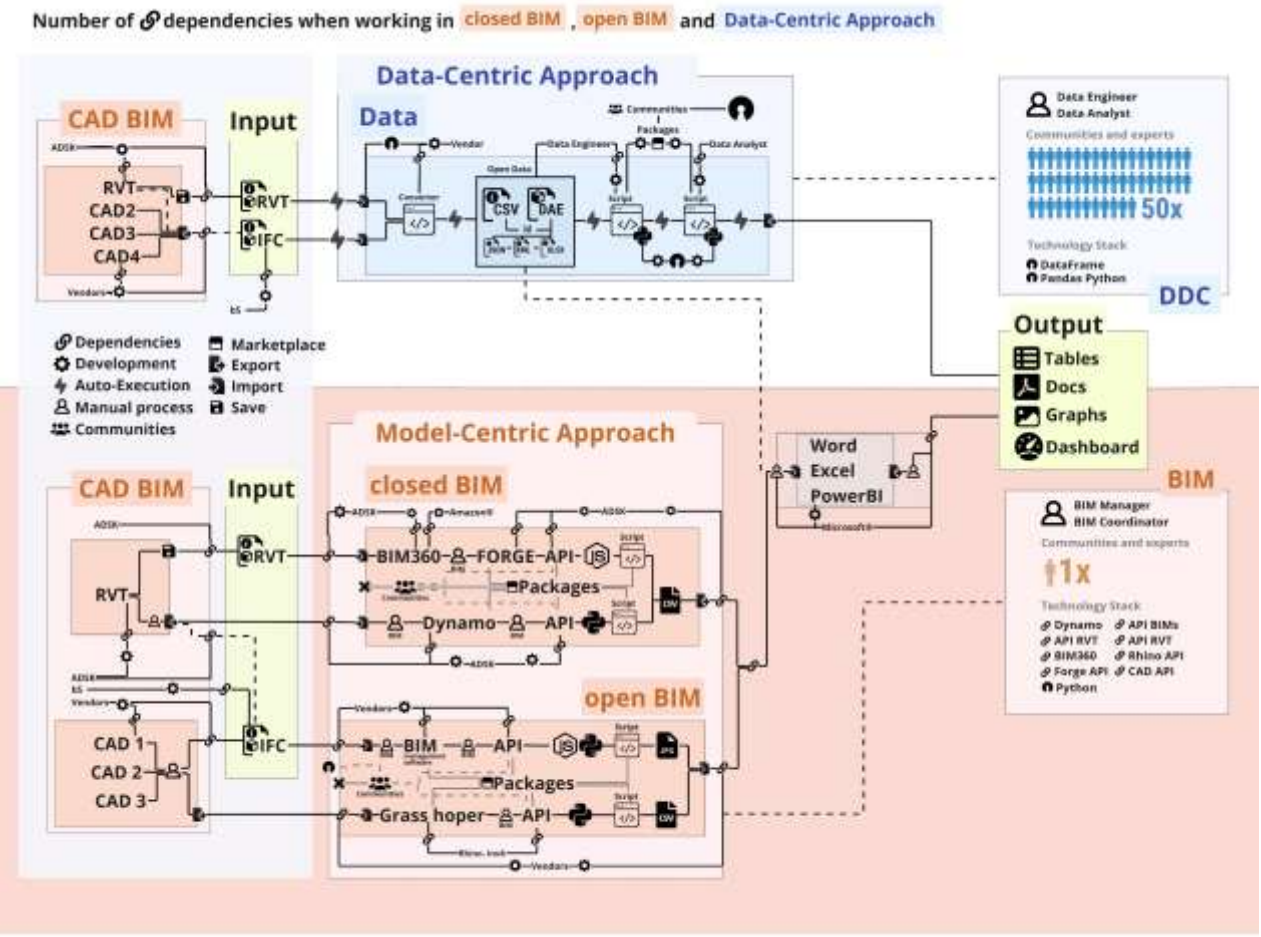
Hình 4.111 Thông tin từ các cơ sở dữ liệu CAD (BIM) được trình bày cho người dùng dưới dạng cây phân loại.

Với sự phát triển của công nghệ reverse engineering và sự xuất hiện của các bộ phát triển phần mềm SDK (Software Development Kit) cho việc chuyển đổi dữ liệu CAD - việc truy cập và chuyển đổi dữ liệu từ các định dạng đóng của phần mềm CAD (BIM) trở nên đơn giản hơn nhiều. Giờ đây có thể hợp pháp và an toàn để chuyển đổi dữ liệu từ các định dạng đóng sang các định dạng linh hoạt, phù hợp cho việc phân tích và sử dụng trong các hệ thống khác. Về lịch sử xuất hiện của các công cụ đầu tiên cho reverse engineering

("Open DWG") và cuộc chiến giành sự thống trị trên các định dạng CAD của các nhà cung cấp, chúng tôi đã đề cập trong chương "Dữ liệu có cấu trúc: nền tảng của chuyển đổi số".

> Các công cụ reverse engineering cho phép nhận được dữ liệu hợp pháp từ các định dạng độc quyền đóng, chia nhỏ thông tin từ định dạng kết hợp CAD (BIM) thành các loại dữ liệu và định dạng cần thiết cho người dùng, đơn giản hóa việc xử lý và phân tích.

Việc sử dụng reverse engineering và truy cập trực tiếp vào thông tin từ các cơ sở dữ liệu CAD làm cho thông tin trở nên dễ tiếp cận, cho phép sử dụng dữ liệu mở và công cụ mở, cũng như phân tích dữ liệu bằng các công cụ tiêu chuẩn, xây dựng báo cáo, hình ảnh hóa và tích hợp với các hệ thống kỹ thuật số khác. -



Hình 4.112 Truy cập trực tiếp vào dữ liệu CAD giúp giảm thiểu sự phụ thuộc vào các nền tảng phần mềm và chuyển sang cách tiếp cận dựa trên dữ liệu.

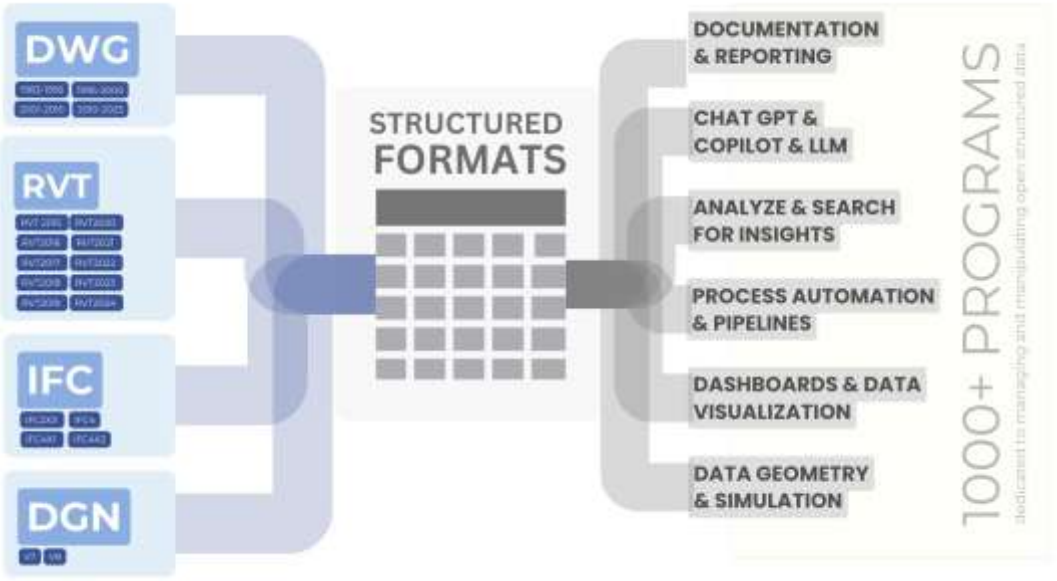
Kể từ năm 1996, đối với định dạng DWG, từ năm 2008 đối với định dạng DGN và từ năm 2018 đối với RVT, việc chuyển đổi tiện lợi và hiệu quả các định dạng CAD dữ liệu ban đầu đóng sang bất kỳ định dạng nào khác, bao gồm cả các định dạng có cấu trúc, đã trở thành khả thi nhờ các công cụ reverse engineering. Ngày nay, hầu hết các công ty CAD (BIM) lớn và các công ty kỹ thuật lớn trên thế giới đều sử dụng các công cụ SDK reverse engineering để trích xuất dữ liệu từ các định dạng đóng của nhà cung cấp CAD (BIM).-



Hình 4.113 Việc sử dụng các công cụ reverse engineering cho phép chuyển đổi cơ sở dữ liệu CAD (BIM) của chương trình thành bất kỳ mô hình dữ liệu nào tiện lợi.

Chuyển đổi dữ liệu từ các định dạng đóng, độc quyền sang các định dạng mở và tách các định dạng CAD (BIM) hỗn hợp thành dữ liệu thuộc tính hình học và siêu thông tin giúp đơn giản hóa quy trình làm việc với chúng, làm cho chúng dễ dàng cho việc phân tích, thao tác và tích hợp với các hệ thống khác (Hình 4.114).-

> Trong công việc hiện đại liên quan đến dữ liệu CAD (BIM), chúng ta đã đạt đến mức độ mà việc truy cập thông tin từ các định dạng CAD không còn cần phải xin phép từ các nhà cung cấp CAD (BIM).



Hình 4.114 Các công cụ SDK hiện đại cho phép chuyển đổi dữ liệu từ các định dạng độc quyền của cơ sở dữ liệu CAD (BIM) một cách hợp pháp.

Các xu hướng hiện đại trong việc xử lý dữ liệu CAD dự án tiếp tục được hình thành dưới ảnh hưởng của các nhân tố chủ chốt trên thị trường — các nhà cung cấp CAD, những người đang làm việc để củng cố vị thế của họ trong thế giới dữ liệu và tạo ra các định dạng và khái niệm mới.

Các nhà cung cấp giải pháp CAD chuyển sang dữ liệu có cấu trúc

Bắt đầu từ năm 2024, lĩnh vực thiết kế và xây dựng đang trải qua sự chuyển mình công nghệ đáng kể trong việc sử dụng và xử lý dữ liệu. Thay vì việc truy cập tự do vào dữ liệu dự án, các nhà sản xuất hệ thống CAD đang tập trung vào việc thúc đẩy các khái niệm mới. Những cách tiếp cận như BIM (được tạo ra vào năm 2002) và open BIM (được tạo ra vào năm 2012) đang dần nhường chỗ cho các giải pháp công nghệ hiện đại mà các nhà cung cấp CAD bắt đầu thúc đẩy:

- Chuyển đổi sang việc sử dụng dữ liệu "hạt" cho phép quản lý thông tin hiệu quả và chuyển sang phân tích dữ liệu
- Sự xuất hiện của định dạng USD và việc áp dụng phương pháp Entity-component-system (ECS) để tổ chức dữ liệu một cách linh hoạt
- Việc sử dụng trí tuệ nhân tạo tích cực trong việc xử lý dữ liệu, tự động hóa quy trình và phân tích dữ liệu
- Sự phát triển của khả năng tương tác - sự cải thiện trong giao tiếp giữa các chương trình, hệ thống và cơ sở dữ liệu khác nhau

Mỗi khía cạnh này sẽ được đề cập chi tiết hơn trong phần sáu của cuốn sách "CAD và BIM: tiếp thị, thực tế và tương lai của dữ liệu dự án trong xây dựng". Trong chương này, chúng tôi chỉ muốn tóm tắt hướng đi tổng quát của những thay đổi: các nhà cung cấp CAD lớn nhất hiện nay đang cố gắng tái định hình cách thức cấu trúc thông tin dự án. Một trong những sự chuyển mình quan trọng là từ bỏ mô hình lưu trữ tệp tin cổ điển để phù hợp với kiến trúc dữ liệu hạt, tập trung vào phân tích, đảm bảo truy cập liên tục vào các thành phần riêng lẻ của mô hình.

> Nội dung diễn ra cho thấy rằng ngành công nghiệp đang dần từ bỏ những định dạng to lớn, chuyên biệt và tham số, yêu cầu sử dụng các lõi hình học, theo hướng những giải pháp linh hoạt, dễ đọc cho máy và đa năng hơn.

Một trong những động lực thay đổi đó là định dạng USD (Universal Scene Description), ban đầu được phát triển trong ngành đồ họa máy tính, nhưng đã nhận được sự công nhận trong các ứng dụng kỹ thuật nhờ sự phát triển của nền tảng NVIDIA Omniverse (và Isaac Sim) cho mô phỏng và trực quan hóa. Khác với IFC tham số, USD cung cấp cấu trúc đơn giản hơn, cho phép mô tả hình học và thuộc tính của các đối tượng theo định dạng JSON, điều này làm dễ dàng việc xử lý thông tin và tăng tốc độ tích hợp vào các quy trình số. Định dạng mới cho phép lưu trữ hình học (ngoài BREP-NURBS – sẽ được trình bày chi tiết hơn trong phần 6 của cuốn sách) dưới dạng đa giác MESH, và các thuộc tính của đối tượng - trong JSON, điều này làm cho nó trở nên tiện lợi hơn cho các quy trình tự động hóa và làm việc trong các hệ sinh thái đám mây.-

Một số nhà cung cấp CAD và ERP đã sử dụng các định dạng tương tự (ví dụ: NWD, SVF, CP2, CPIXML), tuy nhiên, hầu hết trong số đó vẫn còn đóng và không thể truy cập cho việc sử dụng bên ngoài, điều này hạn chế khả năng tích hợp và tái sử dụng dữ liệu. Trong bối cảnh này, USD có thể đóng vai trò tương tự như DXF đã từng làm - một lựa chọn mở cho các định dạng thương mại độc quyền, như DWG.

General Information			Comparison / Notes
Year of format creation	1991	2016	IFC focuses on construction data, USD on 3D graphics
Creator-developer	TU Munich	Pixar	IFC was founded in Germany, USD in America
Prototypes and predecessors	IGES, STEP	PTEX, DAE, GLTF	IFC evolved from IGES/STEP, USD from PTEX/DAE/GLTF
Initiator in Construction	ADSK	ADSK	ADSK initiated the adoption of both formats in construction
Organizer of the Alliance	ADSK	ADSK	ADSK organized both alliances
Name of the Alliance	bS (IAI)	AOUSD	Different alliances for each format
Year of Alliance Formation	1994	2023	The IFC alliance was formed in 1994, AOUSD for USD in 2023
Promoting in the construction	ADSK and Co	ADSK and Co	ADSK and Co actively promotes both formats in bS (IAI) since the introduction

Purpose and Usage			Comparison / Notes
Purpose	Semantic description and interoperability	Data simplification, visualization unification	IFC for semantics and exchange; USD for simplification and visualization
Goals and Objectives	Interoperability and semantics	Unification for visualization and data processing	IFC focuses on semantics; USD on visualization
Use in Other Industries	Predominantly in construction	In film, games, VR/AR, and now in construction	USD is versatile and used in various fields
Supported Data Types	Geometry, object attributes, metadata	Geometry, shaders, animation, light, and camera	USD supports a wider range of data types suitable for complex visualizations; IFC focuses on construction-specific data

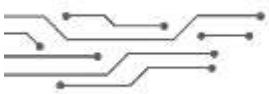
USD định dạng, như một nỗ lực của các nhà cung cấp CAD nhằm đáp ứng yêu cầu về khả năng tương tác và tính độc lập của dữ liệu dự án với các lõi hình học.

> Việc chuyển đổi của các nhà phát triển lớn sang các định dạng mở và đơn giản như USD, GLTF, OBJ, XML (so với những định dạng đóng như NWD, CP2, SVF, SVF2, CPIXML) phản ánh xu hướng toàn cầu và nhu cầu của ngành về việc đơn giản hóa dữ liệu và tăng cường tính khả dụng của chúng. Trong những năm tới, có thể kỳ vọng vào việc dần dần từ bỏ các tiêu chuẩn và định dạng tham số phức tạp, phụ thuộc vào các lõi hình học, để chuyển sang các giải pháp nhẹ hơn và có cấu trúc hơn. Sự chuyển đổi này sẽ thúc đẩy quá trình số hóa ngành xây dựng, dễ dàng hóa tự động hóa quy trình và đơn giản hóa việc trao đổi dữ liệu. -

Mặc dù có các kế hoạch chiến lược của các nhà cung cấp CAD nhằm thúc đẩy các định dạng mở mới, các chuyên gia trong ngành xây dựng vẫn có thể tiếp cận đầy đủ dữ liệu từ các hệ thống CAD đóng mà không cần sử dụng các công cụ CAD (BIM), thông qua các công cụ kỹ thuật ngược.

> Tất cả những xu hướng này chắc chắn dẫn đến việc chuyển từ các mô hình 3D cồng kềnh, khối lượng lớn sang dữ liệu chung, có cấu trúc và việc sử dụng các định dạng đã được chứng minh trong các ngành khác. Khi các đội dự án bắt đầu coi các mô hình CAD không chỉ là các đối tượng hình ảnh hay một tập hợp các tệp, mà là các cơ sở dữ liệu chứa kiến thức và thông tin, thì cách tiếp cận thiết kế và quản lý sẽ thay đổi một cách căn bản.

Sau khi các đội ngũ học cách trích xuất dữ liệu có cấu trúc từ tài liệu, văn bản, bản vẽ và mô hình CAD, và tiếp cận cơ sở dữ liệu, bước quan trọng tiếp theo là mô hình hóa dữ liệu và đảm bảo chất lượng của chúng. Chính từ bước này mà tốc độ xử lý và chuyển đổi thông tin sẽ phụ thuộc rất nhiều, thông tin này cuối cùng sẽ được sử dụng để đưa ra quyết định trong các nhiệm vụ ứng dụng cụ thể.



CHƯƠNG 4.2. PHÂN LOẠI VÀ TÍCH HỢP: NGÔN NGỮ THỐNG NHẤT CHO DỮ LIỆU XÂY DỰNG

Tốc độ ra quyết định phụ thuộc vào chất lượng dữ liệu

Kiến trúc dữ liệu hiện đại đang trải qua những thay đổi cơ bản. Ngành công nghiệp đang chuyển mình từ những mô hình cồng kềnh, biệt lập và các định dạng đóng sang những cấu trúc linh hoạt hơn, có thể đọc được bởi máy móc, chú trọng vào phân tích, tích hợp và tự động hóa các quy trình. Tuy nhiên, việc chuyển sang các định dạng mới không tự nó đảm bảo hiệu quả - chất lượng của chính dữ liệu trở thành yếu tố trung tâm.

> Trên các trang của cuốn sách này, chúng tôi nói rất nhiều về các định dạng, hệ thống và quy trình. Nhưng tất cả những nỗ lực này đều trở nên vô nghĩa nếu thiếu một yếu tố then chốt - đó là dữ liệu đáng tin cậy. Chất lượng dữ liệu là nền tảng của quá trình số hóa, và chúng tôi sẽ trở lại với chủ đề này trong suốt những phần tiếp theo.

Các công ty xây dựng hiện đại - đặc biệt là các công ty lớn - sử dụng hàng chục, thậm chí hàng nghìn hệ thống và cơ sở dữ liệu khác nhau. Những hệ thống này không chỉ cần được cập nhật thông tin mới một cách thường xuyên mà còn phải tương tác hiệu quả với nhau. Tất cả các dữ liệu mới được hình thành từ việc xử lý thông tin đầu vào được tích hợp vào các môi trường này và phục vụ cho việc giải quyết các nhiệm vụ kinh doanh cụ thể.

Nếu như trước đây, quyết định cho các nhiệm vụ kinh doanh cụ thể thường được đưa ra bởi các nhà lãnh đạo cấp cao - hay còn gọi là HiPPO - dựa trên kinh nghiệm và trực giác, thì ngày nay, trong bối cảnh tăng trưởng đáng kể về khối lượng thông tin, cách tiếp cận này trở nên gây tranh cãi. Thay vào đó, phân tích tự động, làm việc với dữ liệu theo thời gian thực đang dần thay thế.

> Cuộc thảo luận "truyền thống - thủ công" về các quy trình kinh doanh ở cấp độ lãnh đạo sẽ chuyển mình theo hướng phân tích hoạt động, yêu cầu những phản hồi nhanh chóng cho các câu hỏi kinh doanh.

Thời kỳ mà các kế toán, chỉ huy công trình và nhân viên lập dự toán phải thủ công hình thành báo cáo và bảng tổng hợp, cũng như các bảng dữ liệu dự án trong vài ngày hay vài tuần đã qua đi. Ngày nay, tốc độ và tính kịp thời trong việc ra quyết định trở thành yếu tố then chốt tạo ra lợi thế cạnh tranh.



Trong ngành xây dựng, việc tính toán và ra quyết định mất vài ngày, khác với các ngành khác, nơi điều này diễn ra chỉ trong vài giờ hoặc vài phút.

Sự khác biệt chính giữa ngành xây dựng và các ngành có sự phát triển kỹ thuật số hơn là ở mức độ chất lượng và chuẩn hóa dữ liệu thấp. Các phương pháp lạc hậu trong việc hình thành, chuyển giao và xử lý thông tin làm chậm lại các quy trình và tạo ra sự hỗn loạn. Việc thiếu các tiêu chuẩn chất lượng thống nhất cản trở việc thực hiện tự động hóa toàn diện.

> Một trong những vấn đề lớn nhất vẫn là chất lượng dữ liệu đầu vào thấp, cũng như thiếu quy trình chuẩn hóa cho việc chuẩn bị và kiểm tra dữ liệu. Không có dữ liệu đáng tin cậy và đồng thuận, việc tích hợp hiệu quả giữa các hệ thống là không thể. Điều này dẫn đến sự chậm trễ, sai sót và tăng chi phí ở mỗi giai đoạn trong vòng đời của dự án.

Trong các phần sau của cuốn sách, chúng tôi sẽ xem xét chi tiết cách nâng cao chất lượng dữ liệu, chuẩn hóa quy trình và rút ngắn thời gian từ việc thu thập thông tin đến khi có dữ liệu chất lượng, đã được kiểm tra và đồng thuận.

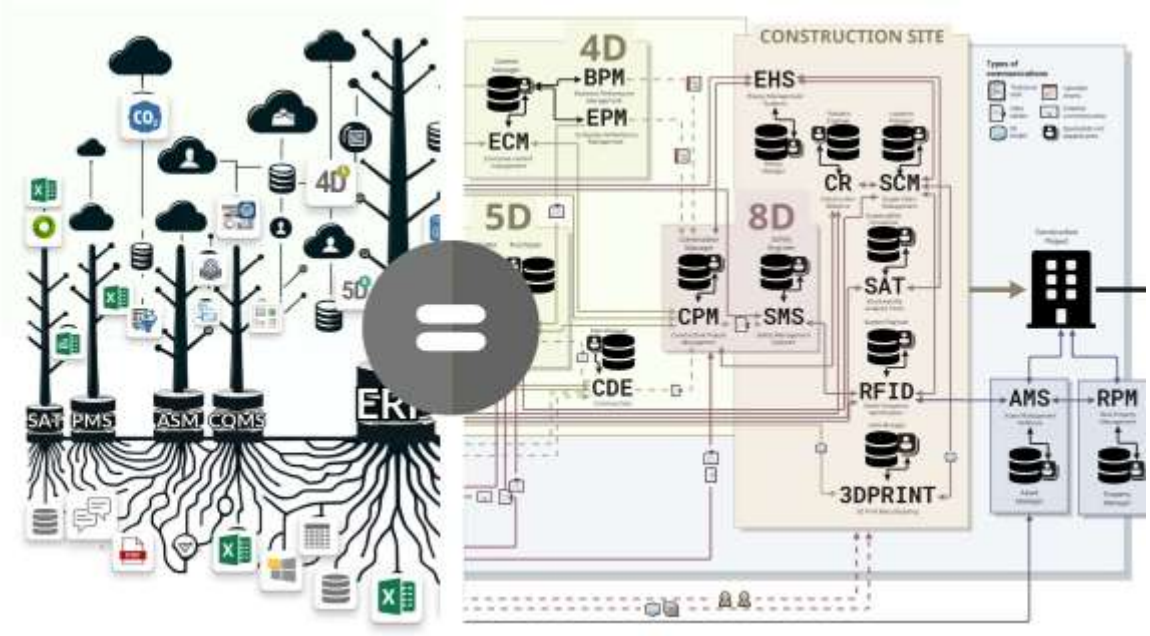
Chuẩn hóa và tích hợp dữ liệu

Việc làm việc hiệu quả với dữ liệu đòi hỏi một chiến lược chuẩn hóa rõ ràng. Chỉ khi có các yêu cầu rõ ràng về cấu trúc và chất lượng dữ liệu, chúng ta mới có thể tự động hóa việc kiểm tra, giảm thiểu số lượng thao tác thủ công và tăng tốc độ ra quyết định dựa trên thông tin đúng đắn ở mọi giai đoạn của dự án.

Trong thực tiễn hàng ngày của công ty xây dựng, họ phải xử lý hàng trăm tệp tin: email, tài liệu PDF, tệp CAD, dữ liệu từ cảm biến IoT mà cần phải tích hợp vào quy trình kinh doanh của công ty.

Hệ sinh thái công ty được cấu thành từ các cơ sở dữ liệu và công cụ (Hình 4.22) cần biết cách hấp thụ các dưỡng chất từ dữ liệu đa định dạng đầu vào để đạt được kết quả mong muốn cho công ty.

Để xử lý hiệu quả dòng dữ liệu, không nhất thiết phải thuê một đội ngũ quản lý đông đảo, mà trước tiên cần phát triển các yêu cầu và tiêu chuẩn nghiêm ngặt cho dữ liệu, đồng thời sử dụng các công cụ phù hợp để kiểm tra, chuẩn hóa và xử lý tự động.



Hình 4.22 Việc duy trì hoạt động khỏe mạnh của hệ sinh thái công ty đòi hỏi cung cấp kịp thời và chất lượng các nguồn lực cho hệ thống của nó.

Để tự động hóa quy trình kiểm tra dữ liệu và chuẩn hóa (để tích hợp tự động sau này), cần bắt đầu bằng cách mô tả các yêu cầu tối thiểu cần có cho dữ liệu trong từng hệ thống cụ thể. Những yêu cầu này xác định:

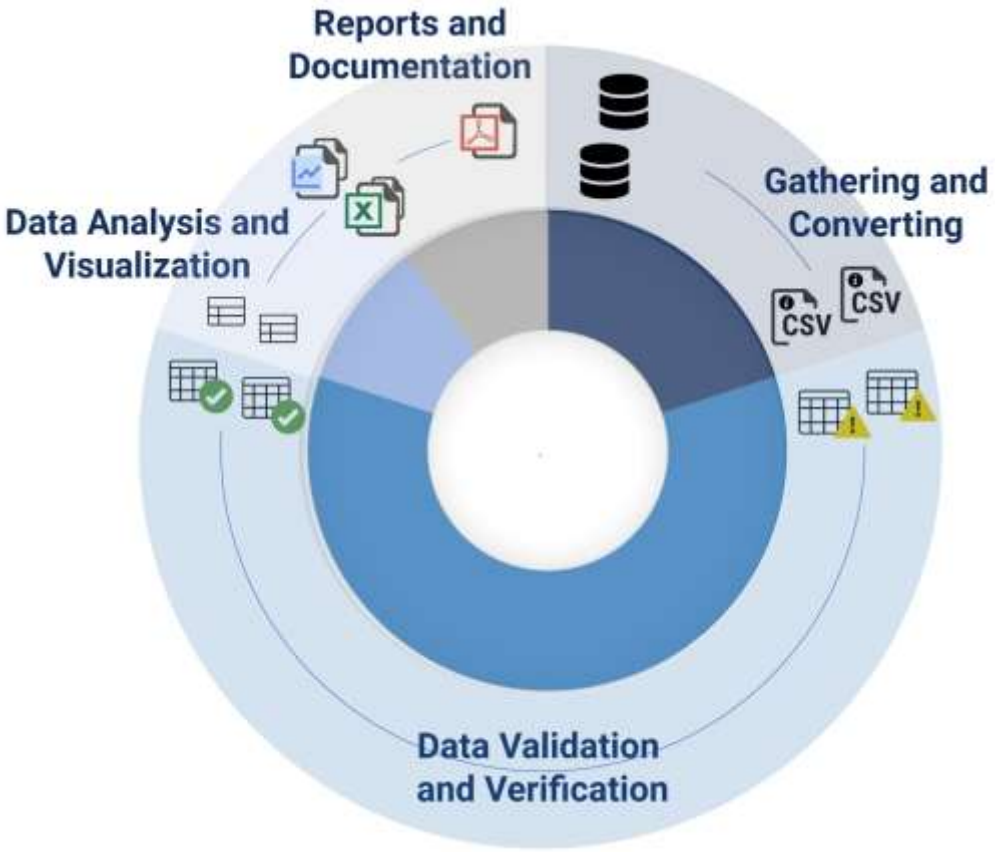
- Cần thu thập gì?
- Dưới dạng nào (cấu trúc, định dạng)?
- Các thuộc tính nào là bắt buộc?
- Các độ sai lệch về độ chính xác và sự đầy đủ nào là chấp nhận được?

Yêu cầu về dữ liệu mô tả các tiêu chí chất lượng, cấu trúc và độ đầy đủ của thông tin thu thập và xử lý. Chẳng hạn, đối với văn bản trong tài liệu PDF, việc trình bày chính xác theo tiêu chuẩn ngành là rất quan trọng (Hình 7.214 - Hình 7.216). Các đối tượng trong mô hình CAD cần có các thuộc tính chính xác (kích thước, mã, liên kết với bộ phân loại) (Hình 7.39, Hình 7.310). Còn đối với các bản quét hợp đồng, các ngày tháng rõ ràng và khả năng trích xuất tự động số tiền và các điều khoản quan trọng là rất cần thiết (Hình 4.17 - Hình 4.110).------

Việc hình thành các yêu cầu về dữ liệu và kiểm tra tự động độ phù hợp của chúng là một trong những bước tốn công nhất, nhưng cũng quan trọng nhất. Đây chính là bước thường chiếm phần lớn thời gian trong quy trình kinh doanh.

Như đã đề cập trong phần ba của cuốn sách, từ 50% đến 90% thời gian làm việc của các chuyên gia phân tích kinh doanh (BI) không dành cho việc phân tích, mà dành cho việc chuẩn bị dữ liệu (Hình 3.25). Quy trình này bao gồm thu thập, xác minh, xác thực, chuẩn hóa và cấu trúc dữ liệu.

Theo một cuộc khảo sát vào năm 2016 [95], các chuyên gia xử lý dữ liệu trong nhiều lĩnh vực khác nhau cho biết họ dành phần lớn thời gian làm việc (khoảng 80%) cho những nhiệm vụ mà họ ít thích nhất: thu thập các bộ dữ liệu hiện có và tổ chức (chuẩn hóa, cấu trúc) chúng. Do đó, họ chỉ còn dưới 20% thời gian cho các nhiệm vụ sáng tạo, chẳng hạn như tìm kiếm quy luật và mẫu, có thể dẫn đến những hiểu biết và phát hiện mới.



Hình 4.23 Kiểm tra và đảm bảo chất lượng dữ liệu là giai đoạn tốn kém, mất thời gian và phức tạp nhất trong việc chuẩn bị dữ liệu để tích hợp vào các hệ thống khác.

Quản lý dữ liệu thành công trong một công ty xây dựng đòi hỏi một cách tiếp cận toàn diện, bao gồm đặc tả các nhiệm vụ, hình thành yêu cầu về chất lượng dữ liệu và sử dụng các công cụ phù hợp để kiểm tra tự động.

Tính tương thích số bắt đầu từ yêu cầu

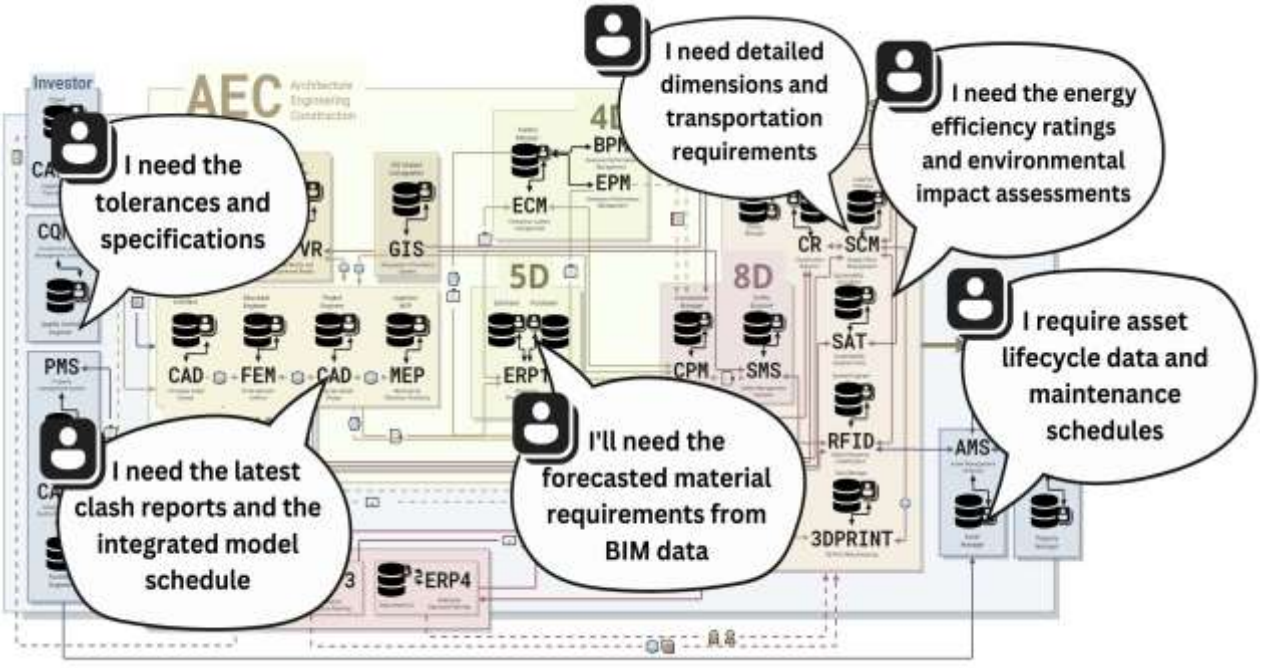
Với sự gia tăng của số lượng hệ thống số trong các công ty, nhu cầu về tính nhất quán dữ liệu giữa chúng cũng tăng lên. Các nhà quản lý phụ trách các hệ thống CNTT khác nhau thường gặp khó khăn trong việc theo kịp với khối lượng thông tin ngày càng tăng và sự đa dạng của các định dạng. Trong những điều kiện như vậy, họ buộc phải yêu cầu các chuyên gia tạo ra dữ liệu dưới hình thức phù hợp để sử dụng trong các

ứng dụng và nền tảng khác.

Điều này, một mặt, yêu cầu các kỹ sư và nhân viên phụ trách tạo lập dữ liệu phải điều chỉnh theo nhiều yêu cầu, thường mà không có sự minh bạch và hiểu biết rõ ràng về nơi và cách mà những dữ liệu này sẽ được áp dụng trong tương lai. Việc thiếu các phương pháp chuẩn hóa trong công việc với thông tin dẫn đến tổn thất về hiệu suất và tăng chi phí trong giai đoạn kiểm tra, thường diễn ra một cách thủ công do sự phức tạp và tính không chuẩn hóa của dữ liệu.

> Vấn đề chuẩn hóa dữ liệu không chỉ là một vấn đề về sự tiện lợi hay tự động hóa. Nó còn liên quan đến những tổn thất tài chính trực tiếp. Theo báo cáo của IBM năm 2016, hàng năm, các tổn thất do chất lượng dữ liệu kém ở Mỹ lên đến 3,1 triệu tỷ đô la. Thêm vào đó, các nghiên cứu của MIT và các công ty tư vấn phân tích khác cho thấy, chi phí do chất lượng dữ liệu kém có thể chiếm từ 15-25% doanh thu của công ty.

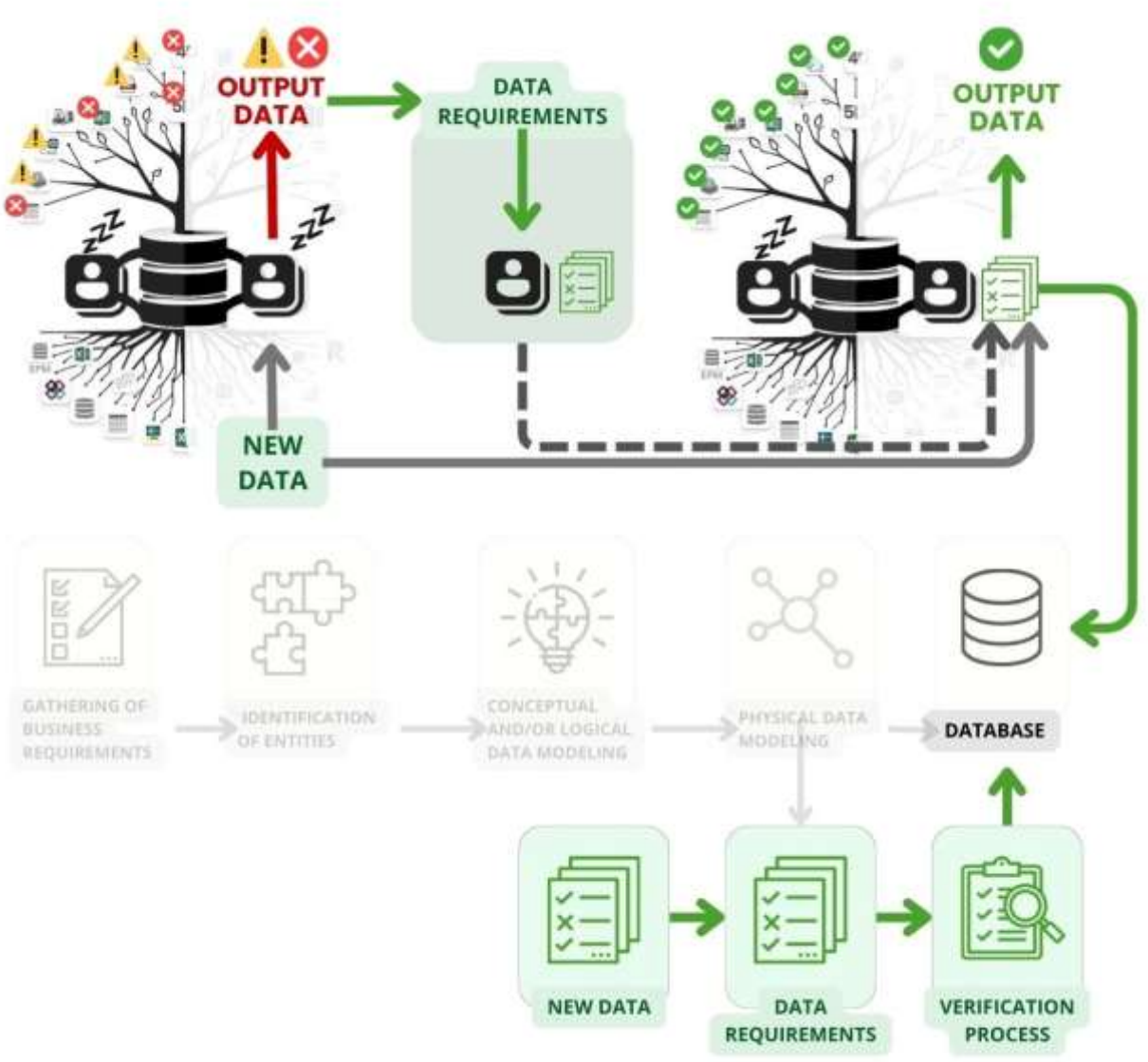
Trong những điều kiện như vậy, việc có các yêu cầu về dữ liệu được định hình rõ ràng và mô tả các tham số, ở định dạng nào và với mức độ chi tiết nào phải được bao gồm trong các đối tượng được tạo ra là vô cùng quan trọng. Nếu không có sự hình thức hóa những yêu cầu này, sẽ không thể đảm bảo chất lượng và tính tương thích của dữ liệu giữa các hệ thống và các giai đoạn của dự án.



Hình 4.24 Kinh doanh dựa trên sự tương tác giữa các vai trò khác nhau, mỗi vai trò yêu cầu những tham số và giá trị nhất định, rất quan trọng để thực hiện các nhiệm vụ kinh doanh.

Để hình thành các yêu cầu chính xác cho dữ liệu, cần phải hiểu các quy trình kinh doanh ở cấp độ dữ liệu. Các dự án xây dựng khác nhau về loại hình, quy mô và số lượng người tham gia, và mỗi hệ thống - bất kể là mô hình hóa (CAD (BIM)), lập kế hoạch lịch (ERP 4D), tính toán chi phí (ERP 5D) hay logistics (SCM) - đều yêu cầu các tham số độc đáo cho các thực thể đầu vào.

Tùy thuộc vào những nhu cầu này, các nhà quản lý kinh doanh phải thiết kế cấu trúc dữ liệu mới theo các yêu cầu đã quy định, hoặc điều chỉnh các bảng và cơ sở dữ liệu hiện có. Chất lượng của dữ liệu được tạo ra sẽ phụ thuộc trực tiếp vào mức độ chính xác và đúng đắn trong việc hình thành các yêu cầu.



Hình 4.25 Chất lượng dữ liệu phụ thuộc vào chất lượng của các yêu cầu được tạo ra cho các trường hợp sử dụng dữ liệu cụ thể.

Bởi vì mỗi hệ thống có những yêu cầu cụ thể của riêng mình đối với dữ liệu, bước đầu tiên trong việc hình thành các yêu cầu chung nên là phân loại tất cả các yếu tố tham gia vào các quy trình kinh doanh. Điều này có nghĩa là cần phải chia tách các đối tượng thành các loại và nhóm loại tương ứng với các hệ thống hoặc nhiệm vụ ứng dụng cụ thể. Đối với mỗi nhóm như vậy, các yêu cầu riêng được phát triển cho cấu trúc, thuộc tính và chất lượng dữ liệu.

Tuy nhiên, trên thực tế, việc triển khai cách tiếp cận này gặp phải một khó khăn nghiêm trọng: sự thiếu hụt ngôn ngữ chung để phân nhóm dữ liệu. Các phân loại rời rạc, việc trùng lặp định danh và sự không tương thích về định dạng dẫn đến việc mỗi công ty, mỗi phần mềm và thậm chí mỗi dự án đều hình thành các mô hình dữ liệu và lớp riêng biệt, cô lập. Kết quả là một "tháp Babel số" xuất hiện, nơi cần có nhiều chuyển đổi sang các mô hình dữ liệu và lớp phù hợp để truyền tải thông tin giữa các hệ thống, thường được thực hiện thủ công. Chỉ có thể vượt qua rào cản này thông qua việc chuyển sang các phân loại phổ quát và bộ yêu cầu

chuẩn hóa.

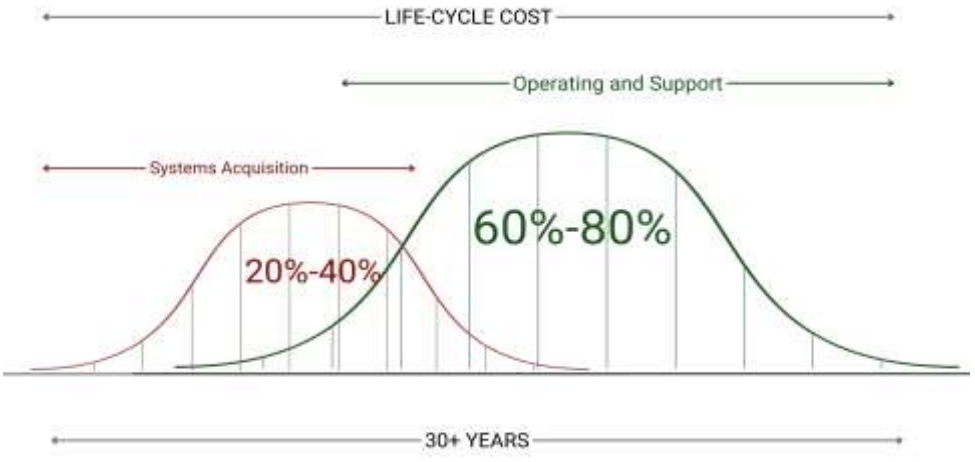
Ngôn ngữ thống nhất xây dựng: vai trò của các phân loại trong chuyển đổi số

Trong bối cảnh số hóa và tự động hóa các quy trình kiểm tra và xử lý, hệ thống phân loại các yếu tố đóng vai trò đặc biệt quan trọng - những "từ điển số" cung cấp sự đồng nhất trong việc mô tả và định tham số cho các đối tượng. Chính các phân loại viên hình thành "ngôn ngữ chung" cho phép nhóm dữ liệu theo ý nghĩa sử dụng và tích hợp dữ liệu giữa các hệ thống khác nhau, các cấp quản lý và các giai đoạn trong vòng đời dự án.

Ảnh hưởng rõ rệt nhất của các bộ phân loại là trong kinh tế vòng đời của tòa nhà, nơi khía cạnh quan trọng nhất là tối ưu hóa chi phí vận hành dài hạn. Các nghiên cứu chỉ ra rằng chi phí vận hành chiếm tới 80% tổng chi phí sở hữu tòa nhà, gấp ba lần chi phí ban đầu cho việc xây dựng (Hình 4.26) [98]. Điều này có nghĩa là quyết định về các chi phí trong tương lai chủ yếu được hình thành ngay từ giai đoạn thiết kế.

> Chính vì lý do đó, các yêu cầu từ các kỹ sư vận hành (CAFM, AMS, PMS, RPM) phải trở thành điểm khởi đầu trong việc xác định yêu cầu về dữ liệu trong giai đoạn thiết kế. Những hệ thống này cần được coi không phải là giai đoạn cuối cùng của dự án, mà là một phần không thể thiếu của toàn bộ hệ sinh thái số của dự án — từ khái niệm đến việc tháo gỡ.-

Phân loại hiện đại không chỉ là một hệ thống mã để nhóm lại. Đây là một cơ chế hiểu biết lẫn nhau giữa các kiến trúc sư, kỹ sư, nhân viên dự toán, nhân viên logistics, các dịch vụ vận hành và các hệ thống CNTT. Tương tự như cách mà hệ thống lái tự động trong ô tô phải nhận diện các đối tượng trên đường một cách chính xác, các hệ thống xây dựng kỹ thuật số và người dùng của chúng cần phải thông qua lớp phần tử để giải thích chính xác cùng một yếu tố dự án cho các hệ thống khác nhau.

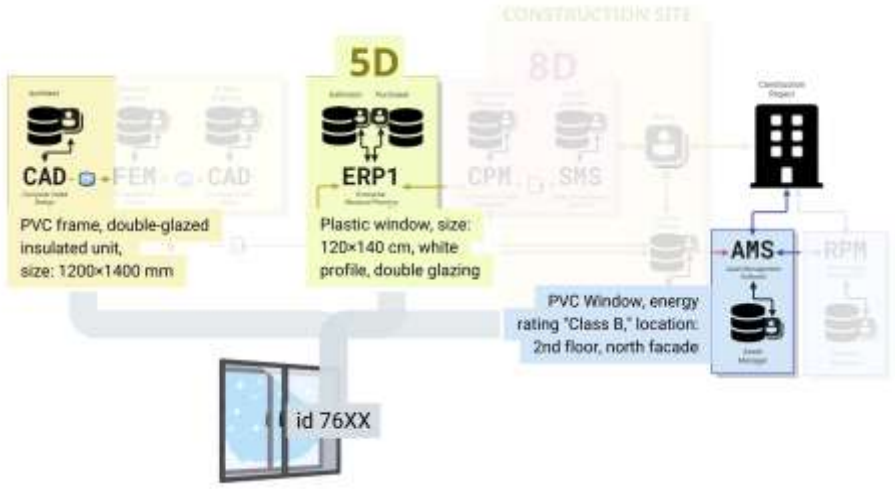


Hình 4.26 Chi phí vận hành và kỹ thuật vượt quá chi phí xây dựng gấp ba lần, chiếm từ 60–80% tổng chi phí vòng đời của tòa nhà (theo tài liệu [99]).

Mức độ phát triển của các bộ phân loại liên quan trực tiếp đến độ sâu số hóa của công ty và sự trưởng thành kỹ thuật số của nó. Các tổ chức có mức độ trưởng thành kỹ thuật số thấp thường gặp phải phân mảnh dữ liệu, sự không tương thích của các hệ thống thông tin và do đó, sự không tương thích và kém hiệu quả của các bộ phân loại. Trong những công ty như vậy, cùng một yếu tố có thể thường xuyên có các mã định

danh nhóm khác nhau trong các hệ thống khác nhau, điều này gây khó khăn nghiêm trọng cho việc tích hợp cuối cùng và làm cho việc tự động hóa quy trình trở nên không khả thi.

Ví dụ, cùng một cửa sổ trong dự án có thể được ký hiệu khác nhau trong mô hình CAD, hệ thống dự toán và hệ thống vận hành (Hình 4.27) do sự đa khía cạnh trong cách mà các yếu tố được nhận thức bởi các bên tham gia khác nhau trong quy trình. Đối với người lập dự toán, trong yếu tố thể loại "Cửa sổ", các khối lượng và chi phí là quan trọng; đối với bộ phận bảo trì, tính khả dụng và khả năng sửa chữa là quan trọng; đối với kiến trúc sư, các đặc tính thẩm mỹ và chức năng là điều cần chú ý. Kết quả là, cùng một yếu tố có thể yêu cầu các thông số khác nhau.-



Hình 4.27 Khi có sự phân loại không đồng nhất giữa các hệ thống, yếu tố sẽ mất một phần thông tin thuộc tính khi chuyển sang hệ thống khác ở mỗi giai đoạn.

Do sự phức tạp trong việc xác định rõ ràng phân loại các yếu tố xây dựng, các chuyên gia đến từ các lĩnh vực khác nhau thường gán cho cùng một yếu tố các lớp không tương thích với nhau. Điều này dẫn đến việc mất đi hình ảnh thống nhất về đối tượng, làm cho việc can thiệp thủ công sau đó cần thiết để đồng bộ hóa các hệ thống phân loại khác nhau và xác lập sự tương ứng giữa các loại và lớp được xác định bởi các chuyên gia khác nhau.

Kết quả của sự không thống nhất như vậy, tài liệu về vận hành, được thu thập bởi bộ phận mua sắm (ERP) khi mua một yếu tố xây dựng từ nhà sản xuất, thường không thể gắn chính xác với phân loại của yếu tố đó tại công trường (PMIS, SCM). Hệ quả là, thông tin quan trọng có khả năng cao không được tích hợp vào các hệ thống quản lý cơ sở hạ tầng và tài sản (CAFM, AMS), điều này tạo ra các vấn đề nghiêm trọng khi đưa đối tượng vào sử dụng, cũng như khi bảo trì tiếp theo (AMS, RPM) hoặc thay thế yếu tố đó.

> Tại các công ty có độ trưởng thành số cao, các bộ phân loại đóng vai trò như một hệ thống thần kinh, liên kết tất cả các luồng thông tin. Cùng một yếu tố nhận được một mã định danh duy nhất, cho phép nó được truyền giữa các hệ thống CAD, ERP, AMS và CAFM cùng với các bộ phân loại của chúng mà không bị biến dạng hoặc mất mát.

Để xây dựng các bộ phân loại hiệu quả, cần hiểu cách mà dữ liệu được sử dụng. Cùng một kỹ sư có thể gọi tên và phân loại một yếu tố khác nhau trong các dự án khác nhau. Chỉ bằng cách thu thập thống kê sử dụng trong nhiều năm, mới có thể phát triển một hệ thống phân loại ổn định. Điều này được hỗ trợ bởi học máy: các thuật toán phân tích hàng nghìn dự án (Hình 9.110), xác định qua học máy các lớp và thông số có khả

năng (Hình 10.16). Phân loại tự động đặc biệt có giá trị trong các điều kiện mà phân loại thủ công là không khả thi do khối lượng dữ liệu. Các hệ thống phân loại tự động có thể phân biệt các loại cơ bản dựa trên các thông số tối thiểu của yếu tố (chi tiết hơn trong phần chín và mười của sách). --

Các hệ thống phân loại phát triển trở thành chất xúc tác cho quá trình số hóa tiếp theo, tạo ra nền tảng cho:

- Đánh giá tự động về chi phí và thời gian thực hiện các dự án.
- Phân tích dự đoán về các rủi ro và xung đột tiềm ẩn.
- Tối ưu hóa quy trình mua sắm và chuỗi cung ứng.
- Tạo ra các bản sao kỹ thuật số của các tòa nhà và công trình.
- Tích hợp với các hệ thống "thành phố thông minh" và Internet vạn vật.

> Thời gian để chuyển đổi là có hạn — với sự phát triển của công nghệ học máy và thị giác máy tính, vấn đề phân loại tự động, điều mà không thể giải quyết trong hàng thập kỷ, sẽ được giải quyết trong những năm tới, và các công ty xây dựng và thiết kế, những người không kịp thích nghi, có nguy cơ gặp phải số phận giống như các hãng taxi bị loại bỏ bởi các nền tảng kỹ thuật số.

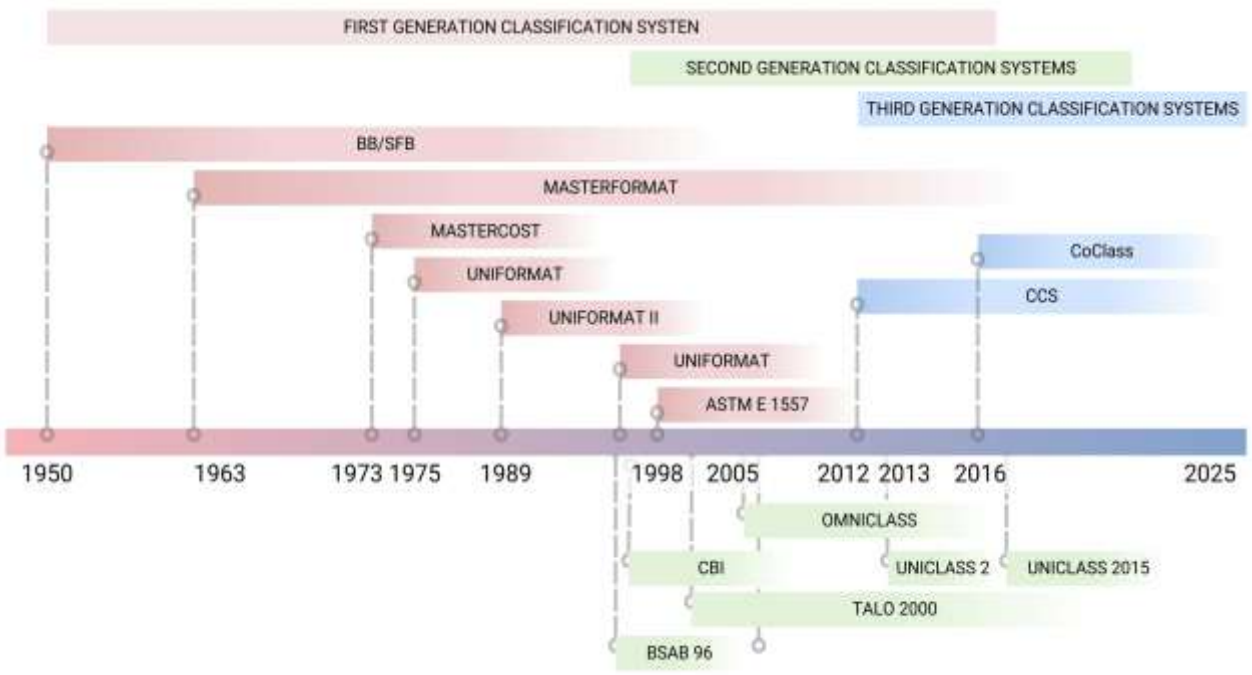
Thông tin chi tiết về tự động hóa trong việc tính toán chi phí và thời gian, cũng như về dữ liệu lớn và học máy sẽ được trình bày trong phần năm và phần chín của cuốn sách. Các vấn đề về rủi ro lặp lại số phận của các hãng taxi và sự "uber hóa" trong ngành xây dựng sẽ được xem xét kỹ lưỡng trong phần mười của cuốn sách.

Hiểu được vai trò then chốt của các bộ phân loại trong quá trình chuyển đổi số của ngành xây dựng, cần phải quay ngược về lịch sử tiến hóa của chúng. Chính ngữ cảnh lịch sử cho phép nhận thức được cách các phương pháp phân loại đã phát triển và những xu hướng nào đang xác định trạng thái hiện tại của chúng.

Masterformat, OmniClass, Uniclass và CoClass: sự tiến hóa của hệ thống phân loại

Về mặt lịch sử, các bộ phân loại cho các yếu tố và công việc xây dựng đã phát triển qua ba thế hệ, mỗi thế hệ phản ánh mức độ công nghệ sẵn có và nhu cầu cấp thiết của ngành trong một khoảng thời gian nhất định.

- Thế hệ đầu tiên (đầu những năm 1950 - cuối những năm 1980) — các thư mục giấy, các bộ phân loại theo cấu trúc hierarchic, được sử dụng tại chỗ (ví dụ: Masterformat, SfB).
- Thế hệ thứ hai (cuối những năm 1990 - giữa những năm 2010) — các bảng và cơ sở dữ liệu có cấu trúc, được triển khai trong Excel và Access (ASTM E 1557, OmniClass, Uniclass 1997).
- Thế hệ thứ ba (từ những năm 2010 đến nay) — các dịch vụ kỹ thuật số và API, tích hợp với CAD (BIM), tự động hóa (Uniclass 2015, CoClass).



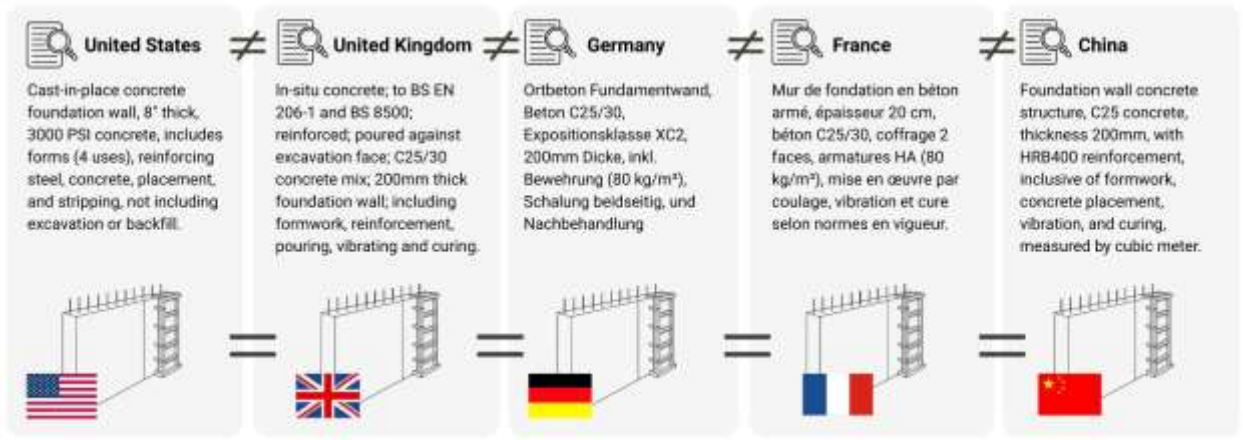
Hình. 4.28 Ba thế hệ bộ phân loại trong ngành xây dựng.

Trong vài thập kỷ qua, đã có sự giảm thiểu độ phức tạp cấu trúc (Hình. 4.29) của các bộ phân loại: nếu các hệ thống đầu tiên, như OmniClass, sử dụng đến 7 cấp độ lồng ghép để mô tả 6887 lớp, thì các giải pháp hiện đại, chẳng hạn như CoClass, giới hạn ở 3 cấp độ với 750 lớp. Điều này làm cho công việc với dữ liệu trở nên dễ dàng hơn, đồng thời vẫn giữ được độ chi tiết cần thiết. Uniclass 2015, thường được sử dụng ở Vương quốc Anh như một tiêu chuẩn, kết hợp 7210 lớp trong chỉ 4 cấp độ, khiến nó trở nên thuận tiện cho các dự án CAD và các cuộc thầu nhà nước.

Classifier	Table / Objects	Number of classes	Nesting depth
OmniClass	Table 23 Products	6887	7 levels
Uniclass 2015	Pr — Products	7210	4 levels
CoClass, CCS	Components	750	3 levels

Hình. 4.29 Với mỗi thế hệ bộ phân loại mới, độ phức tạp của việc phân loại giảm xuống đáng kể.

Trong các hệ thống đánh giá chi phí xây dựng của các quốc gia khác nhau, do sự khác biệt trong phân loại, ngay cả một yếu tố tiêu chuẩn như tường bê tông móng cũng có thể được mô tả hoàn toàn khác nhau (Hình. 4.210). Những khác biệt này phản ánh những đặc điểm dân tộc trong thực hành xây dựng, các hệ thống đo lường được áp dụng, các phương pháp phân loại vật liệu, cũng như các yêu cầu quy định và kỹ thuật hiện hành ở mỗi quốc gia.



Hình. 4.210 Một yếu tố giống nhau ở các quốc gia khác nhau được sử dụng trong các dự án thông qua các mô tả và phân loại khác nhau.

Sự đa dạng của các phân loại các yếu tố giống nhau làm phức tạp sự hợp tác quốc tế, khiến cho việc so sánh chi phí và khối lượng công việc trong các dự án quốc tế trở nên tốn kém và đôi khi là không thể. Hiện tại, ở cấp độ toàn cầu không có một phân loại phổ quát nào - mỗi quốc gia hoặc khu vực phát triển hệ thống riêng của mình, dựa trên các tiêu chuẩn địa phương, ngôn ngữ và văn hóa kinh doanh.

- CCS (Đan Mạch): Hệ thống phân loại chi phí - hệ thống phân loại chi phí trong suốt vòng đời của một đối tượng (thiết kế, xây dựng, vận hành). Tập trung vào logic vận hành và bảo trì, nhưng cũng bao gồm quản lý ngân sách và tài nguyên.
- NS 3451 (Na Uy): phân loại các đối tượng theo chức năng, các yếu tố cấu trúc và giai đoạn vòng đời. Được sử dụng cho quản lý dự án, đánh giá chi phí và lập kế hoạch dài hạn.
- MasterFormat (Mỹ): hệ thống để cấu trúc các thông số xây dựng theo các phần (ví dụ: bê tông, lắp đặt điện, hoàn thiện). Tập trung vào các lĩnh vực và loại công việc, chứ không phải các yếu tố chức năng (khác với UniFormat).
- Uniclass 2 (Vương quốc Anh): một trong những phân loại chi tiết nhất, được áp dụng trong đấu thầu công và các dự án BIM. Kết hợp dữ liệu về các đối tượng, công việc, vật liệu và không gian thành một hệ thống thống nhất.
- OmniClass: tiêu chuẩn quốc tế (được phát triển bởi CSI ở Mỹ) để quản lý thông tin về các đối tượng: từ thư viện thành phần đến các thông số điện tử. Phù hợp cho việc lưu trữ dữ liệu lâu dài, tương thích với CAD (BIM) và các công cụ số khác.
- COBie: Construction-Operation Building Information Exchange - tiêu chuẩn quốc tế về trao đổi dữ liệu giữa các giai đoạn thiết kế, xây dựng và vận hành. Được bao gồm trong BS 1192–4:2014, như một phần của khái niệm "mô hình BIM sẵn sàng cho vận hành". Tập trung vào việc truyền đạt thông tin (ví dụ: thông số kỹ thuật thiết bị, bảo hành, thông tin liên lạc với nhà thầu).

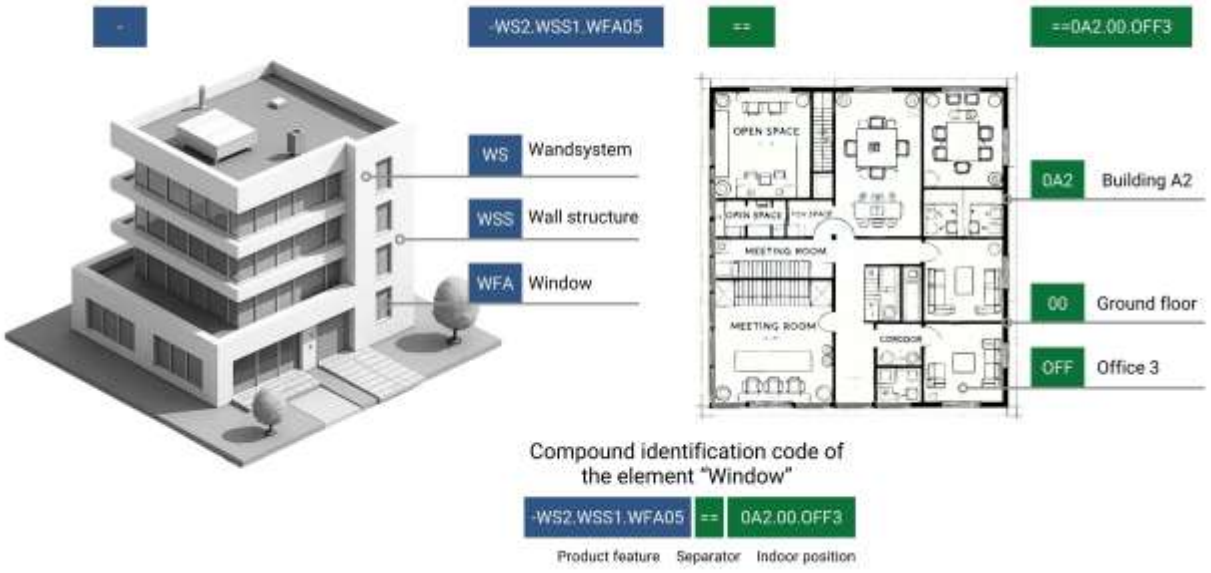
Sự toàn cầu hóa của ngành xây dựng, có khả năng dẫn đến việc dần dần thống nhất các hệ thống phân loại các yếu tố xây dựng, điều này sẽ giảm đáng kể sự phụ thuộc vào các tiêu chuẩn quốc gia địa phương. Quá trình này có thể phát triển theo cách tương tự như sự tiến hóa của các giao tiếp Internet, nơi mà các giao thức truyền dữ liệu phổ quát cuối cùng đã thay thế các định dạng địa phương rời rạc, đảm bảo tính tương thích toàn cầu của các hệ thống.

Một con đường phát triển thay thế có thể là ngay lập tức chuyển sang các hệ thống phân loại tự động dựa trên công nghệ máy học. Những công nghệ này, hiện đang được phát triển chủ yếu trong lĩnh vực giao thông tự động, có tiềm năng đáng kể để ứng dụng cho các khối lượng dữ liệu lớn trong thiết kế CAD.

> Hiện nay, tình huống không chỉ bị giới hạn ở sự phân nhóm phân loại quốc gia. Vì nhiều đặc điểm mà không được tính đến ở cấp quốc gia, mỗi công ty buộc phải tự mình thực hiện việc thống nhất và chuẩn hóa các danh mục yếu tố và tài nguyên mà họ làm việc.

Thông thường, quá trình này bắt đầu từ những cái nhỏ - từ các bảng đối tượng địa phương hoặc hệ thống đặt tên nội bộ. Tuy nhiên, mục tiêu chiến lược trở thành việc chuyển sang một ngôn ngữ thống nhất để mô tả tất cả các yếu tố, mà sẽ được hiểu không chỉ trong công ty mà còn bên ngoài - lý tưởng là được đồng thuận với các phân loại quốc tế hoặc ngành nghề (Hình 4.28). Cách tiếp cận này giúp dễ dàng tích hợp với các đối tác bên ngoài, hệ thống kỹ thuật số và hỗ trợ việc hình thành các quy trình xuyên suốt thống nhất trong vòng đời của các đối tượng.

Trước khi chuyển sang tự động hóa và các hệ thống CNTT có thể mở rộng, cần sử dụng hoặc phân loại ở cấp quốc gia, hoặc xây dựng một cấu trúc xác định các yếu tố một cách hợp lý và rõ ràng. Mỗi đối tượng - cho dù đó là cửa sổ (Hình 4.211), cửa ra vào hay hệ thống kỹ thuật - phải được mô tả theo cách mà nó có thể được nhận diện chính xác trong bất kỳ hệ thống kỹ thuật số nào của công ty. Điều này cực kỳ quan trọng khi chuyển từ bản vẽ phẳng sang các mô hình kỹ thuật số, bao gồm cả giai đoạn thiết kế và vận hành các tòa nhà.



Hình 4.211 Ví dụ về mã định danh thành phần xây dựng cho cửa sổ dựa trên phân loại và vị trí trong tòa nhà.

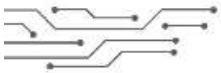
Một trong những ví dụ về phân loại nội bộ có thể là phát triển mã định danh thành phần (Hình 4.211). Mã này kết hợp nhiều cấp thông tin: mục đích chức năng của yếu tố (ví dụ, "cửa sổ trong tường"), loại của nó, cũng như gắn kết không gian chính xác - tòa nhà A2, tầng 0, phòng 3. Cấu trúc đa tầng như vậy cho phép tạo ra một hệ thống dẫn đường duy nhất cho các mô hình kỹ thuật số và tài liệu, đặc biệt là trong các giai đoạn kiểm tra và chuyển đổi dữ liệu, nơi cần phải nhóm các yếu tố một cách rõ ràng. Sự nhận diện chính xác của yếu tố bảo đảm tính nhất quán giữa các phòng ban và giảm thiểu rủi ro về việc lặp lại, sai sót và mất thông tin.

Một phân loại được xây dựng tốt không chỉ là một tài liệu kỹ thuật, mà là nền tảng của hệ sinh thái kỹ thuật số của công ty:

- đảm bảo tính tương thích dữ liệu giữa các hệ thống;
- giảm chi phí tìm kiếm và xử lý thông tin;
- nâng cao tính minh bạch và khả năng quản lý;
- tạo ra cơ sở cho việc mở rộng và tự động hóa.

Mô tả tiêu chuẩn hóa các đối tượng, thông qua việc áp dụng các phân loại quốc gia hoặc mã định danh thành phần riêng của mình, trở thành nền tảng cho dữ liệu nhất quán, trao đổi thông tin đáng tin cậy và sau đó triển khai các dịch vụ thông minh - từ mua sắm tự động đến mô hình kỹ thuật số.

Sau khi hoàn thành giai đoạn cấu trúc hóa dữ liệu khác định dạng và lựa chọn phân loại sẽ được sử dụng để nhận diện và nhóm các yếu tố, bước tiếp theo là mô hình hóa dữ liệu một cách chính xác. Quá trình này bao gồm việc xác định các tham số quan trọng, xây dựng cấu trúc dữ liệu hợp lý và mô tả mối quan hệ giữa



các yếu tố.

CHƯƠNG 4.3.
MÔ HÌNH HÓA DỮ LIỆU VÀ TRUNG TÂM CỦA KINH NGHIỆM HÀNG ĐẦU

Mô hình hóa dữ liệu: mô hình khái niệm, mô hình logic và mô hình vật lý

Quản lý dữ liệu hiệu quả (đã được cấu trúc và phân loại trước đó) không thể thực hiện được mà không có một cấu trúc lưu trữ và xử lý được suy nghĩ kỹ lưỡng. Để đảm bảo quyền truy cập và tính nhất quán của thông tin trong các giai đoạn lưu trữ và xử lý, các công ty sử dụng mô hình hóa dữ liệu - phương pháp cho phép thiết kế bảng, cơ sở dữ liệu và mối quan hệ giữa chúng theo yêu cầu kinh doanh.

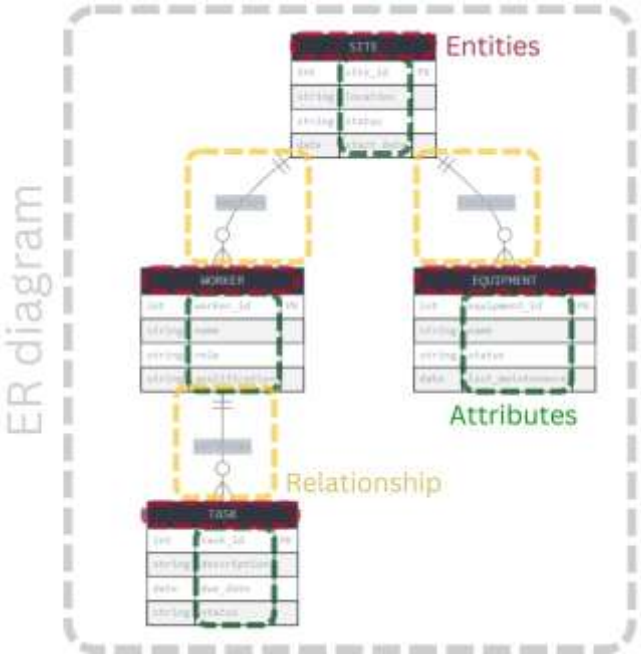
> Mô hình hóa dữ liệu là nền tảng, nơi xây dựng bất kỳ hệ sinh thái kỹ thuật số nào. Nếu không có mô tả về hệ thống, yêu cầu và mô hình hóa dữ liệu, các kỹ sư và chuyên gia tạo ra dữ liệu sẽ không biết và hiểu rõ nơi mà dữ liệu được tạo ra sẽ được sử dụng.

Giống như khi xây dựng một tòa nhà, nơi không thể bắt đầu đặt gạch mà không có kế hoạch, việc tạo ra hệ thống lưu trữ dữ liệu yêu cầu một sự hiểu biết rõ ràng về các dữ liệu sẽ được sử dụng, cách chúng liên kết với nhau và ai sẽ làm việc với chúng. Nếu không có mô tả về các quy trình và yêu cầu, các kỹ sư và chuyên gia tạo ra dữ liệu sẽ mất đi tầm nhìn về nơi và cách mà dữ liệu sẽ được ứng dụng sau này.

Mô hình dữ liệu đóng vai trò là cầu nối giữa doanh nghiệp và CNTT. Nó cho phép yêu cầu, cấu trúc thông tin và đơn giản hóa giao tiếp giữa các bên liên quan. Theo cách này, mô hình hóa dữ liệu giống như công việc của một kiến trúc sư, người thiết kế một kế hoạch cho tòa nhà theo yêu cầu của khách hàng, sau đó chuyển giao nó cho các nhà thầu - quản trị viên và nhà phát triển cơ sở dữ liệu - để hiện thực hóa (tạo ra cơ sở dữ liệu).

Do đó, mỗi công ty xây dựng, ngoài việc cấu trúc và phân loại các yếu tố và tài nguyên, cũng cần thành thạo nghệ thuật "xây dựng" cơ sở dữ liệu (bảng) và học cách tạo ra các mối quan hệ giữa chúng, giống như kết nối các viên gạch để tạo thành một bức tường kiến thức từ dữ liệu của công ty vững chắc. Các khái niệm chính trong mô hình hóa dữ liệu bao gồm:--

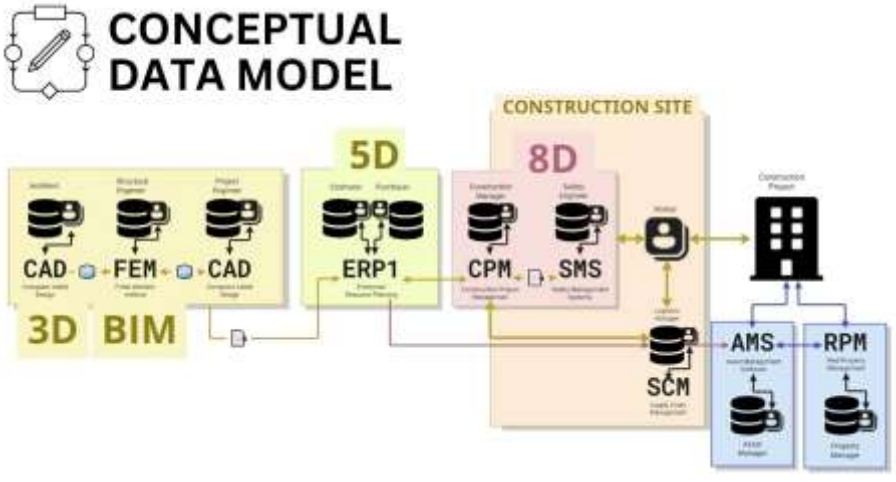
- Thực thể là các đối tượng mà dữ liệu cần phải được thu thập. Trong giai đoạn đầu của thiết kế, một thực thể có thể là một yếu tố riêng lẻ (ví dụ, "cửa"), trong khi trong mô hình ước tính - là một nhóm các yếu tố được kết hợp theo các danh mục (ví dụ, "cửa nội thất").
- Thuộc tính là đặc điểm của các thực thể, mô tả các chi tiết quan trọng: kích thước, tính chất, chi phí lắp ráp, logistics và các thông số khác.
- Mối quan hệ (liên kết) thể hiện cách các thực thể tương tác với nhau. Chúng có thể thuộc một trong các loại: "một với một", "nhiều với một", "nhiều với nhiều".
- Sơ đồ ER (Entity-Relationship diagrams) là các sơ đồ trực quan, thể hiện các thực thể, thuộc tính và mối quan hệ giữa chúng. Sơ đồ ER có thể là sơ đồ khái niệm, logic và vật lý - mỗi cái phản ánh một mức độ chi tiết riêng.



Sơ đồ ER trong cấu trúc cơ sở dữ liệu khái niệm với các thực thể, thuộc tính và mối quan hệ.

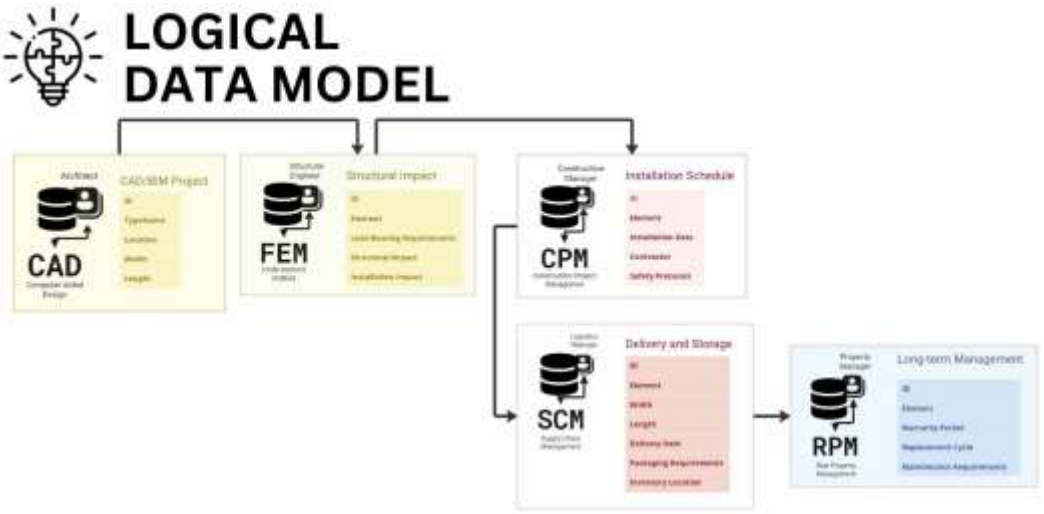
Quy trình thiết kế dữ liệu và xác định các mối liên hệ giữa chúng thường được chia thành ba mô hình chính. Mỗi mô hình thực hiện các chức năng cụ thể, khác nhau về mức độ chi tiết và mức độ trừu tượng trong việc trình bày cấu trúc dữ liệu:

- Mô hình dữ liệu khái niệm: mô hình này mô tả các thực thể chính và mối quan hệ của chúng, không đi vào chi tiết về thuộc tính. Thông thường, nó được sử dụng trong các giai đoạn đầu của lập kế hoạch. Ở giai đoạn này, chúng ta có thể phác thảo từ các cơ sở dữ liệu và hệ thống, để cho thấy mối liên hệ giữa các phòng ban và chuyên gia khác nhau.



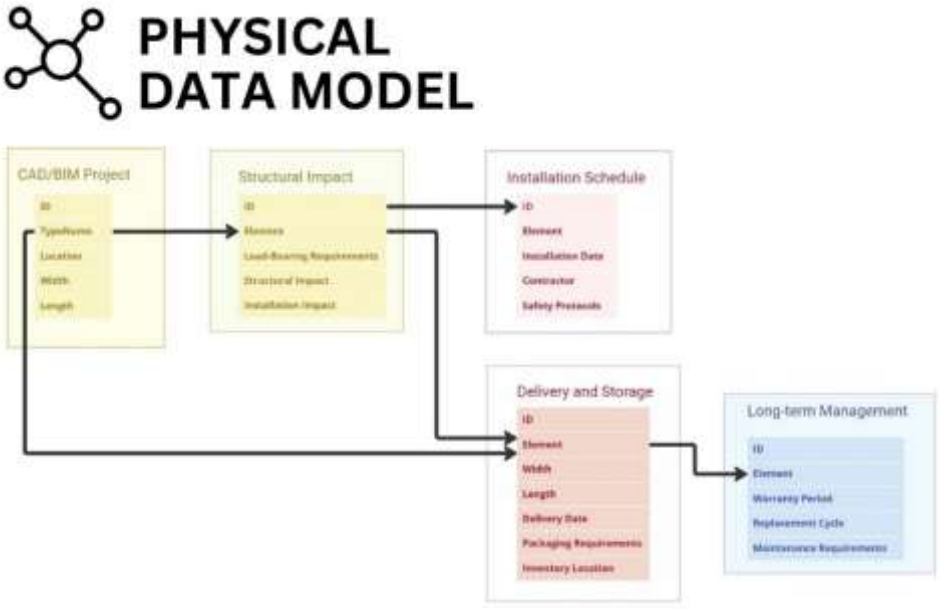
Sơ đồ khái niệm mô tả nội dung của hệ thống: hình ảnh cấp cao về các mối quan hệ, không có chi tiết kỹ thuật.

- Mô hình dữ liệu logic: dựa trên mô hình khái niệm, mô hình dữ liệu logic bao gồm một mô tả chi tiết về các thực thể, thuộc tính, khóa và mối quan hệ, phản ánh thông tin doanh nghiệp và các quy tắc.



Hình 4.33 Mô hình dữ liệu logic mô tả rõ ràng các loại dữ liệu, mối quan hệ và khóa, nhưng không có triển khai hệ thống.

- Mô hình dữ liệu vật lý: mô hình này mô tả các cấu trúc cần thiết để triển khai cơ sở dữ liệu, bao gồm các bảng, cột và mối quan hệ. Nó tập trung vào hiệu suất của cơ sở dữ liệu, các chiến lược lập chỉ mục và lưu trữ vật lý để tối ưu hóa việc triển khai vật lý của cơ sở dữ liệu.

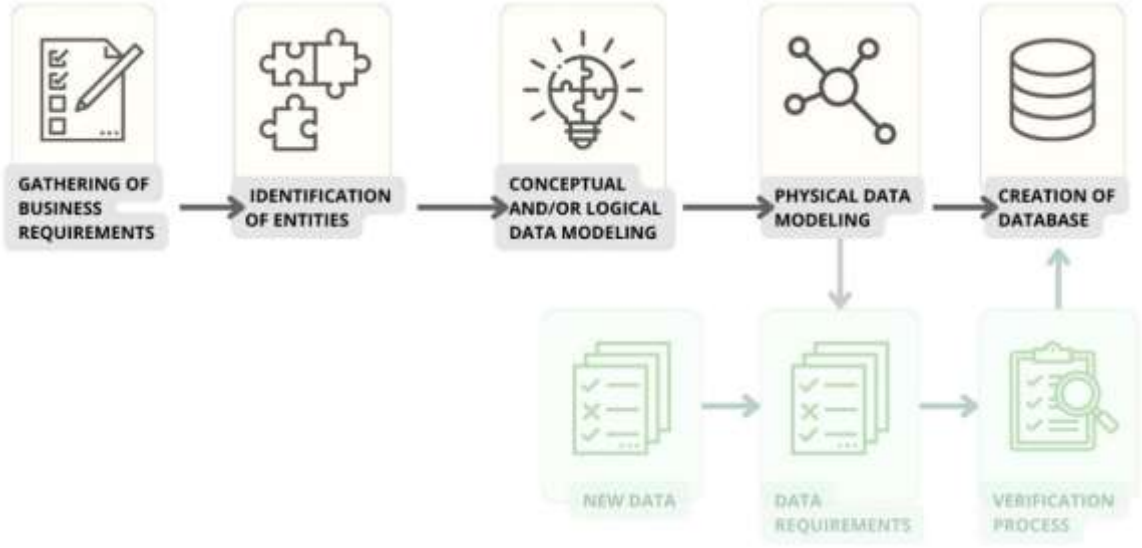


Hình 4.34 Mô hình dữ liệu vật lý xác định cách thức hệ thống sẽ được triển khai, bao gồm các bảng và các chi tiết cụ thể của cơ sở dữ liệu.

Khi phát triển cơ sở dữ liệu và thiết kế mối quan hệ bảng, việc hiểu các cấp độ trừu tượng đóng vai trò quan trọng trong việc xây dựng kiến trúc hệ thống hiệu quả.

Một phương pháp mô hình hóa dữ liệu hiệu quả cho phép kết hợp các nhiệm vụ kinh doanh với việc triển khai kỹ thuật, làm cho toàn bộ chuỗi quá trình trở nên rõ ràng và dễ quản lý hơn. Mô hình hóa dữ liệu không phải là một nhiệm vụ đơn lẻ, mà là một quy trình bao gồm các bước liên tiếp (Hình 4.35):-

- Thu thập yêu cầu kinh doanh: xác định các nhiệm vụ chính, mục tiêu và dòng thông tin. Đây là giai đoạn tương tác tích cực với các chuyên gia và người dùng.
- Xác định các thực thể: xác định các đối tượng chính, loại và phân loại dữ liệu mà cần xem xét trong hệ thống tương lai.
- Phát triển mô hình khái niệm và logic: đầu tiên ghi lại các thực thể chính và mối quan hệ của chúng, sau đó là các thuộc tính, quy tắc và cấu trúc chi tiết.
- Mô hình vật lý: thiết kế triển khai kỹ thuật của mô hình: các bảng, trường, mối quan hệ, ràng buộc, chỉ mục.
- Tạo cơ sở dữ liệu: bước cuối cùng là triển khai mô hình vật lý trong hệ quản trị cơ sở dữ liệu đã chọn, tiến hành kiểm tra và chuẩn bị cho việc vận hành.



Hình 4.35 Việc tạo ra cơ sở dữ liệu và hệ thống quản lý dữ liệu cho các quy trình kinh doanh bắt đầu từ việc hình thành yêu cầu và mô hình hóa dữ liệu.

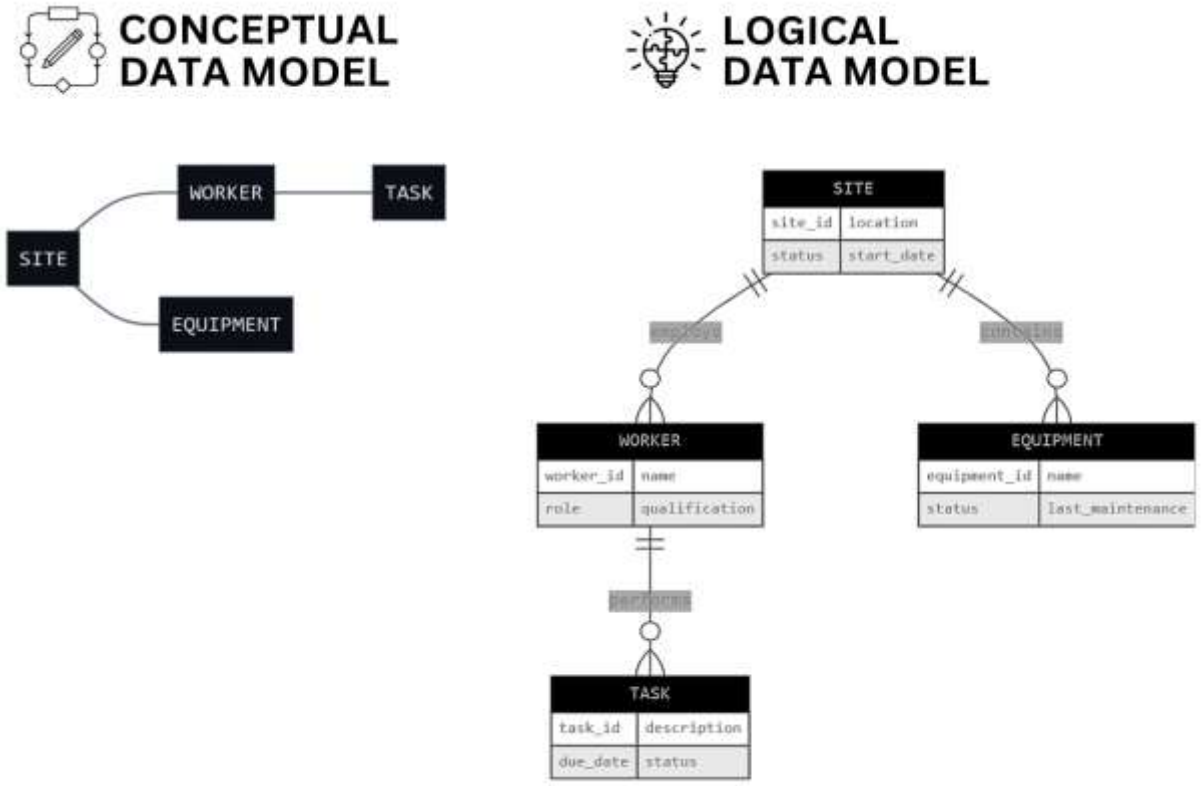
Các quy trình mô hình hóa dữ liệu được xây dựng hợp lý giúp đảm bảo tính minh bạch trong dòng thông tin, điều này đặc biệt quan trọng trong các dự án phức tạp như quản lý dự án xây dựng hoặc công trường xây dựng. Hãy xem cách chuyển từ mô hình khái niệm sang logic, và sau đó sang vật lý, giúp sắp xếp các quy trình.

Mô hình hóa dữ liệu thực tế trong bối cảnh xây dựng

Lấy ví dụ về mô hình hóa dữ liệu cho nhiệm vụ quản lý công trường xây dựng và biến đổi yêu cầu của người chỉ huy thành một mô hình logic có cấu trúc. Dựa trên các nhu cầu cơ bản của việc quản lý xây dựng, chúng

tôi xác định các thực thể chính cho: công trường xây dựng (SITE), công nhân (WORKER), thiết bị (EQUIPMENT), nhiệm vụ (TASK) và việc sử dụng thiết bị (EQUIPMENT_USAGE). Mỗi thực thể chứa một tập hợp các thuộc tính phản ánh các đặc điểm quan trọng. Ví dụ, đối với TASK có thể là mô tả nhiệm vụ, thời gian hoàn thành, trạng thái, ưu tiên; đối với WORKER là tên, vai trò của anh ta tại công trường, tình trạng làm việc hiện tại, v.v.

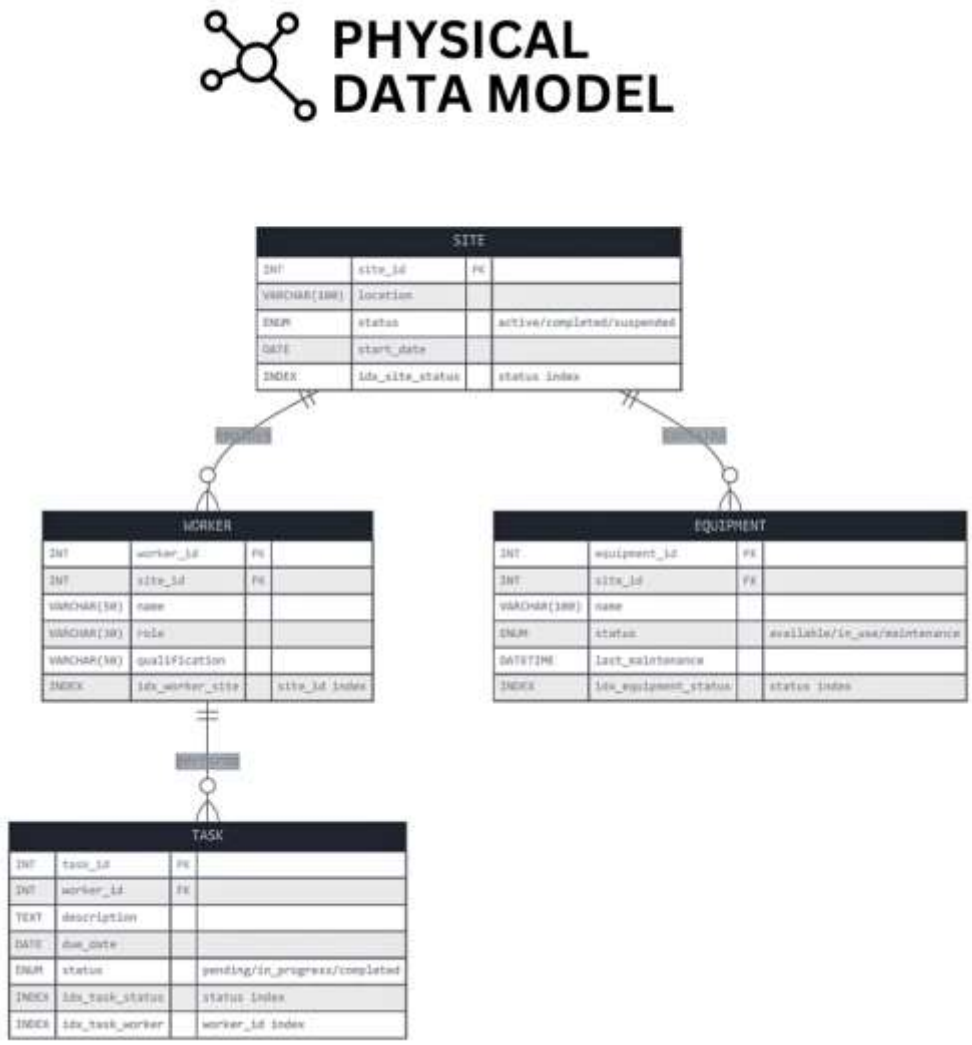
Trong mô hình logic, các mối quan hệ giữa các thực thể được thiết lập, cho thấy cách chúng tương tác với nhau trong các quy trình làm việc thực tế (Hình 4.36). Ví dụ, mối quan hệ giữa trang và công nhân chỉ ra rằng nhiều công nhân có thể làm việc trên một trang, trong khi mối quan hệ giữa công nhân và nhiệm vụ phản ánh rằng một công nhân có thể thực hiện nhiều nhiệm vụ.



Hình 4.36 Mô hình khái niệm và mô hình logic của dữ liệu, được hình thành từ các yêu cầu của kỹ sư trưởng để mô tả các quy trình trên công trường xây dựng.

Khi chuyển sang mô hình vật lý, các chi tiết kỹ thuật triển khai được thêm vào: các loại dữ liệu cụ thể (VARCHAR, INT, DATE), các khóa chính và khóa ngoại cho các mối quan hệ giữa các bảng, cũng như các chỉ mục để tối ưu hóa hiệu suất của cơ sở dữ liệu (Hình 4.37).

Ví dụ, đối với các trạng thái, cần xác định các loại cụ thể với các giá trị có thể, và để cải thiện hiệu suất tìm kiếm, cần thêm chỉ mục cho các trường chính, chẳng hạn như status và worker_id. Điều này biến mô tả logic của hệ thống thành một kế hoạch triển khai cơ sở dữ liệu cụ thể, sẵn sàng để tạo và triển khai.



Hình 4.37 Mô hình vật lý của dữ liệu mô tả các thực thể trên công trường xây dựng thông qua các tham số tối thiểu cần thiết.

> Mô hình vật lý thường khác với mô hình logic. Trung bình, phân bổ thời gian cho việc mô hình hóa trông như sau: khoảng 50% cho mô hình khái niệm (thu thập yêu cầu, thảo luận quy trình, xác định thực thể), 10% cho mô hình logic (làm rõ thuộc tính và liên kết) và 40% cho mô hình vật lý (triển khai, kiểm tra, điều chỉnh cho CSDL).

Cân bằng này được giải thích bởi thực tế rằng giai đoạn khái niệm đặt nền tảng cho cấu trúc dữ liệu, trong khi mô hình logic chỉ làm rõ các mối quan hệ và thuộc tính. Nguồn lực lớn nhất được yêu cầu cho mô hình vật lý, vì chính ở giai đoạn này, dữ liệu được triển khai trên các nền tảng và công cụ cụ thể.

Tạo cơ sở dữ liệu bằng LLM

Có mô hình dữ liệu và mô tả các thực thể thông qua các tham số, chúng ta đã sẵn sàng để tạo ra cơ sở dữ liệu - những nơi lưu trữ thông tin, được tiếp nhận sau giai đoạn cấu trúc theo các quy trình cụ thể.

Hãy thử tạo một ví dụ về cơ sở dữ liệu đơn giản nhưng chức năng với số lượng mã tối thiểu bằng SQLite thông qua ngôn ngữ lập trình Python. Cơ sở dữ liệu quan hệ đã được xem xét chi tiết trong chương "Cơ sở dữ liệu quan hệ có cấu trúc và ngôn ngữ truy vấn SQL".

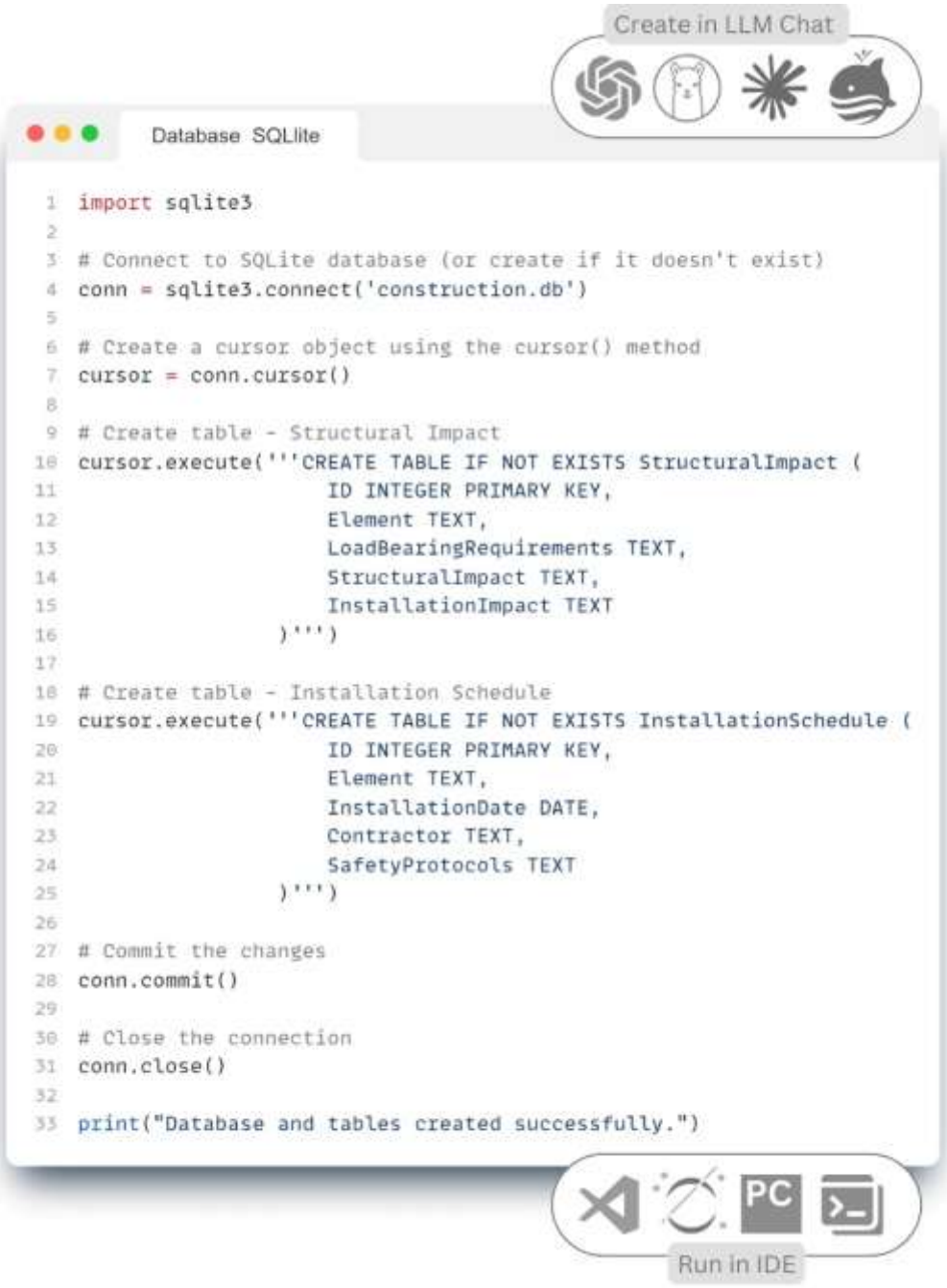
> SQLite là một cơ sở dữ liệu nhẹ, mã nguồn mở, nhúng và đa nền tảng, hỗ trợ các chức năng cơ bản của SQL. SQLite được sử dụng rộng rãi trong CAD (BIM), ERP, EPM và có mặt trong hầu hết các smartphone.

Hãy yêu cầu mô hình LLM tạo cho chúng ta một cơ sở dữ liệu nhỏ SQLlite dựa trên mô hình vật lý đã tạo ở trên (Hình 4.34) cho các hệ thống FEM (tác động cấu trúc) và CPM (lịch trình lắp đặt).-

- Gửi yêu cầu văn bản đến LLM chat (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ công cụ nào khác):

> Tạo cơ sở dữ liệu SQLlite cho các hệ thống "Tác động cấu trúc" và "Lịch trình lắp đặt" từ hình ảnh trong đính kèm (chèn hình ảnh vào chat (Hình 4.34)) ⏎-

Phản hồi từ LLM:



Hình 4.38 Mã được tạo ra tạo ra các cơ sở dữ liệu và bảng từ hình ảnh đã tải lên, vào đó chúng ta bây giờ có thể tải dữ liệu.

Mã này để tạo cơ sở dữ liệu có thể được thực hiện trong môi trường Python hỗ trợ SQLite, thường được bao gồm trong môi trường tiêu chuẩn của Python. Nó có thể được chạy trên bất kỳ máy tính cục bộ nào có Python, có khả năng chạy các tập lệnh và tạo tệp. Sau khi thực thi, mã này sẽ tạo ra một tệp cơ sở dữ liệu SQLite có tên construction.db (tương tự như tệp Excel với các bảng tính), sẽ chứa các bảng "Tác động cấu trúc" và "Lịch trình lắp đặt".

Sau khi tạo ra cơ sở dữ liệu construction.db mới trong các bảng này, có thể thêm dữ liệu thông qua các truy vấn SQL hoặc nhập khẩu, điều này sẽ cho phép bắt đầu tạo ra quy trình xử lý dữ liệu tự động sau này. Dữ liệu có thể được nhập vào cơ sở dữ liệu SQLite từ các tệp CSV, bảng tính Excel hoặc thông qua API xuất khẩu từ các cơ sở dữ liệu và kho lưu trữ khác.

Để thiết lập các quy trình mô hình hóa dữ liệu bền vững và quản lý cơ sở dữ liệu hiệu quả, các công ty cần có một chiến lược được định hình rõ ràng, cũng như phối hợp chặt chẽ giữa các đội ngũ kỹ thuật và kinh doanh. Trong điều kiện các dự án phân tán và nhiều nguồn dữ liệu, thường rất khó để đảm bảo sự đồng nhất, chuẩn hóa và kiểm soát chất lượng ở mọi cấp độ. Một trong những giải pháp quan trọng có thể là thiết lập một Trung tâm Xuất sắc về Mô hình hóa Dữ liệu (Data Modeling Center of Excellence, COE) bên trong công ty.

Trung tâm kinh nghiệm xuất sắc (CoE) về mô hình hóa dữ liệu

Trong bối cảnh dữ liệu trở thành một trong những tài sản chiến lược quan trọng, các công ty không chỉ cần thu thập và lưu trữ thông tin đúng cách — mà còn quan trọng là học cách quản lý dữ liệu một cách hệ thống. Trung tâm Xuất sắc về Phân loại và Mô hình hóa Dữ liệu (Center of Excellence, CoE) là một đơn vị cấu trúc giúp đảm bảo sự đồng nhất, chất lượng và hiệu quả trong tất cả các công việc liên quan đến dữ liệu trong tổ chức.

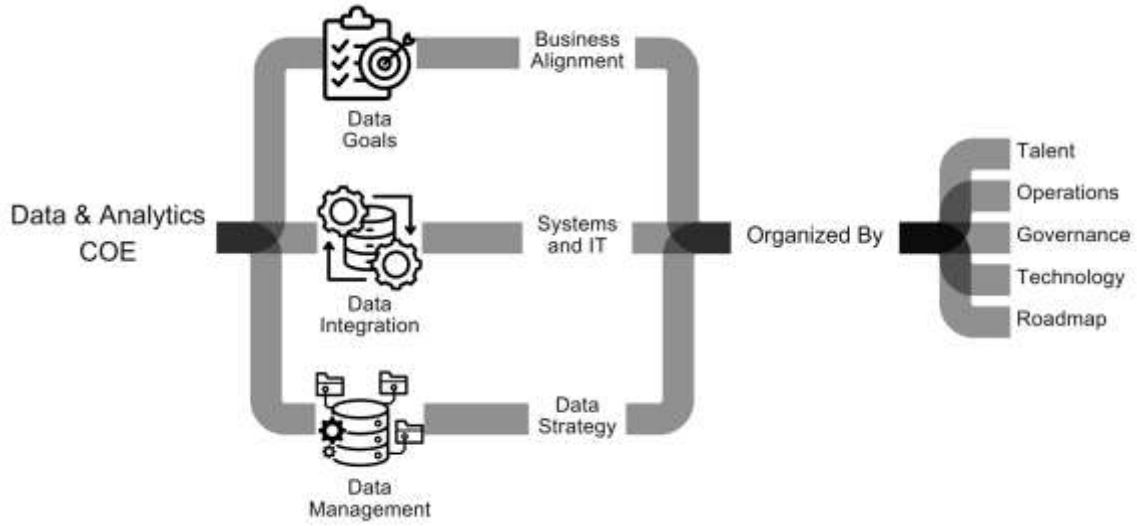
> Trung tâm Xuất sắc (CoE) là hạt nhân hỗ trợ chuyên gia và nền tảng phương pháp cho các chuyển đổi số trong công ty. Nó hình thành văn hóa làm việc với dữ liệu và cho phép các tổ chức xây dựng quy trình ra quyết định không dựa trên trực giác hoặc thông tin cục bộ, mà dựa trên dữ liệu có cấu trúc, đã được kiểm chứng và đại diện.

Trung tâm Xuất sắc về Dữ liệu thường được hình thành từ các đội ngũ đa chức năng làm việc theo nguyên tắc "hai chiếc pizza". Nguyên tắc này, do Jeff Bezos đề xuất, có nghĩa là kích thước của đội ngũ phải đủ nhỏ để có thể được nuôi bằng hai chiếc pizza, tức là không vượt quá 6-10 người. Cách tiếp cận này giúp tránh sự quan liêu quá mức và nâng cao tính linh hoạt trong công việc. Đội ngũ CoE nên bao gồm các nhân viên với nhiều kỹ năng kỹ thuật đa dạng: từ phân tích dữ liệu và học máy đến chuyên môn trong các lĩnh vực kinh doanh cụ thể. Với kiến thức kỹ thuật sâu sắc, các kỹ sư xử lý dữ liệu không chỉ phải tối ưu hóa quy trình và mô hình hóa dữ liệu mà còn hỗ trợ đồng nghiệp, giảm thời gian cho các nhiệm vụ lặp đi lặp lại.-

> Như trong tự nhiên, sự bền vững của hệ sinh thái được đảm bảo bởi sự đa dạng sinh học, cũng như trong thế giới số, tính linh hoạt và khả năng thích ứng được đạt được thông qua sự đa dạng trong cách làm việc với dữ liệu. Tuy nhiên, sự đa dạng này phải dựa trên các quy tắc và khái niệm thống nhất.

Trung tâm Xuất sắc (CoE) có thể được so sánh với "các điều kiện khí hậu" của hệ sinh thái rừng, xác định

loại dữ liệu nào sẽ phát triển mạnh và loại nào sẽ tự động bị loại bỏ. Bằng cách tạo ra một "khí hậu" thuận lợi cho dữ liệu chất lượng, CoE thúc đẩy sự chọn lọc tự nhiên về các thực tiễn và phương pháp tốt nhất, những cái sau này trở thành tiêu chuẩn trong tổ chức.



Hình 4.39 Trung tâm xuất sắc (CoE) về dữ liệu và phân tích kết hợp chuyên môn về các khía cạnh chính của quản lý dữ liệu, tích hợp chúng và phát triển chiến lược.

Để đẩy nhanh chu trình tích hợp và đạt được kết quả tốt hơn, CoE cần cung cấp cho các thành viên của mình mức độ tự chủ nhất định trong việc ra quyết định. Điều này đặc biệt quan trọng trong môi trường năng động, nơi mà phương pháp thử và sai, phản hồi liên tục và các phiên bản thường xuyên có thể mang lại lợi ích đáng kể. Tuy nhiên, sự tự chủ này chỉ hiệu quả khi có sự giao tiếp rõ ràng và hỗ trợ từ ban lãnh đạo cấp cao. Nếu không có tầm nhìn chiến lược và phối hợp từ trên xuống, ngay cả đội ngũ có năng lực nhất cũng có thể gặp phải rào cản trong việc triển khai các sáng kiến của mình.

> Chính CoE hoặc ban lãnh đạo cấp cao của công ty chịu trách nhiệm đảm bảo rằng cách tiếp cận mô hình hóa dữ liệu không chỉ giới hạn trong một hoặc hai dự án mà được tích hợp vào hệ thống quản lý thông tin và quy trình kinh doanh tổng thể.

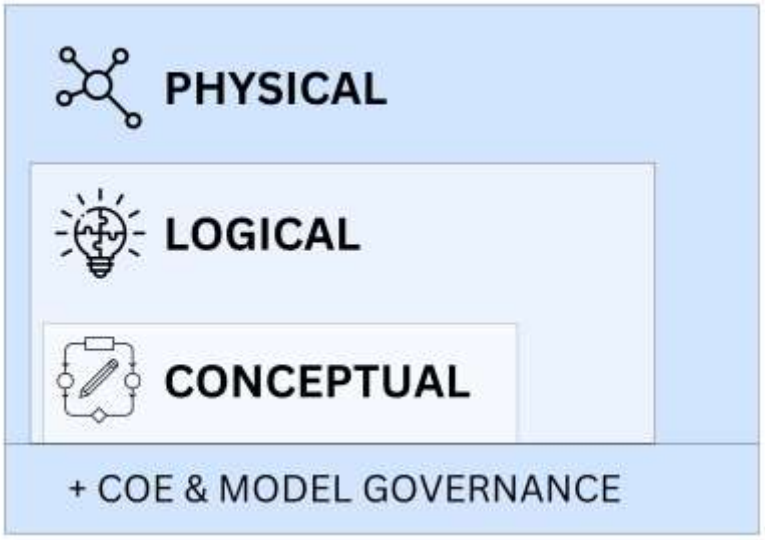
Trung tâm chuyên môn (CoE) không chỉ có nhiệm vụ liên quan đến mô hình hóa dữ liệu và quản lý chúng (Quản lý Dữ liệu), mà còn phụ trách việc thiết lập các tiêu chuẩn và cách tiếp cận thống nhất cho việc triển khai và vận hành hạ tầng dữ liệu. Hơn nữa, nó còn hình thành văn hóa cải tiến liên tục, tối ưu hóa quy trình và sử dụng dữ liệu hiệu quả trong tổ chức (Hình 4.310).

Cách tiếp cận hệ thống trong quản lý dữ liệu và mô hình bên trong CoE có thể được chia thành một số khối chính:

- Chuẩn hóa các quy trình và quản lý vòng đời mô hình: CoE phát triển và triển khai các phương pháp để đồng nhất hóa việc tạo ra và quản lý các mô hình dữ liệu. Điều này bao gồm: hình thành các mẫu cấu trúc, phương pháp kiểm soát chất lượng và hệ thống quản lý phiên bản, đảm bảo tính liên tục của dữ liệu ở tất cả các giai đoạn làm việc.
- Quản lý vai trò và phân bổ trách nhiệm: trong khuôn khổ CoE, các vai trò chính trong quá trình mô hình hóa dữ liệu được xác định. Mỗi thành viên trong dự án đều có chức năng và khu vực trách

nhiệm rõ ràng, điều này góp phần vào việc làm việc nhóm nhịp nhàng và giảm thiểu rủi ro không phù hợp dữ liệu.

- Kiểm soát chất lượng và kiểm toán: việc quản lý dữ liệu xây dựng hiệu quả yêu cầu phải có sự giám sát liên tục về chất lượng của chúng. Các cơ chế kiểm tra tự động được triển khai để phát hiện dữ liệu, phát hiện lỗi, và các thuộc tính thiếu.
- Quản lý siêu dữ liệu và kiến trúc thông tin: CoE chịu trách nhiệm xây dựng một hệ thống phân loại và định danh chung, tiêu chuẩn đặt tên và mô tả đối tượng, điều này rất quan trọng cho việc tích hợp giữa các hệ thống.



Hình 4.310 Mô hình hóa dữ liệu và quản lý chất lượng dữ liệu là một trong những nhiệm vụ chính của CoE.

Trung tâm xuất sắc (CoE) về dữ liệu không chỉ là một nhóm chuyên gia, mà còn là một cơ chế hệ thống tạo ra một văn hóa dựa trên dữ liệu mới và đảm bảo một cách tiếp cận thống nhất trong công việc với dữ liệu trên toàn công ty. Nhờ tích hợp hợp lý quy trình mô hình hóa vào hệ thống quản lý thông tin tổng thể, chuẩn hóa, phân loại và kiểm soát chất lượng dữ liệu, CoE giúp doanh nghiệp cải tiến liên tục sản phẩm và quy trình kinh doanh của mình, phản ứng nhanh chóng với những thay đổi của thị trường và đưa ra quyết định dựa trên phân tích đáng tin cậy.

Các trung tâm như vậy đặc biệt hiệu quả khi kết hợp với các nguyên tắc hiện đại của DataOps - một phương pháp đảm bảo việc cung cấp liên tục, tự động hóa và kiểm soát chất lượng dữ liệu. Chúng ta sẽ nói chi tiết về DataOps trong phần tám, trong chương "Công nghệ hiện đại trong công việc với dữ liệu trong ngành xây dựng".

Trong các chương tiếp theo, chúng ta sẽ chuyển từ chiến lược sang thực tiễn - một cách tương đối "biến hình" thành trung tâm xử lý dữ liệu: xem xét qua một vài ví dụ, cách thức tham số hóa nhiệm vụ, thu thập yêu cầu và quy trình xác thực tự động.



CHƯƠNG 4.4.
HỆ THỐNG CÁC YÊU CẦU VÀ XÁC MINH THÔNG TIN

Thu thập và phân tích yêu cầu: chuyển đổi giao tiếp thành dữ liệu có cấu trúc

Thu thập và quản lý yêu cầu là bước đầu tiên để đảm bảo chất lượng dữ liệu. Mặc dù các công cụ kỹ thuật số đã phát triển, hầu hết các yêu cầu vẫn được diễn đạt dưới dạng không cấu trúc: thông qua email, biên bản cuộc họp, cuộc gọi điện thoại và thảo luận miệng. Hình thức giao tiếp này gây khó khăn cho việc tự động hóa, kiểm tra và tái sử dụng thông tin. Trong chương này, chúng ta sẽ xem xét cách chuyển đổi các yêu cầu văn bản thành các cấu trúc chính thức, đảm bảo tính minh bạch và hệ thống cho các nhiệm vụ kinh doanh.

> Nghiên cứu của công ty Gartner về "Chất lượng dữ liệu: Các thực tiễn tốt nhất để có được thông tin chính xác" nhấn mạnh tầm quan trọng sống còn của chất lượng dữ liệu trong việc thực hiện các sáng kiến thành công về dữ liệu và phân tích. Họ lưu ý rằng chất lượng dữ liệu kém khiến các tổ chức phải chịu chi phí trung bình không dưới 12,9 triệu đô la mỗi năm và rằng dữ liệu đáng tin cậy, chất lượng cao là cần thiết để xây dựng một công ty được điều hành bởi dữ liệu.

Sự thiếu hụt yêu cầu có cấu trúc dẫn đến việc cùng một yếu tố (thực thể) và các tham số của nó có thể được lưu trữ trong các hệ thống khác nhau với nhiều biến thể khác nhau. Điều này không chỉ làm giảm hiệu suất của các quy trình mà còn dẫn đến việc lãng phí thời gian, trùng lặp thông tin và cần kiểm tra lại dữ liệu trước khi sử dụng. Kết quả là, chỉ cần một thiếu sót — tham số bị mất hoặc một yếu tố được mô tả không chính xác — cũng có thể làm chậm quyết định và gây lãng phí tài nguyên không hiệu quả.

> *Vì không có đinh, nên lẻ loi mất đi. Vì không có đinh, nên ngựa đã mất. Vì không có ngựa, nên kỵ binh đã biến mất. Vì không có kỵ binh, nên thông điệp đã bị mất. Vì không có thông điệp, nên trận chiến đã thua. Vì không có trận chiến, nên vương quốc đã mất. Tất cả chỉ vì thiếu một cái đinh trong móng ngựa.*
>
> *— Thành ngữ [101]*

Phân tích và thu thập yêu cầu cho quy trình nhập liệu và lưu trữ dữ liệu bắt đầu bằng việc xác định tất cả các bên liên quan. Giống như trong câu châm ngôn, việc mất một chiếc đinh có thể dẫn đến một chuỗi hậu quả nghiêm trọng, trong kinh doanh, việc mất một người tham gia, bỏ sót một yêu cầu hoặc thậm chí mất một thông số có thể ảnh hưởng đáng kể không chỉ đến một quy trình kinh doanh riêng lẻ mà còn đến toàn bộ hệ sinh thái của dự án và tổ chức nói chung. Do đó, việc xác định ngay cả những yếu tố, thông số và vai trò có vẻ không quan trọng ở cái nhìn đầu tiên nhưng có thể trở nên cực kỳ quan trọng cho sự bền vững của doanh nghiệp là vô cùng cần thiết.

Giả sử rằng một công ty có một dự án, trong đó khách hàng đưa ra yêu cầu mới - "thêm một cửa sổ bổ sung

ở phía bắc của tòa nhà". Trong quy trình nhỏ "yêu cầu của khách hàng về việc thêm cửa sổ mới vào dự án hiện tại", các chuyên gia như kiến trúc sư, nhà thầu, chuyên gia CAD (BIM), quản lý xây dựng, quản lý logistics, phân tích viên ERP, kỹ sư kiểm soát chất lượng, kỹ sư an toàn, quản lý kiểm soát và quản lý bất động sản đều tham gia.

> Ngay cả trong một quy trình nhỏ, có thể có hàng chục chuyên gia khác nhau tham gia. Mỗi thành viên trong quy trình cần hiểu yêu cầu của các chuyên gia mà họ liên quan đến ở mức độ dữ liệu.

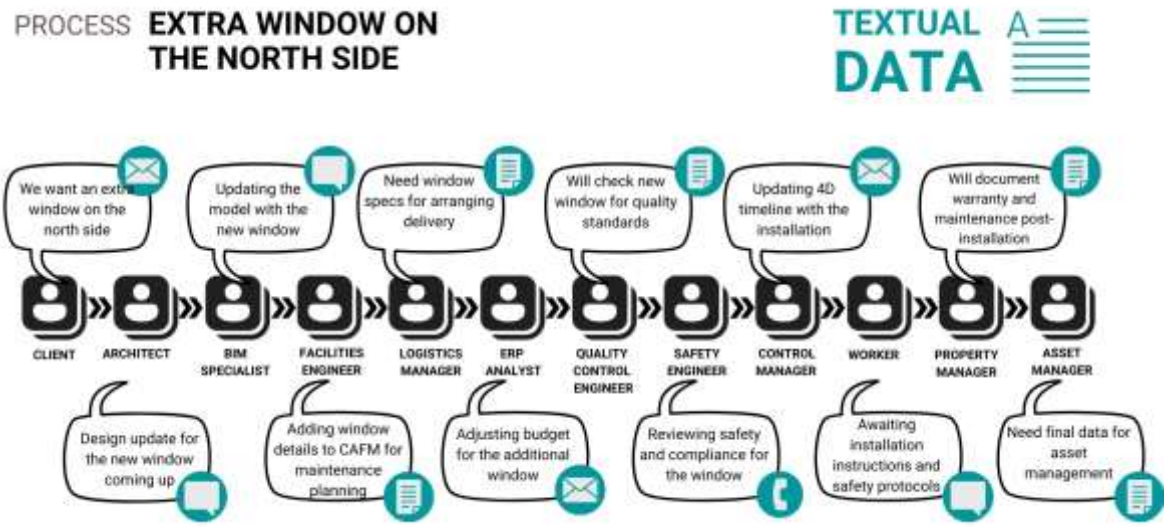
Trên cấp độ văn bản (Hình 4.41), giao tiếp giữa khách hàng và các chuyên gia trong chuỗi quy trình diễn ra như sau:-

- Nhà thầu: "Chúng tôi đã quyết định thêm một cửa sổ bổ sung ở phía bắc để cải thiện ánh sáng. Liệu điều này có thể thực hiện được không?"
- Kiến trúc sư: "Chắc chắn rồi, tôi sẽ xem xét lại dự án để đưa cửa sổ mới vào và gửi các bản kế hoạch CAD (BIM) cập nhật."
- Chuyên gia CAD (BIM): "Tôi đã nhận được dự án mới. Tôi sẽ cập nhật mô hình CAD (BIM) với cửa sổ bổ sung và sau khi đồng ý với kỹ sư FEM, tôi sẽ cung cấp vị trí và kích thước chính xác của cửa sổ mới."
- Quản lý xây dựng: "Dự án mới đã được nhận. Chúng tôi sẽ điều chỉnh thời gian lắp đặt 4D và thông báo cho tất cả các thầu phụ liên quan."
- Kỹ sư hệ thống (CAFM): "Tôi sẽ nhập dữ liệu 6D về cửa sổ mới vào hệ thống CAFM để quản lý tài sản trong tương lai và lập kế hoạch bảo trì."
- Quản lý logistics: "Tôi cần kích thước và trọng lượng của cửa sổ mới để tổ chức vận chuyển cửa sổ đến công trường."
- Phân tích viên ERP: "Tôi cần bảng khối lượng và loại cửa sổ chính xác để cập nhật ngân sách 5D trong hệ thống ERP của chúng tôi, nhằm phản ánh chi phí của cửa sổ mới vào tổng dự toán của dự án."
- Kỹ sư kiểm soát chất lượng: "Khi thông số kỹ thuật của cửa sổ được hoàn thành, tôi sẽ đảm bảo rằng chúng đáp ứng tiêu chuẩn chất lượng và vật liệu của chúng tôi."
- Kỹ sư an toàn: "Tôi sẽ đánh giá các khía cạnh an toàn của cửa sổ mới, đặc biệt chú ý đến việc tuân thủ yêu cầu và sơ đồ thoát hiểm theo quy trình 8D."
- Quản lý kiểm soát: "Dựa trên khối lượng công việc chính xác từ ERP, chúng tôi sẽ cập nhật tiến độ 4D của mình để phản ánh việc lắp đặt cửa sổ mới và lưu trữ dữ liệu mới trong hệ thống quản lý nội dung dự án."
- Công nhân (người lắp đặt): "Tôi cần hướng dẫn về lắp đặt, lắp ráp và thời gian hoàn thành công việc. Hơn nữa, có bất kỳ quy tắc an toàn đặc biệt nào mà tôi cần tuân thủ không?"
- Quản lý tài sản: "Sau khi lắp đặt, tôi sẽ tài liệu thông tin về bảo hành và bảo trì để quản lý lâu dài."
- Quản lý tài sản: "Kỹ sư thiết bị, làm ơn gửi dữ liệu cuối cùng để theo dõi tài sản và quản lý vòng đời."
- Khách hàng: "Chờ đã, có thể tôi đang vội và cửa sổ là không cần thiết. Có nên làm một ban công không?"

Trong những kịch bản như vậy, thường xảy ra, ngay cả một thay đổi nhỏ cũng gây ra phản ứng dây chuyền giữa nhiều hệ thống và vai trò. Tại giai đoạn bắt đầu, gần như toàn bộ giao tiếp được thực hiện qua văn bản: email, chat, biên bản cuộc họp (Hình. 4.41).

Trong một hệ thống giao tiếp văn bản cho dự án xây dựng, hệ thống xác thực và đăng ký pháp lý cho tất cả các hoạt động trao đổi dữ liệu và tất cả các quyết định đã được đưa ra là rất quan trọng. Điều này cần thiết để đảm bảo tính pháp lý và khả năng theo dõi từng quyết định, hướng dẫn hoặc thay đổi, từ đó giảm thiểu

rủi ro xảy ra "sự hiểu lầm" trong tương lai.

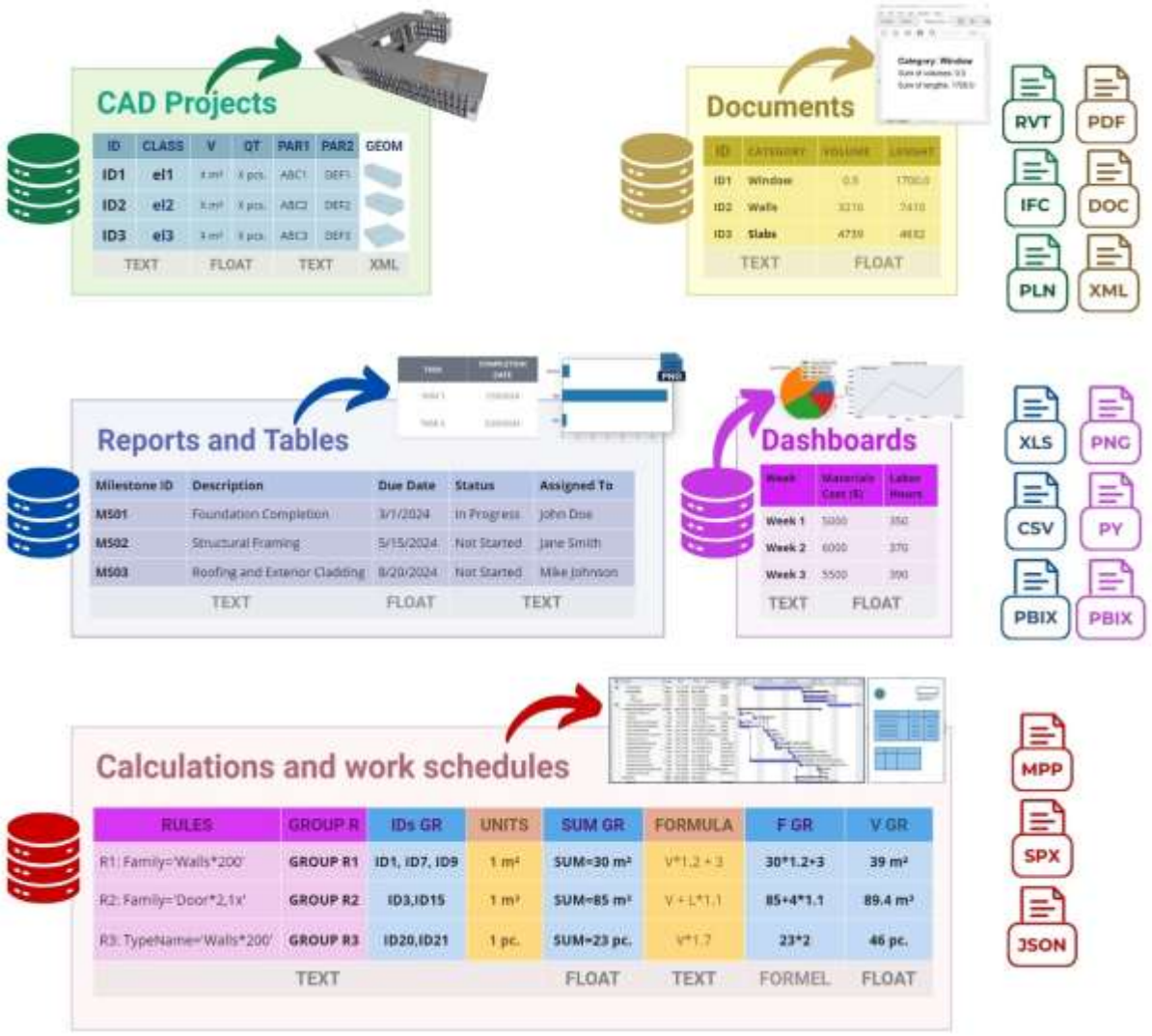


Hình. 4.41 Giao tiếp giữa khách hàng và nhà thầu trong các giai đoạn đầu của dự án thường chứa dữ liệu văn bản đa định dạng.

> Việc thiếu kiểm soát pháp lý và xác nhận các quyết định trong các hệ thống liên quan của dự án xây dựng có thể dẫn đến những vấn đề nghiêm trọng cho tất cả các bên tham gia. Mọi quyết định, chỉ thị hoặc thay đổi được thực hiện mà không có tài liệu hợp lệ và xác nhận có thể dẫn đến tranh chấp (và các vụ kiện tụng).

Việc ghi lại pháp lý cho tất cả các quyết định trong giao tiếp văn bản chỉ có thể được đảm bảo bằng cách có nhiều tài liệu đã được ký, điều này sẽ đè nặng lên ban quản lý, những người có trách nhiệm ghi lại tất cả các giao dịch. Cuối cùng, nếu mỗi thành viên đều bắt buộc phải ký tài liệu cho mỗi hành động, hệ thống sẽ mất đi tính linh hoạt và biến thành một mê cung hành chính. Việc thiếu các xác nhận giao dịch không chỉ làm chậm tiến độ dự án mà còn có thể dẫn đến tổn thất tài chính, cũng như làm xấu đi mối quan hệ giữa các bên tham gia, thậm chí dẫn đến các vấn đề về mặt pháp lý.

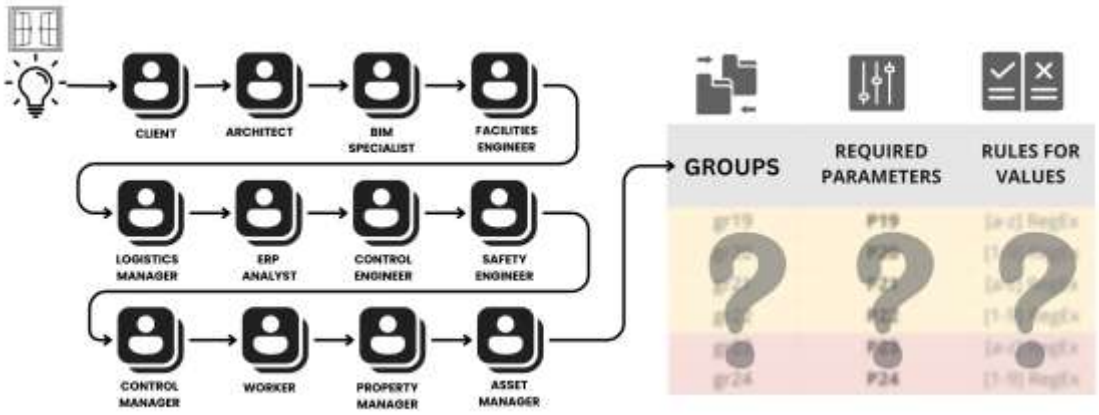
Quy trình phê duyệt và xác nhận giao dịch như vậy, thường bắt đầu với các cuộc thảo luận bằng văn bản, ở các giai đoạn tiếp theo sẽ dần chuyển sang hình thức trao đổi các tài liệu đa định dạng (Hình. 4.42), làm cho giao tiếp trở nên phức tạp hơn nhiều so với việc chỉ sử dụng văn bản. Nếu không có các yêu cầu được xác định rõ ràng, việc tự động hóa những quy trình này, được lấp đầy bởi các dữ liệu đa định dạng và các yêu cầu văn bản phong phú, trở nên gần như không thể.



Hình. 4.42 Mỗi hệ thống trong bối cảnh của công ty xây dựng đều là nguồn tài liệu có giá trị pháp lý ở các định dạng khác nhau.

Giao tiếp văn bản yêu cầu mỗi chuyên gia hoặc phải nắm rõ toàn bộ cuộc trao đổi, hoặc phải tham gia thường xuyên vào tất cả các cuộc họp, để hiểu rõ tình trạng hiện tại của dự án.

Để vượt qua giới hạn này, cần thiết phải chuyển từ giao tiếp văn bản sang mô hình yêu cầu có cấu trúc. Điều này chỉ có thể thực hiện được thông qua phân tích hệ thống, trực quan hóa quy trình và mô tả các tương tác dưới dạng sơ đồ khối và mô hình dữ liệu (Hình. 4.43). Giống như trong việc mô hình hóa dữ liệu (Hình. 4.37), chúng ta đã chuyển từ cấp độ ý tưởng ngữ cảnh sang cấp độ khái niệm, bổ sung các hệ thống và công cụ mà các bên tham gia sử dụng, cũng như các mối liên hệ giữa chúng.-



Hình. 4.43 Để học cách quản lý và tự động hóa quy trình xác thực, cần phải trực quan hóa các quy trình và cấu trúc các yêu cầu.

Bước đầu tiên trong việc hệ thống hóa yêu cầu và quan hệ là hình ảnh hóa tất cả các mối quan hệ và liên kết thông qua các sơ đồ khái niệm. Cấp độ khái niệm không chỉ giúp tất cả các bên tham gia quá trình hiểu rõ toàn bộ chuỗi công nghệ, mà còn cho thấy rõ lý do và đối tượng cần dữ liệu (và yêu cầu) ở mỗi giai đoạn của quá trình.

Sơ đồ quy trình và hiệu quả của các sơ đồ khái niệm

Để vượt qua khoảng cách giữa các phương pháp quản lý dữ liệu truyền thống và hiện đại, các công ty cần ý thức chuyển từ những mô tả văn bản phân mảnh sang việc trình bày cấu trúc các quy trình. Sự tiến hóa của dữ liệu — từ bảng đất sét đến các hệ sinh thái kỹ thuật số — đòi hỏi những công cụ tư duy mới. Một trong những công cụ như vậy là mô hình hóa khái niệm thông qua việc sử dụng sơ đồ. Việc tạo ra các sơ đồ trực quan — sơ đồ khối, biểu đồ quy trình, sơ đồ tương tác — cho phép các bên tham gia dự án nhận thức được cách mà hành động và quyết định của họ ảnh hưởng đến toàn bộ hệ thống ra quyết định.

> Nếu các quy trình không chỉ yêu cầu lưu trữ dữ liệu mà còn cần phân tích hoặc tự động hóa chúng, thì cần phải bắt đầu chú ý đến việc tạo ra một cấp độ yêu cầu khái niệm và hình ảnh.

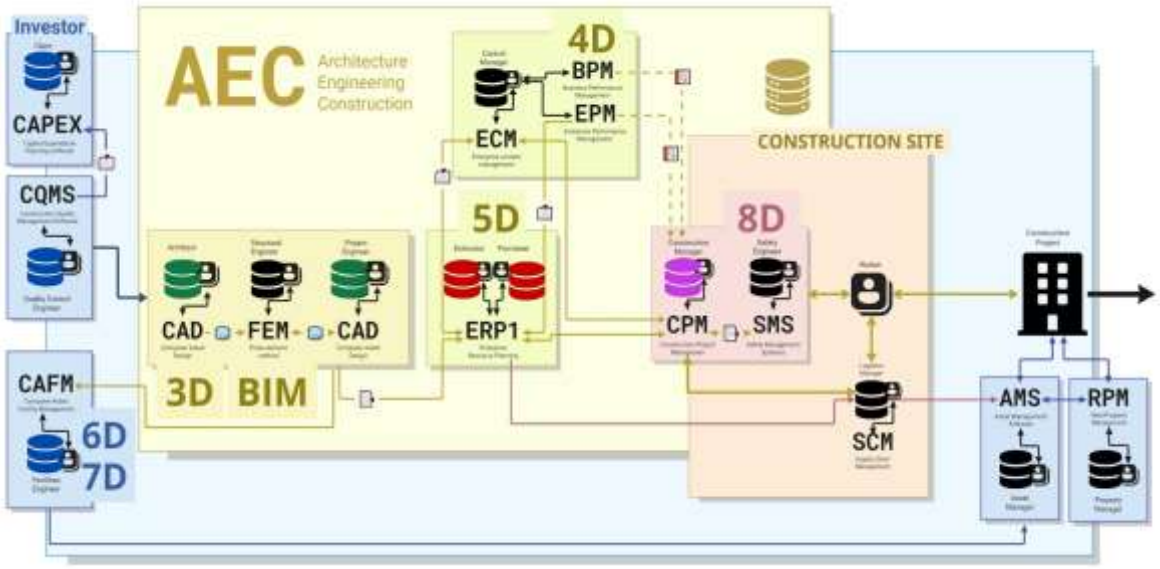
Trong ví dụ của chúng tôi (Hình 4.41), mỗi chuyên gia có thể tham gia không chỉ vào một nhóm nhỏ mà còn vào một phòng ban lớn hơn, gồm nhiều chuyên gia dưới sự quản lý của trưởng phòng. Mỗi phòng ban sử dụng cơ sở dữ liệu chuyên biệt của ứng dụng (Hình 1.24 chẳng hạn như ERP, CAD, MEP, CDE, ECM, CPM, v.v.), thường xuyên được cập nhật thông tin đầu vào cần thiết để tạo ra tài liệu, đăng ký tình trạng pháp lý của các quyết định và quản lý quy trình.

> Quy trình giao dịch tương tự như công việc của các quản lý cổ đại 4000 năm trước, khi bảng đất sét và giấy cói được sử dụng để xác nhận pháp lý các quyết định. Sự khác biệt giữa các hệ thống hiện đại và những người tiền nhiệm bằng đất sét hoặc giấy là các phương pháp hiện đại còn bao gồm cả quy trình chuyển đổi thông tin văn bản thành dạng kỹ thuật số để xử lý tự động trong các hệ thống và công cụ khác.

Việc tạo ra hình ảnh hóa quy trình dưới dạng sơ đồ khái niệm sẽ giúp mô tả mỗi bước và sự tương tác giữa các vai trò khác nhau, làm cho quy trình làm việc phức tạp trở nên dễ hiểu và đơn giản.

> Hình ảnh hóa quy trình đảm bảo tính minh bạch và sẵn có của logic quy trình cho tất cả các thành viên trong nhóm.

Cùng một quy trình giao tiếp thêm cửa sổ vào dự án, được mô tả bằng văn bản, thông điệp (Hình 4.41) và sơ đồ khối, giống như mô hình khái niệm mà chúng tôi đã xem xét trong chương về mô hình hóa dữ liệu (Hình 4.44).



Hình 4.44 Trong sơ đồ khái niệm, các thành viên trong dự án được thể hiện như người dùng cơ sở dữ liệu, nơi các yêu cầu của họ liên kết các hệ thống khác nhau.

Mặc dù các sơ đồ khái niệm là một bước quan trọng, nhiều công ty chỉ giới hạn ở cấp độ này, cho rằng sơ đồ trực quan là đủ để hiểu các quy trình. Điều này tạo ra ảo tưởng về khả năng quản lý: đối với các nhà quản lý, sơ đồ khối như vậy giúp dễ dàng nhận thức bức tranh tổng thể, thấy được mối liên hệ giữa các bên tham gia và các giai đoạn. Tuy nhiên, những sơ đồ này không cung cấp cái nhìn rõ ràng về dữ liệu cần thiết cho từng bên tham gia, định dạng nào chúng cần được truyền tải và các tham số, thuộc tính nào là bắt buộc để thực hiện tự động hóa. Sơ đồ khái niệm giống như một bản đồ lộ trình: nó chỉ ra ai tương tác với ai, nhưng không tiết lộ chính xác những gì được truyền tải trong các tương tác đó.

> Ngay cả khi quy trình được mô tả chi tiết ở cấp độ khái niệm thông qua sơ đồ khối, điều đó không đảm bảo hiệu quả của nó. Hình ảnh hóa thường đơn giản hóa công việc của các nhà quản lý, cho phép họ theo dõi quy trình một cách thuận tiện hơn với hệ thống báo cáo từng bước. Tuy nhiên, đối với các kỹ sư điều hành cơ sở dữ liệu, biểu diễn khái niệm có thể vẫn không rõ ràng và không cung cấp cái nhìn rõ ràng về cách thực hiện quy trình ở cấp độ tham số và yêu cầu.

Khi tiến tới các hệ sinh thái dữ liệu phức tạp hơn, việc áp dụng các công cụ khái niệm và hình ảnh ban đầu

trở nên cực kỳ quan trọng để quy trình xử lý dữ liệu không chỉ hiệu quả mà còn phù hợp với các mục tiêu chiến lược của tổ chức. Để hoàn toàn chuyển quy trình này sang việc bổ sung cửa sổ (Hình. 4.41) lên cấp độ yêu cầu dữ liệu, chúng ta cần đi sâu hơn một bước và chuyển giao hình ảnh khái niệm của quy trình lên cấp độ logic và vật lý của dữ liệu, các thuộc tính yêu cầu và giá trị biên của chúng.

Yêu cầu có cấu trúc và biểu thức chính quy RegEx

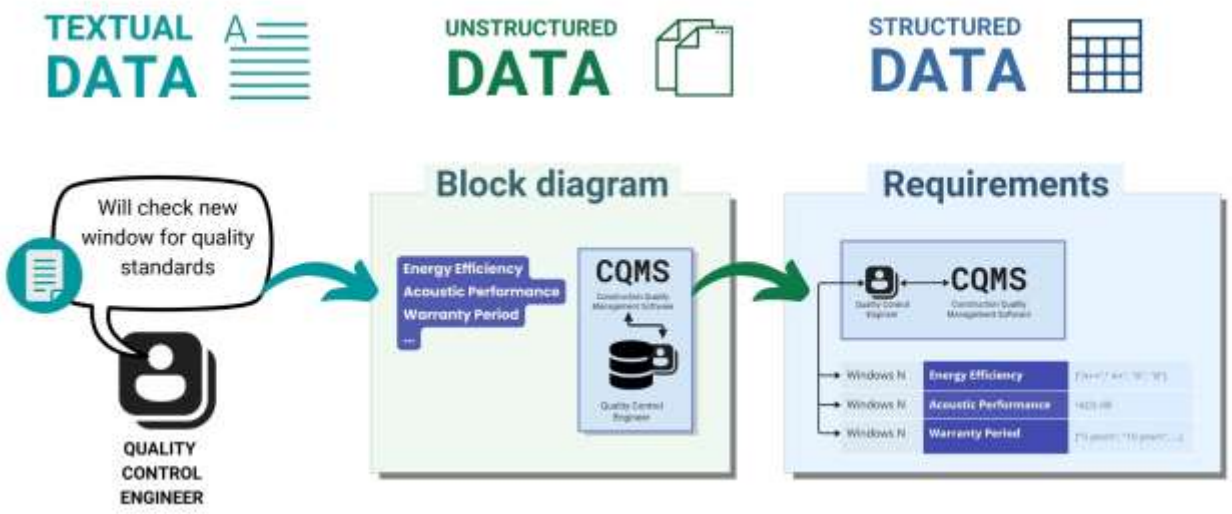
Đến 80% dữ liệu được tạo ra trong các công ty được trình bày ở định dạng không có cấu trúc hoặc nửa cấu trúc — văn bản, tài liệu, thư từ, tệp PDF, cuộc trò chuyện. Những dữ liệu này (Hình. 4.41) rất khó để phân tích, kiểm tra, truyền giữa các hệ thống và sử dụng trong tự động hóa. -

Để đảm bảo khả năng quản lý, tính minh bạch và tự động kiểm tra, cần chuyển đổi các yêu cầu văn bản và nửa cấu trúc thành các định dạng rõ ràng, có cấu trúc. Quá trình cấu trúc không chỉ liên quan đến dữ liệu (điều mà chúng tôi đã xem xét chi tiết trong các chương đầu của phần này của cuốn sách), mà còn liên quan đến chính các yêu cầu mà những người tham gia dự án thường diễn đạt dưới dạng văn bản tự do trong suốt vòng đời của dự án, thường không nghĩ rằng các quy trình này có thể được tự động hóa.

> Tương tự như cách chúng tôi đã chuyển đổi dữ liệu từ hình thức văn bản không có cấu trúc sang có cấu trúc, trong quá trình làm việc với các yêu cầu, chúng tôi sẽ chuyển đổi các yêu cầu văn bản thành định dạng cấu trúc "cấp độ logic và vật lý".

Trong ví dụ về việc bổ sung cửa sổ (Hình. 4.41), bước tiếp theo sẽ là mô tả các yêu cầu dữ liệu dưới dạng bảng. Chúng tôi sẽ cấu trúc thông tin cho mỗi hệ thống được sử dụng bởi các bên tham gia dự án, chỉ ra các thuộc tính chính và giá trị biên của chúng.

Hãy xem xét một trong những hệ thống như vậy (Hình. 4.45) — hệ thống quản lý chất lượng xây dựng (CQMS), được sử dụng bởi kỹ sư kiểm soát chất lượng từ phía khách hàng. Với nó, anh ta kiểm tra xem phần tử mới của dự án — trong trường hợp này là "cửa sổ mới" — có đáp ứng các tiêu chuẩn và yêu cầu đã đặt ra hay không.



Hình 4.45 Chuyển đổi yêu cầu văn bản thành định dạng bảng với mô tả thuộc tính của các thực thể giúp dễ dàng cho những chuyên gia khác hiểu.

Làm ví dụ, hãy xem xét một số yêu cầu quan trọng về thuộc tính của các thực thể loại "hệ thống cửa sổ" trong hệ thống CQMS (Hình 4.46): hiệu suất năng lượng, đặc tính cách âm và thời gian bảo hành. Mỗi danh mục bao gồm các tiêu chuẩn và thông số kỹ thuật nhất định mà cần phải xem xét khi thiết kế và lắp đặt hệ thống cửa sổ.-



Hình 4.46 Kỹ sư kiểm soát chất lượng cần kiểm tra các bộ phận mới loại "Cửa sổ" để đảm bảo chúng đáp ứng các tiêu chuẩn về hiệu suất năng lượng, cách âm và dịch vụ bảo hành.

Các yêu cầu về dữ liệu mà kỹ sư kiểm soát chất lượng đặt ra dưới dạng bảng có, ví dụ, các giá trị biên dạng sau:

- Lớp hiệu suất năng lượng của cửa sổ dao động từ "A++", được coi là hiệu quả cao nhất, đến "B", được coi là mức tối thiểu chấp nhận được, và các lớp này được đại diện bởi danh sách các giá trị hợp lệ ["A++", "A+", "A", "A", "B"].
- Cách âm của cửa sổ, được đo bằng decibel và cho thấy khả năng giảm tiếng ồn từ bên ngoài,

được xác định bằng biểu thức chính quy \d{2}dB.

- Thuộc tính "Thời gian bảo hành" cho thực thể "Loại cửa sổ" bắt đầu từ năm năm, xác định thời gian này là mức tối thiểu chấp nhận được khi lựa chọn sản phẩm; cũng có thể chỉ ra các giá trị thời gian bảo hành, ví dụ ["5 năm", "10 năm", v.v.] hoặc điều kiện logic ">5 (năm)".

Theo các yêu cầu đã được thu thập, trong khuôn khổ các thuộc tính đã được thiết lập, các bộ phận mới thuộc loại hoặc lớp “Cửa sổ” với lớp thấp hơn "B", chẳng hạn như "C" hoặc "D", sẽ không vượt qua kiểm tra về hiệu suất năng lượng. Cách âm của cửa sổ trong dữ liệu hoặc tài liệu gửi đến kỹ sư kiểm soát chất lượng phải được chỉ định bằng một số hai chữ số, theo sau là hậu tố "dB", chẳng hạn như "35 dB" hoặc "40 dB", trong khi các giá trị ngoài định dạng này, như "9 D B" hoặc "100 Decibel", sẽ không được chấp nhận (vì không vượt qua mẫu của chuỗi RegEx). Thời gian bảo hành phải bắt đầu ít nhất từ "5 năm", trong khi cửa sổ có thời gian bảo hành ngắn hơn, chẳng hạn như "3 năm" hoặc "4 năm", sẽ không đáp ứng các yêu cầu mà kỹ sư kiểm soát chất lượng đã mô tả trong định dạng bảng.

Để kiểm tra các giá trị của thuộc tính tham số cho sự phù hợp với các giá trị biên được đưa ra trong yêu cầu trong quá trình xác thực, chúng tôi sử dụng hoặc danh sách các giá trị chấp nhận được (["A", "B", "C"]), từ điển (["A": "H1", "H2"; "B": "W1", "W2"]), các phép toán logic (chẳng hạn như ">", "<", "<=", ">=", "==") cho các giá trị số) và biểu thức chính quy (cho các giá trị chuỗi và văn bản như trong thuộc tính "Hiệu suất cách âm"). Biểu thức chính quy là công cụ cực kỳ quan trọng trong việc làm việc với các giá trị chuỗi.

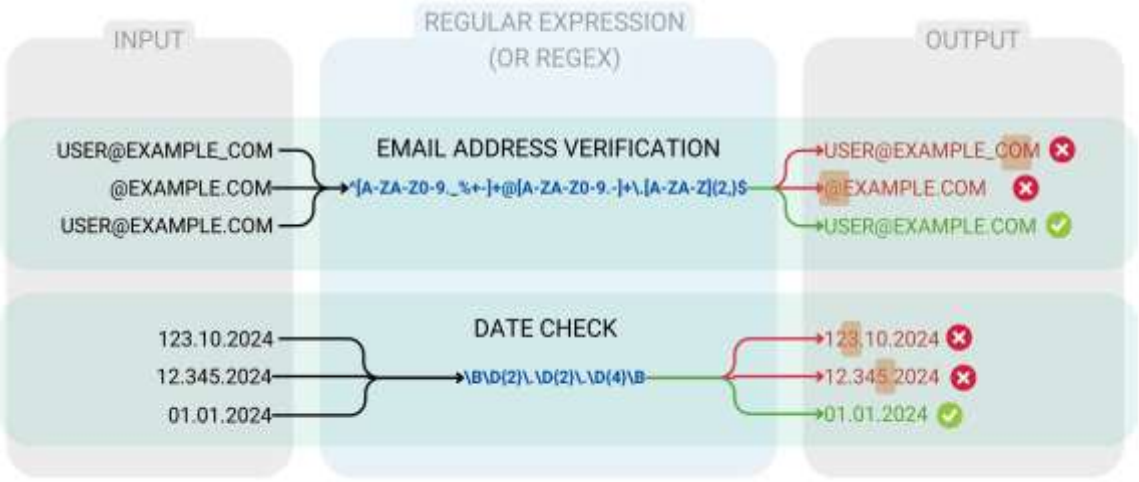
> Biểu thức chính quy (RegEx) được sử dụng trong các ngôn ngữ lập trình, bao gồm Python (thư viện Re), để tìm kiếm và chỉnh sửa chuỗi. Regex giống như một thám tử trong thế giới chuỗi, có khả năng xác định các mẫu văn bản trong văn bản một cách chính xác.

Trong biểu thức chính quy, các chữ cái được mô tả trực tiếp bằng các ký tự tương ứng của bảng chữ cái, trong khi các số có thể được biểu thị bằng ký hiệu đặc biệt \d, tương ứng với bất kỳ chữ số nào từ 0 đến 9. Dấu ngoặc vuông được sử dụng để chỉ định khoảng các chữ cái hoặc chữ số, ví dụ, [a-z] cho bất kỳ chữ cái thường nào của bảng chữ cái Latin hoặc [0-9], tương đương với \d. Đối với các ký tự không phải số và không phải chữ cái, ta sử dụng \D và \W tương ứng.

Một số ví dụ phổ biến về việc sử dụng RegEx (Hình. 4.47): -

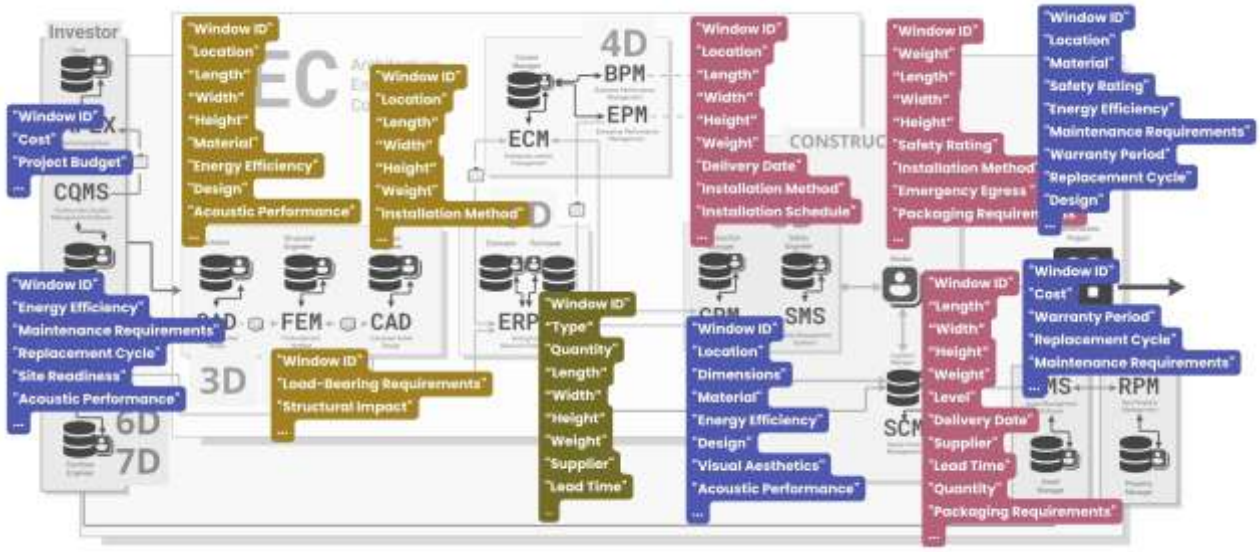
- Kiểm tra địa chỉ email: để kiểm tra xem một chuỗi có phải là địa chỉ email hợp lệ hay không, ta có thể sử dụng mẫu “^[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.]+\.[a-zA-Z]{2,}$”.
- Trích xuất ngày: mẫu “\b\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\.\d{2}\.\d{4}\b” có thể được sử dụng để trích xuất ngày từ văn bản theo định dạng DD.MM.YYYY.
- Kiểm tra số điện thoại: để kiểm tra số điện thoại theo định dạng +49(000)000-0000, mẫu sẽ trông như “\+\d{2}\(\d{3}\)\d{3}-\d{4}”.

Bằng cách chuyển đổi các yêu cầu của kỹ sư kiểm soát chất lượng thành định dạng thuộc tính và các giá trị giới hạn (Hình. 4.46), chúng tôi đã biến chúng từ định dạng văn bản ban đầu (các cuộc trò chuyện, thư từ và tài liệu quy định) thành một bảng có tổ chức và cấu trúc, từ đó cho phép kiểm tra và phân tích tự động bất kỳ dữ liệu nào được gửi đến (ví dụ các yếu tố mới trong danh mục “Cửa sổ”). Việc có yêu cầu cho phép tự động loại bỏ các dữ liệu không đạt tiêu chuẩn, trong khi dữ liệu đã được kiểm tra sẽ tự động được chuyển đến các hệ thống để xử lý tiếp theo.-



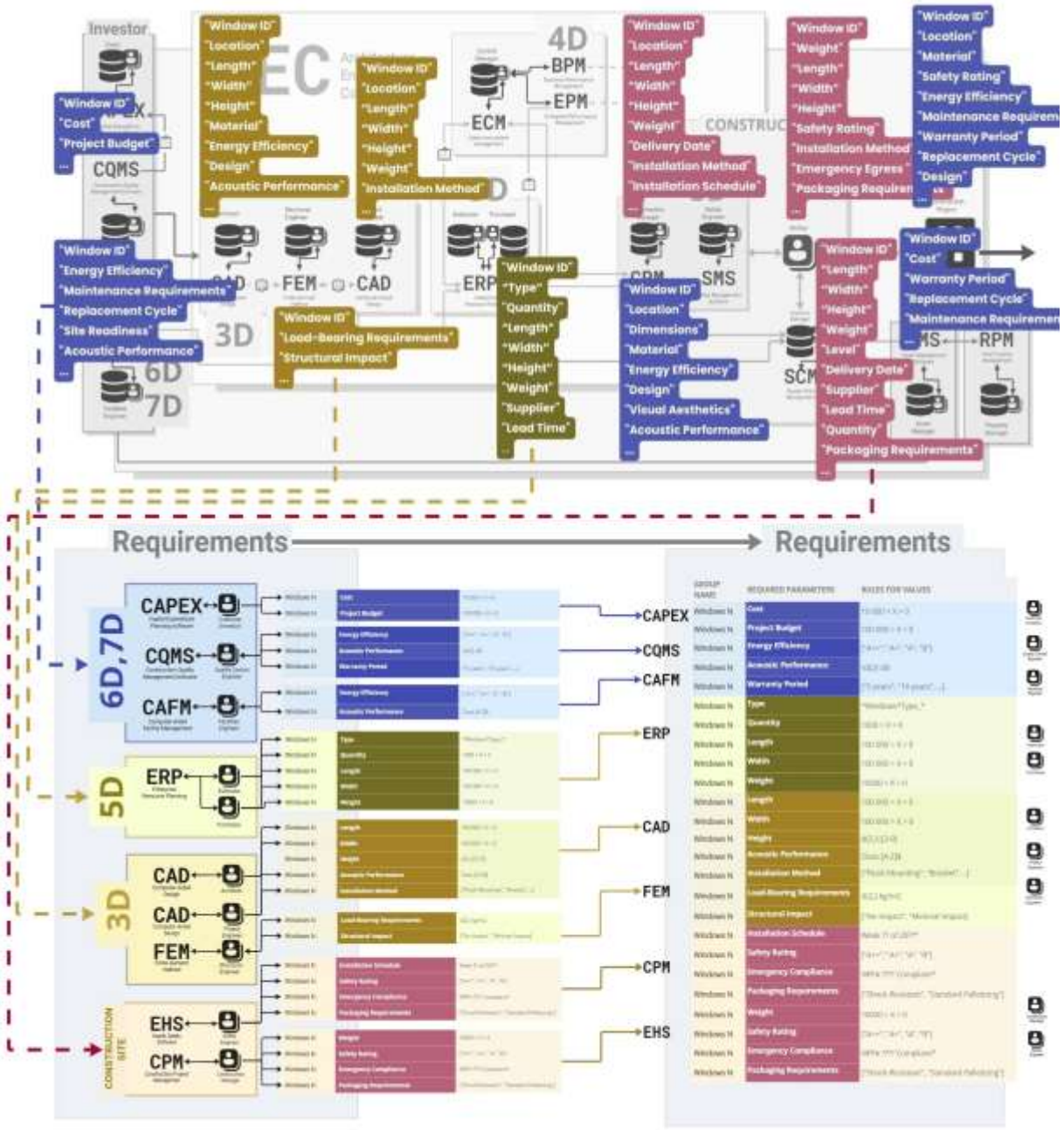
Hình. 4.47 Việc sử dụng biểu thức chính quy là một công cụ cực kỳ quan trọng trong quy trình kiểm tra dữ liệu văn bản.

Giờ đây, chuyển từ mức độ khái niệm sang mức độ logic trong công việc với yêu cầu, chúng tôi sẽ chuyển tất cả các yêu cầu của các chuyên gia trong quy trình lắp đặt cửa sổ mới (Hình. 4.44) thành một danh sách có thứ tự theo định dạng thuộc tính và thêm các danh sách này với các thuộc tính cần thiết vào sơ đồ khối của chúng tôi cho mỗi chuyên gia (Hình. 4.48). --



Hình. 4.48 Ở mức độ logic của quy trình, các thuộc tính mà mỗi chuyên gia xử lý sẽ được thêm vào các hệ thống tương ứng.

Bằng cách thêm tất cả các thuộc tính vào một bảng chung của quy trình, chúng tôi đã chuyển đổi thông tin được trình bày trước đó dưới dạng văn bản và cuộc đối thoại ở mức độ khái niệm (Hình. 4.41) thành một dạng thức có cấu trúc và hệ thống hóa ở mức độ vật lý (Hình. 4.49).



Hình. 4.49 Việc chuyển đổi cuộc đối thoại không có cấu trúc giữa các chuyên gia thành các bảng có cấu trúc giúp hiểu rõ hơn các yêu cầu ở mức độ vật lý.

Bây giờ, các yêu cầu về dữ liệu cần phải được truyền đạt đến các chuyên gia tạo ra thông tin cho các hệ thống cụ thể. Ví dụ, nếu bạn đang làm việc trong cơ sở dữ liệu CAD, thì trước khi bắt đầu mô hình hóa các yếu tố, bạn nên thu thập tất cả các thông số cần thiết dựa trên kịch bản sử dụng cuối cùng cho các dữ liệu này. Thông thường, điều này bắt đầu từ giai đoạn vận hành, sau đó là công trường xây dựng, bộ phận logistics, bộ phận dự toán, bộ phận tính toán kết cấu và cứ như thế. Chỉ sau khi bạn đã tính đến yêu cầu của tất cả các liên kết này, bạn mới có thể bắt đầu tạo dữ liệu - dựa trên các thông số đã thu thập. Điều này sẽ cho phép tự động hóa việc kiểm tra và truyền tải dữ liệu dọc theo chuỗi.

Khi dữ liệu mới phù hợp với các yêu cầu đã đặt ra, chúng sẽ tự động được tích hợp vào hệ sinh thái dữ liệu của công ty, được gửi trực tiếp đến những người dùng và hệ thống mà chúng được thiết kế cho. Việc xác thực dữ liệu về sự hiện diện và tính tương ứng của các thuộc tính và giá trị của chúng đảm bảo rằng thông tin đáp ứng các tiêu chuẩn chất lượng cần thiết và sẵn sàng để sử dụng trong các kịch bản của công ty.

Các yêu cầu về dữ liệu đã được xác định, và bây giờ, trước khi bắt đầu quá trình kiểm tra, cần phải tạo ra, nhận hoặc thu thập những dữ liệu sẽ được kiểm tra, hoặc ghi lại trạng thái hiện tại của thông tin trong cơ sở dữ liệu để sử dụng trong quá trình kiểm tra.

Thu thập dữ liệu cho quy trình kiểm tra

Trước khi bắt đầu kiểm tra, điều quan trọng là phải đảm bảo rằng dữ liệu có sẵn ở dạng phù hợp cho quá trình xác thực. Điều này không chỉ có nghĩa là thông tin có sẵn mà còn là sự chuẩn bị của nó: dữ liệu cần phải được thu thập và chuyển đổi từ các định dạng phi cấu trúc, yếu cấu trúc, văn bản và hình học sang dạng cấu trúc. Quá trình này đã được mô tả chi tiết trong các chương trước, nơi đã đề cập đến các phương pháp chuyển đổi các loại dữ liệu khác nhau. Sau tất cả các bước chuyển đổi, dữ liệu đầu vào sẽ có dạng bảng cấu trúc mở (Hình 4.12, Hình 4.19, Hình 4.113).-

Có được các yêu cầu và bảng cấu trúc với các thông số cần thiết và giới hạn giá trị (Hình 4.49), chúng ta có thể bắt đầu kiểm tra dữ liệu - dưới dạng một quy trình tự động duy nhất (Pipeline) hoặc theo định dạng kiểm tra từng tài liệu đầu vào theo từng bước.

Để khởi động kiểm tra, cần phải nhận một tệp mới hoặc ghi lại trạng thái dữ liệu hiện tại - tạo một ảnh chụp hoặc xuất hiện trạng thông tin hiện tại và đến, hoặc thiết lập kết nối với cơ sở dữ liệu bên ngoài hoặc bên trong. Trong ví dụ được đề cập, ảnh chụp này được tạo ra thông qua việc tự động hóa chuyển đổi dữ liệu CAD sang định dạng cấu trúc, được ghi lại, chẳng hạn, vào lúc 23:00:00 thứ Sáu, ngày 29 tháng 3 năm 2024, sau khi tất cả các kỹ sư thiết kế đã rời văn phòng.



Hình 4.410 Ảnh chụp từ cơ sở dữ liệu CAD (BIM), cho thấy thông tin hiện tại về các thuộc tính cho thực thể mới của lớp "Cửa sổ" trong phiên bản hiện tại của mô hình dự án.

Nhờ vào các công cụ đảo ngược, đã được xem xét trong chương "Chuyển đổi dữ liệu CAD (BIM) sang dạng cấu trúc", thông tin này từ các công cụ và trình chỉnh sửa CAD (BIM) có thể được tổ chức thành các bảng riêng biệt (Hình 4.411) hoặc tích hợp thành một bảng chung, kết hợp các phần khác nhau của dự án (Hình 9.110). --

Trong bảng cơ sở dữ liệu như vậy hiển thị các mã định danh duy nhất của cửa sổ và cửa ra vào (thuộc tính ID), tên kiểu (TypeName), kích thước (Width, Length), vật liệu (Material), cũng như các chỉ số hiệu quả năng lượng và âm thanh, cùng với các đặc điểm khác. Bảng như vậy, được điền trong chương trình CAD (BIM), được thu thập bởi kỹ sư thiết kế từ nhiều bộ phận và tài liệu khác nhau, hình thành mô hình thông tin của dự án.

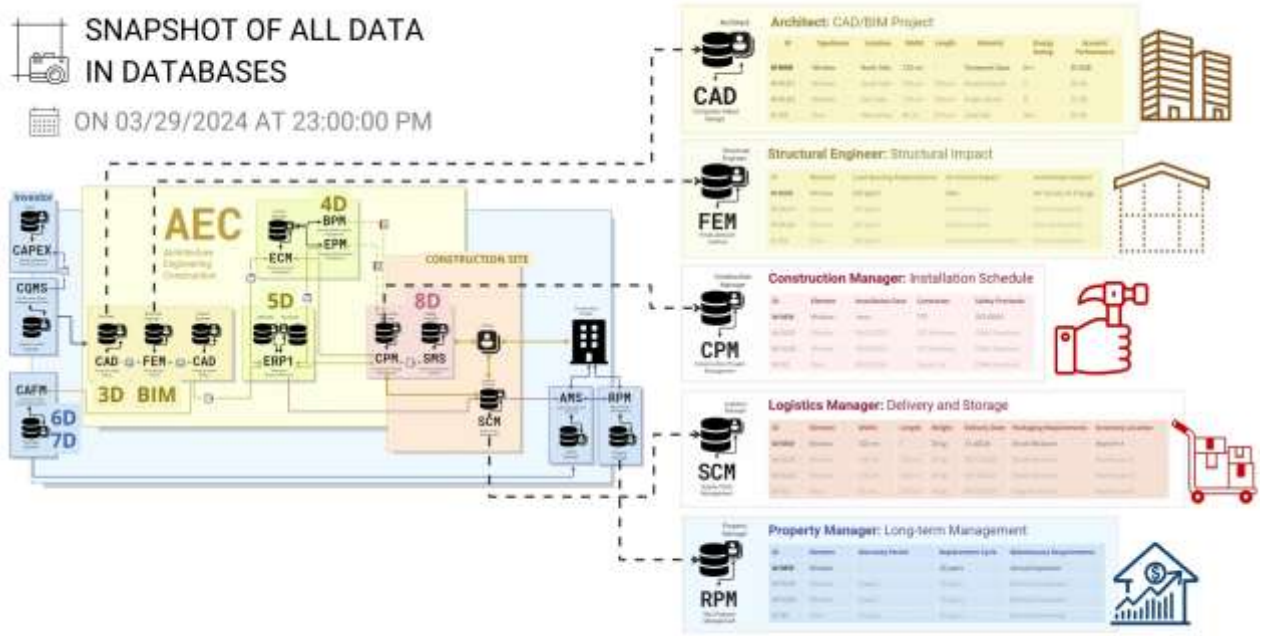


Hình 4.411 Dữ liệu cấu trúc từ các hệ thống CAD có thể là một bảng hai chiều với các cột biểu thị các thuộc tính của các yếu tố.

Các dự án CAD (BIM) thực tế bao gồm hàng chục hoặc hàng trăm ngàn yếu tố (Hình. 9.110). Các yếu tố trong các định dạng CAD được phân loại tự động theo loại và danh mục — từ cửa sổ và cửa ra vào đến gạch, sàn và tường. Các định danh duy nhất (ví dụ, ID gốc, được thiết lập tự động bởi giải pháp CAD) hoặc thuộc tính loại (Tên Loại, Loại, Gia Đình) cho phép theo dõi cùng một đối tượng trong các hệ thống khác nhau. Ví dụ, một cửa sổ mới trên tường phía bắc của tòa nhà có thể được ký hiệu duy nhất bằng một định danh "W-NEW" trong tất cả các hệ thống tương ứng của tổ chức.-

> Mặc dù tên và định danh của các thực thể cần phải nhất quán trong tất cả các hệ thống, tập hợp các thuộc tính và giá trị liên quan đến những thực thể này có thể khác nhau đáng kể tùy thuộc vào ngữ cảnh sử dụng. Các kiến trúc sư, kỹ sư kết cấu, chuyên gia xây dựng, logistics và quản lý bất động sản có cách nhìn khác nhau về cùng một yếu tố. Mỗi người dựa vào các phân loại, tiêu chuẩn và mục tiêu riêng: có người chỉ xem xét cửa sổ từ góc độ thẩm mỹ, đánh giá hình dáng và tỷ lệ của nó, trong khi người khác lại xem xét từ góc độ kỹ thuật hoặc khai thác, phân tích khả năng cách nhiệt, phương pháp lắp đặt, khối lượng hoặc yêu cầu bảo trì. Do đó, khi mô hình hóa dữ liệu và mô tả các yếu tố, cần xem xét tính đa dạng trong việc sử dụng của chúng và đảm bảo tính nhất quán của dữ liệu đồng thời xem xét các đặc thù của ngành.

Đối với mỗi vai trò trong các quy trình của công ty, có các cơ sở dữ liệu chuyên dụng với giao diện người dùng riêng — từ thiết kế và tính toán cho đến logistics, lắp đặt và vận hành tòa nhà (Hình. 4.412). Mỗi hệ thống như vậy được quản lý bởi một đội ngũ chuyên gia thông qua giao diện người dùng đặc biệt hoặc thông qua các truy vấn vào cơ sở dữ liệu, nơi tất cả các quyết định được đưa ra dựa trên các giá trị đầu vào, đứng sau đó là người quản lý hệ thống hoặc trưởng phòng, người chịu trách nhiệm về tính hợp pháp và chất lượng của dữ liệu đầu vào trước các đối tác của mình, phục vụ cho các hệ thống khác.



Hình. 4.412 Một thực thể duy nhất có định danh giống nhau trong các hệ thống khác nhau, nhưng có các thuộc tính khác nhau, điều này chỉ quan trọng trong hệ thống này.

Sau khi thu thập có tổ chức các yêu cầu và dữ liệu có cấu trúc ở cấp độ logic và vật lý, chúng ta cần thiết lập quy trình kiểm tra tự động dữ liệu thu được từ các tài liệu và hệ thống khác nhau, nhằm đảm bảo tuân thủ các yêu cầu đã được thu thập trước đó.

Kiểm tra dữ liệu và kết quả kiểm tra

Tất cả dữ liệu mới được nhập vào hệ thống — dù là tài liệu, bảng tính hay bản ghi trong cơ sở dữ liệu từ khách hàng, kiến trúc sư, kỹ sư, quản lý công trình, logistics hoặc quản lý bất động sản — đều phải trải qua quy trình kiểm tra để đảm bảo tuân thủ các yêu cầu đã được xác định trước (Hình. 4.49). Quy trình xác thực là rất quan trọng: bất kỳ sai sót nào trong dữ liệu có thể dẫn đến tính toán sai, trì hoãn tiến độ và thậm chí thiệt hại tài chính. Để giảm thiểu những rủi ro như vậy, cần tổ chức một quy trình kiểm tra dữ liệu có hệ thống và lặp đi lặp lại, theo từng giai đoạn.

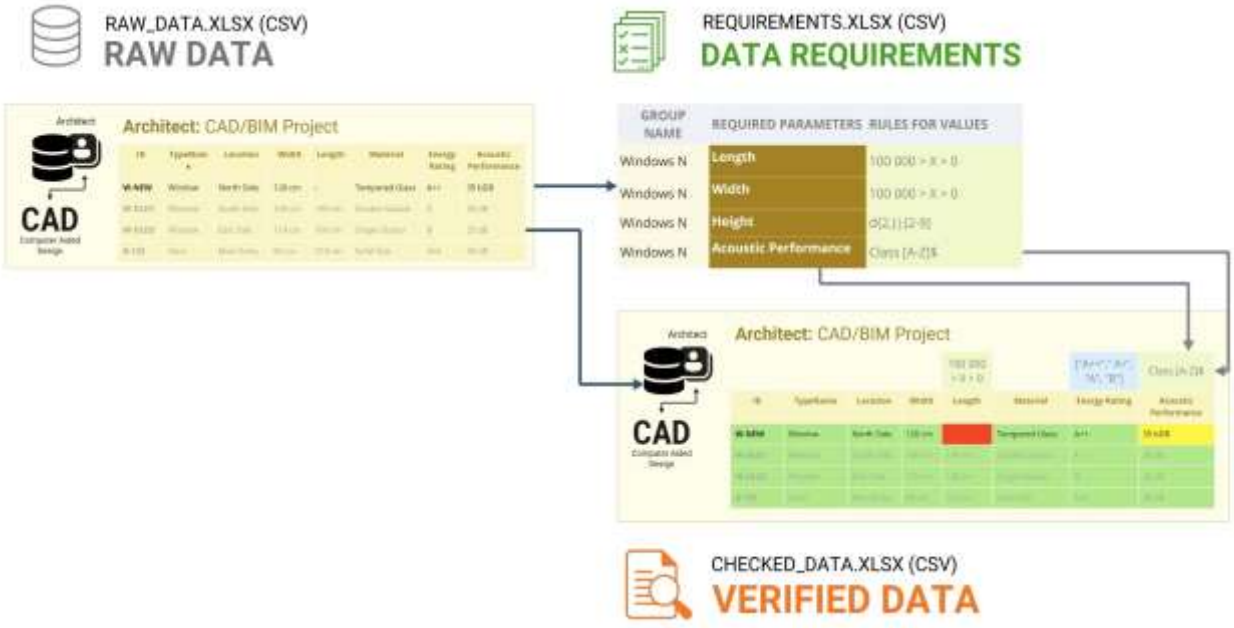
> Để kiểm tra dữ liệu mới vào hệ thống - bao gồm không có cấu trúc, văn bản hoặc hình học, - cần phải chuyển đổi chúng sang định dạng yếu cấu trúc hoặc cấu trúc. Sau đó, trong quy trình kiểm tra, các dữ liệu này cần được kiểm tra để đảm bảo tuân thủ danh sách đầy đủ các thuộc tính yêu cầu và các giá trị hợp lệ của chúng.

Biến đổi các loại dữ liệu khác nhau: văn bản, hình ảnh, tài liệu PDF và dữ liệu hỗn hợp CAD (BIM) thành định dạng có cấu trúc đã được chúng tôi xem xét chi tiết trong chương "Biến đổi dữ liệu thành định dạng có cấu trúc".

Ví dụ có thể dẫn ra là bảng dữ liệu thu được từ dự án CAD (BIM) (Hình 4.411). Nó bao gồm dữ liệu hình học

bán cấu trúc và thông tin thuộc tính có cấu trúc về các thực thể của dự án (Hình 3.114) - chẳng hạn như các phần tử từ lớp "Cửa sổ".--

Để thực hiện kiểm tra, chúng tôi sẽ đối chiếu các giá trị thuộc tính (Hình 4.411) với các giá trị giới hạn chuẩn được xác định bởi các chuyên gia dưới dạng yêu cầu (Hình 4.49). Bảng so sánh cuối cùng (Hình 4.413) sẽ giúp hiểu rõ những giá trị nào là hợp lệ và những giá trị nào cần phải chỉnh sửa trước khi dữ liệu có thể được sử dụng ngoài các ứng dụng CAD (BIM).-



Hình 4.413 Bảng kiểm tra cuối cùng làm nổi bật những giá trị thuộc tính cho thực thể mới thuộc lớp "Cửa sổ" mà cần chú ý.

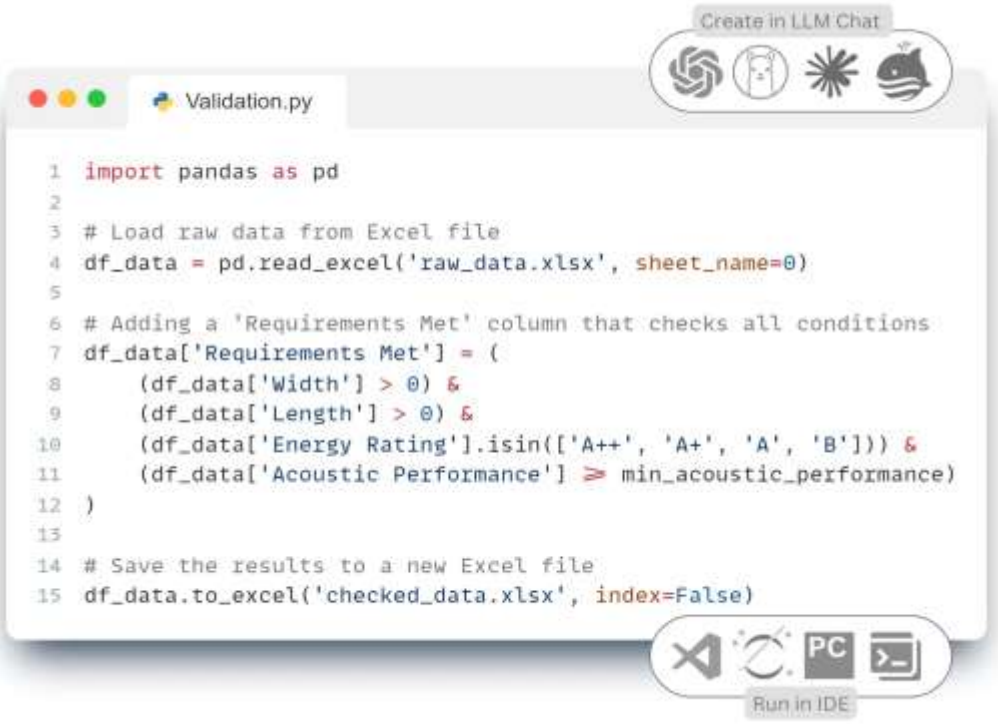
Thực hiện giải pháp tương tự, sử dụng thư viện Pandas mà chúng tôi đã đề cập trước đó trong chương "Pandas: Công cụ không thể thiếu cho phân tích dữ liệu", chúng tôi sẽ kiểm tra dữ liệu từ tệp bảng tính được trích xuất từ tệp CAD (BIM) (RVT, IFC, DWG, NWS, DGN) (Hình 4.411) bằng cách sử dụng các yêu cầu từ tệp bảng khác (Hình 4.49).--

Để lấy mã, chúng tôi cần mô tả trong lời nhắc cho LLM rằng cần tải dữ liệu từ tệp raw_data.xlsx (bộ dữ liệu đầy đủ từ cơ sở dữ liệu CAD (BIM)), kiểm tra chúng và lưu kết quả vào tệp mới checked_data.xlsx (Hình 4.413).-

- Chúng tôi sẽ nhận được mã thông qua LLM mà không đề cập đến thư viện Pandas:

> Viết mã kiểm tra bảng trong tệp raw_data.xlsx và kiểm tra chúng theo các quy tắc sau: giá trị của các cột 'Width' và 'Length' lớn hơn 0, 'Energy Rating' nằm trong danh sách ['A++', 'A+', 'A', 'B'], còn 'Acoustic Performance' như một biến mà chúng tôi sẽ chỉ định sau - kèm theo cột kiểm tra cuối cùng, và lưu bảng kết quả vào tệp Excel mới checked_data.xlsx ⏎

- Phản hồi từ LLM sẽ mô tả một ví dụ ngắn gọn về mã Python mà có thể được làm rõ và bổ sung qua các lời nhắc sau:



Hình 4.414 Mã được tạo ra bởi mô hình LLM kiểm tra dự án CAD (BIM) đã được biến đổi để xác nhận tính tuân thủ của các yêu cầu thuộc tính dưới dạng các giá trị giới hạn.

Mã do mô hình ngôn ngữ LLM tạo ra có thể được sử dụng trong bất kỳ IDE phổ biến nào hoặc trong các công cụ trực tuyến: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse với plugin PyDev, Thonny, Wing IDE, IntelliJ IDEA với plugin Python, JupyterLab hoặc các công cụ trực tuyến phổ biến như Kaggle.com, Google Colab, Microsoft Azure Notebooks, Amazon SageMaker.

Việc thực thi mã (Hình 4.414) sẽ cho thấy rằng "các phần tử-thực thể" W-OLD1, W-OLD2, D-122 (và các phần tử khác) từ cơ sở dữ liệu CAD (BIM) đáp ứng các yêu cầu về thuộc tính: chiều rộng và chiều dài lớn hơn 0, trong khi lớp hiệu suất năng lượng là một trong các giá trị trong danh sách 'A++', 'A', 'B', 'C' (Hình 4.415).-

Phần tử cần thiết và mới được thêm vào W-NEW, đại diện cho một thực thể mới thuộc lớp "Cửa sổ" ở phía bắc, không đáp ứng các yêu cầu (thuộc tính "Requirements Met"), vì chiều dài của nó bằng 0 (giá trị "0.0" được coi là không chấp nhận theo quy tắc 'Width'>0) và nó không chỉ định lớp hiệu suất năng lượng.



CHECKED_DATA.XLSX (CSV)

VERIFIED DATA

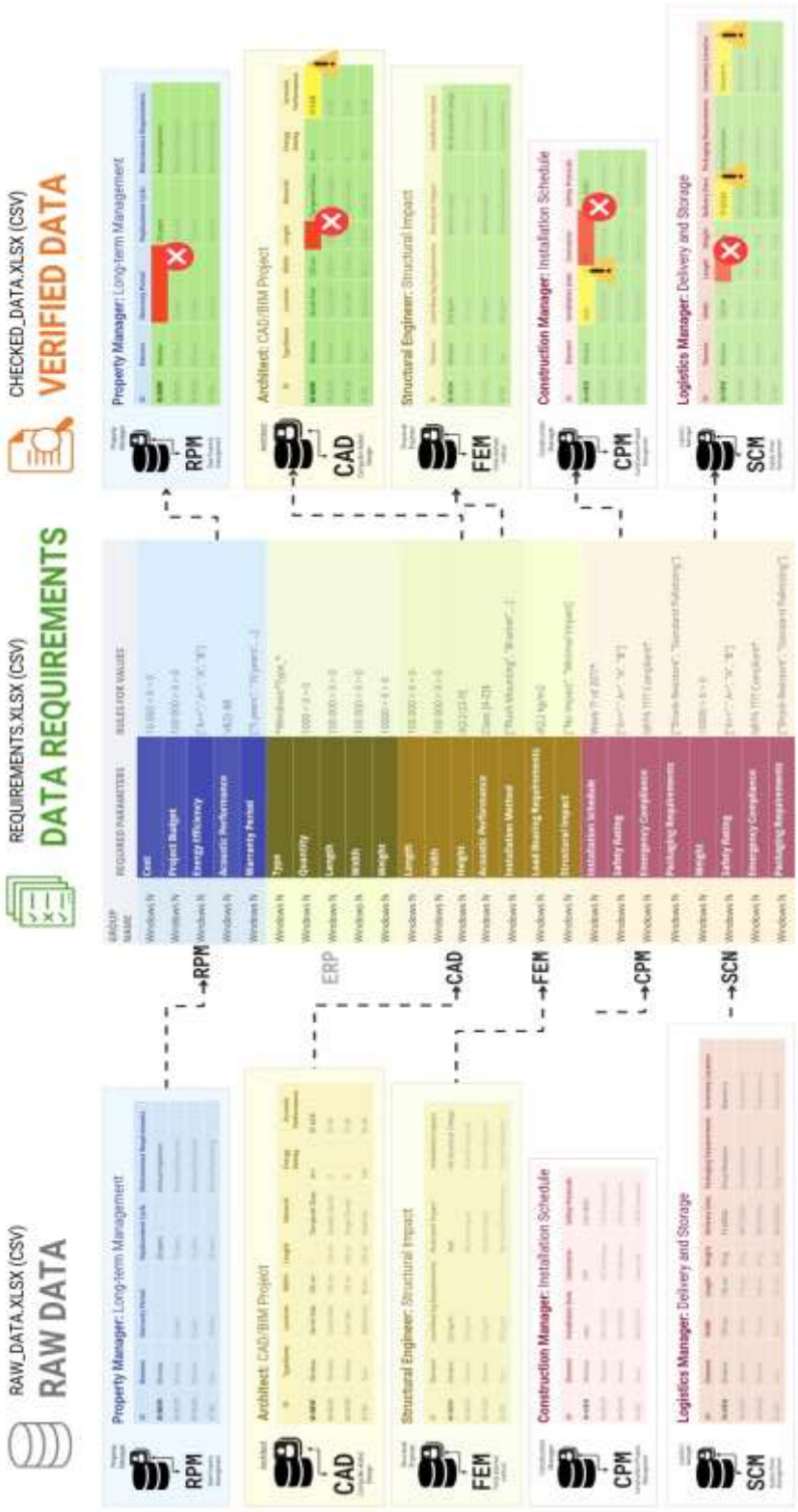
	ID	TypeName	Location	Width	Length	Material	Energy Rating	Acoustic Performance	Requirements Met
0	W-NEW	Window	North Side	120	0.0	Tempered Glass		35	False
1	W-OLD1	Window	South Side	100	140.0	Double Glazed	A++	30	True
2	W-OLD2	Window	East Side	110	160.0	Single Glazed	B	25	True
3	D-122	Door	Main Entry	90	210.0	Solid Oak	B	30	True

Hình 4.415 Kiểm tra phát hiện các thực thể không trải qua quá trình xác minh và bổ sung vào kết quả một thuộc tính mới với các giá trị 'Sai' hoặc 'Đúng'.

Tương tự, chúng tôi kiểm tra tính đồng nhất của tất cả các phần tử trong dự án (các thực thể) và các thuộc tính cần thiết cho mỗi hệ thống, bảng hoặc cơ sở dữ liệu trong tất cả dữ liệu mà chúng tôi nhận được từ các chuyên gia khác nhau (Hình 4.41) trong quá trình thêm cửa sổ vào dự án.

Trong bảng kết quả cuối cùng, để hình ảnh hóa dễ dàng hơn, chúng tôi đánh dấu kết quả kiểm tra bằng màu sắc: màu xanh lá cây đánh dấu các thuộc tính đã kiểm tra thành công, màu vàng đánh dấu các giá trị với sai lệch không nghiêm trọng, và màu đỏ đánh dấu các sự không phù hợp nghiêm trọng (Hình 4.416).

Như một kết quả của việc kiểm tra (Hình 4.416), chúng tôi thu được danh sách các phần tử đáng tin cậy và đã được kiểm tra cùng với các định danh của chúng, được kiểm tra để đảm bảo tính phù hợp với các yêu cầu về thuộc tính. Các phần tử đã được kiểm tra đảm bảo rằng chúng phù hợp với các tiêu chuẩn và thông số đã được công bố cho tất cả các hệ thống tham gia vào quá trình thêm các phần tử thuộc lớp "Cửa sổ" hoặc bất kỳ lớp nào khác (thêm thông tin về tự động hóa kiểm tra dữ liệu và tạo quy trình ETL tự động sẽ được trình bày trong chương "Tự động hóa ETL và kiểm tra dữ liệu").



Hình 4.416 Kết quả kiểm tra thực hiện cho tất cả các hệ thống cho phép xác định những dữ liệu không phù hợp với yêu cầu của công ty.

Các thực thể đã vượt qua kiểm tra thường không cần sự chú ý nhiều hơn. Chúng dễ dàng chuyển sang các giai đoạn tiếp theo của xử lý và tích hợp vào các hệ thống khác. Ngược lại với các phần tử "chất lượng", chính các phần tử không vượt qua kiểm tra lại được quan tâm nhiều nhất. Thông tin về những sai lệch như vậy là vô cùng quan trọng: nó không chỉ cần được truyền đạt dưới dạng báo cáo bảng, mà còn có thể sử dụng nhiều công cụ hình ảnh khác nhau. Việc trình bày đồ họa kết quả kiểm tra giúp đánh giá nhanh chóng tình trạng chất lượng dữ liệu tổng thể, phát hiện các lĩnh vực vấn đề và kịp thời đưa ra các biện pháp sửa chữa hoặc điều chỉnh các thông số.

Hình dung kết quả kiểm tra

Hình ảnh hóa là công cụ quan trọng nhất để giải thích kết quả kiểm tra. Bên cạnh các bảng tổng hợp thông thường, nó cũng có thể bao gồm các bảng thông tin, biểu đồ, và các tài liệu PDF tự động được tạo ra, trong đó các phần tử của dự án được nhóm theo trạng thái kiểm tra. Mã màu có thể đóng vai trò hỗ trợ trong trường hợp này: màu xanh lá cây có thể biểu thị các đối tượng đã được kiểm tra thành công, màu vàng – các phần tử cần sự chú ý thêm, và màu đỏ – những phần tử có lỗi nghiêm trọng hoặc thiếu dữ liệu quan trọng.

Trong ví dụ của chúng tôi (Hình 4.41), chúng tôi phân tích dữ liệu từ mỗi hệ thống theo từng bước: từ CAD (BIM) và quản lý bất động sản đến logistics và lịch trình lắp đặt (Hình 4.416). Sau khi hoàn thành kiểm tra, mỗi chuyên gia sẽ được tự động tạo ra thông báo cá nhân hoặc tài liệu báo cáo, chẳng hạn như định dạng PDF (Hình 4.417). Nếu dữ liệu là chính xác, chuyên gia sẽ nhận được một thông báo ngắn gọn: "Cảm ơn bạn đã hợp tác." Trong trường hợp phát hiện sự không phù hợp, sẽ gửi một báo cáo chi tiết với nội dung: "Tài liệu này liệt kê các phần tử, định danh của chúng, thuộc tính và giá trị không vượt qua kiểm tra sự phù hợp."---



Hình 4.417 Việc xác minh và tự động tạo tài liệu báo cáo giúp tăng tốc quá trình tìm kiếm và hiểu các thiếu sót dữ liệu cho chuyên gia tạo ra dữ liệu.

Nhờ vào quy trình tự động kiểm tra - ngay khi phát hiện ra lỗi hoặc thiếu sót trong dữ liệu, một thông báo sẽ được gửi ngay lập tức dưới dạng tin nhắn trong chat, qua email hoặc tài liệu PDF đến người chịu trách nhiệm tạo lập hoặc xử lý các thực thể và thuộc tính liên quan (Hình 4.418), kèm theo danh sách các phần tử và mô tả thuộc tính không qua kiểm tra.-



Hình 4.418 Các báo cáo tự động về kết quả kiểm tra giúp dễ dàng hiểu các lỗi và tăng tốc độ công việc điền dữ liệu dự án.

Ví dụ, nếu một tài liệu được đưa vào hệ thống quản lý bất động sản (sau khi đã được cấu trúc), trong đó thuộc tính "Thời hạn bảo hành" được điền sai, quản lý bất động sản sẽ nhận được thông báo có danh sách các thuộc tính cần được kiểm tra và sửa chữa.

Tương tự, bất kỳ sai sót nào trong lịch trình lắp đặt hoặc dữ liệu logistics đều dẫn đến việc tự động hình thành một báo cáo và chẳng hạn như gửi thông báo trong chat hoặc email với kết quả kiểm tra đến chuyên gia liên quan.

Bên cạnh tài liệu PDF và biểu đồ kết quả, còn có thể tạo ra bảng điều khiển (dashboard) và mô hình 3D tương tác (Hình 7.16, Hình 7.212) với việc làm nổi bật các phần tử có thiếu sót thuộc tính, cho phép người dùng sử dụng trực quan hình học 3D của các phần tử để lọc và đánh giá chất lượng và tính đầy đủ của dữ liệu các phần tử trong dự án.--

> Việc trực quan hóa kết quả kiểm tra dưới dạng tài liệu, biểu đồ hoặc bảng điều khiển tự động tạo ra giúp đơn giản hóa việc giải thích dữ liệu và thúc đẩy sự tương tác hiệu quả giữa các thành viên trong dự án.

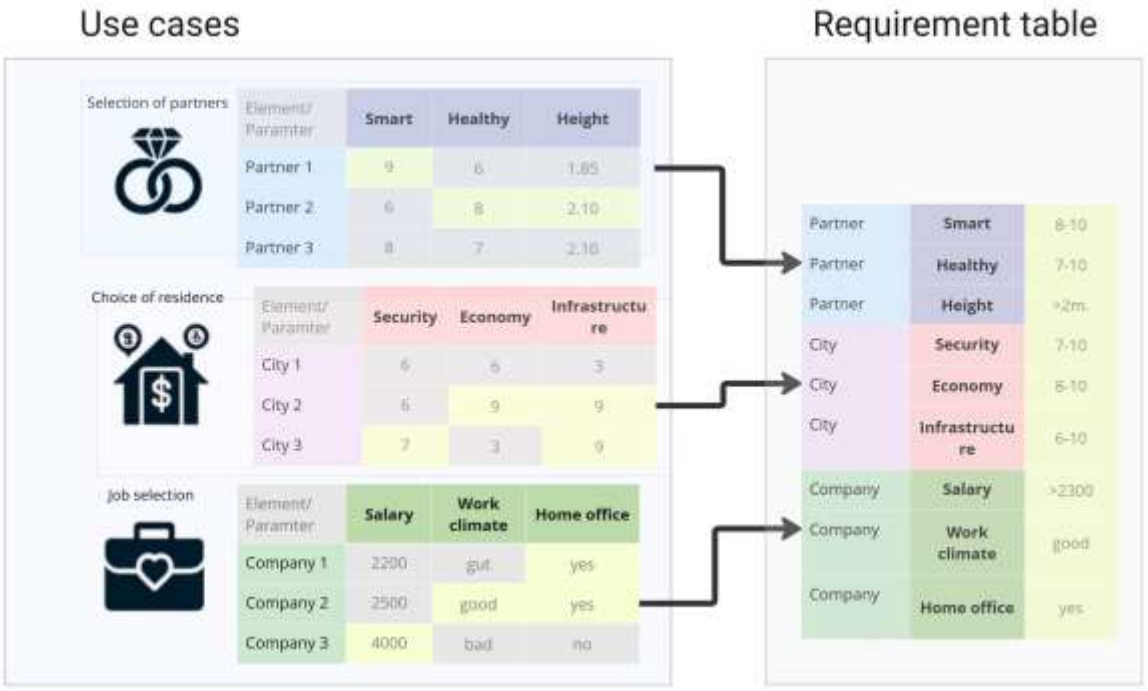
Quy trình kiểm tra dữ liệu tự động từ nhiều hệ thống và nguồn thông tin khác nhau có thể so sánh với việc ra quyết định có ý thức trong đời sống hàng ngày. Tương tự như các công ty trong ngành xây dựng xem xét nhiều biến số - từ độ tin cậy của dữ liệu đầu vào đến ảnh hưởng của nó đến thời gian, chi phí và chất lượng thực hiện dự án, - con người khi đưa ra các quyết định quan trọng, chẳng hạn như lựa chọn nơi cư trú, cân nhắc một loạt các yếu tố: khả năng tiếp cận giao thông, cơ sở hạ tầng, chi phí, an ninh, chất lượng cuộc sống. Tất cả những điều này hình thành nên một hệ thống tiêu chí, dựa trên đó mà các quyết định cuối cùng trong cuộc sống của chúng ta được đưa ra.

So sánh kiểm tra chất lượng dữ liệu với nhu cầu thực tế của con người

Mặc dù các phương pháp và công cụ kiểm soát chất lượng dữ liệu không ngừng phát triển, nguyên tắc cơ bản - sự phù hợp của thông tin với các yêu cầu đã đặt ra - vẫn không thay đổi. Nguyên tắc này được tích hợp vào nền tảng của một hệ thống quản lý trưởng thành, bất kể trong kinh doanh hay trong đời sống hàng ngày.

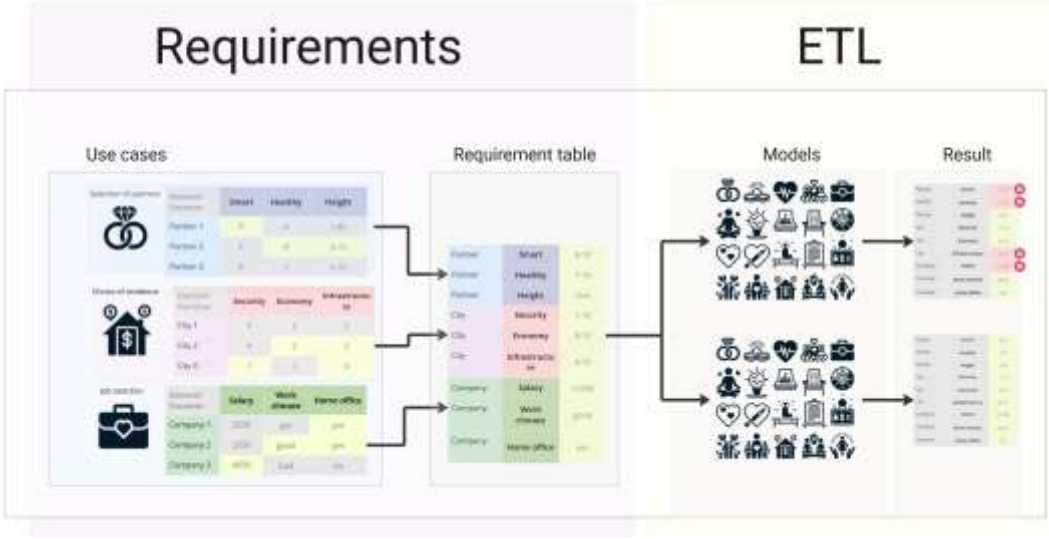
> Quy trình kiểm tra dữ liệu lặp đi lặp lại rất giống với quy trình ra quyết định mà mỗi người trong chúng ta phải đối mặt hàng ngày. Trong cả hai trường hợp, chúng ta dựa vào kinh nghiệm đã tích lũy, dữ liệu có sẵn và thông tin mới tiếp nhận. Hơn nữa, ngày càng nhiều quyết định trong cuộc sống và công việc - từ chiến lược đến hàng ngày - được đưa ra dựa trên dữ liệu.

Ví dụ, khi lựa chọn nơi cư trú hoặc bạn đời, chúng ta vô thức hình thành trong tâm trí một bảng tiêu chí và đặc tính mà chúng ta so sánh các lựa chọn (Hình 4.419). Những đặc tính này - dù là phẩm chất cá nhân hay tham số của bất động sản - đều là những thuộc tính ảnh hưởng đến quyết định cuối cùng.



Hình 4.419 Việc chọn nơi cư trú, công việc hoặc quan hệ đối tác dựa trên các yêu cầu cá nhân về thuộc tính.

Việc sử dụng dữ liệu có cấu trúc và cách tiếp cận chính thức để mô tả yêu cầu (Hình 4.420) hỗ trợ lựa chọn hợp lý hơn và có ý thức hơn cả trong hoạt động nghề nghiệp lẫn cuộc sống cá nhân.

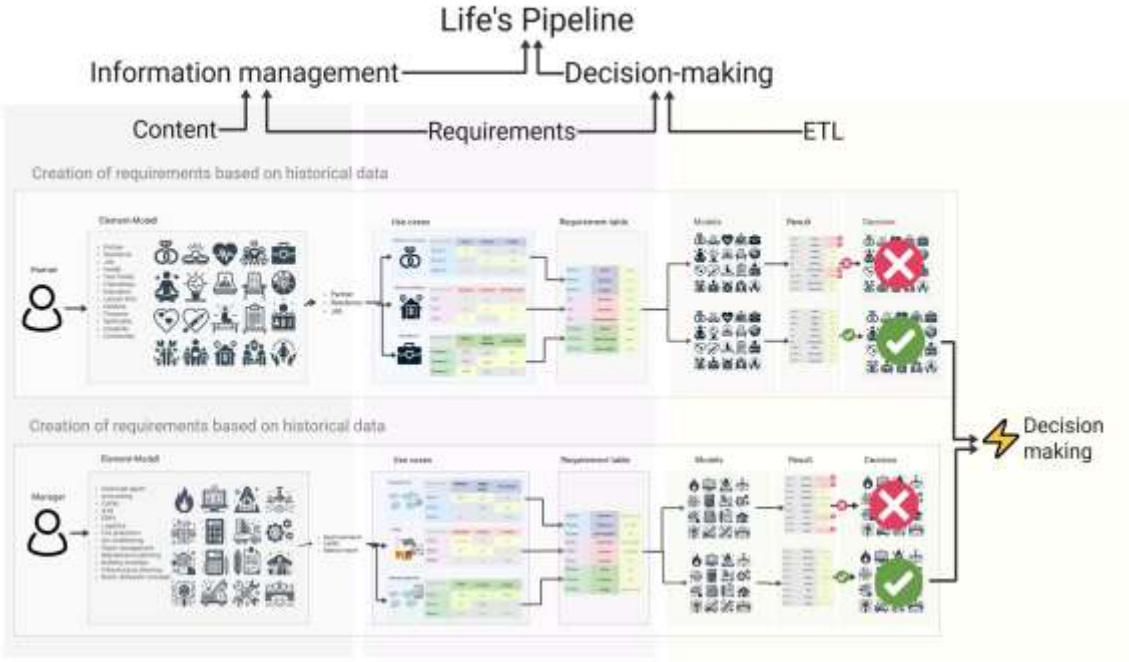


Hình 4.420 Việc chính thức hóa các yêu cầu giúp hệ thống hóa nhận thức về các quyết định trong cuộc sống và kinh doanh.

Cách tiếp cận ra quyết định dựa trên dữ liệu không chỉ là công cụ cho doanh nghiệp. Nó được tích hợp tự nhiên vào cuộc sống hàng ngày, theo các bước xử lý dữ liệu giống như quy trình ETL (Extract, Transform, Load), mà chúng ta đã xem xét ở đầu phần này khi cấu trúc dữ liệu và sẽ được thảo luận chi tiết trong ngữ cảnh tự động hóa các nhiệm vụ trong phần bảy của cuốn sách:-

- Dữ liệu như nền tảng (Extract): trong bất kỳ lĩnh vực nào - dù là công việc hay cuộc sống cá nhân - chúng ta thu thập thông tin. Trong kinh doanh, đó có thể là các báo cáo, chỉ số, dữ liệu thị trường; trong đời sống cá nhân - là kinh nghiệm cá nhân, lời khuyên từ người thân, đánh giá, quan sát.
- Tiêu chí đánh giá (Transform): thông tin thu thập được được diễn giải dựa trên các tiêu chí đã xác định trước. Ở nơi làm việc - đó là chỉ số hiệu suất (KPI), giới hạn ngân sách và tiêu chuẩn; trong cuộc sống cá nhân - các yếu tố như giá cả, vị trí thuận tiện, độ tin cậy, sự quyến rũ, v.v.
- Dự đoán và phân tích rủi ro (Load): ở giai đoạn cuối cùng, quyết định được đưa ra dựa trên phân tích dữ liệu đã được biến đổi và so sánh các hậu quả tiềm năng. Điều này tương tự như các quy trình kinh doanh, nơi dữ liệu được lọc qua lăng kính của logic kinh doanh và rủi ro.

> Các quyết định mà chúng ta đưa ra - từ những sở thích đơn giản như ăn gì cho bữa sáng, đến những sự kiện quan trọng trong cuộc đời như chọn một nghề nghiệp hay bạn đời - bản chất là kết quả của việc xử lý và đánh giá dữ liệu.



Hình 4.421 Kinh doanh và cuộc sống nói chung là một chuỗi quyết định dựa trên dữ liệu, trong đó chất lượng dữ liệu được sử dụng để đưa ra quyết định là yếu tố then chốt.

Tất cả trong cuộc sống của chúng ta đều liên kết với nhau, và giống như các sinh vật sống, bao gồm cả con người, tuân theo các quy luật tự nhiên, tiến hóa và thích ứng với điều kiện thay đổi, các quá trình của con người, bao gồm các phương pháp thu thập và phân tích dữ liệu, phản ánh những nguyên tắc tự nhiên này. Sự tương tác chặt chẽ giữa thiên nhiên và hoạt động của con người không chỉ khẳng định sự phụ thuộc của chúng ta vào thiên nhiên mà còn là mong muốn của chúng ta áp dụng các quy luật được mài giũa qua hàng triệu năm tiến hóa trong việc tạo ra kiến trúc dữ liệu, phát triển các quy trình và hệ thống để ra quyết định.

Các công nghệ mới, đặc biệt trong xây dựng, là ví dụ rõ ràng về việc nhân loại ngày càng được truyền cảm hứng từ thiên nhiên để tạo ra các giải pháp tốt nhất, bền vững và hiệu quả hơn.

Các bước tiếp theo: biến dữ liệu thành tính toán và kế hoạch chính xác

Trong phần này, chúng tôi đã xem xét cách chuyển đổi dữ liệu không cấu trúc thành định dạng có cấu trúc, phát triển mô hình dữ liệu và tổ chức quy trình kiểm tra chất lượng thông tin trong các dự án xây dựng. Quản lý dữ liệu, tiêu chuẩn hóa và phân loại dữ liệu là quá trình cơ bản, đòi hỏi một phương pháp hệ thống và hiểu biết rõ ràng về yêu cầu kinh doanh. Các phương pháp và công cụ đã được xem xét trong phần này cho phép đảm bảo tích hợp đáng tin cậy giữa các hệ thống khác nhau trong suốt vòng đời của dự án.

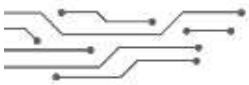
Tổng kết phần này, chúng ta hãy chỉ ra những bước thực tiễn chính giúp áp dụng các cách tiếp cận đã được thảo luận vào công việc hàng ngày của bạn:

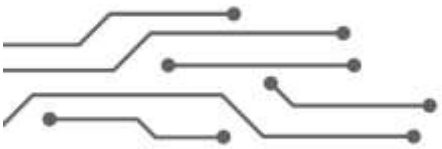
- Bắt đầu bằng việc hệ thống hóa các yêu cầu
  - Tạo dựng một danh mục thuộc tính và tham số cho các yếu tố chính của dự án và quy trình của bạn
  - Tài liệu hóa các giới hạn cho mỗi thuộc tính

- [ ] Hiển thị quy trình và mối quan hệ giữa các lớp, hệ thống và thuộc tính bằng biểu đồ (ví dụ: trong Miro, Canva, Visio)

- Tự động hóa quá trình chuyển đổi dữ liệu
  - [ ] Kiểm tra xem tài liệu nào mà bạn thường sử dụng trong quy trình có thể được số hóa bằng thư viện OCR và chuyển đổi sang dạng bảng
  - [ ] Tìm hiểu về các công cụ đảo ngược kỹ thuật để khai thác dữ liệu từ CAD (BIM)
  - [ ] Thử cấu hình việc tự động lấy dữ liệu từ các tài liệu hoặc định dạng mà bạn thường sử dụng trong công việc sang dạng bảng
  - [ ] Thiết lập những chuyển đổi tự động giữa các định dạng dữ liệu khác nhau
- Tạo một cơ sở tri thức để phân loại
  - [ ] Phát triển một phân loại nội bộ hoặc sử dụng phân loại đã có, phù hợp với các tiêu chuẩn ngành
  - [ ] Tài liệu hóa mối quan hệ giữa các hệ thống phân loại khác nhau
  - [ ] Thảo luận với đội nhóm của bạn về việc sử dụng một hệ thống nhận dạng thống nhất và phân loại rõ ràng các yếu tố
  - [ ] Bắt đầu xây dựng quy trình kiểm tra dữ liệu tự động — cho cả dữ liệu mà bạn làm việc trong đội ngũ và những dữ liệu được truyền sang các hệ thống bên ngoài

Việc sử dụng các cách tiếp cận này sẽ giúp bạn nâng cao chất lượng dữ liệu một cách đáng kể và đơn giản hóa quá trình xử lý và chuyển đổi sau này. Trong các phần tiếp theo của cuốn sách, chúng tôi sẽ xem xét cách áp dụng dữ liệu đã được cấu trúc và chuẩn bị cho các phép toán tự động, ước tính chi phí, lập kế hoạch lịch trình và quản lý dự án xây dựng.





V PHẦN **TÍNH TOÁN CHI PHÍ VÀ THỜI GIAN: TRIỂN KHAI DỮ LIỆU TRONG QUY TRÌNH XÂY DỰNG**

Phần thứ năm đề cập đến các khía cạnh thực tiễn của việc sử dụng dữ liệu để tối ưu hóa việc tính toán giá trị và lập kế hoạch cho các dự án xây dựng. Phương pháp tài nguyên để lập dự toán và tự động hóa quy trình tính toán được phân tích một cách chi tiết. Các phương pháp lấy số lượng tự động (Quantity Take-Off) từ các mô hình CAD (BIM) và tích hợp chúng với các hệ thống tính toán cũng được xem xét. Các công nghệ mô hình hóa 4D và 5D được nghiên cứu nhằm lập kế hoạch cho các tham số thời gian và quản lý chi phí xây dựng, với các ví dụ cụ thể về việc áp dụng chúng. Phân tích các lớp thông tin mở rộng 6D-8D được trình bày, cung cấp cái nhìn tổng thể về việc đánh giá tính bền vững, khai thác và an toàn của các tài sản bất động sản. Các phương pháp tính toán dấu chân carbon và các chỉ số ESG của các dự án xây dựng được xem xét kỹ lưỡng trong ngữ cảnh của các yêu cầu và tiêu chuẩn môi trường hiện đại. Khả năng và hạn chế của các hệ thống ERP và PMIS truyền thống trong quản lý quy trình xây dựng được đánh giá một cách phê phán, cùng với việc phân tích ảnh hưởng của chúng đến tính minh bạch trong định giá. Triển vọng chuyển từ các giải pháp khép kín sang các tiêu chuẩn mở và các công cụ phân tích dữ liệu linh hoạt, có khả năng cung cấp hiệu quả cao hơn cho các quy trình xây dựng được dự đoán.

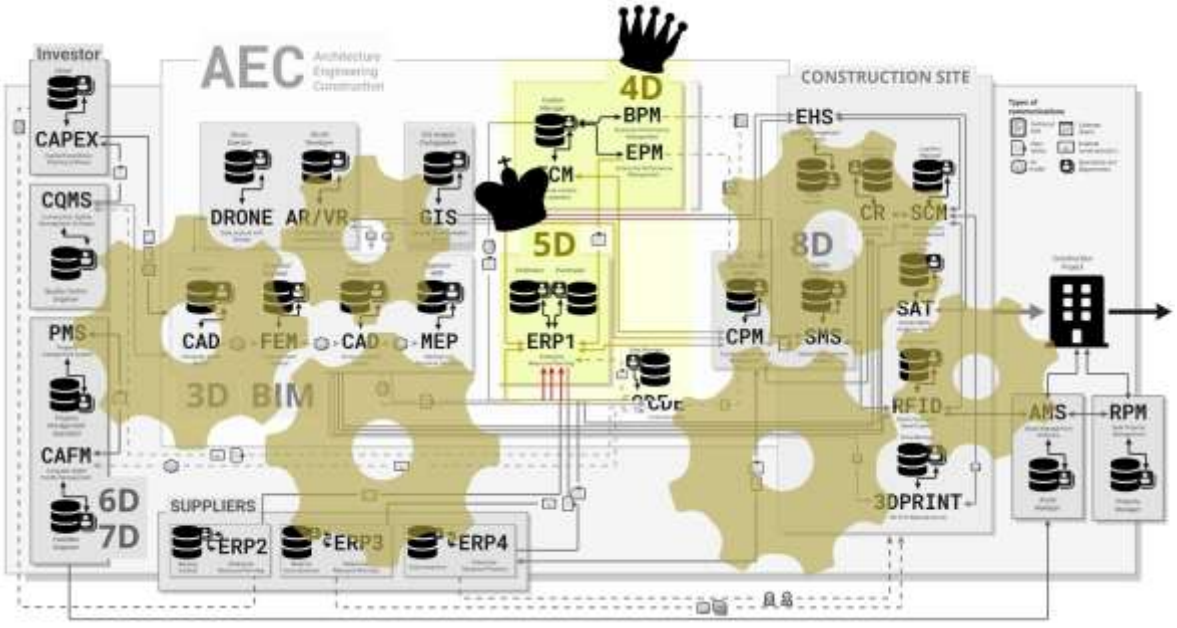
CHƯƠNG 5.1.
TÍNH TOÁN CHI PHÍ VÀ DỰ TOÁN DỰ ÁN XÂY DỰNG

Cơ bản về xây dựng: đánh giá số lượng, chi phí và thời gian

Trong số rất nhiều quy trình kinh doanh xác định sự bền vững của công ty trong ngành xây dựng, quy trình đánh giá chính xác số lượng các yếu tố, chi phí dự án và thời gian hoàn thành giữ vai trò đặc biệt quan trọng — cũng như hàng ngàn năm trước.

Sự phát triển của chữ viết đã trở thành kết quả của một loạt các yếu tố, bao gồm nhu cầu ghi chép các hoạt động kinh tế, phát triển thương mại và quản lý tài nguyên trong các xã hội sơ khai. Các tài liệu pháp lý đầu tiên — những bảng đất sét với các phép tính chi phí nguyên vật liệu và tiền công — đã được sử dụng trong bối cảnh thương mại và xây dựng. Những bảng này ghi lại các nghĩa vụ của các bên trong quá trình xây dựng và được lưu trữ như là bằng chứng cho các thỏa thuận và quan hệ hàng hóa - tiền tệ.

> Trong suốt hàng ngàn năm, phương pháp đánh giá gần như không thay đổi: các phép tính được thực hiện bằng tay, dựa vào kinh nghiệm và trực giác của kỹ sư lập dự toán. Tuy nhiên, với sự xuất hiện của các hệ thống ERP mô-đun và các công cụ CAD, phương pháp truyền thống trong việc đánh giá số lượng, chi phí và thời gian đã bắt đầu được chuyển đổi một cách nhanh chóng. Các công nghệ số hiện đại cho phép tự động hóa hoàn toàn các phép tính chính về thời gian và chi phí, từ đó nâng cao độ chính xác, tốc độ và tính minh bạch trong lập kế hoạch tài nguyên cho các dự án xây dựng.



Hình 5.11 Trong số nhiều hệ thống khác nhau, các công cụ liên quan đến các chỉ số về khối lượng, chi phí và thời gian là quan trọng nhất trong kinh doanh.

Sự chú ý chủ yếu của các công ty xây dựng được tập trung vào các dữ liệu chính xác về thời gian và chi phí của công việc. Những chỉ số này lại phụ thuộc vào khối lượng nguyên vật liệu được sử dụng và thời gian lao động, và tính minh bạch của chúng ảnh hưởng đến khả năng sinh lời. Tuy nhiên, sự phức tạp của các quy trình tính toán và tính không minh bạch của chúng thường dẫn đến việc tăng chi phí dự án, trễ tiến độ và thậm chí là phá sản công ty.

> Theo báo cáo KPMG "Những vấn đề quen thuộc — Các cách tiếp cận mới" (2023), chỉ có 50% các dự án xây dựng hoàn thành đúng hạn, trong khi 87% các công ty báo cáo sự gia tăng kiểm soát đối với kinh tế của các dự án vốn. Các vấn đề chính liên quan đến sự thiếu hụt nguồn nhân lực có trình độ và tính phức tạp trong việc dự đoán rủi ro.

Dữ liệu lịch sử về các phép tính chi phí và thời gian thực hiện quy trình được thu thập trong quá trình xây dựng các dự án trước đây trong suốt cuộc đời của công ty xây dựng và được nhập vào cơ sở dữ liệu của các hệ thống khác nhau (ERP, PMIS, BPM, EPM, v.v.).

> Sự tồn tại của dữ liệu lịch sử chất lượng cao về các phép tính là lợi thế cạnh tranh chính của tổ chức xây dựng, ảnh hưởng trực tiếp đến sự sống còn của nó.

Các phòng ban báo giá và tính toán trong các công ty xây dựng và kỹ thuật được thành lập để thu thập, lưu trữ và cập nhật dữ liệu lịch sử về các phép tính dự án. Chức năng chính của chúng là tích lũy và hệ thống hóa kinh nghiệm của công ty, cho phép theo thời gian nâng cao độ chính xác trong việc đánh giá khối lượng, thời gian và chi phí của các dự án mới. Cách tiếp cận này giúp giảm thiểu sai sót trong các phép tính tương lai, dựa trên thực tiễn và kết quả của các dự án đã hoàn thành.

Phương pháp tính toán dự toán chi phí dự án

Trong công việc của các chuyên gia lập dự toán, sử dụng nhiều phương pháp đánh giá khác nhau, mỗi phương pháp có định hướng vào một loại dữ liệu cụ thể, khả năng tiếp cận thông tin và mức độ chi tiết của dự án. Những phương pháp phổ biến nhất bao gồm:

- Phương pháp tài nguyên: đánh giá chi phí dự toán của dự án dựa trên phân tích chi tiết tất cả các tài nguyên cần thiết như vật liệu, thiết bị và lao động. Phương pháp này yêu cầu danh sách chi tiết tất cả các công việc và tài nguyên cần thiết để thực hiện từng công việc, với việc tính toán chi phí của chúng sau đó. Phương pháp này nổi bật với độ chính xác cao và được sử dụng rộng rãi trong việc lập dự toán.
- Phương pháp tham số: sử dụng các mô hình thống kê để đánh giá chi phí dựa trên các tham số của dự án. Điều này có thể bao gồm phân tích chi phí trên mỗi đơn vị đo lường, chẳng hạn như diện tích xây dựng hoặc khối lượng công việc, và điều chỉnh những chi phí này theo các điều kiện cụ thể của dự án. Phương pháp này đặc biệt hiệu quả trong các giai đoạn đầu, khi thông tin chi tiết chưa có sẵn.
- Phương pháp đơn vị (phương pháp chi phí đơn vị): tính toán chi phí dự toán của dự án dựa trên chi phí trên mỗi đơn vị đo lường (ví dụ: mỗi mét vuông hoặc mỗi mét khối). Điều này đảm bảo so sánh và phân tích chi phí của các dự án hoặc các phần của chúng một cách nhanh chóng và thuận tiện.
- Đánh giá chuyên gia (phương pháp Delphi): dựa trên ý kiến của các chuyên gia, những người sử dụng kinh nghiệm và kiến thức của họ để đánh giá chi phí của dự án. Cách tiếp cận này hữu ích khi không có dữ liệu chính xác hoặc dự án là độc đáo.

Cần lưu ý rằng phương pháp tham số và đánh giá chuyên gia có thể được điều chỉnh cho các mô hình học

máy. Điều này cho phép tự động xây dựng dự đoán về chi phí và thời gian thực hiện dự án dựa trên các mẫu huấn luyện. Các ví dụ về việc áp dụng các mô hình như vậy được xem xét chi tiết hơn trong chương "Ví dụ về việc sử dụng học máy để tìm kiếm chi phí và thời gian của dự án" (Hình 9.35).-

Tuy nhiên, phương pháp phổ biến nhất và được sử dụng rộng rãi trong thực tiễn toàn cầu vẫn là phương pháp tài nguyên. Nó không chỉ đảm bảo đánh giá chính xác chi phí dự toán mà còn cho phép tính toán thời gian thực hiện của từng quy trình trên công trường xây dựng và toàn bộ dự án (xem thêm trong chương "Biểu đồ xây dựng và dữ liệu dự án 4D").

Phương pháp tài nguyên trong lập dự toán và tính toán trong xây dựng

Phương pháp tính toán chi phí dựa trên tài nguyên là một phương pháp quản lý kế toán, trong đó chi phí dự án được hình thành dựa trên việc ghi chép trực tiếp tất cả các tài nguyên được sử dụng. Trong xây dựng, phương pháp này yêu cầu phân tích và đánh giá chi tiết tất cả các tài nguyên vật chất, lao động và kỹ thuật cần thiết để thực hiện công việc.

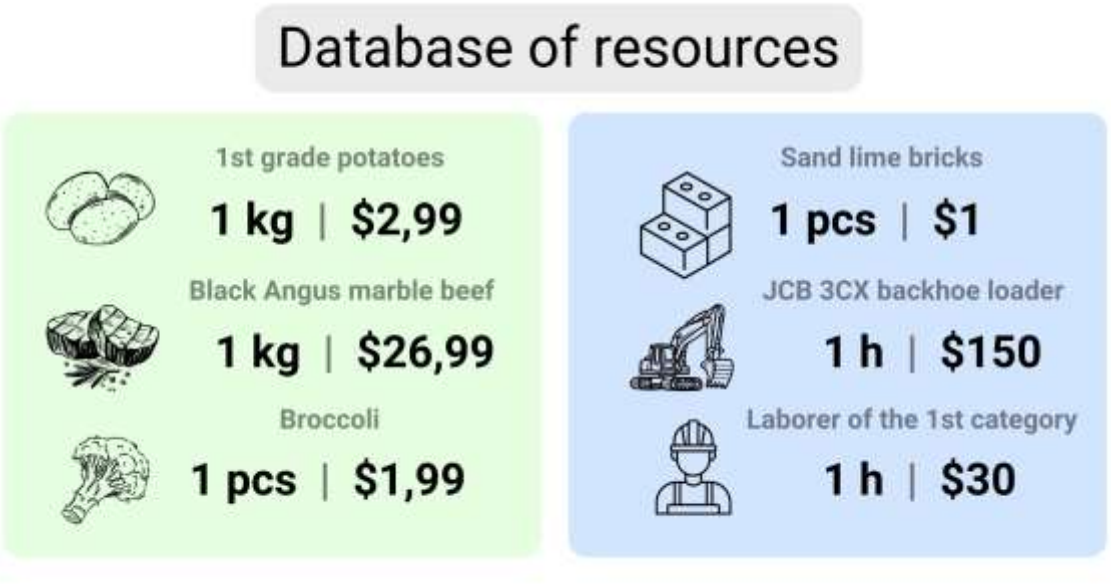
> Phương pháp tài nguyên cung cấp mức độ minh bạch và chính xác cao trong việc lập kế hoạch ngân sách, vì nó dựa trên giá thực tế của các tài nguyên vào thời điểm lập dự toán. Điều này đặc biệt quan trọng trong bối cảnh tình hình kinh tế không ổn định, khi sự biến động giá cả có thể tác động lớn đến tổng chi phí của dự án.

Ở các chương tiếp theo, chúng tôi sẽ phân tích chi tiết quy trình tính toán theo phương pháp tài nguyên. Để hiểu rõ hơn về nguyên tắc của nó trong xây dựng, chúng ta sẽ so sánh với việc tính toán chi phí bữa tối tại nhà hàng. Quản lý nhà hàng, khi lập kế hoạch cho buổi tối, sẽ lập danh sách các nguyên liệu cần thiết, xem xét thời gian chế biến mỗi món ăn, sau đó nhân chi phí với số lượng khách. Trong xây dựng, quy trình này tương tự: cho mỗi danh mục phần tử của dự án (các đối tượng), sẽ hình thành dự toán chi tiết, và tổng chi phí của dự án sẽ được xác định bằng cách cộng tất cả các chi phí trong bản dự toán tổng thể theo các danh mục.

Bước khởi đầu và quan trọng trong cách tiếp cận tài nguyên là hình thành cơ sở dữ liệu ban đầu của công ty. Ở giai đoạn đầu của việc tính toán, một danh sách có cấu trúc về tất cả các vật phẩm, vật liệu, loại công việc và tài nguyên mà công ty có cho các dự án xây dựng được lập - từ chiếc đinh trong kho cho đến mô tả về nhân viên dựa trên trình độ và mức lương theo giờ của họ. Thông tin này được hệ thống hóa thành "Cơ sở dữ liệu tài nguyên và vật liệu xây dựng" duy nhất - một bảng liệt kê chứa dữ liệu về tên gọi, đặc điểm, đơn vị đo lường và giá hiện tại. Chính cơ sở dữ liệu này sẽ trở thành nguồn thông tin chính cho tất cả các tính toán tài nguyên tiếp theo - cả chi phí và thời gian thực hiện công việc.

Cơ sở dữ liệu tài nguyên xây dựng: danh mục vật liệu và công việc xây dựng

Cơ sở dữ liệu hoặc bảng tài nguyên và vật liệu xây dựng bao gồm thông tin chi tiết về từng yếu tố có thể được sử dụng trong một dự án xây dựng - hàng hóa, sản phẩm, vật liệu hoặc dịch vụ, bao gồm tên, mô tả, đơn vị đo lường và giá trị đơn vị, được ghi chép một cách có cấu trúc. Trong bảng này, có thể tìm thấy mọi thứ: từ các loại nhiên liệu và vật liệu khác nhau được sử dụng trong các dự án, đến danh sách chi tiết về các chuyên gia dưới dạng các danh mục khác nhau kèm theo mô tả về mức lương theo giờ.

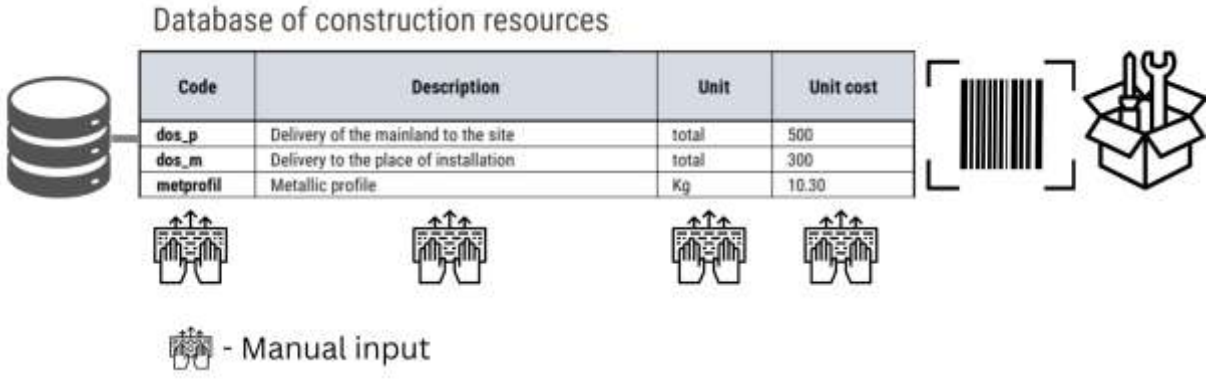


Bảng tài nguyên là danh sách các thành phần, mô tả vật liệu và dịch vụ cùng với giá trị đơn vị.

> Cơ sở dữ liệu tài nguyên tương tự như một danh mục hàng hóa của cửa hàng trực tuyến, nơi mỗi sản phẩm có mô tả chi tiết về các thuộc tính của nó. Điều này giúp cho các chuyên viên dự toán dễ dàng lựa chọn các tài nguyên cần thiết (giống như việc chọn hàng hóa khi thêm vào giỏ) để tính toán các quy trình xây dựng cụ thể dưới dạng dự toán.

Cơ sở dữ liệu tài nguyên cũng có thể được coi như danh sách tất cả các thành phần trong cuốn sách công thức nấu ăn. Mỗi vật liệu xây dựng, thiết bị và dịch vụ giống như các thành phần được sử dụng trong công thức. "Cơ sở dữ liệu tài nguyên" là danh sách chi tiết tất cả các thành phần - vật liệu xây dựng và dịch vụ, bao gồm cả giá trị mỗi đơn vị: cái, mét, giờ, lít, v.v.

Các yếu tố mới của thực thể có thể được thêm vào bảng "Cơ sở dữ liệu tài nguyên xây dựng" bằng hai cách - thủ công (Hình 5.13) hoặc tự động, thông qua việc tích hợp với hệ thống quản lý tồn kho của công ty hoặc cơ sở dữ liệu của nhà cung cấp.



Hình 5.13 Cơ sở dữ liệu tài nguyên được điền thủ công hoặc tự động tiếp nhận dữ liệu từ các cơ sở dữ liệu khác.

Một công ty xây dựng trung bình điển hình sử dụng một cơ sở dữ liệu chứa hàng ngàn, thậm chí hàng chục ngàn yếu tố với mô tả chi tiết, có thể được sử dụng trong các dự án xây dựng. Những dữ liệu này sau đó được sử dụng tự động trong hợp đồng và tài liệu dự án để mô tả chính xác nội dung công việc và quy trình.

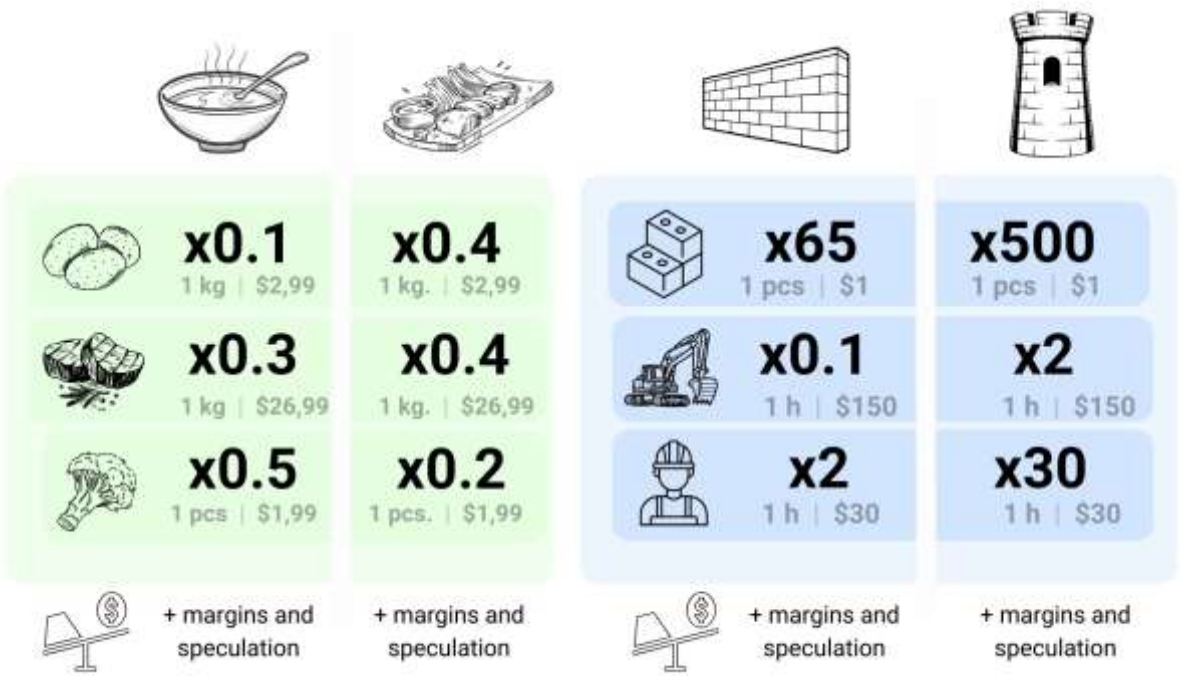
> Để không bị tụt lại so với các điều kiện thị trường thay đổi, chẳng hạn như lạm phát, thuộc tính "giá trị đơn vị sản phẩm" cho mỗi sản phẩm (hàng hóa hoặc dịch vụ) trong cơ sở dữ liệu tài nguyên (Hình 5.13) được cập nhật thường xuyên bằng tay hoặc thông qua việc tự động tải giá cập nhật từ các hệ thống khác hoặc nền tảng trực tuyến.

Việc cập nhật giá trị đơn vị của tài nguyên có thể được thực hiện hàng tháng, hàng quý hoặc hàng năm — tùy thuộc vào bản chất của tài nguyên, lạm phát và bối cảnh kinh tế bên ngoài. Việc cập nhật này là cần thiết để duy trì độ chính xác của các phép tính và ước lượng, vì chính những yếu tố cơ bản này là điểm xuất phát cho công việc của các chuyên gia ước tính. Dựa trên dữ liệu cập nhật, các bảng dự toán, ngân sách và thời gian thực hiện được hình thành, phản ánh các điều kiện thực tế của thị trường và giảm thiểu rủi ro sai sót trong các phép tính dự án tiếp theo.

Lập dự toán và tính toán chi phí công việc dựa trên cơ sở dữ liệu tài nguyên

Sau khi "Cơ sở dữ liệu tài nguyên xây dựng" (Hình 5.13) được lấp đầy bằng các thực thể-minimal elements, có thể tiến hành tạo lập các bảng tính, được tính toán cho mỗi quá trình hoặc công việc tại công trường xây dựng theo các đơn vị đo lường nhất định: ví dụ, cho mỗi mét khối bê tông, mỗi mét vuông tường thạch cao, mỗi mét vỉa hè hoặc cho việc lắp đặt một cửa sổ.

Ví dụ, để xây dựng một bức tường gạch có diện tích 1 m² (Hình 5.14), dựa trên kinh nghiệm của các dự án trước đây, cần khoảng 65 viên gạch (thực thể "Gạch silicate"), có giá 1 đô la mỗi viên (thuộc tính "Giá cho mỗi viên"), tổng cộng là 65 đô la. Ngoài ra, cần sử dụng thiết bị xây dựng (thực thể "Máy xúc JCB 3CX") trong khoảng 10 phút để đặt gạch gần khu vực làm việc. Vì chi phí thuê thiết bị là 150 đô la mỗi giờ, 6 phút sử dụng sẽ tốn khoảng 15 đô la. Bên cạnh đó, cần 2 giờ làm việc của nhà thầu đặt gạch, với mức lương theo giờ là 30 đô la, tổng cộng là 60 đô la.

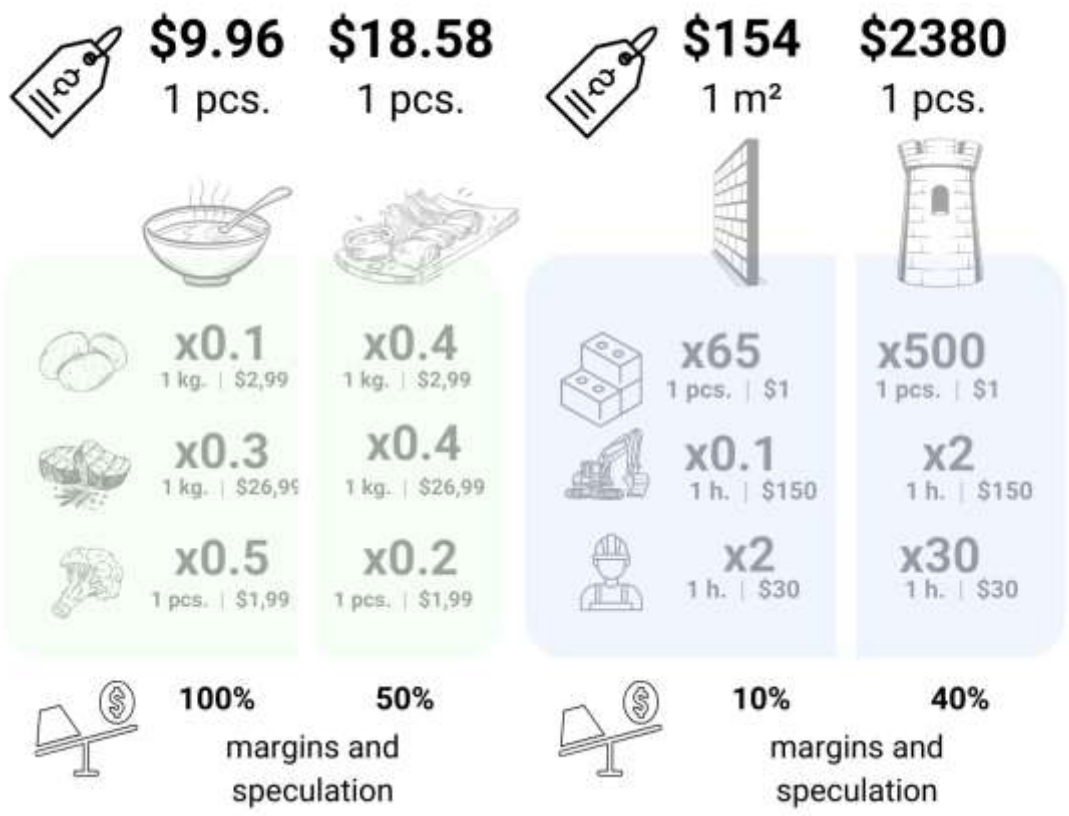


Hình 5.14 Các phép tính chi phí bao gồm danh sách chi tiết các vật liệu xây dựng và dịch vụ cần thiết để thực hiện các công việc và quy trình.

Thành phần của các bảng tính (còn được gọi là "công thức") được hình thành dựa trên kinh nghiệm lịch sử mà công ty đã tích lũy trong quá trình thực hiện khối lượng công việc đồng nhất lớn. Kinh nghiệm thực tế này thường được tích lũy thông qua phản hồi từ công trường. Cụ thể, đội trưởng thu thập thông tin trực tiếp tại công trường, ghi lại thời gian lao động thực tế, lượng nguyên vật liệu sử dụng và những chi tiết về quy trình thực hiện công nghệ. Sau đó, thông tin này sẽ được phối hợp với phòng dự toán để hoàn thiện qua nhiều vòng: mô tả quy trình được làm rõ hơn, thành phần nguồn lực được điều chỉnh, và các bảng tính được cập nhật theo dữ liệu thực tế từ các dự án gần đây.

> CŨng như trong công thức mô tả các thành phần cần thiết và số lượng của chúng để chế biến một món ăn, bảng dự toán cung cấp danh sách chi tiết tất cả các vật liệu xây dựng, nguồn lực và dịch vụ cần thiết để thực hiện một công việc hoặc quy trình cụ thể.

Những công việc được thực hiện thường xuyên cho phép công nhân, trưởng nhóm và nhân viên dự toán xác định được số lượng tài nguyên cần thiết: vật liệu, nhiên liệu, thời gian lao động và các tham số khác cần thiết để thực hiện một đơn vị công việc (Hình 5.15). Dữ liệu này được nhập vào các hệ thống dự toán dưới dạng bảng, trong đó mỗi nhiệm vụ và hoạt động được mô tả thông qua các yếu tố tối thiểu của cơ sở nguồn lực (với giá cả thường xuyên được cập nhật), điều này đảm bảo tính chính xác của các phép tính.



Hình 5.15 Đối với mỗi công việc, các mức giá đơn vị được thu thập, trong đó thuộc tính về khối lượng được nhân với số lượng và thêm vào phần trăm lợi nhuận.

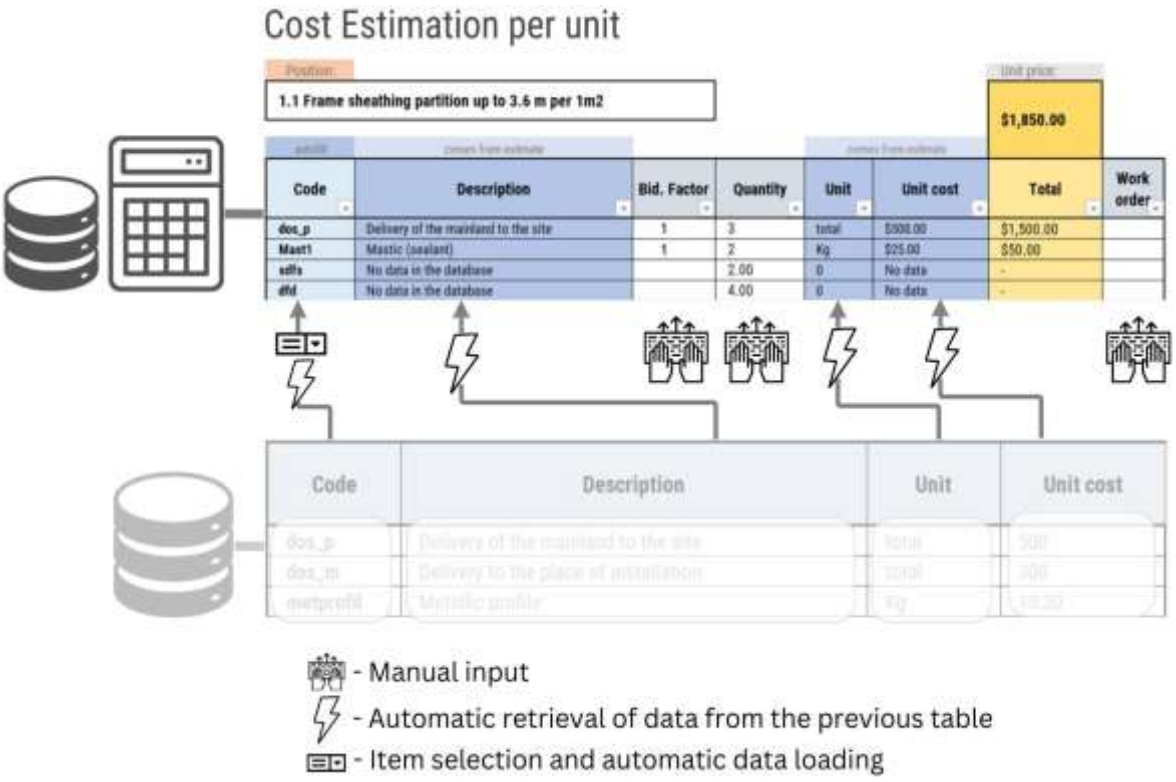
Để có được tổng chi phí cho mỗi quy trình hoặc công việc (đối tượng tính toán), thuộc tính chi phí được nhân với số lượng và các hệ số. Các hệ số này có thể xem xét nhiều yếu tố khác nhau, chẳng hạn như độ phức tạp của công việc, đặc thù vùng miền, mức độ lạm phát, rủi ro tiềm ẩn (phần trăm chi phí chung dự kiến) hoặc đầu cơ (hệ số lợi nhuận bổ sung).

> Nhân viên dự toán, giống như một nhà phân tích, chuyển đổi kinh nghiệm và khuyến nghị của đội trưởng thành các bảng dự toán tiêu chuẩn, mô tả các quy trình xây dựng thông qua các thực thể nguồn lực dưới dạng bảng. Thực chất, nhiệm vụ của nhân viên dự toán là thu thập và cấu trúc thông tin từ công trường thông qua các tham số và hệ số.

Do đó, chi phí cuối cùng cho một đơn vị công việc (chẳng hạn như mét vuông hay mét khối, hoặc lắp đặt một thiết bị) bao gồm không chỉ chi phí trực tiếp cho vật liệu và lực lượng lao động, mà còn bao gồm chi phí tăng thêm của công ty, chi phí chung, bảo hiểm và các yếu tố khác (Hình 5.16).

Nhờ vào đó, chúng ta không cần phải lo lắng về việc cập nhật giá trong (công thức) bảng tính nữa, vì giá thực tế luôn được phản ánh trong "cơ sở nguồn lực" (bảng thành phần). Ở cấp độ bảng tính, dữ liệu sẽ được tự động tải vào bảng từ cơ sở nguồn lực (ví dụ, theo mã thành phần hoặc mã định danh duy nhất), cung cấp mô tả và giá trị hiện tại mỗi đơn vị, mà các dữ liệu này có thể được tự động tải từ các nền tảng trực tuyến hoặc cửa hàng vật liệu xây dựng trên mạng. Nhân viên dự toán chỉ cần mô tả công việc hoặc quy trình thông

qua thuộc tính “số lượng nguồn lực” và các yếu tố bổ sung.



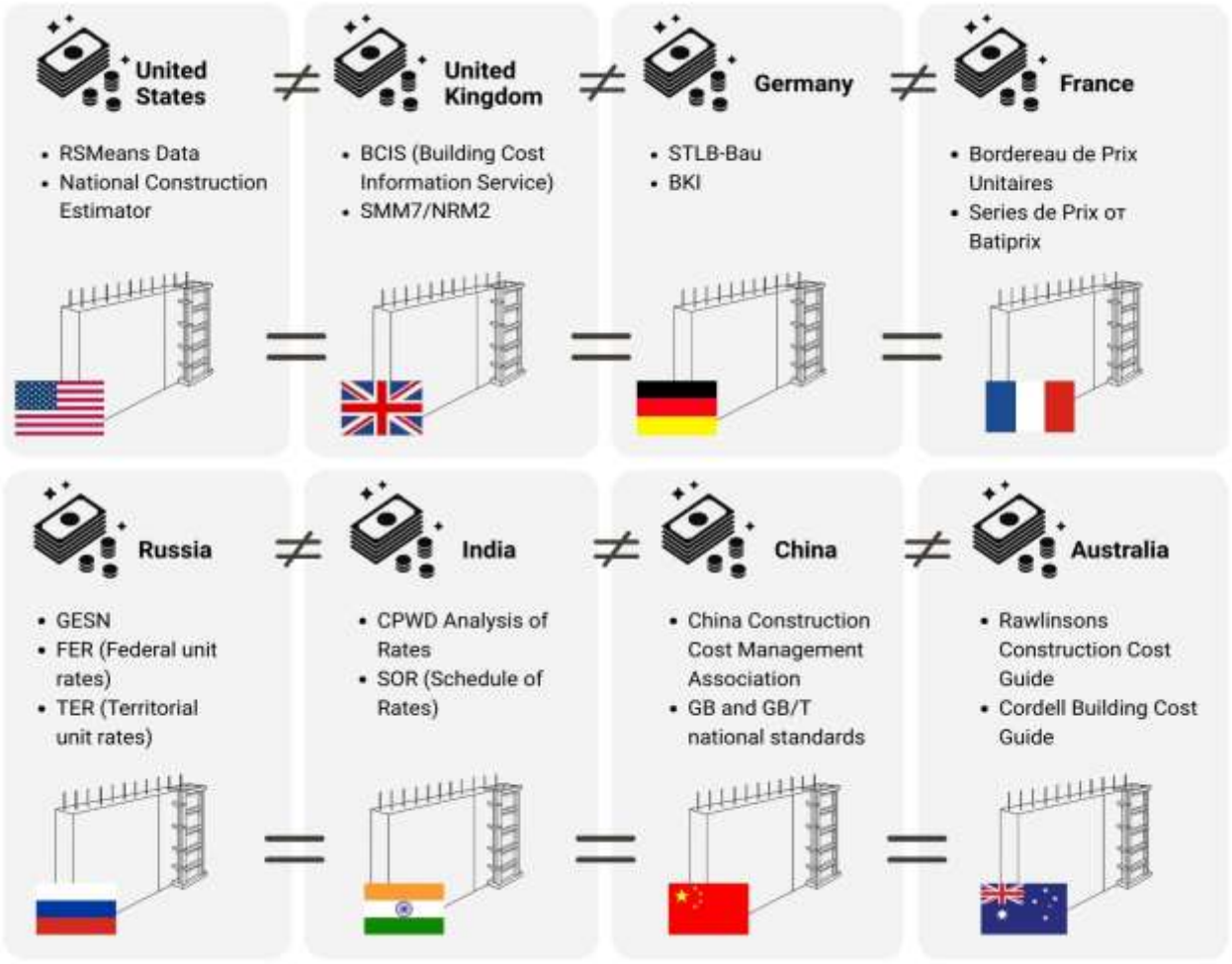
Hình 5.16 Ở giai đoạn tính toán chi phí cho một đơn vị công việc, chỉ cần điền các thuộc tính về số lượng tài nguyên cần thiết, mọi thứ còn lại sẽ tự động được tải từ cơ sở nguồn lực.

Các bảng tính chi phí công việc được tạo ra được lưu trữ dưới dạng các bảng mẫu của các dự án điển hình, mà có liên kết trực tiếp với cơ sở dữ liệu các nguồn lực và vật liệu xây dựng. Những mẫu này đại diện cho các công thức chuẩn hóa để thực hiện các loại công việc lặp đi lặp lại, cho các dự án tương lai, đảm bảo sự nhất quán trong các phép tính ở mức độ toàn công ty.

Khi giá của bất kỳ nguồn lực nào trong cơ sở dữ liệu thay đổi (Hình 5.13) — cho dù là thủ công hay tự động thông qua việc tải lên các mức giá thị trường hiện hành (chẳng hạn như trong điều kiện lạm phát) — các bản cập nhật sẽ ngay lập tức được phản ánh trong tất cả các bảng tính liên quan (Hình 5.16). Điều này có nghĩa là chỉ cần thực hiện thay đổi trong cơ sở dữ liệu nguồn lực, trong khi các mẫu bảng tính và dự toán vẫn giữ nguyên trong một thời gian dài. Cách tiếp cận này đảm bảo tính ổn định và khả năng tái sản xuất của các phép tính khi có bất kỳ biến động giá nào, chỉ được xem xét trong một bảng tài nguyên tương đối đơn giản (Hình 5.13).--

Cho mỗi dự án mới, một bản sao của mẫu bảng tính tiêu chuẩn được tạo ra, cho phép thực hiện các thay đổi và điều chỉnh các hoạt động theo các yêu cầu đặc biệt mà không làm thay đổi mẫu gốc đã được chấp nhận trong công ty. Cách tiếp cận này đảm bảo tính linh hoạt trong việc điều chỉnh các phép tính: có thể xem xét các đặc điểm của địa điểm xây dựng, mong muốn của khách hàng, đưa vào các hệ số rủi ro hoặc lợi nhuận (đầu cơ) — và tất cả điều này mà không vi phạm các tiêu chuẩn của công ty. Điều này giúp công ty tìm ra sự cân bằng giữa việc tối đa hóa lợi nhuận, đáp ứng nhu cầu của khách hàng và duy trì khả năng cạnh tranh.

Tại một số quốc gia, những mẫu bảng tính như vậy, được tích lũy trong nhiều thập kỷ, được chuẩn hóa ở cấp quốc gia và trở thành một phần của các tiêu chuẩn nhà nước trong hệ thống tính toán chi phí xây dựng (Hình 5.17).



Hình 5.17 Ở các quốc gia khác nhau trên thế giới, có quy tắc riêng cho bảng tính của cùng một yếu tố với các (công thức) bộ sưu tập và quy định riêng cho các công việc xây dựng.

Những cơ sở dữ liệu tài nguyên tiêu chuẩn hóa cho bảng tính (Hình 5.17) là bắt buộc đối với tất cả các bên tham gia thị trường, đặc biệt là trong việc thực hiện các dự án được tài trợ bởi nhà nước. Việc chuẩn hóa

như vậy đảm bảo tính minh bạch, khả năng so sánh và công bằng trong việc hình thành giá cả và nghĩa vụ hợp đồng cho khách hàng.

Tính toán tổng chi phí dự án: từ dự toán đến ngân sách

Các tiêu chuẩn dự toán nhà nước và ngành xây dựng đóng vai trò khác nhau trong thực tiễn xây dựng của các quốc gia khác nhau. Trong khi một số quốc gia buộc phải tuân thủ nghiêm ngặt các tiêu chuẩn thống nhất, hầu hết các nền kinh tế phát triển áp dụng cách tiếp cận linh hoạt hơn. Tại các quốc gia có nền kinh tế thị trường, các tiêu chuẩn xây dựng nhà nước thường chỉ đóng vai trò như một điểm xuất phát cơ bản. Các công ty xây dựng điều chỉnh các tiêu chuẩn này cho các mô hình hoạt động của họ hoặc hoàn toàn lược bỏ chúng, bổ sung các hệ số riêng của họ phản ánh các đặc điểm trong hoạt động. Những điều chỉnh này phản ánh kinh nghiệm doanh nghiệp, hiệu quả trong quản lý nguồn lực và thường là các yếu tố, trong đó có thể tính đến lợi nhuận đầu cơ của công ty.

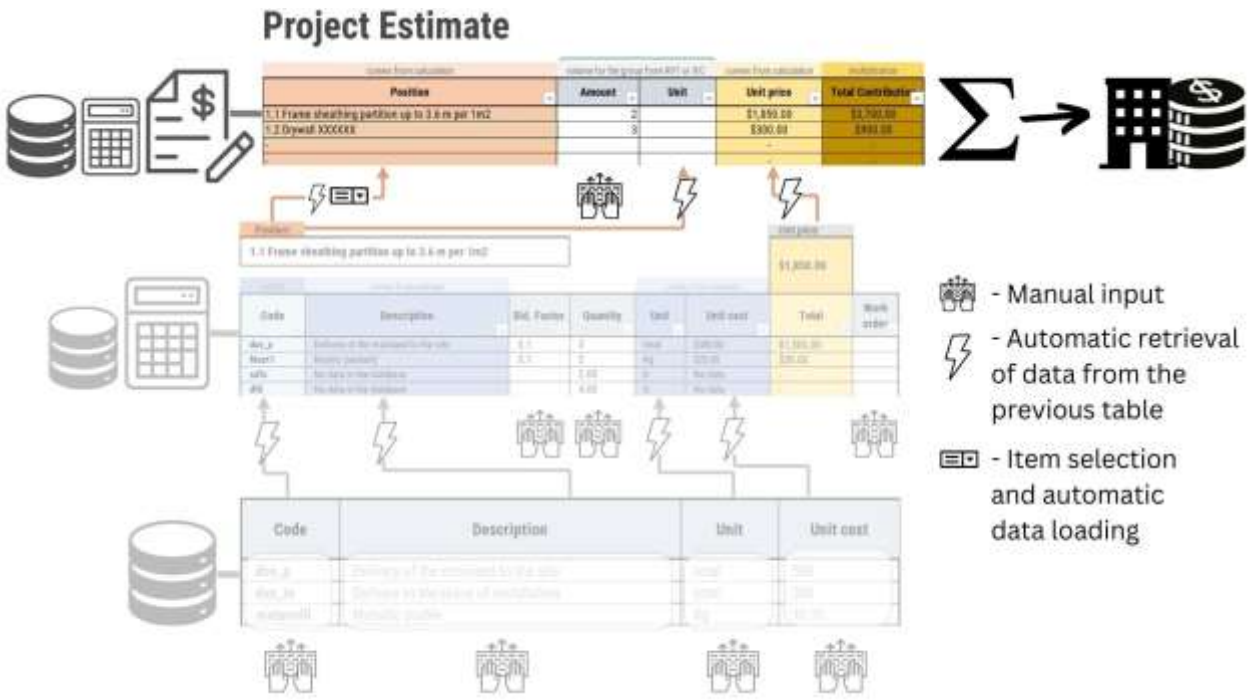
> Cuối cùng, mức độ cạnh tranh, cầu thị trường, biên lợi nhuận mục tiêu và thậm chí mối quan hệ với các khách hàng cụ thể có thể dẫn đến những sai lệch đáng kể so với các tiêu chuẩn chung. Thực tiễn này mang lại tính linh hoạt cho thị trường, nhưng đồng thời cũng làm phức tạp việc so sánh minh bạch giữa các đề xuất của các nhà thầu khác nhau, đưa vào quá trình lập ngân sách một yếu tố định giá đầu cơ trong ngành xây dựng.

Sau khi các mẫu tính toán cho các loại công việc và quy trình riêng lẻ đã được chuẩn bị - hoặc thường là chỉ được sao chép từ các dự toán nhà nước mẫu (Hình 5.17) với các hệ số được chỉnh sửa để phản ánh "đặc thù" của từng công ty cụ thể - ở giai đoạn cuối cùng, chỉ cần nhân giá thành mỗi hạng mục với khối lượng công việc hoặc quy trình tương ứng trong dự án mới.

> Trong việc tính toán tổng chi phí cho một dự án xây dựng mới, một bước quan trọng là tổng hợp chi phí từ tất cả các hạng mục lập ngân sách, nhân với khối lượng của các hạng mục đó trong dự án.

Để tạo ra tổng chi phí của dự án, trong ví dụ đơn giản của chúng tôi, chúng tôi sẽ bắt đầu bằng cách tính toán chi phí xây dựng một mét vuông tường và nhân giá thành của nó (ví dụ, công việc "1m² lắp đặt yếu tố tường tiêu chuẩn") với tổng số mét vuông tường trong dự án (ví dụ, thuộc tính "Diện tích" hoặc "Số lượng" (Hình 5.18) của thực thể loại "Yếu tố tường" từ dự án CAD hoặc từ tính toán của người chỉ huy).

Tương tự, chúng tôi tính toán chi phí cho tất cả các yếu tố của dự án (Hình 5.18): lấy giá thành của đơn vị công việc và nhân nó với khối lượng của yếu tố cụ thể hoặc nhóm của nó trong dự án này. Người lập dự toán chỉ cần nhập số lượng các yếu tố, công việc hoặc quy trình trong dự án dưới dạng khối lượng hoặc số lượng. Điều này cho phép tự động tạo ra một dự toán xây dựng đầy đủ.



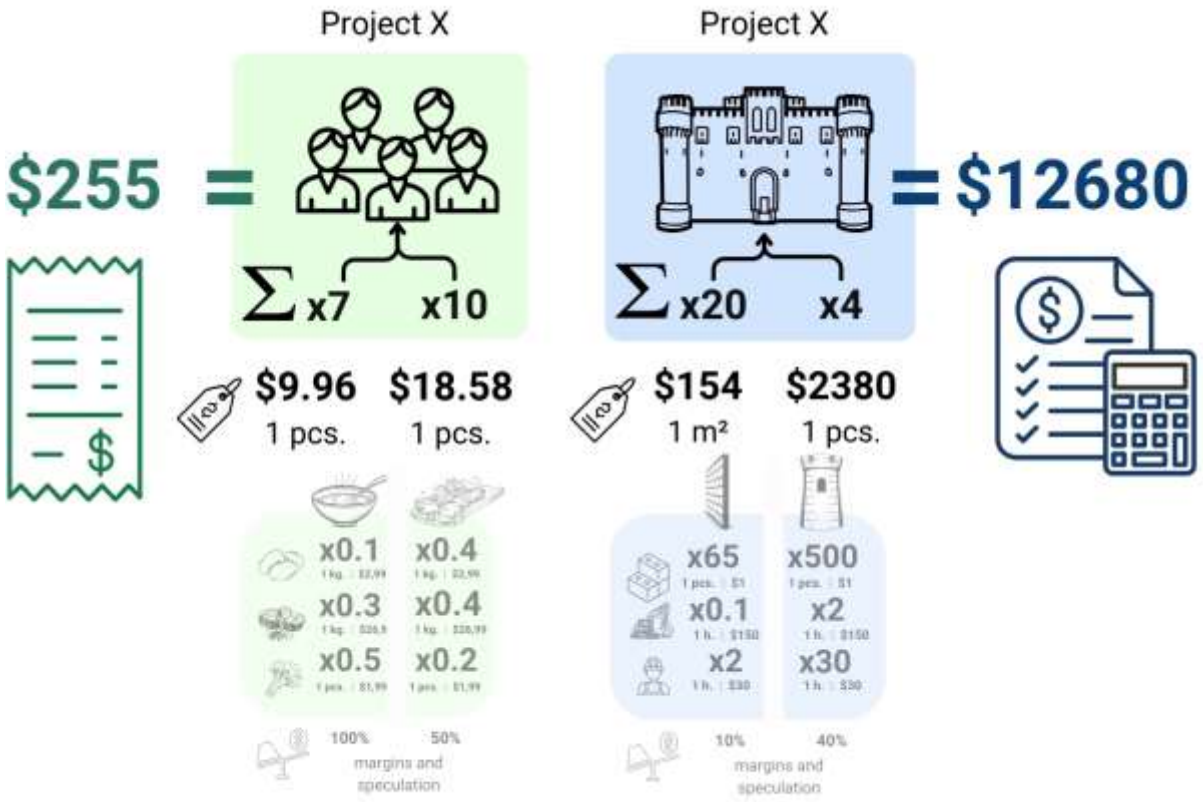
Hình 5.18 Trong quá trình tạo lập dự toán, chúng tôi chỉ nhập khối lượng công việc.

Giống như trong trường hợp lập ngân sách, ở cấp độ này, chúng tôi tự động tải lên các hạng mục đã được tính toán sẵn (từ mẫu lập ngân sách hoặc mới, được sao chép từ mẫu và chỉnh sửa) mà tự động mang theo giá thành hiện tại cho mỗi đơn vị công việc (được cập nhật tự động từ cơ sở dữ liệu tài nguyên (Hình 5.18, bảng dưới)). Do đó, khi có bất kỳ thay đổi nào trong cơ sở dữ liệu tài nguyên hoặc bảng lập ngân sách - dữ liệu trong dự toán sẽ tự động cập nhật đến ngày hiện tại mà không cần thay đổi lập ngân sách hoặc tự thân dự toán.

Trong bối cảnh nhà hàng, chi phí cuối cùng cho sự kiện được tính toán theo cách tương tự và bằng tổng chi phí của toàn bộ bữa tối, trong đó giá thành mỗi món ăn, nhân với số lượng khách, được cộng lại thành tổng chi phí của hóa đơn (Hình 5.19). Và giống như trong xây dựng, công thức chế biến món ăn trong nhà hàng có thể không thay đổi trong nhiều thập kỷ. Khác với giá cả, nơi mà chi phí nguyên liệu có thể thay đổi mỗi giờ.

> Cũng giống như một chủ nhà hàng nhân giá thành mỗi món ăn với số phần và số người để xác định tổng chi phí sự kiện, thì quản lý dự toán tổng hợp chi phí của tất cả các thành phần của dự án để có được dự toán xây dựng đầy đủ.

Do đó, cho mỗi công việc trong dự án, giá trị cuối cùng của nó được xác định (Hình 5.19), và khi nhân với khối lượng thuộc tính của thực thể tương ứng với công việc đó - sẽ cho ra giá trị của nhóm công việc, từ đó dẫn đến giá trị cuối cùng của toàn bộ dự án.



Hình 5.19 Bảng tổng hợp dự toán được tính toán bằng cách cộng tổng thuộc tính giá trị của công việc của mỗi phần tử với thuộc tính khối lượng của nó.

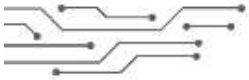
Giá trị cuối cùng của dự án (Hình 5.18) đại diện cho bức tranh tài chính của dự án, cho phép khách hàng, nhà đầu tư hoặc các tổ chức tài chính hiểu được ngân sách tổng thể và các nguồn lực tài chính cần thiết để triển khai dự án vào bất kỳ thời điểm nào, với giá cả hiện tại.

Và nếu các quy trình lập cơ sở tài nguyên, tính toán và dự toán (công thức quy trình) đã được hoàn thiện, tự động hóa và mài dũa qua hàng chục nghìn năm và được ghi nhận ở cấp độ quốc gia, thì việc tự động nhận được thông tin chất lượng về khối lượng và số lượng các phần tử cho giai đoạn cuối cùng của bảng dự toán còn là điểm nghẽn trong các quy trình tính toán thuộc tính giá trị và thời gian của dự án, và nói chung là ngân sách tổng thể của dự án.

> Trong hàng ngàn năm, phương pháp truyền thống để tính toán khối lượng đã là các phương pháp đo lường thủ công các đặc tính khối lượng và số lượng thông qua các bản vẽ phẳng. Với sự xuất hiện của kỷ nguyên kỹ thuật số, các công ty đã phát hiện ra rằng thông tin về khối lượng và số lượng giờ đây có thể được tự động trích xuất từ dữ liệu hình học có trong các mô hình CAD, điều này đã tạo ra một cuộc cách mạng trong các phương pháp truyền thống để có được dữ liệu định lượng.

Các phương pháp hiện đại để đánh giá quy trình và lập dự toán dự kiến việc tự động trích xuất các thuộc tính khối lượng và số lượng từ cơ sở dữ liệu CAD, mà có thể được xuất ra và kết nối với quy trình tính toán,

nhằm nhận được các khối lượng hiện tại của nhóm dự án ở bất kỳ giai đoạn nào, từ thiết kế đến vận hành.



CHƯƠNG 5.2. **QUANTITY TAKE-OFF VÀ TỰ ĐỘNG HÓA LẬP DỰ TOÁN VÀ KẾ HOẠCH THỜI GIAN**

Chuyển từ 3D sang 4D và 5D: sử dụng các tham số khối lượng và số lượng

Có trong tay các bảng tính toán với các quy trình mô tả thông qua tài nguyên (Hình 5.18), bước tiếp theo là tự động nhận được các tham số khối lượng hoặc số lượng cho nhóm các phần tử cần thiết cho các tính toán và lập bảng dự toán cuối cùng.

Các đặc tính khối lượng của các phần tử dự án - chẳng hạn như tường hoặc sàn - có thể được tự động trích xuất từ cơ sở dữ liệu CAD. Các đối tượng tham số được tạo ra trong các chương trình CAD được chuyển đổi bởi các phương tiện của lõi hình học thành các giá trị số của các tham số chiều dài, chiều rộng, diện tích, khối lượng và các thông số khác. Quá trình nhận được khối lượng dựa trên hình học 3D sẽ được xem xét chi tiết trong phần tiếp theo, phần thứ sáu (Hình 6.33), tập trung vào làm việc với CAD (BIM). Ngoài khối lượng, số lượng các phần tử đồng loại cũng có thể được nhận từ cơ sở dữ liệu mô hình CAD thông qua quá trình lọc và phân nhóm các đối tượng theo danh mục và thuộc tính. Những tham số này, cho phép phân nhóm - trở thành nền tảng để liên kết các phần tử của dự án thông qua các tính toán tài nguyên với các tính toán, bảng dự toán cuối cùng và ngân sách tổng thể của dự án.

Do đó, mô hình dữ liệu được trích xuất từ mô hình 3D (CAD) được bổ sung bằng các lớp tham số mới, được gọi là 4D và 5D. Trong các lớp thuộc tính của thực thể mới - 4D (thời gian) và 5D (giá trị) - dữ liệu hình học 3D được sử dụng làm nguồn giá trị cho các thuộc tính khối lượng của các thực thể.

- 4D — lớp thông tin về các tham số, bổ sung cho các tham số 3D của các yếu tố thông tin về thời gian thực hiện các hoạt động xây dựng. Những dữ liệu này cần thiết cho việc lập kế hoạch và quản lý thời gian thực hiện dự án.
- 5D — mức độ mở rộng tiếp theo của mô hình dữ liệu, trong đó các yếu tố được bổ sung bằng các đặc tính giá trị. Một cách như vậy, thông tin hình học được bổ sung với khía cạnh tài chính: chi phí nguyên vật liệu, lao động và thiết bị, cho phép thực hiện các tính toán ngân sách, phân tích tính khả thi và quản lý chi phí trong quá trình xây dựng.

Dữ liệu về chi phí và các thuộc tính 3D, 4D và 5D của nhóm thực thể dự án được mô tả tương tự như các tính toán trong các hệ thống ERP, PIMS (hoặc các công cụ tương tự Excel) và được sử dụng cho việc tính toán chi phí và lập kế hoạch ngân sách tự động cho cả các nhóm riêng lẻ và toàn bộ ngân sách dự án.

Các thuộc tính 5D và thu thập khối lượng thuộc tính từ CAD

Trong quá trình chuẩn bị dự toán cuối cùng cho dự án xây dựng, mà chúng tôi đã đề cập ở các chương trước (Hình 5.18), các thuộc tính khối lượng cho từng loại yếu tố dự án được thu thập bằng tay hoặc trích xuất từ các thông số thuộc tính khối lượng được cung cấp bởi các chương trình CAD.-

> Phương pháp truyền thống để tính toán khối lượng bằng tay yêu cầu các kỹ sư và nhân viên dự toán xem xét các bản vẽ đã được trình bày từ hàng thiên niên kỷ dưới dạng các đường nét trên giấy, và trong 30 năm qua — dưới các định dạng số PDF (PLT) hoặc DWG. Dựa vào kinh nghiệm chuyên môn, họ đo đạc khối lượng công việc và vật liệu cần thiết, thường sử dụng thước kẻ và thước đo góc. Phương pháp này yêu cầu nỗ lực và thời gian đáng kể, cũng như sự chú ý đặc biệt đến từng chi tiết.

Việc xác định các thuộc tính khối lượng công việc theo cách này có thể mất từ vài ngày đến vài tháng, tùy vào quy mô dự án. Hơn nữa, do tất cả các phép đo và tính toán đều được thực hiện bằng tay, có nguy cơ xảy ra sai sót do con người, điều này có thể dẫn đến dữ liệu không chính xác, từ đó ảnh hưởng đến sai sót trong việc ước lượng thời gian và chi phí của dự án, mà công ty sẽ phải chịu trách nhiệm về điều đó.

Các phương pháp hiện đại, dựa trên việc sử dụng cơ sở dữ liệu CAD, đơn giản hóa đáng kể việc tính toán khối lượng. Trong các mô hình CAD, hình học của các yếu tố đã bao gồm các thuộc tính khối lượng, có thể được tính toán tự động (thông qua lõi hình học (Hình 6.33)) và được trình bày hoặc xuất ra dưới dạng bảng.-

Trong kịch bản như vậy, bộ phận dự toán yêu cầu từ nhà thiết kế CAD dữ liệu về các đặc điểm số lượng và khối lượng của các yếu tố dự án. Các dữ liệu này được xuất ra dưới dạng bảng hoặc tích hợp trực tiếp vào các cơ sở dữ liệu tính toán — bất kể đó là Excel, ERP hay các hệ thống PMIS. Quy trình này thường bắt đầu không từ một yêu cầu chính thức, mà từ một cuộc đối thoại ngắn giữa khách hàng (người khởi xướng) và kiến trúc sư cùng nhân viên dự toán từ phía công ty xây dựng hoặc thiết kế. Dưới đây là một ví dụ đơn giản, thể hiện cách thông tin giao tiếp hàng ngày hình thành bảng cấu trúc cho tính toán tự động (QTO):

- Khách hàng – Tôi muốn thêm một tầng nữa vào tòa nhà, với cấu hình giống như tầng hai.
- Kiến trúc sư (CAD) – Thêm tầng ba, cấu hình giống như tầng hai. Sau đó, họ gửi phiên bản mới của dự án CAD đến nhân viên dự toán.
- Người lập dự toán tự động tiến hành nhóm và tính toán (ERP, PMIS, Excel) – “Tôi sẽ đưa dự án qua bảng Excel với quy tắc QTO (ERP, PMIS), nhận khối lượng theo danh mục cho tầng mới và lập dự toán”.

Cuối cùng, cuộc đối thoại văn bản được chuyển thành cấu trúc bảng với quy tắc nhóm:

Phần tử	Danh mục	Tầng
Mặt sàn	OST_Floors	3
Cột	OST_StructuralColumns	3

Phần tử	Danh mục	Tầng
Thang bộ	OST_Stairs	3

Sau quá trình nhóm tự động mô hình CAD từ nhà thiết kế theo quy tắc QTO của người lập dự toán và tự động nhân khối lượng với tính toán tài nguyên (Hình 5.18), chúng ta thu được các kết quả sau đây được gửi đến khách hàng:-

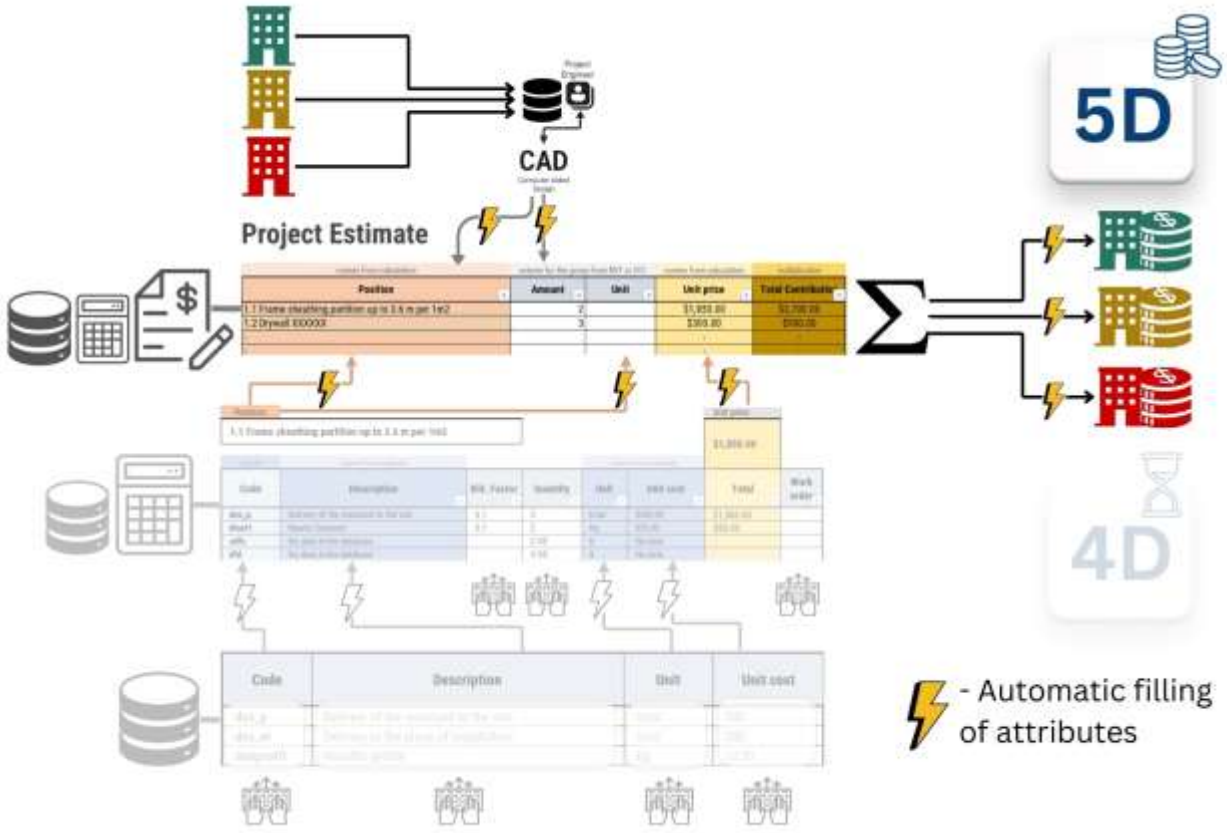
Phần tử	Khối lượng	Tầng	Giá mỗi đơn vị	Tổng chi phí
Mặt sàn	420 m²	3	150 €/m²	**63 000 €**
Cột	4 cái	3	2450 €/cái	**9 800 €**
Thang bộ	2 cái	3	4 300 €/cái	**8 600 €**
TỔNG CỘNG:	-	-	-	**81 400 €**

- Khách hàng – "Cảm ơn, số lượng nhiều quá, cần cắt giảm một vài phòng". Và chu trình lặp lại nhiều lần.

Kịch bản tương tự có thể lặp đi lặp lại nhiều lần, đặc biệt trong giai đoạn thỏa thuận, nơi khách hàng mong đợi phản hồi ngay lập tức. Tuy nhiên, thực tế cho thấy những quy trình như vậy có thể kéo dài nhiều ngày hoặc thậm chí hàng tuần. Ngày nay, nhờ vào việc triển khai các quy tắc nhóm và tính toán tự động, các hành động trước đây mất nhiều thời gian giờ đây chỉ mất vài phút. Việc tự động thu nhận khối lượng qua các quy tắc nhóm không chỉ làm tăng tốc độ tính toán và lập dự toán, mà còn nhờ vào việc giảm thiểu yếu tố con người, làm giảm khả năng xảy ra lỗi, đảm bảo đánh giá chi phí dự án một cách minh bạch và chính xác.

> Nếu trong quá trình tạo mô hình 3D trong hệ thống CAD đã chú ý đến yêu cầu của bộ phận lập dự toán (điều này vẫn khá hiếm gặp trong thực tế), và tên, mã định danh nhóm phần tử cùng các đặc điểm phân loại của chúng được thiết lập dưới dạng tham số trùng khớp với cấu trúc của các nhóm và lớp dự toán, thì các thuộc tính khối lượng có thể được chuyển tự động vào hệ thống dự toán mà không cần chuyển đổi thêm.

Việc trích xuất tự động các thuộc tính khối lượng từ CAD dưới dạng bảng danh mục cho phép nhanh chóng nhận được dữ liệu cập nhật về chi phí các công việc riêng lẻ và tổng thể dự án (Hình 5.21). Bằng cách chỉ cập nhật tệp CAD với khối lượng dự án trong quá trình tính toán hoặc hệ thống định mức, công ty có thể nhanh chóng tính lại dự toán với những thay đổi mới nhất, đảm bảo độ chính xác cao và sự nhất quán của tất cả các tính toán tiếp theo.



Hình 5.21 Các thuộc tính khối lượng từ bảng hoặc cơ sở dữ liệu CAD được nhập tự động vào dự toán, cho phép tính toán nhanh chóng tổng chi phí của dự án.

Trong bối cảnh các dự án vốn ngày càng phức tạp, việc tính toán ngân sách toàn bộ và phân tích tổng chi phí dự án theo kịch bản tương tự (Hình 5.21) trở thành công cụ then chốt cho việc ra quyết định dựa trên cơ sở dữ liệu.-

> Theo nghiên cứu của Accenture "Tạo ra giá trị lớn hơn thông qua các dự án vốn" (2024) [20], các công ty hàng đầu đang tích cực tích hợp phân tích dữ liệu vào các sáng kiến số, sử dụng thông tin lịch sử để dự đoán và tối ưu hóa kết quả. Nghiên cứu cho thấy ngày càng nhiều chủ sở hữu - nhà điều hành áp dụng phân tích dữ liệu lớn để dự đoán xu hướng thị trường và đánh giá tính khả thi thương mại ngay cả trước khi bắt đầu thiết kế. Điều này đạt được nhờ phân tích các kho dữ liệu từ danh mục dự án hiện có. Hơn nữa, 79% chủ sở hữu - nhà điều hành triển khai phân tích dự đoán "đáng tin cậy" để đánh giá hiệu quả của các dự án và hỗ trợ ra quyết định điều hành theo thời gian thực.

Quản lý hiệu quả các dự án xây dựng hiện đại không thể tách rời khỏi việc xử lý và phân tích khối lượng lớn thông tin ở tất cả các giai đoạn thiết kế và các quy trình trước khi thiết kế. Việc sử dụng kho dữ liệu, tính toán tài nguyên, mô hình dự đoán và máy học không chỉ giúp giảm thiểu rủi ro trong các phép toán mà còn cho phép đưa ra các quyết định chiến lược về tài chính cho dự án ngay từ giai đoạn đầu của thiết kế. Về kho dữ liệu và mô hình dự đoán, những thứ sẽ bổ sung cho các phép toán, chúng tôi sẽ nói chi tiết hơn trong phần chín của cuốn sách.

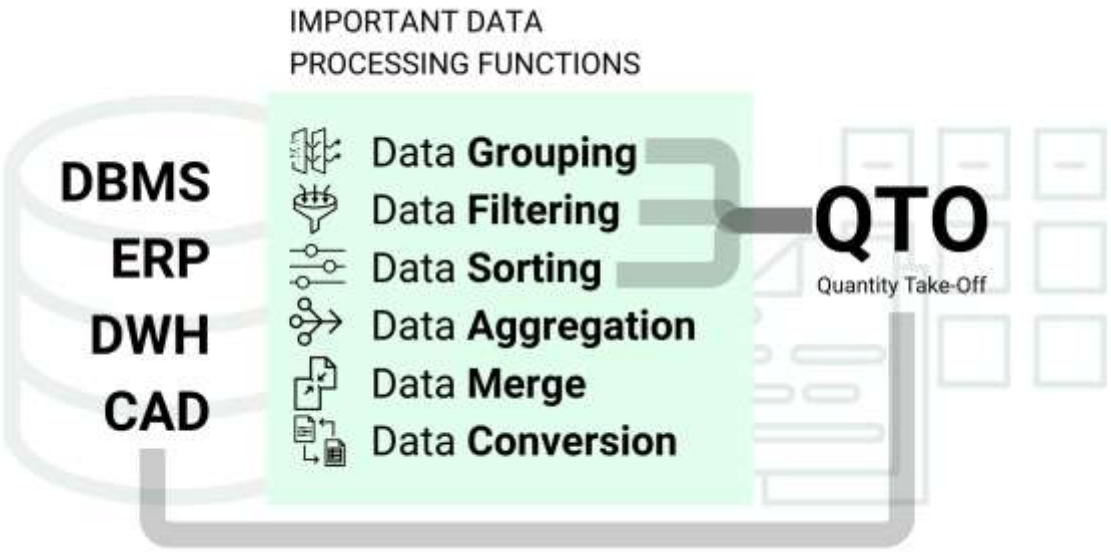
Việc tự động thu thập các thông số khối lượng của các yếu tố từ các dự án CAD, cần thiết cho việc lập dự toán, được thực hiện thông qua các công cụ nhóm QTO (Quantity Take-Off). Các công cụ QTO hoạt động

bằng cách nhóm tất cả các đối tượng của dự án theo các định danh yếu tố hoặc tham số thuộc tính đặc biệt, sử dụng các bảng thông số và bảng được tạo trong cơ sở dữ liệu CAD.

QTO Quantity Take-Off: nhóm dữ liệu dự án theo thuộc tính

Tính toán các tham số khối lượng và số lượng vật liệu (QTO - Quantity Take-Off) trong xây dựng là quá trình trích xuất các đặc điểm số lượng của các yếu tố cần thiết để thực hiện dự án. Trên thực tế, QTO thường vẫn là một quá trình bán tự động, bao gồm việc thu thập dữ liệu từ nhiều nguồn khác nhau: tài liệu PDF, bản vẽ định dạng DWG và mô hình số CAD.

Khi làm việc với dữ liệu được trích xuất từ cơ sở dữ liệu CAD, quá trình đánh giá số lượng (QTO) được thực hiện dưới dạng một chuỗi các thao tác lọc, sắp xếp, nhóm và tổng hợp. Các yếu tố của mô hình được lựa chọn theo các tham số lớp, loại và loại, sau đó các thuộc tính số lượng của chúng - chẳng hạn như thể tích, diện tích, chiều dài hoặc số lượng - được tổng hợp theo logic tính toán (Hình 5.22).



Hình 5.22 Nhóm và lọc dữ liệu là những chức năng phổ biến nhất được áp dụng cho các cơ sở dữ liệu và kho dữ liệu.

Quy trình QTO (lọc và nhóm) cho phép hệ thống hóa dữ liệu, lập các bảng thông số và chuẩn bị thông tin đầu vào cho việc tính toán dự toán, mua sắm và lịch trình thực hiện công việc. Cơ sở của QTO là phân loại các yếu tố theo loại thuộc tính đo lường. Đối với mỗi yếu tố hoặc nhóm yếu tố, một tham số đo lường số lượng tương ứng được chọn. Ví dụ:

- Thuộc tính chiều dài (dải phân cách - tính bằng mét)
- Thuộc tính diện tích (công việc thạch cao - tính bằng mét vuông)
- Thuộc tính thể tích (công việc bê tông - tính bằng mét khối)
- Thuộc tính số lượng (cửa sổ - tính theo từng chiếc)

Bên cạnh các đặc tính khối lượng được tạo ra từ toán học dựa trên hình học, sau khi nhóm QTO, các hệ số dư thừa thường được áp dụng trong các phép tính (Hình 5.212, ví dụ 1,1 để tính 10% cho logistics và lắp

đặt) - các giá trị sửa đổi này xem xét tổn thất, đặc điểm lắp đặt, lưu trữ hoặc vận chuyển. Điều này cho phép ước tính chính xác hơn về lượng vật liệu thực tế và tránh cả tình trạng thiếu hụt và dư thừa hàng hóa trên công trường xây dựng.

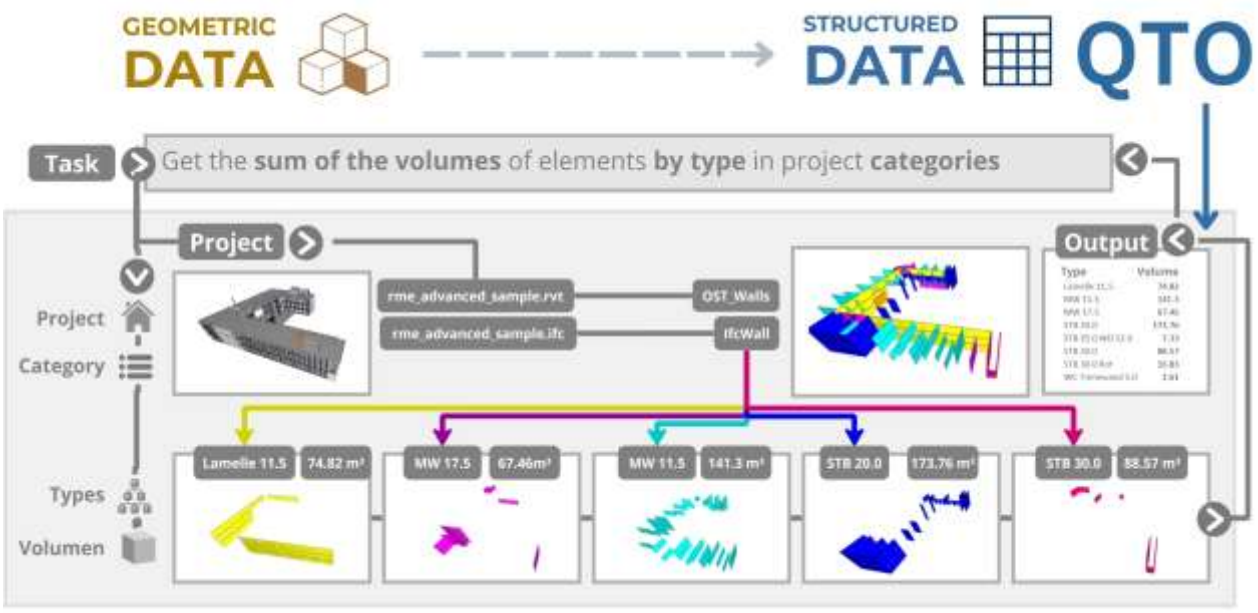
> Quy trình tự động của việc tính toán khối lượng (QTO) là cần thiết để lập các bảng dự toán và dự toán chính xác, giảm thiểu yếu tố con người trong các quá trình xác định các đặc tính khối lượng và ngăn chặn tình trạng thừa hoặc thiếu trong việc đặt hàng vật liệu.

Lấy một ví dụ về quy trình QTO, hãy xem xét trường hợp phổ biến khi cần hiển thị bảng xác định khối lượng từ cơ sở dữ liệu CAD cho các loại thành phần thuộc một danh mục nhất định, các lớp thành phần. Chúng ta sẽ nhóm tất cả các thành phần của dự án theo loại từ danh mục tường của dự án CAD và tổng hợp các thuộc tính khối lượng cho mỗi loại để trình bày kết quả dưới dạng bảng khối lượng QTO (Hình 5.23).-

Trong ví dụ về dự án CAD tiêu chuẩn (Hình 5.23), tất cả các thành phần thuộc danh mục tường trong cơ sở dữ liệu CAD được nhóm lại theo loại tường, ví dụ như "Lamelle 11.5", "MW 11.5" và "STB 20.0", và có các thuộc tính khối lượng được xác định rõ ràng, được trình bày bằng mét khối.

Mục tiêu của người quản lý, đứng ở giao điểm giữa các nhà thiết kế và các chuyên gia tính toán, là nhận được bảng khối lượng tự động theo các loại thành phần trong danh mục đã chọn. Và không chỉ cho dự án cụ thể mà còn ở dạng tổng quát, có thể áp dụng cho các dự án khác có cấu trúc mô hình tương tự. Điều này cho phép mở rộng phương pháp và đảm bảo việc tái sử dụng dữ liệu mà không cần nhân đôi nỗ lực.

> Thời kỳ mà các người điều hành và chuyên gia dự toán dày dạn trang bị thước kẻ, cẩn thận đo lường từng đường nét trên giấy hoặc các bản kế hoạch PDF đã qua. Với sự phát triển của mô hình 3D, nơi hình học của từng yếu tố giờ đây liên kết trực tiếp với các thuộc tính khối lượng được tính toán tự động, quy trình xác định khối lượng và số lượng QTO đã trở thành tự động hóa.



Hình 5.23 Sự thu thập thuộc tính khối lượng và số lượng QTO từ dự án bao gồm việc nhóm và lọc các thành phần của dự án.

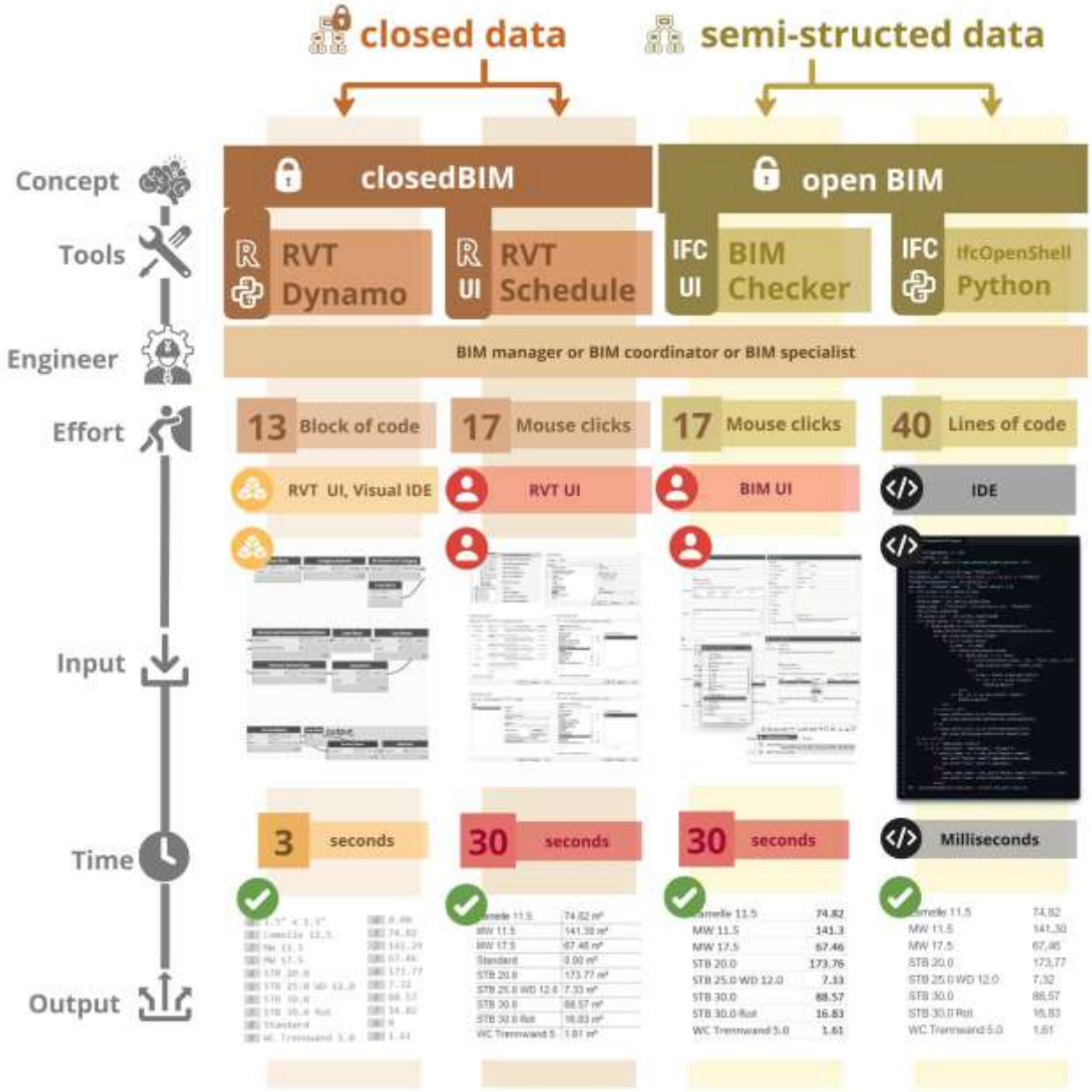
Trong ví dụ của chúng ta, nhiệm vụ là "chọn danh mục tường trong dự án, nhóm tất cả các thành phần theo loại và trình bày thông tin về các thuộc tính khối lượng dưới dạng định dạng bảng có cấu trúc", để bảng này có thể được sử dụng bởi hàng chục chuyên gia khác cho các phép tính, logistics, lịch làm việc và các trường hợp kinh doanh khác (Hình 6.13).

Do tính chất khép kín của dữ liệu CAD, không phải mọi chuyên gia hiện nay đều có thể truy cập trực tiếp vào cơ sở dữ liệu CAD (về nguyên nhân và giải pháp cho vấn đề truy cập được thảo luận chi tiết trong phần sáu của sách). Vì vậy, nhiều người phải sử dụng các công cụ BIM chuyên dụng, dựa trên các khái niệm open BIM và closed BIM [63]. Khi làm việc với các công cụ BIM chuyên dụng hoặc trực tiếp trong môi trường của chương trình CAD, bảng với kết quả QTO (Quantity Take-Off) có thể được hình thành theo nhiều cách khác nhau - tùy thuộc vào việc sử dụng giao diện thủ công hay tự động hóa phần mềm.

Ví dụ, bằng cách sử dụng giao diện người dùng của phần mềm CAD (BIM), chỉ cần thực hiện khoảng 17 hành động (nhấn nút) để có được bảng khối lượng hoàn chỉnh (Hình 5.24). Tuy nhiên, để làm điều này, người dùng cần hiểu rõ cấu trúc của mô hình và các chức năng của phần mềm CAD (BIM).

Nếu áp dụng tự động hóa thông qua mã lập trình hoặc qua các plugin và công cụ API bên trong phần mềm CAD, số lượng hành động thủ công để có được bảng khối lượng sẽ được giảm bớt, nhưng cần viết từ 40 đến 150 dòng mã, tùy thuộc vào thư viện hoặc công cụ được sử dụng:

- IfcOpSh (open BIM) hoặc Dynamo IronPython (closed BIM) - cho phép tạo bảng QTO từ định dạng CAD hoặc phần mềm CAD chỉ với khoảng 40 dòng mã.
- IFC_js (open BIM) - yêu cầu khoảng 150 dòng mã để trích xuất các thuộc tính khối lượng từ mô hình IFC.
- Các công cụ giao diện CAD (BIM) - cho phép đạt được cùng một kết quả một cách thủ công, với 17 cú nhấp chuột.



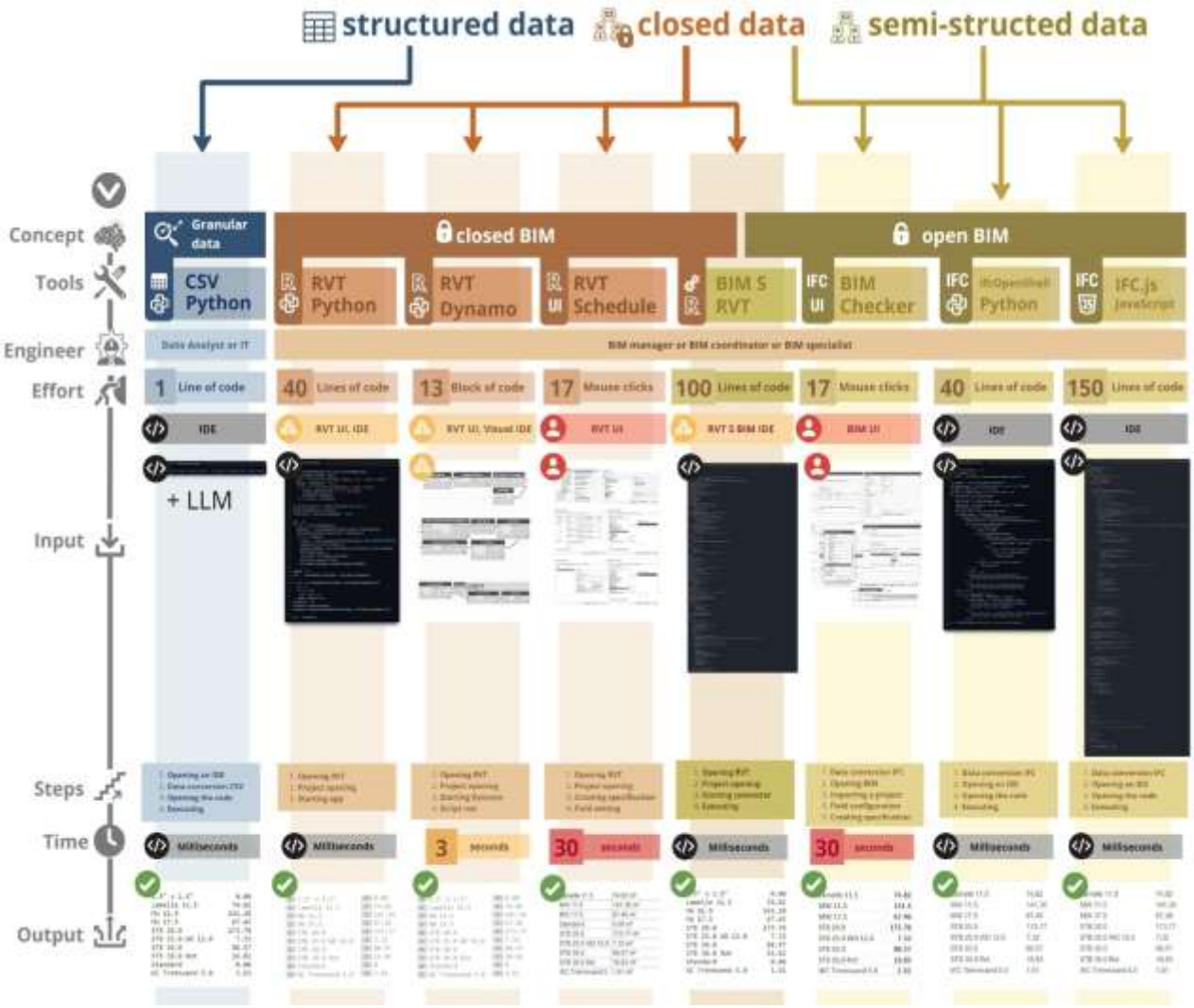
Hình 5.24 Các nhà thiết kế và quản lý CAD (BIM) sử dụng từ 40 đến 150 dòng mã hoặc một chục cú nhấp chuột để tạo bảng QTO.

Kết quả cuối cùng là giống nhau - bảng được cấu trúc với các thuộc tính khối lượng cho nhóm các yếu tố. Sự khác biệt chỉ nằm ở mức độ lao động và trình độ kỹ thuật cần thiết của người dùng (Hình 5.24). Các công cụ hiện đại, so với việc thu thập khối lượng thủ công, đã làm tăng tốc đáng kể quy trình QTO và giảm khả năng xảy ra lỗi. Chúng cho phép trích xuất dữ liệu trực tiếp từ mô hình dự án, loại bỏ nhu cầu tính toán khối lượng theo bản vẽ một cách thủ công, như đã từng làm trước đây.

Bất kể phương pháp được sử dụng - dù là open BIM hay closed BIM - đều có thể tạo ra bảng QTO với khối lượng của các yếu tố trong dự án (Hình 5.24). Tuy nhiên, khi làm việc với dữ liệu dự án trong các khái niệm CAD-(BIM-), người dùng phụ thuộc vào các công cụ chuyên biệt và API do các nhà cung cấp cung cấp (Hình

3.213). Điều này tạo ra các mức độ phụ thuộc bổ sung và yêu cầu nghiên cứu các sơ đồ dữ liệu độc đáo, đồng thời hạn chế quyền truy cập trực tiếp vào dữ liệu. --

Do tính khép kín của dữ liệu CAD, việc thu được bảng QTO và các tham số khác trở nên khó khăn, tự động hóa tính toán và tích hợp với các hệ thống bên ngoài. Bằng cách sử dụng các công cụ truy cập trực tiếp vào cơ sở dữ liệu và chuyển đổi dữ liệu CAD của dự án thông qua các công cụ reverse engineering sang định dạng cấu trúc mở của dataframe (Hình 4.113), bảng QTO tương tự có thể được tạo ra chỉ với một dòng mã (Hình 5.25 - biến thể với dữ liệu phân cấp). --



Hình 5.25 Các công cụ khác nhau cho ra các kết quả giống nhau dưới dạng bảng thuộc tính của các thực thể trong dự án, nhưng với các mức độ lao động khác nhau.

Khi sử dụng dữ liệu mở và có cấu trúc từ các dự án CAD, như đã đề cập trong chương "Chuyển đổi dữ liệu CAD (BIM) thành dạng cấu trúc", quy trình nhóm, QTO, được đơn giản hóa đáng kể.

Các phương pháp dựa trên việc sử dụng dữ liệu có cấu trúc mở hoặc truy cập trực tiếp vào cơ sở dữ liệu của các mô hình CAD, được miễn các hạn chế tiếp thị liên quan đến từ viết tắt BIM. Chúng dựa vào các công cụ đã được kiểm chứng, đã được áp dụng trong các ngành khác (Hình 7.310 quy trình ETL).-

> Theo nghiên cứu của McKinsey "Dữ liệu mở: Khám phá đổi mới và năng suất thông qua thông tin liên tục" [102], được thực hiện vào năm 2013, việc sử dụng dữ liệu mở có thể tạo ra cơ hội tiết kiệm từ 30 đến 50 tỷ đô la mỗi năm trong việc thiết kế, kỹ thuật, mua sắm và xây dựng các cơ sở hạ tầng điện lực. Điều này có nghĩa là tiết kiệm 15% chi phí vốn trong xây dựng.

Làm việc với dữ liệu mở có cấu trúc (granular) giúp việc tìm kiếm và xử lý thông tin trở nên dễ dàng hơn, giảm phụ thuộc vào các nền tảng BIM chuyên dụng và mở đường cho việc tự động hóa mà không cần sử dụng các hệ thống độc quyền hay các mô hình dữ liệu phức tạp và tham số từ các định dạng CAD.

Tự động hóa QTO bằng cách sử dụng LLM và dữ liệu có cấu trúc

Việc chuyển đổi dữ liệu không có cấu trúc sang dạng có cấu trúc nâng cao đáng kể hiệu quả của nhiều quy trình: nó đơn giản hóa việc xử lý dữ liệu (Hình 4.11, Hình 4.12) và tăng tốc quá trình xác thực, làm cho các yêu cầu trở nên rõ ràng và minh bạch, như chúng tôi đã đề cập trong các chương trước. Tương tự, việc chuyển đổi dữ liệu CAD (BIM) sang dạng có cấu trúc mở (Hình 4.112, Hình 4.113) giúp làm cho quy trình nhóm thuộc tính và quy trình QTO trở nên dễ dàng hơn.---

Bảng thuộc tính QTO có dạng có cấu trúc, vì vậy khi sử dụng dữ liệu CAD có cấu trúc, chúng tôi làm việc với một mô hình dữ liệu duy nhất (Hình 5.25), điều này giúp chúng tôi không cần phải thực hiện chuyển đổi và đồng bộ hóa cấu trúc mô hình dữ liệu dự án và quy tắc nhóm lại. Điều này cho phép nhóm dữ liệu theo một hoặc nhiều thuộc tính chỉ với một dòng mã. Ngược lại, trong open BIM và closed BIM, nơi dữ liệu được lưu trữ theo dạng bán cấu trúc, tham số hóa hoặc dạng kín, việc xử lý đòi hỏi hàng chục hoặc thậm chí hàng trăm dòng mã, cũng như việc sử dụng API để tương tác với hình học và thông tin thuộc tính.-

- Ví dụ về việc nhóm QTO của một dự án có cấu trúc theo một thuộc tính. Yêu cầu văn bản trên bất kỳ chat LLM nào (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ cái nào khác):

  > Tôi có một dự án CAD dưới dạng DataFrame – vui lòng lọc dữ liệu dự án để lấy các phần tử có tham số "Type" chỉ chứa giá trị "Type 1" ⏎

- Câu trả lời của LLM có khả năng cao sẽ ở dạng mã Python sử dụng Pandas:



Hình 5.26 Một dòng mã, được viết bằng cách sử dụng LLM, cho phép nhóm toàn bộ dự án CAD theo thuộc tính "Type" và lấy được nhóm phần tử cần thiết.

Nhờ vào cấu trúc đơn giản của DataFrame hai chiều, chúng tôi không cần phải giải thích sơ đồ và mô hình dữ liệu cho LLM, điều này giúp giảm bớt các giai đoạn diễn giải và tăng tốc độ tạo ra các giải pháp cuối cùng. Trước đây, để viết mã ngay cả đơn giản cũng cần học các ngôn ngữ lập trình, nhưng giờ đây các mô hình ngôn ngữ hiện đại (LLM) cho phép tự động chuyển đổi logic quy trình thành mã khi làm việc với dữ liệu có cấu trúc thông qua các yêu cầu văn bản.

> Tự động hóa và các mô hình ngôn ngữ LLM có thể hoàn toàn loại bỏ nhu cầu cho các chuyên gia làm việc với việc nhóm và xử lý dữ liệu CAD (BIM) phải học các ngôn ngữ lập trình hoặc công cụ BIM, tạo điều kiện để giải quyết các nhiệm vụ thông qua các yêu cầu văn bản.

Yêu cầu tương tự — nhóm tất cả các phần tử của dự án trong danh mục "tường" với việc tính toán khối lượng theo từng loại (Hình 5.25) — mà trong môi trường CAD (BIM) yêu cầu 17 lần nhấn chuột trong giao diện hoặc viết 40 dòng mã, trong các công cụ xử lý dữ liệu mở (như SQL hoặc Pandas) lại trở nên như một yêu cầu đơn giản và trực quan.

- Bằng cách sử dụng một dòng lệnh trong Pandas:

```
df[df['Category'].isin(['OST_Walls'])].groupby('Type')['Volume'].sum()
```

Giải thích mã: lấy từ df (DataFrame) các phần tử mà thuộc tính-cột "Category" có giá trị "OST_Walls", nhóm tất cả các phần tử vừa thu được theo thuộc tính-cột "Type" và tính tổng cho nhóm phần tử khoảng "Volume".

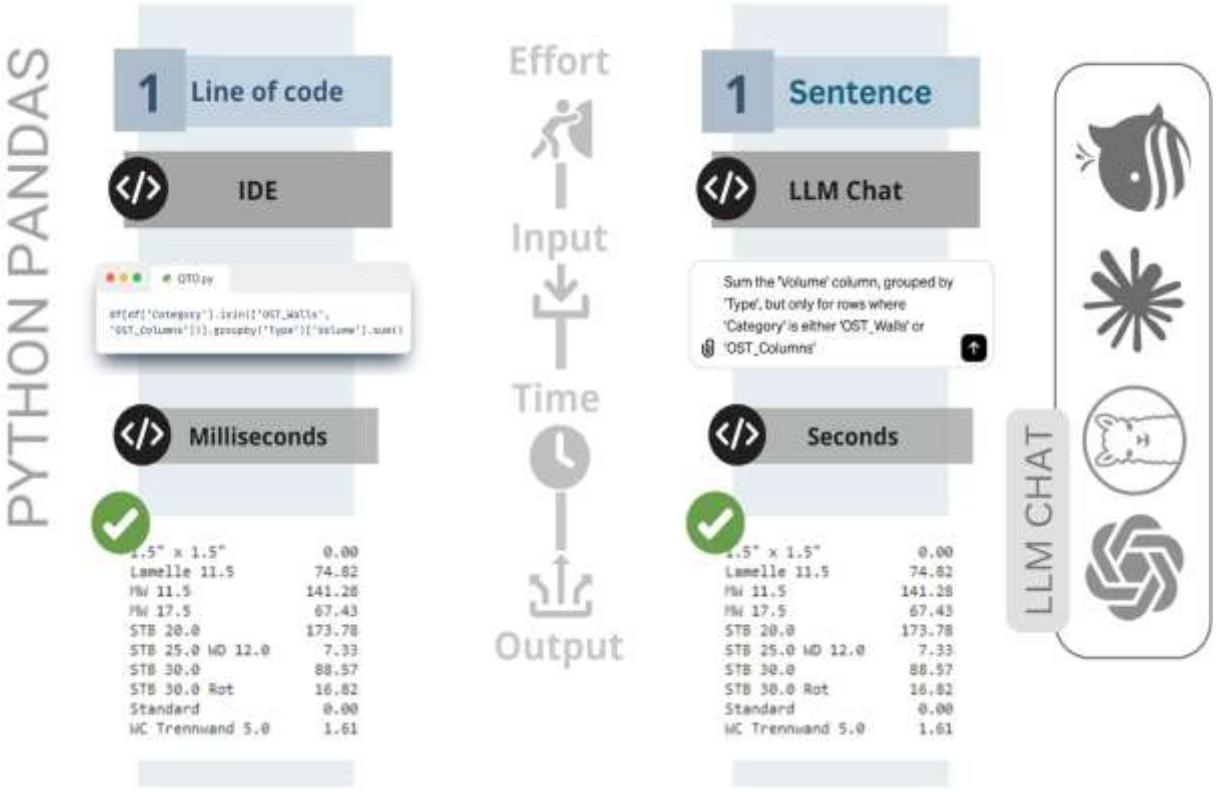
- Nhóm cấu trúc dự án đã được tạo từ CAD bằng cách sử dụng SQL:

```
SELECT Type, SUM(Volume) AS TotalVolumeFROM elements  WHERE Category = 'OST_Walls'
GROUP BY Type;
```

- Bằng cách sử dụng yêu cầu LLM để nhóm cơ sở dữ liệu dự án, chúng ta có thể ghi lại dưới dạng

một yêu cầu ngắn gọn - prompt (Hình 5.27): -

> Đối với DataFrame dự án, hãy nhóm các phần tử theo tham số 'Type', nhưng chỉ cho những phần tử có tham số 'Category' bằng 'OST_Walls' hoặc 'OST_Columns' và vui lòng cộng tổng cột tham số 'Volume' cho nhóm vừa tạo ⏎

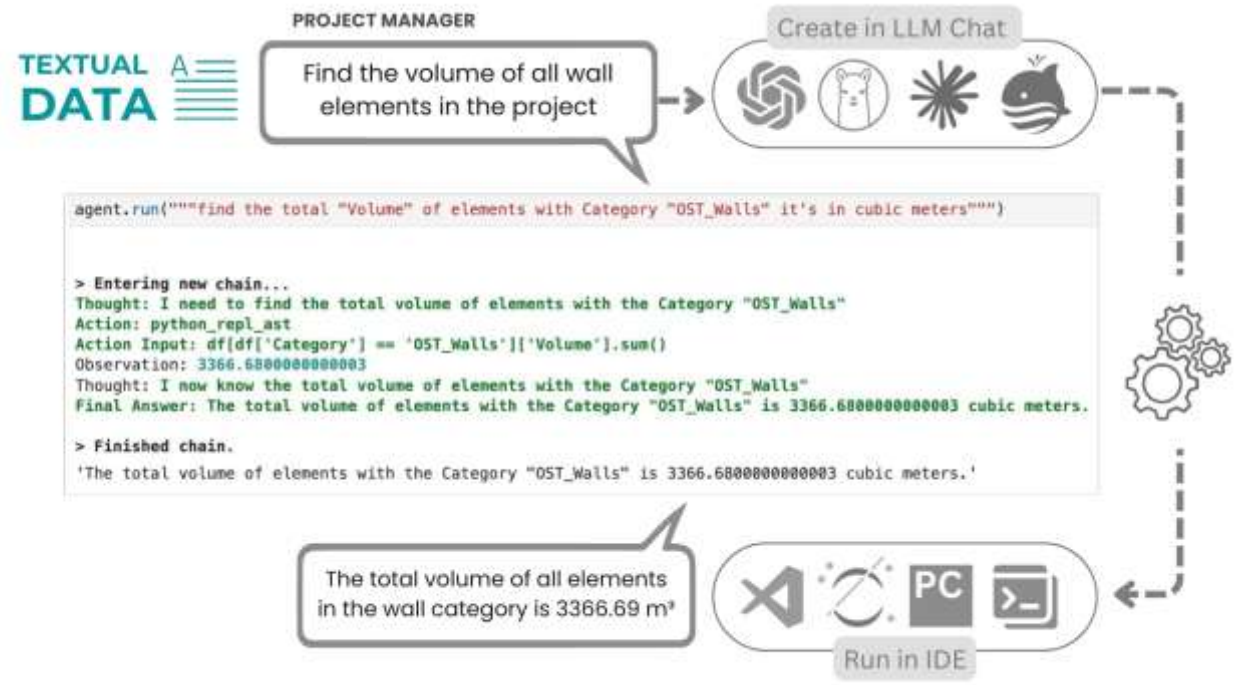


Hình 5.27 Nhờ việc sử dụng SQL, Pandas và LLM, tự động hóa quy trình xử lý dữ liệu giờ đây có thể thực hiện chỉ với một vài dòng mã và yêu cầu bằng văn bản.

Việc nhận diện QTO từ dữ liệu CAD bằng cách sử dụng các công cụ LLM (ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok) đã thay đổi cách truyền thống để trích xuất thông tin thuộc tính, dữ liệu định lượng và thể tích cho các đối tượng riêng lẻ và nhóm.

> Bây giờ ngay cả các nhà quản lý dự án, chuyên gia tính toán chi phí hoặc logistics, những người không có kiến thức sâu về thiết kế và không có phần mềm CAD (BIM) chuyên dụng, chỉ cần truy cập vào cơ sở dữ liệu CAD có thể trong vài giây để lấy tổng thể tích các phần tử thuộc loại tường hay các đối tượng khác, đơn giản chỉ cần viết hoặc đọc to yêu cầu.

Trong các yêu cầu bằng văn bản (Hình 5.28), mô hình LLM xử lý yêu cầu của người dùng để áp dụng một chức năng cụ thể cho một hoặc một vài tham số - cột của bảng. Kết quả, người dùng trong cuộc giao tiếp với LLM nhận được hoặc một cột-tham số mới với các giá trị mới, hoặc một giá trị cụ thể sau khi nhóm lại.



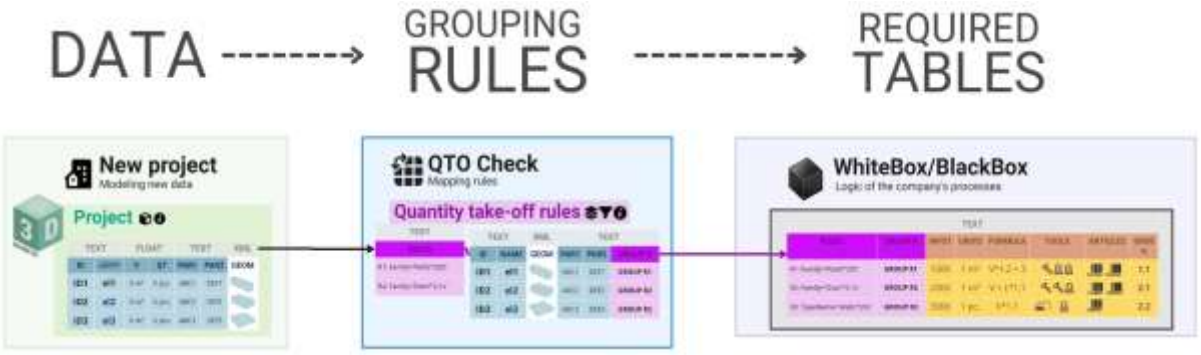
Hình 5.28 Mô hình LLM, khi làm việc với các dữ liệu có cấu trúc, hiểu từ ngữ cảnh của yêu cầu văn bản, người dùng muốn nhóm lại và thuộc tính nào.

Nếu cần thu thập các chỉ số thể tích chỉ cho một nhóm phần tử, chỉ cần thực hiện một yêu cầu QTO đơn giản (Hình 5.27) đến dữ liệu của mô hình CAD. Tuy nhiên, khi tính toán ngân sách hoặc ước lượng cho toàn bộ dự án bao gồm nhiều nhóm phần tử, thường yêu cầu phải trích xuất các đặc tính định lượng cho tất cả các loại phần tử (class), nơi mỗi loại phần tử được xử lý riêng biệt - với việc nhóm theo các thuộc tính tương ứng.

Trong thực tế của chuyên gia ước lượng, người thẩm định sử dụng các quy tắc nhóm và tính toán đặc biệt cho các loại đối tượng khác nhau. Ví dụ, cửa sổ thường được nhóm theo các tầng hoặc khu vực (tham số nhóm - thuộc tính Level, Rooms), còn tường thì nhóm theo loại vật liệu hoặc loại kết cấu (tham số Material, Type). Để tự động hóa quy trình nhóm, các quy tắc này được mô tả trước theo dạng bảng quy tắc nhóm. Những bảng này hoạt động như các mẫu cấu hình, xác định thuộc tính nào nên được sử dụng trong các tính toán cho từng nhóm phần tử trong dự án.

Tính toán QTO cho toàn bộ dự án bằng cách sử dụng quy tắc từ bảng Excel

Trong các dự án xây dựng thực tế, thường xuyên cần thiết phải thực hiện việc tổng hợp theo nhiều thuộc tính cùng một lúc trong cùng một nhóm các phần tử. Ví dụ, khi làm việc với danh mục "Cửa sổ" (nơi thuộc tính Category chứa các giá trị như OST_Windows hoặc IfcWindows), các phần tử có thể được nhóm không chỉ theo loại - chẳng hạn như theo giá trị trong trường Type Name hoặc Type - mà còn theo các đặc điểm bổ sung, chẳng hạn như mức độ dẫn nhiệt được chỉ định trong thuộc tính tương ứng. Việc nhóm đa chiều như vậy giúp thu được kết quả chính xác hơn cho một nhóm cụ thể. Tương tự, trong việc tính toán theo các danh mục tường hoặc sàn, có thể sử dụng các tổ hợp tùy ý của các thuộc tính - chẳng hạn như vật liệu, cấp độ, tầng, khả năng chống cháy và các tham số khác - làm bộ lọc hoặc tiêu chí nhóm.



Đối với mỗi nhóm hoặc danh mục thực thể trong dự án, có một công thức nhóm riêng, bao gồm một hoặc nhiều tiêu chí.

Quy trình xác định các quy tắc nhóm như vậy tương tự như quy trình tạo yêu cầu về dữ liệu, đã được mô tả trong chương "Tạo yêu cầu và kiểm tra chất lượng dữ liệu", nơi chúng tôi đã xem xét kỹ lưỡng công việc với các mô hình dữ liệu. Các quy tắc nhóm và tính toán như vậy đảm bảo độ chính xác và tính thích hợp của kết quả cho việc tính toán tự động tổng hợp các thuộc tính về số lượng hoặc khối lượng của danh mục thực thể, với tất cả các điều kiện cần thiết được xem xét trong các phép tính và tính toán.-

- Ví dụ sau đây của mã lọc bảng dự án theo cách mà tập hợp dữ liệu kết quả chỉ chứa những thực thể mà trong đó thuộc tính-cột "Category" chứa các giá trị "OST_Windows" hoặc "IfcWindows" và đồng thời thuộc tính-cột "Type" chứa giá trị "Type 1":

> Tôi có một DataFrame dự án - hãy lọc dữ liệu sao cho tập dữ liệu chỉ còn lại các phần tử mà thuộc tính "Category" chứa các giá trị "OST_Windows" hoặc "IfcWindows" và đồng thời thuộc tính Type chứa giá trị "Type 1".

- Phản hồi từ LLM:



Một dòng mã, tương tự như công thức Excel, cho phép nhóm tất cả các thực thể dự án theo nhiều đặc điểm.

Mã thu được (hình 5.210) sau khi chuyển đổi dữ liệu CAD sang các định dạng mở có cấu trúc có thể được chạy trong một trong các IDE phổ biến mà chúng tôi đã đề cập ở trên, ở chế độ offline: PyCharm, Visual

Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse với plugin PyDev, Thonny, Wing IDE, IntelliJ IDEA với plugin Python, JupyterLab hoặc các công cụ trực tuyến phổ biến: Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.-

- Để lấy các thực thể của dự án dưới dạng QTO DataFrame theo danh mục “Cửa sổ” chỉ với một giá trị dẫn nhiệt cụ thể, chúng ta có thể sử dụng truy vấn sau cho LLM:

> Tôi có một DataFrame dự án - hãy lọc dữ liệu sao cho tập dữ liệu chỉ còn lại các bản ghi với "Category" chứa các giá trị "OST_Windows" hoặc "IfcWindows", và đồng thời cột ThermalConductivity phải có giá trị 0,5.

- Phản hồi từ LLM:



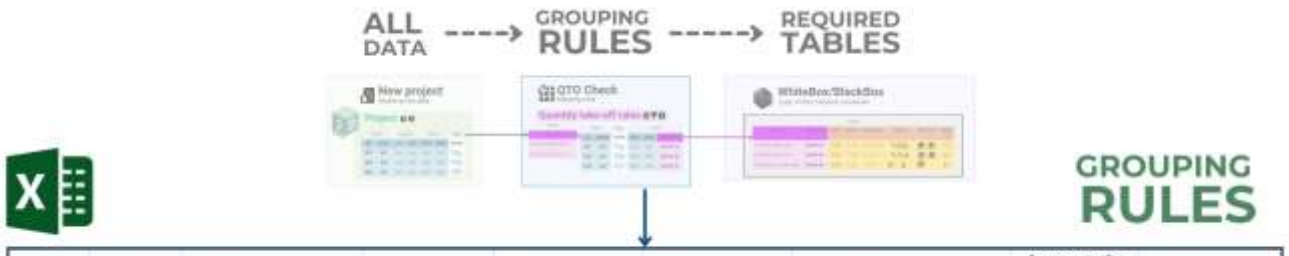
Ngôn ngữ truy vấn cực kỳ đơn giản của Pandas Python cho phép thực hiện QTO cho bất kỳ số lượng dự án nào cùng một lúc.

Trong phản hồi nhận được từ LLM (hình 5.211), sử dụng điều kiện logic "&" để kết hợp hai tiêu chí: giá trị dẫn nhiệt và sự thuộc về một trong hai danh mục. Phương pháp "isin" kiểm tra xem giá trị thuộc tính-cột "Category" có nằm trong danh sách được cung cấp hay không.-

Trong các dự án có nhiều nhóm thành phần với các logic phân nhóm khác nhau - cho mỗi thể loại thực thể của dự án (ví dụ: cửa sổ, cửa ra vào, sàn) cần thiết lập các quy tắc phân nhóm riêng biệt, có thể bao gồm các hệ số bổ sung hoặc công thức tính toán cuối cùng cho các thuộc tính. Những công thức này (Hình 5.212 thuộc tính "formel", ví dụ giá trị x của số lượng và y thể tích của nhóm) và các hệ số tính đến các đặc điểm độc đáo của từng nhóm, ví dụ:-

- phần trăm bổ sung cho thể tích vật liệu để tính đến sự lãng phí
- số lượng vật liệu bổ sung cố định
- các điều chỉnh liên quan đến các rủi ro và sai số có thể trong tính toán dưới dạng công thức

Sau khi các quy tắc lọc và phân nhóm được hình thành dưới dạng các công thức tham số cho từng loại phần tử, chúng có thể được lưu trữ dưới dạng bảng dòng - ví dụ, dưới định dạng Excel (Hình 5.212). Việc lưu trữ các quy tắc này theo cách có cấu trúc cho phép tự động hóa hoàn toàn quá trình trích xuất, lọc và phân nhóm dữ liệu dự án. Thay vì viết tay nhiều truy vấn riêng lẻ, hệ thống chỉ cần đọc bảng tham số và áp dụng các quy tắc tương ứng cho mô hình (dataframe chung của dự án (Hình 4.113)), tạo ra các bảng QTO cuối cùng cho từng thể loại phần tử của dự án.-



pos	code	description	Parametr 1	Expression1	Parametr 2	Expression2	Aggregate Parameter	formel
F1	3	Walls						
W1.1	3.1.1	Interior - 165	Category	OST_Walls	Type Name	Interior - 165*	Area	x*1.5+y*0.1
W1.2	3.1.2	SIP Wall 202mm	Category	OST_Walls	Type Name	SIP 202mm Wall*	Volume	x/1.2-y+0.4
W1.3	3.1.2	CL W1	Category	OST_Walls	Type Name	CL_W1*	Area	x*0.5*y+2
W2.1	3.2.2	Retaining 300mm	Category	OST_Walls	Type Name	Retaining - 300mm*	Volume	x*1.1+8+y*0.2

Hình 5.212 Bảng phân nhóm thuộc tính QTO thiết lập các quy tắc phân nhóm các phần tử của dự án, đảm bảo số lượng và thể tích tổng chính xác cho từng thể loại.

Các quy tắc được tập hợp sẽ cho phép phân nhóm toàn bộ dự án và thực hiện tất cả các tính toán cần thiết, bao gồm việc điều chỉnh thuộc tính thể tích. Kết quả là các thể tích được điều chỉnh thành "thể tích thực tế", được sử dụng cho tính toán và báo giá, không phải là những gì ban đầu tại giai đoạn thiết kế trong mô hình CAD.

> Trong quá trình tự động tạo bảng QTO thể tích cho toàn bộ dự án, ứng dụng cần đi qua tất cả các thể loại của bảng quy tắc phân nhóm, lấy các thuộc tính phân nhóm, phân nhóm tất cả các phần tử của dự án theo chúng và tổng hợp thuộc tính thể tích cho nhóm này, đồng thời nhân nó với yếu tố hoặc hệ số điều chỉnh.

Chúng ta sẽ yêu cầu LLM viết mã cho giải pháp này, trong đó mã sẽ cần tải hai bảng - bảng quy tắc phân nhóm (Hình 5.212) và bảng dữ liệu của chính dự án (Hình 4.113), sau đó áp dụng các quy tắc phân nhóm, phân nhóm các phần tử theo các quy tắc đã chỉ định, tính toán các giá trị tổng hợp và lưu kết quả vào một tệp Excel mới.--

- Gửi yêu cầu văn bản đến LLM chat:

> Tôi cần mã để đọc dữ liệu dự án từ tệp “basic_sample_project.xlsx”, và sau đó là quy tắc từ “Grouping_rules_QTO.xlsx” theo đó phân nhóm tất cả dữ liệu theo 'Parameter 1' và 'Parameter 2', tổng hợp 'Aggregate Parameter', lọc theo 'Expression2', thực hiện các phép tính từ 'Formel1' và lưu bảng QTO vào 'QTQ_table2.xlsx' ⏎

- Phản hồi từ LLM:



Hình 5.213 Mã Python Pandas xử lý dữ liệu dự án từ mô hình CAD, phân nhóm chúng theo các quy tắc được chỉ định trong Excel.

Kết quả cuối cùng của việc thực hiện mã (Hình 5.213) sẽ là bảng nhóm thực thể, chứa không chỉ các thuộc tính tổng hợp thể tích từ mô hình CAD- (BIM-) gốc mà còn cả thuộc tính mới của thể tích thực tế, tính đến tất cả các yêu cầu để tạo ra các phép tính và báo giá chính xác (Ví dụ Hình 5.214).-

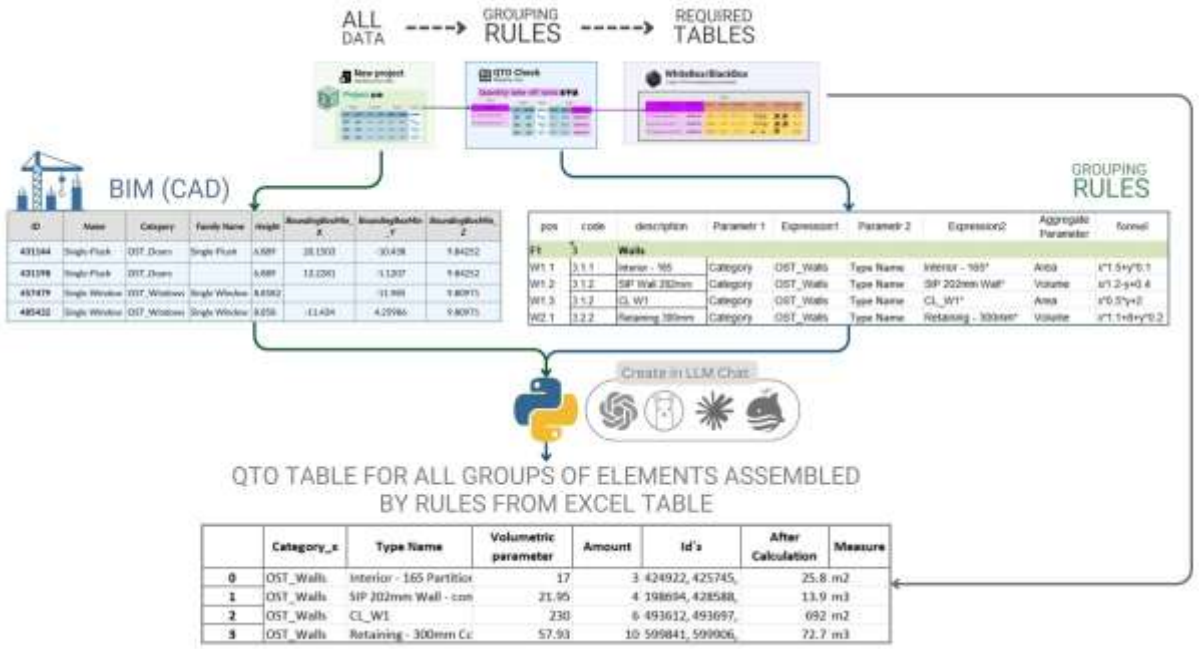
QTO TABLE FOR ALL GROUPS OF ELEMENTS ASSEMBLED BY RULES FROM EXCEL TABLE

	Category_x	Type Name	Volumetric parameter	Amount	Id´s	After Calculation	Measure
0	OST_Walls	Interior - 165 Partitior	17	3	424922, 425745,	25.8	m2
1	OST_Walls	SIP 202mm Wall - con	21.95	4	198694, 428588,	13.9	m3
2	OST_Walls	CL_W1	230	6	493612, 493697,	692	m2
3	OST_Walls	Retaining - 300mm Cc	57.93	10	599841, 599906,	72.7	m3

Hình 5.214 Thuộc tính "Sau khi tính toán" được thêm vào bảng tổng hợp sau khi thực hiện mã, mã sẽ tự động tính toán thể tích thực tế.

Mã còn lại (Hình 5.213) có thể được chạy trên một trong những IDE phổ biến (như chúng ta đã đề cập trước đó) và áp dụng mã này cho bất kỳ số lượng dự án đã tồn tại hoặc mới đến nào (RVT, IFC, DWG, NWS, DGN và các định dạng khác), cho dù là một vài dự án hay có thể hàng trăm dự án ở các định dạng khác nhau,

được đưa vào dạng có cấu trúc (Hình 5.215).-



	Category_x	Type Name	Volumetric parameter	Amount	Id's	After Calculation	Measure
0	OST_Walls	Interior - 165 Partitio	17	3	424922, 425745,	25.8	m2
1	OST_Walls	SIP 202mm Wall - con	21.95	4	198694, 428588,	13.9	m3
2	OST_Walls	CL_W1	230	6	493612, 493697,	692	m2
3	OST_Walls	Retaining - 300mm Co	57.93	10	599841, 599906,	72.7	m3

Hình 5.215 Quy trình tự động nhóm dữ liệu xây dựng liên kết dữ liệu BIM (CAD) với bảng QTO thông qua các quy tắc từ bảng tính Excel.

Quy trình thu thập dữ liệu khối lượng đã được cấu hình và tham số hóa (Hình 5.215) cho phép tự động hóa hoàn toàn việc thu thập dữ liệu về thuộc tính số lượng và khối lượng của các phần tử dự án để sử dụng tiếp theo, bao gồm ước tính chi phí, logistics, lịch trình công việc và tính toán dấu chân carbon cùng các nhiệm vụ phân tích khác.

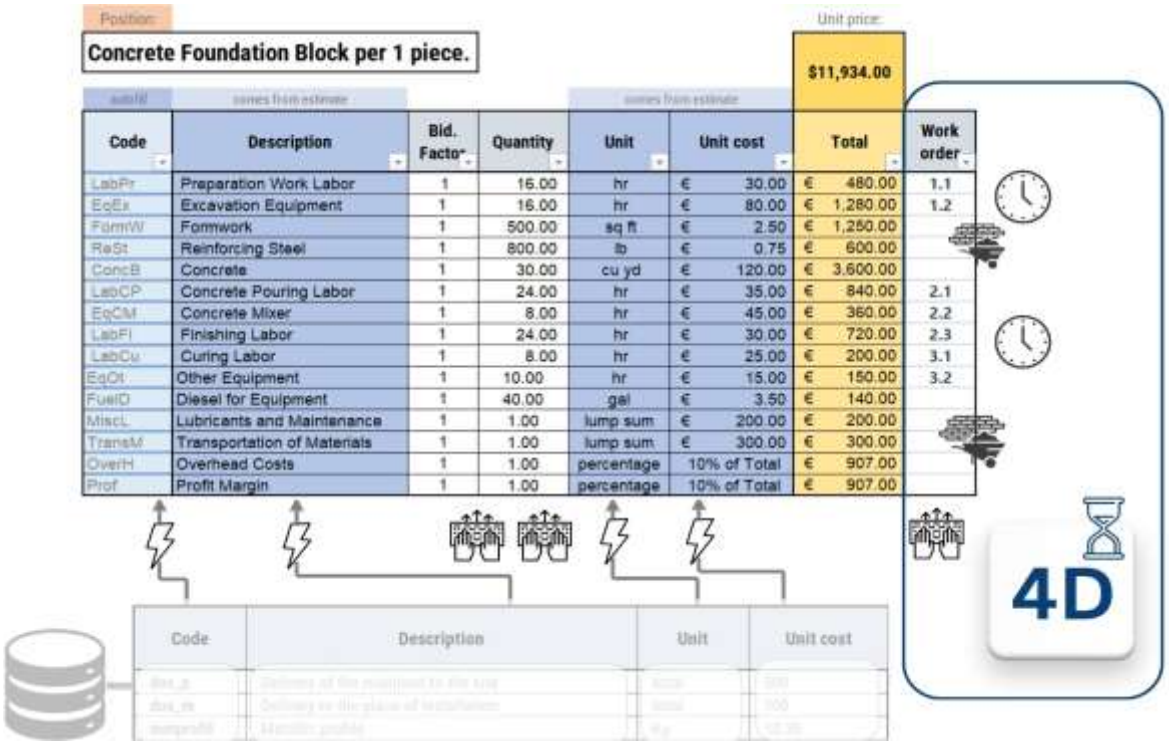
Sau khi nghiên cứu các công cụ cho phép tổ chức và nhóm các nhóm phần tử dự án theo các tiêu chí nhất định, chúng ta đã sẵn sàng tích hợp các dự án đã được nhóm và lọc với các tính toán và kịch bản kinh doanh khác nhau của công ty.

CHƯƠNG 5.3. 4D, 6D-8D VÀ TÍNH TOÁN KHÍ THẢI CO_2

Mô hình 4D: tích hợp thời gian vào dự toán xây dựng

Ngoài việc tính toán chi phí, một trong những hướng chính trong việc áp dụng dữ liệu dự án trong xây dựng là xác định các tham số thời gian — cả cho các hoạt động xây dựng riêng lẻ và cho toàn bộ dự án. Để tính toán tự động các thời gian và lập lịch trình thực hiện công việc, thường dựa trên phương pháp đánh giá tài nguyên và cơ sở dữ liệu tính toán liên quan, đã được xem xét chi tiết trong chương trước với tiêu đề «Tính toán và ước tính cho các dự án xây dựng».

Trong khuôn khổ tiếp cận theo tài nguyên, không chỉ có chi phí vật liệu được xem xét, mà còn cả tài nguyên thời gian. Khi lập bảng tính, mỗi quy trình có thể được gán thuộc tính thứ tự thực hiện công việc (Hình 5.31 – thuộc tính «Work order»), cũng như chỉ định số thời gian và chi phí liên quan đến việc thực hiện quy trình đó. Những tham số này đặc biệt quan trọng để mô tả các hoạt động không có giá cố định trên thị trường và không thuộc về mua sắm trực tiếp — như việc sử dụng thiết bị xây dựng, sự tham gia của công nhân hoặc các quá trình logistics (thường được thể hiện bằng giờ). Trong những trường hợp như vậy, chi phí không được xác định bởi bộ phận mua sắm, mà được xác định trực tiếp bởi công ty thực hiện dựa trên các tiêu chuẩn nội bộ hoặc mức giá sản xuất (Hình 5.31).-



Code	Description	Bid. Factor	Quantity	Unit	Unit cost	Total	Work order
LabPr	Preparation Work Labor	1	16.00	hr	€ 30.00	€ 480.00	1.1
EqEx	Excavation Equipment	1	16.00	hr	€ 80.00	€ 1,280.00	1.2
FormW	Formwork	1	500.00	sq ft	€ 2.50	€ 1,250.00	
ReSt	Reinforcing Steel	1	800.00	lb	€ 0.75	€ 600.00	
ConcB	Concrete	1	30.00	cu yd	€ 120.00	€ 3,600.00	
LabCP	Concrete Pouring Labor	1	24.00	hr	€ 35.00	€ 840.00	2.1
EqCM	Concrete Mixer	1	8.00	hr	€ 45.00	€ 360.00	2.2
LabFi	Finishing Labor	1	24.00	hr	€ 30.00	€ 720.00	2.3
LabCu	Curing Labor	1	8.00	hr	€ 25.00	€ 200.00	3.1
EqOt	Other Equipment	1	10.00	hr	€ 15.00	€ 150.00	3.2
FuelD	Diesel for Equipment	1	40.00	gal	€ 3.50	€ 140.00	
MiscL	Lubricants and Maintenance	1	1.00	lump sum	€ 200.00	€ 200.00	
TransM	Transportation of Materials	1	1.00	lump sum	€ 300.00	€ 300.00	
OverH	Overhead Costs	1	1.00	percentage	10% of Total	€ 907.00	
Prof	Profit Margin	1	1.00	percentage	10% of Total	€ 907.00	

Hình 5.31 Các tính toán công việc trong phương pháp đánh giá tài nguyên bao gồm cả chi phí thời gian của thời gian làm việc.

Do đó, các tính toán ở cấp độ bảng tính không chỉ bao gồm chi phí nhiên liệu và vật liệu (chi phí mua vào), mà còn cả thời gian làm việc của thợ máy, thiết bị và công nhân hỗ trợ trên công trường. Trong ví dụ được lập (Hình 5.31), bảng chi phí đại diện cho chi phí lắp đặt khối bê tông móng, bao gồm các giai đoạn công việc thành phần, chẳng hạn như chuẩn bị, lắp đặt khung và đổ bê tông, cùng với các vật liệu cần thiết và

công sức lao động. Trong đó, các hoạt động riêng lẻ, chẳng hạn như công việc chuẩn bị, có thể không có chi phí vật chất, nhưng lại chứa đựng chi phí thời gian lao động đáng kể, được thể hiện bằng giờ công.

Để lên kế hoạch cho thứ tự công việc (cho lịch trình công việc) tại công trường, ta có thể thêm thủ công một thuộc tính "Work order" vào bảng tính toán (Hình 5.31). Nó được chỉ định trong cột bổ sung chỉ cho các mục có đơn vị đo lường được biểu thị bằng thời gian (giờ, ngày). Thuộc tính này bổ sung cho mã công việc, mô tả, số lượng, đơn vị đo lường (tham số "Unit") và chi phí. Dãy số (tham số "Work order") của các công việc cho phép xác định thứ tự thực hiện các nhiệm vụ trên công trường và sử dụng nó khi lập lịch trình.

Lịch trình xây dựng và tự động hóa dựa trên dữ liệu tính toán

Lịch trình xây dựng là một đại diện trực quan của kế hoạch thực hiện các công việc và quy trình mà cần thực hiện trong khuôn khổ dự án. Nó được tạo ra dựa trên các tính toán nguồn lực chi tiết (Hình 5.31), trong đó mỗi công việc được ghi rõ, bên cạnh chi phí nguồn lực, theo thời gian và thứ tự.

Khác với các phương pháp trung bình, nơi mà tính toán thời gian được xây dựng dựa trên số giờ chuẩn để lắp đặt vật liệu hoặc thiết bị, trong phương pháp tài nguyên, việc lập kế hoạch dựa trên dữ liệu thực tế đã được đưa vào tính toán. Mỗi mục trong dự toán liên quan đến lao động đều dựa vào lịch, trong đó xem xét các điều kiện thực tế của việc sử dụng nguồn lực trong suốt thời gian làm việc. Việc điều chỉnh giờ làm việc hiệu quả thông qua các hệ số ở mức tính toán (Hình 5.31 tham số "Bid. Factor") cho phép xem xét những khác biệt trong năng suất và đặc điểm mùa vụ, ảnh hưởng đến thời gian hoàn thành công việc.

Để xác định ngày bắt đầu và kết thúc quá trình cho lịch trình xây dựng trên biểu đồ Gantt, chúng ta lấy các giá trị thuộc tính thời gian cho mỗi mục từ bảng tính toán các khối móng và nhân chúng với số lượng khối (trong trường hợp này là số lượng khối bê tông móng). Tính toán này cho biết độ dài của mỗi nhiệm vụ. Sau đó, chúng ta đánh dấu những độ dài này trên trục thời gian, bắt đầu từ ngày khởi động dự án, để xây dựng lịch trình và cuối cùng có được một đại diện trực quan cho thấy khi nào mỗi nhiệm vụ sẽ bắt đầu và kết thúc. Tham số "Work order" trong các quy trình cũng cho phép chúng ta hiểu rõ quá trình công việc diễn ra song song ("Work order" ví dụ như 1.1-1.1) hoặc tuần tự (1.1-1.2).

> Biểu đồ Gantt là một công cụ đồ họa dùng để lập kế hoạch và quản lý các dự án, thể hiện các nhiệm vụ dưới dạng các dải ngang trên trục thời gian. Mỗi dải hiển thị độ dài thực hiện nhiệm vụ, thời gian bắt đầu và kết thúc của nó.

Lịch trình công việc, hay biểu đồ Gantt, giúp các nhà quản lý dự án và công nhân hiểu rõ khi nào và theo thứ tự nào các giai đoạn xây dựng khác nhau cần được thực hiện, đảm bảo sử dụng hiệu quả nguồn lực và tuân thủ thời hạn.

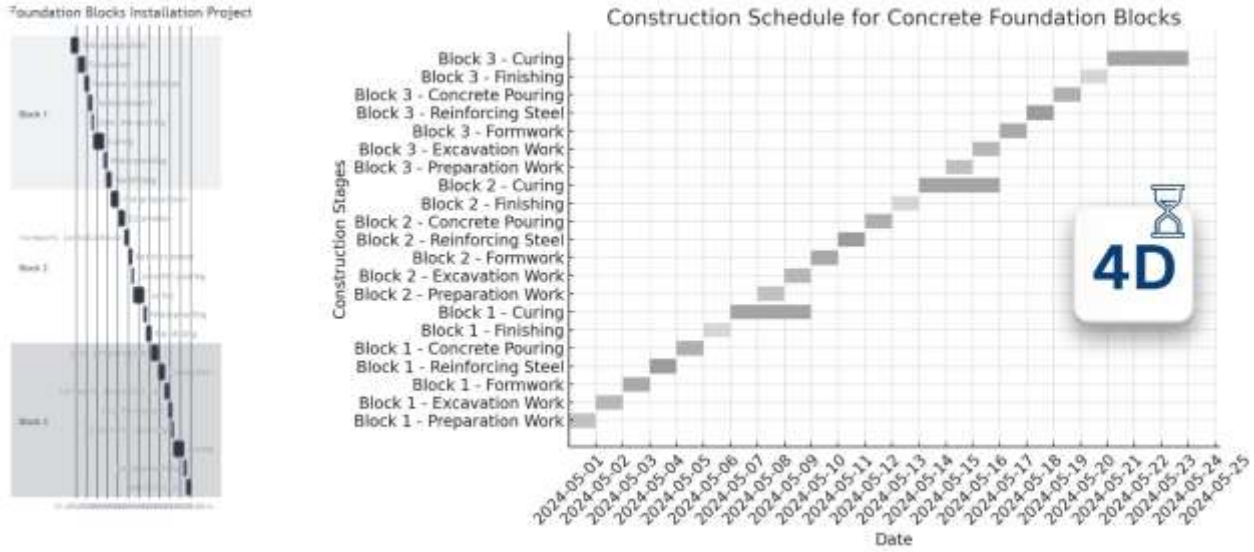
Giả định kế hoạch lịch công việc cho việc lắp đặt ba khối bê tông móng bằng cách sử dụng các tính toán từ bảng trên. Sử dụng bảng chi phí (Hình 5.31) từ ví dụ trên, hãy yêu cầu LLM lập kế hoạch lắp đặt 3 khối bê tông móng, ví dụ vào ngày 1 tháng 5 năm 2024.

Để gửi bảng tính toán cho LLM, chúng ta có thể tải lên bảng tính toán ở định dạng XLSX hoặc đơn giản là chèn một hình chụp màn hình của bảng tính toán ở định dạng JPEG trực tiếp vào trò chuyện với LLM (Hình 5.31). LLM sẽ tự động tìm thư viện để trực quan hóa hình ảnh của bảng và, nhân các thuộc tính thời gian của công việc từ bảng tính với khối lượng của chúng, sẽ tổng hợp tất cả dữ liệu vào lịch trình. -

- Gửi yêu cầu văn bản tới LLM:

> Tạo biểu đồ Gantt để lập kế hoạch công việc sử dụng các giá trị từ bảng liên quan đến thời gian (Hình 5.31 dưới dạng JPEG trong phụ lục) cho việc lắp đặt 3 khối bê tông. Công việc cho mỗi khối được thực hiện theo trình tự. Bắt đầu công việc đặt từ 01/05/2024.

Phản hồi từ LLM:

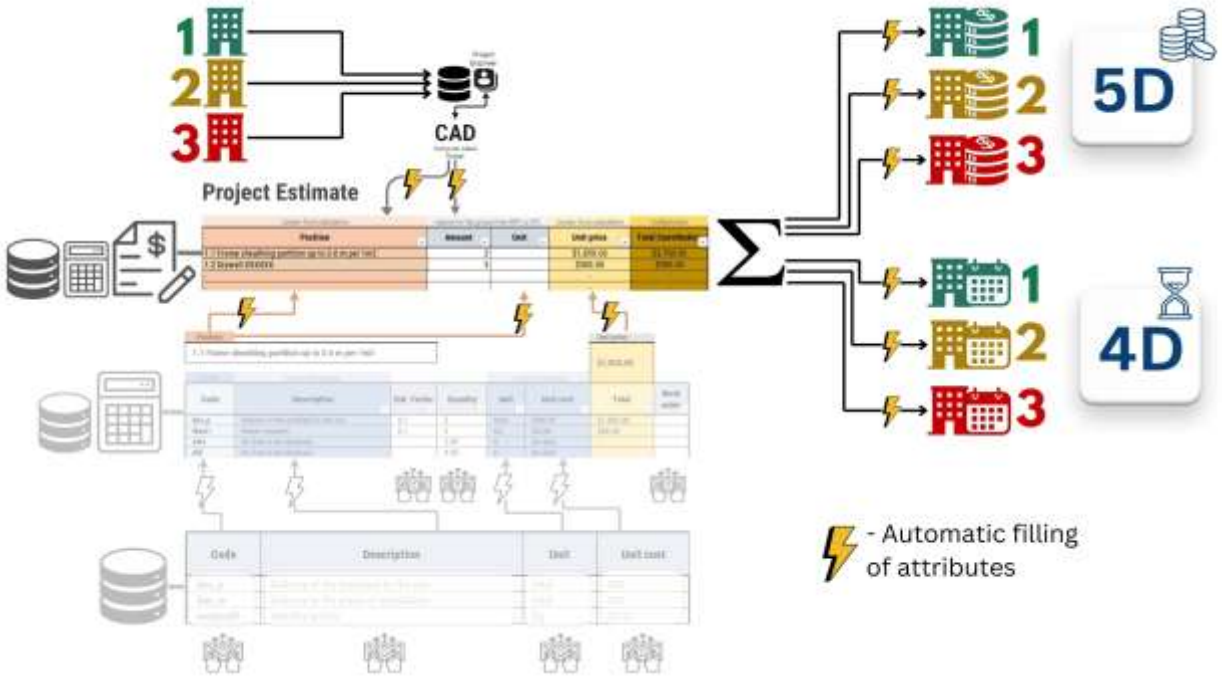


Hình 5.32 Biểu đồ Gantt được tự động tạo bởi một vài LLM cho thấy các giai đoạn xây dựng ba khối bê tông, theo như điều kiện trong prompt.

Biểu đồ nhận được (Hình 5.32) đại diện cho một biểu đồ thời gian, trong đó mỗi dải ngang tương ứng với một giai đoạn nhất định trong việc thực hiện công việc cho khối bê tông và thể hiện thứ tự các hoạt động như chuẩn bị, đào đất, lắp đặt cốp pha, gia cường, đổ bê tông và hoàn thiện, tức là những quy trình mà trong tính toán có các tham số thời gian và trình tự đã được điền đầy đủ.

Biểu đồ tương tự (Hình 5.32) không tính đến các hạn chế liên quan đến ngày làm việc, ca làm việc hay tiêu chuẩn giờ làm, và chỉ nhằm mục đích trực quan hóa quy trình một cách khái niệm. Biểu đồ chính xác hơn, phản ánh tính song song của công việc, có thể được bổ sung bằng các prompt hoặc hướng dẫn bổ sung trong chat.

Sử dụng một bảng tính chi tiết (Hình 5.31), nhờ các thuộc tính thể tích từ hình học 3D, có thể tự động đánh giá cả chi phí dự án thông qua các dự toán tự động và đồng thời tính toán các đặc điểm thời gian của các nhóm dưới dạng bảng hoặc biểu đồ cho các tùy chọn khác nhau của dự án (Hình 5.33).-



Hình 5.33 Tính toán tự động cho phép dự đoán nhanh chóng và tự động chi phí và thời gian cho các tùy chọn dự án khác nhau.

Các hệ thống ERP mô-đun hiện đại (Hình 5.44), khi tải dữ liệu từ các mô hình CAD, sử dụng các phương pháp tính toán tự động tương tự, điều này làm giảm đáng kể thời gian ra quyết định. Điều này cho phép lập kế hoạch biểu đồ làm việc một cách chính xác và tức thì, với giá cả thực tế, để tính toán tổng thời gian cần thiết cho việc thực hiện tất cả các nhiệm vụ trong quá trình thực hiện dự án.-

Các lớp thuộc tính mở rộng 6D-8D: từ hiệu quả năng lượng đến đảm bảo an toàn

6D, 7D và 8D là các cấp độ mở rộng của mô hình thông tin, mỗi cấp độ cung cấp thêm các lớp thuộc tính vào mô hình thông tin tổng thể của dự án, mà trong đó các thuộc tính của mô hình 3D là cơ sở với số lượng và thể tích. Mỗi cấp độ bổ sung mang lại các tham số cụ thể, cần thiết cho việc phân nhóm hoặc xác định sau này trong các hệ thống khác, chẳng hạn như hệ thống quản lý bất động sản (PMS), hệ thống quản lý cơ sở vật chất tự động (CAFM), quản lý dự án xây dựng (CPM) và hệ thống quản lý an toàn (SMS).



Hình 5.34 Các thuộc tính 6D, 7D và 8D trong mô hình dữ liệu thông tin mở rộng việc xem xét các khía cạnh khác nhau của dự án, từ hiệu quả năng lượng đến an toàn.

- Trong 6D, ngoài cơ sở dữ liệu dự án (hoặc khung dữ liệu (Hình 4.113)) với các thuộc tính hình học và thể tích của các phần tử, thông tin (các thuộc tính-cột) về tính bền vững sinh thái được thêm vào. Điều này bao gồm thông tin liên quan đến hiệu quả năng lượng, dấu chân carbon, khả năng tái chế nguyên liệu và việc sử dụng công nghệ thân thiện với môi trường. Những dữ liệu này cho phép đánh giá tác động của dự án đến môi trường, tối ưu hóa các quyết định thiết kế và đạt được các mục tiêu phát triển bền vững (ESG).-
- Các thuộc tính 7D bổ sung các thuộc tính cần thiết để quản lý vận hành tòa nhà. Đây là dữ liệu về lịch bảo trì, tuổi thọ của các thành phần, tài liệu kỹ thuật và lịch sử sửa chữa. Tập hợp thông tin này cung cấp khả năng tích hợp mô hình với các hệ thống vận hành (CAFM, AMS), cho phép lên kế hoạch bảo trì, thay thế thiết bị một cách hiệu quả và hỗ trợ trong toàn bộ vòng đời của dự án.
- 8D là lớp thuộc tính bổ sung, bao gồm thông tin liên quan đến an toàn - cả trong giai đoạn xây dựng và trong quá trình vận hành sau này. Mô hình được bổ sung các biện pháp bảo đảm an toàn cho nhân viên, hướng dẫn xử lý tình huống khẩn cấp, yêu cầu đối với hệ thống thoát hiểm và phòng cháy chữa cháy. Sự tích hợp những dữ liệu này vào mô hình số giúp dự kiến các rủi ro và phát triển các giải pháp kiến trúc, kỹ thuật và tổ chức dựa trên yêu cầu về an toàn lao động và an ninh.

Dưới dạng bảng có cấu trúc, các lớp từ 4D đến 8D đại diện cho các thuộc tính bổ sung dưới dạng các cột với các giá trị được điền (Hình 5.35), được thêm vào các thuộc tính đã điền của mô hình 3D như tên, loại, hạng mục và đặc điểm thể tích. Giá trị trong các lớp thuộc tính 6D, 7D và 8D chứa thêm dữ liệu văn bản và số liệu, như tỷ lệ tái chế, dấu carbon, thời gian bảo hành, chu kỳ thay thế, ngày lắp đặt, biên bản an toàn, v.v.-



ID	Type Name	Width	Length	Recyclability	Carbon Footprint	Warranty Period	Replacement Cycle	Maintenance Schedule	Installation Date	Wellbeing Factors	Safety Protocols
W-NEW	Window	120 cm	-	90%	1622 kgCO2e	8 years	20 years	Annual	-mon	XYZ Windows	ISO 45001
W-OLD1	Window	100 cm	140 cm	90%	1522 kgCO2e	8 years	15 years	Biannual	08/22/2024	XYZ Windows	OSHA Standard
W-OLD2	Window	110 cm	160 cm	90%	1522 kgCO2e	-	15 years	Biannual	08/24/2024	????	OSHA Standard
D-122	Door	90 cm	210 cm	100%	1322 kgCO2e	15 years	25 years	Biennial	08/25/2024	Doors Ltd.	OSHA Standard

Hình 5.35 6D-8D bổ sung các lớp thuộc tính vào mô hình dữ liệu thông tin, vốn đã chứa các thuộc tính hình học và thể tích từ mô hình 3D.

Đối với cửa sổ mới của chúng tôi (Hình 4.41), phần tử với mã định danh W-NEW (Hình 5.35) có thể có các thuộc tính từ 3D đến 8D như sau: --

Các thuộc tính 3D - thông tin hình học, thu được từ các hệ thống CAD:

- Tên loại - phần tử "Cửa sổ"
- "Chiều rộng" - 120 cm
- Thêm vào đó, có thể bổ sung điểm "Bounding Box" của phần tử hoặc "hình học BREP / MESH" của nó như một thuộc tính riêng.

Các thuộc tính 6D - bền vững môi trường:

- Thông số "tái chế" - 90%
- "Dấu carbon" - 1622 kg CO_2

Các thuộc tính 7D - thông tin quản lý cơ sở:

- "Thời gian bảo hành" - 8 năm
- "Chu kỳ thay thế" - 20 năm
- "Bảo trì" - yêu cầu hàng năm

Các thuộc tính 8D - đảm bảo an toàn sử dụng và vận hành tòa nhà:

- Cửa sổ "Được lắp đặt" - bởi công ty "XYZ Windows"
- "Tiêu chuẩn an toàn" - tuân thủ ISO 45001

Tất cả các tham số đã ghi vào cơ sở dữ liệu hoặc tập dữ liệu (Hình 5.35) cần thiết cho các chuyên gia trong các phòng ban khác nhau để nhóm, tìm kiếm hoặc tính toán. Mô tả đa chiều như vậy về các đối tượng của dự án dựa trên các thuộc tính cho phép có cái nhìn toàn diện về vòng đời của chúng, yêu cầu vận hành và nhiều khía cạnh khác cần thiết trong quá trình thiết kế, xây dựng và vận hành dự án.

Đánh giá CO_2 và tính toán khí thải carbon dioxide trong các dự án xây dựng

Cùng với chủ đề bền vững của các dự án xây dựng ở giai đoạn 6D (Hình 5.35), trong xây dựng hiện đại, đặc biệt chú trọng đến tính bền vững về môi trường của các dự án, trong đó một trong những khía cạnh then chốt là đánh giá và giảm thiểu phát thải khí CO_2, diễn ra trong các giai đoạn vòng đời của dự án (ví dụ như trong sản xuất và lắp đặt).-

> Đánh giá và tính toán phát thải carbon của các vật liệu xây dựng là quá trình trong đó tổng phát thải carbon được xác định bằng cách nhân các thuộc tính khối lượng của các thành phần hoặc nhóm thành phần được sử dụng trong dự án với hệ số phát thải carbon phù hợp cho từng danh mục.

Việc tính toán phát thải carbon trong đánh giá các dự án xây dựng, như một phần của các tiêu chí ESG (môi trường, xã hội và quản trị) rộng hơn, thêm một cấp độ mới vào phân tích tổng thể. Điều này đặc biệt quan trọng đối với nhà đầu tư khi nhận chứng nhận tương ứng như LEED® (Lãnh đạo trong Thiết kế Năng lượng và Môi trường), BREEAM® (Phương pháp Đánh giá Môi trường Xây dựng của Tổ chức Nghiên cứu Xây dựng) hoặc DGNB® (Hiệp hội Xây dựng Bền vững Đức). Việc nhận được một trong những chứng nhận này có thể nâng cao đáng kể sức hấp dẫn của thị trường đối với tài sản, đơn giản hóa việc đưa vào vận hành và đảm bảo đáp ứng các yêu cầu của người thuê, từ đó hướng đến sự bền vững (ESG). Tùy thuộc vào yêu cầu của dự án, các tiêu chuẩn HQE (Haute Qualité Environnementale, tiêu chuẩn xây dựng môi trường Pháp), WELL (Tiêu chuẩn Xây dựng WELL, tập trung vào sức khỏe và sự thoải mái của người sử dụng) và GRESB (Chỉ tiêu Bền vững Bất động sản Toàn cầu) cũng có thể được sử dụng.

> ESG (môi trường, xã hội và quản trị) là một tập hợp rộng rãi các nguyên tắc có thể được sử dụng để đánh giá quản trị doanh nghiệp, tác động xã hội và môi trường của kinh doanh cả bên trong và bên ngoài công ty.

ESG, ban đầu được phát triển vào đầu những năm 2000 bởi các quỹ đầu tư để cung cấp thông tin cho các nhà đầu tư về các tiêu chí lớn liên quan đến môi trường, xã hội và quản trị, đã trở thành chỉ số quan trọng để đánh giá các công ty cũng như các dự án, bao gồm cả trong ngành xây dựng. Theo nghiên cứu của các công ty tư vấn hàng đầu, việc tính toán các yếu tố môi trường, xã hội và quản trị (ESG) đang trở thành một phần không thể thiếu của ngành xây dựng.

> Theo dữ liệu từ EY (2023) "Con đường hướng tới trung lập carbon", các công ty tích cực áp dụng các nguyên tắc ESG không chỉ giảm thiểu rủi ro dài hạn mà còn nâng cao hiệu quả của mô hình kinh doanh, điều này đặc biệt quan trọng trong bối cảnh chuyển đổi toàn cầu của các thị trường. Trong báo cáo của PwC "Nhận thức về ESG" chỉ ra rằng mức độ nhận thức của các công ty về tầm quan trọng của các yếu tố ESG dao động từ 67% đến 97%, với phần lớn các tổ chức coi những xu hướng này là chìa khóa cho sự phát triển bền vững trong tương lai và doanh nghiệp phần lớn gặp áp lực đáng kể từ các bên liên quan để tích hợp các nguyên tắc ESG.

> Do đó, việc tích hợp các nguyên tắc ESG vào các dự án xây dựng không chỉ góp phần vào việc đạt được các chứng nhận bền vững quốc tế như LEED, BREEAM, DGNB, mà còn đảm bảo sự bền vững lâu dài và khả năng cạnh tranh của các công ty trong ngành.

Một trong những yếu tố quan trọng nhất ảnh hưởng đến tổng lượng carbon phát thải của dự án xây dựng là

các giai đoạn sản xuất và logistics của vật liệu và cấu kiện xây dựng. Các vật liệu được sử dụng tại công trường thường có ảnh hưởng quyết định đến tổng lượng phát thải CO_2, đặc biệt là ở những giai đoạn đầu của vòng đời dự án - từ khai thác nguyên liệu đến việc giao hàng tại công trường xây dựng.

Để tính toán lượng phát thải theo các loại hoặc kiểu cấu kiện xây dựng, cần phải sử dụng các hệ số tham khảo về phát thải carbon, phản ánh lượng CO_2 phát thải trong quá trình sản xuất các vật liệu khác nhau. Các vật liệu này bao gồm bê tông, gạch, thép tái chế, nhôm và nhiều loại khác. Những giá trị này thường được trích xuất từ các nguồn đáng tin cậy và cơ sở dữ liệu quốc tế như UK ICE 2015 (Inventory of Carbon and Energy) và US EPA 2006 (Cơ quan Bảo vệ Môi trường Hoa Kỳ) [105]. Trong bảng tiếp theo (Hình 5.36) đã cung cấp các hệ số phát thải cơ bản cho một số vật liệu xây dựng phổ biến. Đối với mỗi loại, hai tham số chính đã được chỉ ra: phát thải CO_2 riêng lẻ (tính bằng kilôgam trên kilôgam vật liệu) và hệ số chuyển đổi thể tích sang khối lượng (tính bằng kilôgam trên mét khối), cần thiết để tích hợp tính toán vào mô hình thiết kế và liên kết với nhóm dữ liệu QTO.



Carbon Emitted in Production

Material	Abbreviated	UK ICE Database (2015) USEPA (2006) Process Emissions (kg CO2e/ kg of product) (K1)	UK ICE Database (2015) USEPA (2006) Process Emissions (kg CO2e/ kg of product) (K2)	Coefficient m3 to kg Kg / m3 (K3)
Concrete	Concrete	0.12	0.12	2400
Concrete block	Concrete_block	0.13**	0.14	2000
Brick	Brick	0.24	0.32	2000
Medium density fiberboard (MDF)	MDF	0.39*	0.32	700
Recycled steel (avg recy content)	Recycled_steel	0.47	0.81	7850
Glass (not including primary mfg.)	Glass	0.59	0.6	2500
Cement (Portland, masonry)	Cement	0.95	0.97	1440
Aluminum (virgin)	Aluminum	12.79	16.6	2700

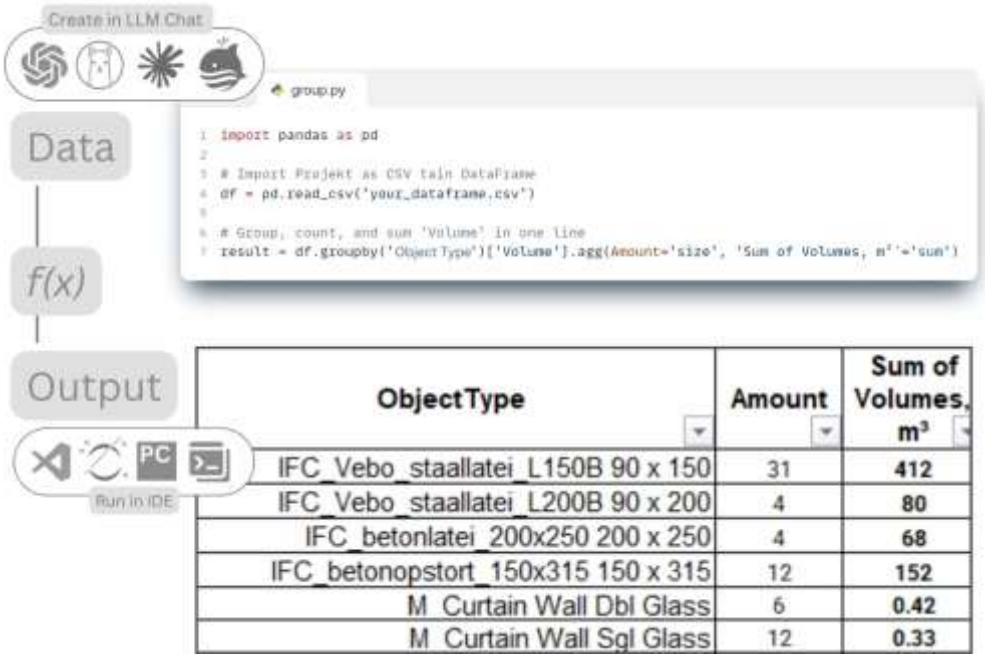
Hình 5.36 Số lượng carbon phát ra trong quá trình sản xuất các vật liệu xây dựng khác nhau, theo cơ sở dữ liệu UK ICE và US EPA.

Để tính tổng lượng phát thải CO_2 của dự án, giống như trong các tính toán 4D và 5D, cần xác định khối lượng thuộc tính của mỗi nhóm đối tượng. Việc này có thể thực hiện bằng cách sử dụng các công cụ phân tích định lượng (QTO), nhận được khối lượng các thuộc tính tính bằng mét khối, như đã được trình bày chi tiết trong phần liên quan đến Quantity take-off. Sau đó, các khối lượng thu được sẽ được nhân với các hệ số tương ứng cho thuộc tính "phát thải công nghệ CO_2" của mỗi nhóm vật liệu.

- Hãy tự động trích xuất bảng khối lượng theo loại cấu kiện từ dự án CAD (BIM), nhóm tất cả dữ liệu của dự án như đã làm trong các chương trước đó. Để thực hiện nhiệm vụ này, chúng ta sẽ sử dụng LLM.

> Vui lòng nhóm bảng DataFrame từ dự án CAD (BIM) theo tham số của cột "Tên Đối Tượng" (hoặc "Loại") và cho biết số lượng đối tượng trong mỗi nhóm, cũng như tính tổng tham số "Khối Lượng" cho tất cả các đối tượng cùng loại. ⏎

Phản hồi từ LLM:



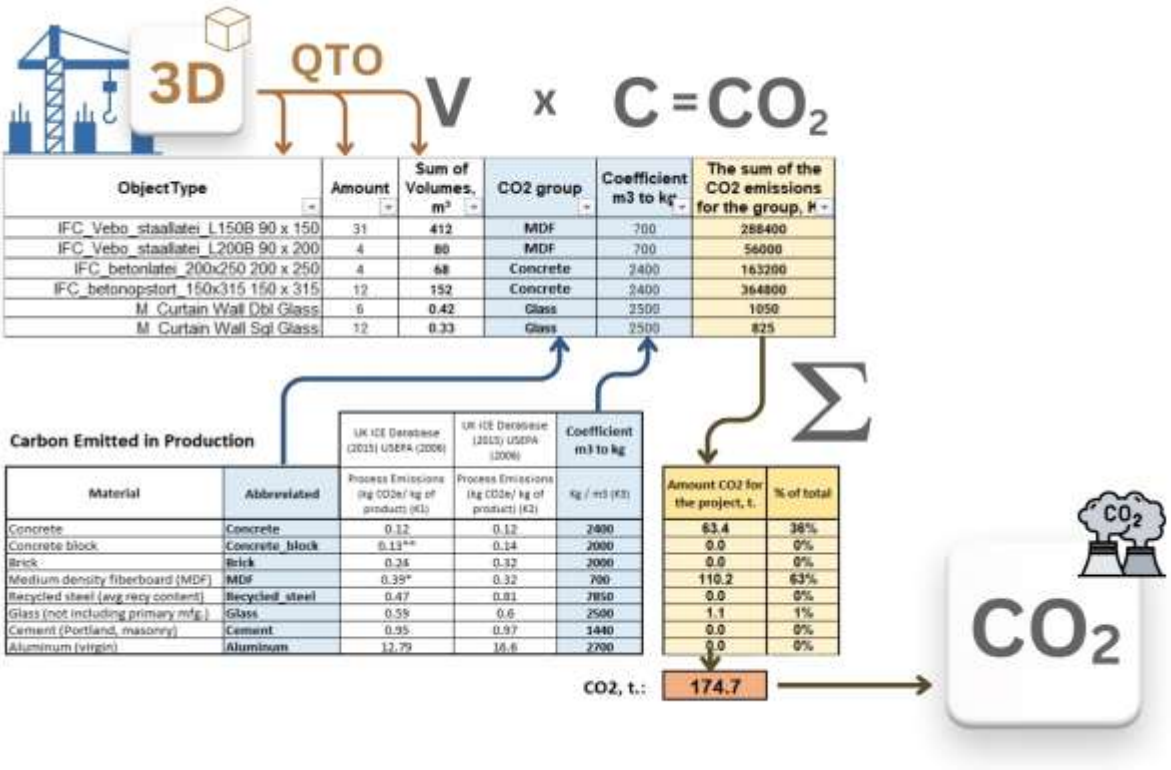
ObjectType	Amount	Sum of Volumes, m^3
IFC_Vebo_staallatei_L150B 90 x 150	31	412
IFC_Vebo_staallatei_L200B 90 x 200	4	80
IFC_betonlatei_200x250 200 x 250	4	68
IFC_betonopstort_150x315 150 x 315	12	152
M_Curtain Wall Dbl Glass	6	0.42
M_Curtain Wall Sgl Glass	12	0.33

Hình 5.37 Mã được sinh ra trong LLM đã nhóm cho chúng ta các thực thể của dự án theo loại (ObjectType) với thuộc tính "Khối Lượng" đã tính tổng.

Để tự động hóa việc tính toán tổng lượng phát thải CO_2 cho toàn bộ dự án, chỉ cần thiết lập việc ánh xạ dữ liệu tự động trong bảng hoặc liên kết thủ công các loại đối tượng (Hình 5.37) với các loại vật liệu tương ứng (Hình 5.36) từ bảng hệ số phát thải. Bảng hoàn chỉnh với các hệ số phát thải và công thức, cũng như mã để lấy khối lượng từ các định dạng CAD (BIM) và tự động xác định CO_2 có thể tìm thấy trên GitHub với từ khóa "CO_2_calculating-the-embodied-carbon. DataDrivenConstruction" [106].--

Như vậy, việc tích hợp dữ liệu sau khi nhóm QTO các phần tử từ cơ sở dữ liệu CAD cho phép tính toán tự động phát thải carbon dioxide (Hình 5.38) cho các phương án thiết kế khác nhau. Điều này cho phép phân tích tác động của các vật liệu khác nhau trong các phương án khác nhau và chỉ chọn những giải pháp đáp ứng yêu cầu của khách hàng về mức phát thải CO_2 để đạt được chứng chỉ khi bàn giao tòa nhà.-

> Đánh giá lượng khí thải CO_2 bằng cách nhân các hệ số với khối lượng của các yếu tố nhóm trong dự án - một ví dụ điển hình về nhiệm vụ trong quá trình nhận được xếp hạng ESG cho công ty xây dựng (ví dụ như chứng nhận LEED) cho công trình.



Hình 5.38 Tích hợp các nhóm QTO từ cơ sở dữ liệu CAD đảm bảo độ chính xác và tự động hóa trong việc thu thập ước lượng khối lượng khí thải CO_2 cuối cùng.

> Tương tự, bằng cách xác định khối lượng của các nhóm yếu tố, chúng ta có thể thực hiện tính toán cho việc kiểm soát và logistics vật liệu, giám sát và quản lý chất lượng, mô hình hóa và phân tích tiêu thụ năng lượng, cũng như nhiều nhiệm vụ khác để đạt được trạng thái thuộc tính mới (tham số trong bảng) cho cả các nhóm yếu tố riêng lẻ và toàn bộ dự án.

Nếu số lượng các quy trình tính toán như vậy trong công ty bắt đầu tăng lên, câu hỏi về nhu cầu tự động hóa các tính toán như vậy và tích hợp các kết quả tính toán vào quy trình công ty và hệ thống quản lý dữ liệu sẽ được đặt ra.

Do độ phức tạp của giải pháp tổng thể, các công ty vừa và lớn hoạt động trong ngành xây dựng chuyển giao việc tự động hóa này cho các công ty chuyên phát triển hệ thống ERP (hoặc PMIS). Các công ty phát triển tạo ra một hệ thống mô-đun tổng thể duy nhất cho các khách hàng lớn, để quản lý nhiều lớp thông tin khác nhau, bao gồm tính toán vật liệu và tài nguyên.

CHƯƠNG 5.4. **HỆ THỐNG ERP VÀ PMIS TRONG XÂY DỰNG**

Hệ thống ERP xây dựng dựa trên các tính toán và dự toán

Các hệ thống ERP mô-đun kết hợp các lớp thuộc tính (thông tin) và lưu lượng dữ liệu khác nhau thành một hệ thống tổng thể duy nhất, cho phép các nhà quản lý dự án quản lý đồng bộ tài nguyên, tài chính, logistics và các khía cạnh khác của dự án trên một nền tảng duy nhất. Hệ thống ERP xây dựng đóng vai trò là "bộ não" của các dự án xây dựng, đơn giản hóa các quy trình lặp đi lặp lại thông qua tự động hóa, cung cấp tính minh bạch và kiểm soát trong toàn bộ quá trình xây dựng.

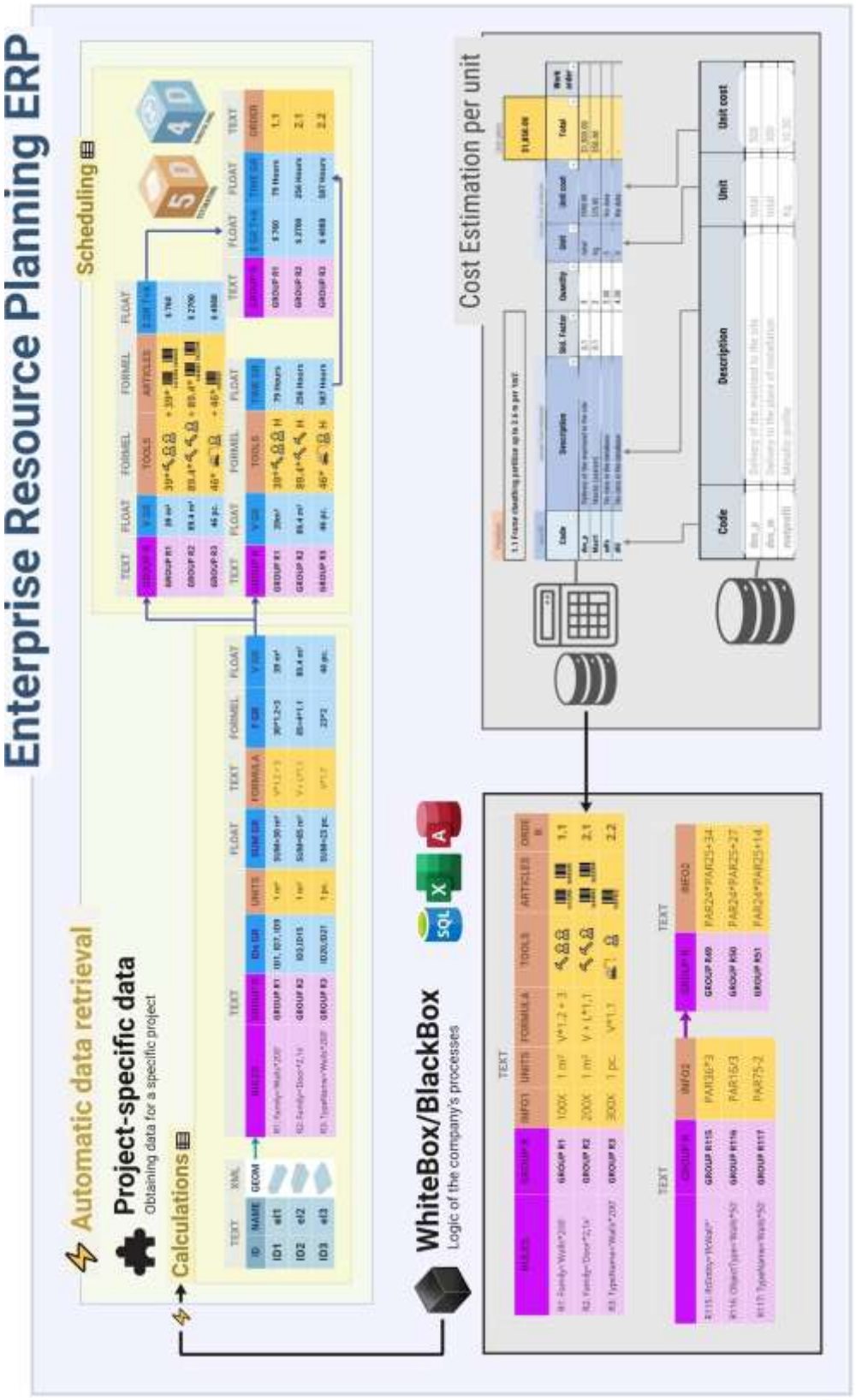
> Các hệ thống ERP xây dựng (Enterprise Resource Planning) là các giải pháp phần mềm tổng hợp, được thiết kế để quản lý và tối ưu hóa các khía cạnh khác nhau của quy trình xây dựng. Cốt lõi của các hệ thống ERP xây dựng là các mô-đun quản lý tính toán chi phí và lập kế hoạch công việc, điều này khiến chúng trở thành công cụ quan trọng cho việc lập kế hoạch tài nguyên hiệu quả.

Các mô-đun của hệ thống ERP cho phép người dùng nhập, xử lý và phân tích dữ liệu, bao quát có cấu trúc các khía cạnh khác nhau của dự án, có thể bao gồm kế toán chi phí vật liệu và lao động, sử dụng thiết bị, quản lý logistics, nguồn nhân lực, các liên hệ và các loại hoạt động xây dựng khác.

Một trong những khối chức năng của hệ thống là mô-đun tự động hóa logic kinh doanh - BlackBox/WhiteBox, đóng vai trò là trung tâm quản lý quy trình.

BlackBox/WhiteBox cho phép các chuyên gia sử dụng hệ thống ERP, thông qua quyền truy cập, linh hoạt quản lý các khía cạnh khác nhau của doanh nghiệp, mà đã được cấu hình trước bởi những người dùng hoặc quản trị viên khác. Trong bối cảnh hệ thống ERP, các thuật ngữ BlackBox và WhiteBox biểu thị các mức độ minh bạch và khả năng kiểm soát logic nội bộ của hệ thống:

- BlackBox ("hộp đen") - người dùng tương tác với hệ thống thông qua giao diện, không có quyền truy cập vào logic nội bộ của quá trình thực hiện. Hệ thống tự động thực hiện các phép tính, dựa trên các quy tắc được thiết lập trước, ẩn sau người dùng cuối. Người dùng nhập dữ liệu và nhận kết quả, không biết những thuộc tính hoặc hệ số nào đã được sử dụng bên trong.
- WhiteBox (hộp trắng) - logic của các quy trình có thể xem, cấu hình và sửa đổi. Người dùng nâng cao, quản trị viên hoặc nhà tích hợp có thể tự tay đặt các thuật toán xử lý dữ liệu, quy tắc tính toán và kịch bản tương tác giữa các thực thể trong dự án.



Hình. 5.41 Kiến trúc hệ thống ERP xây dựng, để nhận được bảng dự toán và tiến độ công việc khi điền thủ công các thuộc tính khối lượng.

Một ví dụ có thể là tình huống mà người dùng có kinh nghiệm hoặc quản trị viên thiết lập quy tắc: các thuộc tính trong bảng dự toán phải được nhân với nhau hoặc nhóm lại theo một tiêu chí nhất định, cũng như nơi lưu trữ kết quả cuối cùng. Sau đó, các chuyên gia ít kinh nghiệm hơn, chẳng hạn như kỹ sư dự toán, chỉ cần tải dữ liệu mới lên ERP qua giao diện người dùng - và nhận được bảng dự toán, tiến độ hoặc thông số kỹ thuật mà không cần viết mã hoặc tìm hiểu các chi tiết kỹ thuật của logic.

> Trong các chương trước, các mô-đun tính toán và logic đã được xem xét trong bối cảnh tương tác với LLM. Trong môi trường hệ thống ERP, các phép tính và chuyển đổi tương tự diễn ra bên trong các mô-đun, được ẩn giấu sau giao diện của các nút và biểu mẫu.

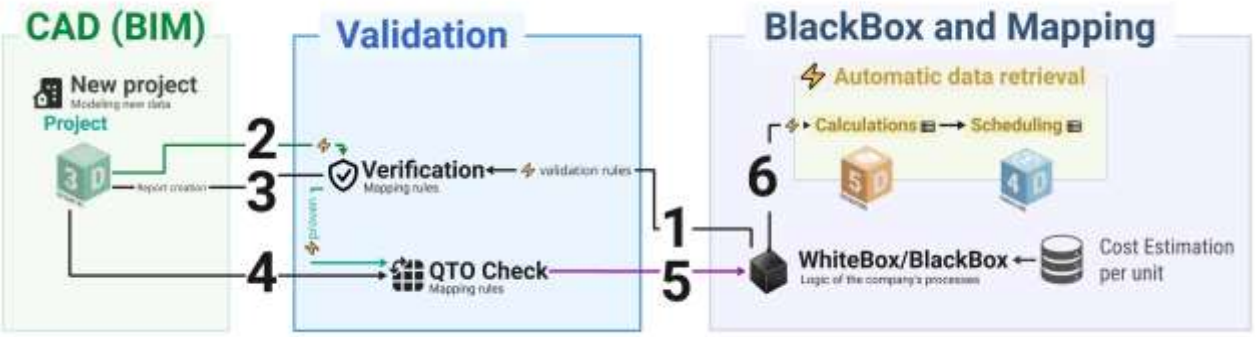
Trong ví dụ tiếp theo (Hình. 5.41), quản trị viên hệ thống ERP trong mô-đun BlackBox/WhiteBox đã thiết lập các quy tắc đối chiếu thuộc tính của các thực thể trong bảng dự toán với các thuộc tính để nhóm QTO. Nhờ vào mô-đun BlackBox/WhiteBox được cấu hình này (bởi quản lý hoặc quản trị viên), người dùng (kỹ sư dự toán hoặc kỹ sư) khi thêm thủ công thuộc tính số lượng hoặc khối lượng qua giao diện người dùng ERP, tự động nhận được bảng dự toán và tiến độ công việc sẵn sàng. Như vậy, các quy trình tính toán và hình thành bảng dự toán, đã được xem xét trong các chương trước bằng cách sử dụng mã, bên trong ERP, chuyển thành một dây chuyền bán tự động.

Kết nối quy trình bán tự động này với các thuộc tính khối lượng từ mô hình CAD (BIM) (Hình. 4.113), thông qua, chẳng hạn như tải dự án CAD, vào mô-đun ERP đã được thiết lập cho mục đích này, biến dòng dữ liệu thành một cơ chế đồng bộ, có khả năng tự động và ngay lập tức cập nhật chi phí của từng nhóm yếu tố hoặc toàn bộ dự án khi có bất kỳ thay đổi nào trong đó trong giai đoạn thiết kế, khi tải mô hình CAD vào ERP. -

Để tạo ra dòng dữ liệu tự động (Hình. 5.42) giữa các hệ thống CAD (BIM) và ERP, cần xác định có cấu trúc các quy trình chính và yêu cầu dữ liệu từ cơ sở dữ liệu của các mô hình CAD (BIM), như chúng tôi đã đề cập trong chương trước "Yêu cầu và đảm bảo chất lượng dữ liệu". Quy trình này trong ERP chia thành các bước tương tự:-

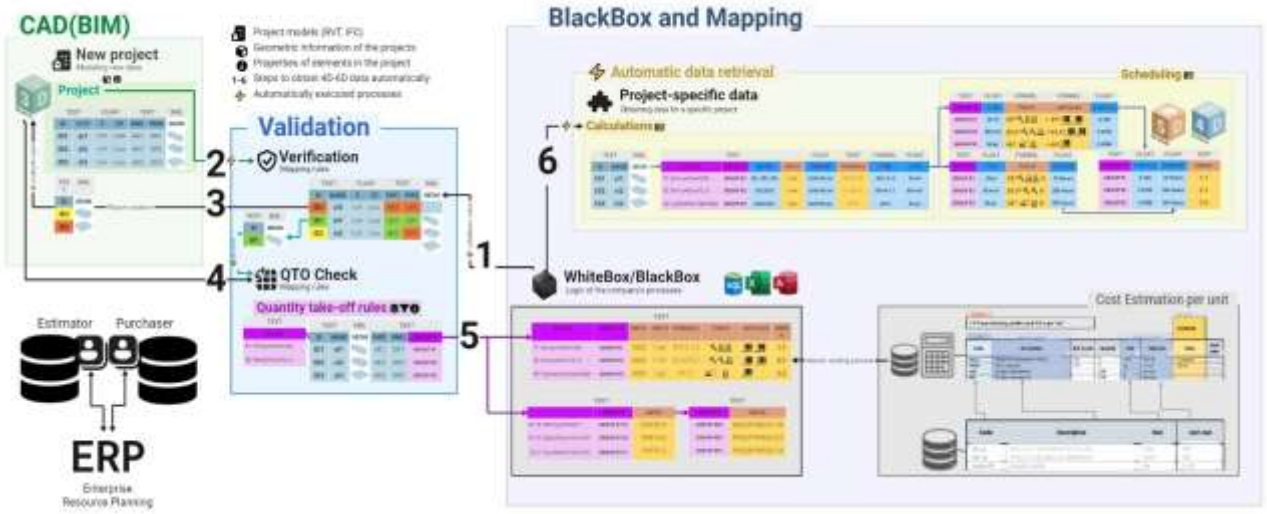
- Tạo quy tắc kiểm tra (1), điều này đóng vai trò quan trọng trong việc đảm bảo độ chính xác của dữ liệu đến hệ thống ERP. Các quy tắc kiểm tra hoạt động như những bộ lọc, kiểm tra các thực thể và thuộc tính của chúng, chỉ cho phép những yếu tố đáp ứng yêu cầu vào hệ thống. Chi tiết hơn về kiểm tra và xác thực trong chương "Tạo yêu cầu và kiểm tra chất lượng dữ liệu".
- Sau đó, bên trong ERP diễn ra quy trình xác minh (2), xác nhận rằng tất cả các yếu tố-thực thể của dự án cùng với các thuộc tính và giá trị của chúng đã được tạo ra đúng cách và sẵn sàng cho các giai đoạn xử lý tiếp theo.
- Nếu xảy ra vấn đề với các dữ liệu thuộc tính không đầy đủ, một báo cáo (3) sẽ được tạo ra và dự án cùng với hướng dẫn khắc phục sẽ được gửi đi chỉnh sửa cho đến khi sẵn sàng cho vòng lặp tiếp theo.
- Sau khi dữ liệu dự án được xác nhận và kiểm tra, chúng được sử dụng trong một mô-đun ERP (4) để tạo ra các bảng Quantity Take-Off (QTO), theo các quy tắc đã được hình thành trước đó (WhiteBox/BlackBox) để tạo ra các thuộc tính số lượng cho các nhóm thực thể, vật liệu và tài nguyên.
- Dữ liệu đã được nhóm theo các quy tắc đối chiếu hoặc QTO được tích hợp tự động với các phép tính (ví dụ: chi phí và thời gian) (5).
- Ở giai đoạn cuối, hệ thống ERP, người dùng, thông qua việc nhân các thuộc tính khối lượng từ bảng QTO với các thuộc tính từ các bảng quy trình (ví dụ: các mục dự toán), tự động tạo ra kết quả

tính toán (6) (ví dụ: dự toán chi phí, lịch trình công việc hoặc phát thải CO_2) cho từng nhóm thực thể và cho toàn bộ dự án.

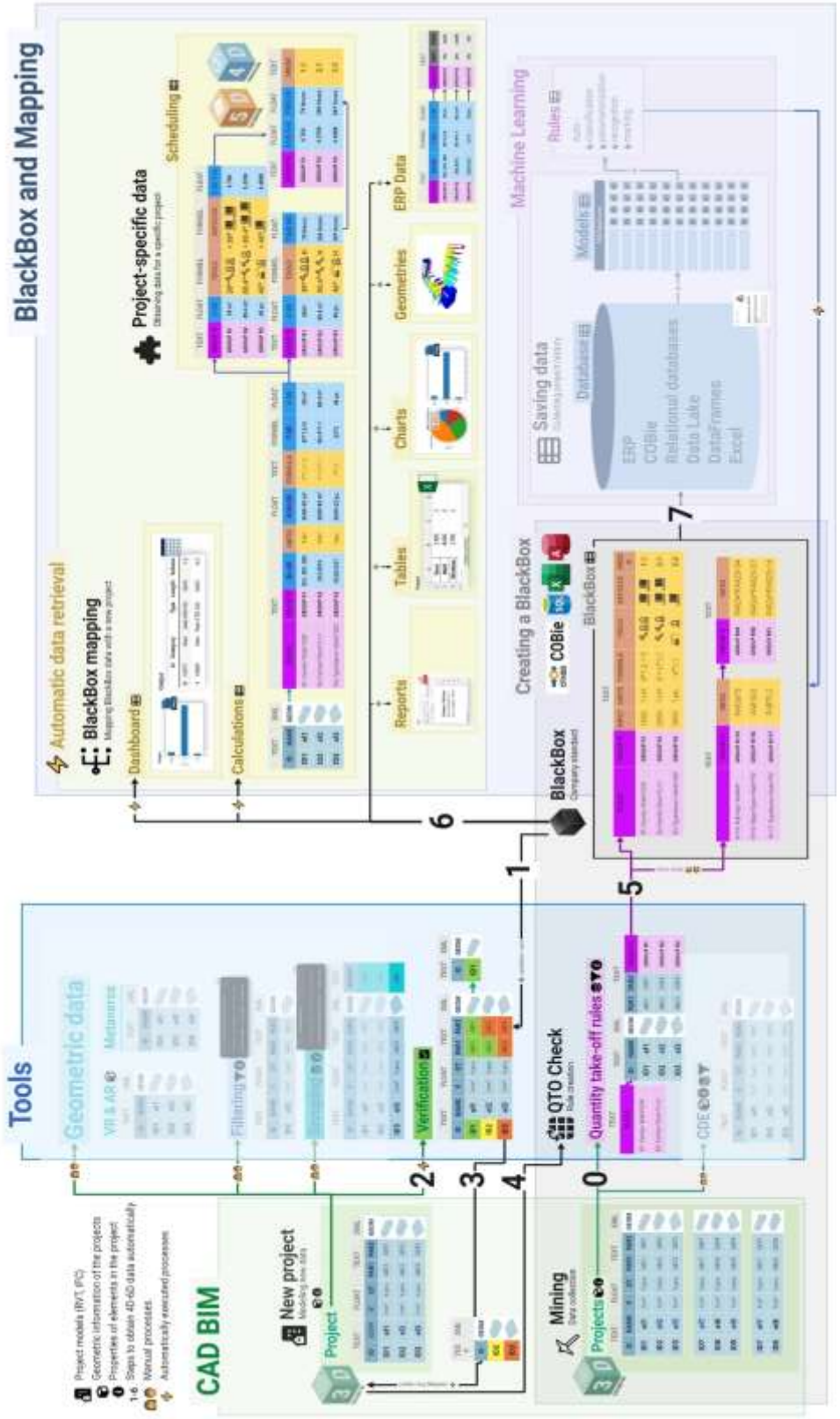


Hình. 5.42 Kiến trúc hệ thống ERP xây dựng với CAD (BIM), từ việc tạo ra các quy tắc xác thực (1) đến việc tính toán chi phí và lịch trình công việc tự động (5-6).

Trong một hệ thống ERP mô-đun, các quy trình được tích hợp thông qua phần mềm, bao gồm giao diện người dùng. Phía sau giao diện là phần nội bộ, nơi các bảng có cấu trúc xử lý dữ liệu, thực hiện nhiều thao tác mà quản lý hoặc điều hành đã thiết lập trước. Do đó, người dùng, nhờ vào logic tự động hóa được xác định và thiết lập trước (trong các mô-đun BlackBox/WhiteBox), nhận được các tài liệu chuẩn bị nửa tự động đáp ứng nhiệm vụ của mình.



Hình. 5.43 Hệ thống ERP giúp các nhà quản lý và người dùng di chuyển giữa các bảng của các chuyên gia để tạo ra dữ liệu mới.



Hình. 5.44 Hệ thống ERP được tích hợp với các công cụ phân tích và tự động hóa quy trình ra quyết định trong công ty.

Tương tự, các quy trình trong hệ thống ERP, từ giai đoạn bắt đầu đến tính toán cuối cùng (các bước 1-6 Hình. 5.43) đại diện cho chuỗi các bước liên kết với nhau, cung cấp sự minh bạch, hiệu quả và chính xác trong lập kế hoạch.-

Các hệ thống ERP xây dựng hiện đại bao gồm không chỉ các mô-đun tính toán chi phí và thời gian mà còn hàng chục mô-đun đã được định cấu hình trước, thường bao phủ các chức năng quản lý tài liệu, theo dõi tiến trình thực hiện dự án, quản lý hợp đồng, chuỗi cung ứng và logistics, cũng như tích hợp với các hệ thống và nền tảng kinh doanh khác. Các công cụ phân tích tích hợp của ERP cho phép người dùng tự động hóa việc tạo bảng điều khiển để giám sát các chỉ số chính (KPI - key performance indicators) của dự án. Điều này đảm bảo quản lý trung tâm và nhất quán tất cả các khía cạnh của dự án xây dựng, với nỗ lực gom lại một lượng lớn ứng dụng và hệ thống trên một nền tảng duy nhất.

Trong tương lai, phân tích ERP sẽ được sử dụng kết hợp với máy học để nâng cao độ chính xác và tối ưu hóa quy trình tính toán các thuộc tính tương lai của dự án. Dữ liệu và thuộc tính, được phân tích và thu thập từ các hệ thống ERP trong Big Data (Hình. 5.44), trong tương lai sẽ trở thành nền tảng cho việc tạo ra các mô hình dự đoán, có khả năng chính xác dự đoán các sự cố tiềm tàng, rủi ro hoặc, chẳng hạn, những thay đổi có thể xảy ra về giá vật liệu.-

Như một lựa chọn thay thế cho ERP, trong ngành xây dựng, hệ thống PMIS (Project Management Information System) thường được sử dụng — hệ thống quản lý dự án, được thiết kế để kiểm soát chi tiết việc thực hiện các nhiệm vụ tại cấp độ từng công trình xây dựng riêng biệt.

PMIS: Liên kết giữa ERP và công trường xây dựng

Khác với ERP, phần mềm quản lý thông tin dự án (PMIS) tập trung vào việc quản lý một dự án cụ thể, đảm bảo theo dõi tiến độ, ngân sách, tài nguyên và tài liệu.

> PMIS (Project Management Information System) là phần mềm dùng để quản lý các dự án xây dựng, được thiết kế cho việc lập kế hoạch, theo dõi, phân tích và báo cáo về tất cả các khía cạnh của dự án.

PMIS cho phép quản lý tài liệu, lịch trình, ngân sách và thoạt nhìn, PMIS có thể có vẻ như là một giải pháp trùng lặp với ERP, nhưng sự khác biệt chính nằm ở cấp độ quản lý:

- ERP tập trung vào các quy trình kinh doanh tổng thể của công ty: quản lý chi phí, hợp đồng, mua sắm, nhân sự và tài nguyên ở cấp độ doanh nghiệp.
- PMIS chú trọng vào việc quản lý từng dự án riêng lẻ, cung cấp lập kế hoạch chi tiết, kiểm soát thay đổi, báo cáo và phối hợp các bên liên quan.

Trong nhiều trường hợp, các hệ thống ERP đã sở hữu đủ chức năng, và việc triển khai PMIS thường chỉ là vấn đề tiện ích và sở thích của công ty. Nhiều nhà thầu và khách hàng sử dụng PMIS không phải vì nó cần thiết, mà vì nó được nhà cung cấp hoặc khách hàng lớn áp đặt, người muốn tập hợp dữ liệu trên một nền tảng cụ thể.

Cũng cần đề cập rằng trong thuật ngữ quốc tế có những khái niệm phổ biến riêng biệt khác cho việc quản lý các dự án xây dựng, như PLM (Quản lý vòng đời sản phẩm), cũng như EPC và EPC-M (Quản lý Kỹ thuật, Mua sắm và Xây dựng) - các phương pháp hợp đồng trong ngành xây dựng.

Nếu công ty đã sử dụng ERP với các mô-đun quản lý dự án, việc triển khai PMIS có thể trở thành một mắt xích thừa, trùng lặp chức năng. Tuy nhiên, nếu các quy trình chưa được tự động hóa và dữ liệu bị rời rạc,

PMIS có thể trở thành một công cụ dễ sử dụng và bảo trì hơn.

Spéculation, lợi nhuận, tính kín đáo và thiếu tính minh bạch trong ERP và PMIS

Mặc dù giao diện và quy trình có vẻ đơn giản, các hệ thống ERP và PMIS trong ngành xây dựng trong hầu hết các trường hợp đều là những giải pháp khép kín và không linh hoạt. Những hệ thống này thường được cung cấp dưới dạng một gói phần mềm được cấu hình trước bởi một nhà cung cấp, với quyền truy cập hạn chế vào cơ sở dữ liệu nội bộ và logic quy trình.

Sự phát triển và kiểm soát các hệ thống như vậy ngày càng được đảm nhiệm bởi các nhà cung cấp sản phẩm CAD-(BIM-), vì chính trong cơ sở dữ liệu của họ chứa thông tin cần thiết cho các hệ thống ERP: Các thuộc tính số lượng và khối lượng của các yếu tố dự án. Tuy nhiên, thay vì cung cấp quyền truy cập vào dữ liệu này ở định dạng mở hoặc có thể đọc bằng máy, các nhà cung cấp chỉ đưa ra kịch bản người dùng hạn chế và logic xử lý khép kín - được xác định trước trong các mô-đun BlackBox. Điều này làm giảm tính linh hoạt của hệ thống và cản trở việc điều chỉnh nó theo các điều kiện cụ thể của dự án.

> Tính minh bạch của dữ liệu hạn chế vẫn là một trong những vấn đề chính trong các quy trình số hóa trong xây dựng. Kiến trúc cơ sở dữ liệu đóng kín, sự thiếu hụt quyền truy cập vào các tập hợp thuộc tính đầy đủ của các yếu tố xây dựng, sự chú trọng vào các mô-đun tự động hóa BlackBox và sự thiếu hụt các giao diện mở làm tăng đáng kể rủi ro về tài liệu và thủ tục. Những hạn chế này tạo ra "điểm nghẽn" trong quá trình ra quyết định, khó khăn trong việc xác minh thông tin và mở ra cơ hội cho việc che giấu dữ liệu hoặc đầu cơ bên trong các hệ thống ERP/PMIS. Người dùng thường chỉ nhận được quyền truy cập hạn chế - có thể là giao diện cắt bớt hoặc API một phần - mà không có khả năng tương tác trực tiếp với nguồn dữ liệu gốc. Điều này đặc biệt quan trọng khi nói đến các thông số được tạo tự động từ các dự án CAD, chẳng hạn như thể tích, diện tích và số lượng được sử dụng để tính toán QTO.

Là hệ quả, thay vì tìm kiếm hiệu quả thông qua tự động hóa quy trình, tính mở của dữ liệu, giảm thiểu chi phí giao dịch và tạo ra các mô hình kinh doanh mới, nhiều công ty xây dựng tập trung vào quản lý các thông số bên ngoài - thao tác với các hệ số, điều chỉnh các yếu tố và phương pháp tính toán tác động đến chi phí dự án trong các nền tảng ERP/PMIS đóng kín. Điều này tạo ra môi trường cho sự đầu cơ, làm lệch lạc chi phí sản xuất thực tế và giảm lòng tin giữa tất cả các bên tham gia quy trình xây dựng.

> Trong xây dựng, lợi nhuận được hình thành từ chênh lệch giữa doanh thu từ dự án hoàn thành và chi phí biến đổi, bao gồm thiết kế, vật liệu, nguồn nhân lực và các chi phí trực tiếp khác liên quan đến việc thực hiện dự án. Tuy nhiên, yếu tố chính ảnh hưởng đến mức độ của những chi phí này không chỉ là công nghệ hay logistics, mà còn là tốc độ và độ chính xác của các phép tính, cũng như chất lượng của các quyết định quản lý bên trong công ty.

Vấn đề trở nên trầm trọng hơn khi ở hầu hết các công ty xây dựng, quy trình tính toán chi phí vẫn không minh bạch không chỉ đối với khách hàng mà còn đối với chính nhân viên của họ, những người không thuộc bộ phận dự toán hoặc tài chính. Tính đóng kín này tạo điều kiện cho việc hình thành một nhóm chuyên gia đặc quyền bên trong công ty - những người sở hữu "chuyên môn tài chính", có quyền sửa đổi các thuộc tính và hệ số điều chỉnh trong các hệ thống ERP/PMIS. Những nhân viên này, cùng với các lãnh đạo công ty, có thể kiểm soát thực tế logic tài chính của dự án.

> Trong những điều kiện như vậy, các nhân viên dự toán trở thành "nhà tung hứng tài chính", phải cân bằng giữa việc tối đa hóa lợi nhuận cho công ty và cần thiết phải giữ giá cả cạnh tranh cho khách hàng. Trong khi đó, họ buộc phải tránh những thao tác rõ ràng và thô thiển để không làm tổn hại đến uy tín của công ty. Chính vào thời điểm này, các hệ số được đặt ra, ngụy trang cho các khối lượng hoặc chi phí vật liệu và công việc bị thổi phồng.

Kết quả là sơ đồ chính để nâng cao hiệu quả và khả năng sinh lời của các công ty hoạt động trong ngành xây dựng không phải là tự động hóa và tăng tốc quá trình ra quyết định, mà là đầu cơ vào giá vật liệu và công việc (Hình 5.45). Việc tăng giá trị của công việc và vật liệu được thực hiện bởi "kế toán mờ" trong các hệ thống ERP/PMIS kín bằng cách tăng tỷ lệ trên giá thị trường trung bình của vật liệu hoặc khối lượng công việc thông qua các hệ số (Hình 5.16), đã được thảo luận trong chương "Lập bảng dự toán và tính toán chi phí công việc dựa trên cơ sở tài nguyên".-

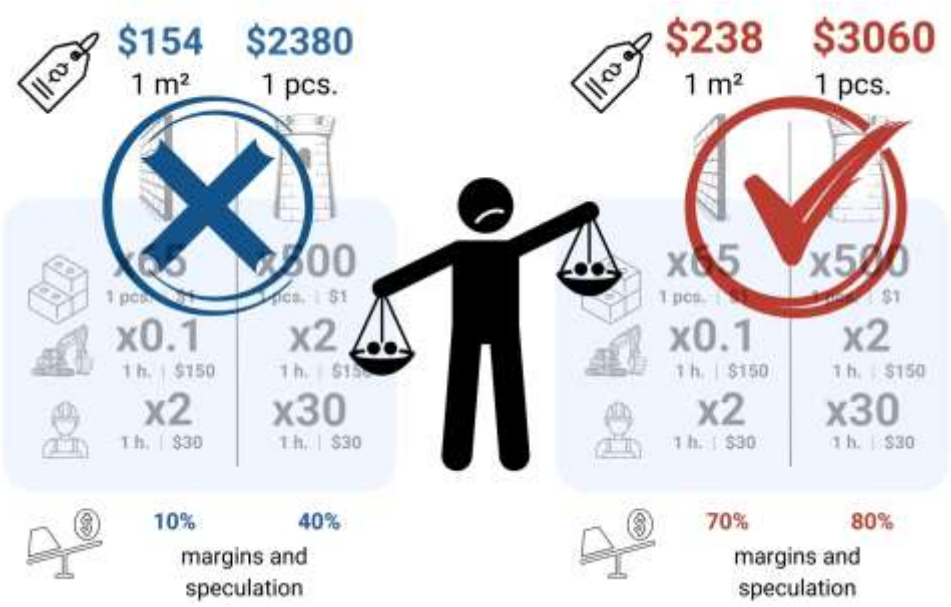
> Kết quả là khách hàng nhận được một bảng tính mà không phản ánh giá trị thực tế hoặc khối lượng công việc, mà là một hệ số phát sinh từ nhiều hệ số nội bộ ẩn. Trong khi đó, các nhà thầu phụ, nhằm phù hợp với các mức giá thấp mà nhà thầu chính đã quy định, thường buộc phải mua vật liệu rẻ hơn và kém chất lượng, làm giảm chất lượng cuối cùng của công trình.

Quy trình đầu cơ nhằm tìm kiếm lợi nhuận trong không khí cuối cùng gây hại cho cả khách hàng, những người nhận dữ liệu không chính xác, lẫn những người thực hiện, những người buộc phải tìm kiếm các mô hình đầu cơ mới.

Kết quả là, dự án càng lớn, thì mức độ quan liêu trong quản lý dữ liệu và quy trình càng cao. Đằng sau mỗi giai đoạn và mỗi mô-đun thường ẩn chứa các hệ số và phụ phí không minh bạch được tích hợp vào các thuật toán tính toán và quy trình nội bộ. Điều này không chỉ gây khó khăn cho việc kiểm toán mà còn làm sai lệch đáng kể bức tranh tài chính của dự án. Trong các dự án xây dựng lớn, những thực tiễn như vậy thường dẫn đến việc tăng chi phí cuối cùng nhiều lần (đôi khi lên tới mười lần), trong khi đó khối lượng và chi phí thực tế nằm ngoài tầm kiểm soát hiệu quả của khách hàng (Hình 2.13 So sánh chi phí dự kiến và thực tế cho các dự án hạ tầng lớn tại Đức).

> Theo báo cáo của McKinsey & Company "Hình dung tương lai số trong ngành xây dựng" (2016), các dự án xây dựng lớn thường hoàn thành muộn hơn kế hoạch 20% và vượt ngân sách lên đến 80%.

Các bộ phận dự toán và ngân sách trở thành yếu tố được bảo vệ nhất trong công ty. Quyền truy cập vào chúng được hạn chế nghiêm ngặt ngay cả đối với các chuyên gia nội bộ, và do sự kín đáo của logic và cấu trúc cơ sở dữ liệu, không thể đánh giá khách quan hiệu quả của các quyết định dự án mà không bị sai lệch. Sự thiếu minh bạch dẫn đến việc các công ty buộc phải không tối ưu hóa quy trình, mà phải chiến đấu để tồn tại bằng cách "quản lý sáng tạo" các con số và hệ số (Hình 5.31, Hình 5.16 - ví dụ như tham số "Bid. Factor").-



Hình 5.45 Các hệ số đầu cơ ở mức độ tính toán - đây là lợi nhuận chính của các công ty và nghệ thuật tung hứng giữa chất lượng công việc và danh tiếng.

Tất cả những điều này đặt ra nghi vấn về tính hợp lý của việc tiếp tục sử dụng các hệ thống ERP/PMIS kín trong ngành xây dựng. Trong bối cảnh chuyển đổi số và yêu cầu ngày càng cao về tính minh bạch từ phía khách hàng (Hình 10.23), khả năng thực hiện các dự án trong tương lai dài hạn sẽ vẫn phụ thuộc vào các giải pháp sở hữu hạn chế tính linh hoạt, cản trở sự tích hợp và kìm hãm sự phát triển của doanh nghiệp là rất thấp.-

Dù cho việc các công ty xây dựng làm việc với silo dữ liệu và các dữ liệu không minh bạch trong các cơ sở dữ liệu đóng kín có lợi thế đến đâu, thì tương lai của ngành xây dựng sẽ không thể tránh khỏi việc chuyển sang các nền tảng mở, cấu trúc dữ liệu có thể đọc được bởi máy và minh bạch, cũng như tự động hóa dựa trên các nguyên tắc tin cậy. Sự chuyển mình này sẽ diễn ra từ "trên xuống" - dưới áp lực của khách hàng, cơ quan quản lý và xã hội, những người ngày càng yêu cầu tính trách nhiệm, bền vững, minh bạch và tính kinh tế hợp lý.

Kết thúc kỷ nguyên ERP/PMIS khép kín: ngành xây dựng cần những cách tiếp cận mới

Việc sử dụng các hệ thống ERP/PMIS mô-đun cồng kềnh, bao gồm hàng chục triệu dòng mã, làm cho bất kỳ sự thay đổi nào trong đó trở nên cực kỳ khó khăn. Trong khi đó, việc chuyển sang một nền tảng mới trong điều kiện đã có các mô-đun được cấu hình sẵn cho công ty, hàng chục ngàn mục trong cơ sở dữ liệu tài nguyên và hàng ngàn bảng tính sẵn có biến trở thành một quá trình tốn kém và kéo dài. Càng nhiều mã và các giải pháp kiến trúc lỗi thời, thì mức độ không hiệu quả nội bộ càng cao, và mỗi dự án mới sẽ chỉ làm tình hình tồi tệ hơn. Tại nhiều công ty, việc chuyển dữ liệu và tích hợp các giải pháp mới trở thành những cuộc ép phê kéo dài nhiều năm, đi kèm với những thay đổi không ngừng và vô tận việc tìm kiếm các thỏa hiệp. Kết quả là, thường có sự quay trở lại các nền tảng cũ, quen thuộc, bất chấp những giới hạn của chúng.--

Như được nhấn mạnh trong báo cáo Đức "Sổ đen", về các sự cố hệ thống trong quản lý dữ liệu xây dựng, việc phân mảnh thông tin và thiếu một cách tiếp cận tập trung trong việc quản lý nó là nguyên nhân chính dẫn đến sự giảm hiệu quả. Nếu không có sự tiêu chuẩn hoá và tích hợp, dữ liệu sẽ mất giá trị, trở thành lưu

trữ thay vì công cụ quản lý.

> Nguyên nhân chính dẫn đến sự suy giảm chất lượng dữ liệu là việc lập kế hoạch và kiểm soát các dự án xây dựng chưa đầy đủ, điều này thường dẫn đến sự gia tăng đáng kể chi phí. Trong phần “Tâm điểm: bùng nổ chi phí” của “Sổ đen”, phân tích các yếu tố chính góp phần vào những hệ quả không mong muốn này. Trong số đó có việc phân tích nhu cầu không đầy đủ, thiếu các báo cáo kỹ thuật - kinh tế và sự lập kế hoạch không thống nhất dẫn đến các chi phí phát sinh mà có thể tránh được.

> Trong một hệ sinh thái CNTT trưởng thành, việc thay thế hệ thống lỗi thời so với việc thay thế một cột chịu lực trong một tòa nhà đã được xây dựng là tương đương. Không chỉ đơn giản là tháo dỡ cái cũ và lắp đặt cái mới - mà còn quan trọng là thực hiện việc này sao cho tòa nhà vẫn giữ nguyên tính ổn định, các tầng không sụp đổ, và tất cả các hệ thống truyền thông vẫn hoạt động. Đây chính là sự phức tạp: bất kỳ sai lầm nào cũng có thể dẫn đến hậu quả nghiêm trọng cho toàn bộ hệ thống của công ty.

Tuy nhiên, các nhà phát triển các sản phẩm ERP lớn cho ngành xây dựng vẫn tiếp tục sử dụng số lượng mã đã viết như một lập luận ủng hộ nền tảng của họ. Tại các hội nghị chuyên ngành, vẫn thường nghe những câu như: "Để tái tạo một hệ thống như vậy sẽ cần đến 150 năm người," mặc dù thực tế, phần lớn chức năng của các hệ thống tương tự ẩn sau các cơ sở dữ liệu và những chức năng khá đơn giản để làm việc với bảng, được đóng gói trong một giao diện người dùng cố định đặc biệt. Thực tế, khối lượng mã "150 năm người" trở thành gánh nặng hơn là lợi thế cạnh tranh. Càng nhiều mã, chi phí duy trì càng cao, việc thích ứng với các điều kiện mới càng khó khăn và ngưỡng gia nhập cho các nhà phát triển và khách hàng mới càng cao hơn.

> Ngày nay, nhiều hệ thống xây dựng mô-đun giống như những "cấu trúc Frankenstein" cồng kềnh và lỗi thời, nơi mà bất kỳ thay đổi không thận trọng nào cũng có thể dẫn đến sự cố. Mỗi mô-đun mới chỉ làm phức tạp thêm hệ thống đã quá tải, biến nó thành một mê cung chỉ có một nhóm nhỏ chuyên gia hiểu được, làm cho việc bảo trì và nâng cấp trở nên khó khăn hơn.

Sự phức tạp được cả những nhà phát triển nhận thức và họ thường xuyên dừng lại để thực hiện refactoring - xem xét lại kiến trúc với sự xuất hiện của các công nghệ mới. Tuy nhiên, ngay cả khi việc refactoring được thực hiện thường xuyên, sự phức tạp không thể tránh khỏi sẽ gia tăng. Các kiến trúc sư của những hệ thống như vậy đã quen với sự phức tạp ngày càng tăng, nhưng đối với người dùng mới và các chuyên gia, đây trở thành một rào cản không thể vượt qua. Kết quả là, mọi chuyên môn tập trung vào tay vài nhà phát triển, và hệ thống không còn khả năng mở rộng. Trong ngắn hạn, những chuyên gia này hữu ích, nhưng về lâu dài, họ trở thành một phần của vấn đề.

> *Các tổ chức sẽ tiếp tục tích hợp dữ liệu "nhỏ" với dữ liệu lớn của họ, và thật ngốc nghếch nếu ai đó tin rằng một ứng dụng - dù có đắt tiền hay đáng tin cậy đến đâu - có thể xử lý tất cả. Phil Simon, người dẫn chương trình podcast Những Cuộc Thảo Luận Về Sự Hợp Tác.*

Xuất hiện một câu hỏi hợp lý: liệu chúng ta có thực sự cần những hệ thống cồng kềnh và khép kín như vậy để tính toán chi phí và thời gian trong hình thức bảng biểu, khi mà trong các ngành khác có cùng nhiệm vụ

đã lâu sử dụng các công cụ phân tích với dữ liệu mở và logic minh bạch?

Hiện tại, các nền tảng mô-đun khép kín vẫn được ưa chuộng trong ngành xây dựng, chủ yếu là do tính chất của kế toán định lượng. Những hệ thống này thường được sử dụng để quản lý các sơ đồ "xám" hoặc không minh bạch, cho phép che giấu chi phí thực tế khỏi khách hàng. Tuy nhiên, khi sự phát triển kỹ thuật số ngày càng chín muồi, đặc biệt là từ phía khách hàng và sự chuyển mình của ngành trong cái gọi là "kỷ nguyên Uber hóa", các bên trung gian, tức là các công ty xây dựng với hệ thống ERP của họ, sẽ mất đi tầm quan trọng trong việc tính toán thời gian và chi phí. Điều này sẽ mãi mãi thay đổi bộ mặt của ngành công nghiệp xây dựng. Chi tiết hơn sẽ được bàn đến trong phần cuối của cuốn sách và trong chương "Xây dựng 5.0: làm cách nào để kiếm tiền khi không thể che giấu được nữa".

> Hàng nghìn giải pháp công nghệ cũ kỹ, tích lũy trong suốt 30 năm qua với hàng triệu giờ công sức đã được đầu tư vào phát triển, sẽ bắt đầu biến mất nhanh chóng. Việc chuyển sang quản lý dữ liệu mở, minh bạch và linh hoạt là điều không thể tránh khỏi. Câu hỏi chỉ là những công ty nào có thể thích ứng với những thay đổi này, và công ty nào sẽ vẫn là nạn nhân của mô hình cũ.

Tình hình tương tự cũng xảy ra trong lĩnh vực công cụ CAD (BIM), dữ liệu của chúng hiện nay đã lấp đầy các tham số lớn của các thực thể dự án trong hệ thống ERP/PMIS. Ban đầu, ý tưởng BIM (được phát triển vào năm 2002) dựa trên khái niệm cơ sở dữ liệu tích hợp duy nhất, tuy nhiên trên thực tế, việc làm việc với BIM hiện đòi hỏi một loạt các phần mềm và định dạng chuyên dụng. Điều lẽ ra nên đơn giản hóa việc thiết kế và quản lý xây dựng đã trở thành một lớp giải pháp sở hữu khác, làm phức tạp hóa việc tích hợp và giảm tính linh hoạt của doanh nghiệp.

Các bước tiếp theo: sử dụng hiệu quả dữ liệu dự án

Ở phần này, chúng tôi đã cho thấy cách mà dữ liệu có cấu trúc trở thành nền tảng cho các phép tính chính xác về chi phí và thời gian của các dự án xây dựng. Tự động hóa các quy trình QTO, lập kế hoạch thời gian và tính toán dự toán giảm bớt khối lượng công việc và tăng cường độ chính xác của kết quả một cách đáng kể.

Tóm lại phần này, cần nhấn mạnh các bước thực tiễn chính sẽ giúp áp dụng những phương pháp đã được thảo luận vào công việc hàng ngày của bạn. Những phương pháp này rất phổ quát — chúng có ích cho cả quá trình chuyển đổi số của công ty lẫn công việc hàng ngày của các chuyên gia làm các phép tính:

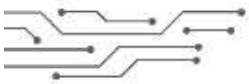
- Tự động hóa các phép tính định kỳ
  - Cố gắng tìm các phép tính tiêu chuẩn về công việc mà bạn có thể liên quan đến trong công việc của mình
  - Phân tích các phương pháp tính toán hoặc định mức công việc hoặc quy trình trên công trường xây dựng ở quốc gia của bạn-
  - Nếu bạn làm việc với các hệ thống CAD - hãy tìm hiểu các tính năng tự động trích xuất các thông số kỹ thuật và dữ liệu QTO trong phần mềm CAD (BIM) của bạn
  - Sử dụng LLM để viết bản nháp mã cho việc tự động hóa các phép tính
- Phát triển các công cụ riêng cho QTO
  - Tạo các kịch bản hoặc bảng tính để tự động hóa việc tính toán khối lượng

- [ ] Chuẩn hóa các loại và nhóm yếu tố để áp dụng một cách đồng nhất trong việc định mức
- [ ] Tài liệu hóa phương pháp tính toán để đảm bảo tính tái lập của kết quả trong các dự án mới

- Tích hợp các khía cạnh khác nhau của dự án vào công việc của bạn
  - [ ] Nếu bạn làm việc với các hệ thống mô-đun, hãy cố gắng hình dung quy trình của bạn không chỉ như sơ đồ hoặc biểu đồ, mà còn ở cấp độ dữ liệu - đặc biệt là dưới dạng bảng
  - [ ] Thành thạo việc tự động kết hợp dữ liệu được trích xuất từ cơ sở dữ liệu CAD với các phép tính - thông qua mã Python, sử dụng nhóm, lọc và tổng hợp
  - [ ] Tạo các hình ảnh trực quan dễ hiểu về các nhóm QTO để trình bày thông tin tổng hợp với đồng nghiệp và khách hàng

Những bước này sẽ giúp bạn xây dựng một hệ thống tính toán bền vững, dựa trên tự động hóa và chuẩn hóa dữ liệu. Cách tiếp cận này sẽ cải thiện độ chính xác, giảm thiểu công việc thường ngày liên quan đến các phép tính.

Các chương tiếp theo tập trung vào các khía cạnh kỹ thuật của các sản phẩm CAD (BIM) và lý do tại sao việc tích hợp cơ sở dữ liệu CAD vào các quy trình kinh doanh của các công ty vẫn còn khó khăn. Nếu bạn không quan tâm đến lịch sử triển khai BIM trong xây dựng, sự tiến hóa của các công cụ CAD và các đặc điểm kỹ thuật liên quan đến việc làm việc với các công nghệ này, bạn có thể chuyển ngay đến phần thứ bảy của cuốn sách "Quyết định dựa trên dữ liệu".



TỐI ĐA TIỆN ÍCH VỚI BẢN IN

Bạn đang cầm trên tay phiên bản kỹ thuật số miễn phí của Data-Driven Construction. Để làm việc thuận tiện hơn và truy cập nhanh chóng vào tài liệu, chúng tôi khuyên bạn nên chú ý đến ấn phẩm in:



- Luôn sẵn sàng: cuốn sách ở định dạng in sẽ trở thành công cụ làm việc đáng tin cậy, cho phép bạn nhanh chóng tìm kiếm và sử dụng các hình ảnh và biểu đồ cần thiết trong bất kỳ tình huống làm việc nào.
- Chất lượng hình ảnh cao: tất cả hình ảnh và đồ họa trong ấn phẩm in được trình bày với chất lượng tốt nhất.
- Truy cập nhanh chóng vào thông tin: điều hướng thuận tiện, khả năng ghi chú, đánh dấu và làm việc với cuốn sách ở bất kỳ đâu.

Khi bạn mua phiên bản in đầy đủ của cuốn sách, bạn nhận được công cụ tiện lợi cho việc làm việc thoải mái và hiệu quả với thông tin: khả năng sử dụng nhanh chóng các tài liệu hình ảnh trong công việc hàng ngày, nhanh chóng tìm thấy các sơ đồ cần thiết và ghi chú. Ngoài ra, việc mua hàng của bạn hỗ trợ việc phổ biến kiến thức mở.

Bạn có thể đặt hàng phiên bản in của cuốn sách tại: datadrivenconstruction.io/books



VI PHẦN

CAD và BIM: TIẾP THỊ, THỰC TẾ VÀ TƯƠI LAI CỦA DỮ LIỆU DỰ ÁN TRONG XÂY DỰNG

Phần thứ sáu của cuốn sách cung cấp một phân tích phê phán về sự tiến hóa của công nghệ CAD và BIM cũng như ảnh hưởng của chúng đối với quy trình quản lý dữ liệu trong xây dựng. Sự chuyển mình lịch sử của khái niệm BIM từ ý tưởng cơ sở dữ liệu tích hợp ban đầu đến các cấu trúc tiếp thị hiện đại được thúc đẩy bởi các nhà cung cấp phần mềm được phân tích. Ảnh hưởng của các định dạng độc quyền và hệ thống khép kín đối với hiệu quả làm việc với dữ liệu dự án và năng suất tổng thể của ngành xây dựng được đánh giá. Những vấn đề về khả năng tương thích của các hệ thống CAD khác nhau cũng như khó khăn trong việc tích hợp chúng vào các quy trình kinh doanh của các công ty xây dựng được phân tích chi tiết. Các xu hướng hiện tại trong việc chuyển sang các định dạng dữ liệu mở đơn giản hơn, chẳng hạn như USD, và tác động tiềm tàng của chúng đối với ngành được xem xét. Các phương pháp tiếp cận thay thế để trích xuất thông tin từ các hệ thống khép kín, bao gồm các phương pháp reverse engineering, cũng được trình bày. Triển vọng ứng dụng trí tuệ nhân tạo và học máy cho tự động hóa quy trình thiết kế và phân tích dữ liệu trong xây dựng được phân tích. Dự báo về sự phát triển của các công nghệ thiết kế, tập trung vào nhu cầu thực tế của người dùng, chứ không phải lợi ích của các nhà cung cấp phần mềm, cũng được nêu ra.

CHƯƠNG 6.1. **SỰ XA XUẤT CỦA CÁC KHÁI NIỆM BIM TRONG NGÀNH XÂY DỰNG**

Phần thứ sáu này, về CAD (BIM), ban đầu không có trong phiên bản đầu tiên của cuốn sách. Các chủ đề về định dạng độc quyền, lõi hình học và hệ thống khép kín thường quá kỹ thuật, nhiều chi tiết và, ở cái nhìn đầu tiên, dường như không hữu ích cho những ai chỉ muốn hiểu về cách thức làm việc với dữ liệu. Tuy nhiên, phản hồi và yêu cầu bổ sung giải thích cho phiên bản đầu tiên của cuốn sách đã chỉ ra rằng: mà không có sự hiểu biết về những phức tạp liên quan đến các cơ chế bên trong của hệ thống CAD, lõi hình học, sự đa dạng của các định dạng và sơ đồ lưu trữ không tương thích cho cùng một dữ liệu, không thể thực sự nhận thức được lý do tại sao các khái niệm được nhà cung cấp thúc đẩy thường cản trở việc làm việc với thông tin và ngăn cản việc chuyển sang thiết kế tham số mở. Chính vì vậy phần này đã lấy được vị trí riêng của mình trong cấu trúc của cuốn sách. Nếu chủ đề CAD (BIM) không phải là ưu tiên của bạn, bạn có thể chuyển ngay đến phần tiếp theo - "PHẦN VII: Quyết định dựa trên dữ liệu, phân tích, tự động hóa và học máy".

Lịch sử xuất hiện của BIM và open BIM như là các khái niệm tiếp thị từ các nhà cung cấp CAD

Với sự xuất hiện của dữ liệu số vào những năm 90, công nghệ máy tính đã được áp dụng không chỉ trong các quy trình kinh doanh mà còn trong các quy trình thiết kế, dẫn đến sự ra đời của các khái niệm như CAD (hệ thống thiết kế tự động) và sau này, BIM (mô hình thông tin xây dựng).

Tuy nhiên, giống như bất kỳ đổi mới nào, chúng không phải là điểm kết thúc của sự phát triển. Các khái niệm như BIM đã trở thành một giai đoạn quan trọng trong lịch sử ngành xây dựng, nhưng rồi một ngày nào đó chúng có thể nhường chỗ cho các công cụ và phương pháp tiên tiến hơn, đáp ứng tốt hơn cho các thách thức của tương lai.

> Bị ảnh hưởng bởi các nhà cung cấp CAD và rơi vào sự phức tạp của việc triển khai riêng, khái niệm BIM, xuất hiện vào năm 2002, có thể sẽ không sống sót đến tuổi ba mươi, giống như một ngôi sao rock bùng nổ nhưng nhanh chóng tàn lụi. Nguyên nhân rất đơn giản: yêu cầu của các chuyên gia xử lý dữ liệu thay đổi nhanh hơn so với khả năng thích ứng của các nhà cung cấp CAD.

Đối mặt với sự thiếu hụt dữ liệu chất lượng, các chuyên gia hiện nay trong ngành xây dựng yêu cầu tính tương thích đa nền tảng và quyền truy cập vào dữ liệu mở từ các dự án CAD để đơn giản hóa việc phân tích và xử lý của họ. Sự phức tạp của dữ liệu CAD và sự rắc rối trong việc xử lý chúng ảnh hưởng tiêu cực đến tất cả các bên tham gia trong quá trình xây dựng: nhà thiết kế, quản lý dự án, công nhân tại công trường và cuối cùng là khách hàng.

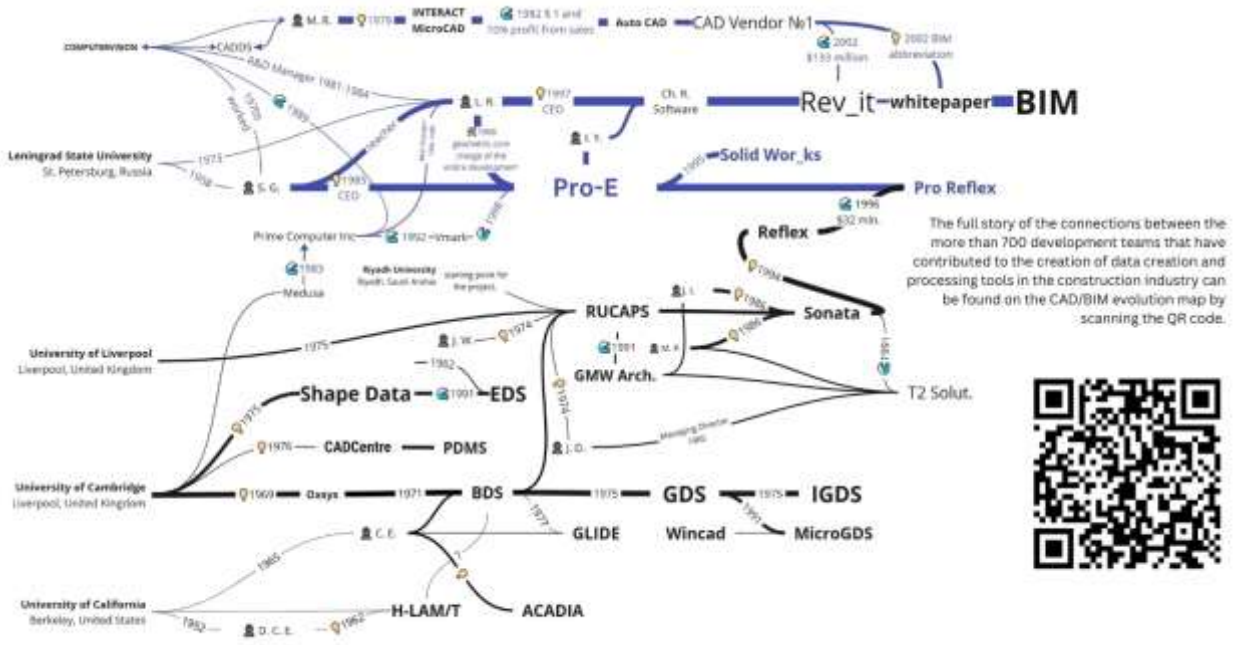
Ngày nay, khách hàng và nhà đầu tư không nhận được bộ dữ liệu hoàn chỉnh cho việc vận hành mà chỉ nhận được các container ở định dạng CAD cần có các lõi hình học phức tạp, sự hiểu biết về sơ đồ dữ liệu, tài liệu API cập nhật hàng năm và phần mềm CAD (BIM) chuyên dụng để làm việc với dữ liệu. Trong khi đó, phần lớn dữ liệu dự án vẫn không được sử dụng.

Hiện nay, trong thế giới thiết kế và xây dựng, sự phức tạp trong việc truy cập dữ liệu CAD dẫn đến việc quản

lý dự án quá mức. Các công ty vừa và lớn, hoạt động với dữ liệu CAD hoặc phát triển các giải pháp BIM, hoặc buộc phải duy trì mối quan hệ chặt chẽ với các nhà cung cấp giải pháp CAD để truy cập dữ liệu qua API, hoặc tìm cách vượt qua các hạn chế của nhà cung cấp CAD bằng cách sử dụng các bộ chuyển đổi SDK đắt tiền để thực hiện reverse engineering nhằm lấy dữ liệu mở.

> Cách tiếp cận sử dụng dữ liệu độc quyền đã lỗi thời và không còn đáp ứng được yêu cầu của môi trường kỹ thuật số hiện đại. Tương lai sẽ chia các công ty thành hai loại: những công ty sử dụng hiệu quả dữ liệu mở và những công ty sẽ rời khỏi thị trường.

Khái niệm BIM (Xây dựng Thông tin Mô hình) đã xuất hiện trong ngành xây dựng với việc phát hành một trong những nhà cung cấp CAD lớn - Whitepaper BIM vào năm 2002 và, kết hợp với khái niệm kỹ thuật BOM (Danh sách Vật liệu), bắt nguồn từ cách tiếp cận tham số trong việc tạo ra và xử lý dữ liệu dự án. Cách tiếp cận tham số này đã được triển khai sớm nhất trong hệ thống Pro-E cho thiết kế cơ khí (MCAD). Hệ thống này đã trở thành nguyên mẫu cho nhiều giải pháp CAD hiện đại, bao gồm cả những giải pháp đang được áp dụng trong ngành xây dựng ngày nay.-



Bản đồ lịch sử xuất hiện của khái niệm BIM và các khái niệm tương tự.

Các nhà báo và cố vấn AEC, đã quảng bá công cụ của các nhà cung cấp CAD cho đến đầu những năm 2000, từ năm 2002 đã chuyển sự chú ý sang Whitepaper BIM. Chính Whitepaper BIM từ năm 2002-2004 và các bài báo được công bố vào các năm 2002, 2003, 2005 và 2007 đã đóng vai trò then chốt trong việc phổ biến khái niệm BIM trong ngành xây dựng.

> *Mô hình hóa thông tin xây dựng — là chiến lược của [tên công ty nhà cung cấp CAD] trong việc áp dụng công nghệ thông tin vào ngành xây dựng. — Whitepaper BIM, 2002*

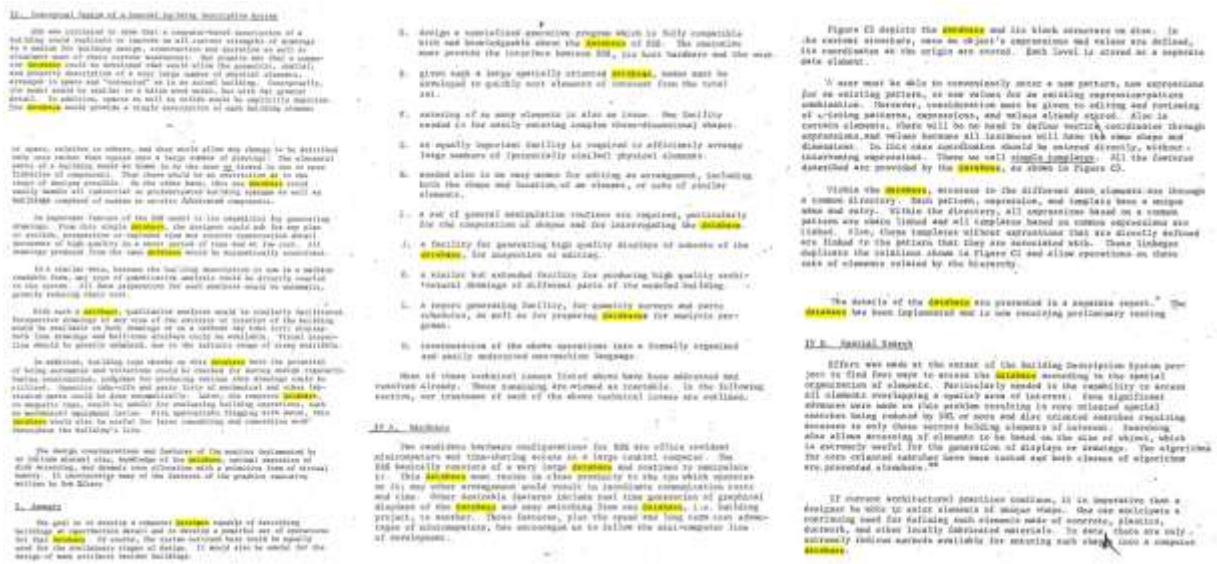
Đến giữa những năm 2000, các "nhà nghiên cứu" bắt đầu liên kết khái niệm BIM, được công bố bởi nhà cung cấp CAD vào năm 2002, với những công trình nghiên cứu trước đó, chẳng hạn như BDS của Charles Eastman, điều này đã trở thành nền tảng cho các hệ thống như GLIDE, GBM, BPM, RUCAPS. Trong công trình đổi mới của mình, "Hệ thống Mô tả Xây dựng" (1974), Charles Eastman đã đặt nền tảng lý thuyết cho mô hình hóa thông tin hiện đại. Thuật ngữ "cơ sở dữ liệu" xuất hiện 43 lần trong tài liệu của ông (Hình 6.12) - thường xuyên hơn bất kỳ từ nào khác, ngoại trừ từ "tòa nhà".-

> Ý tưởng chính của Eastman là tất cả thông tin về một tòa nhà - từ hình học đến các thuộc tính của các phần tử và mối quan hệ của chúng - cần được lưu trữ trong một cơ sở dữ liệu có cấu trúc thống nhất. Chính từ cơ sở dữ liệu này có thể tự động tạo ra bản vẽ, bảng thông số kỹ thuật, tính toán và phân tích sự phù hợp với quy định. Eastman đã chỉ trích trực tiếp các bản vẽ như một phương thức giao tiếp lỗi thời và thừa thãi, chỉ ra rằng nó có sự trùng lặp thông tin, vấn đề cập nhật và cần phải làm mới thủ công khi có thay đổi. Thay vào đó, ông đã đề xuất một mô hình kỹ thuật số thống nhất trong cơ sở dữ liệu, nơi bất kỳ thay đổi nào cũng được nhập một lần và tự động phản ánh trong tất cả các biểu diễn.

Đáng chú ý, trong khái niệm của mình, Eastman không đặt hình ảnh lên hàng đầu. Trung tâm của hệ thống của ông thực sự là thông tin: các tham số, mối liên hệ, các thuộc tính, khả năng phân tích và tự động hóa. Các bản vẽ trong cách hiểu của ông chỉ là một trong những hình thức hiển thị dữ liệu từ cơ sở dữ liệu, chứ không phải là nguồn thông tin dự án.

Trong các tài liệu đầu tiên về BIM từ nhà cung cấp CAD hàng đầu, cụm từ "cơ sở dữ liệu" được sử dụng cũng thường xuyên như trong BDS của Charles Eastman - 23 lần [60] trên bảy trang và là một trong những từ phổ biến nhất trong tài liệu sau "Building", "Information", "Modeling" và "Design". Tuy nhiên, đến năm 2003, trong các tài liệu tương tự, thuật ngữ "cơ sở dữ liệu" chỉ xuất hiện hai lần [61], và đến cuối những năm 2000, chủ đề cơ sở dữ liệu gần như biến mất khỏi cuộc thảo luận về dữ liệu thiết kế. Cuối cùng, khái niệm "cơ sở dữ liệu tích hợp duy nhất cho phân tích hình ảnh và số lượng" đã không bao giờ được thực hiện đầy đủ.

Do đó, ngành xây dựng đã trải qua một hành trình từ khái niệm tiến bộ BDS của Charles Eastman với trọng tâm vào cơ sở dữ liệu và ý tưởng của Samuel Geisberg về việc tự động cập nhật dữ liệu thiết kế từ cơ sở dữ liệu trong sản phẩm máy tính Pro-E (tiền thân của các giải pháp CAD phổ biến được sử dụng ngày nay trong xây dựng) đến BIM tiếp thị hiện đại, nơi quản lý dữ liệu thông qua cơ sở dữ liệu gần như không được đề cập, mặc dù chính khái niệm này đã nằm ở nền tảng của các phát triển lý thuyết ban đầu.



Hình 6.12 Trong khái niệm BDS, được mô tả bởi Charles Eastman vào năm 1974, cụm từ "Cơ sở dữ liệu" (được tô sáng bằng màu vàng) đã được sử dụng 43 lần.

BDS và các khái niệm tương tự trước những năm 2000 đã được phát triển như một cơ sở dữ liệu kỹ thuật số về các tòa nhà, chứ không phải là công cụ hình ảnh. BIM vào năm 2002 đã trở thành công cụ thiết kế, trong đó cơ sở dữ liệu đã trở nên thứ yếu. Những gì chúng ta đã mất khi chuyển từ BDS và các khái niệm tương tự trong những năm 1990 sang BIM vào giữa những năm 2010:

- Cơ sở dữ liệu mở: BDS và các khái niệm tương tự nhấn mạnh vào phân tích, BIM - vào thiết kế.
- Tính linh hoạt trong việc làm việc với dữ liệu: BDS chú trọng đến phân tích dữ liệu, BIM - vào các quá trình mà không rõ dựa trên dữ liệu nào.
- Tính minh bạch: BDS được thiết kế như một cơ sở dữ liệu mở và tích hợp, trong khi các nhà cung cấp CAD trong BIM đã biến các cơ sở dữ liệu của họ thành các hệ thống hoàn toàn khép kín và trong 20 năm qua đã không thành công trong việc chống lại các công cụ đảo ngược kỹ thuật mở khóa các định dạng độc quyền.

Trong 30 năm qua, các nhà thiết kế vẫn chưa có quyền truy cập vào "cơ sở dữ liệu tích hợp" và sau hai mươi năm cơn cuồng nhiệt marketing xoay quanh các công cụ BIM, ngành xây dựng bắt đầu nhận ra hậu quả của sự đam mê này.

Thực tế của BIM: thay vì cơ sở dữ liệu tích hợp - các hệ thống mô-đun kín

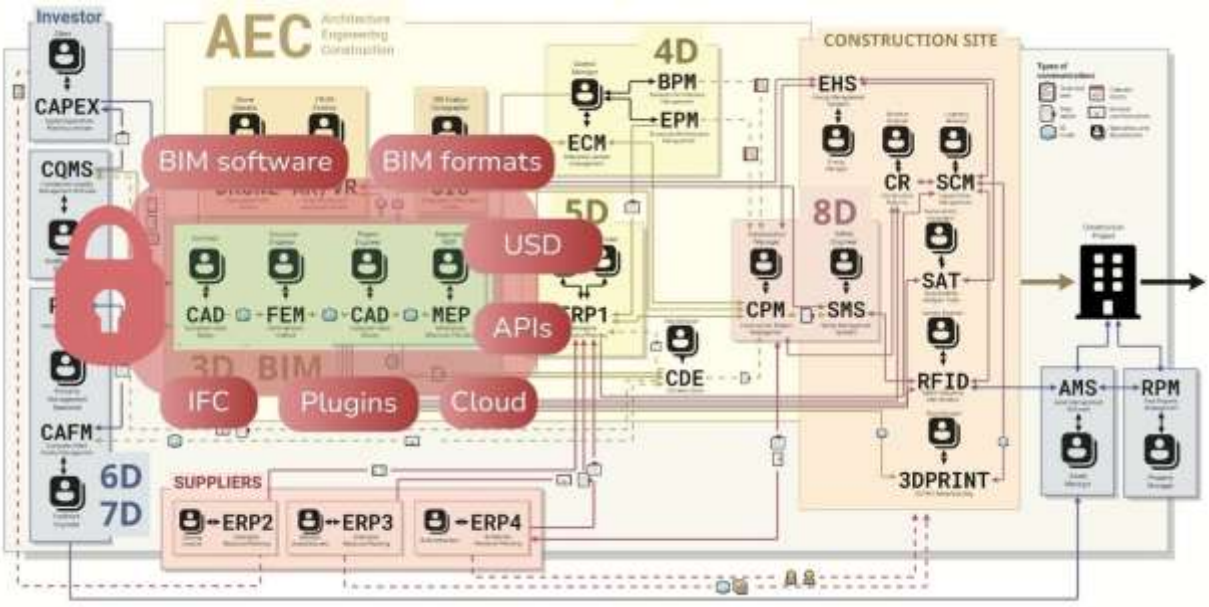
Thay vì tập trung vào dữ liệu, cấu trúc và tích hợp chúng trong các quy trình thống nhất, người dùng hệ thống CAD (BIM) buộc phải làm việc với một tập hợp các giải pháp độc quyền phân mảnh, mỗi giải pháp đều đặt ra các quy tắc riêng.

- Cơ sở dữ liệu thống nhất mà các tài liệu trắng BIM đã đề cập vẫn là một huyền thoại. Bất chấp các tuyên bố lớn lao, quyền truy cập vào dữ liệu vẫn bị hạn chế và phân phối giữa các hệ thống khép kín.
- Các mô hình BIM đã trở thành một hệ sinh thái khép kín thay vì một công cụ. Thay vì chia sẻ thông tin một cách minh bạch, người dùng buộc phải trả tiền cho các gói đăng ký và sử dụng các API độc quyền.

- Dữ liệu thuộc về các nhà cung cấp, chứ không phải là người dùng. Thông tin về các dự án bị khóa trong các định dạng độc quyền hoặc dịch vụ đám mây, thay vì có sẵn trong các định dạng mở và độc lập.

Các kỹ sư thiết kế và quản lý dự án thường không có quyền truy cập vào cơ sở dữ liệu của các hệ thống CAD, cũng như định dạng mà dữ liệu của dự án của họ được lưu trữ. Điều này khiến việc kiểm tra thông tin nhanh chóng hoặc thiết lập yêu cầu về cấu trúc và chất lượng dữ liệu trở nên không khả thi. Quyền truy cập vào các dữ liệu này đòi hỏi một loạt các phần mềm chuyên biệt, liên kết thông qua API và plugin, dẫn đến sự quan liêu thái quá trong các quy trình trong ngành xây dựng. Trong khi đó, những dữ liệu này được sử dụng đồng thời bởi hàng chục hệ thống thông tin và hàng trăm chuyên gia.-

> *Chúng ta cần phải có khả năng quản lý tất cả các dữ liệu này [CAD (BIM)], lưu trữ chúng dưới dạng số và bán phần mềm để quản lý vòng đời và quy trình, bởi vì cho mỗi kỹ sư [nhà thiết kế] tạo ra một cái gì đó [trong chương trình CAD], có mười người làm việc với những dữ liệu này.*
>
> *– Giám đốc điều hành của nhà cung cấp CAD, người đã tạo ra khái niệm BIM, năm 2005.*



Hình 6.13 Cơ sở dữ liệu CAD (BIM) vẫn là một trong những hệ thống khép kín cuối cùng cho các bộ phận CNTT và quản lý dữ liệu trong hệ sinh thái kinh doanh xây dựng.

Khi trở nên rõ ràng rằng BIM chủ yếu là phương tiện thương mại hóa các cơ sở dữ liệu, chứ không phải một công cụ đầy đủ để quản lý chúng, câu hỏi hợp lý đặt ra là: làm thế nào để lấy lại quyền kiểm soát dữ liệu? Câu trả lời là sử dụng các cấu trúc dữ liệu mở, nơi mà người dùng trở thành chủ sở hữu thông tin, không phải nhà cung cấp phần mềm.

Người dùng và các nhà phát triển giải pháp trong ngành xây dựng, giống như các đồng nghiệp của họ trong các ngành khác trong nền kinh tế, sẽ không thể tránh khỏi việc từ chối thuật ngữ mơ hồ của các nhà cung

cấp phần mềm, vốn đã chiếm ưu thế trong 30 năm qua, và tập trung vào các khía cạnh chính của số hóa - "dữ liệu" và "quy trình".

Cuối những năm 1980, định hướng chính của sự phát triển công nghệ kỹ thuật số trong xây dựng được xem như là vấn đề tiếp cận dữ liệu và quản lý thông tin dự án. Tuy nhiên, theo thời gian, trọng tâm đã chuyển dịch. Thay vì phát triển các phương pháp làm việc với dữ liệu minh bạch và dễ tiếp cận, việc thúc đẩy định dạng IFC và khái niệm open BIM đã trở thành một nỗ lực nhằm đánh lạc hướng sự chú ý của các chuyên gia khỏi những vấn đề liên quan đến quản lý cơ sở dữ liệu dự án.

Sự xuất hiện của định dạng mở IFC trong ngành xây dựng

Định dạng mở được gọi là IFC (Industry Foundation Classes) được xác định là tiêu chuẩn cho việc đảm bảo sự tương thích giữa các hệ thống CAD- (BIM-). Việc phát triển nó được thực hiện trong khuôn khổ các tổ chức được thành lập và kiểm soát bởi các nhà cung cấp CAD lớn nhất. Dựa trên định dạng IFC, hai công ty CAD đã phát triển một khái niệm tiếp thị OPEN BIM vào năm 2012.

> IFC (Industry Foundation Classes) – là một tiêu chuẩn mở để trao đổi dữ liệu trong ngành xây dựng, được phát triển nhằm đảm bảo sự tương thích giữa các hệ thống CAD (BIM) khác nhau.

Khái niệm Open BIM đề cập đến việc làm việc với thông tin từ các cơ sở dữ liệu CAD và trao đổi thông tin giữa các hệ thống thông qua định dạng mở để trao đổi dữ liệu CAD - IFC.

> *Chương trình Open BIM là một chiến dịch tiếp thị được khởi xướng bởi … [1 nhà cung cấp CAD], … [2 nhà cung cấp CAD] và các công ty khác nhằm khuyến khích và hỗ trợ việc thúc đẩy có sự phối hợp toàn cầu về khái niệm OPEN BIM trong toàn bộ ngành AEC, với sự giao tiếp thống nhất và thương hiệu chung, có sẵn cho các thành viên trong chương trình.— Từ trang web của nhà cung cấp CAD, Chương trình OPEN BIM, 2012 [113]*

IFC đã được trường Đại học Kỹ thuật Munich chuyển thể từ định dạng STEP trong ngành cơ khí vào cuối những năm 1980, và sau đó được một công ty thiết kế lớn và một nhà cung cấp CAD lớn đăng ký để thành lập liên minh IAI (Liên minh Ngành Công nghiệp về Tính tương tác) vào năm 1994. Định dạng IFC được phát triển nhằm đảm bảo tính tương tác giữa các hệ thống CAD khác nhau và dựa trên các nguyên tắc được thiết lập trong định dạng STEP thuộc ngành cơ khí, định dạng này lại được phát triển từ định dạng IGES, được tạo ra vào năm 1979 bởi một nhóm người dùng và nhà cung cấp CAD với sự hỗ trợ của NIST (Viện Tiêu chuẩn và Công nghệ Quốc gia) và Bộ Quốc phòng Hoa Kỳ.-

Tuy nhiên, cấu trúc phức tạp của IFC, sự phụ thuộc chặt chẽ vào lõi hình học, cũng như những khác biệt trong việc triển khai định dạng bởi các giải pháp phần mềm khác nhau đã dẫn đến nhiều vấn đề trong việc áp dụng thực tế. Các chuyên gia trong lĩnh vực cơ khí cũng đã gặp phải những khó khăn tương tự — mất chi tiết, giới hạn độ chính xác và yêu cầu sử dụng các định dạng trung gian — khi làm việc với các định dạng IGES, STEP, từ đó đã hình thành nên IFC.



Hình 6.14 Bản đồ mối quan hệ giữa các đội ngũ phát triển và sản phẩm CAD (BIM) [116].

> Vào năm 2000, chính nhà cung cấp CAD, người đã đăng ký định dạng IFC và thành lập tổ chức IAI (sau này là bS), đã xuất bản tài liệu Whitepaper "Thiết kế và sản xuất tích hợp: Lợi ích và cơ sở". Trong tài liệu nhấn mạnh tầm quan trọng của việc duy trì đầy đủ chi tiết dữ liệu khi trao đổi giữa các chương trình trong cùng một hệ thống, mà không sử dụng các định dạng trung gian như IGES, STEP (tương tự như IFC). Thay vào đó, đề xuất cung cấp quyền truy cập trực tiếp cho các ứng dụng vào cơ sở dữ liệu chính của CAD, điều này nhằm ngăn chặn việc mất độ chính xác của thông tin.

Vào năm 2002, nhà cung cấp CAD này mua một sản phẩm BOM tham số (Hình. 3.118, xem thêm trong phần ba) và trên cơ sở đó hình thành khái niệm BIM. Kết quả là trong việc trao đổi dữ liệu của các dự án xây dựng hiện nay chỉ sử dụng các định dạng CAD đóng hoặc định dạng IFC (STEP), về sự hạn chế của nó đã được chính nhà cung cấp CAD đề cập vào năm 2000, người đã đưa định dạng này vào ngành xây dựng.-

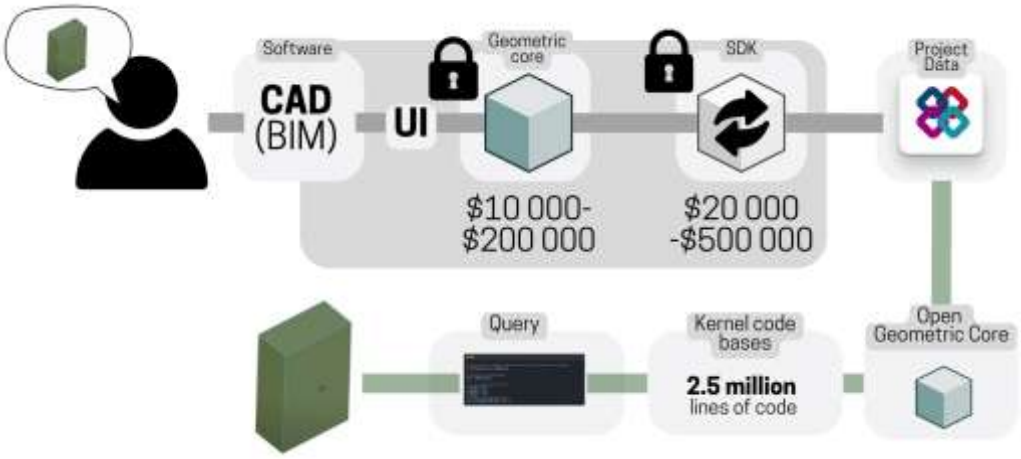
Lịch sử chi tiết về sự tương tác của hơn 700 nhóm phát triển tham gia vào việc tạo ra các công cụ cho việc xây dựng và xử lý dữ liệu xây dựng đã được trình bày trên bản đồ "Evolutie CAD (BIM)" [116].

Định dạng mở IFC bao gồm mô tả hình học của các thành phần của dự án và mô tả siêu thông tin. Để biểu diễn hình học trong định dạng IFC, sử dụng các phương pháp khác nhau như CSG và Swept Solids; tuy nhiên, biểu diễn tham số BREP đã trở thành chuẩn dẫn đầu cho việc truyền tải hình học các thành phần trong định dạng IFC, vì định dạng này được hỗ trợ khi xuất ra từ các chương trình CAD (BIM) và cho phép khả năng chỉnh sửa các thành phần khi nhập lại IFC vào các chương trình CAD.

Vấn đề của định dạng IFC phụ thuộc vào lõi hình học

Trong hầu hết các trường hợp, khi hình học trong IFC được xác định tham số (BREP), việc trực quan hóa hoặc thu nhận các thuộc tính hình học như thể tích hoặc diện tích của các thực thể trong dự án trở nên không thể nếu chỉ có tệp IFC, bởi vì để làm việc với hình học và việc trực quan hóa trong trường hợp này cần có một lõi hình học (Hình. 6.15), mà ban đầu là không có.-

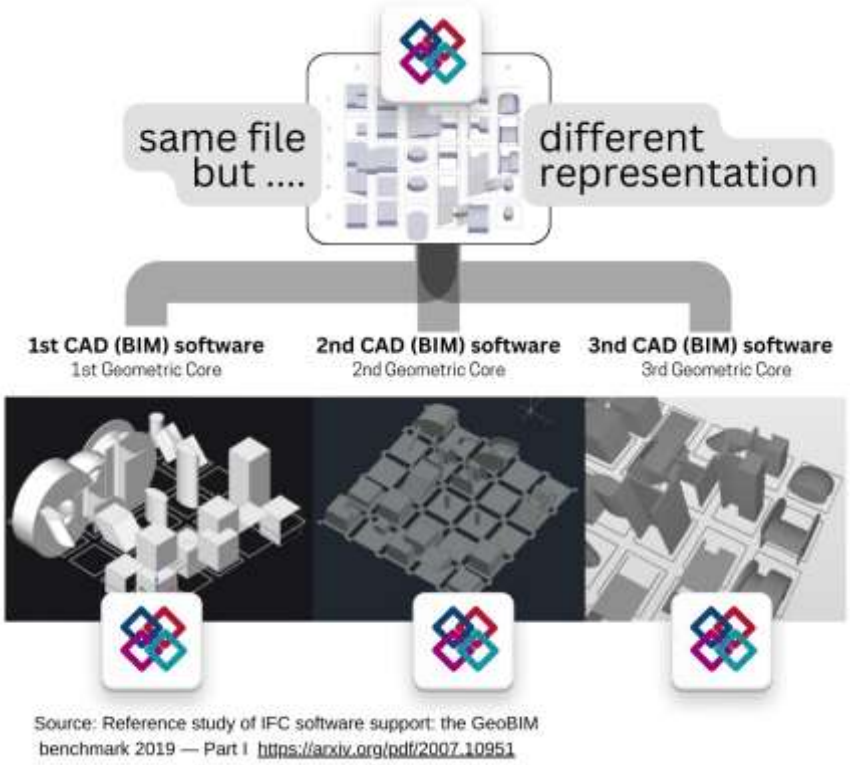
> Lõi hình học là một thành phần phần mềm cung cấp các thuật toán cơ bản cho việc tạo, chỉnh sửa và phân tích các đối tượng hình học trong CAD, BIM và các ứng dụng kỹ thuật khác. Nó chịu trách nhiệm cho việc xây dựng hình học 2D và 3D, cũng như cho các thao tác trên đó, chẳng hạn như: các phép toán Bool, làm mịn, giao nhau, biến đổi và trực quan hóa.



Hình. 6.15 Việc tạo hình học thông qua các chương trình CAD ngày nay đi qua các lõi hình học và SDK độc quyền, mà thường không thuộc sở hữu của các nhà cung cấp CAD.

Mỗi chương trình CAD và bất kỳ chương trình nào làm việc với các định dạng tham số hoặc định dạng IFC đều có lõi hình học riêng của mình hoặc đã mua. Và nếu với các yếu tố nguyên thủy trong định dạng IFC-BREP có thể không phát sinh vấn đề, và trong các chương trình với các lõi hình học khác nhau, các yếu tố này có thể hiển thị tương tự nhau, thì bên cạnh các vấn đề với các động cơ lõi hình học khác nhau, vẫn có rất nhiều yếu tố có các đặc điểm riêng cần hiển thị đúng cách. Vấn đề này được nghiên cứu chi tiết trong nghiên cứu quốc tế "Nghiên cứu tham chiếu hỗ trợ phần mềm IFC", công bố vào năm 2019 [117].

> *Các tập dữ liệu tiêu chuẩn hóa giống nhau đưa ra kết quả mâu thuẫn, trong khi chỉ có ít quy luật chung được phát hiện, và các vấn đề nghiêm trọng được tìm thấy trong việc hỗ trợ tiêu chuẩn [IFC], có thể do độ phức tạp cao của mô hình dữ liệu tiêu chuẩn. Một phần do chính các tiêu chuẩn, vì thường để lại một số chi tiết không xác định, với mức độ tự do cao và các cách diễn giải khác nhau. Chúng cho phép độ phức tạp cao trong việc tổ chức và lưu trữ các đối tượng, điều này không hỗ trợ hiệu quả cho việc hiểu biết chung, các triển khai độc đáo và sự mô hình hóa dữ liệu nhất quán [117].– Nghiên cứu tham chiếu hỗ trợ phần mềm IFC, 2021.*



Hình 6.16 Các lõi hình học khác nhau cung cấp các đại diện khác nhau của cùng một hình học được mô tả theo tham số (từ tài liệu [117]).

Sự hiểu biết đúng đắn về "các vị trí xác định" chỉ dành cho các thành viên trả phí của các tổ chức chuyên môn, những tổ chức này tham gia vào việc phát triển IFC. Do đó, những ai muốn tiếp cận kiến thức quan trọng về những đặc điểm cụ thể của IFC sẽ cố gắng hợp tác với các nhà cung cấp CAD lớn hoặc thu thập kiến thức về những đặc điểm đó qua nghiên cứu cá nhân.

> *Bạn sẽ gặp câu hỏi về việc nhập và xuất dữ liệu qua định dạng IFC và hỏi đồng nghiệp nhà cung cấp: "Tại sao thông tin về việc truyền tham số của các khu vực lại được truyền qua tệp IFC? Trong tài liệu mở không có đề cập gì về việc này." Câu trả lời từ các nhà cung cấp châu Âu "có hiểu biết hơn": "Đúng, không có đề cập, nhưng là được phép."— Từ phỏng vấn nhà phát triển CAD năm 2021 [118]*

IFC mô tả hình học thông qua các nguyên thủy tham số, nhưng không chứa lõi tích hợp — vai trò của nó được thực hiện bởi phần mềm CAD, thông qua lõi hình học, biên dịch hình học. Lõi hình học thực hiện các phép toán toán học và xác định các giao điểm, trong khi IFC chỉ cung cấp dữ liệu để diễn giải. Nếu IFC chứa các mặt không chính xác, các chương trình khác nhau với các lõi hình học khác nhau có thể bỏ qua chúng hoặc phát sinh lỗi, tùy thuộc vào lõi.

> Cuối cùng, để làm việc với định dạng IFC, cần trả lời câu hỏi chính mà rất khó tìm ra câu trả lời thống nhất - công cụ nào, với lõi hình học nào cần sử dụng để đạt được chất lượng dữ liệu mà ban đầu dự án có trong phần mềm CAD từ đó nhận được IFC?

Các vấn đề về chất lượng dữ liệu và sự phức tạp của định dạng IFC không cho phép sử dụng trực tiếp dữ liệu thiết kế để tự động hóa quy trình, phân tích và xử lý dữ liệu, điều này thường dẫn đến việc các nhà phát triển buộc phải sử dụng các giải pháp CAD đóng với "truy cập chất lượng" vào dữ liệu [63], điều mà nhà cung cấp đã đề cập, người đã đăng ký IFC vào năm 1994 [65].

Tất cả các đặc điểm hiển thị và tạo ra các tham số IFC trong lõi hình học chỉ có thể được thực hiện bởi các đội ngũ phát triển lớn có kinh nghiệm làm việc với các lõi hình học. Do đó, thực tiễn hiện nay về các đặc điểm và độ phức tạp của định dạng IFC hầu hết có lợi cho các nhà cung cấp CAD và có nhiều điểm chung với chiến lược của các nhà cung cấp phần mềm lớn "áp dụng, mở rộng, phá hủy", khi sự phức tạp ngày càng tăng của tiêu chuẩn thực tế tạo ra rào cản cho các người chơi nhỏ hơn trên thị trường [94].

Chiến lược của các nhà cung cấp lớn trong chiến lược này có thể nằm ở việc điều chỉnh các tiêu chuẩn mở, thêm các phần mở rộng và chức năng riêng của họ, để tạo ra sự phụ thuộc của người dùng vào sản phẩm của họ nhằm tiêu diệt đối thủ cạnh tranh sau đó.

> Định dạng IFC, được cho là cầu nối phổ quát giữa các hệ thống CAD (BIM) khác nhau, trên thực tế thực hiện vai trò như một chỉ báo cho các vấn đề tương thích giữa các lõi hình học của các nền tảng CAD khác nhau, tương tự như định dạng STEP, từ đó nó đã xuất hiện ban đầu.

Cuối cùng, hiện nay việc triển khai đầy đủ và chất lượng của ontología IFC có thể chịu trách nhiệm những nhà cung cấp CAD lớn, những người có thể đầu tư nguồn lực đáng kể vào việc hỗ trợ tất cả các thực thể và ánh xạ chúng với lõi hình học nội bộ của chính họ, điều mà không tồn tại cho IFC như một tiêu chuẩn. Các nhà cung cấp lớn cũng có khả năng đồng thuận với nhau về các chi tiết kỹ thuật, những điều có thể không khả thi ngay cả đối với những thành viên tích cực nhất trong các tổ chức đang phát triển định dạng IFC.

> Đối với các nhóm độc lập nhỏ và các dự án mã nguồn mở, những người muốn hỗ trợ sự phát triển của các định dạng tương tác, việc thiếu một lõi hình học riêng trở thành vấn đề nghiêm trọng. Không có nó, gần như không thể tính đến tất cả sự phong phú của những chi tiết và sắc thái liên quan đến việc trao đổi dữ liệu giữa các nền tảng.

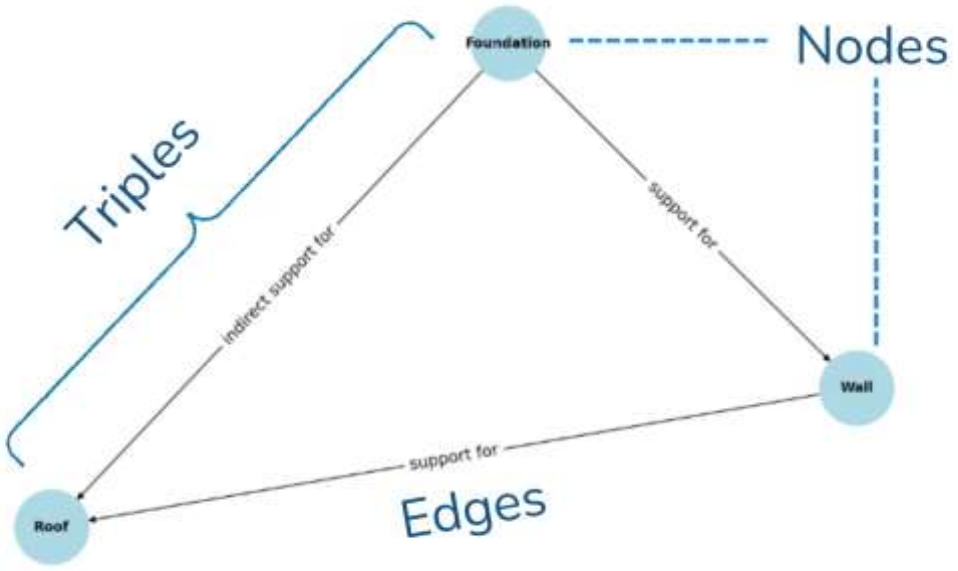
Với sự phát triển của định dạng parametric IFC và khái niệm open BIM, ngành xây dựng đã có những cuộc thảo luận sôi nổi về vai trò của ontología và ngữ nghĩa trong quản lý dữ liệu và quy trình.

Sự xuất hiện của chủ đề ngữ nghĩa và siêu hình trong xây dựng

Nhờ vào các ý tưởng tạo ra internet ngữ nghĩa vào cuối những năm 1990 và những nỗ lực của các tổ chức phát triển định dạng IFC, ngữ nghĩa và ontología đã trở thành một trong những yếu tố chủ chốt của quy chuẩn hóa, được bàn luận trong ngành xây dựng vào giữa những năm 2020.

> Công nghệ ngữ nghĩa bao gồm việc thống nhất, quy chuẩn hóa và điều chỉnh các tập dữ liệu lớn đa dạng, cũng như thực hiện tìm kiếm phức tạp.

Để lưu trữ dữ liệu ngữ nghĩa, ngôn ngữ ontología OWL (Web Ontology Language) được sử dụng, được biểu diễn dưới dạng đồ thị RDF triplet (Resource Description Framework). OWL thuộc về các mô hình đồ thị dữ liệu, về các loại mà chúng ta đã đề cập chi tiết hơn trong chương "Mô hình dữ liệu: quan hệ trong dữ liệu và mối quan hệ giữa các phần tử".-



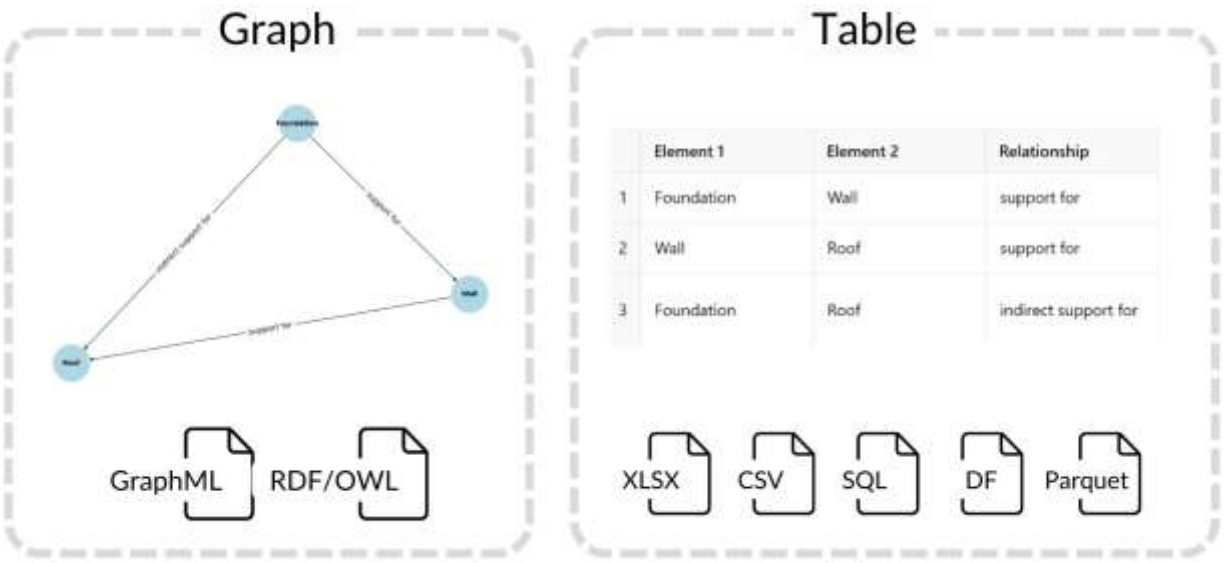
Mô hình dữ liệu RDF: nút (Nodes), liên kết (Edges) và bộ ba (Triples) minh họa các mối quan hệ giữa các phần tử xây dựng.

Về lý thuyết, việc suy luận logic của các chương trình suy diễn (progam để suy diễn logic tự động) cho phép tạo ra các tuyên bố mới dựa trên ontología. Ví dụ, nếu trong ontología xây dựng ghi rằng "móng là điểm tựa cho tường", và "tường là điểm tựa cho mái" (Hình 6.17), các chương trình suy diễn có khả năng tự động kết luận rằng "móng là điểm tựa cho mái".-

Cơ chế tương tự hữu ích cho tối ưu hóa phân tích dữ liệu, vì nó cho phép tránh việc chứng minh rõ ràng tất cả các phụ thuộc. Tuy nhiên, nó không tạo ra kiến thức mới, mà chỉ làm rõ và cấu trúc những sự kiện đã biết.

> Ngữ nghĩa không tự tạo ra ý nghĩa hoặc kiến thức mới và không vượt trội hơn các công nghệ lưu trữ và xử lý dữ liệu khác trong khía cạnh này. Việc trình bày dữ liệu từ các cơ sở dữ liệu quan hệ dưới dạng triplet không làm cho chúng có ý nghĩa hơn. Việc thay thế bảng bằng các cấu trúc đồ thị có thể hữu ích cho việc thống nhất các mô hình dữ liệu, tìm kiếm thuận tiện và chỉnh sửa an toàn, nhưng nó không làm cho dữ liệu "thông minh" hơn — máy tính không bắt đầu hiểu rõ hơn nội dung của chúng.

Các mối liên hệ logic trong dữ liệu có thể được tổ chức mà không cần các công nghệ ngữ nghĩa phức tạp (Hình 6.18). Các cơ sở dữ liệu quan hệ truyền thống (SQL) cũng như các định dạng CSV hoặc XLSX cho phép xây dựng các phụ thuộc tương tự. Ví dụ, trong cơ sở dữ liệu dạng cột, có thể thêm trường "điểm tựa mái" và tự động liên kết mái với móng khi tạo tường. Cách tiếp cận này được thực hiện mà không cần sử dụng RDF, OWL, đồ thị hoặc reasoning, vẫn giữ cho việc lưu trữ và phân tích dữ liệu trở nên đơn giản và hiệu quả.



Hình 6.18 So sánh giữa mô hình dữ liệu đồ thị và dữ liệu bảng đại diện cho cùng một liên hệ logic.

Quyết định của một số công ty xây dựng lớn và tổ chức đang phát triển định dạng IFC [94] theo hướng của khái niệm web ngữ nghĩa, được coi là triển vọng vào cuối những năm 1990, đã có ảnh hưởng đáng kể đến việc phát triển các tiêu chuẩn trong ngành xây dựng.

Tuy nhiên, nghịch lý nằm ở chỗ chính khái niệm web ngữ nghĩa, vốn được thiết kế ban đầu cho internet, đã không được phổ biến rộng rãi ngay cả trong môi trường gốc của nó. Mặc dù đã phát triển RDF và OWL, nhưng một web ngữ nghĩa đầy đủ theo ý tưởng ban đầu vẫn chưa xuất hiện, và việc tạo ra nó hiện nay có vẻ khó xảy ra.

Tại sao các công nghệ ngữ nghĩa không đáp ứng được mong đợi trong xây dựng

Các ngành khác cũng đã gặp phải những hạn chế của công nghệ sử dụng ngữ nghĩa. Trong ngành công nghiệp trò chơi, những nỗ lực mô tả các đối tượng trò chơi và các tương tác của chúng thông qua các ontology đã thất bại do sự thay đổi cao độ của chúng. Cuối cùng, các định dạng dữ liệu đơn giản hơn, chẳng hạn như XML và JSON, cùng với các giải pháp thuật toán, lại trở nên ưa thích hơn. Tình huống tương tự cũng xảy ra trong lĩnh vực bất động sản: do sự khác biệt vùng miền về thuật ngữ và sự biến động thường xuyên của thị trường, việc sử dụng ontology trở nên quá phức tạp, trong khi các cơ sở dữ liệu và tiêu chuẩn đơn giản như RETS [119] lại làm tốt hơn trong việc trao đổi dữ liệu.

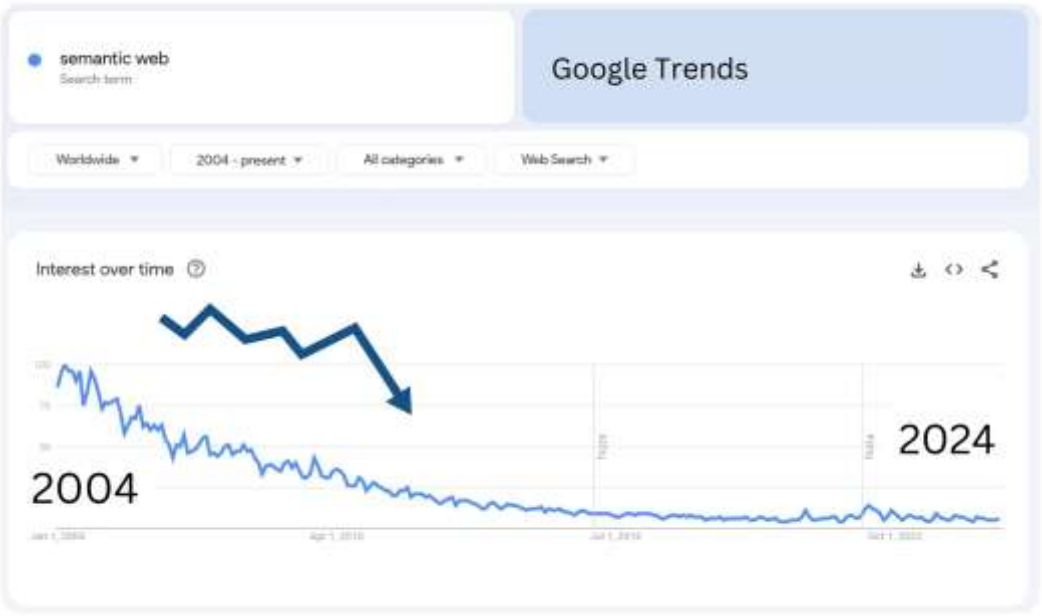
> Những khó khăn kỹ thuật, chẳng hạn như độ phức tạp của việc đánh dấu, khối lượng công việc bảo trì cao và động lực thấp từ phía các nhà phát triển, đã kìm hãm việc áp dụng web ngữ nghĩa ở các lĩnh vực khác của nền kinh tế. RDF (Resource Description Framework) không trở thành một tiêu chuẩn phổ biến, trong khi các ontology quá phức tạp và không kinh tế.

Kết quả là, ý tưởng đầy tham vọng về việc tạo ra một web ngữ nghĩa toàn cầu đã không thành hiện thực. Mặc dù một số yếu tố công nghệ, chẳng hạn như ontology và SPARQL, đã được áp dụng trong các giải pháp doanh nghiệp, nhưng mục tiêu ban đầu là tạo ra một cấu trúc dữ liệu đầy đủ và thống nhất vẫn chưa đạt được.

Khái niệm về một internet mà trong đó các máy tính có thể hiểu được ý nghĩa của nội dung đã trở nên phức tạp về mặt kỹ thuật và không khả thi về mặt thương mại. Đó là lý do tại sao các công ty ủng hộ ý tưởng này theo thời gian đã giảm bớt việc sử dụng nó xuống các công cụ hữu ích riêng lẻ, để lại RDF và OWL cho các nhu cầu đặc thù trong doanh nghiệp, chứ không phải cho internet nói chung. Phân tích Google Trends (Hình 6.19) trong 20 năm qua cho phép đánh giá rằng có thể không còn triển vọng nào cho sự phát triển của web ngữ nghĩa.-

> *Không cần phải nhân rộng các thực thể một cách không cần thiết. Nếu có nhiều lời giải thích logic không mâu thuẫn nào về một hiện tượng, giải thích điều đó một cách tương đương, thì nên, trong điều kiện tương đương, ưu tiên lời giải thích đơn giản nhất.*
> *Lưỡi dao Ockham.*

Ở đây xuất hiện một câu hỏi hợp lý: Tại sao lại sử dụng triplet, risoner và SPARQL trong xây dựng, nếu có thể xử lý dữ liệu bằng các truy vấn cấu trúc phổ biến (SQL, Pandas, Apache®)? Trong các ứng dụng doanh nghiệp, SQL là tiêu chuẩn để làm việc với cơ sở dữ liệu. Ngược lại, SPARQL yêu cầu các cấu trúc đồ thị phức tạp và phần mềm chuyên dụng, và theo xu hướng trên Google, nó không thu hút được sự quan tâm của các nhà phát triển.

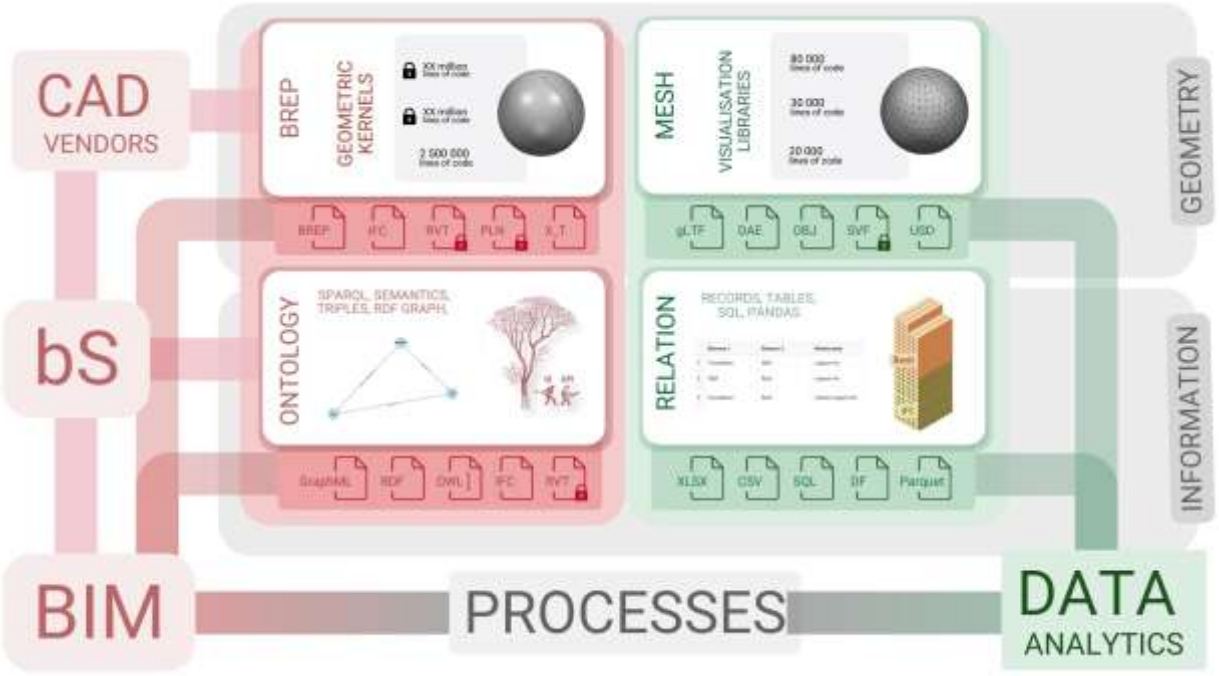


Hình 6.19: Sự quan tâm đến các truy vấn "internet ngữ nghĩa" theo thống kê từ Google.

Cơ sở dữ liệu đồ thị và cây phân loại có thể hữu ích trong một số trường hợp, nhưng việc áp dụng của chúng không phải lúc nào cũng hợp lý cho hầu hết các tác vụ hàng ngày. Cuối cùng, việc tạo ra các đồ thị tri thức và áp dụng công nghệ web ngữ nghĩa chỉ có ý nghĩa trong những trường hợp cần phải thống nhất dữ liệu từ nhiều nguồn khác nhau hoặc thực hiện các suy diễn logic phức tạp.

> Việc chuyển từ bảng sang các mô hình dữ liệu đồ thị giúp cải thiện khả năng tìm kiếm và thống nhất luồng thông tin, nhưng không làm cho dữ liệu trở nên có ý nghĩa hơn với máy móc. Vấn đề không phải là việc nên hay không nên sử dụng công nghệ ngữ nghĩa, mà là nơi nào chúng thực sự mang lại lợi ích. Trước khi triển khai ontologies, ngữ nghĩa và cơ sở dữ liệu đồ thị trong công ty của bạn, hãy xác định các công ty nào đã thành công trong việc áp dụng các công nghệ này và nơi nào chúng không đáp ứng được mong đợi.

Mặc dù có những kỳ vọng tham vọng, công nghệ ngữ nghĩa vẫn chưa trở thành giải pháp toàn diện để cấu trúc dữ liệu trong ngành xây dựng. Trên thực tế, những công nghệ này không mang lại giải pháp phổ quát mà chỉ tạo ra những phức tạp mới, và những nỗ lực này lặp lại những tham vọng chưa được thực hiện của khái niệm internet ngữ nghĩa, nơi mà kỳ vọng vượt xa thực tế.

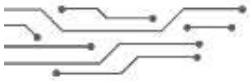


Hình 6.110: Hình học và thông tin trong các quy trình xây dựng: từ các hệ thống CAD và BIM phức tạp đến dữ liệu đơn giản hơn cho phân tích.

Nếu trong lĩnh vực CNTT, những thất bại của web ngữ nghĩa đã được bù đắp bằng sự xuất hiện của các công nghệ mới (dữ liệu lớn, IoT, học máy, AR/VR), thì ngành xây dựng không có những lý do như vậy.

Ngoài những vấn đề liên quan đến việc sử dụng các khái niệm để truyền đạt mối quan hệ dữ liệu giữa các yếu tố của dự án, vẫn còn một vấn đề cơ bản - chính khả năng tiếp cận dữ liệu. Trong ngành xây dựng, các hệ thống đóng vẫn chiếm ưu thế, gây khó khăn cho việc làm việc với dữ liệu, trao đổi thông tin và cải thiện hiệu quả quy trình.

Chính bản chất đóng của dữ liệu trở thành một trong những rào cản chính cản trở sự phát triển của các giải pháp kỹ thuật số trong xây dựng. Khác với ngành CNTT, nơi mà các định dạng dữ liệu mở và thống nhất đã trở thành tiêu chuẩn thì trong lĩnh vực CAD (BIM), mỗi phần mềm sử dụng một định dạng riêng, tạo ra các hệ sinh thái khép kín và hạn chế một cách nhân tạo người dùng.



CHƯƠNG 6.2. ĐỊNH DẠNG KHÓA CỦA CÁC DỰ ÁN VÀ VẤN ĐỀ TÍNH TƯƠI LIÊN TỤC

Dữ liệu đóng và năng suất giảm sút: ngõ cụt của ngành CAD (BIM)

Bản chất sở hữu của các hệ thống CAD đã dẫn đến việc mỗi chương trình có định dạng dữ liệu độc đáo của riêng mình, hoặc là đóng và không thể truy cập từ bên ngoài - RVT, PLN, DWG, NDW, NWD, SKP, hoặc có thể truy cập dưới dạng bán cấu trúc thông qua quá trình chuyển đổi khá phức tạp - JSON, XML (CPIXML), IFC, STEP và ifcXML, IfcJSON, BIMJSON, IfcSQL, CSV và nhiều định dạng khác.

Các định dạng dữ liệu khác nhau, trong đó có thể lưu trữ cùng một dữ liệu về các dự án giống nhau, không chỉ khác nhau về cấu trúc, mà còn bao gồm các phiên bản đánh dấu nội bộ khác nhau mà các nhà phát triển cần phải xem xét để đảm bảo khả năng tương thích của ứng dụng. Ví dụ, định dạng CAD của năm 2025 sẽ mở được trong phần mềm CAD của năm 2026, nhưng dự án đó sẽ không bao giờ mở được trong tất cả các phiên bản phần mềm CAD trước năm 2025.

> Không cung cấp quyền truy cập trực tiếp vào cơ sở dữ liệu, nhà cung cấp phần mềm trong ngành xây dựng thường tạo ra định dạng và công cụ duy nhất của riêng mình, mà chuyên gia (kỹ sư thiết kế hoặc quản lý dữ liệu) phải sử dụng để truy cập, nhập khẩu và xuất khẩu dữ liệu.

Hệ quả là các nhà cung cấp phần mềm CAD (BIM) cơ bản và các giải pháp liên quan (chẳng hạn như ERP/PMIS) liên tục tăng giá sử dụng sản phẩm, và người dùng thông thường phải trả "phí" cho mỗi giai đoạn chuyển dữ liệu qua các định dạng: cho kết nối, nhập, xuất và làm việc với dữ liệu mà người dùng tự tạo.

Chi phí truy cập dữ liệu trong kho đám mây từ các sản phẩm CAD- (BIM-) phổ biến vào năm 2025 sẽ đạt 1 đô la cho mỗi giao dịch, trong khi đăng ký cho các sản phẩm ERP xây dựng cho các công ty vừa sẽ đạt mức năm hoặc sáu con số mỗi năm.

> Bản chất của phần mềm xây dựng hiện đại không phải là tự động hóa hoặc nâng cao hiệu quả, mà là khả năng của các kỹ sư hiểu rõ về phần mềm chuyên ngành cụ thể ảnh hưởng đến chất lượng và chi phí xử lý dữ liệu của dự án xây dựng, cũng như lợi nhuận và khả năng sống sót lâu dài của các công ty thực hiện các dự án xây dựng.

Sự thiếu hụt quyền truy cập vào cơ sở dữ liệu của hệ thống CAD, mà được sử dụng trong hàng chục hệ thống và hàng trăm quy trình, và do đó, sự thiếu hụt giao tiếp chất lượng giữa các chuyên gia riêng lẻ đã đưa ngành xây dựng đến trạng thái trở thành một trong những lĩnh vực không hiệu quả nhất trong nền kinh tế xét về năng suất.

Trong 20 năm qua, việc áp dụng thiết kế CAD- (BIM-), sự xuất hiện của các hệ thống mới (ERP), các công nghệ và vật liệu xây dựng mới, năng suất toàn bộ ngành xây dựng đã giảm xuống 20%, trong khi năng suất chung của tất cả các lĩnh vực kinh tế không gặp phải vấn đề lớn về quyền truy cập vào cơ sở dữ liệu và các khái niệm BIM tương tự đã tăng trưởng 70% (96% trong ngành chế biến).-



Do sự cô lập và phức tạp của dữ liệu dự án, mà hàng chục bộ phận và hàng trăm quy trình trong ngành xây dựng phụ thuộc vào, tốc độ ra quyết định chậm hơn nhiều lần so với các lĩnh vực khác.

Tuy nhiên, vẫn có những ví dụ đơn lẻ về các phương pháp thay thế để tạo sự tương tác giữa các giải pháp CAD. Công ty xây dựng lớn nhất châu Âu với dự án SCOPE, bắt đầu từ năm 2018, cho thấy cách thoát khỏi logic cổ điển của các hệ thống CAD (BIM). Thay vì cố gắng tuân theo IFC hoặc dựa vào các lõi hình học độc quyền, các nhà phát triển SCOPE sử dụng API và SDK reverse engineering để lấy dữ liệu từ các phần mềm CAD khác nhau, chuyển đổi chúng sang các định dạng trung lập như OBJ hoặc CPIXML dựa trên lõi hình học mã nguồn mở OCCT, và sau đó áp dụng chúng trong hàng trăm quy trình kinh doanh của các công ty xây dựng và thiết kế. Tuy nhiên, mặc dù ý tưởng rất tiến bộ, các dự án tương tự vẫn phải đối mặt với những hạn chế và sự phức tạp của các lõi hình học miễn phí, và vẫn ở trong các hệ sinh thái khép kín của một công ty, tái tạo logic của các giải pháp một nhà cung cấp.

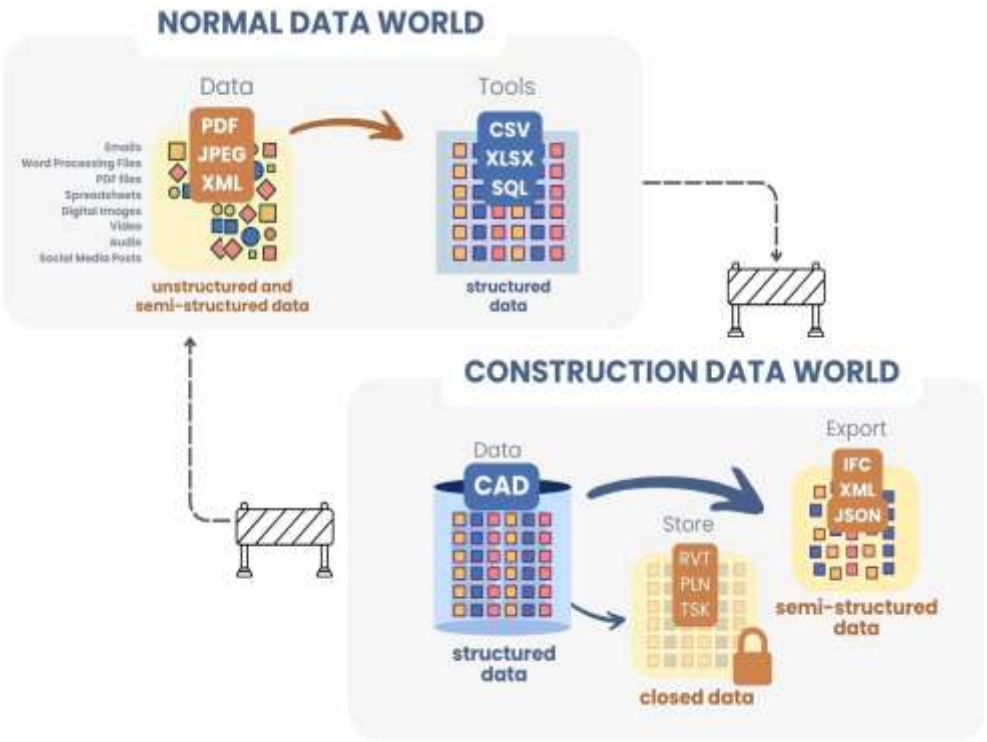
Do những hạn chế của các hệ thống khép kín và sự khác biệt trong các định dạng dữ liệu, cũng như thiếu các công cụ hiệu quả cho việc đồng nhất hóa, các công ty phải làm việc với các định dạng CAD đều gặp phải tình trạng tích lũy một lượng lớn dữ liệu với mức độ cấu trúc và tính khép kín khác nhau. Những dữ liệu này không được sử dụng đúng cách và rơi vào các kho lưu trữ, nơi chúng dần trở nên bị lãng quên và không còn được sử dụng.

> Dữ liệu thu được từ những nỗ lực đáng kể trong giai đoạn thiết kế, do sự phức tạp và tính khép kín của nó trở nên không có sẵn cho việc sử dụng tiếp theo.

Kết quả là trong suốt 30 năm qua, các nhà phát triển trong ngành xây dựng phải liên tục đối mặt với vấn đề tương tự: mỗi định dạng khép kín hoặc giải pháp độc quyền mới lại tạo ra nhu cầu tích hợp với các hệ thống CAD mở và khép kín hiện có. Những nỗ lực liên tục nhằm đảm bảo tính tương tác giữa các giải pháp CAD và BIM khác nhau lại làm phức tạp thêm hệ sinh thái dữ liệu, thay vì thúc đẩy sự đơn giản hóa và tiêu chuẩn hóa.

Huyền thoại về khả năng tương tác giữa các hệ thống CAD

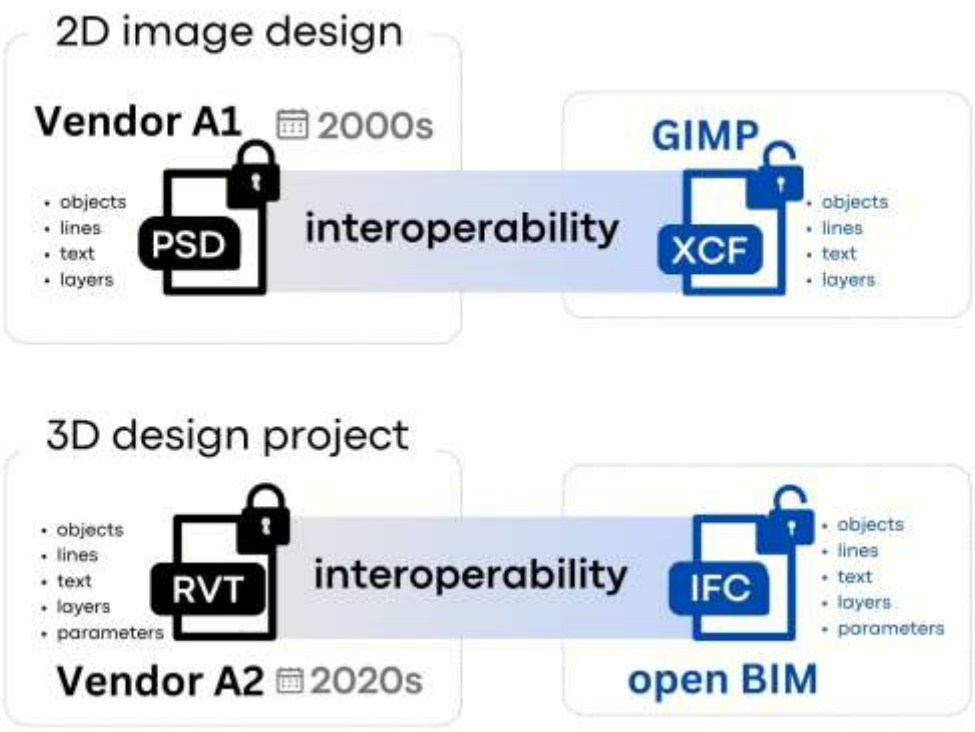
Nếu vào giữa những năm 1990, hướng phát triển chính của tính tương tác trong môi trường CAD là việc khai thác định dạng độc quyền DWG - kết thúc bằng chiến thắng của liên minh Open DWG và việc thực sự mở ra định dạng bản vẽ phổ biến nhất cho toàn ngành xây dựng, thì đến giữa những năm 2020, trọng tâm đã chuyển hướng. Ngành xây dựng đang chứng kiến xu hướng mới: nhiều nhóm phát triển tập trung vào việc tạo ra các "cây cầu" giữa các hệ thống CAD khép kín (closed BIM), định dạng IFC và các giải pháp mở (open BIM). Nền tảng của hầu hết các sáng kiến như vậy là việc sử dụng định dạng IFC và lõi hình học OCCT, tạo ra kết nối kỹ thuật giữa các nền tảng phân tán. Cách tiếp cận này được xem là một hướng đi triển vọng, có khả năng cải thiện đáng kể việc trao đổi dữ liệu và nâng cao tính tương thích của các công cụ phần mềm.



Trong khi các ngành khác làm việc với dữ liệu mở, ngành xây dựng buộc phải làm việc với các định dạng CAD (BIM) khép kín hoặc ít cấu trúc.

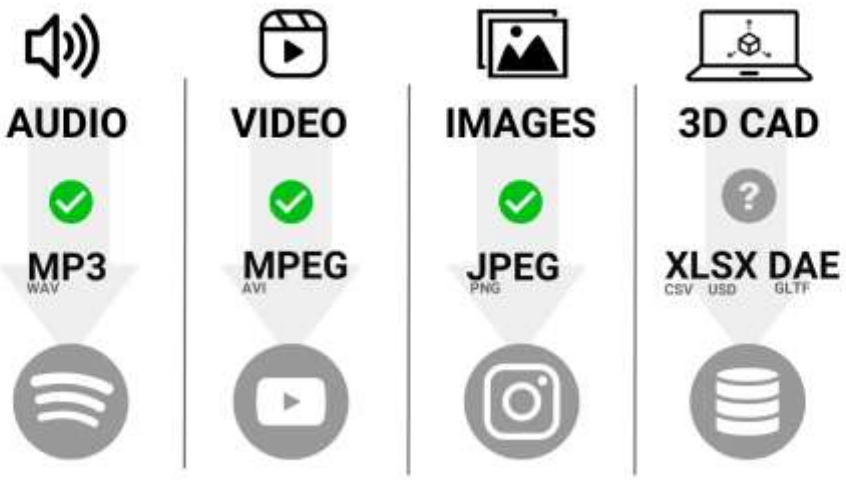
Cách tiếp cận này có những điểm tương đồng lịch sử. Vào những năm 2000, các nhà phát triển, với hy vọng vượt qua sự thống trị của nhà cung cấp phần mềm đồ họa lớn nhất (trên thế giới 2D), đã cố gắng tạo ra sự tích hợp liền mạch giữa giải pháp độc quyền của họ và sự thay thế mã nguồn mở miễn phí GIMP. Khi đó, giống như hiện nay trong lĩnh vực xây dựng, đây là nỗ lực kết nối các hệ thống đóng và mở, trong khi vẫn giữ nguyên các tham số phức tạp, các lớp và logic nội bộ của phần mềm.-

Tuy nhiên, người dùng thực sự tìm kiếm giải pháp đơn giản – dữ liệu phẳng, mở mà không có tính phức tạp thừa thãi của các lớp và tham số trong phần mềm (tương đương với lõi hình học trong CAD). Người dùng tìm kiếm các định dạng dữ liệu đơn giản và mở, không bị ràng buộc bởi logic quá mức. Trong đồ họa, các định dạng như JPEG, PNG và GIF đã trở thành tiêu chuẩn. Ngày nay, chúng được sử dụng rộng rãi trên các mạng xã hội, trang web và ứng dụng – dễ dàng được xử lý và diễn giải, bất kể nền tảng hay nhà cung cấp phần mềm.



Về mặt tương đồng, sự thay đổi định dạng dữ liệu trong xây dựng tương tự như nỗ lực kết hợp sản phẩm độc quyền phổ biến của nhà cung cấp và GIMP mã nguồn mở trong những năm 2000.

Kết quả là, trong ngành công nghiệp hình ảnh ngày nay, hầu như không ai sử dụng các định dạng đóng như PSD hoặc XCF cho ứng dụng, mạng xã hội như Facebook và Instagram hay làm nội dung trên các trang web. Thay vào đó, trong hầu hết các nhiệm vụ, các định dạng phẳng và mở như JPEG, PNG và GIF được sử dụng, cung cấp tính đơn giản và khả năng tương thích rộng rãi. Các định dạng mở như JPEG và PNG đã trở thành tiêu chuẩn cho việc trao đổi hình ảnh nhờ tính linh hoạt và sự hỗ trợ rộng rãi, giúp việc sử dụng chúng trên các nền tảng khác nhau trở nên dễ dàng hơn. Một sự chuyển đổi tương tự cũng đang diễn ra trong các định dạng trao đổi khác, chẳng hạn như video và âm thanh, nơi các định dạng phổ biến như MPEG và MP3 nổi bật nhờ hiệu quả nén và khả năng tương thích rộng rãi. Sự chuyển đổi này hướng tới tiêu chuẩn hóa đã đơn giản hóa việc trao đổi và phát lại nội dung và thông tin, làm cho chúng trở nên dễ tiếp cận với tất cả người dùng trên các nền tảng khác nhau.



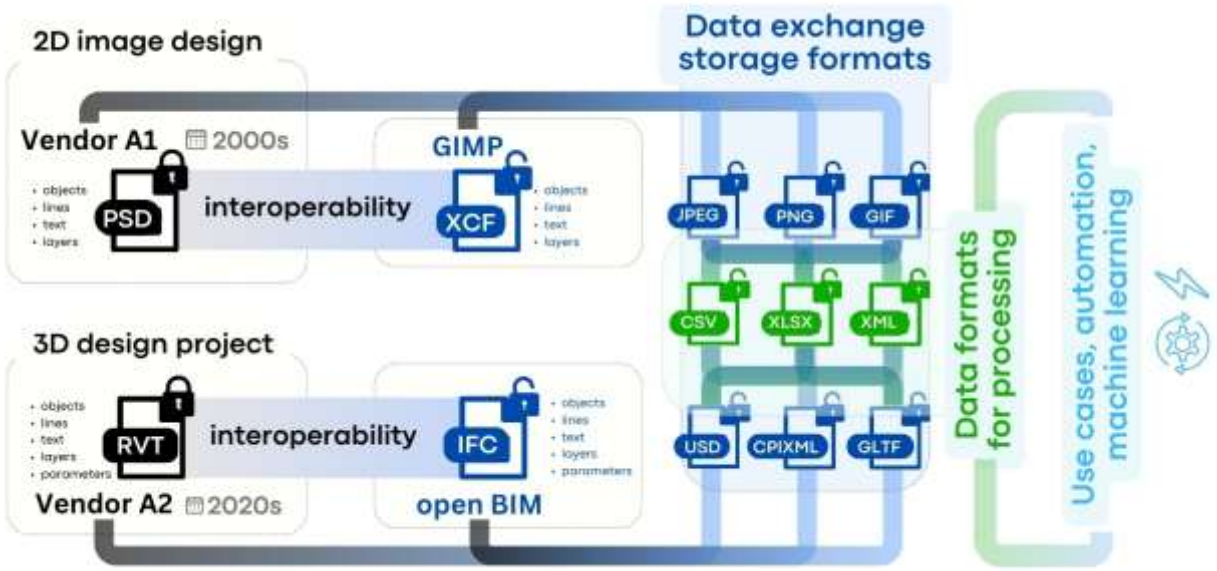
Các định dạng đơn giản không có các chức năng chỉnh sửa phức tạp đã trở nên phổ biến trong việc trao đổi và sử dụng dữ liệu.

Những quy trình tương tự cũng đang diễn ra trong mô hình 3D. Các định dạng đơn giản và mở như USD, OBJ, glTF, DAE, DXF, SQL và XLSX ngày càng được sử dụng trong các dự án để trao đổi dữ liệu ngoài môi trường CAD (BIM). Những định dạng này lưu trữ tất cả thông tin cần thiết, bao gồm hình học và siêu dữ liệu, mà không cần phải thao tác với cấu trúc phức tạp BREP, lõi hình học hay các phân loại nội bộ của các nhà cung cấp cụ thể. Các định dạng độc quyền như NWC, SVF, SVF2, CPIXML và CP2, được cung cấp bởi các nhà cung cấp phần mềm hàng đầu, cũng thực hiện các chức năng tương tự nhưng vẫn giữ tính chất đóng, trái ngược với các tiêu chuẩn mở.

> Đáng chú ý (và cần nhắc lại, như đã đề cập ở chương trước), rằng ý tưởng từ chối các định dạng trung gian và định dạng tham số như IGES, STEP và IFC đã được nhà cung cấp CAD hàng đầu hỗ trợ từ năm 2000, người đã tạo ra Whitepaper BIM và đăng ký định dạng IFC vào năm 1994. Trong Whitepaper năm 2000 "Thiết kế và sản xuất tích hợp" [65], nhà cung cấp CAD nhấn mạnh tầm quan trọng của việc truy cập dữ liệu CAD gốc trong môi trường phần mềm, mà không cần sử dụng các bộ chuyển đổi trung gian và định dạng tham số, để duy trì độ đầy đủ và chính xác của thông tin.

Ngành xây dựng vẫn còn phải đạt được đồng thuận về các công cụ truy cập vào cơ sở dữ liệu CAD, hoặc việc đảo ngược kỹ thuật bắt buộc, hoặc việc chấp nhận một định dạng dữ liệu chung đơn giản hóa để sử dụng ngoài các nền tảng CAD (BIM). Ví dụ, nhiều công ty lớn ở Trung Âu và các khu vực nói tiếng Đức làm việc trong lĩnh vực xây dựng đang sử dụng định dạng CPIXML trong hệ thống ERP của họ [121]. Định dạng độc quyền này, là một dạng của XML, tổ chức dữ liệu dự án CAD (BIM), bao gồm dữ liệu hình học và siêu dữ liệu, thành một cấu trúc tổ chức đơn giản hóa duy nhất. Các công ty xây dựng lớn cũng đang tạo ra các định dạng và hệ thống độc quyền mới, như trong dự án SCOPE mà chúng tôi đã nói đến ở chương trước.

Logic kín của các định dạng CAD tham số hoặc các tệp IFC (STEP) tham số phức tạp trong hầu hết các quy trình kinh doanh trở nên dư thừa. Người dùng tìm kiếm các định dạng đơn giản và phẳng, như USD, CPIXML, XML&OBJ, DXF, glTF, SQLlite, DAE&XLSX, chứa tất cả thông tin cần thiết về các yếu tố, nhưng không bị nặng nề bởi logic dư thừa của việc xây dựng hình học BREP, sự phụ thuộc vào các lõi hình học và phân loại nội bộ của các sản phẩm CAD và BIM cụ thể (Hình 6.25).-



Hình 6.25 Đối với hầu hết các trường hợp sử dụng, người dùng chọn các định dạng đơn giản nhất, không phụ thuộc vào phần mềm của nhà cung cấp.

Sự xuất hiện của các định dạng hình ảnh phẳng, như JPEG, PNG và GIF, không bị ràng buộc bởi logic dư thừa của các động cơ nội bộ của nhà cung cấp, đã góp phần phát triển hàng ngàn giải pháp tương thích để xử lý và sử dụng đồ họa. Điều này đã dẫn đến sự ra đời của nhiều ứng dụng đa dạng: từ các công cụ chỉnh sửa và lọc đến các mạng xã hội như Instagram, Snapchat và Canva, nơi dữ liệu đơn giản hóa này có thể được sử dụng mà không bị ràng buộc vào nhà phát triển phần mềm cụ thể.

> Việc tiêu chuẩn hóa và đơn giản hóa các định dạng CAD cho thiết kế sẽ thúc đẩy sự xuất hiện của nhiều công cụ độc lập và tiện lợi mới cho việc làm việc với các dự án xây dựng.

Việc từ bỏ logic phức tạp của các ứng dụng nhà cung cấp, gắn liền với các lõi hình học kín, và chuyển sang các định dạng mở linh hoạt dựa trên các thư viện các yếu tố đơn giản hóa, tạo ra các điều kiện cho việc làm việc với dữ liệu linh hoạt hơn, minh bạch hơn và hiệu quả hơn. Điều này cũng mở ra quyền truy cập thông tin cho tất cả các bên tham gia trong quá trình xây dựng — từ các nhà thiết kế đến các nhà thầu và các dịch vụ khai thác.

Tuy nhiên, với xác suất cao, trong những năm tới, các nhà cung cấp CAD sẽ thực hiện các nỗ lực để thay đổi lại trọng tâm trong cuộc thảo luận về tính tương tác và truy cập vào cơ sở dữ liệu CAD. Nội dung sẽ xoay quanh những khái niệm "mới" — như dữ liệu phân mảnh, đồ thị thông minh, "mô hình liên bang", và bản sao kỹ thuật số trong các kho lưu trữ đám mây — cũng như việc tạo ra các liên minh và tiêu chuẩn trong ngành, tiếp tục con đường BIM và open BIM. Mặc dù ngôn ngữ hấp dẫn, những sáng kiến như vậy có thể trở thành công cụ để giữ chân người dùng trong các hệ sinh thái sở hữu. Một ví dụ là việc thúc đẩy mạnh mẽ từ năm 2023 định dạng USD (Mô tả Cảnh toàn cầu) như một "tiêu chuẩn mới" cho việc tương tác giữa các nền tảng trong CAD (BIM).

Chuyển sang USD và dữ liệu phân cấp

Sự xuất hiện của liên minh AOUSD vào năm 2023 đánh dấu một bước ngoặt quan trọng trong ngành xây dựng. Chúng ta đang chứng kiến sự khởi đầu của một thực tế mới do các nhà cung cấp CAD hình thành, thông qua một số thay đổi đáng kể trong cách làm việc với dữ liệu xây dựng. Thay đổi quan trọng đầu tiên liên quan đến nhận thức về dữ liệu CAD. Các chuyên gia tham gia vào các giai đoạn đầu của thiết kế khái niệm ngày càng nhận thức được rằng việc tạo ra dự án trong môi trường CAD chỉ là một điểm khởi đầu. Dữ liệu hình thành trong quá trình thiết kế dần trở thành cơ sở cho phân tích, vận hành và quản lý các đối tượng. Điều này có nghĩa là chúng cần phải được truy cập và có thể sử dụng trong các hệ thống vượt ra ngoài các công cụ CAD truyền thống.

Song song với điều này, một cuộc cách mạng đang xảy ra trong cách tiếp cận của các nhà phát triển hàng đầu. Nhà cung cấp CAD hàng đầu trong ngành, người đã tạo ra khái niệm BIM và định dạng IFC, đang thực hiện một bước ngoặt bất ngờ trong chiến lược của mình. Từ năm 2023, công ty sẽ từ bỏ việc lưu trữ dữ liệu trong các tệp riêng biệt, thay vào đó tập trung vào việc làm việc với dữ liệu phân mảnh (đã được chuẩn hóa và cấu trúc) với việc chuyển sang cách tiếp cận dựa trên dữ liệu.

> Các nhà cung cấp đang theo đuổi các xu hướng lịch sử của các ngành khác: hầu hết người dùng không cần các định dạng CAD đóng (giống như PSD) hoặc các tệp tham số phức tạp IFC (giống như GIMP với logic lớp). Họ cần những hình ảnh đơn giản của các đối tượng có thể được sử dụng trong CAFM (Instagram xây dựng), ERP (Facebook) và trong hàng ngàn quy trình khác, đầy ắp bảng tính Excel và tài liệu PDF.

Các xu hướng hiện tại trong ngành xây dựng có khả năng tạo điều kiện cho việc từng bước từ bỏ các định dạng tham số và phức tạp để ủng hộ các định dạng phổ biến và độc lập hơn như USD, GLTF, DAE, OBJ (với siêu thông tin cả bên trong các định dạng hỗn hợp lẫn trong các định dạng cấu trúc hoặc yếu cấu trúc riêng biệt). Những nhà lãnh đạo lịch sử, bao gồm các công ty thiết kế lớn nhất, những người đã từng tích cực thúc đẩy IFC vào giữa những năm 1990, hôm nay đang công khai ủng hộ định dạng USD mới, nhấn mạnh sự đơn giản và tính linh hoạt của nó. Việc áp dụng rộng rãi USD vào các sản phẩm, khả năng tương thích với GLTF và tích hợp chủ động vào các công cụ như Blender, Unreal Engine và Omniverse cho thấy tiềm năng bắt đầu một cách tiếp cận mới trong công việc với dữ liệu. Song song với sự phổ biến của các giải pháp địa phương, như định dạng USD phẳng ở châu Âu — CPIXML, được sử dụng trong các ERP phổ biến ở châu Âu có thể làm tăng vị thế của USD ở Trung Âu. Các tổ chức phát triển định dạng IFC đã điều chỉnh chiến lược của họ theo USD, điều này chỉ xác nhận sự chuyển biến không thể tránh khỏi.-

Technical Specifications			Comparison / Notes
File Structure	Monolithic file	Uses ECS and linked data	IFC stores all data in one file; USD uses Entity-Component-System and linked data for modularity and flexibility
Data Structure	Complex semantics, parametric geometry	Flat format, geometry in MESH, data in JSON	IFC is complex and parametric; USD is simpler and uses flat data
Geometry	Parametric, dependent on BREP	Flat, MESH (triangular meshes)	IFC uses parametrics; USD uses meshes for simplified processing.
Properties	Complex structure of semantic descriptions	Properties in JSON, easy access	Properties in USD are easier to use thanks to JSON
Export/Import	Complex implementation, dependent on third-party SDKs	Easy integration, wide support	USD integrates more easily and is supported in many products
Format Complexity	High, requires deep understanding	Low, optimized for convenience	The time required to understand the structure of the file and the information stored in it
Performance	Can be slow when processing large models	High performance in visualization and processing	USD is optimized for speed and efficiency. Simulations, machine learning, AI, smart cities will be held in the Nvidia Omniverse
Integration with 3D Engines	Limited	High, designed for graphics engines	USD excels with native support for real-time visualization platforms
Support outside CAD Software	BlenderBIM, IfcOpenShell	Unreal Engine, Unity, Blender, Omniverse	USD is widely supported in graphics tools
Cloud Technology Support	Limited	Well-suited for cloud services and online collaboration	USD is optimized for cloud solutions
Ease of Integration into Web Applications	Difficult to integrate due to size and complexity	Easy to integrate, supports modern web technologies	USD is preferable for web applications
Change Management	Versions through separate files	Versioning built into the format core	IFC handles changes via separate files, while USD embeds versioning directly into its structure
Collaboration Support	Supports data exchange between project participants	Designed for collaborative work on complex scenes	USD provides efficient collaboration through layers and variations
Learnability	Steep learning curve due to complexity	Easier to master thanks to a clear structure	USD is easier to learn and implement

Hình. 6.26 So sánh thông số kỹ thuật của định dạng IFC và USD.

> Trong bối cảnh này, USD có khả năng trở thành tiêu chuẩn de facto, hứa hẹn vượt qua nhiều hạn chế hiện tại, chủ yếu liên quan đến sự phức tạp của các định dạng CAD- (BIM-) hiện có và sự phụ thuộc của việc giải thích chúng vào các lõi hình học.

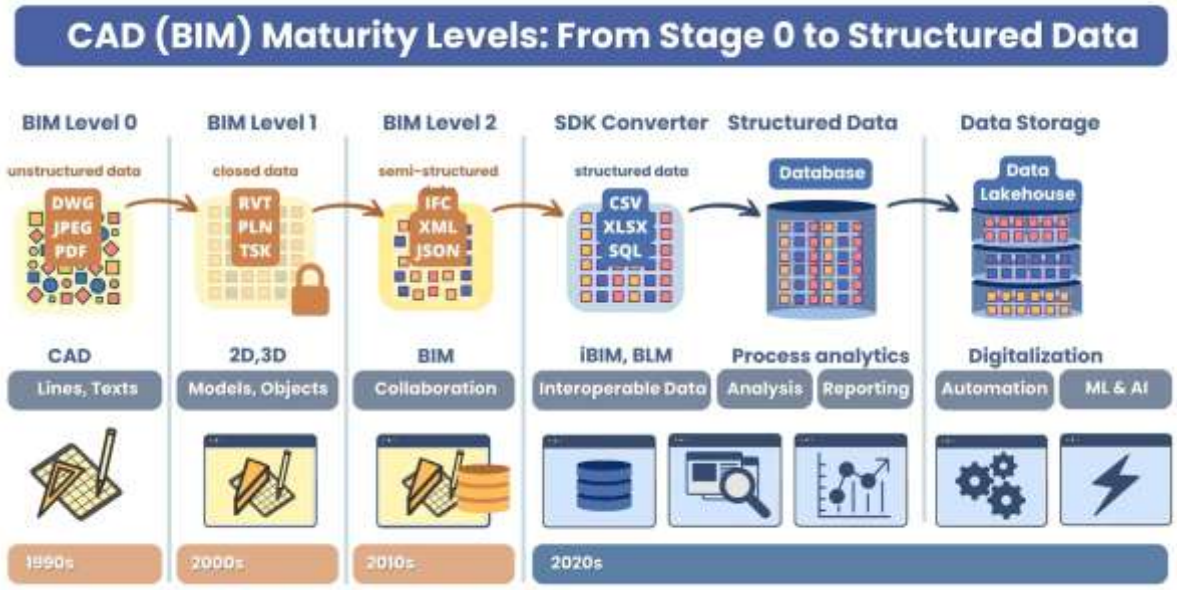
Thay vì các định dạng CAD phức tạp và tham số, các định dạng dữ liệu đơn giản như USD, gLTF, DAE, OBJ với siêu thông tin về các phần tử trong CSV, XLSX, JSON, XML sẽ chiếm ưu thế trong ngành xây dựng nhờ vào sự đơn giản và linh hoạt của chúng.

Những thay đổi hiện tại trong ngành xây dựng thoạt nhìn có vẻ như là một bước đột phá công nghệ, liên quan đến việc chuyển từ IFC lỗi thời sang USD hiện đại hơn. Tuy nhiên, cần lưu ý rằng vào năm 2000, cùng một nhà cung cấp CAD đã phát triển IFC cũng đã chỉ ra các vấn đề của nó và cần thiết phải truy cập vào cơ sở dữ liệu [65], và giờ họ lại đang tích cực thúc đẩy chuyển đổi sang tiêu chuẩn mới - USD.

> Đằng sau lớp "dữ liệu mở" USD và các khái niệm "mới" về quản lý dữ liệu phân cấp thông qua các ứng dụng đám mây mà các nhà cung cấp CAD đang bắt đầu quảng bá, có thể ẩn chứa ý định của các nhà cung cấp trong việc độc quyền quản lý dữ liệu dự án, nơi mà người dùng rơi vào tình thế mà sự lựa chọn định dạng nhiều hơn liên quan đến lợi ích doanh nghiệp hơn là nhu cầu thực tế.

Phân tích các sự kiện chính [93] cho thấy rằng mục tiêu chính của những thay đổi này - không chỉ là sự tiện lợi của người dùng, mà trước hết là duy trì kiểm soát các hệ sinh thái và dòng dữ liệu theo lợi ích của các nhà cung cấp, những người trong suốt 40 năm vẫn chưa thể cung cấp sự truy cập vào cơ sở dữ liệu CAD.

Có lẽ bây giờ là lúc các công ty nên từ bỏ việc chờ đợi các khái niệm mới từ các nhà cung cấp phần mềm và tập trung vào việc phát triển độc lập theo hướng tập trung vào dữ liệu. Một khi được giải phóng khỏi các vấn đề về truy cập dữ liệu, ngành sẽ có thể tự chuyển đổi sang các công cụ hiện đại, miễn phí và thuận tiện để làm việc và phân tích dữ liệu mà không cần các khái niệm mới bị áp đặt.



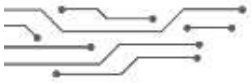
Hình. 6.27 Mức độ trưởng thành của CAD (BIM): từ dữ liệu phi cấu trúc đến dữ liệu có cấu trúc và kho lưu trữ.

Truy cập vào cơ sở dữ liệu, dữ liệu mở và định dạng sẽ không tránh khỏi trở thành tiêu chuẩn trong ngành xây dựng, bất chấp nỗ lực của các nhà cung cấp trong việc làm chậm quá trình này - chỉ là vấn đề thời gian (Hình. 6.27). Tốc độ chuyển đổi này có thể tăng nhanh nếu càng nhiều chuyên gia làm quen với các định dạng mở, các công cụ làm việc với cơ sở dữ liệu và SDK ngược đang có sẵn cho phép tổ chức truy cập trực tiếp vào dữ liệu của các hệ thống CAD [92].-

> Tương lai thuộc về dữ liệu mở, thống nhất và có thể phân tích. Để tránh phụ thuộc vào các giải pháp của nhà cung cấp và không trở thành nạn nhân của các hệ sinh thái kín, các công ty xây dựng và thiết kế sẽ sớm hoặc muộn phải đặt cược vào sự mở cửa và độc lập, lựa chọn các định dạng và giải pháp mà cung cấp quyền kiểm soát hoàn toàn đối với dữ liệu.

Dữ liệu được tạo ra ngày hôm nay trong ngành xây dựng sẽ trở thành tài nguyên chính cho việc đưa ra quyết định kinh doanh trong tương lai. Chúng sẽ đóng vai trò như "nhiên liệu" chiến lược, nuôi dưỡng sự phát triển và hiệu quả của các công ty xây dựng. Tương lai của ngành xây dựng nằm ở khả năng làm việc với dữ liệu, chứ không phải là sự lựa chọn các định dạng hay mô hình dữ liệu.

Để hiểu sự khác biệt giữa các định dạng mở USD, glTF, DAE, OBJ và các định dạng CAD tham số độc quyền, điều quan trọng là xem xét một trong những yếu tố dữ liệu phức tạp và then chốt nhất trong việc trực quan hóa và tính toán dự án - hình học và các quy trình hình thành của nó. Và để hiểu cách dữ liệu hình học trở thành cơ sở cho phân tích và tính toán trong xây dựng, cần phải nghiên cứu sâu hơn các cơ chế tạo hình học, chuyển đổi và lưu trữ của nó.



CHƯƠNG 6.3. **HÌNH HỌC TRONG XÂY DỰNG: TỪ ĐƯỜNG THẲNG ĐẾN KHỐI LƯỢNG**

Khi các đường thẳng trở thành tiền hoặc tại sao hình học lại quan trọng đối với các nhà thầu

Hình học trong xây dựng không chỉ là trực quan hóa, mà còn là cơ sở cho các phép tính định lượng chính xác. Trong mô hình dự án, hình học bổ sung danh sách các tham số của các yếu tố (Hình. 3.116) với các đặc điểm thể tích quan trọng nhất, chẳng hạn như chiều dài, diện tích và thể tích. Những giá trị của các tham số thể tích này được tính toán tự động bằng cách sử dụng các lõi hình học và là điểm khởi đầu cho các ước toán, biểu đồ và mô hình tài nguyên. Như đã thảo luận trong phần thứ năm của cuốn sách và trong chương "Tính toán chi phí và ước toán dự án xây dựng", chính các tham số thể tích của nhóm đối tượng từ các mô hình CAD tạo thành cơ sở cho các hệ thống ERP và PMIS hiện đại. Hình học đóng vai trò nền tảng không chỉ trong giai đoạn thiết kế, mà còn trong việc quản lý thực hiện dự án, kiểm soát thời gian, lập ngân sách và khai thác. Giống như hàng ngàn năm trước khi xây dựng các kim tự tháp ở Ai Cập, độ chính xác của dự án phụ thuộc vào đơn vị đo chiều dài như khuỷu tay và cubit, ngày nay độ chính xác trong việc diễn giải hình học trong các chương trình CAD trực tiếp ảnh hưởng đến kết quả: từ ngân sách và thời gian - đến việc lựa chọn nhà thầu và logistics cung ứng.

> Trong bối cảnh cạnh tranh cao và ngân sách hạn chế, độ chính xác của các phép tính thể tích, trực tiếp phụ thuộc vào hình học, trở thành yếu tố sống còn. Các hệ thống ERP hiện đại phụ thuộc trực tiếp vào các đặc điểm thể tích chính xác, được lấy từ các mô hình CAD và BIM. Đó chính là lý do tại sao mô tả hình học chính xác của các yếu tố không chỉ là trực quan hóa mà còn là công cụ quan trọng để quản lý chi phí và thời gian xây dựng.

Về mặt lịch sử, hình học đã là ngôn ngữ chính trong sự tương tác kỹ thuật. Từ các đường nét trên papyrus đến các mô hình kỹ thuật số - bản vẽ và các biểu diễn hình học đã phục vụ như một phương tiện truyền tải thông tin giữa các nhà thiết kế, giám sát thi công và lập dự toán. Trước khi có máy tính, các phép tính được thực hiện thủ công, bằng thước kẻ và thước đo. Ngày nay, nhiệm vụ này đã được tự động hóa nhờ vào mô hình hóa thể tích: các lõi hình học của các chương trình CAD chuyển đổi các đường thẳng và điểm thành các khối ba chiều, từ đó tự động rút ra tất cả các đặc điểm cần thiết.

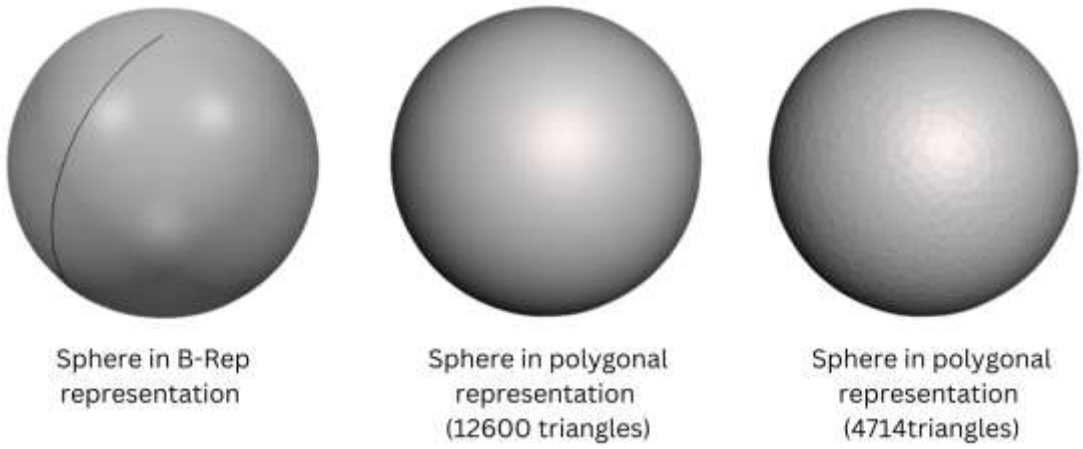
Khi làm việc trong các chương trình CAD, việc tạo ra các yếu tố hình học cho các phép tính diễn ra thông qua giao diện người dùng của các chương trình CAD (BIM). Để chuyển đổi các điểm và đường thành các khối thể tích, một lõi hình học được sử dụng, thực hiện nhiệm vụ then chốt - chuyển đổi hình học thành các mô hình thể tích, từ đó sau khi xấp xỉ tự động tính toán các đặc điểm thể tích của yếu tố.

Từ đường thẳng đến thể tích: cách mà diện tích và thể tích trở thành dữ liệu

Trong thực tiễn kỹ thuật, các thể tích và diện tích được tính toán dựa trên các bề mặt hình học, được mô tả một cách phân tích hoặc thông qua các mô hình tham số, chẳng hạn như NURBS (B-spline hợp lý không đồng nhất) trong khuôn khổ BREP (Biểu diễn ranh giới).

NURBS (Non-Uniform Rational B-Splines) — là một phương pháp toán học để mô tả các đường cong và bề mặt, trong khi BREP — là cấu trúc để mô tả hình học 3D hoàn chỉnh của một đối tượng, bao gồm các đường biên có thể được định nghĩa bằng cách sử dụng NURBS.

Mặc dù BREP và NURBS rất chính xác, nhưng chúng yêu cầu tài nguyên tính toán mạnh mẽ và các thuật toán phức tạp. Tuy nhiên, các phép tính trực tiếp dựa trên những mô tả toán học chính xác như vậy thường rất tốn kém về mặt tính toán, vì vậy trong thực tế, hầu như luôn sử dụng tesselation — chuyển đổi bề mặt thành lưới tam giác, giúp đơn giản hóa các phép tính tiếp theo. Tesselation là quá trình phân chia một bề mặt phức tạp thành các tam giác hoặc đa giác. Trong môi trường CAD/CAE, phương pháp này được sử dụng để hình dung, tính toán thể tích, tìm kiếm va chạm, xuất sang các định dạng như MESH và phân tích va chạm. Một ví dụ từ thiên nhiên là tổ ong, nơi hình dạng phức tạp được phân chia thành một lưới đều (Hình 6.31). -



Hình 6.31 Một hình cầu được mô tả theo dạng tham số BREP và biểu diễn đa giác với số lượng tam giác khác nhau.

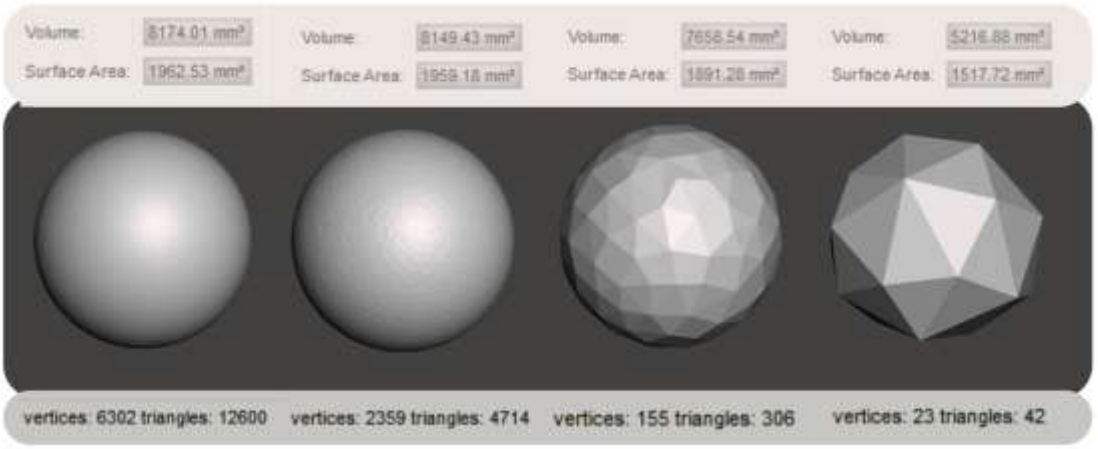
BREP (NURBS) được sử dụng trong CAD không phải là mô hình cơ bản của hình học. Phương pháp này được tạo ra như một công cụ tiện lợi để đại diện cho các đường tròn và splines hữu tỷ, cũng như để tối thiểu hóa việc lưu trữ dữ liệu về hình học. Tuy nhiên, nó có giới hạn — chẳng hạn như không thể mô tả chính xác một hàm sin, nằm trong cơ sở của các đường xoắn và bề mặt, cũng như yêu cầu sử dụng các lõi hình học phức tạp.

Lưới tam giác và tesselation của các hình dạng tham số, ngược lại, khác biệt ở sự đơn giản, việc sử dụng bộ nhớ hiệu quả và khả năng xử lý khối lượng dữ liệu lớn (Hình 6.32). Những ưu điểm này cho phép không cần thiết phải có các lõi hình học phức tạp và đắt đỏ, cùng với hàng chục triệu dòng mã được nhúng bên trong chúng, khi tính toán các hình dạng hình học.

Trong hầu hết các trường hợp xây dựng, không quan trọng kích thước đặc điểm khối lượng được xác định như thế nào — thông qua các mô hình tham số (BREP, IFC) hay thông qua các đa giác (USD, glTF, DAE, OBJ). Hình học vẫn là một dạng xấp xỉ: cho dù thông qua NURBS hay MESH, đó luôn luôn là một mô tả gần đúng về hình dạng.

Hình học được xác định dưới dạng đa giác hoặc BREP (NURBS) vẫn ở mức độ nào đó chỉ là một cách xấp xỉ với mô tả gần đúng về hình dạng liên tục. Giống như các tích phân Fresnel không có biểu thức phân tích chính xác, việc rời rạc hóa hình học thông qua các đa giác hoặc NURBS luôn là một sự xấp xỉ, giống như lưới tam giác MESH.

Hình học tham số ở định dạng BREP chủ yếu cần thiết nơi mà kích thước dữ liệu tối thiểu quan trọng và có khả năng sử dụng các lõi hình học tốn kém và phức tạp cho việc xử lý và hiển thị. Thường thì điều này đặc trưng cho các nhà phát triển phần mềm CAD, những người sử dụng các lõi hình học của các nhà cung cấp MCAD trong sản phẩm của họ. Tuy vậy, thậm chí bên trong các chương trình này, các mô hình BREP trong quá trình tesselation để hình dung và tính toán thường được chuyển đổi thành các tam giác (tương tự như cách mà các tệp PSD được đơn giản hóa thành JPEG).



Hình 6.32 Sự khác biệt về các đặc điểm khối lượng của các hình với số lượng đa giác khác nhau.

Lưới polygonal MESH, giống như BREP tham số, có những ưu điểm và hạn chế riêng, nhưng mục tiêu của chúng giống nhau - mô tả hình học dựa trên nhu cầu của người dùng. Cuối cùng, độ chính xác của mô hình hình học không chỉ phụ thuộc vào phương pháp đại diện của nó, mà còn vào yêu cầu cho nhiệm vụ cụ thể.

> Trong hầu hết các nhiệm vụ xây dựng, nhu cầu về hình học tham số và những lõi hình học phức tạp có thể là thừa thãi.

Trong mỗi nhiệm vụ cụ thể về tự động hóa tính toán, cần xem xét liệu tầm quan trọng của hình học tham số có bị phóng đại bởi các nhà phát triển CAD, những người quan tâm đến việc quảng bá và bán các sản phẩm phần mềm riêng của họ hay không.

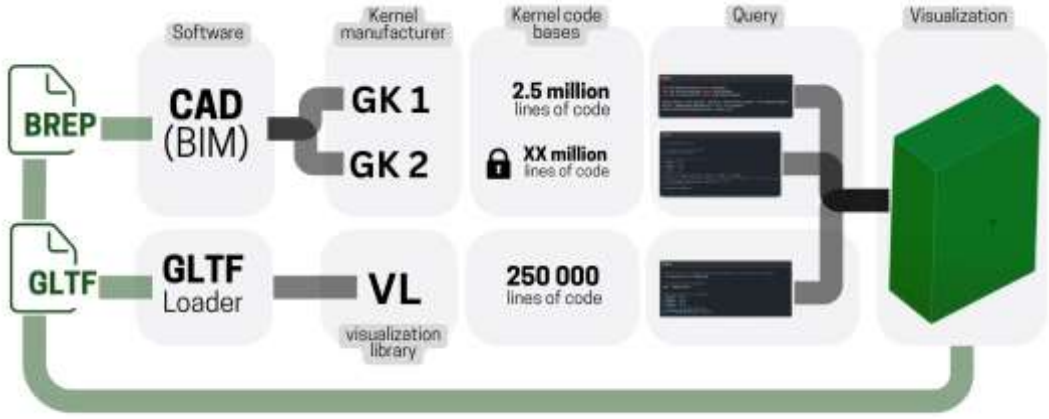
Chuyển sang MESH, USD và đa giác: sử dụng phân mảnh cho hình học

Trong ngành xây dựng, khi làm việc theo chu trình, phát triển hệ thống, cơ sở dữ liệu hoặc tự động hóa các quy trình làm việc với thông tin dự án và hình học của các yếu tố, điều quan trọng là hướng tới sự độc lập khỏi các trình biên tập CAD cụ thể và các lõi hình học.

> Trong định dạng trao đổi sẽ được sử dụng cả ở phòng tính toán và trên công trường - không nên ngầm định bất kỳ chương trình CAD (BIM) cụ thể nào. Thông tin hình học cần được trình bày dưới định dạng trực tiếp thông qua tesselation, không liên kết với lõi hình học hoặc kiến trúc của CAD.

Hình học tham số từ CAD có thể được xem như một nguồn trung gian, nhưng không phải là nền tảng của định dạng phổ quát. Hầu hết các mô tả tham số (bao gồm BREP và NURBS) đều được chuyển đổi thành lưới polygonal MESH cho việc xử lý tiếp theo. Nếu kết quả - giống nhau (tesselation và polygon), và quy trình - đơn giản hơn, thì lựa chọn là hiển nhiên. Điều này tương tự như việc lựa chọn giữa ontology đồ thị và bảng có cấu trúc (như đã đề cập trong phần bốn): độ phức tạp thừa thãi hiếm khi biện minh được.--

Các định dạng mở như: OBJ, STL, glTF, SVF, CPIXML, USD và DAE, sử dụng cấu trúc lưới tam giác phổ quát, điều này mang lại cho chúng những lợi thế đáng kể. Những định dạng này có tính tương thích tuyệt vời - chúng dễ đọc và hình dung bằng cách sử dụng các thư viện mở sẵn có mà không cần đến các lõi hình học chuyên biệt phức tạp, chứa hàng triệu dòng mã. Những định dạng hình học phổ quát này được sử dụng trong nhiều lĩnh vực - từ những công cụ thiết kế bếp tương đối đơn giản ở IKEA™ đến các hệ thống hình dung đối tượng phức tạp trong phim và ứng dụng VR. Một lợi thế quan trọng là sự hiện diện của nhiều thư viện miễn phí và mở để làm việc với những định dạng này, sẵn có cho hầu hết các nền tảng và ngôn ngữ lập trình.-



Sự trình bày hình học giống nhau đạt được thông qua việc sử dụng định dạng tham số và lõi hình học, hoặc thông qua định dạng tam giác và thư viện hình dung mở.

Cũng như người dùng, các nhà cung cấp CAD cũng gặp phải vấn đề trong việc diễn giải các định dạng CAD tham số của người khác hoặc IFC mở do sự khác biệt trong các lõi hình học. Trên thực tế, tất cả các nhà cung cấp CAD, không có ngoại lệ, đều sử dụng SDK đảo ngược kỹ thuật để truyền dữ liệu giữa các hệ thống, và không ai trong số họ dựa vào các định dạng như IFC hoặc USD cho mục đích tương tác.

> Thay vì sử dụng các khái niệm do các liên minh nhà cung cấp CAD thúc đẩy mà chính họ không sử dụng, các nhà phát triển và người dùng giải pháp CAD hiệu quả hơn nên tập trung vào việc hiểu rõ lợi ích của từng phương pháp trong từng ngữ cảnh cụ thể và chọn loại hình hình học phù hợp tùy thuộc vào trường hợp sử dụng. Việc lựa chọn giữa các biểu diễn hình học khác nhau là một sự đánh đổi giữa độ chính xác, hiệu quả tính toán và nhu cầu thực tiễn của từng nhiệm vụ cụ thể.

Sự phức tạp liên quan đến việc sử dụng các lõi hình học, mà các nhà cung cấp lớn thường áp đặt cho ngành xây dựng khi xử lý dữ liệu dự án, thường trở nên dư thừa. Định dạng USD, dựa trên hình học MESH, có thể trở thành một dạng "hộp Pandora" cho ngành công nghiệp, mở ra cho các nhà phát triển những cơ hội mới trong tổ chức việc trao đổi dữ liệu - ngoài các khuôn khổ IFC và cấu trúc BREP tham số đặc trưng cho các nhà cung cấp CAD.

Khi làm quen hơn với cấu trúc USD, DAE, gLTF, OBJ, và các định dạng khác, rõ ràng có những định dạng mở đơn giản hơn cho phép tổ chức hiệu quả việc truyền tải và sử dụng thông tin hình học mà không cần dựa vào các tham số phức tạp và các lõi hình học đóng kín. Cách tiếp cận này không chỉ giảm ngưỡng kỹ thuật cho các nhà phát triển mà còn thúc đẩy sự phát triển của các giải pháp linh hoạt, mở rộng và thực sự mở cho xây dựng kỹ thuật số.

LOD, LOI, LOMD - phân loại chi tiết độc đáo trong CAD (BIM)

Ngoài các định dạng hình học biểu diễn, trong một thế giới mà các ngành khác nhau sử dụng các mức độ chi tiết và độ sâu dữ liệu khác nhau, các phương pháp CAD (BIM) cung cấp các hệ thống phân loại độc đáo của riêng mình, tạo cấu trúc cho việc tiếp cận thông tin trong mô hình tòa nhà.

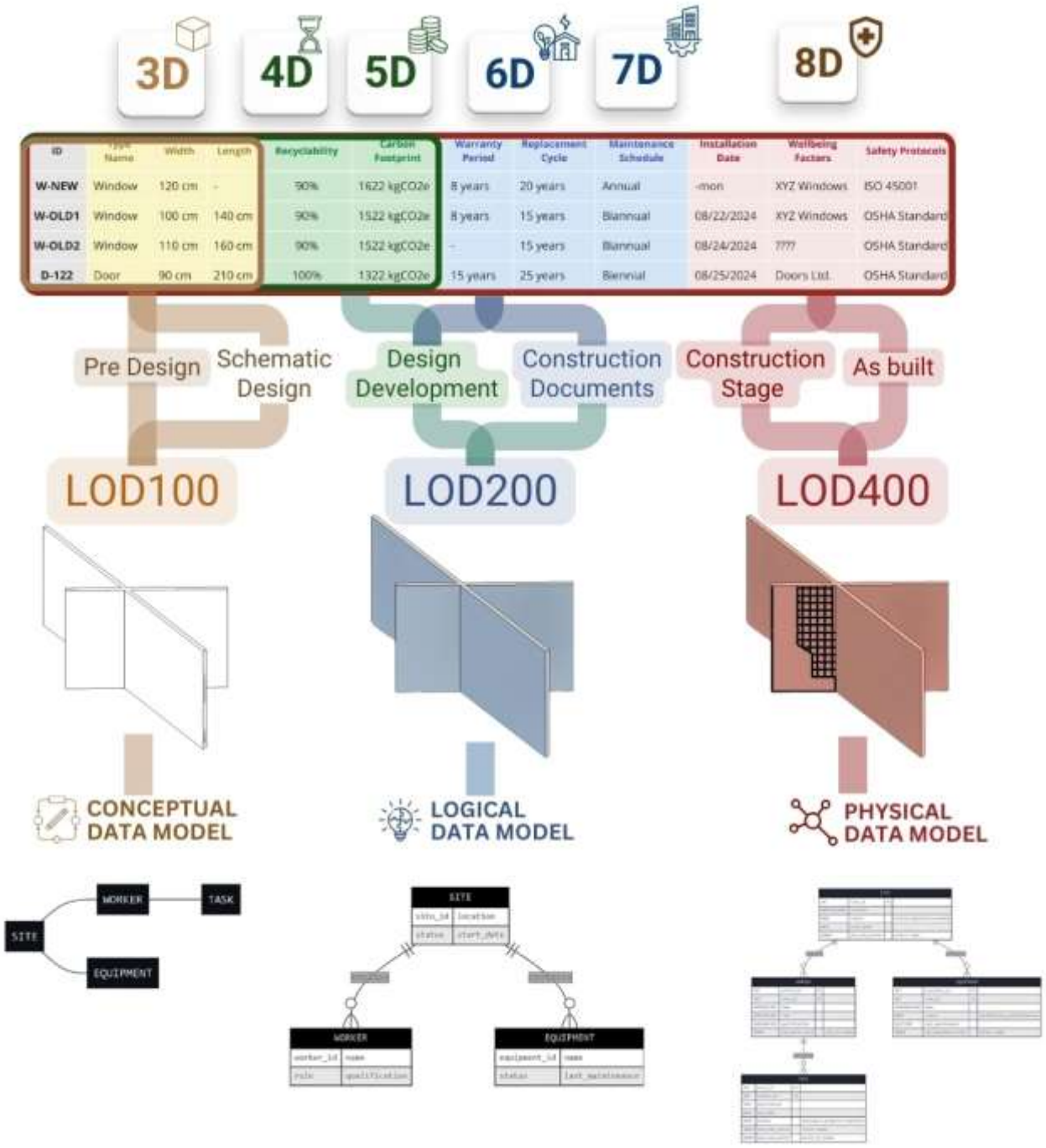
Một trong những ví dụ về các phương pháp mới trong việc tiêu chuẩn hóa là việc giới thiệu các cấp độ phát triển mô hình, phản ánh mức độ sẵn sàng và độ tin cậy của cả thành phần đồ họa và thông tin. Để phân định mức độ thông tin trong công việc với dữ liệu CAD (BIM), một khái niệm gọi là LOD (Level Of Detail) đã xuất hiện - mức độ chi tiết của phần đồ họa của mô hình, và LOI (Level Of Information) - mức độ phát triển của dữ liệu. Thêm vào đó, để có cách tiếp cận toàn diện, khái niệm LOA (Level of Accuracy) đã được giới thiệu - độ chính xác của các yếu tố đã trình bày và LOG (Level of Geometry) để xác định độ chính xác của biểu diễn đồ họa.

Các mức độ chi tiết (LOD) được ký hiệu bằng các số từ 100 đến 500, phản ánh mức độ phát triển của mô hình. LOD 100 là một mô hình khái niệm với các hình dạng và kích thước chung. LOD 200 bao gồm kích thước và hình dạng chính xác hơn, nhưng với độ chi tiết ước lệ. LOD 300 là một mô hình chi tiết với kích thước, hình dạng và vị trí của các yếu tố chính xác. LOD 400 chứa thông tin chi tiết cần thiết cho việc chế tạo và lắp đặt các yếu tố. LOD 500 phản ánh trạng thái thực tế của đối tượng sau khi xây dựng và được sử dụng cho hoạt động và bảo trì. Những cấp độ này mô tả cấu trúc của việc bão hòa thông tin trong mô hình CAD (BIM) ở các giai đoạn khác nhau của vòng đời, bao gồm 3D, 4D, 5D và hơn thế nữa.

Trong các dự án thực tế, độ chi tiết cao (LOD400) thường trở nên dư thừa và có thể chỉ cần sử dụng hình học LOD100 hoặc thậm chí các bản vẽ phẳng, trong khi các dữ liệu khác có thể được thu thập qua tính toán hoặc từ các yếu tố liên quan, không nhất thiết phải có hình học rõ ràng. Ví dụ, không gian và các yếu tố phòng (danh mục các yếu tố "Phòng") có thể không có hình học hình ảnh, nhưng vẫn chứa khối lượng thông tin đáng kể và cơ sở dữ liệu, xung quanh đó nhiều quy trình kinh doanh được xây dựng.

> Do đó, trước khi bắt đầu thiết kế, điều quan trọng là xác định rõ mức độ chi tiết cần thiết. Đối với các trường hợp sử dụng 4D-7D, thường chỉ cần bản vẽ DWG và hình học tối thiểu LOD100. Nhiệm vụ then chốt trong quá trình làm việc với các yêu cầu là tìm ra sự cân bằng giữa độ phong phú và khả năng ứng dụng thực tế của mô hình.

Thực tế thì, nếu xem dữ liệu CAD (BIM) như một cơ sở dữ liệu (mà nó thực sự là), việc mô tả độ phong phú của mô hình qua các viết tắt mới thực chất chỉ là một quá trình mô hình hóa dữ liệu cho các hệ thống thông tin, bắt đầu từ cấp độ khái niệm đến mô hình vật lý (Hình 6.34), điều này đã được xem xét chi tiết trong phần ba và phần bốn của sách. Mỗi sự gia tăng LOD và LOI có nghĩa là việc thêm thông tin cần thiết cho các nhiệm vụ mới: tính toán, quản lý xây dựng, vận hành và được đặc trưng bởi sự làm phong phú mô hình một cách tuần tự với các lớp thông tin bổ sung (3D-8D) dưới dạng các tham số khác nhau, mà chúng tôi đã đề cập trong phần năm của sách.



Hình 6.34 Quy trình tiếp đầy thông tin cho chi tiết dự án tương tự như mô hình hóa dữ liệu từ mô hình khái niệm đến mô hình vật lý.

Hình học chỉ là một phần nhỏ của dữ liệu thiết kế, và sự cần thiết của nó không phải lúc nào cũng được biện minh trong các dự án xây dựng, và câu hỏi then chốt trong công việc với dữ liệu CAD không chỉ nằm ở việc mô hình được hình dung như thế nào mà còn ở việc dữ liệu từ những mô hình này có thể được sử dụng ra sao bên ngoài các chương trình CAD (BIM).

Đến giữa những năm 2000, ngành xây dựng đã đối mặt với một vấn đề chưa từng có liên quan đến sự gia tăng nhanh chóng khối lượng dữ liệu trong các hệ thống quản lý và xử lý dữ liệu, đặc biệt là những dữ liệu

được thu nhận từ các bộ phận CAD (BIM). Sự gia tăng đột ngột này đã khiến các nhà quản lý công ty bất ngờ, và họ không chuẩn bị cho các yêu cầu ngày càng cao về chất lượng và quản lý dữ liệu.

Các tiêu chuẩn mới trong CAD (BIM) - AIA, BEP, IDS, LOD, COBie

Tận dụng việc thiếu quyền truy cập mở vào các cơ sở dữ liệu CAD và sự cạnh tranh hạn chế trên thị trường xử lý dữ liệu, cùng với các chiến dịch tiếp thị liên quan đến viết tắt mới BIM, các tổ chức phát triển các phương pháp làm việc với dữ liệu CAD đã bắt đầu tạo ra các tiêu chuẩn và khái niệm mới, về lý thuyết nhằm cải thiện thực tiễn quản lý dữ liệu.

> Mặc dù hầu hết các sáng kiến, trực tiếp hoặc gián tiếp được hỗ trợ bởi các nhà cung cấp và nhà phát triển CAD (BIM), đã được định hướng vào việc tối ưu hóa quy trình làm việc, chúng đã dẫn đến sự xuất hiện của nhiều tiêu chuẩn được vận động bởi các bên liên quan khác nhau, điều này đã khiến ngành xây dựng rơi vào trạng thái mơ hồ và nhầm lẫn trong quy trình xử lý dữ liệu.

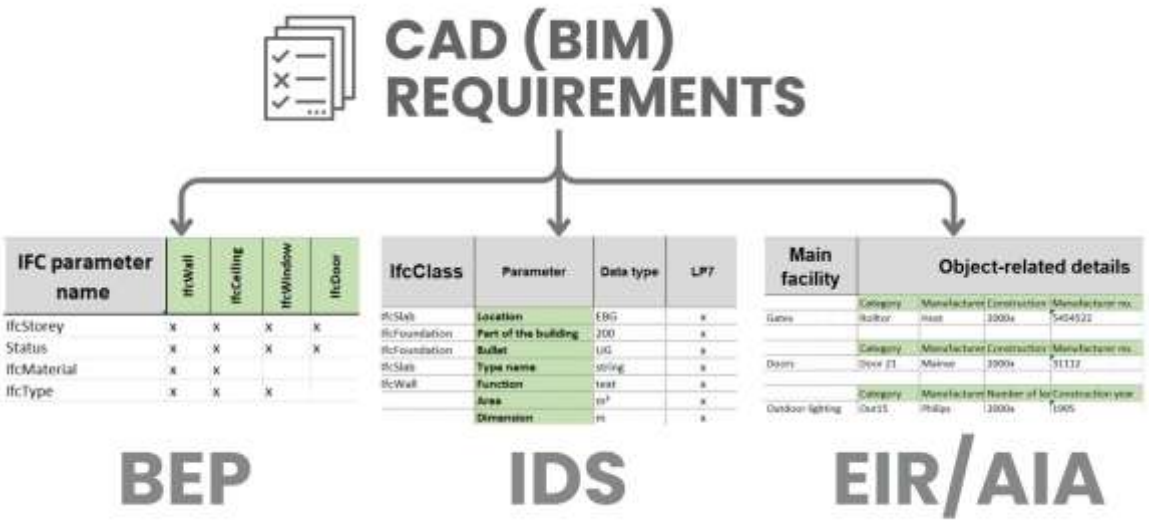
Dưới đây là một số tiêu chuẩn dữ liệu mới xuất hiện trong ngành xây dựng những năm gần đây, bên cạnh LOD, LOI, LOA, LOG:

- BEP (BIM Execution Plan) - mô tả cách tích hợp và sử dụng CAD (BIM) trong dự án, xác định các phương pháp và quy trình xử lý dữ liệu.
- Tài liệu EIR/AIA (Yêu cầu thông tin của khách hàng) - được chuẩn bị bởi khách hàng trước khi công bố thầu và chứa các yêu cầu đối với nhà thầu về việc chuẩn bị và cung cấp thông tin. Nó là cơ sở cho BEP trong dự án tương ứng.
- AIM (Asset Information Model) - một phần của quy trình BIM. Sau khi hoàn thành và bàn giao dự án, mô hình dữ liệu được gọi là "Mô hình thông tin tài sản" hoặc AIM. Mục tiêu của AIM là quản lý, bảo trì và vận hành tài sản đã được hiện thực hóa.
- IDS (Thông số Giao hàng Thông tin) — xác định yêu cầu và dữ liệu cần thiết cũng như định dạng cho các giai đoạn khác nhau của dự án xây dựng.
- iLOD — mức độ chi tiết LOD, giúp thông tin được trình bày trong mô hình BIM. Nó xác định mức độ chi tiết và toàn diện của thông tin trong mô hình, từ các đại diện hình học cơ bản cho đến các đặc điểm kỹ thuật và dữ liệu chi tiết.
- eLOD — mức độ chi tiết LOD của các phần tử riêng lẻ trong mô hình CAD (BIM). Nó xác định mức độ mô hình hóa của từng phần tử và các thông tin liên quan như kích thước, vật liệu, đặc tính vận hành và các thuộc tính phù hợp khác.
- APS (Dịch vụ Nền tảng) và các sản phẩm khác từ các nhà cung cấp CAD lớn (BIM) — mô tả các công cụ và hạ tầng cần thiết để tạo ra các mô hình dữ liệu liên kết và mở.

Mặc dù mục tiêu được tuyên bố khi triển khai các tiêu chuẩn CAD (BIM) như LOD, LOI, LOA, LOG, BEP, EIR, AIA, AIM, IDS, iLOD, eLOD là nhằm nâng cao chất lượng quản lý dữ liệu và mở rộng khả năng tự động hóa, nhưng thực tế việc sử dụng chúng lại thường dẫn đến sự phức tạp và phân mảnh quá mức trong các quy trình. Nếu xem mô hình CAD (BIM) như một loại cơ sở dữ liệu, thì có thể thấy rằng nhiều tiêu chuẩn này đang lặp lại những phương pháp đã hiệu quả và tồn tại lâu đời trong các ngành kinh tế khác khi làm việc với các hệ thống thông tin. Thay vì đơn giản hóa và thống nhất, những sáng kiến này thường tạo ra gánh nặng thuật

ngữ bổ sung và cản trở việc triển khai các giải pháp thực sự mở và linh hoạt.

Điều đáng chú ý là nhiều khái niệm mới này thực sự thay thế cho quy trình mô hình hóa và kiểm tra dữ liệu đã được thảo luận chi tiết trong các phần đầu của cuốn sách và đã được sử dụng từ lâu trong các lĩnh vực khác. Trong ngành xây dựng, quy trình tiêu chuẩn hóa thường diễn ra theo hướng ngược lại — tạo ra các định dạng mới để mô tả dữ liệu, các tiêu chuẩn mới và các khái niệm mới để kiểm tra, mà không phải lúc nào cũng dẫn đến sự thống nhất thực tế và khả năng ứng dụng. Cuối cùng, thay vì đơn giản hóa và tự động hóa việc xử lý, ngành công nghiệp lại phải đối mặt với các cấp quy định và thủ tục thêm vào (Hình. 6.31), điều này không phải lúc nào cũng thúc đẩy hiệu quả.



Hình. 6.31 Các yêu cầu về dữ liệu và thông tin được tóm lược thành mô tả các thuộc tính và các giới hạn của chúng, được mô tả qua bảng biểu.

> Thay vì đơn giản hóa việc xử lý dữ liệu, các khái niệm mới liên quan đến dữ liệu CAD (BIM) thường tạo ra nhiều phức tạp và tranh cãi ngay từ giai đoạn diễn giải và các định nghĩa cơ bản.

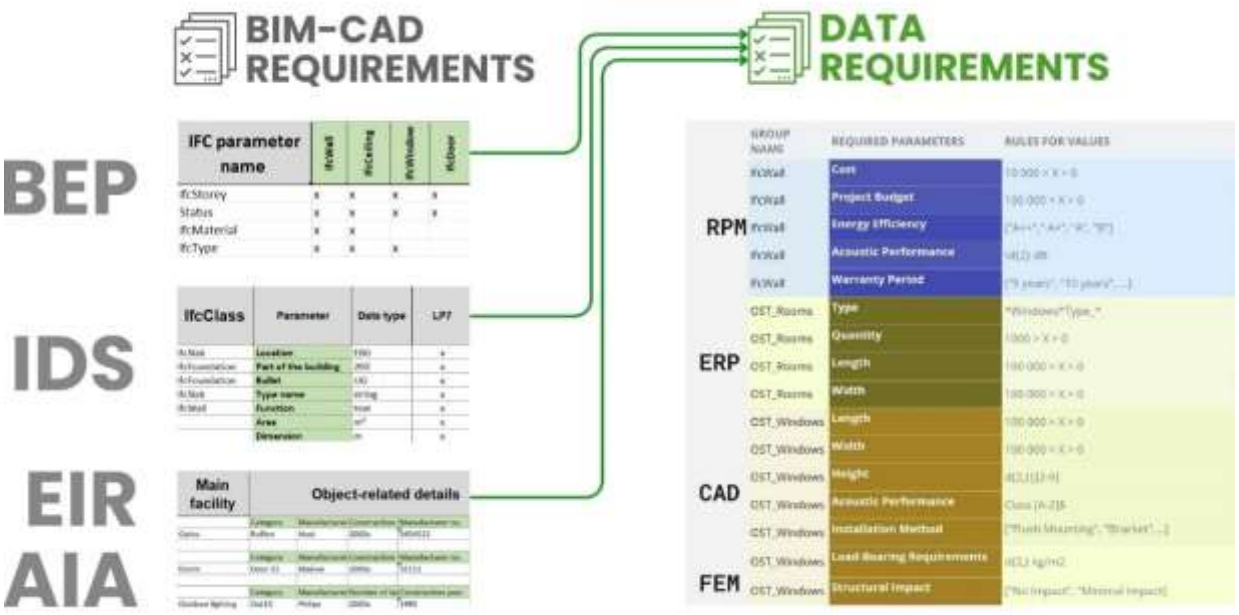
Một trong những ví dụ gần đây về các khái niệm mới là định dạng IDS (xuất hiện năm 2020) cho phép mô tả yêu cầu về thành phần thuộc tính của mô hình thông tin trong khái niệm open BIM. Các yêu cầu IDS mô tả thông tin về các thuộc tính và giới hạn của chúng dưới dạng bảng có cấu trúc (Excel hoặc MySQL), sau đó được chuyển đổi sang đánh dấu định dạng bán cấu trúc XML, được đổi tên từ XML thành viết tắt IDS.

> Không như ý kiến được các nhà cung cấp và các khái niệm BIM và open BIM ủng hộ rằng công việc với dữ liệu trong xây dựng là độc nhất do việc sử dụng các công cụ chuyên biệt như CAD và BIM, các định dạng dữ liệu và phương pháp quản lý dữ liệu trong ngành này không khác biệt so với các định dạng và khái niệm xử lý dữ liệu trong các ngành khác.

Số lượng yêu cầu đối với các dự án và định dạng CAD (BIM) có thể được đơn giản hóa bằng cách sử dụng một bảng yêu cầu thống nhất với các thuộc tính là cột, được mô tả chi tiết trong chương "Chuyển đổi yêu cầu sang dạng cấu trúc", mà không cần phải chuyển đổi các yêu cầu đã được cấu trúc ban đầu sang định

dạng không bảng (IDS được mô tả ban đầu thông qua bảng).

Cách tiếp cận đơn giản hóa (Hình 6.32), bao gồm các cột cho các định danh thực thể, thuộc tính và các giá trị biên, đã được xem xét chi tiết trong các chương trước (Hình 4.49, Hình 4.416, Hình 7.310), cho phép không cần chuyển đổi yêu cầu sang định dạng IDS-XML. Cách này đảm bảo một cơ chế kiểm soát chất lượng dữ liệu trực tiếp, ít cồng kềnh hơn và minh bạch hơn. Nó dựa vào các công cụ phổ biến: từ biểu thức chính quy (RegEx) đến khung dữ liệu, thư viện Pandas và các quy trình ETL tiêu chuẩn - chính những công cụ mà các chuyên gia ở các lĩnh vực khác của nền kinh tế sử dụng khi làm việc với dữ liệu.---



Hình 6.32 Các yêu cầu về dữ liệu ở các ngành khác được đơn giản hóa đến mô tả cấu trúc của các thuộc tính và các giá trị biên của chúng.

Qua thời gian, trong ngành xây dựng, do sự đóng cửa dữ liệu, ngày càng có nhiều phương pháp và kỹ thuật mới để kiểm soát và quản lý dữ liệu đa định dạng này, mặc dù dữ liệu trong các dự án xây dựng về bản chất không khác gì so với dữ liệu ở các lĩnh vực khác. Trong khi các ngành khác thành công với các phương pháp chuẩn hóa trong xử lý dữ liệu, ngành xây dựng tiếp tục phát triển các định dạng dữ liệu, yêu cầu và khái niệm kiểm tra độc đáo mới.

> Các phương pháp và công cụ sử dụng để thu thập, chuẩn bị và phân tích dữ liệu trong ngành xây dựng không nên khác biệt về mặt nguyên tắc so với những gì được áp dụng bởi các chuyên gia trong các lĩnh vực khác của nền kinh tế.

Trong ngành đã hình thành một hệ sinh thái thuật ngữ đặc biệt, cần được xem xét và đánh giá lại một cách có .

- Định dạng STEP được định vị dưới tên gọi mới IFC, được bổ sung bởi phân loại xây dựng, mà không xem xét các hạn chế của chính định dạng STEP.
- Định dạng tham số IFC được áp dụng trong các quy trình truyền dữ liệu, mặc dù thiếu một lõi hình học thống nhất cần thiết cho việc hình dung và tính toán.
- Truy cập vào cơ sở dữ liệu hệ thống CAD được quảng bá dưới thuật ngữ "BIM", mà không thảo luận

về các đặc điểm của các cơ sở dữ liệu này và việc truy cập vào chúng.

- Các nhà cung cấp thúc đẩy tính tương tác thông qua các định dạng IFC và USD, thường không áp dụng chúng trong thực tiễn, sử dụng kỹ thuật đảo ngược tốn kém mà họ từng chống lại.
- Các thuật ngữ LOD, LOI, LOA, LOG, BEP, EIR, AIA, AIM, IDS, iLOD, eLOD được sử dụng rộng rãi để mô tả cùng một tham số của các thực thể, mà không liên quan đến các công cụ mô hình hóa và xác minh, đã được sử dụng từ lâu ở các lĩnh vực khác.

Ngành xây dựng chứng tỏ rằng tất cả những gì được liệt kê trên nghe có vẻ lạ, nhưng lại có thể xảy ra trong ngành xây dựng - đặc biệt nếu mục tiêu chính là kiếm tiền từ mỗi giai đoạn xử lý dữ liệu thông qua việc bán các dịch vụ và phần mềm chuyên biệt. Từ góc độ kinh doanh, điều này không có gì là không thể chấp nhận. Tuy nhiên, câu hỏi rằng liệu những viết tắt và cách tiếp cận liên quan đến CAD (BIM) có thực sự tạo ra giá trị và đơn giản hóa các quy trình chuyên nghiệp hay không, vẫn còn bỏ ngỏ.

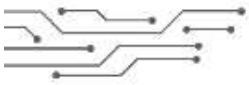
> Trong ngành xây dựng, hệ thống như vậy hoạt động vì chính ngành công nghiệp này thu được lợi nhuận đầu cơ chủ yếu từ những mê cung của hệ thống và từ viết tắt. Các công ty quan tâm đến các quy trình minh bạch và dữ liệu mở là rất hiếm. Tình huống phức tạp này có thể kéo dài vô thời hạn — cho đến khi các nhà thầu, khách hàng, nhà đầu tư, ngân hàng và đại diện của vốn tư nhân bắt đầu yêu cầu các phương pháp quản lý thông tin rõ ràng và hợp lý hơn.

Ngành đã tích lũy một lượng lớn các từ viết tắt, nhưng tất cả chúng đều mô tả các quy trình và yêu cầu dữ liệu tương tự nhau ở những mức độ khác nhau. Lợi ích thực sự của chúng trong việc đơn giản hóa quy trình làm việc vẫn còn gây tranh cãi.

Trong khi các khái niệm và từ viết tắt trong marketing đến và đi, bản thân các quy trình kiểm tra yêu cầu dữ liệu sẽ luôn là một phần không thể thiếu trong quy trình kinh doanh. Thay vì tạo ra ngày càng nhiều định dạng và quy định chuyên biệt mới, ngành xây dựng nên chú ý đến các công cụ đã chứng minh hiệu quả của chúng trong các lĩnh vực khác, chẳng hạn như tài chính, công nghiệp và CNTT.

Sự phong phú của thuật ngữ, từ viết tắt và định dạng tạo ra ảo giác về sự phát triển sâu sắc trong các quy trình xây dựng kỹ thuật số. Tuy nhiên, đằng sau các khái niệm marketing và thuật ngữ phức tạp thường ẩn giấu một sự thật đơn giản nhưng khó chịu: dữ liệu vẫn khó tiếp cận, được tài liệu kém và gắn chặt với các giải pháp phần mềm cụ thể.

Để ra khỏi vòng luẩn quẩn của từ viết tắt và định dạng vì các định dạng, cần nhìn vào các hệ thống CAD (BIM) không phải là những công cụ ma thuật quản lý thông tin, mà là những gì chúng thực sự là — các cơ sở dữ liệu chuyên biệt. Chỉ qua lăng kính này, chúng ta mới có thể hiểu nơi mà marketing kết thúc và công việc thực tế với thông tin bắt đầu.



CHƯƠNG 6.4. TÍNH TOÁN CỦA THIẾT KẾ VÀ VIỆC SỬ DỤNG LLM ĐỂ LÀM VIỆC VỚI CAD

Ảo tưởng về tính độc đáo của dữ liệu CAD (BIM): con đường đến phân tích và định dạng mở

Các nền tảng CAD (BIM) hiện đại đã biến đổi đáng kể cách tiếp cận thiết kế và quản lý thông tin xây dựng. Trước đây, các công cụ này chủ yếu được sử dụng để tạo ra bản vẽ và mô hình 3D, thì ngày nay chúng thực hiện chức năng của một kho dữ liệu dự án toàn diện. Trong khuôn khổ khái niệm Nguồn Thực Thà Nhất (Single Source of Truth), mô hình tham số ngày càng trở thành nguồn thông tin chính và không ít thì nhiều là nguồn thông tin duy nhất về dự án, đảm bảo tính toàn vẹn và độ chính xác của nó trong suốt vòng đời của đối tượng.

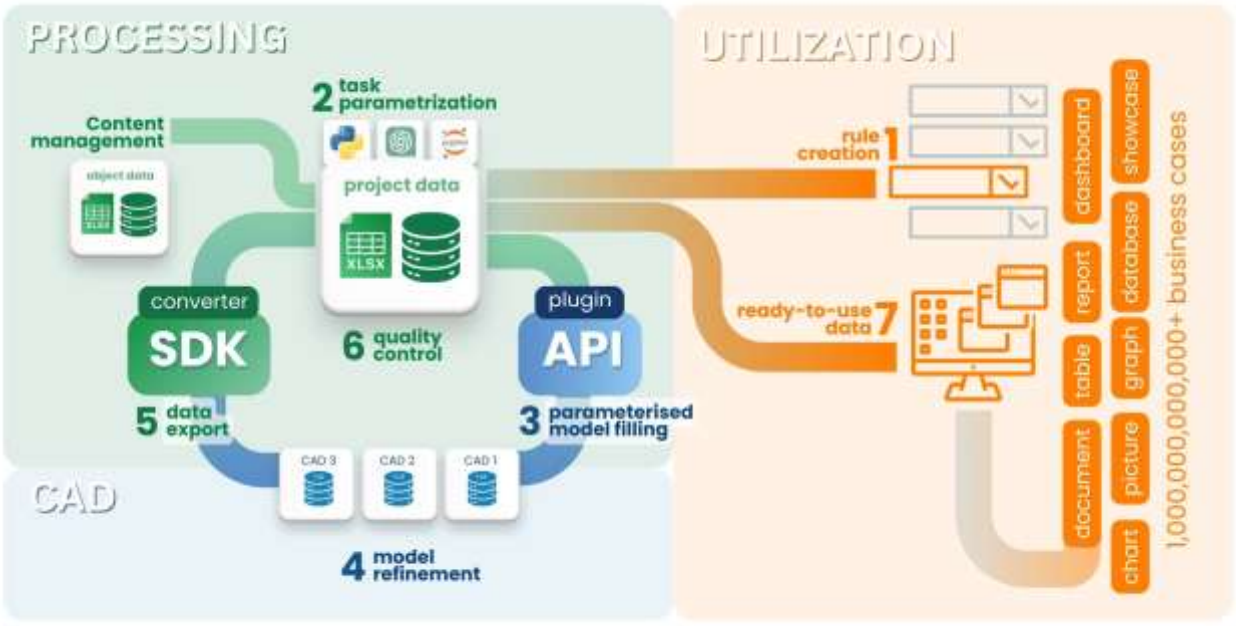
> Sự khác biệt chính giữa các nền tảng CAD- (BIM-) và các hệ thống quản lý dữ liệu xây dựng khác nằm ở sự cần thiết sử dụng các công cụ chuyên biệt và API để truy cập thông tin (nguồn thực thà nhất). Những cơ sở dữ liệu này không phải là phổ quát theo nghĩa truyền thống: thay vì có cấu trúc mở và tích hợp linh hoạt, chúng đại diện cho một môi trường kín, gắn liền chặt chẽ với nền tảng và định dạng cụ thể.

Mặc dù có sự phức tạp trong việc làm việc với dữ liệu CAD, một câu hỏi quan trọng hơn xuất hiện, vượt ra ngoài việc thực hiện kỹ thuật: cơ sở dữ liệu CAD (BIM) thực sự là gì? Để trả lời câu hỏi này, cần phải vượt ra ngoài những từ viết tắt và khái niệm quen thuộc, thường được các nhà phát triển phần mềm áp đặt. Thay vào đó, nên tập trung vào bản chất của việc làm việc với thông tin dự án: dữ liệu và các quy trình xử lý của chúng.

> Quy trình kinh doanh trong xây dựng không bắt đầu với việc làm việc trong các công cụ CAD hoặc BIM, mà là từ việc hình thành yêu cầu cho dự án và mô hình hóa dữ liệu. Đầu tiên, các tham số của nhiệm vụ được xác định: danh sách các thực thể, các đặc điểm ban đầu của chúng và các giá trị biên cần được xem xét khi giải quyết nhiệm vụ cụ thể. Chỉ sau đó, dựa trên các tham số đã được chỉ định, các mô hình và yếu tố được tạo ra trong các hệ thống CAD (BIM).

Quy trình trước khi tạo ra thông tin trong các cơ sở dữ liệu CAD (BIM) hoàn toàn lặp lại quá trình mô hình hóa dữ liệu, đã được xem xét chi tiết trong phần thứ tư của cuốn sách và chương "Mô hình hóa dữ liệu: mô hình khái niệm, logic và vật lý".-

Giống như trong quá trình mô hình hóa dữ liệu, chúng ta tạo ra yêu cầu cho dữ liệu mà sau này chúng ta muốn xử lý trong cơ sở dữ liệu, đối với cơ sở dữ liệu CAD, các nhà quản lý tạo ra yêu cầu thiết kế dưới dạng một số cột bảng hoặc danh sách các cặp "khóa–giá trị". Và chỉ trên cơ sở các tham số ban đầu này, thông qua API một cách tự động hoặc thủ công, các nhà thiết kế tạo ra (hoặc chính xác hơn là xác định lại) các đối tượng trong cơ sở dữ liệu CAD (BIM), sau đó chúng lại được kiểm tra để đảm bảo tuân thủ các yêu cầu ban đầu. Quy trình này — xác định → tạo ra → kiểm tra → điều chỉnh — lặp lại theo chu kỳ cho đến khi chất lượng dữ liệu, giống như trong mô hình hóa dữ liệu, đạt đến mức độ cần thiết cho hệ thống mục tiêu – tài liệu, bảng biểu hoặc bảng điều khiển. -

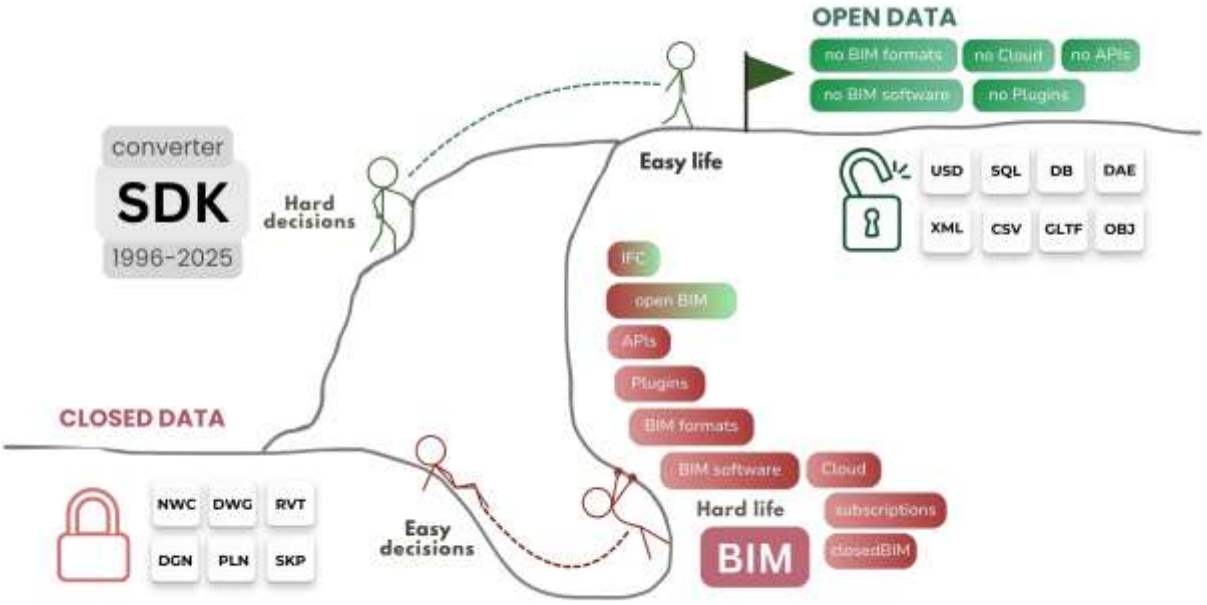


Hình 6.41 Chu trình làm giàu thông tin của cơ sở dữ liệu cho các quy trình kinh doanh trong thực hiện các dự án xây dựng.

Nếu xem CAD (BIM) như một cơ chế truyền tham số dưới dạng một tập hợp các cặp "khóa–giá trị" được hình thành dựa trên các yêu cầu đặt ra bên ngoài môi trường thiết kế, thì trọng tâm của cuộc thảo luận sẽ chuyển từ các giải pháp phần mềm cụ thể và những hạn chế của chúng sang các khía cạnh cơ bản hơn — cấu trúc dữ liệu, mô hình dữ liệu và các yêu cầu đối với chúng. Về cơ bản, đây là việc làm giàu cơ sở dữ liệu với các tham số và quy trình mô hình hóa dữ liệu cổ điển. Điểm khác biệt chỉ nằm ở chỗ, do tính khép kín của các cơ sở dữ liệu CAD và đặc điểm của các định dạng sử dụng, quy trình này được thực hiện với sự hỗ trợ của các công cụ BIM chuyên dụng. Một câu hỏi đặt ra là: sự độc đáo của BIM là gì, nếu trong các ngành kinh tế khác không tồn tại các phương pháp tương tự?-

Trong 20 năm qua, BIM đã được vị trí như một thứ gì đó lớn hơn, không chỉ đơn thuần là một nguồn dữ liệu duy nhất. Về mặt tiếp thị, sự kết hợp CAD-BIM thường được bán như một công cụ tham số với cơ sở dữ liệu tích hợp sẵn, có khả năng tự động hóa các quy trình thiết kế, mô hình hóa và quản lý vòng đời của các đối tượng xây dựng. Tuy nhiên, trên thực tế, BIM đã trở thành nhiều hơn là một công cụ giữ chân người dùng trên nền tảng của các nhà cung cấp, thay vì một phương pháp tiện lợi để quản lý dữ liệu và quy trình.

Cuối cùng, dữ liệu CAD (BIM) bị cô lập bên trong các nền tảng của nó, ẩn giấu thông tin về dự án sau các API độc quyền và các lõi hình học. Điều này đã tước đi khả năng của người dùng trong việc tự mình truy cập vào các cơ sở dữ liệu và khai thác, phân tích, tự động hóa và chuyển giao dữ liệu sang các hệ thống khác mà không cần đi qua hệ sinh thái của các nhà cung cấp.



Hình 6.42 Trong xây dựng, các định dạng hiện đại yêu cầu các lõi hình học phức tạp, API được cập nhật hàng năm và các giấy phép đặc biệt cho các chương trình CAD (BIM).

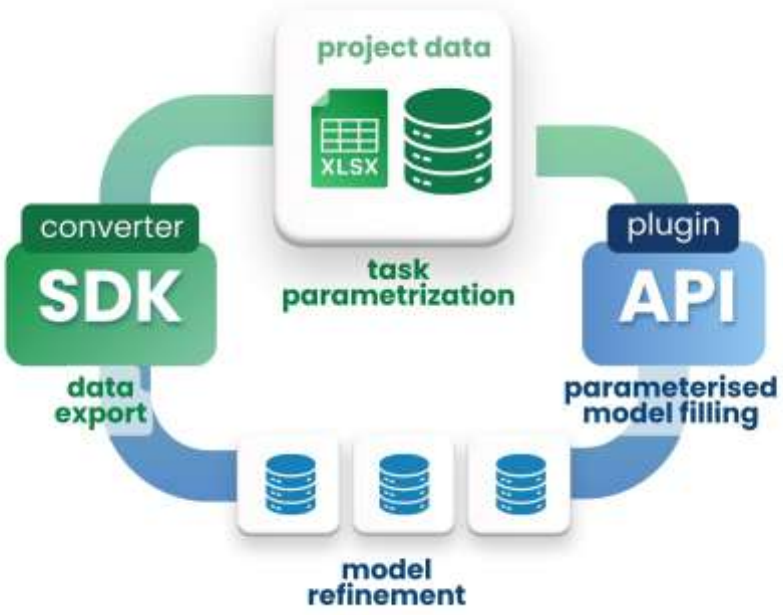
Các công ty đang làm việc với các công cụ CAD hiện đại nên áp dụng cách tiếp cận tương tự để xử lý dữ liệu như tất cả các nhà cung cấp CAD đều sử dụng trong thực tế: chuyển đổi dữ liệu bằng cách sử dụng công cụ SDK cho việc đảo ngược kỹ thuật, mà các nhà cung cấp CAD đã chống lại kể từ năm 1995. Với việc có quyền truy cập đầy đủ vào cơ sở dữ liệu CAD và sử dụng các công cụ đảo ngược kỹ thuật, chúng ta có thể tạo ra một bộ thực thể phẳng với các thuộc tính và xuất chúng sang bất kỳ định dạng mở nào thuận tiện (Hình 6.42), bao gồm cả hình học và thông số của các yếu tố thiết kế. Cách tiếp cận này thay đổi một cách cơ bản cách thức làm việc với thông tin — từ kiến trúc tập trung vào tệp sang kiến trúc tập trung vào dữ liệu.-

- Các định dạng dữ liệu như: RVT, IFC, PLN, DB1, CP2, CPIXML, USD, SQLite, XLSX, PARQUET và những định dạng khác chứa thông tin tương tự về các yếu tố của cùng một dự án. Điều này có nghĩa là việc hiểu biết về định dạng cụ thể và sơ đồ của nó không nên trở thành rào cản cho việc làm việc với chính dữ liệu.
- Dữ liệu từ bất kỳ định dạng nào có thể được kết hợp thành một cấu trúc mở được tổ chức và phân loại (Hình 9.110), chứa hình học ba chiều MESH và thuộc tính của tất cả các thực thể đối tượng, không bị giới hạn bởi các lõi hình học.
- Phân tích dữ liệu hướng tới tính phổ quát: sử dụng dữ liệu mở, có thể làm việc với dữ liệu thiết kế mà không phụ thuộc vào định dạng được sử dụng.
- Giảm thiểu, cũng như sự phụ thuộc vào API và plugin của nhà cung cấp: việc làm việc với dữ liệu không còn phải phụ thuộc vào kỹ năng sử dụng API.

Khi các yêu cầu và dữ liệu CAD được chuyển đổi thành các định dạng dễ phân tích hơn cho việc biểu diễn có tổ chức — các nhà phát triển không còn phải phụ thuộc vào sơ đồ dữ liệu cụ thể và các hệ sinh thái khép kín.

Thiết kế thông qua tham số: tương lai của CAD và BIM

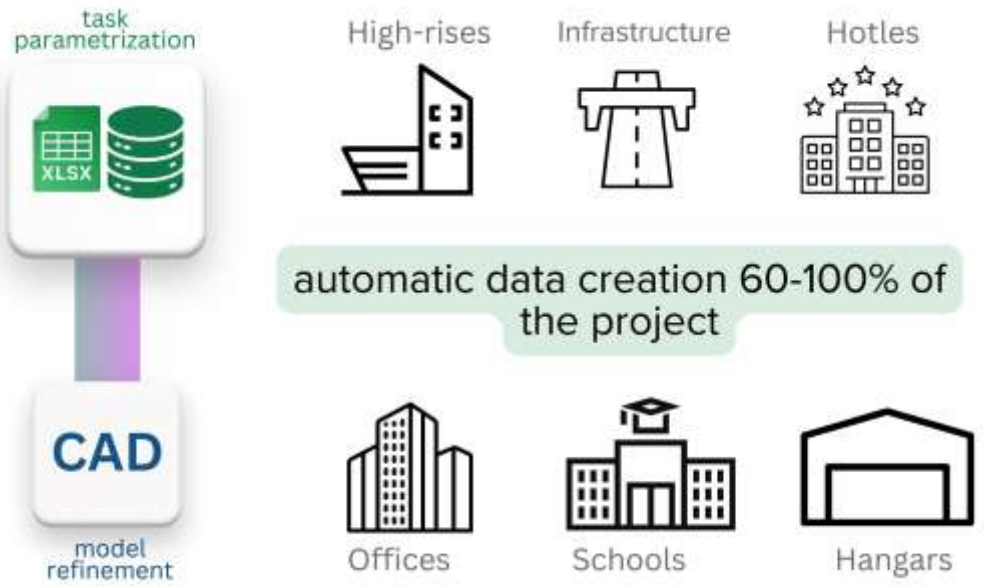
Không một dự án xây dựng nào trên thế giới từng bắt đầu trong chương trình CAD. Trước khi bản vẽ hoặc mô hình được hình thành trong CAD, chúng trải qua giai đoạn hình thành ý tưởng, nơi tập trung vào các thông số xác định ý tưởng cơ bản và logic của đối tượng tương lai. Giai đoạn này tương ứng với cấp độ khái niệm trong mô hình hóa dữ liệu. Các thông số có thể chỉ tồn tại trong tâm trí của người thiết kế, tuy nhiên trong điều kiện lý tưởng, chúng được thể hiện dưới dạng danh sách có cấu trúc, bảng hoặc lưu trữ trong cơ sở dữ liệu, giúp đảm bảo tính minh bạch, khả năng tái tạo và sự tự động hóa tiếp theo trong quy trình thiết kế. ---



Hình 6.43 Quy trình thiết kế là một quy trình lặp đi lặp lại việc làm đầy cơ sở dữ liệu CAD bằng thông tin từ bên ngoài thông qua yêu cầu trong chuỗi tạo ra giá trị.

Trước khi bắt đầu việc mô hình hóa CAD (giai đoạn logic và vật lý của mô hình hóa dữ liệu), điều quan trọng là xác định các tham số biên, đóng vai trò nền tảng cho dự án. Các thuộc tính này, giống như các yêu cầu khác, được thu thập từ phần cuối cùng của chuỗi sử dụng dữ liệu (ví dụ, các hệ thống) và từ đó xác định các giới hạn, mục tiêu và các đặc điểm chính của các đối tượng trong dự án.

Việc lập mô hình, khi có yêu cầu được lập trình hợp lý, có thể hoàn toàn tự động hóa từ 60–100% với sự trợ giúp của các công cụ lập mô hình tham số (Hình 6.43). Khi dự án được mô tả dưới dạng các tham số, việc hình thành nó trở nên khả thi về mặt kỹ thuật, ví dụ thông qua các ngôn ngữ lập trình trực quan như Grasshopper, Dynamo, tích hợp trong các môi trường CAD hiện đại hoặc các giải pháp miễn phí trong các sản phẩm như Blender, UE, Omniverse.



Hình 6.44 Phần lớn các dự án kiểu dáng đã được tạo ra hoàn toàn tự động nhờ vào các công cụ lập trình tham số.

Ngày nay, các dự án công nghiệp lớn và kiểu dáng không được tạo ra bằng tay bởi bộ phận thiết kế, mà thông qua các công cụ tham số và lập trình trực quan. Điều này cho phép xây dựng mô hình dựa trên dữ liệu, không phải từ các quyết định chủ quan của một nhà thiết kế hay quản lý cụ thể.

> *Nội dung phải đi trước thiết kế. Thiết kế mà không có nội dung là không phải thiết kế, mà chỉ là trang trí.*
>
> *– Jeffrey Zeldman, nhà thiết kế web và doanh nhân.*

Quy trình bắt đầu không phải từ việc vẽ hoặc mô hình hóa 3D, mà từ việc hình thành các yêu cầu. Chính các yêu cầu xác định các yếu tố nào sẽ được sử dụng trong dự án, dữ liệu nào cần chuyển đến các phòng ban và hệ thống khác. Chỉ khi nào có các yêu cầu được cấu trúc thì mới có thể tự động kiểm tra các mô hình một cách định kỳ (ví dụ, cứ mỗi 10 phút, mà không làm xao lãng công việc của nhà thiết kế).

> Có thể trong tương lai, hệ thống CAD (BIM) sẽ chỉ trở thành giao diện để điền vào cơ sở dữ liệu, và việc mô hình hóa (ở cấp độ vật lý) trong công cụ CAD nào sẽ không còn quan trọng.

Tương tự như vậy, trong ngành cơ khí, việc mô hình hóa 3D thường được sử dụng, nhưng không phải là yếu tố cần thiết hay bắt buộc trong dự án. Trong phần lớn các trường hợp, tài liệu 2D cổ điển là đủ – trên cơ sở đó, toàn bộ mô hình thông tin cần thiết được tạo ra. Mô hình này được xây dựng từ các thành phần được cấu trúc theo các tiêu chuẩn ngành, và chứa tất cả thông tin cần thiết để hiểu về kết cấu và tổ chức sản xuất. Sau đó, trên cơ sở đó, một mô hình thông tin nhà máy được hình thành, với các sản phẩm cụ thể và các bản đồ quy trình công nghệ, đã được định hướng theo nhu cầu của các kỹ thuật viên. Toàn bộ quá trình có thể được tổ chức một cách đơn giản, không làm quá tải hệ thống 3D với đồ họa nơi mà nó không mang lại lợi ích thực tế.

> Quan trọng là hiểu rằng mô hình 3D và hệ thống CAD không nên đóng vai trò chủ đạo – nó chỉ là công cụ cho phân tích định lượng và hình học. Tất cả các tham số khác, ngoài hình học, mà mô tả thực thể, nên được lưu trữ và xử lý ngoài môi trường CAD (BIM) nếu có thể.

Thiết kế thông qua các tham số không chỉ là một xu hướng, mà là tương lai không thể tránh khỏi của ngành xây dựng. Thay vì tạo ra các mô hình 3D phức tạp bằng tay, các nhà thiết kế sẽ làm việc với dữ liệu, kiểm tra chúng và tự động hóa các quy trình, đưa xây dựng lại gần với thế giới lập trình. Theo thời gian, các quy trình thiết kế sẽ được xây dựng theo các nguyên tắc phát triển phần mềm:

- Tạo yêu cầu → Tạo mô hình → Tải lên máy chủ → Kiểm tra thay đổi → Pull request.
- Trong khuôn khổ Pull request (yêu cầu thêm-gộp), các kiểm tra mô hình theo các yêu cầu đã được tạo ra trước khi bắt đầu hoặc trong quá trình thiết kế sẽ tự động được khởi động.
- Sau khi kiểm tra chất lượng dữ liệu và phê duyệt, các thay đổi sẽ được áp dụng vào dự án, vào cơ sở dữ liệu tổng thể hoặc tự động chuyển đến các hệ thống khác.

Hiện tại, trong ngành chế tạo máy, những thay đổi trong dự án bắt đầu từ việc hình thành thông báo về sự thay đổi. Một sơ đồ tương tự đang chờ đợi ngành xây dựng: quy trình thiết kế sẽ là một quá trình lặp đi lặp lại, trong đó mỗi bước được củng cố bằng các yêu cầu tham số. Hệ thống như vậy sẽ cho phép các nhà thiết kế tạo ra các kiểm tra tự động và các yêu cầu pull request (yêu cầu ghép hoặc thêm) theo các yêu cầu cụ thể.

> Người thiết kế tương lai chủ yếu là một người vận hành dữ liệu, chứ không phải là một người mô phỏng thủ công. Nhiệm vụ của họ là làm phong phú dự án bằng các thực thể tham số, trong đó hình học chỉ là một trong nhiều thuộc tính.

Một vai trò quan trọng trong quá trình chuyển đổi sẽ do hiểu biết về tầm quan trọng của việc mô hình hóa dữ liệu, phân loại và tiêu chuẩn hóa, đã được thảo luận chi tiết trong các chương trước của cuốn sách. Các quy định liên quan đến thiết kế trong tương lai sẽ được trình bày dưới dạng các cặp tham số khóa-giá trị theo định dạng XLSX hoặc XML.

> Tương lai của ngành xây dựng là thu thập dữ liệu, phân tích, kiểm tra và tự động hóa quy trình thông qua các công cụ phân tích. BIM (hoặc CAD) không phải là mục tiêu cuối cùng, mà chỉ là một giai đoạn trong quá trình tiến hóa. Khi các chuyên gia nhận ra rằng họ có thể làm việc trực tiếp với dữ liệu mà không cần qua các công cụ CAD truyền thống, thuật ngữ "BIM" sẽ dần nhường chỗ cho các khái niệm sử dụng dữ liệu có cấu trúc và phân đoạn trong dự án xây dựng.

Một trong những yếu tố chính thúc đẩy sự chuyển đổi là sự xuất hiện của các mô hình ngôn ngữ lớn (LLM) và các công cụ dựa trên chúng. Những công nghệ này đang thay đổi cách tiếp cận trong việc làm việc với dữ liệu dự án, cho phép truy cập thông tin mà không cần biết sâu về API hoặc giải pháp của nhà cung cấp. Nhờ có LLM, quy trình tạo ra yêu cầu và tương tác với dữ liệu CAD trở nên trực quan và dễ tiếp cận.

Sự xuất hiện của LLM trong các quy trình xử lý dữ liệu CAD dự án

Ngoài việc phát triển các công cụ truy cập vào cơ sở dữ liệu CAD và các định dạng CAD mở và đơn giản, sự xuất hiện của các công cụ LLM (Mô hình Ngôn ngữ Lớn) đã mang lại sự thay đổi cách mạng trong xử lý dữ liệu dự án. Trước đây, việc truy cập thông tin chủ yếu thông qua các giao diện phức tạp và yêu cầu kiến thức lập trình cũng như hiểu biết về API, thì giờ đây, có thể tương tác với dữ liệu bằng ngôn ngữ tự nhiên.

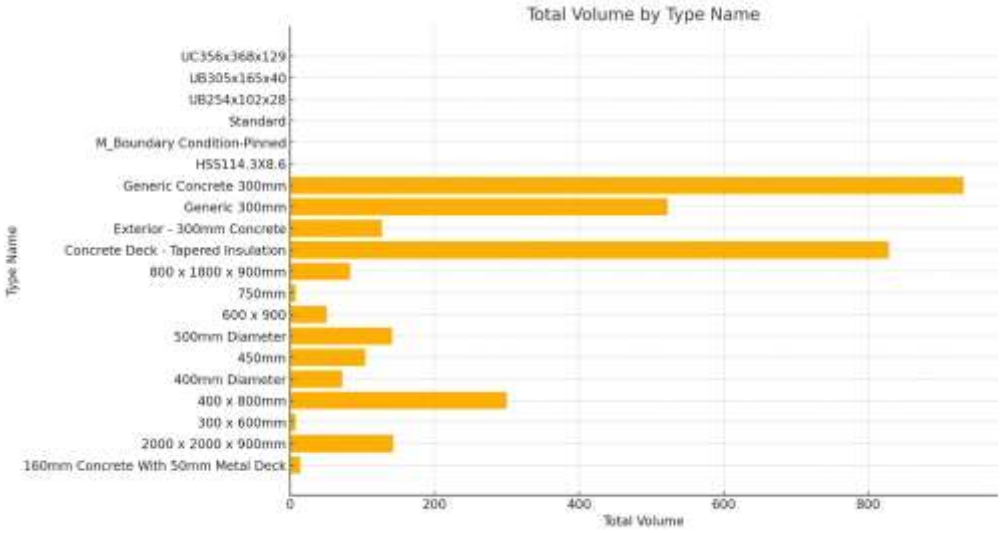
Các kỹ sư, nhà quản lý và nhà thiết kế không có nền tảng kỹ thuật có thể truy xuất thông tin cần thiết từ dữ liệu dự án bằng cách đặt câu hỏi bằng ngôn ngữ thông thường. Miễn là dữ liệu được cấu trúc và có sẵn (Hình 4.113), chỉ cần đặt trong trò chuyện LLM một câu hỏi như: "Hãy hiển thị dưới dạng bảng được nhóm theo loại tất cả các bức tường có thể tích lớn hơn 10 mét khối" - và mô hình sẽ tự động chuyển đổi yêu cầu này thành SQL hoặc mã trên Pandas, tạo ra bảng kết quả, đồ thị hoặc thậm chí một tài liệu hoàn chỉnh.

Dưới đây là một vài ví dụ thực tế về cách mà các mô hình LLM tương tác với dữ liệu dự án được trình bày trong các định dạng CAD (BIM) khác nhau.

- Ví dụ về yêu cầu trong trò chuyện LLM đối với dự án ở định dạng CAD RVT sau khi chuyển đổi (Hình 4.113) sang khung dữ liệu bảng (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ mô hình nào khác):-

> Nhóm dữ liệu trong Dataframe nhận được từ file RVT theo "Tên loại" khi tổng hợp tham số "Thể tích" và cho thấy số lượng thành phần trong nhóm. Và xin hãy hiển thị tất cả điều này dưới dạng biểu đồ thanh ngang mà không có giá trị bằng không.

Phản hồi của LLM dưới dạng biểu đồ thanh ngang (định dạng PNG):



Hình 6.45 Thay vì 17 lần nhấp chuột hoặc 40 dòng mã, thông qua việc sử dụng các plugin, chúng ta nhận được bảng QTO một cách ngay lập tức thông qua yêu cầu văn bản trong LLM.

Để tạo bảng QTO cho các loại tường với tổng diện tích và số lượng từ danh mục "Tường", chúng ta sẽ soạn thảo một yêu cầu văn bản cho LLM-chat:

> Lấy chỉ những phần tử từ dataframe dự án mà trong tham số "Danh mục" có "OST_Walls", nhóm chúng theo "Tên loại", tổng hợp giá trị của cột "Diện tích", thêm số lượng và hiển thị chúng trong bảng, xóa các giá trị bằng không.

Trả lời LLM dưới dạng bảng QTO đã hoàn thiện:

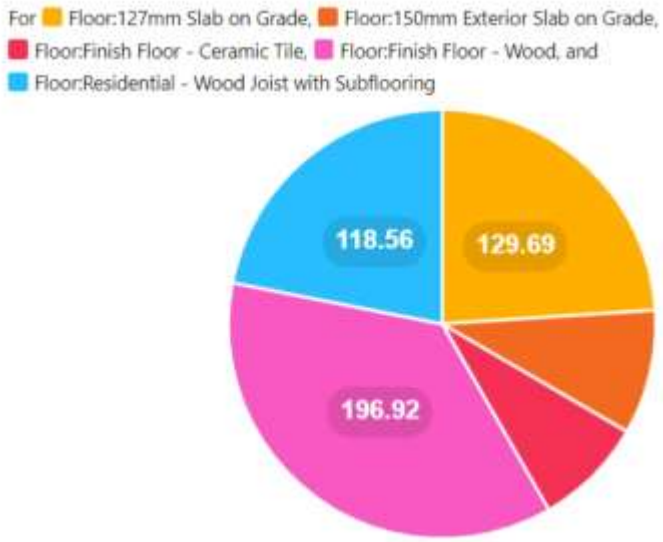
Type Name	Total Area	Count
CL_W1	393.12 sq m	10
Cavity wall_sliders	9.37 sq m	1
Foundation - 300mm Concrete	30.90 sq m	1
Interior - 165 Partition (1-hr)	17.25 sq m	3
Interior - Partition	186.54 sq m	14
Retaining - 300mm Concrete	195.79 sq m	10
SH_Curtain wall	159.42 sq m	9
SIP 202mm Wall - conc clad	114.76 sq m	4
Wall - Timber Clad	162.91 sq m	8

Hình 6.46 Việc tạo bảng QTO bằng ngôn ngữ tự nhiên cung cấp kết quả chất lượng tương tự như khi sử dụng các công cụ CAD- (BIM-).

Chúng ta sẽ thực hiện một yêu cầu đến dự án ở định dạng IFC sau khi chuyển đổi thành dataframe và nhập vào một yêu cầu văn bản tương tự trong bất kỳ LLM chat nào:

> Lấy chỉ những phần tử trong dự án có giá trị Level 1 và Level 2 trong tham số "Parent", và lấy các phần tử có giá trị IfcSlab trong tham số "Danh mục", sau đó nhóm những đối tượng này theo tham số "ObjectType", tổng hợp giá trị trong tham số "PSet_RVT_Dimensions Area" và hiển thị chúng dưới dạng biểu đồ hình tròn.

- Trả lời LLM dưới dạng biểu đồ hình tròn đã hoàn thiện của nhóm các phần tử từ dữ liệu IFC:



Hình 6.47 Kết quả của yêu cầu dữ liệu IFC trong định dạng cấu trúc có thể là bất kỳ loại biểu đồ nào thuận tiện cho việc hiểu dữ liệu.

Đằng sau mỗi giải pháp đã hoàn thiện (Hình 6.45 - Hình 6.47) là một tá dòng mã Python sử dụng thư viện Pandas. Mã thu được có thể được sao chép từ chat LLM và sử dụng trong bất kỳ IDE địa phương hoặc trực tuyến nào để đạt được kết quả giống hệt bên ngoài chat LLM.-

Trong cùng một chat LLM, chúng ta có thể làm việc không chỉ với các dự án lấy từ định dạng CAD 3D (BIM), mà còn với các bản vẽ 2D ở định dạng DWG, mà sau khi chuyển đổi sang định dạng có cấu trúc, chúng ta có thể đặt yêu cầu trong chat LLM để hiển thị, chẳng hạn như dữ liệu theo nhóm các phần tử dưới dạng đường hoặc hình học 3D.

Phân tích tự động hóa các tệp DWG với LLM và Pandas

Quy trình xử lý dữ liệu từ các tệp DWG do thông tin không có cấu trúc - luôn là một nhiệm vụ khó khăn, đòi hỏi phần mềm chuyên dụng và thường xuyên phải phân tích thủ công. Tuy nhiên, với sự phát triển của trí tuệ nhân tạo và các công cụ LLM, việc tự động hóa nhiều giai đoạn của quy trình chủ yếu thủ công này đã trở thành khả thi. Hãy xem xét một Pipeline thực tế từ các yêu cầu đến LLM (trong ví dụ này là ChatGPT) để làm việc với các bản vẽ DWG, giúp trong việc quản lý dự án:

- Lọc dữ liệu DWG theo các lớp, ID và tọa độ
- Hiển thị hình học của các phần tử
- Tự động ghi chú các bản vẽ dựa trên các tham số
- Mở rộng các đường biên tường thành mặt phẳng ngang
- Tạo các hình ảnh 3D tương tác từ dữ liệu 2D

- Cấu trúc và phân tích dữ liệu xây dựng mà không cần các công cụ CAD phức tạp

Trong trường hợp của chúng ta, quy trình xây dựng Pipeline bắt đầu từ việc tạo mã tuần tự thông qua LLM. Đầu tiên, một yêu cầu được tạo ra mô tả nhiệm vụ. ChatGPT sinh ra mã Python, mã này được thực thi và phân tích, hiển thị kết quả trong chat. Nếu kết quả không như mong đợi, yêu cầu sẽ được điều chỉnh và quy trình sẽ lặp lại.

> Pipeline — là một chuỗi các bước tự động được thực hiện để xử lý và phân tích dữ liệu. Trong quá trình này, mỗi bước nhận dữ liệu đầu vào, thực hiện các biến đổi và chuyển kết quả cho bước tiếp theo.

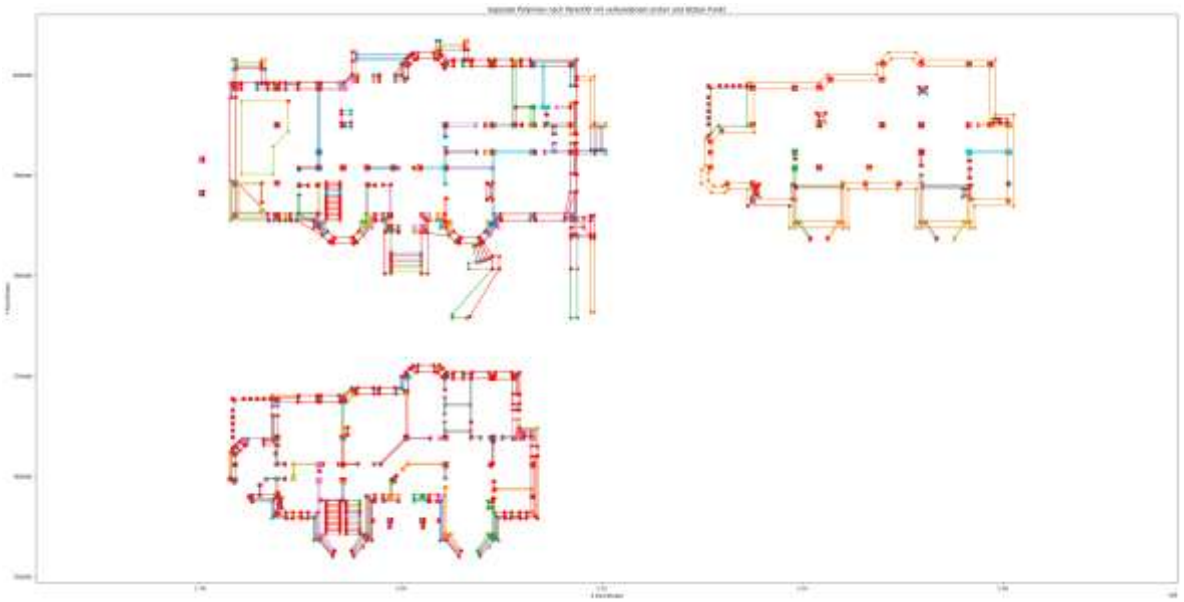
Sau khi có được kết quả mong muốn, mã được sao chép từ LLM và dán vào mã dưới dạng các khối trong bất kỳ IDE nào tiện lợi, trong trường hợp của chúng tôi là trên nền tảng Kaggle.com. Các đoạn mã thu được được kết hợp thành một Pipeline duy nhất, tự động hóa toàn bộ quá trình — từ việc tải dữ liệu cho đến phân tích cuối cùng. Cách tiếp cận này cho phép phát triển nhanh chóng và mở rộng quy trình phân tích mà không cần có chuyên môn sâu về lập trình. Mã hoàn chỉnh của tất cả các đoạn mã được chỉ ra dưới đây, cùng với các ví dụ truy vấn, bạn có thể tìm thấy trên nền tảng Kaggle.com với truy vấn «DWG Analyse with ChatGPT | DataDrivenConstruction» [129].

Chúng ta sẽ bắt đầu quá trình làm việc với dữ liệu DWG, sau khi chuyển đổi sang dạng có cấu trúc (Hình 4.113), từ bước cơ bản – nhóm và lọc tất cả dữ liệu của bản vẽ, cần thiết cho nhiệm vụ của chúng ta là các phần tử tường, cụ thể là các polyline (tham số 'ParentID' cho phép nhóm các dòng trong các nhóm), mà trong tham số (cột của dataframe) «Layer» có giá trị chuỗi chứa tổ hợp ký tự sau (RegEx) – «wall».

- Để có được mã cho nhiệm vụ như vậy và kết quả dưới dạng hình ảnh, cần viết truy vấn sau trong LLM:

> Đầu tiên, kiểm tra xem dataframe nhận được từ DWG có chứa các cột nhất định không: 'Layer', 'ID', 'ParentID' và 'Point'. Sau đó, lọc các định danh từ cột 'Layer' có chứa chuỗi 'wall'. Tìm trong cột 'ParentID' các phần tử tương ứng với các định danh này. Xác định hàm để làm sạch và phân tách dữ liệu trong cột 'Point'. Điều này bao gồm việc loại bỏ dấu ngoặc và phân tách các giá trị thành tọa độ 'x', 'y' và 'z'. Vẽ đồ thị dữ liệu bằng matplotlib. Đối với mỗi «ParentID» duy nhất, vẽ một polyline riêng biệt nối các tọa độ «Point». Đảm bảo rằng các điểm đầu và cuối được kết nối nếu có thể. Thiết lập các nhãn và tiêu đề thích hợp, đảm bảo tỉ lệ đều giữa các trục x và y.

- Phản hồi từ LLM sẽ xuất ra hình ảnh hoàn chỉnh mà ẩn sau đó là mã đã được tạo ra bằng ngôn ngữ Python:

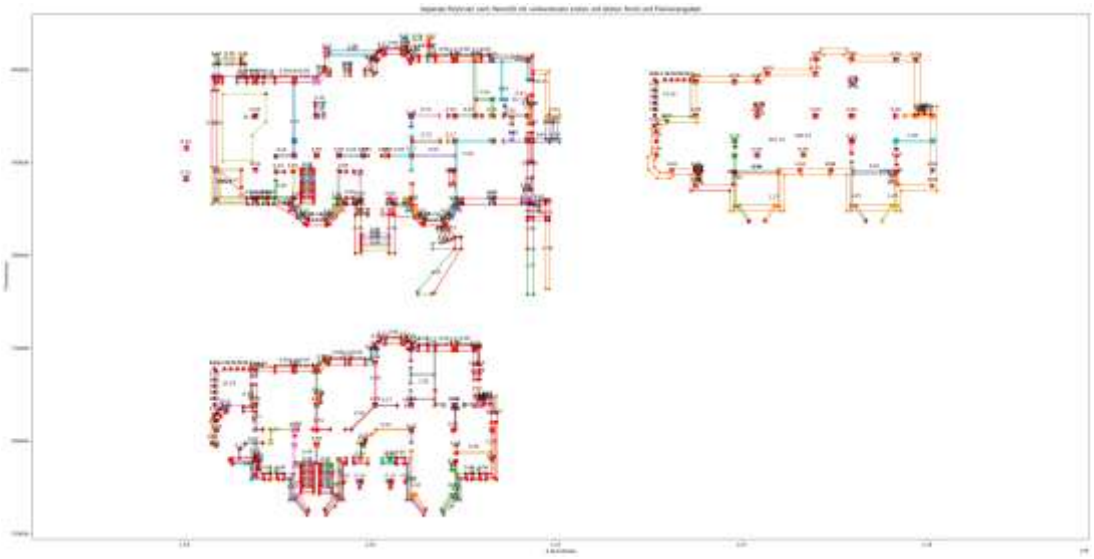


Hình 6.48 Mã LLM đã trích xuất từ tệp DWG tất cả các đường của lớp "wall", làm sạch tọa độ của chúng và vẽ các polyline bằng một trong những thư viện Python.

Bây giờ, chúng ta sẽ thêm vào các đường các tham số diện tích mà mỗi polyline có trong thuộc tính của nó (trong một trong các cột của dataframe):

> Bây giờ hãy lấy từ mỗi polyline chỉ một «ParentID» - tìm định danh này trong cột «ID», lấy giá trị «Area», chia cho 1.000.000 và thêm giá trị này vào đồ thị.

Phản hồi từ LLM sẽ hiển thị đồ thị mới, nơi mỗi polyline sẽ có một nhãn với diện tích của nó:

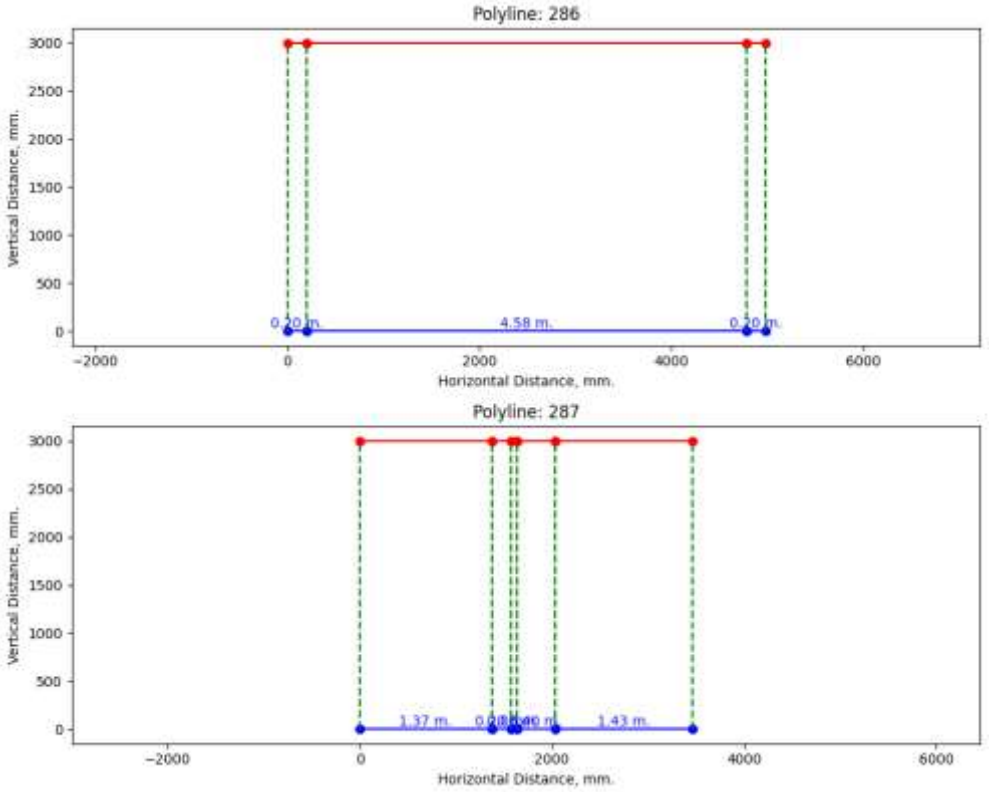


Hình 6.49 LLM đã bổ sung mã mà lấy các giá trị diện tích cho mỗi polyline và thêm nó vào hình ảnh với việc trực quan hóa các đường.

Tiếp theo, chúng ta sẽ chuyển đổi mỗi polyline thành một đường nằm ngang, thêm một đường song song ở độ cao 3000 mm và kết hợp chúng thành một mặt phẳng để thể hiện cách sắp xếp các bề mặt của các phần tử tường.

> Cần lấy tất cả các phần tử từ cột "Layer" có giá trị "tường". Lấy các ID này dưới dạng danh sách từ cột "ID" và tìm các ID này trong toàn bộ dataframe ở cột "ParentID". Tất cả các phần tử là các đường thẳng, được kết nối thành một đường polyline. Mỗi đường thẳng có hình học x, y của điểm đầu tiên trong cột "Point". Cần lấy từng polyline theo thứ tự và từ điểm 0,0 theo chiều ngang xây dựng chiều dài của mỗi đoạn từ polyline. Chiều dài của mỗi đoạn polyline thành một đường thẳng. Sau đó, vẽ các đường thẳng giống hệt ở độ cao 3000, kết nối tất cả các điểm thành một mặt phẳng.

Đáp án LLM sẽ xuất ra mã cho phép xây dựng các bản vẽ tường trong mặt phẳng:

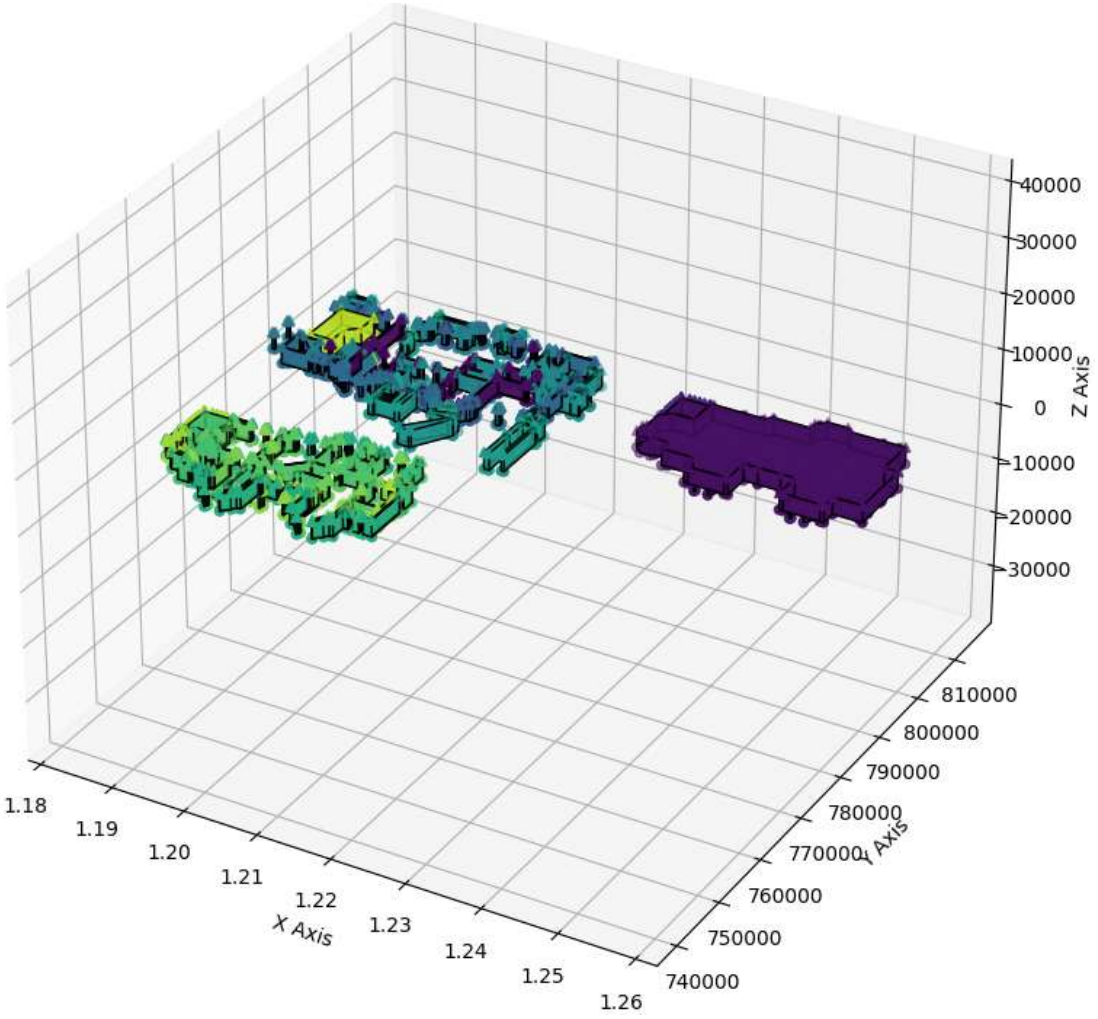


Hình 6.410 Mỗi polyline sẽ được biến đổi qua các prompt thành bản vẽ, trực quan thể hiện các mặt phẳng của tường ngay trong cuộc trò chuyện LLM.

Giờ hãy chuyển từ hình chiếu hai chiều sang mô hình 3D của tường từ các đường phẳng bằng cách kết nối các lớp trên và lớp dưới của các polyline:

> Hãy trực quan hóa các phần tử tường trong 3D, kết nối các polyline ở độ cao z = 0 và z = 3000 mm. Để tạo ra hình học khép kín, đại diện cho các bức tường của tòa nhà. Sử dụng đồ thị 3D Matplotlib.

- LLM sẽ tạo ra một biểu đồ 3D tương tác, trong đó mỗi polyline sẽ được thể hiện dưới dạng một tập hợp các mặt phẳng. Người dùng có thể di chuyển tự do giữa các phần tử bằng chuột máy tính, khám phá mô hình ở chế độ 3D, sao chép mã từ cuộc trò chuyện vào IDE:



Hình 6.411 LLM đã giúp xây dựng mã [129] để trực quan hóa các đường phẳng của bản vẽ ở chế độ 3D, có thể nghiên cứu trong trình xem 3D bên trong IDE.

Để xây dựng một Pipeline hợp lý và có thể tái sản xuất — từ việc chuyển đổi và tải lên file DWG ban đầu đến việc đạt được kết quả cuối cùng — khuyến nghị sau mỗi bước sao chép khối mã được tạo ra bởi LLM vào IDE. Như vậy, bạn không chỉ kiểm tra kết quả trong cuộc trò chuyện mà còn ngay lập tức chạy nó trong môi trường phát triển của bạn. Điều này cho phép xây dựng quy trình một cách tuần tự, gỡ lỗi và điều chỉnh theo nhu cầu.

Mã đầy đủ Pipeline của tất cả các đoạn (Hình 6.48 - Hình 6.411) cùng với các ví dụ yêu cầu có thể được tìm thấy trên nền tảng Kaggle.com với yêu cầu "DWG Analyse with ChatGPT | DataDrivenConstruction" [129]. Trên Kaggle bạn không chỉ có thể xem mã và các prompt được sử dụng, mà còn sao chép và thử nghiệm miễn phí toàn bộ Pipeline với các dataframe DWG gốc trong môi trường đám mây mà không cần cài đặt phần mềm bổ sung hoặc chính IDE.--

Phương pháp được trình bày trong chương này cho phép tự động hóa hoàn toàn việc kiểm tra, xử lý và tạo ra tài liệu dựa trên các dự án DWG. Pipeline được phát triển phù hợp cho cả việc xử lý các bản vẽ riêng lẻ và xử lý hàng loạt hàng chục, hàng trăm và hàng nghìn file DWG với việc tự động tạo ra các báo cáo và hình ảnh cần thiết cho từng dự án.

Quy trình có thể được thiết lập một cách tuần tự và minh bạch: đầu tiên, dữ liệu từ tệp CAD được tự động chuyển đổi sang định dạng XLSX, sau đó tải lên dataframe, sau đó thực hiện nhóm, kiểm tra và tạo kết quả - tất cả những điều này được thực hiện trong một Jupyter Notebook hoặc script Python, trong bất kỳ IDE phổ biến nào. Nếu cần, quy trình dễ dàng mở rộng thông qua việc tích hợp với các hệ thống quản lý tài liệu dự án: Các tệp CAD có thể được tự động trích xuất theo các tiêu chí đã chỉ định, trả lại kết quả vào hệ thống lưu trữ và thông báo cho người dùng về sự sẵn sàng của kết quả - qua email hoặc qua các ứng dụng nhắn tin.

> Việc sử dụng chat và tác nhân LLM để làm việc với dữ liệu dự án giảm thiểu sự phụ thuộc vào các chương trình CAD chuyên biệt và cho phép thực hiện phân tích và trực quan hóa các dự án kiến trúc mà không cần tương tác thủ công với giao diện - không cần nhấp chuột và ghi nhớ cách điều hướng phức tạp trong menu.

Ngày càng có nhiều tin tức về LLM, dữ liệu cấu trúc phân đoạn, DataFrame và cơ sở dữ liệu cột trong ngành xây dựng. Các DataFrame hai chiều thống nhất được hình thành từ các cơ sở dữ liệu và định dạng CAD khác nhau sẽ trở thành nguồn nhiên liệu lý tưởng cho các công cụ phân tích hiện đại mà các chuyên gia trong các lĩnh vực khác của nền kinh tế đang tích cực làm việc.

Quá trình tự động hóa sẽ trở nên đơn giản hơn rất nhiều - thay vì nghiên cứu API của các sản phẩm ngách đóng và viết các script phức tạp để phân tích hoặc chuyển đổi tham số, giờ đây chỉ cần định hình nhiệm vụ dưới dạng một tập hợp các lệnh văn bản riêng lẻ, chúng sẽ được kết hợp thành Pipeline hoặc quy trình Workflow cho ngôn ngữ lập trình cần thiết, có thể chạy miễn phí hầu như trên bất kỳ thiết bị nào. Không còn cần chờ đợi việc phát hành các sản phẩm, định dạng, plugin hoặc cập nhật mới từ các nhà cung cấp công cụ CAD (BIM). Các kỹ sư và nhà xây dựng sẽ có khả năng làm việc độc lập với dữ liệu, sử dụng các công cụ đơn giản, miễn phí và dễ hiểu, với sự trợ giúp của chat và tác nhân LLM.

Các bước tiếp theo: chuyển từ định dạng kín sang dữ liệu mở

Khi làm việc với dữ liệu dự án trong tương lai, không ai thực sự cần phải hiểu sâu về các lõi hình học của các công cụ độc quyền hoặc nghiên cứu hàng trăm định dạng không tương thích, chứa cùng một thông tin. Tuy nhiên, mà không có sự hiểu biết về lý do tại sao việc chuyển sang dữ liệu cấu trúc mở lại quan trọng, sẽ khó để lập luận sự cần thiết phải sử dụng các công cụ miễn phí mới, dữ liệu mở và các phương pháp mà rất có thể không được nhà cung cấp phần mềm thúc đẩy.

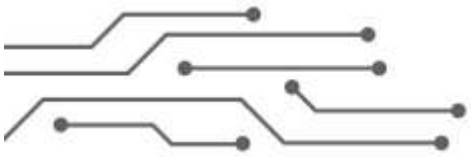
Trong chương này, chúng tôi đã xem xét những đặc điểm chính của dữ liệu CAD (BIM), những hạn chế và cơ hội của nó và rằng mặc dù có những lời hứa marketing từ các nhà cung cấp, các kỹ sư và nhà thiết kế hàng ngày đều gặp phải khó khăn trong việc trích xuất, truyền tải và phân tích thông tin dự án. Hiểu kiến trúc của những hệ thống này và làm quen với các phương pháp thay thế - dựa trên các định dạng mở và tự động hóa bằng cách sử dụng LLM - có thể đơn giản hóa đáng kể công việc của ngay cả một chuyên gia, chưa nói đến các công ty. Tóm tắt lại phần này, đáng để nêu bật những bước thực hành chính có thể giúp áp dụng những phương pháp đã đề cập trong các nhiệm vụ hàng ngày của bạn:

- Mở rộng công cụ của bạn để làm việc với dữ liệu dự án

- [ ] Tìm hiểu các plugin và tiện ích sẵn có để trích xuất dữ liệu từ các hệ thống CAD (BIM) mà bạn đang sử dụng
- [ ] Tìm hiểu các SDK và API có sẵn, cho phép tự động hóa việc trích xuất dữ liệu từ các định dạng đóng mà không cần phải mở thủ công các phần mềm chuyên dụng.
- [ ] Nắm vững kỹ năng cơ bản về làm việc với các định dạng hình học mở không tham số (OBJ, glTF, USD, DAE) và các thư viện mở tương ứng
- [ ] Cố gắng suy nghĩ về hệ thống lưu trữ siêu dữ liệu của dự án riêng biệt với hình học ngoài các giải pháp CAD (BIM) để đơn giản hóa việc phân tích và tích hợp với các hệ thống khác
- [ ] Sử dụng LLM để tự động hóa các câu hỏi về chuyển đổi dữ liệu giữa các định dạng

- [ ] Tạo ra các quy trình riêng cho việc xử lý thông tin dự án
  - [ ] Bắt đầu mô tả các nhiệm vụ và yêu cầu về mô hình hóa thông qua các tham số và giá trị của chúng trong các định dạng đơn giản và có cấu trúc
  - [ ] Tạo thư viện cá nhân các script hoặc đoạn mã cho các thao tác thường xuyên thực hiện
- [ ] Thúc đẩy việc sử dụng các tiêu chuẩn mở trong công việc của bạn
  - [ ] Đề xuất với đồng nghiệp và đối tác rằng họ nên trao đổi dữ liệu bằng các định dạng mở, không bị giới hạn bởi hệ sinh thái của các nhà cung cấp phần mềm
  - [ ] Chứng minh lợi ích của việc sử dụng dữ liệu có cấu trúc qua các ví dụ cụ thể
  - [ ] Khởi xướng các cuộc thảo luận về những vấn đề với các định dạng đóng và các giải pháp khả thi

Ngay cả khi bạn không thể thay đổi chính sách của công ty về các nền tảng CAD- (BIM-), việc hiểu biết cá nhân về các nguyên tắc làm việc với dữ liệu dự án trong các định dạng mở sẽ cho phép bạn tăng cường hiệu quả công việc của mình đáng kể. Bằng cách tạo ra các công cụ và phương pháp của riêng bạn để trích xuất và chuyển đổi dữ liệu từ các định dạng khác nhau, bạn không chỉ tối ưu hóa các quy trình làm việc mà còn có được sự linh hoạt cho phép bạn tránh những giới hạn của các giải pháp phần mềm tiêu chuẩn.



VII PHẦN

QUYẾT ĐỊNH DỰA TRÊN DỮ LIỆU, PHÂN TÍCH, TỰ ĐỘNG HÓA VÀ HỌC MÁY

Phần bảy tập trung vào phân tích dữ liệu và tự động hóa các quy trình trong ngành xây dựng. Tại đây, các dữ liệu được xem là nền tảng cho việc ra quyết định và các nguyên tắc trực quan hóa thông tin cho phân tích hiệu quả cũng được giải thích. Các chỉ số hiệu suất chính (KPI), phương pháp đánh giá hoàn vốn đầu tư (ROI) và việc tạo bảng điều khiển để giám sát các dự án được mô tả chi tiết. Sự chú ý đặc biệt được giành cho các quy trình ETL (Extract, Transform, Load) và việc tự động hóa chúng bằng cách sử dụng các pipeline, cho phép biến dữ liệu rời rạc thành thông tin có cấu trúc dễ phân tích. Các công cụ điều phối quy trình làm việc, chẳng hạn như Apache Airflow, Apache NiFi và n8n, cho phép xây dựng các pipeline dữ liệu tự động mà không cần kiến thức lập trình sâu. Vai trò quan trọng được giao cho các mô hình ngôn ngữ lớn (LLM) và ứng dụng của chúng trong việc đơn giản hóa phân tích dữ liệu và tự động hóa các công việc thường nhật.

CHƯƠNG 7.1.
PHÂN TÍCH DỮ LIỆU VÀ QUYẾT ĐỊNH DỰA TRÊN DỮ LIỆU

Sau các bước thu thập, cấu trúc, làm sạch và kiểm tra thông tin, một khối dữ liệu toàn diện và phù hợp cho phân tích đã được hình thành. Trong các phần trước của cuốn sách, đã xem xét việc hệ thống hóa và cấu trúc các nguồn dữ liệu đa dạng — từ tài liệu PDF và biên bản cuộc họp cho đến các mô hình CAD và dữ liệu hình học. Quá trình kiểm tra và điều chỉnh thông tin để phù hợp với yêu cầu của các hệ thống và phân loại khác nhau, loại bỏ các bản sao và mâu thuẫn được mô tả chi tiết.

Tất cả các phép tính được thực hiện trên những dữ liệu này (phần ba, phần bốn của cuốn sách) — từ các biến đổi đơn giản đến các tính toán thời gian, chi phí và các chỉ số ESG (phần năm) — đều là các bài toán phân tích quy mô lớn. Chúng tạo nền tảng để hiểu tình trạng hiện tại của dự án, đánh giá các tham số của nó và quyết định tiếp theo. Kết quả là, dữ liệu, thông qua các phép tính, trở thành từ một tập hợp các ghi chép rời rạc thành một nguồn tài nguyên có thể quản lý, có khả năng trả lời các câu hỏi then chốt của doanh nghiệp.

Trong các chương trước, quy trình thu thập dữ liệu và kiểm soát chất lượng của chúng để sử dụng trong các trường hợp kinh doanh điển hình và quy trình điển hình trong ngành xây dựng đã được xem xét chi tiết. Phân tích trong bối cảnh này phần nào tương tự như các ứng dụng trong các ngành khác, nhưng cũng có một số đặc điểm riêng.

Trong các chương tiếp theo, quá trình phân tích dữ liệu quy mô lớn sẽ được xem xét một cách chi tiết, bao gồm các giai đoạn tự động hóa — từ việc thu thập thông tin ban đầu đến việc chuyển tiếp vào các hệ thống và tài liệu mục tiêu. Đầu tiên sẽ được trình bày phần lý thuyết, liên quan đến một số khía cạnh của việc ra quyết định dựa trên dữ liệu. Sau đó, trong các chương tiếp theo, phần thực tiễn sẽ bắt đầu, liên quan đến tự động hóa và xây dựng ETL-Pipeline.

Dữ liệu như một nguồn lực trong quyết định

Quy trình ra quyết định dựa trên dữ liệu thường là một quá trình lặp đi lặp lại và bắt đầu từ việc thu thập thông tin từ nhiều nguồn khác nhau một cách có hệ thống. Giống như vòng tuần hoàn tự nhiên, các yếu tố dữ liệu riêng lẻ và các hệ thống thông tin toàn bộ dần dần chìm vào đất — tích lũy trong các kho dữ liệu của công ty (Hình 1.32). Theo thời gian, những dữ liệu này, giống như những chiếc lá và cành cây đã rụng, được biến đổi thành vật liệu quý giá. Sợi nấm của các kỹ sư dữ liệu và nhà phân tích tổ chức và chuẩn bị thông tin cho việc sử dụng trong tương lai, biến dữ liệu và hệ thống rụng thành phân quý giá, để trồng những chồi mới và các hệ thống mới (Hình 1.25).-

Xu hướng sử dụng rộng rãi phân tích trong nhiều ngành đánh dấu sự khởi đầu của một kỷ nguyên mới, nơi mà làm việc với dữ liệu trở thành nền tảng của hoạt động nghề nghiệp (Hình 7.11). Các chuyên gia trong ngành xây dựng cần thích nghi với những thay đổi này và sẵn sàng cho việc chuyển sang kỷ nguyên mới — kỷ nguyên của dữ liệu và phân tích.-

> Việc di chuyển dữ liệu thủ công giữa các bảng và thực hiện tính toán bằng tay dần dần lui vào quá khứ, nhường chỗ cho tự động hóa, phân tích luồng dữ liệu, phân tích và học máy. Những công cụ này trở thành những yếu tố then chốt trong hệ thống hỗ trợ ra quyết định hiện đại.

> Trong cuốn sách của McKinsey "Tái khởi động. Hướng dẫn McKinsey vượt qua cạnh tranh trong kỷ nguyên công nghệ số và trí tuệ nhân tạo" [130], có nghiên cứu được thực hiện vào năm 2022 với sự tham gia của 1.330 lãnh đạo cấp cao từ các khu vực, ngành nghề và chức năng khác nhau. Theo kết quả của nó, 70% lãnh đạo sử dụng phân tích tiên tiến để hình thành ý tưởng của riêng họ, trong khi 50% áp dụng trí tuệ nhân tạo để cải thiện và tự động hóa quy trình ra quyết định.



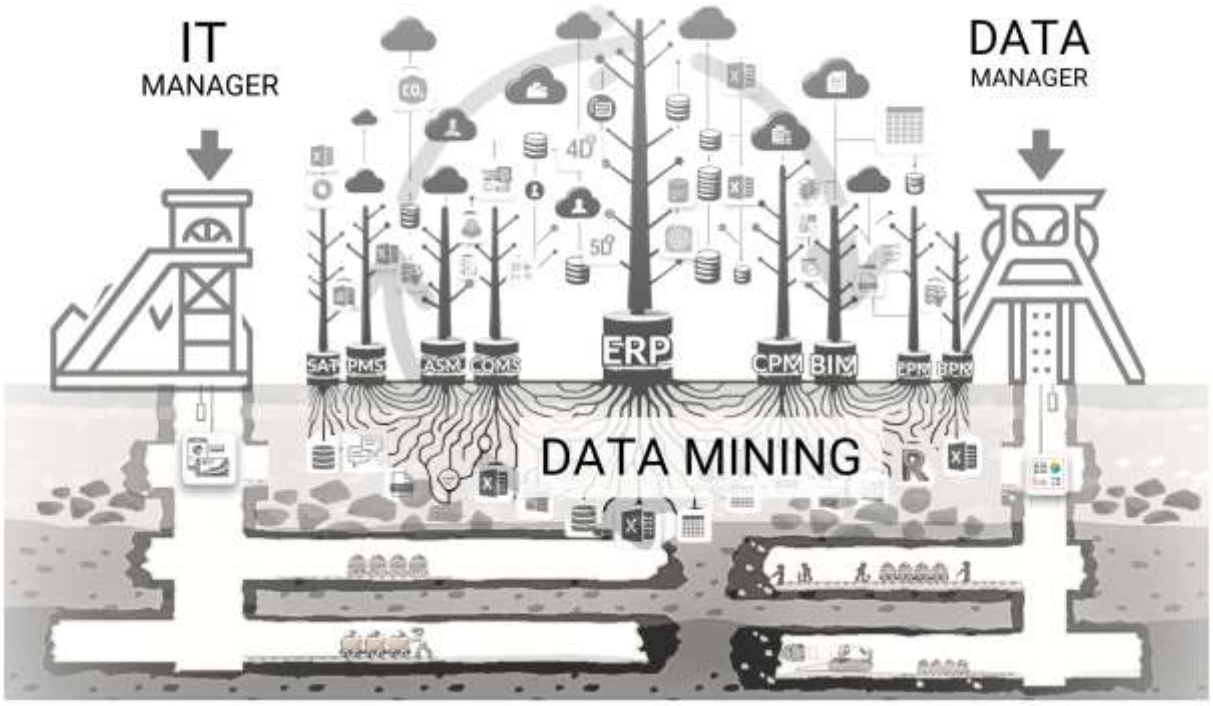
Hình 7.11 Phân tích dữ liệu và phân tích - công cụ chính để tăng tốc độ ra quyết định trong công ty.

Phân tích dữ liệu, giống như sự phát triển của nấm mycelium, xuyên qua lớp mùn của những quyết định trong quá khứ, giúp kết nối các hệ thống riêng lẻ và hướng dẫn các nhà quản lý tới những kiến thức quý giá. Những kiến thức này, giống như các chất dinh dưỡng từ những cây bị phân hủy trong hệ thống dữ liệu, nuôi dưỡng những quyết định mới trong công ty, dẫn đến những thay đổi hiệu quả và sự phát triển thông tin chất lượng, giống như những chồi non và cây con xuất hiện từ đất giàu và khỏe mạnh (Hình 1.25).-

> *Các con số có một câu chuyện quan trọng mà chúng phải kể. Chúng phụ thuộc vào việc bạn cho chúng một giọng nói rõ ràng và thuyết phục [131].*
>
> *– Steven Few, Chuyên gia về trực quan hóa dữ liệu*

Trong các công ty vừa và nhỏ, công việc khai thác và chuẩn bị thông tin để phân tích hiện nay là một quá trình cực kỳ tốn công sức (Hình 7.12), có thể so sánh với việc khai thác than vào thế kỷ XVIII. Cho đến gần đây, việc khai thác và chuẩn bị dữ liệu chủ yếu được dành cho những nhà thám hiểm làm việc trong một ngách hẹp với bộ công cụ nhỏ và hạn chế để làm việc với nhiều loại dữ liệu từ không có cấu trúc, yếu cấu trúc, hỗn hợp đến các nguồn khép kín.

> Các nhà quản lý và nhà điều hành ra quyết định thường thiếu kinh nghiệm trong làm việc với dữ liệu và hệ thống đa dạng, nhưng họ vẫn cần phải đưa ra quyết định dựa trên chúng. Do đó, việc ra quyết định dựa trên dữ liệu trong ngành xây dựng hiện đại trong vài thập kỷ qua ít giống như một quy trình tự động mà nhiều hơn như một công việc thủ công nhiều ngày của thợ mỏ trong những mỏ than đầu tiên.



Hình 7.12 Trong quá trình khai thác dữ liệu, các chuyên gia phải trải qua con đường phức tạp của việc chuẩn bị dữ liệu — từ việc làm sạch đến cấu trúc cho phân tích tiếp theo.

Mặc dù các phương pháp khai thác dữ liệu hiện đại trong ngành xây dựng chắc chắn tiên tiến hơn so với các kỹ thuật thô sơ của thợ mỏ thế kỷ XII, nhưng đây vẫn là một nhiệm vụ phức tạp và rủi ro cao, đòi hỏi nhiều tài nguyên và chuyên môn mà chỉ những công ty lớn mới có thể chi trả. Quá trình khai thác và phân tích dữ liệu từ di sản tích lũy của các dự án trong quá khứ cho đến gần đây chủ yếu được thực hiện bởi các công ty lớn có công nghệ phát triển, những công ty đã liên tục thu thập và lưu trữ dữ liệu trong nhiều thập kỷ.

> Trước đây, các công ty trưởng thành về công nghệ đóng vai trò hàng đầu trong phân tích, tích lũy dữ liệu trong nhiều thập kỷ. Ngày nay, tình hình đang thay đổi: việc tiếp cận dữ liệu và công cụ xử lý đang trở nên dân chủ hóa — những giải pháp phức tạp trước đây giờ đã có sẵn cho tất cả mọi người và miễn phí.

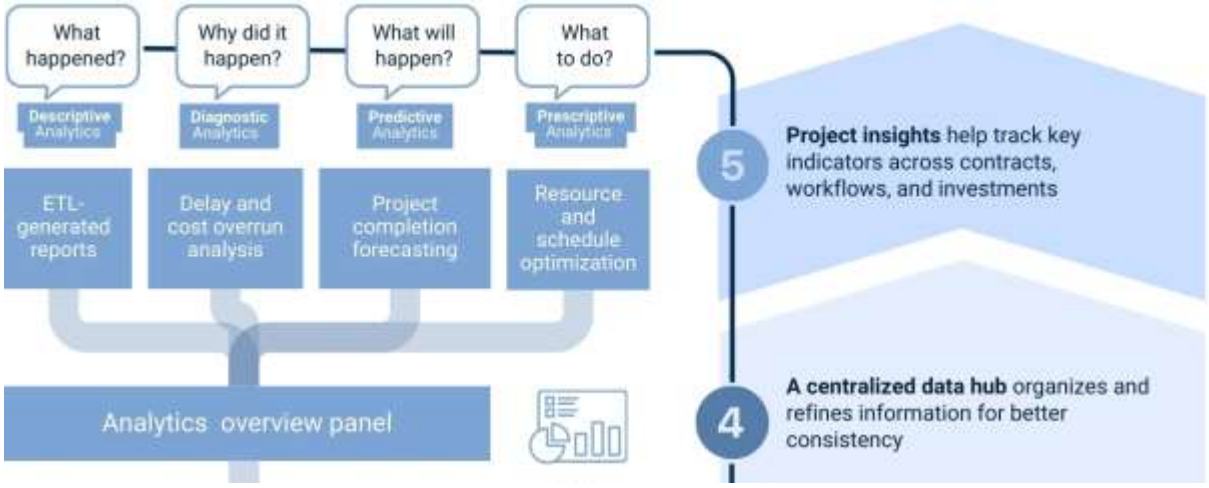
Việc áp dụng phân tích cho phép các công ty đưa ra những quyết định chính xác và có cơ sở hơn trong thời gian thực. Dưới đây là một ví dụ thực tiễn, minh họa cách mà dữ liệu lịch sử giúp đưa ra các quyết định tài chính có cơ sở:

- Nhà quản lý dự án — "Hiện tại giá bê tông trung bình trong thành phố là 82 €/m³, trong dự toán của chúng ta là 95 €/m³."
- Người lập dự toán — "Theo các dự án trước đây, mức vượt chi khoảng 15%, vì vậy tôi đã chuẩn bị sẵn."
- Nhà phân tích dữ liệu hoặc kỹ sư kiểm soát bên nhà thầu — "Chúng ta hãy xem phân tích về ba thầu gần nhất."

Sau khi phân tích DataFrame từ các dự án trước, chúng ta có được:

- **Giá mua thực tế trung bình: 84,80 €/m³**
- **Hệ số tiêu tốn trung bình: +4,7%**
- Tỷ lệ được khuyến nghị trong dự toán: ~85 €/m³

Giải pháp như vậy sẽ không dựa trên cảm nhận chủ quan, mà dựa trên thống kê lịch sử cụ thể, điều này cho phép giảm rủi ro và nâng cao tính hợp lý của tỷ lệ thầu. Phân tích dữ liệu từ các dự án trước trở thành loại "phân bón hữu cơ" từ đó nảy sinh các quyết định mới chính xác hơn.



Hình 7.13 Phân tích dữ liệu trả lời ba câu hỏi chính: điều gì đã xảy ra, tại sao điều đó xảy ra và chúng ta nên làm gì tiếp theo.

Các nhà lãnh đạo và quản lý ra quyết định thường gặp phải nhu cầu làm việc với dữ liệu và hệ thống đa dạng, nhưng không có đủ kiến thức kỹ thuật. Trong những tình huống như vậy, một trong những trợ thủ quan trọng trong quá trình hiểu dữ liệu là trực quan hóa — một trong những giai đoạn đầu tiên và quan trọng nhất của quá trình phân tích. Nó cho phép trình bày thông tin một cách trực quan và dễ hiểu.

Trực quan hóa dữ liệu: chìa khóa để hiểu và ra quyết định

Trong ngành xây dựng hiện đại, nơi dữ liệu dự án có tính phức tạp và cấu trúc nhiều tầng, việc trực quan hóa đóng vai trò quan trọng. Trực quan hóa dữ liệu cho phép các nhà quản lý dự án và kỹ sư hình dung các quy luật và xu hướng phức tạp, ẩn chứa trong khối lượng dữ liệu lớn và đa dạng.

> Trực quan hóa dữ liệu giúp dễ dàng hiểu tình trạng của dự án: phân bổ nguồn lực, động thái chi phí hoặc sử dụng vật liệu. Nhờ đồ thị và biểu đồ, thông tin phức tạp và khô khan trở nên dễ tiếp cận và hiểu biết, cho phép nhanh chóng xác định các lĩnh vực chính cần chú ý và phát hiện các vấn đề tiềm năng.

Trực quan hóa dữ liệu không chỉ đơn thuần giúp việc giải thích thông tin, mà còn là một giai đoạn quan trọng trong quá trình phân tích và ra quyết định quản lý hợp lý, giúp trả lời các câu hỏi "điều gì đã xảy ra?" và "nó

đã xảy ra như thế nào?" (Hình 2.25).-

> *Đồ họa — là phương tiện trực quan để giải quyết các bài toán logic.— Jacques Bertin, "Đồ họa và xử lý thông tin đồ họa"*

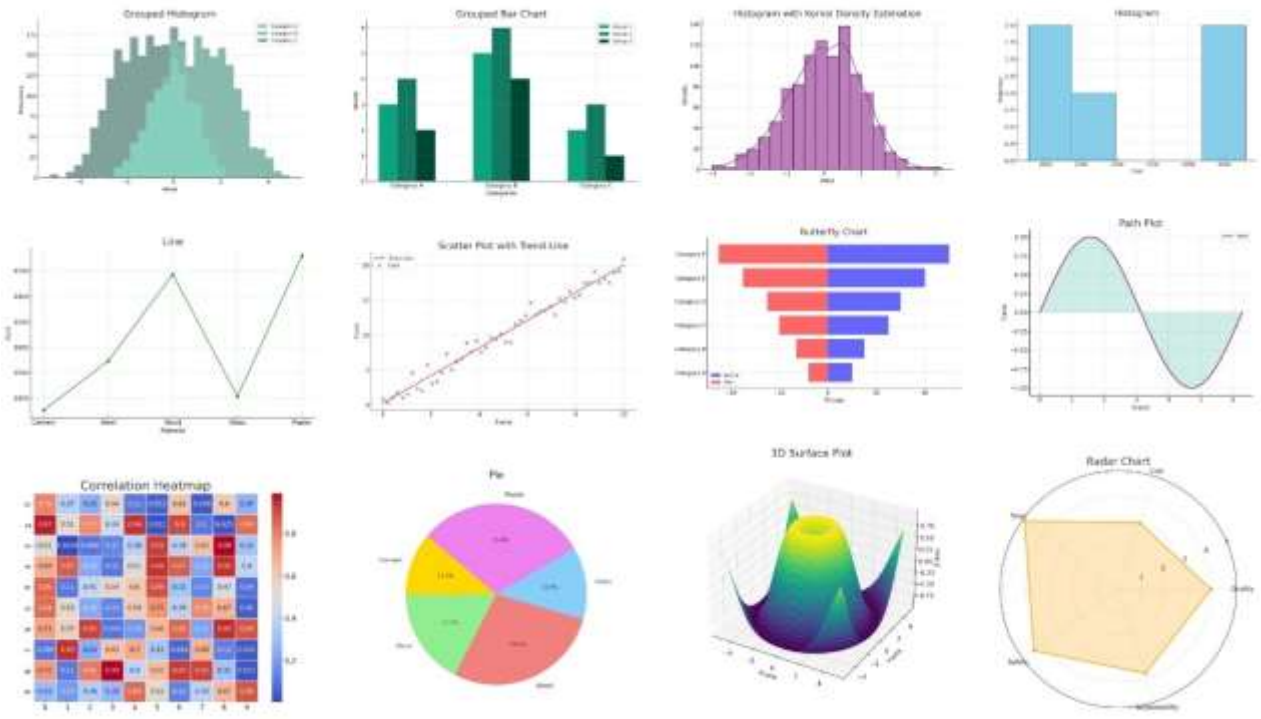
Trước khi đưa ra quyết định quan trọng, các nhà quản lý dự án có nhiều khả năng sẽ sử dụng các biểu diễn trực quan của dữ liệu hơn là những con số khô khan và khó hiểu từ bảng tính điện tử hoặc tin nhắn văn bản.

> Dữ liệu không được hỗ trợ bởi trực quan hóa giống như vật liệu xây dựng, được rải rác lộn xộn trên công trường: tiềm năng của chúng không rõ ràng. Chỉ khi chúng xuất hiện dưới dạng trực quan rõ ràng, giống như từ gạch và bê tông — một ngôi nhà, thì mới rõ ràng về giá trị mà chúng đại diện. Trước khi ngôi nhà được xây dựng, không thể nói rằng đống vật liệu đó sẽ trở thành một căn nhà nhỏ, một biệt thự sang trọng hay một tòa nhà chọc trời.

Các công ty đang nắm giữ dữ liệu từ nhiều hệ thống khác nhau (Hình 1.24 - Hình 2.110), các giao dịch tài chính và hàng loạt dữ liệu văn bản. Tuy nhiên, việc sử dụng những dữ liệu này vì lợi ích của doanh nghiệp thường là một nhiệm vụ phức tạp. Trong những tình huống như vậy, việc trực quan hóa trở thành công cụ quan trọng để truyền tải ý nghĩa của dữ liệu, giúp trình bày thông tin dưới các định dạng dễ hiểu cho bất kỳ chuyên gia nào, chẳng hạn như bảng điều khiển, biểu đồ và đồ thị.-

> Nghiên cứu của PwC "Những gì sinh viên cần để thành công trong thế giới kinh doanh thay đổi nhanh chóng" (2015) nhấn mạnh rằng [9], các công ty thành công không chỉ giới hạn trong việc phân tích dữ liệu mà còn sử dụng tích cực các công cụ trực quan hóa tương tác như biểu đồ, infographic và bảng phân tích để hỗ trợ quyết định. Theo báo cáo - trực quan hóa dữ liệu giúp khách hàng hiểu câu chuyện mà dữ liệu đang kể thông qua các biểu đồ, đồ thị, bảng điều khiển và các mô hình dữ liệu tương tác.

Quá trình chuyển đổi thông tin thành các hình thức đồ họa trực quan, chẳng hạn như biểu đồ, đồ thị, cải thiện khả năng hiểu và diễn giải dữ liệu của bộ não con người. Điều này cho phép các nhà quản lý dự án và các nhà phân tích đánh giá nhanh chóng các kịch bản phức tạp và đưa ra quyết định có thông tin dựa trên các xu hướng và mô hình dễ nhận diện từ trực quan.



Hình 7.14 Các loại hình trực quan khác nhau đã được tạo ra để giúp bộ não con người hiểu và suy nghĩ tốt hơn về những thông tin khô khan của các con số.

Các vấn đề về việc tạo ra các trực quan từ dữ liệu và sử dụng các thư viện trực quan hóa miễn phí khác nhau sẽ được bàn luận trong chương tiếp theo, chuyên về quy trình ETL.

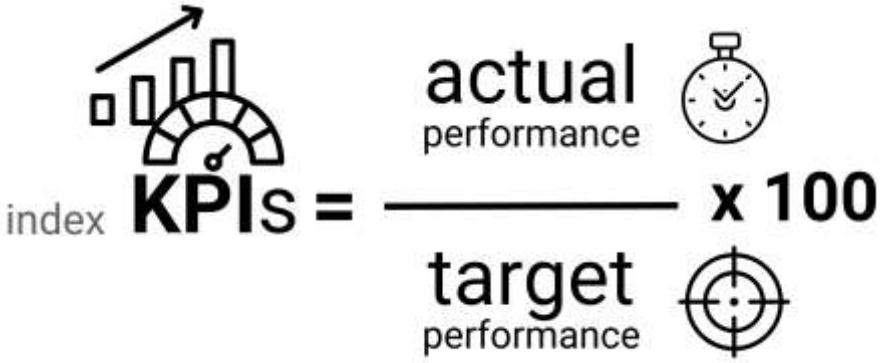
Trực quan hóa trở thành một phần không thể thiếu trong công việc với dữ liệu trong ngành xây dựng - nó giúp không chỉ "thấy" dữ liệu mà còn hiểu được ý nghĩa của chúng trong bối cảnh các nhiệm vụ quản lý. Tuy nhiên, để trực quan hóa thực sự hữu ích, cần xác định trước những gì cần được trực quan hóa và các chỉ số nào thực sự quan trọng để đánh giá hiệu quả của dự án. Tại đây, các chỉ số hiệu suất như KPI và ROI sẽ được đưa vào bàn. Nếu không có chúng, ngay cả những bảng điều khiển đẹp nhất cũng có nguy cơ chỉ còn lại "tín hiệu nhiễu".

Các chỉ số hiệu suất KPI và ROI

Trong ngành xây dựng hiện đại, việc quản lý các chỉ số hiệu suất (KPI và ROI), cũng như việc trực quan hóa chúng thông qua báo cáo và bảng điều khiển (dashboards) đóng vai trò then chốt trong việc nâng cao năng suất và hiệu quả quản lý dự án.

> Giống như trong bất kỳ doanh nghiệp nào, trong xây dựng cần xác định rõ các chỉ số mà qua đó đánh giá thành công, khả năng thu hồi vốn và năng suất. Nhận dữ liệu từ các quy trình khác nhau, tổ chức hoạt động dựa trên dữ liệu trước tiên phải học cách xác định các KPI chính (Key Performance Indicators) - các chỉ số định lượng phản ánh mức độ đạt được các mục tiêu chiến lược và vận hành.

Để tính toán KPI thường sử dụng công thức, bao gồm các chỉ số thực tế và chỉ số kế hoạch. Ví dụ, để tính toán KPI cá nhân cho một dự án, nhân viên hoặc quy trình, cần chia các chỉ số thực tế cho các chỉ số đã lập kế hoạch và nhân kết quả thu được với 100%.-



KPI được sử dụng để đo lường thành công của dự án hoặc quy trình trong việc đạt được các mục tiêu chính.

Ở cấp độ công trường xây dựng, có thể sử dụng các chỉ số KPI chi tiết hơn:

- Thời gian hoàn thành các giai đoạn chính (móng, lắp đặt, hoàn thiện) - cho phép kiểm soát việc tuân thủ kế hoạch công việc.
- Tỷ lệ vượt chi phí vật liệu - giúp quản lý việc mua sắm và giảm thiểu lãng phí.
- Số lượng thời gian ngừng hoạt động không theo kế hoạch của máy móc - ảnh hưởng đến năng suất và chi phí.

Việc chọn chỉ số không phù hợp có thể dẫn đến quyết định sai lầm về việc "cần làm gì?" (Hình 2.25). Ví dụ, nếu công ty chỉ tập trung vào chi phí mỗi mét vuông, nhưng không xem xét chi phí cho việc điều chỉnh, việc tiết kiệm vật liệu có thể dẫn đến giảm chất lượng và tăng chi phí trong các dự án tương lai.

Khi đặt ra mục tiêu, điều quan trọng là xác định rõ ràng cái gì cụ thể được đo lường. Những diễn đạt mờ mịt dẫn đến những kết luận không chính xác và làm phức tạp việc kiểm soát. Hãy xem xét các ví dụ về KPI thành công và không thành công trong xây dựng.

KPI tốt:

- "Đến cuối năm, giảm tỷ lệ điều chỉnh công việc hoàn thiện xuống 10%"
- "Tăng tốc độ lắp đặt mặt tiền lên 15% mà không giảm chất lượng vào quý sau"
- "Giảm thời gian ngưng hoạt động của máy móc xuống 20% thông qua việc tối ưu hóa lịch làm việc vào cuối năm"

Những chỉ số này có thể đo lường một cách rõ ràng, có giá trị cụ thể và khung thời gian.

KPI kém:

- "Chúng tôi sẽ xây dựng nhanh hơn" (Nhanh hơn bao nhiêu? "Nhanh hơn" có nghĩa là gì?)
- "Chúng tôi sẽ nâng cao chất lượng công trình bê tông" (Chất lượng được đo lường như thế nào?)

- "Chúng tôi sẽ cải thiện sự phối hợp của các nhà thầu trên công trường" (Tiêu chí nào sẽ cho thấy sự cải thiện?)

> Một KPI tốt là cái mà có thể đo lường và đánh giá một cách khách quan. Trong xây dựng, điều này đặc biệt quan trọng, vì nếu không có các chỉ số rõ ràng sẽ không thể kiểm soát hiệu quả làm việc và đạt được kết quả ổn định.

Ngoài KPI, còn có một chỉ số khác để đánh giá hiệu quả đầu tư: ROI (Return on Investment) - chỉ số hoàn vốn, phản ánh tỷ lệ giữa lợi nhuận và tiền đầu tư. ROI cho phép đánh giá mức độ hợp lý của việc áp dụng các phương pháp, công nghệ hoặc công cụ mới: từ các giải pháp số và tự động hóa (ví dụ Hình 7.32) đến việc sử dụng các vật liệu xây dựng mới. Chỉ số này giúp đưa ra quyết định có căn cứ về các khoản đầu tư tiếp theo, dựa trên ảnh hưởng thực tế của chúng đến tính sinh lời của doanh nghiệp.-

> Trong bối cảnh quản lý dự án xây dựng, ROI (hoàn vốn) có thể được sử dụng như một trong những chỉ số chính về hiệu suất (KPI), nếu mục tiêu của công ty là đánh giá khả năng thu hồi vốn đầu tư vào dự án, công nghệ hoặc cải tiến quy trình. Ví dụ, nếu một phương pháp quản lý xây dựng mới được triển khai, ROI có thể cho thấy nó đã tăng cường tính sinh lời như thế nào.

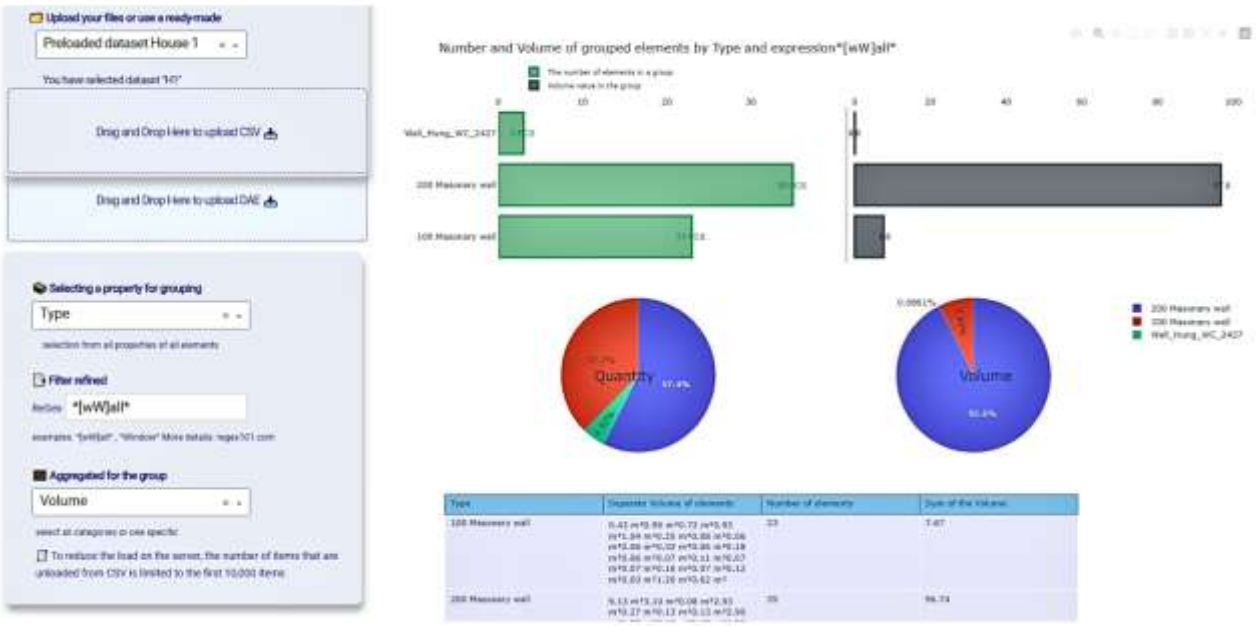
Việc đo lường thường xuyên các chỉ số KPI và ROI dựa trên dữ liệu thu thập từ nhiều nguồn khác nhau, chẳng hạn như chi phí vật liệu, giờ làm việc và chi phí, cho phép ban quản lý dự án quản lý tài nguyên một cách hiệu quả và nhanh chóng đưa ra quyết định. Lưu trữ các dữ liệu này trong dài hạn cho phép phân tích các xu hướng tương lai và tối ưu hóa quy trình.

Để trực quan hóa KPI, ROI và các chỉ số khác, nhiều loại biểu đồ và đồ thị khác nhau được sử dụng, thường được tổng hợp lại trong các bảng điều khiển.

Bảng thông tin và bảng điều khiển: trực quan hóa các chỉ số để quản lý hiệu quả

Để trực quan hóa các chỉ số và số liệu, nhiều loại biểu đồ và đồ thị phong phú được sử dụng, thường được tổng hợp trong các kho dữ liệu và bảng điều khiển (dashboards). Những bảng điều khiển này cung cấp cái nhìn tập trung về trạng thái của dự án hoặc các phần riêng lẻ của nó, hiển thị các chỉ số chính (lý tưởng là trong thời gian thực). Các bảng điều khiển hiện tại và được cập nhật liên tục cho phép nhóm nhanh chóng phản ứng với các thay đổi.

> Bảng điều khiển là công cụ giúp trực quan hóa các đánh giá định lượng, làm cho chúng dễ tiếp cận và hiểu rõ đối với tất cả các thành viên trong dự án.



Hình 7.16 Quản lý KPI và trực quan hóa chúng thông qua các bảng điều khiển - chìa khóa để nâng cao hiệu suất và hiệu quả của dự án.

Dưới đây là một số ví dụ về các công cụ phổ biến, nơi có thể tạo ra các bảng điều khiển:

- Power BI - công cụ của Microsoft để tạo các báo cáo tương tác và bảng điều khiển.
- Tableau và Google Data Studio - các công cụ mạnh mẽ cho việc trực quan hóa dữ liệu và tạo bảng điều khiển mà không cần viết mã.
- Plotly (Hình 7.16, Hình 7.212) - là một thư viện để tạo các biểu đồ tương tác, và Dash - framework để tạo các ứng dụng web cho phân tích dữ liệu. Chúng có thể được sử dụng kết hợp để tạo ra các bảng điều khiển tương tác.--
- Nhiều thư viện Python (Hình 7.29 - Hình 7.211) - trong Python có rất nhiều thư viện mở và miễn phí để trực quan hóa dữ liệu như Matplotlib, Seaborn, Plotly, Bokeh và nhiều thư viện khác. Chúng có thể được sử dụng để tạo các biểu đồ và tích hợp chúng vào ứng dụng web thông qua các framework như Flask hoặc Django.--
- Thư viện JavaScript: cho phép tạo các bảng điều khiển tương tác bằng cách sử dụng các thư viện JavaScript mã nguồn mở như D3.js hoặc Chart.js, và tích hợp chúng vào ứng dụng web.

> Để đánh giá KPI và tạo ra các bảng điều khiển, cần có dữ liệu hiện tại và một lịch trình rõ ràng cho việc thu thập và phân tích thông tin.

Tổng thể, KPI, ROI và bảng điều khiển trong ngành xây dựng hình thành nền tảng cho cách tiếp cận phân tích trong quản lý dự án. Chúng không chỉ giúp theo dõi và đánh giá trạng thái hiện tại, mà còn cung cấp thông tin quý giá cho việc lập kế hoạch trong tương lai và tối ưu hóa quy trình - những quy trình phụ thuộc trực tiếp vào việc giải thích dữ liệu và khả năng đặt ra các câu hỏi đúng và kịp thời.

Phân tích dữ liệu và nghệ thuật đặt câu hỏi

Giải thích dữ liệu là giai đoạn cuối cùng của phân tích, nơi thông tin có được ý nghĩa và bắt đầu "nói". Đây là nơi các câu trả lời cho các câu hỏi then chốt được đưa ra: "phải làm gì?" và "làm như thế nào?" (Hình 2.25). Giai đoạn này cho phép tổng quát hóa kết quả, phát hiện các quy luật, thiết lập các mối quan hệ nhân quả và rút ra kết luận dựa trên việc trực quan hóa và phân tích thống kê.

> *Có thể không còn xa nữa, thời điểm mà việc nhận thức rằng để trở thành một công dân hiệu quả trong một trong những quốc gia phức tạp mới lớn đang phát triển, cũng cần có khả năng tính toán, suy nghĩ bằng các đại lượng trung bình, cực đại và cực tiểu, giống như việc hiện nay phải biết đọc và viết.*
>
> *– Samuel C. Wilks, trích dẫn từ bài phát biểu của Tổng thống vào năm 1951 trước Hiệp hội Thống kê Hoa Kỳ*

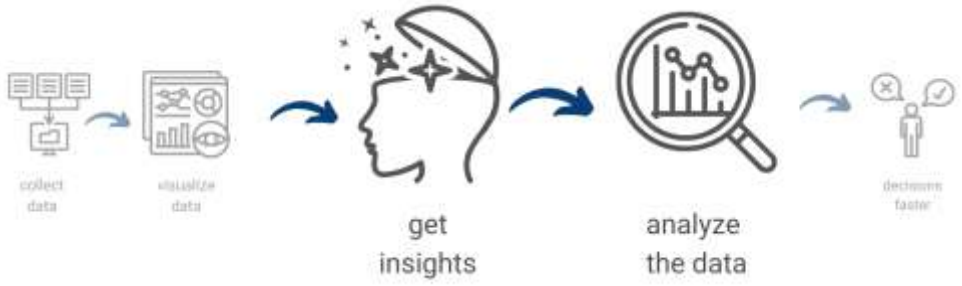
> Theo báo cáo được Chính phủ Vương quốc Anh công bố có tiêu đề "Phân tích dữ liệu và trí tuệ nhân tạo trong thực hiện các dự án chính phủ" (2024) [83], việc áp dụng phân tích dữ liệu và trí tuệ nhân tạo (AI) có thể cải thiện đáng kể quy trình quản lý dự án, tăng độ chính xác trong việc dự đoán thời gian và chi phí, cũng như giảm thiểu rủi ro và sự không chắc chắn. Tài liệu nhấn mạnh rằng các tổ chức nhà nước sử dụng các công cụ phân tích tiên tiến đạt được hiệu suất cao hơn trong việc thực hiện các sáng kiến hạ tầng.

Ngành xây dựng hiện đại, hoạt động trong bối cảnh cạnh tranh khốc liệt và biên lợi nhuận thấp trong cuộc cách mạng công nghiệp lần thứ tư, có thể so sánh với chiến tranh. Tại đây, sự sống sót và thành công của công ty phụ thuộc vào tốc độ thu thập tài nguyên và thông tin chất lượng — tức là, vào việc ra quyết định kịp thời và hợp lý (Hình 7.17).

Nếu trực quan hóa dữ liệu là "trinh sát", cung cấp cái nhìn tổng quan, thì phân tích dữ liệu là "đạn dược", cần thiết cho hành động. Chính nó trả lời các câu hỏi: cần làm gì? và làm như thế nào?, tạo ra nền tảng cho việc đạt được lợi thế cạnh tranh trên thị trường.

> Phân tích biến đổi dữ liệu rời rạc thành thông tin có cấu trúc và có nội dung, trên cơ sở đó quyết định được đưa ra.

Nhiệm vụ của các nhà phân tích và quản lý không chỉ là diễn giải thông tin, mà còn đề xuất các quyết định dựa trên cơ sở, phát hiện xu hướng, xác định mối quan hệ giữa các loại dữ liệu khác nhau và phân loại chúng theo mục tiêu và đặc thù của dự án. Sử dụng các công cụ trực quan hóa và phương pháp phân tích thống kê, họ biến dữ liệu thành tài sản chiến lược của công ty.



Hình 7.17 Chính phân tích dữ liệu cuối cùng chuyển đổi thông tin đã thu thập thành nguồn cho việc ra quyết định.

Để ra quyết định thực sự dựa trên cơ sở trong quá trình phân tích, cần học cách đặt câu hỏi đúng về dữ liệu. Chất lượng của những câu hỏi này ảnh hưởng trực tiếp đến độ sâu của những hiểu biết thu được và, do đó, đến chất lượng của các quyết định quản lý.

> *Quá khứ chỉ tồn tại bởi vì nó hiện hữu trong những ghi chép của ngày hôm nay. Và điều mà những ghi chép này thể hiện phụ thuộc vào những câu hỏi mà chúng ta đặt ra. Không có câu chuyện nào khác ngoài câu chuyện này [134].*
>
> *– John Archibald Wheeler, nhà vật lý năm 1982*

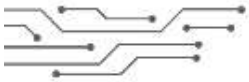
Nghệ thuật đặt ra những câu hỏi sâu sắc và tư duy phản biện là kỹ năng quan trọng nhất khi làm việc với dữ liệu. Hầu hết mọi người có xu hướng đặt ra những câu hỏi đơn giản, nông cạn, mà câu trả lời không đòi hỏi nhiều nỗ lực. Tuy nhiên, phân tích thực sự bắt đầu từ những câu hỏi có nội dung và được suy nghĩ kỹ lưỡng, có khả năng khám phá các mối quan hệ ẩn và các mối liên hệ nhân quả trong thông tin, có thể ẩn mình sau nhiều lớp lập luận.

> Theo nghiên cứu "Chuyển đổi dữ liệu: Thúc đẩy quy mô ngay bây giờ" (BCG, 2017) [135], quá trình chuyển đổi số thành công đòi hỏi các khoản đầu tư vào năng lực phân tích, chương trình quản lý thay đổi và sự phối hợp giữa mục tiêu kinh doanh và các sáng kiến CNTT. Các công ty xây dựng văn hóa định hướng dữ liệu cần đầu tư vào khả năng sử dụng dữ liệu phân tích và triển khai các chương trình quản lý thay đổi để khuyến khích tư duy, hành vi và cách thức làm việc mới.

Nếu không đầu tư vào việc phát triển văn hóa phân tích, cải thiện công cụ làm việc với dữ liệu và đào tạo nhân viên, các công ty trong tương lai có nguy cơ đưa ra quyết định dựa trên thông tin lỗi thời hoặc không đầy đủ — hoặc dựa vào các ý kiến chủ quan của các nhà quản lý HiPPO (Hình 2.19).

> Nhận thức về sự cần thiết và tầm quan trọng của việc liên tục cập nhật phân tích và bảng thông tin không thể tránh khỏi dẫn đến sự hiểu biết của ban lãnh đạo về tầm quan trọng của việc tự động hóa các quy trình phân tích. Tự động hóa nâng cao tốc độ ra quyết định, giảm thiểu ảnh hưởng của yếu tố con người và đảm bảo tính chính xác của dữ liệu. Trong bối cảnh khối lượng thông tin tăng trưởng theo cấp số nhân, tốc độ không chỉ trở thành lợi thế cạnh tranh mà còn là yếu tố then chốt cho sự thành công bền vững.

Tự động hóa quy trình phân tích và xử lý dữ liệu liên quan chặt chẽ đến chủ đề ETL (Extract, Transform, Load). Cũng giống như trong quá trình tự động hóa, chúng ta cần chuyển đổi dữ liệu, trong quá trình ETL dữ liệu được trích xuất từ các nguồn khác nhau, được chuyển đổi theo các yêu cầu cần thiết và được tải vào các hệ thống mục tiêu để sử dụng tiếp theo.



CHƯƠNG 7.2. **DÒNG DỮ LIỆU KHÔNG CẦN NỖ LỰC THỦ CÔNG: TẠI SAO CẦN ETL**

Tự động hóa ETL: giảm chi phí và tăng tốc độ làm việc với dữ liệu

Khi các chỉ số hiệu suất chính (KPI) không còn tăng trưởng, mặc dù có sự gia tăng về khối lượng dữ liệu và số lượng đội ngũ, ban lãnh đạo các công ty không thể không nhận ra sự cần thiết của việc tự động hóa các quy trình. Sớm hay muộn, sự nhận thức này trở thành động lực để khởi động tự động hóa toàn diện, với mục tiêu chính là giảm bớt độ phức tạp của các quy trình, tăng tốc độ xử lý và giảm sự phụ thuộc vào yếu tố con người.

> Theo nghiên cứu của McKinsey "Làm thế nào để xây dựng kiến trúc dữ liệu để thúc đẩy đổi mới - hôm nay và ngày mai" (2022) [136], các công ty sử dụng kiến trúc dữ liệu theo luồng nhận được lợi thế đáng kể, vì họ có thể phân tích thông tin theo thời gian thực. Công nghệ luồng cho phép phân tích trực tiếp các thông điệp theo thời gian thực và áp dụng bảo trì dự đoán trong sản xuất nhờ vào việc phân tích dữ liệu cảm biến theo thời gian thực.

> Đơn giản hóa quy trình là tự động hóa, trong đó các chức năng quản lý thủ công truyền thống được thay thế bằng các thuật toán và hệ thống.

Vấn đề tự động hóa, hay chính xác hơn, "giảm thiểu vai trò của con người trong xử lý dữ liệu", là một quá trình không thể đảo ngược và cực kỳ nhạy cảm đối với mỗi công ty. Các chuyên gia trong bất kỳ lĩnh vực nghề nghiệp nào thường không dám công khai hoàn toàn các phương pháp và những tinh tế của công việc cho các đồng nghiệp tối ưu hóa, nhận thức được rủi ro mất việc trong môi trường công nghệ phát triển nhanh chóng.

> *Nếu bạn muốn tạo ra kẻ thù cho mình, hãy thử thay đổi điều gì đó [137].*
>
> *— Woodrow Wilson, phát biểu tại hội nghị bán hàng, Detroit, năm 1916.*

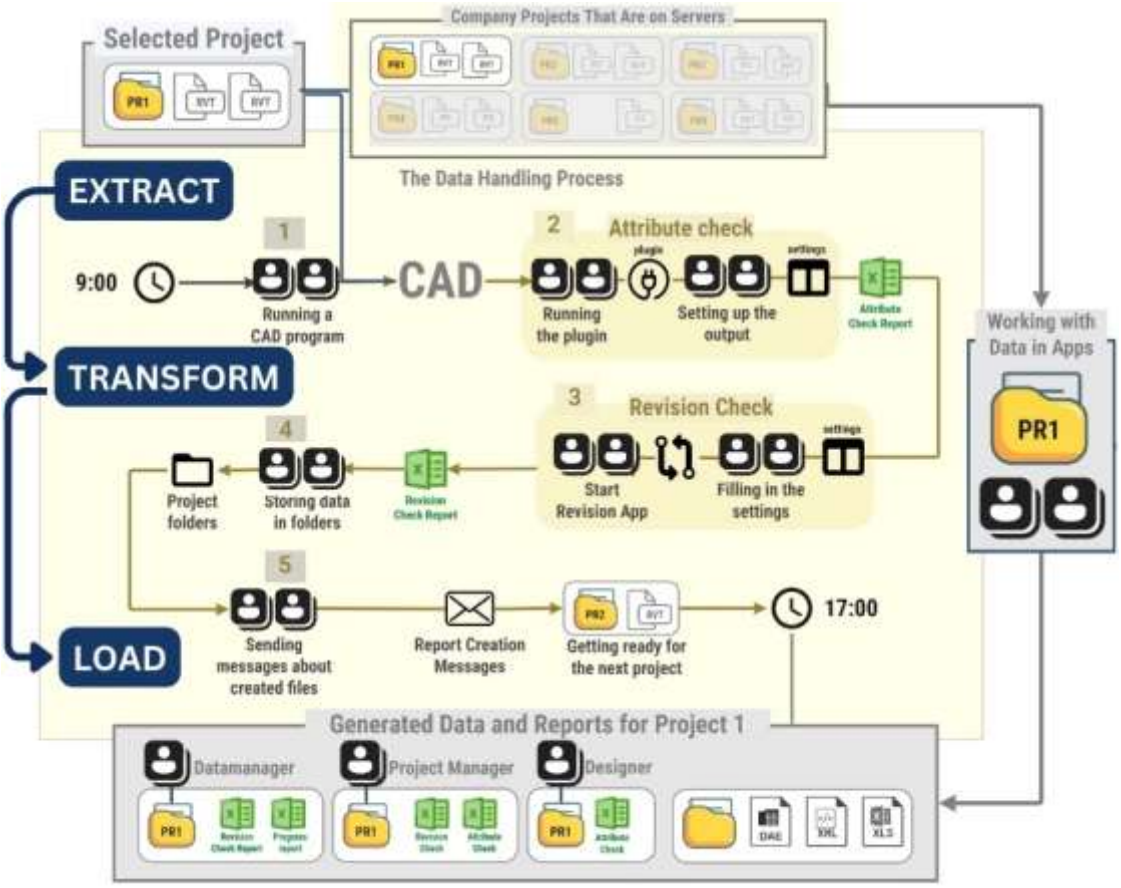
Mặc dù có những lợi ích rõ ràng của việc tự động hóa, trong thực tế hàng ngày của nhiều công ty vẫn tồn tại một lượng công việc thủ công cao, đặc biệt là trong lĩnh vực xử lý dữ liệu kỹ thuật. Để minh họa một cách rõ ràng tình hình hiện tại, hãy xem xét một ví dụ điển hình về quy trình xử lý dữ liệu tuần tự trong các quy trình tương tự.

Quy trình làm việc thủ công với dữ liệu có thể được minh họa qua việc tương tác với thông tin thu được từ các cơ sở dữ liệu CAD. Việc xử lý dữ liệu truyền thống ("quy trình ETL thủ công") trong các bộ phận CAD (BIM) để tạo các bảng thuộc tính hoặc tài liệu dựa trên dữ liệu thiết kế diễn ra theo thứ tự sau (Hình 7.21):-

1. Trích xuất thủ công (Extract): người dùng thủ công mở dự án - bằng cách khởi động ứng dụng CAD (BIM) (Hình 7.21 bước 1). -
2. Xác minh: ở giai đoạn tiếp theo, thường có một vài plugin hoặc ứng dụng hỗ trợ được khởi động một cách thủ công nhằm chuẩn bị dữ liệu và đánh giá chất lượng dữ liệu (Hình 7.21 bước 2-3).
3. Chuyển đổi thủ công (Transform): sau khi chuẩn bị, quá trình xử lý dữ liệu bắt đầu, yêu cầu quản lý

thủ công các công cụ phần mềm khác nhau trong đó dữ liệu được chuẩn bị để xuất (Hình 7.21 bước 4). -

4. Xuất thủ công (Load): việc xuất dữ liệu đã chuyển đổi vào các hệ thống bên ngoài, định dạng dữ liệu và tài liệu (Hình 7.21 bước 5). -



Hình 7.21 Quy trình ETL thủ công truyền thống bị giới hạn bởi mong muốn và khả năng vật lý của từng chuyên gia kỹ thuật.

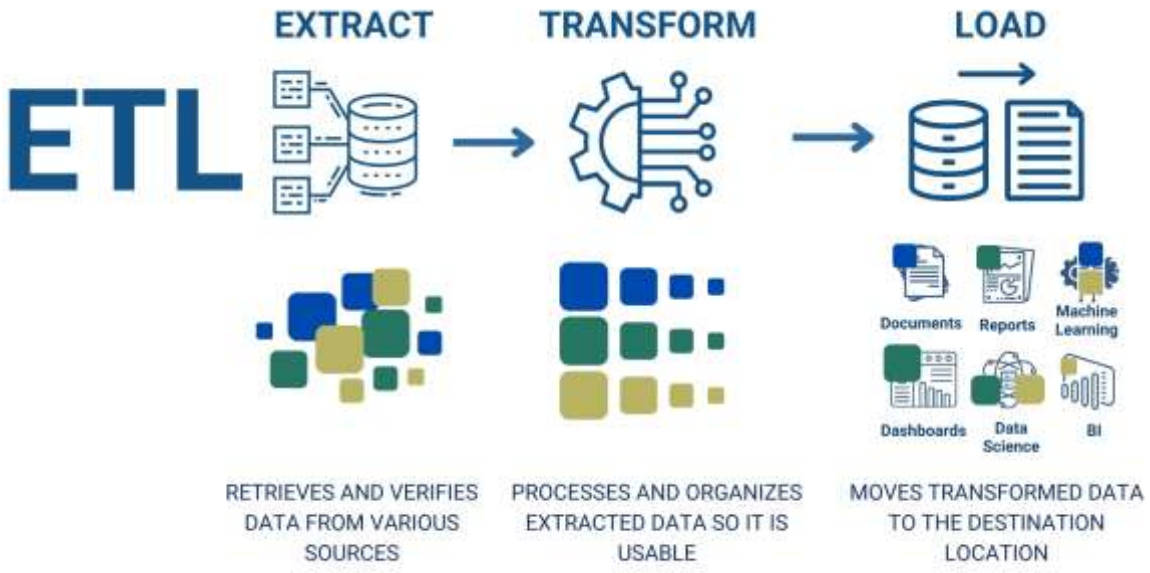
Quy trình làm việc như vậy là ví dụ về quy trình ETL cổ điển - trích xuất, chuyển đổi và tải (ETL). Khác với các ngành khác, nơi mà các pipeline ETL tự động đã trở thành tiêu chuẩn từ lâu, trong ngành xây dựng vẫn chiếm ưu thế là lao động thủ công, làm chậm các quy trình và tăng chi phí.

> ETL (Extract, Transform, Load) là quá trình trích xuất dữ liệu từ nhiều nguồn khác nhau, chuyển đổi chúng sang định dạng cần thiết và tải vào hệ thống đích để phân tích và sử dụng tiếp theo.

ETL là quy trình chỉ ba thành phần chính trong xử lý dữ liệu: Trích xuất, Chuyển đổi và Tải (Hình 7.22):-

- Trích xuất (Extract) - trích xuất dữ liệu từ các nguồn khác nhau (tệp, cơ sở dữ liệu, API).
- Chuyển đổi (Transform) - làm sạch, tổng hợp, chuẩn hóa và xử lý logic dữ liệu.
- Tải (Load) - tải thông tin đã cấu trúc vào kho dữ liệu, báo cáo hoặc hệ thống BI.

Trước đây trong cuốn sách, khái niệm ETL chỉ được đề cập một cách thoáng qua: khi chuyển đổi tài liệu không cấu trúc đã quét thành định dạng bảng có cấu trúc (Hình 4.11), trong bối cảnh chính thức hóa các yêu cầu, cho phép hệ thống hóa nhận thức về các quá trình sống và công việc (Hình 4.420), cũng như trong khuôn khổ tự động hóa kiểm tra dữ liệu và xử lý dữ liệu từ các giải pháp CAD. Giờ đây, chúng ta sẽ xem xét ETL chi tiết hơn trong bối cảnh các quy trình làm việc điển hình.-



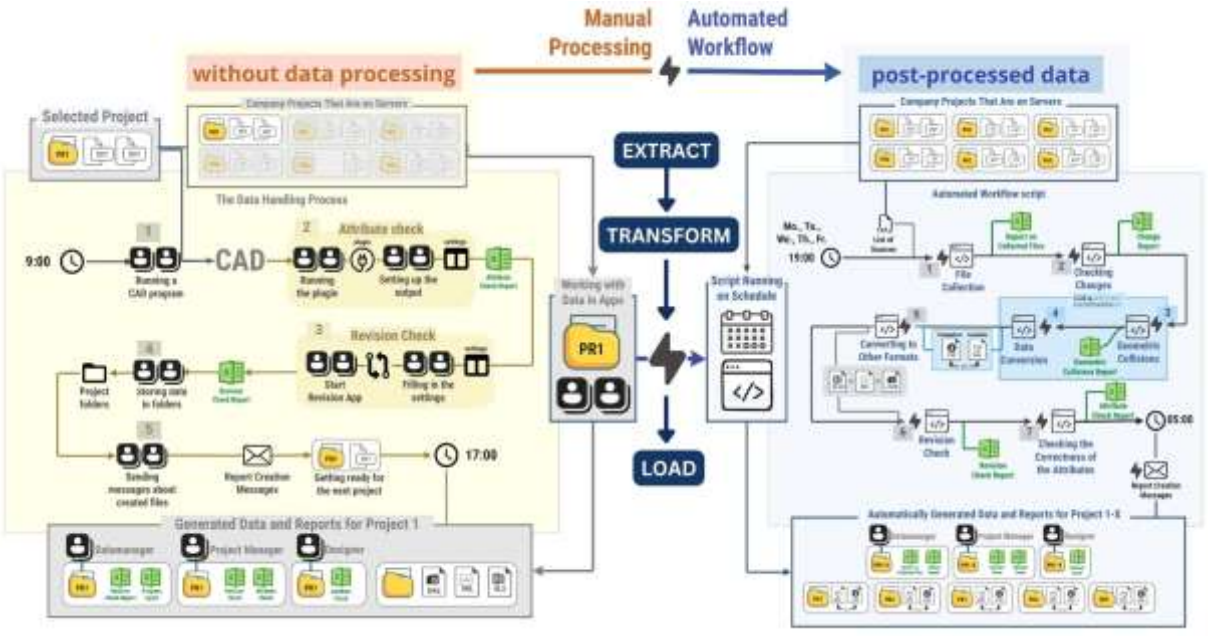
Hình 7.22 ETL tự động hóa các nhiệm vụ xử lý dữ liệu lặp đi lặp lại.

Quy trình ETL thủ công hoặc bán tự động yêu cầu có một người quản lý hoặc chuyên gia kỹ thuật, người điều hành tất cả các bước bằng tay - từ thu thập dữ liệu đến tạo báo cáo. Quy trình này mất một khoảng thời gian đáng kể, đặc biệt trong điều kiện một ngày làm việc hạn chế (ví dụ, từ 9:00 đến 17:00).

Thường thì các công ty cố gắng giải quyết vấn đề hiệu suất thấp và tốc độ chậm bằng cách mua các giải pháp tích hợp mô-đun (ERP, PMIS, CPM, CAFM, v.v.), sau đó được tinh chỉnh bởi các nhà cung cấp và tư vấn bên ngoài. Nhưng các nhà cung cấp và nhà phát triển bên ngoài thường trở thành điểm phụ thuộc quan trọng: những hạn chế kỹ thuật của họ ảnh hưởng trực tiếp đến hiệu suất của toàn bộ hệ thống và của doanh nghiệp nói chung, điều này đã được mô tả chi tiết trong các chương trước về hệ thống và định dạng sở hữu. Về các vấn đề do sự phân mảnh và phụ thuộc gây ra, chúng ta đã nói rõ trong chương "Cách ngành xây dựng chìm trong hỗn loạn dữ liệu".

> Nếu công ty không sẵn sàng triển khai một nền tảng mô-đun lớn từ một nhà cung cấp nào đó, họ bắt đầu tìm kiếm các phương án thay thế cho tự động hóa. Một trong số đó là phát triển các quy trình ETL mô-đun mở, trong đó mỗi bước (trích xuất, chuyển đổi, xác thực, tải lên) được triển khai dưới dạng các đoạn mã được thực thi theo lịch trình.

Trong phiên bản tự động hóa của cùng một quy trình ETL (Hình 7.21), quy trình làm việc trông giống như mã mô-đun, bắt đầu từ việc xử lý dữ liệu và chuyển đổi chúng thành định dạng có cấu trúc mở. Sau khi nhận được dữ liệu có cấu trúc, các kịch bản hoặc mô-đun khác nhau được tự động khởi động theo lịch trình để kiểm tra các thay đổi, chuyển đổi và gửi thông báo (Hình 7.23).--



Hình 7.23 Bên trái là xử lý thủ công, bên phải là quy trình tự động, mà khác với xử lý thủ công truyền thống, không bị hạn chế bởi khả năng của người dùng.

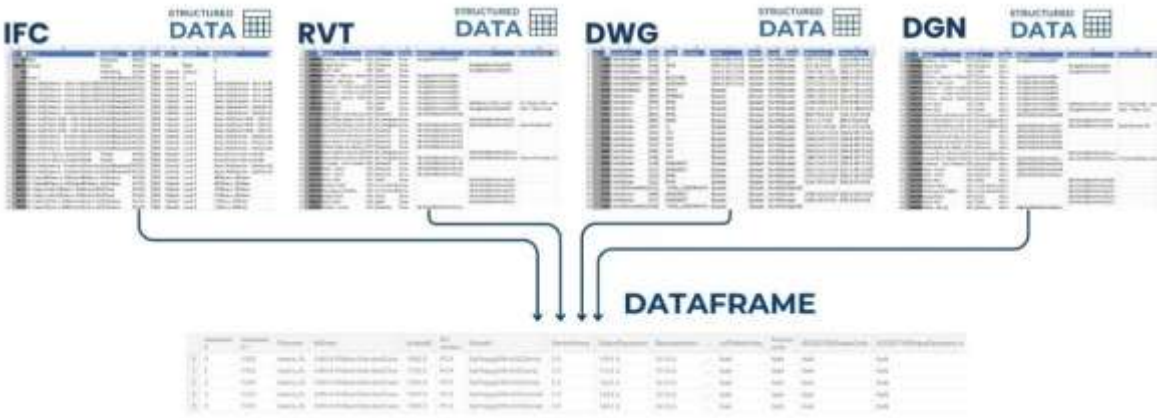
Trong quy trình làm việc tự động hóa, việc xử lý dữ liệu được đơn giản hóa nhờ vào việc xử lý dữ liệu ET(L) trước: cấu trúc hóa và đồng nhất hóa.

> Với các phương pháp xử lý truyền thống, các chuyên gia làm việc với dữ liệu "như nó có" — ở hình thức mà chúng được trích xuất từ các hệ thống hoặc phần mềm. Ngược lại, trong các quy trình tự động hóa, dữ liệu thường trải qua quy trình ETL, nơi chúng được chuẩn hóa thành cấu trúc và định dạng nhất quán phù hợp cho việc sử dụng và phân tích tiếp theo.

Xem xét một ví dụ thực tiễn về ETL, minh họa quy trình kiểm tra bảng dữ liệu như đã mô tả trong chương "Kiểm tra dữ liệu và kết quả kiểm tra" (Hình 4.413). Để làm điều này, chúng tôi sử dụng thư viện Pandas kết hợp với LLM cho các quy trình phân tích và xử lý dữ liệu tự động hóa.-

ETL Extract: thu thập dữ liệu

Giai đoạn đầu tiên của quy trình ETL — trích xuất (Extract) — bắt đầu bằng việc viết mã để thu thập tập hợp dữ liệu cần được kiểm tra và xử lý sau này. Để thực hiện điều này, chúng tôi quét tất cả các thư mục trên máy chủ làm việc, thu thập tài liệu ở định dạng và nội dung cụ thể, sau đó chuyển đổi chúng thành dạng có cấu trúc. Quá trình này đã được trình bày chi tiết trong các chương "Chuyển đổi dữ liệu không cấu trúc và văn bản thành dạng có cấu trúc" và "Chuyển đổi dữ liệu CAD (BIM) thành dạng có cấu trúc" (Hình 4.11 - Hình 4.112).--



Hình 7.24 Chuyển đổi dữ liệu CAD (BIM) thành một khung dữ liệu lớn, sẽ chứa tất cả các phần của dự án.

Là một ví dụ minh họa, ở bước tải dữ liệu Extract và thu được bảng tất cả các dự án CAD (BIM) (Hình 7.24) sử dụng các bộ chuyển đổi hỗ trợ kỹ thuật đảo ngược [138] cho các định dạng RVT và IFC, để thu thập bảng có cấu trúc từ tất cả các dự án và hợp nhất chúng thành một bảng DataFrame lớn.-



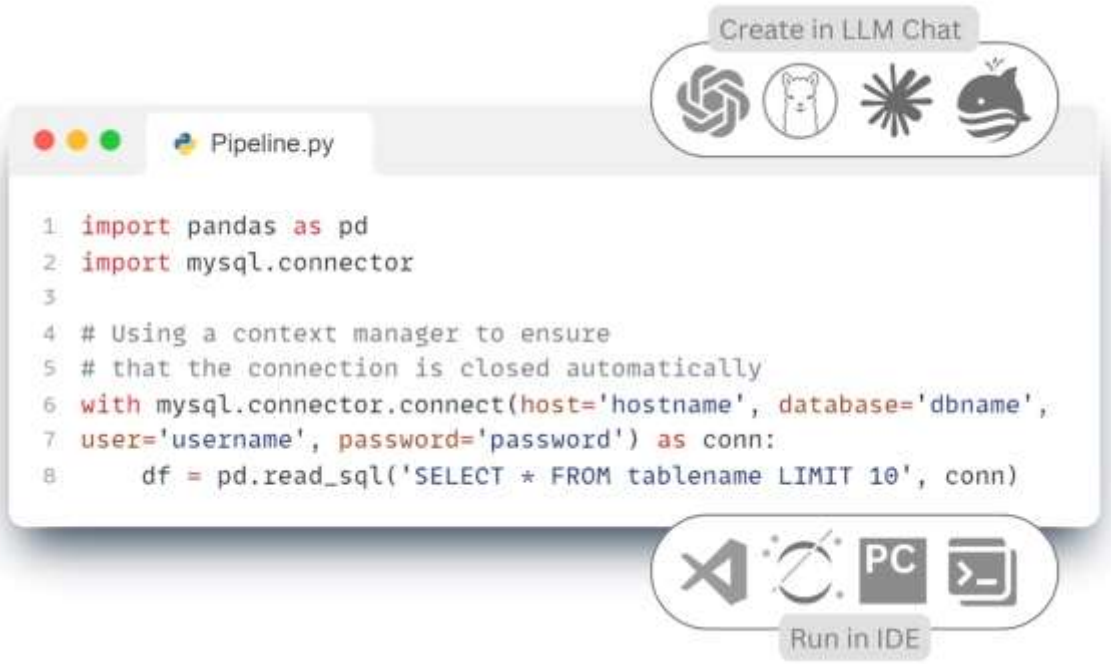
Hình 7.25 Chuyển đổi thông qua mã Python và SDK công cụ đảo ngược kỹ thuật của tập tin RVT và IFC thành một DataFrame lớn có cấu trúc (df).

Trong DataFrame Pandas, bạn có thể tải dữ liệu từ các nguồn khác nhau, bao gồm các tệp văn bản CSV, bảng Excel, tệp JSON và XML, các định dạng lưu trữ khối lượng lớn như Parquet và HDF5, cũng như từ các cơ sở dữ liệu như MySQL, PostgreSQL, SQLite, Microsoft SQL Server, Oracle và những cơ sở dữ liệu khác. Hơn nữa, Pandas hỗ trợ tải dữ liệu từ API, trang web, dịch vụ đám mây và hệ thống lưu trữ như Google BigQuery, Amazon Redshift và Snowflake.

- Để viết mã kết nối và thu thập thông tin từ các cơ sở dữ liệu, bạn có thể gửi yêu cầu văn bản tương tự trong cuộc trò chuyện với LLM (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ cái nào khác):

> Xin vui lòng viết một ví dụ kết nối tới MySQL và chuyển đổi dữ liệu thành DataFrame ⏎

Phản hồi từ LLM:



Hình 7.26 Ví dụ kết nối qua Python tới cơ sở dữ liệu MySQL và nhập dữ liệu từ cơ sở dữ liệu MySQL vào DataFrame.

Mã thu được (Hình 7.25, Hình 7.26) có thể được chạy trong một trong những IDE phổ biến (môi trường phát triển tích hợp), mà chúng ta đã đề cập ở trên, ở chế độ ngoại tuyến: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse với plugin PyDev, Thonny, Wing IDE, IntelliJ IDEA với plugin Python, JupyterLab hoặc các công cụ trực tuyến phổ biến: Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.-

Sau khi tải dữ liệu đa định dạng vào biến "df" (Hình 7.25 - dòng thứ 25; Hình 7.26 - dòng thứ 8), chúng tôi đã chuyển đổi dữ liệu vào định dạng DataFrame của Pandas - một trong những cấu trúc phổ biến nhất để xử lý dữ liệu, đại diện cho một bảng hai chiều với các hàng và cột. Về các định dạng lưu trữ khác được sử dụng trong ETL-Pipelines, chẳng hạn như Parquet, Apache ORC, JSON, Feather, HDF5, cũng như các kho dữ liệu hiện đại, chúng tôi sẽ thảo luận chi tiết hơn trong chương "Lưu trữ và quản lý dữ liệu trong ngành xây dựng" (Hình 8.12).--

Sau giai đoạn trích xuất và cấu trúc dữ liệu (Extract), một mảng thông tin duy nhất được hình thành (Hình 7.25, Hình 7.26), sẵn sàng cho việc xử lý tiếp theo. Tuy nhiên, trước khi tải những dữ liệu này vào các hệ thống mục tiêu hoặc sử dụng cho phân tích, cần phải đảm bảo về chất lượng, toàn bộ và đáp ứng các yêu cầu đã đặt ra. Chính tại giai đoạn này, quá trình chuyển đổi dữ liệu (Transform) diễn ra - một bước quan trọng đảm bảo độ tin cậy của các kết luận và quyết định sau này.--

ETL Transform: áp dụng quy tắc kiểm tra và chuyển đổi

Tại giai đoạn Transform, dữ liệu được xử lý và chuyển đổi. Quá trình này có thể bao gồm việc kiểm tra tính chính xác, chuẩn hóa, điền các giá trị thiếu và xác thực bằng các công cụ tự động hóa.

> Theo Nghiên cứu của PwC "Dữ liệu định hướng. Những gì sinh viên cần để thành công trong thế giới kinh doanh đang thay đổi nhanh chóng" (2015) [9], các công ty kiểm toán hiện đại đang từ bỏ việc kiểm tra mẫu dữ liệu và chuyển sang phân tích khối lượng thông tin bằng cách sử dụng các công cụ tự động hóa. Cách tiếp cận này không chỉ cho phép phát hiện các sự không thống nhất trong báo cáo mà còn đưa ra các khuyến nghị tối ưu hóa quy trình kinh doanh.

Trong xây dựng, các phương pháp tương tự có thể được áp dụng, chẳng hạn như để tự động xác thực dữ liệu thiết kế, kiểm soát chất lượng xây dựng và đánh giá hiệu quả công việc của nhà thầu. Một trong những công cụ cho phép tự động hóa và tăng tốc độ xử lý dữ liệu là việc sử dụng biểu thức chính quy (RegEx) trong giai đoạn chuyển đổi dữ liệu (Transform) trong quy trình ETL. RegEx cho phép kiểm tra hiệu quả các chuỗi dữ liệu, phát hiện sự không phù hợp và đảm bảo tính toàn vẹn của thông tin với chi phí tài nguyên tối thiểu. Chi tiết về RegEx (Hình 4.47) chúng tôi đã đề cập trong chương "Chuyển đổi yêu cầu thành dạng có cấu trúc".-

Hãy xem xét một ví dụ thực tế: trong hệ thống quản lý tài sản bất động sản (RPM), người quản lý thiết lập yêu cầu cho các thuộc tính chính của tài sản (Hình 7.27). Trong giai đoạn chuyển đổi, cần thực hiện xác thực các tham số sau:-

- kiểm tra định dạng của các định danh tài sản (thuộc tính "ID")
- kiểm soát giá trị của thời gian bảo hành thay thế (thuộc tính "Thời gian bảo hành")
- kiểm tra chu kỳ thay thế các phần tử (thuộc tính "Yêu cầu bảo trì")



Hình 7.27 Kiểm tra chất lượng bắt đầu từ việc thiết lập yêu cầu cho các thuộc tính và giá trị biên của chúng.

Để thiết lập giá trị biên cho việc kiểm tra các tham số, ví dụ, giả sử từ kinh nghiệm của chúng tôi, chúng tôi biết rằng các giá trị cho thuộc tính "ID" có thể chỉ bao gồm các giá trị chuỗi "W-NEW", "W-OLD1" hoặc "D-122" hoặc các giá trị tương tự, trong đó ký tự đầu tiên là chữ cái, theo sau là dấu gạch nối, và sau đó là ba ký tự chữ 'NEW', 'OLD' hoặc bất kỳ số ba chữ số nào (Hình 7.27). Để xác thực các định danh này, có thể sử dụng biểu thức chính quy (RegEx) sau:-

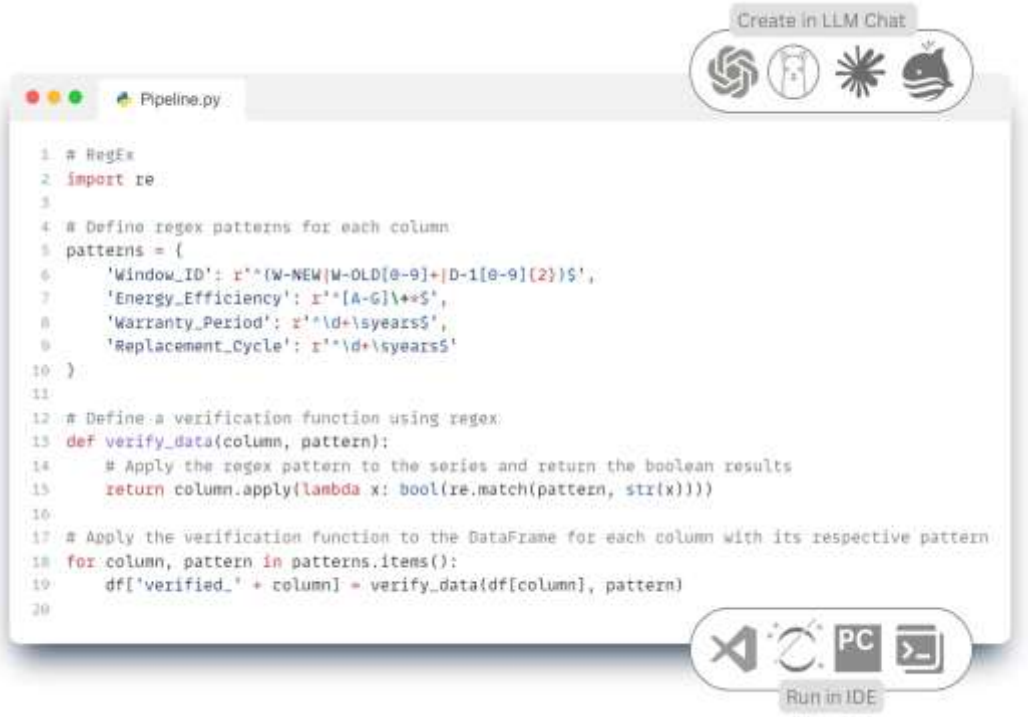
```
^W-NEW$|^W-OLD[0-9]+$|^D-1[0-9]{2}$
```

Mẫu này cho phép đảm bảo rằng tất cả các định danh trong dữ liệu đều phù hợp với các tiêu chí đã định. Nếu bất kỳ giá trị nào không vượt qua kiểm tra, hệ thống sẽ ghi lại lỗi. Để tạo mã Python cho việc chuyển đổi dữ liệu và sử dụng dữ liệu thu được để tạo bảng kết quả, chỉ cần xây dựng truy vấn trong LLM chat.

- Truy vấn văn bản trong LLM:

> Viết mã để kiểm tra các cột DataFrame bằng cách sử dụng biểu thức chính quy, kiểm tra các định danh ở định dạng 'W-NEW' hoặc 'W-OLD' qua RegEx, hiệu suất năng lượng với các chữ cái từ 'A' đến 'G', thời gian bảo hành và chu kỳ thay thế với các giá trị số trong năm ⏎

- Phản hồi từ LLM:



Hình 7.28 Mã tự động hóa quy trình kiểm tra bằng việc áp dụng các mẫu RegEx cho các cột tham số của DataFrame.

Mã Python được tạo tự động (Hình 7.28) sử dụng thư viện “re“ (biểu thức chính quy RegEx) để xác định một hàm, hàm này kiểm tra mỗi thuộc tính của phần tử dữ liệu trong DataFrame. Đối với mỗi cột (thuộc tính) được chỉ định, hàm áp dụng mẫu RegEx để kiểm tra sự phù hợp của từng bản ghi với định dạng mong đợi và thêm kết quả dưới dạng các giá trị mới (False/True) vào một cột thuộc tính mới trong DataFrame.-

> Kiểm tra tự động hóa tương tự đảm bảo dữ liệu tuân thủ các yêu cầu được đặt ra và có thể được sử dụng như một phần của hệ thống kiểm soát chất lượng trong giai đoạn chuyển đổi.

Sau khi hoàn thành thành công giai đoạn Transform và kiểm tra chất lượng, dữ liệu đã sẵn sàng để tải lên

các hệ thống mục tiêu. Dữ liệu đã được chuyển đổi và xác minh có thể được xuất ra các định dạng như CSV, JSON, Excel, cơ sở dữ liệu và nhiều định dạng khác để sử dụng sau này. Ngoài ra, tùy thuộc vào yêu cầu, kết quả có thể được trình bày dưới dạng báo cáo, biểu đồ hoặc bảng điều khiển phân tích.

ETL Load: Hình ảnh hóa kết quả dưới dạng biểu đồ và đồ thị

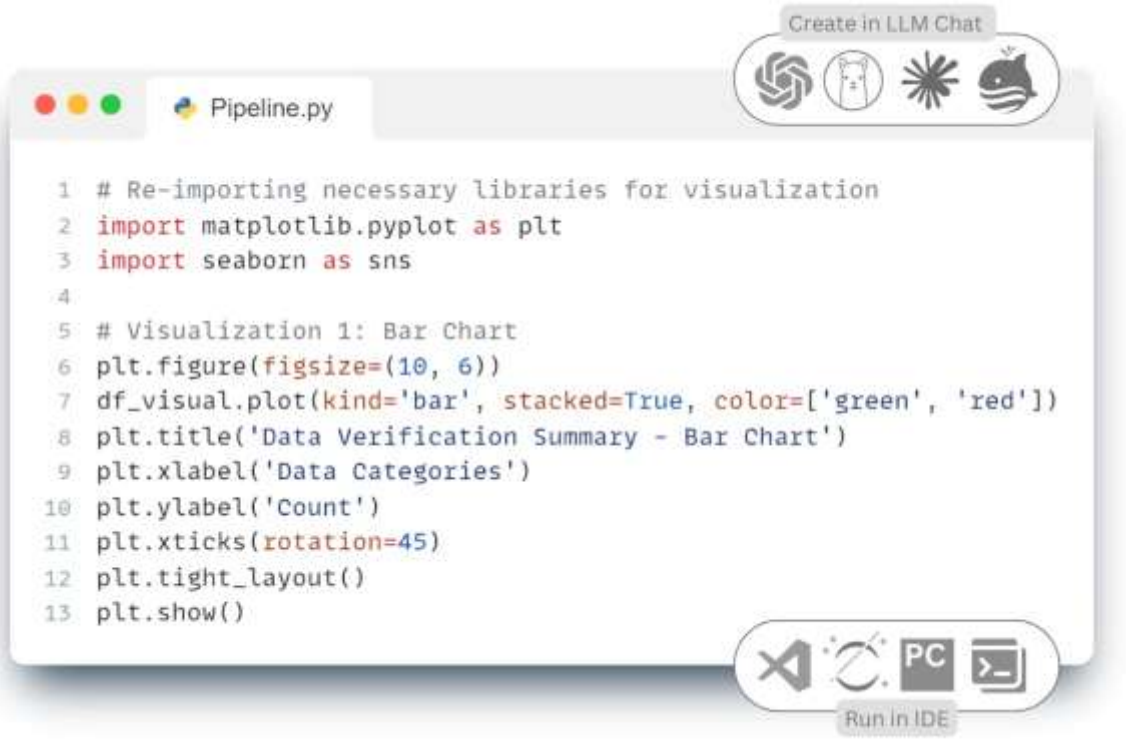
Sau khi hoàn thành giai đoạn Transform, khi dữ liệu đã được cấu trúc và kiểm tra, giai đoạn cuối cùng — Load — diễn ra, tại đây dữ liệu có thể được tải lên hệ thống mục tiêu hoặc được trực quan hóa để phân tích. Việc thể hiện dữ liệu một cách trực quan giúp phát hiện kịp thời các sai lệch, phân tích sự phân bố và truyền đạt các kết luận chính đến tất cả các bên liên quan trong dự án, bao gồm cả những người không có nền tảng kỹ thuật.

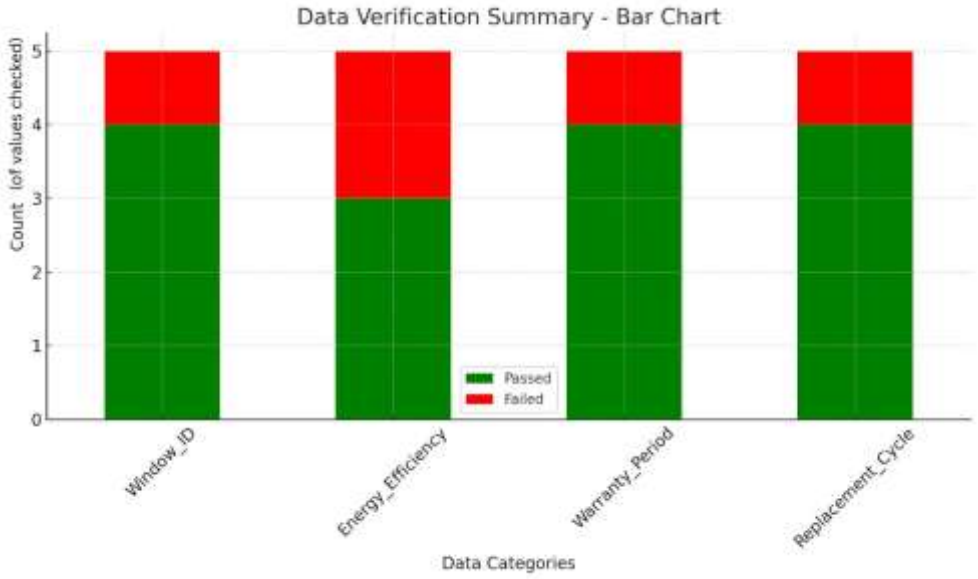
Thay vì trình bày thông tin dưới dạng bảng và số, chúng ta có thể sử dụng đồ họa thông tin, biểu đồ và bảng điều khiển (dashboards). Một trong những công cụ trực quan hóa dữ liệu cấu trúc phổ biến và linh hoạt nhất trong ngôn ngữ Python là thư viện Matplotlib (Hình 7.29, Hình 7.210). Nó cho phép tạo ra các biểu đồ tĩnh, động và tương tác, đồng thời hỗ trợ nhiều loại biểu đồ khác nhau. --

- Để trực quan hóa kết quả kiểm tra thuộc tính từ hệ thống RPM (Hình 7.27), có thể sử dụng truy vấn sau đối với mô hình ngôn ngữ:-

> Viết mã để trực quan hóa dữ liệu DataFrame, như đã được đề cập ở trên (Hình 7.27), với histogram để hiển thị tần suất lỗi trong thuộc tính ⏎-

Phản hồi của LLM dưới dạng mã và trực quan hóa sẵn sàng trong cuộc trò chuyện của LLM về kết quả thực thi mã:





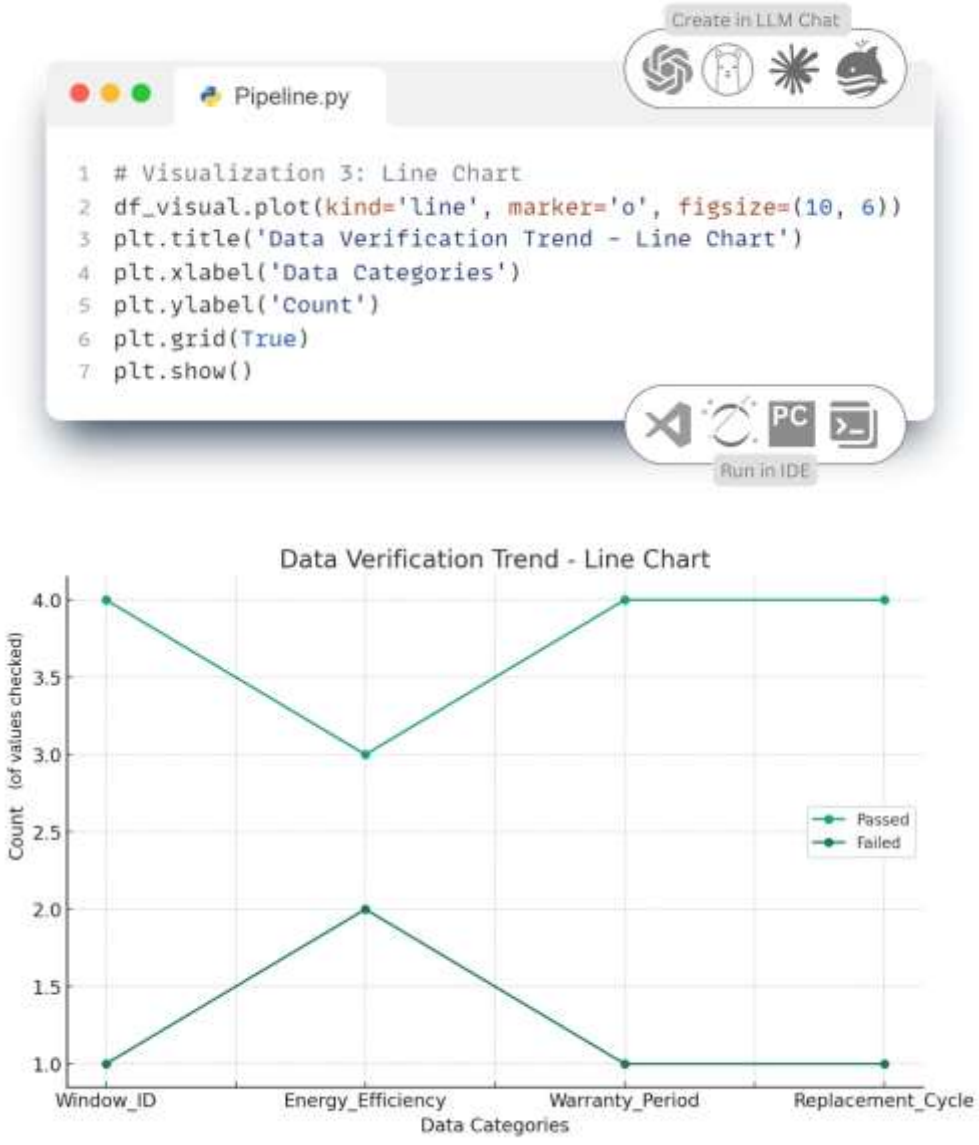
Hình 7.29 Trực quan hóa kết quả giai đoạn Transform liên quan đến việc kiểm tra giá trị thuộc tính từ hệ thống RPM (Hình 7.27) dưới dạng histogram ở giai đoạn Load.-

Có rất nhiều thư viện trực quan hóa miễn phí và mã nguồn mở cho phép trình bày dữ liệu cấu

trúc ở nhiều định dạng khác nhau. Chúng ta sẽ tiếp tục trực quan hóa kết quả bằng một dạng biểu đồ khác trong một truy vấn tiếp theo trong cuộc trò chuyện:

> Hãy thể hiện các dữ liệu tương tự dưới dạng biểu đồ đường ⏎

Phản hồi từ LLM:



Hình 7.210 Trực quan hóa dữ liệu kiểm tra (Hình 7.28) dưới dạng biểu đồ đường được tạo ra bằng thư viện Matplotlib.-

Có rất nhiều thư viện trực quan hóa miễn phí và mã nguồn mở, chẳng hạn như:

- Seaborn — cho các biểu đồ thống kê (Hình 7.211)-
- Plotly — cho các trực quan hóa web tương tác (Hình 7.212, Hình 7.16)--
- Altair — cho trực quan hóa khai báo
- Dash hoặc Streamlit — để tạo ra các bảng điều khiển hoàn chỉnh

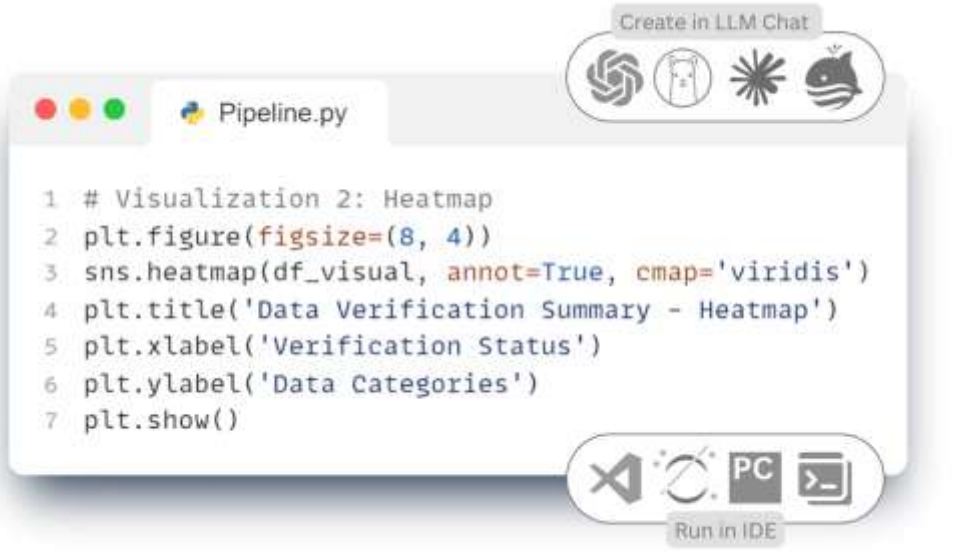
Tuy nhiên, việc biết các thư viện cụ thể cho việc trực quan hóa không phải là điều bắt buộc – các công cụ hiện đại, bao gồm cả LLM, cho phép tự động tạo mã để xây dựng biểu đồ và các ứng dụng hoàn chỉnh dựa trên mô tả nhiệm vụ.

Việc lựa chọn công cụ phụ thuộc vào yêu cầu của dự án: cho dù đó là báo cáo, thuyết trình hay bảng theo dõi trực tuyến. Ví dụ, thư viện mã nguồn mở Seaborn đặc biệt tốt cho việc làm việc với dữ liệu phân loại, giúp phát hiện các quy luật và xu hướng.

- Để xem thư viện Seaborn hoạt động, bạn có thể yêu cầu LLM sử dụng trực tiếp thư viện cần thiết hoặc gửi một yêu cầu văn bản tương tự trong quá trình làm việc với LLM:

> Hãy hiển thị bản đồ nhiệt cho các kết quả ⏎

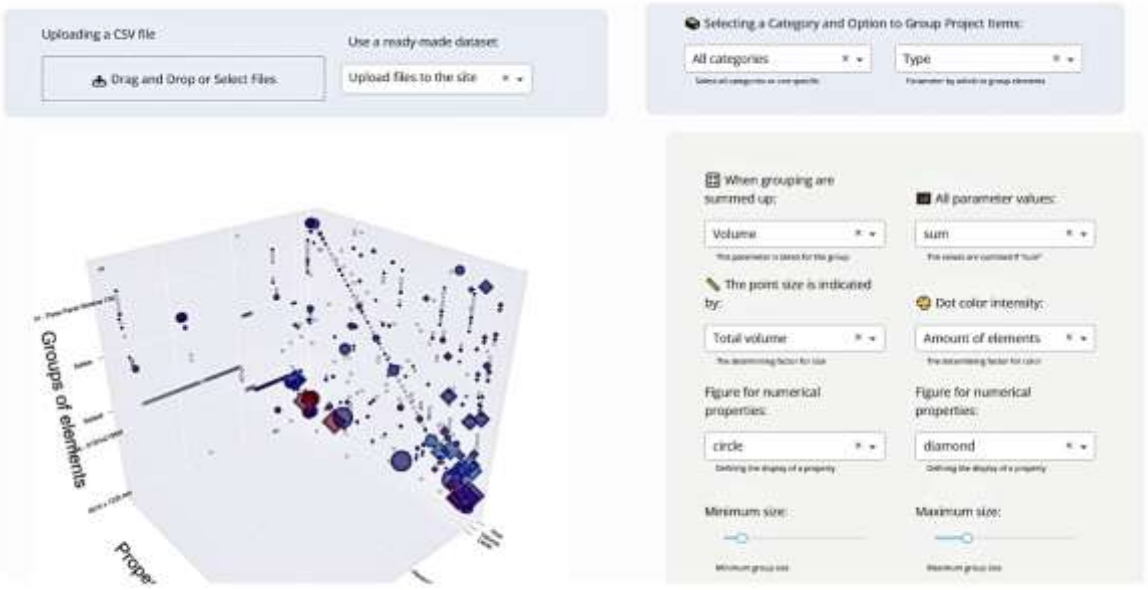
- Phản hồi LLM dưới dạng mã và biểu đồ đã sẵn sàng, mã để dựng biểu đồ giờ đây có thể được sao chép vào IDE, và biểu đồ có thể được sao chép hoặc lưu để chèn vào tài liệu:





Hình 7.211 Hiển thị kết quả kiểm tra (Hình 7.28) dữ liệu bằng thư viện Seaborn.-

Đối với những ai thích phương pháp tương tác, có các công cụ cho phép tạo biểu đồ và bảng điều khiển động với khả năng tương tác. Thư viện Plotly (Hình 7.16, Hình 7.212) cung cấp khả năng tạo các biểu đồ và bảng điều khiển có tính tương tác cao, có thể nhúng vào các trang web và cung cấp cho người dùng khả năng tương tác với dữ liệu trong thời gian thực.--



Hình 7.212 Hiển thị 3D tương tác của các thuộc tính thành phần từ dự án CAD- (BIM-) bằng thư viện Plotly.

Các thư viện mã nguồn mở chuyên biệt như Bokeh, Dash và Streamlit cung cấp cách thuận tiện để trình bày dữ liệu mà không cần kiến thức sâu về phát triển web. Bokeh phù hợp cho các biểu đồ tương tác phức tạp, Dash được sử dụng để xây dựng các bảng điều khiển phân tích đầy đủ, trong khi Streamlit cho phép tạo nhanh các ứng dụng web cho phân tích dữ liệu.

Nhờ vào những công cụ hình ảnh hóa này, các nhà phát triển và nhà phân tích có thể hiệu quả trong việc phát tán kết quả đến đồng nghiệp và các bên liên quan, đảm bảo tương tác trực quan với dữ liệu và đơn

giản hóa quy trình ra quyết định.

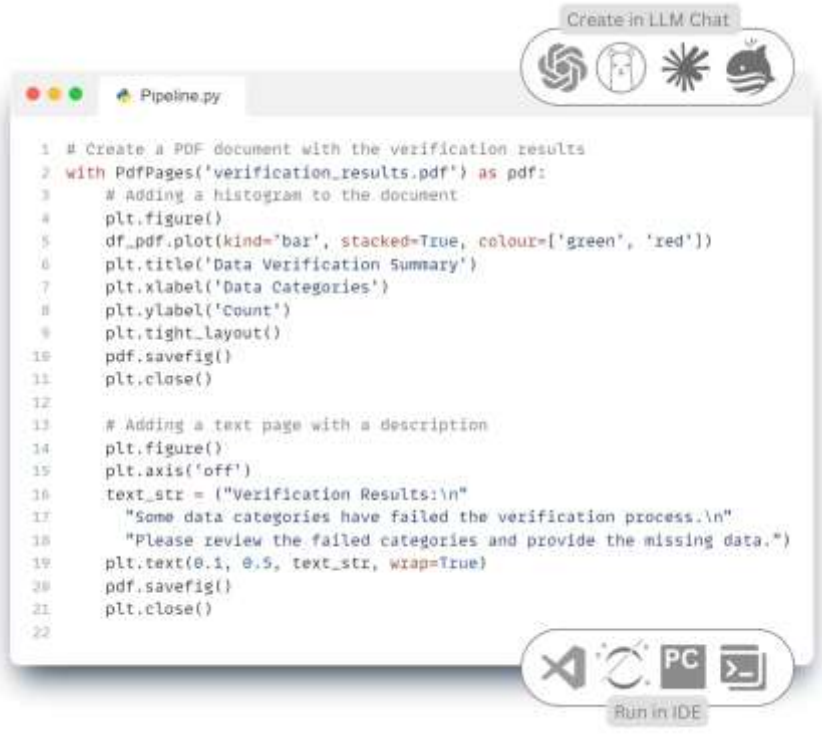
ETL Load: Tạo tài liệu PDF tự động

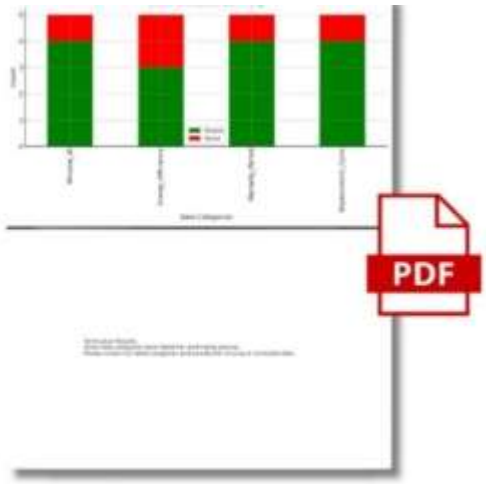
Trong giai đoạn tải lên, không chỉ có thể hình ảnh hóa dữ liệu, xuất chúng vào bảng hoặc cơ sở dữ liệu, mà còn tự động tạo báo cáo bao gồm các biểu đồ, đồ thị và các chỉ số phân tích chính mà nhà quản lý hoặc chuyên gia cần nhận được kết quả kiểm tra. Các báo cáo tự động có thể chứa cả bình luận và diễn giải văn bản về dữ liệu lẫn các yếu tố hình ảnh – bảng biểu, đồ thị.

- Để tạo báo cáo PDF với biểu đồ (Hình 7.29) và mô tả phân tích dựa trên kết quả kiểm tra mà chúng tôi thực hiện trong các chương trước, chỉ cần soạn thảo một yêu cầu trong cuộc hội thoại tiếp theo với LLM, chẳng hạn:-

> Viết mã để tạo tệp PDF với biểu đồ và mô tả kết quả kiểm tra dữ liệu đã được nhận trước (trong chat), và viết cảnh báo dưới dạng văn bản rằng một số danh mục không qua kiểm tra và cần điền dữ liệu còn thiếu ⏎

- Phản hồi LLM dưới dạng mã và tệp PDF đã tạo có kết quả:





Hình 7.213 Mã tự động tạo ra tài liệu PDF, chứa biểu đồ với dữ liệu thử nghiệm và văn bản với kết quả kiểm tra.

Giải pháp tự động được viết chỉ với 20 dòng mã bằng LLM ngay lập tức tạo ra tài liệu PDF (hoặc DOC) cần thiết với hình ảnh hóa dưới dạng biểu đồ thể hiện các thuộc tính (Hình 7.213), cho thấy số lượng dữ liệu đã qua và không qua kiểm tra, cũng như thêm một khối văn bản mô tả ngắn về kết quả và các khuyến nghị về các hành động tiếp theo.-

Việc tạo tài liệu tự động hóa là yếu tố chính của giai đoạn Tải lên, đặc biệt trong điều kiện hoạt động dự án, nơi tốc độ chuẩn bị báo cáo và độ chính xác của nó là điều cực kỳ quan trọng.

ETL Load: Tự động tạo tài liệu bằng FPDF

Tự động hóa báo cáo trong giai đoạn ETL Load là một bước quan trọng trong xử lý dữ liệu, đặc biệt khi kết quả phân tích cần được trình bày ở định dạng dễ truyền tải và tiếp nhận. Trong ngành xây dựng, điều này thường liên quan đến các báo cáo về tiến độ công trình, thống kê dữ liệu dự án, báo cáo kiểm tra chất lượng hoặc tài liệu tài chính.

Một trong những công cụ tiện lợi nhất cho các tác vụ như vậy là thư viện mã nguồn mở FPDF, có sẵn cho cả Python và PHP.

> Thư viện mã nguồn mở FPDF cung cấp một cách linh hoạt để tạo tài liệu thông qua mã, cho phép thêm tiêu đề, văn bản, bảng và hình ảnh. Việc sử dụng mã thay vì chỉnh sửa thủ công giảm thiểu lỗi và tăng tốc độ chuẩn bị các báo cáo ở định dạng PDF.

Một trong những bước quan trọng trong việc tạo tài liệu PDF là thêm tiêu đề và văn bản chính dưới dạng nhận xét hoặc mô tả. Tuy nhiên, khi tạo báo cáo, không chỉ cần thêm văn bản mà còn phải cấu trúc nó một cách hợp lý. Các tiêu đề, khoảng cách, khoảng cách giữa các dòng - tất cả đều ảnh hưởng đến độ dễ đọc của tài liệu. Sử dụng FPDF, người dùng có thể đặt các tham số định dạng, quản lý vị trí của các yếu tố và tùy chỉnh kiểu tài liệu.

FPDF hoạt động rất giống với HTML. Những ai đã quen thuộc với HTML sẽ dễ dàng tạo ra các tài liệu PDF phức tạp bằng cách sử dụng FPDF, vì cấu trúc mã trong FPDF phần nào giống với đánh dấu HTML: việc thêm tiêu đề, văn bản, hình ảnh và bảng diễn ra theo cách tương tự. Những ai không quen thuộc với HTML

cũng đừng lo lắng - có thể sử dụng LLM, công cụ này sẽ ngay lập tức giúp tạo mã cần thiết để định dạng tài liệu.

- Ví dụ tiếp theo minh họa cách tạo một báo cáo với tiêu đề và văn bản chính. Việc chạy mã này trong bất kỳ IDE nào hỗ trợ Python sẽ tạo ra một file PDF chứa tiêu đề và văn bản mong muốn:

```
từ fpdf nhập FPDF  # Nhập thư viện FPDF  pdf = FPDF()  # Tạo tài liệu PDF
pdf.add_page() # Thêm trang

pdf.set_font("Arial", style='B', size=16)  # Đặt font: Arial, đậm, kích thước
16pdf.cell(200, 10, "Báo cáo dự án", ln=True, align='C')  # Tạo tiêu đề và căn
giữa  pdf.set_font("Arial", size=12) # Thay đổi font thành Arial thông thường, kích thước 12
pdf.multi_cell(0, 10, "Tài liệu này chứa dữ liệu về kết quả kiểm tra các tệp dự
án...")  # Thêm văn bản nhiều dòng  pdf.output(r"C:\reports\report.pdf") # Lưu file PDF
```



Hình. 7.214 Bằng cách sử dụng vài dòng mã Python, chúng ta có thể tự động sinh ra tài liệu PDF cần thiết với văn bản.

Khi chuẩn bị báo cáo, điều quan trọng là phải xem xét rằng dữ liệu được sử dụng để tạo ra tài liệu hiếm khi giữ nguyên. Tiêu đề và các khối văn bản (Hình. 7.214) thường được tạo ra một cách động, nhận các giá trị ngay trong giai đoạn Transform trong quá trình ETL.

> Việc sử dụng mã cho phép tạo ra các tài liệu chứa thông tin, như tên dự án, ngày tạo báo cáo, cũng như thông tin về các thành viên hoặc trạng thái hiện tại. Việc sử dụng biến trong mã cho phép tự động chèn những dữ liệu này vào các vị trí cần thiết trong báo cáo, hoàn toàn loại bỏ nhu cầu chỉnh sửa thủ công trước khi gửi.

Ngoài văn bản và tiêu đề đơn giản, bảng trong tài liệu dự án chiếm một vị trí đặc biệt. Hầu như trong mỗi tài liệu đều có dữ liệu có cấu trúc: từ mô tả đối tượng đến kết quả kiểm tra. Việc tự động tạo bảng dựa trên dữ liệu từ giai đoạn Transform không chỉ giúp tăng tốc quá trình chuẩn bị tài liệu mà còn giảm thiểu lỗi khi chuyển thông tin. FPDF cho phép chèn bảng vào các tệp PDF (dưới dạng văn bản hoặc hình ảnh), định nghĩa các ranh giới ô, kích thước cột và phông chữ (Hình 7.215). Điều này đặc biệt tiện lợi khi làm việc với dữ liệu động, khi số lượng hàng và cột có thể thay đổi tùy thuộc vào yêu cầu của tài liệu.-

- Ví dụ tiếp theo cho thấy cách tự động hóa việc tạo bảng, chẳng hạn như danh sách vật liệu, tính toán dự toán hoặc kết quả kiểm tra thông số:

```
dữ liệu = [   ["Phần tử", "Số lượng", "Giá"],  # Tiêu đề cột    ["Bê tông", "10
m³", "$500."],  # Dữ liệu dòng đầu tiên

Armaturo, 2 tấn, $600, Vữa, 5000 chiếc, $750



pdf = FPDF()  # Tạo tài liệu PDFpdf.them_trang()  # Thêm
trangpdf.set_font("Arial", size=12)  # Thiết lập phông chữ

đối với hàng trong dữ liệu:  # Lặp qua các hàng của bảng    đối với mục trong
hàng:  # Lặp qua các ô trong hàng
pdf.cell(60, 10, item, border=1)  # Tạo ô với viền, chiều rộng 60 và chiều cao
10pdf.ln() # Chuyển sang dòng tiếp theopdf.output(r"C:\reports\table.pdf")  # Lưu tệp PDF
```

Item	Quantity	Price
Concrete	10 m³	$500
Rebar	2 t.	$600
Brick	5000 pcs.	$750



table.pdf

Hình 7.215 Có thể tự động tạo ra PDF không chỉ văn bản mà còn bất kỳ thông tin bảng nào được thu thập trong giai đoạn Chuyển đổi.

Trong các kịch bản báo cáo thực tế, bảng thường là thông tin được tạo động, thu được trong giai đoạn chuyển đổi dữ liệu. Trong ví dụ được đưa ra (Hình 7.215), bảng được chèn vào tài liệu PDF dưới dạng tĩnh: trong từ điển data (dòng đầu tiên của mã), dữ liệu cho ví dụ đã được đưa vào, trong điều kiện thực tế, biến data tương tự sẽ được tự động điền sau khi, chẳng hạn, nhóm lại dataframe.

Trong thực tế, các bảng như vậy thường được xây dựng dựa trên dữ liệu có cấu trúc từ các nguồn động khác nhau: cơ sở dữ liệu, tệp Excel, giao diện API hoặc kết quả của các phép tính phân tích. Thường thì, trong giai đoạn Chuyển đổi (ETL), dữ liệu được tổng hợp, nhóm lại hoặc lọc - và chỉ sau đó được chuyển đổi thành các kết quả dưới dạng đồ thị hoặc bảng hai chiều, được hiển thị trong các báo cáo. Điều này có nghĩa là nội dung của bảng có thể thay đổi tùy thuộc vào các tham số đã chọn, khoảng thời gian phân tích, bộ lọc dự án hoặc cài đặt của người dùng.

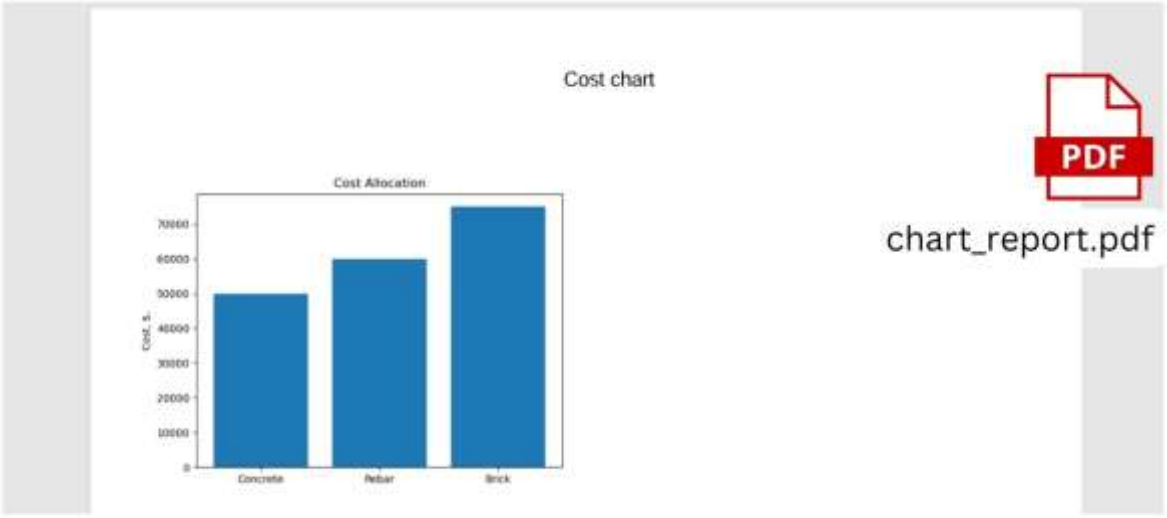
> Việc sử dụng các khung dữ liệu và tập dữ liệu động trong giai đoạn Chuyển đổi giúp quá trình tạo báo cáo ở giai đoạn Tải trở nên linh hoạt, có thể mở rộng và dễ dàng lặp lại mà không cần can thiệp thủ công.

Ngoài bảng và văn bản, FPDF cũng hỗ trợ việc thêm đồ thị từ dữ liệu bảng, cho phép nhúng vào báo cáo các hình ảnh được tạo ra bằng Matplotlib hoặc các thư viện trực quan hóa khác mà chúng ta đã đề cập ở trên. Bạn có thể bổ sung tài liệu bằng mã với bất kỳ đồ thị, biểu đồ và sơ đồ nào.

- Sử dụng thư viện FPDF của Python, chúng ta sẽ thêm vào tài liệu PDF một biểu đồ được tạo ra

trước đó bằng Matplotlib.

```
Nhập khẩu matplotlib.pyplot như plt  # Nhập khẩu Matplotlib để tạo đồ thị.fig, ax = plt.subplots() # Tạo một hình và trục cho đồ thịcategories = ["Bê tông", "Thép", "Gạch"]  # Tên các danh mục values = [50000, 60000, 75000]  # Giá trị theo danh mục ax.bar(categories, values) # Tạo biểu đồ cộtGiá trị, $.  # Ghi chú trục YĐặt tiêu đề "Phân phối chi phí"plt.savefig(r"C:\reports\chart\chart.png")  # Lưu đồ thị dưới dạng hình ảnhpdf = FPDF() # Tạo tài liệu PDFpdf.them_trang()  # Thêm trangpdf.set_font("Arial", size=12)  # Thiết lập phông chữpdf.cell(200, 10, "Biểu đồ chi phí", ln=True, align='C')  # Thêm tiêu đềpdf.image(r"C:\reports\chart\chart.png", x=10, y=30, w=100)  # Chèn hình ảnh vào PDF (x, y - tọa độ, w - chiều rộng)pdf.output(r"C:\reports\chart_report.pdf")  # Lưu tệp PDF
```



Hình 7.216 Chỉ với một chục dòng mã, bạn có thể tạo ra một biểu đồ, lưu nó lại và sau đó chèn vào vị trí cần thiết trong tài liệu PDF.

Nhờ vào FPDF, quy trình chuẩn bị và logic của tài liệu trở nên minh bạch, nhanh chóng và tiện lợi. Các mẫu được tích hợp trong mã cho phép tạo ra các tài liệu với dữ liệu cập nhật, loại bỏ nhu cầu điền tay.

> Sử dụng tự động hóa ETL - thay vì phải tốn công sức để tạo báo cáo thủ công, các chuyên gia có thể tập trung vào việc phân tích dữ liệu và đưa ra quyết định, thay vì phải chọn công cụ phù hợp để làm việc với từng kho dữ liệu với giao diện người dùng dễ hiểu.

Do đó, thư viện FPDF cung cấp một công cụ linh hoạt cho việc tự động tạo ra các tài liệu với mọi mức độ phức tạp - từ các báo cáo kỹ thuật ngắn gọn đến các tóm tắt phân tích phức tạp với bảng biểu và đồ thị, điều này không chỉ giúp tăng tốc độ luân chuyển tài liệu mà còn giảm thiểu đáng kể khả năng xảy ra lỗi liên quan đến việc nhập liệu và định dạng dữ liệu bằng tay.

ETL Load: Lập báo cáo và tải lên các hệ thống khác

Tại giai đoạn Load, các kết quả đã được hình thành dưới dạng bảng, biểu đồ và báo cáo PDF cuối cùng, được chuẩn bị theo các yêu cầu đã thiết lập. Tiếp theo, có thể xuất dữ liệu này sang các định dạng có thể đọc được bởi máy (ví dụ: CSV), điều này là cần thiết cho việc tích hợp với các hệ thống bên ngoài - chẳng hạn như ERP, CAFM, CPM, các nền tảng BI và các giải pháp doanh nghiệp hoặc ngành khác. Ngoài CSV, việc xuất dữ liệu có thể được thực hiện sang các định dạng XLSX, JSON, XML hoặc trực tiếp vào các cơ sở dữ liệu hỗ trợ trao đổi thông tin tự động.

- Để tạo mã tương ứng cho việc tự động hóa giai đoạn Load, chỉ cần gửi yêu cầu đến giao diện LLM, chẳng hạn như: ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude hoặc QWEN:

> Viết mã để tạo báo cáo về kết quả kiểm tra dữ liệu trong DataFrame, trong đó các cột có tiền tố 'verified_' được đếm, đổi tên thành 'Passed' và 'Failed', các giá trị bị thiếu được thay thế bằng 0, và sau đó xuất ra tệp CSV chỉ những dòng đã vượt qua tất cả các kiểm tra.

- Phản hồi từ LLM:



Hình 7.217 Dữ liệu đã được xác minh, thu được từ giai đoạn Transform của dataframe cuối cùng, được xuất ra tệp CSV để tích hợp với các hệ thống khác.

Trong đoạn mã được cung cấp (Hình 7.217), giai đoạn cuối cùng của quy trình ETL – Load được thực hiện, trong đó dữ liệu đã được kiểm tra được lưu dưới định dạng CSV, tương thích với hầu hết các hệ thống và cơ sở dữ liệu bên ngoài. Như vậy, chúng ta đã hoàn thành vòng đời hoàn chỉnh của quy trình ETL, bao gồm việc trích xuất, chuyển đổi, trực quan hóa, tài liệu hóa và xuất dữ liệu sang các hệ thống và định dạng cần thiết, điều này đảm bảo tính tái sản xuất, minh bạch và tự động hóa công việc với thông tin.

Quy trình ETL đã được thiết kế có thể được sử dụng cho cả việc xử lý các dự án đơn lẻ, cũng như cho các ứng dụng quy mô lớn – khi phân tích hàng trăm và hàng nghìn dữ liệu đầu vào dưới dạng tài liệu, hình ảnh, bản quét, dự án CAD, đám mây điểm, tệp PDF hoặc các nguồn khác từ các hệ thống phân tán. Khả năng tự động hóa hoàn toàn quy trình biến ETL không chỉ thành công cụ xử lý kỹ thuật mà còn thành nền tảng của

cơ sở hạ tầng thông tin trong xây dựng số.

ETL với sự hỗ trợ của LLM: Hình ảnh hóa dữ liệu từ tài liệu PDF

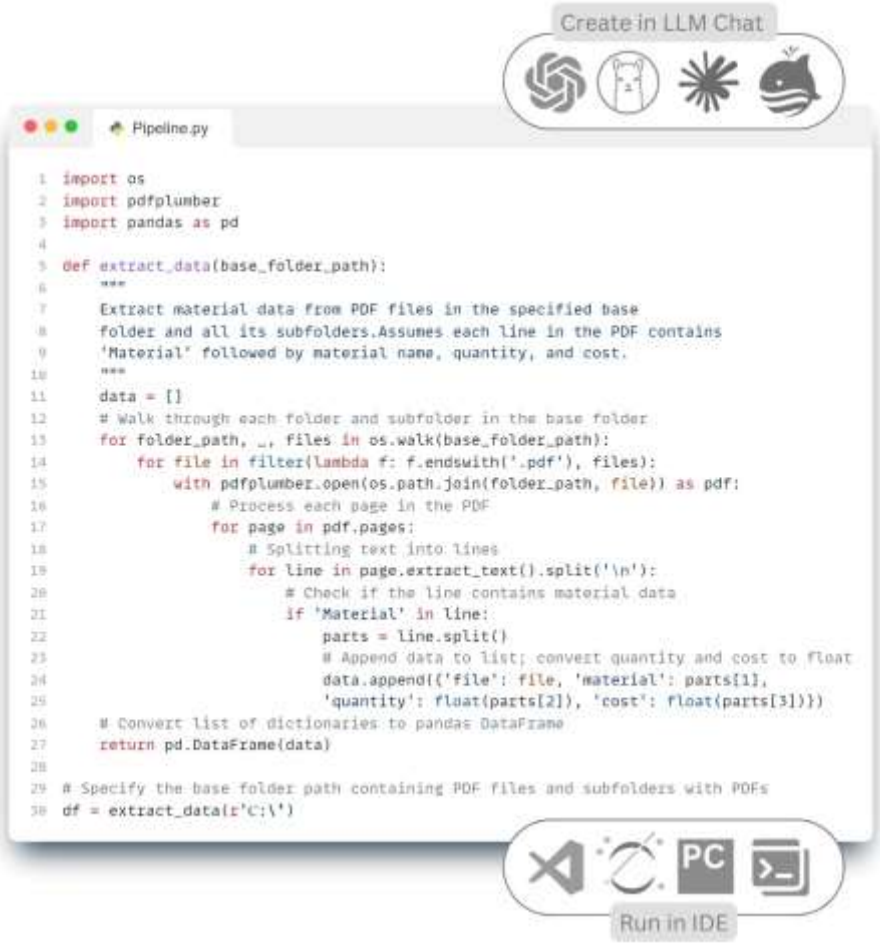
Đến lúc xây dựng một quy trình ETL hoàn chỉnh, bao gồm tất cả các bước chính trong làm việc với dữ liệu trong một kịch bản – trích xuất, chuyển đổi và tải. Chúng ta sẽ xây dựng một ETL-Pipeline tự động, cho phép xử lý các tài liệu PDF mà không cần làm thủ công – trích xuất dữ liệu từ các tài liệu, trực quan hóa, phân tích và truyền sang các hệ thống khác.

Quy trình ETL trong ví dụ của chúng tôi sẽ được mô tả qua các yêu cầu, mà phải giải thích cho mô hình ngôn ngữ (LLM) tất cả các quy trình ETL với mô tả kết quả cuối cùng cần đạt được. Trong trường hợp này, nhiệm vụ là tìm tất cả các tệp PDF trong thư mục được chỉ định và các thư mục con của nó, trích xuất thông tin liên quan từ chúng – chẳng hạn như tên vật liệu, số lượng và giá cả – và trình bày kết quả dưới dạng bảng có cấu trúc (DataFrame) để phân tích sau này.

- Yêu cầu văn bản đầu tiên tới LLM để tự động trích xuất dữ liệu từ nhiều tài liệu PDF và tạo DataFrame cho giai đoạn Extract:

> Viết mã để trích xuất thông tin về vật liệu từ các tệp PDF trong thư mục chỉ định và các thư mục con của nó. Dữ liệu trong PDF bao gồm tên vật liệu, số lượng và giá cả. Kết quả cần lưu vào DataFrame ⏎

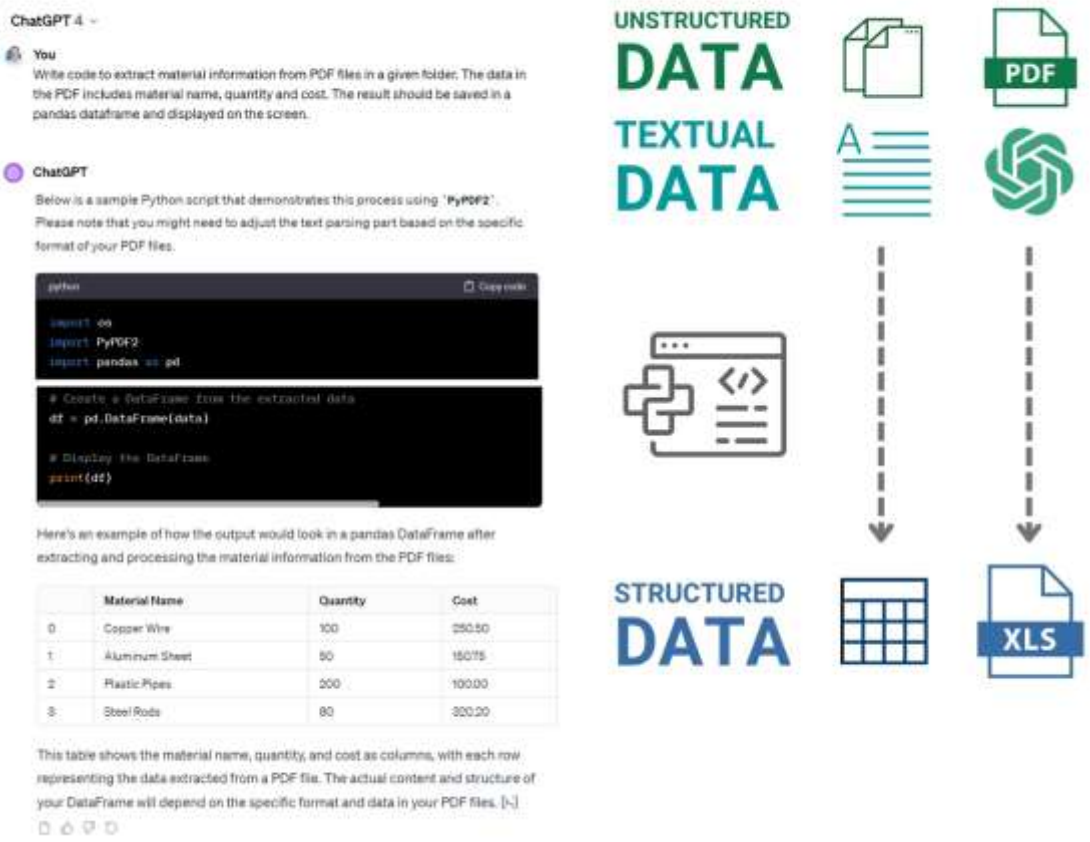
- Phản hồi từ LLM:



Hình 7.218 LLM tạo mã Python để trích xuất dữ liệu từ các tệp PDF trong thư mục cụ thể và tất cả các thư mục con của nó.

Phản hồi của LLM (Hình 7.218) là một mã Python hoàn chỉnh, tự động duyệt qua tất cả các thư mục, mở các tệp PDF tìm thấy, trích xuất thông tin văn bản từ chúng và chuyển đổi nó thành bảng. Mã thu được trong cuộc trò chuyện có thể được chạy trên một trong những IDE phổ biến như PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse với plugin PyDev, Thonny, Wing IDE, IntelliJ IDEA với plugin Python, JupyterLab hoặc các công cụ trực tuyến phổ biến như Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.-

Như một lựa chọn, để đơn giản hóa quy trình thu thập, thay vì sao chép mã từ LLM và sử dụng mã trong IDE, chúng ta cũng có thể tải trực tiếp một chục tệp PDF vào trò chuyện LLM (Hình 7.219) và nhận được một bảng kết quả ra, mà không cần xem mã hoặc chạy nó. Kết quả của việc thực hiện mã này sẽ là một bảng với các thuộc tính mà chúng tôi đã chọn.-

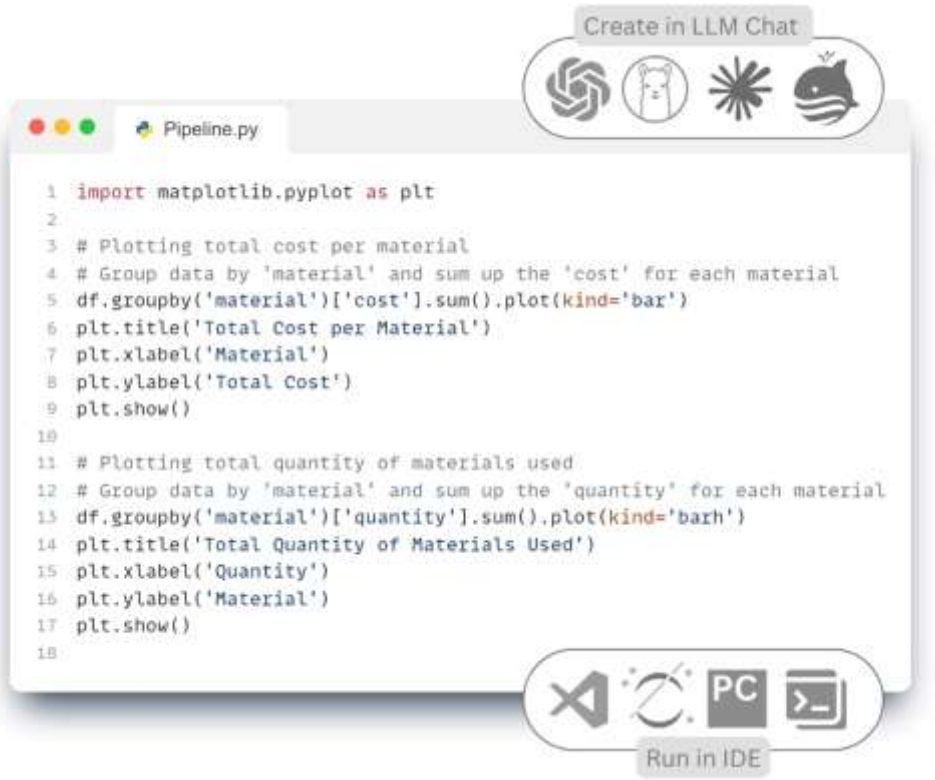


Hình 7.219 Kết quả thực hiện mã trong LLM, trích xuất dữ liệu từ các tệp PDF dưới dạng có cấu trúc của DataFrame với các thuộc tính đã chọn.

Ở giai đoạn tiếp theo, chúng ta yêu cầu mô hình ngôn ngữ phân tích dữ liệu thu được – ví dụ, so sánh giá cả và khối lượng sử dụng vật liệu và tạo một số ví dụ trực quan hóa, sẽ là cơ sở cho phân tích sau này.

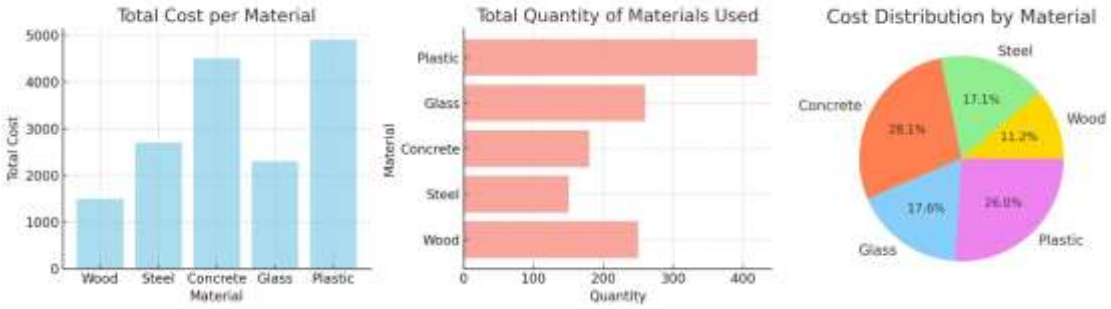
- Yêu cầu trong cuộc trò chuyện tiếp theo với LLM xây dựng một số biểu đồ từ bảng được lấy ở giai đoạn Transform (Hình 7.218): -

> Hình dung tổng chi phí và số lượng của từng loại vật liệu từ DataFrame (Hình 7.218) ⏎-



Hình 7.220 Phản hồi của mô hình LLM dưới dạng mã Python để trực quan hóa dữ liệu từ khung dữ liệu bằng thư viện matplotlib.

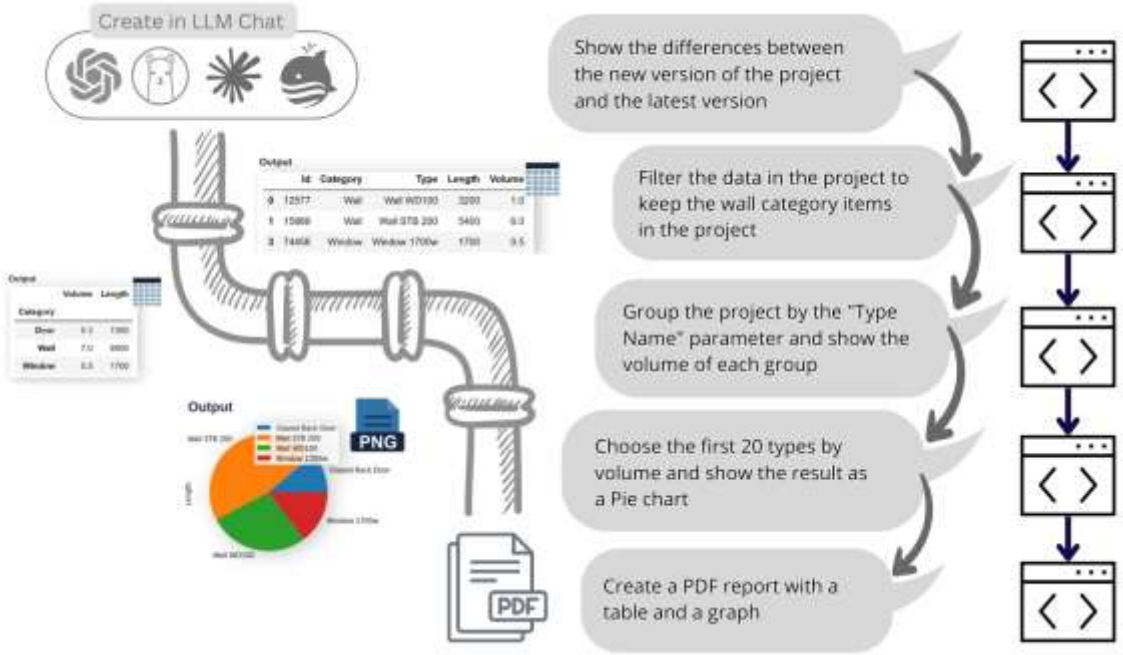
LLM tự động tạo và thực thi mã Python (Hình 7.220) sử dụng thư viện matplotlib. Sau khi thực hiện mã này, chúng ta nhận được biểu đồ chi phí và khối lượng sử dụng vật liệu trong các dự án xây dựng trực tiếp trong cuộc trò chuyện (Hình 7.221), điều này làm cho công việc phân tích trở nên đơn giản hơn nhiều.--



Hình 7.221 Trực quan hóa phản hồi của LLM dưới dạng biểu đồ dựa trên dữ liệu được thu thập trong DataFrame.

Hỗ trợ trong việc phát triển ý tưởng viết mã ETL, phân tích và thực hiện mã, trực quan hóa kết quả có sẵn thông qua các yêu cầu văn bản đơn giản trong LLM, mà không cần phải học những điều cơ bản về lập trình. Sự xuất hiện của những công cụ AI như LLM chắc chắn thay đổi cách tiếp cận lập trình và tự động hóa xử lý dữ liệu (Hình 7.222). -

> Theo báo cáo của PwC "Giá trị thực sự của trí tuệ nhân tạo đối với doanh nghiệp của bạn là gì và làm thế nào bạn có thể thu lợi từ nó?" (2017) [139], tự động hóa quy trình và nâng cao năng suất sẽ trở thành những động lực chính cho tăng trưởng kinh tế. Dự kiến, việc tăng năng suất lao động sẽ đóng góp hơn 55% toàn bộ mức tăng GDP nhờ AI trong giai đoạn 2017-2030."



Hình 7.222 AI LLM hỗ trợ tạo bản nháp mã, được áp dụng trong các dự án tương lai mà không cần sử dụng LLM.

> Sử dụng những công cụ như ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok, cùng với dữ liệu mở và phần mềm mã nguồn mở, chúng ta có thể tự động hóa những quy trình vốn chỉ được thực hiện bằng các hệ thống mô-đun độc quyền phức tạp và tốn kém.

Trong bối cảnh xây dựng, điều này có nghĩa là các công ty tiên phong trong việc triển khai các quy trình Pipeline tự động hóa xử lý dữ liệu sẽ nhận được lợi ích đáng kể — từ việc nâng cao hiệu quả quản lý dự án đến việc giảm thiểu mất mát tài chính và loại bỏ các ứng dụng phân tán và kho dữ liệu cô lập.

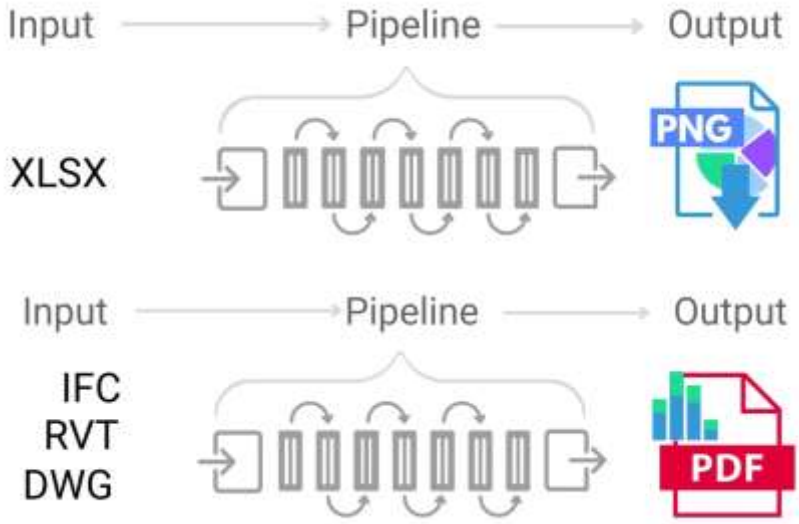
Logic được mô tả về việc thực hiện các nhiệm vụ kinh doanh trong quy trình ETL là một phần quan trọng của tự động hóa quy trình phân tích và xử lý dữ liệu, là một dạng cụ thể của khái niệm rộng hơn — các pipeline.

CHƯƠNG 7.3. **DÂY CHUYỀN ETL TỰ ĐỘNG (PIPELINE)**

Pipeline: Dây chuyền ETL tự động dữ liệu

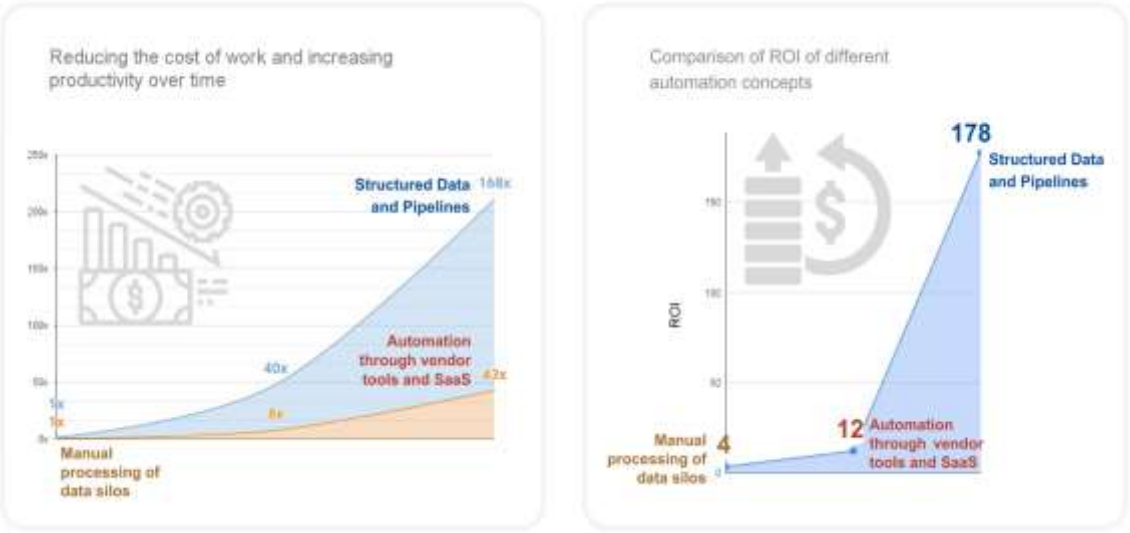
Quy trình ETL truyền thống được sử dụng để xử lý dữ liệu trong các hệ thống phân tích, bao gồm cả các nguồn dữ liệu có cấu trúc và không có cấu trúc. Tuy nhiên, trong môi trường kỹ thuật số hiện đại, thuật ngữ rộng hơn — Pipeline — ngày càng được sử dụng, mô tả bất kỳ chuỗi xử lý liên tục nào mà đầu ra của một giai đoạn trở thành đầu vào cho giai đoạn tiếp theo.

Cách tiếp cận này không chỉ áp dụng cho dữ liệu mà còn dành cho các hình thức tự động hóa khác: xử lý nhiệm vụ, xây dựng báo cáo, tích hợp với phần mềm và quản lý tài liệu điện tử (Hình 7.31).-



Hình 7.31 Pipeline — là một chuỗi xử lý, trong đó đầu ra của một giai đoạn trở thành đầu vào cho giai đoạn tiếp theo.

Việc áp dụng Pipeline là một trong những yếu tố chính của tự động hóa, đặc biệt trong điều kiện làm việc với một lượng lớn dữ liệu đa dạng. Kiến trúc đường ống cho phép tổ chức các giai đoạn xử lý phức tạp theo định dạng mô-đun, tuần tự và có thể quản lý, điều này nâng cao khả năng đọc, đơn giản hóa việc bảo trì mã và cho phép gỡ lỗi từng bước cũng như kiểm tra có thể mở rộng.



Hình 7.32 ROI Pipeline của quy trình kiểm tra dữ liệu giảm thời gian thực hiện hàng chục và hàng trăm lần so với việc xử lý bằng các công cụ truyền thống.

Khác với công việc thủ công trong các hệ thống độc quyền (ERP, PMIS, CAD, v.v.), xử lý dữ liệu theo đường ống cho phép tăng tốc độ thực hiện nhiệm vụ một cách đáng kể (Hình 7.32), tránh công việc lặp lại và tự động hóa việc khởi động các quy trình vào thời điểm cần thiết (Hình 7.33).--

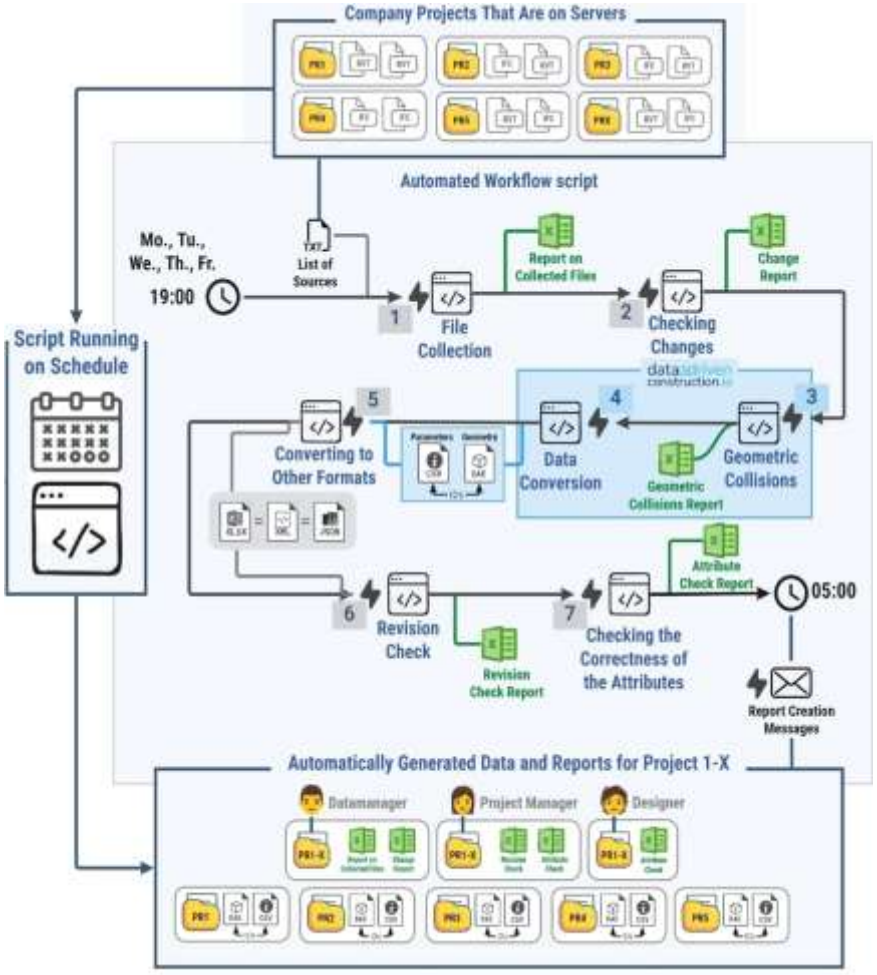


Hình 7.33 Ví dụ về ETL Pipeline để tự động lấy lịch trình từ dữ liệu bảng trong tệp XLSX mà không cần mở Excel.

Để xử lý dữ liệu theo luồng và xây dựng Pipeline tự động, tương tự như quy trình ETL, cần xác định trước các nguồn dữ liệu cũng như khung thời gian thu thập chúng - cho một quy trình kinh doanh cụ thể hoặc trong toàn bộ công ty.

> Trong các dự án xây dựng, dữ liệu đến từ nhiều nguồn khác nhau với tần suất cập nhật khác nhau. Để hình thành một kho dữ liệu đáng tin cậy, việc ghi lại thời điểm trích xuất và cập nhật thông tin là rất quan trọng. Điều này giúp đảm bảo quyết định kịp thời và nâng cao tính linh hoạt trong quản lý dự án.

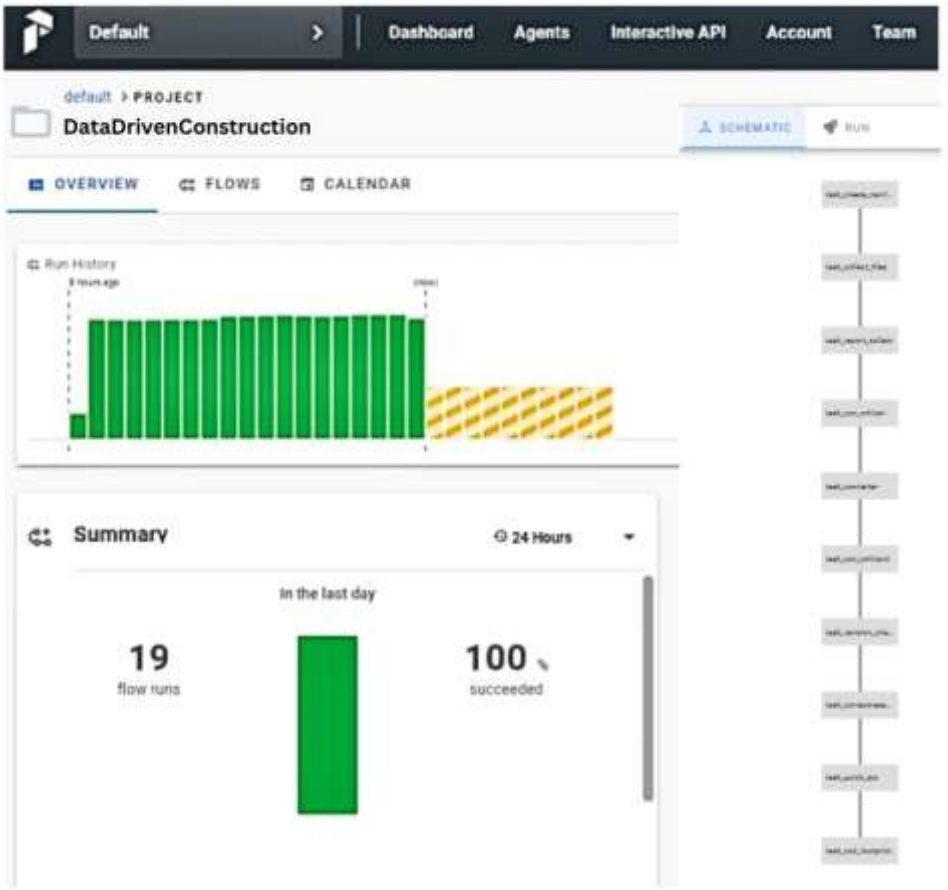
Một trong những lựa chọn là khởi động quy trình thu thập vào một thời điểm cố định - chẳng hạn như lúc 19:00, sau khi kết thúc ngày làm việc. Vào thời điểm này, kịch bản đầu tiên được kích hoạt, chịu trách nhiệm tổng hợp dữ liệu từ các hệ thống và kho lưu trữ khác nhau (Hình 7.34 bước 1). Tiếp theo là xử lý tự động và chuyển đổi dữ liệu thành định dạng có cấu trúc, phù hợp cho phân tích (Hình 7.34 bước 2-4). Ở giai đoạn cuối, sử dụng dữ liệu đã chuẩn bị, các báo cáo, bảng điều khiển và các sản phẩm khác được tự động hình thành như đã mô tả trong các chương trước (Hình 7.34 bước 6-7). Kết quả là vào lúc 05:00 sáng, các nhà quản lý đã có báo cáo cập nhật về tình trạng dự án ở định dạng cần thiết (Hình 7.35).---



Hình 7.34 Dữ liệu trong Pipeline, được thu thập tự động vào buổi tối, được xử lý vào ban đêm, để vào buổi sáng các nhà quản lý có báo cáo cập nhật và báo cáo mới.

> Việc thu thập dữ liệu kịp thời, xác định KPI, tự động hóa quy trình chuyển đổi và trực quan hóa thông qua bảng thông tin - là những yếu tố chính cho việc ra quyết định thành công dựa trên dữ liệu.

Các quy trình tự động hóa tương tự (Hình 7.34) có thể được thực hiện với sự tự chủ hoàn toàn: chúng được khởi động theo lịch trình, xử lý dữ liệu mà không cần sự tham gia của người vận hành và có thể được triển khai cả trên đám mây lẫn trên máy chủ riêng của công ty (Hình 7.35). Điều này cho phép tích hợp các đường ống ETL như vậy vào hạ tầng CNTT hiện có, giữ lại quyền kiểm soát dữ liệu và đảm bảo tính linh hoạt khi mở rộng.



Hình 7.35 ETL Pipeline tự động của các quy trình (Hình 7.34) trên nền tảng Prefect, trong đó 10 kịch bản python được khởi động lần lượt sau 19:00 mỗi ngày làm việc.-

Tự động hóa quy trình làm việc không chỉ nâng cao năng suất của đội ngũ, giải phóng thời gian cho những nhiệm vụ quan trọng hơn và ít lặp đi lặp lại hơn, mà còn là bước đầu tiên quan trọng trong việc triển khai công nghệ trí tuệ nhân tạo (AI) vào các quy trình kinh doanh, điều mà chúng ta sẽ bàn luận chi tiết hơn trong chương "Dự đoán và học máy".

Quá trình kiểm tra dữ liệu của Pipeline-ETL với LLM

Trong các chương trước, tập trung vào việc tạo yêu cầu dữ liệu và tự động hóa ETL, chúng ta đã phân tích từng bước quy trình chuẩn bị, chuyển đổi, xác thực và trực quan hóa dữ liệu. Những hành động này đã được thực hiện dưới dạng các khối mã riêng biệt (Hình 7.218 - Hình 7.220), mỗi khối thực hiện một nhiệm vụ cụ thể.--

Bây giờ, mục tiêu tiếp theo của chúng ta là kết hợp những yếu tố này thành một quy trình xử lý dữ liệu tự động và liên kết — quy trình ETL-Pipeline — trong đó tất cả các bước (tải lên, kiểm tra, trực quan hóa, xuất khẩu) được thực hiện tuần tự trong một tập lệnh tự động.

Trong ví dụ tiếp theo, sẽ được thực hiện chu trình xử lý dữ liệu hoàn chỉnh: từ việc tải lên tệp CSV gốc → đến việc kiểm tra cấu trúc và giá trị bằng cách sử dụng biểu thức chính quy → đếm kết quả → tạo báo cáo trực quan dưới định dạng PDF.

- Để nhận được mã tương ứng, bạn có thể sử dụng truy vấn văn bản sau cho LLM:

> Vui lòng viết một ví dụ mã, thực hiện việc tải dữ liệu từ CSV, kiểm tra dữ liệu DataFrame bằng cách sử dụng biểu thức chính quy, kiểm tra các định danh theo định dạng 'W-NEW' hoặc 'W-OLD', hiệu suất năng lượng với các chữ cái từ 'A' đến 'G', thời gian bảo hành và chu kỳ thay thế với các giá trị số bằng năm và cuối cùng tạo báo cáo với số lượng đã vượt qua và không vượt qua kiểm tra, tạo PDF với biểu đồ kết quả và thêm mô tả văn bản ⏎

Phản hồi từ LLM:



Hình 7.36 Pipeline (ETL) tự động hóa chu trình xử lý dữ liệu hoàn chỉnh: từ việc tải lên và kiểm tra đến việc tạo báo cáo có cấu trúc dưới định dạng PDF.

Mã tự động (Hình 7.36) bên trong chat LLM hoặc trong DIE, sau khi sao chép mã, sẽ thực hiện xác thực dữ

liệu từ tệp CSV bằng cách sử dụng các biểu thức chính quy đã chỉ định, tạo báo cáo về số lượng bản ghi đã vượt qua và không vượt qua kiểm tra, và sau đó lưu kết quả kiểm tra dưới dạng tệp PDF.

Cấu trúc ETL-Pipeline như vậy, nơi mỗi bước — từ việc tải dữ liệu đến việc tạo báo cáo — được thực hiện như một mô-đun riêng biệt, đảm bảo tính minh bạch, khả năng mở rộng và khả năng tái tạo. Việc trình bày logic kiểm tra dưới dạng mã Python dễ đọc giúp quy trình trở nên minh bạch và dễ hiểu không chỉ cho các nhà phát triển mà còn cho các chuyên gia trong lĩnh vực quản lý dữ liệu, chất lượng và phân tích.

> Việc sử dụng phương pháp Pipeline để tự động hóa xử lý dữ liệu cho phép chuẩn hóa các quy trình, nâng cao khả năng lặp lại và đơn giản hóa việc thích ứng với các dự án mới. Nhờ đó, một phương pháp phân tích dữ liệu thống nhất được hình thành, bất kể nguồn gốc hay loại nhiệm vụ — cho dù là kiểm tra sự tuân thủ tiêu chuẩn, tạo báo cáo hay truyền dữ liệu sang các hệ thống bên ngoài.

Tự động hóa như vậy giảm thiểu ảnh hưởng của yếu tố con người, giảm sự phụ thuộc vào các giải pháp độc quyền và nâng cao độ chính xác và độ tin cậy của kết quả, làm cho chúng phù hợp cho cả phân tích hoạt động ở cấp độ dự án và phân tích chiến lược ở cấp độ công ty.

Pipeline-ETL: kiểm tra dữ liệu và thông tin các phần tử dự án trong CAD (BIM)

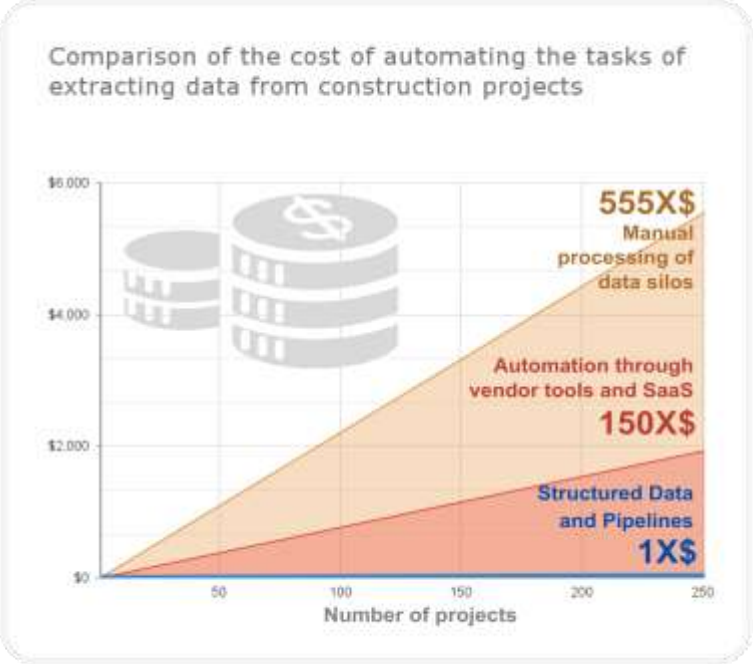
Dữ liệu từ các hệ thống và cơ sở dữ liệu CAD (BIM) là một trong những nguồn dữ liệu phức tạp và thường xuyên được cập nhật nhất trong ngành xây dựng. Những ứng dụng này không chỉ mô tả dự án bằng hình học mà còn bổ sung thông tin văn bản đa lớp: khối lượng, tính chất vật liệu, mục đích sử dụng các phòng, mức độ hiệu quả năng lượng, dung sai, thời gian sử dụng và các thuộc tính khác.

> Các thuộc tính được gán cho các thực thể trong mô hình CAD được hình thành trong giai đoạn thiết kế và trở thành cơ sở cho các quy trình kinh doanh tiếp theo, bao gồm tính toán chi phí, lập lịch trình, đánh giá vòng đời và tích hợp với các hệ thống ERP và CAFM, nơi mà hiệu quả của các quy trình phụ thuộc nhiều vào chất lượng dữ liệu từ các phòng thiết kế.

Cách tiếp cận truyền thống để kiểm tra các thuộc tính trong mô hình CAD (BIM) yêu cầu xác thực thủ công, điều này trở thành một quá trình dài và tốn kém khi khối lượng mô hình lớn. Với khối lượng và số lượng các dự án xây dựng hiện đại và sự cập nhật thường xuyên của chúng, quá trình kiểm tra và chuyển đổi dữ liệu trở nên không ổn định và khó khăn.-

> Các nhà thầu chính và quản lý dự án phải đối mặt với nhu cầu xử lý một lượng lớn dữ liệu dự án, bao gồm nhiều phiên bản và mảnh ghép của cùng một mô hình. Dữ liệu được cung cấp từ các tổ chức thiết kế dưới các định dạng RVT, DWG, DGN, IFC, NWD và các định dạng khác và yêu cầu kiểm tra thường xuyên để đảm bảo tuân thủ các tiêu chuẩn ngành và doanh nghiệp.

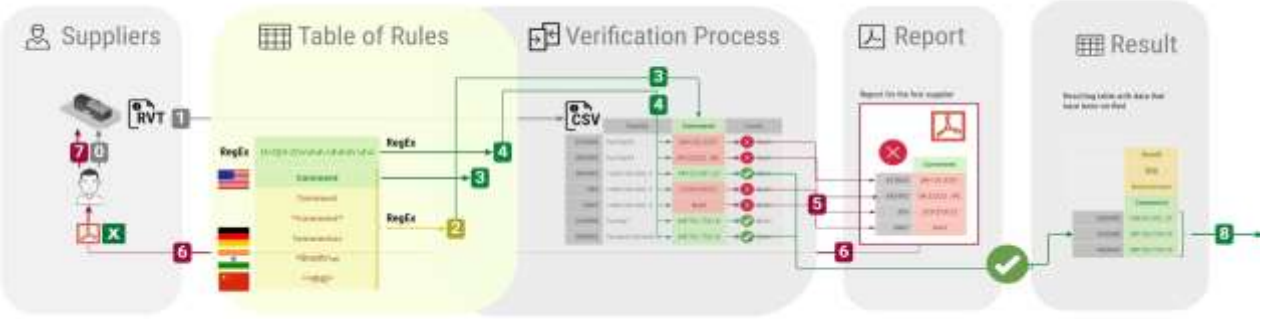
Do phụ thuộc vào các hành động thủ công và phần mềm chuyên dụng, quá trình xác thực dữ liệu trở thành điểm nghẽn trong các quy trình làm việc liên quan đến dữ liệu từ các mô hình cho toàn công ty. Tự động hóa và áp dụng các yêu cầu có cấu trúc cho phép loại bỏ sự phụ thuộc này, tăng tốc độ và độ tin cậy của việc kiểm tra dữ liệu lên nhiều lần.



Tự động hóa tăng tốc độ kiểm tra và xử lý dữ liệu, giúp giảm chi phí công việc xuống hàng chục lần.

Quá trình kiểm tra dữ liệu CAD bao gồm việc xuất dữ liệu (giai đoạn ETL Extract) từ các định dạng đóng (RVT, DWG, DGN, NWS, v.v.) hoặc các định dạng bán cấu trúc và tham số mở (IFC, CPXML, USD), trong đó các bảng quy tắc có thể được áp dụng cho từng thuộc tính và giá trị của nó (giai đoạn Transform) bằng cách sử dụng biểu thức chính quy RegEx, quá trình mà chúng tôi đã xem xét chi tiết trong phần bốn của cuốn sách. -

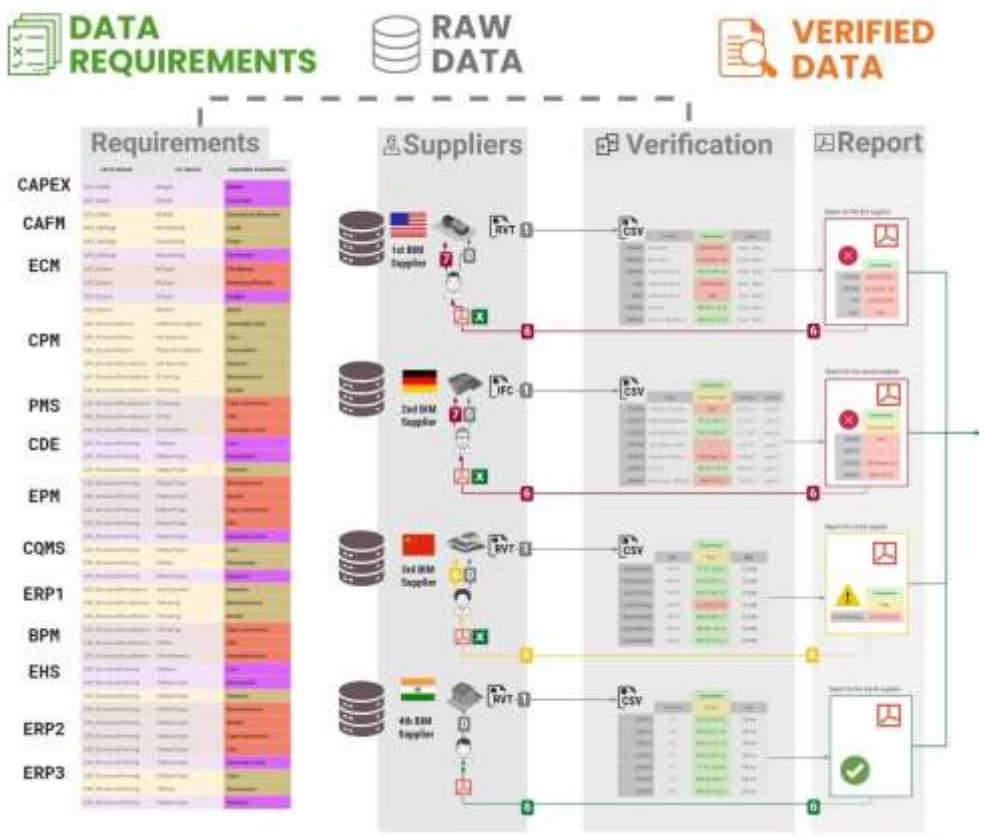
Việc tạo báo cáo lỗi ở định dạng PDF và các bản ghi đã được kiểm tra thành công phải kết thúc bằng việc xuất (giai đoạn Load) sang các định dạng có cấu trúc, chỉ bao gồm các thực thể đã được kiểm tra, có thể được sử dụng cho các quy trình tiếp theo.



Quá trình kiểm tra dữ liệu từ các nhà cung cấp dữ liệu thiết kế đến báo cáo cuối cùng, đã được kiểm tra bằng cách sử dụng biểu thức chính quy.

Tự động hóa việc kiểm tra dữ liệu từ các hệ thống CAD (BIM) khi có các yêu cầu có cấu trúc và khi có dòng dữ liệu mới liên tục được xử lý qua các ETL-Pipelines giảm thiểu sự cần thiết phải tham gia thủ công trong

quá trình xác thực (mỗi quy trình kiểm tra và lập yêu cầu dữ liệu đã được xem xét trong các chương trước).
-



Tự động hóa việc kiểm tra dữ liệu qua ETL đơn giản hóa việc quản lý các dự án xây dựng nhờ vào việc tăng tốc các quy trình.

Truyền thống, việc kiểm tra các mô hình do nhà thầu và các chuyên gia CAD (BIM) cung cấp có thể mất từ vài ngày đến vài tuần. Tuy nhiên, với việc triển khai các quy trình ETL tự động hóa, thời gian này có thể rút ngắn xuống chỉ còn vài phút. Trong tình huống điển hình, nhà thầu tuyên bố: "Mô hình đã được kiểm tra và đáp ứng các yêu cầu". Tuyên bố như vậy khởi động một chuỗi kiểm tra tuyên bố của nhà thầu về chất lượng dữ liệu:

- Quản lý dự án - "Nhà thầu khẳng định: 'Mô hình đã được kiểm tra, mọi thứ đều ổn'".
- Quản lý dữ liệu - "Tải lên xác thực":
  - Một đoạn mã đơn giản trong Pandas phát hiện vi phạm chỉ trong vài giây. Tự động hóa loại bỏ tranh cãi:
    - Danh mục: OST_StructuralColumns, Tham số: FireRating IS NULL.
    - Tạo danh sách ID vi phạm → xuất sang Excel/PDF.

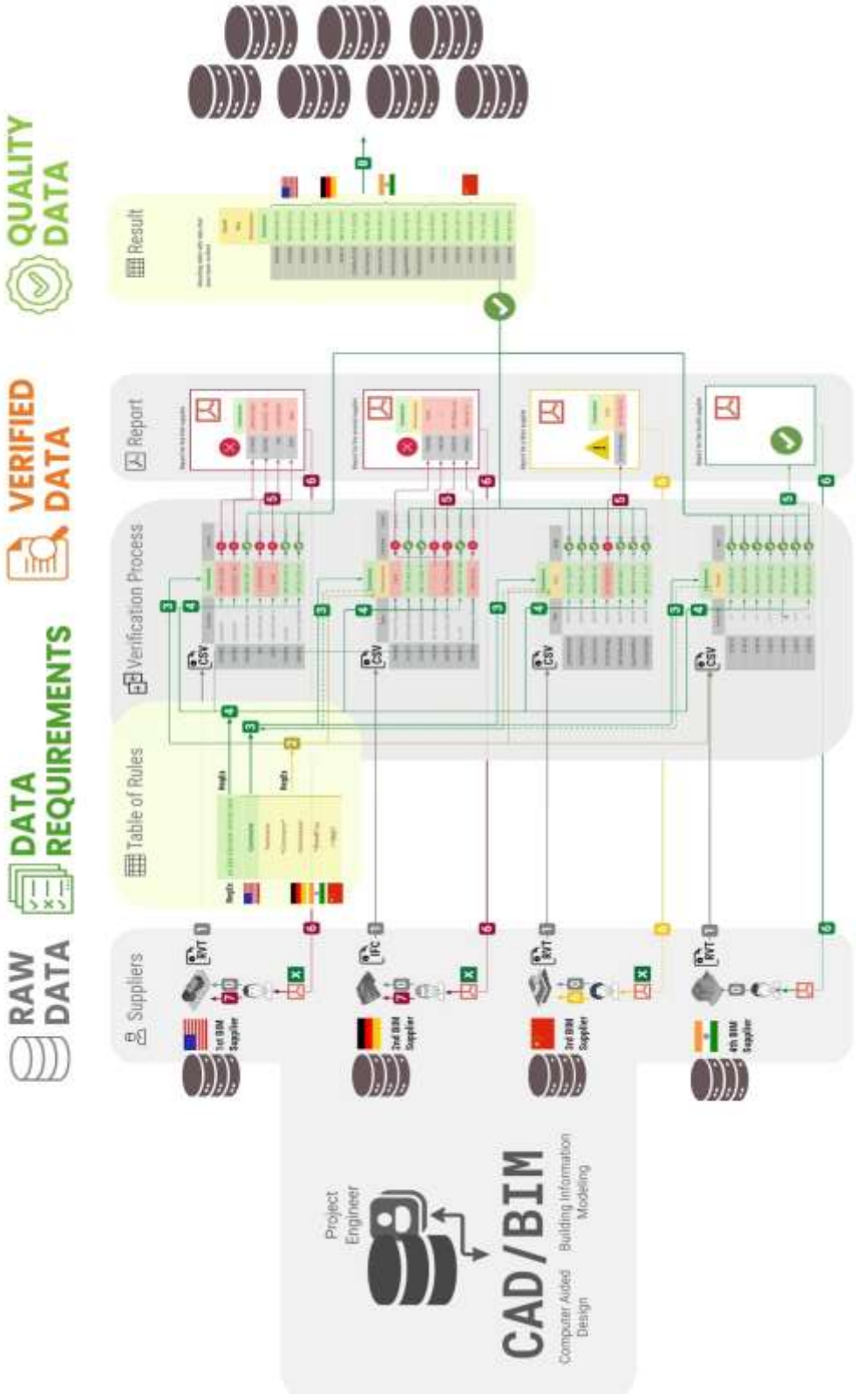
Một đoạn mã đơn giản trong Pandas phát hiện vi phạm chỉ trong vài giây:

```
df = model_data[model_data["Category"] == "OST_StructuralColumns"]  # Lọc  issues
= df[df["FireRating"].isnull()]  # Giá trị trống
issues[["ElementID"]].to_excel("fire_rating_issues.xlsx")  # Xuất ID
```

- Quản lý dữ liệu cho quản lý dự án - "Kiểm tra cho thấy rằng 18 cột không có tham số FireRating".
- Quản lý dự án cho nhà thầu - "Mô hình được trả lại để chỉnh sửa: tham số FireRating là bắt buộc, không có nó thì không thể tiếp nhận".

Kết quả là mô hình CAD không vượt qua kiểm tra chất lượng, tự động hóa loại bỏ tranh cãi, và nhà thầu gần như ngay lập tức nhận được báo cáo có cấu trúc với danh sách ID của các phần tử gặp vấn đề. Do đó, quy trình xác thực trở nên minh bạch, có thể tái tạo và được bảo vệ khỏi yếu tố con người (Hình. 7.310).

> Cách tiếp cận này biến quy trình kiểm tra dữ liệu thành một chức năng kỹ thuật, chứ không phải là kiểm soát chất lượng thủ công. Điều này không chỉ nâng cao năng suất mà còn cho phép áp dụng cùng một logic cho tất cả các dự án của công ty, đảm bảo chuyển đổi số toàn diện trong các quy trình, từ thiết kế đến vận hành.



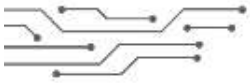
Hình. 7.310 Tự động hóa kiểm tra thuộc tính của các phần tử loại bỏ yếu tố con người và giảm khả năng xảy ra lỗi.

Nhờ vào việc áp dụng các dây chuyền tự động hóa (Hình. 7.310), người dùng hệ thống, những người mong đợi dữ liệu chất lượng từ các hệ thống CAD- (BIM-) có thể ngay lập tức nhận được các đầu ra cần thiết - bảng, tài liệu, hình ảnh - và nhanh chóng tích hợp chúng vào các nhiệm vụ công việc của mình.

> Tự động hóa kiểm soát, xử lý và phân tích dẫn đến sự thay đổi trong các phương pháp quản lý dự án xây dựng, đặc biệt là trong việc tương tác giữa các hệ thống khác nhau, mà không cần sử dụng các hệ thống mô-đun phức tạp và đắt tiền hoặc các giải pháp đóng từ các nhà cung cấp.

Trong khi các khái niệm và từ viết tắt tiếp thị đến và đi, chính các quy trình kiểm tra yêu cầu dữ liệu sẽ mãi mãi là một phần không thể thiếu trong các quy trình kinh doanh. Thay vì tạo ra ngày càng nhiều định dạng và tiêu chuẩn chuyên biệt mới, ngành xây dựng nên chú ý đến các công cụ đã chứng minh hiệu quả trong các lĩnh vực kinh tế khác. Ngày nay, có những nền tảng mạnh mẽ cho việc tự động hóa xử lý dữ liệu và tích hợp quy trình, cho phép các công ty giảm đáng kể thời gian cho các hoạt động thường ngày và tối thiểu hóa lỗi trong các quy trình Extract, Transform và Load.

Một trong những ví dụ phổ biến về giải pháp tự động hóa và điều phối quy trình ETL là Apache Airflow, cho phép tổ chức các quy trình tính toán phức tạp và quản lý các pipeline ETL. Bên cạnh Airflow, còn có nhiều giải pháp tương tự khác được sử dụng tích cực, chẳng hạn như Apache NiFi cho việc định tuyến và xử lý dữ liệu theo luồng, cũng như n8n cho việc tự động hóa quy trình kinh doanh.



CHƯƠNG 7.4. **ĐIỀU PHỐI ETL VÀ QUY TRÌNH LÀM VIỆC: GIẢI PHÁP THỰC TẾ**

DAG và Apache Airflow: tự động hóa và điều phối quy trình làm việc

Apache Airflow là một nền tảng mã nguồn mở miễn phí, được thiết kế để tự động hóa, điều phối và giám sát các quy trình làm việc (pipeline ETL).

Mỗi ngày, trong công việc với khối lượng dữ liệu lớn, cần phải:

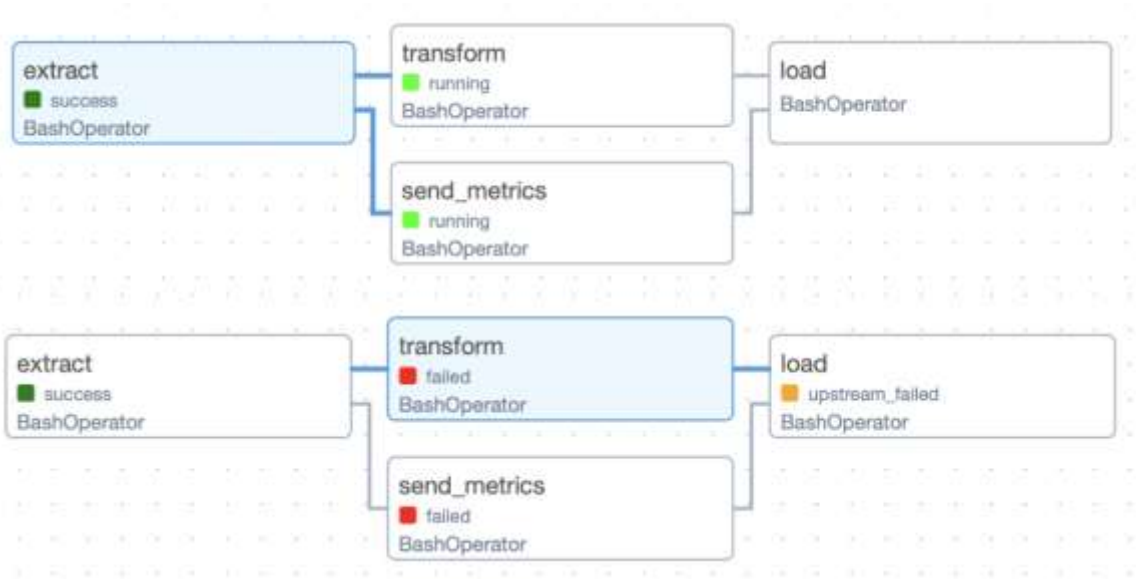
- Tải xuống các tệp từ nhiều nguồn khác nhau - Extract (ví dụ, từ nhà cung cấp hoặc khách hàng).
- Chuyển đổi dữ liệu này sang định dạng cần thiết - Transform (cấu trúc, làm sạch và kiểm tra).
- Gửi kết quả để kiểm tra và tạo báo cáo - Load (xuất vào các hệ thống, tài liệu, cơ sở dữ liệu hoặc bảng điều khiển).

Việc thực hiện thủ công các quy trình ETL như vậy tốn nhiều thời gian và dẫn đến rủi ro sai sót liên quan đến yếu tố con người. Việc thay đổi nguồn dữ liệu hoặc sự cố ở một trong các giai đoạn có thể gây ra sự chậm trễ và kết quả không chính xác.

Các công cụ tự động hóa, chẳng hạn như Apache Airflow, cho phép xây dựng một pipeline ETL đáng tin cậy, giảm thiểu sai sót, rút ngắn thời gian xử lý dữ liệu và đảm bảo tính chính xác của chúng ở mỗi giai đoạn. Cốt lõi của Apache Airflow là khái niệm DAG (Directed Acyclic Graph) - đồ thị có hướng không chu trình, trong đó mỗi nhiệm vụ (operator) liên kết với các phụ thuộc khác và được thực hiện theo một trình tự nhất định. DAG loại trừ các chu trình, điều này đảm bảo cấu trúc thực hiện nhiệm vụ hợp lý và có thể dự đoán.

Airflow đảm nhận việc điều phối - quản lý các phụ thuộc giữa các nhiệm vụ, kiểm soát lịch trình thực hiện, theo dõi trạng thái và phản ứng tự động với các sự cố. Cách tiếp cận này giảm thiểu sự can thiệp thủ công và đảm bảo độ tin cậy của toàn bộ quy trình.

> Bộ điều phối nhiệm vụ - công cụ hoặc hệ thống được thiết kế để quản lý và kiểm soát việc thực hiện các nhiệm vụ trong các môi trường tính toán và thông tin phức tạp. Nó giúp đơn giản hóa quy trình triển khai, tự động hóa và quản lý việc thực hiện các nhiệm vụ, từ đó nâng cao hiệu quả công việc và tối ưu hóa tài nguyên.



Hình 7.41 Apache Airflow cung cấp giao diện thân thiện, nơi có thể trực quan hóa DAG-ETL, xem nhật ký thực hiện, trạng thái khởi động nhiệm vụ và nhiều thông tin khác.

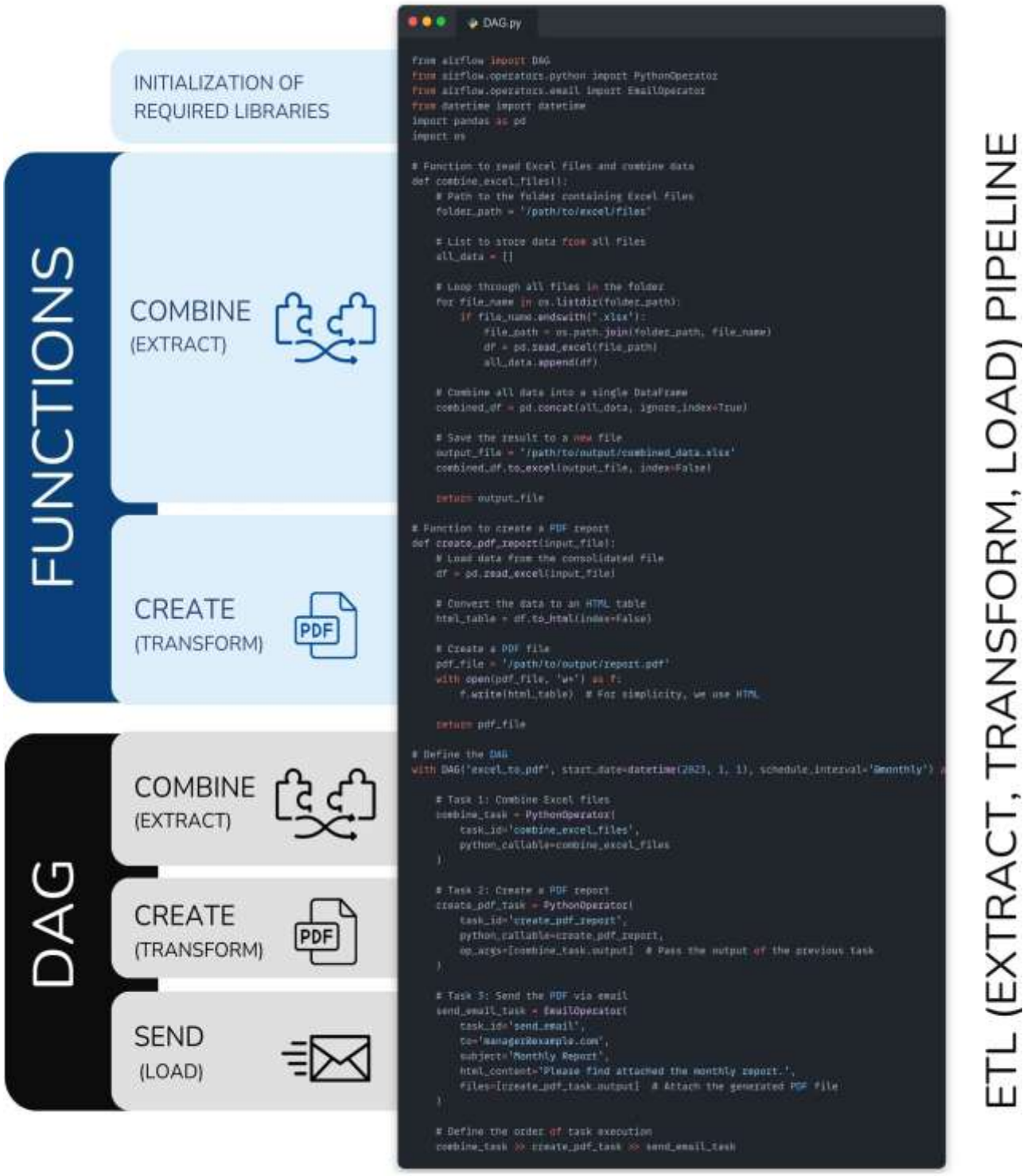
Airflow được sử dụng rộng rãi để điều phối và tự động hóa các tính toán phân tán, xử lý dữ liệu, quản lý quy trình ETL (Extract, Transform, Load), lập lịch nhiệm vụ và các kịch bản làm việc với dữ liệu khác. Theo mặc định, Apache Airflow sử dụng SQLite làm cơ sở dữ liệu.

Ví dụ về một DAG đơn giản, tương tự như ETL, bao gồm các nhiệm vụ - Extract, Transform và Load. Trong đồ thị, được quản lý qua giao diện người dùng (Hình 7.41), đã xác định thứ tự thực hiện các nhiệm vụ (đoạn mã): ví dụ, trước tiên thực hiện extract, sau đó là transform (và sending_metrics), và cuối cùng là nhiệm vụ load. Khi tất cả các nhiệm vụ được thực hiện, quá trình tải dữ liệu được coi là thành công.-

Apache Airflow: ứng dụng thực tiễn trong tự động hóa ETL

Apache Airflow được sử dụng rộng rãi để tổ chức các quy trình xử lý dữ liệu phức tạp, cho phép xây dựng các pipeline ETL linh hoạt. Apache Airflow có thể được khởi chạy thông qua giao diện web hoặc lập trình thông qua mã Python (Hình 7.42). Trong giao diện web (Hình 7.43), các quản trị viên và nhà phát triển có thể theo dõi trực quan các DAG, khởi động các tác vụ và phân tích kết quả thực hiện.--

Sử dụng DAG, có thể xác định một trình tự rõ ràng cho việc thực hiện các tác vụ, quản lý các phụ thuộc giữa chúng và tự động phản ứng với những thay đổi trong dữ liệu đầu vào. Hãy xem xét một ví dụ về việc sử dụng Airflow để tự động hóa quy trình báo cáo (Hình 7.42).-



Hình 7.42 Khái niệm pipeline ETL để xử lý dữ liệu bằng Apache Airflow.

Trong ví dụ này (Hình 7.42), một DAG được xem xét, thực hiện các tác vụ chính trong khuôn khổ pipeline ETL:-

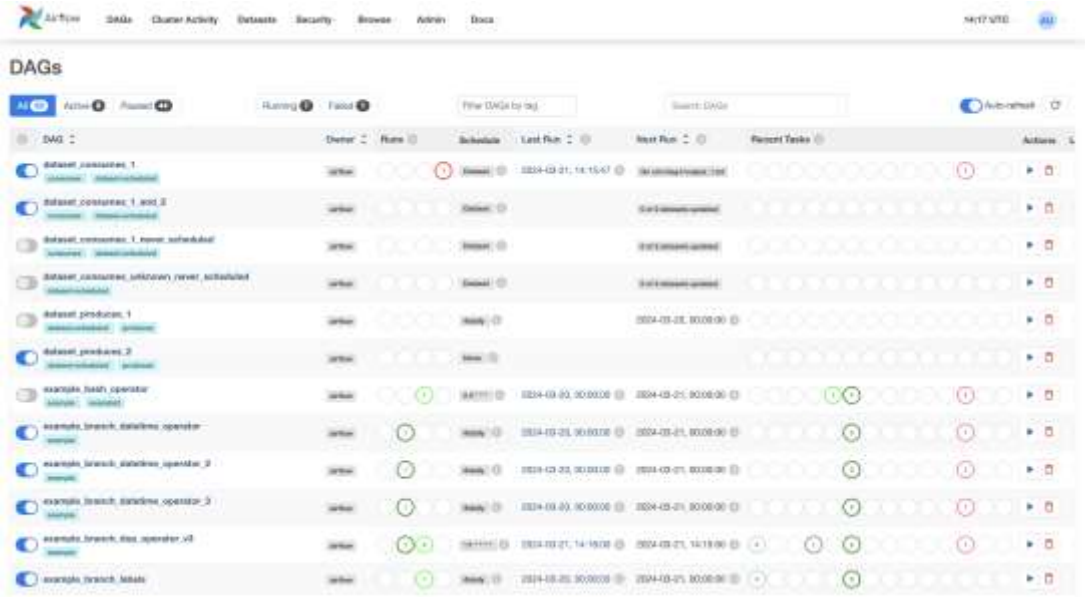
- Đọc các tệp Excel (Extract): • Duyệt tuần tự tất cả các tệp trong thư mục đã chỉ định. • Đọc dữ liệu từ mỗi tệp bằng cách sử dụng thư viện pandas. • Kết hợp tất cả dữ liệu vào một DataFrame duy nhất.

- Tạo tài liệu PDF (Transform): • Chuyển đổi DataFrame đã kết hợp thành bảng HTML. • Lưu bảng ở định dạng PDF (trong phiên bản demo - thông qua HTML).

- Gửi báo cáo qua email (Load): • Sử dụng EmailOperator để gửi tài liệu PDF qua email.

- Cấu hình DAG: • Xác định trình tự thực hiện các tác vụ: trích xuất dữ liệu → tạo báo cáo → gửi đi. • Đặt lịch khởi động (@monthly - ngày đầu tiên của mỗi tháng).

Trong ví dụ ETL tự động (Hình 7.42), cho thấy cách thu thập dữ liệu từ các tệp Excel, tạo tài liệu PDF và gửi nó qua email. Đây chỉ là một trong nhiều kịch bản có thể sử dụng Airflow. Ví dụ này có thể được điều chỉnh cho bất kỳ nhiệm vụ cụ thể nào, nhằm đơn giản hóa và tự động hóa các quy trình xử lý dữ liệu.-



Hình 7.43 Tổng quan về tất cả các nhóm DAG trong môi trường với thông tin về các lần khởi động gần đây.

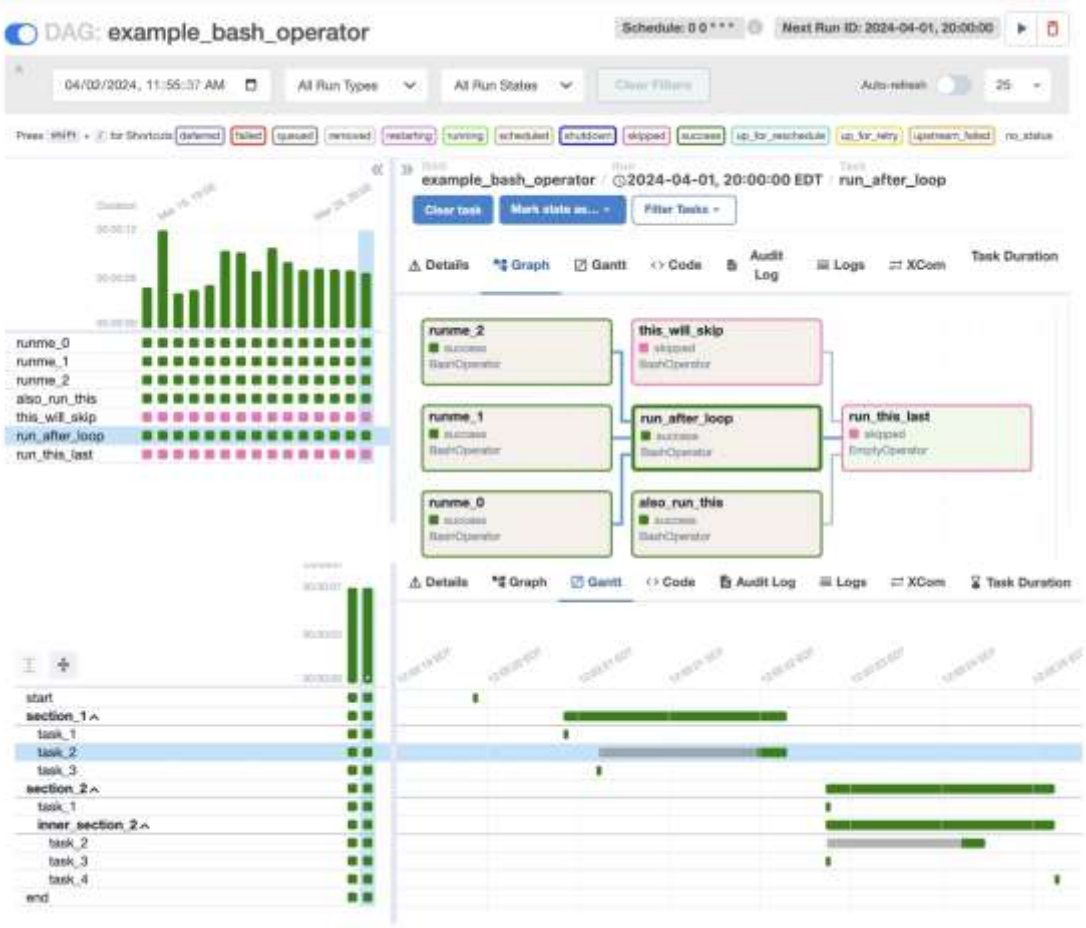
Giao diện web của Apache Airflow (Hình 7.43) cung cấp một môi trường trực quan toàn diện để quản lý các quy trình dữ liệu. Nó hiển thị các DAG dưới dạng đồ thị tương tác, trong đó các nút đại diện cho các tác vụ và các cạnh là các phụ thuộc giữa chúng, cho phép dễ dàng theo dõi các quy trình xử lý dữ liệu phức tạp. Giao diện bao gồm bảng điều khiển với thông tin về trạng thái thực hiện các tác vụ, lịch sử khởi động, nhật ký chi tiết và các chỉ số hiệu suất. Các quản trị viên có thể khởi động các tác vụ thủ công, khởi động lại các hoạt động không thành công, tạm dừng các DAG và cấu hình các biến môi trường - tất cả đều thông qua giao diện người dùng trực quan.

Kiến trúc tương tự có thể được bổ sung bằng việc xác thực dữ liệu, thông báo về trạng thái thực hiện, tích hợp với các API hoặc cơ sở dữ liệu bên ngoài. Airflow cho phép linh hoạt điều chỉnh DAG: thêm các tác vụ

mới, thay đổi thứ tự của chúng, kết hợp các chuỗi - điều này làm cho nó trở thành một công cụ hiệu quả để tự động hóa các quy trình xử lý dữ liệu phức tạp. Khi khởi động DAG trong giao diện web của Airflow (Hình 7.43, Hình 7.44), có thể theo dõi trạng thái thực hiện các tác vụ. Hệ thống sử dụng chỉ báo màu sắc:--

- Màu xanh - tác vụ đã hoàn thành thành công.
- Màu vàng - quá trình đang được thực hiện.
- Màu đỏ - có lỗi trong quá trình thực hiện tác vụ.

Trong trường hợp xảy ra sự cố (ví dụ, tệp bị thiếu hoặc cấu trúc dữ liệu bị vi phạm), hệ thống sẽ tự động khởi động việc gửi thông báo.



Hình 7.44 Apache Airflow giúp đơn giản hóa việc chẩn đoán sự cố, tối ưu hóa quy trình và hợp tác giữa các nhóm trong việc xử lý các pipeline dữ liệu phức tạp.

> Apache Airflow rất tiện lợi vì tự động hóa các tác vụ lặp đi lặp lại, giúp loại bỏ nhu cầu thực hiện chúng một cách thủ công. Nó đảm bảo độ tin cậy thông qua việc giám sát quá trình thực hiện và thông báo ngay lập tức về các lỗi. Tính linh hoạt của hệ thống cho phép dễ dàng thêm các tác vụ mới hoặc thay đổi các tác vụ hiện có, điều chỉnh quy trình làm việc theo các yêu cầu thay đổi.

Ngoài Apache Airflow, còn có những công cụ tương tự để điều phối quy trình làm việc. Ví dụ, Prefect mã

nguồn mở và miễn phí (Hình 7.35) cung cấp cú pháp đơn giản hơn và tích hợp tốt hơn với Python, Luigi, được phát triển bởi Spotify, cung cấp chức năng tương tự và hoạt động tốt với dữ liệu lớn. Cũng cần lưu ý đến Kronos và Dagster, những công cụ này cung cấp các phương pháp hiện đại để tạo Pipeline với trọng tâm vào tính mô-đun và khả năng mở rộng. Việc lựa chọn công cụ điều phối tác vụ phụ thuộc vào nhu cầu cụ thể của dự án, nhưng tất cả đều giúp tự động hóa các quy trình ETL phức tạp trong xử lý dữ liệu.-

Apache NiFi xứng đáng được đề cập riêng — một nền tảng mã nguồn mở, được thiết kế cho việc xử lý và định tuyến dữ liệu theo luồng (streaming). Khác với Airflow, vốn tập trung vào xử lý theo lô và quản lý phụ thuộc, NiFi tập trung vào thời gian thực, chuyển đổi dữ liệu ngay lập tức và định tuyến linh hoạt giữa các hệ thống.

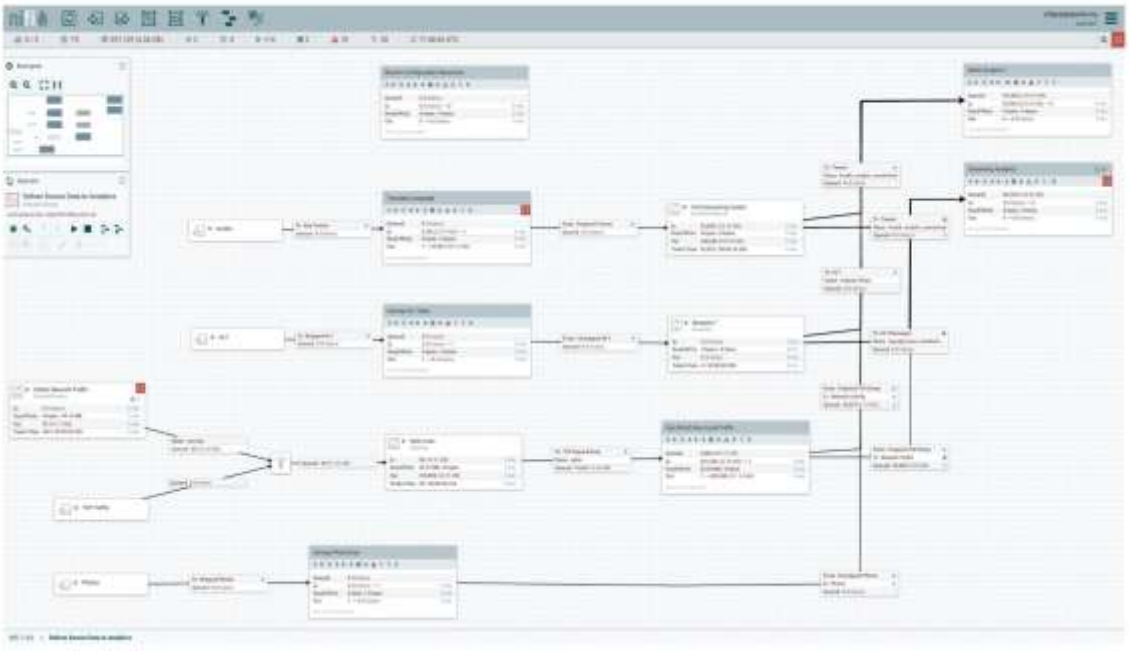
Apache NiFi cho việc định tuyến và chuyển đổi dữ liệu

Apache NiFi là một nền tảng mạnh mẽ mã nguồn mở, được thiết kế để tự động hóa các luồng dữ liệu giữa các hệ thống khác nhau. Ban đầu được phát triển vào năm 2006 bởi Cơ quan An ninh Quốc gia Hoa Kỳ (NSA) với tên gọi "Niagara Files" cho các nhu cầu nội bộ. Năm 2014, dự án đã được mở và chuyển giao cho Quỹ Phần mềm Apache, trở thành một phần trong các sáng kiến chuyển giao công nghệ của họ.

Apache NiFi được thiết kế để thu thập, xử lý và truyền dữ liệu theo thời gian thực. Khác với Airflow, vốn làm việc với các tác vụ theo lô và yêu cầu lịch trình rõ ràng, NiFi hoạt động theo chế độ xử lý theo luồng, cho phép truyền dữ liệu liên tục giữa các dịch vụ khác nhau.

> Apache NiFi hoàn hảo cho việc tích hợp với các thiết bị IoT, cảm biến công trình, hệ thống giám sát, và ví dụ như kiểm tra luồng các định dạng CAD trên máy chủ, nơi có thể yêu cầu phản ứng ngay lập tức với các thay đổi trong dữ liệu.

Nhờ vào các công cụ lọc, chuyển đổi và định tuyến tích hợp sẵn, NiFi cho phép chuẩn hóa dữ liệu (giai đoạn Transform) trước khi truyền (Load) vào các kho lưu trữ hoặc hệ thống phân tích. Một trong những lợi thế chính của nó là hỗ trợ bảo mật và kiểm soát truy cập tích hợp, điều này làm cho nó trở thành một giải pháp đáng tin cậy cho việc xử lý thông tin nhạy cảm.



Hình 7.45 Biểu diễn đồ họa của luồng dữ liệu trong giao diện Apache NiFi.

Apache NiFi hiệu quả trong việc giải quyết các nhiệm vụ truyền tải, lọc và định tuyến dữ liệu theo thời gian thực. Nó hoàn hảo cho các kịch bản công nghệ cao, nơi việc truyền thông tin ổn định giữa các hệ thống và băng thông cao là rất quan trọng.

Tuy nhiên, trong những trường hợp mà mục tiêu chính là tích hợp các dịch vụ đa dạng, tự động hóa các quy trình lặp đi lặp lại và nhanh chóng thiết lập quy trình làm việc mà không cần kiến thức sâu về lập trình, các giải pháp có ngưỡng đầu vào thấp và tính linh hoạt tối đa được ưa chuộng. Một trong những công cụ như vậy là n8n - nền tảng Low-Code/No-Code, hướng đến tự động hóa doanh nghiệp và điều phối quy trình một cách trực quan.

n8n Low-Code, No-Code cho việc điều phối quy trình

n8n là nền tảng Open Source Low-Code/No-Code để xây dựng các quy trình làm việc tự động, nổi bật với tính dễ sử dụng, linh hoạt và khả năng tích hợp nhanh chóng với nhiều dịch vụ bên ngoài.

> No-Code là phương pháp tạo ra các sản phẩm kỹ thuật số mà không cần viết mã. Tất cả các yếu tố của quy trình - từ logic đến giao diện - được thực hiện hoàn toàn bằng các công cụ trực quan. Các nền tảng No-Code nhắm đến người dùng không có nền tảng kỹ thuật và cho phép nhanh chóng tạo ra các tự động hóa, biểu mẫu, tích hợp và ứng dụng web. Ví dụ: người dùng có thể thiết lập việc gửi thông báo tự động hoặc tích hợp với Google Sheets thông qua giao diện kéo và thả mà không cần kiến thức lập trình.

Nhờ mã nguồn mở và khả năng triển khai cục bộ, n8n trong các quy trình tự động hóa và tạo ETL Pipelines mang lại cho các công ty quyền kiểm soát hoàn toàn dữ liệu của họ, đảm bảo an toàn và độc lập khỏi các nhà cung cấp đám mây.

Khác với Apache Airflow, tập trung vào các tác vụ tính toán với điều phối nghiêm ngặt và yêu cầu kiến thức về Python, n8n cung cấp một trình chỉnh sửa trực quan cho phép tạo kịch bản mà không cần kiến thức về ngôn ngữ lập trình (Hình. 7.46). Mặc dù giao diện của nó cho phép tạo ra các quy trình tự động mà không cần viết mã (No-Code), trong các kịch bản phức tạp hơn, người dùng có thể thêm các chức năng JavaScript và Python của riêng họ để mở rộng khả năng (Low-Code).-

> Low-Code là một phương pháp phát triển phần mềm, trong đó logic chính của ứng dụng hoặc quy trình được tạo ra bằng cách sử dụng giao diện đồ họa và các yếu tố trực quan, trong khi mã lập trình chỉ được áp dụng để cấu hình hoặc mở rộng chức năng. Các nền tảng Low-Code cho phép tăng tốc đáng kể quá trình phát triển giải pháp, thu hút không chỉ lập trình viên mà còn cả người dùng doanh nghiệp có kỹ năng kỹ thuật cơ bản. Ví dụ: người dùng có thể xây dựng quy trình kinh doanh từ các khối có sẵn, và khi cần, thêm mã của riêng họ bằng JavaScript hoặc Python.

Mặc dù n8n được định vị là nền tảng có ngưỡng đầu vào thấp, nhưng để tạo ra các kịch bản tự động hóa phức tạp, kiến thức cơ bản về lập trình, hiểu biết về công nghệ web và kỹ năng làm việc với API là rất hữu ích. Tính linh hoạt của hệ thống cho phép điều chỉnh nó cho nhiều loại nhiệm vụ - từ xử lý dữ liệu tự động đến tích hợp với các ứng dụng nhắn tin, thiết bị IoT và dịch vụ đám mây.

Các đặc điểm và lợi ích chính khi sử dụng n8n:

- **Mã nguồn mở và khả năng triển khai cục bộ đảm bảo quyền kiểm soát hoàn toàn dữ liệu, tuân thủ các yêu cầu về an ninh và độc lập khỏi các nhà cung cấp đám mây.**
- **Tích hợp với hơn 330 dịch vụ, bao gồm CRM, ERP, thương mại điện tử, nền tảng đám mây, ứng dụng nhắn tin và cơ sở dữ liệu.**
- **Tính linh hoạt của kịch bản: từ các thông báo đơn giản đến các chuỗi phức tạp với xử lý yêu cầu API, logic ra quyết định và kết nối với các dịch vụ AI.**
- **Hỗ trợ JavaScript và Python: nếu cần, người dùng có thể nhúng mã tùy chỉnh, mở rộng khả năng tự động hóa.**
- **Giao diện trực quan trực quan: cho phép nhanh chóng cấu hình và trực quan hóa tất cả các giai đoạn của quy trình.**

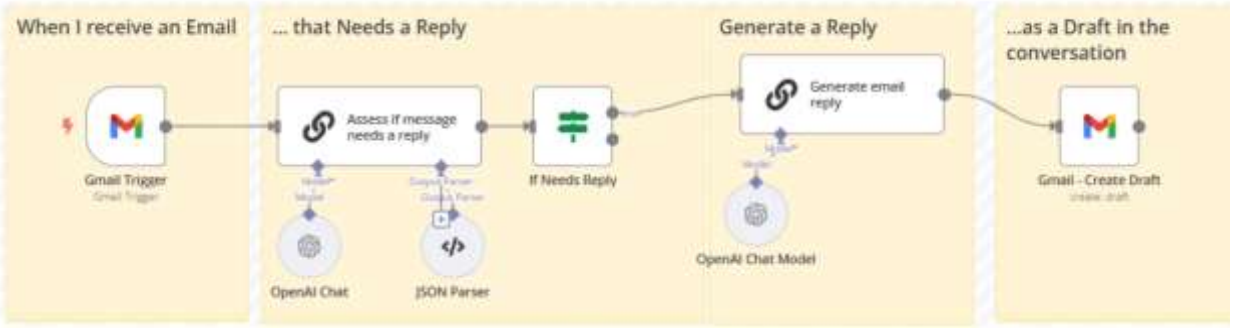
Các nền tảng Low-Code cung cấp công cụ để tạo ra các giải pháp kỹ thuật số với lượng mã tối thiểu, điều này làm cho chúng trở nên lý tưởng cho các nhóm không có chuyên môn kỹ thuật sâu nhưng cần tự động hóa quy trình.

Trong xây dựng, n8n có thể được sử dụng để tự động hóa nhiều quy trình khác nhau, chẳng hạn như tích hợp với các hệ thống quản lý dự án, kiểm tra luồng, viết báo cáo và thư sẵn sàng, tự động cập nhật dữ liệu về tồn kho vật liệu, gửi thông báo cho các nhóm về trạng thái nhiệm vụ và nhiều hơn nữa. Pipeline được cấu hình trong n8n cho phép giảm đáng kể số lượng thao tác thủ công, giảm khả năng xảy ra lỗi và tăng tốc độ ra quyết định cho việc thực hiện các dự án.

> Bạn có thể chọn một trong gần hai nghìn Pipeline miễn phí và mã nguồn mở có sẵn trên trang web: n8n.io/workflows, để tự động hóa cả quy trình làm việc trong xây dựng và các nhiệm vụ cá nhân, giảm thiểu các hoạt động lặp đi lặp lại.

Hãy lấy một trong những mẫu Pipeline có sẵn miễn phí trên trang web n8n.io, tự động tạo bản nháp phản hồi trong Gmail, giúp người dùng nhận được khối lượng lớn email hoặc gặp khó khăn trong việc soạn thảo phản hồi.-

Mẫu n8n này "Auto-Responder Gmail AI: Tạo bản nháp phản hồi cho các email đến" phân tích các tin nhắn đến bằng cách sử dụng LLM từ ChatGPT, xác định nhu cầu phản hồi, tạo bản nháp với ChatGPT và chuyển đổi văn bản thành HTML và thêm nó vào chuỗi tin nhắn trong Gmail. Trong quá trình này, email không được gửi tự động, cho phép chỉnh sửa và phê duyệt phản hồi một cách thủ công. Việc cấu hình mất khoảng 10 phút và bao gồm cấu hình OAuth cho Gmail API và tích hợp OpenAI API. Cuối cùng, chúng ta có một giải pháp tiện lợi và miễn phí để tự động hóa giao tiếp email hàng ngày mà không mất kiểm soát nội dung của các email.



Hình 7.46 Quy trình tự động hóa tạo phản hồi cho email bằng n8n.

Một ví dụ khác về tự động hóa với n8n là tìm kiếm các giao dịch có lợi trên thị trường bất động sản. Pipeline n8n "Tự động hóa giao dịch bất động sản hàng ngày với Zillow API, Google Sheets và Gmail", hàng ngày thu thập các đề xuất hiện tại phù hợp với các tiêu chí đã đặt, sử dụng Zillow API. Nó tự động tính toán các chỉ số đầu tư chính (Cash on Cash ROI, Monthly Cash Flow, Down Payment), cập nhật Google Sheets và gửi báo cáo tổng hợp qua email, cho phép các nhà đầu tư tiết kiệm thời gian và nhanh chóng phản ứng với các đề xuất tốt nhất.-



Hình 7.47 Quy trình tự động hóa đánh giá tính hấp dẫn đầu tư bất động sản.

Nhờ vào tính linh hoạt và khả năng mở rộng, n8n trở thành công cụ quý giá cho các công ty đang tìm kiếm chuyển đổi số và nâng cao khả năng cạnh tranh trên thị trường thông qua các công cụ tương đối đơn giản và miễn phí với mã nguồn mở.

Các công cụ như Apache NiFi, Airflow và n8n có thể được xem như ba cấp độ xử lý dữ liệu (Hình 7.48). NiFi quản lý luồng dữ liệu, đảm bảo việc giao nhận và chuyển đổi dữ liệu, Airflow điều phối việc thực hiện các tác vụ, kết hợp dữ liệu trong các quy trình xử lý, trong khi n8n tự động hóa việc tích hợp với các dịch vụ bên ngoài và quản lý logic kinh doanh. -





	The main task	Approach
Apache NiFi	Streaming and data transformation	Real-time stream processing
Apache Airflow	Task orchestration, ETL pipelines	Batch planning, DAG processes
n8n	Integration, automation of business logic	Low-code visual orchestration

Hình 7.48 Apache Airflow, Apache NiFi và n8n có thể được xem như ba cấp độ bổ sung cho nhau trong kiến trúc quản lý dữ liệu hiện đại.

Cùng nhau, những công cụ miễn phí và mã nguồn mở này có khả năng tạo thành một hệ sinh thái hiệu quả cho việc quản lý dữ liệu và quy trình trong ngành xây dựng, cho phép các công ty sử dụng thông tin một cách hiệu quả để ra quyết định và tự động hóa quy trình.

Các bước tiếp theo: chuyển từ các thao tác thủ công sang giải pháp dựa trên phân tích

Các công ty xây dựng hiện đại hoạt động trong điều kiện không chắc chắn cao: sự thay đổi giá nguyên liệu, sự chậm trễ trong giao hàng, thiếu hụt lao động và thời hạn dự án nghiêm ngặt. Việc sử dụng bảng điều khiển phân tích, quy trình ETL và hệ thống BI giúp các công ty nhanh chóng xác định các vấn đề, đánh giá hiệu quả tài nguyên và dự đoán các thay đổi trước khi chúng dẫn đến tổn thất tài chính.

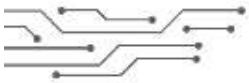
Tóm lại phần này, cần nêu rõ các bước thực tiễn chính giúp áp dụng các công nghệ đã đề cập vào các nhiệm vụ hàng ngày của bạn:

- Triển khai trực quan hóa dữ liệu và bảng điều khiển phân tích
  - Làm quen với quy trình tạo bảng điều khiển thông tin để theo dõi các chỉ số hiệu suất chính (KPI)
  - Sử dụng các công cụ trực quan hóa cho dữ liệu của bạn (Power BI, Tableau, Matplotlib, Plotly)
- Tự động hóa xử lý dữ liệu thông qua quy trình ETL
  - Thiết lập thu thập dữ liệu tự động từ các nguồn khác nhau (tài liệu, bảng tính, CAD) thông qua quy trình ETL
  - Tổ chức chuyển đổi dữ liệu (ví dụ kiểm tra qua biểu thức chính quy hoặc tính toán) bằng cách sử dụng các kịch bản Python
  - Thử thiết lập tự động tạo báo cáo ở định dạng PDF (hoặc DOC) bằng thư viện FPDF, sử dụng dữ liệu từ các tệp Excel hoặc trích xuất thông tin từ các tài liệu PDF khác

- Sử dụng mô hình ngôn ngữ (LLM) để tự động hóa
  - [ ] Sử dụng các mô hình ngôn ngữ lớn (LLM) để tạo mã giúp trích xuất và phân tích dữ liệu từ các tài liệu không cấu trúc
  - [ ] Làm quen với công cụ tự động hóa n8n và tìm hiểu các mẫu và trường hợp sẵn có trên trang web của họ. Xác định các quy trình trong công việc của bạn có thể được tự động hóa hoàn toàn bằng cách tiếp cận No-Code/Low-Code

Cách tiếp cận phân tích dữ liệu và tự động hóa quy trình không chỉ giảm thời gian cho các hoạt động lặp đi lặp lại mà còn nâng cao chất lượng quyết định. Các công ty áp dụng các công cụ phân tích trực quan và quy trình ETL có khả năng phản ứng nhanh chóng với những thay đổi.

Tự động hóa quy trình kinh doanh bằng cách sử dụng các công cụ như n8n, Airflow và NiFi chỉ là bước đầu tiên hướng tới sự trưởng thành số. Bước tiếp theo là lưu trữ và quản lý dữ liệu chất lượng, vốn là nền tảng của tự động hóa. Ở phần thứ tám, chúng ta sẽ xem xét chi tiết cách các công ty xây dựng có thể xây dựng kiến trúc lưu trữ dữ liệu bền vững, chuyển từ sự hỗn loạn của tài liệu và các tệp định dạng khác nhau sang các kho lưu trữ tập trung và các nền tảng phân tích.





VIII PHẦN
LƯU TRỮ VÀ QUẢN LÝ DỮ LIỆU TRONG XÂY DỰNG

Phần thứ tám khám phá các công nghệ hiện đại trong việc lưu trữ và quản lý dữ liệu trong ngành xây dựng. Tại đây, chúng ta phân tích các định dạng hiệu quả để làm việc với khối lượng thông tin lớn - từ CSV và XLSX đơn giản đến Apache Parquet và ORC hiệu suất cao hơn, với sự so sánh chi tiết về khả năng và hạn chế của chúng. Các khái niệm về kho dữ liệu (DWH), hồ dữ liệu (Data Lakes) và các giải pháp lai của chúng (Data Lakehouse) cũng như các nguyên tắc quản lý dữ liệu (Data Governance) và chủ nghĩa tối giản trong việc xử lý thông tin (Data Minimalism) được xem xét. Các vấn đề về "đầm lầy dữ liệu" (Data Swamp) và các chiến lược để ngăn chặn sự hỗn loạn trong các hệ thống thông tin cũng được làm rõ. Các phương pháp mới trong việc làm việc với dữ liệu, bao gồm cơ sở dữ liệu vector và ứng dụng của chúng trong xây dựng thông qua khái niệm Bounding Box, cũng được đề cập. Phần này cũng đề cập đến các phương pháp DataOps và VectorOps như các tiêu chuẩn mới trong việc tổ chức quy trình làm việc với dữ liệu.

CHƯƠNG 8.1. HẠ TẦNG DỮ LIỆU: TỪ ĐỊNH DẠNG LƯU TRỮ ĐẾN KHO LƯU TRỮ SỐ

Nguyên tử dữ liệu: nền tảng của quản lý thông tin hiệu quả

Tất cả trong vũ trụ đều được cấu thành từ những khối xây dựng nhỏ nhất - nguyên tử và phân tử, và theo thời gian, mọi thứ sống và không sống đều trở về trạng thái ban đầu này. Trong tự nhiên, quá trình này diễn ra với tốc độ đáng kinh ngạc mà chúng ta cố gắng áp dụng vào các quy trình do con người điều khiển.

> Trong rừng, bất kỳ sinh vật sống nào theo thời gian cũng sẽ biến thành chất dinh dưỡng, làm nền tảng cho các cây mới. Những cây này, theo đó, trở thành thức ăn cho những sinh vật sống mới, được cấu thành từ những nguyên tử mà hàng triệu năm trước đã tạo ra vũ trụ.

Trong thế giới kinh doanh, cũng quan trọng việc phân tách các cấu trúc phức tạp thành những đơn vị cơ bản nhất, có thể xử lý tối thiểu - giống như nguyên tử và phân tử trong tự nhiên. Điều này cho phép lưu trữ và quản lý hiệu quả các nguyên tử dữ liệu, biến chúng thành một nền tảng phong phú, màu mỡ, trở thành nguồn tài nguyên chính cho sự phát triển của phân tích và chất lượng ra quyết định.



Hình 8.11 Phân tích và ra quyết định dựa trên dữ liệu có thể tái sử dụng, vốn đã được xử lý và lưu trữ trước đó.

Các tác phẩm âm nhạc được tạo thành từ các nốt nhạc, khi kết hợp lại, tạo ra những tác phẩm âm nhạc phức tạp, trong khi từ ngữ được tạo ra từ đơn vị nguyên thủy - chữ cái âm thanh. Dù là thiên nhiên, khoa học, kinh tế, nghệ thuật hay công nghệ, thế giới thể hiện sự thống nhất và hài hòa đáng kinh ngạc trong nỗ lực

của nó đối với sự phá hủy, cấu trúc, chu kỳ và sáng tạo. Tương tự, các quy trình trong hệ thống tính toán chi phí được chia nhỏ thành các đơn vị cấu trúc nhỏ nhất - các mục tài nguyên - ở cấp độ tính toán và lịch trình. Sau đó, những đơn vị này, giống như các nốt nhạc, được sử dụng để hình thành các phép tính và biểu đồ phức tạp hơn. Theo cùng một nguyên tắc, các hệ thống thiết kế tự động hoạt động, trong đó các dự án kiến trúc và kỹ thuật phức tạp được xây dựng từ các yếu tố cơ bản - các thành phần riêng lẻ và các thành phần thư viện, từ đó tạo ra mô hình 3D hoàn chỉnh của dự án một tòa nhà hoặc công trình phức tạp.

> Khái niệm chu kỳ và cấu trúc, vốn có trong thiên nhiên và khoa học, cũng phản ánh trong thế giới dữ liệu hiện đại. Như trong thiên nhiên, tất cả các sinh vật sống đều trở về với các nguyên tử và phân tử, trong thế giới các công cụ xử lý dữ liệu hiện đại, thông tin cũng có xu hướng trở về hình thức nguyên thủy nhất.

Các yếu tố nhỏ nhất với tính không thể chia tách của chúng là các khối xây dựng cơ bản của các quy trình kinh doanh. Việc lên kế hoạch cẩn thận từ đầu về cách thu thập, cấu trúc (chia nhỏ thành nguyên tử) và lưu trữ những khối xây dựng nhỏ nhất này từ các nguồn khác nhau là rất quan trọng. Trong đó, tổ chức và lưu trữ dữ liệu không chỉ là vấn đề chia nhỏ chúng thành các phần cấu thành. Cũng quan trọng không kém là đảm bảo tích hợp và lưu trữ có cấu trúc, để dữ liệu có thể được dễ dàng truy xuất, phân tích và sử dụng cho việc ra quyết định bất cứ khi nào cần thiết.

Để xử lý thông tin hiệu quả, cần lựa chọn cẩn thận định dạng và phương pháp lưu trữ dữ liệu - giống như đất phải được chuẩn bị cho sự phát triển của cây cối. Các kho dữ liệu cần được tổ chức sao cho đảm bảo chất lượng và tính cập nhật cao của thông tin, loại bỏ dữ liệu dư thừa hoặc không liên quan. Càng tốt hơn khi "đất thông tin" này được cấu trúc, người dùng càng nhanh chóng và chính xác hơn trong việc tìm kiếm dữ liệu cần thiết và giải quyết các nhiệm vụ phân tích.

Kho lưu trữ thông tin: tệp hoặc dữ liệu

Các kho dữ liệu cho phép các công ty thu thập và kết hợp thông tin từ nhiều hệ thống khác nhau, tạo ra một trung tâm duy nhất cho phân tích sau này. Dữ liệu lịch sử được thu thập không chỉ cho phép phân tích sâu hơn các quy trình mà còn phát hiện các quy luật có thể ảnh hưởng đến hiệu quả kinh doanh.

Giả sử một công ty đang quản lý nhiều dự án cùng lúc. Kỹ sư muốn hiểu xem đã đổ bao nhiêu bê tông và còn bao nhiêu khối lượng cần mua thêm. Với cách tiếp cận truyền thống, anh ta sẽ phải tìm kiếm thủ công trên máy chủ và mở nhiều bảng dự toán, so sánh chúng với các biên bản nghiệm thu và kiểm tra tồn kho hiện tại. Điều này mất hàng giờ, thậm chí hàng ngày. Ngay cả khi có các quy trình ETL và các kịch bản tự động, nhiệm vụ vẫn còn bán thủ công: kỹ sư vẫn phải chỉ định thủ công đường dẫn đến các thư mục hoặc tệp cụ thể trên máy chủ. Điều này làm giảm hiệu quả tổng thể của việc tự động hóa, vì nó vẫn chiếm thời gian làm việc quý giá.

Khi chuyển sang quản lý dữ liệu thay vì làm việc với hệ thống tệp của máy chủ, kỹ sư có quyền truy cập vào một cấu trúc lưu trữ thống nhất, nơi thông tin được cập nhật theo thời gian thực. Một truy vấn — dưới dạng mã, truy vấn SQL hoặc thậm chí là yêu cầu đến đại lý LLM — cho phép ngay lập tức nhận được dữ liệu chính xác về số dư hiện tại, khối lượng công việc đã hoàn thành và các lô hàng sắp tới, nếu dữ liệu đã được chuẩn bị lại và kết hợp thành một kho dữ liệu, nơi không cần phải lục lọi qua các thư mục, mở hàng chục tệp và đối chiếu giá trị một cách thủ công.

Trong một thời gian dài, các công ty xây dựng đã sử dụng tài liệu PDF, bản vẽ DWG, mô hình RVT và hàng

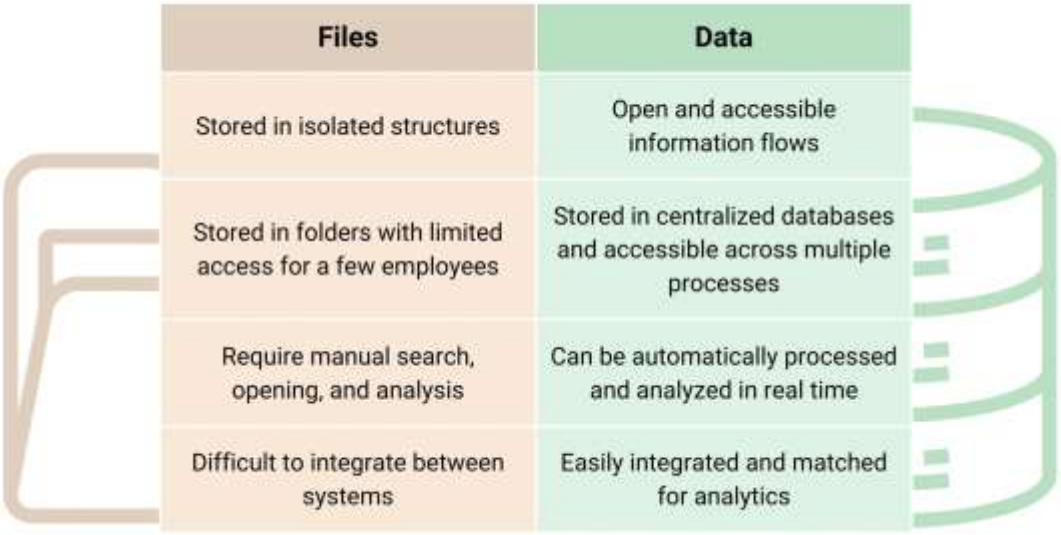
trăm, hàng nghìn bảng Excel cùng các định dạng rời rạc khác, được lưu trữ trong các thư mục nhất định trên máy chủ của công ty, điều này làm phức tạp việc tìm kiếm thông tin, kiểm tra và phân tích. Cuối cùng, các tệp còn lại sau khi hoàn thành dự án thường được chuyển trở lại máy chủ vào các thư mục lưu trữ lưu trữ, mà hầu như không được sử dụng sau đó. Phương pháp lưu trữ dữ liệu truyền thống như vậy, với sự gia tăng lưu lượng dữ liệu, ngày càng trở nên lỗi thời, do tính dễ bị tổn thương với các lỗi do yếu tố con người gây ra.

> Tệp chỉ là một container cô lập, nơi lưu trữ dữ liệu. Các tệp được tạo ra cho con người, không phải cho hệ thống, vì vậy chúng yêu cầu phải mở, đọc và diễn giải bằng tay. Ví dụ có thể là bảng Excel, tài liệu PDF hoặc bản vẽ CAD, mà cần phải mở bằng một công cụ cụ thể để truy cập thông tin cần thiết. Nếu không có việc trích xuất và xử lý có cấu trúc, thông tin trong đó sẽ vẫn không được sử dụng.
>
> Dữ liệu, ngược lại, là thông tin có thể đọc được bởi máy móc, có liên kết, được cập nhật và phân tích tự động. Trong một kho dữ liệu thống nhất (ví dụ, cơ sở dữ liệu, DWH hoặc Data Lake), thông tin được trình bày dưới dạng bảng, bản ghi và mối quan hệ. Điều này đảm bảo khả năng lưu trữ đồng nhất, thực hiện các truy vấn tự động, phân tích giá trị và xây dựng báo cáo theo thời gian thực.

Việc sử dụng dữ liệu thay vì tệp (Hình. 8.11) cho phép từ bỏ quy trình tìm kiếm thủ công và thống nhất các quy trình xử lý. Các công ty đã triển khai phương pháp này đang có lợi thế cạnh tranh nhờ tốc độ truy cập thông tin và khả năng tích hợp nhanh chóng vào các quy trình kinh doanh.

> Sự chuyển đổi từ việc sử dụng tệp sang dữ liệu là một thay đổi không thể tránh khỏi, sẽ xác định tương lai của ngành xây dựng.

Mỗi công ty trong ngành xây dựng sẽ phải đối mặt với một lựa chọn quan trọng: tiếp tục lưu trữ thông tin trong các tệp rời rạc và silo, mà phải được đọc bởi con người thông qua các chương trình đặc biệt hoặc chuyển đổi nó trong các giai đoạn đầu của xử lý thành dữ liệu có cấu trúc, tạo ra một nền tảng kỹ thuật số tích hợp duy nhất cho việc quản lý dự án tự động.



Hình. 8.11 Sự tiến hóa của dòng thông tin: từ các tệp cô lập đến dữ liệu tích hợp.

Trong bối cảnh sự gia tăng nhanh chóng của khối lượng thông tin, các phương pháp lưu trữ và xử lý tệp truyền thống ngày càng trở nên kém hiệu quả. Trong ngành xây dựng, cũng như trong các lĩnh vực khác, không còn đủ để dựa vào các thư mục rời rạc với các tệp ở các định dạng khác nhau hoặc các cơ sở dữ liệu không liên kết với nhau.

Các công ty muốn duy trì tính cạnh tranh trong kỷ nguyên công nghệ số sẽ không thể tránh khỏi việc chuyển sang các nền tảng kỹ thuật số tích hợp, sử dụng công nghệ dữ liệu lớn và các hệ thống phân tích tự động.

Việc chuyển từ lưu trữ tệp sang làm việc với dữ liệu sẽ yêu cầu suy nghĩ lại về các phương pháp quản lý thông tin và lựa chọn có ý thức các định dạng phù hợp cho việc tích hợp vào các kho lưu trữ tập trung. Sự lựa chọn này sẽ ảnh hưởng đến mức độ hiệu quả trong việc xử lý dữ liệu, tốc độ truy cập và độ đơn giản trong việc tích hợp vào các quy trình số của công ty.

Lưu trữ dữ liệu lớn: phân tích các định dạng phổ biến và hiệu quả của chúng

Các định dạng lưu trữ đóng vai trò then chốt trong việc đảm bảo khả năng mở rộng, độ tin cậy và hiệu suất của hạ tầng phân tích. Để phân tích và xử lý dữ liệu — chẳng hạn như lọc, nhóm và tổng hợp — trong các ví dụ của chúng tôi, chúng tôi đã sử dụng Pandas DataFrame — cấu trúc phổ biến để làm việc với dữ liệu trong bộ nhớ.

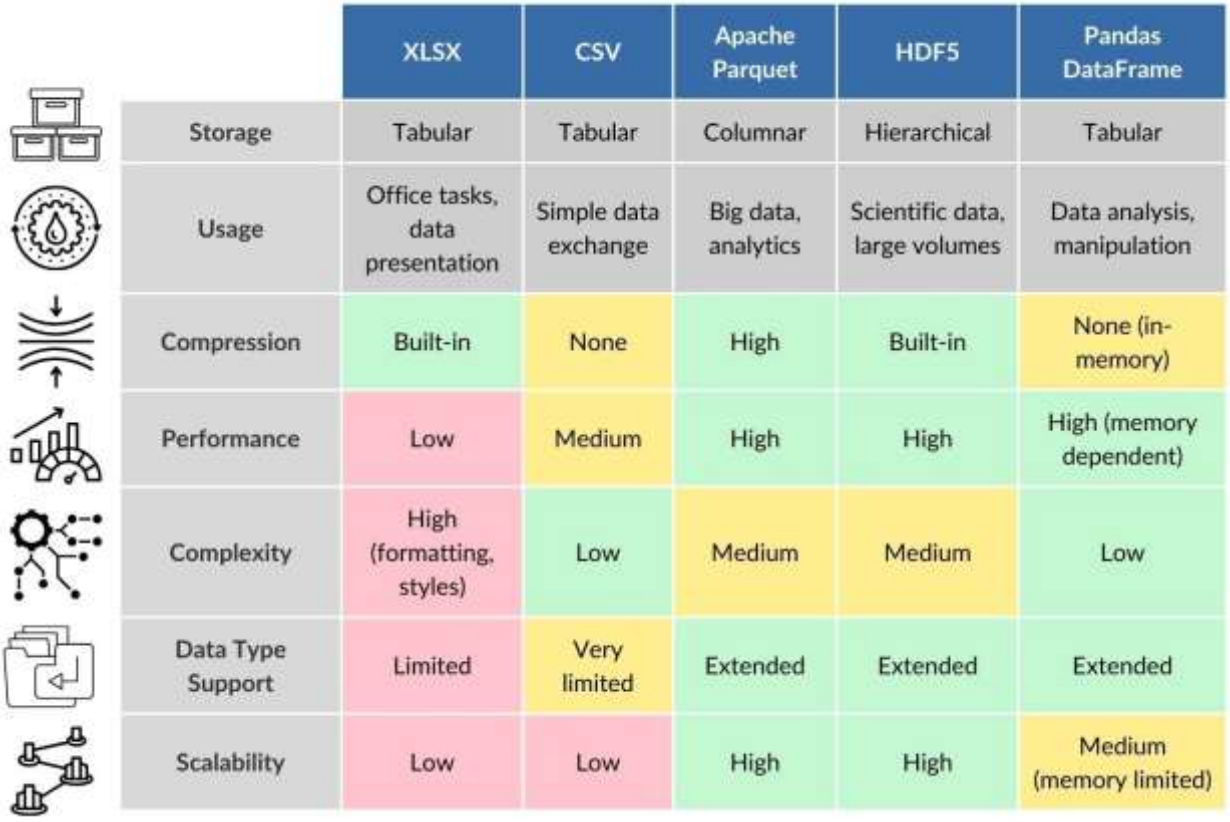
Tuy nhiên, Pandas DataFrame không có định dạng lưu trữ riêng, vì vậy sau khi hoàn tất xử lý, dữ liệu sẽ được xuất sang một trong các định dạng bên ngoài — thường là CSV hoặc XLSX. Những định dạng bảng này rất tiện lợi cho việc trao đổi và tương thích với hầu hết các hệ thống bên ngoài, nhưng có một số hạn chế: hiệu quả lưu trữ thấp, không có nén và hỗ trợ phiên bản kém.

- CSV (Comma-Separated Values): định dạng văn bản đơn giản, được hỗ trợ rộng rãi bởi nhiều nền tảng và công cụ khác nhau. Nó dễ sử dụng, nhưng không hỗ trợ các loại dữ liệu phức tạp và nén.
- XLSX (Excel Open XML Spreadsheet): định dạng tệp của Microsoft Excel, hỗ trợ các chức năng phức tạp như công thức, biểu đồ và định dạng. Mặc dù nó tiện lợi cho việc phân tích và trực quan

hóa dữ liệu thủ công, nhưng không được tối ưu hóa cho việc xử lý dữ liệu quy mô lớn.

Ngoài các định dạng bảng phổ biến XLSX và CSV, còn có một số định dạng phổ biến khác cho việc lưu trữ dữ liệu có cấu trúc hiệu quả (Hình. 8.12), mỗi định dạng đều có những lợi thế độc đáo tùy thuộc vào yêu cầu cụ thể về lưu trữ và phân tích dữ liệu:-

- Apache Parquet: định dạng tệp cho lưu trữ dữ liệu theo cột, được tối ưu hóa cho việc sử dụng trong các hệ thống phân tích dữ liệu. Nó cung cấp các sơ đồ nén và mã hóa dữ liệu hiệu quả, làm cho nó trở thành lựa chọn lý tưởng cho các cấu trúc dữ liệu phức tạp và xử lý dữ liệu lớn.
- Apache ORC (Optimized Row Columnar - cột hàng tối ưu hóa): tương tự như Parquet, ORC cung cấp mức độ nén cao và lưu trữ dữ liệu hiệu quả. Nó được tối ưu hóa cho các thao tác đọc nặng và rất phù hợp cho việc lưu trữ hồ dữ liệu.
- JSON (JavaScript Object Notation): mặc dù JSON không hiệu quả trong việc lưu trữ dữ liệu so với các định dạng nhị phân như Parquet hoặc ORC, nhưng nó rất dễ tiếp cận và dễ làm việc, làm cho nó trở thành lựa chọn lý tưởng cho các kịch bản mà tính dễ đọc và tương thích với công nghệ web là quan trọng.
- Feather: định dạng lưu trữ dữ liệu nhị phân theo cột nhanh, nhẹ và dễ sử dụng, tập trung vào phân tích. Nó được thiết kế để truyền dữ liệu hiệu quả giữa Python (Pandas) và R, làm cho nó trở thành lựa chọn tuyệt vời cho các dự án liên quan đến các môi trường lập trình này.
- HDF5 (Hierarchical Data Format version 5): được thiết kế để lưu trữ và tổ chức khối lượng lớn dữ liệu. Nó hỗ trợ một loạt các loại dữ liệu và rất phù hợp cho việc làm việc với các bộ sưu tập dữ liệu phức tạp. HDF5 đặc biệt phổ biến trong tính toán khoa học nhờ khả năng lưu trữ và truy cập hiệu quả các tập dữ liệu lớn.



	XLSX	CSV	Apache Parquet	HDF5	Pandas DataFrame
Storage	Tabular	Tabular	Columnar	Hierarchical	Tabular
Usage	Office tasks, data presentation	Simple data exchange	Big data, analytics	Scientific data, large volumes	Data analysis, manipulation
Compression	Built-in	None	High	Built-in	None (in-memory)
Performance	Low	Medium	High	High	High (memory dependent)
Complexity	High (formatting, styles)	Low	Medium	Medium	Low
Data Type Support	Limited	Very limited	Extended	Extended	Extended
Scalability	Low	Low	High	High	Medium (memory limited)

Hình. 8.12 So sánh các định dạng dữ liệu với các khác biệt chính trong các khía cạnh lưu trữ và xử lý.

Để thực hiện phân tích so sánh các định dạng được sử dụng trong giai đoạn Load của quy trình ETL, một bảng đã được lập để trình bày kích thước tệp và thời gian đọc của chúng (Hình 8.13). Trong nghiên cứu này, các tệp có dữ liệu giống hệt nhau: bảng chứa 10.000 dòng và 10 cột, được điền bằng các giá trị ngẫu nhiên.

Nghiên cứu bao gồm các định dạng lưu trữ sau: CSV, Parquet, XLSX và HDF5, cũng như các phiên bản nén của chúng trong các tệp ZIP. Dữ liệu gốc được tạo ra bằng cách sử dụng thư viện NumPy và được trình bày dưới dạng cấu trúc Pandas DataFrame. Quy trình thử nghiệm bao gồm các bước sau:

- Lưu tệp: DataFrame đã được lưu dưới bốn định dạng khác nhau: CSV, Parquet, XLSX và HDF5. Mỗi định dạng có những đặc điểm riêng về cách lưu trữ dữ liệu, ảnh hưởng đến kích thước tệp và tốc độ đọc của nó.
- Nén tệp trong ZIP: để phân tích hiệu quả của việc nén tiêu chuẩn, mỗi tệp đã được nén thêm vào tệp ZIP.
- Đọc tệp (ETL – Load): thời gian đọc được đo cho mỗi tệp sau khi nó được giải nén từ ZIP. Điều này cho phép đánh giá tốc độ truy cập dữ liệu sau khi được lấy ra từ kho lưu trữ.

Quan trọng cần lưu ý rằng Pandas DataFrame không được sử dụng trực tiếp trong việc phân tích kích thước hoặc thời gian đọc, vì nó không phải là một định dạng lưu trữ độc lập. Nó chỉ đóng vai trò là cấu trúc trung gian để tạo ra và lưu trữ dữ liệu vào các định dạng khác nhau.



Hình 8.13 So sánh các định dạng lưu trữ dữ liệu theo kích thước và tốc độ đọc.

Tệp CSV và HDF5 cho thấy (Hình 8.13) hiệu quả nén cao, giảm đáng kể kích thước của chúng khi được nén vào ZIP, điều này có thể đặc biệt hữu ích trong các kịch bản yêu cầu tối ưu hóa lưu trữ. Tệp XLSX, ngược lại, gần như không thể nén, và kích thước của nó trong ZIP vẫn tương đương với bản gốc, điều này làm cho nó kém lợi thế hơn khi sử dụng trong khối lượng dữ liệu lớn hoặc trong các điều kiện mà tốc độ truy cập dữ liệu là quan trọng. Hơn nữa, thời gian đọc cho XLSX cao hơn đáng kể so với các định dạng khác, điều này làm cho nó kém ưu tiên cho các thao tác đọc dữ liệu nhanh. Apache Parquet nhờ vào cấu trúc cột của nó đã thể hiện hiệu quả cao cho các nhiệm vụ phân tích và khối lượng dữ liệu lớn.-

Tối ưu hóa lưu trữ dữ liệu với Apache Parquet

Một trong những định dạng phổ biến để lưu trữ và xử lý dữ liệu lớn là Apache Parquet. Định dạng này được phát triển đặc biệt cho lưu trữ cột (tương tự như Pandas), cho phép giảm đáng kể dung lượng bộ nhớ và tăng tốc độ truy vấn phân tích. Khác với các định dạng truyền thống như CSV và XLSX, Parquet hỗ trợ nén tích hợp và được tối ưu hóa để làm việc với các hệ thống xử lý dữ liệu lớn, bao gồm Spark, Hadoop và các kho lưu trữ đám mây.

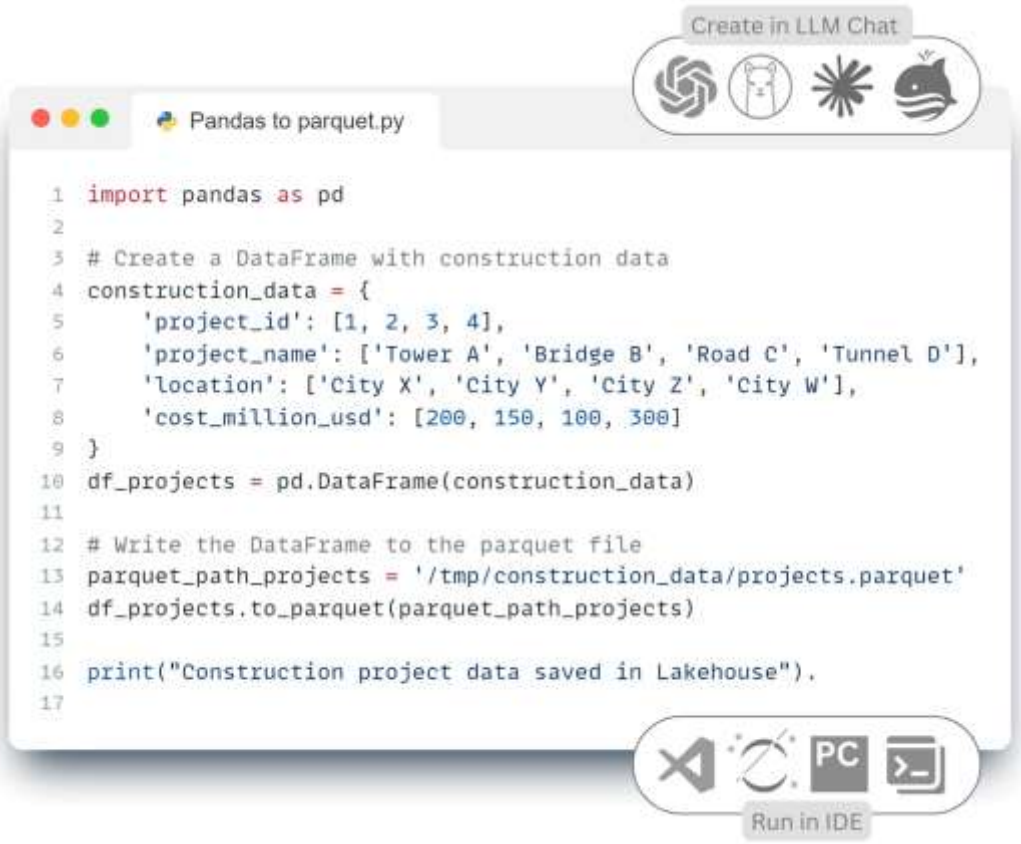
> Các đặc điểm chính của Parquet bao gồm hỗ trợ nén và mã hóa dữ liệu, điều này giảm đáng kể kích thước lưu trữ và tăng tốc độ các thao tác đọc dữ liệu bằng cách làm việc trực tiếp với các cột cần thiết, thay vì tất cả các dòng dữ liệu.

Để có một ví dụ trực quan về cách dễ dàng nhận được mã cần thiết để chuyển đổi dữ liệu sang Apache Parquet, hãy sử dụng LLM.

- Gửi yêu cầu văn bản vào trò chuyện LLM (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN):

> Viết mã để lưu dữ liệu từ Pandas DataFrame vào Apache Parquet. ⏎

- Phản hồi từ LLM:



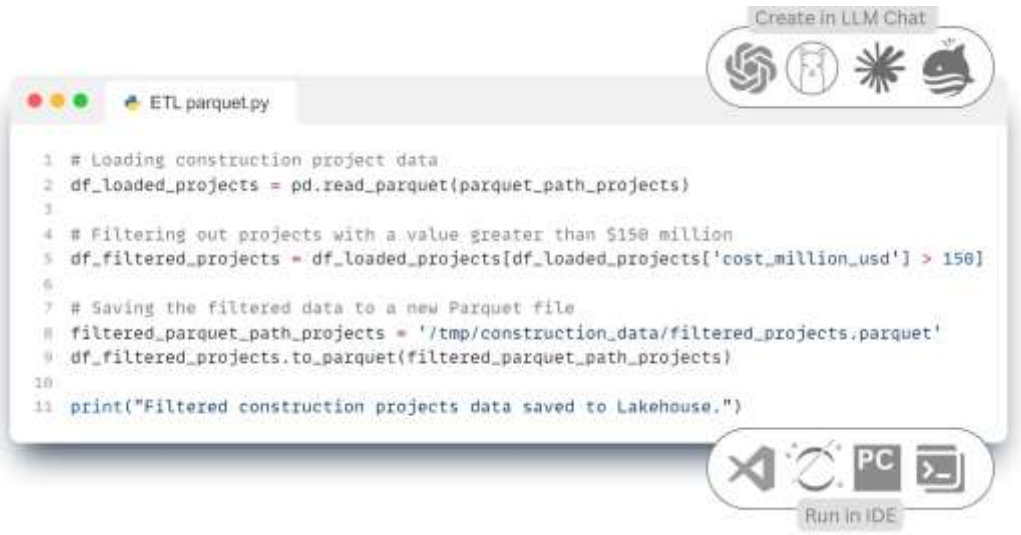
Hình 8.14 Chuyển dữ liệu DataFrame từ bộ nhớ vào định dạng lưu trữ hiệu quả Apache Parquet bằng vài dòng Python.

Ví dụ tiếp theo: mô phỏng quy trình ETL với dữ liệu được lưu trữ ở định dạng Parquet để lọc các dự án theo một giá trị nhất định của thuộc tính "cost_million_usd" (Hình 8.14).-

- Trong phần tiếp theo của cuộc trò chuyện, hãy gửi yêu cầu văn bản đến LLM:

> Viết mã, trong đó chúng ta muốn lọc dữ liệu trong bảng và chỉ lưu lại những dự án (dòng bảng) từ dữ liệu Apache Parquet, có giá trị (tham số cost_million_usd) vượt quá 150 triệu đô la. ⏎

- Phản hồi từ LLM:



Hình 8.15 Quy trình ETL khi làm việc với dữ liệu ở định dạng Apache Parquet cũng giống như với các định dạng có cấu trúc khác.

Việc sử dụng định dạng Parquet (so với XLSX, CSV và các định dạng khác) giảm đáng kể khối lượng thông tin lưu trữ và tăng tốc độ tìm kiếm. Nhờ đó, nó rất phù hợp cho cả việc lưu trữ và phân tích dữ liệu. Parquet tích hợp với nhiều hệ thống xử lý, đảm bảo truy cập hiệu quả trong các kiến trúc lai.

Tuy nhiên, định dạng lưu trữ hiệu quả chỉ là một trong những yếu tố của việc làm việc với dữ liệu một cách toàn diện. Để tạo ra một môi trường bền vững và có thể mở rộng, cần có một kiến trúc quản lý dữ liệu được thiết kế rõ ràng. Chính chức năng này được thực hiện bởi các hệ thống loại DWH (Data Warehouse). Chúng đảm bảo việc tổng hợp dữ liệu từ các nguồn khác nhau, tính minh bạch của các quy trình kinh doanh và khả năng phân tích tổng hợp bằng cách sử dụng các công cụ BI và thuật toán học máy.

DWH: Kho dữ liệu Data Warehouse

Tương tự như cách mà định dạng Parquet được tối ưu hóa để lưu trữ hiệu quả khối lượng thông tin lớn, Data Warehouse được tối ưu hóa để tích hợp và cấu trúc dữ liệu nhằm hỗ trợ phân tích, dự đoán và ra quyết định quản lý.

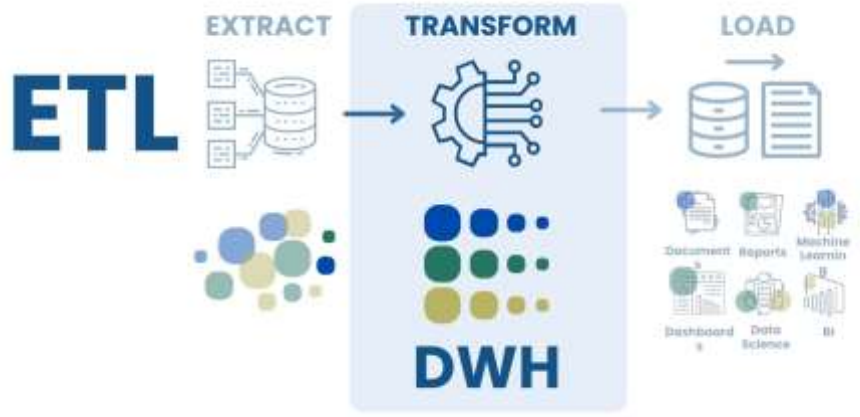
Trong các công ty hiện đại, dữ liệu đến từ nhiều nguồn khác nhau: ERP, CAFM, CPM, hệ thống CRM, kế toán và quản lý kho, mô hình CAD kỹ thuật số của các tòa nhà, cảm biến IoT và các giải pháp khác. Để có được bức tranh toàn cảnh, không chỉ đơn giản là thu thập dữ liệu - mà cần phải tổ chức, chuẩn hóa và tập trung chúng trong một kho lưu trữ duy nhất. Chính chức năng này được thực hiện bởi DWH - hệ thống lưu trữ trung tâm, cho phép tổng hợp thông tin từ các nguồn khác nhau, cấu trúc nó và làm cho nó có sẵn cho phân tích và quản lý chiến lược.

> DWH (Data Warehouse) là một hệ thống lưu trữ dữ liệu tập trung, tổng hợp thông tin từ nhiều nguồn, cấu trúc nó và làm cho nó có sẵn cho phân tích và báo cáo.

Trong nhiều công ty, dữ liệu bị phân tán qua các hệ thống khác nhau mà chúng ta đã xem xét trong các phần đầu của cuốn sách (Hình 1.24). DWH tích hợp các nguồn này, đảm bảo tính minh bạch và độ tin cậy của thông tin. Kho dữ liệu DWH là một cơ sở dữ liệu chuyên dụng (cơ sở dữ liệu lớn), thu thập, xử lý và lưu trữ dữ liệu từ nhiều nguồn. Các đặc điểm chính của DWH:-

- Sử dụng quy trình ETL (Extract, Transform, Load) - trích xuất dữ liệu từ các nguồn, làm sạch, chuyển đổi, tải vào kho lưu trữ và tự động hóa các quy trình này, như đã đề cập trong phần bảy của cuốn sách.
- Độ granularity của dữ liệu – dữ liệu trong DWH có thể được lưu trữ dưới dạng tổng hợp (báo cáo tổng hợp) hoặc chi tiết (dữ liệu thô). Từ năm 2024, các nhà cung cấp CAD bắt đầu nói về dữ liệu granularity, điều này có thể cho thấy ngành công nghiệp đang chuẩn bị chuyển sang sử dụng các kho lưu trữ đám mây chuyên dụng để làm việc với dữ liệu của các mô hình kỹ thuật số của tòa nhà.
- Hỗ trợ phân tích và dự đoán – kho dữ liệu cung cấp nền tảng cho các công cụ BI, phân tích Big Data và học máy.

DWH đóng vai trò nền tảng cho phân tích kinh doanh, cho phép phân tích các chỉ số hiệu suất chính, dự đoán doanh số, mua sắm và chi phí, cũng như tự động hóa báo cáo và trực quan hóa dữ liệu.-



Hình 8.16 Trong quy trình ETL, DWH có thể đóng vai trò là kho lưu trữ trung tâm, nơi dữ liệu được trích xuất từ các hệ thống khác nhau trải qua các giai đoạn biến đổi và xuất dữ liệu.

DWH đóng vai trò quan trọng trong việc tích hợp, làm sạch và cấu trúc thông tin, tạo ra nền tảng vững chắc cho phân tích kinh doanh và quy trình ra quyết định. Tuy nhiên, trong điều kiện hiện đại, khi khối lượng dữ liệu tăng nhanh và các nguồn dữ liệu ngày càng đa dạng, phương pháp truyền thống trong việc lưu trữ thông tin DWH thường cần được mở rộng bằng các phương pháp ELT và Data Lake.

Data Lake - sự tiến hóa của ETL trong ELT: từ làm sạch truyền thống đến xử lý linh hoạt

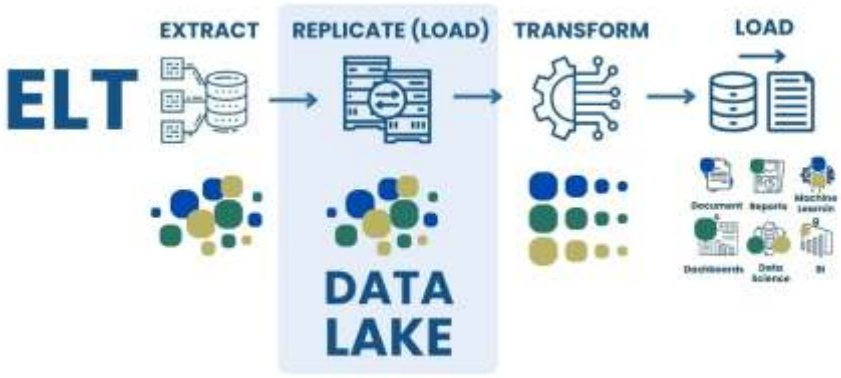
Các kho dữ liệu DWH cổ điển, được thiết kế để lưu trữ dữ liệu có cấu trúc theo định dạng tối ưu cho các truy vấn phân tích, đã gặp phải những hạn chế trong việc xử lý dữ liệu không có cấu trúc và khả năng mở rộng.

Để đáp ứng những vấn đề này, các hồ dữ liệu (Data Lakes) đã xuất hiện, cung cấp khả năng lưu trữ linh hoạt cho khối lượng lớn dữ liệu đa dạng.

> Data Lake cung cấp một phương pháp thay thế cho DWH, cho phép làm việc với dữ liệu không có cấu trúc, bán cấu trúc và dữ liệu thô mà không cần sơ đồ cứng trước. Phương pháp lưu trữ này thường phù hợp cho việc xử lý dữ liệu theo thời gian thực, học máy và phân tích nâng cao. Khác với DWH, nơi dữ liệu được cấu trúc và tổng hợp trước khi tải lên, Data Lake cho phép lưu trữ thông tin ở dạng gốc, từ đó đảm bảo tính linh hoạt và khả năng mở rộng.

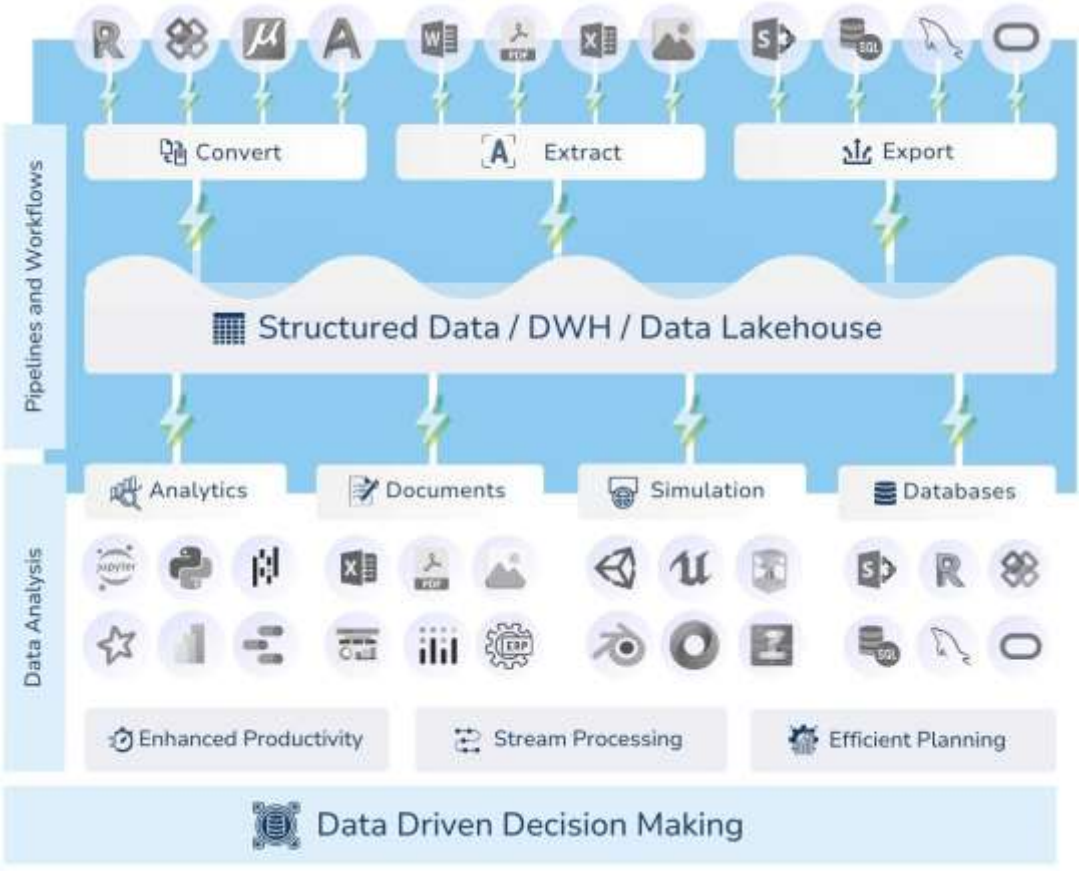
Chính sự thất vọng với các kho dữ liệu truyền thống (RDBMS, DWH) và sự quan tâm đến "dữ liệu lớn" đã dẫn đến sự xuất hiện của các hồ dữ liệu, nơi mà thay vì quy trình ETL phức tạp, dữ liệu chỉ cần được tải vào kho lưu trữ có cấu trúc lỏng lẻo, và việc xử lý diễn ra ở giai đoạn phân tích.

- Trong các kho dữ liệu truyền thống, dữ liệu thường trải qua quá trình xử lý trước, biến đổi và làm sạch (ETL - Extract, Transform, Load) trước khi được tải vào kho lưu trữ. Điều này có nghĩa là dữ liệu được cấu trúc và tối ưu hóa cho các nhiệm vụ phân tích và báo cáo cụ thể trong tương lai. Sự chú trọng chính là duy trì hiệu suất truy vấn cao và tính toàn vẹn của dữ liệu. Tuy nhiên, phương pháp này có thể tốn kém và kém linh hoạt trong việc tích hợp các loại dữ liệu mới và các sơ đồ dữ liệu đang thay đổi nhanh chóng.-
- Hồ dữ liệu, mặt khác, được thiết kế để lưu trữ khối lượng lớn dữ liệu thô ở định dạng gốc của chúng (Hình 8.17). Thay vì quy trình ETL (Extract, Transform, Load), ELT (Extract, Load, Transform) xuất hiện, khi dữ liệu được tải vào kho "như nó là" và chỉ sau đó có thể được chuyển đổi và phân tích khi cần thiết. Điều này cung cấp tính linh hoạt cao hơn và khả năng lưu trữ dữ liệu đa dạng, bao gồm cả dữ liệu phi cấu trúc như văn bản, hình ảnh và nhật ký.



Hình 8.17 Khác với ETL, trong Hồ Dữ Liệu sử dụng ELT, trong đó thông tin được tải lên ở dạng "thô", và việc chuyển đổi được thực hiện ở giai đoạn xuất.

Các kho dữ liệu truyền thống tập trung vào việc xử lý dữ liệu trước để đảm bảo hiệu suất truy vấn cao, trong khi ở hồ dữ liệu, ưu tiên được đặt vào tính linh hoạt: chúng lưu trữ dữ liệu thô và chuyển đổi chúng khi cần thiết (Hình 8.18).-



Hình 8.18 Các khái niệm hiện đại về kho lưu trữ nhằm mục đích lưu trữ và xử lý tất cả các loại dữ liệu cho mục đích ra quyết định.

Tuy nhiên, bất chấp tất cả những lợi ích, hồ dữ liệu cũng không thiếu nhược điểm. Việc thiếu cấu trúc nghiêm ngặt và độ phức tạp trong quản lý thông tin có thể dẫn đến hỗn loạn, trong đó dữ liệu bị trùng lặp, mâu thuẫn với nhau hoặc mất tính cập nhật. Hơn nữa, việc tìm kiếm và phân tích dữ liệu trong một kho như vậy đòi hỏi nỗ lực đáng kể, đặc biệt là khi làm việc với thông tin đa dạng. Để vượt qua những hạn chế này và kết hợp những điểm mạnh của kho dữ liệu truyền thống và hồ dữ liệu, kiến trúc Data Lakehouse đã được phát triển.

Kiến trúc Data Lakehouse: sự hợp lực giữa kho lưu trữ và hồ dữ liệu

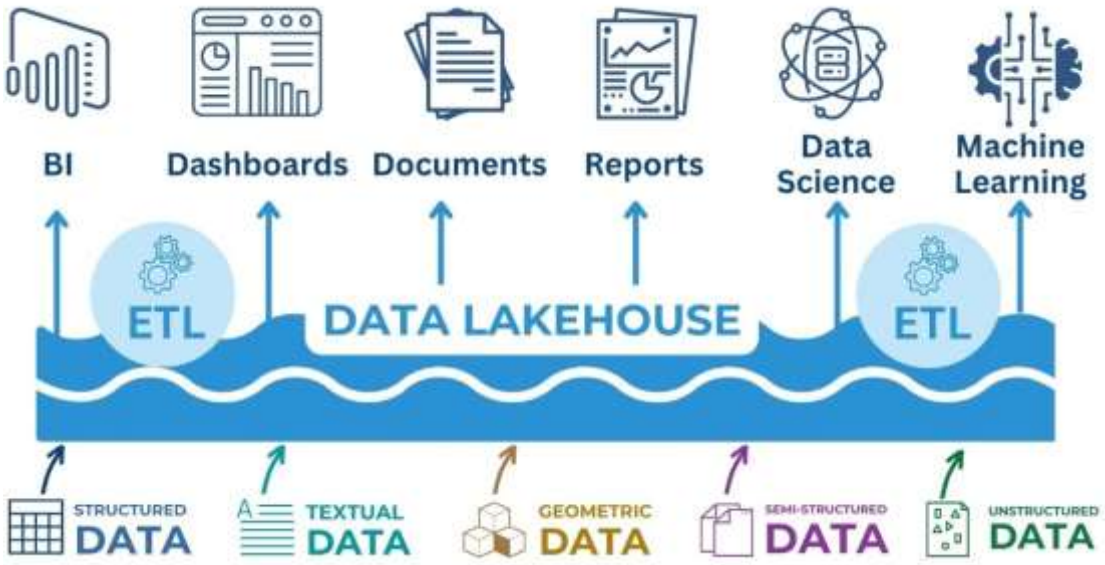
Để kết hợp những đặc điểm tốt nhất của DWH (cấu trúc, khả năng quản lý, hiệu suất phân tích cao) và Hồ Dữ Liệu (khả năng mở rộng, làm việc với dữ liệu đa dạng), một phương pháp Data Lakehouse đã được phát triển. Kiến trúc này kết hợp tính linh hoạt của hồ dữ liệu với các công cụ xử lý và quản lý mạnh mẽ, đặc trưng cho các kho dữ liệu truyền thống, đảm bảo sự cân bằng giữa lưu trữ, phân tích và học máy. Data Lakehouse là sự tổng hợp giữa hồ dữ liệu và kho dữ liệu, kết hợp tính linh hoạt và khả năng mở rộng của hồ dữ liệu với khả năng quản lý và tối ưu hóa truy vấn của kho dữ liệu.

Data Lakehouse là một phương pháp kiến trúc nhằm kết hợp tính linh hoạt và khả năng mở rộng của hồ dữ liệu với khả năng quản lý và hiệu suất truy vấn trong các kho dữ liệu (Hình 8.19).-

Các đặc điểm chính của Data Lakehouse bao gồm:

- Định dạng lưu trữ dữ liệu mở: việc sử dụng các định dạng mở để lưu trữ dữ liệu, chẳng hạn như Apache Parquet, đảm bảo hiệu quả và tối ưu hóa truy vấn.
- Sơ đồ chỉ đọc: khác với phương pháp truyền thống về sơ đồ chỉ ghi trong DWH, Lakehouse hỗ trợ sơ đồ chỉ đọc, cho phép quản lý cấu trúc dữ liệu một cách linh hoạt hơn.
- Tính linh hoạt và khả năng mở rộng: hỗ trợ lưu trữ và phân tích dữ liệu có cấu trúc và không có cấu trúc, đảm bảo hiệu suất truy vấn cao thông qua tối ưu hóa ở cấp độ kho.

Data Lakehouse cung cấp một giải pháp thỏa hiệp, kết hợp những lợi ích của cả hai phương pháp, làm cho nó trở thành lý tưởng cho các tải phân tích hiện đại, yêu cầu tính linh hoạt trong xử lý dữ liệu.



Hình 8.19 Data Lakehouse - thế hệ tiếp theo của các hệ thống lưu trữ dữ liệu, được tạo ra để đáp ứng các yêu cầu phức tạp và luôn thay đổi.

Ý tưởng về các kho dữ liệu hiện đại có vẻ đơn giản: nếu tất cả dữ liệu nằm ở một nơi, việc phân tích sẽ dễ dàng hơn. Tuy nhiên, trên thực tế, mọi thứ không diễn ra suôn sẻ như vậy. Hãy tưởng tượng rằng một công ty quyết định từ bỏ hoàn toàn các hệ thống quản lý và kế toán quen thuộc (ERP, PMIS, CAFM, v.v.), thay thế chúng bằng một hồ dữ liệu khổng lồ mà mọi người đều có quyền truy cập. Điều gì sẽ xảy ra? Có lẽ sẽ bắt đầu một sự hỗn loạn: dữ liệu sẽ bị trùng lặp, mâu thuẫn với nhau, và thông tin quan trọng sẽ bị mất hoặc bị sai lệch. Ngay cả khi hồ dữ liệu chỉ được sử dụng cho phân tích, nếu không có sự quản lý hợp lý, sẽ gặp phải những khó khăn nghiêm trọng.

- Dữ liệu khó hiểu: trong các hệ thống thông thường, dữ liệu có cấu trúc rõ ràng, trong khi ở hồ dữ liệu chỉ là một đống lớn các tệp và bảng. Để tìm kiếm thông tin, chuyên gia phải tìm hiểu - mỗi dòng và cột đại diện cho điều gì.

- Dữ liệu có thể không chính xác: nếu nhiều phiên bản của cùng một thông tin được lưu trữ ở một nơi, thật khó để xác định phiên bản nào là chính xác. Kết quả là, các quyết định được đưa ra dựa trên dữ liệu lỗi thời hoặc sai sót.
- Khó khăn trong việc chuẩn bị dữ liệu để làm việc: dữ liệu không chỉ cần được lưu trữ mà còn phải được trình bày dưới dạng báo cáo, đồ thị, bảng biểu. Trong các hệ thống truyền thống, điều này được thực hiện tự động, trong khi ở hồ dữ liệu cần thêm xử lý.

Cuối cùng, mỗi khái niệm lưu trữ dữ liệu có những đặc điểm riêng, cách tiếp cận xử lý và ứng dụng trong kinh doanh. Cơ sở dữ liệu truyền thống tập trung vào các giao dịch, kho dữ liệu (DWH) cung cấp cấu trúc cho phân tích, hồ dữ liệu (Data Lake) lưu trữ thông tin ở dạng thô, trong khi các kho dữ liệu lai (Data Lakehouse) kết hợp những lợi ích của DWH và Data Lake.-

	Traditional Approach	Data Warehouse	Data Lake	Data Lakehouse
Data Types	Relational Databases	Structured, ready for analytics	Raw, semi-structured, or unstructured	Mix of structured and unstructured
Use Cases	Transactional Systems	Reporting, dashboards, BI	Big data storage, AI, advanced analytics	Hybrid analytics, AI, real-time data
Processing	OLTP – real-time transactions	ETL – clean and structure before analysis	ELT – store raw data, transform later	ELT with optimized storage and real-time processing
Storage	On-premise servers	Centralized, SQL-based	Decentralized, flexible formats	Combines advantages of DWH and DL
Common Tools	MySQL, PostgreSQL	Snowflake, Redshift, BigQuery	Hadoop, AWS S3, Azure Data Lake	Databricks, Snowflake, Google BigLake

Hình 8.110 DWH, Data Lake và Data Lakehouse: sự khác biệt chính trong các loại dữ liệu, kịch bản sử dụng, phương pháp xử lý và cách tiếp cận lưu trữ.

Việc lựa chọn kiến trúc lưu trữ dữ liệu là một quá trình phức tạp, phụ thuộc vào nhu cầu kinh doanh, khối lượng thông tin và yêu cầu phân tích. Mỗi giải pháp đều có ưu và nhược điểm riêng: DWH cung cấp tính cấu trúc, Data Lake - tính linh hoạt, và Lakehouse - sự cân bằng giữa chúng. Các tổ chức hiếm khi chỉ sử dụng một kiến trúc dữ liệu duy nhất.

> Bất kể kiến trúc nào được chọn, các hệ thống quản lý dữ liệu tự động vượt trội hơn nhiều so với các phương pháp thủ công. Chúng giúp giảm thiểu sai sót do con người, tăng tốc độ xử lý thông tin, đảm bảo tính minh bạch và khả năng truy xuất dữ liệu ở tất cả các giai đoạn của quy trình kinh doanh.

Và nếu các kho dữ liệu tập trung đã trở thành tiêu chuẩn ngành trong nhiều lĩnh vực kinh tế, thì trong ngành xây dựng, tình hình vẫn còn phân mảnh. Dữ liệu ở đây được phân tán giữa các nền tảng khác nhau (CDE, PMIS, ERP, v.v.), điều này gây khó khăn cho việc tạo ra một bức tranh tổng thể và yêu cầu các kiến trúc có khả năng kết hợp những nguồn này thành một môi trường số toàn diện, có thể phân tích được.

CDE, PMIS, ERP hoặc DWH và Data Lake

Trong một số công ty hoạt động trong lĩnh vực xây dựng và thiết kế, khái niệm môi trường dữ liệu chung (Common Data Environment, CDE) đã được áp dụng theo tiêu chuẩn ISO 19650. Về cơ bản, CDE thực hiện những chức năng tương tự như kho dữ liệu (DWH) trong các ngành khác: tập trung thông tin, đảm bảo kiểm soát phiên bản, cung cấp quyền truy cập vào thông tin đã được xác minh.

> Môi trường dữ liệu chung (CDE) là một không gian kỹ thuật số tập trung, được sử dụng để quản lý, lưu trữ, chia sẻ và hợp tác với thông tin dự án ở tất cả các giai đoạn của vòng đời công trình. CDE thường được triển khai bằng cách sử dụng công nghệ đám mây và tích hợp với các hệ thống CAD (BIM).

Ngành tài chính, bán lẻ, logistics và công nghiệp đã sử dụng các hệ thống quản lý dữ liệu tập trung trong nhiều thập kỷ, kết hợp thông tin từ nhiều nguồn khác nhau, kiểm soát tính chính xác của nó và cung cấp phân tích. CDE phát triển những nguyên tắc này, điều chỉnh chúng cho các nhiệm vụ thiết kế và quản lý vòng đời của các tòa nhà.

Giống như DWH, CDE cấu trúc dữ liệu, ghi lại các thay đổi và đảm bảo quyền truy cập duy nhất vào thông tin đã được xác minh. Với sự chuyển đổi sang công nghệ đám mây và tích hợp với các công cụ phân tích, sự khác biệt giữa chúng ngày càng trở nên mờ nhạt. Bằng cách thêm dữ liệu chi tiết vào CDE, khái niệm mà các nhà cung cấp CAD đã thảo luận từ năm 2023, có thể thấy thêm nhiều điểm tương đồng với các DWH cổ điển.

Trước đó trong chương "Hệ thống ERP và PMIS trong xây dựng", chúng ta đã xem xét PMIS (Hệ thống Thông tin Quản lý Dự án) và ERP (Kế hoạch Tài nguyên Doanh nghiệp). Trong các dự án xây dựng, CDE và PMIS hoạt động cùng nhau: CDE phục vụ như một kho dữ liệu, bao gồm bản vẽ, mô hình và tài liệu dự án, trong khi PMIS quản lý các quy trình như kiểm soát thời gian, nhiệm vụ, tài nguyên và ngân sách.

ERP, chịu trách nhiệm quản lý doanh nghiệp tổng thể (tài chính, mua sắm, nhân sự, sản xuất), có thể tích hợp với PMIS, đảm bảo kiểm soát chi phí và ngân sách ở cấp độ công ty. Để phân tích và báo cáo, có thể sử dụng DWH, giúp thu thập, cấu trúc và tổng hợp dữ liệu từ CDE, PMIS và ERP, từ đó đánh giá các chỉ số tài chính KPI (ROI) và phát hiện các mẫu. Ngược lại, Data Lake (DL) có thể bổ sung cho DWH, lưu trữ dữ liệu thô và không cấu trúc (ví dụ: nhật ký, dữ liệu cảm biến, hình ảnh). Những dữ liệu này có thể được xử lý và tải vào DWH để phân tích tiếp theo.

> Do đó, CDE và PMIS tập trung vào quản lý dự án, ERP tập trung vào quy trình kinh doanh, trong khi DWH và Data Lake tập trung vào phân tích và làm việc với dữ liệu.

So sánh các hệ thống CDE, PMIS và ERP với DWH và Data Lake, có thể nhận thấy sự khác biệt đáng kể về độ độc lập với nhà cung cấp, chi phí, tính linh hoạt trong tích hợp, độc lập dữ liệu, tốc độ thích ứng với thay đổi, cũng như khả năng phân tích. Các hệ thống truyền thống như CDE, PMIS và ERP thường liên quan đến các giải pháp và tiêu chuẩn cụ thể của nhà cung cấp, điều này làm cho chúng kém linh hoạt hơn và tăng chi phí do giấy phép và hỗ trợ. Hơn nữa, dữ liệu trong các hệ thống như vậy thường bị khóa trong các định dạng đóng độc quyền, điều này hạn chế việc sử dụng và phân tích của chúng.-

		CDE, PMIS, ERP	DWH, Data Lake
	**Vendor Dependency**	High (tied to specific solutions and standards of vendors)	Low (flexibility in tool and platform choice)
	**Integration Flexibility**	Limited (integration depends on vendor solutions)	High (easily integrates with various data sources)
	**Cost**	High (licensing and support costs)	Relatively lower (use of open technologies and platforms)
	**Data Independence**	Low (data often locked in proprietary formats)	High (data stored in open and accessible formats)
	**Adaptability to Changes**	Slow (changes require vendor approval and integration)	Fast (adaptation and data structure modification without intermediaries)
	**Analytical Capabilities**	Limited (dependent on vendor-provided solutions)	Extensive (support for a wide range of analytical tools)

DWH và Data Lake cung cấp tính linh hoạt và độc lập dữ liệu lớn hơn so với các hệ thống như CDE, PMIS và ERP.

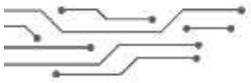
Ngược lại, DWH và Data Lake cung cấp tính linh hoạt lớn hơn trong việc tích hợp với các nguồn dữ liệu khác nhau, và việc sử dụng công nghệ và nền tảng mở giúp giảm tổng chi phí sở hữu. Hơn nữa, DWH và Data Lake hỗ trợ một loạt các công cụ phân tích, mở rộng khả năng phân tích và quản lý.

> Với sự phát triển của các công cụ reverse engineering cho các định dạng CAD và việc có sẵn quyền truy cập vào các cơ sở dữ liệu ứng dụng CAD, vấn đề càng trở nên cấp thiết: liệu có hợp lý không khi tiếp tục sử dụng các nền tảng đóng, biệt lập, khi mà dữ liệu dự án cần phải được truy cập bởi một đội ngũ chuyên gia rộng lớn, làm việc trong hàng chục tổ chức thầu và thiết kế khác nhau?

Sự phụ thuộc công nghệ vào một nhà cung cấp cụ thể có thể hạn chế đáng kể tính linh hoạt trong việc quản lý dữ liệu, làm chậm phản ứng với các thay đổi trong dự án và cản trở sự hợp tác hiệu quả giữa các bên tham gia.

Các phương pháp truyền thống trong việc quản lý dữ liệu — bao gồm DWH, Data Lake, CDE và PMIS — chủ yếu tập trung vào việc lưu trữ, cấu trúc và xử lý thông tin. Tuy nhiên, với sự phát triển của trí tuệ nhân tạo và học máy, nhu cầu về các phương pháp tổ chức dữ liệu mới đang gia tăng, cho phép không chỉ tập hợp mà còn phát hiện các mối quan hệ phức tạp, tìm ra những quy luật tiềm ẩn và cung cấp truy cập ngay lập tức đến thông tin liên quan nhất.

Trong lĩnh vực này, các cơ sở dữ liệu vector đang bắt đầu đóng vai trò đặc biệt — đây là một loại kho lưu trữ mới, được tối ưu hóa cho việc làm việc với các nhúng có nhiều chiều.



CHƯƠNG 8.2.
QUẢN LÝ KHO DỮ LIỆU VÀ NGĂN NGỪA HỖN LOẠN

Cơ sở dữ liệu vector và Bounding Box

Cơ sở dữ liệu vector là một lớp lưu trữ mới, không chỉ lưu giữ dữ liệu mà còn cho phép tìm kiếm theo nghĩa, so sánh các đối tượng theo gần gũi ngữ nghĩa và xây dựng các hệ thống thông minh: từ gợi ý đến phân tích tự động và tạo bối cảnh. Khác với các cơ sở dữ liệu truyền thống, được định hướng vào các sự trùng hợp chính xác, các cơ sở dữ liệu vector tìm kiếm các đối tượng tương tự trên cơ sở đặc điểm — ngay cả khi không có trùng hợp chính xác.

> Cơ sở dữ liệu vector là một loại cơ sở dữ liệu chuyên biệt, lưu giữ dữ liệu dưới dạng các vector đa chiều, mỗi vector thể hiện các đặc tính hoặc phẩm chất cụ thể. Những vector này có thể có số lượng chiều khác nhau, tùy thuộc vào độ phức tạp của dữ liệu (trong một số trường hợp, có thể chỉ vài chiều, trong khi ở trường hợp khác có thể lên đến hàng nghìn).

Lợi thế chính của cơ sở dữ liệu vector là tìm kiếm dựa trên ý nghĩa ngữ nghĩa, thay vì sự trùng hợp chính xác của các giá trị. Thay vì sử dụng các truy vấn SQL và Pandas với các bộ lọc "bằng" hoặc "chứa", tìm kiếm dựa trên hàng xóm gần nhất (k-NN) được sử dụng trong không gian đặc điểm (k-NN sẽ được nói rõ hơn trong phần tiếp theo của cuốn sách).

Với sự phát triển của Mô hình Ngôn ngữ Lớn (LLM) và các mô hình sinh, cách thức tương tác với các cơ sở dữ liệu bắt đầu thay đổi. Giờ đây, người dùng có thể truy xuất dữ liệu bằng ngôn ngữ tự nhiên, nhận được tìm kiếm ngữ nghĩa trong các tài liệu, tự động trích xuất các thuật ngữ chính và xây dựng các mối liên hệ bối cảnh giữa các đối tượng — tất cả đều không cần phải thành thạo SQL hoặc biết cấu trúc của các bảng. Chi tiết về điều này đã được đề cập trong phần "LLM và vai trò của chúng trong việc xử lý dữ liệu và quy trình kinh doanh".

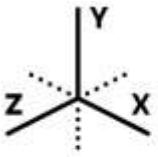
Tuy nhiên, điều quan trọng là phải hiểu rằng LLM không tự động cấu trúc thông tin và không tổ chức nó một cách hợp lý. Mô hình chỉ "bơi lội" trong một biển dữ liệu và tìm kiếm đoạn phù hợp nhất, dựa trên ngữ cảnh của yêu cầu. Nếu dữ liệu chưa được làm sạch hoặc chuyển đổi trước, việc tìm kiếm sâu (deep search) sẽ giống như là cố gắng tìm một câu trả lời trong "rác kỹ thuật số" — có thể làm việc được, nhưng chất lượng của kết quả sẽ kém. Lý tưởng là nếu dữ liệu có thể được cấu trúc trước (ví dụ, chuyển đổi tài liệu thành Markdown) và tải vào cơ sở dữ liệu vector. Điều này sẽ nâng cao đáng kể độ chính xác và sự liên quan của kết quả tìm kiếm.

> Ban đầu, các cơ sở dữ liệu vector được áp dụng trong học máy, nhưng ngày nay chúng ngày càng được sử dụng rộng rãi hơn ngoài lĩnh vực này - trong các hệ thống tìm kiếm, cá nhân hóa nội dung và phân tích thông minh.

Một trong những ví dụ rõ ràng nhất về cách tiếp cận vector trong xây dựng là Bounding Box (hộp giới hạn). Đây là một cấu trúc hình học mô tả ranh giới của vật thể trong không gian ba chiều. Bounding Box được xác định bởi các tọa độ tối thiểu và tối đa theo các trục X, Y và Z, tạo thành một "hộp" xung quanh vật thể. Phương pháp này cho phép đánh giá kích thước và vị trí của thành phần mà không cần phân tích toàn bộ

hình học.

Mỗi Bounding Box có thể được đại diện như một vector trong không gian đa chiều: ví dụ, [x, y, z, chiều rộng, chiều cao, chiều sâu] - đã có 6 chiều (Hình 8.21).-



Bounding Box

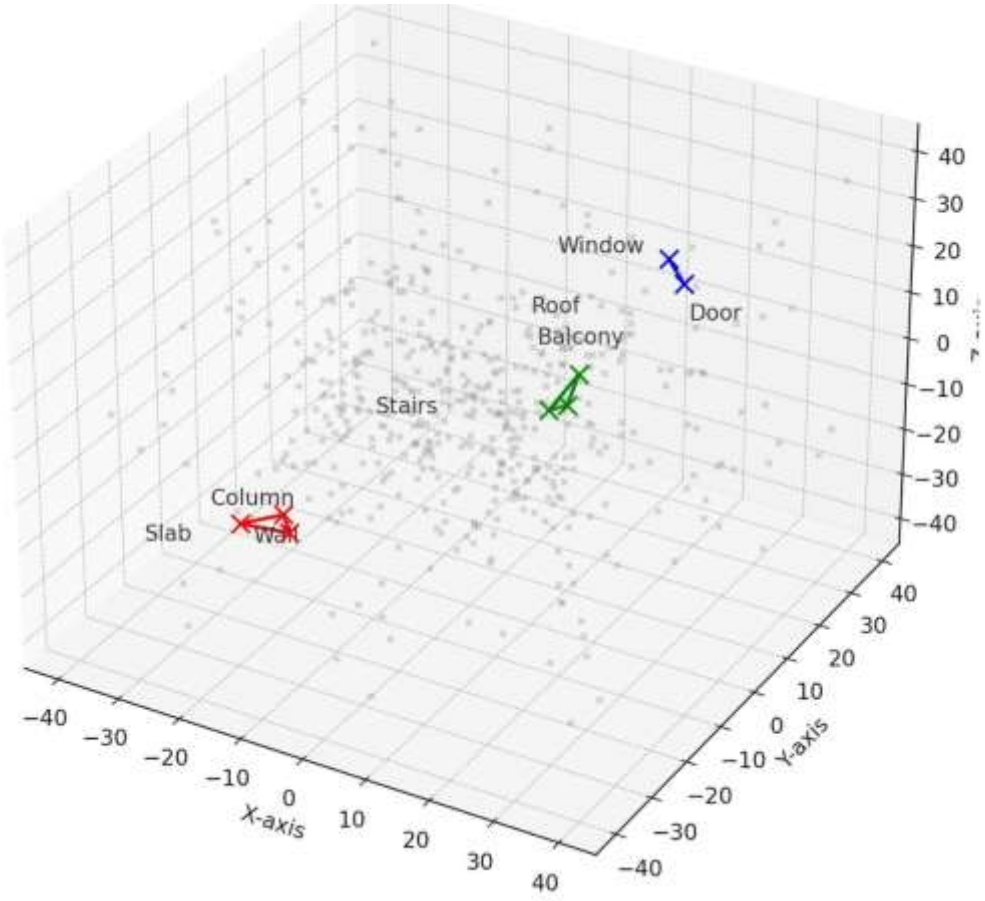
	minX	maxX	minY	maxY	minZ	maxZ	Width	Height	Depth
Column	-15	-5	-25	-15	0	10	10	10	20
Stairs	-5	5	-15	-5	0	10	10	10	10
Door	5	15	5	15	0	10	10	10	10
Window	25	35	-35	-25	10	30	10	20	20
Balcony	15	25	-5	5	20	40	10	20	20

Hình 8.21 Thông tin về tọa độ của các phần tử Bounding Box và vị trí của chúng trong mô hình dự án là một phép ẩn dụ cho cơ sở dữ liệu vector.

Việc trình bày dữ liệu như vậy giúp đơn giản hóa nhiều nhiệm vụ, bao gồm kiểm tra sự giao nhau giữa các đối tượng, lập kế hoạch phân bổ không gian cho các yếu tố của tòa nhà và thực hiện các phép tính tự động. Bounding Box có thể đóng vai trò là cầu nối giữa các mô hình ba chiều phức tạp và các cơ sở dữ liệu vector truyền thống, cho phép sử dụng hiệu quả lợi thế của cả hai phương pháp trong mô hình hóa kiến trúc và kỹ thuật.

> Bounding Box là "vecto hóa hình học", còn embedding (cách chuyển đổi một thứ gì đó trừu tượng) là "vecto hóa ý nghĩa". Cả hai phương pháp cho phép chuyển từ việc tìm kiếm thủ công sang tìm kiếm thông minh, dù là các đối tượng 3D trong mô hình dự án hay các khái niệm trong văn bản.

Tìm kiếm đối tượng trong dự án (ví dụ, "tìm tất cả các cửa sổ có chiều rộng > 1,5 m") tương tự như tìm kiếm hàng xóm gần nhất (k-NN) trong cơ sở dữ liệu vectơ, nơi các tiêu chí xác định "khu vực" trong không gian đặc trưng. (chúng ta sẽ bàn thêm về tìm kiếm k-NN trong phần tiếp theo về học máy) (Hình 8.22). Nếu thêm các tham số bổ sung vào bounding box như vật liệu, trọng lượng, thời gian sản xuất, thì bảng trở thành một vectơ nhiều chiều, trong đó mỗi thuộc tính là một chiều mới. Điều này đã gần gũi hơn với các cơ sở dữ liệu vectơ hiện đại, nơi các chiều được đo bằng hàng trăm hoặc hàng nghìn (ví dụ, nhúng từ mạng nơ-ron).-



Hình 8.22 Tìm kiếm đối tượng trong dự án bằng cách sử dụng cơ sở dữ liệu vector.

Phương pháp được sử dụng trong Bounding Box không chỉ áp dụng cho các đối tượng hình học mà còn cho việc phân tích văn bản và ngôn ngữ. Các đại diện vector của dữ liệu đã được sử dụng tích cực trong xử lý ngôn ngữ tự nhiên (NLP). Tương tự như việc các đối tượng trong một dự án xây dựng có thể được nhóm lại theo sự gần gũi về không gian (Hình 8.22), các từ trong văn bản cũng có thể được phân tích dựa trên sự gần gũi về nghĩa và ngữ cảnh.-

Ví dụ, các từ "kiến trúc sư", "xây dựng", "thiết kế" trong không gian vector sẽ nằm gần nhau, vì có ý nghĩa tương tự. Trong LLM, cơ chế này cho phép tự động, mà không cần phân loại thủ công:

- Xác định chủ đề của văn bản
- Thực hiện tìm kiếm ngữ nghĩa theo nội dung tài liệu.
- Tạo ra các chú thích và tóm tắt tự động cho văn bản.
- Tìm kiếm từ đồng nghĩa và các thuật ngữ liên quan.

Cơ sở dữ liệu vector cho phép phân tích văn bản và tìm kiếm các thuật ngữ liên quan giống như Bounding Box giúp phân tích các đối tượng không gian trong mô hình 3D. Ví dụ về Bounding Box của các phần tử trong dự án giúp hiểu rằng đại diện vector không phải chỉ là một khái niệm "nhân tạo" trong ML, mà là một cách tự nhiên để cấu trúc dữ liệu nhằm giải quyết các vấn đề thực tiễn, bất kể là tìm kiếm các cột trong dự án CAD hay các hình ảnh có ngữ nghĩa gần gũi trong cơ sở dữ liệu.

> Các chuyên gia làm việc với cơ sở dữ liệu nên chú ý đến các kho lưu trữ vector. Việc gia tăng sử dụng của chúng chỉ ra một giai đoạn mới trong sự phát triển của cơ sở dữ liệu, nơi mà các hệ thống quan hệ cổ điển và các công nghệ AI bắt đầu đan xen nhau, hình thành các giải pháp lai của tương lai.

Người dùng phát triển các ứng dụng AI phức tạp và quy mô lớn sẽ sử dụng các cơ sở dữ liệu chuyên biệt cho việc tìm kiếm vector. Trong khi đó, những ai chỉ cần một số chức năng AI riêng lẻ để tích hợp vào các ứng dụng hiện có sẽ có xu hướng chọn các khả năng tìm kiếm vector đã được tích hợp trong các cơ sở dữ liệu mà họ đang sử dụng (PostgreSQL, Redis).

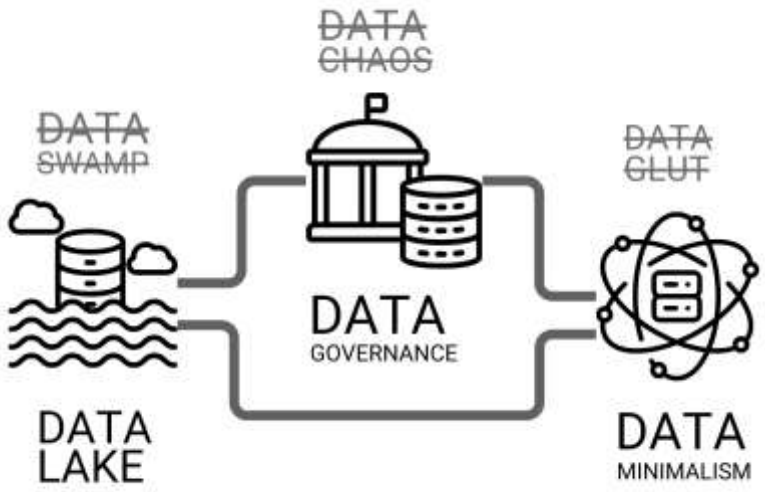
Mặc dù các hệ thống như DWH, Data Lake, CDE, PMIS, cơ sở dữ liệu vector và các hệ thống khác cung cấp các cách tiếp cận khác nhau cho việc lưu trữ và quản lý dữ liệu, hiệu quả của chúng không chỉ được xác định bởi kiến trúc mà còn bởi cách mà dữ liệu được tổ chức và quản lý hợp lý. Ngay cả khi sử dụng các giải pháp hiện đại — bất kể là cơ sở dữ liệu vector, cơ sở dữ liệu quan hệ cổ điển hay kho dữ liệu kiểu Data Lake — việc thiếu các quy tắc rõ ràng trong quản lý, cấu trúc và cập nhật dữ liệu có thể dẫn đến những khó khăn tương tự mà người dùng gặp phải khi làm việc với các tệp rời rạc và dữ liệu đa định dạng.

Nếu không có quản lý dữ liệu hợp lý (Data Governance), ngay cả các giải pháp mạnh nhất cũng có thể trở thành những tập hợp thông tin hỗn loạn và không có cấu trúc, biến các hồ dữ liệu thành đầm lầy dữ liệu (Data Swamp). Để tránh điều này, các công ty cần không chỉ chọn kiến trúc kho lưu trữ phù hợp mà còn triển khai các chiến lược tối giản dữ liệu (Data Minimalism), quản lý quyền truy cập và kiểm soát chất lượng, cho phép biến dữ liệu thành công cụ hiệu quả cho việc ra quyết định.

Quản lý dữ liệu (Data Governance), tối giản dữ liệu (Data Minimalism) và đầm lầy dữ liệu (Data Swamp)

Việc hiểu và triển khai các khái niệm quản lý dữ liệu (Data Governance), tối giản dữ liệu (Data Minimalism) và ngăn chặn sự hình thành đầm lầy dữ liệu (Data Swamp) là những yếu tố then chốt cho việc quản lý kho dữ liệu thành công và đảm bảo giá trị của chúng cho doanh nghiệp.-

> Theo nghiên cứu của Gartner (2017), 85% các dự án trong lĩnh vực dữ liệu lớn gặp thất bại, và một trong những lý do chính là việc quản lý chất lượng dữ liệu không đầy đủ và quản lý chúng.



Một trong những khía cạnh quan trọng của quản lý dữ liệu là Data Governance và Data Minimalism.

Quản lý dữ liệu (Data Governance) là một thành phần cơ bản của quản lý dữ liệu, đảm bảo việc sử dụng dữ liệu một cách đúng đắn và hiệu quả trong tất cả các quy trình kinh doanh. Nó không chỉ liên quan đến việc thiết lập các quy tắc và thủ tục mà còn đảm bảo tính khả dụng, độ tin cậy và an toàn của dữ liệu.

- Xác định và phân loại dữ liệu: việc xác định và phân loại rõ ràng các thực thể cho phép các tổ chức hiểu được các thực thể nào cần thiết trong công ty và xác định cách sử dụng chúng.
- Quyền truy cập và quản lý: phát triển các chính sách và thủ tục truy cập và quản lý dữ liệu đảm bảo rằng chỉ những người dùng đã được ủy quyền mới có thể truy cập vào các dữ liệu nhất định.
- Bảo vệ dữ liệu khỏi các mối đe dọa bên ngoài: bảo vệ dữ liệu khỏi các mối đe dọa bên ngoài là một trong những khía cạnh then chốt trong quản lý dữ liệu. Điều này không chỉ bao gồm các biện pháp kỹ thuật mà còn cả việc đào tạo nhân viên về các nguyên tắc bảo mật thông tin.

Chủ nghĩa tối giản dữ liệu (Data Minimalism) - đây là cách tiếp cận để giảm thiểu dữ liệu đến các thông tin quý giá và có ý nghĩa nhất khi hình thành các thuộc tính và thực thể (Hình 8.24), điều này giúp tiết kiệm chi phí và nâng cao hiệu quả sử dụng dữ liệu: -

- Đơn giản hóa quá trình ra quyết định: giảm số lượng đối tượng và các thuộc tính của chúng xuống những thông tin có ý nghĩa nhất giúp đơn giản hóa quá trình ra quyết định bằng cách giảm thời gian và nguồn lực cần thiết cho việc phân tích và xử lý dữ liệu.
- Tập trung vào điều quan trọng: lựa chọn các thực thể và thuộc tính liên quan nhất cho phép tập trung vào thông tin thực sự quan trọng cho doanh nghiệp, loại bỏ tiếng ồn và dữ liệu không cần thiết.
- Phân bổ tài nguyên hiệu quả: việc giảm thiểu dữ liệu cho phép phân bổ tài nguyên hiệu quả hơn, giảm chi phí lưu trữ và xử lý dữ liệu, tăng cường chất lượng và bảo mật dữ liệu.

Lógica làm việc với dữ liệu nên bắt đầu không phải từ việc tạo ra chúng như vậy (Hình 8.24), mà từ việc hiểu các kịch bản sử dụng dữ liệu đã thấy trước trước khi bắt đầu quá trình tạo ra. Cách tiếp cận này cho phép xác định trước các yêu cầu tối thiểu cần thiết cho các thuộc tính, kiểu loại và các giá trị biên. Chính những yêu cầu này tạo nền tảng cho việc tạo ra các thực thể hợp lệ và bền vững trong mô hình thông tin. Việc suy nghĩ trước về các mục tiêu và cách sử dụng dữ liệu góp phần hình thành cấu trúc thích hợp cho phân tích.

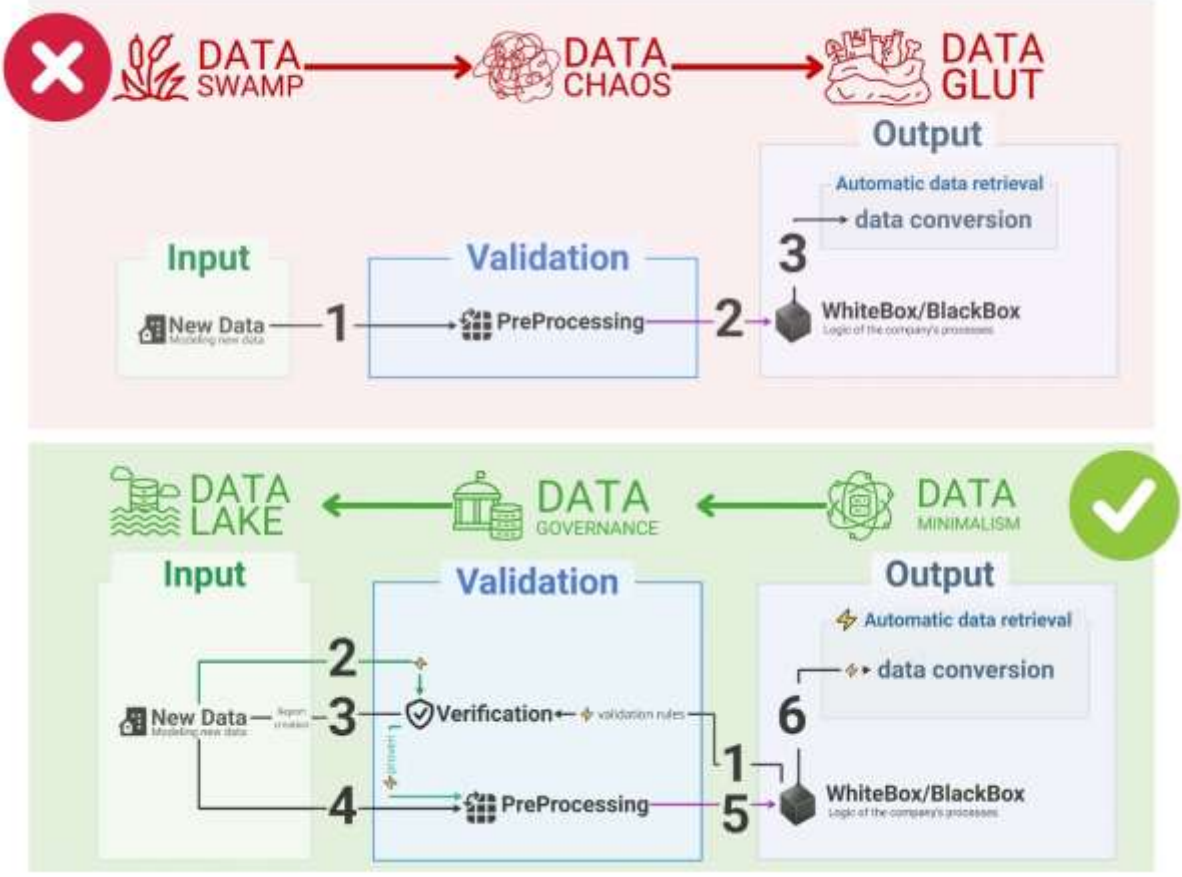
Chi tiết về các phương pháp mô hình hóa dữ liệu ở các cấp độ khái niệm, logic và vật lý đã được đề cập trong chương "Mô hình hóa dữ liệu: mô hình khái niệm, logic và vật lý".

> Trong các quy trình kinh doanh truyền thống của các công ty xây dựng, việc xử lý dữ liệu thường giống như việc thả dữ liệu vào một cái ao, nơi mà trước tiên dữ liệu được tạo ra, sau đó các chuyên gia cố gắng tích hợp chúng vào các hệ thống và công cụ khác.

Ao dữ liệu (Data Swamp) - là kết quả của việc thu thập và lưu trữ dữ liệu không kiểm soát mà không có tổ chức, cấu trúc và quản lý hợp lý, dẫn đến dữ liệu trở nên không có cấu trúc, khó sử dụng và có ít giá trị.

Làm thế nào để ngăn chặn việc biến dòng thông tin thành ao dữ liệu:

- Quản lý cấu trúc dữ liệu: đảm bảo cấu trúc và phân loại dữ liệu giúp ngăn chặn "ao dữ liệu", làm cho chúng được tổ chức và dễ dàng truy cập.
- Hiểu và diễn giải dữ liệu: mô tả rõ ràng nguồn gốc của dữ liệu, các thay đổi của chúng và các giá trị đảm bảo rằng dữ liệu sẽ được hiểu và diễn giải đúng cách.
- Duy trì chất lượng dữ liệu: bảo trì và làm sạch dữ liệu thường xuyên giúp duy trì chất lượng, tính thời sự và giá trị của chúng đối với các quy trình phân tích và kinh doanh.



Hình 8.24 Để tránh sự lộn xộn trong kho dữ liệu, quy trình tạo dữ liệu cần bắt đầu từ việc thu thập yêu cầu về các thuộc tính.

> Bằng cách tích hợp các nguyên tắc quản lý dữ liệu và chủ nghĩa tối giản dữ liệu vào trong quy trình quản lý dữ liệu, cũng như tích cực ngăn chặn việc biến kho dữ liệu thành ao dữ liệu, các tổ chức có thể tối đa hóa tiềm năng của dữ liệu của mình.

Giai đoạn tiếp theo trong sự tiến hóa của công việc với dữ liệu sau khi giải quyết các vấn đề về quản lý và tối giản là chuẩn hóa việc xử lý tự động, đảm bảo chất lượng và triển khai các phương pháp làm cho dữ liệu dễ dàng phân tích, chuyển đổi và ra quyết định. Chính điều này đang được thực hiện bởi các phương pháp DataOps và VectorOps, những công cụ ngày càng quan trọng cho các công ty làm việc với khối lượng thông tin lớn và học máy.

DataOps và VectorOps: tiêu chuẩn mới trong làm việc với dữ liệu

Nếu Data Governance chịu trách nhiệm kiểm soát và tổ chức dữ liệu, thì DataOps giúp đảm bảo tính chính xác, tính nhất quán và luồng dữ liệu liên tục trong nội bộ công ty. Điều này đặc biệt quan trọng cho một số trường hợp kinh doanh trong xây dựng, nơi dữ liệu được tạo ra liên tục và cần được xử lý kịp thời. Ví dụ, trong các tình huống mà các mô hình thông tin của tòa nhà, yêu cầu thiết kế và báo cáo phân tích cần được đồng bộ hóa giữa các hệ thống khác nhau trong một ngày làm việc, vai trò của DataOps có thể trở nên quyết định. Nó cho phép thiết lập quy trình xử lý dữ liệu ổn định và có thể tái lập, giảm thiểu rủi ro về độ trễ và mất tính cập nhật của thông tin.

Quản lý ở cấp độ Data Governance tự nó là không đủ - quan trọng là dữ liệu không chỉ được lưu trữ mà còn được sử dụng tích cực trong các hoạt động hàng ngày. Chính ở đây DataOps nổi bật - phương pháp luận tập trung vào tự động hóa, tích hợp và đảm bảo dòng chảy dữ liệu liên tục.

> DataOps tập trung vào việc cải thiện sự hợp tác, tích hợp và tự động hóa các luồng dữ liệu trong các tổ chức. Việc triển khai các thực tiễn DataOps góp phần vào tính chính xác, tính nhất quán và khả năng tiếp cận dữ liệu, điều này cực kỳ quan trọng cho các ứng dụng định hướng dữ liệu.

Những công cụ chính trong hệ sinh thái DataOps là Apache Airflow (Hình 7.44) - để điều phối quy trình làm việc, và Apache NiFi (Hình 7.45) - để định tuyến và chuyển đổi luồng dữ liệu. Cùng nhau, các công nghệ này cho phép xây dựng các đường ống dữ liệu linh hoạt, đáng tin cậy và có thể mở rộng, đảm bảo xử lý tự động, kiểm soát và tích hợp thông tin giữa các hệ thống (chi tiết hơn trong chương «Đường ống ETL tự động»). Khi triển khai phương pháp DataOps trong quy trình xây dựng, cần xem xét bốn khía cạnh cơ bản.--

1. Con người và công cụ quan trọng hơn dữ liệu: các kho dữ liệu phân tán có thể được coi là vấn đề chính, nhưng trên thực tế tình huống thì phức tạp hơn. Ngoài việc phân mảnh dữ liệu, sự tách biệt của các đội nhóm và sự phân tán các công cụ họ sử dụng cũng đóng vai trò quan trọng. Trong xây dựng, dữ liệu được xử lý bởi các chuyên gia từ nhiều lĩnh vực khác nhau: kỹ sư dữ liệu và nhà phân tích, các đội BI và trực quan hóa, cũng như các chuyên gia quản lý dự án và chất lượng. Mỗi người trong số họ có phương pháp làm việc riêng, do đó, một yếu tố quan trọng là xây dựng một hệ sinh thái, nơi dữ liệu tự do được chuyển giao giữa các bên tham gia, đảm bảo một phiên bản thông tin thống nhất, nhất quán.
2. Tự động hóa kiểm thử và phát hiện lỗi: dữ liệu xây dựng luôn chứa đựng những sai sót, cho dù đó là sự không chính xác trong các mô hình, lỗi trong tính toán hay thông số kỹ thuật lỗi thời. Việc kiểm

thử dữ liệu thường xuyên và loại bỏ các lỗi lặp lại có thể nâng cao đáng kể chất lượng dữ liệu. Trong khuôn khổ DataOps, cần triển khai các cơ chế tự động kiểm soát và xác minh để theo dõi tính chính xác của dữ liệu, phân tích lỗi và phát hiện các quy luật, đồng thời ghi lại và khắc phục các sự cố hệ thống trong mỗi chu kỳ làm việc. Mức độ tự động hóa kiểm tra càng cao, chất lượng dữ liệu càng tốt và xác suất xảy ra lỗi trong các giai đoạn cuối càng thấp.

3. Dữ liệu cũng cần được kiểm thử giống như mã nguồn: hầu hết các ứng dụng xây dựng đều dựa trên việc xử lý dữ liệu, tuy nhiên việc kiểm soát chúng thường rơi vào vai trò thứ yếu. Nếu các mô hình học máy được đào tạo trên dữ liệu không chính xác, sẽ dẫn đến các dự đoán sai lệch và tổn thất tài chính. Trong khuôn khổ DataOps, dữ liệu cần trải qua quy trình kiểm tra nghiêm ngặt như mã nguồn: kiểm tra logic, stress test, đánh giá hành vi của các mô hình khi thay đổi các giá trị đầu vào. Chỉ có dữ liệu đã được kiểm định và đáng tin cậy mới có thể được sử dụng làm cơ sở cho việc ra quyết định quản lý.
4. Quan sát dữ liệu mà không ảnh hưởng đến hiệu suất: theo dõi dữ liệu không chỉ đơn thuần là thu thập số liệu, mà còn là một công cụ chiến lược trong quản lý chất lượng. Để DataOps hoạt động hiệu quả, tính quan sát cần được xây dựng vào tất cả các giai đoạn xử lý dữ liệu, từ thiết kế đến vận hành. Đồng thời, rất quan trọng để việc theo dõi không làm chậm hệ thống. Trong các dự án xây dựng, không chỉ việc thu thập dữ liệu là quan trọng, mà còn cần thực hiện điều này một cách mà công việc của các chuyên gia (ví dụ như các nhà thiết kế) - những người tạo ra dữ liệu này - không bị gián đoạn. Cân bằng này cho phép kiểm soát chất lượng dữ liệu mà không ảnh hưởng đến hiệu suất.

DataOps không phải là gánh nặng bổ sung cho các chuyên gia dữ liệu, mà là nền tảng cho công việc của họ. Bằng cách triển khai DataOps, các công ty xây dựng có thể chuyển từ việc quản lý hỗn loạn sang một hệ sinh thái hiệu quả, nơi dữ liệu phục vụ cho doanh nghiệp.

VectorOps là bước tiếp theo trong tiến trình phát triển của DataOps, tập trung vào xử lý, lưu trữ và phân tích dữ liệu vector đa chiều (như đã thảo luận trong chương trước). Điều này đặc biệt quan trọng trong các lĩnh vực như mô hình đôi số, các mô hình mạng nơ-ron và tìm kiếm ngữ nghĩa, những thứ đang bắt đầu xuất hiện trong ngành xây dựng. VectorOps dựa trên các cơ sở dữ liệu vector cho phép lưu trữ, lập chỉ mục và tìm kiếm hiệu quả các biểu diễn đa chiều của các đối tượng.

> VectorOps là bước tiếp theo sau DataOps, tập trung vào xử lý, phân tích và ứng dụng dữ liệu vector trong xây dựng. Khác với DataOps, vốn tập trung vào dòng chảy, tính nhất quán và chất lượng dữ liệu, VectorOps chú trọng vào việc quản lý các biểu diễn đa chiều của các đối tượng cần thiết cho máy học.

Khác với các phương pháp truyền thống, VectorOps cho phép đạt được mô tả chính xác hơn về các đối tượng, điều này cực kỳ quan trọng cho các mô hình đôi số, hệ thống thiết kế sinh điều kiện và phát hiện lỗi tự động trong dữ liệu CAD đã được chuyển đổi sang định dạng vector. Việc triển khai đồng thời DataOps và VectorOps tạo nên nền tảng vững chắc cho việc làm việc với khối lượng thông tin lớn một cách tự động, từ các bảng tĩnh truyền thống cho đến các mô hình không gian phong phú về ngữ nghĩa.

Các bước tiếp theo: từ lưu trữ hỗn loạn đến các kho lưu trữ có cấu trúc

Các phương pháp lưu trữ dữ liệu xây dựng truyền thống thường dẫn đến việc tạo ra các "kho thông tin" rời rạc, nơi mà những thông tin quan trọng trở nên không thể truy cập cho phân tích và quyết định. Các khái

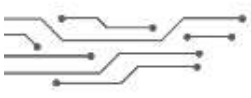
niệm lưu trữ dữ liệu hiện đại, chẳng hạn như Data Warehouse, Data Lake và các biến thể của chúng, cho phép kết hợp thông tin rời rạc và làm cho nó có sẵn một cách tập trung cho quá trình xử lý dữ liệu theo dòng và phân tích kinh doanh. Việc chọn một kiến trúc lưu trữ phù hợp là rất quan trọng, nhưng cũng cần thực hiện các nguyên tắc quản lý dữ liệu (Data Governance) và tối thiểu hóa dữ liệu (Data Minimalism) để ngăn chặn việc biến các kho dữ liệu thành những "vũng dữ liệu" không thể kiểm soát (Data Swamp).

Tóm lại, phần này nên làm nổi bật những bước thực tiễn chính giúp áp dụng các khái niệm đã thảo luận vào những nhiệm vụ hàng ngày của bạn:

- Chọn định dạng lưu trữ dữ liệu hiệu quả
  - [ ] Chuyển từ CSV và XLSX sang các định dạng hiệu quả hơn (Apache Parquet, ORC) để lưu trữ một lượng lớn dữ liệu.
  - [ ] Triển khai hệ thống phiên bản dữ liệu để theo dõi các thay đổi.
  - [ ] Sử dụng siêu dữ liệu để mô tả cấu trúc và nguồn gốc của thông tin.
- Tạo ra một kiến trúc dữ liệu thống nhất cho công ty.
  - [ ] So sánh các kiến trúc lưu trữ dữ liệu khác nhau: RDBMS, DWH và Data Lake. Chọn cái nào phù hợp nhất với các nhiệm vụ của bạn về khả năng mở rộng, tích hợp nguồn và xử lý phân tích.
  - [ ] Thiết kế bản đồ quy trình trích xuất, tải lên và biến đổi dữ liệu (ETL) từ nhiều nguồn khác nhau cho các nhiệm vụ của bạn. Sử dụng các công cụ trực quan hóa như Miro, Lucidchart hoặc Draw.io để thể hiện một cách trực quan các bước quan trọng và điểm tích hợp.
- Triển khai các thực hành Quản lý Dữ liệu và Tối thiểu hóa Dữ liệu.
  - [ ] Thực hiện nguyên tắc Tối thiểu hóa Dữ liệu - chỉ lưu trữ và xử lý những gì thực sự có giá trị.
  - [ ] Triển khai các nguyên tắc Quản lý Dữ liệu - xác định trách nhiệm về dữ liệu, đảm bảo chất lượng và tính minh bạch.
  - [ ] Tìm hiểu thêm về chính sách quản lý dữ liệu và các khái niệm DataOps, VectorOps.
  - [ ] Xác định các tiêu chí chất lượng dữ liệu và quy trình kiểm tra chúng trong khuôn khổ DataOps.

Việc lưu trữ dữ liệu được tổ chức tốt tạo ra nền tảng cho việc tập trung hóa quy trình phân tích trong công ty. Việc chuyển từ việc tích lũy tệp tin một cách hỗn độn sang các kho dữ liệu có cấu trúc cho phép biến thông tin thành một tài sản chiến lược, giúp đưa ra những quyết định hợp lý và nâng cao hiệu quả của các quy trình kinh doanh.

Sau khi các quy trình thu thập, biến đổi, phân tích và lưu trữ dữ liệu có cấu trúc được tự động hóa và chuẩn hóa, bước tiếp theo trong việc chuyển đổi số sẽ là làm việc đầy đủ với dữ liệu lớn (Big Data).





IX PHẦN
DỮ LIỆU LỚN, HỌC MÁY VÀ DỰ ĐOÁN

Phần thứ chín tập trung vào dữ liệu lớn, học máy và dự đoán trong ngành xây dựng. Tại đây, quá trình chuyển đổi từ quyết định dựa trên trực giác sang phân tích khách quan dựa trên dữ liệu lịch sử được xem xét. Những ví dụ thực tiễn cho thấy phân tích dữ liệu lớn trong lĩnh vực xây dựng — từ việc phân tích tập dữ liệu cấp phép xây dựng tại San Francisco đến xử lý các dự án CAD với hàng triệu yếu tố. Đặc biệt chú ý đến các phương pháp học máy để dự đoán chi phí và thời gian của các dự án xây dựng, với phân tích chi tiết các thuật toán hồi quy tuyến tính và k láng giềng gần nhất. Điều này cho thấy cách mà dữ liệu có cấu trúc trở thành nền tảng cho các mô hình dự đoán, cho phép đánh giá rủi ro, tối ưu hóa tài nguyên và cải thiện hiệu quả quản lý dự án. Phần này cũng chứa các khuyến nghị về việc lựa chọn mẫu dữ liệu đại diện và giải thích tại sao để phân tích hiệu quả không phải lúc nào cũng cần khối lượng thông tin khổng lồ.

CHƯƠNG 9.1. **DỮ LIỆU LỚN VÀ PHÂN TÍCH CHÚNG**

Dữ liệu lớn trong xây dựng: từ trực giác đến khả năng dự đoán.

Thuật ngữ "dữ liệu lớn" không có định nghĩa chặt chẽ. Ban đầu, khái niệm này xuất hiện khi khối lượng thông tin vượt quá khả năng xử lý của các phương pháp truyền thống. Ngày nay, khối lượng và độ phức tạp của dữ liệu trong nhiều ngành, bao gồm cả xây dựng, đã tăng lên đáng kể đến mức không thể chứa trong bộ nhớ cục bộ của máy tính và cần sử dụng các công nghệ mới để xử lý.

Bản chất của việc làm việc với dữ liệu lớn không chỉ nằm ở việc lưu trữ và xử lý, mà còn ở khả năng dự đoán. Trong ngành xây dựng, Big Data mở ra con đường từ những quyết định trực quan, dựa trên cách diễn giải chủ quan của các bảng và hình ảnh (như đã bàn trước đó), đến những dự đoán có căn cứ, được củng cố bởi các quan sát và thống kê thực tế.

Trái với nhận thức phổ biến, mục tiêu của việc làm việc với dữ liệu lớn không phải là "khiến máy tính suy nghĩ như con người", mà là áp dụng các mô hình toán học và thuật toán để phân tích các tập dữ liệu với mục đích phát hiện các quy luật, dự đoán sự kiện và tối ưu hóa quy trình.

> Dữ liệu lớn (Big Data) không phải là thế giới lạnh lẽo của các thuật toán, thiếu vắng ảnh hưởng con người. Ngược lại, dữ liệu lớn hoạt động kết hợp với bản năng, sai sót và sự sáng tạo của chúng ta. Chính sự không hoàn hảo của tư duy con người cho phép tìm ra những giải pháp phi tiêu chuẩn và tạo ra những đột phá.

Với sự phát triển của công nghệ số, ngành xây dựng đã bắt đầu tích cực áp dụng các phương pháp xử lý dữ liệu từ lĩnh vực CNTT. Nhờ các công cụ như Pandas và Apache Parquet, dữ liệu có cấu trúc và không có cấu trúc có thể được kết hợp, đơn giản hóa việc truy cập thông tin và giảm thiểu thiệt hại trong phân tích, trong khi các tập dữ liệu lớn từ tài liệu hoặc dự án CAD (Hình. 9.210 - Hình. 9.212) cho phép thu thập, phân tích và dự đoán dữ liệu ở tất cả các giai đoạn vòng đời của dự án.–-

Dữ liệu lớn có tác động chuyển mình đến ngành xây dựng, ảnh hưởng đến nhiều khía cạnh khác nhau. Việc áp dụng công nghệ Big Data mang lại kết quả trong một số lĩnh vực chính, bao gồm các lĩnh vực như sau:

- Phân tích tiềm năng đầu tư — dự đoán khả năng sinh lời và thời gian thu hồi vốn của các dự án dựa trên dữ liệu từ các công trình trước đó.
- Bảo trì dự đoán — phát hiện các sự cố thiết bị có khả năng xảy ra trước khi chúng thực sự xảy ra, giúp giảm thời gian ngừng hoạt động.
- Tối ưu hóa chuỗi cung ứng - dự đoán sự cố và nâng cao hiệu quả logistics.
- Phân tích hiệu quả năng lượng - hỗ trợ trong thiết kế các tòa nhà tiêu thụ năng lượng thấp.
- Giám sát an toàn - sử dụng cảm biến và thiết bị đeo để theo dõi điều kiện trên công trường xây dựng.
- Kiểm soát chất lượng - giám sát sự tuân thủ các tiêu chuẩn công nghệ trong thời gian thực.
- Quản lý nguồn nhân lực - phân tích năng suất và dự đoán nhu cầu nhân sự.

Rất khó để tìm thấy lĩnh vực nào trong xây dựng mà phân tích dữ liệu và dự đoán không được yêu cầu. Lợi

thế chính của các thuật toán dự đoán là khả năng tự học và cải tiến liên tục khi dữ liệu được tích lũy.

> Trong tương lai gần, trí tuệ nhân tạo sẽ không chỉ hỗ trợ các nhà xây dựng mà còn đưa ra các quyết định quan trọng - từ quy trình thiết kế đến các vấn đề vận hành tòa nhà.

Chi tiết về cách thức hình thành các dự đoán và sử dụng các mô hình học máy sẽ được trình bày trong phần tiếp theo của cuốn sách: "Học máy và dự đoán".

Để chuyển sang làm việc hiệu quả với dữ liệu lớn, cần thay đổi cách tiếp cận phân tích. Nếu trong các hệ thống cổ điển mà chúng ta đã xem xét trước đó, trọng tâm chủ yếu là các mối quan hệ nguyên nhân - kết quả, thì trong phân tích dữ liệu lớn, sự chú ý chuyển sang việc tìm kiếm các quy luật thống kê và tương quan, cho phép phát hiện các mối liên hệ ẩn và dự đoán hành vi của các đối tượng ngay cả khi không hiểu đầy đủ tất cả các yếu tố.

Vấn đề về tính hợp lý của dữ liệu lớn: tương quan, thống kê và mẫu dữ liệu.

Truyền thống, xây dựng dựa vào các giả thuyết chủ quan và kinh nghiệm cá nhân. Các kỹ sư đã giả định - với một xác suất nhất định - cách mà vật liệu sẽ hành xử, tải trọng mà cấu trúc có thể chịu đựng và thời gian dự án sẽ kéo dài bao lâu. Những giả định này thường được kiểm tra trong thực tế, thường với giá phải trả là thời gian, tài nguyên và rủi ro trong tương lai.

> Với sự xuất hiện của dữ liệu lớn, cách tiếp cận đã thay đổi một cách căn bản: các quyết định giờ đây không được đưa ra dựa trên những phỏng đoán trực giác, mà là kết quả của việc phân tích các khối lượng thông tin lớn. Ngành xây dựng dần dần không còn là nghệ thuật của trực giác mà trở thành một khoa học chính xác của dự đoán.

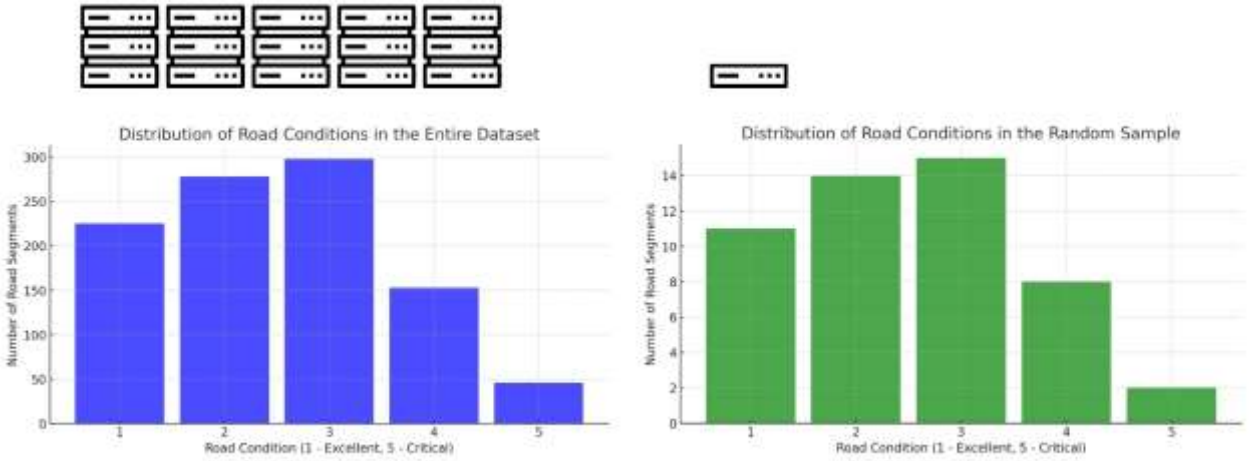
Việc chuyển sang ý tưởng sử dụng dữ liệu lớn không thể tránh khỏi đặt ra câu hỏi quan trọng: khối lượng dữ liệu có quan trọng đến mức nào và bao nhiêu thông tin thực sự cần thiết cho phân tích dự đoán đáng tin cậy? Quan niệm phổ biến rằng "càng nhiều dữ liệu - độ chính xác càng cao" thực tế không phải lúc nào cũng được chứng minh là hợp lý từ góc độ thống kê.

> Ngay từ năm 1934, nhà thống kê Jerzy Neyman đã chứng minh rằng chìa khóa cho độ chính xác của các kết luận thống kê không nằm ở khối lượng dữ liệu mà ở tính đại diện và ngẫu nhiên của mẫu.

Điều này đặc biệt quan trọng trong ngành xây dựng, nơi mà một khối lượng lớn dữ liệu được thu thập thông qua các cảm biến IoT, máy quét, camera giám sát, máy bay không người lái và thậm chí là các mô hình CAD đa dạng, điều này làm tăng nguy cơ xuất hiện "vùng mù", sự phát thải và biến dạng trong dữ liệu.

Xem xét ví dụ về việc giám sát tình trạng mặt đường. Tập dữ liệu đầy đủ về tất cả các đoạn đường có thể chiếm X GB và yêu cầu khoảng một ngày để xử lý. Trong khi đó, một mẫu ngẫu nhiên chỉ bao gồm mỗi đoạn đường thứ 50 sẽ chỉ chiếm X/50 GB và được xử lý trong nửa giờ, đồng thời cung cấp độ chính xác tương tự

cho một số tính toán nhất định (Hình 9.11).



Hình 9.11 Biểu đồ trạng thái mặt đường: tập dữ liệu đầy đủ và mẫu ngẫu nhiên cho thấy kết quả giống hệt nhau.

Do đó, yếu tố then chốt cho việc phân tích dữ liệu thành công thường không phải là khối lượng dữ liệu, mà là tính đại diện của mẫu và chất lượng của các phương pháp xử lý được áp dụng. Việc chuyển sang mẫu ngẫu nhiên và cách tiếp cận chọn lọc hơn đòi hỏi phải thay đổi tư duy trong ngành xây dựng. Lịch sử cho thấy các công ty thường tuân theo logic: "càng nhiều dữ liệu càng tốt", cho rằng việc bao quát tất cả các chỉ số có thể sẽ đảm bảo độ chính xác tối đa.

Cách tiếp cận này giống như một hiểu lầm phổ biến trong quản lý dự án: "càng nhiều chuyên gia tôi thu hút, công việc sẽ càng hiệu quả". Tuy nhiên, cũng như trong trường hợp nhân sự, chất lượng và công cụ quan trọng hơn số lượng. Nếu không xem xét các mối quan hệ (tương quan) giữa các dữ liệu hoặc các thành viên trong dự án, sự gia tăng khối lượng có thể chỉ dẫn đến tiếng ồn, sai lệch, trùng lặp và các giá trị ngoại lai không hợp lý.

Cuối cùng, thường thì việc có một tập dữ liệu nhỏ hơn nhưng được chuẩn bị chất lượng tốt sẽ mang lại dự đoán ổn định và có cơ sở hơn là dựa vào thông tin khổng lồ nhưng hỗn loạn, chứa nhiều tín hiệu mâu thuẫn.

> Khối lượng dữ liệu quá lớn không chỉ không đảm bảo độ chính xác cao hơn mà còn có thể dẫn đến những kết luận sai lệch - do sự hiện diện của tiếng ồn, các đặc trưng dư thừa, các tương quan ẩn và thông tin không liên quan. Trong những điều kiện như vậy, nguy cơ quá khớp mô hình tăng lên và độ tin cậy của kết quả phân tích giảm xuống.

Trong xây dựng, thách thức chính trong việc làm việc với dữ liệu lớn là xác định số lượng và chất lượng dữ liệu tối ưu. Ví dụ, trong việc giám sát tình trạng của các cấu trúc bê tông, việc sử dụng hàng ngàn cảm biến và thu thập thông tin mỗi phút có thể làm quá tải hệ thống lưu trữ và phân tích. Tuy nhiên, nếu thực hiện phân tích tương quan và chọn 10% cảm biến thông tin nhất, có thể đạt được độ chính xác dự đoán gần như giống hệt, tiêu tốn ít tài nguyên hơn nhiều, đôi khi là hàng chục và hàng trăm lần.

Việc sử dụng một tập con dữ liệu nhỏ hơn giảm thiểu cả khối lượng lưu trữ cần thiết và thời gian xử lý, điều này làm giảm đáng kể chi phí lưu trữ và phân tích dữ liệu và thường làm cho mẫu ngẫu nhiên trở thành giải pháp lý tưởng cho phân tích dự đoán, đặc biệt trong các dự án hạ tầng lớn hoặc khi làm việc trong thời gian thực. Cuối cùng, hiệu quả của các quy trình xây dựng không được xác định bởi khối lượng dữ liệu thu thập

được, mà là chất lượng của việc phân tích chúng. Nếu không có cách tiếp cận phê phán và phân tích kỹ lưỡng, dữ liệu có thể dẫn đến những kết luận sai lầm.

> Sau một khối lượng dữ liệu nhất định, mỗi đơn vị thông tin mới mang lại ít kết quả hữu ích hơn. Thay vì thu thập thông tin vô tận, quan trọng là tập trung vào tính đại diện của nó và các phương pháp phân tích (Hình 9.22).

Hiện tượng này được Allen Wallis mô tả rõ ràng, người đã minh họa việc sử dụng các phương pháp thống kê thông qua việc thử nghiệm hai thiết kế đạn dược thay thế của Hải quân Hoa Kỳ.

> *Hải quân đã thử nghiệm hai thiết kế đạn dược (A và B) bằng cách thực hiện một loạt các phát bắn cặp. Trong mỗi vòng, A nhận được 1 hoặc 0 tùy thuộc vào việc hiệu suất của nó tốt hơn hay kém hơn so với B, và ngược lại. Phương pháp thống kê tiêu chuẩn yêu cầu thực hiện một số lượng thử nghiệm cố định (ví dụ, 1000) và xác định người chiến thắng dựa trên phân phối phần trăm (ví dụ, nếu A nhận được 1 nhiều hơn 53% số lần, nó được coi là tốt hơn). Khi Allen Wallis thảo luận về vấn đề này với Đại tá Hải quân Garrett L. Shyler, Đại tá đã phản biện rằng thử nghiệm như vậy, theo lời kể của Allen, có thể trở nên vô ích. Nếu một sĩ quan pháo binh khôn ngoan và dày dạn kinh nghiệm như Shyler ở trong tình huống đó, ông sẽ nhận thấy sau vài trăm phát bắn rằng không cần phải hoàn thành thí nghiệm, vì hoặc phương pháp mới rõ ràng kém hơn, hoặc nó rõ ràng vượt trội hơn những gì mà họ hy vọng. Nhóm nghiên cứu thống kê chính phủ Hoa Kỳ tại Đại học Columbia, trong thời kỳ Thế chiến II.*

Nguyên tắc này được sử dụng rộng rãi trong nhiều lĩnh vực. Trong y tế, chẳng hạn, các nghiên cứu lâm sàng về các loại thuốc mới được thực hiện trên các mẫu ngẫu nhiên của bệnh nhân, cho phép thu được các kết quả có ý nghĩa thống kê mà không cần thử nghiệm thuốc trên toàn bộ dân số sống trên hành tinh. Trong kinh tế và xã hội học, các cuộc khảo sát đại diện được thực hiện để phản ánh ý kiến của xã hội mà không cần phải khảo sát từng cư dân của quốc gia.

Tương tự như cách mà các quốc gia và tổ chức nghiên cứu thực hiện khảo sát các nhóm dân cư nhỏ để hiểu các xu hướng xã hội chung, các công ty trong ngành xây dựng có thể sử dụng các mẫu ngẫu nhiên dữ liệu để theo dõi hiệu quả và tạo ra dự đoán cho việc quản lý dự án.

> *Dữ liệu lớn có thể thay đổi cách tiếp cận đối với các khoa học xã hội, nhưng không thể thay thế được sự hợp lý thống kê.*
>
> *- Thomas Landsall-Welfair, "Dự đoán tâm trạng của quốc gia hiện nay", Significance v. 9(4), năm 2012.*

Từ góc độ tiết kiệm tài nguyên, khi thu thập dữ liệu cho các dự đoán và quyết định trong tương lai, điều quan trọng là phải trả lời câu hỏi: có hợp lý khi chi tiêu một số tiền lớn cho việc thu thập và xử lý các khối dữ liệu khổng lồ, khi có thể sử dụng một bộ dữ liệu thử nghiệm nhỏ hơn và rẻ hơn, có thể mở rộng dần dần? Hiệu quả của các mẫu ngẫu nhiên cho thấy rằng các công ty có thể giảm chi phí hàng chục hoặc thậm chí hàng nghìn lần trong việc thu thập và đào tạo các mô hình, bằng cách chọn các phương pháp thu thập dữ liệu không yêu cầu phải bao phủ toàn diện, nhưng vẫn đảm bảo độ chính xác và tính đại diện đủ. Cách tiếp cận

này cho phép ngay cả những công ty nhỏ đạt được kết quả ở mức độ của các tập đoàn lớn, sử dụng tài nguyên và khối lượng dữ liệu nhỏ hơn nhiều, điều này rất quan trọng đối với các công ty đang cố gắng tối ưu hóa chi phí và tăng tốc quá trình ra quyết định hợp lý, sử dụng các nguồn lực nhỏ. Trong các chương tiếp theo, chúng ta sẽ xem xét các ví dụ về phân tích và dự đoán dựa trên các bộ dữ liệu công khai bằng cách sử dụng các công cụ dữ liệu lớn.

Dữ liệu lớn: phân tích dữ liệu bộ dữ liệu một triệu giấy phép xây dựng ở San Francisco

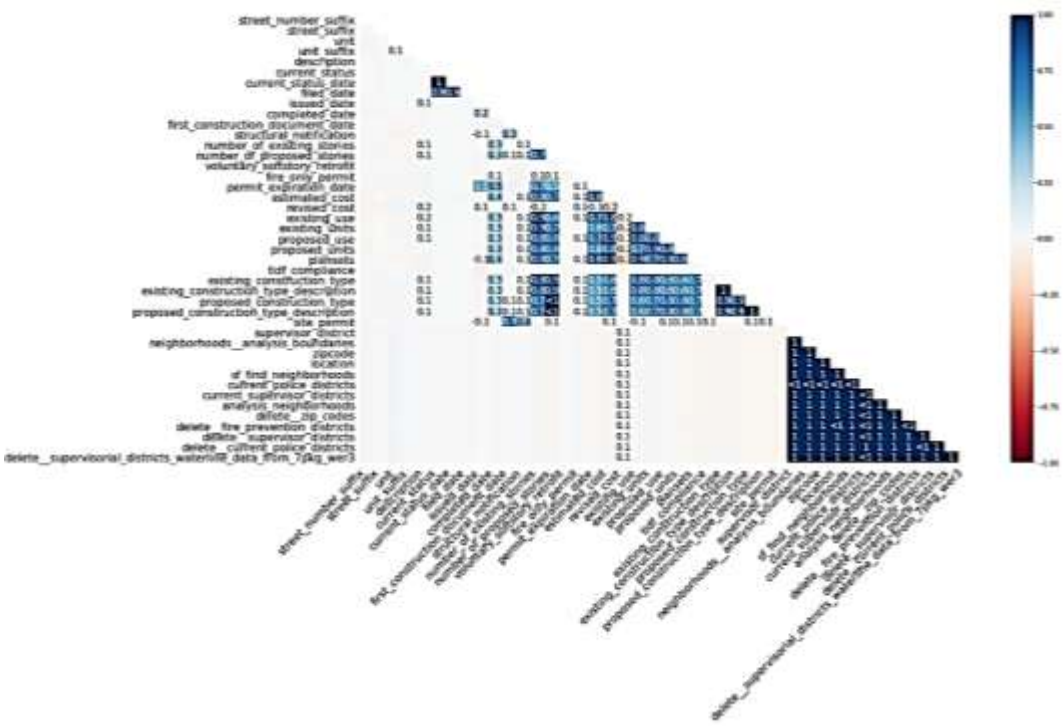
Làm việc với các bộ dữ liệu mở cung cấp cơ hội độc đáo để áp dụng vào thực tiễn các nguyên tắc đã được thảo luận trong các chương trước: lựa chọn đặc trưng hợp lý, mẫu đại diện, trực quan hóa và phân tích phản biện. Trong chương này, chúng ta sẽ xem xét cách nghiên cứu các hiện tượng phức tạp, chẳng hạn như hoạt động xây dựng ở một thành phố lớn, bằng cách sử dụng dữ liệu mở - cụ thể là hơn một triệu bản ghi về giấy phép xây dựng ở San Francisco.

Dữ liệu công khai về hơn một triệu giấy phép (Hình 9.12) xây dựng (bản ghi trong hai bộ dữ liệu định dạng CSV) từ "Sở Xây dựng San Francisco" cho phép chúng ta sử dụng bảng CSV thô để phân tích không chỉ hoạt động xây dựng trong thành phố mà còn để phân tích phản biện các xu hướng và lịch sử gần đây của ngành xây dựng San Francisco trong 40 năm qua, từ 1980 đến 2019.

Các ví dụ mã được sử dụng để tạo ra các trực quan hóa của bộ dữ liệu (Hình 9.13 - Hình 9.18), cũng như các đồ thị trực quan với mã, giải thích và bình luận có thể được tìm thấy trên nền tảng Kaggle với từ khóa "San Francisco. Ngành xây dựng 1980–2019." --



Hình 9.12 Các bộ dữ liệu chứa thông tin về các giấy phép xây dựng đã được cấp với các thuộc tính khác nhau của các đối tượng.

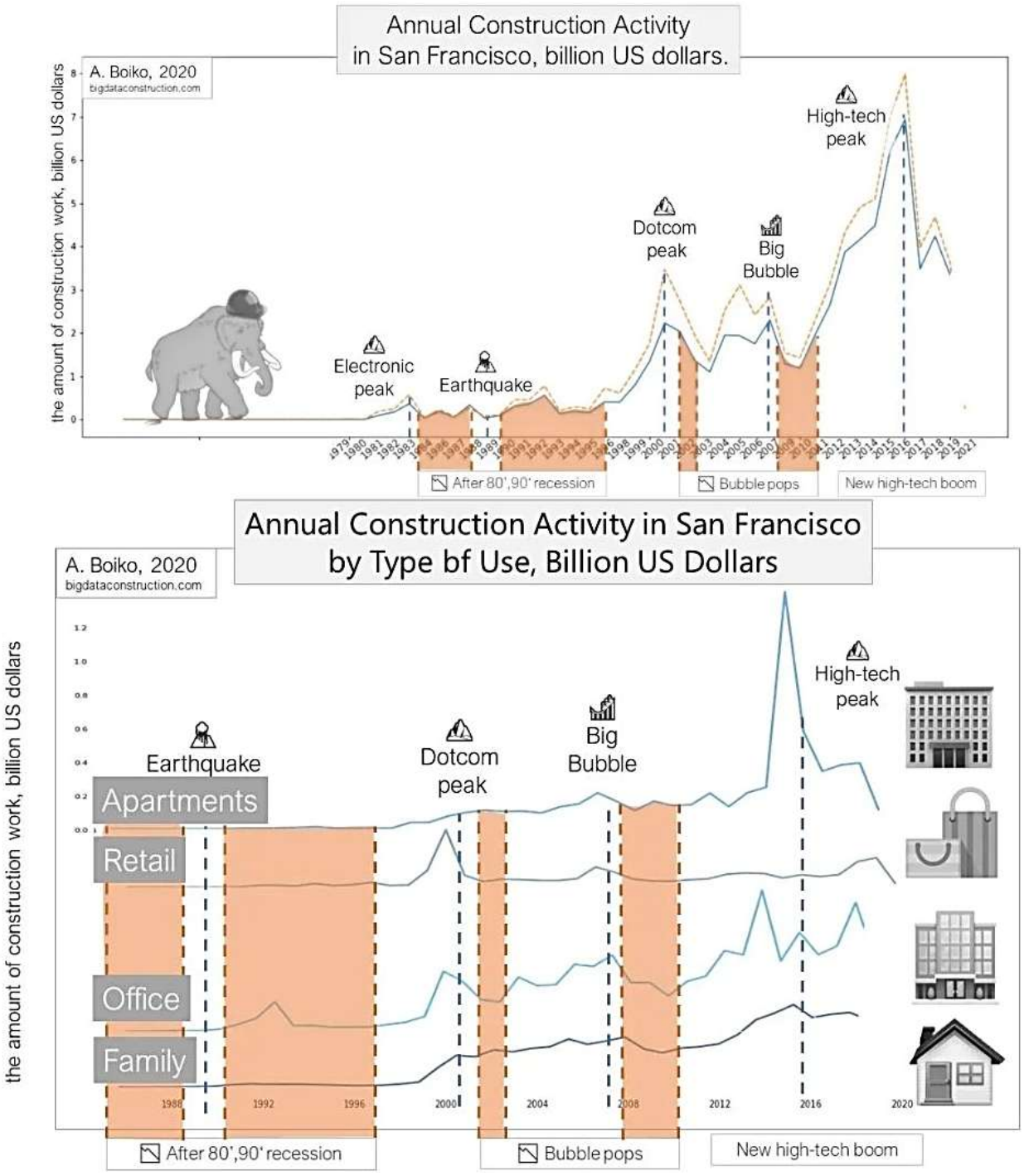


Hình 9.13 Bản đồ nhiệt (Pandas và Seaborn), trực quan hóa tất cả các thuộc tính của bộ dữ liệu và giúp phát hiện mối quan hệ giữa các cặp thuộc tính.

Từ bảng dữ liệu do Sở Xây dựng San Francisco cung cấp (Hình 9.12), không có xu hướng hay kết luận nào được nhìn thấy. Những con số khô khan trong dạng bảng không phải là cơ sở để đưa ra quyết định. Để làm cho dữ liệu trở nên dễ hiểu về mặt trực quan, như đã được thảo luận chi tiết trong các chương về trực quan hóa dữ liệu, chúng cần được trực quan hóa bằng cách sử dụng các thư viện khác nhau đã được đề cập trong phần bảy của cuốn sách về chủ đề “ETL và trực quan hóa kết quả dưới dạng đồ thị".

Phân tích dữ liệu cho thấy, với sự trợ giúp của Pandas DataFrame và các thư viện trực quan hóa Python, về giá trị của 1.137.695 giấy phép [148], có thể kết luận rằng hoạt động xây dựng ở San Francisco có mối liên hệ chặt chẽ với các chu kỳ kinh tế, đặc biệt là trong ngành công nghệ đang phát triển mạnh mẽ của Thung lũng Silicon (Hình 9.14). -

Các đợt bùng nổ và suy thoái kinh tế có ảnh hưởng đáng kể đến số lượng và giá trị của các dự án xây dựng. Ví dụ, đợt bùng nổ đầu tiên của hoạt động xây dựng trùng với cơn sốt điện tử vào giữa những năm 1980 (đã sử dụng Pandas và Matplotlib), trong khi các đợt bùng nổ và suy thoái tiếp theo liên quan đến bong bóng dot-com và cơn sốt công nghệ trong những năm gần đây.

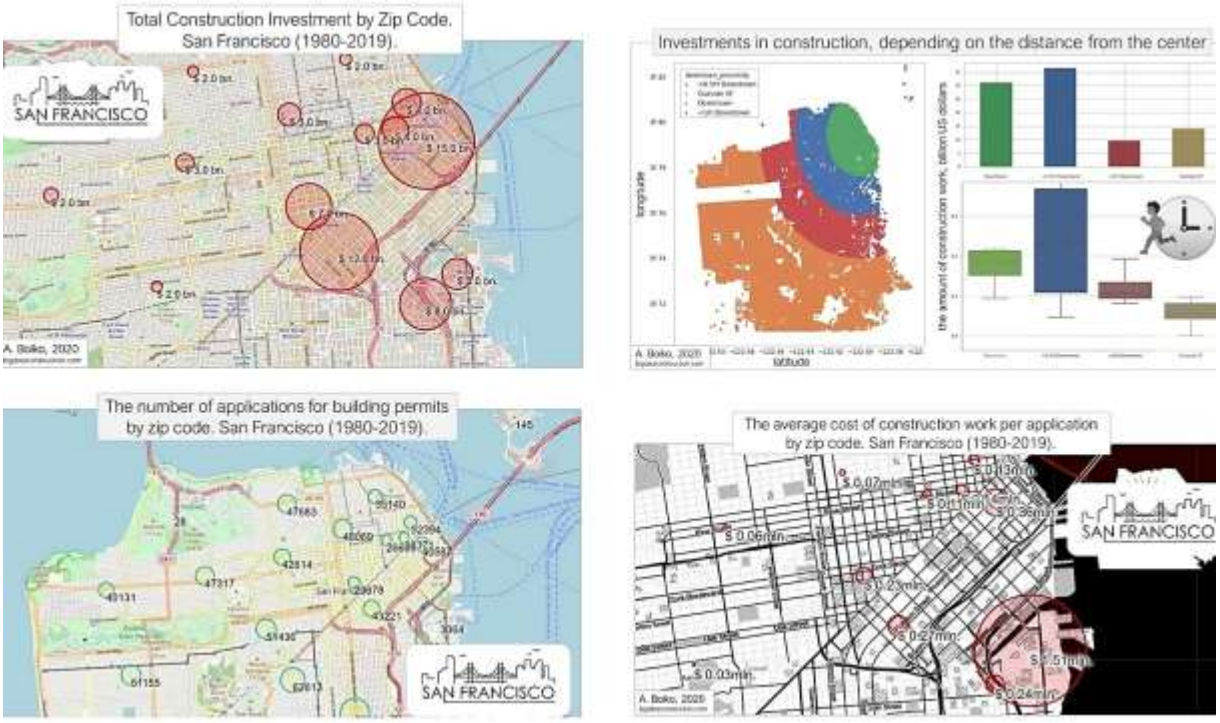


Hình 9.14 Trong lĩnh vực bất động sản, các khoản đầu tư ở San Francisco có mối tương quan với sự phát triển công nghệ của Thung lũng Silicon.

Phân tích dữ liệu cho thấy rằng ở San Francisco, phần lớn trong số 91,5 tỷ đô la được đầu tư vào xây dựng và cải tạo trong thập kỷ qua - gần 75% - tập trung ở trung tâm thành phố (Hình 9.15 - đã sử dụng Pandas và thư viện trực quan hóa Folium) và trong bán kính 2 km từ trung tâm thành phố, điều này phản ánh mật độ đầu tư cao hơn trong các khu vực trung tâm này.

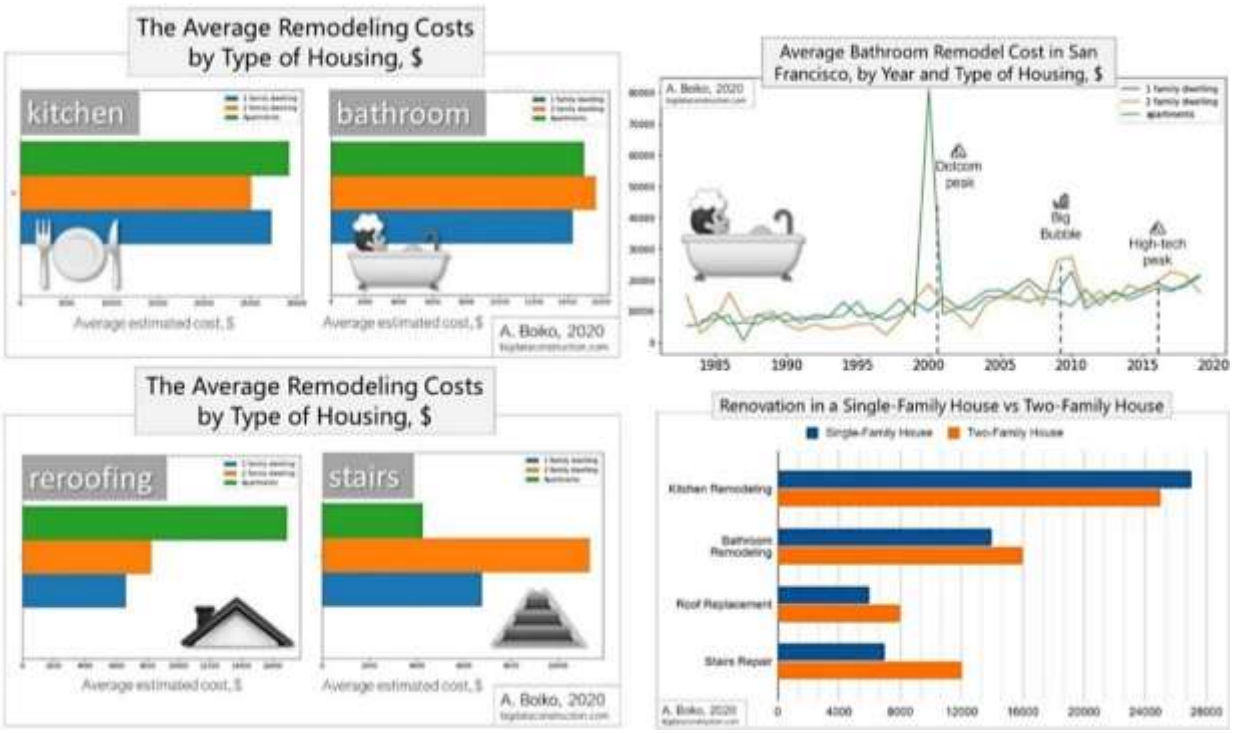
Chi phí trung bình cho giấy phép xây dựng thay đổi đáng kể tùy thuộc vào khu vực, trong đó các đơn xin ở

trung tâm thành phố có giá cao gấp ba lần so với bên ngoài, do chi phí đất đai, lao động, vật liệu cao hơn và các quy định xây dựng nghiêm ngặt yêu cầu sử dụng vật liệu đắt tiền hơn để nâng cao hiệu quả năng lượng.



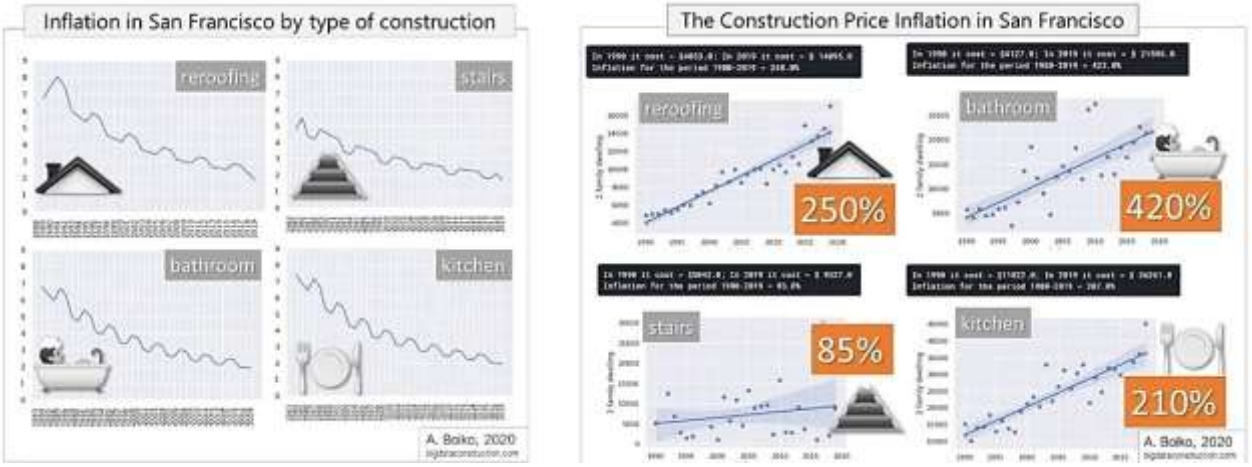
Hình 9.15 Tại San Francisco, 75% đầu tư vào xây dựng (91,5 tỷ đô la) tập trung ở trung tâm thành phố.

Tập dữ liệu cũng cho phép tính toán giá trung bình cho việc sửa chữa không chỉ theo loại nhà mà còn theo khu vực của thành phố và các địa chỉ cụ thể (mã bưu chính). Tại San Francisco, động thái giá sửa chữa nhà ở cho thấy những xu hướng rõ ràng cho các loại sửa chữa và nhà ở khác nhau (Hình 9.16 - đã sử dụng Pandas và Matplotlib). Sửa chữa bếp đắt hơn đáng kể so với sửa chữa phòng tắm: sửa chữa bếp trung bình trong một ngôi nhà đơn lập khoảng 28.000 đô la so với 25.000 đô la trong một ngôi nhà hai gia đình.



Hình 9.16 Tại SF, sửa chữa bếp tốn kém gần gấp đôi so với sửa chữa phòng tắm và các chủ nhà cần tiết kiệm 350 đô la mỗi tháng trong 15 năm để trang trải chi phí sửa chữa chính cho nhà ở.

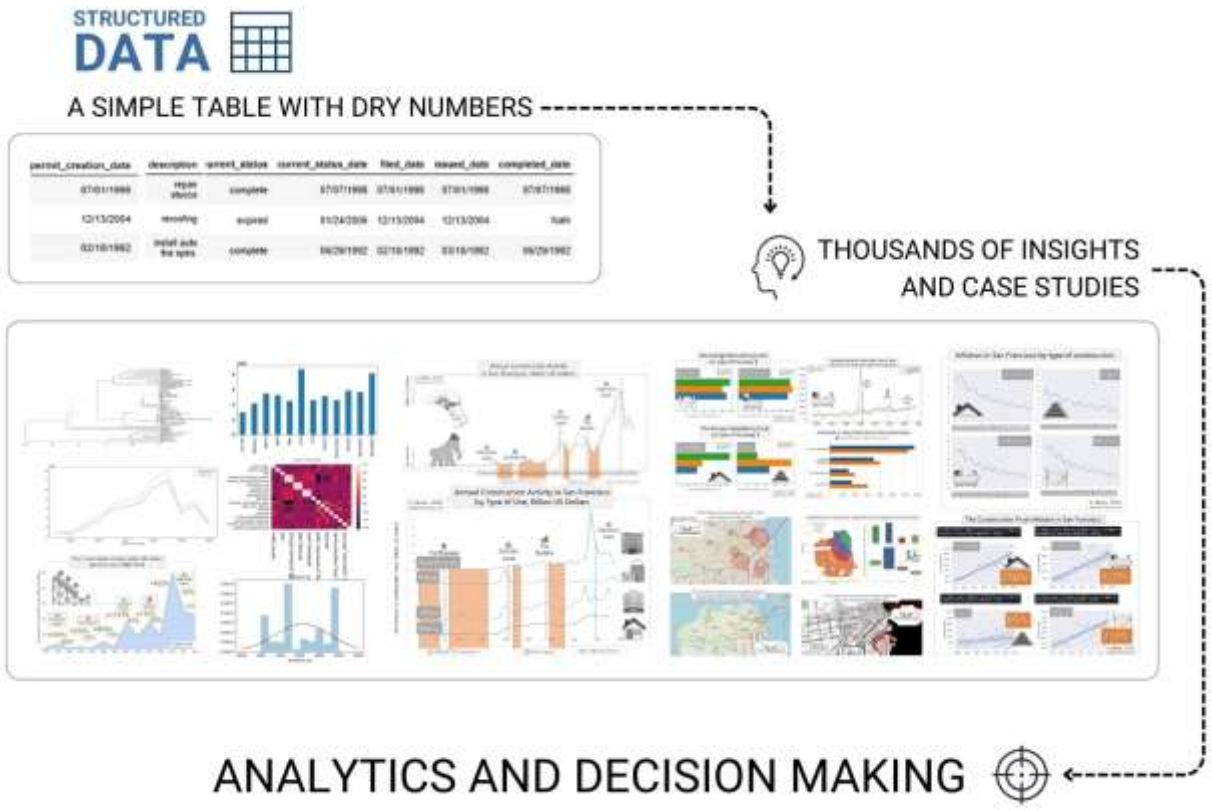
Có thể theo dõi lạm phát chi phí xây dựng ở San Francisco theo thời gian bằng cách phân tích dữ liệu được nhóm theo loại nhà và năm (Hình 9.17 - đã sử dụng Pandas và Seaborn), cho thấy sự gia tăng liên tục của chi phí sửa chữa trung bình từ năm 1990 và phát hiện ra các chu kỳ ngắn hạn ba năm trong chi phí sửa chữa nhà chung cư.



Hình 9.17 Từ năm 1980 đến 2019, chi phí sửa chữa phòng tắm ở SF đã tăng gấp năm lần, trong khi sửa chữa mái và bếp tăng gấp ba lần, còn sửa chữa cầu thang chỉ tăng 85%.

Nghiên cứu dữ liệu mở từ sở xây dựng San Francisco (Hình 9.13) cho thấy chi phí xây dựng trong thành phố cực kỳ biến động và thường không thể dự đoán, chịu ảnh hưởng của nhiều yếu tố. Trong số các yếu tố này có tăng trưởng kinh tế, đổi mới công nghệ và các yêu cầu độc đáo của các loại nhà khác nhau.

Trước đây, để thực hiện phân tích như vậy cần có kiến thức sâu về lập trình và phân tích. Tuy nhiên, với sự xuất hiện của các công cụ LLM, quy trình đã trở nên dễ tiếp cận và dễ hiểu cho một loạt các chuyên gia trong ngành xây dựng, từ kỹ sư trong các phòng thiết kế đến ban lãnh đạo cấp cao của các công ty.



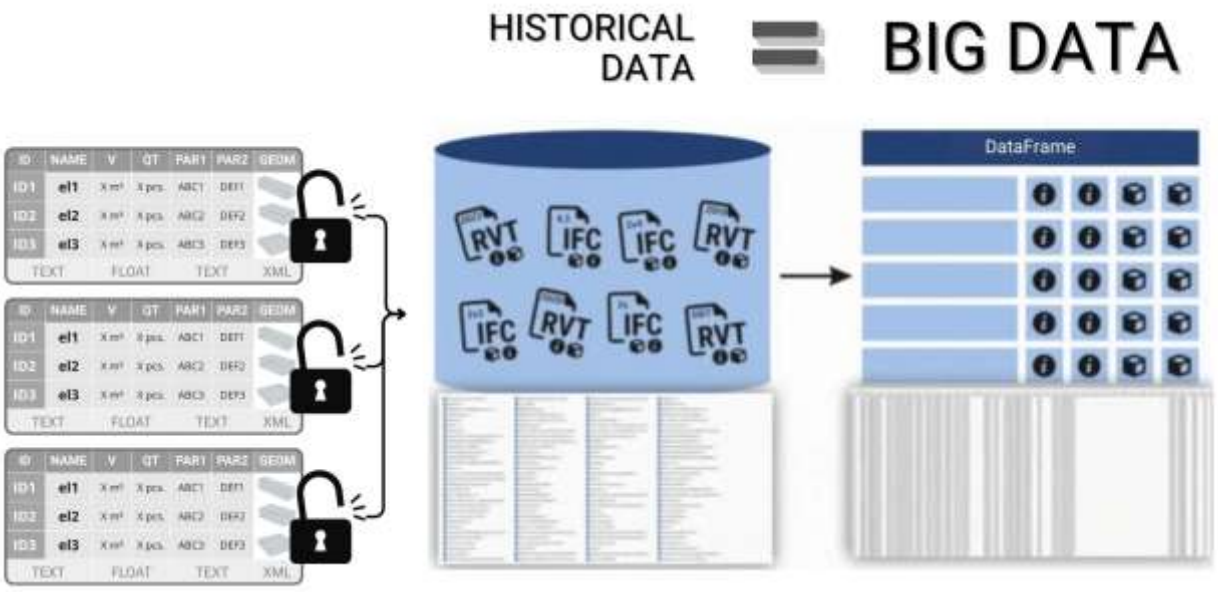
Hình 9.18 Việc chuyển sang dữ liệu trực quan giúp tự động hóa quá trình ra quyết định thông qua việc nhận diện các mẫu ẩn.

Tương tự như cách chúng ta phân tích dữ liệu từ tập dữ liệu "sở xây dựng San Francisco", chúng ta có thể trực quan hóa và phân tích bất kỳ tập dữ liệu nào - từ hình ảnh và tài liệu đến dữ liệu IoT, hoặc dữ liệu thu được từ cơ sở dữ liệu CAD.

Ví dụ về dữ liệu lớn dựa trên dữ liệu CAD (BIM)

Trong ví dụ tiếp theo, chúng ta sẽ phân tích một tập dữ liệu lớn, sử dụng dữ liệu từ các công cụ CAD (BIM) khác nhau. Để thu thập và tạo ra một tập dữ liệu lớn, một trình thu thập thông tin web tự động chuyên dụng (kịch bản) đã được sử dụng, được cấu hình để tự động tìm kiếm và thu thập các tệp dự án từ các trang web cung cấp mô hình kiến trúc miễn phí ở định dạng RVT và IFC. Trong vài ngày, trình thu thập thông tin đã thành công trong việc tìm thấy và tải xuống 4.596 tệp IFC và 6.471 tệp RVT cùng với 156.024 tệp DWG.

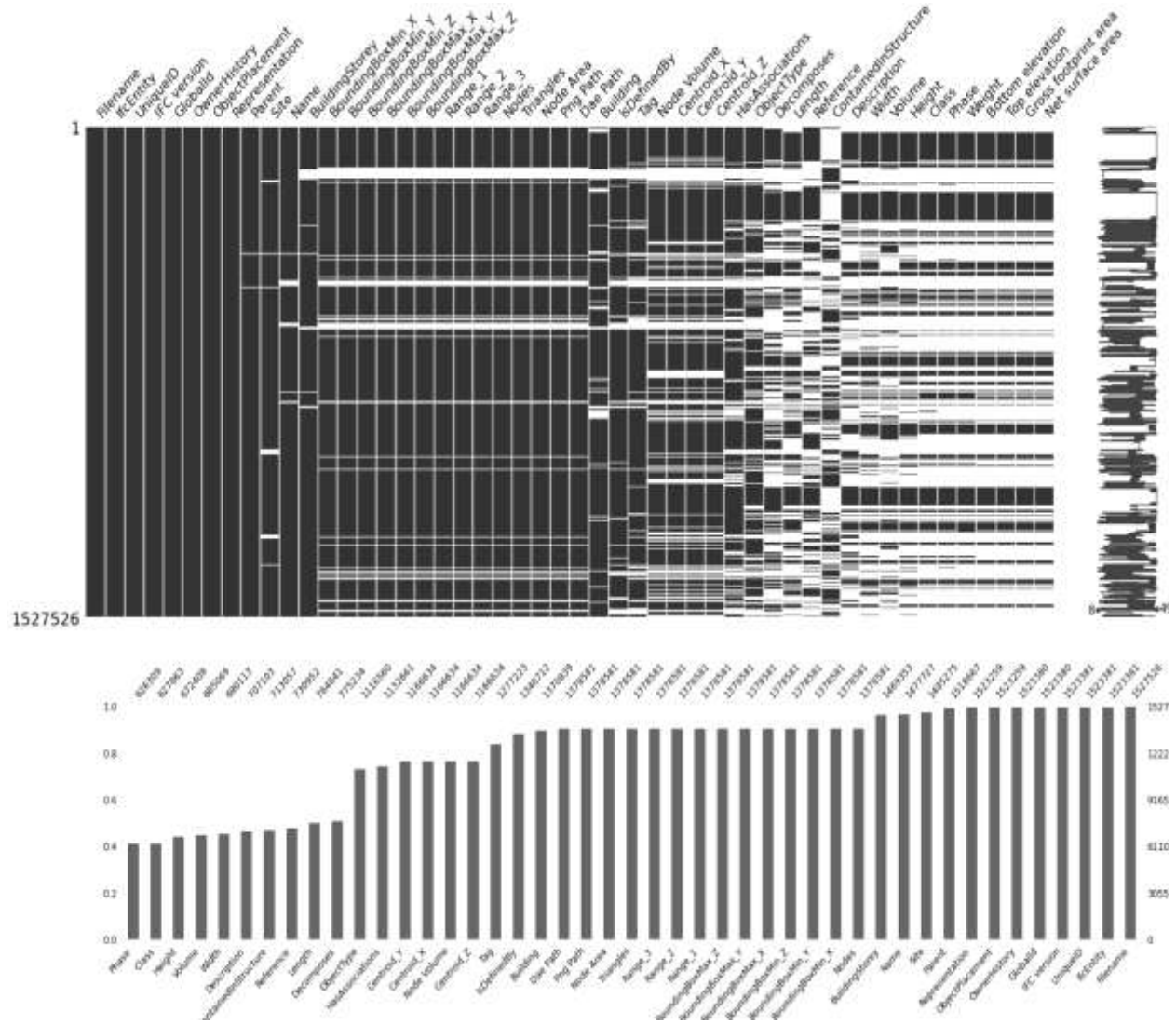
Sau khi thu thập các dự án ở định dạng RVT và IFC với nhiều phiên bản khác nhau và chuyển đổi chúng sang định dạng có cấu trúc CSV bằng cách sử dụng SDK miễn phí cho kỹ thuật đảo ngược, gần 10.000 dự án RVT và IFC đã được tập hợp thành một tệp bảng lớn duy nhất là Apache Parquet và được tải lên để phân tích trong Pandas DataFrame.-



Dữ liệu có cấu trúc của dự án cho phép kết hợp bất kỳ số lượng dự án nào thành một bảng hai chiều.

Dữ liệu từ bộ sưu tập quy mô lớn này chứa các thông tin sau: bộ tệp IFC chứa khoảng 4 triệu thực thể (dòng) và 24.962 thuộc tính (cột), trong khi bộ tệp RVT, bao gồm khoảng 6 triệu thực thể (dòng), chứa 27.025 thuộc tính khác nhau (cột).

Các tập thông tin này bao gồm hàng triệu yếu tố, cho mỗi yếu tố đã được thu thập thêm và thêm vào bảng tổng thể - tọa độ hình học của Bounding Box (hình chữ nhật xác định ranh giới của đối tượng trong dự án) và tạo ra hình ảnh của từng yếu tố ở định dạng PNG và hình học ở định dạng mở XML - DAE (Collada).

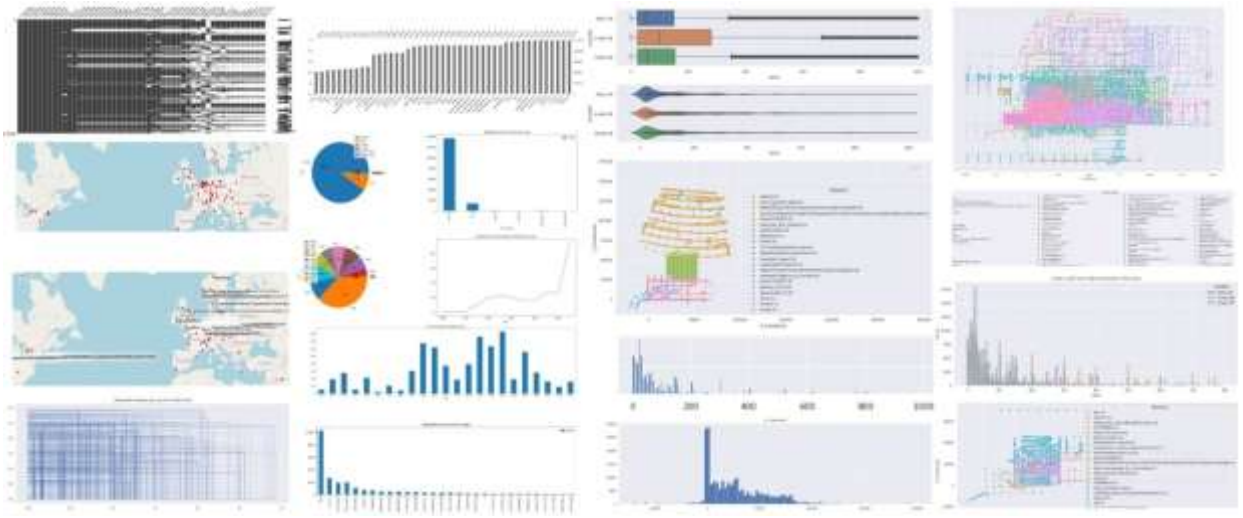


Một tập con gồm 1,5 triệu yếu tố và hình ảnh hóa (thư viện missingno) về độ đầy đủ của 100 thuộc tính đầu tiên dưới dạng biểu đồ.

Như vậy, chúng ta đã thu thập được toàn bộ thông tin về hàng chục triệu yếu tố từ 4.596 dự án IFC và 6.471 dự án RVT, nơi tất cả các thuộc tính - đặc điểm của tất cả các yếu tố - thực thể và hình học của chúng (Bounding Box) đã được chuyển đổi thành dạng có cấu trúc của một bảng (DataFrame).-

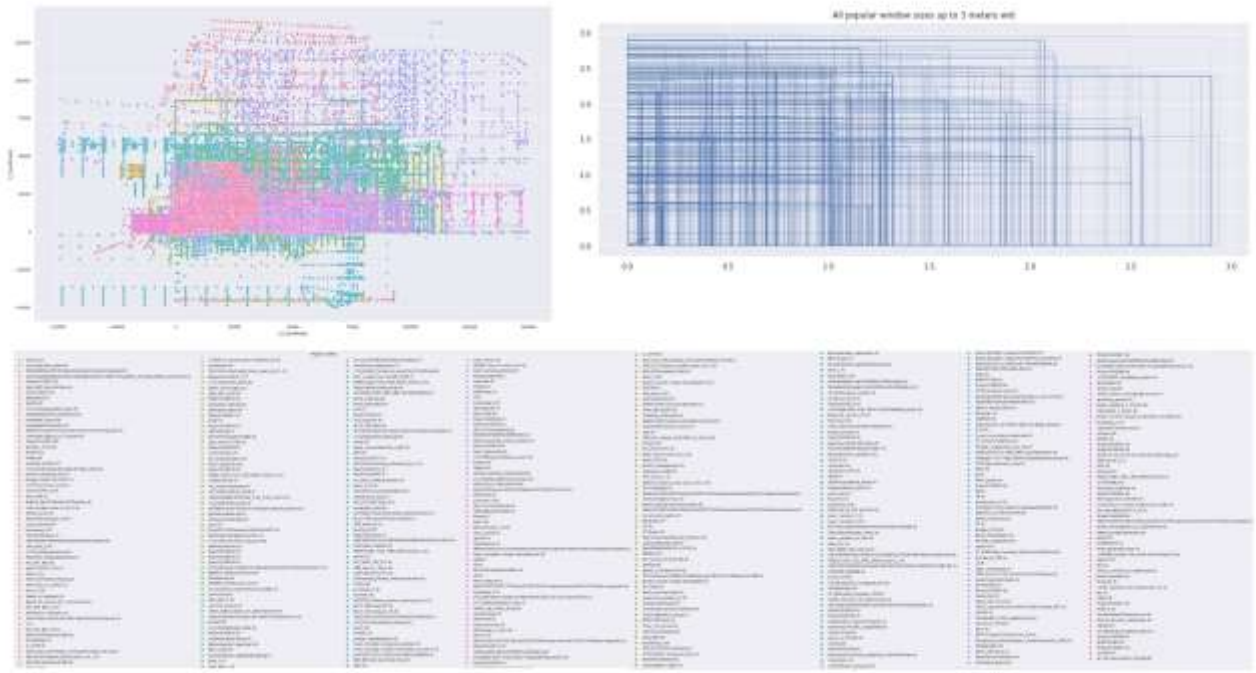
> Các biểu đồ (hình 9.110, hình 9.26, hình 9.27) được xây dựng trong quá trình phân tích cho phép đánh giá nhanh độ dày của dữ liệu và tần suất xuất hiện của các giá trị trong các cột. Điều này cung cấp cái nhìn đầu tiên về phân phối các đặc điểm, sự hiện diện của các giá trị thiếu và khả năng hữu ích tiềm năng của các thuộc tính riêng lẻ trong phân tích và xây dựng các mô hình học máy.-

Một trong những ví dụ về việc sử dụng thực tiễn của tập dữ liệu này là dự án "5000 dự án IFC và RVT", có sẵn trên nền tảng Kaggle. Trong đó, một Jupyter Notebook được trình bày với giải pháp Pipeline hoàn chỉnh: từ tiền xử lý và phân tích dữ liệu đến hình ảnh hóa kết quả bằng cách sử dụng các thư viện Python - pandas, matplotlib, seaborn, folium và các thư viện khác.-



Các ví dụ về phân tích dữ liệu từ các định dạng CAD (BIM) bằng cách sử dụng các thư viện hình ảnh hóa Python và thư viện pandas.

Dựa trên siêu thông tin, có thể xác định các thành phố nơi các dự án đã được phát triển và hiển thị điều này trên bản đồ (ví dụ, bằng cách sử dụng thư viện folium). Ngoài ra, các dấu thời gian trong dữ liệu cho phép khám phá các quy luật theo thời gian lưu trữ hoặc chỉnh sửa tệp: theo ngày trong tuần, thời gian trong ngày và tháng.



Hình 9.112 Hiển thị vị trí hình học của tất cả các cột và kích thước của tất cả các cửa sổ cao đến 3 mét trong các dự án từ danh sách ở dưới cùng của biểu đồ.

Các tham số hình học dưới dạng Bounding Box, được trích xuất từ các mô hình, cũng có thể được phân tích tổng hợp. Ví dụ, trong Hình 9.112 có hai biểu đồ: bên trái cho thấy phân bố khoảng cách giữa các cột trong tất cả các dự án so với điểm gốc, trong khi bên phải hiển thị kích thước của tất cả các cửa sổ cao đến 3 mét trong mẫu từ hàng chục nghìn phần tử cửa sổ (sau khi nhóm toàn bộ tập dữ liệu theo tham số "Category"

với giá trị "OST_Windows", "IfcWindows").

Mã phân tích Pipeline cho ví dụ này và bộ dữ liệu tự nó có sẵn trên trang web Kaggle với tên "5000 dự án IFC và RVT | DataDrivenConstruction.io" [149]. Pipeline hoàn chỉnh này cùng với bộ dữ liệu có thể được sao chép và chạy miễn phí trực tuyến trên Kaggle hoặc ngoại tuyến trong một trong những IDE phổ biến: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse với plugin PyDev, Thonny, Wing IDE, IntelliJ IDEA với plugin Python, JupyterLab hoặc các công cụ trực tuyến phổ biến như Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

> Dữ liệu phân tích, thu được từ việc xử lý và nghiên cứu khối lượng lớn dữ liệu có cấu trúc, sẽ đóng vai trò quyết định trong các quy trình ra quyết định trong ngành xây dựng.

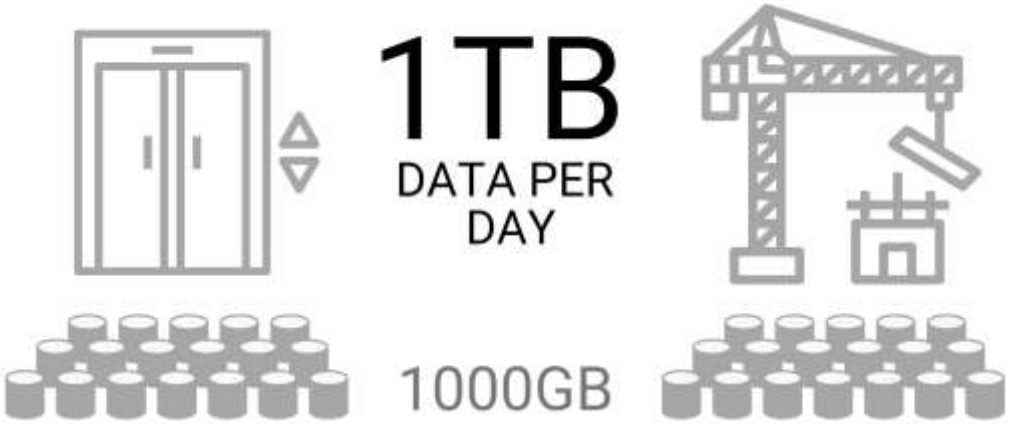
Nhờ vào việc phân tích thông tin dựa trên dữ liệu của các dự án trước, các chuyên gia có thể dự đoán hiệu quả nhu cầu về vật liệu và lao động, cũng như tối ưu hóa các giải pháp thiết kế ngay cả trước khi bắt đầu xây dựng.

Tuy nhiên, trong khi dữ liệu thiết kế hoặc giấy phép xây dựng là thông tin tương đối tĩnh, thay đổi tương đối chậm, thì chính quá trình xây dựng đang nhanh chóng được làm phong phú bởi nhiều cảm biến và thiết bị IoT: camera, hệ thống giám sát tự động, tất cả đều truyền dữ liệu theo thời gian thực - tất cả điều này biến công trường xây dựng thành một môi trường số động, nơi dữ liệu cần được phân tích theo thời gian thực.

IoT Internet of Things và hợp đồng thông minh.

IoT (Internet of Things) đại diện cho một làn sóng mới của chuyển đổi số, trong đó mỗi thiết bị nhận được một địa chỉ IP riêng và trở thành một phần của mạng toàn cầu. IoT là một khái niệm cho phép kết nối các đối tượng vật lý với internet để thu thập, xử lý và truyền dữ liệu. Trong xây dựng, điều này có nghĩa là khả năng kiểm soát các quy trình xây dựng theo thời gian thực, giảm thiểu lãng phí vật liệu, dự đoán sự hao mòn của thiết bị và tự động hóa việc ra quyết định.

> Theo bài viết của CFMA "Chuẩn bị cho tương lai với xây dựng kết nối" [150], trong thập kỷ tới, ngành xây dựng sẽ trải qua một cuộc chuyển đổi số quy mô lớn, đỉnh điểm là khái niệm Xây dựng Kết nối - một công trường xây dựng hoàn toàn tích hợp và tự động hóa.



Hình 9.113 Các thiết bị IoT hoặc thiết bị truyền dữ liệu trên công trường xây dựng có thể sản xuất và truyền tải hàng terabyte dữ liệu mỗi ngày.

Nền tảng xây dựng kỹ thuật số dự kiến rằng tất cả các yếu tố của quá trình xây dựng - từ lập kế hoạch và logistics đến thực hiện công việc và kiểm soát chất lượng trên công trường thông qua camera cố định và drone - sẽ được kết hợp thành một hệ sinh thái kỹ thuật số động duy nhất. Trước đó, trong phần bảy của cuốn sách, chúng ta đã xem xét khả năng của công cụ miễn phí và mã nguồn mở Apache NiFi, cho phép tổ chức xử lý dữ liệu theo thời gian thực - từ việc thu thập từ các nguồn khác nhau đến việc truyền tải vào kho lưu trữ hoặc nền tảng phân tích.

Dữ liệu về tiến độ xây dựng, tiêu thụ vật liệu, tình trạng thiết bị và an toàn sẽ được truyền tải theo thời gian thực vào các hệ thống phân tích. Điều này cho phép dự đoán các rủi ro tiềm ẩn, phản ứng kịp thời với các sai lệch và tối ưu hóa quy trình trên công trường. Các thành phần chính của nền tảng xây dựng kỹ thuật số bao gồm:-

- Cảm biến IoT - theo dõi các thông số môi trường, giám sát thiết bị xây dựng và kiểm soát điều kiện làm việc.
- Mô hình kỹ thuật số - các mô hình ảo của tòa nhà và cơ sở hạ tầng, cho phép dự đoán các sai lệch có thể xảy ra và ngăn chặn lỗi.
- Hệ thống logistics tự động - quản lý chuỗi cung ứng theo thời gian thực để giảm thiểu thời gian chết và chi phí.
- Các hệ thống xây dựng robot - sử dụng máy móc tự động để thực hiện các nhiệm vụ lặp đi lặp lại và nguy hiểm.

Robot hóa, việc sử dụng rộng rãi IoT và khái niệm nền tảng xây dựng "Connected Site (Construction)" không chỉ nâng cao hiệu quả và giảm chi phí, mà còn mở ra một kỷ nguyên mới về an toàn, xây dựng bền vững và quản lý dự án dự đoán.

> Một trong những thành phần quan trọng của các thành phần IoT cũng là các thẻ RFID (Nhận dạng tần số vô tuyến). Chúng được sử dụng để xác định và theo dõi vật liệu, thiết bị và thậm chí cả nhân viên trên công trường, nâng cao tính minh bạch và khả năng quản lý tài nguyên của dự án.

Công nghệ RFID được sử dụng để nhận diện tự động các đối tượng thông qua tín hiệu vô tuyến. Nó bao gồm ba thành phần chính:

- Thẻ RFID (thụ động hoặc chủ động) - chứa một mã định danh duy nhất và được gắn lên vật liệu, công cụ hoặc thiết bị.
- Máy quét - thiết bị đọc thông tin từ các thẻ và truyền tải nó vào hệ thống.
- Cơ sở dữ liệu tập trung - lưu trữ thông tin về vị trí, tình trạng và chuyển động của các đối tượng.

Ứng dụng của RFID trong xây dựng:

- Kiểm kê tự động vật liệu - các thẻ trên các sản phẩm bê tông hoàn thiện, thép gia cố hoặc các gói panel sandwich cho phép kiểm soát tồn kho và ngăn chặn trộm cắp.
- Kiểm soát công việc của nhân viên - thẻ RFID của nhân viên ghi lại thời gian bắt đầu và kết thúc ca làm việc, đảm bảo việc theo dõi thời gian làm việc.
- Giám sát thiết bị - hệ thống RFID theo dõi sự di chuyển của thiết bị, ngăn chặn thời gian chết và nâng cao hiệu quả logistics.

Bổ sung cho bộ công nghệ này là các hợp đồng thông minh dựa trên công nghệ blockchain, cho phép tự động hóa thanh toán, kiểm soát cung ứng và tuân thủ các điều kiện hợp đồng mà không cần trung gian, giảm thiểu rủi ro gian lận và chậm trễ.

> Hôm nay, khi không có một mô hình dữ liệu thống nhất, hợp đồng thông minh chỉ đơn thuần là mã mà các bên tham gia đồng ý. Tuy nhiên, với cách tiếp cận dựa trên dữ liệu, có thể tạo ra một mô hình chung cho các tham số của hợp đồng, mã hóa nó vào blockchain và tự động hóa việc thực hiện các điều kiện.

Chẳng hạn, trong hệ thống quản lý chuỗi cung ứng, hợp đồng thông minh có thể theo dõi việc giao hàng thông qua các cảm biến IoT và thẻ RFID, và tự động chuyển tiền khi hàng đến nơi. Tương tự, trên công trường xây dựng, hợp đồng thông minh có thể ghi nhận việc hoàn thành một giai đoạn công việc — chẳng hạn như lắp đặt cốt thép hoặc đổ bê tông — dựa trên dữ liệu từ drone hoặc cảm biến xây dựng và tự động khởi động thanh toán tiếp theo cho nhà thầu mà không cần kiểm tra thủ công và các tài liệu giấy.

Nhưng bất chấp các công nghệ mới và nỗ lực của các tổ chức quốc tế trong việc tiêu chuẩn hóa, nhiều tiêu chuẩn cạnh tranh làm phức tạp bối cảnh IoT.

> Theo một nghiên cứu của Cisco được công bố vào năm 2017, gần 60% sáng kiến trong lĩnh vực Internet of Things (IoT) dừng lại ở giai đoạn chứng minh khái niệm, và chỉ 26% công ty coi các dự án IoT của họ hoàn toàn thành công. Hơn nữa, một phần ba các dự án đã hoàn thành không đạt được mục tiêu đã đề ra và không được công nhận là thành công ngay cả sau khi triển khai.

Một trong những lý do chính là sự thiếu tương thích giữa các nền tảng xử lý dữ liệu từ các cảm biến khác nhau. Kết quả là, dữ liệu vẫn bị cô lập trong các giải pháp riêng lẻ. Giải pháp thay thế cho cách tiếp cận này, giống như trong các trường hợp tương tự (mà chúng tôi đã xem xét trong cuốn sách này), là kiến trúc được xây dựng xung quanh chính dữ liệu như một tài sản chính.

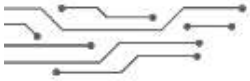
> Các cảm biến IoT đóng vai trò quan trọng không chỉ trong việc giám sát tình trạng kỹ thuật của thiết bị mà còn trong phân tích dự đoán, cho phép giảm thiểu rủi ro trên công trường xây dựng và nâng cao hiệu suất tổng thể của các quy trình, nhờ vào việc dự đoán các sự cố và sai lệch.

Dữ liệu được thu thập thông qua các cảm biến IoT và thẻ RFID có thể được xử lý theo thời gian thực bởi các

thuật toán học máy, có khả năng phát hiện các bất thường và thông báo trước cho kỹ sư về các sự cố tiềm ẩn. Điều này có thể là sự xuất hiện của các vết nứt nhỏ trong các cấu trúc bê tông, cũng như các khoảng dừng không điển hình trong hoạt động của cần cẩu tháp, cho thấy các sự cố kỹ thuật hoặc vi phạm quy định. Hơn nữa, các thuật toán phân tích hành vi tiên tiến cho phép ghi nhận các mẫu hành vi, có thể cho thấy, chẳng hạn, sự mệt mỏi về thể chất của nhân viên, nâng cao mức độ quản lý pro-active về an toàn và phúc lợi của nhân viên trên công trường.

Trong ngành xây dựng, tai nạn và sự cố — dù là thiết bị hay con người — hiếm khi xảy ra đột ngột. Thông thường, chúng được dẫn dắt bởi những sai lệch nhỏ mà không được chú ý. Phân tích dự đoán và học máy cho phép phát hiện những tín hiệu này ở giai đoạn sớm, trước khi xảy ra các hậu quả nghiêm trọng.

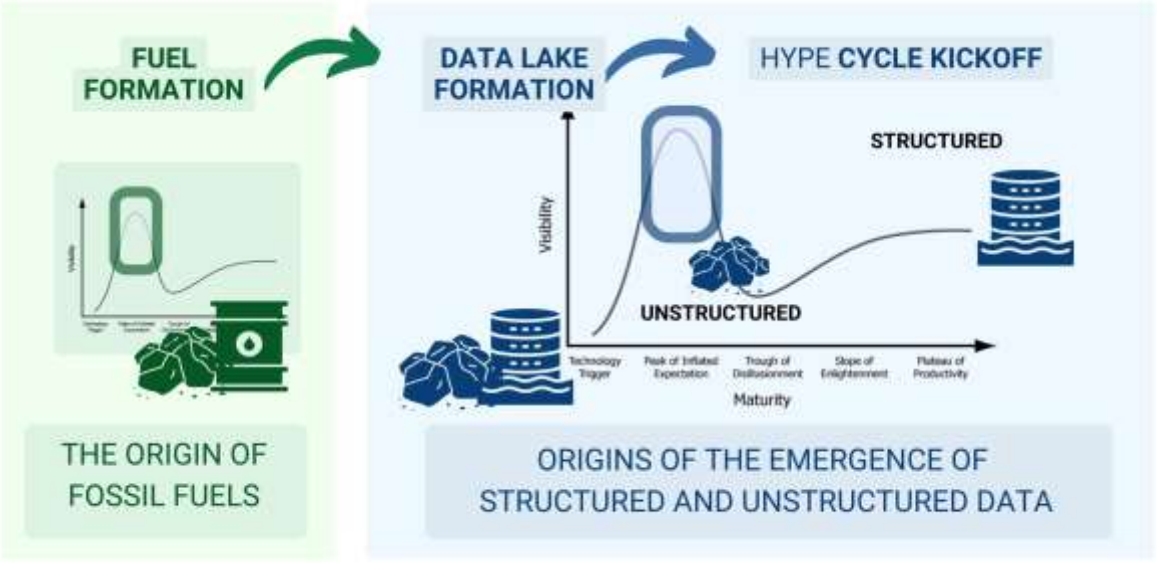
Nếu tài liệu, tệp dự án và dữ liệu từ các thiết bị IoT và thẻ RFID tạo thành dấu vết số của các công trình xây dựng, thì học máy giúp khai thác những kiến thức hữu ích từ đó. Với sự gia tăng khối lượng dữ liệu và sự dân chủ hóa quyền truy cập vào dữ liệu, ngành xây dựng đang nhận được những cơ hội mới trong lĩnh vực phân tích, dự đoán và ứng dụng trí tuệ nhân tạo.



CHƯƠNG 9.2. **HỌC MÁY VÀ DỰ ĐOÁN**

Học máy và trí tuệ nhân tạo sẽ thay đổi cách chúng ta xây dựng.

Cơ sở dữ liệu của các hệ thống khác nhau trong ngành xây dựng - với cơ sở hạ tầng không thể tránh khỏi sự lão hóa và phức tạp hóa - trở thành môi trường màu mỡ cho các giải pháp tương lai. Các máy chủ của công ty, giống như một khu rừng, chứa đựng khối lượng thông tin quan trọng, thường bị ẩn giấu dưới lòng đất, trong các thư mục và máy chủ. Khối lượng dữ liệu từ các hệ thống khác nhau được tạo ra ngày hôm nay - sau khi sử dụng, rơi xuống đáy máy chủ và sau nhiều năm hóa thạch - sẽ trở thành nhiên liệu cho các mô hình học máy và mô hình ngôn ngữ trong tương lai. Trên các mô hình nội bộ này sử dụng kho lưu trữ tập trung sẽ được xây dựng các cuộc trò chuyện nội bộ của công ty (ví dụ, một phiên bản ChatGPT, LlaMa, Mistral, DeepSeek được cấu hình cục bộ), cho phép nhanh chóng và thuận tiện truy xuất thông tin và tạo ra các biểu đồ, bảng điều khiển và tài liệu cần thiết.



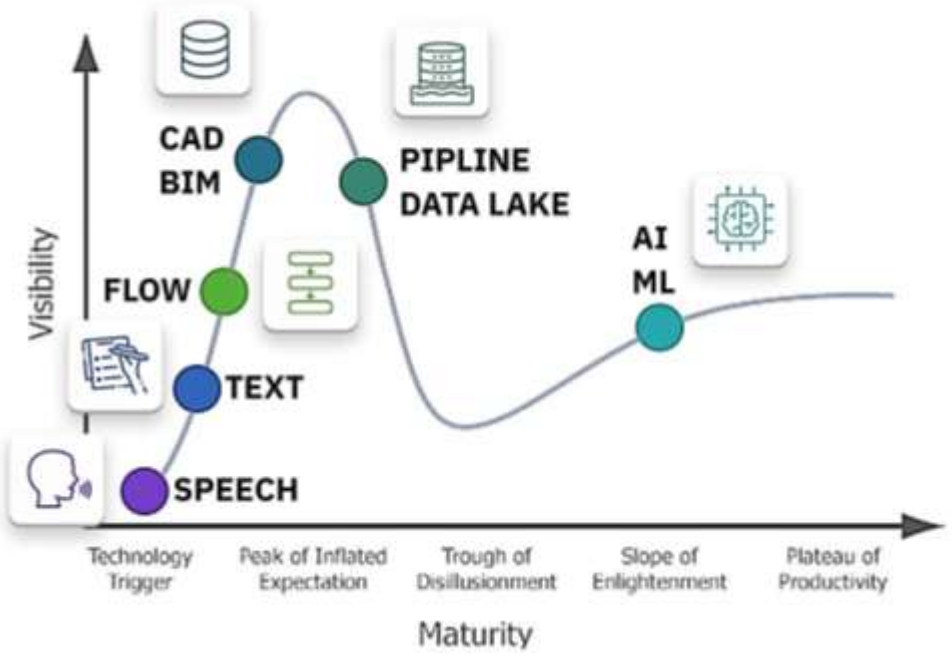
Hình 9.21 Như cây cối biến thành than, thông tin theo thời gian dưới áp lực của thời gian và các nhà phân tích trở thành nguồn năng lượng quý giá cho doanh nghiệp.

Sự hóa thạch của khối lượng thực vật kết hợp với áp lực và nhiệt độ tạo ra một khối đồng nhất và có cấu trúc độc đáo của các loại cây khác nhau, sống vào những thời điểm khác nhau - than gỗ. Tương tự, thông tin được ghi lại trên các ổ cứng ở các định dạng khác nhau và vào những thời điểm khác nhau dưới áp lực của các phòng phân tích và nhiệt độ của quản lý chất lượng, cuối cùng tạo thành một khối thông tin có cấu trúc đồng nhất và quý giá (Hình 9.21).

Những lớp (hoặc thường là những viên kim cương cô lập) thông tin này được tạo ra thông qua công việc tỉ mỉ của các nhà phân tích có kinh nghiệm, những người bắt đầu dần dần khai thác thông tin quý giá từ những dữ liệu dường như đã lỗi thời.

Vào thời điểm mà những lớp dữ liệu đã chín này không còn chỉ "cháy" trong các báo cáo, mà bắt đầu lưu thông trong các quy trình kinh doanh, làm phong phú thêm các quyết định và cải thiện quy trình, công ty trở nên sẵn sàng cho bước tiếp theo - chuyển sang học máy và trí tuệ nhân tạo (Hình 9.22).

Học máy (ML - Machine Learning) là một lớp các phương pháp để giải quyết các nhiệm vụ của trí tuệ nhân tạo. Các thuật toán học máy nhận diện các quy luật trong các khối dữ liệu lớn và sử dụng chúng để tự học. Mỗi bộ dữ liệu mới cho phép các thuật toán toán học cải thiện và thích ứng theo thông tin nhận được, điều này cho phép liên tục nâng cao độ chính xác của các khuyến nghị và dự đoán.



Hình 9.22 Sự suy giảm công nghệ tạo dữ liệu và việc áp dụng các công cụ phân tích mở ra con đường cho chủ đề học máy.

Như một giám đốc điều hành có ảnh hưởng của quỹ đầu tư lớn nhất thế giới đã nói trong một cuộc phỏng vấn năm 2023 (quỹ này sở hữu các cổ phần chủ chốt của hầu hết các công ty phần mềm xây dựng lớn nhất, cũng như các công ty sở hữu nhiều bất động sản nhất thế giới) - học máy sẽ thay đổi thế giới xây dựng.

> *Trí tuệ nhân tạo có tiềm năng to lớn. Nó sẽ thay đổi cách chúng ta làm việc, cách chúng ta sống. Trí tuệ nhân tạo và robot sẽ thay đổi cách chúng ta làm việc và xây dựng, và chúng ta có thể sử dụng trí tuệ nhân tạo và robot như một công cụ để tạo ra năng suất lớn hơn nhiều.- Giám đốc điều hành của quỹ đầu tư lớn nhất thế giới, phỏng vấn, tháng 9 năm 2023.*

Học máy (ML) hoạt động nhờ vào việc xử lý khối lượng lớn dữ liệu, sử dụng các phương pháp thống kê để mô phỏng các khía cạnh của tư duy con người. Tuy nhiên, hầu hết các công ty không có những bộ dữ liệu như vậy, và nếu có, thì thường không được gán nhãn đầy đủ. Các công nghệ ngữ nghĩa và học chuyển giao có thể giúp ích ở đây - một phương pháp cho phép ML hoạt động hiệu quả hơn khi làm việc với khối lượng dữ liệu nhỏ, tính khả thi của nó đã được thảo luận trong các chương trước của phần này.

Cốt lõi của học chuyển giao là thay vì xử lý mỗi nhiệm vụ từ đầu, ta có thể sử dụng kiến thức thu được từ các lĩnh vực liên quan. Cần hiểu rằng các mẫu và phát hiện từ các lĩnh vực kinh tế khác có thể được điều

chỉnh và áp dụng trong ngành xây dựng. Ví dụ, các phương pháp tối ưu hóa quy trình logistics được phát triển trong ngành bán lẻ giúp nâng cao hiệu quả quản lý chuỗi cung ứng trong xây dựng. Phân tích dữ liệu lớn, được sử dụng tích cực trong tài chính, có thể được áp dụng để dự đoán chi phí và quản lý rủi ro trong các dự án xây dựng. Các công nghệ thị giác máy tính và robot, đang phát triển trong ngành công nghiệp, đã tìm thấy ứng dụng trong việc kiểm soát chất lượng tự động, giám sát an toàn và quản lý các đối tượng trên công trường xây dựng.

> Học chuyển giao không chỉ giúp tăng tốc độ triển khai đổi mới mà còn giảm chi phí phát triển, sử dụng kinh nghiệm đã tích lũy từ các lĩnh vực khác.



Hình 9.23 Các công nghệ trí tuệ nhân tạo và robot sẽ trở thành động lực chính cho tương lai nhằm nâng cao năng suất trong ngành xây dựng.

Tư duy con người được tổ chức theo nguyên tắc tương tự: chúng ta dựa vào kiến thức đã có để giải quyết các nhiệm vụ mới. Trong học máy, phương pháp này cũng hoạt động - bằng cách đơn giản hóa mô hình dữ liệu và làm cho nó trở nên tinh tế hơn, ta có thể giảm độ phức tạp của nhiệm vụ cho các thuật toán ML. Điều này, ngược lại, giảm nhu cầu về khối lượng dữ liệu lớn và cắt giảm chi phí tính toán.--

Từ đánh giá chủ quan đến dự đoán thống kê

Thời đại mà các quyết định chiến lược phụ thuộc vào trực giác của một số lãnh đạo (Hình 9.24) đang dần qua đi. Trong bối cảnh cạnh tranh ngày càng tăng và điều kiện kinh tế phức tạp, cách tiếp cận chủ quan trở nên quá rủi ro và không hiệu quả. Các công ty tiếp tục dựa vào ý kiến cá nhân thay vì phân tích dữ liệu khách quan sẽ mất khả năng phản ứng kịp thời với những thay đổi.

> Môi trường cạnh tranh yêu cầu độ chính xác và khả năng tái sản xuất, dựa trên dữ liệu, quy luật thống kê và xác suất có thể tính toán. Các quyết định không thể chỉ dựa vào cảm giác, mà phải dựa vào các mối tương quan, xu hướng và mô hình dự đoán được thu thập thông qua phân tích và học máy. Đây không chỉ là sự thay đổi công cụ - mà là sự thay đổi trong logic tư duy: từ giả định - đến bằng chứng, từ xác suất chủ quan - đến độ lệch được tính toán thống kê, từ cảm giác - đến sự thật.



Hình 9.24 Thời đại quyết định được đưa ra bởi HiPPO (ý kiến của nhân viên được trả lương cao nhất) với sự xuất hiện của dữ liệu lớn và học máy sẽ đi vào quá khứ.

Các nhà lãnh đạo, những người quen dựa vào cảm giác cá nhân, sẽ không thể tránh khỏi thực tế mới: uy tín không còn quyết định sự lựa chọn. Trung tâm quản lý giờ đây là các hệ thống phân tích hàng triệu tham số và vector, phát hiện các quy luật tiềm ẩn và đề xuất các chiến lược tối ưu.

Nguyên nhân chính khiến các công ty ngày nay vẫn tránh xa việc triển khai ML là tính không minh bạch của nó. Hầu hết các mô hình đối với các nhà quản lý hoạt động như "hộp đen", không giải thích cách mà chúng đưa ra kết luận. Điều này dẫn đến các vấn đề: các thuật toán có thể củng cố các khuôn mẫu và thậm chí tạo ra những tình huống hài hước, như trong trường hợp của chatbot Microsoft, nhanh chóng trở thành một công cụ giao tiếp độc hại.

> Trong cuốn sách "Deep Thinking", Garry Kasparov, cựu vô địch thế giới cờ vua, suy ngẫm về thất bại của mình trước máy tính IBM Big Blue. Ông cho rằng giá trị thực sự của AI không nằm ở việc sao chép trí thông minh của con người, mà là bổ sung cho khả năng của chúng ta. AI nên thực hiện các nhiệm vụ mà con người yếu kém, trong khi con người mang lại sự sáng tạo. Máy tính đã thay đổi cách tiếp cận truyền thống trong phân tích cờ vua. Thay vì tạo ra những câu chuyện hấp dẫn về các ván cờ, các chương trình cờ vua máy tính đánh giá từng nước đi một cách khách quan, chỉ dựa trên sức mạnh hoặc điểm yếu thực tế của nó. Kasparov lưu ý rằng xu hướng của con người trong việc nhìn nhận các sự kiện như những câu chuyện liên kết, chứ không phải là các hành động riêng lẻ, thường dẫn đến những kết luận sai lầm - không chỉ trong cờ vua mà còn trong cuộc sống nói chung.

Do đó, nếu bạn dự định sử dụng học máy để dự đoán và phân tích, điều quan trọng là phải hiểu các nguyên tắc cơ bản của nó - cách mà các thuật toán hoạt động và cách mà dữ liệu được xử lý, trước khi bắt đầu sử dụng các công cụ học máy và AI trong công việc của mình. Cách tốt nhất để bắt đầu là thông qua kinh nghiệm thực hành.

Một trong những công cụ tiện lợi nhất để làm quen với chủ đề học máy và dự đoán là Jupyter Notebook và

bộ dữ liệu cổ điển Titanic phổ biến, cho phép bạn nắm vững các phương pháp phân tích dữ liệu và xây dựng mô hình ML một cách trực quan.

Titanic dataset: Xin chào thế giới trong lĩnh vực phân tích dữ liệu và dữ liệu lớn

Một trong những ví dụ nổi tiếng nhất về việc sử dụng ML trong phân tích dữ liệu là phân tích bộ dữ liệu "Titanic", thường được sử dụng để nghiên cứu xác suất sống sót của hành khách. Việc nghiên cứu bảng dữ liệu này tương tự như chương trình "Hello World" khi học các ngôn ngữ lập trình.

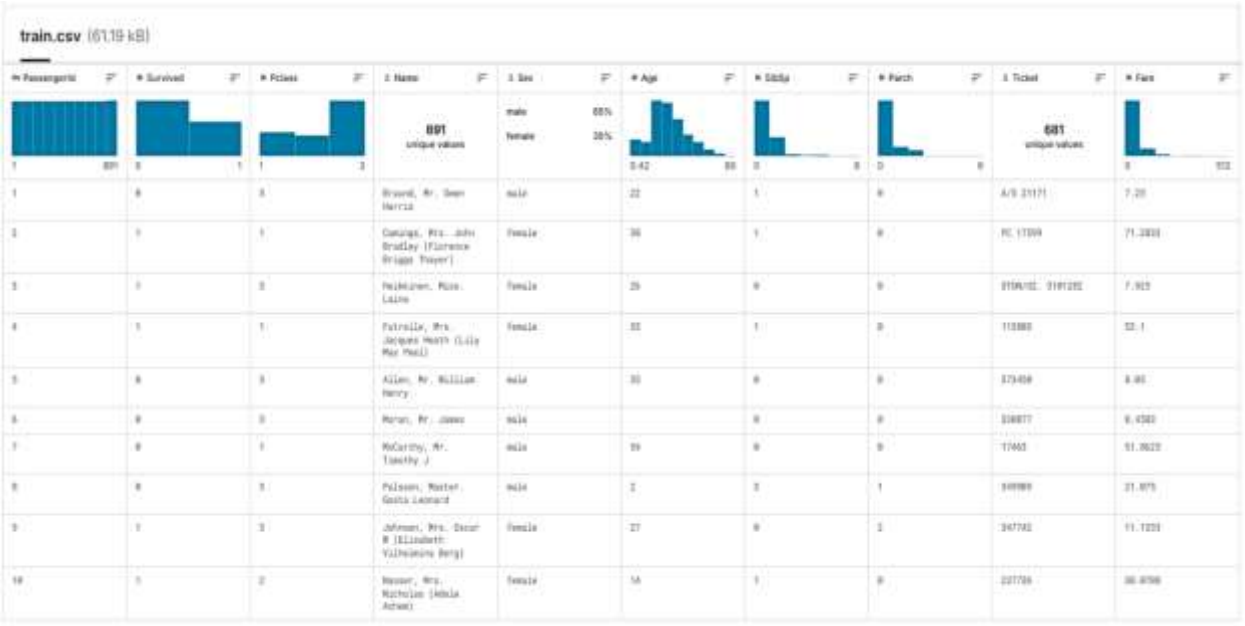
Việc RMS Titanic bị chìm vào năm 1912 đã dẫn đến cái chết của 1502 trong số 2224 người. Bộ dữ liệu Titanic không chỉ chứa thông tin về việc hành khách có sống sót hay không, mà còn có các thuộc tính như: tuổi, giới tính, hạng vé và các tham số khác. Bộ dữ liệu này có sẵn miễn phí và có thể được mở và phân tích trên nhiều nền tảng cả ngoại tuyến và trực tuyến.

```
Liên kết đến bộ dữ liệu Titanic:
https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv
```

Trước đó trong chương "IDE hỗ trợ LLM và những thay đổi trong lập trình tương lai", đã thảo luận về Jupyter Notebook - một trong những môi trường phát triển phổ biến nhất cho phân tích dữ liệu và học máy. Các nền tảng miễn phí tương tự Jupyter Notebook bao gồm Kaggle và Google Collab, cho phép chạy mã Python mà không cần cài đặt phần mềm và cung cấp quyền truy cập miễn phí vào các tài nguyên tính toán.

> Kaggle – nền tảng lớn nhất cho phân tích dữ liệu, các cuộc thi về học máy với môi trường thực thi mã tích hợp. Tính đến tháng 10 năm 2023, Kaggle có hơn 15 triệu người dùng từ 194 quốc gia.

Tải lên và sử dụng bộ dữ liệu Titanic trên nền tảng Kaggle (Hình 9.25), để lưu trữ bộ dữ liệu (bản sao của nó) và chạy mã Python với các thư viện đã được cài đặt sẵn trực tiếp trong trình duyệt, mà không cần phải cài đặt IDE riêng.-



Hình 9.25 Thống kê bảng Titanic – bộ dữ liệu học tập phổ biến nhất để nghiên cứu phân tích dữ liệu và học máy.

Bộ dữ liệu Titanic bao gồm thông tin về 2224 hành khách có mặt trên tàu RMS Titanic trong thời gian xảy ra vụ đắm vào năm 1912. Bộ dữ liệu được trình bày dưới dạng hai bảng riêng biệt – tập huấn luyện (train.csv) và tập kiểm tra (test.csv), cho phép sử dụng nó để đào tạo mô hình cũng như đánh giá độ chính xác của chúng trên dữ liệu mới.

Tập dữ liệu huấn luyện chứa cả các thuộc tính của hành khách (tuổi, giới tính, hạng vé và các thông tin khác), cũng như thông tin về việc ai đã sống sót (cột với các giá trị nhị phân "Sống sót"). Tập dữ liệu huấn luyện (Hình 9.26 – tệp train.csv) được sử dụng để đào tạo mô hình. Tập dữ liệu kiểm tra (Hình 9.27 – tệp test.csv) chỉ bao gồm các thuộc tính của hành khách mà không có thông tin về việc sống sót (thiếu một cột duy nhất "Sống sót"). Tập dữ liệu kiểm tra được thiết kế để kiểm tra hoạt động của mô hình trên dữ liệu mới và đánh giá độ chính xác của nó.-

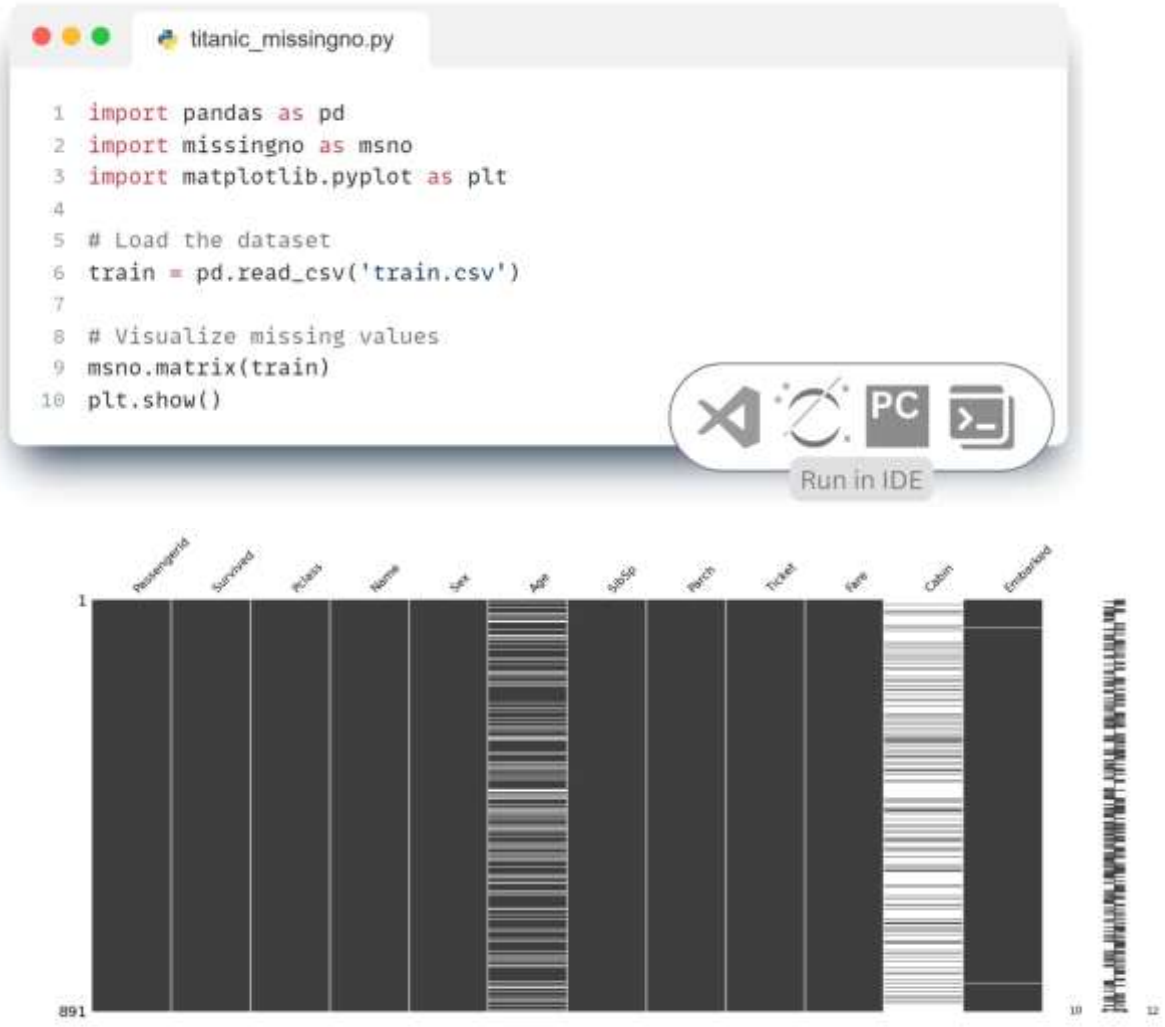
Do đó, chúng ta có các thuộc tính hành khách gần như giống hệt nhau trong các tập dữ liệu huấn luyện và kiểm tra. Sự khác biệt chính là trong tập dữ liệu kiểm tra, chúng ta có danh sách hành khách mà thiếu cột "Sống sót" – biến mục tiêu mà chúng ta muốn học cách dự đoán bằng các thuật toán toán học khác nhau. Và sau khi xây dựng mô hình, chúng ta có thể so sánh đầu ra của mô hình với tham số thực tế "Sống sót" từ tập dữ liệu kiểm tra, mà chúng ta sẽ xem xét để đánh giá kết quả.

Các cột chính trong bảng, các tham số của hành khách trong tập dữ liệu huấn luyện và kiểm tra:

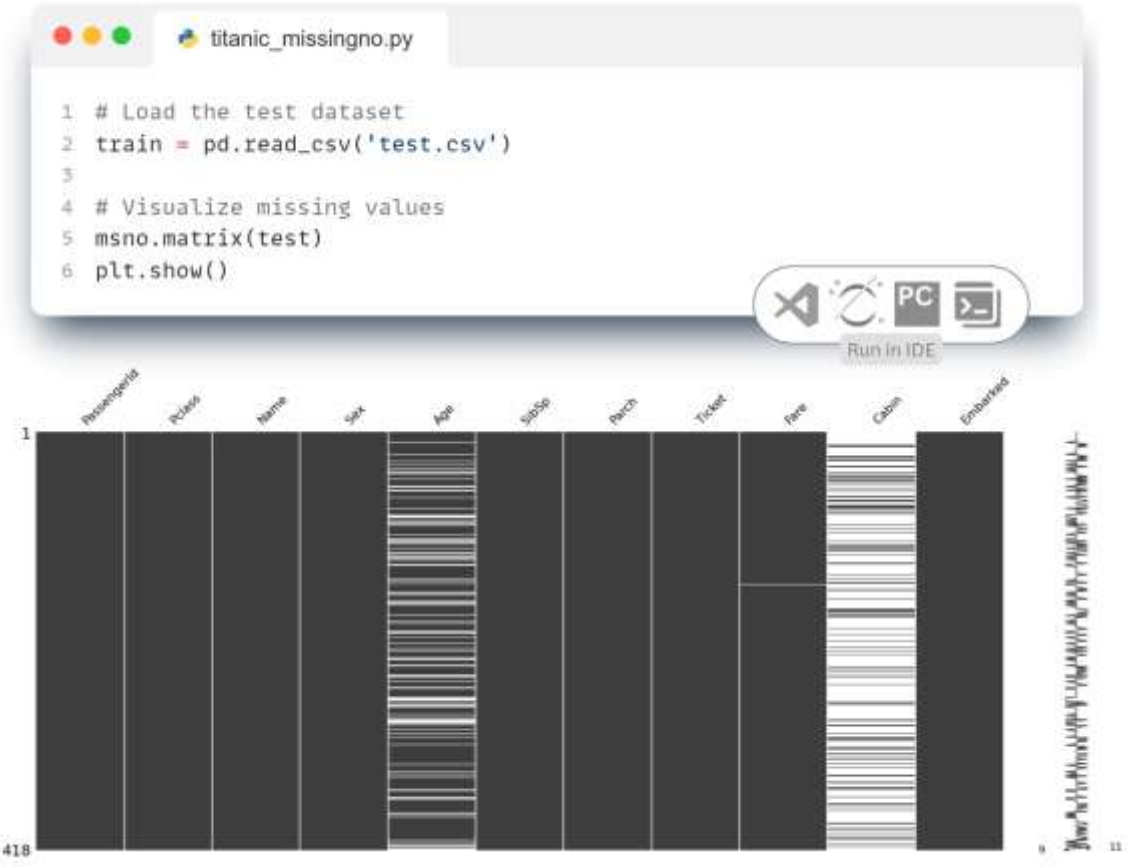
- PassengerId – mã định danh duy nhất của hành khách
- Survived – 1, nếu hành khách sống sót, 0 – nếu đã chết (không có trong tập kiểm tra)
- Pclass – hạng vé (1, 2 hoặc 3)
- Name – tên hành khách
- Sex – giới tính hành khách (nam/nữ)
- Age – tuổi
- SibSp – số lượng anh/chị/em hoặc vợ/chồng trên tàu

- Parch – số lượng cha/mẹ hoặc con cái trên tàu
- Ticket – số vé
- Fare – giá vé
- Cabin – số phòng (nhiều dữ liệu bị thiếu)
- Embarked – cảng lên tàu (C = Cherbourg, Q = Queenstown, S = Southampton)

Để trực quan hóa dữ liệu bị thiếu trong cả hai bảng, có thể sử dụng thư viện missingno (Hình 9.26, Hình 9.27), hiển thị các giá trị bị thiếu dưới dạng biểu đồ, trong đó các vùng trắng cho thấy dữ liệu bị thiếu. Trực quan hóa như vậy cho phép nhanh chóng đánh giá chất lượng dữ liệu trước khi xử lý.



Hình 9.26 Bằng cách sử dụng một vài dòng mã, dữ liệu bị thiếu trong tập dữ liệu huấn luyện Titanic được trực quan hóa, trong đó tham số chính để đào tạo là tham số "Sống sót".



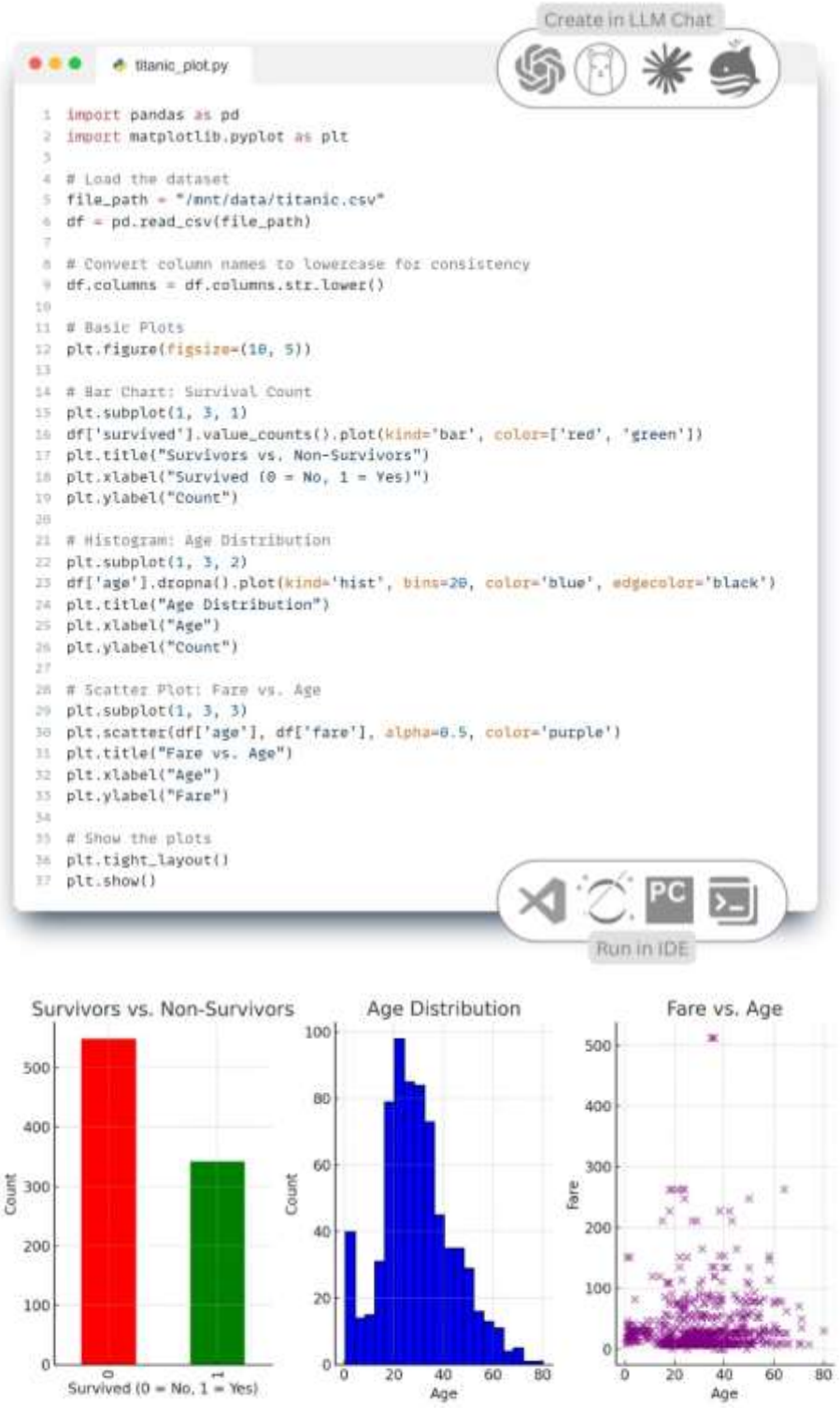
Hình 9.27: Hình ảnh hóa dữ liệu bị thiếu trong tập dữ liệu thử nghiệm Titanic, chỉ chứa các đặc điểm của hành khách mà không có thông tin.

Trước khi xây dựng giả thuyết và dự đoán dựa trên tập dữ liệu, phân tích hình ảnh giúp phát hiện các quy luật chính trong dữ liệu, đánh giá chất lượng của chúng và xác định các mối quan hệ có thể có. Có nhiều phương pháp hình ảnh hóa giúp hiểu rõ hơn về tập dữ liệu Titanic. Bạn có thể sử dụng biểu đồ phân phối để phân tích các nhóm tuổi của hành khách, biểu đồ sống sót theo giới tính và hạng, cũng như ma trận dữ liệu bị thiếu để đánh giá chất lượng thông tin và hiểu dữ liệu.

- Hãy nhờ LLM giúp chúng ta hình ảnh hóa dữ liệu từ tập dữ liệu Titanic, để làm điều này, chúng ta sẽ gửi yêu cầu văn bản sau đến bất kỳ mô hình LLM nào (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN hoặc bất kỳ mô hình nào khác):

> Làm ơn hãy cho tôi xem một vài biểu đồ đơn giản cho tập dữ liệu Titanic. Tự động tải xuống tập dữ liệu và hiển thị các biểu đồ ⏎

- Phản hồi của LLM dưới dạng mã và biểu đồ sẵn sàng, hình ảnh hóa các tham số của tập dữ liệu:



Hình 9.28: LLM giúp nhanh chóng có được hình ảnh hóa dữ liệu của tập dữ liệu.

Hình ảnh hóa dữ liệu là một bước quan trọng, cho phép chuẩn bị tập dữ liệu cho việc xây dựng mô hình học máy tiếp theo, mà chỉ có thể thực hiện sau khi hiểu rõ dữ liệu.

Machine learning in action: from Titanic passengers to project management.

Giả thuyết chính được sử dụng để nghiên cứu các nguyên tắc của học máy dựa trên tập dữ liệu Titanic là một số nhóm hành khách có khả năng sống sót cao hơn.

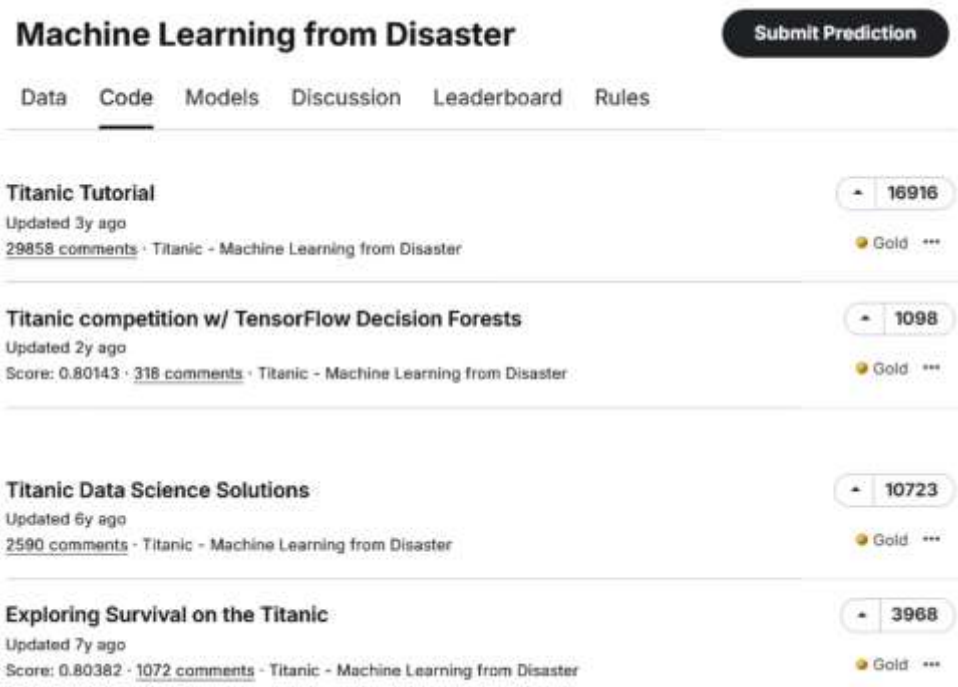
> Một bảng nhỏ về hành khách Titanic đã trở nên phổ biến trên toàn thế giới, và hàng triệu người sử dụng nó để học tập, thử nghiệm và kiểm tra các mô hình nhằm tìm ra các thuật toán và giả thuyết nào sẽ xây dựng được mô hình dự đoán sống sót chính xác nhất dựa trên tập dữ liệu huấn luyện cho hành khách Titanic.

Sự hấp dẫn của tập dữ liệu Titanic được giải thích bởi tính gọn nhẹ: với vài trăm dòng và mười hai cột (Hình 9.26), nó cung cấp nhiều cơ hội cho phân tích. Tập dữ liệu này là một ví dụ tương đối đơn giản về giải quyết phân loại nhị phân, trong đó mục tiêu của nhiệm vụ - sống sót - được thể hiện dưới dạng 0 hoặc 1.

> John Wheeler trong "It from Bit" [7] khẳng định rằng nền tảng của vũ trụ là các quyết định nhị phân. Tương tự, doanh nghiệp, được điều hành bởi những con người cấu thành từ các phân tử, thực sự được xây dựng trên một chuỗi các lựa chọn nhị phân.

Hơn nữa, dữ liệu dựa trên một sự kiện lịch sử có thật, điều này làm cho chúng có giá trị cho nghiên cứu, khác với các ví dụ được tạo ra nhân tạo. Chỉ trên nền tảng Kaggle, một trong những nền tảng lớn nhất cho công việc với Data Pipeline và ETL, đã có 1.355.998 người tham gia giải quyết các bài toán dựa trên tập dữ liệu Titanic, phát triển 53.963 giải pháp Data Pipeline độc đáo [157] (Hình 9.29).-

> Có vẻ không thể tin được, nhưng chỉ với 1000 dòng dữ liệu về hành khách "Titanic" với 12 tham số đã trở thành một lĩnh vực cho hàng triệu giả thuyết, chuỗi logic và các Data-Pipelines độc đáo. Từ một tập dữ liệu nhỏ, vô số thông tin, giả thuyết và diễn giải được sinh ra - từ các mô hình sống sót đơn giản đến các tổ hợp phức tạp, xem xét các quy luật ẩn và những mê cung phức tạp của lý luận.



Hình 9.29 Năm giải pháp đầu tiên trong tổng số 53.963 Pipeline đã hoàn thành và mở. Gần 1,5 triệu người đã thử giải quyết bài toán này chỉ trên Kaggle [157].

Nếu ngay cả một bảng nhỏ như vậy có thể tạo ra hàng triệu giải pháp độc đáo (Hình 9.29), thì nói gì đến các tập dữ liệu xây dựng công nghiệp thực tế, nơi các tham số được đo bằng hàng chục nghìn? -

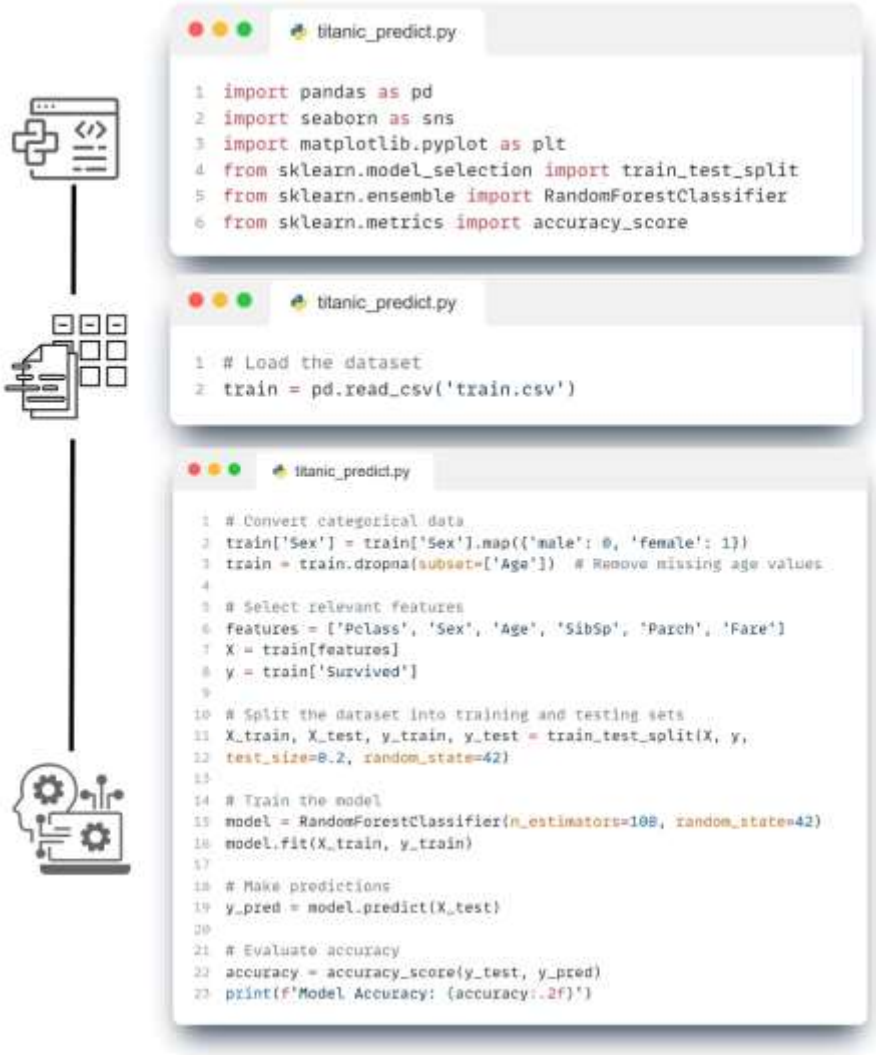
> Một dự án CAD tiêu chuẩn cho một tòa nhà nhỏ chứa hàng chục nghìn thực thể với hàng nghìn tham số - từ các đặc điểm hình học đến các thuộc tính về chi phí và thời gian. Hãy tưởng tượng có bao nhiêu thông tin tiềm năng, mối quan hệ, dự đoán và giả thuyết quản lý ẩn chứa trong dữ liệu của tất cả các dự án của công ty bạn, được thu thập trong những năm qua. Dữ liệu dự án lịch sử không chỉ là một kho lưu trữ - đó là ký ức sống của tổ chức, dấu vết kỹ thuật số của nó, có thể được phân tích để xây dựng một lượng lớn giả thuyết độc đáo.

Điều quan trọng nhất là không cần phải chờ đợi cộng đồng Kaggle quan tâm đến công ty hoặc dữ liệu của bạn. Ngày hôm nay, bạn có thể bắt đầu làm việc với những gì bạn có: thực hiện phân tích trên dữ liệu của chính mình, đào tạo mô hình trên dữ liệu của chính mình, phát hiện các lặp lại, bất thường và mẫu. Nơi mà trước đây cần nhiều năm thử nghiệm và tư vấn tốn kém, ngày nay chỉ cần có sáng kiến, LLM, cách tiếp cận mở đối với dữ liệu và sẵn sàng học hỏi.

- Để xây dựng một thuật toán học máy có thể dự đoán khả năng sống sót của hành khách dựa trên tập dữ liệu huấn luyện hành khách train.csv, chúng ta sẽ yêu cầu LLM giải quyết bài toán này cho chúng ta:

> Dựa trên tập dữ liệu huấn luyện hành khách Titanic, hãy xây dựng mô hình học máy để dự đoán khả năng sống sót ⏎

- Phản hồi từ LLM:



Hình 9.210 LLM đã xây dựng dự đoán về những người sống sót trên Titanic bằng cách sử dụng thuật toán học máy Random Forest.

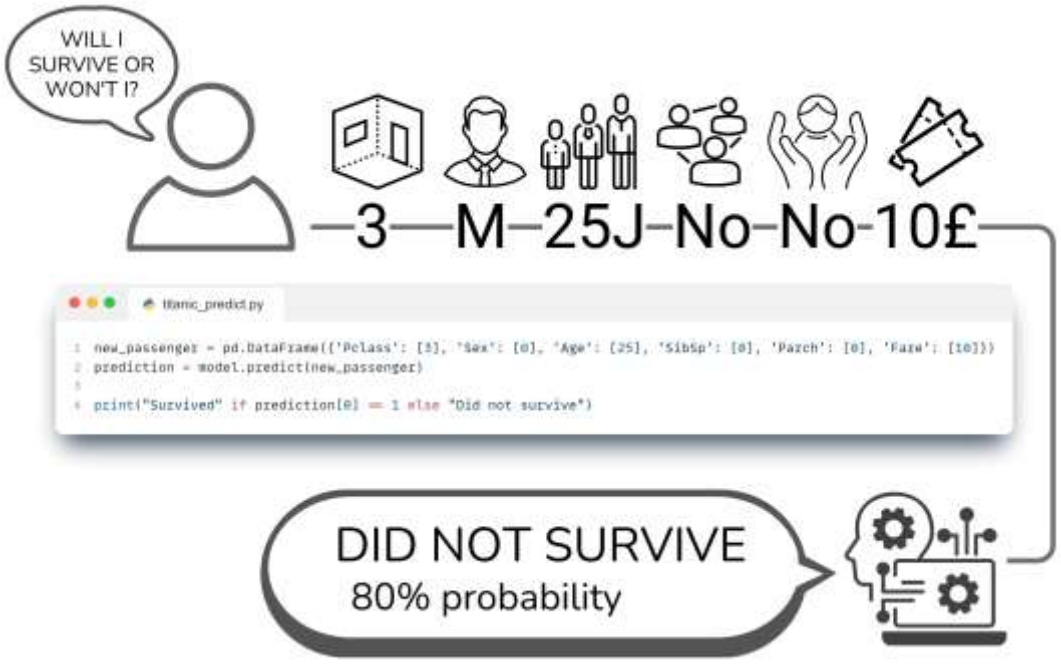
Mã nhận được từ LLM (Hình 9.210) tải dữ liệu về hành khách "Titanic", làm sạch chúng, chuyển đổi các biến phân loại (ví dụ: giới tính sang định dạng số) và đào tạo mô hình thông qua thuật toán RandomForestClassifier để dự đoán xem hành khách có sống sót hay không (chúng ta sẽ nói thêm về các thuật toán phổ biến trong các chương tiếp theo).-

Bằng cách sử dụng mã, dữ liệu trong quá trình đào tạo được chia thành tập huấn luyện và tập kiểm tra (trên trang web Kaggle, đã có sẵn test.csv để đào tạo (Hình 9.27) và train.csv (Hình 9.26), sau đó mô hình được đào tạo trên dữ liệu huấn luyện và được kiểm tra trên dữ liệu kiểm tra, để hiểu mô hình dự đoán nào là tốt nhất. Sau khi đào tạo, dữ liệu kiểm tra từ test.csv (với dữ liệu thực tế về những người sống sót hoặc không sống sót) được đưa vào mô hình, và nó dự đoán ai sống sót và ai không. Trong trường hợp của chúng tôi, độ chính xác của mô hình học máy mà chúng tôi nhận được khoảng 80%, cho thấy rằng nó khá tốt trong

việc nắm bắt các quy luật.--

> Học máy có thể được so sánh với một đứa trẻ đang cố gắng đưa một khối hình chữ nhật vào một lỗ tròn. Ở giai đoạn đầu, thuật toán thử nhiều cách tiếp cận, gặp phải lỗi và sự không phù hợp. Quá trình này có thể có vẻ không hiệu quả, nhưng nó cung cấp một bài học quan trọng: bằng cách phân tích từng lỗi, mô hình cải thiện dự đoán của mình và đưa ra các quyết định ngày càng chính xác hơn.

Bây giờ mô hình này (Hình 9.210) có thể được sử dụng để dự đoán khả năng sống sót của những hành khách mới và ví dụ, nếu cung cấp thông tin về một hành khách thông qua hàm model.predict với các tham số: "nam", "hạng 3", "25 tuổi", "không có người thân trên tàu", mô hình sẽ đưa ra dự đoán - rằng hành khách có 80% khả năng không sống sót trong thảm họa, nếu anh ta ở trên tàu Titanic vào năm 1912 (Hình 9.211).-



Hình 9.211 Mô hình mà chúng tôi đã tạo ra, giờ đây có thể dự đoán với 80% khả năng hành khách nào đó trên Titanic sẽ sống sót hay không.

Mô hình dự đoán khả năng sống sót của hành khách "Titanic" minh họa một khái niệm rộng lớn hơn: hàng ngày, hàng ngàn chuyên gia trong ngành xây dựng đưa ra những quyết định "đôi" tương tự - quyết định sống hay chết của dự án, dự toán, công cụ, lợi nhuận hay thua lỗ, an toàn hay rủi ro. Giống như trong ví dụ với "Titanic", nơi kết quả phụ thuộc vào các yếu tố (giới tính, tuổi tác, hạng), trong xây dựng, mỗi khía cạnh của quyết định bị ảnh hưởng bởi nhiều yếu tố và biến số riêng (các cột trong bảng): chi phí vật liệu, trình độ công nhân, thời gian, thời tiết, logistics, rủi ro kỹ thuật, nhận xét và hàng trăm ngàn tham số khác.

> Trong ngành xây dựng, học máy được áp dụng theo những nguyên tắc tương tự như trong các lĩnh vực khác: các mô hình được đào tạo trên dữ liệu lịch sử - từ các dự án, hợp đồng, dự toán - để kiểm tra các giả thuyết khác nhau và tìm kiếm các giải pháp hiệu quả nhất. Quá trình này phần nào giống như việc dạy trẻ em thông qua phương pháp thử và sai: với mỗi chu kỳ, các mô hình thích nghi và trở nên chính xác hơn.

Việc sử dụng dữ liệu tích lũy mở ra những chân trời mới cho ngành xây dựng. Thay vì các phép tính thủ công tốn thời gian, có thể đào tạo các mô hình có khả năng dự đoán với độ chính xác cao các đặc điểm chính của các dự án tương lai. Do đó, phân tích dự đoán biến ngành xây dựng thành một không gian mà không chỉ có thể lập kế hoạch mà còn tự tin dự đoán sự phát triển của các sự kiện.

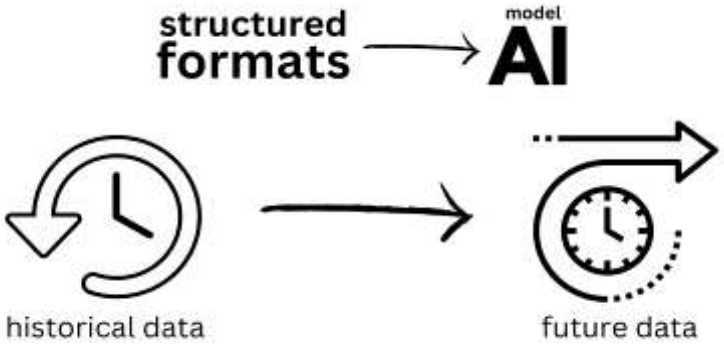
Dự đoán và dự báo dựa trên dữ liệu lịch sử

Dữ liệu thu thập được về các dự án của công ty mở ra khả năng xây dựng các mô hình có thể dự đoán chi phí và các đặc điểm thời gian của các đối tượng tương lai chưa được thực hiện - mà không cần các phép tính và so sánh thủ công tốn thời gian. Điều này cho phép tăng tốc và đơn giản hóa đáng kể các quy trình đánh giá, dựa vào các dự đoán toán học có cơ sở thay vì các giả định chủ quan.

Trước đây, trong phần thứ tư của cuốn sách, chúng tôi đã xem xét chi tiết các phương pháp truyền thống để tính toán chi phí dự án, bao gồm phương pháp tài nguyên, cũng như đề cập đến các phương pháp tham số và chuyên gia. Những phương pháp này vẫn còn phù hợp, nhưng trong thực tiễn hiện đại, chúng bắt đầu được làm phong phú bằng các công cụ phân tích thống kê và học máy, cho phép tăng cường đáng kể độ chính xác và khả năng tái tạo của các đánh giá.

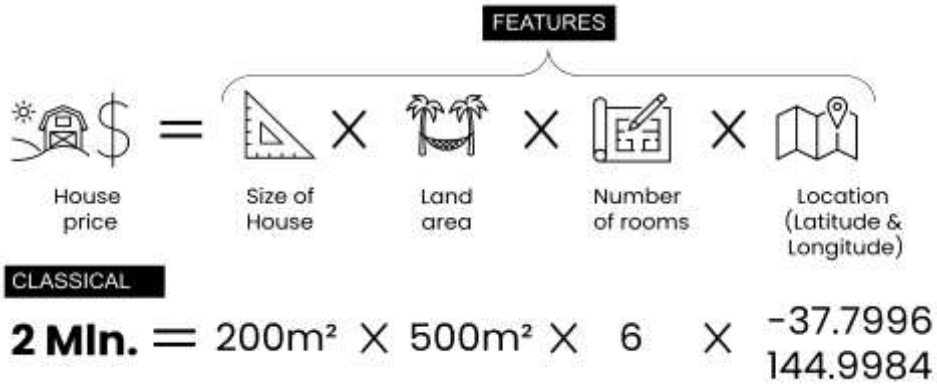
Các quy trình tính toán giá cả và các thuộc tính thời gian bằng tay và bán tự động sẽ trong tương lai được bổ sung bằng ý kiến và dự đoán của các mô hình học máy, có khả năng phân tích dữ liệu lịch sử, tìm ra các quy luật ẩn và đề xuất các giải pháp có cơ sở. Dữ liệu và kịch bản mới sẽ được tạo ra tự động từ thông tin đã có - tương tự như cách mà các mô hình ngôn ngữ (LLM) tạo ra văn bản, hình ảnh và mã dựa trên dữ liệu được thu thập trong nhiều năm từ các nguồn mở.

> Cũng giống như ngày nay con người dựa vào kinh nghiệm, trực giác và thống kê nội bộ khi đánh giá các sự kiện trong tương lai, trong những năm tới, tương lai của các dự án xây dựng sẽ ngày càng được xác định bởi sự kết hợp giữa kiến thức tích lũy và các mô hình toán học của học máy.



Hình 9.212 Dữ liệu lịch sử chất lượng và có cấu trúc của công ty - nguyên liệu để xây dựng các mô hình học máy và dự đoán.

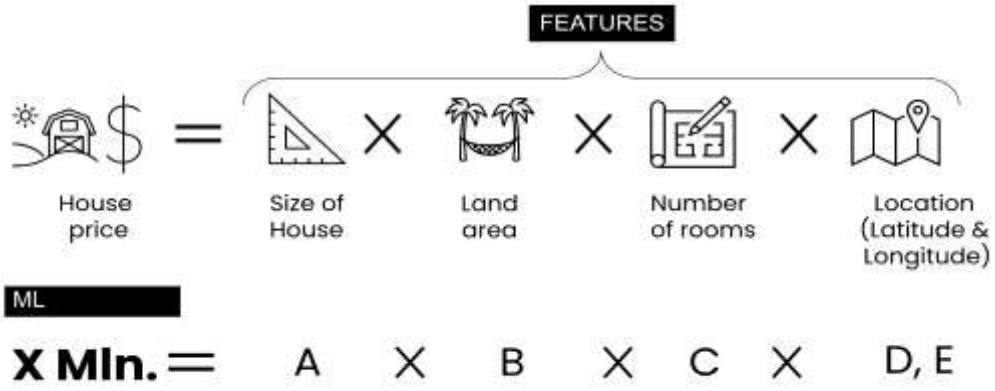
Hãy xem xét một ví dụ đơn giản: dự đoán giá nhà dựa trên diện tích, kích thước lô đất, số lượng phòng và vị trí địa lý. Một trong những cách tiếp cận là xây dựng một mô hình cổ điển, phân tích các tham số này và tính toán giá dự kiến (Hình 9.213). Cách tiếp cận này yêu cầu một công thức chính xác và đã biết trước, điều này trong thực tế gần như không thể.



Hình 9.213 Để đánh giá giá trị của một ngôi nhà, có thể sử dụng thuật toán cổ điển với công thức cố định, mà cần phải tìm ra.

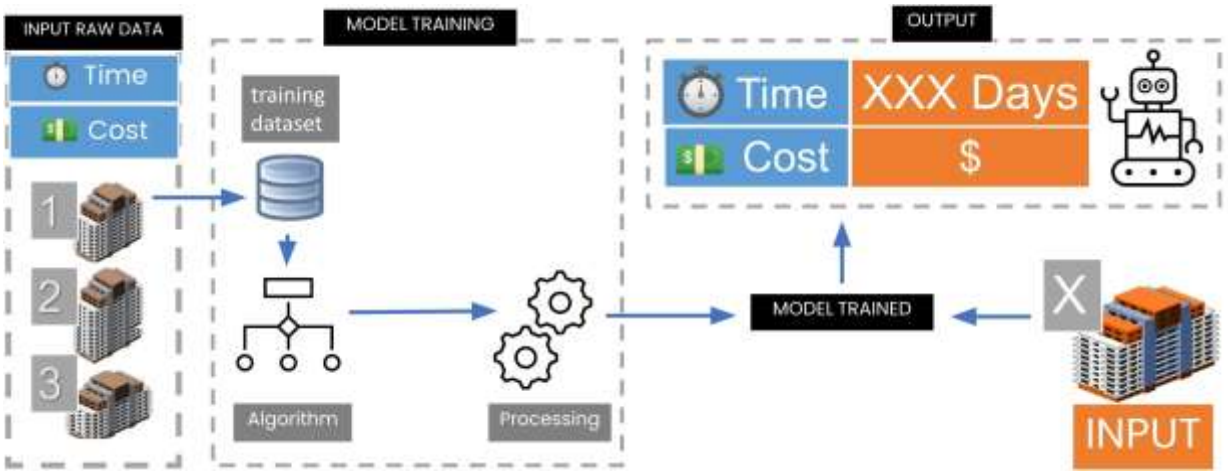
Học máy cho phép từ bỏ việc tìm kiếm công thức thủ công và thay thế chúng bằng các thuật toán có thể học, tự phát hiện các mối quan hệ, vượt trội hơn về độ chính xác so với bất kỳ phương trình đã cho nào. Như một sự thay thế, chúng ta sẽ tạo ra một thuật toán học máy, sẽ tạo ra mô hình dựa trên sự hiểu biết trước về vấn đề và dữ liệu lịch sử, có thể không đầy đủ (Hình 9.214).-

> Trong ví dụ về vấn đề định giá, học máy cho phép tạo ra các loại mô hình toán học khác nhau, không yêu cầu biết cơ chế chính xác của việc hình thành giá trị. Mô hình "học" từ dữ liệu về các dự án trước đó, điều chỉnh theo các quy luật thực tế giữa các tham số của tòa nhà, giá trị của chúng và thời gian thực hiện.



Hình 9.214 Khác với việc đánh giá cổ điển theo công thức, thuật toán học máy được đào tạo trên dữ liệu lịch sử.

Trong bối cảnh học máy có giám sát (supervised machine learning), mỗi dự án trong tập dữ liệu huấn luyện chứa cả các thuộc tính đầu vào (ví dụ: dữ liệu về giá trị và thời gian xây dựng của các tòa nhà tương tự) và các giá trị đầu ra mong đợi (ví dụ: giá trị hoặc thời gian). Tập dữ liệu như vậy được sử dụng để tạo ra và điều chỉnh mô hình học máy (Hình 9.215). Càng nhiều dữ liệu và chất lượng dữ liệu càng cao, mô hình sẽ càng chính xác và kết quả dự đoán sẽ càng chính xác.-



Hình 9.215 Mô hình ML, được đào tạo trên dữ liệu về giá trị và tiến độ thực hiện của các dự án trước đó, sẽ xác định giá trị và tiến độ thực hiện của một dự án mới với một xác suất nhất định.

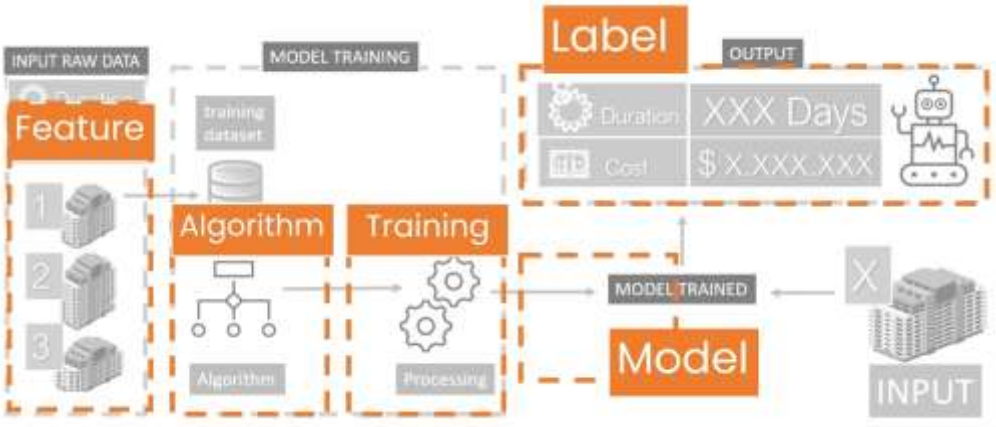
Sau khi tạo ra và đào tạo mô hình để đánh giá xây dựng một dự án mới, chỉ cần cung cấp cho mô hình các thuộc tính mới cho dự án mới, và mô hình sẽ cung cấp kết quả tính toán dựa trên các quy luật đã học trước đó với một xác suất nhất định.

Các khái niệm chính trong học máy

Học máy không phải là phép thuật, mà chỉ là toán học, dữ liệu và tìm kiếm các quy luật. Nó không có trí thông minh thực sự, mà là một chương trình được đào tạo trên dữ liệu để nhận diện các mẫu và đưa ra quyết định mà không cần sự tham gia liên tục của con người.

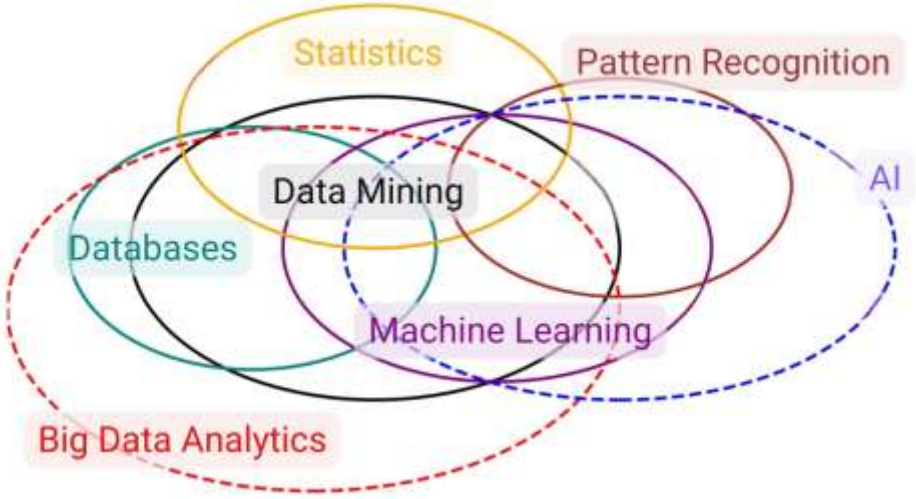
Để mô tả cấu trúc của nó, học máy sử dụng một loạt các khái niệm chính (Hình 9.216):-

- Nhãn (Labels) là các biến mục tiêu hoặc thuộc tính (tham số "Sống sót" trong tập dữ liệu Titanic) mà mô hình cần dự đoán. Ví dụ: chi phí xây dựng (chẳng hạn, bằng đô la), thời gian xây dựng (chẳng hạn, bằng tháng).
- Đặc điểm (Features) là các biến độc lập hoặc thuộc tính, đóng vai trò là đầu vào cho mô hình. Trong mô hình dự đoán, chúng được sử dụng để dự đoán các nhãn. Ví dụ: diện tích lô đất (tính bằng mét vuông), số tầng của tòa nhà, tổng diện tích tòa nhà (tính bằng mét vuông), vị trí địa lý (vĩ độ và kinh độ), loại vật liệu được sử dụng trong xây dựng. Số lượng đặc điểm cũng xác định kích thước của dữ liệu.
- Mô hình (Model) là một tập hợp các giả thuyết khác nhau, trong đó một giả thuyết gần nhất với hàm mục tiêu mà cần dự đoán hoặc xấp xỉ. Ví dụ: mô hình học máy sử dụng các phương pháp phân tích hồi quy để dự đoán chi phí và thời gian xây dựng.
- Thuật toán học (Learning Algorithm) là quá trình tìm kiếm giả thuyết tốt nhất trong mô hình, phù hợp chính xác với hàm mục tiêu, sử dụng một tập hợp dữ liệu huấn luyện. Ví dụ: Thuật toán hồi quy tuyến tính, KNN hoặc rừng ngẫu nhiên, phân tích dữ liệu về chi phí và thời gian xây dựng để phát hiện mối quan hệ và quy luật.
- Huấn luyện (Training) - trong quá trình huấn luyện, thuật toán phân tích dữ liệu huấn luyện, tìm ra các quy luật tương ứng với mối quan hệ giữa các thuộc tính đầu vào và các nhãn mục tiêu. Kết quả của quá trình này là một mô hình học máy đã được huấn luyện, sẵn sàng cho việc dự đoán. Ví dụ: quá trình mà thuật toán phân tích dữ liệu lịch sử về xây dựng (chi phí, thời gian, đặc điểm của đối tượng) để tạo ra một mô hình dự đoán.



Hình 9.216 ML sử dụng nhãn và thuộc tính để tạo ra các mô hình, được huấn luyện trên dữ liệu thông qua các thuật toán để dự đoán kết quả.

Học máy không tồn tại trong sự cô lập, mà là một phần của hệ sinh thái rộng lớn hơn của các lĩnh vực phân tích, bao gồm thống kê, cơ sở dữ liệu, phân tích dữ liệu thông minh, nhận dạng mẫu, phân tích dữ liệu lớn và trí tuệ nhân tạo. Hình 9.217 cho thấy cách các lĩnh vực này giao thoa và bổ sung cho nhau, tạo ra một nền tảng toàn diện cho các hệ thống ra quyết định và tự động hóa hiện đại.-

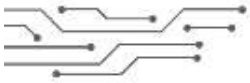


Hình 9.217 Mối quan hệ giữa các lĩnh vực phân tích dữ liệu khác nhau: thống kê, học máy, trí tuệ nhân tạo, dữ liệu lớn, nhận dạng mẫu và phân tích dữ liệu thông minh.

> Mục tiêu chính của học máy là trang bị cho máy tính khả năng tự động tiếp thu kiến thức mà không cần sự can thiệp hoặc trợ giúp của con người và điều chỉnh hành động của chúng một cách phù hợp.

Do đó, trong tương lai, vai trò của con người sẽ chỉ là cung cấp cho máy những khả năng nhận thức - họ sẽ đặt ra các điều kiện, trọng số và tham số, trong khi mô hình học máy sẽ thực hiện tất cả các phần còn lại.

Trong chương tiếp theo, chúng ta sẽ xem xét các ví dụ cụ thể về việc áp dụng các thuật toán. Trên các bảng thực tế và các mô hình đơn giản, sẽ được trình bày cách từng bước xây dựng dự đoán.



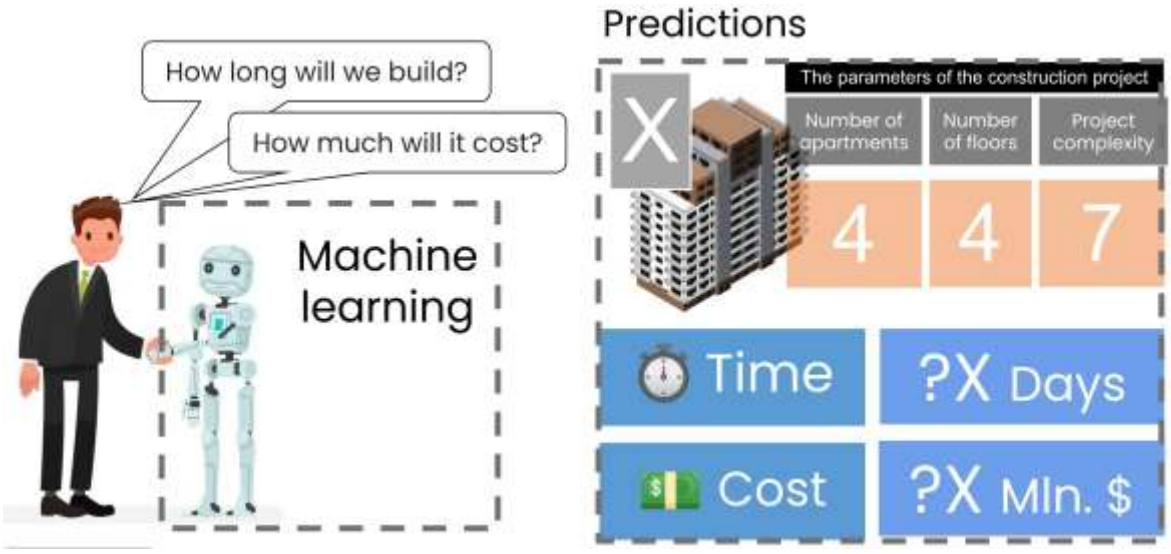
CHƯƠNG 9.3.
DỰ ĐOÁN CHI PHÍ VÀ THỜI GIAN BẰNG HỌC MÁY.

Ví dụ về việc sử dụng học máy để xác định chi phí và thời gian của dự án.

Đánh giá thời gian và chi phí xây dựng là một trong những quy trình then chốt trong hoạt động của công ty xây dựng. Truyền thống, những đánh giá này được thực hiện bởi các chuyên gia dựa trên kinh nghiệm, tài liệu tham khảo và cơ sở quy định. Tuy nhiên, trong bối cảnh chuyển đổi số và sự gia tăng khả năng tiếp cận dữ liệu, có cơ hội sử dụng các mô hình học máy (ML) để nâng cao độ chính xác và tự động hóa các đánh giá này.

Việc áp dụng học máy vào quy trình tính toán chi phí và thời gian xây dựng không chỉ nâng cao hiệu quả lập kế hoạch mà còn trở thành điểm khởi đầu cho việc tích hợp các mô hình thông minh vào các quy trình kinh doanh khác — từ quản lý rủi ro đến tối ưu hóa logistics và mua sắm.

> Quan trọng là phải nhanh chóng xác định thời gian cần thiết để xây dựng một công trình và tổng chi phí của nó. Những câu hỏi về thời gian và chi phí của dự án từ lâu đã chiếm vị trí trung tâm trong tâm trí của cả khách hàng và các công ty xây dựng kể từ khi ngành xây dựng ra đời.



Rủi ro 9.31 Trong các dự án xây dựng, các yếu tố then chốt cho sự thành công là tốc độ và chất lượng đánh giá thời gian và chi phí xây dựng.

Trong ví dụ tiếp theo, các dữ liệu chính sẽ được trích xuất từ các dự án trước đó, và dựa trên đó, một mô hình học máy sẽ được phát triển, cho phép chúng ta sử dụng mô hình này để đánh giá chi phí và thời gian thực hiện các dự án xây dựng mới với các tham số mới (Rủi ro 9.31).-

Hãy xem xét ba dự án với ba thuộc tính chính: số lượng căn hộ (trong đó 100 căn hộ tương đương với số 10 để đơn giản hóa việc hình dung), số tầng và mức độ phức tạp xây dựng theo thang điểm từ 1 đến 10, trong đó 10 là mức độ phức tạp cao nhất. Trong học máy, quá trình chuyển đổi và đơn giản hóa các giá trị như 100 thành 10 hoặc 50 thành 5 được gọi là "chuẩn hóa".

> Chuẩn hóa trong học máy là quá trình đưa các dữ liệu số khác nhau về cùng một thang đo để dễ dàng xử lý và phân tích. Quá trình này đặc biệt quan trọng khi dữ liệu có các thang đo và đơn vị đo lường khác nhau.

Giả sử rằng trong dự án đầu tiên (Rủi ro 9.32) có 50 căn hộ (sau khi chuẩn hóa – 5), 7 tầng và đánh giá độ phức tạp là 2, điều này có nghĩa là xây dựng tương đối đơn giản. Trong dự án thứ hai có 80 căn hộ, 9 tầng và dự án tương đối phức tạp. Trong những điều kiện như vậy, việc xây dựng hai tòa nhà chung cư đầu tiên và thứ hai mất 270 và 330 ngày, và tổng chi phí của dự án lần lượt là 4,5 và 5,8 triệu đô la.



Rủi ro 9.32 Ví dụ về tập hợp các dự án trước đó sẽ được sử dụng để đánh giá thời gian và chi phí cho dự án tương lai X.

Khi xây dựng mô hình học máy cho những dữ liệu như vậy, nhiệm vụ chính là xác định các thuộc tính (hoặc nhãn) quan trọng để dự đoán, trong trường hợp này là thời gian và chi phí xây dựng. Với một tập dữ liệu nhỏ, chúng tôi sẽ sử dụng thông tin từ các dự án xây dựng trước đó để lập kế hoạch cho các dự án mới: bằng cách sử dụng các thuật toán học máy, chúng tôi phải dự đoán chi phí và thời gian xây dựng của dự án mới X dựa trên các thuộc tính đã cho của dự án mới, chẳng hạn như 40 căn hộ, 4 tầng và độ phức tạp tương đối cao của dự án - 7. Trong thực tế, số lượng tham số đầu vào có thể lớn hơn nhiều - từ vài chục đến hàng trăm yếu tố. Chúng có thể bao gồm: loại vật liệu xây dựng, khu vực khí hậu, trình độ tay nghề của nhà thầu, sự hiện diện của các mạng lưới kỹ thuật, loại nền móng, mùa bắt đầu công việc, nhận xét của giám sát viên, v.v. -

> Để tạo ra một mô hình dự đoán học máy, chúng ta cần chọn một thuật toán để xây dựng nó. Thuật toán trong học máy giống như một công thức toán học, dạy cho máy tính cách thực hiện dự đoán (trộn các tham số theo đúng thứ tự) hoặc đưa ra quyết định dựa trên dữ liệu.

Để phân tích dữ liệu về các dự án xây dựng trước đó và dự đoán thời gian và chi phí cho các dự án tương lai, có thể sử dụng một trong những thuật toán học máy phổ biến.

- Hồi quy tuyến tính (Linear regression): thuật toán này cố gắng tìm mối quan hệ tuyến tính giữa các

thuộc tính, chẳng hạn như giữa số tầng và chi phí xây dựng. Mục tiêu của thuật toán là tìm ra phương trình tuyến tính mô tả tốt nhất mối quan hệ này, cho phép thực hiện dự đoán.

- Thuật toán k-láng giềng gần nhất (K-nearest neighbors (k-NN)): thuật toán này so sánh dự án mới với các dự án trước đó có kích thước hoặc độ phức tạp tương tự. k-NN phân loại dữ liệu dựa trên những ví dụ huấn luyện nào trong số k (số lượng) gần nhất với chúng. Trong bối cảnh hồi quy, kết quả là giá trị trung bình hoặc trung vị từ k láng giềng gần nhất.
- Cây quyết định (Decision Trees): đây là mô hình dự đoán phân tích, chia dữ liệu thành các tập con dựa trên các điều kiện khác nhau, sử dụng cấu trúc dạng cây. Mỗi nút của cây đại diện cho một điều kiện hoặc câu hỏi, dẫn đến việc chia nhỏ dữ liệu hơn nữa, và mỗi lá là dự đoán hoặc kết quả cuối cùng. Thuật toán chia dữ liệu thành các nhóm nhỏ hơn dựa trên các đặc điểm khác nhau, chẳng hạn như đầu tiên theo số tầng, sau đó theo độ phức tạp, v.v., để thực hiện dự đoán.

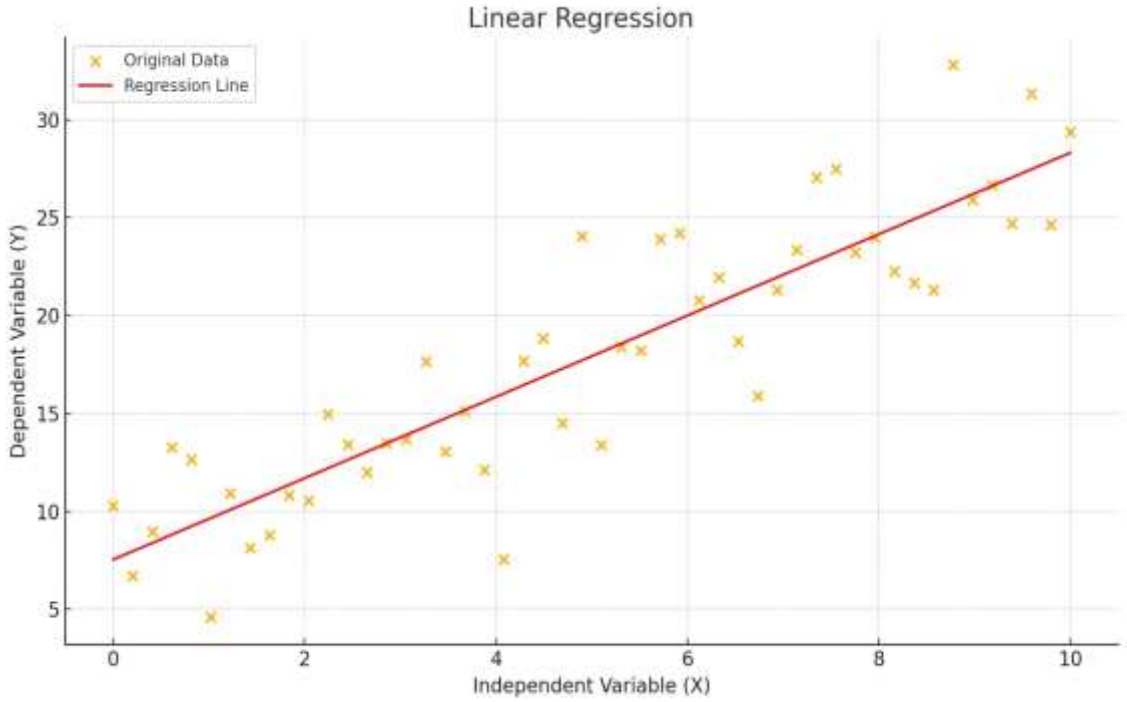
Hãy xem xét các thuật toán học máy để đánh giá chi phí của dự án mới dựa trên hai thuật toán phổ biến: hồi quy tuyến tính và thuật toán k-láng giềng gần nhất.

Dự đoán chi phí và thời gian của dự án bằng cách sử dụng hồi quy tuyến tính.

Hồi quy tuyến tính là một thuật toán cơ bản trong phân tích dữ liệu, cho phép dự đoán giá trị của một biến dựa trên mối quan hệ tuyến tính với một hoặc nhiều biến khác. Mô hình này giả định rằng có một mối quan hệ tuyến tính giữa biến phụ thuộc và một hoặc nhiều biến độc lập, và mục tiêu của thuật toán là tìm ra mối quan hệ này.

> Sự đơn giản và dễ hiểu của hồi quy tuyến tính đã khiến nó trở thành một công cụ phổ biến trong nhiều lĩnh vực. Khi làm việc với một biến, hồi quy tuyến tính liên quan đến việc tìm ra đường thẳng phù hợp nhất đi qua các điểm dữ liệu.

Hồi quy tuyến tính tìm ra đường thẳng tốt nhất (đường màu đỏ) để xấp xỉ mối quan hệ giữa biến đầu vào X và biến đầu ra Y. Đường này cho phép dự đoán giá trị Y cho các giá trị X mới dựa trên mối quan hệ tuyến tính đã được xác định (Hình 9.33).-



Hình 9.33 Nguyên tắc hoạt động của hồi quy tuyến tính trong việc tìm ra đường thẳng tốt nhất đi qua các giá trị huấn luyện.

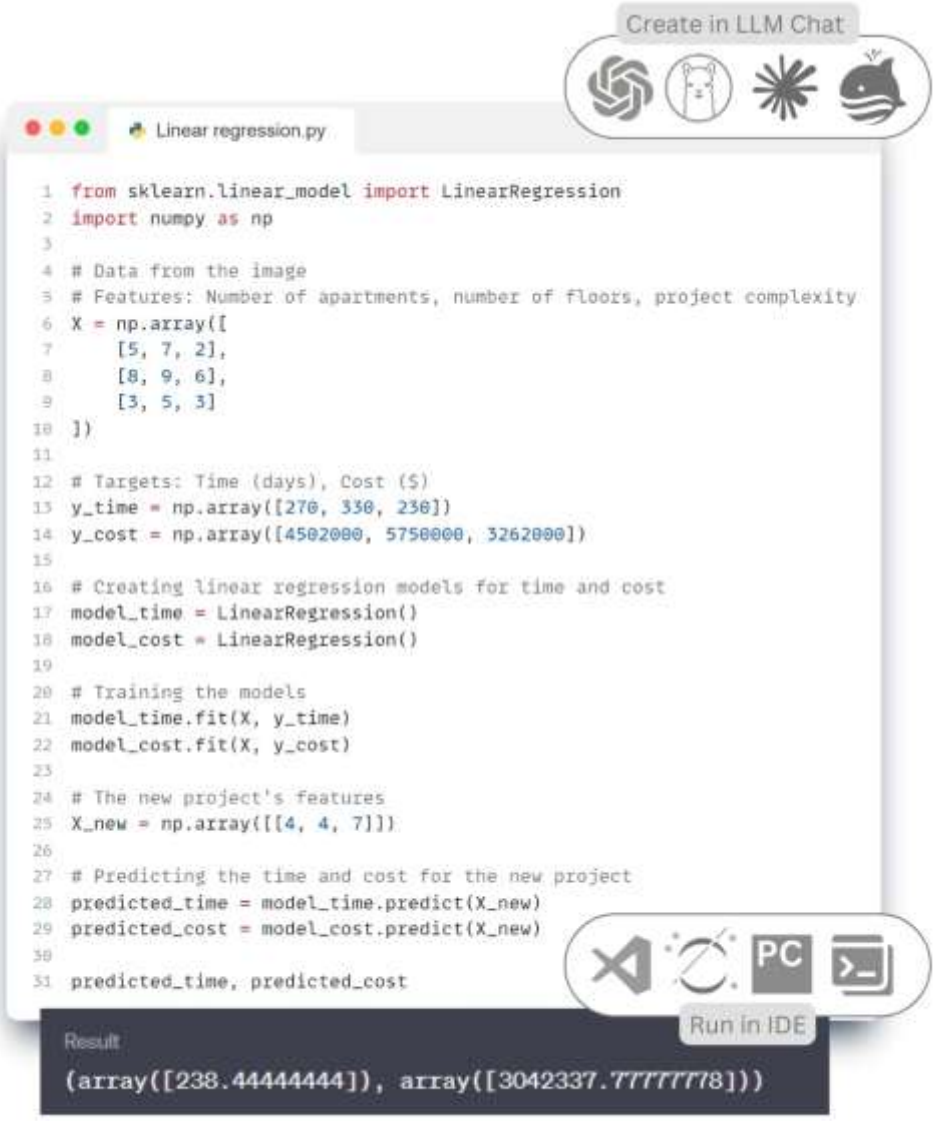
Đường này được biểu diễn bằng một phương trình, trong đó khi nhập giá trị của biến độc lập (X) sẽ cho ra giá trị dự đoán của biến phụ thuộc (Y). Quá trình này cho phép dự đoán hiệu quả Y dựa trên các giá trị X đã biết, sử dụng mối quan hệ tuyến tính giữa chúng. Ví dụ về việc tìm ra đường trung bình thống kê như vậy có thể thấy trong việc đánh giá dữ liệu về giấy phép xây dựng ở San Francisco (Hình 9.17), nơi mà lạm phát được tính toán bằng hồi quy tuyến tính cho các loại đối tượng khác nhau.

Hãy tải bảng dữ liệu của dự án (Hình 9.32 từ chương trước) trực tiếp vào LLM và yêu cầu nó xây dựng cho chúng ta một mô hình học máy đơn giản.

- Gửi yêu cầu văn bản vào trò chuyện LLM (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN:

> Cần phải thể hiện việc xây dựng một mô hình học máy đơn giản để dự đoán chi phí và thời gian thực hiện dự án mới X (Hình 9.32 như hình đính kèm) ⏎ -

- Phản hồi từ LLM:



Hình 9.34 LLM đã chọn hồi quy tuyến tính để tạo ra mô hình học máy cho việc dự đoán chi phí và thời gian của dự án.

LLM đã tự động nhận diện bảng từ hình ảnh đính kèm và chuyển đổi dữ liệu từ định dạng hình ảnh sang mảng bảng (Hình 9.34 – dòng 6). Mảng này đã được sử dụng làm cơ sở để tạo ra các đặc trưng và nhãn, từ đó mô hình học máy đã được xây dựng (Hình 9.34 – dòng 17-22), trong đó sử dụng hồi quy tuyến tính.--

Với mô hình hồi quy tuyến tính cơ bản, được huấn luyện trên một tập dữ liệu "cực kỳ nhỏ", đã đưa ra dự đoán cho một dự án xây dựng giả định mới, được ký hiệu là Dự án X. Trong bài toán của chúng ta, dự án này được đặc trưng bởi việc có 40 căn hộ, 4 tầng và mức độ phức tạp là 7 (Hình 9.32).

Theo các dự đoán được thực hiện bằng mô hình hồi quy tuyến tính, dựa trên tập dữ liệu hạn chế và nhỏ cho Dự án X mới (Hình 9.34 - dòng 24-29): -

- Thời gian xây dựng sẽ khoảng 238 ngày (238,4444444)
- Tổng chi phí sẽ khoảng $3 042 338 (3042337,777)

Để nghiên cứu thêm về giả thuyết chi phí của dự án, sẽ hữu ích khi thử nghiệm với các thuật toán và phương pháp học máy khác nhau. Do đó, chúng ta sẽ dự đoán các giá trị chi phí và thời gian tương tự cho Dự án X mới dựa trên một tập dữ liệu lịch sử nhỏ bằng cách sử dụng thuật toán K-Nearest Neighbours (k-NN).

Dự đoán chi phí và thời gian dự án bằng thuật toán K-nearest neighbor (k-NN)

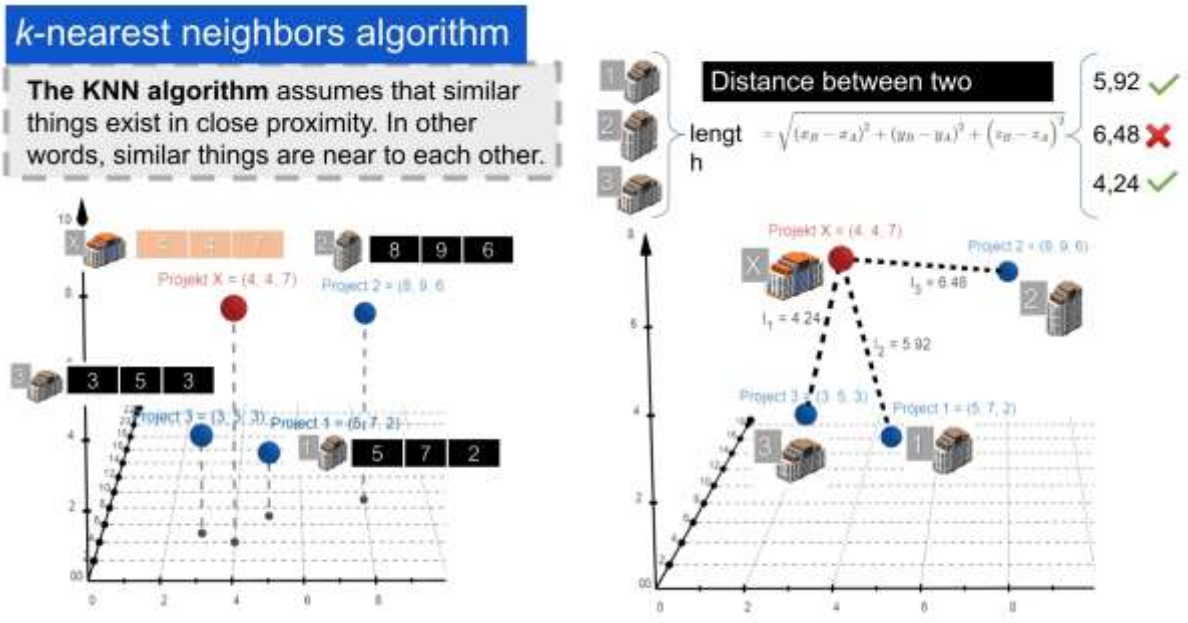
Là một dự đoán bổ sung để đánh giá chi phí và thời gian của dự án mới, chúng tôi sử dụng thuật toán k-Nearest Neighbours (k-NN). Thuật toán K-Nearest Neighbors (k-NN) là một phương pháp học máy có giám sát (supervised machine learning), được áp dụng cho cả phân loại và hồi quy. Thuật toán k-NN cũng đã được chúng tôi xem xét trước đó trong bối cảnh tìm kiếm trong cơ sở dữ liệu vector (Hình 8.22), nơi nó được sử dụng để tìm các vector gần nhất (ví dụ, văn bản, hình ảnh hoặc mô tả kỹ thuật). Trong phương pháp này, mỗi dự án được đại diện như một điểm trong không gian đa chiều, nơi mỗi chiều tương ứng với một thuộc tính cụ thể của dự án.-

Trong trường hợp của chúng tôi, với ba thuộc tính của mỗi dự án, chúng tôi sẽ đại diện chúng như những điểm trong không gian ba chiều (Hình 9.35). Do đó, dự án sắp tới của chúng tôi X sẽ được định vị trong không gian này với tọa độ (x=4, y=4, z=7). Cần lưu ý rằng trong điều kiện thực tế, số lượng điểm và kích thước không gian có thể lớn hơn nhiều.

> Thuật toán K-NN (k-nearest neighbors) hoạt động bằng cách đo khoảng cách giữa dự án mong muốn X và các dự án trong cơ sở dữ liệu huấn luyện. Bằng cách so sánh những khoảng cách này, thuật toán xác định các dự án gần nhất với điểm của dự án mới X.

Ví dụ, nếu dự án thứ hai (x=8, y=9, z=6) trong tập dữ liệu ban đầu của chúng tôi nằm xa hơn nhiều so với X (Hình 9.35) so với các dự án khác, nó có thể được loại trừ khỏi phân tích tiếp theo. Kết quả là, chỉ hai (k=2) dự án gần nhất có thể được sử dụng cho các phép tính, dựa trên đó sẽ xác định giá trị trung bình.

Phương pháp tương tự, thông qua việc tìm kiếm hàng xóm, cho phép đánh giá sự tương đồng giữa các dự án, điều này giúp đưa ra kết luận về chi phí và thời gian thực hiện dự án mới dựa trên các dự án tương tự đã được thực hiện trước đó.



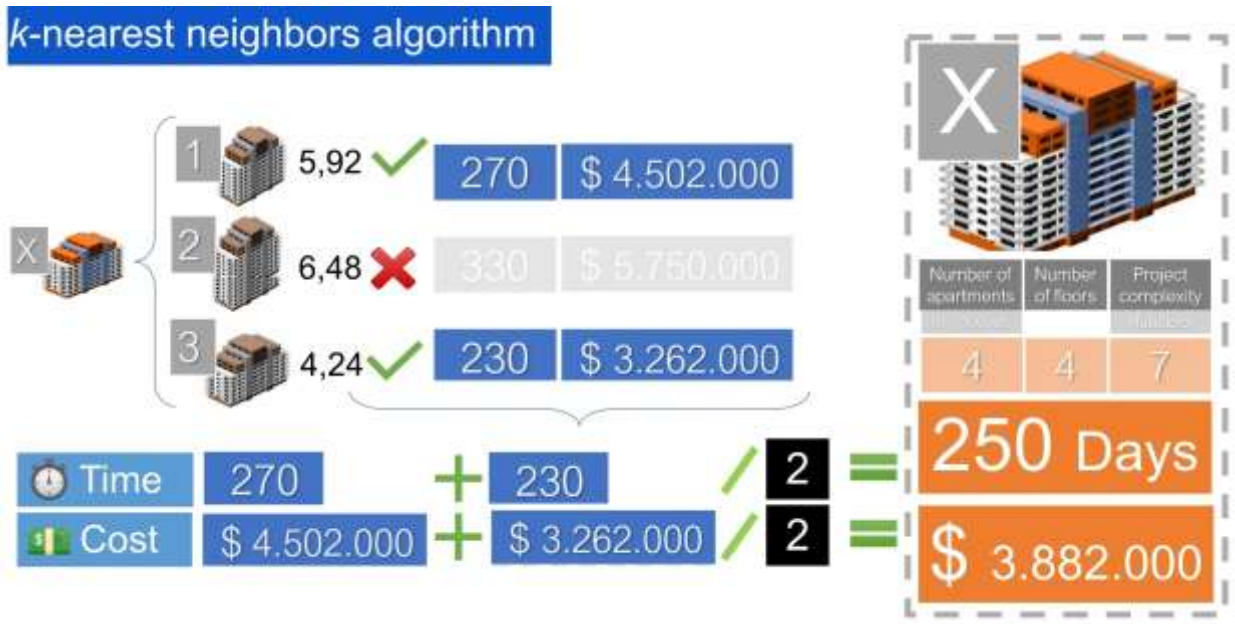
Hình 9.35 Trong thuật toán K-NN, các dự án được đại diện như những điểm trong không gian đa chiều, và để đánh giá sự tương đồng và dự đoán, các dự án gần nhất được chọn dựa trên khoảng cách.

Công việc của k-NN bao gồm một số bước chính:

- Chuẩn bị dữ liệu: trước tiên, tải các tập dữ liệu huấn luyện và kiểm tra. Dữ liệu huấn luyện được sử dụng để "đào tạo" thuật toán, trong khi dữ liệu kiểm tra được sử dụng để kiểm tra hiệu quả của nó.
- Chọn tham số K: chọn số K, chỉ ra số lượng hàng xóm gần nhất (các điểm dữ liệu) cần được xem xét trong thuật toán. Giá trị "K" rất quan trọng, vì nó ảnh hưởng đến kết quả.
- Quy trình phân loại và hồi quy cho dữ liệu kiểm tra:
  - Tính toán khoảng cách: cho mỗi phần tử trong dữ liệu kiểm tra, tính toán khoảng cách đến từng phần tử trong dữ liệu huấn luyện (Hình 9.35). Để làm điều này, có thể sử dụng các phương pháp đo khoảng cách khác nhau, chẳng hạn như khoảng cách Euclid (phương pháp phổ biến nhất), khoảng cách Manhattan hoặc khoảng cách Hamming.
  - Sắp xếp và chọn K hàng xóm gần nhất: sau khi tính toán khoảng cách, chúng được sắp xếp và chọn K điểm gần nhất với điểm kiểm tra.
  - Xác định lớp hoặc giá trị của điểm kiểm tra: nếu đây là bài toán phân loại, lớp của điểm kiểm tra được xác định dựa trên lớp xuất hiện nhiều nhất trong số K hàng xóm được chọn. Nếu đây là bài toán hồi quy, thì giá trị trung bình (hoặc một thước đo xu hướng trung tâm khác) của K hàng xóm sẽ được tính toán.
- Hoàn thành quá trình: ngay khi tất cả dữ liệu kiểm tra được phân loại hoặc dự đoán, quá trình sẽ được hoàn tất.

> Thuật toán k-nearest neighbors (k-NN) hiệu quả trong nhiều ứng dụng thực tiễn và là một trong những công cụ chính trong kho vũ khí của các chuyên gia học máy. Thuật toán này phổ biến nhờ vào sự đơn giản và hiệu quả của nó, đặc biệt trong các bài toán mà mối quan hệ giữa các dữ liệu dễ dàng được diễn giải.

Trong ví dụ của chúng tôi, sau khi áp dụng thuật toán K-b, hai dự án (từ mẫu nhỏ của chúng tôi) với khoảng cách nhỏ nhất đến dự án X đã được xác định (Hình 9.35). Dựa trên các dự án này, thuật toán xác định giá trị trung bình của giá cả và thời gian xây dựng của chúng. Sau khi phân tích (Hình 9.36), thuật toán, thông qua việc trung bình hóa các hàng xóm gần nhất, kết luận rằng dự án X sẽ có giá khoảng 3.800.000 đô la và sẽ mất khoảng 250 ngày.-



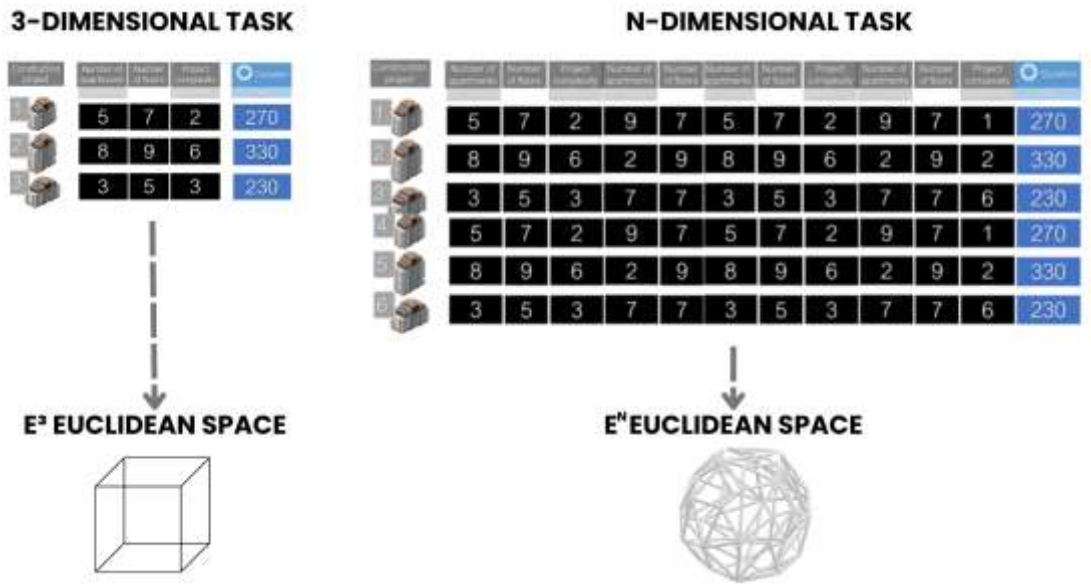
Hình 9.36 Thuật toán K-nearest neighbors xác định chi phí và lịch trình của dự án X, phân tích hai dự án gần nhất trong mẫu.

Thuật toán k-Nearest Neighbors (k-NN) đặc biệt phổ biến trong các bài toán phân loại và hồi quy, chẳng hạn như trong các hệ thống gợi ý, nơi nó được sử dụng để đề xuất hàng hóa hoặc nội dung dựa trên sở thích tương tự với sở thích của người dùng cụ thể. Hơn nữa, k-NN được sử dụng rộng rãi trong chẩn đoán y tế để phân loại các loại bệnh dựa trên triệu chứng của bệnh nhân, trong nhận dạng hình ảnh và trong lĩnh vực tài chính để đánh giá khả năng tín dụng của khách hàng.

> Ngay cả khi có một lượng dữ liệu hạn chế, các mô hình học máy vẫn có thể đưa ra dự đoán hữu ích và tăng cường đáng kể khả năng phân tích trong quản lý các dự án xây dựng. Khi mở rộng và làm sạch dữ liệu lịch sử, có thể chuyển sang các mô hình phức tạp hơn — chẳng hạn như xem xét loại công trình, vị trí, mùa bắt đầu xây dựng và các yếu tố khác.

Trong bài toán đơn giản của chúng tôi để trực quan hóa trong không gian ba chiều, ba thuộc tính đã được sử dụng, nhưng các dự án thực tế, trung bình, bao gồm hàng trăm hoặc hàng nghìn thuộc tính (xem tập dữ

liệu trong chương "Ví dụ về dữ liệu lớn dựa trên dữ liệu CAD (BIM)"), điều này làm tăng đáng kể độ lớn của không gian và độ phức tạp trong việc đại diện các dự án dưới dạng vector (Hình 9.37).-



Hình 9.37 Trong ví dụ đơn giản, ba thuộc tính đã được sử dụng cho trực quan hóa 3D, trong khi các dự án thực tế có số lượng lớn hơn.

Việc áp dụng các thuật toán khác nhau cho cùng một tập dữ liệu cho dự án X, trong đó có 40 căn hộ, 4 tầng và mức độ phức tạp là 7, đã cho ra các giá trị dự đoán khác nhau. Thuật toán hồi quy tuyến tính dự đoán thời gian hoàn thành là 238 ngày và chi phí là 3.042.338 đô la (Hình 9.34), trong khi thuật toán k-NN dự đoán 250 ngày và 3.882.000 đô la (Hình 9.36).--

Độ chính xác của các dự đoán được thực hiện bằng các mô hình học máy phụ thuộc trực tiếp vào khối lượng và chất lượng của dữ liệu đầu vào. Càng nhiều dự án tham gia vào quá trình huấn luyện, và càng đầy đủ và chính xác các đặc điểm (đặc trưng) và kết quả (nhãn) của chúng được trình bày, thì xác suất nhận được các dự đoán đáng tin cậy với các giá trị sai số tối thiểu càng cao.

Các phương pháp xử lý dữ liệu sơ bộ đóng vai trò quan trọng trong quá trình này, bao gồm:

- Chuẩn hóa, cho phép đưa các đặc trưng về cùng một thang đo;
- Phát hiện và loại bỏ các điểm ngoại lai, không làm sai lệch mô hình;
- Mã hóa các thuộc tính phân loại, cho phép làm việc với dữ liệu văn bản;
- Điền vào các giá trị bị thiếu, tăng cường độ ổn định của mô hình.

Ngoài ra, để đánh giá khả năng tổng quát của mô hình và sự ổn định của nó đối với các bộ dữ liệu mới, các phương pháp kiểm tra chéo được sử dụng để phát hiện hiện tượng học thuộc quá mức và nâng cao độ tin cậy của dự đoán.

> *Hỗn loạn là trật tự cần được giải mã.— José Saramago, "Người đôi"*

Ngay cả khi bạn cảm thấy rằng sự hỗn loạn của các nhiệm vụ của bạn không thể được mô tả một cách chính thức, hãy biết rằng bất kỳ sự kiện nào trong thế giới và đặc biệt là các quy trình xây dựng đều tuân theo các quy luật toán học, mà có thể cần hỗ trợ tính toán giá trị không thông qua các công thức nghiêm ngặt mà bằng cách sử dụng thống kê và dữ liệu lịch sử.

Cả các phương pháp ước lượng truyền thống do các phòng kế toán thực hiện và các mô hình machine learning đều gặp phải sự không chắc chắn và các nguồn sai sót tiềm ẩn. Tuy nhiên, với đủ dữ liệu chất lượng, các mô hình machine learning có thể thể hiện độ chính xác dự đoán tương đương, thậm chí đôi khi cao hơn so với các ước lượng của chuyên gia.

> Học máy, có khả năng cao, sẽ trở thành một công cụ bổ trợ đáng tin cậy trong phân tích, cho phép: làm rõ các phép toán, đề xuất các kịch bản thay thế và phát hiện các mối quan hệ tiềm ẩn giữa các tham số của dự án. Các mô hình tương tự sẽ không có tham vọng phổ quát, nhưng sẽ sớm chiếm một vị trí quan trọng trong các phép tính và quy trình ra quyết định dự án. Công nghệ học máy sẽ không loại bỏ sự tham gia của kỹ sư, chuyên viên dự toán và nhà phân tích, mà ngược lại, sẽ mở rộng khả năng của họ, cung cấp một góc nhìn bổ sung dựa trên dữ liệu lịch sử.

Khi được tích hợp một cách hiệu quả vào các quy trình kinh doanh của các công ty xây dựng, học máy có tiềm năng trở thành một yếu tố quan trọng trong hệ thống hỗ trợ quyết định quản lý — không phải là một sự thay thế cho con người, mà là một sự mở rộng cho trực giác chuyên môn và logic kỹ thuật của họ.

Các bước tiếp theo: từ lưu trữ đến phân tích và dự đoán.

Các phương pháp hiện đại trong công việc với dữ liệu đang bắt đầu thay đổi các nguyên tắc ra quyết định trong ngành xây dựng. Việc chuyển từ các đánh giá mang tính trực giác sang phân tích dữ liệu khách quan không chỉ nâng cao độ chính xác mà còn mở ra những cơ hội mới để tối ưu hóa quy trình. Tóm tắt lại phần này, cần nêu rõ các bước thực tiễn chính có thể giúp áp dụng các phương pháp đã xem xét vào các nhiệm vụ hàng ngày của bạn:

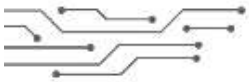
- Xây dựng hạ tầng lưu trữ dữ liệu bền vững.
  - Hãy cố gắng kết hợp các tài liệu rời rạc và dữ liệu dự án thành một mô hình bảng duy nhất, tổng hợp thông tin chính trong một dataframe để phân tích tiếp theo.
  - Sử dụng các định dạng lưu trữ dữ liệu hiệu quả — chẳng hạn như định dạng cột như Apache Parquet thay vì CSV hoặc XLSX — đặc biệt cho những bộ dữ liệu có thể được sử dụng trong tương lai cho việc huấn luyện các mô hình học máy.
  - Tạo ra một hệ thống phiên bản dữ liệu cho phép theo dõi các thay đổi trong suốt dự án.
- Triển khai các công cụ phân tích và tự động hóa.
  - Bắt đầu phân tích dữ liệu lịch sử của các dự án — theo tài liệu, mô hình, dự toán — để phát hiện các quy luật, xu hướng và bất thường.
  - Làm quen với các quy trình ETL (Extract, Transform, Load) để tự động tải lên và chuẩn bị dữ liệu.
  - Học cách trực quan hóa các chỉ số chính bằng cách sử dụng các thư viện Python miễn phí cho

việc trực quan hóa.

- [ ] Bắt đầu áp dụng các phương pháp thống kê và mẫu ngẫu nhiên để có được các kết luận phân tích đại diện và có thể tái tạo.

- Nâng cao độ trưởng thành trong công việc với dữ liệu.
  - [ ] Nghiên cứu một số thuật toán học máy cơ bản qua các ví dụ đơn giản và dễ hiểu, như bộ dữ liệu Titanic.
  - [ ] Phân tích các quy trình hiện tại và xác định nơi có thể chuyển từ logic nguyên nhân - kết quả cứng nhắc sang các phương pháp thống kê dự đoán và đánh giá.
  - [ ] Bắt đầu xem dữ liệu như một tài sản chiến lược, không phải là sản phẩm phụ: xây dựng các quy trình ra quyết định dựa trên các mô hình dữ liệu, không phải xung quanh các giải pháp phần mềm cụ thể.

Các công ty xây dựng nhận ra giá trị của dữ liệu đang bước vào một giai đoạn phát triển mới, nơi lợi thế cạnh tranh không được xác định bởi khối lượng tài nguyên mà bởi tốc độ ra quyết định dựa trên phân tích.



TỐI ĐA TIỆN ÍCH VỚI BẢN IN

Bạn đang cầm trên tay phiên bản kỹ thuật số miễn phí của Data-Driven Construction. Để làm việc thuận tiện hơn và truy cập nhanh chóng vào tài liệu, chúng tôi khuyên bạn nên chú ý đến ấn phẩm in:



- Luôn sẵn sàng: cuốn sách ở định dạng in sẽ trở thành công cụ làm việc đáng tin cậy, cho phép bạn nhanh chóng tìm kiếm và sử dụng các hình ảnh và biểu đồ cần thiết trong bất kỳ tình huống làm việc nào.
- Chất lượng hình ảnh cao: tất cả hình ảnh và đồ họa trong ấn phẩm in được trình bày với chất lượng tốt nhất.
- Truy cập nhanh chóng vào thông tin: điều hướng thuận tiện, khả năng ghi chú, đánh dấu và làm việc với cuốn sách ở bất kỳ đâu.

Khi bạn mua phiên bản in đầy đủ của cuốn sách, bạn nhận được công cụ tiện lợi cho việc làm việc thoải mái và hiệu quả với thông tin: khả năng sử dụng nhanh chóng các tài liệu hình ảnh trong công việc hàng ngày, nhanh chóng tìm thấy các sơ đồ cần thiết và ghi chú. Ngoài ra, việc mua hàng của bạn hỗ trợ việc phổ biến kiến thức mở.

Bạn có thể đặt hàng phiên bản in của cuốn sách tại: datadrivenconstruction.io/books



X PHẦN **NGÀNH XÂY DỰNG TRONG THỜI ĐẠI DỮ LIỆU SỐ. CƠ HỘI VÀ THÁCH THỨC.**

Phần kết thúc thứ mười cung cấp cái nhìn tổng thể về tương lai của ngành xây dựng trong thời đại chuyển đổi số. Tại đây, phân tích sự chuyển đổi từ phân tích nguyên nhân - kết quả sang làm việc với các tương quan của dữ liệu lớn. Các mối tương quan giữa sự tiến hóa của nghệ thuật hình ảnh và sự phát triển của công việc với dữ liệu trong xây dựng được đưa ra, cho thấy ngành này đang chuyển từ kiểm soát chi tiết sang hiểu biết toàn diện về các quy trình. Khái niệm "uber hóa" ngành xây dựng được xem xét, nơi tính minh bạch của dữ liệu và tự động hóa các phép tính có thể thay đổi một cách triệt để các mô hình kinh doanh truyền thống, loại bỏ sự cần thiết của các trung gian và giảm khả năng đầu cơ. Các vấn đề chưa được giải quyết, chẳng hạn như phân loại phổ quát các yếu tố, được thảo luận chi tiết, cho các công ty xây dựng thời gian để thích ứng với các điều kiện mới. Phần này kết thúc với các khuyến nghị cụ thể về việc hình thành chiến lược chuyển đổi số, bao gồm phân tích các điểm yếu và mở rộng phạm vi dịch vụ để duy trì tính cạnh tranh trong một ngành đang thay đổi.

CHƯƠNG 10.1.
Chiến lược sinh tồn: Hình thành lợi thế cạnh tranh

Sự tương quan thay vì tính toán: tương lai của phân tích xây dựng.

Do sự thúc đẩy nhanh chóng của việc số hóa thông tin, ngành xây dựng hiện đại đang trải qua một sự chuyển hóa sâu sắc, trong đó dữ liệu không chỉ đơn thuần là công cụ mà còn là tài sản chiến lược có khả năng thay đổi mạnh mẽ các phương pháp quản lý dự án và doanh nghiệp truyền thống.

Trong suốt hàng ngàn năm, hoạt động xây dựng dựa vào các phương pháp xác định — tính toán chính xác, chi tiết hóa và kiểm soát nghiêm ngặt các thông số. Trong những thế kỷ đầu của kỷ nguyên Công Nguyên, các kỹ sư La Mã đã áp dụng các nguyên tắc toán học để xây dựng các công trình như aqueduct và cầu. Trong thời Trung Cổ, các kiến trúc sư cố gắng đạt được tỷ lệ lý tưởng cho các nhà thờ Gô-tích, và trong kỷ nguyên công nghiệp thế kỷ 20, các hệ thống tiêu chuẩn hóa và quy định đã được hình thành, trở thành nền tảng của xây dựng hàng loạt.

Ngày nay, hướng phát triển đang chuyển từ việc tìm kiếm các mối quan hệ nguyên nhân - kết quả rõ ràng sang phân tích xác suất, tìm kiếm các tương quan và quy luật tiềm ẩn. Ngành xây dựng bước vào một giai đoạn mới — dữ liệu trở thành tài nguyên thiết yếu, và phân tích dựa trên chúng đang thay thế các phương pháp trực giác và tối ưu hóa cục bộ.



Hình 10.11 Tiềm năng ẩn giấu của dữ liệu xây dựng: các phép tính hiện có trong công ty chỉ là phần nổi của tảng băng, có thể truy cập cho quản lý để phân tích.

> Hệ thống thông tin của công ty giống như một tảng băng: quản lý công ty chỉ nhìn thấy một phần nhỏ tiềm năng của dữ liệu, trong khi giá trị chủ yếu bị ẩn giấu sâu bên dưới. Quan trọng là đánh giá dữ liệu không chỉ theo cách sử dụng hiện tại của chúng mà còn theo các cơ hội mà chúng sẽ mở ra trong tương lai. Chỉ những công ty nào biết cách khai thác các quy luật ẩn giấu và tạo ra kiến thức mới từ dữ liệu mới có thể xây dựng được lợi thế cạnh tranh bền vững.

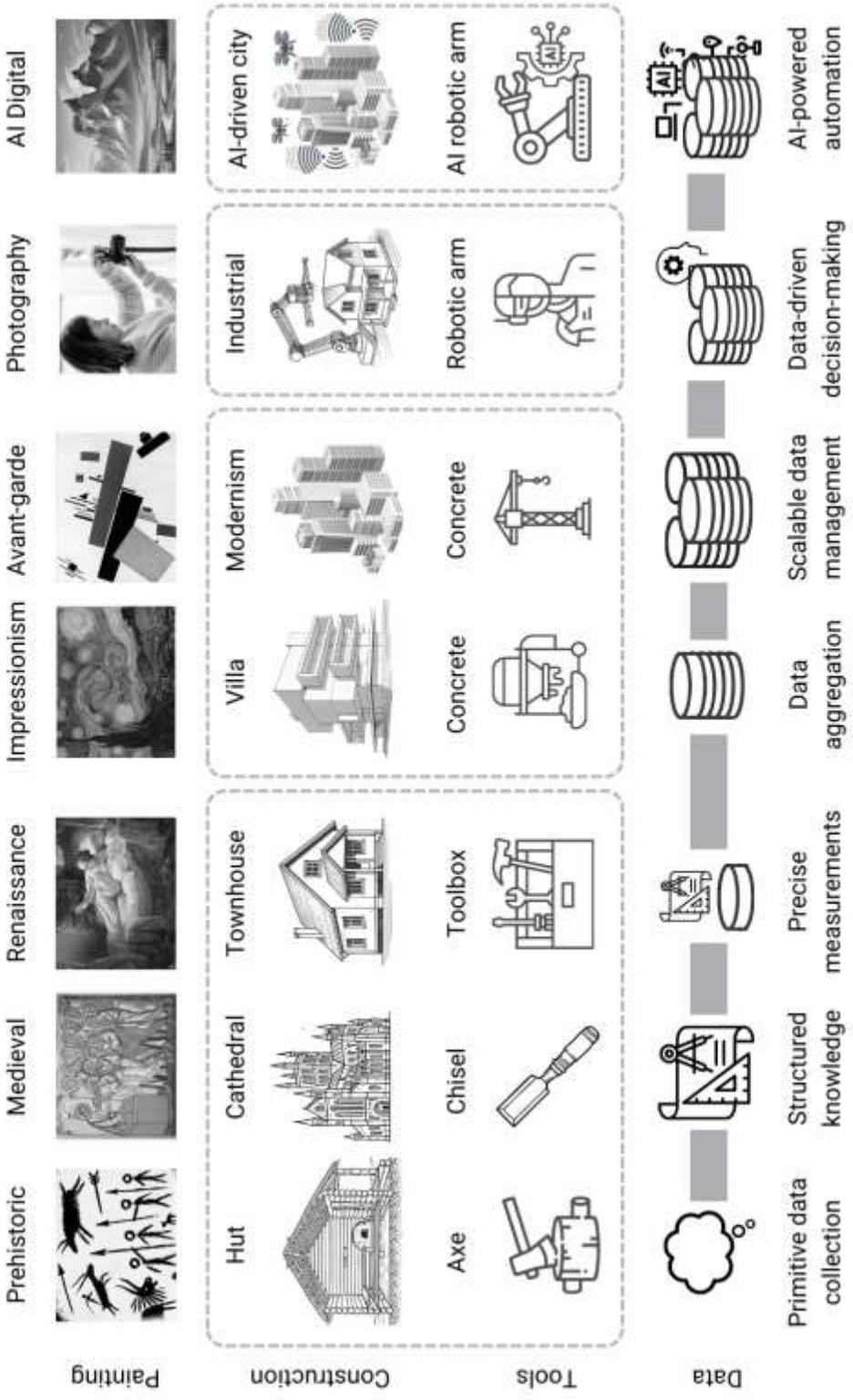
Tìm kiếm các quy luật ẩn giấu và suy ngẫm dữ liệu không chỉ đơn thuần là làm việc với số liệu, mà còn là một quá trình sáng tạo yêu cầu tư duy trừu tượng và khả năng nhìn ra bức tranh toàn cảnh từ những yếu tố rời rạc. Trong ý nghĩa này, sự phát triển của công việc với dữ liệu có thể so sánh với sự tiến hóa của nghệ thuật thị giác.

Sự phát triển của ngành xây dựng thật sự giống với sự tiến bộ của nghệ thuật thị giác. Trong cả hai trường hợp, nhân loại đã trải qua hành trình từ các phương pháp nguyên thủy đến các công nghệ phức tạp trong việc hình dung và phân tích. Trong thời kỳ đồ đá, con người đã sử dụng tranh vẽ trên đá và các công cụ thô sơ để giải quyết các vấn đề hàng ngày. Trong thời Trung cổ và thời kỳ Phục hưng, mức độ phức tạp trong kiến trúc và nghệ thuật đã tăng lên đáng kể. Đến đầu thời kỳ Trung cổ, các công cụ xây dựng đã tiến hóa từ cái rìu đơn giản đến các bộ công cụ phong phú, biểu thị sự gia tăng kiến thức kỹ thuật.

Thế kỷ hiện thực trở thành cuộc cách mạng đầu tiên trong nghệ thuật thị giác: các nghệ sĩ học cách tái tạo những chi tiết nhỏ nhất, đạt được độ chân thực tối đa trong hình ảnh. Trong xây dựng, sự tương tự của giai đoạn này là các phương pháp kỹ thuật chính xác, các bản vẽ chi tiết và các phép tính nghiêm ngặt, đã trở thành nền tảng của thực hành thiết kế suốt hàng thế kỷ.

> Muộn hơn, chủ nghĩa ấn tượng đã thay đổi cách cảm nhận hiện thực nghệ thuật: thay vì chuyển tải hình thức một cách chính xác, các nghệ sĩ bắt đầu ghi lại tâm trạng, ánh sáng và động lực, nhằm phản ánh ấn tượng tổng thể thay vì độ chính xác tuyệt đối. Tương tự, học máy trong phân tích xây dựng đã chuyển từ các mô hình logic cứng nhắc sang nhận diện các mẫu và quy luật xác suất, cho phép "nhìn thấy" các mối liên hệ ẩn trong dữ liệu, điều mà phân tích cổ điển không thể tiếp cận được. Cách tiếp cận này gợi nhớ đến những ý tưởng về chủ nghĩa tối giản và chức năng của Bauhaus, nơi mà ý nghĩa (chức năng) quan trọng hơn hình thức. Bauhaus tìm cách loại bỏ những thứ không cần thiết, từ bỏ trang trí vì sự rõ ràng, tính tiện dụng và tính đại trà. Các vật phẩm phải rõ ràng và hữu ích, không có sự thừa thãi — thẩm mỹ được sinh ra từ logic của cấu trúc và mục đích.

Với sự xuất hiện của nhiếp ảnh vào cuối thế kỷ 19, nghệ thuật đã nhận được một công cụ mới để ghi lại thực tế với độ chính xác chưa từng có và đã đảo lộn cách nhìn nhận về nghệ thuật hình họa. Tương tự, trong xây dựng, cuộc cách mạng công nghiệp thế kỷ 21 dẫn đến việc áp dụng các công nghệ robot, laser, IoT, RFID và các khái niệm như "Xây dựng Kết nối", nơi việc thu thập các thông số riêng lẻ đã phát triển thành việc ghi lại sự thật quy mô lớn và thông minh về toàn bộ hiện trường xây dựng.



Hình 10.12 Các thời kỳ tiến hóa của nghệ thuật hình họa tương ứng với sự phát triển trong cách tiếp cận làm việc với dữ liệu trong ngành xây dựng.

Ngày nay, cũng như nghệ thuật hình họa đang trải qua sự tái định hình với sự xuất hiện của các công cụ AI và LLM, ngành xây dựng đang trải qua một bước nhảy lượng tử khác: các hệ thống thông minh được điều khiển bởi trí tuệ nhân tạo (AI), các trò chuyện LLM cho phép dự đoán, tối ưu hóa và tạo ra các giải pháp với sự can thiệp tối thiểu từ con người.

Vai trò của dữ liệu trong thiết kế và quản lý đã thay đổi một cách triệt để. Nếu trước đây kiến thức được truyền tải bằng lời nói và mang tính chất kinh nghiệm — giống như trước thế kỷ 19, thực tế được ghi lại bằng các bức tranh vẽ tay — thì ngày nay, sự chú ý tập trung vào việc ghi nhận đầy đủ một "bức tranh" xây dựng trong không gian số. Thông qua các thuật toán học máy, bức tranh số này được chuyển đổi thành một hình ảnh ấn tượng về thực tại xây dựng — không phải bản sao chính xác, mà là một sự hiểu biết tổng quát và xác suất về các quá trình.

> Chúng ta đang tiến gần đến một kỷ nguyên mà các quy trình thiết kế, xây dựng và vận hành tòa nhà không chỉ được bổ sung, mà còn bị điều khiển một cách đáng kể bởi các hệ thống trí tuệ nhân tạo. Tương tự như nghệ thuật số hiện đại được tạo ra mà không cần cọ — thông qua các yêu cầu văn bản và các mô hình sinh tạo, các giải pháp kiến trúc và kỹ thuật trong tương lai sẽ được hình thành dựa trên các yêu cầu và tham số chính được người dùng đặt ra.

Trong thế kỷ XXI, việc truy cập vào dữ liệu, sự diễn giải và chất lượng phân tích trở thành những điều kiện không thể thiếu cho sự thành công của dự án. Hơn nữa, giá trị của dữ liệu không được xác định bởi khối lượng của chúng, mà bởi khả năng của các chuyên gia trong việc phân tích, kiểm tra và biến chúng thành hành động.

Cách tiếp cận dựa trên dữ liệu trong xây dựng: hạ tầng cấp độ mới

Trong lịch sử nhân loại, mỗi bước nhảy công nghệ như vậy đều mang lại những thay đổi cơ bản cho nền kinh tế và xã hội. Hôm nay, chúng ta đang chứng kiến một làn sóng chuyển mình mới, có quy mô có thể so sánh với cách mạng công nghiệp thế kỷ 19. Tuy nhiên, trong khi một trăm năm trước, động lực chính của sự thay đổi là sức mạnh cơ khí và công nghệ năng lượng, thì bây giờ – đó là dữ liệu và trí tuệ nhân tạo.

Học máy, LLM và các tác nhân AI đang thay đổi bản chất của các ứng dụng, khiến cho các công nghệ phần mềm truyền thống (như đã thảo luận trong phần thứ hai của cuốn sách) trở nên không cần thiết. Tất cả logic làm việc với dữ liệu được tập trung vào các tác nhân AI, chứ không phải vào các quy tắc kinh doanh được mã hóa cứng.--

> Trong thời đại dữ liệu, những quan niệm truyền thống về ứng dụng đang biến đổi một cách căn bản. Chúng ta đang tiến tới một mô hình mà những hệ thống mô-đun doanh nghiệp cồng kềnh sẽ không thể tránh khỏi việc nhường chỗ cho những giải pháp nhẹ, chuyên biệt và mở.

> *Trong tương lai, chỉ còn lại cấu trúc dữ liệu cơ bản, và tất cả sự tương tác với nó sẽ diễn ra thông qua các tác nhân làm việc trực tiếp với cơ sở dữ liệu. Tôi thực sự tin rằng toàn bộ công nghệ ứng dụng sẽ biến mất, vì không còn cần thiết khi trí tuệ nhân tạo tương tác trực tiếp với cơ sở dữ liệu chính. Trong suốt sự nghiệp của mình, tôi đã làm việc trong lĩnh vực SaaS – đã tạo ra các công ty, làm việc tại đó, và nếu nói thật, bây giờ có lẽ tôi sẽ không bắt đầu một doanh nghiệp SaaS mới. Và có lẽ tôi cũng sẽ không đầu tư vào các công ty SaaS ngay bây giờ. Tình hình quá không chắc chắn. Điều này không có nghĩa trong tương lai sẽ không còn các công ty phần mềm, chỉ là chúng sẽ mang hình thức hoàn toàn khác. Các hệ thống tương lai sẽ là cơ sở dữ liệu với logic kinh doanh được chuyển ra ngoài cho các tác nhân AI. Những tác nhân này sẽ làm việc với nhiều kho dữ liệu cùng một lúc, không bị giới hạn chỉ trong một cơ sở dữ liệu duy nhất. Tất cả logic sẽ được chuyển lên cấp độ trí tuệ nhân tạo.– Matthew Berman, CEO Forward Future*

Sự khác biệt chính của mô hình mới là sự tối thiểu hóa gánh nặng công nghệ. Thay vì những bộ phần mềm phức tạp và đóng kín đồ sộ, chúng ta sẽ có những mô-đun linh hoạt, mở và dễ dàng tùy chỉnh, sống ngay trong dòng dữ liệu. Kiến trúc quản lý quy trình tương lai dự kiến sẽ sử dụng các ứng dụng vi mô - những công cụ nhỏ gọn, mục tiêu, khác biệt hoàn toàn với các hệ thống ERP, PMIS, CDE, CAFM khổng lồ và đóng kín. Những tác nhân mới sẽ phải thích ứng tối đa, tích hợp và định hướng vào những nhiệm vụ kinh doanh cụ thể (Ví dụ như Low-Code/No-Code).-

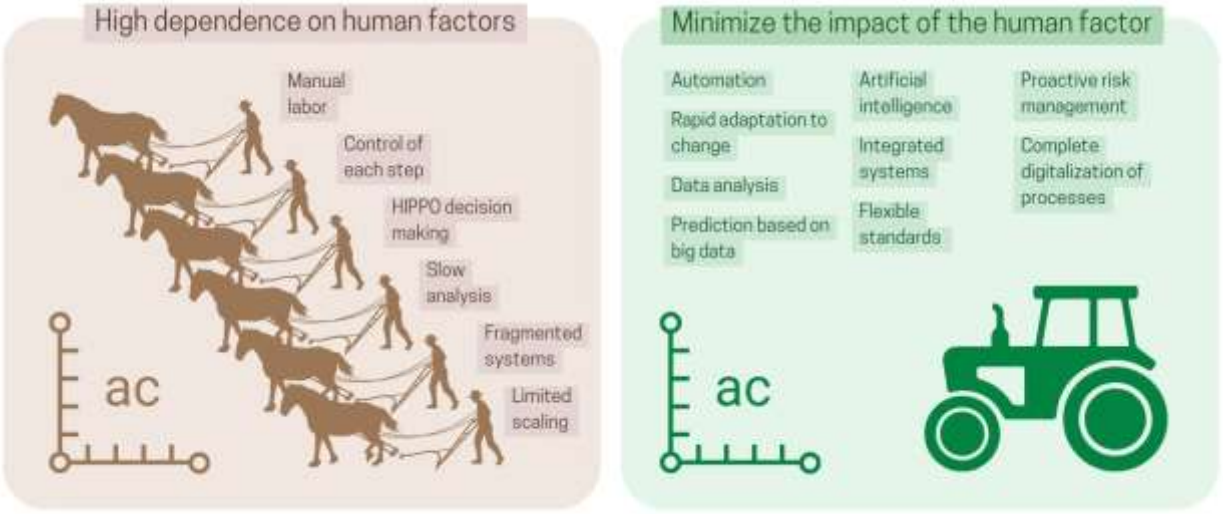
> *Tất cả logic kinh doanh sẽ chuyển sang các tác nhân [AI] này, và các tác nhân này sẽ thực hiện các thao tác CRUD [Tạo, Đọc, Cập nhật và Xóa] trong nhiều kho dữ liệu, nghĩa là chúng sẽ không phân biệt xem backend cụ thể nào đang được sử dụng. Chúng sẽ cập nhật nhiều cơ sở dữ liệu, và toàn bộ logic sẽ nằm ở cái gọi là cấp độ AI. Và khi cấp độ AI trở thành nơi chứa toàn bộ logic, con người sẽ bắt đầu thay thế các backend. Chúng ta đã quan sát thấy một tỷ lệ chiến thắng khá cao trên thị trường backend Dynamics và việc sử dụng các tác nhân, và chúng ta sẽ tích cực hướng tới điều này, cố gắng gom tất cả lại. Dù là trong lĩnh vực dịch vụ khách hàng hay trong các lĩnh vực khác, chẳng hạn như không chỉ CRM, mà còn cả các giải pháp tài chính và hoạt động của chúng tôi. Bởi vì mọi người muốn có nhiều ứng dụng kinh doanh dựa trên AI, nơi mà cấp độ logic có thể được quản lý bởi AI và các tác nhân AI. Một trong những điều thú vị nhất đối với tôi là Excel với Python, có thể so sánh với GitHub với Copilot. Nghĩa là, điều chúng tôi đã làm: bây giờ khi bạn có Excel, bạn chỉ cần mở nó ra, khởi động Copilot và bắt đầu chơi với nó. Nó không chỉ đơn thuần là hiểu các con số có sẵn — nó sẽ tự động lập kế hoạch. Giống như GitHub Copilot Workspace tạo ra một kế hoạch và sau đó thực hiện nó, điều này cũng giống như công việc của một nhà phân tích dữ liệu sử dụng Excel như là một công cụ trực quan hóa dòng và cột để phân tích. Như vậy, Copilot sử dụng Excel như là một công cụ với tất cả các khả năng của nó, bởi vì nó có thể tạo ra dữ liệu và có một trình thông dịch Python.— Satya Nadella, CEO, Microsoft, phỏng vấn kênh BG2 tháng 12 năm 2024.*

Chuyển biến mà chúng ta đang chứng kiến trong logic của các ứng dụng văn phòng — việc chuyển từ các hệ thống mô-đun, khép kín sang các tác nhân AI hoạt động trực tiếp với dữ liệu mở — chỉ là một phần của một quá trình quy mô lớn hơn nhiều. Không chỉ là sự thay đổi giao diện hoặc kiến trúc phần mềm: những

thay đổi này sẽ tác động đến các nguyên tắc cơ bản của tổ chức lao động, ra quyết định và quản lý kinh doanh. Trong xây dựng, điều này sẽ dẫn đến việc hình thành logic dựa trên dữ liệu, trong đó dữ liệu sẽ trở thành yếu tố trung tâm của các quy trình — từ thiết kế đến quản lý tài nguyên và kiểm soát tiến độ xây dựng.

Văn phòng kỹ thuật số thế hệ tiếp theo: cách AI thay đổi không gian làm việc

Gần một thế kỷ trước, nhân loại đã trải qua một cuộc cách mạng công nghệ tương tự. Việc chuyển từ động cơ hơi nước sang động cơ điện đã mất hơn bốn thập kỷ, nhưng cuối cùng đã trở thành chất xúc tác cho sự tăng trưởng năng suất chưa từng có — chủ yếu nhờ vào sự phân quyền của năng lượng và tính linh hoạt của các giải pháp mới. Sự thay đổi này không chỉ thay đổi dòng lịch sử, mà còn chuyển phần lớn dân số từ các làng đến các thành phố, và đặt nền tảng cho nền kinh tế hiện đại. Lịch sử công nghệ là hành trình từ lao động thể chất sang tự động hóa và hệ thống thông minh. Cũng giống như máy kéo thay thế hàng chục nông dân, các công nghệ số hiện đại đang thay thế các phương pháp quản lý xây dựng truyền thống (Hình. 10.13). Còn vào đầu thế kỷ XX, đại đa số dân số Trái đất đã làm việc trên cánh đồng một cách thủ công, cho đến khi vào những năm 1930, việc cơ giới hóa lao động bắt đầu diễn ra với sự trợ giúp của máy móc và máy kéo.



Hình. 10.13 Như máy kéo đã thay thế hàng chục người vào đầu thế kỷ 20, thì máy học sẽ thay thế các phương pháp quản lý doanh nghiệp và dự án truyền thống trong thế kỷ 21.

Cũng như nhân loại một trăm năm trước đã chuyển từ việc canh tác các mảnh đất riêng lẻ bằng những công cụ thô sơ sang nông nghiệp quy mô lớn sử dụng máy móc, hôm nay chúng ta đang thực hiện sự chuyển đổi từ việc xử lý các "siêu dữ liệu" thông tin rời rạc sang làm việc với những khối lượng dữ liệu lớn bằng cách sử dụng "máy kéo" mạnh mẽ - ETL-pipeline và các thuật toán trí tuệ nhân tạo.

Chúng ta đang đứng trước một bước nhảy vọt tương tự - nhưng ở lĩnh vực số: từ việc quản lý doanh nghiệp truyền thống, thủ công sang các mô hình dựa trên dữ liệu.

Con đường hướng tới kiến trúc dữ liệu đầy đủ sẽ đòi hỏi thời gian, đầu tư và nỗ lực tổ chức. Nhưng con đường này mở ra khả năng không chỉ cải tiến dần dần mà còn là một bước nhảy chất lượng - hướng tới hiệu quả, tính minh bạch và khả năng quản lý cao hơn trong các quá trình xây dựng. Tất cả điều này - với điều kiện áp dụng hệ thống các công cụ số và từ bỏ các thực tiễn kinh doanh lỗi thời.

> Việc tham số hóa các nhiệm vụ, ETL, LLM, các thành phần IoT, RFID, mã hóa, dữ liệu lớn và học máy sẽ biến đổi ngành xây dựng truyền thống thành xây dựng dựa trên dữ liệu, nơi mỗi chi tiết của dự án và doanh nghiệp xây dựng sẽ được kiểm soát và tối ưu hóa bằng dữ liệu.

Trước đây, để phân tích thông tin cần hàng ngàn giờ công sức. Giờ đây, những nhiệm vụ này được thực hiện bởi các thuật toán và LLM, biến đổi các khối dữ liệu rời rạc thành các nguồn chiến lược thông qua các thông điệp. Trong thế giới công nghệ đang xảy ra điều tương tự như những gì đã xảy ra với nông nghiệp: từ cái cuốc, chúng ta chuyển sang các tổ hợp nông nghiệp tự động. Cũng giống như công việc văn phòng trong xây dựng - từ các tệp Excel và báo cáo thủ công - chuyển sang hệ thống thông minh, nơi dữ liệu được thu thập, làm sạch, cấu trúc và biến thành thông tin chi tiết.

Ngày nay, các công ty cần bắt đầu "cày cấy" các cánh đồng thông tin thông qua việc thu thập dữ liệu chất lượng và cấu trúc thông tin, và "bón phân" cho chúng bằng các công cụ làm sạch và chuẩn hóa, sau đó "thu hoạch" - dưới dạng phân tích dự đoán và các giải pháp tự động hóa. Nếu một nông dân hiện đại với máy móc có thể thay thế hàng trăm người làm đất, thì các thuật toán thông minh cũng có thể loại bỏ sự nhàm chán khỏi công việc của nhân viên và chuyển họ sang vai trò quản lý chiến lược dòng thông tin.

Tuy nhiên, điều quan trọng là phải hiểu rằng việc xây dựng một tổ chức thực sự dựa trên dữ liệu - là một quá trình không nhanh chóng. Đây là một định hướng chiến lược dài hạn, giống như việc tạo ra một khu vực mới để trồng một khu rừng mới, nơi mỗi "cây" trong hệ sinh thái này - là một quy trình, năng lực hoặc công cụ riêng biệt, cần thời gian để phát triển và lớn lên. Và như trong trường hợp của một khu rừng thực sự, thành công không chỉ phụ thuộc vào chất lượng giống cây trồng (công nghệ), mà còn phụ thuộc vào đất (văn hóa doanh nghiệp), khí hậu (môi trường kinh doanh) và chế độ chăm sóc (cách tiếp cận hệ thống).-

Các công ty không còn có thể chỉ dựa vào các giải pháp đóng kín "hộp" nữa. Khác với các giai đoạn phát triển công nghệ trước đây, quá trình chuyển đổi hiện tại - sang việc truy cập dữ liệu mở, sử dụng trí tuệ nhân tạo và phổ biến mã nguồn mở - khó có thể nhận được sự hỗ trợ từ các nhà cung cấp lớn, vì nó trực tiếp đe dọa các mô hình kinh doanh và nguồn thu nhập chính của họ.

Nghiên cứu của Trường Kinh doanh Harvard cho thấy rằng chi phí để phát triển các giải pháp mã nguồn mở được sử dụng nhiều nhất từ đầu cho tất cả các công ty sẽ khoảng 4,15 tỷ đô la. Tuy nhiên, nếu tưởng tượng rằng mỗi công ty phát triển các giải pháp thay thế riêng mà không có quyền truy cập vào các công cụ mã nguồn mở hiện có, như đã xảy ra trong vài thập kỷ qua, tổng chi phí của doanh nghiệp có thể lên tới con số khổng lồ 8,8 triệu tỷ đô la — đó là giá của nhu cầu không hợp lý mà thị trường phần mềm có thể được định giá.

Tiến bộ công nghệ chắc chắn sẽ dẫn đến việc xem xét lại các mô hình kinh doanh đã được thiết lập. Nếu trước đây các công ty có thể kiếm lời từ những quy trình phức tạp, không minh bạch và dữ liệu đóng, thì với sự phát triển của trí tuệ nhân tạo và phân tích, cách tiếp cận này ngày càng trở nên kém khả thi.

> Kết quả là, việc dân chủ hóa quyền truy cập vào dữ liệu và công cụ có thể làm giảm đáng kể thị trường bán phần mềm truyền thống. Tuy nhiên, cùng với đó, một thị trường mới sẽ nổi lên — thị trường chuyên môn số, tùy chỉnh, tích hợp và thiết kế giải pháp. Tại đây, giá trị sẽ được tạo ra không phải từ việc bán giấy phép, mà từ khả năng xây dựng các quy trình số linh hoạt, mở và có thể thích ứng. Giống như cách mà việc điện khí hóa và ra đời của máy kéo đã tạo ra các ngành công nghiệp mới, vậy việc ứng dụng dữ liệu lớn, AI và mô hình ngôn ngữ lớn mở ra những chân trời hoàn toàn mới cho doanh nghiệp trong ngành xây dựng, đòi hỏi không chỉ đầu tư công nghệ mà còn sự chuyển đổi sâu sắc trong tư duy, quy trình và cấu trúc tổ chức. Những công ty và chuyên gia nào nhận ra điều này và bắt đầu hành động ngay hôm nay sẽ trở thành những người dẫn đầu trong tương lai.

Trong một thế giới mà dữ liệu mở trở thành tài sản chính, khả năng tiếp cận thông tin sẽ thay đổi quy tắc của trò chơi. Các nhà đầu tư, khách hàng và nhà quản lý sẽ ngày càng yêu cầu tính minh bạch, và các thuật toán học máy có thể tự động phát hiện sự không nhất quán trong các dự toán, thời gian và chi phí. Điều này tạo điều kiện cho một giai đoạn chuyển đổi số mới, từ từ đưa chúng ta đến "uber hóa" ngành xây dựng.

Dữ liệu mở và sự uber hóa là mối đe dọa đối với ngành xây dựng hiện tại.

Ngành xây dựng đang biến thành một quá trình quản lý thông tin. Càng chính xác, chất lượng và đầy đủ dữ liệu, thì việc thiết kế, tính toán, dự toán giá cả, xây dựng và vận hành các tòa nhà sẽ càng hiệu quả. Trong tương lai, nguồn lực quan trọng không phải là việc sở hữu cần cẩu, bê tông và thép gia cố, mà là khả năng thu thập, phân tích và sử dụng thông tin.

> Khách hàng của các công ty xây dựng - các nhà đầu tư và khách hàng tài trợ cho việc xây dựng - trong tương lai chắc chắn sẽ sử dụng giá trị của dữ liệu mở và phân tích dữ liệu lịch sử. Điều này sẽ mở ra cơ hội cho việc tự động hóa tính toán thời gian và chi phí của các dự án, mà không cần phải mời các công ty xây dựng tham gia vào các vấn đề dự toán, từ đó giúp kiểm soát chi phí và nhanh chóng phát hiện các khoản chi thừa.

Hãy tưởng tượng một công trường xây dựng nơi mà các máy quét laser, drone và hệ thống photogrammetry thu thập dữ liệu chính xác về khối lượng bê tông được sử dụng trong thời gian thực. Thông tin này tự động được chuyển đổi thành các mô hình MESH phẳng đơn giản với siêu dữ liệu, không cần phải qua các hệ thống CAD (BIM) cồng kềnh, mà không phụ thuộc vào các lõi hình học phức tạp, ERP hay PMIS. Dữ liệu được thu thập từ công trường này được chuyển trung tâm vào các kho lưu trữ có cấu trúc thống nhất, có thể truy cập bởi bên thuê, phục vụ cho phân tích độc lập, nơi mà giá thực tế từ nhiều cửa hàng xây dựng và các tham số khác - từ tỷ lệ tài chính đến các yếu tố biến động như điều kiện thời tiết, tỷ giá vật liệu xây dựng, chi phí logistics và biến động mùa vụ về giá lao động được tải lên. Trong bối cảnh này, bất kỳ sự khác biệt nào giữa khối lượng vật liệu thiết kế và thực tế sẽ rõ ràng ngay lập tức, làm cho việc thao túng ngân sách trở nên không thể có cả trong giai đoạn thiết kế và lúc bàn giao công trình. Kết quả là, tính minh bạch trong quá trình xây dựng đạt được không phải thông qua một đội ngũ giám sát hay quản lý mà nhờ vào những dữ liệu kỹ thuật số khách quan, giảm thiểu yếu tố con người và khả năng đầu cơ.

Tương lai công việc kiểm soát dữ liệu như vậy sẽ chủ yếu được thực hiện bởi các quản lý dữ liệu từ phía bên thuê (Hình 1.24 CQMS manager). Đặc biệt, điều này liên quan đến việc tính toán và ước lượng ngân sách

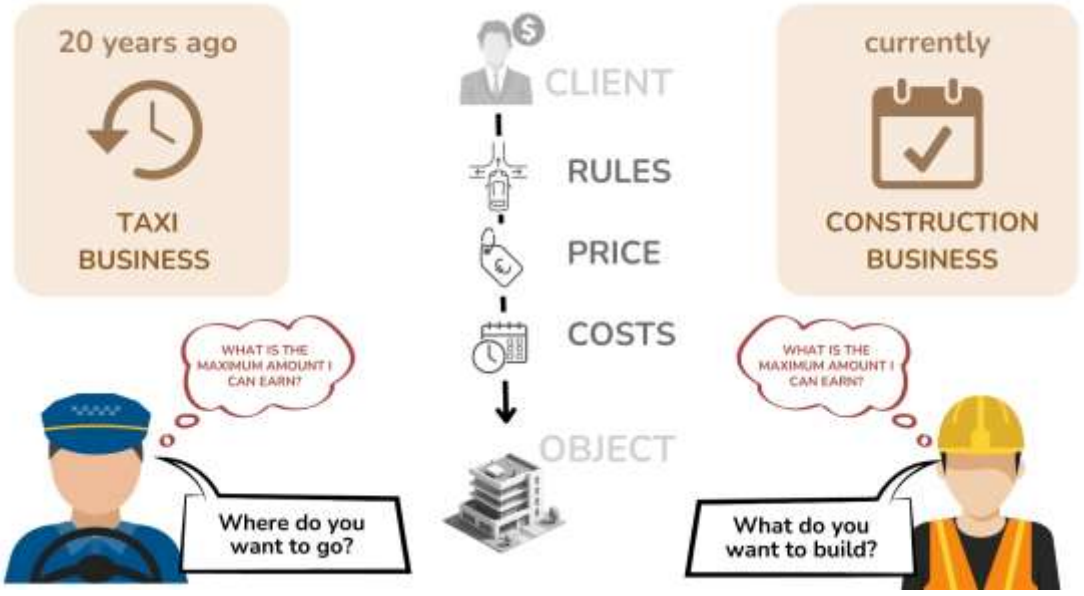
của các dự án: nơi mà trước đây có một bộ phận ước tính, ngày mai sẽ xuất hiện các công cụ học máy và dự đoán, thiết lập các khung giá mà các công ty xây dựng cần phải tuân thủ.

> *Xét đến tính phân mảnh của ngành xây dựng, khi mà phần lớn các hệ thống và tiểu hệ thống được cung cấp bởi các doanh nghiệp nhỏ và vừa, chiến lược số nên được khởi xướng từ phía khách hàng. Khách hàng cần tạo ra các điều kiện và cơ chế để mở ra các cơ hội số trong chuỗi cung ứng.*
>
> *– Andrew Davis và Giuliano DeNicola, Accenture "Tạo ra giá trị cao hơn thông qua các dự án vốn".*

Sự minh bạch và cởi mở như vậy đối với dữ liệu là một mối đe dọa đối với các công ty xây dựng, những công ty đã quen kiếm lời từ sự không minh bạch của các quy trình và báo cáo phức tạp, nơi mà có thể che giấu đầu cơ và chi phí ẩn sau các định dạng phức tạp và các nền tảng truyền tải dữ liệu độc quyền. Do đó, các công ty xây dựng, giống như trong trường hợp thúc đẩy các giải pháp mã nguồn mở của các nhà cung cấp, khó có thể quan tâm đến việc triển khai dữ liệu mở hoàn toàn trong các quy trình kinh doanh của họ. Nếu dữ liệu có sẵn và dễ xử lý cho bên thuê, chúng sẽ có thể được kiểm tra tự động, điều này sẽ loại trừ khả năng thổi phồng khối lượng và thao túng ngân sách.

> Theo báo cáo của Diễn đàn Kinh tế Thế giới "Hình thành tương lai của ngành xây dựng" (2016), một trong những vấn đề chính của ngành vẫn là vai trò thụ động của bên thuê. Tuy nhiên, chính các bên thuê cần phải chịu trách nhiệm lớn hơn về kết quả của các dự án - bắt đầu từ việc lập kế hoạch sớm, lựa chọn mô hình tương tác bền vững và kết thúc bằng việc kiểm soát việc thực hiện. Nếu không có sự tham gia tích cực từ phía chủ sở hữu dự án, việc chuyển đổi hệ thống trong ngành xây dựng là không thể.

Sự mất kiểm soát trong việc tính toán khối lượng và chi phí đã dẫn đến sự chuyển biến trong 20 năm qua ở nhiều ngành kinh tế khác, cho phép khách hàng tiếp cận trực tiếp, không qua trung gian, để đạt được mục tiêu của mình. Sự số hóa và minh bạch dữ liệu đã biến đổi nhiều mô hình kinh doanh truyền thống, giống như điều đã xảy ra với các tài xế taxi sau sự xuất hiện của Uber, với các nhà nghỉ sau khi Airbnb ra đời, và với các nhà bán lẻ cũng như cửa hàng do sự phát triển của Amazon, cũng như ngành ngân hàng - nhờ sự gia tăng của các ngân hàng kỹ thuật số và các hệ sinh thái fintech phi tập trung, nơi việc truy cập thông tin trực tiếp và tự động hóa trong việc tính toán thời gian và chi phí đã giảm thiểu vai trò của các trung gian.



Ngành xây dựng sẽ phải đối mặt với sự "uberization", giống như những gì tài xế taxi, chủ khách sạn và người bán hàng đã trải qua 10 năm trước.

Quá trình dân chủ hóa quyền truy cập vào dữ liệu và các công cụ xử lý chúng là điều không thể tránh khỏi, và theo thời gian, dữ liệu mở về tất cả các thành phần của dự án sẽ trở thành yêu cầu của khách hàng và tiêu chuẩn mới. Do đó, các vấn đề về việc triển khai các định dạng mở và tính toán minh bạch sẽ được thúc đẩy từ phía các nhà đầu tư, khách hàng, ngân hàng và quỹ đầu tư tư nhân - những người cuối cùng là người sử dụng các công trình đã được xây dựng và sau đó vận hành chúng trong nhiều năm.

> Các nhà đầu tư lớn, khách hàng và ngân hàng đã yêu cầu sự minh bạch trong ngành xây dựng. Theo nghiên cứu của Accenture "Tạo ra giá trị lớn hơn từ các dự án vốn" (2020), dữ liệu minh bạch và đáng tin cậy trở thành yếu tố quyết định cho các quyết định đầu tư trong xây dựng. Như các chuyên gia đã chỉ ra, việc quản lý dự án hiệu quả và đáng tin cậy không thể diễn ra nếu thiếu tính minh bạch, đặc biệt trong các tình huống khủng hoảng. Hơn nữa, các chủ sở hữu tài sản và nhà thầu ngày càng chuyển hướng sang các hợp đồng thúc đẩy việc trao đổi dữ liệu và phân tích hợp tác, phản ánh yêu cầu ngày càng tăng từ phía các nhà đầu tư, ngân hàng và cơ quan quản lý về trách nhiệm và tính minh bạch.

> Hành trình của nhà đầu tư, khách hàng từ ý tưởng đến tòa nhà hoàn thiện trong tương lai sẽ giống như một chuyến đi tự động - không có tài xế là công ty xây dựng, hứa hẹn sẽ trở nên độc lập với những sự đầu cơ và không chắc chắn.

Kỷ nguyên của dữ liệu mở và tự động hóa sẽ không thể tránh khỏi việc thay đổi ngành xây dựng giống như những gì đã xảy ra trong lĩnh vực ngân hàng, thương mại, nông nghiệp và logistics. Ở những ngành này, vai trò của các trung gian và các phương thức kinh doanh truyền thống đang nhường chỗ cho tự động hóa và robot hóa, không để lại chỗ cho việc tăng giá vô lý và đầu cơ.

Dữ liệu và quy trình trong tất cả các loại hoạt động kinh tế của con người không khác gì so với những gì các chuyên gia trong ngành xây dựng phải đối mặt. Trong dài hạn, các công ty xây dựng hiện đang thống trị thị trường, thiết lập tiêu chuẩn về giá cả và chất lượng dịch vụ, có thể sẽ đánh mất vị trí là trung gian chủ chốt

giữa khách hàng và dự án xây dựng của họ.

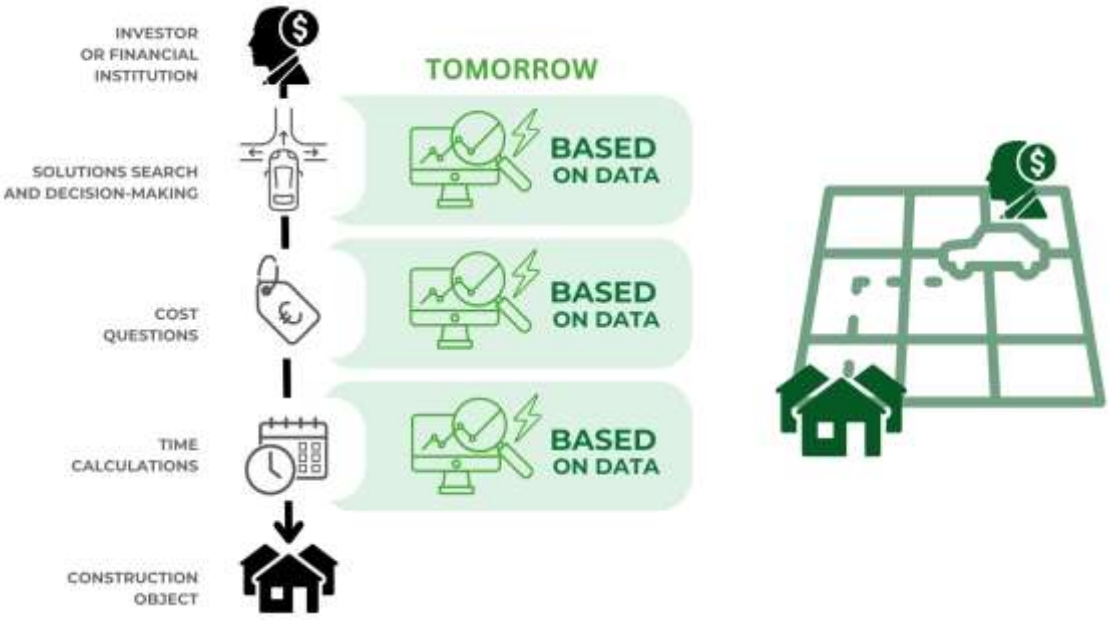
Các vấn đề chưa được giải quyết của mô hình Uber hóa như là cơ hội cuối cùng để tận dụng thời gian cho việc chuyển đổi.

Nhưng hãy quay trở lại với thực tế của ngành xây dựng. Trong khi một số lĩnh vực kinh tế xuất hiện xe tự lái, hệ thống tài chính phi tập trung và giải pháp dựa trên trí tuệ nhân tạo, một phần lớn các công ty xây dựng vẫn là những tổ chức giấy tờ, nơi các quyết định then chốt thường được đưa ra dựa trên trực giác và kinh nghiệm của một số chuyên gia riêng lẻ.

Trong mô hình này, một công ty xây dựng hiện đại có thể được so sánh với một đội taxi 20 năm trước, kiểm soát tài nguyên, lộ trình và thời gian giao hàng, chịu trách nhiệm về thời hạn và chi phí "chuyến đi" — từ ý tưởng thiết kế (quy trình logistics và lắp đặt) cho đến khi bàn giao dự án. Giống như cách mà GPS (trong xây dựng IoT, RFID) và các thuật toán máy học đã thay đổi lĩnh vực giao thông, dữ liệu, thuật toán và các tác nhân AI có khả năng chuyển đổi quản lý xây dựng — từ các ước lượng trực giác sang mô hình dự đoán, có kiểm soát. Trong 20 năm qua, trong nhiều lĩnh vực — tài chính, nông nghiệp, bán lẻ và logistics — khả năng đầu cơ dựa trên sự không minh bạch của dữ liệu dần dần biến mất. Giá, chi phí giao hàng hoặc giao dịch tài chính được tính toán tự động và dựa trên thống kê — chỉ trong vài giây trên các nền tảng kỹ thuật số.

Nhìn về tương lai, các công ty xây dựng cần nhận thức rằng việc dân chủ hóa quyền truy cập vào dữ liệu và các công cụ phân tích của chúng sẽ làm gián đoạn cách tiếp cận truyền thống đối với việc đánh giá chi phí và thời gian thực hiện các dự án, cũng như loại bỏ khả năng đầu cơ trên dữ liệu không minh bạch về khối lượng và giá cả.

> Giống như việc di chuyển trên đường có điều tiết mà không cần sự can thiệp của tài xế, các quy trình xây dựng trong tương lai sẽ ngày càng giống một hệ thống “Uber hóa” — với việc đánh giá tự động về thời gian và chi phí, phân bổ nhiệm vụ minh bạch và độ phụ thuộc tối thiểu vào yếu tố con người. Điều này sẽ thay đổi bản chất của "cuộc hành trình" từ ý tưởng đến thực hiện — khiến nó trở nên dễ dự đoán hơn, có quản lý và dựa trên dữ liệu.



R. 10.15 Chi phí và thời gian "chặng đường" trong quá trình xây dựng sẽ được xác định bằng cách sử dụng máy học và các công cụ thống kê.

Khi ngày càng có nhiều quy chuẩn và yêu cầu mới được đưa vào gần như mọi quốc gia trên thế giới, yêu cầu chuyển giao mô hình CAD- (BIM) cho khách hàng hoặc ngân hàng tài trợ cho các dự án xây dựng, khách hàng và nhà thầu có khả năng tự đảm bảo tính minh bạch của các phép tính dữ liệu về chi phí và khối lượng công việc. Điều này đặc biệt quan trọng đối với các khách hàng lớn và nhà đầu tư, những người có đủ năng lực và công cụ để phân tích khối lượng kịp thời và theo dõi giá thị trường. Đối với các công ty thực hiện các dự án quy mô lớn — cửa hàng, tòa nhà văn phòng, khu chung cư — những thực tiễn như vậy trở thành tiêu chuẩn.

Khi thông tin trong các mô hình trở nên đầy đủ và chuẩn hóa hơn, khả năng thao tác và đầu cơ gần như biến mất. Chuyển đổi số đang dần thay đổi các quy tắc trò chơi trong ngành xây dựng, và những công ty không thích nghi với những thay đổi này có thể sẽ phải đối mặt với những thách thức nghiêm trọng.

> Sự gia tăng cạnh tranh, khoảng cách công nghệ và sự suy giảm lợi nhuận có thể ảnh hưởng đến sự bền vững của doanh nghiệp. Trong bối cảnh thanh khoản hạn chế, ngày càng nhiều thành viên trong ngành đang chuyển sang tự động hóa, phân tích và công nghệ xử lý dữ liệu như một cách để nâng cao hiệu quả và tính minh bạch của các quy trình. Những công cụ này đang trở thành tài nguyên quan trọng để duy trì khả năng cạnh tranh trong môi trường kinh tế đang thay đổi.

Có lẽ không nên chờ đợi cho đến khi các hoàn cảnh bên ngoài buộc phải thực hiện các bước khẩn cấp - hiệu quả hơn nhiều nếu bắt đầu chuẩn bị ngay hôm nay, củng cố năng lực số, triển khai các giải pháp hiện đại và xây dựng văn hóa tập trung vào công việc với dữ liệu.

Một trong những rào cản công nghệ chính cuối cùng trên con đường chuyển đổi số quy mô lớn của ngành xây dựng, mà sẽ ảnh hưởng đến từng công ty trong những năm tới, vẫn là vấn đề phân loại tự động các yếu tố trong các dự án xây dựng.

> Không có phân loại đáng tin cậy, chính xác và có thể mở rộng, rất khó để tạo ra nền tảng cho phân tích đầy đủ, tự động hóa quy trình và quản lý vòng đời của các đối tượng với việc áp dụng trí tuệ nhân tạo và các mô hình dự đoán. Khi mà việc phân loại các đối tượng vẫn phụ thuộc vào việc giải thích thủ công từ các chuyên gia dày dạn kinh nghiệm — như các kỹ sư, nhà thiết kế và nhân viên định giá — ngành xây dựng vẫn còn một khoảng thời gian để tận dụng. Thời gian này có thể được sử dụng để chuẩn bị cho những thay đổi không thể tránh khỏi: sự gia tăng yêu cầu về tính minh bạch, dân chủ hóa công cụ và dữ liệu, cũng như sự xuất hiện của các hệ thống phân loại tự động, sẽ thay đổi quy tắc trò chơi một cách triệt để.

Nhiệm vụ phân loại tự động các yếu tố trong thế giới xây dựng có độ phức tạp tương đương với việc nhận diện đối tượng trong các hệ thống lái xe tự động, vốn là một thách thức lớn. Hãy tưởng tượng một chiếc ô tô tự lái di chuyển từ điểm A đến điểm B. Các hệ thống lái xe tự động hiện đại gặp phải vấn đề phân loại các đối tượng, được nhận diện bằng lidar và camera. Chiếc ô tô không chỉ cần "nhìn thấy" trở ngại hoặc điểm đánh dấu - nó cần phải hiểu một cách chính xác rằng điều gì đang trước mặt: người đi bộ, biển báo giao thông hay thùng rác.

Một vấn đề cơ bản tương tự đang đứng trước toàn bộ ngành xây dựng. Các yếu tố của dự án - như cửa sổ, cửa ra vào hay cột - có thể được ghi lại trong tài liệu, được trình bày trong mô hình CAD, được chụp ảnh trên công trường xây dựng hoặc được nhận diện trong các đám mây điểm từ quét laser. Tuy nhiên, để xây dựng một hệ thống quản lý dự án thực sự tự động hóa, không chỉ cần nhận diện hình ảnh hoặc hình dạng cơ bản của chúng. Cần phải đảm bảo phân loại chính xác và ổn định cho từng yếu tố theo loại mà sẽ được xác định rõ ràng trong tất cả các quy trình tiếp theo - từ dự toán và bảng thông số kỹ thuật đến logistics, quản lý kho và quan trọng nhất - khai thác.

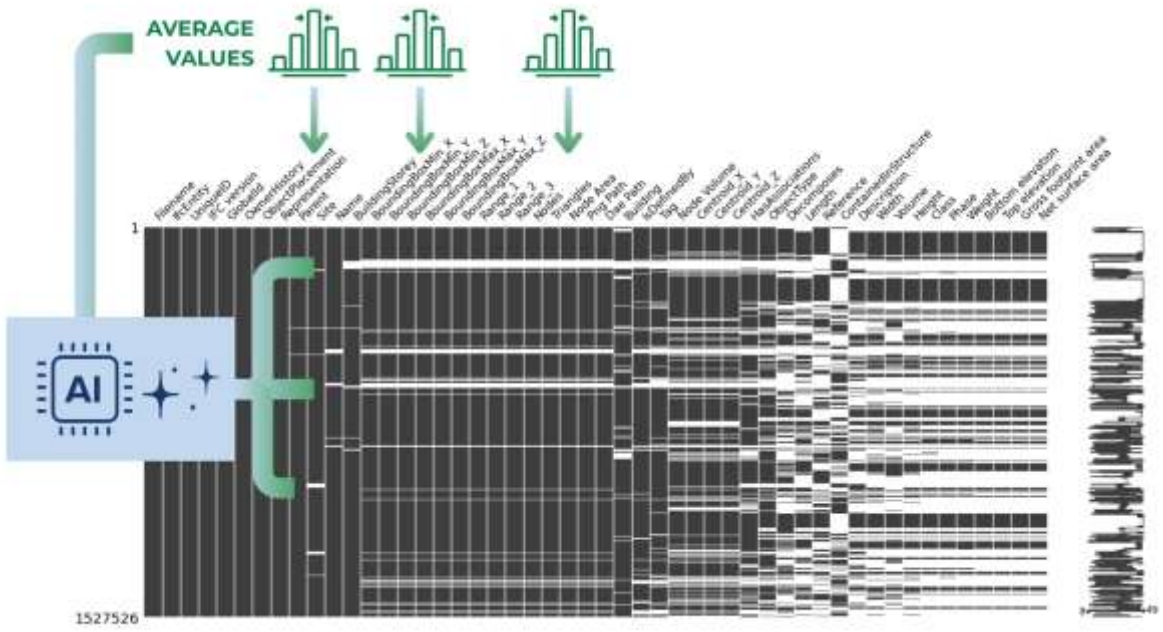
Chính vào giai đoạn này — chuyển tiếp từ nhận diện đến phân loại có ý nghĩa — một trong những trở ngại then chốt xuất hiện. Ngay cả khi các hệ thống kỹ thuật số về mặt kỹ thuật có khả năng phân tách và xác định đối tượng trong mô hình và tại công trường, sự khó khăn chính nằm ở việc định nghĩa loại yếu tố một cách chính xác và bền vững trong bối cảnh cho các môi trường phần mềm khác nhau. Ví dụ, một cánh cửa có thể được nhà thiết kế ký hiệu trong mô hình CAD là một yếu tố thuộc loại "cửa", nhưng khi chuyển giao vào hệ thống ERP hoặc PMIS, nó có thể bị phân loại sai — trong trường hợp lỗi từ nhà thiết kế hoặc do sự không khớp giữa các hệ thống. Hơn nữa, yếu tố thường mất một phần quan trọng của thuộc tính hoặc thậm chí biến mất khỏi hồ sơ hệ thống trong quá trình xuất và nhập dữ liệu. Điều này dẫn đến một khoảng trống trong dòng dữ liệu và làm tổn hại nguyên tắc số hóa đồng bộ trong quy trình xây dựng. Do đó, một khoảng cách nghiêm trọng hình thành giữa ý nghĩa "nhìn thấy" và "hiểu được" ngữ nghĩa, điều này làm tổn hại tính toàn vẹn của dữ liệu và gây khó khăn đáng kể cho tự động hóa quy trình trong suốt vòng đời của công trình xây dựng.

Giải quyết nhiệm vụ phân loại đồng nhất các yếu tố xây dựng bằng cách sử dụng công nghệ dữ liệu lớn và học máy (Hình 10.16) sẽ trở thành chất xúc tác cho việc chuyển đổi toàn bộ ngành — và có thể là một phát hiện bất ngờ cho nhiều công ty xây dựng. Hệ thống phân loại thống nhất, có khả năng học hỏi sẽ trở thành nền tảng cho phân tích có thể mở rộng, quản lý số và triển khai trí tuệ nhân tạo vào thực tiễn hàng ngày của các tổ chức xây dựng.-

NVIDIA và các tập đoàn công nghệ hàng đầu khác hiện nay đã cung cấp các giải pháp trong những lĩnh vực kinh tế khác, có khả năng tự động phân loại và cấu trúc khối lượng lớn thông tin văn bản và hình ảnh.

> Mô hình NeMo Curator [161] từ NVIDIA, chẳng hạn, chuyên về phân loại tự động và phân phối dữ liệu theo các loại đã được định trước, đóng vai trò quan trọng trong việc tối ưu hóa quy trình xử lý thông tin cho các nhiệm vụ tinh chỉnh và đào tạo trước các mô hình AI sinh sinh. Nền tảng Cosmos được đào tạo trên các video thực tế và các cảnh 3D [162], tạo ra cơ sở cho các hệ thống tự động và bản sao kỹ thuật số, đã được phát triển trong hệ sinh thái của NVIDIA. NVIDIA Omniverse, dự kiến sẽ trở thành công cụ hàng đầu làm việc với định dạng USD — mô tả cảnh phổ quát, có khả năng thay thế định dạng IFC trong quy trình truyền tải thông tin dự án trong tương lai. Kết hợp với Isaac Sim — mô phỏng quy trình robot [163] — những giải pháp như NeMo Curator, Cosmos và Omniverse đại diện cho một cấp độ tự động hóa mới: từ việc làm sạch và lọc dữ liệu cho đến việc tạo ra bộ dữ liệu đào tạo, mô phỏng các thuộc tính của đối tượng và đào tạo robot tại công trường xây dựng. Tất cả các công cụ này đều được phát tán miễn phí và có sẵn để truy cập, điều này làm giảm đáng kể rào cản trong việc áp dụng vào thực tiễn kỹ thuật và xây dựng.

Phân loại tự động dữ liệu ở cấp độ bảng có cấu trúc không phải là nhiệm vụ quá phức tạp như có vẻ ban đầu. Như chúng tôi đã chỉ ra trong chương trước (Hình 9.110), với sự sẵn có của dữ liệu lịch sử, có thể điền vào các giá trị thiếu hoặc không chính xác của các lớp dựa trên các tham số tương tự từ các yếu tố khác. Nếu trong một vài dự án hoàn thành, các yếu tố có đặc điểm tương tự đã được phân loại chính xác, thì hệ thống có thể với độ chính xác cao đề xuất giá trị phù hợp cho một yếu tố mới hoặc chưa đầy đủ (Hình 10.16). Logic như vậy, dựa trên các giá trị trung bình và phân tích ngữ cảnh, có thể đặc biệt hiệu quả khi xử lý hàng loạt dữ liệu bảng từ các ước tính, đặc tả hoặc mô hình CAD.-

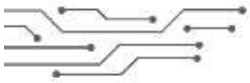


Hình 10.16 Machine Learning sẽ giúp tự động tìm các giá trị trung bình cho các thông số bảng chưa được điền (các ô trắng) dựa trên các dự án trong quá khứ.

> Trên nền tảng sự phát triển nhanh chóng trong lĩnh vực Machine Learning, điều trở nên rõ ràng: vào năm 2025 sẽ thật ngây thơ khi nghĩ rằng vấn đề phân loại tự động các yếu tố xây dựng sẽ lâu dài chưa được giải quyết. Đúng vậy, các thuật toán hiện đại vẫn chưa đạt đến độ trưởng thành hoàn toàn, đặc biệt là trong điều kiện dữ liệu không đầy đủ hoặc không đồng nhất, nhưng cơ hội để thích ứng đang nhanh chóng bị khép lại.

Các công ty đã đầu tư vào việc thu thập, làm sạch và hệ thống hóa dữ liệu của họ, cũng như âm các công cụ tự động hóa ETL, sẽ ở vào vị trí có lợi hơn rõ rệt. Những đối thủ khác có thể sẽ không theo kịp — giống như nhiều công ty đã không thành công trong việc đối phó với những thách thức của sự chuyển đổi số trong các lĩnh vực giao thông và tài chính.

Những người tiếp tục phụ thuộc vào việc quản lý dữ liệu bằng tay và các phương pháp truyền thống để đánh giá chi phí và thời gian có nguy cơ rơi vào tình huống giống như các hãng taxi vào những năm 2000, không kịp thích ứng với kỷ nguyên ứng dụng di động và các tính toán lộ trình tự động vào đầu những năm 2020.



CHƯƠNG 10.2.
HƯỚNG DẪN THỰC HÀNH VỀ VIỆC TRIỂN KHAI CÁCH TIẾP CẬN DỮ LIỆU HƯỚNG DẪN

Từ lý thuyết đến thực tiễn: bản đồ đường đi của chuyển đổi số trong xây dựng.

Ngành xây dựng đang từ từ bước vào một giai đoạn phát triển mới, nơi mà các quy trình quen thuộc ngày càng được bổ sung — hoặc đôi khi thay thế — bằng các nền tảng số và các mô hình tương tác minh bạch. Điều này không chỉ mang lại thách thức mà còn mở ra những cơ hội đáng kể cho các công ty. Những tổ chức đã xây dựng được chiến lược số lâu dài ngay hôm nay sẽ không chỉ giữ vững vị thế trên thị trường mà còn mở rộng nó, bằng cách cung cấp cho khách hàng những phương pháp hiện đại và giải pháp công nghệ đáng tin cậy.

> Tuy nhiên, điều quan trọng là phải hiểu rằng: kiến thức về các khái niệm và công nghệ chỉ là điểm khởi đầu. Các lãnh đạo và chuyên gia đang đối mặt với một câu hỏi thực tế: nên bắt đầu từ đâu để triển khai và làm thế nào để biến những ý tưởng lý thuyết thành giá trị thực tế. Hơn nữa, ngày càng có nhiều câu hỏi nổi lên: doanh nghiệp sẽ được xây dựng trên nền tảng nào nếu các phương pháp truyền thống để tính toán chi phí và thời gian có thể bị khách hàng xem xét lại bất cứ lúc nào.

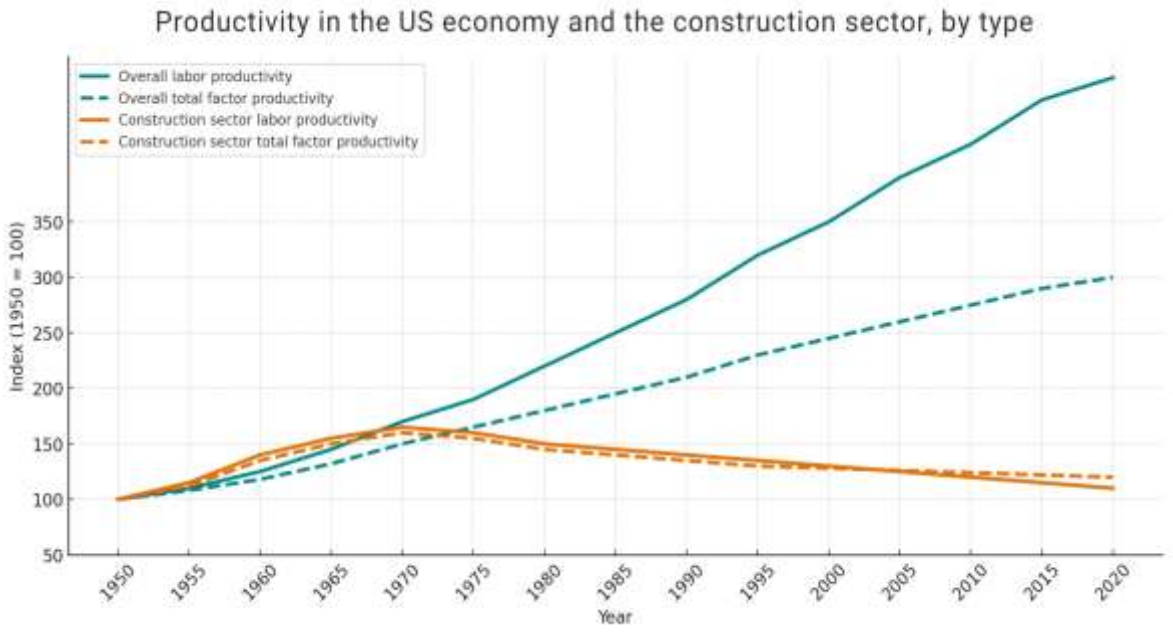
Câu trả lời có thể không nằm ở công nghệ, mà ở việc hình thành một nền văn hóa nghề nghiệp mới, nơi mà việc làm việc với dữ liệu được xem là một phần không thể tách rời của thực tiễn hàng ngày. Chính sự thiếu chú ý đến công nghệ số và đổi mới đã dẫn đến sự tụt hậu nghiêm trọng của ngành xây dựng, điều này đã được quan sát trong nhiều thập kỷ qua [43].

> Theo dữ liệu từ McKinsey, chi phí cho Nghiên cứu và Phát triển (R&D) trong ngành xây dựng chỉ chiếm dưới 1% doanh thu, trong khi đó ở ngành ô tô và hàng không vũ trụ tỷ lệ này đạt từ 3,5-4,5%. Tương tự, chi phí cho công nghệ thông tin trong xây dựng vẫn ở mức dưới 1% tổng doanh thu.

Cuối cùng, không chỉ mức độ tự động hóa mà năng suất lao động trong ngành xây dựng cũng giảm, và đến năm 2020, một công nhân xây dựng đã sản xuất ít hơn so với nửa thế kỷ trước.

Những vấn đề tương tự về năng suất trong ngành xây dựng đặc trưng cho hầu hết các nước phát triển và đang phát triển (năng suất xây dựng đã giảm ở 16 trong số 29 quốc gia OECD) và cho thấy không chỉ sự thiếu hụt công nghệ mà còn cần cải cách hệ thống trong các phương pháp quản lý, đào tạo và triển khai đổi mới.

> Thành công trong chuyển đổi số không chỉ phụ thuộc vào số lượng và sự hiện diện của các công cụ, mà còn vào khả năng của tổ chức trong việc xem xét lại quy trình của mình và phát triển văn hóa cởi mở trước những thay đổi. Vai trò quan trọng không thuộc về công nghệ mà là con người và quy trình được thiết lập, đảm bảo việc ứng dụng hiệu quả, hỗ trợ học tập liên tục và khuyến khích việc tiếp thu những ý tưởng mới.



-

Trong những phần đầu của cuốn sách, mô hình môi trường kinh doanh được so sánh với hệ sinh thái rừng. Trong một khu rừng khỏe mạnh, những trận cháy định kỳ, mặc dù có sức tàn phá lớn, nhưng đóng vai trò then chốt trong việc tái tạo lâu dài. Chúng làm sạch đất khỏi thực vật cũ, phục hồi các chất dinh dưỡng tích lũy và tạo không gian cho sự sống mới. Một số loài thực vật thậm chí đã tiến hóa đến mức hạt của chúng chỉ nảy mầm dưới tác động của nhiệt độ cao từ lửa – đây là cơ chế tự nhiên đảm bảo thời điểm lý tưởng cho sự nảy mầm.-

Tương tự trong kinh doanh: khủng hoảng có thể đóng vai trò như "đốt cháy có kiểm soát", giúp hình thành những phương pháp và công ty mới không gắn liền với các hệ thống lỗi thời. Những giai đoạn như vậy buộc phải từ bỏ các thực tiễn không hiệu quả, giải phóng nguồn lực cho đổi mới. Giống như rừng sau khi cháy bắt đầu với những cây tiên phong, doanh nghiệp sau khủng hoảng hình thành những quy trình mới, linh hoạt, trở thành nền tảng cho môi trường thông tin trưởng thành.

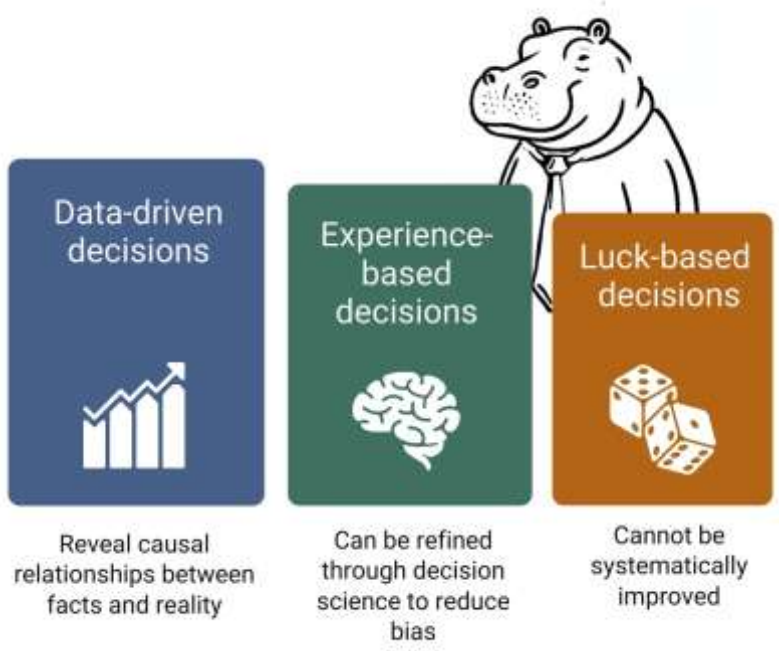
Các công ty đã biết cách giải mã những "giặc lửa tín hiệu" này và biến năng lượng phá hủy của chúng thành những thay đổi xây dựng sẽ đạt được mức hiệu quả mới – với các quy trình xử lý dữ liệu minh bạch hơn, thích ứng tốt hơn, tăng cường khả năng tự đổi mới và phát triển tự nhiên của tổ chức.

> Ảnh hưởng ngày càng tăng của trí tuệ nhân tạo (AI) và học máy (machine learning) đối với môi trường kinh doanh đã không còn nghi ngờ gì nữa. Đây không chỉ là một xu hướng tạm thời, mà là một nhu cầu chiến lược. Các công ty phớt lờ AI có nguy cơ mất đi khả năng cạnh tranh trên thị trường, nơi ngày càng khuyến khích sự đổi mới và linh hoạt.

Tương lai thuộc về những ai nhìn thấy trí tuệ nhân tạo không chỉ là công cụ, mà là cơ hội để xem xét lại từng khía cạnh của hoạt động của họ - từ tối ưu hóa quy trình đến ra quyết định quản lý.

Xây dựng nền tảng số: 1-5 bước đến sự trưởng thành số.

Ở chương này, chúng ta sẽ xem xét lộ trình chuyển đổi số và xác định các bước chính cần thiết để triển khai phương pháp dựa trên dữ liệu, điều này có thể giúp biến đổi cả văn hóa doanh nghiệp lẫn hệ sinh thái thông tin của công ty.



Hình 10.22 Cập nhật có kiểm soát và lựa chọn chiến lược: tình huống, kinh nghiệm hay dữ liệu.

Theo nghiên cứu của McKinsey "Tại sao chiến lược số thất bại" (2018), có ít nhất năm lý do mà các công ty không đạt được mục tiêu chuyển đổi số.

- Định nghĩa mơ hồ: các nhà lãnh đạo và quản lý hiểu khác nhau về cái gọi là "công nghệ số", dẫn đến sự hiểu lầm và không nhất quán trong hành động.
- Hiểu sai về kinh tế công nghệ số: nhiều công ty đánh giá thấp quy mô thay đổi mà chuyển đổi số mang lại cho mô hình kinh doanh và động lực ngành (Hình 10.16).
- Bỏ qua hệ sinh thái: các công ty tập trung vào các giải pháp công nghệ riêng lẻ (ngăn dữ liệu), bỏ lỡ sự cần thiết phải tích hợp vào các hệ sinh thái số rộng hơn (Hình 2.22, Hình 4.112).-
- Đánh giá thấp chuyển đổi số từ phía đối thủ: các nhà lãnh đạo không xem xét rằng các đối thủ cũng đang tích cực triển khai công nghệ số, điều này có thể dẫn đến mất đi lợi thế cạnh tranh.
- Bỏ lỡ sự hai mặt của chuyển đổi số: giám đốc điều hành ủy quyền trách nhiệm chuyển đổi số cho các giám đốc khác, điều này quan liêu hóa quản lý và làm chậm quá trình thay đổi.

Để giải quyết những vấn đề này, cần có sự hiểu biết và thống nhất rõ ràng về chiến lược số ở tất cả các cấp của tổ chức. Trước khi xây dựng chiến lược số, quan trọng là phải hiểu điểm khởi đầu. Nhiều tổ chức mong muốn áp dụng các công cụ và nền tảng mới mà không có cái nhìn toàn diện về tình trạng hiện tại.

Bước 1. Tiến hành kiểm toán các hệ thống và dữ liệu hiện có.

Trước khi thay đổi quy trình, quan trọng là phải hiểu rõ những gì đã có. Việc thực hiện kiểm toán giúp xác định điểm yếu trong quản lý dữ liệu và hiểu rõ những nguồn lực nào có thể được sử dụng. Kiểm toán tương tự như một "X-quang" của quy trình kinh doanh. Nó sẽ giúp bạn xác định các khu vực rủi ro và xác định dữ liệu nào là quan trọng đối với dự án hoặc doanh nghiệp của bạn, và dữ liệu nào là thứ yếu.

Các hành động chính:

- Tạo bản đồ môi trường CNTT (trong Draw.io, Lucidchart, Miro, Visio hoặc Canva). Liệt kê các hệ thống đang sử dụng (ERP, CAD, CAFM, CPM, SCM và các hệ thống khác) có liên quan đến quy trình của bạn và đã được thảo luận trong chương "Công nghệ và hệ thống quản lý trong xây dựng hiện đại" (Hình 1.24).-
- Đánh giá vấn đề chất lượng dữ liệu cho mỗi hệ thống dựa trên tần suất tồn tại các bản sao, các giá trị có thể bị bỏ lỡ và sự không nhất quán trong định dạng ở mỗi hệ thống.
- Xác định các "điểm đau" - những nơi quy trình có thể bị gián đoạn hoặc thường xuyên yêu cầu can thiệp thủ công - nhập khẩu, xuất khẩu và các quy trình kiểm tra bổ sung.

> Nếu bạn muốn đội ngũ tin tưởng vào các báo cáo, cần phải theo dõi độ chính xác của dữ liệu ngay từ đầu.

Một cuộc kiểm toán dữ liệu được thực hiện chất lượng sẽ chỉ ra những dữ liệu nào:

- Cần được chỉnh sửa (cần thiết lập các quá trình tự động để làm sạch hoặc biến đổi thêm)
- Là "rác", chỉ làm tắc nghẽn hệ thống và có thể loại bỏ, không sử dụng lại trong các quy trình.

Cuộc kiểm toán như vậy có thể thực hiện độc lập. Nhưng đôi khi hữu ích khi thu hút một chuyên gia bên ngoài - đặc biệt là từ các lĩnh vực kinh tế khác: cái nhìn mới mẻ và độc lập với các "đặc thù" xây dựng sẽ giúp đánh giá đúng trạng thái hiện tại và tránh các cạm bẫy thường thấy do thiên lệch đối với một số giải pháp và công nghệ.

Bước 2. Xác định các tiêu chuẩn chính để thống nhất dữ liệu.

Sau khi thực hiện kiểm toán, cần tạo ra các quy tắc chung để làm việc với dữ liệu. Như chúng ta đã thảo luận trong chương "Các tiêu chuẩn: từ các tệp ngẫu nhiên đến mô hình dữ liệu được thiết kế", điều này sẽ giúp loại bỏ sự phân tán của các luồng thông tin.

> Nếu không có một tiêu chuẩn thống nhất, mỗi đội sẽ tiếp tục làm việc "theo cách riêng của mình", và bạn sẽ giữ lại "zoo" các tích hợp, nơi dữ liệu bị mất mát trong mỗi lần biến đổi.

Các hành động chính:

- Chọn các tiêu chuẩn dữ liệu để trao đổi thông tin giữa các hệ thống:
  - Đối với dữ liệu bảng, có thể sử dụng các định dạng có cấu trúc như CSV, XLSX hoặc các định dạng hiệu quả hơn như Parquet
  - Đối với việc trao đổi dữ liệu hơi có cấu trúc và tài liệu: JSON hoặc XML
- Nắm vững việc làm việc với các mô hình dữ liệu:

- Bắt đầu với việc tham số hóa các nhiệm vụ ở cấp độ mô hình khái niệm dữ liệu - như mô tả trong chương "Mô hình hóa dữ liệu: mô hình khái niệm, logic và vật lý" (Hình. 4.32)-
- Khi đi sâu vào logic của các quá trình kinh doanh, chuyển sang việc chính thức hóa các yêu cầu bằng cách sử dụng các tham số trong các mô hình logic và vật lý (Hình. 4.36) -
- Xác định các thực thể chính, các thuộc tính của chúng và các mối quan hệ trong các quy trình, cũng như hình dung những mối quan hệ này - cả giữa các thực thể và giữa các tham số (Hình. 4.37)-

- Sử dụng biểu thức chính quy (RegEx) để xác thực và chuẩn hóa dữ liệu (Hình. 4.47), như chúng ta đã thảo luận trong chương "Các yêu cầu có cấu trúc và biểu thức chính quy RegEx". RegEx - không phải là một chủ đề phức tạp, nhưng cực kỳ quan trọng trong việc xây dựng các yêu cầu ở cấp độ các mô hình dữ liệu vật lý.-

Nếu không có tiêu chuẩn ở cấp độ dữ liệu và hình ảnh hóa quy trình, không thể đảm bảo một môi trường kỹ thuật số nhất quán và có thể mở rộng. Hãy nhớ: "dữ liệu kém giá cả cao". Và giá trị của lỗi tăng lên khi dự án hoặc tổ chức trở nên phức tạp hơn. Thống nhất định dạng, xác định quy tắc đặt tên, cấu trúc và xác thực - đó là những khoản đầu tư vào sự ổn định và khả năng mở rộng của các giải pháp trong tương lai.

Bước 3. Triển khai DataOps và tự động hóa quy trình.

Nếu không có một kiến trúc rõ ràng, công ty chắc chắn sẽ phải đối mặt với các dữ liệu phân tán, bị khóa trong các hệ thống thông tin cô lập. Dữ liệu sẽ không được tích hợp, sẽ bị trùng lặp ở nhiều nơi khác nhau và sẽ đòi hỏi chi phí đáng kể để duy trì.

> Hãy tưởng tượng rằng dữ liệu giống như nước, và kiến trúc dữ liệu là hệ thống ống dẫn phức tạp, nơi nước này được vận chuyển từ nguồn lưu trữ đến nơi sử dụng. Chính kiến trúc dữ liệu xác định cách thông tin được thu thập, lưu trữ, chuyển đổi, phân tích và cung cấp cho người dùng cuối hoặc các ứng dụng.

DataOps (Hoạt động Dữ liệu) là một phương pháp kết hợp việc thu thập, làm sạch, kiểm tra và sử dụng dữ liệu thành một quy trình tự động hóa duy nhất, như chúng tôi đã thảo luận chi tiết trong phần tám của cuốn sách.

Các hành động chính:

- Tạo và cấu hình các pipeline ETL để tự động hóa quy trình:
  - Extract: tổ chức thu thập dữ liệu tự động từ tài liệu PDF, bảng tính Excel, mô hình CAD, hệ thống ERP và các nguồn khác mà bạn đang làm việc với.----
  - Transform: thiết lập các quy trình tự động chuyển đổi dữ liệu về một định dạng có cấu trúc duy nhất và tự động hóa các phép tính sẽ diễn ra bên ngoài các ứng dụng khép kín.-
  - Load: thử tạo một cách xuất dữ liệu tự động vào bảng kết quả, tài liệu hoặc kho lưu trữ trung tâm.---
- Tự động hóa các quy trình tính toán và QTO (Quantity Take-Off), như chúng tôi đã thảo luận trong chương "QTO Quantity Take-Off: phân nhóm dữ liệu dự án theo thuộc tính".

- Thiết lập tự động trích xuất khối lượng từ mô hình CAD bằng cách sử dụng API, plugin hoặc công cụ ngược kỹ thuật.-
- Tạo quy tắc phân nhóm các phần tử cho các lớp khác nhau theo thuộc tính dưới dạng bảng.
- Thử tự động hóa các phép tính khối lượng và chi phí lặp lại bên ngoài các hệ thống khép kín mô-đun.

- Bắt đầu sử dụng Python và Pandas để xử lý dữ liệu, như chúng tôi đã đề cập trong chương "Python Pandas: công cụ không thể thiếu cho việc làm việc với dữ liệu".
  - Áp dụng DataFrame để làm việc với các tệp XLSX và tự động hóa quá trình xử lý dữ liệu bảng.-
  - Tự động hóa việc tổng hợp và chuyển đổi thông tin thông qua các thư viện Python khác nhau.
  - Sử dụng LLM để đơn giản hóa việc viết các khối mã hoàn chỉnh và toàn bộ Pipeline.-
  - Thử xây dựng một Pipeline trên Python, tìm các lỗi hoặc phát hiện các bất thường và gửi thông báo cho người chịu trách nhiệm (ví dụ: người quản lý dự án).-

Tự động hóa dựa trên các nguyên tắc DataOps cho phép chuyển từ công việc thủ công và phân mảnh với dữ liệu sang các quy trình bền vững và tái sản xuất được. Điều này không chỉ giảm tải cho nhân viên thường xuyên phải thực hiện các chuyển đổi giống nhau mà còn nâng cao đáng kể tính đáng tin cậy, khả năng mở rộng và tính minh bạch của toàn bộ hệ thống thông tin.

Bước 4. Tạo hệ sinh thái quản lý dữ liệu mở.

Mặc dù sự phát triển của các hệ thống mô-đun khép kín và sự tích hợp của chúng với các công cụ mới, các công ty đang đối mặt với một vấn đề nghiêm trọng - sự gia tăng độ phức tạp của các hệ thống này vượt quá khả năng hữu ích của chúng. Ý tưởng ban đầu của việc tạo ra một nền tảng sở hữu độc quyền duy nhất bao quát tất cả các quy trình kinh doanh đã dẫn đến sự tập trung quá mức, nơi mọi thay đổi yêu cầu tài nguyên và thời gian đáng kể để điều chỉnh.

Như chúng ta đã đề cập trong chương "Mạch giống doanh nghiệp: cách dữ liệu kết nối các quy trình kinh doanh", việc làm việc hiệu quả với dữ liệu đòi hỏi phải tạo ra một hệ sinh thái mở và thống nhất, kết nối tất cả các nguồn thông tin.

Các yếu tố chính của hệ sinh thái:

- Chọn kho dữ liệu phù hợp:
  - Đối với bảng và các phép tính, hãy sử dụng cơ sở dữ liệu — chẳng hạn như PostgreSQL hoặc MySQL (Hình. 3.17) -
  - Đối với tài liệu và báo cáo, có thể sử dụng các kho lưu trữ đám mây (Google Drive, OneDrive) hoặc các hệ thống hỗ trợ định dạng JSON.
  - Tìm hiểu về khả năng của Data Warehouse, Data Lakes và các công cụ khác để lưu trữ và phân tích khối lượng thông tin lớn một cách tập trung (Hình. 8.18) -
- Triển khai các giải pháp để truy cập dữ liệu độc quyền:

- Nếu bạn đang sử dụng các hệ thống độc quyền, hãy thiết lập truy cập vào chúng thông qua API hoặc SDK để nhận dữ liệu cho xử lý bên ngoài (Hình. 4.12)-
- Tìm hiểu về tiềm năng của các công cụ đảo ngược kỹ thuật cho định dạng CAD (Hình. 4.113)-
- Thiết lập ETL-Pipeline, thu thập dữ liệu định kỳ từ các ứng dụng hoặc máy chủ, chuyển đổi chúng thành các định dạng có cấu trúc mở và lưu trữ trong các kho (Hình. 7.23) -
- Thảo luận trong nhóm về việc đảm bảo truy cập vào dữ liệu mà không cần sử dụng phần mềm độc quyền.
- Hãy nhớ rằng: dữ liệu quan trọng hơn các giao diện. Giá trị lâu dài nằm ở cấu trúc và tính khả dụng của thông tin, chứ không phải là các công cụ giao diện người dùng cụ thể.

- Hãy cân nhắc việc tạo ra một trung tâm xuất sắc (CoE) về dữ liệu, như chúng ta đã thảo luận trong chương "Trung tâm xuất sắc (CoE) về mô hình dữ liệu" hoặc các cách khác để đảm bảo chuyên môn trong công việc với dữ liệu (Hình. 4.39) -

Hệ sinh thái quản lý dữ liệu tạo ra một không gian thông tin thống nhất, nơi mà tất cả các thành viên trong dự án làm việc với thông tin đồng nhất, cập nhật và đã được kiểm tra. Đây là nền tảng cho các quy trình số có thể mở rộng, linh hoạt và đáng tin cậy.

Khám phá tiềm năng dữ liệu: 5-10 bước đến sự trưởng thành số.

Ngoài việc tích hợp công nghệ, một yếu tố quan trọng khác trong việc triển khai các giải pháp kỹ thuật số thành công là sự chấp nhận từ phía người dùng cuối. Việc thu hút khách hàng hoặc người dùng vào các vấn đề là rất quan trọng cho việc cải thiện trải nghiệm người dùng và quản lý thay đổi trong công ty. Nếu giải pháp không phù hợp với quy trình làm việc quen thuộc hoặc không giải quyết được những vấn đề thực sự của người dùng hoặc khách hàng, nó sẽ không được sử dụng, và không có biện pháp hay động viên bổ sung nào có thể khắc phục điều này.

Chuyển đổi là một quá trình lặp đi lặp lại, dựa trên việc phân tích dữ liệu về sự tương tác của người dùng với các quy trình mới, với các chu trình thử nghiệm thường xuyên, phản hồi liên tục và cải tiến.

Bước 5. Xây dựng văn hóa làm việc với dữ liệu, đào tạo nhân viên và thu thập phản hồi.

Ngay cả hệ thống tiên tiến nhất cũng sẽ không hoạt động nếu không có sự tham gia của nhân viên. Cần tạo ra một môi trường mà trong đó dữ liệu được sử dụng hàng ngày và đội ngũ hiểu rõ giá trị của chúng.

> Trong báo cáo được công bố của chính phủ Vương quốc Anh "Phân tích dữ liệu và AI trong việc triển khai dự án chính phủ" năm 2024, có ghi [83] rằng để triển khai thành công phân tích dữ liệu và AI, chuẩn bị chuyên gia có năng lực cần thiết trong xử lý và giải thích dữ liệu là rất quan trọng.

Thiếu kiến thức trong lĩnh vực phân tích dữ liệu là một trong những vấn đề chính hạn chế quá trình chuyển đổi số. Các nhà điều hành đã quen với các quy trình cố định: các chu kỳ hàng quý, các sáng kiến ưu tiên và các con đường truyền thống để thực hiện dự án. Để thay đổi, cần có một nhà lãnh đạo đặc biệt — đủ cấp cao để có ảnh hưởng, nhưng không đủ cao để có thời gian và động lực điều hành một dự án chuyển đổi dài hạn.

Các hành động chính:

- Nhận thức về sự cần thiết phải chuyển từ các quyết định dựa trên ý kiến của nhân viên có mức lương cao (HiPPO) sang một nền văn hóa ra quyết định dựa trên sự kiện và dữ liệu, như đã thảo luận trong chương "HiPPO hoặc nguy hiểm của ý kiến trong việc ra quyết định".-
- Tổ chức đào tạo hệ thống:
    - Tiến hành các khóa đào tạo về việc sử dụng dữ liệu có cấu trúc và mời các chuyên gia từ các lĩnh vực kinh tế khác, những người không bị thiên lệch về các sản phẩm và khái niệm phổ biến hiện nay trong ngành xây dựng.
    - Thảo luận các phương pháp và công cụ phân tích dữ liệu với đồng nghiệp, cũng như tự học cách làm việc thực tế với các công cụ như Python, pandas và LLM.-
    - Tạo thư viện tài liệu học tập (tốt hơn cả là với các video ngắn) về chủ đề cấu trúc dữ liệu và tạo mô hình dữ liệu.--
- Sử dụng công nghệ học tập hiện đại:
    - Sử dụng các mô hình ngôn ngữ (LLM) để hỗ trợ khi làm việc với mã và dữ liệu, bao gồm tạo, tái cấu trúc và phân tích mã, cũng như xử lý và diễn giải thông tin bảng.
    - Nghiên cứu cách mã được tạo ra bằng LLM có thể được điều chỉnh và tích hợp vào giải pháp Pipeline hoàn chỉnh khi làm việc trong môi trường phát triển ngoại tuyến (IDE).-

> Khi nhà lãnh đạo tiếp tục ra quyết định "theo cách cũ", không có khóa đào tạo nào có thể khiến mọi người nghiêm túc xem xét phân tích.

Việc hình thành văn hóa làm việc với dữ liệu là không thể nếu thiếu phản hồi thường xuyên. Phản hồi cho phép phát hiện những thiếu sót trong các quy trình, công cụ và chiến lược mà không thể phát hiện qua các báo cáo nội bộ hoặc các chỉ số KPI chính thức. Những nhận xét khen ngợi từ người dùng giải pháp của bạn sẽ không mang lại giá trị thực tiễn. Giá trị đến từ phản hồi mang tính phê bình, đặc biệt nếu nó dựa trên những quan sát và sự kiện cụ thể. Tuy nhiên, việc nhận được thông tin như vậy đòi hỏi nỗ lực: cần xây dựng các quy trình mà ở đó người tham gia — cả nội bộ và bên ngoài — có thể chia sẻ ý kiến (có thể là ẩn danh) mà không bị bóp méo và không lo lắng rằng ý kiến của họ có thể ảnh hưởng đến công việc của chính họ. Quan trọng là họ phải thực hiện điều này mà không bị bóp méo và không sợ các hậu quả tiêu cực cho bản thân.

> *Bất kỳ hình thức đào tạo nào cuối cùng cũng là tự học.*
>
> *– Milton Friedman, nhà kinh tế và thống kê người Mỹ.*

Việc triển khai các công cụ phân tích cần phải kèm theo việc xác thực định kỳ hiệu quả của chúng trong thực tế (ROI, KPI), điều này chỉ có thể đạt được thông qua phản hồi có cấu trúc từ nhân viên, khách hàng và đối tác. Điều này cho phép các công ty không chỉ tránh lặp lại sai lầm mà còn nhanh chóng thích ứng với những thay đổi của môi trường. Sự hiện diện của cơ chế thu thập và phân tích phản hồi là một trong những dấu hiệu chứng minh sự trưởng thành của tổ chức, chuyển từ những sáng kiến số theo từng giai đoạn sang mô hình bền vững của sự cải tiến liên tục (Hình. 2.25).

Bước 6. Từ các dự án thí điểm đến mở rộng quy mô

> *Hãy chọn những trận đánh đủ lớn để có ý nghĩa, và đủ nhỏ để có thể chiến thắng.– Jonathan Kozol*

Việc khởi động chuyển đổi số "ngay lập tức và ở mọi nơi" là cực kỳ rủi ro. Cách tiếp cận hiệu quả hơn là bắt đầu với các dự án thí điểm và dần dần mở rộng quy mô các thực tiễn thành công.

Các hành động chính:

- Chọn một dự án phù hợp cho thí điểm:
  - Xác định một nhiệm vụ kinh doanh hoặc quy trình cụ thể với các kết quả có thể đo lường (KPI, ROI) (Hình. 7.15)-
  - Chọn quy trình tự động hóa ETL, chẳng hạn như kiểm tra dữ liệu tự động hoặc tính toán khối lượng công việc (QTO) bằng Python và Pandas (Hình. 5.210)-
  - Thiết lập các chỉ số thành công rõ ràng (chẳng hạn - giảm thời gian chuẩn bị đặc tả kiểm tra hoặc báo cáo kiểm tra dữ liệu từ một tuần xuống một ngày)
- Áp dụng các phương pháp lặp lại:
  - Bắt đầu với các quy trình đơn giản trong việc chuyển đổi dữ liệu và tạo ra quy trình chuyển đổi dữ liệu đa định dạng thành các định dạng cần thiết cho quy trình của bạn (Hình. 4.12, Hình. 4.15)--
  - Tăng dần độ phức tạp của các nhiệm vụ và mở rộng tự động hóa quy trình, tạo ra một Pipeline hoàn chỉnh trong môi trường phát triển (IDE) dựa trên các khối mã đã được tài liệu hóa (Hình. 4.17, Hình. 7.218)--
  - Tài liệu hóa và ghi lại (tốt nhất bằng cách sử dụng video ngắn) các giải pháp thành công và chia sẻ chúng với đồng nghiệp hoặc trong các cộng đồng chuyên nghiệp
- Phát triển các mẫu và tài liệu đi kèm để có thể nhân rộng các giải pháp tương tự, để đồng nghiệp của bạn (hoặc các thành viên trong cộng đồng chuyên nghiệp, bao gồm cả người dùng trên các mạng xã hội) có thể sử dụng hiệu quả

Cách tiếp cận “từng bước” cho phép duy trì chất lượng cao của các thay đổi và không rơi vào hỗn loạn của các triển khai đồng thời. Chiến lược "từ nhỏ đến lớn" giảm thiểu rủi ro và cho phép học hỏi từ những sai lầm nhỏ, không cho phép chúng trở thành các vấn đề nghiêm trọng.

Việc chuyển từ phương pháp dự án, trong đó nhân viên chỉ tham gia một phần, sang việc hình thành các đội thường trực (chẳng hạn như trung tâm chuyên môn - CoE) cho phép đảm bảo sự phát triển bền vững của sản phẩm ngay cả sau khi phiên bản đầu tiên ra mắt. Những đội ngũ này không chỉ hỗ trợ các giải pháp hiện có mà còn tiếp tục cải tiến chúng.

Điều này làm giảm sự phụ thuộc vào các phê duyệt kéo dài: các thành viên trong đội được quyền đưa ra

quyết định trong phạm vi trách nhiệm của họ. Kết quả là, các nhà quản lý được giải phóng khỏi việc quản lý vi mô, và các đội có thể tập trung vào việc tạo ra giá trị thực sự.

> Phát triển các giải pháp mới không phải là một cuộc chạy nước rút, mà là một cuộc thi marathon. Những ai thành công trong lĩnh vực này là những người có định hướng từ đầu vào công việc liên tục và lâu dài.

Quan trọng là phải hiểu rằng công nghệ đòi hỏi sự phát triển liên tục. Đầu tư vào sự phát triển lâu dài của các giải pháp công nghệ là nền tảng cho công việc thành công.

Bước 7. Sử dụng định dạng dữ liệu và giải pháp mở.

Như chúng ta đã thảo luận trong các chương liên quan đến các nền tảng mô-đun (ERP, PMIS, CAFM, CDE, v.v.), việc tập trung vào các định dạng dữ liệu mở và phổ quát là rất quan trọng, điều này đảm bảo tính độc lập khỏi các giải pháp của nhà cung cấp và tăng cường khả năng tiếp cận thông tin cho tất cả các bên tham gia quy trình.

Các hành động chính:

- Chuyển từ các định dạng đóng sang mở:
  - Sử dụng định dạng mở thay vì định dạng độc quyền, hoặc tìm cách thiết lập xuất tự động hoặc chuyển đổi các định dạng đóng sang mở.
  - Triển khai các công cụ để làm việc với Parquet, CSV, JSON, XLSX, đây là các tiêu chuẩn trao đổi giữa hầu hết các hệ thống hiện đại.-
  - Nếu công việc với hình học 3D đóng vai trò quan trọng trong các quy trình của bạn, hãy xem xét việc sử dụng các định dạng mở như USD, glTF, DAE hoặc OBJ.
- Sử dụng cơ sở dữ liệu vector để phân tích và tìm kiếm thông tin hiệu quả:
  - Sử dụng Bounding Box và các phương pháp khác để đơn giản hóa công việc với hình học 3D.-
  - Nghĩ về nơi có thể triển khai vector hóa dữ liệu — chuyển đổi văn bản, đối tượng hoặc tài liệu thành các đại diện số.
- Áp dụng các công cụ phân tích dữ liệu lớn:
  - Tổ chức lưu trữ dữ liệu lịch sử tích lũy (ví dụ PDF, XLSX, CAD) trong các định dạng phù hợp để phân tích (Apache Parquet, CSV, ORC).-
  - Bắt đầu áp dụng các phương pháp thống kê cơ bản và làm việc với các mẫu đại diện — hoặc ít nhất, hãy tìm hiểu các nguyên tắc cơ bản của thống kê.
  - Triển khai và tìm hiểu các công cụ trực quan hóa dữ liệu và mối liên hệ giữa các dữ liệu để thể hiện rõ ràng kết quả phân tích. Mà không có trực quan hóa chất lượng, sẽ không thể hiểu một cách đầy đủ về cả dữ liệu và các quy trình dựa trên chúng.

Việc chuyển sang các định dạng dữ liệu mở và triển khai các công cụ để phân tích, lưu trữ và trực quan hóa thông tin tạo nền tảng cho quản lý kỹ thuật số bền vững và độc lập. Điều này không chỉ giảm sự phụ thuộc vào nhà cung cấp mà còn đảm bảo quyền truy cập bình đẳng vào dữ liệu cho tất cả các bên tham gia quy

trình.

Bước 8. Bắt đầu triển khai máy học để dự đoán.

Nhiều công ty đã tích lũy được những khối dữ liệu khổng lồ - một dạng "geyser thông tin", vẫn còn chưa được khai thác. Những dữ liệu này được thu thập trong hàng trăm và hàng nghìn dự án, nhưng thường chỉ được sử dụng một lần hoặc hoàn toàn không được đưa vào các quy trình tiếp theo. Các tài liệu và mô hình được lưu trữ trong các định dạng và hệ thống kín thường bị coi là vật cản lỗi thời và vô dụng. Tuy nhiên, chính chúng lại là nguồn tài nguyên quý giá - nền tảng cho việc phân tích các lỗi đã xảy ra, tự động hóa các hoạt động hàng ngày và phát triển các giải pháp sáng tạo về phân loại tự động và nhận diện các yếu tố trong các dự án tương lai.

Nhiệm vụ chính là học cách trích xuất những dữ liệu này và chuyển đổi chúng thành giá trị thực tiễn. Như đã được thảo luận trong chương "Học máy và dự đoán", các phương pháp học máy có thể tăng đáng kể độ chính xác của các đánh giá và dự đoán trong các quy trình liên quan đến xây dựng. Việc khai thác toàn diện dữ liệu đã tích lũy mở ra con đường dẫn đến tăng hiệu quả, giảm rủi ro và xây dựng các quy trình số bền vững.

Các hành động chính:

- Bắt đầu với các thuật toán đơn giản:
  - Hãy thử áp dụng hồi quy tuyến tính - với sự hỗ trợ từ LLM - để dự đoán các chỉ số lặp lại trong các tập dữ liệu, nơi các mối quan hệ từ nhiều yếu tố không có hoặc tối thiểu.
  - Xem xét các giai đoạn trong quy trình của bạn mà lý thuyết có thể áp dụng thuật toán k-láng giềng gần nhất (k-NN) - ví dụ, cho các nhiệm vụ phân loại, đánh giá sự tương đồng của các đối tượng hoặc dự đoán dựa trên các trường hợp lịch sử.
- Thu thập và cấu trúc dữ liệu để đào tạo các mô hình:
  - Tập hợp dữ liệu lịch sử về các dự án tại một nơi và định dạng đồng nhất.
  - Làm việc để nâng cao chất lượng và tính đại diện của các tập dữ liệu huấn luyện, thông qua ETL tự động.-
  - Hãy học cách phân chia dữ liệu thành các tập huấn luyện và thử nghiệm, như chúng ta đã làm trong ví dụ với tập dữ liệu Titanic.-
- Xem xét các cơ hội mở rộng áp dụng các phương pháp học máy cho một loạt các nhiệm vụ - từ dự đoán thời gian hoàn thành dự án đến tối ưu hóa logistics, quản lý nguồn lực và phát hiện sớm các vấn đề tiềm ẩn.

Học máy là một công cụ cho phép biến dữ liệu lưu trữ thành một tài sản quý giá cho việc dự đoán, tối ưu hóa và ra quyết định có căn cứ. Bắt đầu với các tập dữ liệu nhỏ và các mô hình đơn giản, từ từ tăng cường độ phức tạp.

Bước 9. Tích hợp IoT và công nghệ thu thập dữ liệu hiện đại.

Thế giới xây dựng đang nhanh chóng trở nên số hóa: mỗi bức ảnh từ công trường, mỗi tin nhắn trong Teams - tất cả đã trở thành một phần của một quy trình lớn về tham số hóa và token hóa thực tại. Như GPS đã từng cách mạng hóa logistics, công nghệ IoT, RFID và thu thập dữ liệu tự động đang thay đổi ngành xây dựng.

Như đã đề cập trong chương "IoT Internet vạn vật và hợp đồng thông minh", công trường số với cảm biến và giám sát tự động là tương lai của ngành này.

Các hành động chính:

- Triển khai các thiết bị IoT, thẻ RFID và chi tiết hóa các quy trình liên quan đến chúng:
  - Đánh giá xem ở những khu vực hoặc giai đoạn nào của dự án việc lắp đặt cảm biến có thể mang lại lợi tức đầu tư cao nhất (ROI) - chẳng hạn, để giám sát nhiệt độ, rung động, độ ẩm hoặc chuyển động của thiết bị.
  - Xem xét việc ứng dụng RFID để theo dõi vật liệu, công cụ và thiết bị trên tất cả các giai đoạn của chuỗi cung ứng.
  - Tìm hiểu cách tích hợp dữ liệu thu thập được vào một hệ thống thông tin duy nhất, chẳng hạn như Apache NiFi, để xử lý và phân tích tự động theo thời gian thực (Hình 7.45).-
- Tạo ra hệ thống giám sát thời gian thực:
  - Phát triển bảng điều khiển để theo dõi các chỉ số chính của quy trình hoặc dự án bằng cách sử dụng các công cụ trực quan hóa như Streamlit, Flask hoặc Power BI.
  - Thiết lập thông báo tự động để cảnh báo về những sai lệch nghiêm trọng so với kế hoạch hoặc tiêu chuẩn (Hình 7.42).
  - Đánh giá tiềm năng bảo trì dự đoán thiết bị dựa trên dữ liệu thu thập và các mẫu đã phát hiện (Hình 9.36).
- Kết hợp dữ liệu từ nhiều nguồn khác nhau:
  - Bắt đầu với việc trực quan hóa mô hình dữ liệu ở cấp độ vật lý - phản ánh cấu trúc luồng thông tin và các thông số chính thu thập từ hệ thống CAD, thiết bị IoT và nền tảng ERP (Hình 4.31).
  - Bắt đầu bằng việc tạo ra một mô tả thô về nền tảng duy nhất, dành cho phân tích dữ liệu và hỗ trợ ra quyết định quản lý. Ghi lại các chức năng chính, nguồn dữ liệu, người dùng, cũng như các kịch bản ứng dụng dự kiến (Hình 4.37).

Càng sớm kết nối quy trình thực tế với thế giới số, bạn càng nhanh chóng có thể quản lý chúng bằng dữ liệu - một cách hiệu quả, minh bạch và theo thời gian thực.

Bước 10. Chuẩn bị cho tương lai những thay đổi trong ngành.

Các công ty xây dựng luôn chịu áp lực từ môi trường bên ngoài: các cuộc khủng hoảng kinh tế, những bước nhảy công nghệ, và các thay đổi quy định. Giống như một khu rừng phải chịu đựng mưa, tuyết, hạn hán và ánh nắng gay gắt, các công ty sống trong điều kiện liên tục thích ứng. Và giống như những cây cối phát triển khả năng chống lạnh và hạn hán nhờ vào hệ thống rễ sâu, chỉ những tổ chức có nền tảng vững chắc bằng các quy trình tự động hóa, khả năng dự đoán sự thay đổi và thích ứng linh hoạt với chiến lược mới, mới có thể giữ vững sự sống còn và cạnh tranh.

Như đã đề cập trong chương "Chiến lược sinh tồn: xây dựng lợi thế cạnh tranh", ngành xây dựng đang bước vào giai đoạn chuyển mình mạnh mẽ. Tương tác giữa khách hàng và nhà thầu đang tiến tới mô hình Uber hóa, nơi tính minh bạch, tính dự đoán và các công cụ số đang thay thế phương pháp truyền thống. Trong thực tế mới này, lợi thế không thuộc về những người lớn nhất mà thuộc về những người linh hoạt và có công

nghệ trưởng thành.

Các hành động chính:

- Phân tích điểm yếu của doanh nghiệp trong bối cảnh dữ liệu mở:
  - Đánh giá cách mà việc dân chủ hóa quyền truy cập dữ liệu trong khuôn khổ Uber hóa có thể tác động tiêu cực đến lợi thế cạnh tranh và doanh nghiệp của bạn (Hình 10.15).
  - Hãy suy nghĩ về chiến lược chuyển đổi từ các quy trình không rõ ràng và cô lập sang các mô hình kinh doanh dựa trên các giải pháp mở, khả năng tương thích hệ thống và tính minh bạch của dữ liệu (Hình 2.25)-
- Phát triển chiến lược số lâu dài:
  - Xác định xem bạn có muốn trở thành người dẫn đầu trong đổi mới hay thích kịch bản "theo sau", trong đó bạn sẽ tiết kiệm nguồn lực của mình.
  - Phác thảo các giai đoạn: ngắn hạn (tự động hóa quy trình, cấu trúc hóa dữ liệu), trung hạn (triển khai LLM và ETL), dài hạn (hệ sinh thái số, kho dữ liệu tập trung).
- Suy nghĩ về việc mở rộng danh mục dịch vụ:
  - Xem xét khả năng cung cấp các dịch vụ mới (hướng tới hiệu quả năng lượng, ESG, dịch vụ xử lý dữ liệu). Chi tiết về các mô hình kinh doanh mới sẽ được thảo luận ở chương tiếp theo.
  - Cố gắng định vị bản thân như một đối tác công nghệ đáng tin cậy, đồng hành suốt vòng đời của vật thể - từ thiết kế đến vận hành. Niềm tin vào bạn phải dựa trên cách tiếp cận hệ thống, tính minh bạch của quy trình và khả năng cung cấp các giải pháp công nghệ bền vững.

Trong bối cảnh chuyển đổi, người thắng cuộc không phải là những người chỉ phản ứng với thay đổi, mà là những người hành động trước. Sự linh hoạt, cởi mở và độ trưởng thành số là nền tảng cho sự bền vững trong xây dựng tương lai.

Bản đồ đường đi của quá trình chuyển đổi: từ hỗn loạn đến công ty dựa trên dữ liệu.

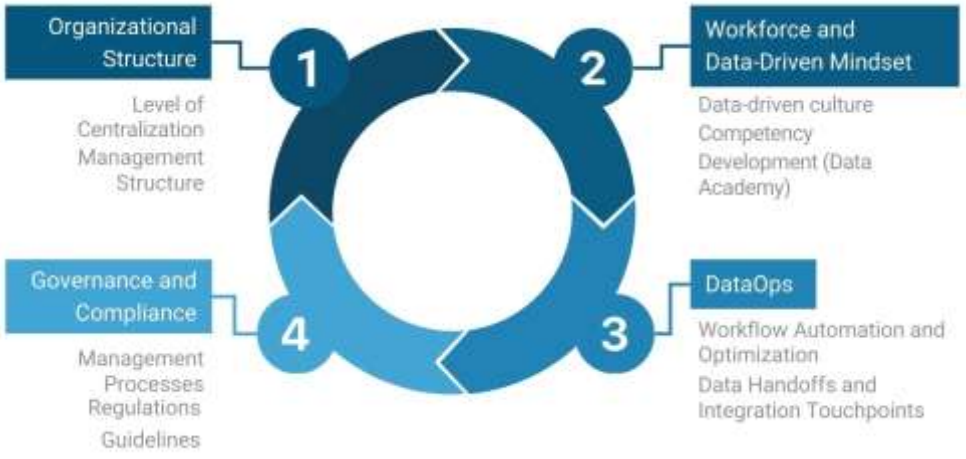
Kế hoạch tiếp theo có thể phục vụ như một định hướng sơ bộ - điểm khởi đầu cho việc hình thành chiến lược chuyển đổi số dựa trên dữ liệu của riêng bạn:

- Kiểm toán và tiêu chuẩn: phân tích tình trạng hiện tại, chuẩn hóa dữ liệu.
- Cấu trúc và phân loại dữ liệu: tự động hóa việc chuyển đổi dữ liệu phi cấu trúc và kém cấu trúc.
- Tự động hóa nhóm, tính toán và chi phí: sử dụng các công cụ và thư viện mở để tự động hóa.
- Hệ sinh thái và COE: tạo một đội ngũ nội bộ để hình thành hệ sinh thái dữ liệu thống nhất trong công ty.
- Văn hóa và đào tạo: rời xa các quyết định dựa trên HiPPO sang các quyết định dựa trên dữ liệu.
- Thử nghiệm, phản hồi và mở rộng: hành động theo từng giai đoạn: thử nghiệm các phương pháp mới trong quy mô hạn chế, thu thập phản hồi hợp lý và từ từ mở rộng các giải pháp.

- Định dạng mở: sử dụng các định dạng phổ quát và mở để độc lập với nhà cung cấp phần mềm.
- Machine Learning: triển khai các thuật toán ML vào quy trình để dự đoán và tối ưu hóa.
- IoT và công trường xây dựng số: tích hợp các công nghệ thu thập dữ liệu hiện đại vào quy trình.
- Thích ứng chiến lược: chuẩn bị cho các thay đổi trong ngành trong tương lai.

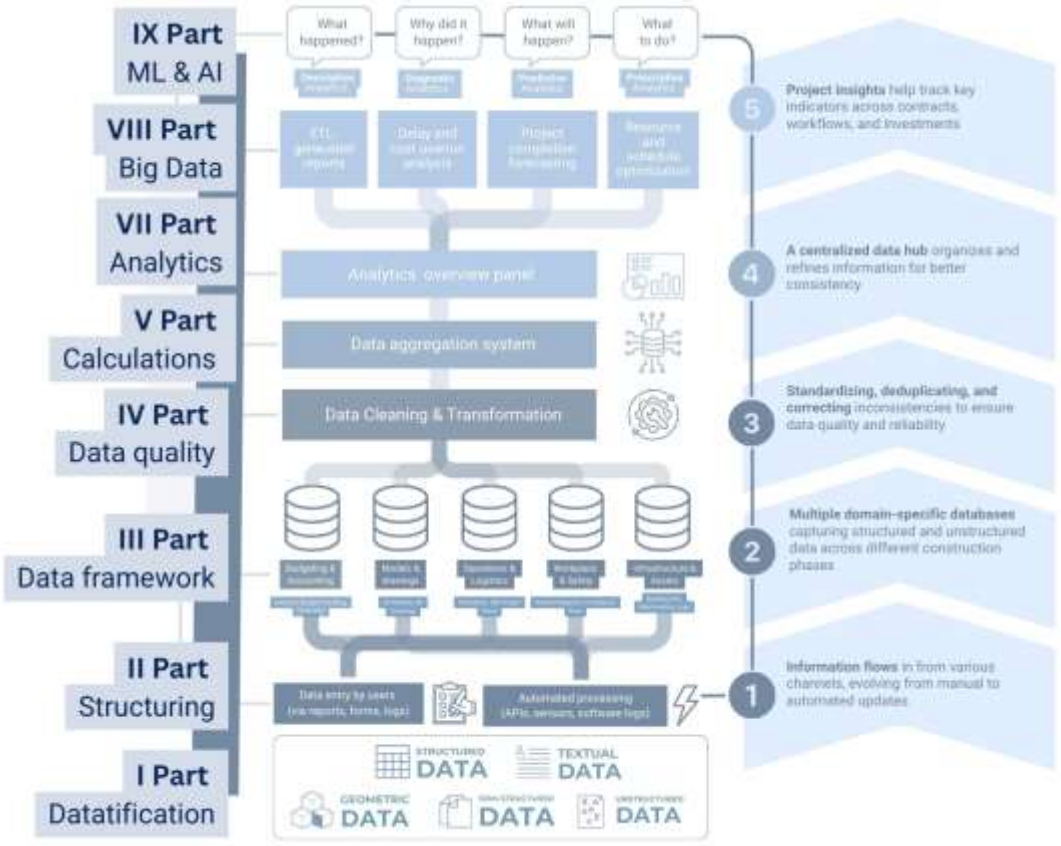
> Điều quan trọng - hãy nhớ rằng "dữ liệu một mình không thay đổi công ty: chính con người biết cách làm việc với dữ liệu đó mới thay đổi công ty". Đặt cược vào văn hóa, quy trình minh bạch và cam kết cải tiến liên tục.

Phương pháp hệ thống cho phép chuyển từ các sáng kiến số rời rạc sang một mô hình quản lý dựa trên dữ liệu hoàn chỉnh, trong đó quyết định không dựa trên cảm tính hay giả định, mà dựa trên dữ liệu, sự thật và xác suất tính toán bằng toán học. Chuyển đổi số trong ngành xây dựng không chỉ là việc triển khai công nghệ, mà là việc hình thành một hệ sinh thái kinh doanh, nơi thông tin về dự án được truyền tải một cách liền mạch và lặp đi lặp lại giữa các hệ thống khác nhau. Các thuật toán máy học cung cấp phân tích tự động, liên tục, dự đoán và tối ưu hóa quy trình. Trong môi trường như vậy, suy đoán và dữ liệu ẩn trở nên kém quan trọng - chỉ còn lại các mô hình đã được chứng minh, các phép tính minh bạch và kết quả có thể dự đoán được.



Hình 10.23 Các yếu tố chính của quản lý dữ liệu thành công ở cấp độ công ty.

Mỗi phần của cuốn sách tương ứng với một giai đoạn cụ thể trong quá trình xử lý và phân tích dữ liệu trong các dự án xây dựng (Hình 2.25). Nếu bạn muốn quay lại một trong những chủ đề đã được đề cập trước đây và nhìn nhận nó từ góc độ hiểu biết tổng thể về luồng sử dụng dữ liệu - bạn có thể tham khảo các tiêu đề phần được chỉ ra trong Hình 10.24.-



Hình 10.24 Các phần của cuốn sách trong bối cảnh quy trình xử lý dữ liệu (Hình 2.25): từ số hóa thông tin đến phân tích và trí tuệ nhân tạo.-

Bất kể quy mô tổ chức, mức độ trưởng thành công nghệ hay ngân sách của bạn, bạn có thể bắt đầu tiến tới một cách tiếp cận dựa trên dữ liệu ngay hôm nay. Ngay cả những bước nhỏ trong hướng đi đúng sẽ dẫn đến những kết quả theo thời gian.

> Chuyển đổi dựa trên dữ liệu không phải là một dự án đơn lần, mà là một quá trình cải tiến liên tục, lặp đi lặp lại, bao gồm việc triển khai các công cụ mới, xem xét lại các quy trình và phát triển văn hóa ra quyết định dựa trên dữ liệu.

Xây dựng trong ngành công nghiệp 5.0: cách kiếm tiền khi không thể giấu diếm nữa.

Trong một thời gian dài, các công ty xây dựng kiếm tiền từ sự thiếu minh bạch của quy trình. Mô hình kinh doanh chủ yếu trở thành sự suy đoán - nâng cao giá trị vật liệu, khối lượng công việc và phần trăm phụ phí trong các hệ thống ERP, PMIS kín, không thể kiểm toán bởi bên ngoài. Việc hạn chế tiếp cận của khách hàng và các bên đáng tin cậy của họ đến dữ liệu dự án ban đầu đã tạo điều kiện cho những sơ hở, trong đó việc kiểm tra độ tin cậy của các phép tính trở nên gần như không thể.

Tuy nhiên, mô hình này đang nhanh chóng mất đi tính. Với sự dân chủ hóa quyền truy cập vào dữ liệu, sự xuất hiện của LLM, việc ra mắt các dữ liệu mở, cũng như các công cụ ETL tự động hóa, ngành công nghiệp đang chuyển mình sang một tiêu chuẩn công việc mới.

> Kết quả là, sự không minh bạch không còn là lợi thế cạnh tranh - sớm muộn gì nó cũng trở thành gánh nặng khó từ bỏ. Sự minh bạch từ tùy chọn trở thành điều kiện bắt buộc để tồn tại trên thị trường.

Vậy ai sẽ làm việc với các khách hàng - ngân hàng, nhà đầu tư, khách hàng cá nhân, vốn tư nhân, nhà thầu nhà nước - trong thực tế số mới? Câu trả lời rõ ràng: với những người có khả năng cung cấp không chỉ kết quả, mà còn là lý do cho từng bước đi đến nó. Trong bối cảnh khối lượng dữ liệu mở ngày càng tăng, các đối tác và khách hàng sẽ lựa chọn các công ty đảm bảo tính minh bạch, chính xác và dự đoán được các kết quả.

Trên nền tảng đó, các mô hình kinh doanh mới đang hình thành, trong đó không dựa vào sự suy đoán, mà là quản lý dữ liệu và niềm tin.

- Bán quy trình thay vì mét vuông: tài sản then chốt trở thành niềm tin và hiệu quả, không phải là các thỏa thuận giảm giá bê tông. Giá trị chính sẽ là sự dự đoán kết quả dựa trên dữ liệu đáng tin cậy và đã được xác thực. Các công ty hiện đại sẽ không bán đối tượng xây dựng như vậy, mà là:
  - thời gian chính xác và lịch làm việc minh bạch;
  - bảng dự toán hợp lý, được xác nhận bằng tính toán;
  - khả năng theo dõi và kiểm soát số hóa toàn bộ ở mọi bước của dự án.
- Kỹ thuật và phân tích như một dịch vụ: mô hình “Dữ liệu như Dịch vụ” (phương thức cung cấp dữ liệu sẵn có cho người dùng qua internet như một dịch vụ), nơi mỗi dự án trở thành một phần của chuỗi dữ liệu số, và giá trị doanh nghiệp nằm ở khả năng quản lý chuỗi đó. Các công ty chuyển mình thành những nền tảng thông minh, cung cấp các giải pháp dựa trên tự động hóa và phân tích:
  - lập bảng dự toán và kế hoạch tự động và minh bạch;
  - đánh giá rủi ro và thời gian dựa trên thuật toán học máy;
  - tính toán các chỉ số môi trường (ESG, CO_2, hiệu quả năng lượng);
  - hình thành báo cáo từ các nguồn mở có thể kiểm tra.
- Sản phẩm hóa kinh nghiệm kỹ thuật: những sản phẩm công ty phát triển có thể được tái sử dụng bên trong công ty và phát tán như một sản phẩm riêng biệt — tạo ra nguồn thu nhập bổ sung thông qua các dịch vụ số. Trong điều kiện mới, các công ty không chỉ tạo ra dự án mà còn tạo ra tài sản

số:

- thư viện thành phần và mẫu dự toán;
- các mô-đun kiểm tra tự động;
- plug-in và script mã nguồn mở (bán tư vấn) cho việc xử lý dữ liệu.

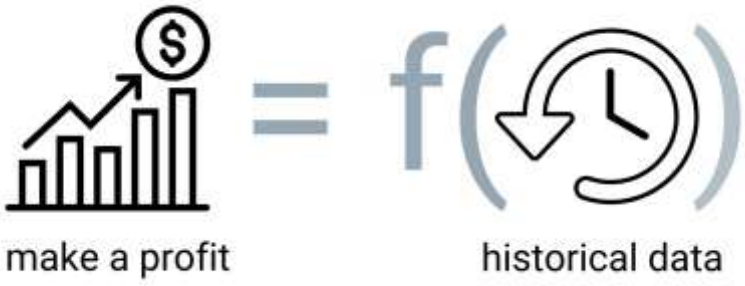
- Loại hình công ty mới: Tích hợp dựa trên Dữ liệu: một bên tham gia thị trường không phụ thuộc vào các nhà cung cấp phần mềm hoặc hệ thống mô-đun cụ thể và không "bị kẹt" trong giao diện của một phần mềm duy nhất. Họ hoạt động tự do với dữ liệu — và dựa vào điều này để xây dựng tính cạnh tranh. Công ty xây dựng của tương lai không chỉ là nhà thầu, mà còn là tích hợp thông tin, có khả năng thực hiện các chức năng sau cho khách hàng:
  - kết hợp dữ liệu từ các nguồn rời rạc và thực hiện phân tích;
  - đảm bảo tính minh bạch và độ tin cậy của quy trình;
  - tư vấn về tối ưu hóa quy trình kinh doanh;
  - phát triển công cụ hoạt động trong hệ sinh thái dữ liệu mở, LLM, ETL và Pipelines.

Ngành công nghiệp 5.0 (Hình. 2.112) đánh dấu sự kết thúc của "thời đại của các hệ số trung bình thủ công" và các cuộc họp buổi tối của giám đốc điều hành với bộ phận dự toán và kế toán. Tất cả những gì trước đây được giấu kín — tính toán, bảng dự toán, khối lượng — trở nên minh bạch, có thể kiểm tra và hiểu được ngay cả với những người không phải chuyên gia. Những người đầu tiên chuyển hướng sẽ hưởng lợi. Tất cả những người còn lại sẽ bị bỏ lại phía sau trong nền kinh tế số mới của lĩnh vực xây dựng.

KẾ LUẬN

Ngành xây dựng bước vào kỷ nguyên thay đổi căn bản. Từ những ghi chép đầu tiên trên các bảng đất sét đến hàng loạt dữ liệu số được đưa ra từ các máy chủ dự án và công trường xây dựng, lịch sử làm việc với thông tin trong lĩnh vực này luôn phản ánh mức độ trưởng thành của công nghệ trong từng thời kỳ. Ngày nay, với sự xuất hiện của tự động hóa, các định dạng mở và hệ thống phân tích thông minh, ngành này đang đối mặt không chỉ với sự tiến hóa dần dần mà còn với một cuộc chuyển đổi số cấp tốc.

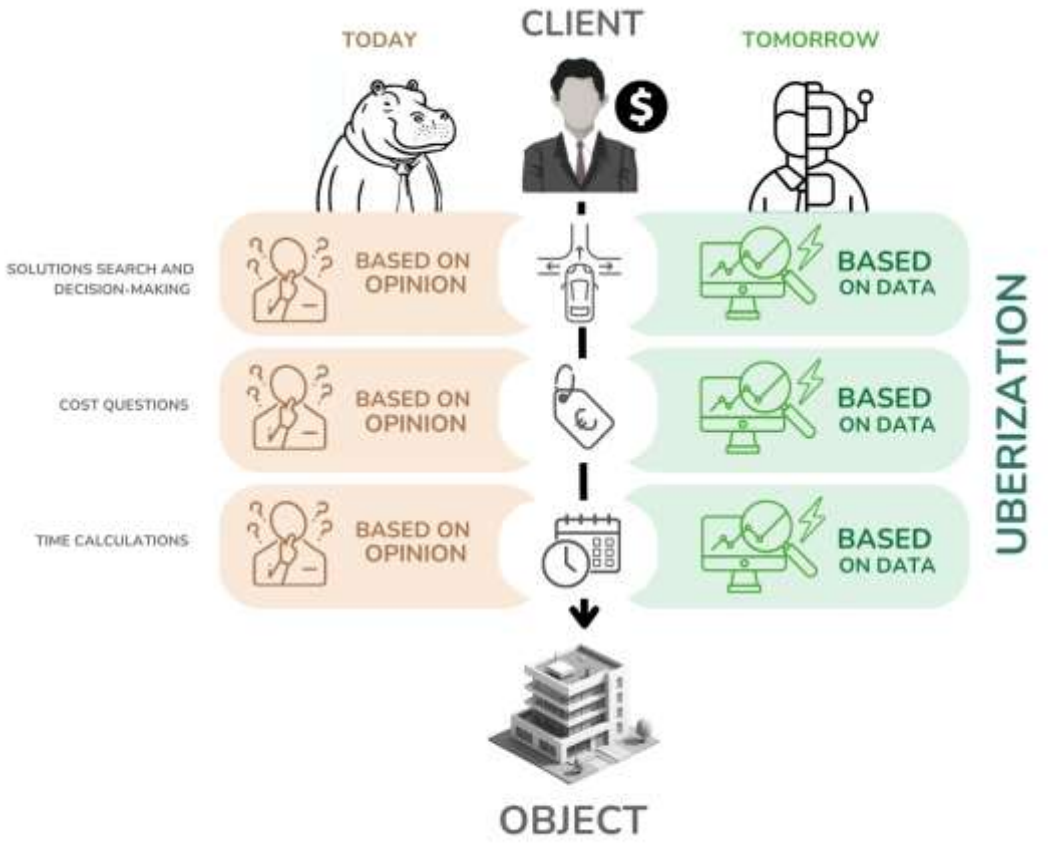
Giống như các lĩnh vực khác trong nền kinh tế, ngành xây dựng cần phải suy nghĩ lại không chỉ về các công cụ mà còn cả các nguyên tắc hoạt động. Các công ty từng chi phối thị trường và đóng vai trò là trung gian chính giữa khách hàng và dự án đang mất đi vị thế độc nhất của mình. Niềm tin và khả năng làm việc với dữ liệu đang được chú trọng: từ việc thu thập và cấu trúc dữ liệu đến phân tích, dự đoán và tự động hóa quyết định.



Rửa. 10.21 Dữ liệu lịch sử được cấu trúc - nhiên liệu cho một doanh nghiệp hiệu quả và có thể quản lý.

Trong cuốn sách này, chúng tôi đã xem xét chi tiết các nguyên tắc chính về quản lý dữ liệu trong ngành xây dựng - từ kiểm toán và chuẩn hóa cho đến tự động hóa quy trình, sử dụng công cụ trực quan hóa và triển khai các thuật toán thông minh. Chúng tôi đã bàn luận về cách mà ngay cả với nguồn lực hạn chế, có thể xây dựng một kiến trúc dữ liệu hoạt động và bắt đầu đưa ra quyết định không dựa trên trực giác mà dựa trên các sự kiện có thể kiểm chứng. Làm việc với dữ liệu không còn là nhiệm vụ chỉ dành cho bộ phận CNTT - nó trở thành nền tảng của văn hóa quản lý, từ đó phụ thuộc vào khả năng linh hoạt, tính thích ứng và độ bền vững lâu dài của công ty.

Việc áp dụng công nghệ học máy, các hệ thống xử lý tự động, các mô hình số và các định dạng mở đã cho phép chúng ta loại bỏ yếu tố con người ở những nơi mà trước đây nó là rất quan trọng. Ngành xây dựng đang tiến tới tự động hóa và khả năng quản lý, nơi mà quá trình từ ý tưởng đến hiện thực hóa dự án có thể so sánh với việc điều hướng trong chế độ tự lái: không phụ thuộc vào các quyết định chủ quan, không cần can thiệp thủ công ở mỗi giai đoạn, nhưng vẫn đảm bảo được sự theo dõi và kiểm soát hoàn toàn bằng số (Rửa. 10.22).-



Рыа. 10.22 Сыа чуыа đổi тыа việc ra quыа́т định dựа трêн ý киến của нhыа́ng chуêн gia quan трọng (HiPPO) sang việc phân тích dữ liệu sẽ đượ thúc đẩy bởi кhách hàng.

Bằng cách nghiên cứu các phương pháp, nguyên tắc và công cụ được trình bày trong cuốn sách này, bạn sẽ có thể bắt đầu đưa ra quyết định trong công ty của mình dựa trên dữ liệu thay vì trực giác. Bạn cũng có thể khởi chạy chuỗi các mô-đun trong LLM, sao chép các quy trình ETL đã sẵn sàng vào môi trường phát triển của mình (IDE) và tự động xử lý dữ liệu, thu được thông tin ở dạng mà bạn cần. Tiến xa hơn nữa, dựa vào các chương trong cuốn sách liên quan đến dữ liệu lớn và học máy, bạn có thể hiện thực hóa các kịch bản phức tạp hơn - khai thác các kiến thức mới từ dữ liệu lịch sử và áp dụng các thuật toán học máy vào việc dự đoán và tối ưu hóa quy trình của mình.

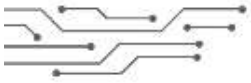
Dữ liệu và quy trình mở sẽ trở thành nền tảng cho những ước lượng chính xác về chi phí và thời gian thực hiện dự án, loại bỏ khả năng các công ty xây dựng có thể đầu cơ ở những dữ liệu không minh bạch. Đây đồng thời là một thách thức và cơ hội cho ngành này để suy ngẫm lại vai trò của mình và thích ứng với môi trường mới, nơi mà tính minh bạch và hiệu quả sẽ trở thành những yếu tố chủ chốt cho sự thành công.

> Sự sẵn sàng tiếp nhận và áp dụng kiến thức vào thực tiễn là chìa khóa dẫn đến thành công trong kỷ nguyên chuyển đổi số.

Các công ty nhận ra điều này sớm nhất sẽ có lợi thế trong bối cảnh cạnh tranh số mới. Nhưng điều quan trọng là phải hiểu rằng dữ liệu tự nó không thay đổi gì cả. Nhiều người sẽ cần thay đổi cách nghĩ truyền thống, và để làm điều đó, cần có động lực. Công ty của bạn nên xem xét lại cách tiếp cận việc chia sẻ dữ liệu.

Những người có khả năng làm việc với dữ liệu, diễn giải chúng, sử dụng để tối ưu hóa và xây dựng kiến trúc quy trình mới từ chúng, chính là những người thay đổi công ty.

Nếu bạn đang đọc những dòng này, bạn đã sẵn sàng cho sự thay đổi và bạn đã tiến một bước trước. Cảm ơn bạn đã chọn con đường này. Chào mừng đến với kỷ nguyên chuyển đổi số!



VỀ TÁC GIẢ



Tôi tên là Artem Boiko. Con đường của tôi tại công trường xây dựng bắt đầu vào năm 2007 - với công việc thợ mỏ tại một mỏ đá phiến, ở thành phố quê hương tôi, trong khi học tại Đại học Mỏ Saint Petersburg chuyên ngành "Xây dựng mỏ và ngầm". Ở mặt sau của bìa sách này, bạn có thể thấy tôi là một thợ nổ trong khu vực mà chúng tôi khai thác và nổ hàng trăm mét khối đá phiến có thể cháy. Sự nghiệp của tôi đã phát triển theo nhiều hướng khác nhau - từ làm công nhân mỏ và xây dựng tàu điện ngầm đến việc leo núi công nghiệp, lắp đặt mái nhà và thiết bị thang máy. Tôi đã có cơ hội tham gia vào nhiều dự án với quy mô khác nhau: từ xây dựng nhà riêng đến các công trình công nghiệp lớn ở nhiều vùng khác nhau trên thế giới.

Theo thời gian, công việc của tôi đã chuyển từ xây dựng vật lý sang quản lý thông tin và quy trình kỹ thuật số. Kể từ năm 2013, tôi đã làm việc ở nhiều vị trí trong các công ty xây dựng nhỏ, vừa và lớn ở nhiều vùng của Đức, từ nhà thiết kế đến quản lý dữ liệu. Về quản lý dữ liệu, kinh nghiệm của tôi bao gồm làm việc với dữ liệu trong các hệ thống ERP, CAD (BIM), MEP, FEM, CMS. Tôi đã tham gia tối ưu hóa, tự động hóa quy trình, cũng như phân tích, máy học và xử lý dữ liệu ở các giai đoạn lập kế hoạch, tính toán và thực hiện công việc xây dựng trong các công ty công trình công nghiệp, nhà ở, hạ tầng và tiện ích.

Từ năm 2003, tôi đã làm việc với phần mềm mã nguồn mở và dữ liệu mở. Trong thời gian đó, tôi đã thực hiện nhiều dự án web - từ trang web và cửa hàng trực tuyến đến ứng dụng web hoàn chỉnh, sử dụng các giải pháp mã nguồn mở và CMS mở. Những nền tảng này, về cơ bản giống với các ERP xây dựng hiện đại, có kiến trúc mô-đun, khả năng thích ứng cao và tính khả dụng. Kinh nghiệm này đã hình thành nên phương pháp làm việc chuyên nghiệp của tôi - tập trung vào công nghệ mở và văn hóa phát triển cộng đồng. Tôi cố gắng thúc đẩy sự tôn trọng đối với mã nguồn mở và việc trao đổi kiến thức tự do trong ngành xây dựng. Công việc của tôi nhằm nâng cao khả năng truy cập dữ liệu trong ngành xây dựng đã thể hiện qua việc tạo ra một số cộng đồng trên mạng xã hội để thảo luận về vấn đề mở dữ liệu và sử dụng mã nguồn mở trong xây dựng, cũng như khởi động một số dự án khởi nghiệp phát triển giải pháp để đảm bảo truy cập vào dữ liệu từ các hệ thống và nền tảng đóng khác nhau.

Đóng góp của tôi cho cộng đồng chuyên nghiệp được thể hiện qua việc tham gia với vai trò diễn giả tại các hội nghị liên quan đến vấn đề khả năng tương tác CAD (BIM), ERP, 4D-5D, và LLM trong machine learning và trí tuệ nhân tạo, cũng như thông qua các bài viết được công bố trên các ấn phẩm châu Âu về ngành xây dựng. Một trong những thành tựu đáng chú ý của tôi là việc tạo ra "Lịch sử BIM", một bản đồ toàn diện về các giải pháp phần mềm quan trọng để quản lý dữ liệu trong ngành xây dựng. Chuỗi bài viết 7 phần của tôi mang tên "Sự phát triển của BIM và các hoạt động vận động hành lang" đã được dịch sang nhiều ngôn ngữ và nhận được sự công nhận rộng rãi như một nỗ lực để làm sáng tỏ những động lực ẩn giấu trong sự phát triển của các tiêu chuẩn số.

Tôi đã trải qua một hành trình từ việc khai thác khoáng sản đến việc thu thập và hệ thống hóa dữ liệu xây dựng. Tôi luôn mở cửa cho các cuộc đối thoại chuyên nghiệp, các ý tưởng mới và các dự án hợp tác. Tôi sẽ rất biết ơn đối với mọi phản hồi và sẽ rất vui nếu bạn gửi tin nhắn cho tôi hoặc theo dõi tôi trên các mạng xã hội. Cảm ơn bạn rất nhiều vì đã đọc cuốn sách này đến cuối!

Phản hồi

Ý kiến của độc giả đóng vai trò quan trọng trong việc phát triển các ấn phẩm và chọn lựa các chủ đề ưu tiên. Những nhận xét về việc những ý tưởng nào hữu ích và những ý tưởng nào gây nghi ngờ và cần thêm giải thích hoặc chỉ dẫn nguồn là vô cùng quý giá. Cuốn sách bao gồm một loạt các tài liệu và đánh giá phân tích, một số trong số đó có thể gây tranh cãi hoặc chủ quan. Nếu trong quá trình đọc bạn phát hiện sai sót, nguồn được chỉ định không đúng, sự không nhất quán logic hoặc lỗi chính tả — tôi sẽ rất biết ơn nếu bạn gửi cho tôi ý kiến, suy nghĩ hoặc phê bình của bạn qua địa chỉ: boikoartem@gmail.com. Hoặc qua tin nhắn trên LinkedIn: linkedin.com/in/boikoartem

Tôi sẽ rất biết ơn nếu bạn đề cập đến cuốn sách Data-Driven Construction trên các mạng xã hội - việc trao đổi kinh nghiệm đọc giúp phổ biến thông tin về dữ liệu mở và các công cụ và hỗ trợ cho công việc của tôi.

NHẬN XÉT VỀ BẢN DỊCH

Cuốn sách này được dịch nhờ vào công nghệ trí tuệ nhân tạo. Điều này đã giúp tăng tốc đáng kể tiến trình dịch thuật. Tuy nhiên, như trong bất kỳ hoạt động công nghệ nào, có thể xảy ra sai sót hoặc không chính xác. Nếu bạn thấy điều gì đó có vẻ sai hoặc dịch không chính xác, xin vui lòng viết cho tôi. Những nhận xét của bạn sẽ giúp nâng cao chất lượng dịch thuật.

CỘNG ĐỒNG DATADRIVENCONSTRUCTION

Đây là nơi bạn có thể tự do đặt câu hỏi và chia sẻ những vấn đề và giải pháp của mình:

DataDrivenConstruction.io: https://datadrivenconstruction.io

LinkedIn: https://www.linkedin.com/company/datadrivenconstruction/

Twitter: https://twitter.com/datadrivenconst

Telegram: https://t.me/datadrivenconstruction

YouTube: https://www.youtube.com/@datadrivenconstruction

KỸ NĂNG VÀ KHÁI NIỆM KHÁC

Ngoài các nguyên tắc chính trong việc làm việc với dữ liệu trong ngành xây dựng, cuốn sách DataDriven-Construction còn đề cập đến một loạt các khái niệm, chương trình và kỹ năng bổ sung cần thiết cho một chuyên gia làm việc với dữ liệu. Một số trong số đó chỉ được trình bày một cách tổng quan, nhưng lại đóng vai trò quan trọng trong thực tiễn.

Người đọc quan tâm có thể truy cập trang web DataDrivenConstruction.io, nơi cung cấp liên kết đến các tài liệu bổ sung về các kỹ năng chính. Những tài liệu này bao gồm làm việc với Python và Pandas, xây dựng quy trình ETL, ví dụ về xử lý dữ liệu trong các dự án CAD xây dựng, hệ thống xử lý dữ liệu lớn, cũng như các phương pháp hiện đại trong việc trực quan hóa và phân tích dữ liệu xây dựng.

Trong việc chuẩn bị cho cuốn sách “DataDrivenConstruction” và tất cả các ví dụ thực tế, đã sử dụng nhiều công cụ và phần mềm nguồn mở. Tác giả xin cảm ơn các nhà phát triển và đồng tác giả của các giải pháp sau:

- Python và Pandas – nền tảng làm việc với dữ liệu và tự động hóa.
- Scipy, NumPy, Matplotlib và Scikit-Learn – thư viện dành cho phân tích dữ liệu và học máy.
- SQL và Apache Parquet – công cụ để lưu trữ và xử lý khối lượng lớn dữ liệu xây dựng.
- Open Source CAD (BIM) – công cụ nguồn mở làm việc với dữ liệu ở định dạng mở.
- N8n, Apache Airflow, Apache NiFi – hệ thống điều phối và tự động hóa quy trình làm việc.
- DeepSeek, LlaMa, Mistral – Open Source LLM.

Đặc biệt cảm ơn tất cả những người tham gia thảo luận về dữ liệu mở và công cụ trong các cộng đồng nghề nghiệp và mạng xã hội, những người đã giúp cải thiện nội dung và cấu trúc của cuốn sách này bằng những phản hồi, ý kiến và ý tưởng của họ.

Hãy theo dõi sự phát triển của dự án trên trang web DataDrivenConstruction.io, nơi không chỉ công bố các cập nhật cho cuốn sách và các sửa lỗi, mà còn có các chương mới, tài liệu đào tạo và ví dụ thực tế về việc áp dụng các phương pháp được mô tả.

TỐI ĐA TIỆN ÍCH VỚI BẢN IN

Bạn đang cầm trên tay phiên bản kỹ thuật số miễn phí của Data-Driven Construction. Để làm việc thuận tiện hơn và truy cập nhanh chóng vào tài liệu, chúng tôi khuyên bạn nên chú ý đến ấn phẩm in:



- Luôn sẵn sàng: cuốn sách ở định dạng in sẽ trở thành công cụ làm việc đáng tin cậy, cho phép bạn nhanh chóng tìm kiếm và sử dụng các hình ảnh và biểu đồ cần thiết trong bất kỳ tình huống làm việc nào.
- Chất lượng hình ảnh cao: tất cả hình ảnh và đồ họa trong ấn phẩm in được trình bày với chất lượng tốt nhất.
- Truy cập nhanh chóng vào thông tin: điều hướng thuận tiện, khả năng ghi chú, đánh dấu và làm việc với cuốn sách ở bất kỳ đâu.

Khi bạn mua phiên bản in đầy đủ của cuốn sách, bạn nhận được công cụ tiện lợi cho việc làm việc thoải mái và hiệu quả với thông tin: khả năng sử dụng nhanh chóng các tài liệu hình ảnh trong công việc hàng ngày, nhanh chóng tìm thấy các sơ đồ cần thiết và ghi chú. Ngoài ra, việc mua hàng của bạn hỗ trợ việc phổ biến kiến thức mở.

Bạn có thể đặt hàng phiên bản in của cuốn sách tại: datadrivenconstruction.io/books



CƠ HỘI ĐỘC NHẤT ĐỂ ĐỊNH VỊ CHIẾN LƯỢC.

Chúng tôi mời bạn đặt tài liệu quảng cáo trong phiên bản miễn phí của ấn phẩm DataDrivenConstruction. Phiên bản trả phí của ấn phẩm trong năm đầu tiên sau khi phát hành đã thu hút sự chú ý của các chuyên gia từ hơn 50 quốc gia trên thế giới — từ Latin America đến khu vực Châu Á - Thái Bình Dương. Để thảo luận về các điều khoản hợp tác cụ thể và nhận thông tin chi tiết về các cơ hội đặt quảng cáo, vui lòng điền vào mẫu phản hồi trên cổng thông tin chính thức datadrivenconstruction.io hoặc gửi email theo các thông tin liên hệ được chỉ định ở cuối cuốn sách.



CÁC CHƯƠNG CỦA CUỐN SÁCH CÓ SẴN TRÊN TRANG WEB DATADRIVENCONSTRUCTION.IO.

Bạn có thể đọc các chương của cuốn sách Data-Driven Construction trên trang web, nơi các phần của cuốn sách sẽ được công bố dần dần, để bạn có thể nhanh chóng tìm thấy thông tin cần thiết và sử dụng nó trong công việc của mình. Hơn nữa, trên trang web, bạn sẽ tìm thấy nhiều ấn phẩm khác về các chủ đề tương tự, cũng như các ví dụ ứng dụng và giải pháp giúp bạn phát triển kỹ năng và ứng dụng dữ liệu trong xây dựng.



TẢI XUỐNG CÁC PHIÊN BẢN MỚI NHẤT CỦA CUỐN SÁCH TỪ TRANG WEB CHÍNH THỨC.

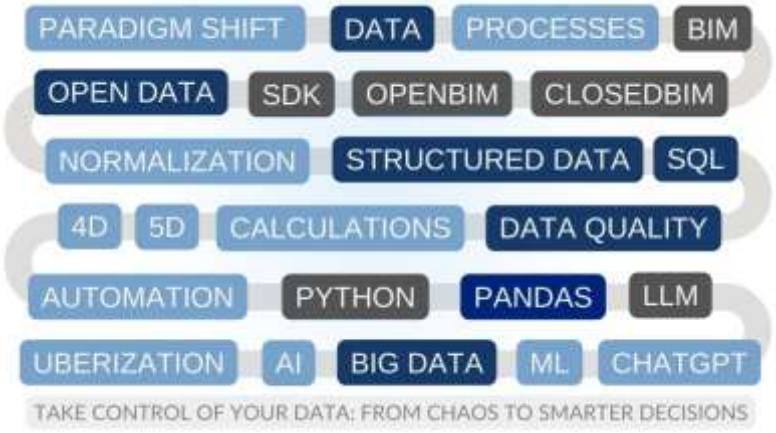
Các phiên bản cập nhật và mới nhất của cuốn sách DataDrivenConstruction có sẵn để tải xuống trên trang web datadrivenconstruction.io. Nếu bạn muốn nhận thông tin cập nhật với các chương mới của cuốn sách, mẹo thực tiễn hoặc đánh giá về các ứng dụng mới, hãy đăng ký nhận bản tin:

- Bạn sẽ là người đầu tiên được tiếp cận các phần mới của cuốn sách.
- Nhận các trường hợp thực tiễn và mẹo về phân tích và tự động hóa trong xây dựng.
- Theo dõi các xu hướng, bài viết và ví dụ về ứng dụng

Truy cập datadrivenconstruction.io để đăng ký!

DATADRIVENCONSTRUCTION: TƯ VẤN, HỘI THẢO VÀ ĐÀO TẠO

Các chương trình đào tạo của DataDrivenConstruction cùng với tư vấn đã giúp hàng chục công ty xây dựng hàng đầu trên toàn cầu nâng cao hiệu quả, giảm chi phí và cải thiện chất lượng quyết định. Trong số những khách hàng của DataDrivenConstruction có những người chơi lớn trong ngành với doanh thu hàng tỷ euro, bao gồm các công ty xây dựng, tư vấn và công nghệ thông tin.



Tại sao nên chọn chúng tôi?

- Tính cập nhật: chúng tôi cung cấp thông tin về các xu hướng và thông tin chi tiết chính của ngành
- Thực tiễn: chúng tôi giúp các chuyên gia giải quyết hiệu quả những nhiệm vụ hàng ngày thông qua PoC.
- Cách tiếp cận cá nhân hóa: chúng tôi xem xét các đặc điểm của doanh nghiệp của bạn, đảm bảo lợi ích tối đa từ đào tạo và tư vấn

Các lĩnh vực chính của đội ngũ DataDrivenConstruction:

- Quản lý chất lượng dữ liệu: chúng tôi giúp xác định thông số cho các nhiệm vụ, thu thập yêu cầu, kiểm tra và chuẩn bị dữ liệu cho xử lý tự động.
- Data Mining - khai thác và cấu trúc dữ liệu: chúng tôi thiết lập quy trình ETL và khai thác dữ liệu từ email, PDF, Excel, hình ảnh và các nguồn khác.
- Phân tích BIM và CAD: chúng tôi thu thập, cấu trúc và phân tích thông tin từ các tệp RVT, IFC, DWG và các định dạng CAD (BIM) khác.
- Phân tích và biến đổi dữ liệu: chúng tôi chuyển đổi thông tin rời rạc thành dữ liệu có cấu trúc, phân tích, kết luận và giải pháp.
- Tích hợp dữ liệu và tự động hóa quy trình: từ việc tự động tạo tài liệu đến tích hợp với các hệ thống nội bộ và cơ sở dữ liệu bên ngoài.

Liên hệ với DataDrivenConstruction.io để tìm hiểu cách sử dụng tự động hóa có thể giúp doanh nghiệp của bạn đạt được những kết quả kinh doanh rõ rệt.

TỪ ĐIỂN

AI (Trí tuệ nhân tạo) - khả năng của các hệ thống máy tính thực hiện các nhiệm vụ thường yêu cầu trí tuệ con người, chẳng hạn như nhận diện hình ảnh, học tập và ra quyết định.

Apache Airflow - nền tảng mở để quản lý quy trình làm việc, cho phép tạo, lên lịch và theo dõi quy trình làm việc và ETL bằng cách sử dụng DAG (đồ thị vô hướng có chu trình).

Apache NiFi - công cụ tự động hóa luồng dữ liệu giữa các hệ thống, chuyên môn về định tuyến và biến đổi dữ liệu.

Apache Parquet - định dạng tệp hiệu quả cho lưu trữ dữ liệu theo cột, được tối ưu hóa cho sử dụng trong các hệ thống phân tích dữ liệu lớn. Đảm bảo nén lớn và xử lý nhanh.

API (Giao diện lập trình ứng dụng) - giao diện chuẩn cho phép một chương trình tương tác với chương trình khác mà không cần truy cập mã nguồn, trao đổi dữ liệu và chức năng qua các yêu cầu và phản hồi tiêu chuẩn hóa.

Thuộc tính - đặc điểm hoặc tính chất của một đối tượng mô tả các đặc điểm của nó (ví dụ: diện tích, thể tích, chi phí, vật liệu).

Cơ sở dữ liệu - cấu trúc tổ chức để lưu trữ, quản lý và truy cập thông tin, được sử dụng để tìm kiếm và xử lý dữ liệu một cách hiệu quả.

Kế hoạch Thực hiện BIM (BEP) - kế hoạch triển khai mô hình thông tin xây dựng, xác định mục tiêu, phương pháp và quy trình áp dụng BIM trong dự án.

Dữ liệu lớn (Big Data) - tập hợp thông tin có khối lượng, tính đa dạng và tốc độ cập nhật đáng kể, yêu cầu công nghệ đặc biệt để xử lý và phân tích.

Trí tuệ doanh nghiệp (BI) - phân tích kinh doanh; quy trình, công nghệ và công cụ chuyển đổi dữ liệu thành thông tin có ý nghĩa cho việc ra quyết định.

Mô hình thông tin xây dựng (BIM) - quy trình tạo ra và quản lý các đại diện kỹ thuật số về các đặc điểm vật lý và chức năng của các đối tượng xây dựng, bao gồm không chỉ mô hình 3D mà còn thông tin về đặc điểm, vật liệu, thời gian và chi phí.

BlackBox/WhiteBox - các phương pháp hiểu biết hệ thống: trong trường hợp đầu tiên, logic bên trong bị ẩn đi, chỉ thấy đầu vào và đầu ra; trong trường hợp thứ hai - quy trình xử lý là minh bạch và có thể phân tích.

Hộp giới hạn (Bounding Box) - cấu trúc hình học mô tả biên giới của đối tượng trong không gian ba chiều thông qua tọa độ tối thiểu và tối đa trên các trục X, Y và Z, tạo ra "hộp" xung quanh đối tượng.

BREP (Biểu diễn Biên giới) - biểu diễn hình học của các đối tượng, xác định chúng thông qua các biên giới bề mặt.

CAD (Thiết kế hỗ trợ máy tính) - hệ thống thiết kế tự động, được sử dụng để tạo, chỉnh sửa và phân tích các bản vẽ chính xác và mô hình 3D trong kiến trúc, xây dựng, kỹ thuật cơ khí và các ngành khác.

CAFM (Quản lý cơ sở hỗ trợ máy tính) - phần mềm quản lý bất động sản và hạ tầng, bao gồm lập kế hoạch không gian, quản lý tài sản, bảo trì kỹ thuật và giám sát chi phí.

CDE (Môi trường Dữ liệu Chung) - không gian kỹ thuật số trung tâm để quản lý, lưu trữ, trao đổi và hợp tác với thông tin dự án ở tất cả các giai đoạn vòng đời của đối tượng.

Trung tâm xuất sắc (CoE) - cấu trúc chuyên biệt trong tổ chức, chịu trách nhiệm phát triển một lĩnh vực kiến thức cụ thể, phát triển tiêu chuẩn và thực tiễn tốt nhất, đào tạo nhân viên và hỗ trợ triển khai đổi mới.

CoClass - hệ thống phân loại hiện đại cho các yếu tố xây dựng thế hệ thứ ba.

Mô hình dữ liệu khái niệm - đại diện cấp cao về các thực thể chính và mối quan hệ của chúng mà không có chi tiết về thuộc tính, được sử dụng trong các giai đoạn đầu tiên của thiết kế cơ sở dữ liệu.

CRM (Quản lý Quan hệ Khách hàng) - hệ thống quản lý tương tác với khách hàng, được sử dụng để tự động hóa các quy trình bán hàng và dịch vụ.

DAG (Đồ thị Acyclic Hướng) - đồ thị có hướng không chu trình, được sử dụng trong các hệ thống điều phối dữ liệu (Airflow, NiFi) để xác định thứ tự và các mối quan hệ phụ thuộc của các nhiệm vụ.

Dash - khung Python để tạo ra các hình ảnh web tương tác cho dữ liệu.

Bảng điều khiển (Dashboard) - bảng thông tin hiển thị trực quan các chỉ số hiệu suất chính và các chỉ số theo thời gian thực.

Phương pháp tiếp cận Dữ liệu-Centric - phương pháp đặt dữ liệu lên hàng đầu, không phải ứng dụng hoặc mã phần mềm, làm cho dữ liệu trở thành tài sản trung tâm của tổ chức.

Quản trị Dữ liệu - tập hợp các thực tiễn, quy trình và chính sách đảm bảo việc sử dụng dữ liệu đầy đủ và hiệu quả trong tổ chức, bao gồm kiểm soát quyền truy cập, chất lượng và an ninh.

Data Lake — kho lưu trữ được thiết kế để lưu trữ khối lượng lớn dữ liệu thô ở định dạng gốc của chúng cho đến khi được sử dụng.

Data Lakehouse — cách tiếp cận kiến trúc kết hợp sự linh hoạt và khả năng mở rộng của hồ dữ liệu (Data Lake) với tính quản lý và hiệu suất của kho dữ liệu (DWH).

Data-Driven Construction — đây là cách tiếp cận chiến lược, trong đó mỗi giai đoạn của vòng đời dự án — từ thiết kế đến vận hành — được hỗ trợ bởi các hệ thống tự động hóa, liên kết với nhau. Cách tiếp cận này đảm bảo việc học tập liên tục dựa trên dữ liệu thực tế, giảm thiểu sự không chắc chắn và giúp các công ty đạt được vị thế lãnh đạo bền vững trong ngành.

Data-Driven integrator — công ty chuyên về việc kết hợp dữ liệu từ nhiều nguồn khác nhau và phân tích chúng để đưa ra quyết định quản lý.

Data-Driven approach (Cách tiếp cận dựa trên dữ liệu) — phương pháp mà trong đó dữ liệu được coi là tài sản chiến lược, và các quyết định được đưa ra dựa trên phân tích khách quan của thông tin, không phải ý kiến chủ quan.

Data Minimalism — cách tiếp cận nhằm giảm thiểu dữ liệu xuống còn những phần có giá trị và ý nghĩa nhất, cho phép đơn giản hóa việc xử lý và phân tích thông tin.

Data Swamp — một tập hợp dữ liệu không được tổ chức, phát sinh từ việc thu thập và lưu trữ thông tin không kiểm soát.

DataOps — phương pháp luận kết hợp các nguyên tắc DevOps, dữ liệu và phân tích, tập trung vào việc cải thiện sự hợp tác, tích hợp và tự động hóa quy trình dữ liệu.

Số hóa thông tin — quá trình chuyển đổi tất cả các khía cạnh của hoạt động xây dựng sang dạng kỹ thuật số, phù hợp cho việc phân tích, giải thích và tự động hóa.

DataFrame (Khung dữ liệu) — cấu trúc dữ liệu dạng bảng hai chiều trong thư viện Pandas, trong đó các hàng đại diện cho các bản ghi hoặc đối tượng riêng lẻ, còn các cột — là các đặc điểm hoặc thuộc tính của chúng.

Descriptive Analytics (Phân tích mô tả) — phân tích dữ liệu lịch sử để hiểu điều gì đã xảy ra trong quá khứ.

Diagnostic Analytics (Phân tích chẩn đoán) — phân tích dữ liệu để xác định nguyên nhân tại sao một điều gì đó đã xảy ra.

Biểu đồ Gantt — công cụ lập kế hoạch dự án, thể hiện các nhiệm vụ dưới dạng các thanh ngang trên một mốc thời gian, cho phép hiển thị trực quan thứ tự và thời gian thực hiện công việc.

DWH (Data Warehouse) — hệ thống lưu trữ dữ liệu tập trung, thu thập thông tin từ nhiều nguồn, cấu trúc nó và làm cho nó dễ dàng truy cập cho phân tích và báo cáo.

ESG (Môi trường, Xã hội, Quản trị) — một tập hợp các tiêu chí để đánh giá tác động về môi trường, xã hội và quản trị của một công ty hoặc dự án.

ELT (Trích xuất, Tải lên, Chuyển đổi) — quy trình mà dữ liệu được trích xuất từ các nguồn và tải lên kho lưu trữ, sau đó được chuyển đổi cho các mục đích phân tích.

ETL (Trích xuất, Chuyển đổi, Tải lên) — quy trình trích xuất dữ liệu từ nhiều nguồn khác nhau, chuyển đổi chúng sang định dạng cần thiết và tải lên kho lưu trữ mục tiêu để phân tích.

ER-diagram (Sơ đồ thực thể-mối quan hệ) — sơ đồ hình ảnh hiển thị các thực thể, thuộc tính của chúng và mối quan hệ giữa chúng, được sử dụng trong việc mô hình hóa dữ liệu.

ERP (Lập kế hoạch nguồn lực doanh nghiệp) — hệ thống lập kế hoạch nguồn lực doanh nghiệp đa mô-đun, được sử dụng để quản lý và tối ưu hóa các khía cạnh khác nhau của quy trình xây dựng.

Đặc điểm (Features) - trong học máy, các biến độc lập hoặc thuộc tính được sử dụng làm dữ liệu đầu vào cho mô hình.

Mô hình dữ liệu vật lý - đại diện chi tiết cấu trúc của cơ sở dữ liệu, bao gồm bảng, cột, kiểu dữ liệu, khóa và chỉ mục, được tối ưu hóa cho hệ quản trị cơ sở dữ liệu cụ thể.

FPDF - thư viện Python để tạo tài liệu PDF.

Nhân hình học - thành phần phần mềm cung cấp các thuật toán cơ bản để tạo, chỉnh sửa và phân tích các đối tượng hình học trong CAD, BIM và các ứng dụng kỹ thuật khác.

HiPPO (Ý kiến của Người được trả lương cao nhất) - phương pháp ra quyết định dựa trên ý kiến của người có mức lương cao nhất trong tổ chức, thay vì dựa trên dữ liệu khách quan.

IDE (Môi trường phát triển tích hợp) - công cụ toàn diện để viết, kiểm tra và gỡ lỗi mã (ví dụ: PyCharm, VS Code, Jupyter Notebook).

IDS (Chỉ dẫn chuyển giao thông tin) - đặc tả chuyển giao thông tin, định nghĩa yêu cầu về dữ liệu ở các giai đoạn khác nhau của dự án.

IFC (Các lớp nền công nghiệp) - định dạng trao đổi dữ liệu BIM, đảm bảo tính tương thích giữa các giải pháp phần mềm khác nhau.

Industry 5.0 - khái niệm phát triển công nghiệp, kết hợp các khả năng số hóa, tự động hóa và trí tuệ nhân tạo với tiềm năng con người và tính bền vững môi trường.

Tích hợp dữ liệu - quá trình kết hợp dữ liệu từ nhiều nguồn khác nhau thành một hệ thống hoàn chỉnh, để bảo đảm một cái nhìn thống nhất về thông tin.

Silos thông tin - các hệ thống lưu trữ dữ liệu bị cô lập, không chia sẻ thông tin với các hệ thống khác, tạo ra rào cản cho việc sử dụng dữ liệu hiệu quả.

IoT (Internet of Things) - khái niệm kết nối các đối tượng vật lý với internet để thu thập, xử lý và truyền dữ liệu.

k-NN (k-Láng giềng gần nhất) - thuật toán học máy, phân loại các đối tượng dựa trên sự tương đồng với các láng giềng gần nhất trong tập huấn luyện.

Kaggle - nền tảng cho phân tích dữ liệu và các cuộc thi học máy.

Tính toán - việc tính toán chi phí xây dựng hoặc quy trình theo một đơn vị đo lường nhất định (ví dụ: cho 1 m^2 tường thạch cao, 1 m^3 bê tông).

KPI (Chỉ số hiệu suất chính) - chỉ số đo lường hiệu suất định lượng, được sử dụng để đánh giá mức độ thành công của hoạt động của công ty hoặc một dự án cụ thể.

Nhãn (Labels) - trong học máy, các biến mục tiêu hoặc thuộc tính mà mô hình phải dự đoán.

Thuật toán học (Learning Algorithm) - quá trình tìm kiếm giả thuyết tốt nhất trong mô hình, phù hợp với hàm mục tiêu, sử dụng bộ dữ liệu huấn luyện.

Hồi quy tuyến tính (Linear Regression) - phương pháp thống kê mô hình hóa mối quan hệ giữa biến phụ thuộc và một hoặc nhiều biến độc lập.

LLM (Mô hình ngôn ngữ lớn) - mô hình ngôn ngữ lớn, trí tuệ nhân tạo được đào tạo để hiểu và sinh ra văn bản dựa trên khối lượng dữ liệu khổng lồ, có khả năng phân tích ngữ cảnh và viết mã chương trình.

LOD (Cấp độ chi tiết/phát triển) - cấp độ chi tiết của mô hình, xác định mức độ chính xác hình học và thông tin đi kèm.

Mô hình dữ liệu logic — mô tả chi tiết về thực thể, thuộc tính, khoá và mối quan hệ, phản ánh thông tin và quy tắc kinh doanh, là giai đoạn trung gian giữa mô hình khái niệm và mô hình vật lý.

Machine Learning (Học máy) — một lớp các phương pháp trí tuệ nhân tạo cho phép hệ thống máy tính học hỏi và đưa ra dự đoán dựa trên dữ liệu mà không cần lập trình rõ ràng.

Masterformat — hệ thống phân loại thế hệ đầu tiên được sử dụng để cấu trúc các thông số xây dựng theo các phần và ngành.

MEP (Cơ khí, Điện, Nước) — các hệ thống kỹ thuật của tòa nhà bao gồm các thành phần cơ khí, điện và cấp thoát nước.

Mesh — đại diện lưới của các đối tượng 3D, bao gồm các đỉnh, cạnh và mặt.

Model (Mô hình) — trong học máy, là tập hợp các giả thuyết khác nhau, một trong số đó gần nhất với hàm mục tiêu cần dự đoán hoặc xấp xỉ.

Mô hình hóa dữ liệu — quá trình tạo ra biểu diễn cấu trúc của dữ liệu và mối quan hệ của chúng để triển khai trong các hệ thống thông tin, bao gồm các cấp độ khái niệm, logic và vật lý.

n8n — công cụ mã nguồn mở để tự động hóa quy trình làm việc và tích hợp ứng dụng thông qua phương pháp thấp mã (low-code).

Chuẩn hóa — trong học máy, quá trình đưa các dữ liệu số khác nhau về một thang đo thống nhất để dễ dàng xử lý và phân tích.

Kỹ thuật đảo ngược — quá trình nghiên cứu cấu trúc, chức năng và công nghệ sản xuất của một đối tượng thông qua việc phân tích cấu trúc, chức năng và hoạt động của nó. Trong bối cảnh dữ liệu — việc trích xuất thông tin từ các định dạng độc quyền để sử dụng trong các hệ thống mở.

OCR (Nhận dạng ký tự quang học) — công nghệ nhận dạng ký tự quang học cho phép chuyển đổi hình ảnh văn bản (tài liệu quét, ảnh) thành định dạng văn bản mà máy có thể đọc được.

OmniClass — tiêu chuẩn phân loại quốc tế thế hệ thứ hai để quản lý thông tin về các đối tượng xây dựng.

Ontology (Nhận thức luận) — hệ thống các mối quan hệ khái niệm, hình thức hóa một lĩnh vực kiến thức nhất định.

Mã nguồn mở — mô hình phát triển và phân phối phần mềm với mã nguồn mở, miễn phí cho việc sử dụng, nghiên cứu và sửa đổi.

Open BIM — khái niệm BIM mở, yêu cầu sử dụng các tiêu chuẩn và định dạng mở để trao đổi dữ liệu giữa các giải pháp phần mềm khác nhau.

Tiêu chuẩn mở — các thông số công khai có sẵn để đạt được một nhiệm vụ cụ thể, cho phép các hệ thống khác nhau tương tác và trao đổi dữ liệu.

Pandas — thư viện Python mã nguồn mở dành cho xử lý và phân tích dữ liệu, cung cấp các cấu trúc dữ liệu DataFrame và Series để làm việc hiệu quả với thông tin bảng.

Paradigm dữ liệu mở — cách tiếp cận xử lý dữ liệu, trong đó thông tin trở nên miễn phí để sử dụng, tái sử dụng và phát tán bởi bất kỳ ai.

Phương pháp tham số — phương pháp đánh giá dự án xây dựng, sử dụng các mô hình thống kê để đánh giá chi phí dựa trên các tham số của dự án.

PIMS (Mô hình thông tin dự án) — hệ thống kỹ thuật số dành cho tổ chức, lưu trữ và trao đổi toàn bộ thông tin dự án.

Pipeline — chuỗi quy trình xử lý dữ liệu, từ việc trích xuất và chuyển đổi đến phân tích và trực quan hóa.

PMIS (Hệ thống Quản lý Thông tin Dự án) — hệ thống quản lý dự án, nhằm kiểm soát chi tiết việc thực hiện các nhiệm vụ ở cấp độ từng công trình xây dựng.

Phân tích Dự đoán (Predictive Analytics) — lĩnh vực phân tích sử dụng các phương pháp thống kê và machine learning để dự đoán kết quả trong tương lai dựa trên dữ liệu lịch sử.

Phân tích Tư vấn (Prescriptive Analytics) — lĩnh vực phân tích không chỉ dự đoán các kết quả trong tương lai mà còn đề xuất các hành động tối ưu để đạt được kết quả mong muốn.

Định dạng độc quyền — các định dạng dữ liệu khép kín, được kiểm soát bởi một công ty cụ thể, hạn chế khả năng trao đổi thông tin và gia tăng sự phụ thuộc vào phần mềm cụ thể.

QTO (Lấy số lượng) — quy trình trích xuất đặc điểm định lượng của các yếu tố từ tài liệu thiết kế để tính toán khối lượng vật liệu cần thiết cho việc thực hiện dự án.

Hệ thống Quản lý Chất lượng — hệ thống quản lý chất lượng đảm bảo rằng các quy trình và kết quả đáp ứng các yêu cầu đã được thiết lập.

RAG (Generation Tăng cường Truy xuất) — phương pháp kết hợp khả năng sinh ra của các mô hình ngôn ngữ với việc trích xuất thông tin liên quan từ các cơ sở dữ liệu doanh nghiệp, nâng cao tính chính xác và chất lượng của các câu trả lời.

RDBMS (Hệ thống Quản lý Cơ sở Dữ liệu Quan hệ) — hệ thống quản lý cơ sở dữ liệu quan hệ, tổ chức thông tin dưới dạng các bảng liên kết với nhau.

RegEx (Biểu thức Chính quy) — ngôn ngữ chính thức để tìm kiếm và xử lý chuỗi, cho phép xác định các mẫu để kiểm tra dữ liệu văn bản theo các tiêu chí nhất định.

Hồi quy (Regression) — phương pháp phân tích thống kê về sự phụ thuộc giữa các biến.

Tính toán CO_2 — phương pháp đánh giá lượng khí thải carbon dioxide liên quan đến sản xuất và sử dụng vật liệu và quy trình xây dựng.

Phương pháp Tài nguyên — phương pháp lập dự toán dựa trên phân tích chi tiết tất cả các tài nguyên cần thiết (vật liệu, lao động, thiết bị) để thực hiện công việc xây dựng.

RFID (Nhận dạng Tần số Radio) — công nghệ nhận dạng tự động các đối tượng bằng sóng radio, được sử dụng để theo dõi vật liệu, thiết bị và nhân sự.

ROI (Lợi tức đầu tư) — chỉ số phản ánh tỷ lệ giữa lợi nhuận và tài chính đã đầu tư, được sử dụng để đánh giá hiệu quả của các khoản đầu tư.

SaaS (Phần mềm dưới dạng Dịch vụ) — mô hình cung cấp phần mềm như một dịch vụ, trong đó các ứng dụng được lưu trữ bởi nhà cung cấp và có sẵn cho người dùng qua internet.

SCM (Quản lý Chuỗi Cung ứng) — quản lý chuỗi cung ứng, bao gồm việc phối hợp và tối ưu hóa tất cả các quy trình từ mua sắm vật liệu đến giao hàng sản phẩm hoàn chỉnh.

Silos Dữ liệu — các kho lưu trữ thông tin tách biệt trong tổ chức, không được tích hợp với các hệ thống khác, gây khó khăn trong việc trao đổi dữ liệu và giảm hiệu quả.

SQL (Ngôn ngữ Truy vấn Cấu trúc) — ngôn ngữ truy vấn cấu trúc được sử dụng để làm việc với các cơ sở dữ liệu quan hệ.

SQLite — hệ quản trị cơ sở dữ liệu nhúng, nhẹ, đa nền tảng, không cần máy chủ riêng và hỗ trợ các chức năng cơ bản của SQL, được sử dụng phổ biến trong các ứng dụng di động và hệ thống nhúng.

Dữ liệu có cấu trúc — thông tin được tổ chức theo một định dạng nhất định với cấu trúc rõ ràng, ví dụ như trong các cơ sở dữ liệu quan hệ hoặc bảng.

Dữ liệu yếu cấu trúc là thông tin có tổ chức một phần và cấu trúc linh hoạt, chẳng hạn như JSON hoặc XML, trong đó các phần tử khác nhau có thể chứa các tập hợp thuộc tính khác nhau.

Thực thể (tiếng Anh: entity) là một đối tượng cụ thể hoặc trừu tượng của thế giới thực, có thể được phân biệt rõ ràng, mô tả và thể hiện dưới dạng dữ liệu.

Học Có Giám Sát (Supervised Learning) là loại hình học máy, trong đó thuật toán được đào tạo trên dữ liệu được gán nhãn, nơi mà đối với mỗi ví dụ, kết quả mong muốn đã được biết đến.

Phân loại - một hệ thống phân loại phân tầng, được sử dụng để phân phối hệ thống các yếu tố theo danh mục dựa trên các đặc điểm chung.

Bộ dữ liệu Titanic - một bộ dữ liệu phổ biến cho việc đào tạo và kiểm tra các mô hình học máy.

Đào tạo (Training) - quá trình mà thuật toán học máy phân tích dữ liệu để phát hiện những quy luật và xây dựng mô hình.

Chuyển giao học — phương pháp học máy, trong đó mô hình đã được đào tạo cho một nhiệm vụ được sử dụng như một điểm khởi đầu cho nhiệm vụ khác.

Biến đổi dữ liệu (Transformation) - quá trình thay đổi định dạng, cấu trúc hoặc nội dung của dữ liệu để sử dụng sau này.

Các yêu cầu về dữ liệu là các tiêu chí được hình thức hóa, xác định cấu trúc, định dạng, độ đầy đủ và chất lượng thông tin cần thiết để hỗ trợ các quy trình kinh doanh.

Uber hóa ngành xây dựng — quá trình chuyển đổi các mô hình kinh doanh truyền thống trong xây dựng dưới ảnh hưởng của các nền tảng số, cho phép tương tác trực tiếp giữa khách hàng và người thực hiện mà không cần trung gian.

Uniclass — hệ thống phân loại các yếu tố xây dựng thế hệ thứ hai và thứ ba, được sử dụng rộng rãi tại Vương quốc Anh.

USD (Universal Scene Description) — định dạng dữ liệu được phát triển cho đồ họa máy tính, nhưng đã được ứng dụng trong các hệ thống kỹ thuật nhờ vào cấu trúc đơn giản và tính độc lập với các lõi hình học.

Xác thực dữ liệu là quá trình kiểm tra thông tin để đảm bảo tính chính xác, đầy đủ và nhất quán theo các tiêu chí và yêu cầu đã được thiết lập.

Cơ sở dữ liệu vector (Vector Database) - một loại cơ sở dữ liệu chuyên biệt, lưu trữ dữ liệu dưới dạng các vector đa chiều để tìm kiếm và so sánh đối tượng một cách hiệu quả về mặt ngữ nghĩa.

Biểu diễn véc tơ (embedding) là phương pháp chuyển đổi dữ liệu thành các véc tơ số đa chiều, cho phép các thuật toán máy tính xử lý và phân tích thông tin một cách hiệu quả.

VectorOps - phương pháp tập trung vào việc xử lý, lưu trữ và phân tích dữ liệu vectơ đa chiều, đặc biệt phù hợp trong các lĩnh vực như mô hình số (digital twins) và tìm kiếm ngữ nghĩa (semantic search).

Trực quan hóa - biểu diễn đồ họa dữ liệu để tăng cường hiệu quả trong việc tiếp nhận và phân tích thông tin.

Phân loại thuật ngữ được thực hiện theo tên gọi của chúng bằng tiếng Anh.

DANH MỤC TÀI LIỆU VÀ TÀI NGUYÊN TRỰC TUYẾN

[1] Gartner, "Dữ liệu Chỉ số Các chỉ số CNTT 2017: Danh mục Tài liệu và Chỉ số đã công bố," 12 tháng 12 năm 2016. [Trực tuyến]. Có sẵn: https://www.gartner.com/en/documents/3530919. [Ngày truy cập: 1 tháng 3 năm 2025].

[2] KPMG, "Những thách thức quen thuộc - những phương pháp mới. Khảo sát Xây dựng Toàn cầu 2023," 1 tháng 1 năm 2023. [Trên Internet]. Có sẵn: https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2023/06/familiar-challenges-new-solutions-1.pdf. [Ngày truy cập: 5 tháng 3 năm 2025].

[3] F. R. Barnard, A picture is worth a thousand words, 10 tháng 3 năm 1927. [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words. [Ngày truy cập: 15 tháng 3 năm 2025].

[4] M. Bastian, Giám đốc điều hành Microsoft Satya Nadella cho biết việc tự công nhận AGI là "hacking chuẩn mực phi lý", 21 tháng 2 năm 2025. [Trực tuyến]. Có sẵn: https://the-decoder.com/microsoft-ceo-satya-nadella-says-self-claiming-agi-is-nonsensical-benchmark-hacking/. [Ngày truy cập: 15 tháng 3 năm 2025].

[5] W. E. Forum, Diễn Đàn Hình Thành Tương Lai Ngành Xây Dựng – Một Cảnh Quan Đang Biến Đổi, 1 tháng 1 năm 2016. [Trên Mạng]. Có sẵn: https://www3.weforum.org/docs/WEF_Shaping_the_Future_of_Construction.pdf. [Ngày truy cập: 2 tháng 3 năm 2025].

[6] C. D. Gillespie, Gạch: sự phức tạp của Trái Đất trong kỷ nguyên gạch, 2024. [Trên Internet]. Có sẵn: https://ufl.pb.unizin.org/imos/chapter/clay/.

[7] Papyrus thế kỷ III trước Công nguyên. Ngôn ngữ – tiếng Hy Lạp, 2024. [Trên Internet]. Có sẵn: https://www.facebook.com/429710190886668/posts/595698270954525.

[8] Giám sát: tận dụng các công cụ hiện có, 1980. [Trên Internet]. Có sẵn: https://pubmed.ncbi.nlm.nih.gov/10246720/. [Ngày truy cập: 15 tháng 3 năm 2025].

[9] PWC, "Dữ liệu dựa vào những gì sinh viên cần để thành công trong một thế giới kinh doanh đang thay đổi nhanh chóng," 15 tháng 2 năm 2015. [Trên Internet]. Có sẵn: https://www.pwc.com/us/en/faculty-resource/assets/PwC-Data-driven-paper-Feb2015.pdf. [Ngày truy cập: 15 tháng 3 năm 2025].

[10] Skanska USA, Xu hướng thị trường xây dựng mùa thu, 2 tháng 11 năm 2023. [Trực tuyến]. Có sẵn: https://x.com/SkanskaUSA/status/1720167220817588714.

[11] Oxford Essential Quotations (4 ed.), Oxford University Press, 2016. [Trực tuyến]. Có sẵn: https://www.oxfordreference.com/display/10.1093/acref/9780191826719.001.0001/q-oro-ed4-

00006236. [Ngày truy cập: 1 tháng 3 năm 2025].

[12] Sondergaard về Phân tích Dữ liệu, [Trên Internet]. Có sẵn: https://www.causeweb.org/cause/resources/library/r2493. [Ngày truy cập: 15 Tháng 3 năm 2025].

[13] Cách mà sự quan tâm toàn cầu đến AI đang thúc đẩy thị trường quản lý dữ liệu, 28 tháng 5 năm 2024. [Trên Internet]. Có sẵn: https://iot-analytics.com/how-global-ai-interest-is-boosting-data-management-market/. [Ngày truy cập: 15 tháng 3 năm 2025].

[14] I. McCue, "Lịch sử ERP," 2024. [Trên Internet]. Có sẵn: https://www.netsuite.com/portal/resource/articles/erp/erp-history.shtml.

[15] erpscout, «Giá ERP: Hệ thống ERP có giá bao nhiêu?,» [Trên Internet]. Có sẵn: https://erpscout.de/en/erp-costs/. [Ngày truy cập: 15 Tháng 3 năm 2025].

[16] softwarepath, 1.384 dự án ERP cho chúng ta biết gì về việc chọn ERP (báo cáo ERP 2022), 18 tháng 1 năm 2022. [Trên Internet]. Có sẵn: https://softwarepath.com/guides/erp-report. [Ngày truy cập: 15 tháng 3 năm 2025].

[17] Deloitte, Quản lý Dữ liệu trong các Dự án Vốn Kỹ thuật số, 16 tháng 12 năm 2016. [Trên Internet]. Có sẵn: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Real%20Estate/us-engineering-construction-data-driven-management-digital-capital-projects.pdf. [Ngày truy cập: 1 tháng 3 năm 2025].

[18] Mckinsey, "Doanh nghiệp dựa trên dữ liệu năm 2025," 28 tháng 1 năm 2022. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025. [Ngày truy cập: 22 tháng 5 năm 2024].

[19] Wikipedia, «Định luật Moore,» [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/Moore%27s_law. [Ngày truy cập: 15 Tháng 3 năm 2025].

[20] Accenture, Xây Dựng Giá Trị Nhiều Hơn Với Các Dự Án Vốn, 1 Tháng 1 Năm 2020. [Trên Internet]. Có sẵn: https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-143/accenture-industryx-building-value-capital-projects-highres.pdf. [Ngày truy cập: 3 Tháng 3 Năm 2024].

[21] B. Marr, "Chúng ta tạo ra bao nhiêu dữ liệu mỗi ngày? Những số liệu gây sốc mà mọi người nên đọc," 2018. [Trên Internet]. Có sẵn: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read.

[22] Bao nhiêu dữ liệu được sản xuất mỗi ngày?, 2024. [Trên Internet]. Có sẵn: https://graduate.northeastern.edu/resources/how-much-data-produced-every-day/.

[23] T. Sullivan, AI và "dữ liệu toàn cầu": bao nhiêu thông tin sẽ có cho nhân loại vào năm 2025?, 2024. [Trên Internet]. Có sẵn: https://www.datauniverseevent.com/en-us/blog/general/AI-and-the-Global-Datasphere-How-Much-Information-Will-Humanity-Have-By-2025.html.

[24] Statista, Tổng số sách in được sản xuất ở các khu vực Tây Âu trong mỗi nửa thế kỷ từ năm 1454 đến 1800, [Trên Internet]. Có sẵn: https://www.statista.com/statistics/1396121/europe-book-production-half-century-region-historical/. [Ngày truy cập: 1 tháng 3 năm 2025].

[25] Ví dụ về định giá, 2024. [Trên Internet]. Có sẵn: https://cloud.google.com/storage/pricing-examples.

[26] M. Ashare, Doanh nghiệp thuê ngoài lưu trữ dữ liệu khi độ phức tạp gia tăng, 10 tháng 5 năm 2024. [Trên Internet]. Có sẵn: https://www.ciodive.com/news/enterprises-outsource-data-storage-complexity-rises/715854/. [Ngày truy cập: 15 tháng 3 năm 2025].

[27] JETSOFTPRO, "SaaS đã chết? Dự đoán gây sốc của CEO Microsoft được giải thích," 13 tháng 1 năm 2025. [Trên Internet]. Có sẵn: https://jetsoftpro.com/blog/saas-is-dead/.

[28] BG2 Pod, Satya Nadella | BG2 với Bill Gurley & Brad Gerstner, 12 tháng 12 năm 2024. [Trên Internet]. Có sẵn: https://www.youtube.com/watch?v=9NtsnzRFJ_o. [Ngày truy cập: 15 tháng 3 năm 2025].

[29] GoodReads, «Tim Berners-Lee,» [Trên Internet]. Có sẵn: https://www.goodreads.com/quotes/8644920-data-is-a-precious-thing-and-will-last-longer-than. [Ngày truy cập: 15 Tháng Ba 2025].

[30] KPMG, Cue Construction 4.0: Thời điểm quyết định, 1 tháng 1 năm 2023. [Trên Internet]. Có sẵn: https://kpmg.com/ca/en/home/insights/2023/05/cue-construction-make-or-break-time.html. [Ngày truy cập: 5 tháng 3 năm 2025].

[31] I. Daininger, B. Koch, R. Bauknecht và M. Langhans, Sử dụng mô hình số để giảm carbon hóa địa điểm sản xuất: Ví dụ về việc kết nối mô hình tòa nhà, mô hình sản xuất và mô hình năng lượng, 2024. [Trên Internet]. Có sẵn: https://www.researchgate.net/publication/374023998_Using_Digital_Models_to_Decarbonize_a_Production_Site_A_Case_Study_of_Connecting_the_Building_Model_Production_Model_and_Energy_Model.

[32] McKinsey, REINVENTING CONSTRUCTION: MỘT HƯỚNG ĐI ĐẾN NĂNG SUẤT CAO HƠN, 1 Tháng 2 năm 2017. [Trên Internet]. Có sẵn: https://www.mckinsey.com/~/media/mckinsey/business%20functions/operations/our%20insights/reinventing%20construction%20through%20a%20productivity%20revolution/mgi-reinventing-construction-a-route-to-higher-productivity-full-report.pdf.

[33] Tổ công tác xây dựng gửi Phó Thủ tướng, Rethinking Construction, 1 tháng 10 năm 2014. [Trên Internet]. Có sẵn: https://constructingexcellence.org.uk/wp-content/uploads/2014/10/rethinking_construction_report.pdf.

[34] Forbes, Không có ý kiến, bạn chỉ là một người khác có dữ liệu, 15 tháng 3 năm 2016. [Trên Internet]. Có sẵn: https://www.forbes.com/sites/silberzahnjones/2016/03/15/without-an-opinion-youre-just-another-person-with-data/. [Ngày truy cập: 15 tháng 3 năm 2025].

[35] Wikiquote, Charles Babbage, [Trên Internet]. Có sẵn: https://en.wikiquote.org/wiki/Charles_Babbage. [Ngày truy cập: 15 Tháng Ba 2025].

[36] SAP, Nghiên cứu mới phát hiện rằng gần một nửa số giám đốc tin tưởng AI hơn chính họ, 12 tháng 3 năm 2025. [Trên Internet]. Có sẵn: https://news.sap.com/2025/03/new-research-executive-trust-ai/. [Ngày truy cập: 15 tháng 3 năm 2025].

[37] Hiệp hội Xây dựng Canada và KPMG tại Canada, 2021, Xây dựng trong thế giới số, 1 tháng 5 năm 2021. [Trực tuyến]. Có sẵn: https://assets.kpmg.com/content/dam/kpmg/ca/pdf/2021/05/construction-in-the-digital-age-report-en.pdf. [Ngày truy cập: 5 tháng 3 năm 2025].

[38] ZCCS, "Giải mã Cách mạng Công nghiệp lần thứ năm," [Trên Internet]. Có sẵn: https://www.pwc.in/decoding-the-fifth-industrial-revolution.html. [Ngày truy cập: 15 Tháng Ba 2025].

[39] M. K, Quyền riêng tư và Vấn đề công cộng: Kinh tế toàn cầu của, Viện Kinh tế Quốc tế Peterson, 2012.

[40] F. N. a. Y. Z. Trường Kinh doanh Harvard: Manuel Hoffmann, Giá trị của Phần mềm Mã nguồn Mở, 24 tháng 1 năm 2024. [Trên Internet]. Có sẵn: https://www.hbs.edu/faculty/Pages/item.aspx?num=65230. [Ngày truy cập: 15 tháng 3 năm 2025].

[41] Trung tâm Phân tích Chi phí Hải quân Cơ quan Phân tích Chi phí Không quân, Tài liệu Hướng dẫn Ước lượng Chi phí Phát triển Phần mềm, 1 tháng 9 năm 2008. [Trên Internet]. Có sẵn: https://www.dau.edu/sites/default/files/Migrated/CopDocuments/SW%20Cost%20Est%20Manual%20Vol%20I%20rev%2010.pdf.

[42] McKinsey, "Cải thiện năng suất xây dựng," [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/operations/our-insights/improving-construction-productivity. [Ngày truy cập: 15 Tháng Ba 2025].

[43] A. G. a. C. Syverson, "Con đường kỳ lạ và khủng khiếp của năng suất trong lĩnh vực xây dựng của Hoa Kỳ," 19 tháng 1 năm 2023. [Trên Internet]. Có sẵn: https://bfi.uchicago.edu/insight/research-summary/the-strange-and-awful-path-of-productivity-in-the-us-construction-sector/. [Ngày truy cập: 1 tháng 3 năm 2025].

[44] McKinsey, "Đưa năng suất xây dựng vào thực tiễn không còn là tùy chọn," 9 tháng 8 năm 2024. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/operations/our-insights/delivering-on-construction-productivity-is-no-longer-optional. [Ngày truy cập: 5 tháng 3 năm 2025].

[45] ING Group, "Năng suất chậm chạp trong ngành xây dựng đang đẩy chi phí xây dựng lên cao," 12 tháng 12 năm 2022. [Trên Internet]. Có sẵn: https://think.ing.com/articles/lagging-productivity-drives-up-building-costs-in-many-eu-countries/. [Ngày truy cập: 15 tháng 3 năm 2025].

[46] M. Berman, Giám đốc điều hành Microsoft đưa ra dự đoán gây sốc: “Các tác nhân sẽ thay thế TẤT CẢ phần mềm”, 19 tháng 12 năm 2024. [Trên Internet]. Có sẵn:

https://www.youtube.com/watch?v=uGOLYz2pgr8. [Ngày truy cập: 15 tháng 3 năm 2025].

[47] Business Insider, Giám đốc điều hành của Anthropic cho biết rằng trong 3 đến 6 tháng tới, AI sẽ viết 90% mã mà các nhà phát triển phần mềm phụ trách, 15 tháng 3 năm 2025. [Trực tuyến]. Có sẵn: https://www.businessinsider.com/anthropic-ceo-ai-90-percent-code-3-to-6-months-2025-3. [Ngày truy cập: 30 tháng 3 năm 2025].

[48] Statista, So sánh độ phổ biến của các hệ thống quản lý cơ sở dữ liệu (DBMS) trên toàn cầu tính đến tháng 6 năm 2024, theo danh mục, tháng 6 năm 2024. [Trực tuyến]. Có sẵn: https://www.statista.com/statistics/1131595/worldwide-popularity-database-management-systems-category/. [Ngày truy cập: 15 tháng 3 năm 2025].

[49] DB-Engines, Xếp hạng DB-Engines, [Trực tuyến]. Có sẵn: https://db-engines.com/en/ranking. [Ngày truy cập: 15 tháng 3 năm 2025].

[50] Khảo sát Nhà phát triển Stack Overflow 2023, 2024. [Trực tuyến]. Có sẵn: https://survey.stackoverflow.co/2023/.

[51] SQL, 2024. [Trực tuyến]. Có sẵn: https://en.wikipedia.org/wiki/SQL.

[52] Dữ liệu có cấu trúc và không có cấu trúc: Sự khác biệt là gì?, 2024. [Trực tuyến]. Có sẵn: https://www.ibm.com/blog/structured-vs-unstructured-data/.

[53] DataDrivenConstruction, SO SÁNH ĐỊNH DẠNG DỮ LIỆU CHO CÁC DỰ ÁN XÂY DỰNG, 23 tháng 4 năm 2024. [Trực tuyến]. Có sẵn: https://datadrivenconstruction.io/wp-content/uploads/2024/10/COMPARISON-OF-DATA-FORMATS-FOR-CONSTRUCTION-PROJECTS-1.pdf.

[54] Tài liệu trắng về Mô hình thông tin xây dựng, 2003. [Trực tuyến]. Có sẵn: https://web.archive.org/web/20030711125527/http:/usa.autodesk.com/adsk/servlet/item?id=2255342&siteID=123112.

[55] A. Boyko, Cuộc chiến vận động hành lang và sự phát triển của BIM. Phần 5: BlackRock – bậc thầy của tất cả công nghệ. Cách các tập đoàn kiểm soát mã nguồn mở, 2024. [Trực tuyến]. Có sẵn: https://bigdataconstruction.com/autodesk-oracle-blackrock-open-source/.

[56] D. Ushakov, Mô hình trực tiếp - Ai và Tại sao cần nó? Một cái nhìn tổng quan về các công nghệ cạnh tranh, 14 tháng 11 năm 2011. [Trực tuyến]. Có sẵn: https://isicad.net/articles.php?article_num=14805. [Ngày truy cập: 02 tháng 2025].

[57] C. Eastman và A. Cthers, Eastman, Charles; Và Cthers, tháng 9 năm 1974. [Trực tuyến]. Có sẵn: https://files.eric.ed.gov/fulltext/ED113833.pdf. [Ngày truy cập: 15 tháng 3 năm 2025].

[58] D. Ushakov, Mô hình trực tiếp - Ai và Tại sao cần nó? Một cái nhìn tổng quan về các công nghệ cạnh tranh, 11 tháng 11 năm 2011. [Trực tuyến]. Có sẵn: https://isicad.net/articles.php?article_num=14805. [Ngày truy cập: 15 tháng 3 năm 2025].

[59] D. Weisberg, Lịch sử CAD, 12 tháng 12 năm 2022. [Trực tuyến]. Có sẵn: https://www.shapr3d.com/blog/history-of-cad?utm_campaign=cadhistorynet. [Ngày truy cập: 15 tháng 3 năm 2025].

[60] ADSK, Tài liệu trắng về Mô hình thông tin xây dựng, 2002. [Trực tuyến]. Có sẵn: https://web.archive.org/web/20060512180953/http:/images.adsk.com/apac_sapac_main/files/4525081_BIM_WP_Rev5.pdf#expand. [Ngày truy cập: 15 tháng 3 năm 2025].

[61] ADSK, Tài liệu trắng về Mô hình thông tin xây dựng trong thực tiễn, [Trực tuyến]. Có sẵn: https://web.archive.org/web/20060512181000/http:/images.adsk.com/apac_sapac_main/files/4525077_BIM_in_Practice.pdf. [Ngày truy cập: 15 tháng 3 năm 2025].

[62] A. Boyko, "Cuộc chiến vận động hành lang và sự phát triển của BIM. Phần 2: BIM mở so với BIM đóng. Châu Âu so với phần còn lại của thế giới," 2024. [Trên Internet]. Có sẵn: https://bigdataconstruction.com/lobbyist-wars-and-the-development-of-bim-part-2-open-bim-vs-closed-bim-revit-vs-archicad-and-europe-vs-the-rest-of-the-world/.

[63] A. Boyko, "Cuộc chiến vận động hành lang về dữ liệu trong xây dựng | Công nghệ phong kiến và lịch sử của BIM," 2024. [Trên Internet]. Có sẵn: https://youtu.be/S-TNdUgfHxk?si=evM_v28KQbGOG0k&t=1360.

[64] ADSK, Tài liệu trắng BIM, 2002. [Trên Internet]. Có sẵn: https://web.archive.org/web/20060512180953/http:/images.autodesk.com/apac_sapac_main/files/4525081_BIM_WP_Rev5.pdf#expand. [Ngày truy cập: 15 Tháng 3 năm 2025].

[65] ADSK, Tích hợp Thiết kế-Qua-Sản xuất: Lợi ích và Lý do, [Trên Internet]. Có sẵn: https://web.archive.org/web/20010615093351/http:/www3.adsk.com:80/adsk/files/734489_Benefits_of_MAI.pdf. [Ngày truy cập: 15 Tháng 3 Năm 2025].

[66] M. Sheklet, "Dữ liệu có cấu trúc và không có cấu trúc: Những khác biệt chính," 2024. [Trên Internet]. Có sẵn: https://www.datamation.com/big-data/structured-vs-unstructured-data/.

[67] K. Vullard, "Hiểu về sự gia tăng dữ liệu không có cấu trúc," 2024. [Trên Internet]. Có sẵn: https://automationhero.ai/blog/making-sense-of-the-rise-of-unstructured-data/.

[68] A. C. O. J. L. D. J. a. L. T. G. Michael P. Gallaher, Phân tích chi phí của sự không tương tác đầy đủ trong, 2004. [Trên Internet]. Có sẵn: https://nvlpubs.nist.gov/nistpubs/gcr/2004/nist.gcr.04-867.pdf. [Ngày truy cập: 02 2025].

[69] CrowdFlower, Báo cáo Khoa học Dữ liệu 2016, 2016. [Trên Internet]. Có sẵn: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf. [Ngày truy cập: 15 Tháng Ba 2025].

[70] Analyticsindiamag, 6 Nhiệm Vụ Tốn Thời Gian Nhất Cho Các Nhà Khoa Học Dữ Liệu, 15 Tháng 5 2019. [Trên Internet]. Có sẵn: https://analyticsindiamag.com/ai-trends/6-tasks-data-scientists-

spend-the-most-time-doing/.

[71] BizReport, Báo cáo: Các nhà khoa học dữ liệu dành phần lớn thời gian để làm sạch dữ liệu, 06 tháng 7 năm 2015. [Trên Internet]. Có sẵn: https://web.archive.org/web/20200824174530/http://www.bizreport.com/2015/07/report-data-scientists-spend-bulk-of-time-cleaning-up.html. [Ngày truy cập: 5 tháng 3 năm 2025].

[72] S. Hawking, "Chuỗi AMA Khoa Học: Các câu trả lời của Stephen Hawking trong AMA," 27 tháng 7 năm 2015. [Trên Internet]. Có sẵn: https://www.reddit.com/r/science/comments/3nyn5i/science_ama_series_stephen_hawking_ama_answers/. [Ngày truy cập: 15 tháng 3 năm 2025].

[73] B. Cyphers và K. Doctorow, "Quyền riêng tư không độc quyền: Bảo vệ dữ liệu và khả năng tương thích," 2024. [Trên Internet]. Có sẵn: https://www.eff.org/wp/interoperability-and-privacy.

[74] Viện Toàn cầu McKinsey, Open data: Khai thác đổi mới và hiệu suất với thông tin linh hoạt, 1 tháng 10 năm 2013. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information. [Ngày truy cập: 15 tháng 3 năm 2025].

[75] A. Boyko, Cuộc chiến vì dữ liệu mở trong ngành xây dựng. Lịch sử của AUTOLISP, intelliCAD, openDWG, ODA và openCASCADE, 15 05 2024. [Trên Internet]. Có sẵn: https://boikoartem.medium.com/the-struggle-for-open-data-in-the-construction-industry-2b97200e6393. [Ngày truy cập: 16 02 2025].

[76] Wikipedia, «Microsoft và mã nguồn mở,» [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/Microsoft_and_open_source. [Ngày truy cập: 15 Tháng 3 năm 2025].

[77] Thời gian, «Khoảng cách giữa các mô hình AI mở và đóng có thể đang thu hẹp. Đây là lý do tại sao điều đó quan trọng,» 5 tháng 11 năm 2024. [Trên Internet]. Có sẵn: https://time.com/7171962/open-closed-ai-models-epoch/. [Ngày truy cập: 15 tháng 3 năm 2025].

[78] The Verge, «Hơn một phần tư mã mới tại Google được tạo ra bởi AI,» 29 tháng 10 năm 2024. [Trên Internet]. Có sẵn: https://www.theverge.com/2024/10/29/24282757/google-new-code-generated-ai-q3-2024. [Ngày truy cập: 15 tháng 3 năm 2025].

[79] McKinsey Digital, "Lý do kinh doanh để sử dụng GPU nhằm tăng tốc xử lý phân tích," 15 tháng 12 năm 2020. [Trực tuyến]. Có sẵn: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/the-business-case-for-using-gpus-to-accelerate-analytics-processing. [Ngày truy cập: 15 tháng 3 năm 2025].

[80] PWC, Báo cáo Giám sát Mã nguồn Mở 2019, 2019. [Trên Internet]. Có sẵn: https://www.pwc.de/de/digitale-transformation/open-source-monitor-research-report-2019.pdf. [Ngày truy cập: 15 Tháng 3 năm 2025].

[81] Travers Smith, Tài liệu Mở: Phần mềm Mã Nguồn Mở, 2024. [Trên Internet]. Có sẵn:

https://www.traverssmith.com/knowledge/knowledge-container/the-open-secret-open-source-software/. [Ngày truy cập: 15 Tháng Ba 2025].

[82] Deloitte, Quy trình chuyển giao dữ liệu trong các chuyển đổi doanh nghiệp, 2021. [Trên Internet]. Có sẵn: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/finance/us-the-data-transfer-process-in-corporate-transformations.pdf. [Ngày truy cập: 15 Tháng 3 năm 2025].

[83] gov.uk, Dự án Phân tích Dữ liệu và AI trong việc thực hiện Chính phủ, 20 tháng 3 năm 2024. [Trên Internet]. Có sẵn: https://www.gov.uk/government/publications/data-analytics-and-ai-in-government-project-delivery/data-analytics-and-ai-in-government-project-delivery. [Ngày truy cập: 5 tháng 3 năm 2025].

[84] Nguồn gốc câu trích: Mọi thứ nên được làm đơn giản nhất có thể, nhưng không đơn giản hơn. 13 tháng 5 năm 2011. [Trên Internet]. Có sẵn: https://quoteinvestigator.com/2011/05/13/einstein-simple/. [Ngày truy cập: 15 tháng 3 năm 2025].

[85] Transformer (kiến trúc học sâu), [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture). [Ngày truy cập: 15 Tháng 3 năm 2025].

[86] Python Packages Download Stats, 2024. [Trên Internet]. Có sẵn: https://www.pepy.tech/projects/pandas.

[87] Interview Bit, «Top 10 Thư viện Python,» 2023. [Trên Internet]. Có sẵn: https://www.interviewbit.com/blog/python-libraries/#:~:text=With%20more%20than%20137%2C000%20libraries,data%20manipulation%2C%20and%20many%20more. [Ngày truy cập: 30 Tháng 3 2025].

[88] NVIDIA và HP Tăng Cường Khoa Học Dữ Liệu và AI Tạo Sinh Trên Các Trạm Làm Việc, 7 Tháng 3 Năm 2025. [Trên Internet]. Có sẵn: https://nvidianews.nvidia.com/news/nvidia-hp-supercharge-data-science-generative-ai-workstations. [Ngày truy cập: 15 Tháng 3 Năm 2025].

[89] R. Oraak, Cách xử lý DataFrame với hàng triệu dòng trong vòng vài giây, 2024. [Trên Internet]. Có sẵn: https://towardsdatascience.com/how-to-process-a-dataframe-with-millions-of-rows-in-seconds.

[90] Ç. Uslu, "Kaggle là gì?", 2024. [Trên Internet]. Có sẵn: https://www.datacamp.com/blog/what-is-kaggle.

[91] NVIDIA CEO Jensen Huang Keynote tại COMPUTEX 2024, 2 tháng 6 năm 2024. [Trên Internet]. Có sẵn: https://www.youtube.com/live/pKXDVsWZmUU?si=Z3Rj1Las8wiPII2w. [Ngày truy cập: 15 tháng 3 năm 2025].

[92] Các thành viên: Những người sáng lập và thành viên doanh nghiệp, 2024. [Trên Internet]. Có sẵn: https://www.opendesign.com/member-showcase.

[93] A. Boyko, "Thời đại thay đổi: IFC là một thứ của quá khứ hay tại sao ADSK và các nhà cung cấp CAD khác sẵn sàng từ bỏ IFC để chuyển sang USD trong 14 sự thật chính," 24 tháng 11 năm 2024. [Trên Internet]. Có sẵn: https://boikoartem.medium.com/the-age-of-change-ifc-is-a-thing-of-the-past-or-why-adsk-and-other-cad-vendors-are-willing-to-3f9a82ccd10a. [Ngày truy cập: 23 tháng 2 năm 2025].

[94] A. Boiko, "Thế giới hậu BIM. Chuyển đổi sang dữ liệu và quy trình và liệu ngành xây dựng có cần ngữ nghĩa, định dạng và khả năng tương tác hay không," 20 tháng 12 năm 2024. [Trên Internet]. Có sẵn: https://boikoartem.medium.com/the-post-bim-world-7e35b7271119. [Ngày truy cập: 23 tháng 2 năm 2025].

[95] N. I. o. Sức khỏe, Kế hoạch chiến lược của NIH cho Khoa học Dữ liệu, 2016. [Trên Internet]. Có sẵn: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf. [Ngày truy cập: 23 tháng 2 năm 2025].

[96] Harvard Business Review, Bad Data Costs the U.S. $3 Trillion Per Year, 22 tháng 9 năm 2016. [Trên Internet]. Có sẵn: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

[97] Delpha, Tác động của Chất lượng Dữ liệu, 1 tháng 1 năm 2025. [Trên Internet]. Có sẵn: https://delpha.io/impacts-of-data-quality/.

[98] W. B. D. Guide, Thiết kế cho khả năng bảo trì: Tầm quan trọng của các yếu tố vận hành và bảo trì trong giai đoạn thiết kế của các dự án xây dựng, [Trên Internet]. Có sẵn: https://www.wbdg.org/resources/design-for-maintainability. [Ngày truy cập: 15 Tháng Ba 2025].

[99] O. o. D. C. P. a. Giám sát, Hướng dẫn lập kế hoạch phòng ngừa và kiểm soát ăn mòn cho các hệ thống và thiết bị quân sự, Tháng 4 năm 2014. [Trên Internet]. Có sẵn: https://www.dau.edu/sites/default/files/Migrated/CopDocuments/CPC%20Planning%20Guidebook%204%20Feb%2014.pdf. [Ngày truy cập: 15 Tháng 3 năm 2025].

[100] Gartner, "Chất lượng Dữ liệu: Các Thực tiễn Tốt nhất cho Những Thông tin Chính xác," 1 tháng 1 năm 2025. [Trên Internet]. Có sẵn: https://www.gartner.com/en/data-analytics/topics/data-quality.

[101] Để Thiếu Một Đinh, [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/For_Want_of_a_Nail. [Ngày truy cập: 15 Tháng Ba 2025].

[102] Viện Toàn cầu McKinsey, "Dữ liệu mở: Khai thác đổi mới và hiệu suất với thông tin linh hoạt," Tháng 10 năm 2013. [Trên Internet]. Có sẵn: https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/open%20data%20unlocking%20innovation%20and%20performance%20with%20liquid%20information/mgi_open_data_fullreport_oct2013.pdf. [Ngày truy cập: 15 Tháng 3 năm 2025].

[103] EY, "Con đường đến trung hòa carbon," 10 tháng 3 năm 2023. [Trực tuyến]. Có sẵn: https://www.ey.com/ru_kz/services/consulting/the-path-to-carbon-neutrality. [Ngày truy cập: 15 tháng 3 năm 2025].

[104] PWC, «Nhận thức về ESG,» 1 tháng 7 năm 2024. [Trên Internet]. Có sẵn: https://www.pwc.com/kz/en/assets/esg-awareness/kz-esg-awareness-rus.pdf. [Ngày truy cập: 15 tháng 3 năm 2025].

[105] G. Hammond, Embodied Carbon - Danh mục Carbon và Năng lượng (ICE), 2024. [Trên Internet]. Có sẵn: https://greenbuildingencyclopaedia.uk/wp-content/uploads/2014/07/Full-BSRIA-ICE-guide.pdf.

[106] CO2_tính toán carbon được chứa, 2024. [Trên Internet]. Có sẵn: https://github.com/datadrivenconstruction/CO2_calculating-the-embodied-carbon.

[107] McKinsey, Tưởng Tượng Tương Lai Kỹ Thuật Số Của Ngành Xây Dựng, 24 Tháng 6 2016. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/operations/our-insights/imagining-constructions-digital-future. [Ngày truy cập: 25 Tháng 2 2025].

[108] Bund der Steuerzahler Deutschland e.V., Das Schwarzbuch, 10 tháng 10 năm 2024. [Trên Internet]. Có sẵn: https://steuerzahler.de/aktuelles/detail/das-schwarzbuch-202425/. [Ngày truy cập: 15 tháng 3 năm 2025].

[109] SAS, "Hồ dữ liệu và kho dữ liệu - hiểu sự khác biệt," [Trên Internet]. Có sẵn: https://www.sas.com/en_is/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html. [Ngày truy cập: 15 Tháng Ba 2025].

[110] ADSK, Mô hình thông tin xây dựng, 2002. [Trên Internet]. Có sẵn: https://www.laiserin.com/features/bim/autodesk_bim.pdf. [Ngày truy cập: 15 Tháng 3 năm 2025].

[111] A. Boyko, "Bản đồ lịch sử BIM," 2024. [Trên Internet]. Có sẵn: https://bigdataconstruction.com/history-of-bim/.

[112] A. S. Borkowski, Định nghĩa về BIM bởi các tổ chức và tiêu chuẩn, 27 tháng 12 năm 2023. [Trên Internet]. Có sẵn: https://encyclopedia.pub/entry/53149. [Ngày truy cập: 5 tháng 3 năm 2025].

[113] Nhà cung cấp CAD, Chương trình «OPEN BIM», 2012. [Trên Internet]. Có sẵn: https://web.archive.org/web/20140611075601/http://www.graphisoft.com/archicad/open_bim/. [Ngày truy cập: 30 Tháng Ba 2025].

[114] Wikipedia, «Các lớp cơ sở ngành», [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/Industry_Foundation_Classes. [Ngày truy cập: 15 tháng 3 năm 2025].

[115] Wikipedia, «IGES,» [Trên Internet]. Có sẵn: https://en.wikipedia.org/wiki/IGES. [Ngày truy cập: 30 Tháng 3 năm 2025].

[116] A. Boyko, "Lịch sử CAD (BIM)," 15 tháng 12 năm 2021. [Trên Internet]. Có sẵn: https://miro.com/app/board/o9J_laML2cs=/. [Ngày truy cập: 24 tháng 2 năm 2025].

[117] T. K. K. A. O. F. B. C. E. L. H. H. E. L. P. N. S. H. T. J. v. L. H. G. D. H. T. K. C. L. A. W. J. S. Francesca

Noardo, Nghiên cứu tham chiếu về hỗ trợ phần mềm IFC: tiêu chuẩn GeoBIM 2019 — Phần I, 8 tháng 1 năm 2021. [Trên Internet]. Có sẵn: https://arxiv.org/pdf/2007.10951. [Ngày truy cập: 5 tháng 3 năm 2025].

[118] I. Rogachyov, "Chúng ta hãy nói về BIM: Maxim Nechiporenko | Renga | IFC | BIM trong nước," 13 tháng 4 năm 2021. [Trên Internet]. Có sẵn: https://www.youtube.com/watch?t=3000&v=VO3Y9uuzF9M&feature=youtu.be. [Ngày truy cập: 5 tháng 3 năm 2025].

[119] D. Ares, RETS trong Bất động sản: Tại sao nó quan trọng cho hiệu quả và tăng trưởng, 17 tháng 12 năm 2024. [Trên Internet]. Có sẵn: https://www.realpha.com/blog/rets-importance-in-real-estate-explained. [Ngày truy cập: 5 tháng 3 năm 2025].

[120] Chi phí token Flex, 2024. [Trên Internet]. Có sẵn: https://www.adsk.com/buying/flex?term=1-YEAR&tab=flex.

[121] A. Boyko, "Quên đi BIM và dân chủ hóa quyền truy cập dữ liệu (17. Kolloquium Nhà đầu tư – Trường đại học – Ngành xây dựng)," 2024. [Trên Internet]. Có sẵn: https://www.bim.bayern.de/wp-content/uploads/2023/06/Kolloquium-17-TUM-Bauprozessmanagment-und-Bay-Bauindustrie.pdf.

[122] D. Hill, D. Foldesi, S. Ferrer, M. Friedman, E. Loch và F. Plaschke, "Giải quyết câu đố về năng suất của ngành xây dựng," 2015. [Trên Internet]. Có sẵn: https://www.bcg.com/publications/2015/engineered-products-project-business-solving-construction-industrys-productivity-puzzle.

[123] SCOPE – Môi trường dữ liệu dự án và mô hình hóa các sản phẩm xây dựng đa chức năng với trọng tâm là lớp vỏ công trình, 1 tháng 1 năm 2018. [Trên Internet]. Có sẵn: https://www.ise.fraunhofer.de/de/forschungsprojekte/scope.html. [Ngày truy cập: 2 tháng 3 năm 2025].

[124] Apple.com, Pixars, Adobe, Apple và NVIDIA thành lập Liên minh OpenUSD để thúc đẩy các tiêu chuẩn mở cho nội dung 3D, 1 tháng 8 năm 2023. [Trên Internet]. Có sẵn: https://www.apple.com/newsroom/2023/08/pixar-adobe-apple-adsk-and-nvidia-form-alliance-for-openusd/. [Ngày truy cập: 2 tháng 3 năm 2025].

[125] AECmag, Chiến lược dữ liệu chi tiết của ADSK, 25 tháng 7 năm 2024. [Trên Internet]. Có sẵn: https://aecmag.com/technology/autodesks-granular-data-strategy/. [Ngày truy cập: 15 tháng 3 năm 2025].

[126] A. Boyko, "Thời đại thay đổi: IFC là một thứ của quá khứ hay tại sao ADSK và các nhà cung cấp CAD khác sẵn sàng từ bỏ IFC để chuyển sang USD trong 14 sự thật chính," 24 tháng 11 năm 2024. [Trên Internet]. Có sẵn: https://boikoartem.medium.com/the-age-of-change-ifc-is-a-thing-of-the-past-or-why-adsk-and-other-cad-vendors-are-willing-to-3f9a82ccd10a. [Ngày truy cập: 23 tháng 2 năm 2025].

[127] A. Boyko, ENG BIM Cluster 2024 | Cuộc chiến về Dữ liệu và Ứng dụng của LLM và ChatGPT trong Xây dựng, 7 tháng 8 năm 2024. [Trên Internet]. Có sẵn: ENG BIM Cluster 2024 | Cuộc chiến về Dữ liệu và

Ứng dụng của LLM và ChatGPT trong Xây dựng. [Ngày truy cập: 15 tháng 3 năm 2025].

[128] Jeffrey Zeldman Trình bày, 6 tháng 5 năm 2008. [Trên Internet]. Có sẵn: https://zeldman.com/2008/05/06/content-precedes-design/. [Ngày truy cập: 15 tháng 3 năm 2025].

[129] A. Boiko, DWG Phân Tích với ChatGPT | Xây Dựng Dữ Liệu, 5 Tháng 3 2024. [Trên Internet]. Có sẵn: https://www.kaggle.com/code/artemboiko/dwg-analyse-with-chatgpt-datadrivenconstruction. [Ngày truy cập: 15 Tháng 3 2025].

[130] McKinsey, Hướng dẫn McKinsey về việc cạnh tranh vượt trội trong kỷ nguyên số và trí tuệ nhân tạo, 2023. [Trên Internet]. Có sẵn: https://www.mckinsey.com/featured-insights/mckinsey-on-books/rewired. [Ngày truy cập: 30 Tháng Ba 2025].

[131] Forbes, Kể chuyện dữ liệu: Kỹ năng khoa học dữ liệu thiết yếu mà mọi người cần, 31 tháng 3 năm 2016. [Trên Internet]. Có sẵn: https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/. [Ngày truy cập: 15 tháng 3 năm 2025].

[132] J. Bertin, Đồ họa và Xử lý Thông tin Đồ họa, 8 tháng 9 năm 2011. [Trên Internet]. Có sẵn: https://books.google.de/books/about/Graphics_and_Graphic_Information_Process.html?id=csqX_xnm4tcC&redir_esc=y. [Ngày truy cập: 15 tháng 3 năm 2025].

[133] [133]

[134] CauseWeb, Wells/Wilks về Tư duy Thống kê, [Trên Internet]. Có sẵn: https://www.causeweb.org/cause/resources/library/r1266. [Ngày truy cập: 15 tháng 3 năm 2025].

[135] [134]

[136] Ministrymagazine, Cách khoa học phát hiện ra Sáng tạo, tháng 1 năm 1986. [Trên Internet]. Có sẵn: https://www.ministrymagazine.org/archive/1986/01/how-science-discovered-creation. [Ngày truy cập: 15 tháng 3 năm 2025].

[137] [135]

[138] BCG, Chuyển đổi Dữ liệu: Tăng tốc quy mô ngay bây giờ, 23 tháng 5 năm 2017. [Trên Internet]. Có sẵn: https://www.bcg.com/publications/2017/digital-transformation-transformation-data-driven-transformation. [Ngày truy cập: 15 tháng 5 năm 2024].

[139] [136]

[140] Cách xây dựng kiến trúc dữ liệu để thúc đẩy đổi mới - hôm nay và ngày mai, 3 tháng 6 năm 2020. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/how-to-build-a-data-architecture-to-drive-innovation-today-and-tomorrow. [Ngày truy cập: 15 tháng 3 năm 2025].

[141] [137]

[142] Oxford, Woodrow Wilson 1856–1924, [Trên Internet]. Có sẵn: https://www.oxfordreference.com/display/10.1093/acref/9780191866692.001.0001/q-oro-ed6-00011630. [Ngày truy cập: 15 tháng 3 năm 2025].

[143] [138]

[144] Chuyển đổi, 2024. [Trên Internet]. Có sẵn: https://datadrivenconstruction.io/index.php/convertors/.

[145] [139]

[146] PWC, Kích thước giải thưởng Giá trị thực sự của AI cho doanh nghiệp của bạn và cách bạn có thể tận dụng?, 1 tháng 1 năm 2017. [Trên Internet]. Có sẵn: https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf. [Ngày truy cập: 18 tháng 2 năm 2025].

[147] [140]

[148] Đường ống trong xây dựng, 2024. [Trên Internet]. Có sẵn: https://datadrivenconstruction.io/index.php/pipeline-in-construction/.

[149] [141]

[150] Wikipedia, Apache NiFi, 1 tháng 1 năm 2025. [Trên Internet]. Có sẵn: https://de.wikipedia.org/wiki/Apache_NiFi. [Ngày truy cập: 5 tháng 3 năm 2025].

[151] [142]

[152] n8n, Trả lời tự động Gmail AI: Tạo bản nháp trả lời cho các email đến, 1 tháng 5 năm 2024. [Trên Internet]. Có sẵn: https://n8n.io/workflows/2271-gmail-ai-auto-responder-create-draft-replies-to-incoming-emails/. [Ngày truy cập: 15 tháng 3 năm 2025].

[153] [143]

[154] n8n, Tự động hóa giao dịch hàng ngày bất động sản với API Zillow, Google Sheets và Gmail, 1 tháng 3 năm 2025. [Trên Internet]. Có sẵn: https://n8n.io/workflows/3030-real-estate-daily-deals-automation-with-zillow-api-google-sheets-and-gmail/. [Ngày truy cập: 15 tháng 3 năm 2025].

[155] [144]

[156] B. T. O'Neill, Tỷ lệ thất bại cho các dự án phân tích, AI và dữ liệu lớn = 85% – thật đáng sợ!, 1 tháng 1 năm 2025. [Trên Internet]. Có sẵn: https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/.

[157] [145]

[158] J. Neyman, Về Hai Khía Cạnh Khác Nhau của Phương Pháp Đại diện: Phương Pháp Lấy Mẫu Tầng và Phương Pháp Lựa Chọn Có Mục Đích, Oxford University Press, 1934.

[159] [146]

[160] T. J. S. và J. S. Jesse Perla, Một Vấn Đề Gây Khó Khăn cho Milton Friedman, Kinh tế Định lượng với Julia, 1 tháng 1 năm 2025. [Trên Internet]. Có sẵn: https://julia.quantecon.org/dynamic_programming/wald_friedman.html. [Ngày truy cập: 1 tháng 5 năm 2024].

[161] [147]

[162] T. Landsall-Welfer, Dự đoán tâm trạng của quốc gia hiện nay, Significance, 2012.

[163] NVIDIA, NVIDIA Isaac Sim, [Trực tuyến]. Có sẵn: https://developer.nvidia.com/isaac/sim. [Ngày truy cập: 25 Tháng 3 2025].

[164] M. Quarterly, Tại sao các chiến lược số thất bại, 25 tháng 1 năm 2018. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-digital-strategies-fail. [Ngày truy cập: 15 tháng 3 năm 2025].

[165] M. J. Perry, My Favorite Milton Friedman Quotes, 17 tháng 11 năm 2006. [Trên Internet]. Có sẵn: https://www.aei.org/carpe-diem/my-favorite-milton-friedman-quotes/. [Ngày truy cập: 1 tháng 3 năm 2025].

[166] J. A. Wheeler, Thông tin, vật lý, lượng tử: tìm kiếm mối liên hệ, 1990.

[169] A. Boiko, "Cuộc chiến vận động hành lang và sự phát triển của BIM. Phần 5: BlackRock - bậc thầy của tất cả các công nghệ. Cách các tập đoàn kiểm soát mã nguồn mở," 2024. [Trên Internet]. Có sẵn: https://boikoartem.medium.com/lobbyist-wars-and-the-development-of-bim-d72ad0111a7d.

[170] T. Krijnen và J. Beetz, A SPARQL query engine for binary-formatted IFC building models, Advanced Engineering Informatics, 2024.

[171] Số lượng doanh nghiệp trong lĩnh vực xây dựng tại Vương quốc Anh vào năm 2021, theo kích thước doanh nghiệp, 2024. [Trực tuyến]. Có sẵn: https://www.statista.com/statistics/677151/uk-construction-businesses-by-size/.

[172] 5000 dự án IFC&RVT, 2024. [Trên Internet]. Có sẵn: https://www.kaggle.com/code/artemboiko/5000-projects-ifc-rvt-datadrivenconstruction-io.

[173] M. Popova, It from Bit: Nhà vật lý tiên phong John Archibald Wheeler về Thông tin, Bản chất của Thực tại, và Tại sao Chúng ta sống trong một Vũ trụ Tham gia, 2008. [Trên Internet]. Có sẵn:

https://www.themarginalian.org/2016/09/02/it-from-bit-wheeler/. [Ngày truy cập: Tháng Hai 2025].

[174] *Cuộc chiến vận động hành lang về dữ liệu trong xây dựng | Công nghệ phong kiến và lịch sử ẩn giấu của BIM. [Phim]. Đức: Artem Boyko, 2023.*

[175] A. Boyko, CHATGPT VỚI REVIT VÀ IFC | Tự động truy xuất tài liệu và dữ liệu từ các dự án, 16 Tháng 11 2023. [Trên Internet]. Có sẵn: https://www.youtube.com/watch?v=ASXolti_YPs&t. [Ngày truy cập: 2 Tháng 3 2025].

[176] M. & Công ty, Ba nhiệm vụ mới để nắm bắt giá trị đầy đủ của chuyển đổi số, 22 tháng 1 năm 2022. [Trên Internet]. Có sẵn: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/three-new-mandates-for-capturing-a-digital-transformations-full-value. [Ngày truy cập: 15 tháng 2 năm 2025].

[177] KPMG, "Xây dựng trong Thế giới Kỹ thuật số," 1 tháng 5 năm 2021. [Trên Internet]. Có sẵn: https://assets.kpmg.com/content/dam/kpmg/ca/pdf/2021/05/construction-in-the-digital-age-report-en.pdf. [Ngày truy cập: 5 tháng 4 năm 2024].

[178] LLP, KPMG, Cue Construction 4.0: Thời điểm quyết định, 17 tháng 3 năm 2023. [Trên Internet]. Có sẵn: https://kpmg.com/ca/en/home/insights/2023/05/cue-construction-make-or-break-time.html. [Ngày truy cập: 15 tháng 2 năm 2025].

[179] O. Business, Satya Nadella tiết lộ 'Cách mà các tác nhân AI sẽ làm gián đoạn các mô hình SaaS,' 10 tháng 1 năm 2025. [Trực tuyến]. Có sẵn: https://www.outlookbusiness.com/artificial-intelligence/microsoft-ceo-satya-nadella-reveals-how-ai-agents-will-disrupt-saas-models. [Ngày truy cập: 15 tháng 3 năm 2025].

[180] Forbes, "Dọn dẹp Dữ liệu Lớn: Nhiệm vụ Khoa học Dữ liệu Tốn Thời gian Nhất, Ít Thú vị Nhất, Theo Khảo sát," 23 tháng 3 năm 2016. [Trên Internet]. Có sẵn: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. [Ngày truy cập: 15 tháng 3 năm 2025].

[181] Bộ Ngoại giao, Bộ Các vấn đề Khối thịnh vượng chung và Phát triển Vương quốc Anh, Chiến lược phát triển kỹ thuật số 2024 đến 2030, 18 tháng 3 năm 2024. [Trực tuyến]. Có sẵn: https://www.gov.uk/government/publications/digital-development-strategy-2024-to-2030/digital-development-strategy-2024-to-2030. [Ngày truy cập: 15 tháng 3 năm 2025].

[182] Tầm nhìn và chiến lược trong ngành thiết kế xây dựng, 7 tháng 11 năm 2003. [Trên Internet]. Có sẵn: https://web.archive.org/web/20030711125527/http:/usa.adsk.com/adsk/servlet/item?id=2255342&siteID=123112. [Ngày truy cập: 5 tháng 3 năm 2025].

[183] M. Bocharov, "Mô hình thông tin," Tháng 3 năm 2025. [Trên Internet]. Có sẵn: https://www.litres.ru/book/mihail-evgenevich-bocharov/informacionnoe-modelirovanie-v-rossii-71780080/chitat-onlayn/?page=5. [Ngày truy cập: 15 Tháng 3 năm 2025].

[184] Thiết kế tích hợp qua sản xuất: Lợi ích và lý do, 2000. [Trên Internet]. Có sẵn:

https://web.archive.org/web/20010615093351/http:/www3.autodesk.com:80/adsk/files/734489_Benefits_of_MAI.pdf. [Ngày truy cập: 25 tháng 3 năm 2025].

[185] Nhà cung cấp CAD, Chương trình Open BIM là một chiến dịch tiếp thị, 12 tháng 3 năm 2012. [Trên Internet]. Có sẵn: https://web.archive.org/web/20120827193840/http:/www.graphisoft.com/openbim/. [Ngày truy cập: 30 tháng 3 năm 2025].

MỤC LỤC THEO CHỦ ĐỀ

3D, 8, 14, 71, 73, 84, 191, 210, 215, 232, 234, 263, 276, 277, 287, 298, 299, 302, 303, 306, 307, 337, 375, 393, 394, 448, 468, 480

4D, 84, 172, 196, 199, 210, 229, 234, 237, 287
Cách mạng công nghiệp 4.0, 43

5D, 84, 172, 196, 210, 237, 287, 492

6D, 172, 196, 229, 232, 233, 234, 235

7D, 232, 233, 234, 287

8D, 172, 196, 229, 232, 233, 234, 287

A

AI, 3, 50, 52, 100, 102, 103, 106, 107, 116, 395, 457, 459, 461, 477
AIA, 289
MỤC TIÊU, 3, 289
AMS, 14, 84, 151, 153
Apache Airflow, 311, 361, 362, 363, 364, 366, 367, 369, 371, 399, 494
Apache NiFi, 116, 311, 361, 367, 368, 371, 399, 482, 494
Apache ORC, 62, 330, 378
Apache Parquet, 62, 67, 380, 381, 494
API, 54, 94, 95, 97, 109, 137, 138, 154, 168, 218, 219, 256, 257, 260, 271, 273, 294, 295, 296, 297, 300, 308, 326, 329, 342, 366, 369, 370, 488

B

BDS, 258, 259
Dữ liệu lớn, 9, 67, 245, 383
BIM, 2, 3, 4, 6, 3, 14, 17, 24, 56, 58, 60, 63, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 84, 92, 137, 138, 139, 140, 141, 142, 144, 150, 154, 156, 166, 172, 183, 184, 186, 187, 190, 195, 196, 208, 210, 211, 213, 216, 217, 218, 219, 220, 221, 222, 227, 228, 237, 238, 239, 242, 243, 246, 250, 251, 252, 255, 256, 257, 258, 259, 260, 261, 262, 263, 266, 271, 272, 273, 275, 276, 277, 278, 279, 280, 282, 285, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 308, 309, 324, 328, 337, 351, 356, 358, 361, 389, 413, 416, 448, 462, 466, 475, 476, 492, 494, 497.
BlackBox, 240, 242, 243
BMS, 8
Bokeh, 320, 337
BOM, 76, 77, 79, 257, 263
Hộp bao, 234, 373, 392, 393, 394, 414, 415, 480
BREP, 142, 234, 263, 264, 276, 283, 284

C

CAD, 6, 14, 18, 24, 56, 57, 58, 63, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 84, 85, 95, 97, 111, 126, 137, 138, 139, 140, 141, 142, 144, 146, 147, 152, 153, 155, 156, 166, 172, 175, 183, 184, 186, 187, 190, 195, 196, 206, 208, 210, 211, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 224, 227, 228, 232, 234, 237, 238, 239, 242, 243, 251, 252, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 303, 308, 309, 324, 328, 337, 344, 351, 356, 358, 361, 367, 372, 376, 383, 389, 402, 403, 405, 413, 416, 431, 448, 462, 466, 475, 476, 480, 492, 494, 497.
CAE, 16, 283
CAFM, 14, 24, 62, 84, 151, 153, 172, 233, 278, 326, 356, 387, 458, 473
CAM, 16, 78
Chi phí đầu tư, 14, 82
CDE, 84, 175, 388, 389, 390
ChatGPT, 103, 104, 109, 110, 124, 129, 221, 303, 307, 343, 349, 421
Claude, 103, 104, 109, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 429, 443
CO2, 72
CO_2, 229, 234, 235, 236, 237, 238, 239, 243
COBie, 156, 289, 292
CoE, 56, 168, 169, 170, 477
Copilot, 114, 116, 459
CPIXML, 143, 272, 273, 276, 277, 279, 285, 296
CPM, 14, 17, 62, 166, 175, 233, 326, 473
CQMS, 14, 84, 177, 178, 462
CRM, 109, 369, 459
CRUD, 51, 459
CSG, 263
CSV, 61, 62, 88, 89, 120, 123, 128, 129, 130, 131, 135, 168, 268, 272, 280, 329, 333, 344, 354, 356, 373, 377, 378, 379, 380, 401, 407, 414, 474, 480

D

DAE, 276, 277, 278, 280, 281, 284, 285, 414
DAG, 362, 363, 365, 366
Dash, 320, 336, 337
Quản trị dữ liệu, 373, 395, 396, 398, 400, 401
Hồ dữ liệu, 214, 373, 376, 384, 385, 386, 387, 388, 389, 390, 400
Hồ dữ liệu nhà, 67, 373, 386, 387, 388
Tối giản dữ liệu, 373, 395, 396, 400, 401
Đầm dữ liệu, 373, 395, 397, 401
Kho dữ liệu, 382, 383, 400
Dữ liệu như một dịch vụ, 487
dựa trên dữ liệu, 50, 170, 460, 461, 484, 486
DataFrame, 67, 117, 121, 122, 123, 125, 129, 130, 131, 133, 134, 135, 137, 219, 220, 224, 237, 308, 328, 329, 330, 332, 333, 344, 345, 347, 348, 354, 365, 377, 381, 409, 414, 415, 475
DataOps, 170, 373, 398, 399, 400, 401, 475
DeepSeek, 103, 104, 107, 109, 110, 124, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 421, 429, 443, 494
DGN, 8, 140, 186, 227, 357
DWG, 8, 70, 71, 73, 97, 140, 186, 211, 227, 272, 287, 302, 303, 304, 307, 357, 376, 497
DWH, 67, 373, 376, 382, 383, 384, 386, 387, 388, 389, 390
DXF, 8, 73, 277

E

ECM, 58, 175
ECS, 142
EIR, 289
eLOD, 289
ELT, 384, 385
EPM, 14, 166, 198
ERP, 2, 11, 12, 14, 17, 18, 24, 25, 58, 62, 109, 153, 166, 172, 175, 196, 198, 210, 232, 239, 240, 241, 242, 243, 244, 245, 246, 247, 249, 272, 277, 278, 279, 282, 326, 351, 356, 361, 369, 387, 388, 389, 390, 458, 462, 473, 475, 487, 492
ESG, 196, 235, 236, 238
ETL, 6, 1.1-8, 32, 49, 81, 113, 116, 119, 128, 188, 193, 219, 291, 311, 312, 317, 323, 324, 325, 326, 327, 328, 330, 331, 333, 338, 339, 340, 343, 344, 348, 349, 350, 351, 353, 354, 356, 361, 362, 363, 364, 365, 367, 371, 372, 381, 382, 383, 384, 385, 399, 409, 430, 475, 481, 494
Excel, 57, 61, 62, 65, 66, 85, 88, 111, 120, 123, 125, 132, 154, 167, 187, 210, 223, 224, 226, 227, 228, 278, 291, 329, 333, 342, 351, 365, 376, 378, 459, 475
Trích xuất, 81, 128, 134, 193, 311, 323, 324, 326, 328, 330, 345, 361, 362, 363, 365, 383, 384, 450, 475

F

Lông vũ, 62, 123, 330, 378
FPDF, 339, 340, 341, 342, 343

G

GDPR, 109
GIS, 58
GLTF, 143, 278
Google Sheets, 368, 370
Grok, 103, 104, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 429, 443

H

HDF5, 62, 67, 123, 329, 330, 378, 379, 380
HiPPO, 29, 37, 95, 424, 477, 484, 490
HTML, 123, 340, 365, 370

I

IDS, 289, 290, 291
IFC, 8, 73, 138, 142, 186, 227, 261, 262, 263, 264, 265, 266, 267, 268, 272, 273, 276, 277, 278, 279, 280, 284, 286, 292, 296, 302, 329, 357, 414, 415, 417, 497
IGES, 262, 263, 276
iLOD, 289
IoT, 10, 18, 67, 271, 367, 369, 405, 413, 417, 418, 419, 455, 460, 465, 482, 484
ISO 19650, 388

J

JavaScript, 320, 369, 378
JSON, 88, 89, 90, 92, 123, 128, 142, 269, 272, 280, 329, 330, 333, 378, 474, 480
Jupyter Notebook, 114, 115, 116, 130, 187, 224, 330, 346, 417, 425

K

Kaggle, 115, 121, 130, 187, 224, 303, 307, 330, 346, 408, 415, 417, 425, 426, 430, 431, 433
k-NN, 392, 393, 442, 445, 446, 447, 448
KPI, 245, 311, 317, 318, 319, 320, 321, 324, 353, 372, 389, 478, 479

L

LEED, 235, 236, 238
LlaMa, 103, 104, 120, 124, 129, 133, 166, 219, 300, 329,

343, 349, 381, 421, 429, 443, 494
LLM, 3, 4, 24, 29, 50, 51, 52, 55, 56, 92, 95, 99, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 118, 120, 124, 125, 126, 129, 131, 133, 134, 135, 166, 187, 219, 220, 221, 222, 224, 225, 226, 231, 237, 238, 251, 294, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 328, 329, 332, 333, 334, 335, 336, 338, 339, 340, 343, 344, 345, 346, 347, 348, 349, 354, 370, 372, 376, 381, 382, 392, 394, 399, 413, 425, 429, 430, 432, 433, 435, 443, 444, 457, 460, 461, 476, 488, 494
Tải, 81, 128, 134, 193, 311, 323, 325, 326, 333, 334, 338, 339, 342, 343, 361, 362, 363, 365, 367, 383, 384, 450, 475
LOD, 287, 289
LOI, 287
LOMD, 287
Low-Code, 368, 369

M

Matplotlib, 123, 306, 320, 333, 335, 342, 372, 409, 411, 494
MCAD, 77, 78, 257, 284
MEP, 14, 175, 492
MESH, 234, 283, 284, 285, 296, 462
Microsoft SQL, 65
Mistral, 103, 104, 107, 110, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 421, 429, 443, 494
MRP, 11, 12
MS Project, 70
MySQL, 63, 64, 65, 291, 329, 330

N

n8n, 116, 311, 361, 368, 369, 370, 371, 372
NLP, 69, 394
No-Code, 368, 369
NURBS, 142, 282, 283, 284, 285
NWC, 8, 276

O

OBJ, 143, 273, 276, 277, 278, 280, 281, 284, 285
OCCT, 273
OCR, 69, 128, 132, 134
OmniClass, 154, 155, 156
open BIM, 142, 216, 219, 256, 261, 278, 291
Open Source, 45, 55, 97, 98, 107, 108, 273, 275, 494
OWL, 267, 268, 269

P

Pandas, 56, 67, 103, 117, 118, 119, 120, 121, 122, 123, 125, 130, 134, 138, 186, 187, 220, 221, 225, 227, 269, 300, 303, 328, 329, 330, 377, 378, 380, 381, 403, 408, 409, 410, 411, 412, 414, 475, 479, 494
Parquet, 67, 123, 329, 330, 373, 378, 379, 380, 381, 382, 386, 401, 403, 414, 474, 480
PDF, 69, 70, 71, 85, 111, 126, 127, 128, 129, 130, 131, 132, 134, 146, 147, 177, 186, 190, 191, 211, 215, 278, 338, 339, 340, 341, 342, 344, 345, 346, 347, 354, 355, 356, 357, 365, 376, 475, 480, 497
PDM, 16
PHP, 63, 339
Pipeline, 44, 53, 115, 128, 183, 303, 307, 308, 311, 312, 349, 350, 351, 352, 354, 355, 356, 367, 370, 417, 430, 476
PLM, 16, 246
PLN, 8, 272, 296
Plotly, 320, 336, 337, 372
PMIS, 3, 24, 32, 62, 152, 196, 198, 211, 239, 240, 245, 246, 247, 248, 249, 250, 272, 326, 351, 387, 388, 389, 390, 458, 462, 487
PMS, 84, 151, 233
PostgreSQL, 63, 64, 65, 329, 395
Power BI, 320, 372, 482
vốn tư nhân, 464, 487
Python, 56, 63, 103, 105, 112, 114, 115, 117, 118, 119, 129, 130, 131, 133, 134, 135, 166, 167, 179, 187, 219, 224, 225, 227, 303, 304, 308, 320, 329, 330, 332, 333, 339, 340, 346, 348, 356, 367, 369, 372, 378, 409, 417, 426, 459, 475, 476, 479, 494

Q

QTO, 72, 196, 214, 215, 216, 217, 218, 219, 221, 223, 225, 226, 228, 237, 238, 242, 243, 251, 301, 302, 475, 479
QWEN, 103, 104, 124, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 429, 443

R

RAG, 111, 116
RDBMS, 63, 64, 65, 82, 89
RDF, 267, 268, 269
RegEx, 126, 136, 177, 179, 331, 332, 333, 357, 474
RFID, 8, 18, 58, 84, 418, 419, 455, 460, 465, 482
ROI, 311, 317, 319, 321, 351, 370, 389, 479
RPM, 14, 84, 151, 331, 333, 334
RVT, 8, 73, 77, 140, 186, 227, 272, 296, 300, 302, 329, 357, 376, 414, 415, 417, 497

S

SaaS, 24, 50, 51, 52, 458
SCOPE, 273, 277
SDK, 139, 141, 257, 264, 273, 281, 286, 296, 329, 414
Seaborn, 123, 320, 336, 337, 408, 412
SPARQL, 269
SQL, 63, 65, 66, 88, 89, 103, 105, 119, 123, 166, 168, 268, 269, 276, 277, 300, 329, 376, 392, 494
SQLite, 63, 64, 65, 166, 167, 296, 329, 363
STEP, 261, 262, 263, 266, 272, 276, 277, 292
Streamlit, 336, 337
SVF, 142, 276, 285

T

Transform, 128, 193, 311, 323, 325, 326, 330, 331, 333, 334, 340, 341, 342, 344, 347, 361, 362, 363, 365, 367, 383, 384, 450, 475

U

Uniclass, 154, 155, 156
USD, 142, 143, 255, 276, 277, 278, 279, 280, 281, 284, 285, 286, 292, 296

V

VectorOps, 373, 398, 400, 401
VR, 84, 271, 285

W

WhiteBox, 240, 242, 243

X

XLSX, 8, 61, 62, 123, 128, 129, 231, 268, 276, 277, 280, 296, 299, 308, 351, 373, 377, 378, 379, 380, 401, 474, 480
XML, 61, 88, 89, 92, 128, 143, 269, 272, 277, 280, 291, 299, 329, 378, 414, 474