第二版，修订和补充

数据驱动建设

在建筑行业中穿越数据时代

+ 人工智能和大型语言模型的使用实例

Artem Boiko



MORE LANGUAGES ON THE SITE DATADRIVENCONSTRUCTION.IO









数据驱动建设

在建筑行业中穿越数据时代

第二版，修订和补充

ARTEM BOIKO

"Boiko 是 IT 界的詹姆斯-卡维尔（James Carville）--在后者广为引用的 "傻瓜，是经济问题 "中，这本著名的书只需要换一个词。"愚蠢的是数据"。(要想在数据世界中找到方向，古罗马人的一句话可以追溯到希腊语，至今仍然有效："Navigare necesse est"）。作者用一双精准的手和一个坚定不移的指南针，带领读者穿越数据海洋的深浅，更不用说全面的历史方法，最后但并非最不重要的是，他还拥有非常新颖的图表和幽默感，这些都不是一眼就能看出来的。国际上对博伊科这本书的反响不一，既有欣喜若狂的赞许，也有嗤之以鼻的怀疑。博伊科是一位独创的、不信奉教条的数据思想家。他向读者展示了令人兴奋的见解，并总是勇敢地、甚至是挑衅性地提出能激发读者进一步思考的论点。这本书是治疗德国人潜伏的共识主义疾病的良药。顺便提一句，上述拉丁谚语有一个补充："vivere non est necesse"。这句话并不适用于博伊科处理数据世界的方法--数据是有生命的，其生命是必要的，更不用说是至关重要的"。

- Burkhard Talebitari 博士，自由编辑--包括自 2013 年起由 Ernst & Sohn 出版的年度期刊：BIM。

"Artem Boiko 的著作是建筑行业数字化民主化的里程碑，真正改变了中小型企业（SMEs）的游戏规则。尤其具有开创性的是：通过使用现代开源低代码和无代码工具，企业已经可以有效地将数据整合到业务流程中，并对其进行盈利分析，而无需任何深入的编程知识。这使得使用繁琐的商业软件包变得多余。本书呼吁人们行动起来！对于那些不仅想了解建筑行业的数字化转型，而且还想以务实、高效和前瞻性的方式积极推动数字化转型的人来说，这是一本非常有价值的指南。现在，我们应该共同努力，分享这些知识，可持续地提高建筑行业的生产力"。

-Michael Max Buehler 博士，康斯坦茨 HTWG 建筑管理教授，GemeinWerk Ventures 联合所有人，DevvStream 独立董事。

"数据驱动建筑 "一书是超越建筑商常规世界界限的第一步，他们拥有复杂的设计和管理系统，而数据的复杂性和饱和度似乎根本无法从根本上简化和提高建筑数据工作的透明度。阿特姆在书中用简洁的语言向我们展示了现代数据工作技术为我们带来的机遇，并给出了可以立即应用到工作中的具体步骤。我建议每一个想了解自动化系统在建筑行业的发展方向的人仔细研读这本书，以便认识到建筑行业的数据革命已经来临。现在只有极客才会对它感兴趣，但再过几年，就像 BIM 一样，这样的方法和软件将无处不在！"

- Ihor Rogachew，RGD IMT 能力中心、BIM 和数字化转型负责人，InfraBIM.Pro 创始人。

"我强烈推荐 DataDrivenConstruction 一书，正如书名所言，它涉及 AECO 的数据驱动信息管理方法。我目前正在使用这本书来帮助与不同的团体开展一些讨论。我发现这是一本非常容易理解的参考书。除了全面概述 AECO 工具的历史背景、数据和介绍几种关键技术外，书中还包含了许多非常有用的图表，概述了数据源的范围和带有工作流程示例的最终用户人工制品。我认为，在制定和监控信息战略时，我们需要更多这种类型的图表，并为 BEP 做出贡献--定义整体企业数据模型，并将 PIM 和 AIM 的边界叠加到该模型上"。

- Paul Ransley，Acmena 首席顾问，伦敦交通局系统集成工程师。

"如果说 "数据是新的石油"，那么我们就需要学会定义它、发现它、挖掘它、提炼它，使它变得有价值。我认为《数据驱动的建筑》一书内容丰富，见解独到。这本书提供了有用的历史背景，并用通俗易懂的语言解释了如何使用数据。对于那些对数字化转型感兴趣的人来说，这本书让他们很好地了解了数据--它是如何工作的、它是如何结构化的以及它是如何被使用的

- Ralph Montague，ArcDox 总监，BIM 协调员峰会总监，爱尔兰国家标准管理局 BIM 国家镜像委员会主席。

"正如书中所强调的那样，信息是建筑行业的重要资产，以易于获取的形式提供信息可极大地促进准确决策，加快项目进度。该书提供了一种中立、高效的方法来获取和利用这一决策来源。书中介绍的方法利用了一种现代方法，将人工智能驱动的编程与可访问的开源工具相结合。通过利用人工智能，并利用开源软件，该方法旨在提高自动化程度、优化流程，并促进该领域内的可访问性和协作。本书语言清晰，通俗易懂"。

- Salih Ofluoğlu 博士，安塔利亚比林大学美术与建筑学院院长，欧亚 BIM 论坛组织者。

我只能说 "哇"！你将历史、法学硕士、图形和你的观点的整体易懂性融为一体的方式真的非常了不起。本书的流畅性令人惊叹。这本书有很多精彩之处；它真正改变了游戏规则。这是一个很好的信息来源，我对你为之付出的努力和热情表示赞赏。祝贺你创作出如此杰出的作品。我可以继续说下去，但我只想说，我对你印象深刻！"

- Natasha Prinsloo，energylab_数字实践负责人。

"对于建筑行业的任何人来说，从新手到经验丰富的专家，这本书都是改变游戏规则的利器！这不是一本典型的尘封读物，它充满了洞察力、策略和幽默感，让你读得津津有味。从古老的数据记录方法到尖端的数字技术，本书涵盖了建筑数据使用的演变过程。它就像一台时光机，穿越了建筑数据的发展历程。无论您是建筑师、工程师、项目经理还是数据分析师，这本全面的指南都将彻底改变您处理项目的方式。准备好以前所未有的方式优化流程、加强决策和管理项目吧！"

- Pierpaolo Vergati，罗马萨皮恩扎大学讲师，Fintecna 高级建筑项目经理。

"我一口气读完了这本书，用了不到 6 个小时。这本书的制作质量非常

好，纸张光洁致密，色彩搭配合理，字体悦目。书中提供了大量关于如何使用建筑行业专用的 LLM 的实用范例，这将为您节省数月甚至数年的自学时间。工作示例多种多样，从简单到复杂，无需购买复杂昂贵的软件。本书将使建筑行业任何企业的所有者重新审视其业务战略、数字化和发展前景。对于规模较小的公司来说，也可以利用经济实惠的免费工具提高效率"。

- Mikhail Kosarev，TIM-ASG 建筑行业数字化转型讲师兼顾问。

"数据驱动的建筑业》一书对于任何想了解数据时代建筑业走向的人来说，都是一本改变游戏规则的书。Artem 并没有浅尝辄止，而是深入探讨了建筑业当前的发展、挑战和大有可为的机遇。本书的与众不同之处在于它的通俗易懂--Artem 用通俗易懂的比喻来解释复杂的观点，使内容易于掌握。我发现这本书信息量非常大，而且引人入胜。总之，Artem 精心制作了一本宝贵的资料，不仅能提供信息，还能激发灵感。无论您是经验丰富的专业人士，还是建筑行业的新手，这本书都将拓宽您的视野，加深您对行业发展方向的理解。强烈推荐！"

- Moayad Saleh，TMM GROUP Gesamtplanungs GmbH 建筑师兼 BIM 实施经理。

"应该说，《数据驱动的建筑》不愧为大学的教科书，是一本对发展中的 BIM 领域做出宝贵贡献的书。数据驱动的建筑》一书中的技术词汇表很好地解释了相关概念。书中用非常优美的视觉语言将极难解释的主题变得简单易懂。我认为，视觉效果所要解释的内容应该向读者表达出来，哪怕是简短的表达。有些视觉效果的可理解性，换句话说，阅读视觉效果需要单独的信息。我还想说，我很高兴在大学的讲座和研讨会上介绍阿尔腾-博伊科的宝贵作品"。

- Ediz Yazicioglu 博士，ArchCube 公司所有人，伊斯坦布尔技术大学和 Medipol 大学建筑系建筑项目管理讲师。

"数据驱动的建筑 "生动地传达了以建筑数据为基础的信息化工作的基本原理。这本书涉及信息流和基本经济概念，因此有别于其他 BIM 书籍，因为它不仅代表了软件制造商的视角，还试图传达基本概念。这本书值得一读，值得一看"。

- Jakob Hirn，Build Informed GmbH 首席执行官兼联合创始人，"On Top With BIM "创新论坛发起人。

俗话说 "数据是新的石油"，因此数据勘探者或开采者应该拥有正确的工具和思维方式，以便从这一 21 世纪的资源中获取价值。长期以来，建筑行业一直处于基于 "三维信息 "流程的滑坡状态，即项目交付是基于他人提供的信息（例如，他们已经绘制了饼图或柱状图），而底层 "数据"（例如原始电子表格）能够提供更多信息，尤其是因为多数据融合和人工智能带来了无限潜力。如果你正在从事建筑工程的交付（或教学/研究）工作，这本书是你驾驭数据驱动的世界的最佳资源，也是迄今为止唯一的资源。

- 英国 LSBU 建筑战略信息技术副教授 **Zulfikar Adamu** 博士。

Artem Boiko 的 "数据驱动的建筑 "是一部令人印象深刻的作品，在技术和信息不断发展的时代，它为建筑行业提供了坚实的基础。博伊科设法以易于理解的方式介绍了复杂的主题，同时还引入了富有远见的观点。该书是一本经过深思熟虑的汇编，不仅突出了当前的发展，还对未来的创新进行了展望。强烈推荐给想要了解数据驱动型建筑规划和执行的任何人。

- Markus Eiberger，斯图加特应用科学大学讲师，Konstruktionsgruppe Bauen 高级项目经理兼分公司副经理，巴登-符腾堡州 BIM 群协会董事会成员。

第二版，2025 年 4 月。

© 2025 | Artem Boiko | 卡尔斯鲁厄

ISBN 978-3-9827303-7-0







Artem Boyko 版权

boikoartem@gmail.com
info@datadrivenconstruction.io

未经作者书面许可，不得以任何形式或通过任何电子或机械手段，包括影印、录制或任何信息存储和检索系统，复制或传播本书的任何部分--未经改动的非商业性分发除外。本书免费发行，可出于个人、教育或研究目的与其他用户自由共享，但须保留作者姓名和原文出处。作者保留对文本的所有非专有权利，不作任何明示或暗示的保证。书中提到的公司、产品和名称可能是虚构的或用作示例。作者对使用所提供信息的任何后果不承担任何责任。本书所含信息按 "原样 "提供，不保证完整性或相关性。对于因使用本书所含信息、代码或程序而造成的附带或间接损失，作者不承担任何责任。本书中提供的代码示例仅用于教育目的。读者需自行承担使用风险。作者建议，在生产环境中使用所有软件解决方案之前，应对其进行测试。文中提及的所有商标和产品名称均为其各自公司的商标、注册商标或服务标志，是其各自所有者的财产。书中使用这些名称并不意味着与它们的所有者有任何关系或得到了它们的认可。提及第三方产品或服务并不构成推荐或暗示认可。示例中使用的公司和产品名称可能是其所有者的商标。提供第三方网站链接是为了方便读者，并不意味着作者认可这些网站提供的信息。所有引用的统计数据、引文和研究均为撰写本文时的最新数据。数据可能会随时间发生变化

本书采用知识共享署名-非商业性-禁止衍生 4.0 国际许可协议 (CC BY-NC-ND 4.0) 发布。您可以出于非商业目的复制和分发本书，但必须保留作者姓名且不得做任何改动。

cc by-nc-nd 4.0

© 2024 Artyom Boyko。第一版。
© 2025 阿尔乔姆-博伊科。第二版，修订和补充。
保留所有权利

第二版序言

本书是与专业人士热烈对话的成果。它基于在各种专业平台和社交媒体平台上就建筑行业数据管理进行的大量专业讨论。这些讨论成为文章、出版物和视觉材料的基础，在专业界引起了广泛反响。作者的内容每年在各种平台和语言上吸引数百万次浏览，汇聚了建筑数字化领域的专业人士。

在第一版出版后的一年内，就有来自巴西、秘鲁、毛里求斯和日本等 50 多个国家的专家订购了该书。根据专家的反馈意见、对第一版的批评以及专业界的讨论，我们对大家现在手中的这本书的第二版进行了修订和扩充。根据反馈意见，第二版的内容得到了大幅扩充：新增了关于 CAD (BIM) 技术和创建有效 ETL 流程的章节。实用范例和案例研究的数量也大幅增加。在本书第一版出版前后，建筑行业、咨询公司和大型 IT 公司的领导向作者提出了有关数字化和互操作性 的问题，他们的反馈意见尤为宝贵。他们中的许多人已经应用了书中介绍的方法，或计划在不久的将来这样做

您手中的这本书是通过讨论和积极交换意见而创作的。进步源于对话，源于观点的碰撞和对新方法的开放。感谢您参与这场对话。您的建设性批评意见是今后改进工作的基础。如果文中有错误，或者您愿意分享想法和建议，欢迎提供任何反馈意见。本书末尾提供了联系方式

为什么这本书是免费的?

本书是一本开放式教育资源，旨在传播建筑行业数据管理的现代方法。本书的第一版是收集专业人士意见和建议的基础，这些意见和建议有助于改进教材的结构和内容。我们对所有意见、建议和想法进行了认真分析，并将其纳入了本修订版。本书的目的是帮助建筑专业人员了解数据工作的重要性：系统地、有意识地并着眼于信息的长期价值。作者从十多年的建筑数字化工作中收集了大量实例、插图和实践经验。这些资料大多来自实际项目、与工程师和开发商的讨论、参与国际活动和培训研讨会。本书试图将积累的经验条理化，并以易于理解的方式与读者分享。如果您想支持本书观点的进一步传播，并以方便的形式阅读、使用实例和视觉资料，可以购买印刷版。

使用权

本书的所有资料、插图和片段均可以任何形式在任何媒体上复制、引用或使用，但必须注明出处：作者：Artem Boyko，书名：《数据驱动的建筑》。感谢您对劳动和知识传播的尊重。

我怀着诚挚的感激之情，将献给我的家人，他们从小就灌输给我对建筑的深深热爱；献给我的家乡矿区，因为他们教会了我坚韧不拔的精神；也献给我的测量师妻子，她不离不弃的支持一直激励着我。

本书的读者对象

本书以通俗易懂的语言编写，面向建筑行业的广大读者--从希望掌握现代建筑流程基础知识的学生和新手，到需要最新建筑数据管理方法的专业人士。无论您是建筑师、工程师、工头、施工经理还是数据分析师，这本包含大量独特插图和图表的综合指南都能为您提供宝贵的见解，帮助您了解如何在业务中使用数据来优化和自动化流程、改进决策以及使用现代工具管理不同级别的施工项目

本书是一本综合性指南，结合了将数据管理技术融入建筑流程的理论基础和实践建议。书中重点介绍了如何战略性地使用信息来优化运营、实现流程自动化、改进决策以及使用现代数字工具有效管理项目。

本书涵盖了建筑行业信息工作的理论和实践方面。通过详细的实例，它探讨了任务参数化、需求收集、处理非结构化和多格式数据以及将其转化为建筑公司有效解决方案的方法。

读者可以从需求的形成、基本数据模型的开发，到更复杂的异构信息源整合过程、ETL 流程的创建、信息管道的构建和机器学习模型的建立，逐一了解。这种循序渐进的方法可以清晰地展示建筑行业业务流程和决策支持系统的组织和自动化机制。本书每一部分最后都有一个实用章节，其中包含分步指导，可以将所学知识立即应用到实际项目中。

部分内容简介

本书围绕价值链中的数据转换概念展开：从数据收集和质量保证到分析处理，以及利用现代工具和方法提取有价值的实用解决方案。

第 **1** 部分：建筑业的数字化发展 - 追溯数据管理从泥板到现代数字系统的历史变迁，分析模块化系统的出现以及信息数字化在工业革命背景下日益增长的重要性。

第 **2** 部分：建筑行业面临的信息挑战--探讨了数据分散、"信息孤岛"、HiPPO 方法对决策的影响以及专有格式的局限性等问题，并建议考虑向人工智能和 LLM 生态系统转变。

第 **3** 部分：建筑数据的系统化--对建筑数据进行分类，介绍其组织方法、与企业系统的整合，并讨论建立能力中心以实现信息流程的标准化。

第 **4** 部分：数据质量保证--揭示将不同信息转化为高质量结构化数据的方法，包括从各种来源提取数据、使用 LLM 进行验证和建模。

第 **5** 部分：成本和时间计算--涉及成本和规划计算的数字化、从 CAD（BIM）模型获取工程量的自动化、4D-8D 建模技术以及建筑项目的 ESG 计算。

第 **6** 部分：计算机辅助设计（**CAD**）和建筑信息模型（**BIM**）--批判性地分析设计技术的演变、系统互操作性问题、开放数据格式的发展趋势以及将人工智能应用于设计的前景。

第 **7** 部分：数据分析与自动化 - 介绍信息可视化原理、关键性能指标、ETL 流程、工作流协调工具以及应用语言模型实现日常任务自动化。

第 **8** 部分：数据存储和管理--探讨数据存储格式、数据仓库和数据湖概念、数据管理原则和新方法，包括矢量数据库以及 DataOps 和 VectorOps 方法。

第 **9** 部分：大数据和机器学习--重点关注向基于历史数据的客观分析过渡、建筑工地上的物联网，以及应用机器学习算法预测项目成本和工期。

第 **10** 部分：数字数据时代的建筑业--展望建筑业的未来，分析从因果分析到关联分析的转变、建筑业 "优步化 "的概念以及数字化转型战略。

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



引言

在技术飞速发展的世界里，从时间和成本计算到风险分析，业务的方方面面都在通过机器学习模型实现自动化 ，在这种情况下，您的公司还能保持多久的竞争力?

与人类历史一样悠久的建筑业正处于一场革命性变革的风口浪尖，这场变革有望彻底改变我们对传统建筑业的看法。在其他经济领域，数字化不仅改变了规则，还将那些无法适应新的数据处理环境、无法提高决策速度的公司无情地逐出市场（图 1）。

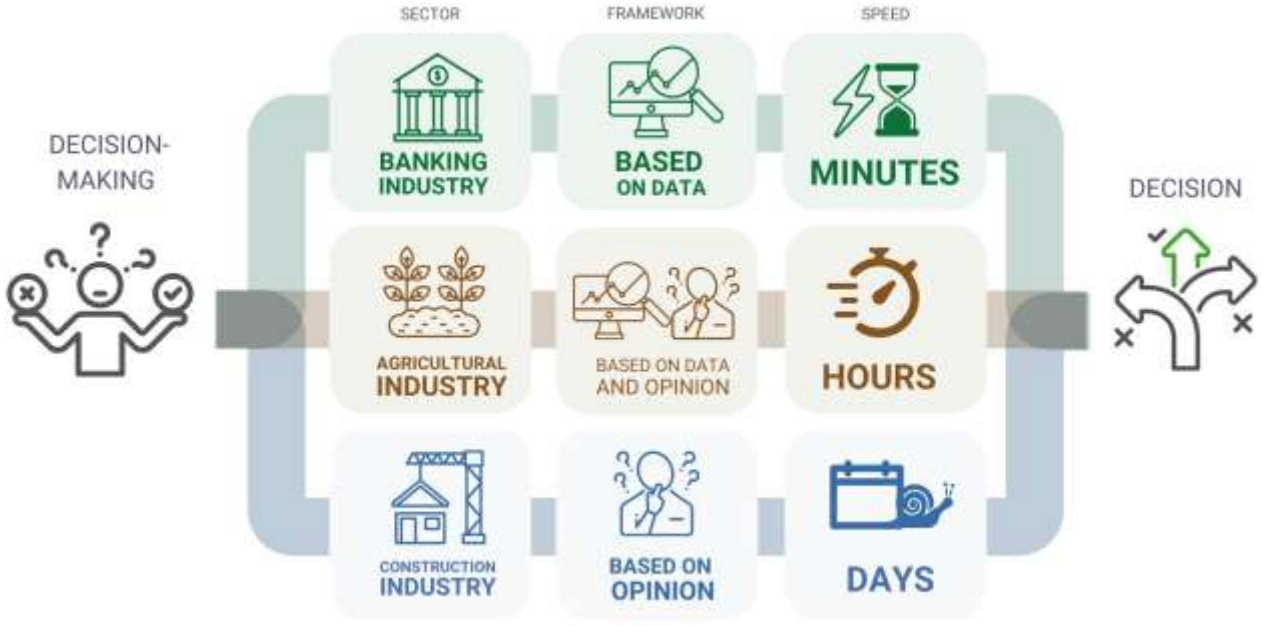


图 1 与其他行业相比，建筑行业的决策速度更多地取决于人的因素。

银行业、零售业、物流业和农业综合企业正迅速向全面数字化迈进，不准确和主观意见在这里已无立足之地。现代算法能够分析海量数据，为客户提供准确的预测--无论是偿还贷款的可能性、最佳送货路线还是风险预测。

建筑业是最后一个不可避免地从基于高薪专家意见的解决方案向数据驱动型解决方案过渡的行业之一。推动这一转变的不仅有新的技术能力，还有市场和客户对透明度、准确性和速度日益增长的需求。

机器人化 、流程自动化、开放数据 以及基于数据的预测 - 所有这些都不再只是可能，而是不可避免的。 建

筑行业的大多数公司最近都要对客户负责，计算工程量、成本、工期和质量控制，但现在却有可能沦为单纯的订单执行者，无法做出关键决策（图 2）。

随着计算能力、机器学习算法和民主化数据访问 的进步，现在可以自动组合来自不同来源的数据，从而在建筑项目的讨论阶段实现更深入的流程分析、风险预测和成本优化。这些技术有可能从根本上提高整个行业的效率并降低成本。

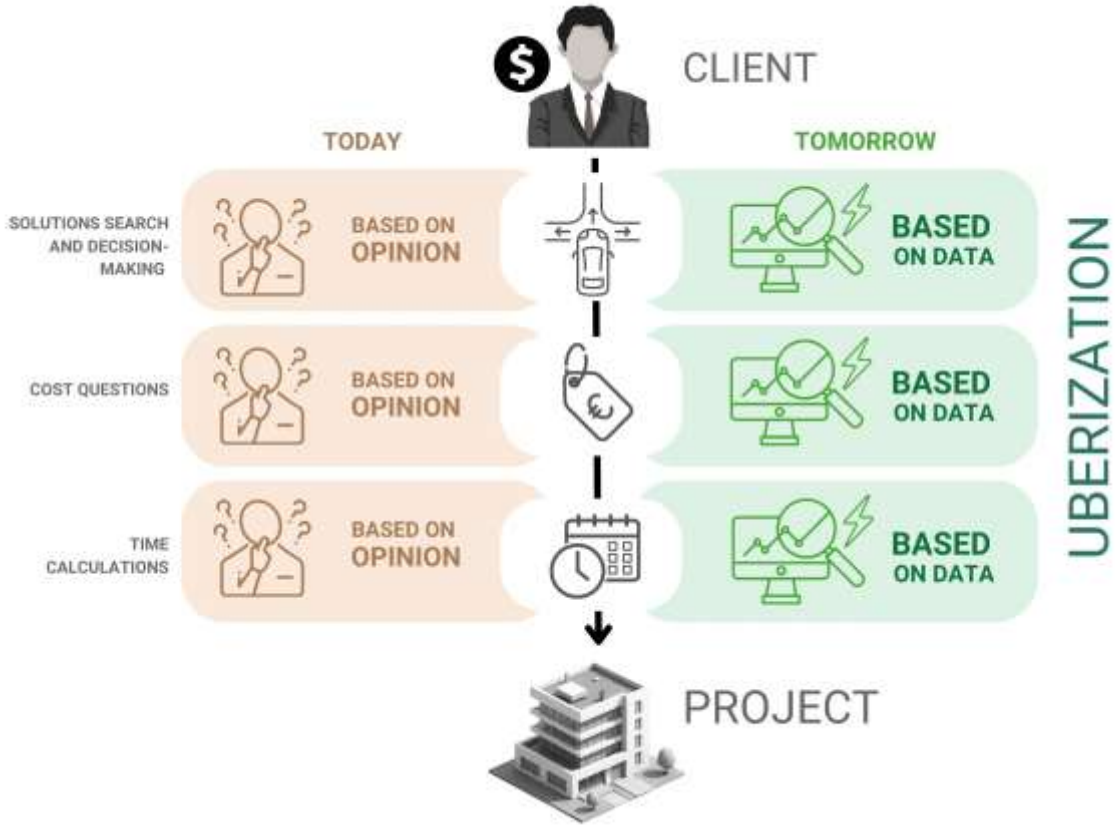


图 2 客户对项目实施过程中过多的人为因素不感兴趣。

尽管新工具和新概念具有种种优势，但建筑业在采用新技术方面却远远落后于其他经济部门。

> 根据《IT Metrics Key Data 2017》报告，在其他 19 个行业中，建筑业的 IT 支出排名倒数第一[1]。

数据量的快速增长和流程的复杂性正成为公司管理层头疼的问题，而使用新技术的主要问题是，尽管数据量很大，但仍然是零散的、非结构化的，而且不同系统和软件产品之间往往不兼容。这就是为什么建筑行业的许多公司现在主要关注数据质量问题 ，只有实施高效、自动化的管理和分析系统才能解决这个问题。

> 根据毕马威会计师事务所（KPMG®）在 2023 年对建筑经理进行的一项调查[2]，项目管理信息系统（PMIS ）、高级和基础数据分析 和建筑信息模型（BIM ）在提高项目投资回报率方面潜力最大（图 3）。

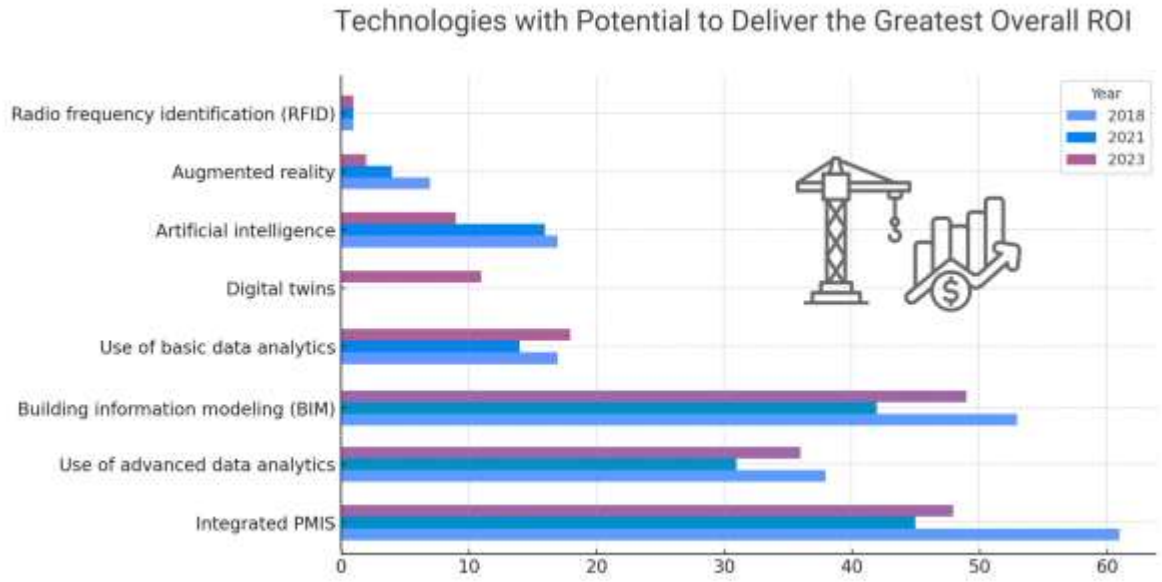


图 3 在建筑公司经理中开展的调查：哪些技术将为资本项目提供最高的投资回报（ROI）？(根据材料 [2]）。

要应对将数据整合到业务流程中的挑战，就要确保高质量的信息，使用合适的数据格式，并采用有效的方法来创建、存储、分析和处理数据。

数据价值的实现正迫使各行各业摒弃孤立的应用程序和复杂的官僚管理结构。取而代之的是，重点正在转向创建新的信息架构方法，将公司转变为以数据为驱动的现代企业。建筑行业自身迟早也会迈出这一步，从逐步的数字化演变转变为影响所有公司的真正的数字化革命。

> 向数据驱动型业务流程过渡并非易事。许多公司将面临挑战，因为高管们并不总是了解如何利用混乱的数据集来提高效率和业务增长。

本书深入探讨了数据世界，信息正在成为决定业务流程效率和可持续性的关键战略资源。随着信息的快速增长，企业正面临着新的挑战。数字化转型不再只是一个流行词，它正成为一种必然。



图 4 数据和流程是建筑的基础。

理解变革意味着能够用简单的语言解释复杂的事物。因此，本书采用通俗易懂的语言编写，并配有作者专门为清晰解释关键概念而创作的插图。这些示意图、图表和可视化设计旨在消除感知障碍，使以前认为此类主题过于复杂的人也能理解这些材料。本书中的所有插图、图表和图形均由作者创作，专门用于将文中描述的关键概念形象化。

> 一图胜千言 [3]。
>
> - Fred R. Barnard，英国插画家，1927 年。

为了将理论与实践相结合，我们将使用人工智能工具（特别是语言模型），让您无需深厚的编程知识即可开发解决方案。如果你偏向于实用资料，对数据的实际操作更感兴趣，可以跳过第一部分的介绍，直接进入本书的第二部分，这里开始介绍具体的例子和案例。

但是，不要对 AI（人工智能）、机器学习和 LLM（大型语言模型）工具寄予过高的期望。如果没有高质量的输入数据和对主题的深刻理解，即使是最先进的算法也无法提供可靠而有意义的结果。

> 微软首席执行官萨蒂亚-纳德拉在 2025 年初的 [4] 中警告人工智能存在泡沫风险，并将当前的炒作与互联网泡沫相提并论。他强调，在没有适当理由的情况下宣称达到 AGI（人工通用智能）里程碑是 "毫无意义的指标操纵"。纳德拉认为，人工智能 ，真正的成功应通过其对全球 GDP 增长的贡献来衡量，而不是过分强调流行语。

在所有关于新技术和新概念的流行语背后，是一项复杂而艰苦的工作，以确保数据质量 、业务流程参数化和工具适应实际任务。

> 数据驱动型方法不是一种可以随便下载或购买的产品。它是一项必须建立的战略。它始于对现有流程和问题的重新审视，然后需要朝着选定的方向有条不紊地前进。

领先的软件开发商和应用软件供应商不会成为建筑业变革的引擎，对他们中的许多人来说，数据驱动的方法是对其既有商业模式的威胁。

> 其他行业（与建筑业不同），如汽车业，已经经历了翻天覆地的颠覆性变革，其数字化转型也在顺利进行中。建筑公司需要迅速果断地采取行动：灵活的公司将获得巨大回报，而犹豫不决的公司将面临严重风险。想想数码摄影给这个行业带来的动荡吧 [5]。
>
> - 世界经济论坛报告《塑造建筑业的未来》，2016 年

那些及时认识到新方法的机遇和好处的公司将获得可持续的竞争优势，并能够在不依赖大型供应商解决方案的情况下发展壮大。

这是你的机会，你不仅能经受住即将到来的信息数字化风暴，还能掌控它。在本书中，您不仅可以看到对行业现状的分析，还可以找到重新思考和重组流程及业务的具体建议，从而成为新建筑时代的领导者，并提升您的专业经验。

建筑业数字化的未来不仅仅是使用新技术和程序，而是从根本上重新思考数据处理和业务模式。

您的公司准备好迎接这一战略变革了吗？

目录

引言 1

目录 I

II 部分 从泥板到数字革命：信息如何在建筑业中演变 2

章 2.1. 建筑业使用数据的演变 3

建筑业数据时代的诞生 3

从粘土和纸莎草纸到数字技术 4

将流程作为数据驱动体验的工具 5

施工过程信息数字化 7

章 2.2. 现代建筑技术和管理系统 11

数字化革命和模块化 MRP/ERP 系统的出现 11

数据管理系统：从数据挖掘到业务挑战 13

企业菌丝：数据如何连接业务流程 16

章 2.3. 数字革命和数据爆炸 20

数据繁荣的开端是一次进化浪潮 20

现代公司产生的数据量 22

数据存储的成本：经济方面 23

数据积累的前沿：从质量到意义 25

下一步：从数据理论到实际变革 27

III 部分 建筑业如何淹没在混乱的数据中 29

章 3.1. 数据分散和孤岛 30

工具越多，业务效率就越高？ 30

数据孤岛及其对公司业绩的影响 32

重复和缺乏数据质量 因为不统一的结果 36

HiPPO 或决策中的意见危险 38

业务流程的复杂性和动态性不断增加 41

第四次工业革命 (工业 4.0) 和建筑业的第五次工业革命 (工业 5.0) 44

章 3.2. 化繁为简 47

冗余代码和封闭系统是提高生产力的障碍 47

从孤岛到单一数据仓库 49

集成存储系统实现了向人工智能的过渡 agent 50

从数据收集到决策：自动化之路 53

下一步：将混乱变为可管理的系统 55

IV 部分 建筑业务流程中的数据框架 **57**

章 4.1. 建筑数据类型 58

建筑业最重要的数据类型 58

结构化数据 62

关系数据库 RDBMS 和 SQL 查询语言 64

SQL - 数据库中的查询和新趋势 66

非结构化数据 68

文本数据：介于非结构化混沌和结构化混沌之间 y 69

半结构化和松散结构化数据 70

几何数据 及其应用 71

CAD 数据：从设计到数据存储 74

BIM (BOM)概念的出现 以及在流程中使用 CAD 77

章 4.2. 数据统一和结构化 84

在建筑业系统中填入数据 84

数据转换：现代业务分析的重要基础 87

数据模型：数据中的关系和元素之间的关系 90

专有格式及其对数字流程的影响 95

开放格式正在改变数字化方式 98
范式转变：开放源码 作为软件供应商主导时代的终结者 100
结构化开放数据：数字化转型的基础 102
章 4.3. LLM 及其在数据处理和业务流程中的作用 105
LLM 聊天室：ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok 实现数据处理自动化 105
大型 LLM 语言模型： 工作原理 106
使用本地 LLM 查找公司敏感数据 109
全面控制公司的人工智能 以及如何部署自己的 LLM 111
RAG: Intelligent LLM - 可访问公司数据的助理 113
章 4.4. 与 LLM 支持的 IDE 和未来的程序更改 116
选择集成开发环境：从 LLM 实验到商业解决方案 116
支持 LLM 的集成开发环境 和未来的编程变更 119
Python Pandas： 处理数据不可或缺的工具 120
DataFrame： 通用表格数据格式 124
下一步：建立可持续的数据框架 127
V 部分 数据质量：组织、结构化、建模 **129**
章 5.1. 将数据转换为结构化格式 130
学习如何将文档、PDF 、图片和文本转化为结构化格式 130
将 PDF 文件转换为表格的示例 131
将 JPEG、PNG 图像转换为结构化形式 135
将文本数据转换为结构化形式 138
将 CAD 数据 (BIM) 转换为结构化形式 141
CAD 解决方案供应商正在向结构化数据方向发展 146
章 5.2. 分类与整合：建筑数据的通用语言 149
决策速度取决于数据质量 149
数据标准化和集成 150

数字互操作性始于需求 153

共同的建筑语言：分类器在数字化转型中的作用 155

Masterformat, OmniClass, Uniclass 和 CoClass：分类系统的演变 158

章 5.3. 数据建模和示范中心 163

数据建模：概念模型、逻辑模型和物理模型 163

建筑工程中的实用数据建模 167

使用 LLM 创建数据库 170

数据建模卓越中心（CoE） 172

章 5.4. 要求系统化和信息验证 176

需求收集和分析： 将通信转换为结构化数据 176

流程图和概念框架的有效性 180

结构化需求 和 RegEx 正则表达式 182

数据收集 187

核实 数据和核实结果 189

验证结果的可视化 194

数据质量检查与人类生活需求的比较 196

下一步：将数据转化为精确的计算和计划 198

VI 部分 成本和时间计算：将数据纳入施工流程 **200**

章 6.1. 建筑项目的成本计算和估算 201

建筑基础知识：估算数量、成本和时间 201

项目估计费用的计算方法 203

基于资源的方法 建筑成本计算和估算 203

建筑资源数据库：建筑材料和工程目录 204

计算汇编 并根据资源基础计算工程成本 206

最终项目成本计算：从估算到预算 211

章 6.2. 工程量核算以及自动创建估算和进度表 215

从三维到四维和五维：使用体积和定量参数 215

5D 属性，并从 CAD 获取属性体积 215

QTO 工程量清单：按属性对项目数据进行分组 219

使用 LLM 和结构化数据的 QTO 自动化 226

QTO 利用 Excel 电子表格中的分组规则对整个项目进行计算 230

章 6.3. 4D, 6D -8D 和二氧化碳排放量的计算 CO_2 237

4D 模型：将时间纳入施工估算 237

施工进度及其基于成本计算数据的自动化 238

扩展属性层 6D -8D：从能源效率到安全保证 240

估算二氧化碳排放量 并计算建筑项目的二氧化碳排放量 243

章 6.4. 建筑 ERP 和 PMIS 系统 248

建筑企业资源规划系统 - 计算和估算实例系统 248

PMIS：ERP 与施工现场之间的中介 254

投机、利润、孤立和缺乏透明度 企业资源规划系统 和项目管理信息系统缺乏透明度 255

封闭式企业资源规划系统 /PMIS 时代终结：建筑业需要新方法 258

下一步：有效利用项目数据 260

VII 部分 **CAD** 和 **BIM**：市场、现实和建筑设计数据的未来 **263**

章 7.1. BIM- 概念在建筑业的兴起 264

BIM 和开放式 BIM 作为 CAD- 供应商营销概念出现的历史 264

BIM 的现实：取代集成数据库 - 封闭式模块系统 267

开放格式 IFC 在建筑行业的出现 269

IFC 格式问题取决于几何核心 271

出现在语义 和本体 的主题构建中 274

语义技术为何辜负了建筑行业的期望 276

章 7.2. 封闭项目格式和互操作性问题 280

封闭数据和生产力下降：CAD 行业的死胡同 (BIM) 280

CAD 系统之间互操作性的神话 282

转至美元 和粒度数据 m 286

章 7.3. 建筑几何：从线条到立方米 290

当线条变成金钱或建筑商为何需要几何图形 290

从线条到体积：面积和体积如何成为数据 290

转到 MESH、USD 和多边形：使用细分 绘制几何图形 293

LOD, LOI, LOMD - CAD 中 细部的唯一分类 (BIM) 294

新的 CAD 标准 (BIM) - AIA, BEP, IDS, LOD, COBie 297

章 7.4. 设计参数化和使用 LML 进行 CAD 操作 302

CAD 数据唯一性的假象 (BIM): 通往分析和开放格式之路 302

通过参数进行设计：CAD 和 BIM 的未来 305

设计 CAD 数据处理过程中出现的 LLM 308

利用 LLM 和 Pandas 自动分析 DWG - 文件 311

下一步：从封闭格式转向开放数据 317

VIII 部分 数据驱动决策、分析、自动化和机器学习 **319**

章 8.1. 数据分析和数据驱动决策 320

数据作为决策资源 320

可视化 数据：理解和决策的关键 324

关键绩效指标 和投资回报率 325

仪表板和仪表盘：将指标可视化，实现有效管理 327

数据分析 和提问的艺术 329

章 8.2. 无需人工的数据流：为什么需要 ETL 332

ETL 自动化 ： 降低成本，加快数据处理 332

ETL 提取 ： 数据收集 336

ETL Transform ： 应用验证和转换规则 339

ETL Load ： 通过图表和图形可视化 结果 342

ETL Load: 自动创建 PDF 文档 347
ETL Load： 从 FPDF 自动生成文件 349
ETL Load： 报告和加载到其他系统 353
ETL with LLM: 可视化 数据来自 PDF -文档 354
章 8.3. 自动 ETL 输送机（管道） 360
管道： 自动 ETL 输送机 数据 360
管道 -ETL 使用 LLM 进行数据验证过程 364
管道 -ETL: 核查 CAD 中项目元素的数据和信息 (BIM) 367
章 8.4. 协调 ETL 和工作流程：实用解决方案 373
DAG 和 Apache Airflow： 工作流自动化和协调 373
Apache Airflow： ETL 自动化的实际应用 374
Apache NiFi 用于路由选择和数据转换 379
n8n 低代码, 无代码 流程协调 380
下一步：从人工操作转向基于分析的解决方案 383
IX 部分 建筑数据的存储和管理 **385**
章 9.1. 数据基础设施：从存储格式到数字资料库 386
数据原子：有效信息管理的基础 386
信息存储：文件或数据 387
大数据存储：分析流行格式及其有效性 389
使用 Apache Parquet 优化存储 392
DWH: 数据仓库 数据仓库 394
数据湖 - 从 ETL 到 ELT 的演变：从传统清理到灵活处理 396
数据湖架构： 仓库和数据湖的协同作用 398
CDE, PMIS, ERP 或 DWH 和数据湖 401
章 9.2. 数据仓库管理和混乱预防 404
矢量数据库和边界框 404

数据治理)、数据最小化) 和数据沼泽)....407

DataOps 和 VectorOps： 新数据标准....410

下一步：从混乱存储到结构化存储....412

X 部分 大数据、机器学习和预测 **414**

章 10.1. 大数据及其分析....415

建筑业大数据：从直觉到可预测性....415

质疑大数据的可行性：相关性、统计和数据抽样....416

大数据：分析旧金山百万建筑许可数据集的数据....419

基于 CAD 数据的大数据示例 (BIM)....426

IoT 物联网和智能合约....430

章 10.2. 机器学习和预测....434

机器学习 和人工智能将改变我们的构建方式....434

从主观评估到统计预测....437

泰坦尼克数据集：分析世界中的 "你好世界" 数据和大数据....438

机器学习 在行动：从泰坦尼克号乘客到项目管理....444

基于历史数据的预测和预报....448

机器学习的关键概念....451

章 10.3. 利用机器学习预测成本和期限....454

使用机器学习 查找项目成本和进度的示例....454

利用线性回归预测项目成本和时间....456

使用 K 近邻算法预测项目成本和时间（k-NN)....459

下一步：从存储到分析和预测....463

XI 部分 数字时代的建筑业。机遇与挑战**466**

章 11.1. 生存战略：建立竞争优势....467

关联而非计算：建筑分析的未来....467

数据驱动的建筑方法：基础设施的新水平....470

下一代数字办公室：人工智能 如何改变工作空间 472

开放数据和 Uber 化对现有建筑企业构成威胁 475

uberisation 的未决问题 作为利用时间进行转型的最后机会 477

章 11.2. 实施数据驱动方法的实用指南 482

从理论到实践：建筑业数字化转型路线图 482

奠定数字化基础：迈向数字化成熟的 1-5 个步骤 484

释放数据潜力：实现数字化成熟的 5-10 个步骤 489

转型路线图：从混乱到数据驱动型公司 496

工业 5.0 中的建筑：无法再躲藏时如何赚钱 499

结 论 **501**

关于作者 **504**

进一步关系 **505**

译注 **505**

其他技能和概念 **506**

术语表 **510**

参考文献和在线资料清单 **516**

主题索引 **537**

印刷版带来的最大便利

您现在持有的是《数据驱动的建筑》的免费电子版。为了更方便地工作和快速查阅资料，我们建议您关注印刷版：



- 随手可得：印刷版书籍将成为可靠的工作工具，让您在任何工作环境中都能快速找到并使用正确的可视化图示和图表
- 高质量插图：印刷版中的所有图像和图形均以最高质量呈现
- 快速获取信息：导航简便，可做笔记、书签，随时随地使用图书。

通过购买本书的完整印刷版，您将获得一个舒适高效地处理信息的便捷工具：能够在日常工作中快速使用视觉材料，快速查找必要的方案并做笔记。此外，您的购买还支持了开放知识的传播。

您可以从以下网址订购该书的印刷版： datadrivenconstruction.io/books



I 部分

从泥板到数字革命：信息如何在建筑业中演变

本书第一部分探讨了建筑行业数据管理的历史演变，从物理介质上的原始记录到现代数字生态系统。书中分析了信息管理技术的变革、ERP 系统的出现以及数据碎片化对业务流程效率的影响。其中特别关注了信息数字化的进程，以及客观分析取代主观专家判断的日益重要性。详细探讨了现代建筑业所面临的信息量指数级增长以及企业系统所面临的相关挑战。探讨了建筑业在第四次和第五次工业革命背景下的定位，以及利用人工智能 和以数据为中心的方法创造可持续竞争优势的潜力。

章 1.1.

建筑业使用数据的演变

建筑业数据时代的诞生

大约 1 万年前，在新石器时代，人类的发展发生了革命性的转变，放弃了游牧生活方式，转而开始定居生活，从而出现了第一批用粘土、木材和石头建造的原始建筑[6]。从此，建筑业的历史开始了。

随着文明的发展，建筑变得越来越复杂，出现了最早的祭祀神庙和公共建筑。建筑设计日趋复杂，要求古代的工程师和管理人员创造出第一批记录和计算方法。泥板和纸莎草纸上的第一批记录通常包括计算所需建筑材料的数量、成本以及计算工程款的逻辑[7]。由此开启了建筑业使用数据的时代--远早于现代数字技术的出现（图 1.1-1）。

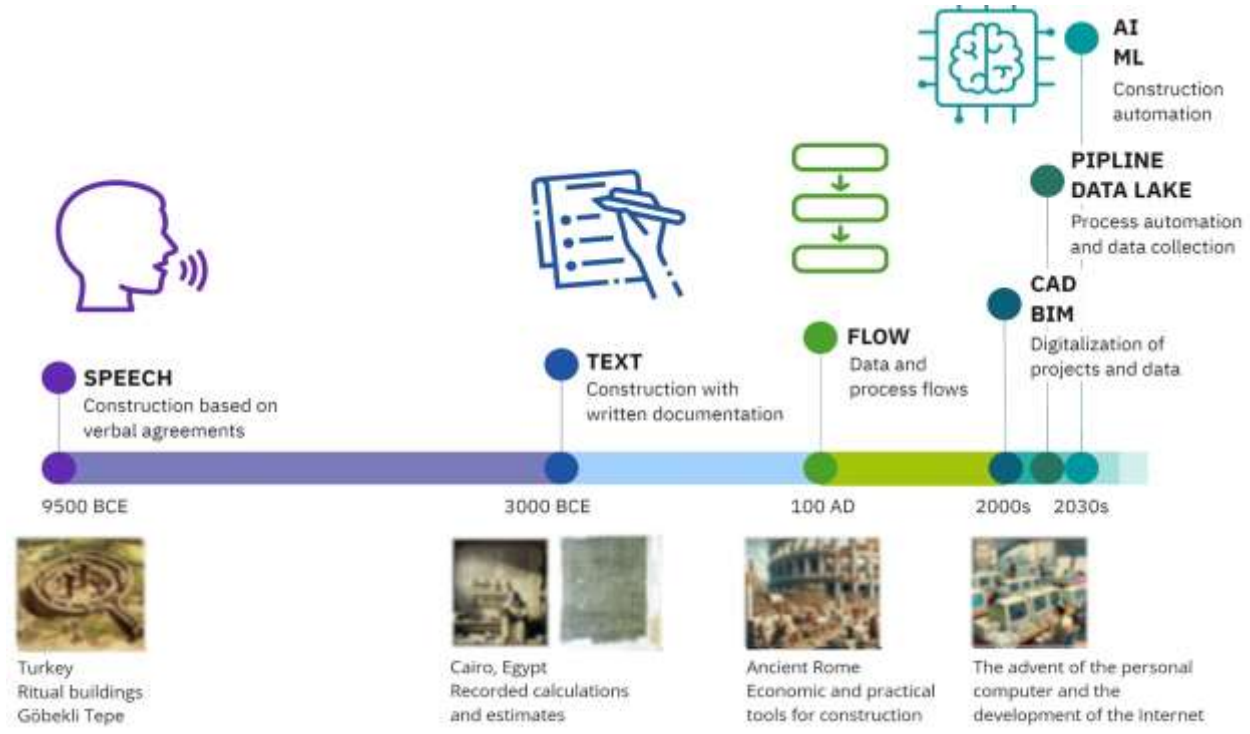


图 1.1-1 建筑信息技术发展大事记：从语言信息到人工智能.

从粘土和纸莎草纸到数字技术

最早的建筑文献可追溯到公元前 3000-4000 年左右的金字塔建筑时期[7]。从那时起，文字记录的保存促进并伴随着建筑业的进步，使知识得以积累和系统化，并在随后的一万年中促成了建筑方法和建筑学的重大创新。

建筑业最初使用泥板、几千年前的纸莎草纸（图 1.1-2）或 20 世纪 80 年代的 "A0 "纸等物理介质来记录数据，最初并不是为了将这些信息应用到新项目中。这些记录的主要目的是详细说明项目的现状，包括所需材料和工程成本的计算。同样，在当今世界，数字设计数据和模型的可用性并不总能保证其在未来项目中的应用，通常主要是作为当前所需材料和建筑成本计算的信息。

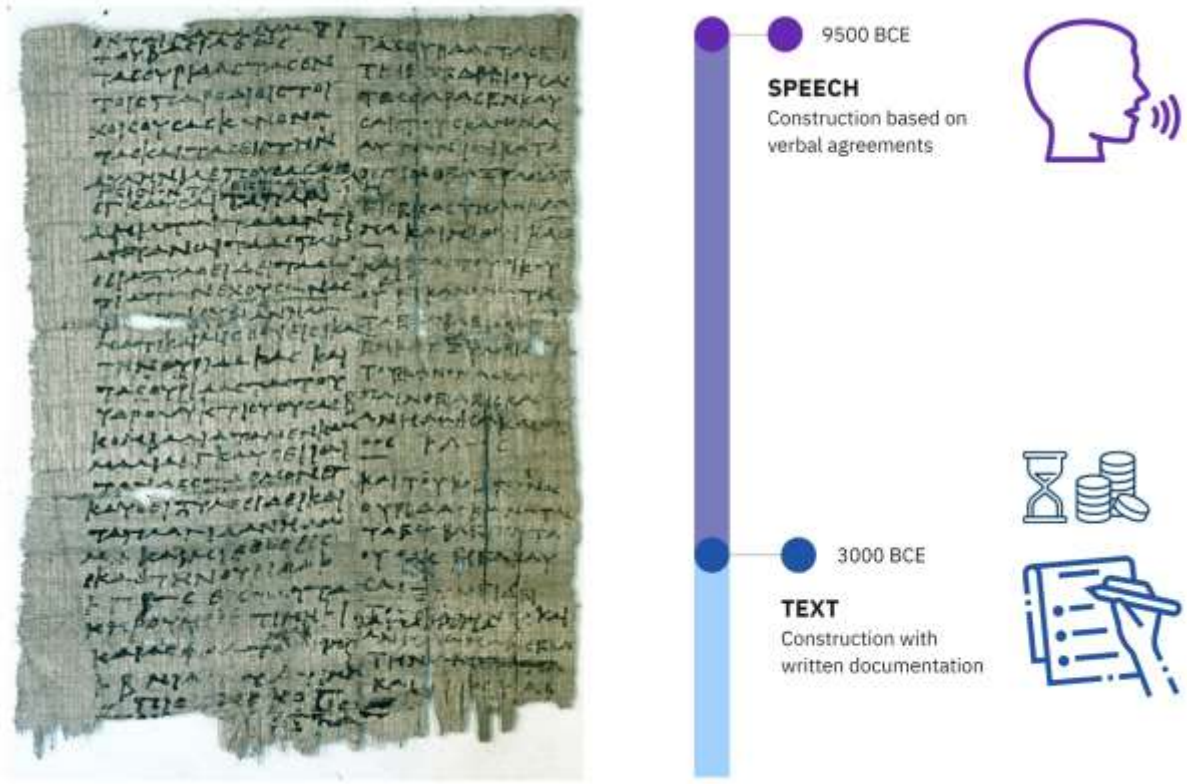


图 1.1-2 一张公元前 3 世纪的纸莎草纸，上面描述了在皇家宫殿中使用漆绘技术绘制各类窗户的成本。

> 人类用了大约 5000 年的时间，在建筑项目管理中从口头交谈转变为书面文件，也用了同样的时间，从纸张转变为数字数据作为规划和控制的主要资源。

正如贸易和货币关系的发展促进了文字和第一批解决纠纷的律师的出现一样，第一批材料成本和建筑工程范围的记录也导致了建筑行业第一批管理人员的出现，他们的职责包括记录、监督和负责有关项目时间表和成本的关键信息。

如今，数据发挥着更为重要的作用：它不仅记录决策，还成为预测和模拟未来的工具。这是现代项目管理流程方法的基础--将积累的经验转化为基于结构化和可验证数据的决策系统。

将流程作为数据驱动体验的工具

任何流程的核心都是将过去的经验转化为规划未来的工具。现代意义上的经验是一组结构化的数据，通过分析可以做出明智的预测。

> 历史数据是预测的基础，因为它清楚地显示了已完成工作的结果，并提供了对影响这些结果的因素的深入了解。

让我们以单体建筑为例：通常在计划施工时间时，会考虑混凝土量、结构复杂程度和天气条件。假设某个工地领班或公司过去三年（2023-2025 年）的历史数据显示，在雨天浇筑 200 平方米的整体结构需要 4.5 至 6 天（图 1.1-3）。正是这些累积的统计数据成为了在规划未来项目类似工程时预测准备时间和计算资源成本的基础。根据这些历史数据，工头或估算人员可以根据经验对 2026 年在类似条件下完成未来类似工程所需的时间做出明智的预测。

就时间估算而言，分析流程是将不同数据转化为结构化经验的机制，然后再转化为精确的规划工具。数据和流程是一个生态系统，缺一不可。

> 计算可计算的东西，测量可测量的东西，让不可测量的东西变得可测量[8]。
>
> - 伽利略-伽利莱

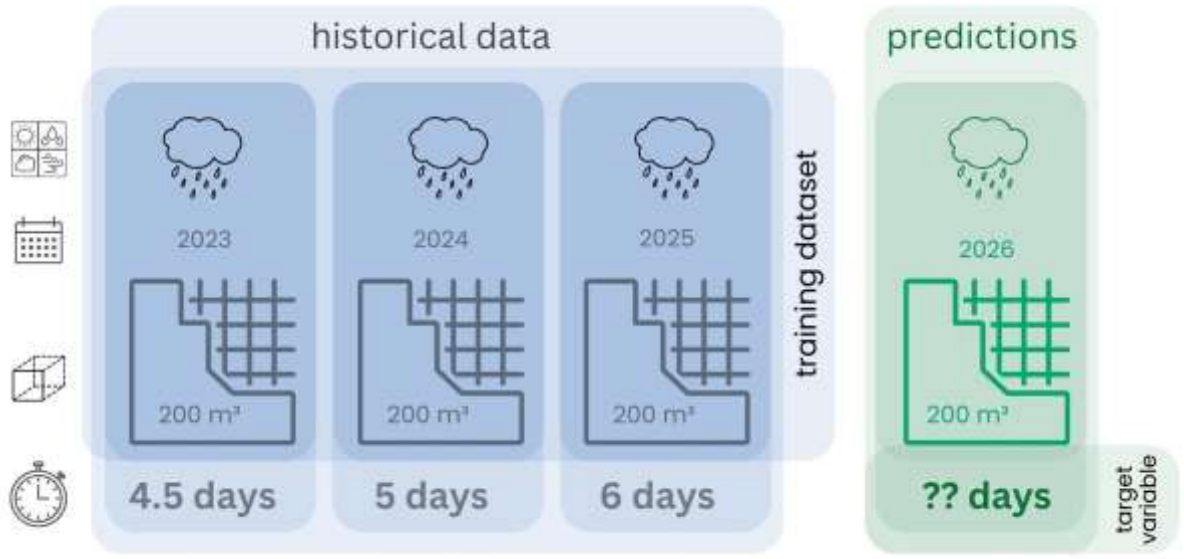


图 1.1-3 历史数据作为预测未来某一数值的训练数据集

在当今的商业环境中，数据分析 正在成为有效项目管理、流程优化和战略决策的重要组成部分。建筑行业正在逐步掌握四个关键层次的分析，每个层次都能回答特定的问题并提供独特的优势（图 1.1-4）：

- 描述性分析 - 回答 "发生了什么？"的问题，并提供有关过去事件和结果的历史数据和报告：过去三年（2023-2025 年），在雨天浇筑 200 平方米的整体结构需要 4.5 到 6 天时间。
- 诊断分析--通过找出问题的原因来回答 "为什么会发生这种情况？"的问题：分析表明，由于阴雨天气，混凝土的养护过程减慢，整体结构的浇筑时间增加了
- 预测分析--前瞻性，通过回答 "会发生什么？"的问题来预测可能的风险和准备时间：根据历史数据，考虑到所有已知因素和趋势，预测在 2026 年的雨天浇筑一个类似的 200 平方米的整体结构需要大约 5.5 天。
- 规定性分析 - 提供自动建议并回答 "做什么？"的问题，使公司能够选择最佳行动：例如，为了优化施工，建议：使用特殊添加剂，加快高湿度条件下的混凝土养护；计划在降水概率最低的时段进行浇筑；为结构安排临时遮蔽物，即使在恶劣天气条件下，也能将施工时间缩短至 4-4.5 天。

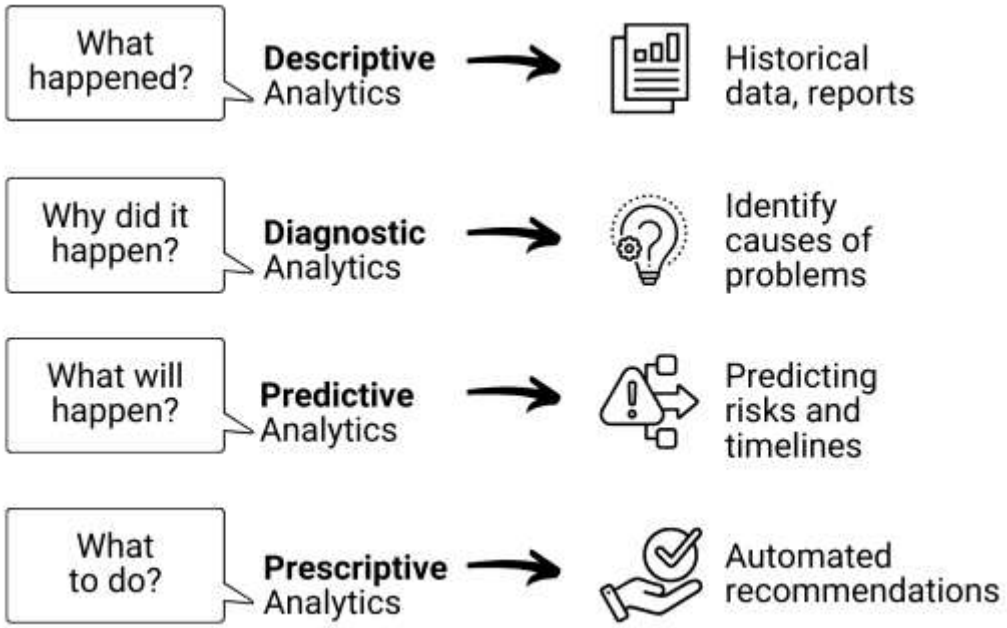


图 1.1-4 分析的主要类型： 从过去的描述到自动决策。

全面的数字化转型意味着向系统分析和数据驱动管理的过渡，这不仅需要外包，还需要组建一支有能力的内部团队。这种团队的主要成员应该是产品经理、数据工程师、分析师 和开发人员，他们将与业务部门密切合作（图 4.3-9）。这种合作对于提出智能分析问题和有效参数化业务决策任务是必不可少的。在信息社会中，数据已不仅仅是辅助工具，而是预测和优化的基础。

在建筑领域，数字化转型正在从根本上改变设施的设计、管理和运营方式。这一过程被称为信息数字化--施工过程的所有方面都被数字化为适合分析的数字形式。

施工过程信息数字化

千百年来，建筑中记录的信息量几乎没有变化，但近几十年来却迅速增长（图 1.1-5）。

> 根据普华永道的研究® 《管理数据。学生在瞬息万变的商业世界中取得成功所需的条件》（2015 年）[9]，世界上 90% 的数据是在过去两年中创建的（截至 2015 年）。然而，大多数公司并没有充分利用这些数据，因为这些数据要么被保存在孤立的系统中，要么只是被归档而没有进行真正的分析。

近年来，数据量的增长速度越来越快，从 2015 年的 15 ZB 翻了一番，到 2025 年将达到 181 ZB [10]。每天，建筑和设计公司的服务器上都充斥着项目文档、工作计划、计算和计算结果、财务报告。二维/三维 - 图纸使用 DWG 、DXF 和 DGN 格式，三维模型使用 RVT 、NWC 、PLN 和 IFC™ 格式。文本文件、表格和演示文稿保存为 DOC、XLSX 和 PPT 格式。除了来自施工现场的 MPG 和 JPEG 格式的视频和图像外，还有来自物联网组件、RFID® 标签（识别和跟踪）和 BMS 建筑管理系统（监测和控制）的实时数据

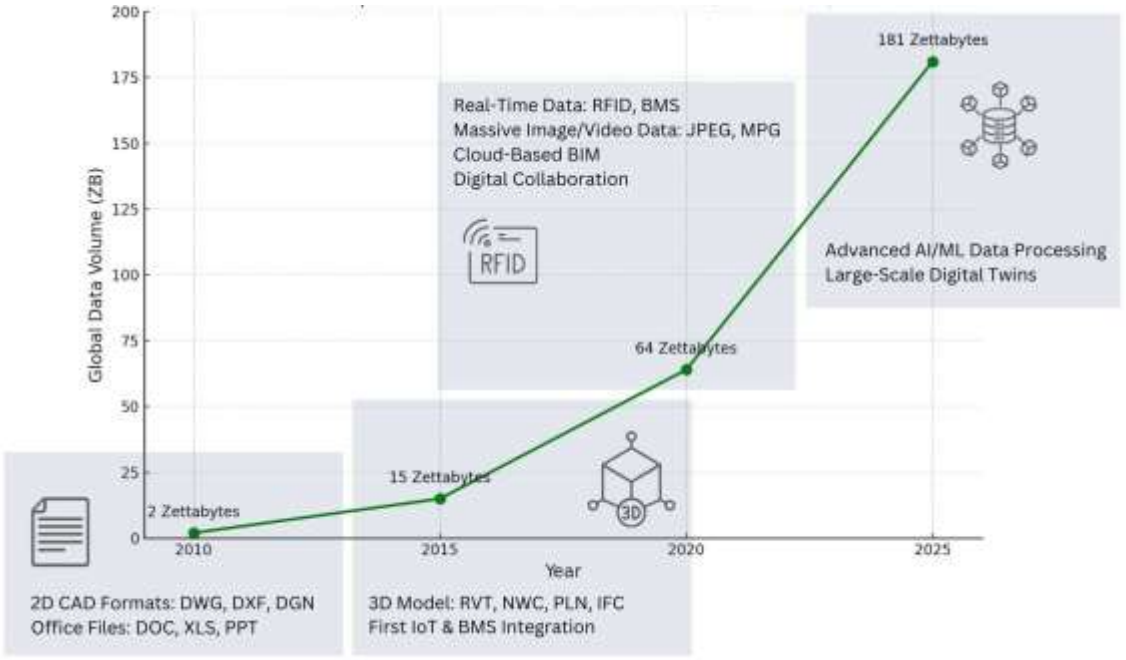


图 1.1-5 2010-2025 年抛物线数据增长（基于 [10]）。

随着信息的快速增长，建筑行业不仅需要收集和存储数据，还需要确保数据的验证、确认、可测量性和分析处理。如今，该行业正处于信息数字化的活跃阶段--将建筑活动的所有方面系统地转化为适合分析、解释和自动化的数字形式。

> 信息数字化意味着将建筑项目的所有实体和要素以及建筑过程本身的信息--包括那些我们以前根本不认为是信息的信息--转换成数据格式，使信息可以量化并易于分析。

就建筑业而言，这意味着要捕捉项目的所有要素和所有流程的信息并将数字化--从建筑工地上机械和人员的移动，到建筑工地上的天气和气候条件、当前的材料价格和中央银行利率--以便生成分析模型。

> 如果你能测量你所谈论的内容并用数字表达出来，那么你就对该主题有所了解。但是，如果你不能定量地表达它，你的知识就是极其有限的，不能令人满意的。它可能是一个起点，但不是真正的科学知识水平。[11].
>
> - 汤姆逊（开尔文勋爵），1824-1907 年，英国科学家

信息数字化远远超越了传统的信息收集方法，在传统方法中，只记录工时或实际材料成本等基本指标。如今，几乎任何事件都可以通过先进的分析工具和机器学习技术转化为适合深入分析的数据流。建筑行业经历了从纸质图纸、Excel 电子表格和口头指令到数字系统的根本性转变（图 1.2-4），在数字系统中，项目的每个元素都成为数据源。即使是员工，从工程师到现场施工人员，现在也被视为数字变量和数据集的集合。

> 根据毕马威会计师事务所的《熟悉的挑战--新方法：2023 年全球建筑业调查》，数字双胞胎、人工智能（AI）和大数据，正在成为提高项目盈利能力的主要驱动力[2]。

现代技术不仅简化了信息收集工作，使其在很大程度上实现了自动化，而且还从根本上降低了数据存储的成本。因此，企业不再采用选择性的方法，而是倾向于存储所有信息，以便日后进行分析（图 2.1-5），这为今后优化流程提供了潜在的机会。

> 信息数字化和数字化使我们有可能发掘信息中隐藏的、以前未曾开发的价值。如果组织得当，数据可以重新利用、重新解释并整合到新的服务和解决方案中。

未来，信息的数字化很可能导致文件管理的全面自动化，引入自我管理的施工流程，并出现新的职业-- 建筑数据分析师、人工智能项目管理专家和数字工程师。建筑项目将成为动态信息源，决策将不再基于直觉或主

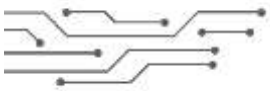
观经验，而是基于可靠且可复制的数字事实

> 信息是 21 世纪的石油，而分析 则是内燃机 [12]。
>
> - Peter Sondergaard，Gartner 高级副总裁®

> 根据物联网 Analytics 2024 [13]，全球数据管理和分析支出预计将从 2023 年的 1,855 亿美元大幅增长到 2030 年的 5,133 亿美元，复合年增长率为 16%。然而，并非所有组成部分都以相同的速度增长：分析 增长迅速，而存储的增长却在放缓。在数据管理生态系统中，分析的增长速度最快：预计将从 2023 年的 606 亿美元增长到 2030 年的 2279 亿美元，复合年增长率为 27%。

随着信息数字化进程的加快和信息量的快速增长，建筑项目和公司管理部门面临着系统存储、分析和处理各种数据（通常是异构数据）的需求。为了应对这一挑战，从 20 世纪 90 年代中期开始，建筑行业开始大规模转向电子文档的创建、存储和管理--从电子表格和设计计算到图纸和合同。

传统的纸质文件需要签名、实际存储、定期修订和存档，现在正逐渐被数字系统所取代，后者以结构化的方式将数据存储在专门应用程序的数据库中。



章 1.2.

现代建筑技术和管理系统

数字化革命和模块化 **MRP/ERP** 系统的出现

现代数字数据存储和处理时代始于 20 世纪 50 年代磁带的出现，它为存储和使用大量信息提供了可能。下一个突破是光盘驱动器的出现，它从根本上改变了建筑行业的数据管理方法。

随着数据仓库 的发展，大量公司进入解决方案市场，开始开发模块化软件，用于创建、存储、处理数据和自动执行日常任务

> 信息和工具的指数级增长促使人们需要集成式模块化解决方案，这种解决方案不仅能处理单个文件，还能帮助管理和控制跨流程和跨项目的数据流。

第一批综合平台工具不仅要存储文档，还要记录流程中的所有变更请求和操作：由 谁发起、请求的范围以及最终记录为值或属性的内容。为此，需要一个能够跟踪精确计算和决策的系统（图 1.2-1）。这种平台就是 20 世纪 90 年代初开始流行的 MRP （物料需求计划）和 ERP （企业资源计划）系统[14]

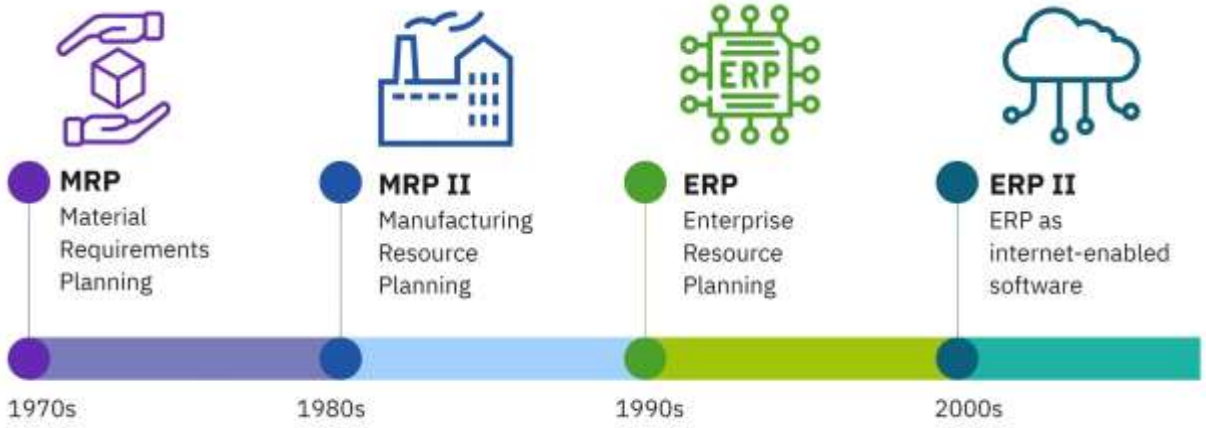


图 1.2-1 数据存储技术的进步导致了 20 世纪 80 年代企业资源规划 系统的出现。

最早的 MRP 和 ERP 系统为业务流程和施工项目管理的数字化时代奠定了基础。模块化系统最初是为实现关键业务流程自动化而设计的，最终与附加的、更加灵活和适应性更强的软件解决方案集成在一起。

这些附加解决方案是为数据处理和项目内容管理而设计的（图 1.2-2），它们或者取代了大型系统 的某些模块，或者对其进行了有效补充，扩展了整个系统的功能。

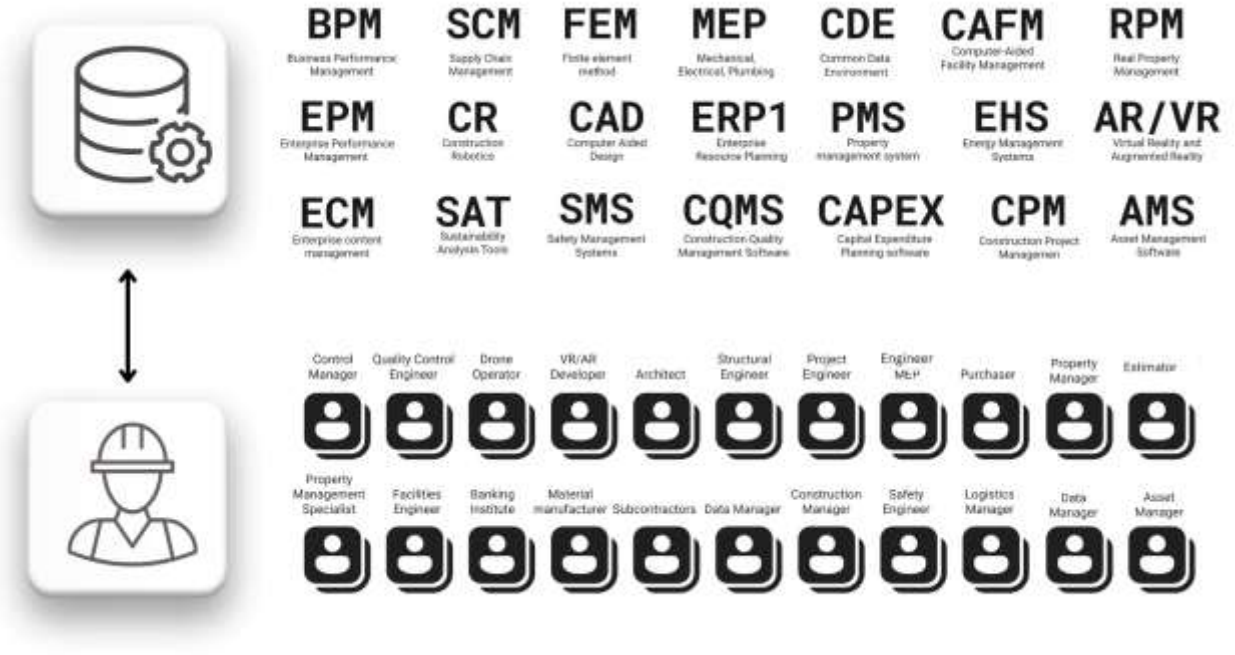


图 1.2-2 新的软件解决方案吸引了一大批管理人员进入企业管理数据流。

在过去的几十年里，企业对模块化系统进行了大量投资[15]，将其视为长期的综合解决方案。

> 根据 2022 年的 Software Path 报告[16]，ERP 系统每个用户的平均预算为 9,000 美元。平均而言，约有 26% 的公司员工使用此类系统。因此，对于拥有 100 名用户的企业来说，实施企业资源规划系统的总成本约为 90 万美元。

在灵活开放的现代技术飞速发展的背景下，投资于专有的、封闭的、模块化的解 决方案已越来越不合理。如果已经进行了这种投资，就必须客观地重新评估现有系统的作用：从长远来看，这些系统是否仍然必不可少，或者是否可以对其功能进行修改，并以更高效、更透明的方式加以实施。

当今模块化数据平台的一个主要问题是，它们将数据管理集中在封闭的应用程序中。因此，数据--公司的核心资产--变得依赖于特定的软件解决方案，而不是相反。这限制了信息的重复使用，使 迁移变得复杂，并降低了企业在快速变化的数字环境中的灵活性。

如果封闭式模块架构的相关性或实用性在未来可能会降低，那么将今天产生的成本视为沉没成本，并将重点

放在向更加开放、可扩展和适应性更强的数字生态系统的战略转变上，是非常有意义的。

> 专有软件的特点是，开发者独家控制源代码和在使用这些解决方案过程中产生的用户数据。与开放源码软件不同，用户不能访问应用程序的内部结构，也不能独立审查、修改或根据自己的需要进行调整。相反，用户必须购买许可证，才能在供应商规定的范围内使用软件。

以数据为中心的现代方法提供了一种不同的模式：数据应被视为一种重要的战略资产--独立、持久且独立于特定的软件解决方案。反过来，应用程序只是数据工具，可以自由更换，不会有丢失关键信息的风险。

20 世纪 90 年代，ERP 和 MRP 系统的发展（图 1.2-1）为企业提供了强大的流程管理工具，但也带来了意想不到的后果，即大大增加了参与维护信息流的人员数量。这些系统非但没有自动化和简化操作任务，反而常常造成新的复杂性、官僚主义和对内部 IT 资源的依赖。

数据管理系统：从数据挖掘到业务挑战

当今的公司面临着整合多个数据管理系统的需求。选择数据管理系统、明智地管理这些系统以及整合不同的数据源对业务绩效至关重要。

在 2020 年代中期，你可以找到数百个（在大型建筑公司则有数千个）不同的系统（图 1.2-3），这些系统必须协调工作，才能使施工过程的所有方面都顺畅、连贯地运行。

> 根据德勤 2016 年的研究® "数字资本项目中的数据驱动管理"--建筑专业人员平均每天使用 3.3 个软件应用程序，但其中只有 1.7 个相互集成[17]。

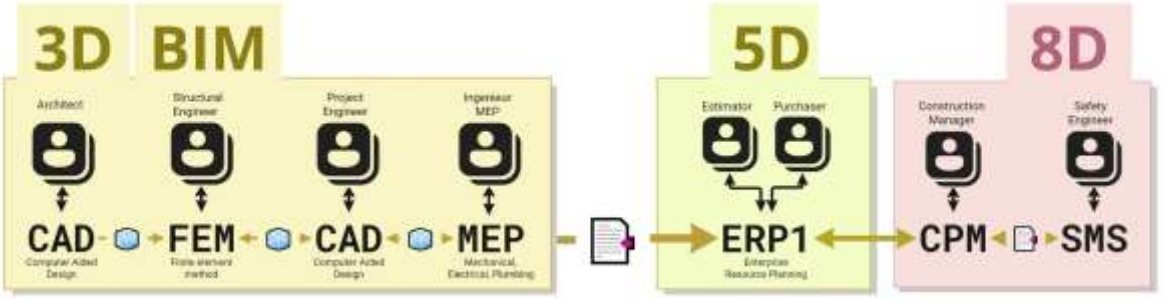


图 1.2-3 每个业务系统都需要一个专业团队和一个负责任的管理者来进行高质量的数据管理。

以下是建筑行业大中型企业在有效管理建筑项目时常用的系统列表：

- **ERP** （企业资源规划）--提供包括会计、采购和项目管理在内的业务流程集成。
- **CAPEX** （资本支出规划软件）- 用于编制预算和管理建筑项目的财务投资，有助于确定固定资产和长期资产投资的成本。
- **CAD** （计算机辅助设计）和 **BIM** （建筑信息建模）用于绘制详细准确的技术图纸和项目的三维 - 模型。这些系统的重点是处理几何信息。
- **MEP** （机械、电气、给排水）- 包括机械、电气和给排水组件的工程系统，并详细说明项目的内部 "循环 "系统。
- 地理信息系统（**GIS**）--用于地形分析和规划，包括制图和空间分析。
- **CQMS** （施工质量管理软件）--确保施工过程符合既定标准和规定，帮助消除缺陷。
- **CPM** （施工项目管理）--包括施工过程的规划、协调和控制。
- **CAFM** （计算机辅助设施管理）--楼宇管理和维护系统。
- 要优化供应商和施工现场之间的材料和信息流，就需要供应链管理（**SCM**）。
- **EPM** （企业绩效管理）--旨在改进业务流程和绩效。
- **AMS** （资产管理软件）--用于在整个资产生命周期内优化设备和基础设施的使用、管理和维护。
- **RPM** （不动产管理）--包括与建筑物和土地以及相关资源和资产的管理和运作有关的任务和流程。

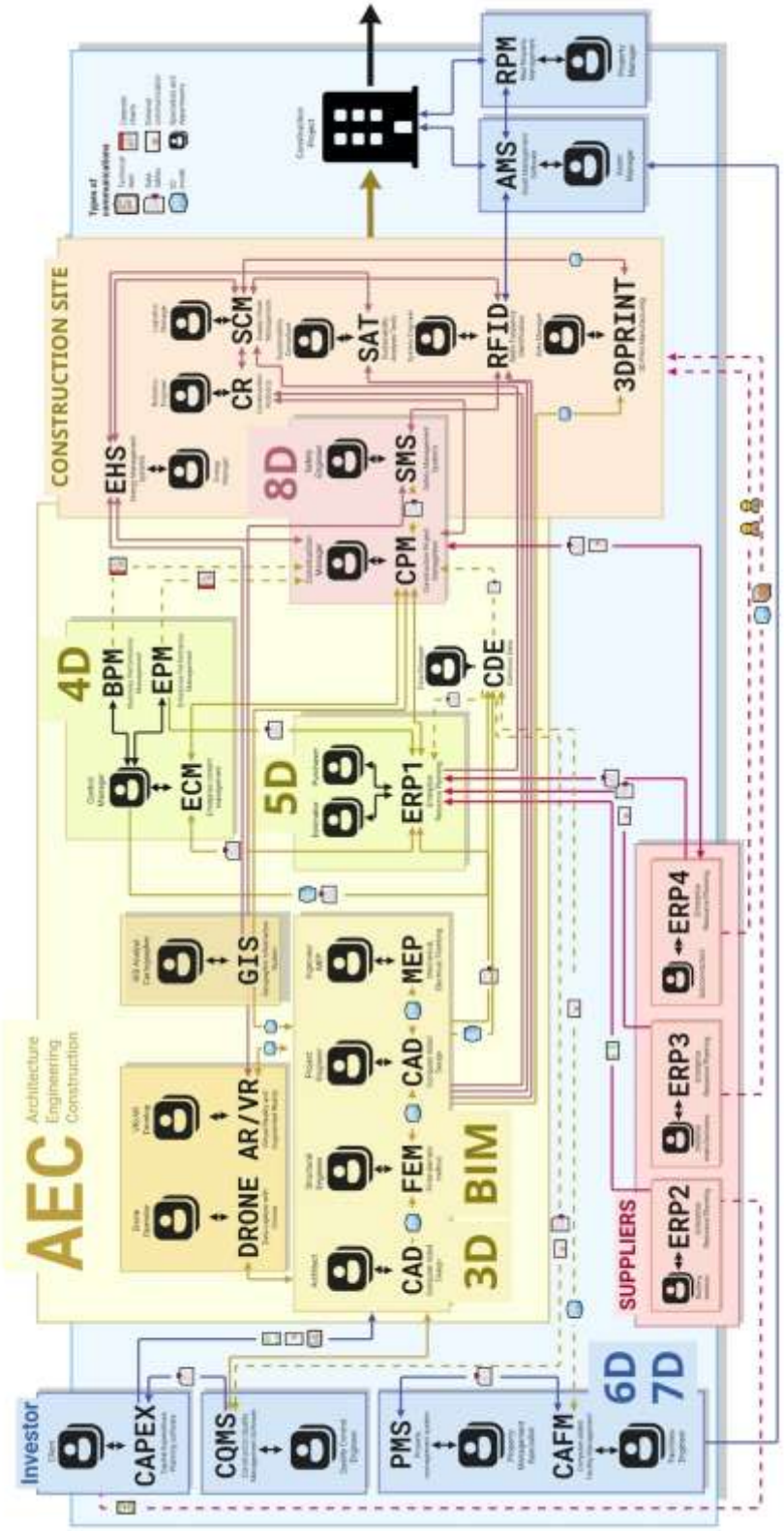


图 1.2-4 将公司流程与不同部门之间的信息流连接起来的系统互联性。

- **CAE**（计算机辅助工程）- 计算机辅助工程，包括计算和模拟系统，如有限元分析 **(FEA)** 和计算流体动力学 **(CFD)**。
- **CFD**（计算流体动力学） - 计算流体动力学，流体和气体流动建模。CAE 子类别。
- **CAPP**（计算机辅助流程规划）- 计算机辅助流程规划。它用于创建路线和流程图。
- **CAM**（计算机辅助制造）--计算机辅助制造，为数控机床生成控制程序。
- **PDM**（产品数据管理）--产品数据管理，一种存储和管理技术文档的系统。
- **MES**（制造执行系统）是一种实时制造过程控制系统。
- **PLM**（产品生命周期管理）--项目元素的生命周期管理，集成 PDM 、CAPP、CAM 和其他系统，实现从开发到报废的完整产品控制。

这些系统和许多其他系统结合了各种软件解决方案，已成为现代建筑业不可或缺的一部分（图 1.2-4）。这些系统的核心是带有直观界面的专业数据库，可在设计和施工的各个阶段提供高效的信息输入、处理和分析。数字工具之间的集成不仅有助于优化工作流程，还能显著提高决策的准确性，从而对项目实施的时间和质量产生积极影响。

但有一半的情况是没有集成。据统计，每两个应用程序或系统中就有一个与其他解决方案集成[17]。这表明数字环境仍在继续分裂，并强调有必要制定开放标准和统一接口，以确保建筑项目内端到端的信息交换。

现代企业在集成方面面临的主要挑战之一仍然是数字系统的高度复杂性，以及有效检索和解释信息所需的用户能力要求。在企业中实施的每个系统都需要一个以主要管理人员为首的专家团队提供支持（图 1.2-2）。

> 关键系统管理员对数据流的正确方向起着决定性作用，并对最终信息的质量负责，就像几千年前的第一批管理者对写在纸莎草纸或泥板上的数字负责一样。

要将分散的信息流转化为管理工具，必须具备系统整合和管理数据的能力。在这种架构中，管理人员必须充当单一网络中的元素，就像菌丝体一样，将公司的各个部分连接成一个能够适应和发展的连贯的生命体。

企业菌丝：数据如何连接业务流程

将数据整合到应用程序和数据库的过程是基于对不同来源信息的汇总，包括不同部门和专家的信息（图 1.2-4）。专家们搜索相关数据，对其进行处理，并将其传输到各自的系统和应用程序中供进一步使用。

每个公司系统都由一系列工具、技术和数据库组成，是一棵扎根于历史数据土壤的知识之树，并以成品解决方案的形式结出新的果实：文档、计算、表格、图表和仪表板（图 1.2-5）。公司的各个系统就像某片森林中的树木一样，相互影响、相互沟通，构成了一个由专业管理人员支持和管理的复杂而结构合理的系统。

> 公司的信息检索和传输系统就像一个复杂的森林网络，由树木（系统）和菌丝蘑菇（管理者）组成，它们起着导体和回收器的作用，确保信息传输并流向正确的系统。这有助于在公司内部保持健康、高效的数据流动和分配。

专家就像树根一样，在项目初始阶段吸收原始数据，将其转化为企业生态系统的养分。数据和内容管理系统（图 1.2-4 - ERP, CPM, BIM 等）是强大的信息高速公路，通过它，这些知识在公司的各个层面循环。

正如在自然界中，生态系统中的每个元素都发挥着各自的作用一样，在公司的业务环境中，每个流程参与者--从工程师到分析师 --都为信息环境的成长和肥沃做出了贡献。这些系统性的 "数据树"（图 1.2-5）不仅是收集信息的机制，也是确保公司可持续发展的竞争优势。

森林生态系统出人意料地准确反映了数字公司的组织结构。就像森林的分层结构一样--从灌木丛到树梢--企业管理将任务分配给各级责任人和职能部门。

> 扎根深厚、枝繁叶茂的树根可提供韧性和养分。同样，稳固的组织结构和稳定的优质数据工作流程可支持公司的整个信息生态系统，即使在（高风速）市场不稳定和危机时期，也能促进公司的可持续增长和发展。

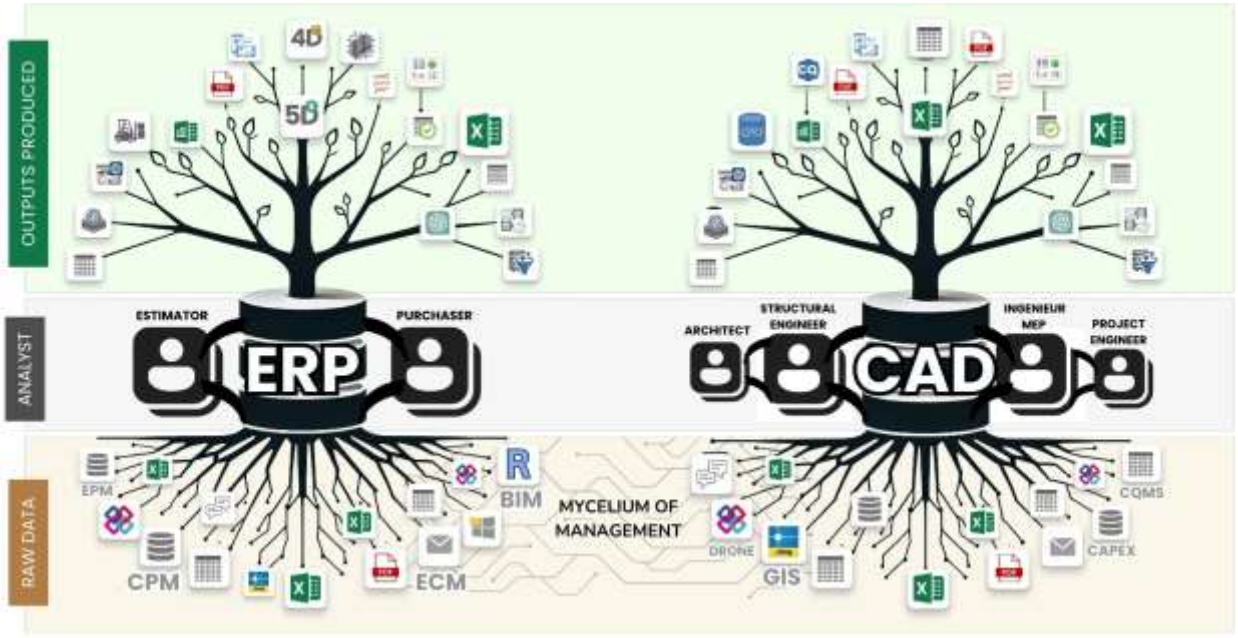


图 1.2-5 通过不同系统整合数据就像菌丝一样，将管理人员和专家连接成一个信息网络。

现代企业对规模的理解已经发生了变化。如今，一家公司的价值不仅取决于其可见部分--最终文件和报告形式的 "皇冠"，还取决于定性收集和系统处理数据的 "根系统 "的深度。收集和处理的信息越多，商业价值就越高。有条不紊地积累已处理数据的 "堆肥"，并能从中提取有用见解的公司将获得战略优势

> 历史信息正在成为一种新的资本，促进增长、优化流程和竞争优势。在数据驱动的世界里，获胜的不是谁拥有更多，而是谁知道更多。

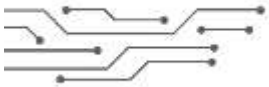
对于建筑行业来说，这意味着转向实时项目管理，从设计、采购到承包商协调的所有流程都将以每日更新的相关数据为基础。通过整合不同来源的信息（ERP - 系统、CAD - 模型、传感器 建筑工地上的物联网 、RFID ），可以做出更准确的预测，对变化做出快速反应，并避免因缺乏最新数据而造成的延误。

> 根据麦肯锡公司的《2025 年数据驱动型企业》（McKinsey & Company®, 2022 [18]），未来成功的企业将在其运营的所有关键方面依赖数据，从战略决策到运营互动。

数据将不再仅仅是一种分析工具， ，而将成为所有业务流程的组成部分，提供透明度、控制和自动化管理 。 下的数据驱动将使企业最大限度地减少人为因素的影响，降低运营风险，提高决策的透明度和效率。

21 世纪的经济模式正在发生翻天覆地的变化：过去，石油被称为 "黑金"，因为它能为机械和运输提供动力

，而如今，在时间压力的压缩下，历史数据正在成为一种新的战略资源，它不是机器的动力，而是推动业务发展的决策算法的动力。



章 1.3.

数字革命和数据爆炸

数据繁荣的开端是一次进化浪潮

建筑行业正在经历前所未有的信息爆炸。如果我们将企业视为一棵由数据喂养的知识树（图 1.2-5），那么当前的数字化阶段可以比作石炭纪时期植被的快速生长，在那个时代，地球生物圈因生物量的快速积累而发生了巨大变化（图 1.3-1）。

随着全球数字化的发展，建筑行业的信息量每年都在翻番。现代技术可以在后台收集数据，并对其进行实时分析和大规模使用，而这在短时间内似乎是不可能实现的。

> 根据戈登-摩尔（英特尔公司创始人之一）提出的摩尔定律，集成电路的密度和复杂性以及处理和存储的数据量大约每两年翻一番[19]。

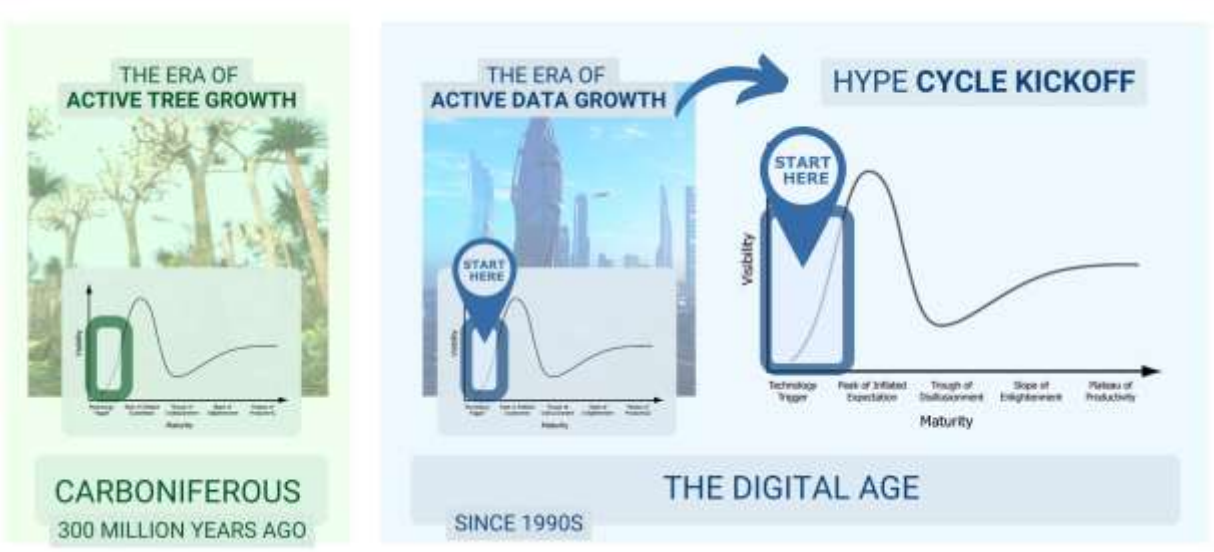


图 1.3-1 数字化的出现导致了数据的指数式增长，就像煤炭时代植被的激增一样。

土耳其的戈贝克利特佩（Göbekli Tepe）等古代巨石建筑没有留下适合重复使用的文献知识，而今天的数字技术使信息的积累和重复使用成为可能。这可以比作从孢子植物到种子植物（被子植物）的进化转变：种子的出现促成了生命在地球上的广泛传播。(图 1.3-2）。

同样，过去项目中的数据也成为一种 "数字种子"--DNA 知识载体，可在新项目和亲 duces 中扩展和使用。现代人工智能工具--机器学习和大型语言模型（LLMs），如 ChatGPT、LlaMa、Mistral、Claude、DeepSeek、QWEN、Grok--的出现使得数据可以自动提取、解释并应用于新的语境中

> 正如种子彻底改变了最初没有生命的星球上的生命传播一样，"数据种子 "正在成为自动产生新信息结构和知识的基础，使数字生态系统能够独立发展，适应不断变化的用户需求。

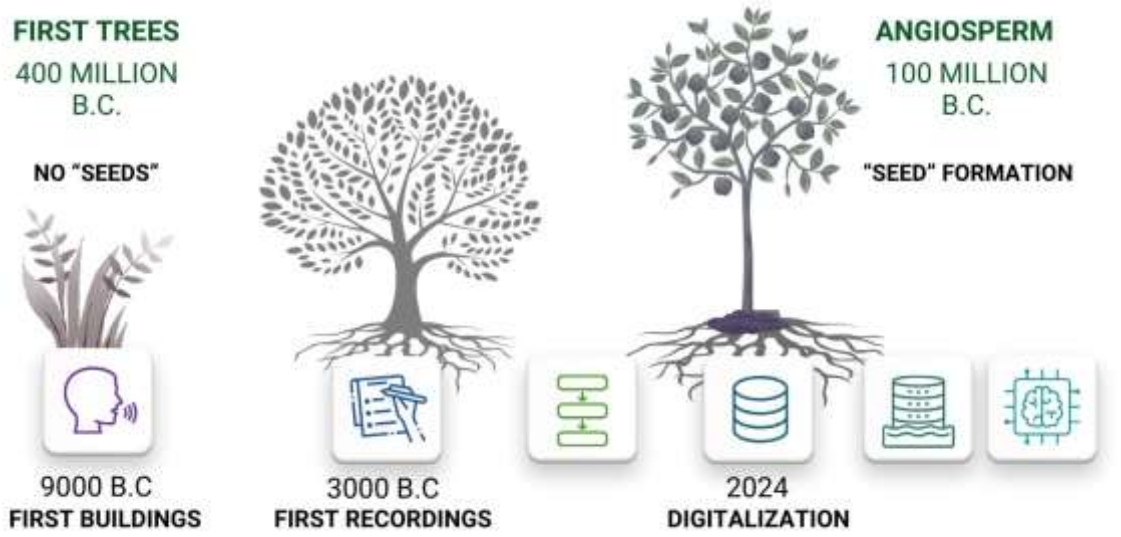


图 1.3-2 数字 "数据种子 "在进化过程中扮演着与被子植物相同的角色，被子植物是改变地球生态系统的开花植物。

我们正站在建筑业新时代的门槛上，数据爆炸和 "数据种子"（来自过去和正在进行的项目的结构化信息）的积极传播正在为建筑业的数字化未来奠定基础。通过大数据语言模型（LLMs）对其进行 "授粉"，我们不仅可以观察数字变化，还可以积极参与创建自学习、自适应的生态系统。这不是进化，而是一场数字革命，在这场革命中，数据正在成为新现实的主要组成部分

> 在建筑项目的整个生命周期中，来自不同学科的信息使建筑行业的数据量急剧增加。这种巨大的数据积累将建筑行业推向了大数据时代[20]。
>
> - 杨杭教授，武汉理工大学土木工程与建筑系，武汉，

信息时代的数据增长让人联想到自然界的进化过程：正如森林的发展改变了地球的古老面貌一样，当前的信息爆炸正在改变整个建筑行业的面貌。

现代公司产生的数据量

在过去两年中，全球现有数据的 90% 都是创建的[21]。截至 2023 年，包括建筑业专业人员在内的每个人每秒产生的数据量约为 1.7 兆字节[22]，2023 年全球数据总量将达到 64 兆字节，预计到 2025 年将超过 180 兆字节，即 180*10^15 兆字节[23]。

> 这种信息爆炸有一个历史先例--约翰内斯-古腾堡（Johannes Gutenberg）在 15 世纪发明了印刷机。印刷机问世后仅 50 年，欧洲的书籍数量就翻了一番：在几十年的时间里，印刷的书籍数量相当于过去 1200 年中手工制作的书籍数量[24]。今天，我们正目睹着更加迅猛的增长：世界上的数据量每三年翻一番。

> 鉴于目前的数据增长速度，建筑业在未来几十年内产生的信息量有可能相当于其以往整个历史所积累的信息量

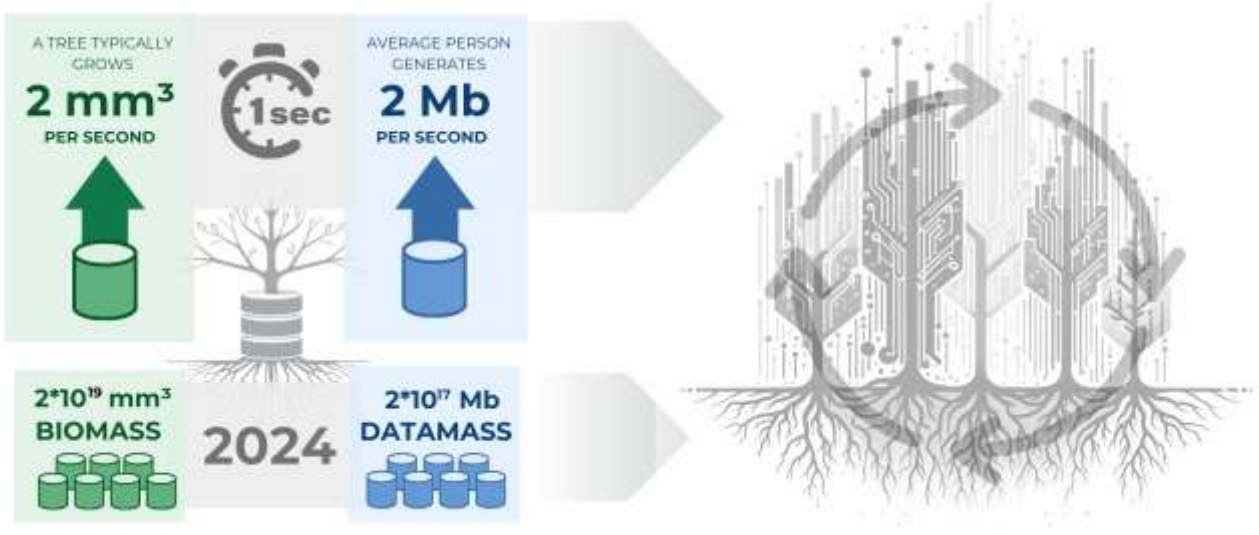


图 1.3-3 每位员工每天在公司服务器上存储数据，导致数据不断增长。

在当今的建筑商业世界中，即使是小公司每天也会产生大量的多格式信息，从模型和图纸到现场的照片记录和传感器，即使是一家小型建筑公司的数字足迹每天也可达数十 GB。如果我们假设每个技术人员平均每秒产生约 1.7 MB 的数据，这相当于每天约 146 GB，或每年 53 TB（图 1.3-3）。

如果一个 10 人团队每天只工作 3 个小时，那么每天产生的信息量累计可达 180 千兆字节（图 1.3-4）。

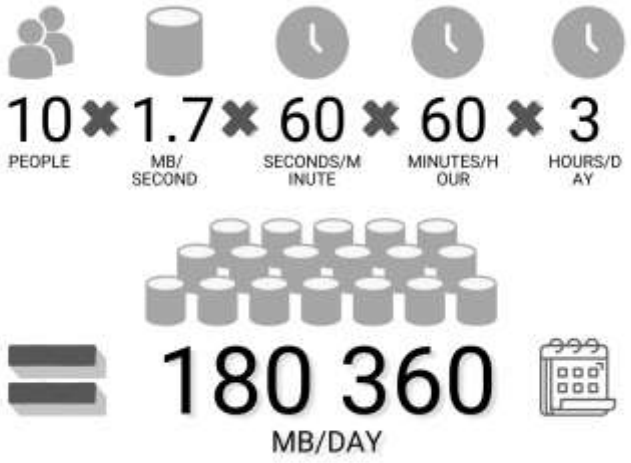


图 1.3-4 一家 10 人公司每天大约产生 50-200 千兆字节的数据。

假设 30% 的工作数据是新数据（其余数据被覆盖或删除），一家 10 人公司每月可创建数百千兆字节的新数据（实际数字取决于公司的业务类型）

由此可见，我们不仅在产生越来越多的数据，而且对数据的高效管理、存储和长期可用性的需求也在不断增长。以前，数据可以免费 "躺 "在本地服务器上，而在数字化转型的背景下，越来越多的公司开始使用云解决方案作为其信息基础设施的基础。

数据存储的成本：经济方面

近年来，越来越多的公司将数据存储外包给云服务。例如，如果一家公司将其一半的数据托管在云端，以平均每月每千兆字节 0.015 美元的价格计算，其存储成本每月可能会增加 10-50 美元[25]。

对于具有典型数据生成模式的小型公司来说，云存储成本在 几年内可能从每月数百美元到可能超过千美元

不等（图 1.3-5），这可能会造成巨大的财务负担。

> Forrester 的研究 "复杂性增加，企业外包数据存储"[26]对 214 位技术基础设施决策者进行了调查。[26]，超过三分之一的企业正在外包存储，以应对日益增长的数据业务量和复杂性，其中近三分之二的企业倾向于采用基于订阅的模式。



图 1.3-5 将数据迁移到云端可使每月存储成本增加高达 2000 美元，即使是只有 10 名员工的公司也是如此。

CAD (BIM)、CAFM 、PMIS 和 ERP 等云技术的加速应用进一步增加了数据存储和处理成本。因此，企业不得不寻找优化成本和减少对云提供商依赖的方法。

2023 年以来，随着大型语言模型（LLM ）的积极发展，数据存储的方法开始发生变化。越来越多的公司开始考虑收回对数据的控制权，因为在自己的服务器上处理信息变得更加安全和有利可图。

在这种情况下，从基于云的存储和处理必要数据转向本地部署企业 LLM 和人工智能 解决方案的趋势凸显出来。正如微软公司首席执行官在一次访谈中指出的那样[27]，人工智能代理将管理数据库中的流程，自动执行不同系统的功能，而不是依靠几个单独的应用程序或基于云的 SaaS 解决方案来执行不同的任务。

> [.....]解决这个[数据处理]问题的旧方法是：如果你回想一下不同的业务应用程序是如何处理集成的，它们使用的是连接器。公司出售这些连接器的许可证，商业模式就是围绕这一点形成的。SAP [ERP] 就是一个典型的例子：只有拥有合适的连接器，才能访问 SAP 数据。因此，在我看来，[人工智能]代理互动[.....]也会出现类似的情况。至少我们采取的方法是：我认为，在[人工智能]代理时代，业务应用程序存在的概念可能会崩溃。因为仔细想想，它们本质上就是带有大量业务逻辑的数据库
>
> - 微软首席执行官萨提亚-纳德拉（Satya Nadella），2024 年接受 BG2 频道采访。[28]

在这种模式下，数据驱动的 LLM 方法超越了传统系统。人工智能成为用户和数据之间的中介（图 2.2-3，图 2.2-4），无需多个中介界面，提高了业务流程的效率。我们将在 "化混乱为有序，降低复杂性 "一章中详细介绍这种处理数据的方法。

虽然未来的架构仍在形成之中，但企业已经面临着过去决策所带来的后果。近几十年来的大规模数字化，伴随着不同系统的引入和无节制的数据积累，导致了一个新的问题--信息超载。

数据积累的前沿：从质量到意义

当数据量和应用程序数量与信息技术部门和管理人员的能力相平衡时，现代公司系统就能在可控增长的情况下成功发展和运行。然而，近几十年来，数字化导致了数据量和复杂性的无节制增长，给企业的信息生态系统带来了过饱和效应。

如今，服务器和存储设备正承受着前所未有的大量未经处理和多格式化的信息，这些信息来不及变成肥料，就迅速变得无关紧要。有限的公司资源无法应对这股洪流，数据堆积在孤立的筒仓（即所谓的 "筒仓 "）中，需要人工处理才能提取有用的信息。

> 因此，现代公司管理系统就像长满常春藤和霉菌的森林一样，经常出现信息超载的问题。企业生态系统的核心不是滋养信息的腐殖质，而是形成了不同格式数据的孤立区域，这不可避免地导致业务流程整体效率的降低。

在过去 40 年中，数据呈指数级增长的漫长时期将不可避免地进入饱和及随后的冷却阶段。当存储达到极限时，将发生质的转变：数据将不再仅仅是一种存储对象，而是一种战略资源。

随着人工智能和机器学习的发展，企业有机会降低信息处理成本，并从数据的定量增长转向定性使用。未来十年，建筑行业必须将重点从创建越来越多的数据转向确保数据的结构、完整性和分析价值。

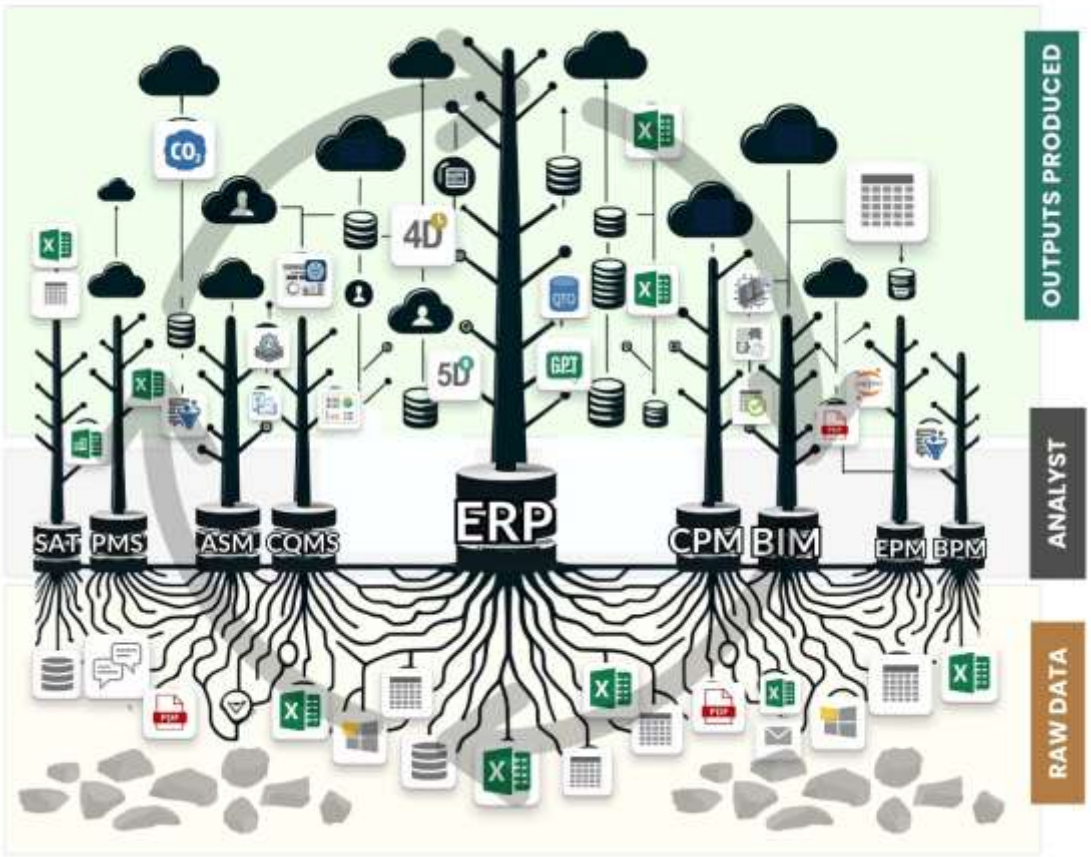


图 1.3-6 孤立的数据源阻碍了数据系统之间的信息共享。

> 主要价值不再在于信息量，而在于自动解读信息并将其转化为对管理决策有用的应用知识的能力。为了使数据真正有用，必须对其进行适当的管理：根据具体的业务任务进行收集、验证、结构化、存储和分析。

公司的数据分析流程类似于森林中树木的生老病死和新的年轻力壮的树木的出现：成熟的树木死亡、腐烂，并成为新的树木生长的滋养环境。已完成和已完成的流程在完成后会成为公司信息生态系统的一部分，最终成为信息腐殖质，为新系统和新数据的未来发展提供动力。

然而，在实践中，这种循环往往被打破。新系统没有得到有机的更新，而是像地质层一样，在旧系统的基础上层层叠加，没有进行深入的整合和结构化。结果，出现了互不关联的信息 "孤岛"，阻碍了知识的流通，使数据管理复杂化。

下一步：从数据理论到实际变革

建筑数据的发展历程是从泥板到现代模块化平台的过程。当今的挑战不是收集信息，而是创建一个框架，将不同的数据转化为战略资源。无论您的角色是公司经理还是工程师，了解数据的价值以及如何使用数据都将是未来的一项关键技能。

在总结这部分内容时，值得强调的是主要的实际步骤，这些步骤将有助于您在日常工作中应用所讨论的方法：

- 对信息流进行个人审计
  - [ ] 列出你日常使用的所有系统和应用程序清单
  - [ ] 标出您花最多时间搜索或重复检查数据的地方
  - [ ] 确定关键信息来源
  - [ ] 分析当前的应用程序状况，查找冗余和重复功能
- 努力按照分析成熟度的等级来推进流程
  - [ ] 从描述性分析开始您的任务 （发生了什么？）

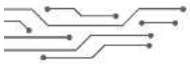
- ☐ 逐步引入诊断（为什么会发生这种情况？）
- ☐ 思考如何在流程中转向预测性分析（会发生什么？

■ 开始构建工作数据

- ☐ 为工作中常用的文件和文件夹实施统一的命名系统
- ☐ 为常用文件和报告创建模板
- ☐ 定期将已完成的项目归档，并建立清晰的结构

即使无法改变团队或公司的整个信息基础设施，也可以从自己的流程和日常工作中的微小改进开始。请记住，数据的真正价值不在于其数量，而在于从中提取可行见解的能力。即使是规模较小但结构和分析得当的数据集，在整合到决策流程中后也能产生重大影响。

在本书接下来的部分中，我们将介绍处理数据的具体方法和工具，探讨将非结构化信息转化为结构化信息的方法，探索分析自动化技术，并详细介绍如何在建筑公司建立有效的分析生态系统。





II 部分

建筑业如何淹没在混乱的数据中

第二部分专门批判性地分析了建筑公司在处理日益增长的数据量时所面临的挑战。详细分析了信息碎片化的后果和 "孤岛数据 "现象，这种现象阻碍了有效决策。研究了 HiPPO -approach（最高薪酬人员意见）的问题及其对建筑项目管理决策质量的影响。评估了动态业务流程及其日益增长的复杂性对信息流和运营效率的影响。举例说明了过度复杂的系统如何增加成本和降低组织的灵活性。特别关注了专有格式造成的限制以及在建筑行业使用开放标准的前景。 ，将过度复杂性和技术障碍降至最低。

章 2.1.

数据分散和孤岛

工具越多，业务效率就越高?

乍看起来，数字化工具越多，效率就越高。但实际上并非如此。每推出一种新的解决方案，无论是云服务、传统系统还是 Excel 报告，公司的数字化版图就会增加一层，而这一层往往无法与其他层集成（图 2.1-1）。

数据可以比作煤炭或石油：它需要数年的时间才能积累起来，在混乱、错误、非结构化流程和被遗忘的格式的层层压迫下被压缩。要从中提取真正有用的信息，企业必须从层层叠叠的过时解决方案和数字噪音中跋涉。

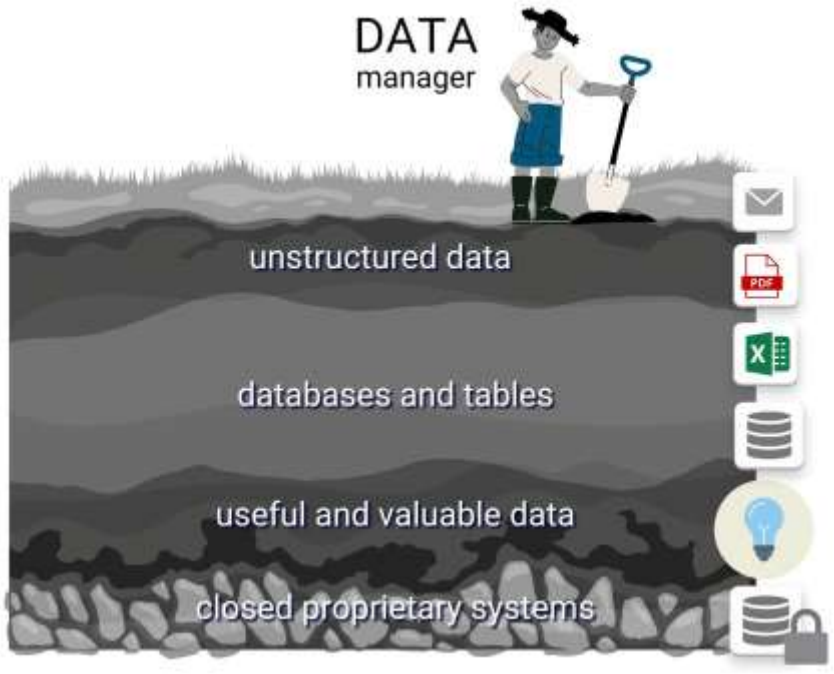


图 2.1-1 多样化的数据形成条块分割的层次--即使是 "黄金 "洞察力也会在系统复杂性的地质岩石中丢失。

每个新的应用程序都会留下痕迹：一个文件、一个表格或服务器上的整个孤立的 "筒仓"。一层是粘土（过时

和被遗忘的数据），另一层是沙子（不同的表格和报告），第三层是花岗岩（无法集成的封闭式专有格式）。随着时间的推移，公司的数字环境越来越像一个不受控制的信息积累库，价值在公司服务器的深处流失。

随着每个新项目和新系统的推出，不仅基础设施变得更加复杂，而且获取有用优质数据的途径也变得更加复杂。要获得有价值的 "岩石"，就需要对信息进行深度清理、结构化、"分块"、将其归类为有意义的块，并通过分析和数据建模提取具有战略意义的见解。

> 数据是有价值的东西，它比[处理数据的]系统本身更持久[29]。
>
> - 蒂姆-伯纳斯-李，万维网之父，第一个网站的创建者

在数据成为 "有价值的东西 "和可靠的决策依据之前，必须对其进行精心准备。只有经过适当的预处理，才能将分散的数据转化为结构化的经验和有用的信息腐殖质，进而成为预测和优化工具。

> 有一种误解是，你需要完全干净的数据才能开始分析，但实际上，能够处理脏数据是分析过程中必不可少的一部分。

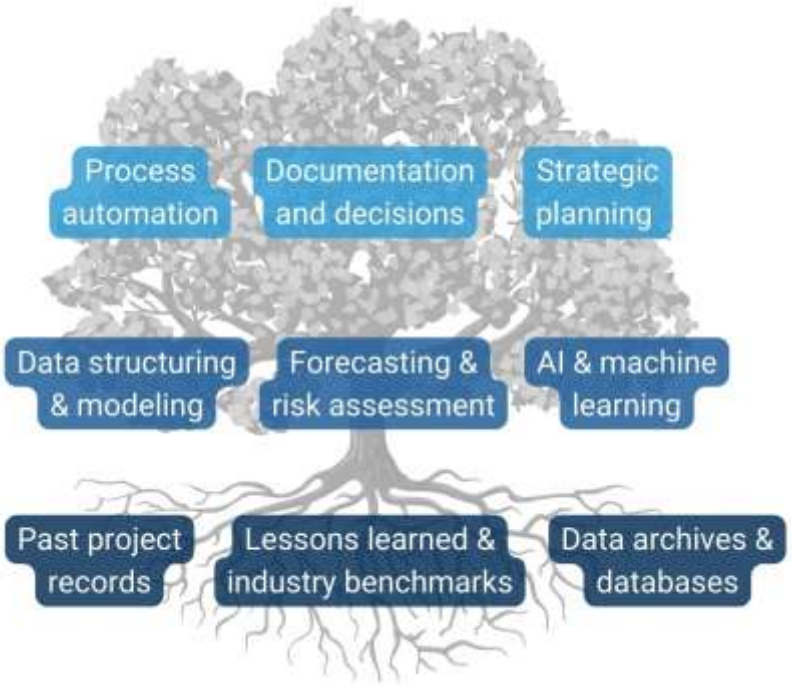


图 2.1-2 数据是业务的根系和基础，而业务又以决策过程为基础。

随着技术的不断进步，您的企业也必须向前迈进，学习如何从数据中创造价值。正如石油和煤炭公司正在建设开采矿物的基础设施一样，企业也必须学会如何管理自己服务器上的新信息流，并从闲置、未格式化和过时的数据中提取有价值的见解，将其转化为战略资源。

创建字段（数据仓库 ）是第一步。如果企业继续在孤岛式系统中运行，即使是最强大的工具也无法解决数据隔离和多格式数据的问题。如果数据彼此独立，没有交叉和共享信息，企业就会面临 "数据孤岛 "效应。企业不得不花费资源来合并和同步数据，而不是使用单一、一致的基础设施。

数据孤岛及其对公司业绩的影响

想象一下，你正在建造一个住宅区，但每个团队都有自己的项目。有的人在砌墙，有的人在铺设通信线路，还有的人在铺设道路，却没有相互核对。结果，管道与墙壁上的开口不匹配，电梯井与楼层不一致，道路不得不拆除并重新铺设。

这种情况并不只是一种假设，而是许多现代建筑项目的现实。由于使用不同系统的总承包商和分包商数量众多，又没有一个统一的协调中心，整个过程变成了一系列无休止的审批、返工和冲突。所有这些都会导致严

重的延误和项目成本的增加。

建筑工地上的一个典型情况很简单：模板已经准备就绪，但钢筋却没有按时运到。在检查各种系统的信息时，沟通情况大致如下：

- 20 日，建筑工地的工头写信给项目经理："我们已经完成了模板的安装，钢筋在哪里？
- 项目经理（PMIS）致采购部： - "模板已经准备好了。在我的系统 [PMIS] 中，钢筋应该在 18 日到达。钢筋在哪里？
- 供应链专家（ERP）： - 我们的企业资源规划系统说 25 日交货。
- 数据工程师或 IT 部门（负责集成）： - 在 PMIS 中，日期是 18 日，而在 ERP 中是 25 日。ERP 和 PMIS 之间没有订单 ID 链接，因此数据不同步。这是一个典型的信息差距例子。
- 项目经理致总经理--"配件交付延迟，工地停工，责任人不明确"。

事件的起因是不同系统中数据的隔离。通过整合和统一数据源，创建单一的信息库，并通过 ETL 工具（Apache NiFi、Airflow 或 n8n）实现自动化，可以消除系统间的孤岛。本书后面的章节将详细讨论这些方法和工具。

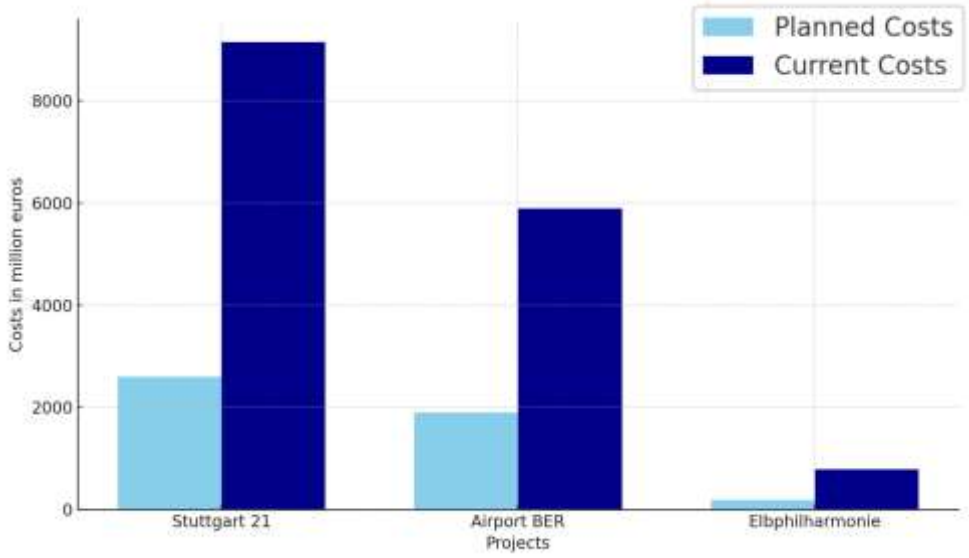


图 2.1-3 德国主要基础设施项目计划成本和实际成本的比较。

同样的情况也发生在企业系统中：先是创建了孤立的解决方案，然后又不得不花费巨额预算来整合和协调这些解决方案。如果从一开始就考虑到数据和通信模式，就根本不需要整合。孤立的数据会给数字世界带来混

乱，就像不协调的施工过程一样。

> 毕马威会计师事务所的 2023 年研究报告《提示建设 4.0：成败在此一举》指出，只有 36% 的公司在各部门之间有效共享数据，61% 的公司因孤立的数据 "孤岛 "而面临严重问题[30]。

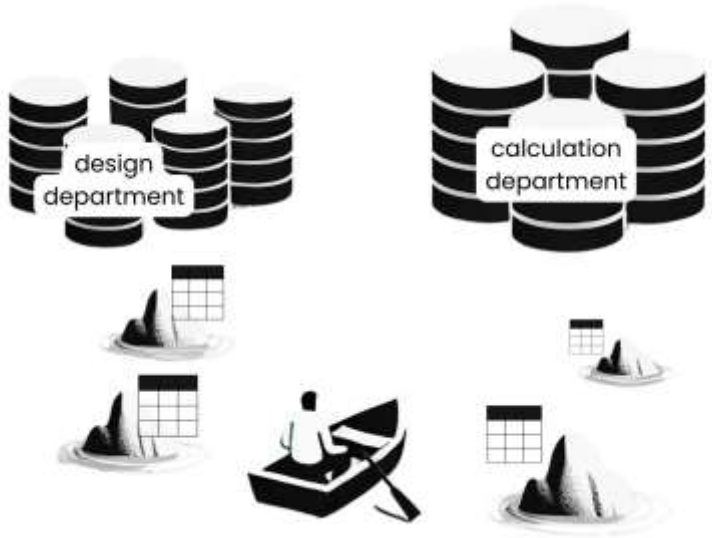


图 2.1-4 多年来难以收集的数据积累在孤立的存储 "筒仓 "中，有可能永远不会被使用。

公司数据存储在孤立的系统中，就像一棵棵散落在各地的树木。每一棵树都包含有价值的信息，但它们之间缺乏联系，无法形成一个互联互通的生态系统。这种各自为政的情况阻碍了数据的流动，限制了企业了解全局的能力。连接这些孤岛是一个极其漫长而复杂的过程，需要在管理层面培育蘑菇菌丝，学习如何在系统之间传输单个信息。

根据 2016 年世界经济论坛的一项研究，数字化转型的主要障碍之一是缺乏通用数据标准和数据分散。

> 建筑业是世界上最分散的行业之一，有赖于价值链上所有参与者的顺利互动 [5]。
>
> - 2016 世界经济论坛：塑造的未来

设计人员、管理人员、协调人员和开发人员往往喜欢自主工作，避免复杂的协调工作。这种自然倾向导致信息 "孤岛 "的产生，数据被隔离在不同的系统中。这种孤立的系统越多，让它们协同工作就越困难。随着时间的推移，每个系统都有自己的数据库和专门的管理人员支持部门（图 1.2-4），使整合工作更加复杂。

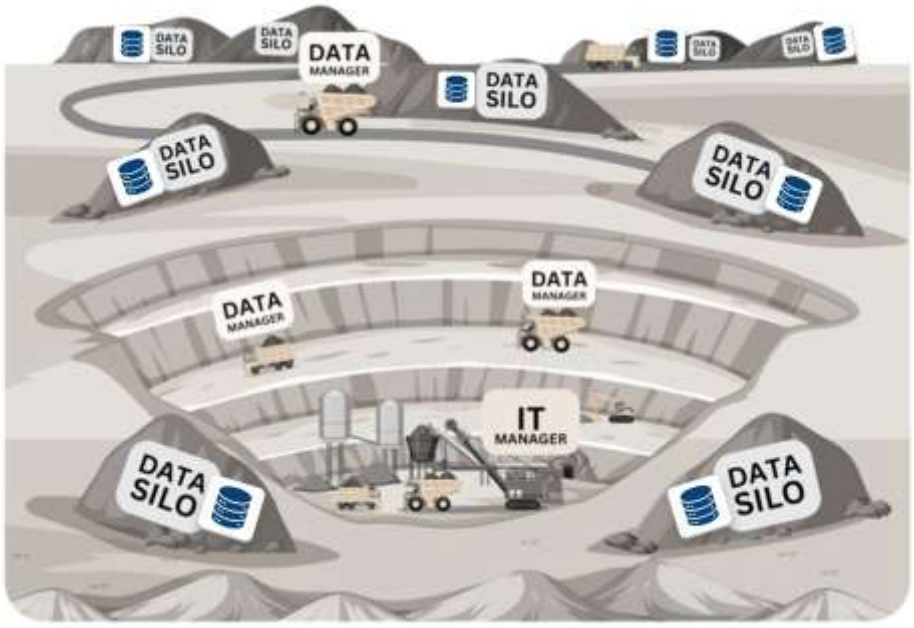


图 2.1-5 每个系统往往会创建自己独特的数据筒仓，需要用合适的工具来处理 [31]。

企业系统的恶性循环是这样的：企业投资于复杂的 ISO leated 解决方案，然后面临高昂的集成成本，而开发人员意识到系统组合的复杂性，宁愿在封闭的生态系统中工作。所有这些都加剧了 IT 环境的分散性，增加了迁移到新解决方案的难度（图 2.1-5）。管理人员最终会批评数据孤岛，却很少分析其原因以及如何加以预防。管理人员抱怨 IT 系统过时，但更换系统需要大量投资，而且很少能取得预期效果。因此，即使试图解决这个问题，也往往会使情况变得更糟。

造成脱节的主要原因是应用优先于数据。公司首先开发独立的系统或从供应商处购买现成的解决方案，然后试图通过创建重复且不兼容的存储和数据库来统一这些系统。

要克服各自为政的问题，就必须采取一种全新的方法--优先考虑数据而不是应用程序。企业必须首先制定数据管理战略和数据模型，然后建立系统或购买解决方案，以处理单一信息集，而不是制造新的障碍。

> 我们正在进入一个数据可能比软件更重要的新世界。
>
> - Tim O'Reilly，O'Reilly 媒体公司首席执行官

> 麦肯锡全球研究院的研究报告《反思建筑业：提高生产力之路》（2016 年）表明，建筑业在数字化转型方面落后于其他行业[32]。根据该报告，采用自动化数据管理和数字平台可显著提高生产率，减少因流程不一致而造成的损失。Egan（英国，1998 年）的报告[33]也强调了数字化转型的必要性，该报告强调了集成流程和协作方法在建筑业中的关键作用。

> 因此，在过去的一万年里，数据管理人员面临的主要问题是缺乏数据，而随着数据和数据管理系统的大量涌现，用户和管理人员面临着一个问题--数据过多，难以找到合法正确和高质量的信息。

数据孤岛不可避免地会导致数据质量下降的严重问题。在多个独立系统中，相同的数据可能存在不同的版本，其数值往往相互矛盾，这给需要确定哪些信息是相关和可靠的用户带来了更多的困难。

重复和缺乏数据质量 因为不统一的结果

由于数据孤岛问题，管理人员不得不花费大量时间搜索和核对数据。为了避免出现质量问题，公司建立了复杂的信息管理结构，由垂直的管理人员负责搜索、核实和核对数据。然而，这种做法只会增加官僚主义，减缓决策速度。数据越多，分析和解释的难度就越大，尤其是在没有统一的数据存储和处理标准的情况下。

近十年来，软件应用程序和系统如雨后春笋般不断涌现，对于最终用户来说，孤岛和不适当的数据质量 问题变得越来越重要。现在，在不同的系统和应用程序中可以找到相同的数据，但数值却不同（图 2.1-6）。这给终端用户在众多可用数据中确定哪个版本是相关和正确的数据带来了困难。这就导致了分析和最终决策的错误。

为了防止出现找不到正确数据的问题，公司经理们建立了一个多层次的核查经理官僚机构。他们的任务是能够以表格和报告的形式，在不同系统的迷宫中快速查找、检查和发送所需的数据。

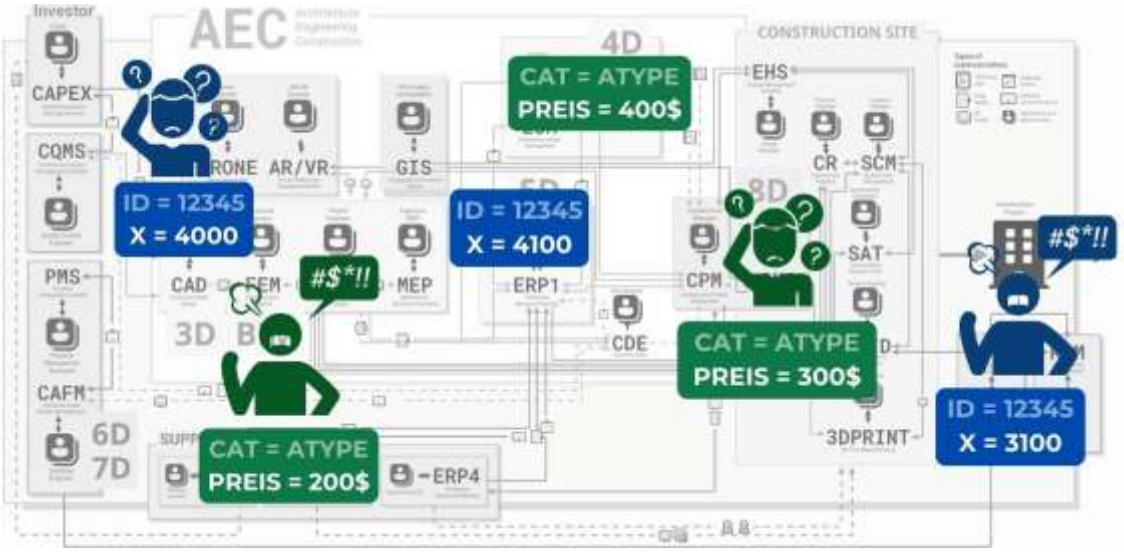


图 2.1-6 在试图找到正确的数据时，管理人员必须确保不同系统之间数据的质量和法律可靠性。

然而，在实践中，这种模式产生了新的复杂性。当数据由人工管理，信息分散在许多不相关的决策中时，通过金字塔形的决策者（图 2.1-7）获取准确和最新信息的每一次尝试都会成为一个瓶颈，既耗时又容易出错。

数字化解决方案的大量涌现加剧了这一局面。软件市场继续充斥着看似很有前途的新工具。但是，，如果没有明确的数据管理战略，这些解决方案就无法集成到一个统一的系统中，反而会增加复杂性和重复性。结果，企业不但没有简化流程，反而陷入了更加分散和混乱的信息环境。

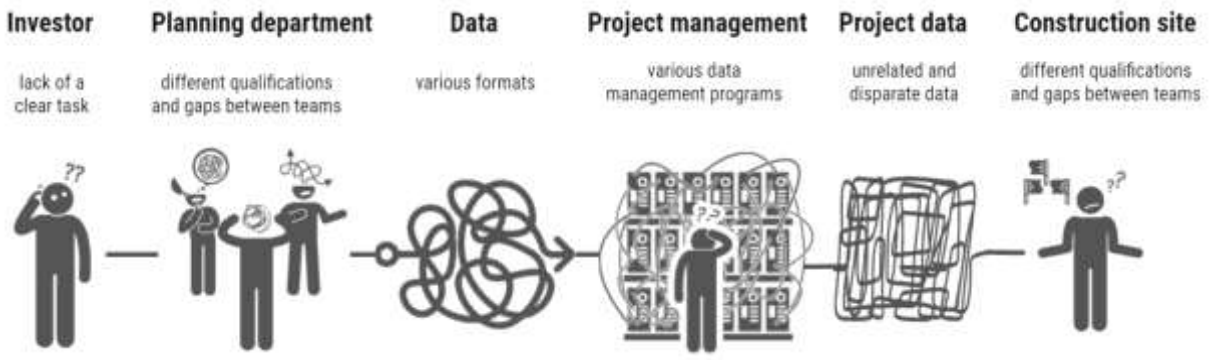


图 2.1-7 系统的复杂性和数据格式的多样性导致施工过程失去一致性。

与管理众多不同解决方案相关的所有这些问题迟早会让公司管理层意识到一个重要问题：问题不在于数据量，也不在于寻找下一个 "放之四海而皆准 "的数据处理工具。真正的原因在于数据的质量以及企业如何创建

、接收、存储和使用数据。

> 持续成功的关键不在于追逐新的 "神奇 "应用，而在于在公司内部建立数据文化。这意味着要将数据视为一种战略资产，并将数据质量、完整性和相关性作为企业各个层面的优先事项。

解决质量与数量两难问题的办法在于建立一个统一的数据结构，以消除重复、不一致和统一信息流。这种结构提供了单一、可靠的数据源，可据此做出明智、准确和及时的决策。

否则，就会像现在经常出现的情况一样，企业继续依赖 HiPPO 专家的主观意见和直观评估，而不是可靠的事实。在传统上专业知识发挥重要作用的建筑行业，这种情况尤为明显。

HiPPO 或决策中的意见危险

传统上，建筑行业的关键决策都是根据经验和主观判断做出的。由于缺乏及时可靠的数据，公司管理人员不得不盲目行事，依靠薪酬最高员工的直觉（HiPPO - 最高薪酬人员的意见）而非客观事实（图 2.1-8）。

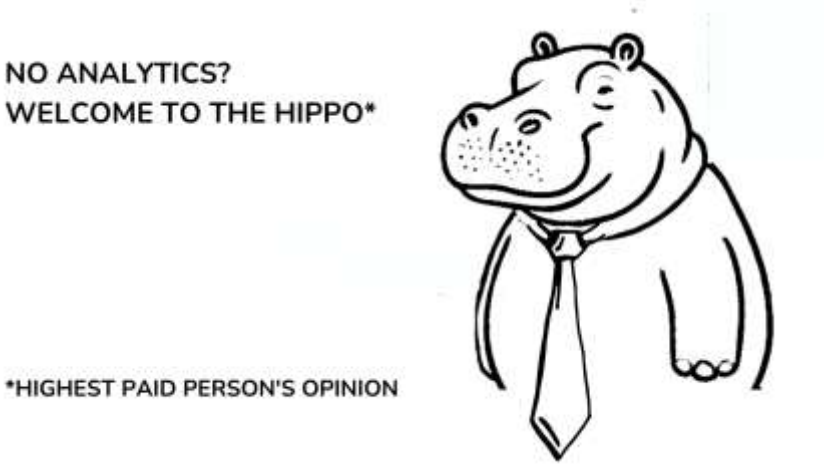


图 2.1-8 在没有分析方法的情况下， 业务依赖于经验丰富的专业人士的主观意见。

在稳定和缓慢变化的环境中，这种方法可能是合理的，但在数字化转型时代，它就成了一种严重的风险。基于直觉和猜测的决策很容易失真，往往是建立在毫无根据的假设之上，而且没有考虑到数据所反映的复杂情况

> 在公司的决策层，所谓的智慧辩论往往没有任何具体的依据。公司的成功不应取决于专家的权威和薪酬，而应取决于有效利用数据、识别模式和做出明智决策的能力。

必须摒弃权威或经验自动意味着决策正确的观念。数据驱动的方法改变了游戏规则：数据和分析，而不是职位和薪水，现在是决策的基础。大数据、机器学习和可视化分析使我们能够识别模式，并依靠事实而不是猜测（图 1.1-4）。

> 没有数据，你只是另一个有主见的人[34]。
>
> - 学者兼管理顾问 W. Edwards Deming

现代数据管理方法还能确保公司内部知识的连续性。清晰的流程、自动化和系统化的方法使得即使是关键岗位的转移也不会降低效率。

然而，盲目相信数据也可能导致严重错误。数据本身只是数字的集合。如果没有适当的分析、背景和识别模式的能力，它们就没有价值，也无法推动流程。成功的关键在于，而不是在 HiPPO 直觉和分析之间做出选择，而是建立智能工具，将不同的信息转化为可管理的明智决策。

> 在数字化施工环境中，决定成败的因素不是资历和职位，而是反应速度、决策准确性和资源效率

数据是工具，不是绝对真理。它应该是人类思维的补充，而不是替代。尽管分析，但数据不能完全取代人的直觉和经验。它们的作用是帮助做出更准确、更明智的决策。

竞争优势将不仅仅是通过达到标准来实现的，而是通过在有效利用对每个人都相同的资源方面超越竞争对手来实现的。未来，数据技能的重要性将不亚于识字或数学技能。能够分析和解释数据的专业人员将能够做出更准确的决策，从而取代那些只依靠个人经验的人员（图 2.1-9）。

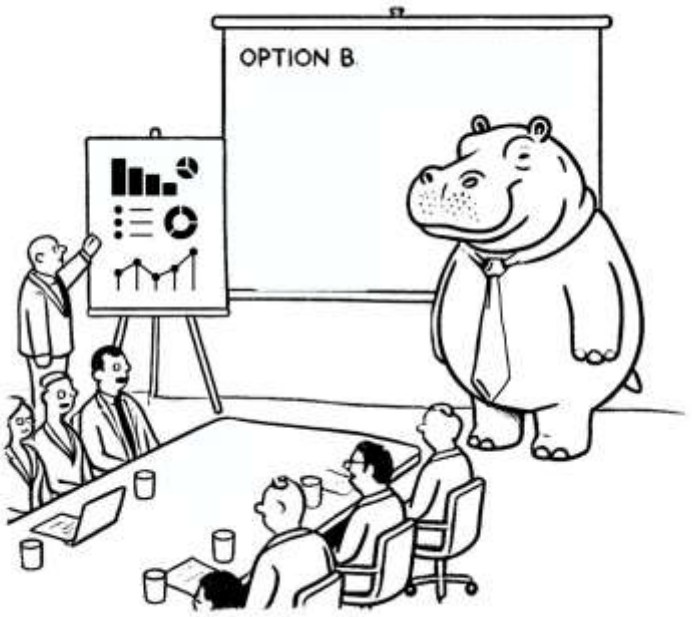


图 2.1-9 决策应基于客观分析，而不是薪酬最高的员工的意见。

管理人员、专家和工程师将充当 数据分析师，研究项目的结构、动态和关键指标。人力资源将成为系统的要素，需要灵活的数据驱动定制，以最大限度地提高效率。

> 使用不充分数据时的误差比不使用数据时的误差要小得多 [35]。
>
> - 查尔斯-巴贝奇，第一台分析计算器的发明者

大数据的出现和 LLM（大型语言模型）的引入不仅彻底改变了我们的分析方式，也改变了决策的本质。以前，我们关注的重点是因果关系（为什么事情会发生--诊断分析法 ）（图 1.1-4），而如今，预测未来（预测分析法）的能力，以及机器学习和人工智能在决策过程中提出最佳选择建议的规范分析法，正逐渐凸显出来。

> 根据 SAP™ 的最新研究报告《新研究发现近一半高管更信任人工智能》2025 [36]，44% 的高级管理人员愿意根据人工智能建议改变他们之前的决策，38% 的高级管理人员愿意信任人工智能代表他们做出商业决策。同时，74% 的高管表示，他们对人工智能建议的信任程度超过了他们的朋友和家人，55% 的高管所在的公司用人工智能的洞察力取代或经常绕过传统的决策方法--尤其是在年收入超过 50 亿美元的组织中。此外，48% 的受访者每天都使用人工智能生成工具，其中 15% 的受访者每天使用多次。

随着 LLM 和自动化数据管理系统的发展，一个新的挑战出现了：如何有效地使用信息，而又不至于在格式不兼容和来源不统一的混乱局面中失去其价值，同时业务流程的复杂性和动态性也在不断增加。

业务流程的复杂性和动态性不断增加

如今，建筑行业在数据和流程管理方面面临严峻挑战。主要挑战是信息系统各自为政、官僚作风严重以及数字工具之间缺乏整合。在技术、不断变化的客户需求和不断发展的法规的推动下，业务流程本身变得更加复杂，这些挑战也随之加剧。

建筑项目的独特性不仅在于其技术上的特殊性，还在于不同国家在国家标准和法规要求上的差异（图 4.2-10，图 5.1-7）。这就要求对每个项目采取灵活、个性化的方法，而这在传统的模块化控制系统中很难实现。由于流程复杂、数据量大，许多公司转向提供专业解决方案的供应商。但是，市场已经超载--许多初创公司提供类似的产品，专注于狭窄的任务。因此，数据管理的整体方法往往会丢失。

> 适应不断涌现的新技术和市场需求已成为竞争力的关键因素。然而，现有的专有应用程序和模块化系统的适应性较低，任何更改往往都需要开发人员进行冗长而昂贵的修改，因为他们并不总是了解施工流程的具体情况。

公司发现自己受制于技术的滞后性，只能等待新技术的更新，而不能迅速实施创新的综合方法。因此，建筑组织的内部结构往往是一个复杂的生态系统，由相互关联的等级系统组成，而且往往是封闭的，通过多级管理人员网络进行协调（图 2.1-10）。

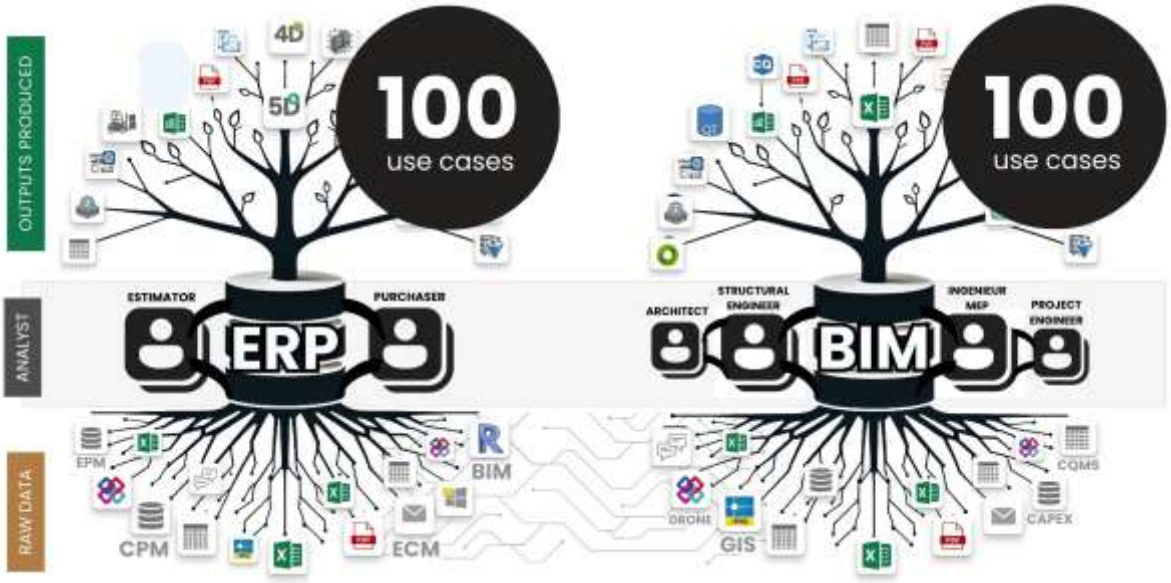


图 2.1-10 公司由相互连接的系统组成，这些系统的相互连接形成了需要自动化的流程。

> 根据加拿大建筑协会和毕马威加拿大 2021 年进行的一项调查[37]，只有 25% 的公司认为，在采用技术或数字解决方案方面，他们与竞争对手相比处于显著或不同的地位。只有 23% 的受访者表示，他们的解决方案明显或大量采用了数据驱动。与此同时，大多数受访者认为他们对一系列其他技术的使用纯属试验性质，或者承认根本没有使用这些技术。

这种不愿参与技术实验的态度在大型基础设施项目中尤为明显，因为在这些项目中，失误可能会造成数百万美元的损失。即使是最先进的技术--数字孪生、预测分析 --也经常遇到阻力，不是因为它们的有效性，而是因为在实际项目中缺乏经过验证的可靠性。

> 根据世界经济论坛（WEF）的报告 "塑造建筑业的未来"[5]，在建筑业中引入新技术不仅面临技术上的困难，而且还面临客户的心理障碍。世界经济论坛（WEF）的报告 "塑造建筑业的未来"[5]指出，建筑业引进新技术不仅面临技术上的困难，而且还面临客户的心理障碍。许多客户担心，使用先进的解决方案会使他们的项目成为实验场地，使他们成为 "小白鼠"，不可预测的后果可能会导致额外的成本和风险。

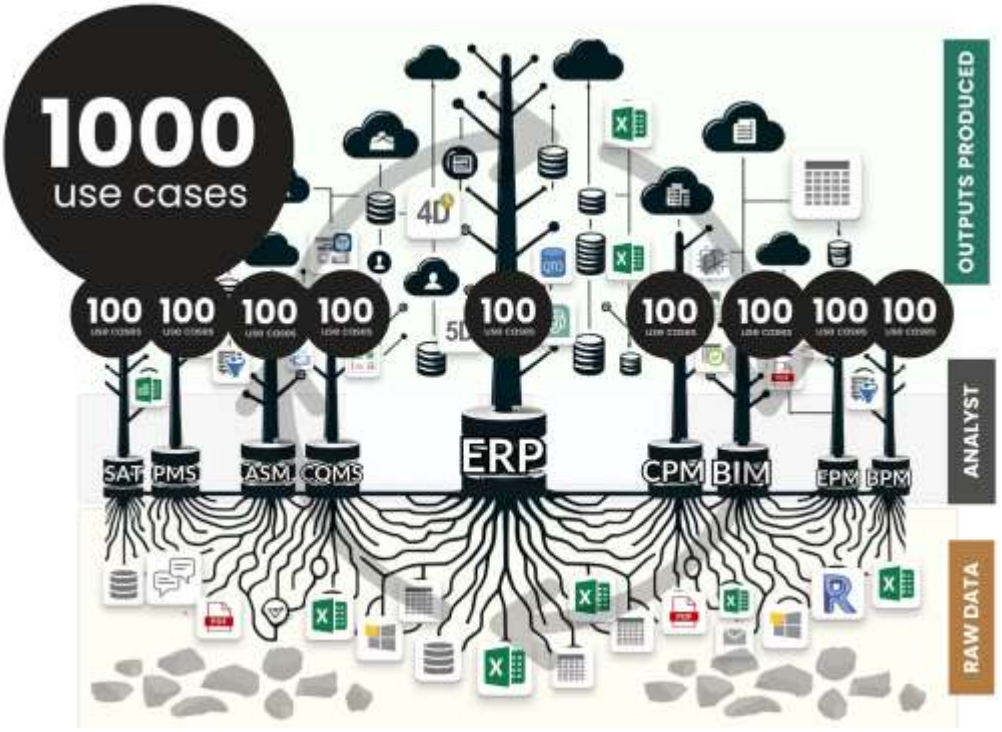


图 2.1-11 针对每种数据使用情况，解决方案市场都提供了用于优化和自动化流程的应用程序。

> 建筑行业千差万别：不同的项目有不同的要求 、地区特点、法定分类规定（图 4.2-10）、计算标准（图 5.1-7）等。因此，实际上不可能创建一个专有的通用应用程序或系统来完全满足所有这些要求 和项目的具体要求。

为了应对日益复杂的系统和对软件供应商的依赖，人们越来越意识到，有效数据管理的关键不仅在于开放性和标准化，还在于简化流程架构本身。业务流程的复杂性和动态性不断增加，这就需要采用新的方法，将工作重点从积累数据转移到结构化和组织数据上。正是这种转变将成为建筑行业发展的下一步，标志着软件供应商主导时代的结束和有意义的信息组织时代的开始。

由于意识到 "一刀切 "解决方案的局限性以及易受日益复杂性影响的脆弱性，人们正在将工作重点从封闭平台和数据囤积转向透明度、适应性和结构化信息处理。这种思维转变反映了全球经济和技术的广泛变化，通过所谓 "工业革命 "的视角进行描述。要了解建筑业的发展方向和未来走向，就必须考虑该行业在第四次和第五次工业革命中的地位--从自动化和数字化到个性化、开放标准和基于服务的数据模型。

第四次工业革命 **(**工业 **4.0)** 和建筑业的第五次工业革命 **(**工业 **5.0)**

技术和经济阶段是用来描述和分析社会和经济在不同发展阶段演变情况的理论概念。不同的研究人员和专家可以对它们做出不同的解释。

- 第四次工业革命 （4IR 或工业 4.0）与信息技术、自动化、数字化和全球化有关。其关键要素之一是创建专有软件解决方案，即为特定任务和公司设计的专业化数字产品。这些解决方案往往成为信息技术基础设施的重要组成部分，但如果不进行额外的修改，其可扩展性就很差。
- 与 4IR 相比，第五次工业革命 **(5IR)** 目前正处于概念化和发展的早期阶段。其核心原则包括提高产品和服务的个性化程度。第五次工业革命是一场向更具适应性、灵活性和个性化的经济活动发展的运动，其重点是个性化、咨询和以服务为导向的模式。第五种经济模式的一个关键方面是利用数据进行决策，如果不使用开放数据和开放工具，这几乎是不可能的（图 2.1-12）。

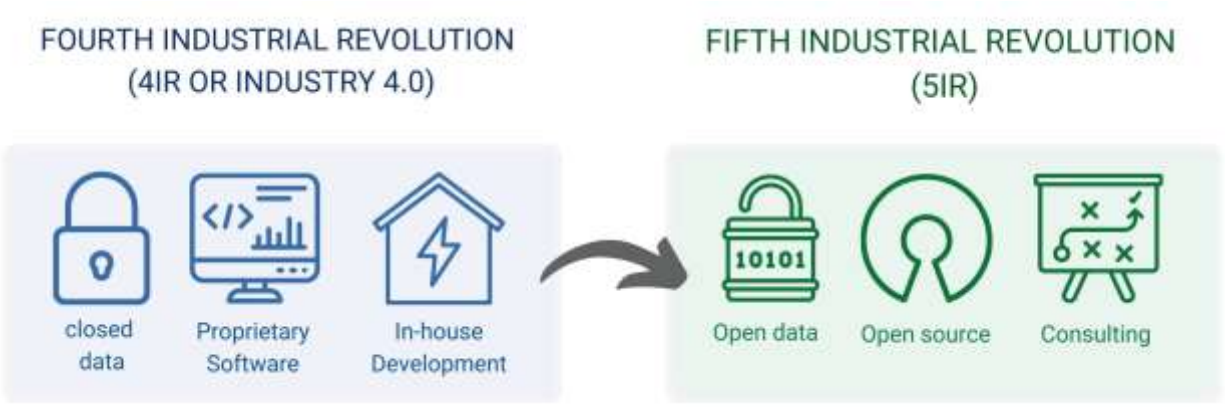


图 2.1-12 第四种模式侧重于解决方案，第五种模式侧重于个性化和数据。

> 为建筑行业的公司创建一个应用程序，供十个或一百个组织使用，并不能保证在没有重大修改和改进的情况下成功推广到其他公司、地区或国家。成功推广此类解决方案的可能性仍然很低，因为每个组织都有独特的流程、要求和条件，可能需要进行个性化调整。

重要的是要明白，如今技术解决方案的成功整合已经意味着对每个流程、项目和公司都要采取高度个性化的方法。这就意味着，即使开发出了通用框架、工具或方案，也需要对其进行详细的调整和定制，以满足每个特定公司和项目的独特要求 和条件。

> 根据普华永道的报告《解码第五次工业革命》[38]。[38]，今年各行各业约 50%的高级管理人员依赖于先进技术与人类专业知识的融合。这种方法使他们能够迅速适应产品设计或客户要求的变化，创造个性化生产。

鉴于全球建筑业的规模和项目的多样性，每个流程都需要开发独特的功能或应用程序，这就导致存在大量的业务案例，每次都代表着独特的管道 逻辑（图 2.1-13）。每个案例都有其自身的特点，需要采用定制的方法。我们将在专门讨论机器学习和泰坦尼克号数据集解析的章节（图 9.2-9）中更详细地探讨在不同方法的背景下，同一分析问题可能存在的各种解决方案。

> 在数字流程的背景下，管道 是一连串的活动、流程和工具，能够在项目生命周期的不同阶段实现数据和工作的自动化或结构化流动。

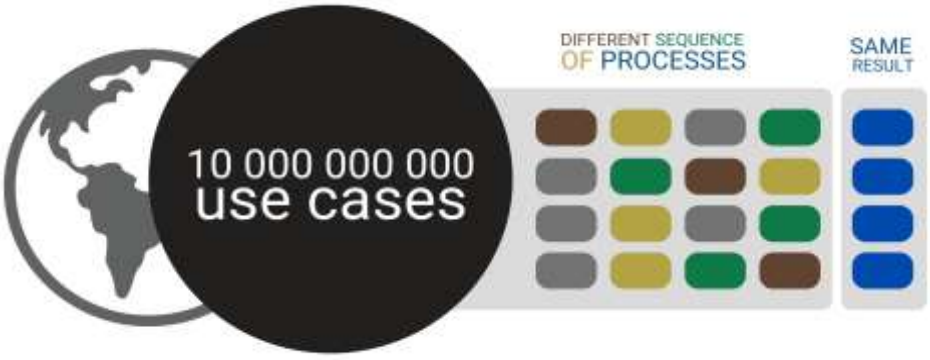


图 2.1-13 业务案例的个性化和多变性使得试图创建可扩展的封闭式平台和工具成为不可能。

在数字化转型的影响下，我们的生活已经发生了许多变化，今天我们可以说建筑业的经济发展已经进入了一个新阶段。在这个 "新经济 "中，竞争将按照不同的规则进行：谁能有效地将公共知识和开放数据 转化为人们所需要的产品和服务，谁就能在第五次工业革命的条件下获得关键优势。

> 经济学家凯特-马斯库斯在《私人权利与公共问题：21 世纪的全球知识产权经济》一书中指出[39]：“我们生活在全球知识经济中，未来属于那些知道如何把科学发现变成商品的人。2012 年[39]，“我们生活在全球知识经济时代，未来属于那些懂得如何将科学发现转化为商品的人”。

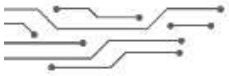
向第五种经济模式的过渡意味着重点将从封闭的信息技术解决方案转向开放的标准和平台。企业将开始放弃

传统的软件产品，转而采用以服务为导向的模式，在这种模式下，数据而非专有技术将成为主要资产。

> 哈佛商学院 2024 年研究[40]显示，开放源码 软件（Open Source Software, OSS）具有巨大的经济价值。根据这项研究，96% 的软件代码都包含开放源码软件，一些商业软件 99.9% 包含开放源码软件组件。如果没有开放源码软件，公司在软件上的花费将增加 3.5 倍。

> 随着全球趋势的发展，建立公司生态系统将逐渐转向第五种经济模式，在这种模式下，以数据为中心的分析和咨询服务将比具有严格定义的使用场景的孤立、封闭解决方案更受重视。

数字化时代将改变行业的力量平衡：企业将不再依赖供应商的解决方案，而是将其竞争力建立在有效利用数据的能力之上。因此，建筑行业将从传统的僵化系统转向灵活、适应性强的生态系统，开放标准和互操作工具将成为项目管理的基础。应用软件供应商主导时代的结束将创造一个新的环境，在这个环境中，价值将不是由拥有封闭源代码和专用连接器来定义，而是由将数据转化为战略优势的能力来定义。



章 2.2.

化繁为简

冗余代码和封闭系统是提高生产力的障碍

过去几十年来，IT 领域的技术变革主要是由软件供应商推动的。它们确定了发展方向，决定了公司应采用哪些技术，哪些技术应被抛在后面。在从孤岛式解决方案向集中式数据库和集成系统转变的时代，供应商推广授权产品，提供对访问和可扩展性的控制。后来，随着云技术和软件即服务（SaaS）模式的出现，这种控制权演变成了订阅模式，巩固了用户作为数字服务忠实客户的地位。

这种做法产生了一个悖论：尽管创建的程序代码数量空前巨大，但实际使用的只是其中的一小部分。代码的数量可能是所需数量的数百或数千倍，因为同样的业务流程在数十或数百个程序中以不同的方式进行描述和重复，甚至在同一公司内也是如此。同时，开发成本已经支付，而且无法收回。尽管如此，该行业仍在继续复制这种循环，为最终用户创造附加值极低的新产品，更多时候是迫于市场期望而非实际需求的压力。

> 根据国防采办大学（DAU）的《软件开发成本估算指南》[41]，软件开发成本可能会因若干因素（包括系统的复杂性和所选技术）的不同而有很大差异。从历史上看，2008 年的开发成本约为每行源代码（SLOC）100 美元，而维护成本可高达每行 SLOC 4,000 美元。

仅 CAD 应用程序的一个组件--几何核心--就可能有数千万行代码（图 6.1-5）。ERP 系统中也有类似的情况（图 5.4-4），我们将在本书的第五部分再讨论其复杂性。然而，仔细观察就会发现，大部分代码并没有增加价值，而只是充当了 "邮递员 "的角色--在数据库、应用程序接口、用户界面和系统中的其他表之间机械地移动数据。尽管所谓的业务逻辑至关重要的说法广为流传，但残酷的现实却要平实得多：现代代码库中充满了过时的模板块（遗留代码），其唯一目的就是确保数据在表和组件之间传输，而不会影响决策或业务效率。

因此，处理各种来源数据的封闭式解决方案不可避免地会变成混乱的 "意大利面生态系统"。这些错综复杂、盘根错节的系统只能由一大批管理人员以半例行公事的方式来处理。这种数据管理组织方式不仅在资源方面效率低下，而且还会在业务流程中造成关键漏洞，使公司只能依赖于了解这种技术迷宫运作方式的少数专家。

代码量、应用程序数量的不断增加，以及供应商所提供概念的日益复杂，自然导致了一个结果--建筑业 IT 生态系统复杂性的增加。这使得通过增加行业应用软件数量来实现数字化的实际效果大打折扣。在没有充分考虑用户需求的情况下开发的软件产品往往需要大量的资源来实施和支持，但却不能带来预期的回报。

> 麦肯锡 "提高建筑业生产率 "研究报告指出[42]，在过去二十年里，全球建筑业的劳动生产率平均每年仅增长 1%，而世界经济整体增长 2.8%，制造业增长 3.6%。在美国，自 20 世纪 60 年代以来，每名工人的建筑劳动生产率下降了一半[43]。

日益复杂的系统、孤立和封闭的数据阻碍了专业人员之间的交流，使建筑业成为效率最低的行业之一（图 2.2-1）。到 2040 年，建筑业的总产值将达到 22 万亿美元，这就需要大幅提高效率。

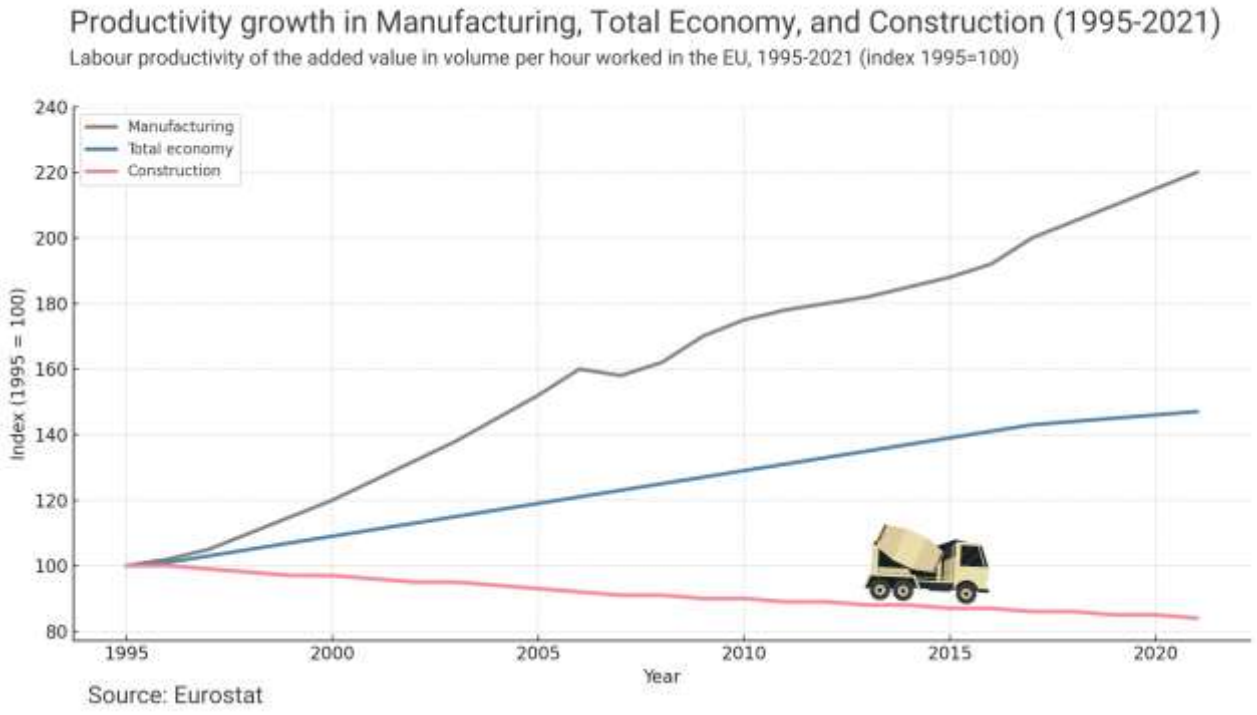


图 2.2-1 数据封闭而复杂，因此专家之间沟通不畅，导致建筑业成为经济中效率最低的行业之一（基于 [44], [45]）。

> 正如麦肯锡（2024 年）研究报告《确保建筑业的生产率不再是可有可无的》中所强调的，随着资源的日益稀缺和建筑业翻番增长的雄心，建筑业再也不能停留在目前的生产率水平上了[44]。全球建筑成本预计将从 2023 年的 13 万亿美元上升到本十年末更高的水平，这使得效率问题不仅具有现实意义，而且至关重要。

提高效率的关键方法之一是不可避免地统一和简化应用结构和数据生态系统架构。这种合理化方法将消除企业系统中多年积累的多余抽象层和不必要的复杂性。

从孤岛到单一数据仓库

企业积累的数据越多，就越难从中提取真正的价值。由于将信息存储在孤立的筒仓中具有分散性，现代公司的业务流程就像建筑工人试图用存储在成千上万个不同仓库中的材料建造一座摩天大楼。过多的信息不仅难以获得合法的相关信息，而且还会拖慢决策速度：每一步都需要反复检查和确认。

> 每个任务或流程都与单独的表格或数据库硬连接，系统间的数据交换需要复杂的集成。一个系统中的错误和不一致会导致其他系统的连锁故障。不正确的数值、延迟更新和重复信息迫使员工花费大量时间手动核对和调节数据。因此，与开发和优化流程相比，企业要花费更多的时间来处理碎片化带来的后果

这个问题是普遍存在的：一些公司继续在混乱中挣扎，而另一些公司则在整合中找到了解决方案--将信息流转移到一个集中存储系统中。把它想象成一张大表，可以存储与任务、项目和对象相关的任何实体。这样，就不再需要数十个不同的表格和格式，而是可以实现以下功能的单一、有凝聚力的存储库（图 2.2-2）：

- 最大限度地减少数据丢失；
- 无需不断统一信息；
- 提高数据的可用性和质量；
- 简化分析处理和机器学习

将数据纳入通用标准意味着，无论数据来源如何，信息都会被转换成统一的机器可读格式。通过这种数据组织方式，可以检查其完整性，进行实时分析，并及时用于管理决策。

集成存储系统的概念及其在分析和机器学习中的应用将在 "大数据存储和机器学习 "一章中详细讨论。数据建模和结构化主题将在 "将数据转化为结构化形式 "和 "标准如何改变游戏规则：从随机文件到精心设计的数

据模型 "两章中详细介绍。

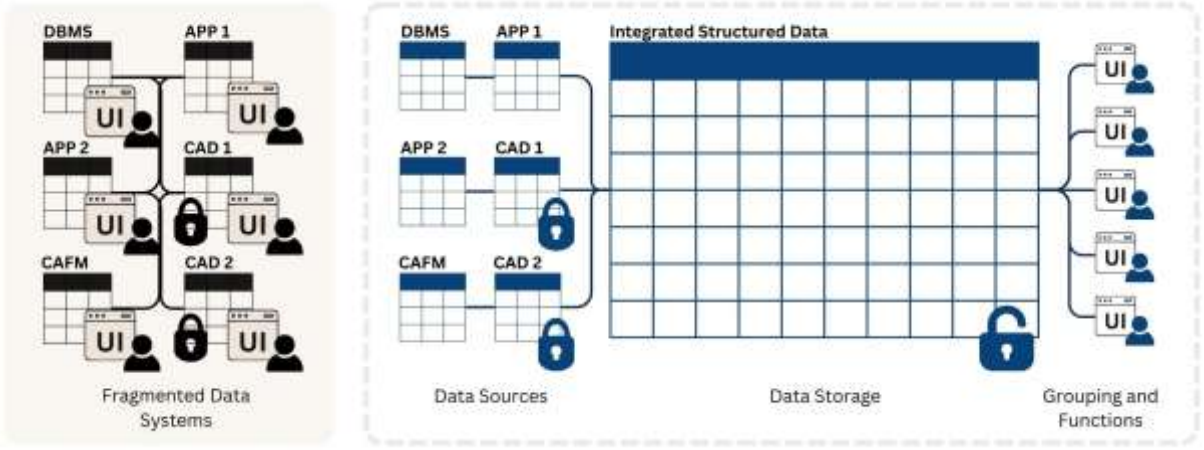


图 2.2-2 数据集成消除了孤岛，提高了信息可用性，优化了业务流程。

一旦对数据进行了结构化和合并，下一个合乎逻辑的步骤就是对其进行验证。有了单一的集成存储库，这一过程就大大简化了：不再有多个不一致的模式、重复的结构和表间复杂的关系。所有信息都与单一数据模型保持一致，消除了内部不一致，加快了验证过程。验证 和确保数据质量 是所有业务流程的基石，我们将在本书相关章节中详细介绍。

在最后阶段，对数据进行分组、过滤和分析。对数据应用各种功能：汇总（加法、乘法），表、列或行之间的计算（图 2.2-4）。数据处理变成了一系列步骤：收集、结构化、验证、转换、分析处理和卸载到最终应用，在最终应用中，信息被用来解决实际问题。我们将在有关 ETL 流程和数据管道方法的章节中详细讨论如何构建此类方案、实现步骤自动化和构建处理流程。

因此，数字化转型不仅仅是简化信息处理。它是要消除数据管理中的过度复杂性，从混乱转向可预测性，从多个系统转向可管理的流程。架构的复杂性越低，支持架构所需的代码就越少。在未来，代码可能会完全消失，取而代之的是能够独立分析、系统化和转换数据的智能代理。

集成存储系统实现了向人工智能的过渡 **agent**

数据和系统越不复杂，需要编写和维护的代码就越少。而节省开发时间的最简单方法就是完全摒弃代码，代之以数据。当应用程序代码开发从代码转向数据模型时，不可避免地会转向以数据为中心（数据驱动 ）的

方法，因为这些概念背后有着完全不同的思维方式。

> 当人们选择以数据为中心开展工作时，就会开始以不同的方式看待数据的作用。数据不再仅仅是应用程序的 "原材料"，它现在是构建架构、逻辑和交互的基础。

传统的数据管理方法通常从应用层面开始，在结构上类似于一个繁琐的官僚系统：多级审批、人工检查、通过相关软件产品无休止地修改文件版本。随着数字技术的发展，越来越多的公司将被迫转而采用最小化原则，即只存储和使用真正必要和会用到的东西。

最小化的逻辑已被供应商采纳。为了简化数据存储和处理，用户的工作正在从离线应用程序和工具转向云服务和所谓的 SaaS 解决方案。

> SaaS 概念 （软件即服务，或 "软件即服务"）是现代 IT 基础设施的主要趋势之一，它允许用户通过互联网访问应用程序，而无需在自己的计算机上安装和维护软件。

一方面，SaaS 为扩展、版本控制以及降低支持和维护成本提供了便利，但另一方面，除了依赖于特定应用程序的逻辑之外，它还使用户完全依赖于提供商的云基础设施。如果服务出现故障，对数据和业务流程的访问可能会暂时甚至永久受阻。此外，在使用 SaaS 应用程序时，所有用户数据都存储在提供商的服务器上，这就产生了安全和合规风险。资费或使用条款的变化也可能导致成本增加或需要紧急迁移。

人工智能 、LLM 代理和以数据为中心的方法的发展，对传统形式的应用程序和 SaaS 执行的未来提出了质疑。以前，应用程序和服务需要管理业务逻辑和处理数据，而随着人工智能代理的出现，这些功能可能会转移到直接处理数据的智能系统上。

这就是为什么 IT 部门和管理层越来越多地讨论混合架构的原因，在混合架构中，人工智能 - 代理和内部部署解决方案与云服务相辅相成，减少了对 SaaS - 平台的依赖。

> 我们采用的方法认识到，传统的商业应用程序或 SaaS 应用程序在代理时代可能会发生巨大变化。这些应用程序本质上是具有业务逻辑的 CRUD [创建、读取、更新和删除] 数据库。但在未来，这些逻辑将由人工智能代理[46]。

> - 2024 年，微软首席执行官萨提亚-纳德拉。

以数据为中心的方法和人工智能/LLM 代理的使用可以减少冗余流程，从而减轻员工的工作量。如果数据组织得当，分析、可视化和应用于决策就会变得更加容易。专家们只需点击几下，或在 LLM 代理的帮助下，就能以现成文件和仪表板的形式自动获取最新信息，而不再需要无休止的报告和检查。

我们将通过人工智能工具 （AI ）和 LLM 聊天来协助我们进行数据操作。近年来，数据管理的趋势已经从传统的 CRUD 操作 （创建、读取、更新、删除）转向使用大型语言模型（LLM）。LLM 能够解释自然语言并自动生成适当的数据库查询，从而简化了与数据管理系统的交互（图 2.2-3）。

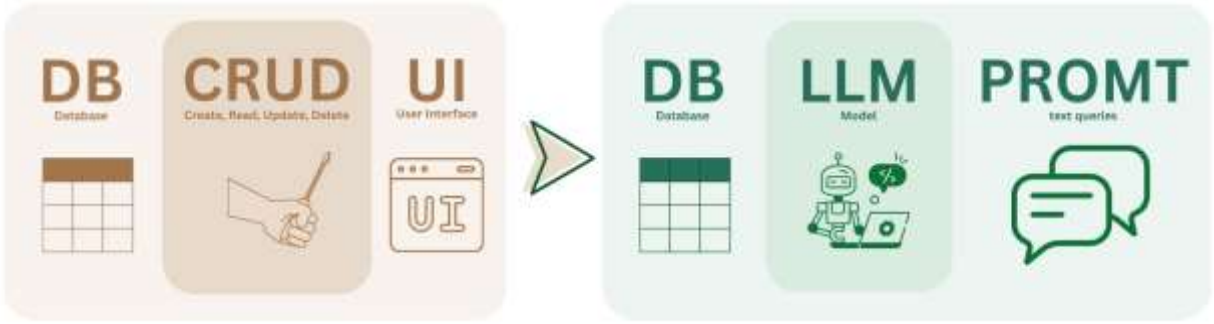


图 2.2-3 人工智能 将取代并整合存储和数据库解决方案，逐渐取代传统应用程序和 CRUD -操作。

> 在未来 3-6 个月内，人工智能 将编写 90% 的代码，而在 12 个月内，几乎所有代码都可能由人工智能生成[47]。
>
> - 达里奥-阿莫代（Dario Amodei），LLM Anthropic 首席执行官，2025 年 3 月。

尽管人工智能开发工具（如 GitHub Copilot）发展迅速，但在 2025 年，开发人员仍将在这一过程中扮演关键角色。人工智能代理正成为越来越有用的助手：它们能自动解释用户 查询、生成 SQL 和 Pandas 查询（下文将详细介绍），或编写代码来分析数据。这样，人工智能正逐渐取代传统的应用程序用户界面。

人工智能模型 ，如语言模型，将推动混合架构的发展。我们可能会看到云服务与本地数据管理系统的整合，而不是完全放弃云解决方案和 SaaS 产品。例如，联合学习可实现强大的人工智能 模型，而无需将敏感数

据转移到云中。通过这种方式，企业可以在获得先进技术的同时保持对数据的控制。

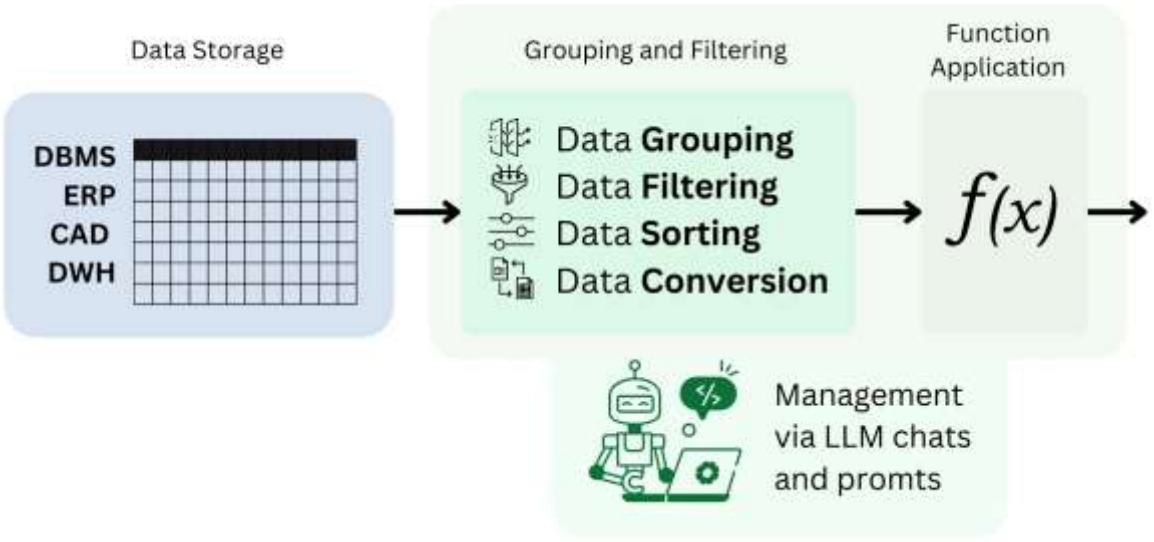


图 2.2-4 分组、筛选和排序等基本操作以及函数应用将由 LLM 聊天处理。

建筑行业的未来将基于企业内部解决方案、云计算和智能模型的结合，共同打造高效、安全的数据管理系统。LLM 将使没有深厚技术知识的用户能够通过使用自然语言进行查询来与数据库和数据仓库进行交互。我们将在 "LLM 代理和结构化数据格式 "一章中详细介绍 LLM 和人工智能 代理及其工作原理。

妥善组织的数据和简单易用的支持 LLM 的分析工具不仅能让信息处理变得更容易，还有助于最大限度地减少错误、提高效率和实现流程自动化。

从数据收集到决策：自动化之路

在本书的后半部分，我们将详细介绍专家之间如何互动，以及数据如何成为决策、自动化和运行效率的基础。图 2.2-5 提供了一个示例图，显示了以数据为中心的方法中数据处理步骤的顺序。该图说明了持续改进管道），本书稍后将详细讨论其中的部分内容。

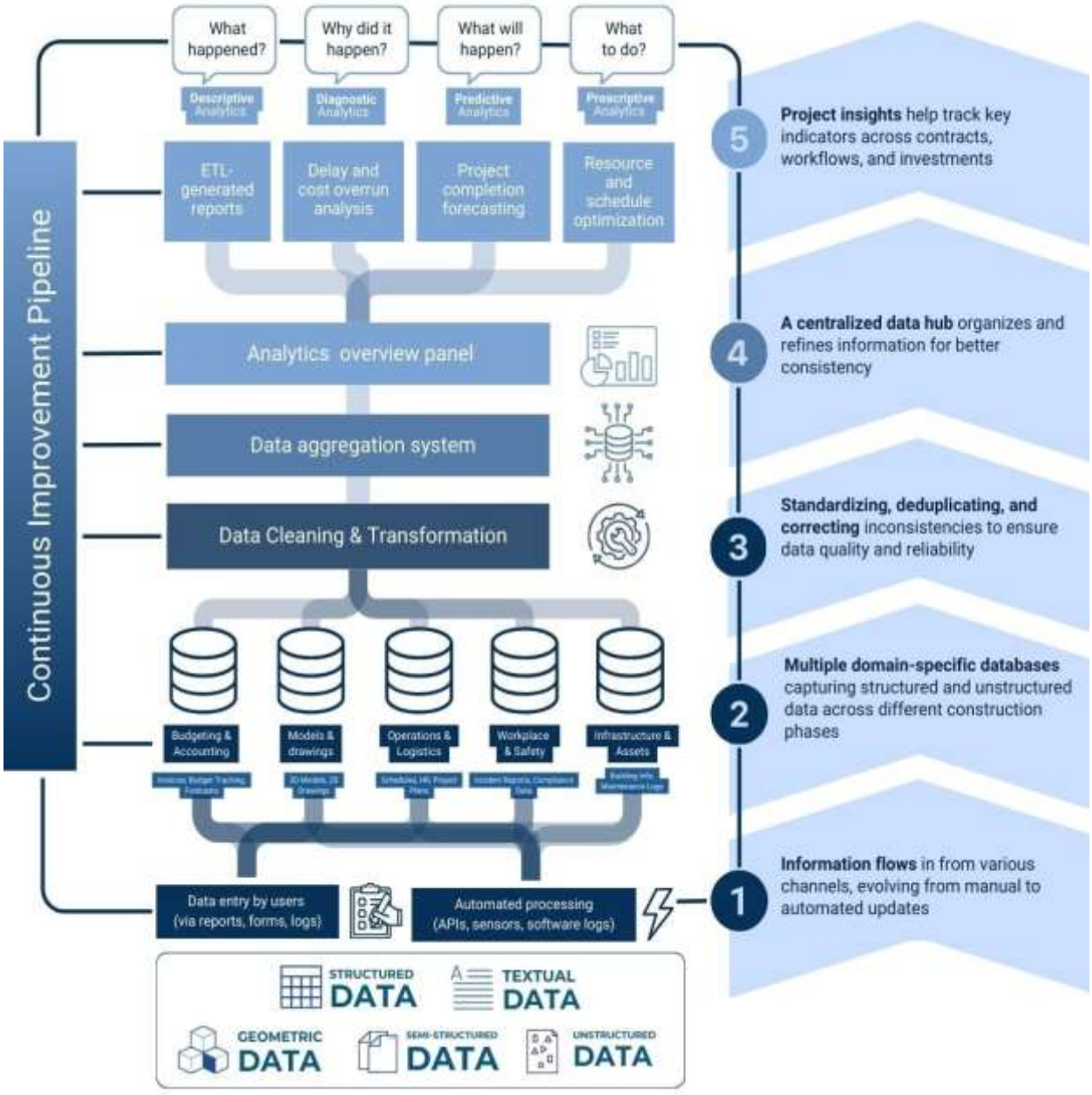


图 2.2-5 持续数据改进管道的一个例子：建筑项目中的数据处理和分析流程。

描述一家中型公司业务流程的系统是根据多层次原则建立的。它包括：数据收集、清理、汇总、分析处理和基于结果的决策。我们将在本书的后面部分对所有这些阶段进行研究--既有理论方面的，也有实际案例：

- 第一层是数据录入（图 3.1-1）。信息既可以通过人工方式（报告、表格、日志）接收，也可以通过自动方式（来自 API 、传感器、软件系统）接收。数据可以是不同的结构：几何、文本、非结构化。在这一阶段，需要对信息流进行标准化、结构化和统一化处理。

- 下一级是数据处理和转换。它包括清理、删除重复数据、纠正错误以及为进一步分析准备信息等过程（图 4.2-5）。这一阶段至关重要，因为分析 的质量直接取决于数据的清洁度和准确性。
- 然后，数据进入按职能领域划分的专门表格、数据框架或数据库：预算和会计、模型和图纸、后勤、安全和基础设施。通过这种划分，可以方便地访问和交叉分析信息。
- 然后将数据汇总并显示在分析仪表板（展示台）上。描述性、诊断性、预测性和规范性分析方法在这里得到应用。这有助于回答关键问题（图 1.1-4）：发生了什么、为什么会发生、将来会发生什么以及需要采取什么行动。例如，系统可以识别延误、预测项目完成或优化资源。
- 最后，最后一个层次产生的分析结论和关键指标有助于监督合同履行情况、管理投资和改进业务流程（图 7.4-2）。这些信息成为决策和公司发展战略的基础。

同样，数据也要从收集到用于战略管理。在本书的以下部分，我们将详细介绍每个阶段，重点是数据类型、数据处理技术、分析工具 以及这些方法在建筑行业中的实际应用案例。

下一步：将混乱变为可管理的系统

在这一部分中，我们探讨了信息孤岛带来的挑战，研究了系统过度复杂对业务绩效的影响，分析了从第四次工业革命向第五次工业革命的过渡，在第五次工业革命中，数据而非应用程序成为核心。我们看到孤岛式信息系统是如何为知识共享制造障碍的，以及信息技术环境的持续复杂性是如何降低生产力和抑制建筑业创新的。

在总结这部分内容时，值得强调的是主要的实际步骤，这些步骤将有助于您在日常工作中应用所讨论的方法：

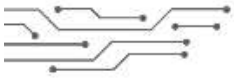
- 可视化您的信息环境
    - ☐ 为你经常使用的数据源创建可视化地图（Miro、Figma、Canva）。
    - ☐ 将您在工作中使用的系统和应用程序添加到此地图中
    - ☐ 识别潜在的重复功能和冗余解决方案
    - ☐ 确定系统间传输数据时可能发生数据丢失或损坏的关键点
- 采用个性化数据管理方法

- 将重点从应用程序转移到作为流程关键资产的数据上
- 记录数据来源和处理方法，确保透明度
- 建立评估和改进数据质量的机制
- 努力确保数据一次输入，反复使用 - 这是高效流程组织的基础

- 在团队中推广以数据为中心（数据驱动）的方法
  - 建议使用标准化的统一格式进行点对点数据交换
  - 定期在团队会议上提出与数据质量和可用性有关的问题
  - 了解开放源码 替代您用来解决问题的工具

从小处着手 - 挑选一个对您的工作至关重要的特定流程或数据集，并对其采用以数据为中心的方法，将重点从工具转移到数据上。通过单个试点取得成功，您不仅可以获得实践经验，还可以向团队清楚地展示新方法的好处。在完成大部分步骤的过程中，如果您有问题，可以向任何最新的法律硕士 寻求澄清和帮助。

在本书的后续部分，我们将更详细地介绍数据结构和协调技术，并探讨整合异构信息的实用方法。我们将特别关注从分散的孤岛到统一数据生态系统的转变，这在建筑行业的数字化转型中发挥着关键作用。





III 部分

建筑业务流程中的数据框架

第三部分是对建筑数据类型及其有效组织方法的全面了解。分析了建筑项目中结构化数据、非结构化数据、半结构化数据、文本数据和几何数据的特点和特性。回顾了在行业内使用的不同系统之间交换信息的现代存储格式和协议。介绍了将多格式数据转换为单一结构化环境的实用工具和技术，包括如何集成 CAD (BIM) 数据。提出了通过标准化和验证确保 数据质量的方法，这对施工计算的准确性至关重要。详细分析了使用现代技术（Python Pandas, LLM -models）和代码示例解决建筑行业典型问题的实践方面。建立能力中心（CoE ）作为协调和标准化信息管理方法的组织结构的价值得到了证实。

章 3.1.

建筑数据类型

建筑业最重要的数据类型

在现代建筑业中， 公司的系统、应用软件和数据仓库都积极充斥着各种类型和格式的信息和数据（图 3.1-1 ）。让我们来详细了解一下构成现代建筑业公司信息图景的主要数据类型：

- 结构化数据： 这种数据有明确的组织结构，如 Excel 电子表格 和关系数据库。
- 非结构化数据： 这是没有按照严格规则组织的信息。此类数据的例子包括文本、视频、照片和录音 。
- 松散结构化数据： 这些数据介于结构化数据和非结构化数据之间。它们包含结构元素，但这种结构并不总是很清晰，或经常通过不同的模式来描述。建筑业半结构化数据的例子有：技术规范、项目文件或进度报告。
- 文本数据： 包括来自口头和书面通信的任何内容，如电子邮件、会议和约会记录。
- 几何数据： 这些数据来自 CAD 程序 ，专家在这些程序中创建项目元素的几何数据，用于可视化、体积值确认或碰撞检查。

值得注意的是，几何数据和文本（字母数字）数据不是一个单独的类别，而是可以出现在所有 三类数据中。例如，几何数据 既可以是结构化数据（参数化 CAD 格式）的一部分，也可以是非结构化数据（扫描图纸）的一部分。文本数据 同样既可以组织在数据库中（结构化数据），也可以作为没有明确结构的文档存在。

建筑公司的每一种数据都是公司信息资产马赛克中的一个独特元素。从非结构化数据（如施工现场的图像和会议录音）到结构化记录（包括表格和数据库），每个元素都在塑造公司的信息景观方面发挥着重要作用。

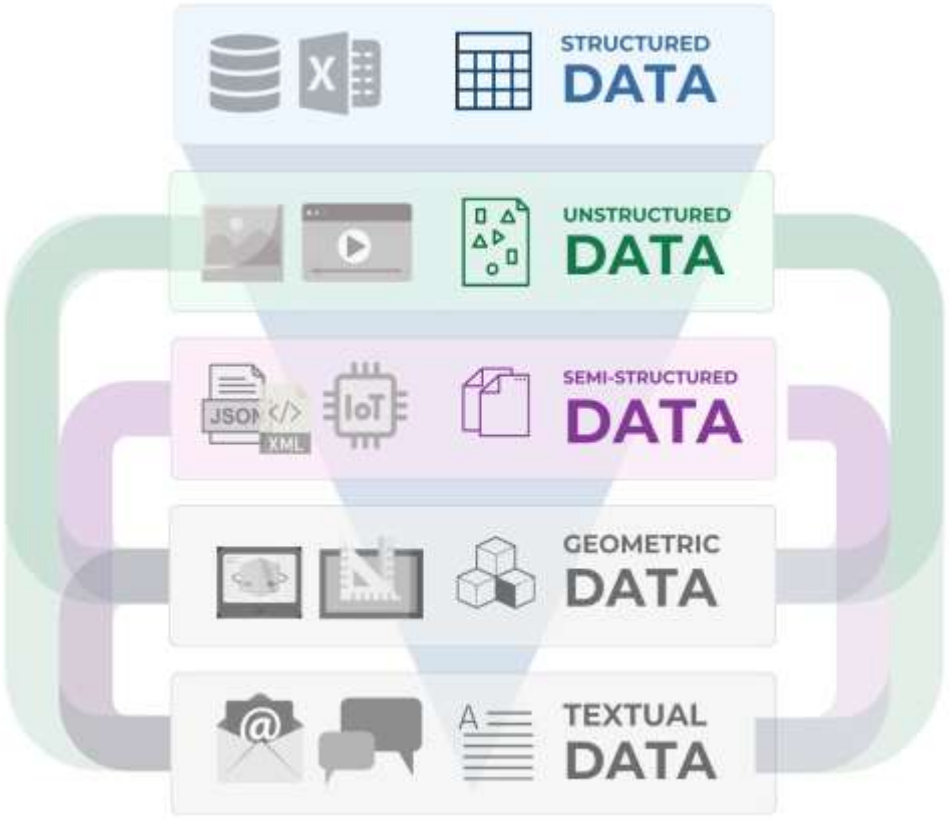


图 3.1-1 工程师和数据管理人员必须学会处理建筑行业使用的各类数据。

以下是建筑中使用的部分系统和相关数据类型的示例列表（图 3.1-2）：

- **ERP** （企业资源规划）--处理一般结构化数据，帮助管理企业资源和整合各种业务流程。
- **CAD** （计算机辅助设计）与 **BIM** （建筑信息模型）相结合--使用几何和半结构化数据对建筑项目进行设计和建模，确保设计阶段信息的准确性和一致性。
- **GIS** （地理信息系统）--使用几何和结构化数据创建和分析制图数据和空间关系。
- **RFID** （射频识别）--利用半结构化数据，通过射频识别有效跟踪建筑工地上的材料和设备。
- **ECM** （工程内容管理系统）是一个管理工程数据和文档的系统，包括半结构化和非结构化数据，如技术图纸和设计文档。

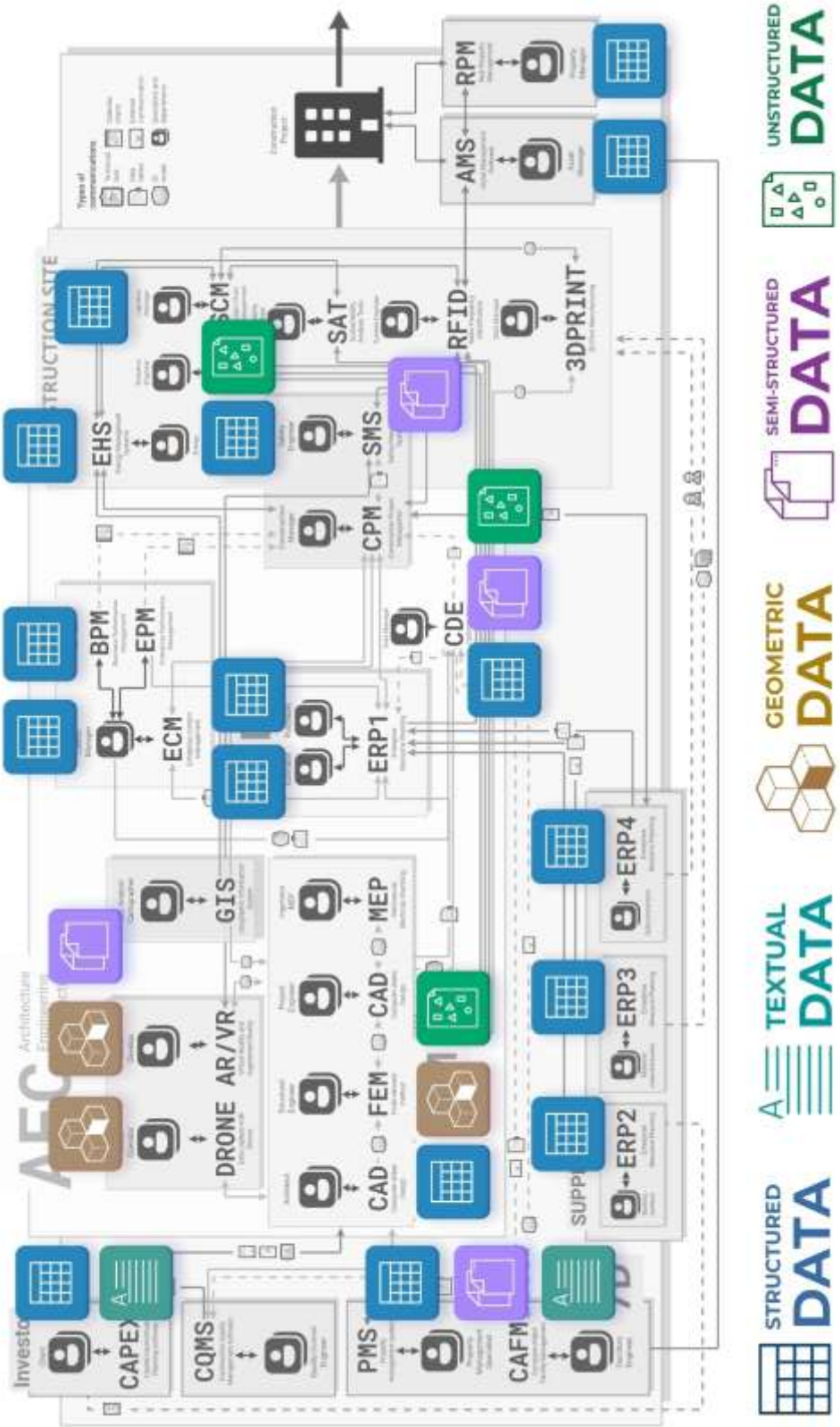


图 3.1-2 不同的格式和数据充斥着不同的系统，需要翻译成适合复杂集成的形式。

这些系统和公司的许多其他系统可管理从结构化表格数据到复杂几何模型等各种数据，在设计、规划和施工管理过程中提供综合协作。

以简化对话为例（图 3.1-3），建筑项目专家之间交换不同类型的数据：

- 建筑师: "考虑到客户的愿望，我在屋顶增加了一个座位区。请看一下新的设计"（几何数据 - 模型）。
- 结构工程师："项目已收到。我正在计算新休闲区屋顶的承重能力"（结构化和半结构化数据--计算表）。
- 采购经理："需要娱乐区材料的规格和数量，以便组织采购"（文本和半结构化数据 - 清单和规格）。
- 健康与安全工程师："收到了关于新区域的数据。我正在评估风险并更新安全计划"（半结构化数据 - 文件和计划）。
- **BIM -**建模专家："对整个项目模型进行修改，以调整工作文件"（几何数据和半结构化数据）。
- 项目经理："我正在将新的休息区纳入工作计划。我正在更新项目管理系统中的时间表和资源"（结构化和半结构化数据 - 时间表和计划）。
- 设施维护专家（**FM**）："我为娱乐区未来的维护工作准备数据，并将其输入财产管理系统"（结构化和半结构化数据--说明和维护计划）。

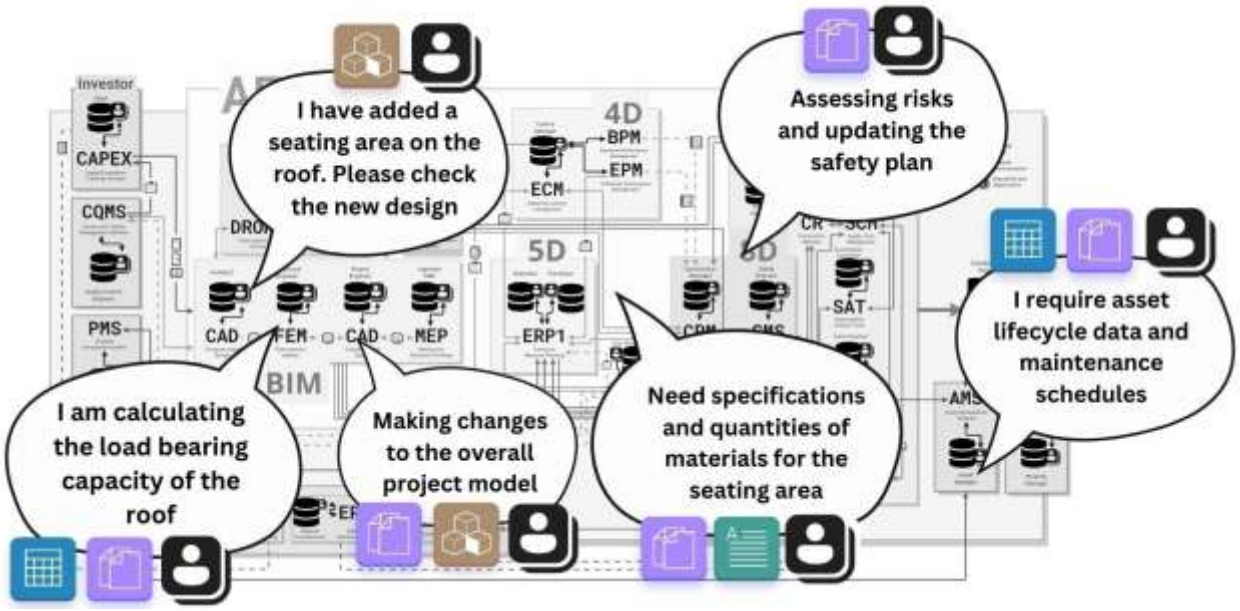


图 3.1-3 专家之间的交流在文本和数据两个层面上进行。

每个专业人员都要处理不同类型的数据，以确保有效的团队协作和项目成功。了解结构化数据、半结构化数据和非结构化数据之间的区别，可以让您认识到每种类型的数据在数字业务流程中发挥的独特作用。不仅要知道有不同形式的数据，还要了解它们的使用方式、地点和原因。

> 不久前，将如此多样的数据结合起来的想法似乎雄心勃勃，但却难以实现。如今，这已成为日常实践的一部分。整合不同模式和结构的数据已成为现代信息系统架构不可分割的一部分。

在接下来的章节中，我们将详细介绍能够将结构化数据、半结构化数据和非结构化数据整合为一个统一视图的主要标准和方法。我们将特别关注结构化数据和关系数据库，它们是建筑行业存储、处理和分析信息的主要机制。

结构化数据

在建筑行业中，信息有很多来源--图纸、规范、时间表和报告。为了有效管理这些信息流，需要对其进行结构化处理。结构化数据 可以让您以方便、可读和可访问的形式组织信息。

> 根据 JB Knowledge 的《第五次年度建筑技术报告》[17]，67% 的建筑项目管理专业人员使用手动或电子表格跟踪和评估工作绩效。

最常见的结构化数据格式有 XLSX 和 CSV。它们广泛用于在电子表格中存储、处理和分析信息。在这些电子表格中，数据以行和列的形式呈现，便于阅读、编辑和分析。

XLSX 是微软公司创建的一种格式，基于 XML 结构，并使用 ZIP 算法进行归档。该格式的主要特点是

- 支持复杂的公式、图表和宏。
- 能够将数据存储在不同的表单中，并对信息进行格式化。
- 优化用于 Microsoft Excel ，但与其他办公套件兼容。

CSV 格式 是一种纯文本文件，其中的值用逗号、分号或其他分隔符分隔。主要优点

- 与各种程序和操作系统普遍兼容。
- 轻松导入/导出至数据库和分析系统。

- 即使在文本编辑器中也能轻松处理。

不过，CSV 不支持公式和格式化，因此其主要应用是系统间的数据交换和信息的大量更新。由于其通用性和平台独立性，CSV 已成为异构 IT 环境中数据传输的常用工具。

XLSX 和 CSV 这两种格式是处理结构化数据的不同系统之间的纽带（图 3.1-4）。它们在可读性、手动编辑和基本兼容性非常重要的任务中特别有用

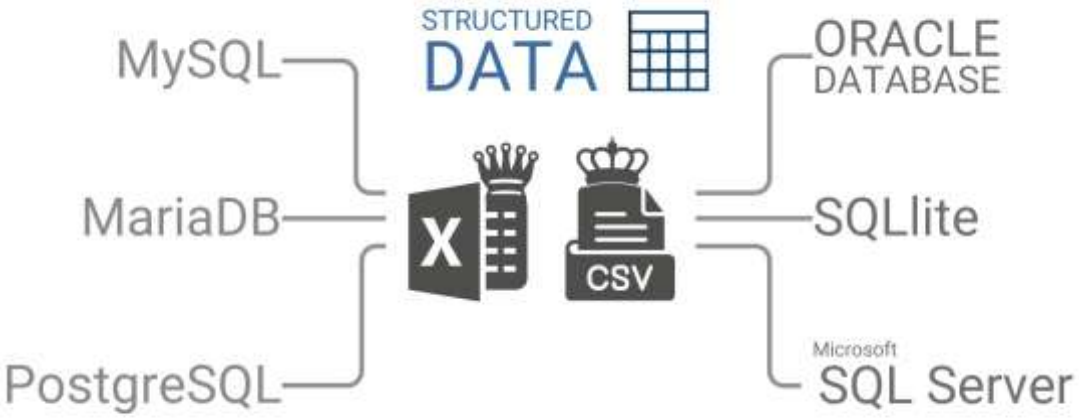


图 3.1-4 XLSX 和 CSV 格式是连接处理结构化数据的不同系统的纽带。

> 平台独立性使 CSV 成为异构 IT 环境和系统中最常用的数据传输格式。

不过，XLSX 和 CSV 并不是为高性能计算或长期存储大量数据而设计的。更现代的结构化格式，如 Apache Parquet 、Apache ORC 、Feather 、HDF5 都可用于此类目的。这些格式将在本书第 9 部分 "存储大数据：分析流行格式及其有效性 "一章中详细讨论

在实践中，Excel 与 XLSX 格式 更常用于小型任务和日常流程自动化。更复杂的情况需要使用数据管理系统，如 ERP, PMIS CAFM, CPM, SCM 等（图 3.2-1）。这些系统存储结构化数据，是组织和管理公司信息流的基础。

建筑行业使用的现代数据管理信息系统依赖于以表格形式组织的结构化数据。为了对大量信息进行可靠、

、可扩展的整体管理，应用程序和系统开发人员正在转向关系数据库管理系统（RDBMS ）。

关系数据库 **RDBMS** 和 **SQL** 查询语言

为了有效地存储、处理和分析数据，关系数据库（**RDBMS** ）是一种数据存储系统，它将信息组织到表格中，并定义了表格之间的关系。

> 组织在数据库（RDBMS ）中的数据不仅仅是数字信息，它们还是不同系统之间交易和交互的基础。

下面是几个最常见的关系数据库管理系统（RDBMS）（图 3.1-5）：

- **MySQL**（开源） 是最流行的 RDBMS 之一，是 LAMP 堆栈（Linux、Apache、MySQL、PHP /Perl/Python ）的一部分。由于其简单和高性能，它被广泛用于网络开发。
- **PostgreSQL** （开源） 是一个功能强大的对象关系系统，以其可靠性和高级功能而著称。它适用于复杂的企业解决方案。
- **Microsoft SQL Server** 是微软公司推出的商业系统，因其与公司其他产品的集成性和高度安全性，在企业环境中被广泛使用。
- **Oracle** 数据库是大型企业和关键任务应用程序中使用的最强大、最可靠的 DBMS 之一。
- **IBM DB2** - 面向大型企业，提供高性能和容错功能。
- **SQLite** (Open Source) 是一种轻量级嵌入式数据库 ，非常适合移动应用和独立系统，如 CAD 设计软件 (BIM)。

建筑企业中流行的数据库管理系统--MySQL 、PostgreSQL 、Microsoft SQL Server、Oracle® Database 、IBM® DB2 和 SQLite --可处理结构化数据。所有这些 DBMS 都是强大而灵活的解决方案，可用于管理从小型网站到大型企业系统的各种业务流程和应用程序（图 3.2-1）。

> 根据 Statista [48]，关系数据库管理系统（RDBMS）约占 2022 年使用的 DBMS 总量的 72%。



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

图 3.1-5 在 DBMS 排名中使用结构化数据库（蓝色标记）的受欢迎程度（基于 [49]）。

即使没有丰富的技术知识，安装开源数据库 也非常容易。开源系统 ，如 PostgreSQL 、MySQL 或 SQLite ，都是免费提供的，可在 Windows、macOS 和 Linux 等大多数操作系统上运行。您只需访问项目的官方网站，下载安装程序并按说明操作即可。大多数情况下，安装时间不超过 10-15 分钟。我们将在本书的第四部分建立一个这样的数据库（图 4.3-8）。

如果贵公司使用云服务（例如，亚马逊网络服务、谷歌云或微软 Azure），只需点击几下即可部署数据库--平台将为您提供现成的安装模板。由于代码的开放性，此类数据库很容易根据您的任务进行定制，而且庞大的用户社区总能帮助您找到任何问题的解决方案。

RDBMS 仍是大量业务应用程序和分析平台的基础（图 3.1-6），这些应用程序和平台使公司能够高效地存储、处理和分析数据，从而及时做出明智的决策。

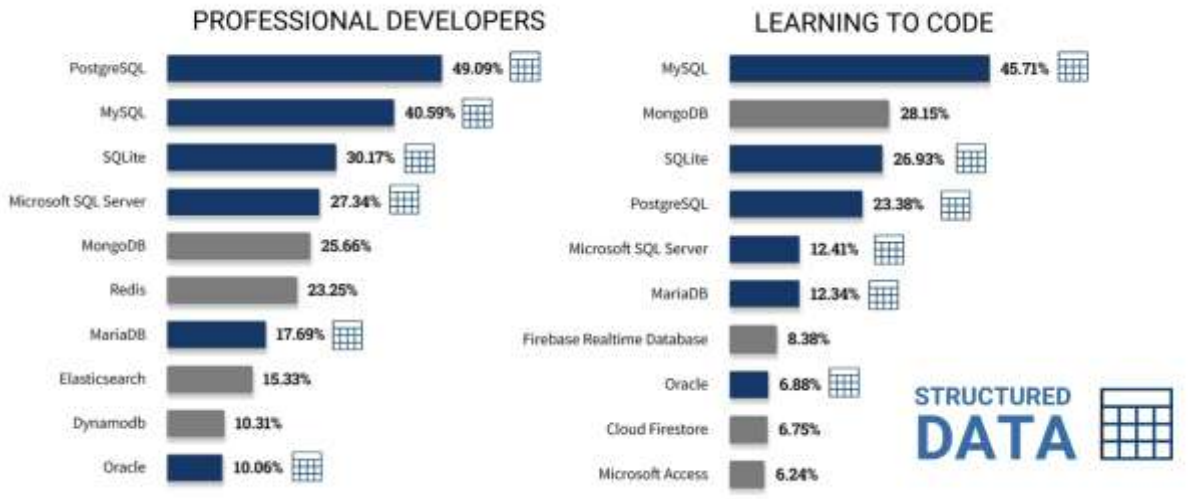


图 3.1-6 在 StackOverFlow（最大的 IT 论坛）上对开发人员进行调查，了解他们去年使用了哪些数据库，明年想使用哪些数据库（RDBMS 用蓝色标出）（基于 [50]）。

RDBMS 提供可靠性、数据一致性和事务支持，并使用功能强大的查询语言--SQL （结构化查询语言），该语言常用于分析，可让您轻松获取、修改和分析数据库中存储的信息。SQL 是在关系系统中处理数据的主要工具。

SQL - 数据库中的查询和新趋势

SQL 语言 通常用于关系数据库，与其他类型的信息管理相比（例如，借助传统的 Excel 电子表格 ），其主要优势在于可以高速处理查询，支持超大容量的数据库。

> 结构化查询语言（SQL）是一种专门的编程语言，用于存储、处理和分析关系数据库中的信息。SQL 用于创建、管理和访问数据，让您可以高效地查找、过滤、组合和汇总信息。它是访问数据的重要工具，为与信息存储进行交互提供了一种方便、正规的方式。

SEQUEL-SQL 系统的发展经历了 Oracle、IBM DB2、Microsoft SQL Server、SAP、PostgreSQL 和 MySQL 等重要产品和公司，最终出现了 SQLite 和 MariaDB [51]。SQL 提供 Excel 中没有的电子表格功能，使数据处理更具可扩展性、更安全、更易于自动化：

- 创建和管理数据结构（**DDL**）：在 SQL 中，您可以创建、修改和删除数据库中的表格，在表格之间

建立链接，并定义数据存储结构。而 Excel 则使用固定的工作表和单元格，工作表和数据集之间没有 明确定义的关系。

- 数据操作 **(DML)**：SQL 通过执行带有筛选、排序和表连接的复杂查询，可以高速大量添加、修改、删除和检索数据（图 3.1-7）。在 Excel 中，处理大量信息需要手动操作或使用特殊的宏，这会减慢处理速度并增加出错的可能性。
- 访问控制（**DCL**）：SQL 允许您区分不同用户对数据的访问权限，限制编辑或查看信息的能力。而在 Excel 中，访问权限要么是共享的（传输文件时），要么需要通过云服务进行权限共享的复杂设置。

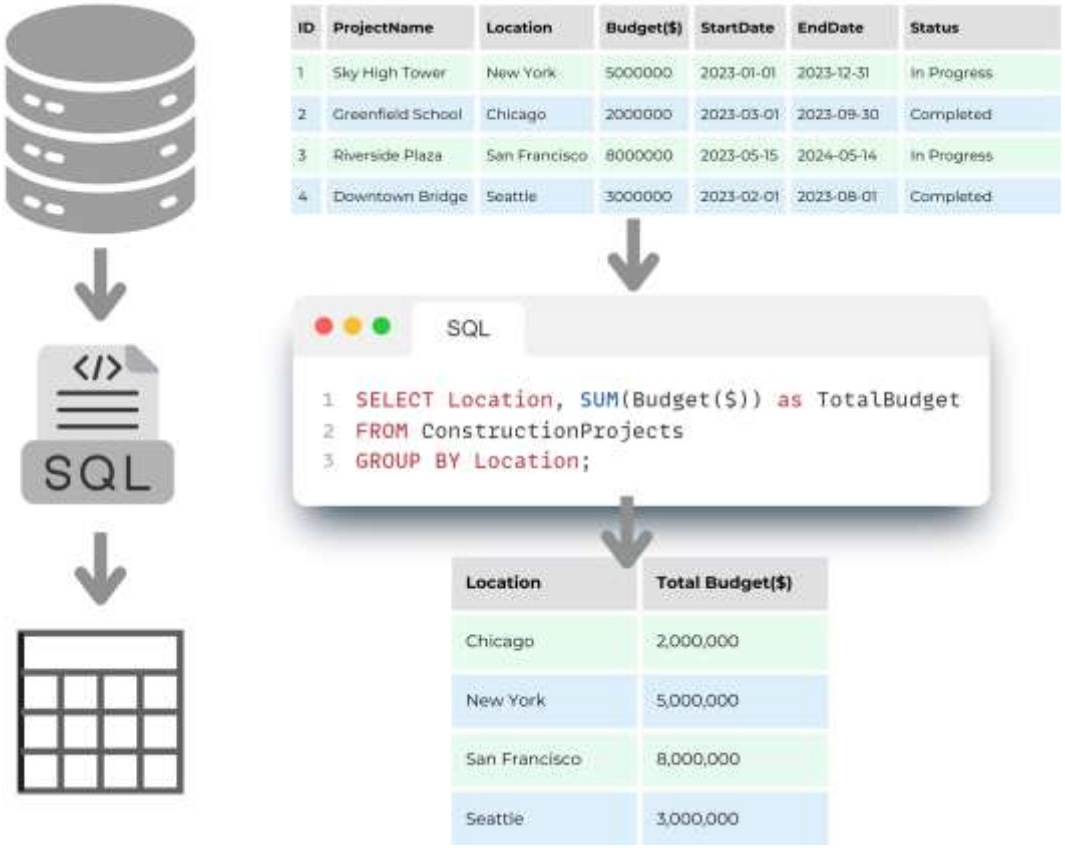


图 3.1-7 SQL 中的 DML 示例： 只需几行代码即可实现快速处理、分组和聚合，实现自动数据处理。

Excel 以其可视化和直观的结构使数据处理变得更容易。然而，随着数据量的增加，Excel 的性能也在下降。Excel 还面临着数据存储量的限制--最多只能存储 100 万行，而在达到这个限制之前，性能早已下降。因

此，Excel 看起来更适合可视化和处理少量数据，而 SQL 则更适合处理大型数据集。

结构化数据发展的下一阶段是列式数据库（Columnar Databases）的出现，它是传统关系数据库 的替代品，尤其是在数据量显著增大和分析计算时。与逐行存储数据的行式数据库不同，列式数据库按列记录信息。与传统数据库相比，列式数据库允许

- 通过有效压缩列中的统一数据，减少存储空间。
- 加快分析查询速度，因为只读取所需列，而不是整个表。
- 优化大数据 和数据仓库，如 Data Lakehouse Architecture。

关于列式数据库、Pandas DataFrame 、Apache Parquet 、HDF5 以及基于它们创建用于数据分析和处理目的的大数据 - 存储，我们将在本书的后续章节--"DataFrame：一种通用的表格数据格式 "和 "数据存储格式和使用 Apache Parquet: DWH -数据仓库 和数据湖泊架构 "中详细介绍。

非结构化数据

虽然应用软件和信息系统中使用的大部分数据都是结构化形式的，但建筑工程中产生的大部分信息都是非结构化数据--图像、视频、文本文档、录音和其他形式的内容。尤其是在施工、运行和技术监督阶段，视觉和文本信息占主导地位。

> 非结构化数据是指没有预定义模型或结构的信息，不像数据库或表格那样按照传统的行和列进行组织。

一般来说，非结构化数据可分为两类：

- 人工生成的非结构化数据，包括各类人工生成的内容：文本文档、电子邮件、图像、视频等。
- 机器生成的非结构化数据由设备和传感器创建：例如，这些数据包括日志文件、全球定位系统数据、物联网结果（物联网 ）和来自建筑工地的其他遥测信息。

非结构化数据与结构化数据不同，结构化数据可以方便地组织到表格和数据库中，而非结构化数据则需要额外的处理步骤才能整合到信息系统中（图 3.1-8）。使用自动收集、分析和转换此类数据的技术为提高施工效率、减少错误和最大限度地降低人为因素的影响提供了新的机遇。

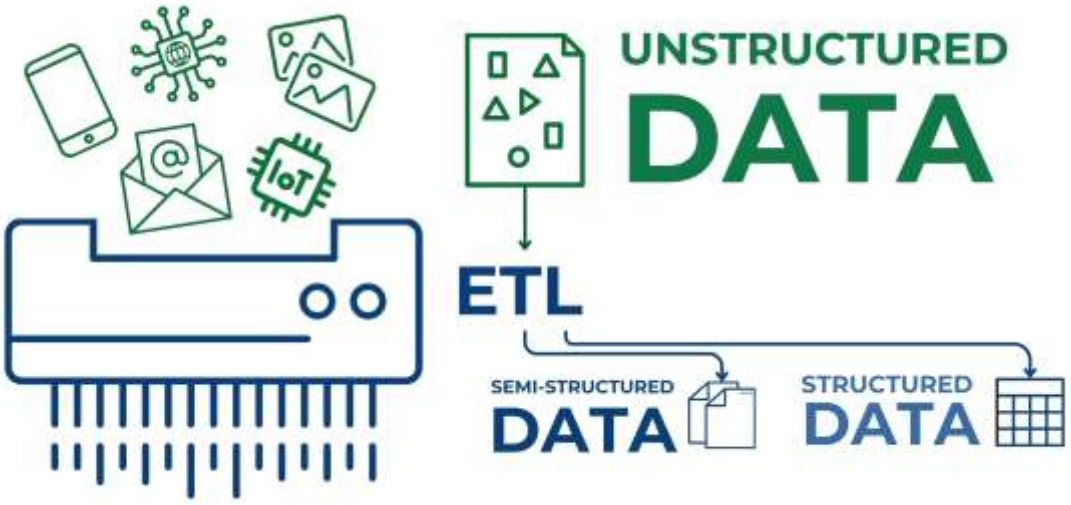


图 3.1-8 处理非结构化数据首先要将其转换为半结构化数据和结构化数据。

非结构化数据占公司专业人员接触到的所有信息的 80% [52]，因此我们将在本书接下来的章节中结合实例详细讨论非结构化数据的类型和处理方法。

为便于讨论，文本数据单独分类。虽然文本数据属于非结构化数据的一种，但其在建筑业中的重要性和普遍性需要特别关注。

文本数据：介于非结构化混沌和结构化混沌之间 **y**

建筑行业的文本数据 涵盖多种格式和类型的信息，从纸质文件到非正式的沟通方式，如信件、对话、工作信函和施工现场的口头会议。所有这些文本数据都承载着管理建筑项目的重要信息，从设计决策和计划变更的细节，到安全问题的讨论以及与承包商和客户的谈判（图 3.1-9）。



图 3.1-9 文本数据 ，是项目参与者之间交流中最常用的信息类型之一。

文本信息既可以是正式的，也可以是非结构化的。正式数据包括 Word 文档（.doc、.docx）、PDF 以及会议记录文本文件（.txt）。非正式数据包括信使和电子邮件通信、会议记录（Teams、Zoom、Google Meet）以及需要转换为文本的讨论录音（.mp3、.wav）。

但是，虽然正式请求、合同条款和条件以及电子邮件等书面文件通常已经具有一定的结构，但口头沟通和工作信函往往仍然是非结构化的，因此很难对其进行分析并将其整合到项目管理系统中。

有效管理文本数据的关键是将其转换为结构化格式。这样，经过处理的信息就可以自动集成到已经可以处理结构化数据的现有系统中。



图 3.1-10 将文本内容转换为结构化数据。

要有效利用文本信息，必须将其自动转换为结构化形式（图 3.1-10）。这一过程通常包括几个步骤：

- 文本识别（**OCR**）--将文件和图纸的图像转换成机器可读的格式。
- 文本分析（**NLP**）--自动识别关键参数（与项目相关的日期、金额和数字）。
- 数据分类 - 信息分类（财务、物流、风险管理）。

经过识别和分类已经结构化的数据可以整合到数据库中，并用于自动报告和管理系统。

半结构化和松散结构化数据

半结构化数据 包含一定程度的组织，但没有严格的模式或结构。虽然此类信息包含结构化元素（如 日期、员工姓名和已完成任务清单），但不同项目甚至不同员工的表述格式可能会有很大不同。时间日志、进度报告和计划表就是此类数据的例子，它们可以以多种格式呈现。

半结构化数据 比非结构化数据更容易分析，但需要额外的处理才能整合到标准化的项目管理系统中。

> 处理结构不断变化的半结构化数据是一项重大挑战。这是因为数据结构的多变性要求我们采用不同的方法来处理和分析每个半结构化数据源。

但是，处理非结构化数据需要耗费大量精力，而处理半结构化数据则可以使用相对简单的方法和工具。

> 弱结构化数据 是一个更笼统的术语，用于描述结构极小或不完整的数据。这些数据通常是文本文档、聊天记录、电子邮件，其中可以找到一些元数据（如日期、发件人），但大部分信息都是以混乱的方式呈现的。

在建筑业中，松散结构数据存在于各种流程中。例如，它们可能包括

- 估算 和报价单 - 包含材料、数量和成本数据的表格，但没有统一格式。
- 图纸和工程图表 - PDF 或 DWG 格式的文件，包含文本注释和元数据，但没有严格固定的结构。
- 工作计划 - MS Project 、Primavera P6 或其他系统的数据，这些系统可能有不同的导出结构。
- CAD (BIM -models) - 包含结构元素，但数据表示取决于软件和项目标准。

由 CAD 系统产生的几何数据 ，可以与半结构化数据进行相同的分类。不过，我们将把几何 CAD（BIM）数据归为单独的数据类型，因为它与文本数据一样，在公司流程中通常可作为单独的数据类型处理。

几何数据 及其应用

有关项目要素的元数据几乎总是以表格、结构化或松散结构化格式的形式存储，而 中项目要素的几何数据在大多数情况下是使用专门的 CAD 工具创建的（图 3.1-11），可以将项目要素详细地可视化为一组线条（2D）或几何体（3D ）。



图 3.1-11 CAD 工具有助于将几何信息从物理介质转移到数据库形式。

在施工和建筑中使用几何数据时，可以确定几何数据的三种主要应用（图 3.1-12）：

- 体积的确认：需要在 CAD 程序 (BIM) 中使用特殊几何核 生成几何数据 ，以自动准确地确定项目元素的体积和尺寸。这些数据包括自动计算的面积、体积、长度以及规划、预算和订购资源和材料所需的其他重要属性
- 项目的可视化 ： 如果项目发生任何变化，元素的可视化可自动生成不同平面的更新图纸。在初始阶段对项目进行可视化有助于加快所有参与者之间的理解，从而在施工过程中节省时间和资源。
- 检查 碰撞：在复杂的建筑和工程项目中，多类元素（如管道和墙壁）之间的相互作用而不发生 "几何冲突 "至关重要，因此碰撞检查起着关键作用。使用 碰撞检测软件可以主动识别项目元素之间潜在的几何冲突，防止在施工过程中出现代价高昂的错误。

从工程设计院成立之初，从建造第一个复杂结构开始，结构工程师就以图纸、线条和平面几何元素（纸莎草纸、"A0 "纸或 DWG 、PDF 、PLT 格式）的形式提供几何信息��工头 和估算师（图 3.1-11) ，在过去的几千年里，借助标尺和量角器收集元素和元素组的属性体积或数量。3.1-11) ，在过去的几千年里，他们借助标尺和运输工具收集元素和元素组的属性体积或数量。

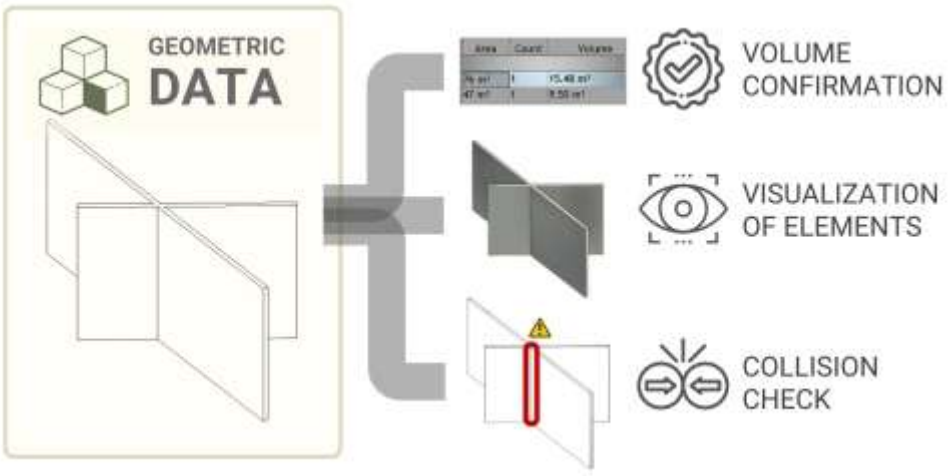


图 3.1-12 几何图形是获得元素体积参数的基础，然后用这些参数来计算项目的成本和时间。

如今，由于现代 CAD 工具 (BIM) 中体积建模功能的出现，这项耗时的手动任务得以全自动解决，它可以借助特殊的几何内核自动获取任何元素的体积属性，而无需手动计算体积参数。

现代 CAD 工具 还可以对项目要素进行分类和归类，以便从项目数据库上传规格表，供成本估算、进度安排或 CO_2 计算等各种系统使用 （图 3.1-13）。我们将在 "获取工程量和量化 "一章中讨论获取规格、QTO 表 和工程量以及实际例子。

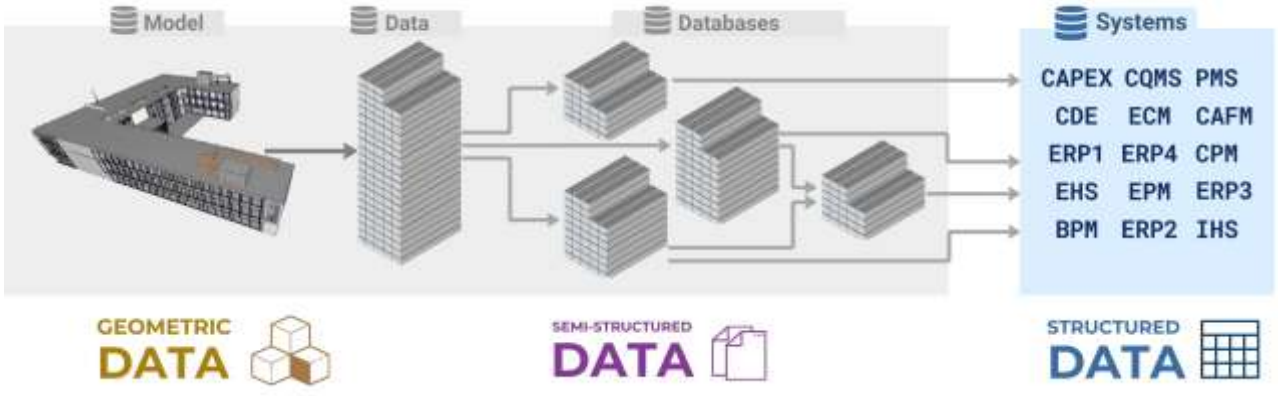


图 3.1-13 CAD 工具 (BIM) 将数据存储在数据库中，旨在与其他系统集成和交互。

由于 CAD 环境中使用的数据库和格式的封闭性，CAD 解决方案中创建的几何数据实际上已成为一种独立的信息类型。它结合了元素的几何形状和元信息（结构化或半结构化），并封装在专门的文件和格式中。

CAD 数据：从设计到数据存储

现代 CAD 和 BIM 系统通常以自己的专有格式存储数据：DWG、DXF、RVT、DGN、PLN 等。这些格式支持对象的二维和三维表示，不仅保留了几何图形，还保留了与对象相关的属性。以下是最常见的几种格式：

- **DWG** 是一种二进制文件格式，用于存储二维（三维较少见）设计数据和元数据。
- **DXF** 是一种文本格式，用于在 CAD 系统之间交换 2D 和 3D 图纸。它包含几何图形、图层和属性数据，支持 ASCII 和二进制表示。
- **RVT** 是一种二进制格式，用于存储 CAD 模型，包括三维 - 几何图形、元素属性、关系和设计参数。
- **IFC** 是一种开放文本格式，用于在 CAD (BIM) 系统之间交换建筑数据。它包括几何图形、对象属性及其关系信息。

除此之外，还使用了其他格式：PLN、DB1、SVF、NWC、CPIXML、BLEND、BX3、USD、XLSX、DAE。尽管它们的目的和开放程度不同（图 3.1-14），但它们都能以不同的形式表示相同的项目信息模型。在复杂的项目中，从绘制项目模型到协调项目模型，这些格式经常并行使用。

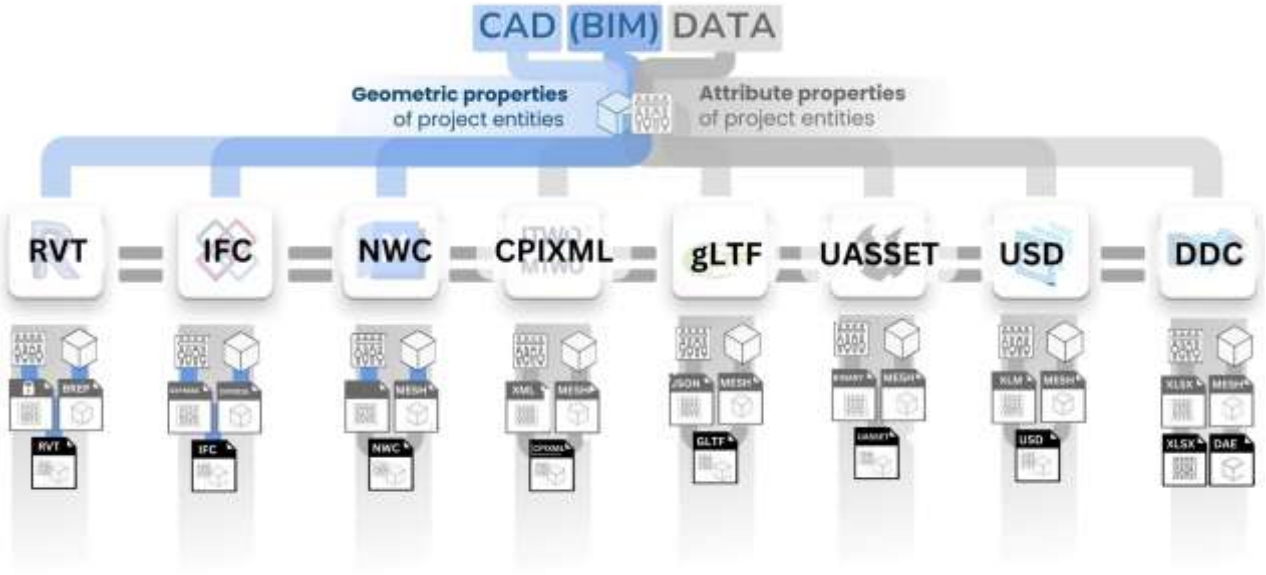


图 3.1-14 常用的 CAD 存储格式通过 BREP 或 MESH 参数描述几何体，并辅以属性数据。

上述所有格式都允许您存储有关建筑项目各要素的数据，而且上述所有格式都包含两类关键数据：

- 几何参数 - 描述对象的形状、位置和尺寸。本书第六部分 "CAD(BIM)解决方案 "将详细讨论几何参数

及其使用；

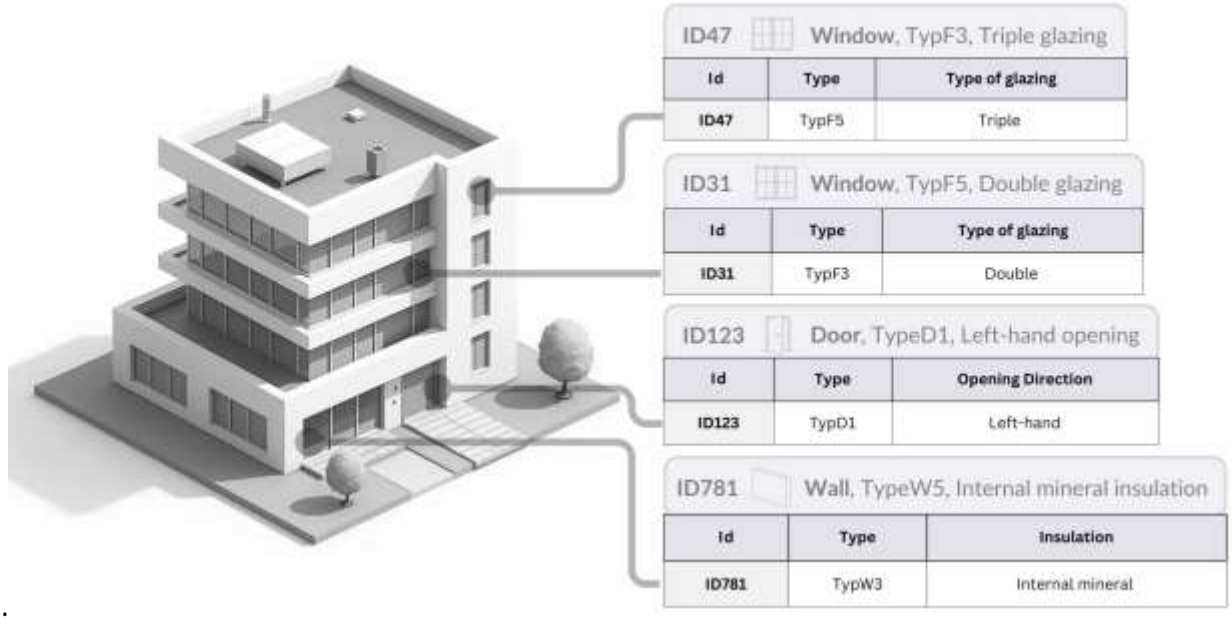
- 属性属性--包含各种信息：材料、要素类型、技术特征、唯一标识符以及项目要素可能具有的其他属性。

属性数据在现代项目中尤为重要，因为它们定义了对象的运行特征，可用于工程和成本计算，并提供设计、施工和运行参与者之间的端到端交互。例如

- 门窗：结构类型、玻璃类型、开启方向（图 3.2-1）。
- 对于墙壁，记录了有关材料、隔热和隔音性能的信息。
- 对于工程系统，可存储管道、导管、电缆线路及其连接的参数。

这些参数既可以存储在 CAD(BIM -)文件本身中，也可以存储在外部数据库中--作为导出、转换或通过逆向工程工具直接访问内部 CAD 结构的结果。这种方法有助于将设计信息与其他企业系统和平台整合

> 计算机辅助设计（BIM）中的逆向工程是提取和分析数字模型内部结构的过程，目的是在无法访问原始算法或文档的情况下，重新创建其逻辑、数据结构和依赖关系。



.

图 3.1-15 项目元素除了描述参数或多边形几何图形外，还包含有关元素参数和属性的信息。

因此，围绕每个要素形成了一套独特的参数和属性，包括每个对象的独特特征（如标识符和尺寸）和要素组

的共同属性。这样，不仅可以分析项目的单个元素实体，还可以将它们组合成逻辑组，供其他专家在系统和数据库中执行任务和进行计算时使用。

> 实体是现实世界中的一个具体或抽象的物体，可以以数据的形式进行唯一识别、描述和表示。

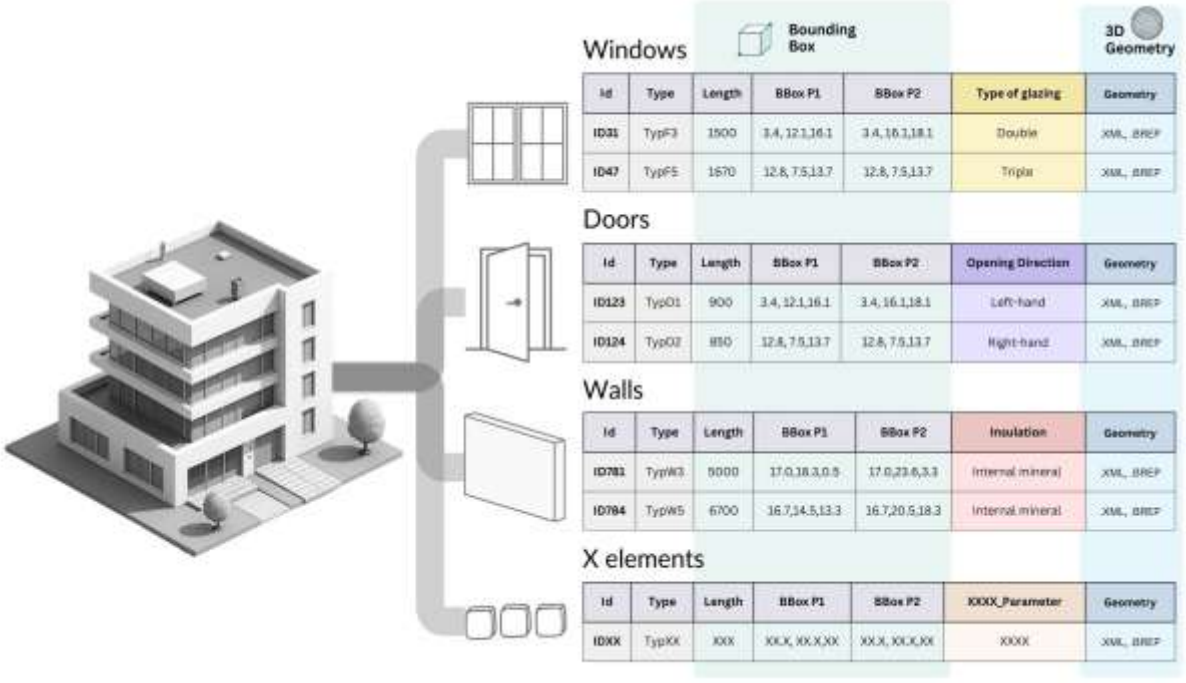


图 3.1-16 每个项目元素都包含由设计人员输入或在 CAD 程序中计算得出的属性。

过去几十年来，建筑行业开发了许多新的 CAD(BIM)格式，简化了数据的创建、存储和传输。这些格式既可以是封闭式的，也可以是开放式的，既可以是表格格式，也可以是参数格式或图形格式。然而，这些格式的多样性和分散性大大增加了项目生命周期各阶段数据管理的复杂性。图 3.1-17 列出了用于建筑信息交换的主要格式对照表（完整版可通过二维码获取）。

为了解决互操作性 和获取 CAD 数据 的问题，需要 BIM 管理人员（BIM ）和协调人员的参与，他们的任务是控制导出、检查数据质量并将部分 CAD（BIM）数据集成到其他系统中。

然而，由于格式的封闭性和复杂性，这一过程很难实现自动化，专家们不得不手动执行许多操作，无法建立完整的在线数据处理流程（管道）。

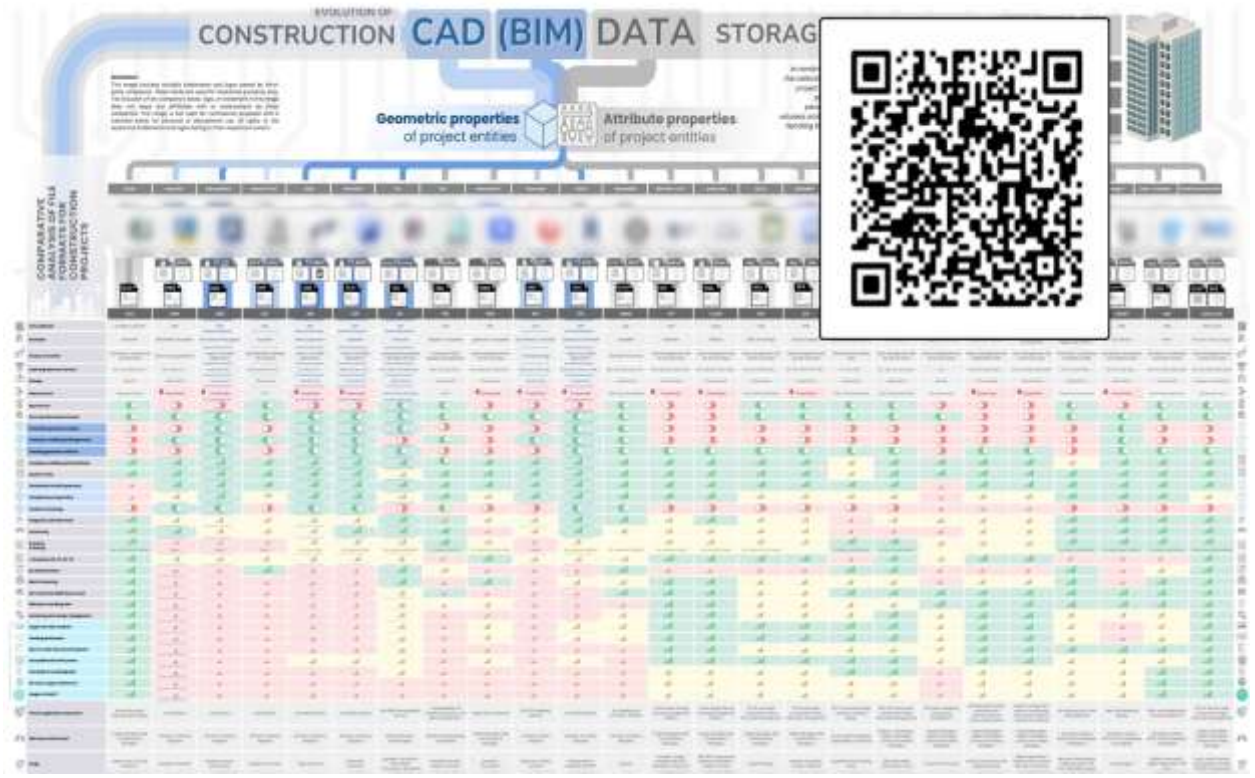


图 3.1-17 项目要素信息主要数据格式比较表[53]。

要了解为什么会有这么多不同的数据格式，以及为什么其中大多数都是封闭的，就必须深入研究 CAD (BIM) 程序内部的流程，本书第六部分将对此进行详细探讨。

CAD 系统开发商以 BIM 概念 （建筑信息模型）的形式为几何图形添加了一个额外的信息层，这是建筑行业自 2002 年以来积极推广的一个营销术语[54]。

BIM (BOM)概念的出现 以及在流程中使用 CAD

建筑信息模型（BIM ）的概念在 2002 年的 BIM 白皮书[54]中首次提出，源于 CAD 软件制造商的市场营销活动。它源自 CAD 软件开发商的市场推广活动，试图将机械工程中早已确立的原则应用于建筑行业的需求。

BIM 的灵感来自 BOM（物料清单）--物料清单的概念，自 20 世纪 80 年代末以来，BOM 在工业领域得到了广泛应用。在机械工程领域，BOM 可以将 CAD 系统中的数据与 PDM（产品数据管理）、PLM（产品生命周期管理）和 ERP 系统连接起来，从而对整个产品生命周期的工程信息进行整体管理（图 3.1-8）。

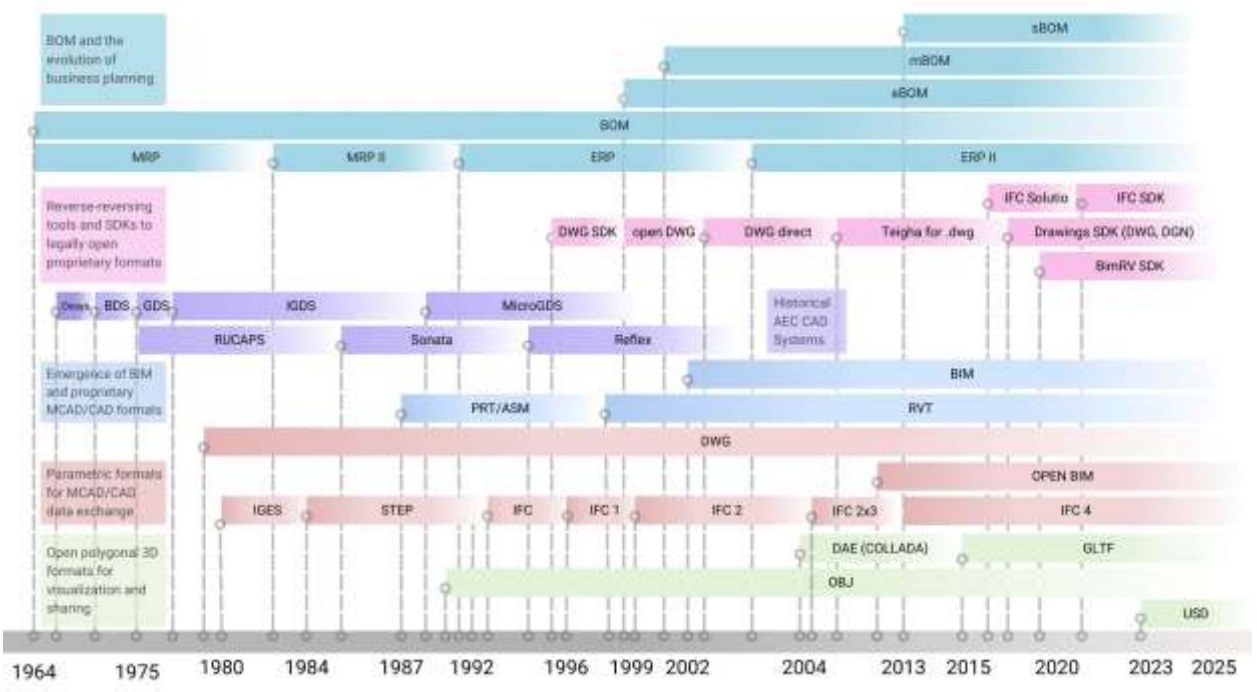


图 3.1-18 工程建筑行业规格（BOM ）、信息建模（BIM ）和数字格式的演变。

BOM 概念的现代发展导致出现了一个扩展框架 - XBOM（扩展 BOM），其中不仅包括产品构成，还包括行为情景、操作要求、可持续性参数和用于预测分析的数据。XBOM 在本质上与 BIM 在建筑领域的作用相同：这两种方法都旨在将数字模型转化为项目生命周期内所有项目参与者的 "单一真实来源"。

BOM 在建筑业兴起的一个重要里程碑是 2002 年推出的第一个专门针对建筑业的参数化 CAD（MCAD）。该系统由创建 Pro-E® 的团队开发，Pro-E® 是 20 世纪 80 年代末出现的用于机械工程的革命性 MCAD 系统，已成为行业标准[55]。

早在 20 世纪 80 年代末，人们就开始致力于消除当时的 CAD 程序的局限性[56]。其主要目标是减少更改设计元素参数所需的人力，并通过数据库[57]使根据 CAD 程序之外的数据更新模型成为可能。其中最重要的作用是参数化：从数据库中自动检索特征，并利用这些特征在 CAD 系统内更新模型。

> Pro-E 及其所依据的要素参数化建模 c BOM 概念对 CAD 和 MCAD 市场的发展产生了重大影响[58]。25 年来，该模型一直在行业中使用，许多现代系统已成为其概念的继承者。

> 我们的目标是创建一个足够灵活的系统，鼓励工程师轻松考虑不同的设计。对设计进行更改的成本应尽可能接近于零。传统的 CAD / CAM 软件仅在设计流程的最开始阶段才限制进行低成本的修改，这是不现实的[59]。
>
> - Samuel Heisenberg，参数技术公司® 的创始人，MCAD 产品 Pro-E 的开发者，以及使用 RVT 格式的 CAD 产品创造者的老师

在机械工程领域，PDM、PLM、MRP 和 ERP 系统已成为关键平台。它们在数据和流程管理中发挥着核心作用，从 CAx 系统（CAD、CAM、CAE）收集信息，并根据产品结构（BOM：eBOM、pBOM、mBOM）组织设计活动（图 3.1-18）。这种集成可减少错误，避免数据重复，并确保从设计到生产的端到端可追溯性。

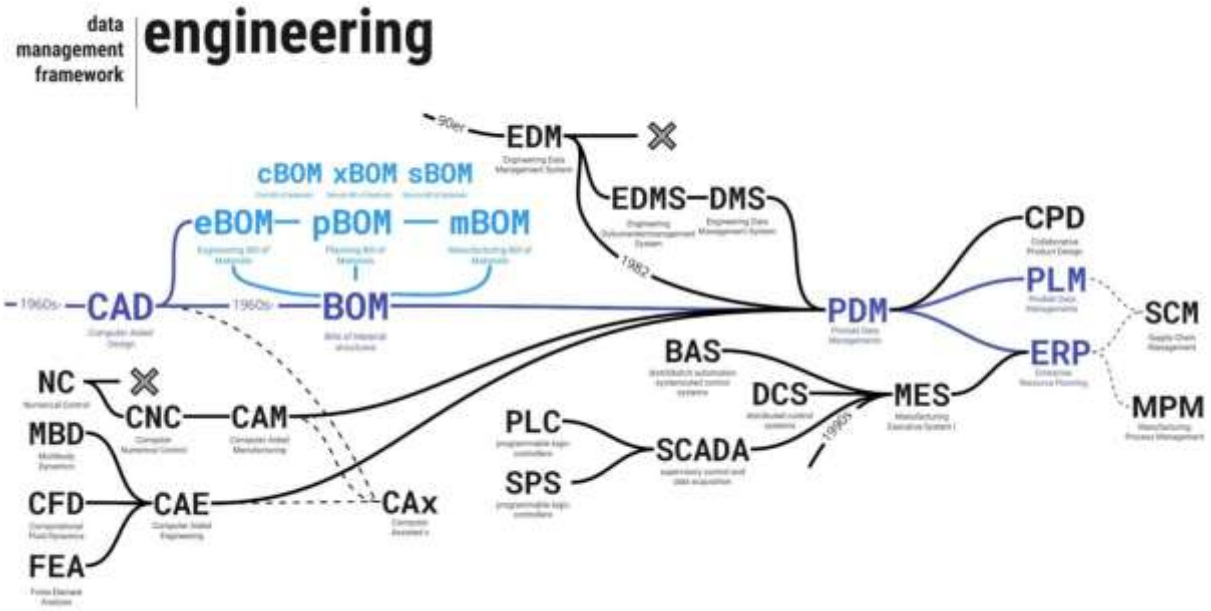


图 3.1-19 从历史上看，BOM 出现于 20 世纪 60 年代，是将 CAx 系统中的数据结构化并传递给控制系统的一种方法。

一家领先的供应商购买了由前 Pro-E 团队开发的基于 BOM 方法的 CAD 解决方案，其标志是 BIM 白皮书系列 （2002-2003 年）的出版[60][61]。从 2000 年代中期开始，建筑行业就开始积极推广 BIM 概念，这显著

提高了人们对参数化软件的兴趣。其普及速度如此之快，以至于机械工程 Pro-E 的建筑分叉（该供应商推广的参数化 CAD）实际上已经取代了建筑和结构设计领域的竞争对手（图 3.1-20）。到 2020 年代初，该公司事实上已经巩固了其在 BIM（CAD）市场的全球主导地位[62]。



图 3.1-20 谷歌搜索查询流行度（RVT 与 IFC）：前 Pro-E 团队创建的参数化 CAD，支持 BOM - BIM 在世界上几乎大多数国家都得到了普及。

在过去的 20 年中，BIM 这一缩写获得了多种解释，其多重含义源于本世纪初出现的最初营销概念。ISO 19650 标准在普及该术语方面发挥了重要作用，实际上确保了 BIM 作为一种 "以科学为基础 "的信息管理方法的地位。然而，在专门针对使用 BIM 的对象的整个生命周期的数据管理的标准文本本身中，虽然提到了缩写 BIM，但从未对其进行明确定义

该供应商最初的网站在 2002 年[60] 和 2003 年[61] 发布了一系列关于 BIM 的白皮书，实际上转载了早在 20 世纪 90 年代 Pro-E 机械工程软件中使用的 BOM（物料清单）和 PLM（产品生命周期管理）概念的营销材料[63]。

> 建筑信息模型是..... [CAD 供应商公司名称] 于 2002 年推出的一种创新的建筑设计、施工和管理新方法，它改变了全球业界专业人士对如何将技术应用于建筑设计、施工和管理的看法。
>
> - BIM 白皮书, 2003 [61]

这些早期出版物将 BIM 与中央集成数据库的概念直接联系起来。正如 2003 年白皮书所述，BIM 是一种建筑信息管理，所有更新都在一个单一的存储库中进行，确保所有图纸、切割和规格（BOM - 材料清单）同步。

> BIM 被描述为建筑信息管理，所有更新和更改都在数据库中进行。因此，无论您处理的是示意图、剖面图还是图纸，一切都始终是协调、一致和最新的。
>
> - CAD 公司网站 供应商与 BIM 白皮书，2003 年 [54].

早在 20 世纪 80 年代的研究中，就已经广泛讨论了通过单一集成数据库管理设计的想法。例如，查尔斯-伊士曼的 BDS 概念[57]中有 43 处提到 "数据库 "一词（图 6.1-2）。到 2004 年，这一数字几乎减半，在 2002 年的 BIM 白皮书 [64]中只有 23 次。而到了 2000 年代中期，数据库的话题几乎从供应商的营销材料和整个数字化议程中消失了。

虽然数据库和对数据库的访问最初被认为是 BIM 系统的核心，但随着时间的推移，重点已转移到几何、可视化和 3D 方面。1994 年 IFC 标准的注册者，也就是 2002 年发布 BIM 白皮书的供应商，在本世纪初的白皮书中明确指出了 IGES、STEP 和 IFC 等中性格式的局限性以及直接访问 CAD 数据库的必要性：

> 不同的应用程序可能不兼容，重新输入的数据可能不准确[……]。传统计算机辅助设计（CAD）的结果是：成本增加、上市时间延长、产品质量降低。如今，所有主要应用程序都使用行业标准接口进行底层数据交换。通过使用旧的 IGES 标准或新的 STEP [IFC 是 STEP/IGES 格式事实上和法律上的复制品] 在不同供应商的应用程序之间交换数据，用户可以在同类产品之间实现一定的数据兼容性。但是，IGES 和 STEP 只能在低层次上工作，它们无法交换与当今领先的应用程序所生成的信息一样丰富的数据[……]。虽然这些标准和其他标准几乎每天都在改进，但它们在数据丰富性方面始终落后于当今的供应商产品。[……]应用程序中的程序必须能够交换和保留丰富的数据，而无需使用 IGES、STEP [IFC] 或 PATRAN 等中性转换器。相反，框架应用程序应能直接访问底层 CAD 数据库，这样就不会丢失信息的细节和准确性。
>
> - CAD 供应商白皮书（IFC，BIM ）"集成设计与制造：优势与理由"，2000 年 [65]。

因此，早在 20 世纪 80 年代和 21 世纪初，CAD 环境中数字设计的关键要素就被认为是数据库，而不是格式文件或中立的 IFC 格式。有人建议，应放弃翻译器，应用程序应能直接访问数据。但实际上，到 20 世纪 20 年代中期，BIM 的概念开始类似于 "分而治之 "的战略，即使用封闭几何内核的软件供应商的利益优先于开放信息交换的发展。

如今，BIM 已被视为建筑行业不可或缺的一部分。但在过去二十年中，简化协作和数据集成的承诺在很大程度上未能实现。大多数解决方案仍然受制于封闭或中立的格式和专用工具。我们将在《CAD 和 BIM：建筑设计数据的营销、现实和未来》一书的第 6 部分详细介绍 BIM 的历史、开放式 BIM 和 IFC，以及互操作性和几何内核问题。

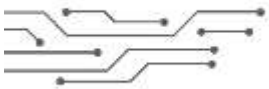
> 如今，该行业面临着一项重要挑战，即从传统意义上的将 CAD (BIM) 作为建模工具，转变为将其作为一个完整的数据库来使用。这就需要采用新的方法来处理信息，放弃对封闭生态系统的依赖，实施开放式解决方案。

随着可访问 CAD 数据库的逆向工程工具的开发，以及开放源代码和 LLM 技术的普及，建筑行业的用户和开

发人员正逐渐摆脱软件供应商的模糊术语。相反，重点正在转向真正重要的方面：数据（数据库）和流程。

> 在这些时髦的缩略语和可视化的背后，是标准的数据管理实践：存储、传输和转换--即经典的 **ETL** 流程 （提取 、转换、加载 ）。与其他行业一样，建筑业的数字化不仅需要交换标准，还需要对异构信息进行清晰的结构化处理。

为了充分利用 **CAD (BIM)** 数据的潜力，企业需要重新思考其信息管理方法。这将不可避免地导致数字化转型的一个关键因素--统一、标准化和有意义地构建建筑专业人员日常工作的数据。



章 3.2.

数据统一和结构化

在建筑业系统中填入数据

无论是大型企业还是中型企业，专家们每天都要在软件系统和数据库的各种界面上填充多格式信息（图 3.2-1），这些系统和数据库必须在管理人员的帮助下相互配合。正是这些相互作用的系统和流程最终为公司创造了收入和利润。

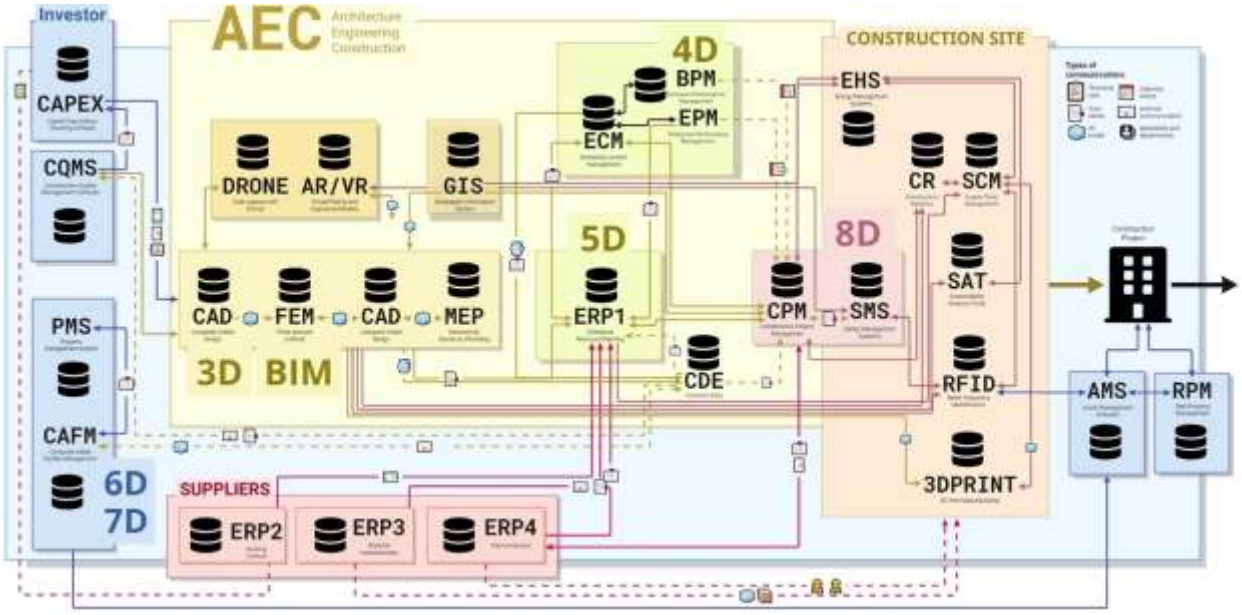


图 3.2-1 实际上，建筑行业中使用的每个系统或应用程序的核心都是一种流行的 RDBMS 数据库。

前面提到的应用于建筑行业的各类系统都有自己的数据类型，与这些系统的功能作用相对应。从抽象层面到具体层面，我们从数据类型到数据格式和文件的表示。

在之前提供的系统列表（图 1.2-4）中，我们现在增加了它们经常使用的特定格式和文件类型：

- 投资者 **(CAPEX)**
  - 财务数据：预算、支出预测（结构化数据）。
  - 市场趋势数据：市场分析（结构化和非结构化数据）。
  - 法律和合同数据：合同（文本数据 ）。

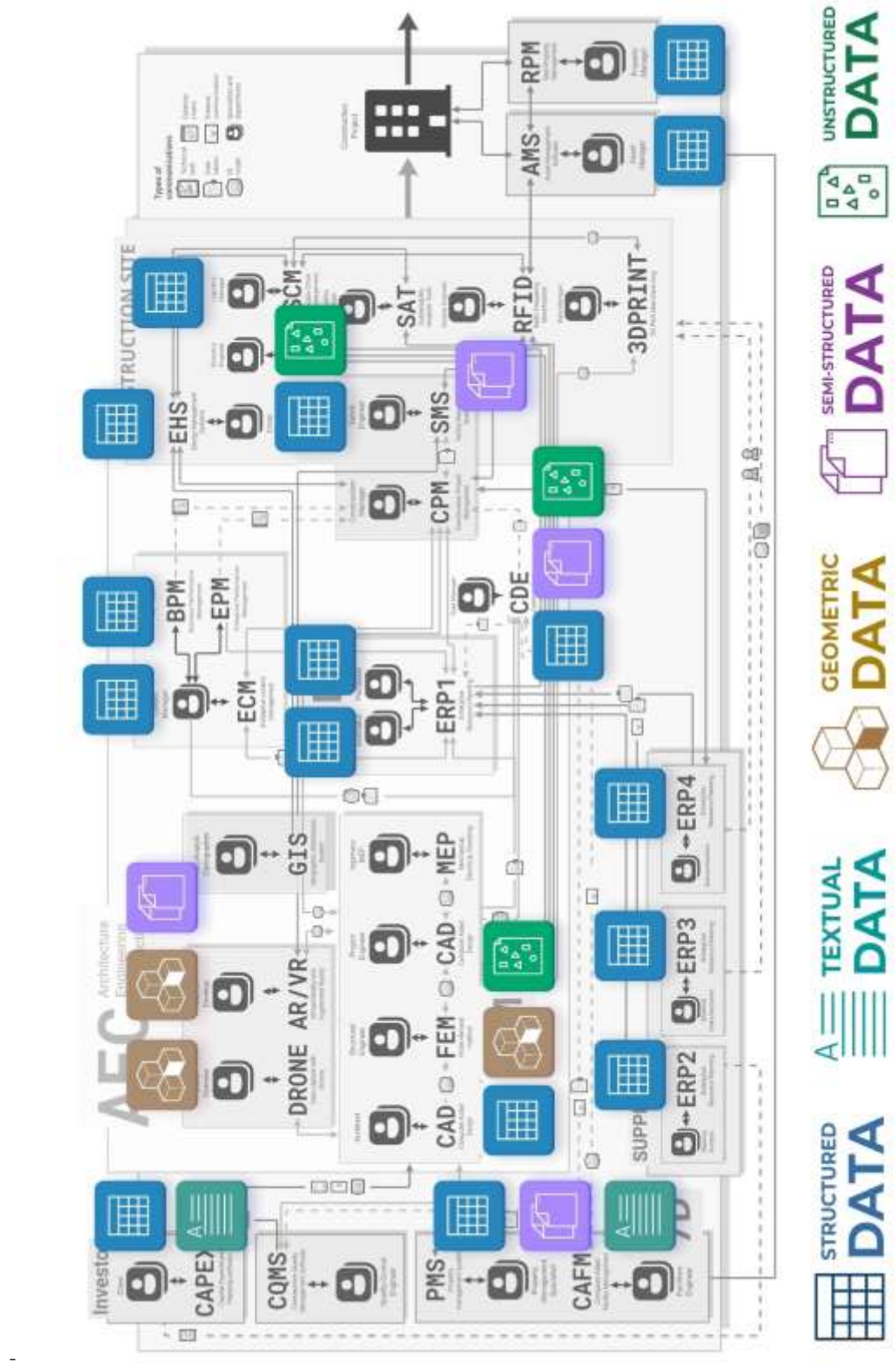


图 3.2-2 建筑行业使用许多具有不同界面的系统来处理不同类型的数据。

- 管理系统（**PMS, CAFM, CQMS)**
  - 项目数据：图表、任务（结构化数据）。
  - 设施维护数据：维护计划（文本和半结构化数据）。
  - 质量控制数据：标准、检验报告（文本和非结构化数据）。
- **CAD** 、**FEM** 和 **BIM**
  - 技术图纸：建筑、结构图（几何数据、非结构化数据）。
  - 建筑模型：3D - 模型、材料数据（几何和半结构化数据）。
  - 工程计算：载荷分析（结构化数据）。
- 施工现场管理系统（**EHS**、**SCM**）
  - 安全与健康数据：安全协议（文本和结构化数据）。
  - 供应链数据：库存、订单（结构化数据）。
  - 每日报告：工作时间、生产率（结构化数据）。
- 无人机、**AR/VR** 、地理信息系统、**3D -** 打印
  - 地理数据：地形图（几何数据和结构化数据）。
  - 实时数据：视频和照片（非结构化数据）。
  - 用于 3D - 打印的模型：数字图纸（几何数据）。
- 其他管理系统（**4D BPM**、**5D ERP1**）
  - 时间和成本数据：时间表、估算 （结构化数据）。
  - 变更管理：项目变更记录（文本和结构化数据）。
  - 绩效报告：成功指标（结构化数据）。
- 数据整合与通信（**CDE, RFID, AMS, RPM)**
  - 数据交换：文件交换、数据模型（结构化数据和文本数据）。
  - RFID 和跟踪数据：物流、资产管理（结构化数据）。
  - 监测和控制：现场传感器 （结构化和非结构化数据）。

因此，建筑行业的每个系统--从工地管理系统到操作数据库--都有自己的信息类型：结构化信息、文本信息、几何信息和其他信息。专业人员每天必须处理的 "数据景观 "极其多样化。然而，简单列举格式并不能揭

示实际信息工作的复杂性。

在实践中，公司面临着这样一个事实，即数据即使从系统中检索出来，也不能 "原封不动 "地使用。对于文本、图像、PDF、CAD 文件和其他难以用标准工具分析的格式，情况尤其如此。这就是为什么下一个关键步骤是数据转换--没有这个过程，处理、分析、可视化和决策就无法有效实现自动化。

数据转换：现代业务分析的重要基础

如今，大多数公司发现自己面临着一个悖论：大约 80% 的日常流程仍然依赖于传统的结构化数据--熟悉的 Excel 电子表格 和关系数据库 (RDBMS)[66]。但与此同时，进入公司数字生态系统的 80% 的新信息都是非结构化或松散结构化的（图 3.2-3）[52]。这包括文本、图形、几何图形、图像、CAD -模型、PDF 文件 、音频和视频记录、电子信件等等。

此外，非结构化数据量持续快速增长--年增长率估计为 55-65%[67]。这种动态变化给将新信息整合到现有业务流程中造成了严重困难。忽视这种多格式数据流会导致信息缺口的形成，并降低公司整个数字环境的可管理性。



图 3.2-3 非结构化数据的逐年增长给将流式信息集成到业务流程中带来了挑战。

> 在自动化流程中忽视复杂的非结构化数据和令人困惑的松散结构化数据，会导致公司的信息环境出现重大漏洞。当今世界的信息流动无法控制，犹如雪崩一般，公司需要采用混合数据管理方法，其中包括处理所有类型数据的有效方法。

有效数据管理的关键在于组织、结构化 和分类 不同类型的数据 "巴别"（包括非结构化、文本和几何格式，

分为结构化或松散结构化数据）。这一过程可将杂乱无章的数据集转化为有组织的结构，以便整合到系统中，从而根据这些数据做出决策（图 3.2-4）。

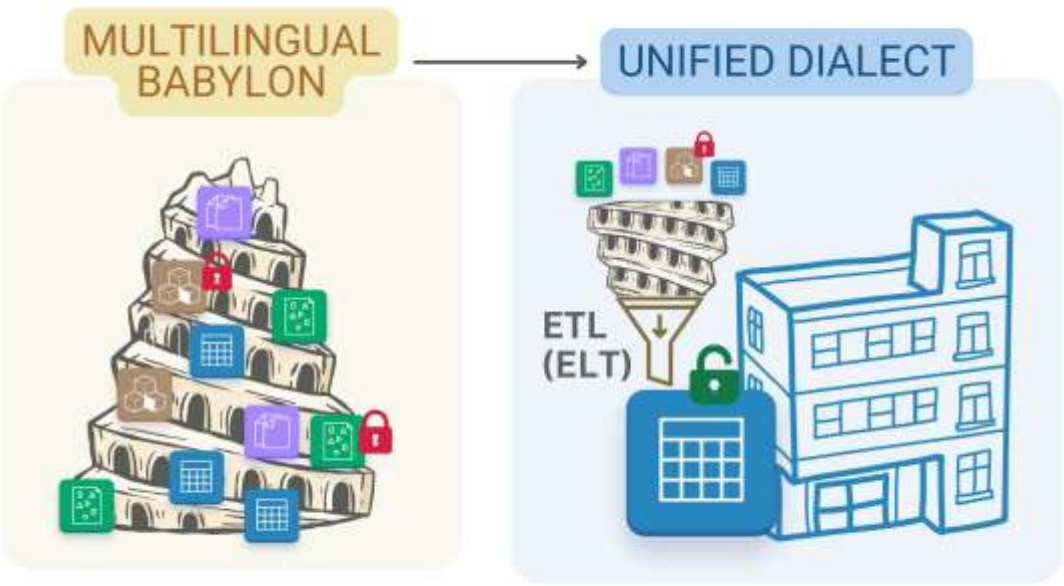


图 3.2-4 数据管理部门的主要任务是将多样化和多格式数据的 "巴比伦 "转化为结构化和分类系 统。

这种协调的主要障碍之一仍然是不同数字平台之间的互操作性较低，也就是我们在前几章中讨论过的 "孤岛 "问题。

> 根据该报告，美国国家标准与技术研究院（NIST）强调[68]，不同建筑平台之间数据兼容性差会导致信息丢失和大量额外成本。仅在 2002 年，软件互操作性问题给美国基本建设造成的损失每年就高达 158 亿美元，其中三分之二的损失由建筑物业主和运营商承担，尤其是在运行和维护期间[68]。该研究还指出，数据格式的标准化可以减少这些损失，并提高整个设施生命周期的效率。

2016 年 CrowdFlower 的研究[69]覆盖了全球 16000 名数据科学家，根据该研究，主要问题仍然是 "脏 "和多格式数据。根据这项研究，最有价值的资源不是最终数据库或机器学习模型 ，而是准备信息所花费的时间。

清理、格式化和组织工作占用了分析师 和数据管理员 60% 的时间。近五分之一的时间用于搜索和收集正确的数据集，而这些数据集往往隐藏在筒仓中，无法用于分析。只有约 9% 的时间直接用于建模、分析、预测

和假设检验。其余时间则用于交流、可视化、报告和研究辅助信息源

管理人员的数据工作平均分配如下（图 3.2-5）：

- 清理和整理数据 **(60%)**：拥有干净和结构化的数据可以大大减少 分析员的工作时间，加快完成任务的进程。
- 数据收集 **(19%)**：数据科学专业人员面临的一大挑战是寻找相关数据集。通常情况下，公司数据都堆放在组织混乱的 "筒仓 "中，很难获取所需的信息。
- 建模**/**机器学习（**9%**）：客户的业务目标不明确往往会阻碍建模**/**机器学习的发展。即使是最好的模型，如果缺乏明确的任务说明，也会失去其潜力。
- 其他任务（**5%**）：除处理数据外，分析员还必须进行研究，从不同角度探索数据，通过可视化和报告交流结果，并提出优化流程和战略的建议。

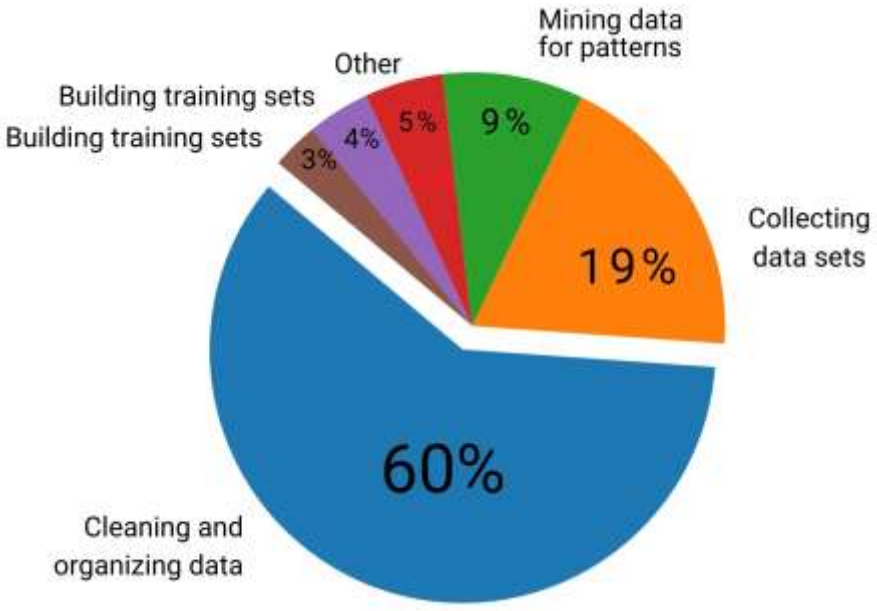


图 3.2-5 数据管理人员处理数据时花费时间最多的工作（基于 [70]）。

> 这些估计得到了其他研究的支持。根据 2015 年《商业报告》（BizReport）上发表的 Xplenty 研究报告[71]，商业智能（BI）专业人员 50%到 90% 的时间 （BI）都花在准备分析数据上。

清理、验证和组织 数据是所有下游数据和分析流程的重要基础，占用了数据科学家多达 90% 的时间。

这种最终用户看不到的艰苦劳动至关重要。原始数据中的错误不可避免地会扭曲分析结果，产生误导，并可能导致代价高昂的管理失误。这就是为什么数据清理和标准化流程 - 从消除重复和填补遗漏，到统一计量单位和调整为通用模型 - 正在成为现代数字战略的基石。

因此，数据的彻底转换、清理和标准化不仅占用了专家的大部分时间（多达 80% 的数据工作），还决定了在现代业务流程框架内有效使用数据的可能性。然而，仅靠数据组织和清理并不能完成对公司信息流进行优化管理的任务。在组织和结构化阶段，如何选择合适的数据模型，直接影响到后续处理阶段信息工作的便利性和效率。

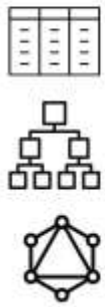
由于数据和业务目标不同，了解数据模型的特点并能够选择或创建正确的结构非常重要。根据结构化程度和元素间关系的描述方式，主要有三种模型：结构化、松散结构化和图形化。每种模式都适用于不同的任务，各有优缺点。

数据模型：数据中的关系和元素之间的关系

信息系统中的数据是以不同的方式组织起来的--这取决于存储、处理和传输信息的任务和要求。数据模型类型和信息存储形式之间的主要区别在于结构化程度和描述元素之间关系的方式。

> 结构化数据 有一个清晰和可重复的模式：它被组织成具有固定列的表格。在执行 SQL 查询、过滤和聚合时，这种格式提供了可预测性、易处理性和效率。例如 - 数据库 (RDBMS)、Excel 、CSV。

松散结构数据 允许灵活的结构：不同的元素可包含不同的属性集，并以层次结构存储。例如 JSON 、XML 或其他文档格式。当需要对嵌套对象和它们之间的关系建模时，这种数据很方便，但另一方面，它也使数据分析和标准化变得复杂（图 3.2-6）。



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

图 3.2-6 数据模型是一种逻辑结构，用于描述数据在系统中的组织、存储和处理方式。

选择适当的形式取决于目标：

- 如果过滤和分析的速度非常重要， - 关系表（SQL 、CSV、RDBMS、列式数据库 ）也可以。
- 如果需要结构的灵活性，最好使用 JSON 或 XML。
- 如果数据关系复杂，图形数据库 可提供可视性和可扩展性。

在传统的关系数据库 (RDBMS) 中，每个实体（如门）用一行表示，其属性用表格列表示。例如，"门 "类别的项目表可能包含 ID、高度、宽度、耐火度和表示房间的房间 ID 字段（图 3.2-7）。

在传统的关系数据库（RDBMS ）中，关系是以表格的形式形成的，其中每条记录代表一个对象，列代表其参数。在表格式中，项目中有关门的数据是这样的：每一行代表一个单独的元素--门及其唯一标识符和属性，通过参数 "房间 ID "与房间建立联系。

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

图 3.2-7 以表格结构形式提供项目 "门 "类别三个要素的信息。

在结构松散的格式（如 JSON 或 XML ）中，数据以分层或嵌套的形式存储，其中元素可能包含其他对象，其结构也可能各不相同。这样就可以模拟元素之间的复杂关系。项目中以结构化形式记录的有关门的类似信息（图 3.2-7）以松散的结构化格式（JSON ）表示（图 3.2-8），它们成为房间（Rooms - ID）中的嵌套对象，这在逻辑上反映了层次结构。

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"}
      ]
    }
  ]
}
```

图 3.2-8 JSON 格式的项目 "门 "类别元素信息。

在图模型中，数据以节点（顶点）和节点之间的链接（边）来表示。这样就可以直观地显示对象及其属性之

间的复杂关系。就项目中的门和房间数据而言， 的图表示如下：

- 节点（结点）代表主要实体：房间（101 室、102 室）和门（ID1001、ID1002、ID1003）
- 肋条（链接）显示这些实体之间的关系，例如，某扇门属于某个房间
- 属性映射到节点，包含实体属性（高度、宽度、门的防火性能）



图 3.2-9 图表视图中的项目门实体信息。

在门描述的图数据模型中，每个房间和每扇门都是独立的节点。门通过边与房间相连，表示门属于某个房间。门的属性（高度、宽度、防火性能）作为相应节点的属性存储。有关图格式 以及图语义如何在建筑业中出现的更多详情，将在 "建筑业中 语义和本体的出现 "一章中讨论。

当重要的不是数据本身，而是数据之间的关系时，如在推荐系统、路由系统或设施管理项目的复杂关系建模中，图形数据库 就非常有效。图形格式允许在不改变存储结构的情况下将新的数据类型添加到图形中，从而简化了新关系的创建。不过，与关系表和结构化格式相比，图形中没有额外的数据连接--将二维数据库数据传输到图形中不会增加关系数量，也无法获取新信息。

> 数据的形式和模式应根据具体的使用情况和要执行的任务来定制。要想在业务流程中有效工作，就必须使用那些有助于尽可能快速、轻松地获得结果的工具和数据模型。

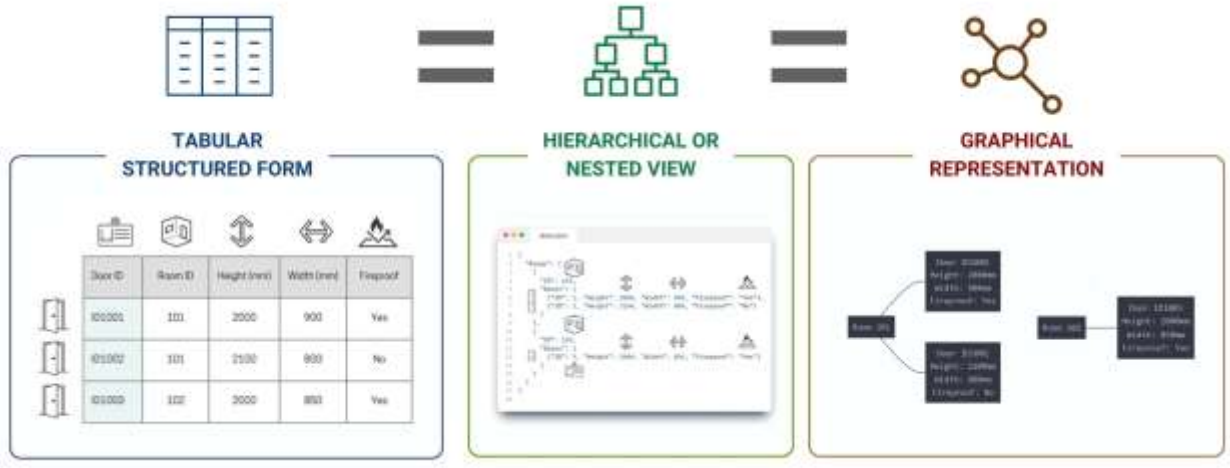


图 3.2-10 有关项目要素的相同信息可以使用不同的数据模型以不同的格式存储。

> 如今，大多数大公司都面临着数据过于复杂的问题。成百上千个应用程序中的每一个都使用自己的数据模型，这就造成了过度的复杂性--单个模型的复杂性往往是所需复杂性的数十倍，而所有模型的总和则是所需复杂性的数千倍。这种过度的复杂性极大地妨碍了开发人员和最终用户的工作。

这种复杂性严重限制了公司系统的开发和维护。模型中的每一个新元素都需要额外的代码、新逻辑的实施、全面的测试以及对现有解决方案的调整。所有这一切都增加了成本，减缓了公司自动化团队的工作速度，即使是简单的任务也会变成昂贵而耗时的过程。

复杂性影响着数据架构的各个层面。在关系数据库中，复杂性表现为表格和列的数量不断增加，而且往往是多余的。在面向对象系统中，复杂性因类和相互关联属性的多重性而增加。在 XML 或 JSON 等格式中，复杂性表现为混乱的嵌套结构、唯一键和不一致的模式。

> 数据模型过于复杂，不仅降低了系统的效率，而且难以被最终用户理解，将来也难以被大型语言模型和 LLM 代理。正是数据模型和数据处理的理解和复杂性问题提出了这样一个问题：如何使数据足够易于使用，从而使其迅速发挥作用。

即使明智地选择了数据模型，如果数据访问受到限制，其实用性也会大打折扣。专有格式和封闭式平台阻碍

了集成，使自动化复杂化，并剥夺了对专有信息的控制权，不仅造成了新数据的孤岛，还造成了只能通过供应商授权才能访问的锁定孤岛。要了解问题的严重性，就必须考虑封闭系统对建筑业数字化流程的确切影响。

专有格式及其对数字流程的影响

建筑公司在数字化过程中面临的主要挑战之一是数据访问受限。这导致系统难以整合，信息质量下降，难以组织高效的流程。使用专有格式和封闭式软件解决方案往往是造成这些困难的根源。

遗憾的是，迄今为止，建筑行业使用的许多程序只允许用户以专有格式或云存储方式保存数据，而且只能通过严格限制的接口访问。而且，这些解决方案往往是依赖于大型供应商提供的更加封闭的系统而构建的。因此，即使是那些希望提供更开放架构的开发人员，也不得不遵守大型供应商规定的规则。

> 虽然现代建筑数据管理系统越来越多地支持开放格式和标准（图 3.1-5），但基于 CAD (BIM)的数据库以及相关的 ERP 和 CAFM 系统仍然是行业数字景观中孤立的专有 "孤岛"（图 3.2-11）。



图 3.2-11 数据的封闭性和专有性为数据集成和访问制造了障碍。

封闭和垄断的格式和协议不仅仅是建筑行业的问题。在许多经济领域，与封闭标准和数据获取受限的斗争始于创新放缓（图 3.2-12）、新参与者进入市场的人为障碍以及对大型供应商的依赖加深。随着数据重要性的快速增长，竞争管理机构根本没有时间应对新数字市场带来的挑战，因此，封闭格式和限制数据访问实质上成为了限制信息流动和增长的数字 "边界"[63]。

> 如果机器能生产我们所需的一切，那么我们的处境将取决于这些商品的分配方式。只有共享机器生产的财富，每个人才能过上富裕的生活。或者，如果汽车拥有者能够成功游说反对财富再分配，那么大多数人最终将生活在赤贫之中。到目前为止，情况似乎正朝着第二种方向发展，技术导致了越来越严重的不平等[72]。
>
> - 斯蒂芬-霍金，天体物理学家，2015 年

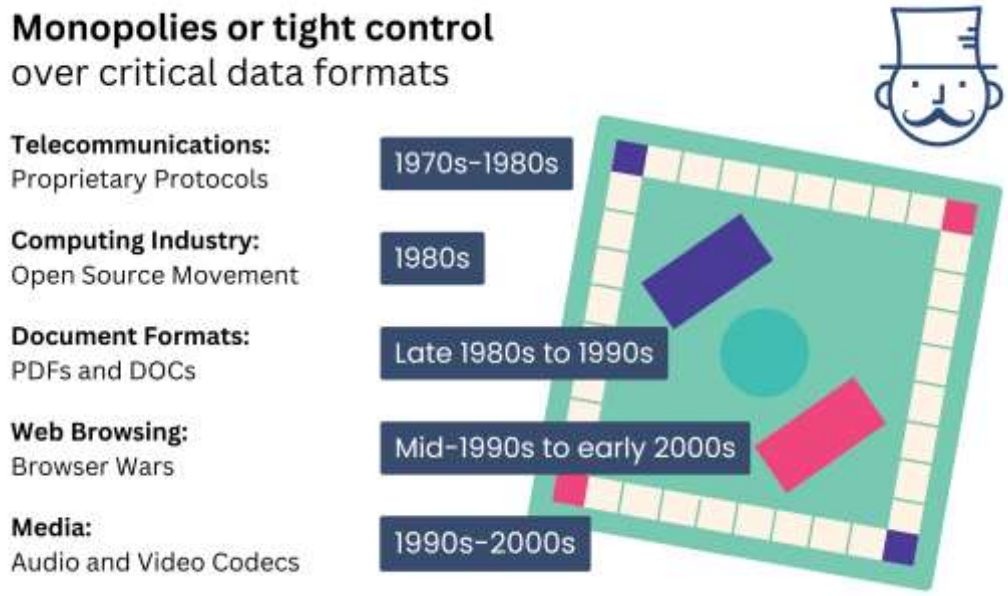


图 3.2-12 对关键数据格式和协议的垄断所有权并不是建筑业独有的问题。

因此，由于对数据库 程序的封闭访问，数据管理人员、数据分析人员 、IT 专家和开发人员在创建建筑业数据访问、处理和自动化应用程序时，面临着对软件供应商的诸多依赖（图 3.2-13）。这些额外访问层形式的依赖性要求创建具有专门 API - 连接及特殊工具和软件的解决方案。

> 应用程序接口（API）是一个正式的接口，通过它，一个程序可以与另一个程序交互，交换数据和功能，而无需访问源代码。应用程序接口描述了外部系统可以提出的请求、请求的格式以及将收到的响应。它是软件模块之间的标准化 "契约"。

对封闭式解决方案的大量依赖会导致公司的整个代码架构和业务流程逻辑成为一个由各种工具组成的 "意大利面条架构"，这些工具依赖于软件供应商提供高质量数据访问的政策。

依赖封闭的解决方案和平台不仅会丧失灵活性，还会带来实际的业务风险。更改许可条款、关闭数据访问、更改格式或 API 结构 - 所有这些都会阻碍关键流程。突然发现，更新一个表需要重新制作一整块集成和连接器（图 3.2-13），软件或其 API 供应商的任何大规模更新都会对整个公司系统的稳定性构成潜在威胁。

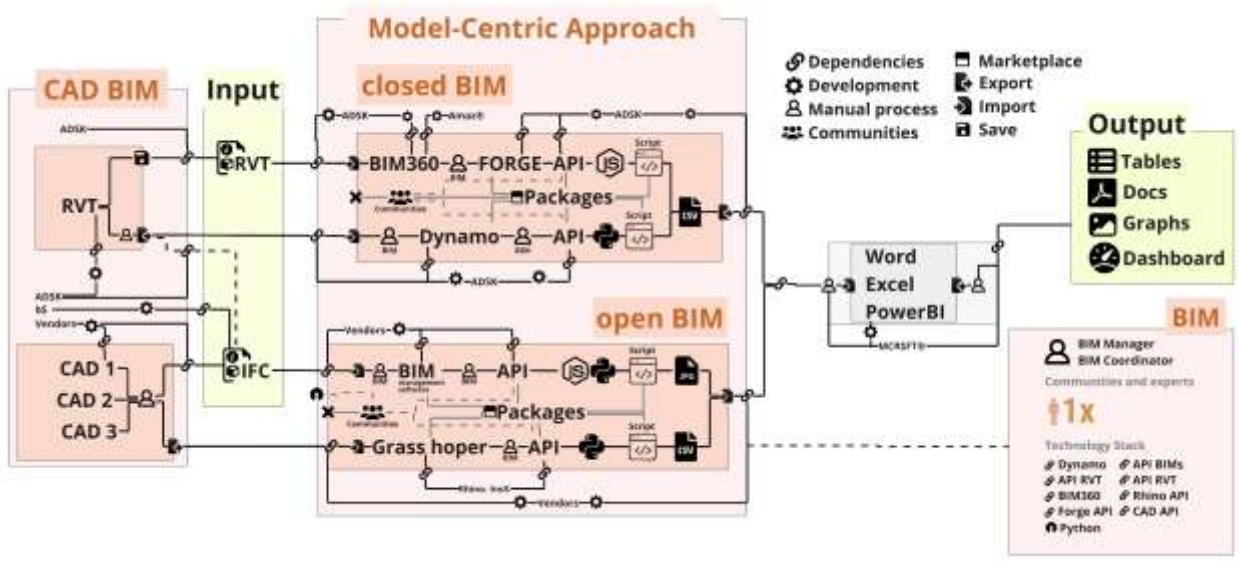


图 3.2-13 CAD 处理过程中存在大量依赖关系的示例 -数据为建筑公司生态系统中的数据集成制造了障碍。

> 在这种情况下，开发人员和系统架构师被迫工作，不是为了预期，而是为了生存。他们不去实施新的解决方案，而是去适应。他们不去开发，而是努力保持兼容性。他们不是自动化和加快流程，而是把时间花在研究下一个封闭接口、API 文档和无休止的代码重建上。

使用封闭的格式和系统工作不仅是一项技术挑战，也是一项战略限制。尽管现代自动化、人工智能 、LLM 和预测分析 带来了显而易见的机遇，但许多公司未能充分发挥其潜力。专有格式设置的障碍（图 3.2-13）

使企业无法访问自己的数据。这或许是对建筑业数字化转型的讽刺。

> 数据透明和系统开放不是奢侈品，而是提高速度和效率的先决条件。没有开放性，业务流程中就会充斥着不必要的官僚主义、多层审批链以及对 HiPPO 原则的日益依赖 ，即根据薪酬最高者的意见做出决策。

然而，一种模式转变正在地平线上形成。尽管专有解决方案占据主导地位，但越来越多的公司正在意识到第四次工业革命架构的局限性。如今，矢量正在转向第五次革命的原则，即以数据为战略资产、开放接口（API）和系统间真正的互操作性为中心。

这一转变标志着从封闭的生态系统向灵活、模块化的数字架构转变，开放格式、标准和透明的数据交换在其中发挥着关键作用。

开放格式正在改变数字化方式

建筑业是最晚解决封闭和专有数据问题的行业之一。与其他经济部门不同，这里的数字化发展缓慢。究其原因，包括该行业的传统保守性、不同本地解决方案的普遍性以及纸质文件管理的根深蒂固性。几十年来，主要的施工流程都依赖于实物图纸、电话和不同步的数据库。在这种情况下，封闭格式长期以来一直被视为规范而非障碍。

其他行业的经验表明，消除封闭数据的障碍会导致创新激增、发展加速和竞争加剧[73]。在科学领域，开放数据的交换有助于加速发现和发展国际合作。在医学领域，它可以提高诊断和治疗的效率。在软件工程领域--创建共同创造和快速改进产品的生态系统。

> 麦肯锡报告《开放数据：2013 年麦肯锡报告《开放数据：通过信息流释放创新和生产力》[74]指出，开放数据 有潜力在建筑、交通、医疗保健和能源等七个关键行业每年释放 3 万亿至 5 万亿美元。[74]，开放数据 有可能每年为建筑、交通、医疗保健和能源等七个关键行业释放 3 至 5 万亿美元。根据同一份研究报告，分散的数据生态系统使大型建筑公司和承包商能够降低软件开发和维护成本，加快数字化应用。

向开放式架构的过渡早已在其他经济部门开始，并逐渐蔓延到建筑行业。大公司和公共客户，尤其是控制建

筑项目投资的金融组织，越来越多地要求使用开放数据和获取计算、计算和应用程序的源代码。开发人员不再仅仅被期望创建数字解决方案和显示项目的最终数据，他们还被期望做到透明、可复制和独立于第三方应用程序供应商。

> 使用开放源码解决方案 可使客户确信，即使外部开发人员停止合作或离开项目，也不会影响进一步开发工具和系统的能力。开放数据的主要好处之一是能够消除应用程序开发人员对特定平台访问数据的依赖。

如果公司不能完全放弃专有解决方案，一种可能的折衷办法是使用逆向工程技术。这些在法律和技术上 合理的方法可以将封闭格式转换为更易于访问、结构化和适合集成的格式。当需要连接传统系统或将信息从一个软件环境迁移到另一个软件环境时，这一点尤为重要。

> 在计算机辅助设计系统（CAD ）中广泛使用的 DWG 格式 ，是建筑业向开放格式过渡和应用逆向工程（合法入侵专有系统）历史上最耀眼的例子之一。1998 年，为了应对一家软件供应商的垄断 ，其他 15 家 CAD 供应商组成了一个名为 "开放 DWG "的新联盟，为开发人员提供免费和独立的工具，使其无需专有软件或封闭的 API 即可使用 DWG 格式（图纸传输的事实标准）。这一事件是一个转折点，从 20 世纪 80 年代末至今，数以万计的公司得以免费使用流行 CAD 解决方案的封闭格式，并创建兼容的解决方案，从而促进了 CAD 市场的竞争 [75]。如今，在 DWG 格式开发者的官方应用程序之外，几乎所有可以导入、编辑和导出 DWG 格式的解决方案都使用了早在 1996 年就已创建的 "开放 DWG "SDK。

其他科技巨头也在进行类似的转型。曾经是专有方法象征的微软开放了.NET Framework 源代码，开始在 Azure 云服务基础设施中使用 Linux，并收购了 GitHub 以加强其在开源社区 中的地位。[76].Meta（前身为 Facebook）发布了开源人工智能模型 ，如 Llama 系列，以促进人工智能代理开发的创新与合作。首席执行官马克-扎克伯格预计，开源 平台将在未来十年引领技术进步 [77]。

> 开放源码 是一种软件开发和发布模式，其源代码开放供自由使用、研究、修改和发布。

开放数据和开源解决方案不仅是一种趋势，也是数字可持续发展的基础。它们为企业提供了灵活性、复原力、对自身决策的控制力，以及在不依赖供应商政策的情况下扩展数字流程的能力。同样重要的是，它们让企

业重新掌控 21 世纪最宝贵的资源--数据。

范式转变：开放源码 作为软件供应商主导时代的终结者

建筑行业正在经历一场无法以常规方式盈利的变革。数据驱动的概念、以数据为中心的方法以及 Open Source 工具的使用正在促使人们重新思考市场上软件巨头赖以生存的游戏规则。

> 与以往的技术转型不同，这次转型不会得到供应商的积极推动。这种模式的转变威胁到它们基于许可、订阅和咨询的传统业务模式。新的现实并不涉及开箱即用的产品或付费订阅，而是需要重新调整流程和思维。

为了管理和开发基于开放技术的数据中心解决方案，企业需要重新思考内部流程。来自不同部门的专家不仅要相互协作，还要重新思考如何协同工作。

新模式意味着使用开放数据和开源解决方案，在创建软件代码的过程中，基于人工智能和大型语言模型（LLM）的工具而不是程序员将发挥特殊作用。到 2024 年中期，谷歌已有超过 25% 的新代码是用人工智能创建的[78]。未来，使用 LLM 编码将只需 20% 的时间就能完成 80% 的工作（图 3.2-14）。

> 根据麦肯锡的 2020 年研究报告[79]，由于 GPU 性能高，而且有现代开源工具的支持，因此在分析领域，GPU 正在逐渐取代 CPU。这样，企业就可以加快数据处理速度，而无需在昂贵的软件上进行大量投资，也无需聘请稀缺的专家。

麦肯锡、普华永道和德勤等知名咨询公司都强调，开放标准、开放源码 应用程序在各行各业的重要性与日俱增。

> 根据普华永道《开源 Monitor 2019》报告[80]，在员工人数达到或超过 100 人的公司中，69% 的公司有意识地使用开源解决方案。开放源码软件在大公司的使用尤为活跃：在拥有 200-499 名员工的公司中，71% 的公司使用开放源码软件；在拥有 500-1999 名员工的公司中，78% 的公司使用开放源码软件；在拥有 2000 名以上员工的公司中，使用开放源码软件的比例高达 86%。根据 Synopsys OSSRA 2023 报告，96% 的分析代码库包含开源组件 [81]。

未来，开发人员的角色不再是手动编写代码，而是设计数据模型、流程架构和管理人工智能代理，按需创建

正确的计算。用户界面将变得简约，交互将以对话为基础。传统编程将让位于数字解决方案的高级设计和协调（图 3.2-14）。当前的趋势--如低代码平台（图 7.4-6）和支持 LLM 的生态系统（图 7.4-4）--将大大降低开发和维护 IT 系统的成本。

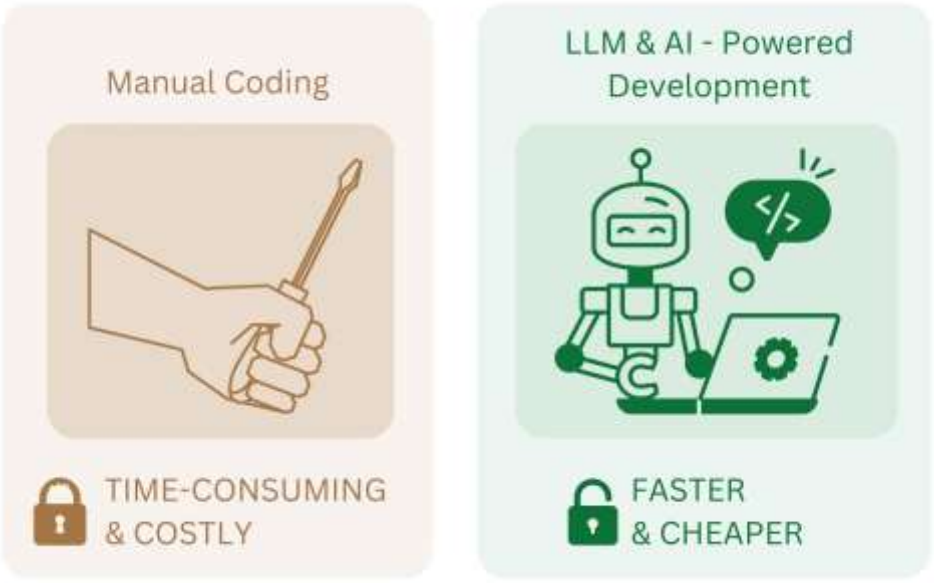


图 3.2-14 虽然现在的应用程序是由程序员手动创建的，但在未来，很大一部分代码将由人工智能和基于 LLM 的解决方案生成。

这次转型与以往不同，大型软件厂商不太可能成为催化剂。

> 哈佛商学院的研究报告《2024 年开源软件的价值》[40]从两个角度估算了开源软件的总价值。一方面，如果我们计算一下从零开始构建所有现有的开放源码解决方案需要多少钱，那么金额大约为 41.5 亿美元。另一方面，如果我们设想每家公司在无法使用现有工具的情况下，自行开发类似的开源解决方案（这种情况随处可见），那么企业的总成本将达到 8.8 万亿美元之巨--这就是需求成本。

> 不难猜测，没有一家大型软件厂商有兴趣将潜在价值达 8.8 万亿美元的软件市场缩减至 41.5 亿美元。这意味着需求量将减少 2000 多倍。对于那些商业模式建立在多年来维持客户对封闭式解决方案的依赖上的供应商来说，这样的转型根本无利可图。因此，期望有人为其提供便捷、开放的交钥匙解决方案的公司可能会失望--这些供应商根本不会出现。

向开放式数字架构的转变并不意味着工作或收入的损失。相反，它为灵活、适应性强的商业模式创造了条件，最终可能取代传统的许可证和盒装软件市场。

不再出售许可证，而是提供服务；不再是封闭的格式，而是开放的平台；不再依赖供应商，而是独立自主地为实际需求提供解决方案。过去只是简单使用工具的人将能够成为工具的共同创造者。而那些能够处理数据、模型、方案和逻辑的人，将发现自己处于行业新数字经济的中心。我们将在本书的最后，也就是第十部分，详细讨论这些变化，以及围绕开放数据正在出现的新角色、商业模式和合作形式。

> 基于开放数据和开放代码的解决方案将使企业能够专注于业务流程的效率，而不是纠结于过时的应用程序接口和集成封闭系统。有意识地过渡到开放式架构可以大大提高生产率，减少对供应商的依赖。

向新现实的过渡不仅仅是软件开发方法的改变，也是对数据工作原理的重新思考。这一转变的核心不是代码，而是信息：信息的结构、可访问性和可解释性。这就是开放式结构化数据的重要性所在，它已成为新数字架构不可分割的一部分。

结构化开放数据：数字化转型的基础

在过去的几十年里，企业的可持续发展在很大程度上取决于软件解决方案的选择和对特定供应商的依赖，而在当今的数字经济时代，关键因素则是数据质量和有效利用数据的能力。开放源代码是新技术范式的重要组成部分，但只有当数据可理解、有组织且机器可读时，才能真正释放其潜力。在所有类型的数据模型中，结构化开放数据正成为可持续数字化转型的基石。

结构化开放数据的主要优点是解释清晰，可以自动处理。这样，无论是在单个业务层面，还是在整个组织层面，都能大大提高效率。

> 根据德勤公司的报告《企业转型中的数据传输过程》[82]，与 IT 部门合作管理结构化数据的传输至关重要。[82]，与信息技术部门合作管理结构化数据的传输至关重要。根据英国政府的报告《政府项目交付中的数据分析和人工智能 》（2024 年）[83]，消除不同项目和组织之间的数据共享障碍是提高项目管理效率的关键。该文件强调，数据格式的标准化和开放数据原则的引入可以避免信息重复、最大限度地减少时间损失并提高预测的准确性。

对于传统上高度分散且格式多样的建筑行业来说，结构化统一流程和结构化开放数据 在形成一致且可管理的流程方面发挥着至关重要的作用（图 4.1-14）。它们使项目参与者能够专注于提高生产率，而不是解决与封闭平台、数据模型和格式之间不兼容有关的技术问题。

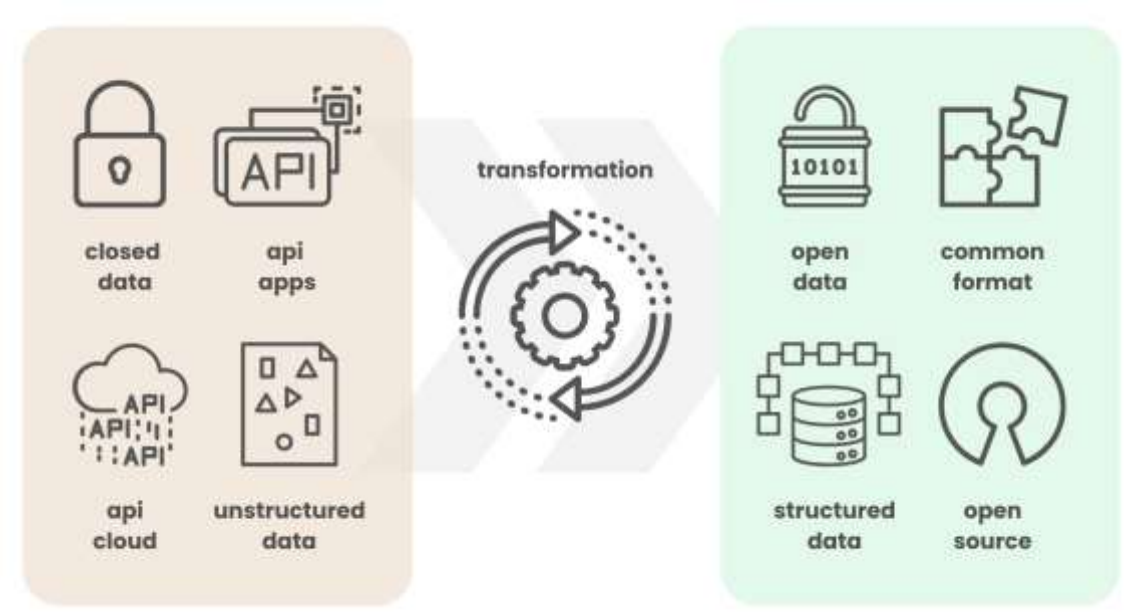


图 3.2-15 开放式结构化数据可减少对软件解决方案和平台的依赖，加快创新。

我们将在本书后面详细讨论的现代技术工具不仅可以收集信息，还可以自动清理信息：消除重复、纠正错误和规范数值。这意味着，分析师和工程师在工作中处理的不是分散的文件，而是适合分析、自动化和决策的有组织的知识库。

> 尽量简单，但不能再简单了。
>
> - 阿尔伯特-爱因斯坦，理论物理学家（该名言的作者有争议 [84]）

如今，大多数处理数据的用户界面都可以自动创建--无需为每个业务案例手动编写代码。这就需要一个基础结构层，它能理解数据结构、模型和逻辑，而不需要额外的指令（图 4.1-15）。结构化数据使这种方法成为可能：只需极少的编程工作就能自动生成表单、表格、过滤器和视图。

最重要的用户关键界面可能仍然需要人工完善。但在大多数情况下（占工作场景的 50%至 90%），自动生成应用程序和计算，而无需为此使用专门的应用程序 就足够了（图 3.2-16），这样可以大大降低开发和维护成本，减少错误并加快数字化解决方案的实施。

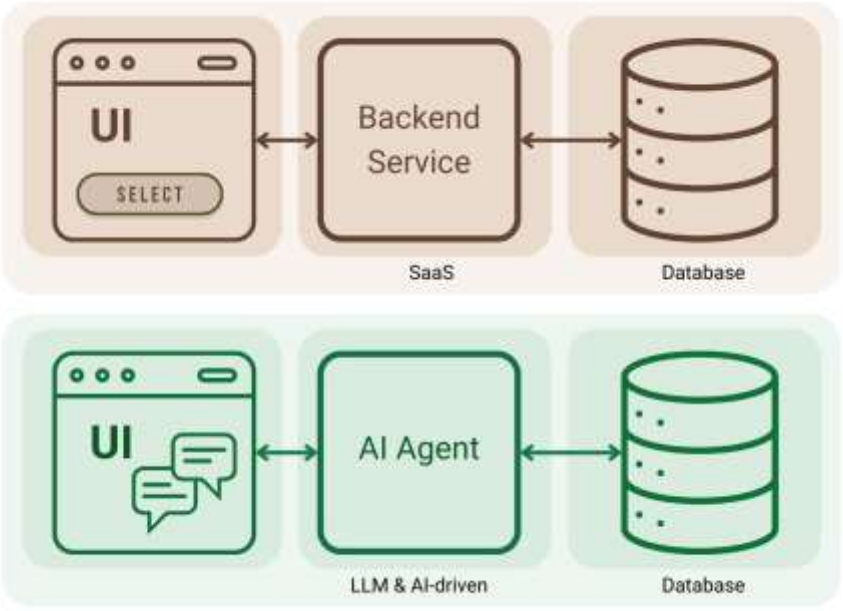
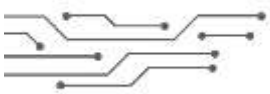


图 3.2-16 处理数据的架构模型：传统应用架构和面向人工智能 与 LLM 的模型。

从基于单个应用程序的架构到基于语言模型（LLM）的智能管理系统的转变，是数字化发展的下一步。在这种架构中，结构化数据不仅是存储对象，而且还是与人工智能工具进行交互的基础，人工智能工具能够根据上下文进行分析、解释和推荐操作。

在接下来的章节中，我们将介绍基于开放式结构化数据实施架构的实际案例，并展示如何应用语言模型来自动解释、验证和处理数据。这些实际案例将帮助您更好地理解新数字逻辑的实际运作方式，以及它能为准备转型的企业带来哪些好处。



章 3.3.

llm 及其在数据处理和业务流程中的作用

LLM 聊天室：**ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok** 实现数据处理自动化

大型语言模型（LLM）的出现是结构化开放数据运动和开源理念的自然延伸。当数据变得有组织、可访问和机器可读时，下一步就是一个无需编写复杂代码或掌握专业技术知识就能与这些信息交互的工具。

> LLM 是开放性的直接产物：大型开放数据集、出版物和开源运动。没有开放的学术文章、公开可用的文本数据以及合作开发的文化，就不会有 ChatGPT 或其他 LLM。从某种意义上说，LLM 是人类积累的数字知识的 "提炼物"，是通过开放原则收集和教育而来的。

现代大型语言模型（LLM - Large Language Models），如 ChatGPT ® (OpenAI), LlaMa ™ (Meta AI), Mistral DeepSeek™, Grok ™ (xAI), Claude ™ (Anthropic), QWEN™ 为用户提供了用自然语言查询数据的能力。这样，不仅开发人员可以使用信息，分析师、工程师、规划师、经理和其他以前远离编程的专业人员也可以使用信息

> **LLM** （大型语言模型）是一种人工智能，经过训练，可以根据从互联网上收集的大量数据理解和生成文本。它能够分析上下文、回答问题、参与对话、编写文本和生成软件代码。

如果说早期的数据可视化、处理或分析需要掌握一种特殊的编程语言：Python 、SQL 、R 或 Scala，以及使用 Pandas 、Polars 或 DuckDB 等库的能力，那么从 2023 年开始，情况发生了翻天覆地的变化。现在，用户只需描述他想要得到什么，模型本身就会生成代码、执行代码、显示表格或图形并解释结果。几十年来，技术的发展第一次没有走上复杂化的道路，而是走上了彻底简化和无障碍的道路。

这一原则--"用文字（提示）处理数据"。- 这标志着信息工作的发展进入了一个新阶段，有效地将解决方案的创建提升到了一个更高的抽象水平。正如用户曾经不再需要了解互联网的技术基础，就可以使用 WordPress、Joomla 和其他开源模块系统经营网上商店或创建网站一样（av tor 这本书从 2005 年起就开始使用

这些系统，包括教育和工程在线平台）。- 这反过来又促进了数字内容和在线业务的蓬勃发展--如今，工程师、分析师和管理人员无需编程语言知识就能实现工作流程自动化。这得益于功能强大的 LLM（免费和开源），如 LLaMA、Mistral、Qwen、DeepSeek 等，它们让最广泛的受众也能使用先进技术。

大型 **LLM** 语言模型： 工作原理

大语言模型（ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok ）是在互联网、书籍、文章和其他来源的海量文本数据上训练出来的神经网络。它们的主要任务是理解人类语音的上下文，并生成有意义的回应。

现代 LLM 基于谷歌研究人员于 2017 年提出的 Transformer 架构[85]。该架构的关键组件是关注机制，它允许模型考虑单词之间的关系，而不管它们在文本中的位置如何。

> LLM 的学习过程 与人类学习语言的方式极为相似，只是规模大了数百万倍。该模型分析数十亿个单词和表达方式的例子，识别语言结构和语义转换逻辑中的模式。整个文本被分解成标记--最小语义单位（词语或其部分），然后在多维空间中转换成向量（图 8.2-2）。这些向量表示允许机器 "理解 "概念之间的隐藏关系，而不是简单地将文本作为一连串符号来处理。

大语言模型不仅仅是生成文本的工具。它们能够识别意义、找到概念之间的联系并处理数据，即使数据是以不同格式呈现的。最重要的是，应将信息分解为可理解的模型，并表示为大语言模型可以处理的标记。

同样的方法也适用于建筑项目。如果我们将项目视为一种文本，其中每个建筑、元素或构造都是一个标记，那么我们就可以开始以类似的方式处理这些信息。可以将建筑项目比作书籍，书籍分为类别、章节和段落组，由最小的标记--建筑项目的元素组成（图 3.3-1）。通过将数据模型转化为结构化格式，我们还可以将结构化数据转化为向量库（图 8.2-2)，这是机器学习 和 LLM 等技术的理想来源。

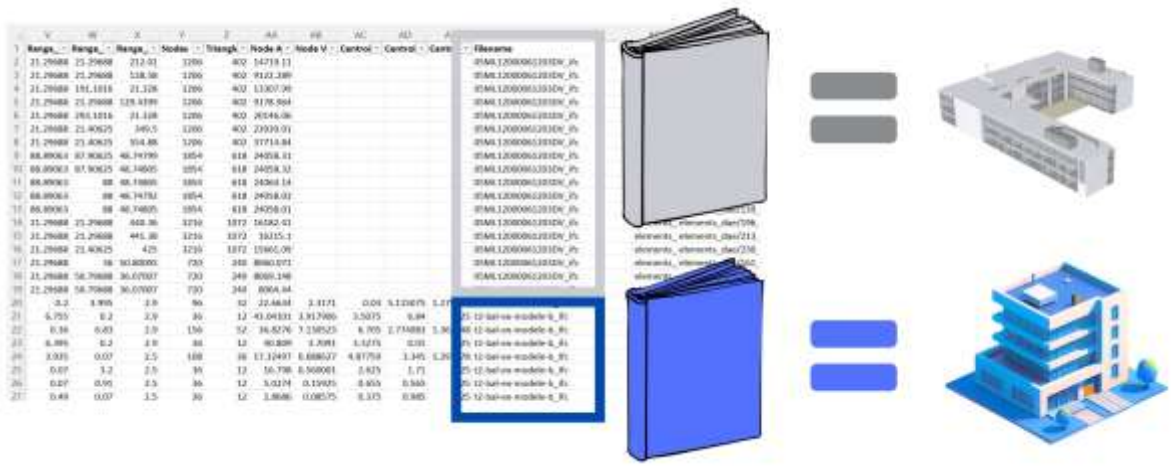


图 3.3-1 建筑项目要素就像文本中的一个标记：一个最小单位，由它组成整个项目的各组（段落）各节（类别）。

如果将建筑项目数字化，并将其元素表示为标记或矢量，那么就可以用自然语言而不是死板的形式化查询来访问它们。这就是 LLM 的关键优势之一--理解查询的含义并将其与相关数据联系起来的能力。

工程师不再需要编写 SQL -query 或 Python 代码 来获取所需数据，只需了解 LLM 和数据结构，就能以通常的方式制定任务："查找所有混凝土等级高于 B30 的钢筋混凝土结构并计算其总体积"。模型将识别查询的含义，将其转化为机器可读的形式，查找数据（分组和转换）并返回最终结果。

文件、表格、项目模型被转换成矢量表示（嵌入 ）并存储在数据库中。当用户提问时，查询也会被转换成矢量，系统会找到最相关的数据。这样，LLM 不仅可以依靠自己训练有素的知识，还可以依靠实际的企业数据，即使这些数据是在模型训练结束后出现的。

> LLM 在建筑领域最重要的优势之一就是能够生成程序代码。专家可以用自然语言描述任务，而不是将技术任务交给程序员，模型将生成必要的代码，这些代码可以（通过从聊天中复制）用于创建流程自动化代码。LLM - 模型可以让没有深厚编程知识的专家为公司业务流程的自动化和改进做出贡献。

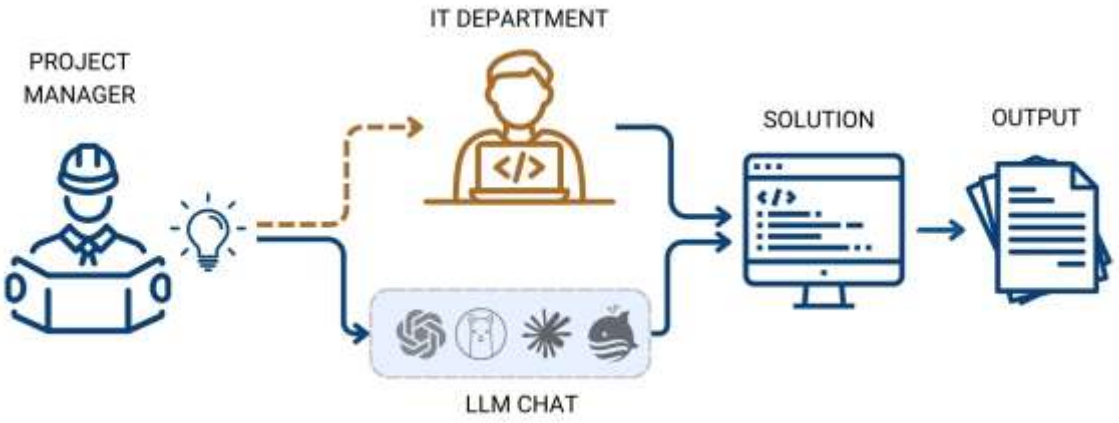


图 3.3-2 LLM 为用户提供了编写代码和获取结果的能力，而无需编程技能。

> 2024 年，由 Wakefield Research 和 SAP 赞助的一项研究[36]对美国年收入至少 10 亿美元的 300 家公司的高级管理人员进行了调查：52% 的高级管理人员相信人工智能 可以分析数据并为决策提供建议。另有 48% 的人使用人工智能来识别以前未曾考虑到的风险，47% 的人使用人工智能来建议替代计划。此外，40%的人将人工智能用于新产品开发、预算规划和市场调研。研究还显示了人工智能对个人生活的积极影响，39% 的受访者表示工作与生活的平衡得到了改善，38% 的受访者表示心理健康得到了改善，31% 的受访者表示压力水平降低。

然而，尽管有如此强大的功能，法律硕士仍然是一种必须有意识地使用的工具。与任何技术一样，它们也有局限性。其中一个最著名的问题就是所谓的 "幻觉"--即模型自信地得出了一个看似合理但实际上不正确的答案。因此，了解模型的工作原理至关重要：它可以无误地解释哪些数据和数据模型，如何解释查询，以及从哪里获取信息。同样值得记住的是，LLM 的知识仅限于其培训日期，如果不与外部数据连接，模型可能无法考虑当前的规范、标准、价格或技术。

解决这些问题的办法是定期更新矢量数据库、连接相关资源并开发自主人工智能 --这些代理不仅能回答问题，还能主动利用数据进行培训、管理任务、识别风险、提出优化方案并监控项目绩效。

> 建筑业向 LLM 界面的过渡不仅仅是一项技术创新。它是一种模式转变，消除了人与数据之间的障碍。它使我们能够像彼此交谈一样轻松地处理信息，并获得准确、可验证和可操作的结果。

那些比其他公司更早开始使用此类工具的公司将获得显著的竞争优势。这包括加快工作速度、降低成本，以

及通过快速访问数据分析和快速找到复杂问题答案的能力来提高设计方案的质量。但同时也要考虑安全问题。使用基于云的 LLM 服务可能存在数据泄漏的风险。因此，越来越多的企业开始寻找替代解决方案，以便在本地部署 LLM 工具，并对信息进行全面保护和控制。

使用本地 **LLM** 查找公司敏感数据

2022 年，首款聊天式人工智能 的出现标志着人工智能的发展进入了一个新阶段。然而，在这些模型被广泛采用后，一个合理的问题随即出现：将公司相关数据和查询转移到云端有多安全？大多数基于云的语言模型都会在服务器上存储通信历史和上传文件，对于处理敏感信息的公司来说，这是采用人工智能的一个严重障碍。

> 解决这一问题的最可持续、最合理的方法之一，就是在企业 IT 基础设施内本地部署开源 LLM。与云服务不同，本地模式无需互联网连接即可工作，不向外部服务器传输数据，并使公司能够完全控制信息

> 目前最好的开放模型[开源 LLM]与封闭模型[如 ChatGPT、Claude]性能相当，但落后约一年[77]。
>
> - 2024 年，非营利研究机构 Epoch AI 首席研究员 Ben Cottier

主要技术公司已开始将其 LLM 提供给本地使用。Meta 的开源 LLaMA 系列和中国快速发展的 DeepSeek 项目就是向开放式架构迈进的范例。与此同时，Mistral 和 Falcon 也发布了功能强大的模型，摆脱了专有平台的束缚。这些举措不仅加速了全球人工智能 的发展，也为注重隐私的公司提供了真正的独立、灵活和安全合规的选择。

在企业环境中，尤其是在建筑行业，数据保护不仅是一个方便的问题，而且是一个遵守法规的问题。在处理投标文件、估算、图纸和机密信函时需要严格控制。当地的 LLM 在这方面提供了必要的保证，确保数据不被泄露。

	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

图 3.3-3 本地模式可提供全面控制和安全性，而基于云的解决方案则可提供轻松集成和自动更新。

本地开源 LLM 的主要优势：

- 完全控制数据。所有信息都保留在公司内部，杜绝了未经授权的访问和数据泄漏。
- 离线操作。无需依赖互联网连接，这对于孤立的 IT 基础设施尤为重要。这还能确保在受到制裁或云服务受阻时不间断运行。
- 应用灵活。该模型可用于文本生成、数据分析、程序代码编写、设计支持和业务流程管理。
- 适应企业目标。LLM 可根据内部文件进行培训，从而使您能够考虑到公司工作的具体情况及其行业

特点。本地 LLM 可连接至 CRM 、ERP 或 BI 平台 ，从而实现客户查询分析、报告创建甚至趋势预测的自动化。

> 在服务器上部署 DeepSeek 的免费开源模型 -R1-7B，供整个团队的用户访问，每月成本为 1000 美元，其成本可能低于云 API 的年费 ，如 ChatGPT 或 Claude ，并使公司能够完全控制其数据，无需将数据传输到互联网，有助于遵守 GDPR 等监管要求

在其他行业，本地龙 8 国际 pt 娱乐官网 已经在改变自动化方式。在支持服务部门，他们对客户的频繁询问做出回应，从而减少操作员的工作量。在人力资源部门，它们分析简历并挑选相关候选人。在电子商务领域，它们在不泄露用户数据的情况下生成个性化的报价。

预计在建筑领域也会产生类似的效果。通过将 LLM 与项目数据和标准相结合，可以加快文件编制速度，实现估算和预测成本分析的自动化。将 LLM 与结构化表格和数据框架结合使用正在成为一个特别有前景的领域。

全面控制公司的人工智能 以及如何部署自己的 **LLM**

现代工具允许公司在本地部署大型语言模型（LLM），只需几个小时。这就实现了对数据和基础设施的完全控制，消除了对外部云服务的依赖，最大限度地降低了信息泄露的风险。这种解决方案对于处理敏感项目文档或商业敏感数据的组织尤为重要。

根据任务和资源的不同，可以采用不同的部署方案，从开箱即用的解决方案到更加灵活和可扩展的架构。最简单的工具之一是 Ollama，它可以让您一键运行语言模型，无需深厚的技术知识。Ollama 快速入门：

1. 从官方网站下载适用于你的操作系统（Windows/Linux/macOS）的发行版：ollama.com
2. 通过命令行安装模型。例如，安装 Mistral 模型：

```
奥拉马跑米斯特拉尔
```

3. 运行模型后，它就可以工作了--您可以通过终端发送文本查询，或将其集成到其他工具中。运行模型并执行查询

```
ollama run mistral "如何计算安装 100 毫米宽石膏板隔墙的所有施工资源?
```

对于那些喜欢在熟悉的可视化环境中工作的人来说，LM Studio 是一款免费应用程序，其界面让人想起 ChatGPT

- 从官方网站 lmstudio.ai 下载分发包，安装 LM Studio。
- 通过内置目录选择型号（如 Falcon 或 GPT-Neo-X）并下载
- 通过类似 ChatGPT 的直观界面使用模型，但完全本地化

	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

图 3.3-4 流行的本地开源 LLM - 模型比较。

模型的选择取决于对速度、精度和可用硬件能力的要求（图 3.3-4）。Mistral 7B 和百川 7B 等小型模型适用于轻量级任务和移动设备，而 DeepSeek -V3 等功能强大的模型则需要大量计算资源，但性能高且支持多种语言。未来几年，LLM 市场将快速增长--我们将看到越来越多的轻量级和专用模型。我们将看到越来越多

的轻量级和专业化模型出现，而不是覆盖所有 人类内容的通用 LLM，而是根据狭窄领域的专业知识训练的模型。例如，我们有望看到专门用于处理工程计算、建筑估算或 CAD 格式数据的模型出现。这种专业化的模型使用起来将更快、更准确、更安全--尤其是在对高可靠性和主题深度要求很高的专业环境中。

本地 LLM 启动后，可根据公司的具体任务进行调整。为此，可采用微调技术，根据内部文件、技术说明、合同模板或项目文件对模型进行进一步培训。

RAG: Intelligent LLM - 可访问公司数据的助理

LLM 在企业中应用的下一个发展阶段是将模型与实际的实时企业数据相结合。这种方法被称为 RAG（Retrieval-Augmented Generation）--检索-增强生成。在这种架构中，语言模型不仅仅是一个对话界面，而是一个全面的智能助手，能够浏览文档、图纸和数据库，并提供准确的上下文答案。

> RAG 的主要优势在于能够利用公司内部数据，无需预先训练模型，同时保持信息处理的高准确性和灵活性。

RAG 技术由两个主要部分组成：

- 检索：模型连接到数据存储--文档、表格、PDF -文件、图纸--并根据用户要求检索相关信息。
- 增强生成：根据提取的数据，该模型可生成准确、明智的回复，同时考虑到查询的上下文和特殊性。

要在 RAG 支持下运行 LLM ，需要遵循几个步骤：

- 数据准备：收集必要的文件、图纸、规格和表格。它们可以是不同的格式和结构，从 PDF 到 Excel。
- 索引和矢量化：使用 LlamaIndex 或 LangChain 等工具，将数据转换为矢量表示法，以便查找文本片段之间的语义链接（有关矢量数据库 和将大型数组转换为矢量表示法的更多信息，包括 CAD 项目，请参见第 8 部分）。
- 查询助手：上传数据后，您可以向模型提问，它将 在企业框架内搜索答案，而不是从互联网上收集

常识。

假设一家公司有一个 constructionite_docs 文件夹，其中存储了合同、说明、估算 和表格。使用 Python 脚本（图 3.3-5），我们可以扫描该文件夹并建立矢量索引：每份文档都将转换成一组反映文本语义内容的矢量。这样，文档就变成了一种 "意义地图"，模型可以在上面高效地导航并找到术语和短语之间的联系。

例如，该模型 "记住"，"退货 "和 "投诉 "这两个词经常出现在合同中有关将材料运往施工现场的部分。然后，如果有人提问，例如 "我们的退货期限是多长？(图 3.3-5 - 代码第 11 行） - LLM 将分析内部文件并找到准确信息，就像一个智能助手，能够阅读并理解所有公司文件的内容。

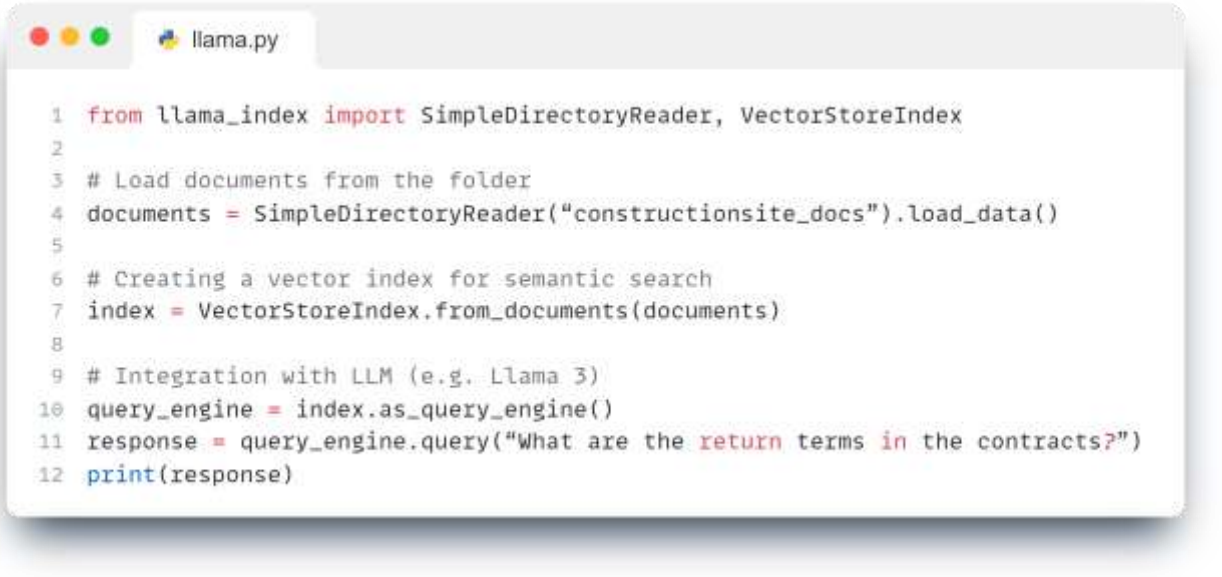
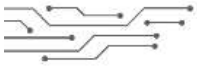


图 3.3-5 LM 读取文件夹中的文件--类似于人打开文件夹并搜索所需文件的方式

代码可以在任何安装了 Python 的计算机上运行。我们将在下一章详细介绍如何使用 Python 和集成开发环境来运行代码。

本地部署 LLM 不仅是一种趋势，也是重视安全性和灵活性的公司的战略解决方案。然而，无论是在本地公司计算机上还是使用在线解决方案，，部署 LLM 只是第一步。为了将 LLM 功能应用到实际任务中，公司必须使用一些工具，使其不仅能接收聊天回复，还能将创建的逻辑以代码的形式存储起来，以便在使用 LLM 的环境之外运行。这对于扩展解决方案非常重要--合理组织的流程可以让人工智能开发 同时应用于多个项目，甚至整个公司。

在这种情况下，选择合适的开发环境（IDE）起着重要作用。现代 编程工具不仅可以开发基于 LLM 的解决方案，还可以将其集成到现有的业务流程中，将其转化为自动化 ETL -Pipeline



章 3.4.

与 **llm** 支持的 **ide** 和未来的程序更改

选择集成开发环境：从 **LLM** 实验到商业解决方案

在进入自动化、数据分析和人工智能领域时，尤其是在使用大型语言模型（LLM）时，选择合适的集成开发环境（IDE）至关重要。集成开发环境将是您的主要工作工具：LLM 生成的代码将在本地计算机和公司网络上运行。集成开发环境的选择不仅决定了您工作的便利性，还决定了您从 LLM 实验性查询到嵌入实际业务流程的成熟解决方案的速度。

> 集成开发环境（**IDE**）是计算机上的一个多功能构件，用于流程自动化和数据处理。您不再需要单独存放锯子、锤子、钻头和其他工具，只需一台设备就能完成所有工作--切割、紧固、钻孔，甚至检查材料质量。程序员的集成开发环境是一个单一的空间，您可以在这里编写代码（与建筑类比--创建图纸）、测试其工作（建筑模型组装）、查找错误（如检查建筑结构的强度）并运行已完成的项目（房屋试运行）。

流行集成开发环境概述

- **PyCharm®** (JetBrains) 是一款功能强大的 Python 专业集成开发环境。由于具有大量内置功能，它非常适合严肃项目。不过，只有付费版本才提供对交互式 Jupyter 文件 (IPYNB) 的基本支持，初学者可能会发现界面过于复杂。

> 扩展名为 IPYNB 的文件（交互式 Python Notebook ）是一种用于交互式 Jupyter® Notebook 的格式（图 3.4-1），其中代码、可视化和解释都被整合到一个文档中。这种格式非常适合构建报告、分析 和培训场景。

- **VS Code®** (Microsoft) 是一款快速、灵活、可定制的工具，免费支持 IPYNB 和许多插件。适合初学者和专业人士使用。允许集成 GitHub Copilot 和语言模型插件，是人工智能 和数据科学项目的最佳选择。

- **Jupyter Notebook** - 用于实验和学习的经典流行选择。通过它，您可以在一个界面上编写代码、添加解释并将结果可视化（图 3.4-1）。它是快速测试假设、使用 LLM 和创建可重现的 野生数据分析步骤的理想工具。要管理依赖关系和库，我们建议使用 Anaconda Navigator，这是一个管理 Python 环境的可视化界面。

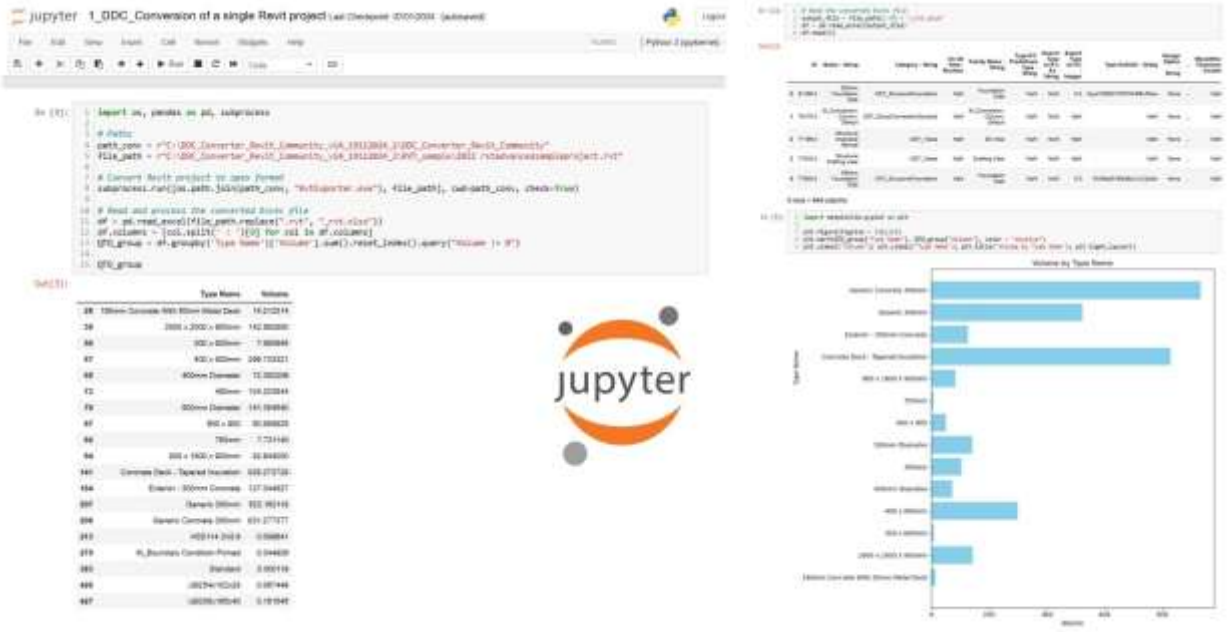


图 3.4-1 Jupyter Notebook 创建 Pipeline 流程最方便、最流行的工具之一。

- **Google Collab™** （以及 Kaggle 平台 （图 9.2-5））是基于云的 Jupyter 替代方案，提供免费 GPU/TPU 访问。它是一个很好的入门解决方案--无需安装本地软件，可以直接从浏览器工作。它支持与 Google Drive 集成，最近还支持与 Gemini（Google 的 LLM ）集成。

	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

图 3.4-2 集成开发环境比较：Jupyter Notebook 创建管道 流程最方便、最简单的工具之一。

集成开发环境的选择取决于您的任务。如果您想快速开始使用人工智能，请尝试使用 Jupyter Notebook 或 Google Collab。对于严肃的项目，最好使用 PyCharm 或 VS Code。最重要的是开始工作。现代工具可让您快速将实验转化为工作解决方案。

上述所有集成开发环境都允许您创建数据处理流水线，即代码块模块链（可由 LLM 生成），例如，每个模块负责不同的阶段：

- 分析方案、
- 从文件中提取信息的链条、
- 基于 RAG 的自动响应 、
- 生成报告和可视化效果。

得益于模块化结构，每个步骤都可以表示为一个单独的模块：数据加载→ 过滤→ 分析→ 可视化→ 输出结果。这些模块可以重复使用、改编并组装成新的数据链，就像一个构造函数。

对于工程师、管理人员和分析师来说，这就为以代码形式记录决策逻辑提供了可能性，而这些代码可通过 LLM 生成。这种方法有助于加快例行任务的速度，实现典型操作的自动化，并创建可重复的流程，其中的每一步都有清晰的文档记录，并对所有团队成员透明。

自动化 ETL Pipelines（图 7.2-3）、Apache Airflow（图 7.4-4）、Apache NiFi（图 7.4-5） 和 n8n（图 7.4-6）用于构建流程自动化逻辑块的工具将在本书第 7 部分和第 8 部分中详细讨论。

支持 **LLM** 的集成开发环境 和未来的编程变更

人工智能 与开发流程的整合正在改变编程领域的格局。现代环境不再只是带有语法高亮的文本编辑器，它们正在变成能够理解项目逻辑、完成代码，甚至解释特定代码片段如何工作的智能助手。市场上出现了一些利用人工智能突破传统开发界限的产品：

- **GitHub Copilot** （与 VS Code、PyCharm 集成）：人工智能 助手，可根据注释或部分描述生成代码，将文本提示转化为成品解决方案。
- **Cursor**（VS Code 的分叉，内核为 AI ）：不仅可以完成代码编写，还可以向项目提问、查找依赖关系并从代码库中学习。
- **JetBrains AI Assistant**：JetBrains IDE（包括 PyCharm）的插件，具有解释复杂代码、优化和创建测试的功能。
- **Amazon CodeWhisperer**：类似于 Copilot ，侧重于安全性和对亚马逊 AWS 服务的支持。

未来几年，编程将发生巨变。主要重点将从日常代码编写转向模型和数据架构设计--开发人员将更多地参与系统设计，而人工智能 将接管模板任务：代码生成、测试、文档和基本功能。编程的未来是人类与人工智能之间的合作，机器接管日常技术工作，人类则专注于创造力。

自然语言编程将成为家常便饭。集成开发环境的个性化将达到一个新的水平--开发环境将通过预测模式、提供上下文解决方案和学习以前的项目，学会适应用户的工作方式和公司

> 这不会取消开发人员的角色，但会彻底改变其角色：从编写代码到知识、质量和流程管理。这种演变还将影响到商业智能，在人工智能和 LLM、聊天和代理界面的帮助下，报告、可视化和决策支持应用程序的创建将越来越多地通过生成代码和逻辑来完成。

一旦公司建立了 LLM 聊天系统并选择了合适的开发环境，下一个重要步骤就是组织数据。这一过程包括从不同来源中提取信息、清理信息、将信息转化为结构化形式并将其整合到企业系统中。

在以数据为中心的现代数据管理方法中，一个关键目标是将数据转换成与大量工具和应用程序兼容的单一通用形式。处理结构化流程和结构化数据需要专门的库。其中功能最强大、最灵活、最受欢迎的是 Pandas 库 for Python。它可以方便地处理表格数据：过滤、分组、清理、追加、聚合和报告。

Python Pandas：处理数据不可或缺的工具

Pandas 在数据分析和自动化领域占有特殊地位。它是 Python 编程语言 [86]中最流行、使用最广泛的库之一，专为处理结构化数据而设计。

程序库就像是一套现成的工具：函数、模块、类。就像在建筑工地上，你不需要每次都发明锤子或水平仪一样，编程库也能让你快速解决问题，而无需重新发明基本功能和解决方案。

> Pandas 是一个开源 Python 库 ，提供高性能和直观的数据结构，特别是 DataFrame ，一种处理表格的通用格式。Pandas 是分析师、工程师和开发人员处理数据的瑞士刀。

Python 是一种高级编程语言，语法简单，被广泛应用于分析、自动化、机器学习和网络开发等领域。其受欢迎的原因在于代码的可读性、跨平台性和丰富的库生态系统。迄今为止，为 Python 创建的开源软件包已超过 137,000 个[87]，而且这个数字几乎每天都在增长。每一个这样的库都是一个现成的函数库：从简单的数学运算到图像处理、大数据分析、神经网络和与外部服务集成的复杂工具。

> 换句话说，想象一下，您可以免费、开放地访问成千上万个开箱即用的软件解决方案--可以直接嵌入业务流程的库和工具。这就像一个庞大的自动化、分析、可视化、集成等应用程序目录，而所有这些都可以在安装 Python 后立即使用。

Pandas 是 Python 生态系统中最流行的软件包之一。2022 年，Pandas 库 的平均下载量达到每天 400 万次（图 3.4-3），而到 2025 年初，这一数字已增加到每天 1200 万次，这反映出它在数据分析和 LLM 聊天[86]越来越受欢迎和广泛使用。

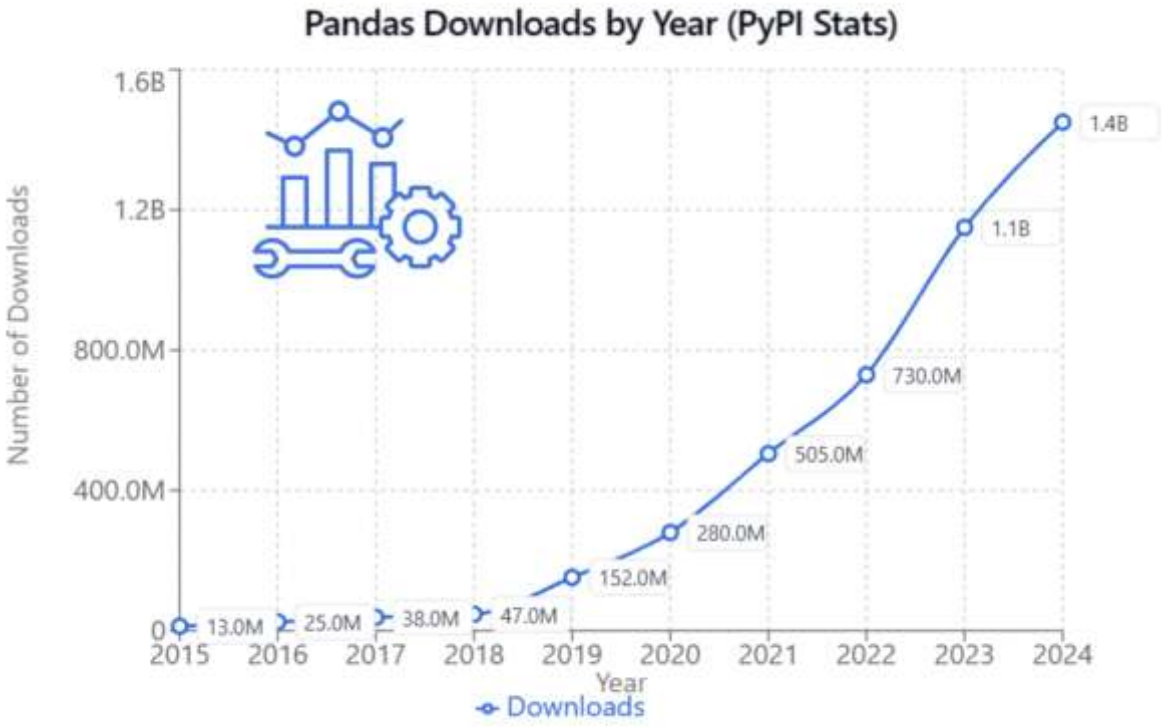


图 3.4-3 Pandas 是下载量最大的库之一。2024 年，其年下载量超过 14 亿次。

Pandas 库 中的查询语言在功能上与 SQL 查询语言 相似，我们在 "关系数据库 和 SQL 查询语言 "一章中讨论过 SQL 查询语言。

> 在分析 和结构化数据管理领域，Pandas 以其简单、快速和强大而脱颖而出，为用户提供了多种有效分析和处理信息的工具。

SQL 和 Pandas 这两种工具都具有强大的数据处理能力，尤其是与传统 Excel 相比。它们都支持选择、过滤等操作（图 3.4-4），唯一不同的是，SQL 针对关系数据库进行了优化，而 Pandas 则在 RAM 中处理数据，因此可以在任何计算机上运行，无需创建数据库和部署单独的基础设施。



图 3.4-4 Pandas，与 SQL 不同，它可以灵活地处理各种数据格式，而不仅限于数据库。

Pandas 通常是科学研究、流程自动化、创建管道（包括 ETL ）和在 Python 中进行数据操作的首选，而 SQL 是一种数据库管理标准，通常用于企业环境中处理大量数据。

> Python 编程语言 的 Pandas 库 不仅可以执行读写表等基本操作，还可以执行更复杂的任务，包括合并数据、分组数据和执行复杂的分析计算。

如今，Pandas 库不仅用于学术研究和商业分析，还与 LLM - 模型结合使用。例如，Meta® 部门（Facebook™）在 2024 年发布新的开源模型 LlaMa 3.1 时，特别关注结构化数据的处理，其发布的关键和首批案例之一正是处理 CSV 格式的结构化数据帧（图 3.4-5），并在聊天中直接与 Pandas 库集成。

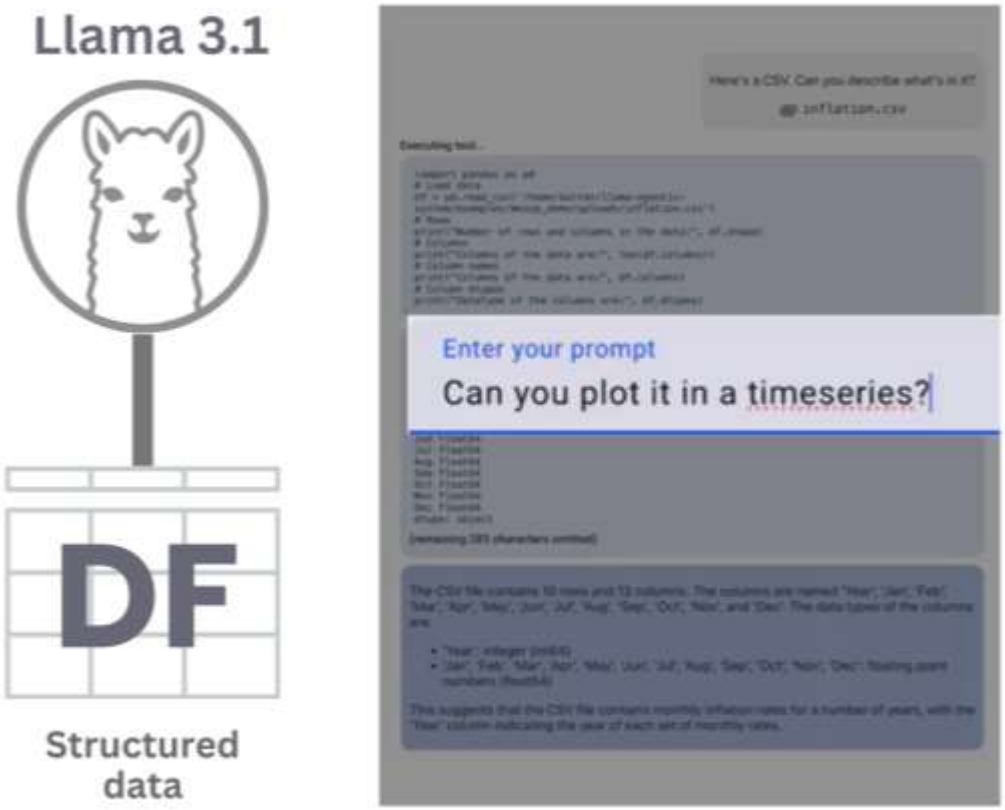


图 3.4-5 2024 年，Meta 团队在 LlaMa 3.1 中提出的第一个主要案例之一是使用 Pandas 构建应用程序。

> Pandas 是数百万数据科学家处理和准备生成式人工智能数据的重要工具。在不修改代码的情况下加速 Pandas 将是一个巨大的进步。数据科学家将能够在几分钟而不是几小时内处理数据，并获得更多数量级的数据来训练生成式人工智能模型[88]。
>
> - 英伟达公司创始人兼首席执行官黄仁勋

使用 Pandas ，管理和分析数据集的能力远远超过 Excel。Excel 通常能够处理多达 100 万行的数据，而 Pandas 可以轻松处理包含 千万行的数据集（图 9.1-2，图 9.1-10）[89]。这一功能允许用户对大型数据集进行复杂的数据分析和可视化，提供深刻的见解，促进数据驱动的决策制定。此外，Pandas 还拥有强大的社区支持[90]：全球数以亿计的开发人员和分析师（Kaggle.com、Google Collab、Microsoft® Azure™ Notebooks、Amazon SageMaker）每天在线或离线使用它，为任何业务问题提供大量开箱即用的解决方

案。

大多数 Python 分析流程的核心是一种名为 DataFrame 的结构化数据形式，由 Pandas 库 提供。它是组织、分析和可视化表格数据的强大而灵活的工具。

DataFrame：通用表格数据格式

DataFrame 是 Pandas 库中的核心结构，它是一个二维表格（图 3.4-6），其中行对应单个对象或记录，列对应其特征、参数或类别。这种结构在视觉上类似于 Excel 电子表格，但在灵活性、可扩展性和功能性方面要优越得多。

> 数据帧是一种表示和处理存储在计算机 RAM 中的表格数据的方法。

DataFrame 是一种表示和处理存储在计算机 RAM 中的表格数据的方式。例如，在表格中，行可以反映建筑项目的要素，列则反映其属性：类别、尺寸、坐标、 成本、术语等。此外，这种表格既可以包含一个项目的信息（图 4.1-13），也可以包含来自数千个不同项目的数百万个对象的数据（图 9.1-10）。有了 Pandas 的矢量化操作，就可以轻松地高速过滤、分组和汇总这些海量信息。

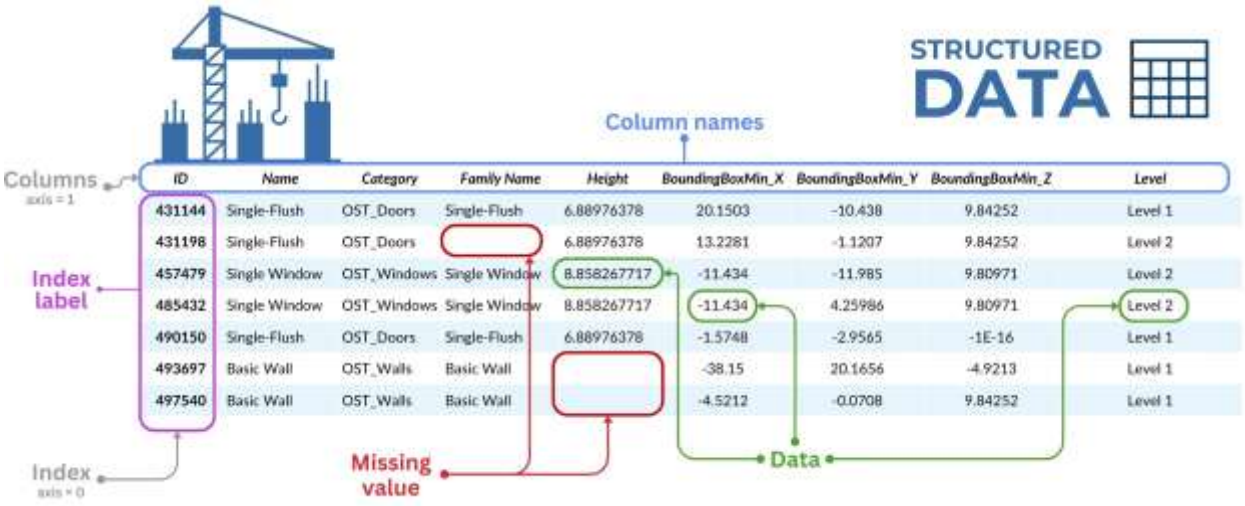


图 3.4-6 作为数据帧的建筑项目 是一个二维表格，行中有元素，列中有属性。

据 Nvidia 估计，目前已有高达 30% 的计算资源用于处理结构化数据（数据帧），而且这一比例还在继续增长。

> 世界上大概三分之一的计算工作都是在每家公司进行的数据处理。大多数公司的数据处理和数据都是以 DataFrame 表格格式进行的
>
> - Jensen Huang，Nvidia 首席执行官 [91]

让我们列出 Pandas 中 DataFrame 的一些主要功能：

- 列：在 DataFrame 中，数据被组织成列，每列都有一个唯一的名称。属性列可以包含不同类型的数据，类似于数据库中的列或表格中的列。
- **Pandas** 系列是 Pandas 中的一维数据结构，类似于表格中的列表或列，其中每个值对应不同的索引

> Pandas 系列拥有 400 多个属性和方法，使数据处理变得异常灵活。您可以直接将四百个可用函数中的一个应用到列、执行数学运算、过滤数据、替换值、处理日期、字符串等。此外，Series 还支持矢量化操作，与循环计算相比，这大大加快了处理大型数据集的速度。例如，你可以轻松地将所有数值乘以一个数字，替换缺失的数据，或应用复杂的变换，而无需编写复杂的循环。

- 行：在 DataFrame 中，可以用唯一值建立索引。通过该索引，您可以快速更改和调整特定行中的数据。
- 索引：默认情况下，当您创建 DataFrame Pandas 时，会为每一行分配一个从 0 到 N-1 的索引（其中 N 是 DataFrame 中所有行的数量）。不过，索引可以更改，以包含日期或唯一特征等特殊指定。
- 对 DataFrame 中的行进行索引是指为每一行分配一个唯一的名称或标签，称为 DataFrame 索引。
- 数据类型：DataFrame 支持多种数据类型，包括 int"、"float"、"bool"、"datetime64 "和用于文本数据的 "obect"。每个 DataFrame 列都有自己的数据类型，这些数据类型决定了可以对其内容执行哪些操作。
- 数据操作：DataFrame 支持多种数据处理操作，包括聚合（"groupby"）、合并（"merge "和 "join"）、连接（"concat"）、拆分-应用-合并以及许多其他数据转换技术。

- 大小操作：DataFrame 允许添加和删除列以及 行，使其成为一种可根据数据分析需要进行修改的动态结构。
- 可视化 数据：使用内置可视化技术或与 Matplotlib 或 Seaborn 等流行的数据可视化库接口，DataFrame 可轻松转换成图形和图表，以图形方式展示数据。
- 数据输入和输出：Pandas 提供了读取导入和导出数据到各种文件格式的函数，如 CSV 、Excel 、JSON 、HTML 和 SQL ，可能使 DataFrame 成为数据收集和分发的中心枢纽。

与 CSV 和 XLSX 不同，Pandas DataFrame 在处理数据时提供了更高的灵活性和性能：它可以在 RAM 中处理大量信息，支持扩展数据类型（包括日期、逻辑值和时间序列），并提供过滤、聚合、合并和可视化数据的广泛功能。虽然 CSV 不能存储数据类型和结构信息，XLSX 通常格式化过多且可扩展性低，但 DataFrame 仍然是快速分析 、流程自动化以及与人工智能 - 模型集成的最佳选择（图 3.4-7）。以下各章将详细探讨数据的上述各个方面，本书第 8 部分还将详细讨论类似格式，如 Parquet 、Apache Orc、JSON 、Feather 、HDF5 和数据仓库（图 8.1-2）。

	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)

图 3.4-7 DataFrame 是数据处理的最佳选择，具有高性能和高级数据类型支持。

凭借其灵活性、强大功能和易用性，Pandas 库和 DataFrame 格式已成为 Python 数据分析的事实标准。

它们是创建简单报告和构建复杂分析管道的理想，尤其是与 LLM 模型结合使用时。

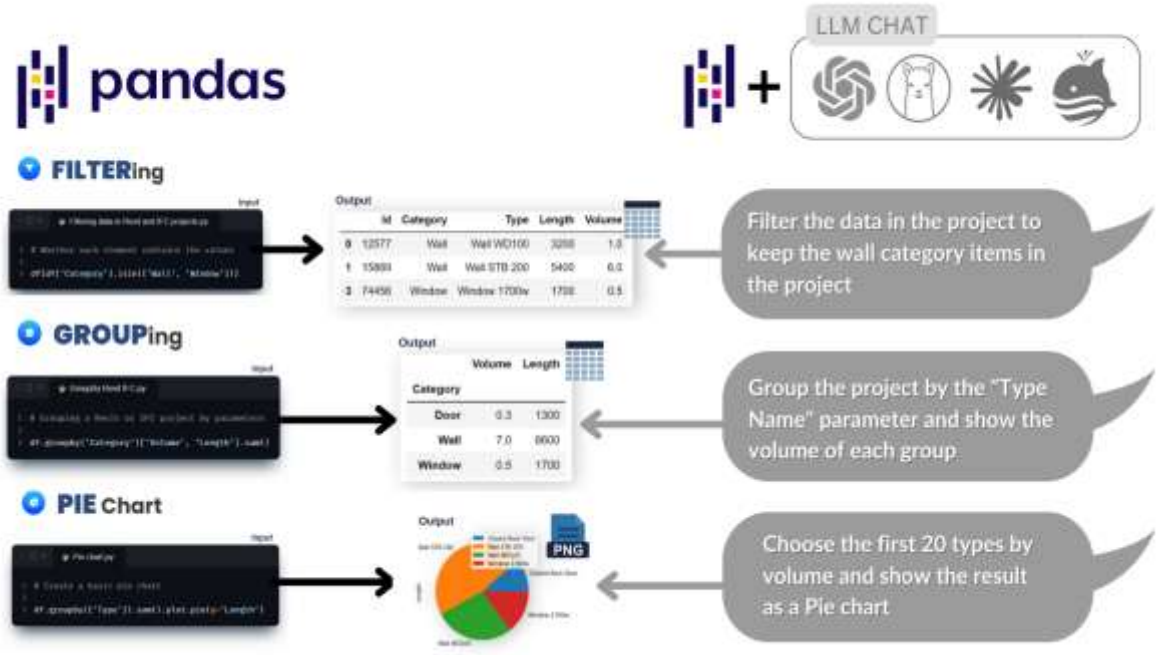


图 3.4-8 LLMs 简化了与 Pandas 的交互：无需代码，只需文本查询即可。

如今，Pandas 已被积极用于基于 LLM 的聊天室，如 ChatGPT 、LlaMa 、DeepSeek 、QWEN 等。在许多情况下，当模型接收到与表格处理、数据验证或分析有关的查询时，它会使用 Pandas 库准确生成代码。这使得 DataFrame 成为在与人工智能对话中表示数据的自然 "语言"（图 3.4-8）。

Pandas 等现代数据技术使数据分析、自动化和集成到业务流程中变得更加容易。它们能快速交付结果，减少专家的工作量，并确保可重复操作。

下一步：建立可持续的数据框架

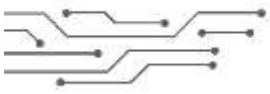
在这一部分中，我们了解了建筑行业使用的主要数据类型，学习了存储这些数据的不同格式，并分析了现代工具（包括 LLM 和集成开发环境）在处理信息方面的作用。我们了解到，有效的数据管理是明智决策和业务流程自动化的基础。能够对数据进行结构化和组织的组织在数据处理和转换阶段会获得显著的竞争优势。

在总结这部分内容时，值得强调的是主要的实际步骤，这些步骤将有助于您在日常工作中应用所讨论的方法：

- 对流程进行数据审计
  - ☐ 清点项目中使用的所有数据类型
  - ☐ 确定哪些数据类型和模型对您的业务流程最为关键
  - ☐ 确定信息经常缺乏条理、结构混乱或无法获取的问题领域
- 开始形成数据管理战略
  - ☐ 提出处理不同数据类型的政策问题和标准
  - ☐ 分析将非结构化数据转换为结构化数据可改进哪些工作流程
  - ☐ 制定考虑安全性和保密性的数据存储和访问政策
- 安装并掌握处理数据的基本工具
  - ☐ 根据任务选择合适的集成开发环境（例如安装 VS Code 或 Jupyter Notebook）
  - ☐ 尝试建立本地法律硕士，以保密方式处理您的个人数据
  - ☐ 开始尝试使用 Pandas 库 处理 XLSX 表格数据
  - ☐ 向 LLM 描述您在电子表格工具或数据库中处理的典型任务，并要求 LLM 使用 Pandas 自动完成这些工作

应用这些步骤将使您逐步转变处理数据的方法，从分散、非结构化的数据集转变为统一的生态系统，使数据成为可访问、可理解的资产。从小事做起--在 Pandas 中创建第一个 DataFrame，运行本地 LLM，使用 Python 自动执行第一个常规任务（如 Excel 电子表格）。

本书第四部分的重点是数据质量 、数据组织、结构化和建模。我们将重点介绍将不同数据源（从 PDF 和文本到图像和 CAD 模型）转换为适合分析和自动化的结构化数据集的方法。我们还将探讨如何将数据要求正规化，如何在建设项目中建立概念和逻辑模型，以及现代语言模型 (LLM) 如何在这一过程中提供帮助。





IV 部分

数据质量：组织、结构化、建模

第四部分侧重于将不同信息转化为高质量结构化数据集的方法和技术。详细讨论了形成和记录数据需求的过程，以此作为建筑项目中有效信息架构的基础。介绍了从各种来源（PDF -文档、图像、文本文件、CAD -模型）提取结构化信息的实用方法和实施实例。分析了使用正则表达式 (RegEx) 和其他工具自动验证和核实数据的方法。考虑到建筑行业的特殊性，逐步介绍了概念、逻辑和物理层面的数据建模过程。还展示了使用语言模型 (LLM) 自动构建和验证信息的具体实例。此外，还提出了分析结果可视化的有效方法，为各级建筑项目管理人员提供更多的分析信息

章 4.1.

将数据转换为结构化格式

在数据驱动经济时代，数据正在成为决策的基础而不是障碍。企业不再根据每个新系统及其格式不断调整信息，而是越来越多地寻求形成一个单一的结构化数据模型，作为所有流程的通用真相来源。现代信息系统不是围绕格式和界面设计的，而是围绕数据的意义设计的--因为结构可能会改变，但信息的意义在更长的时间内是不变的。

> 有效使用数据的关键不是无休止地转换和变换数据，而是从一开始就正确地组织数据：创建一个通用结构，能够在项目生命周期的各个阶段提供透明度、自动化和集成。

每实施一个新平台，传统方法都要进行手动调整：迁移数据、更改属性名称和调整格式。这些步骤并不能提高数据本身的质量，只会掩盖问题，造成无休止转换的恶性循环。结果，企业变得依赖于特定的软件解决方案，数字化转型也随之放缓。

在接下来的章节中，我们将探讨如何正确构建数据结构，然后如何创建通用模型，最大限度地减少对平台的依赖，并将重点放在最重要的事情上--将数据作为一种战略资源，围绕它建立可持续流程。

学习如何将文档、**PDF** 、图片和文本转化为结构化格式

在建筑项目中，绝大多数信息都是以非结构化形式存在的：技术文件、工作说明书、图纸、规范、时间表和协议。这些信息在格式和内容上都多种多样，因此很难实现整合和自动化。

> 根据输入数据的类型和所需的处理结果，转换为结构化或半结构化格式的过程可能会有所不同。

将非结构化数据转换为结构化数据既是一门艺术，也是一门科学。这一过程因输入数据的类型和分析目的而异，通常占数据工程师（图 3.2-5）和分析师 工作的很大一部分，目的是生成干净、有序的数据集。



图 4.1-1 将非结构化扫描文档转换为结构化表格格式。

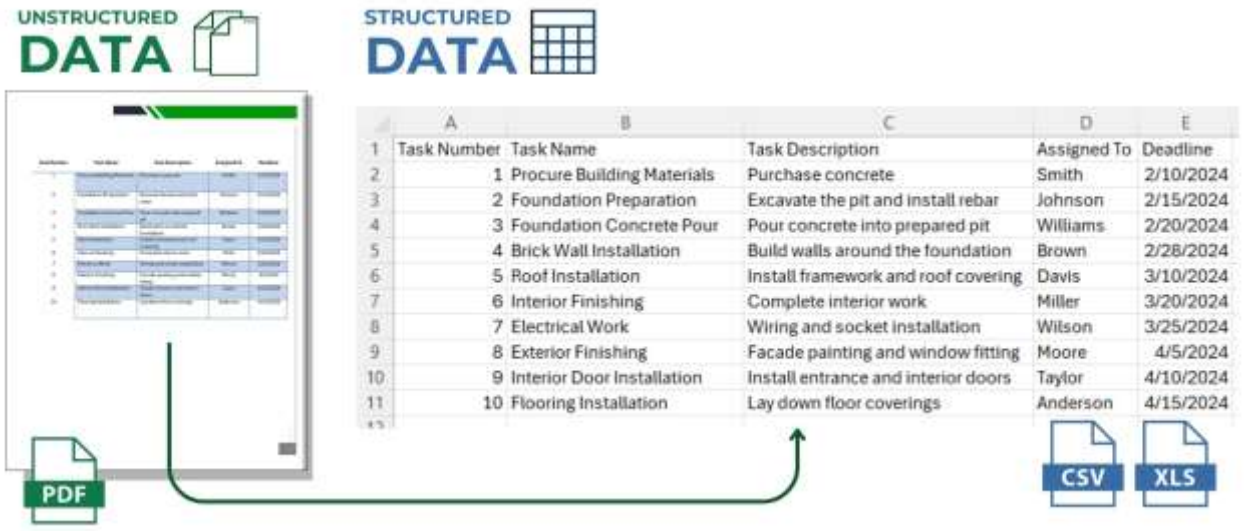
将文档、PDF 、图片和文本转换为结构化格式（图 4.1-1）是一个循序渐进的过程，包括以下步骤：

- 提取 ）：在此步骤中，将加载包含非结构化数据的源文件或图像。例如，可以是 PDF 文档、照片、图纸或示意图。
- 数据转换（Transform ）：接下来是将非结构化数据转换为结构化格式的步骤。例如，这可能涉及使用光学字符识别（OCR ）或其他处理方法识别和解释图像中的文本。
- 加载和保存数据（加载 ）：最后一步涉及将处理过的数据保存为各种格式，如 CSV 、XLSX 、XML 、JSON ，以便进一步工作，其中格式的选择取决于具体要求和偏好。

这一过程被称为 ETL （提取 ，转换 ，加载 ），在自动数据处理中起着关键作用，将在 "ETL 和管道 ：提取 、转换、加载 "一章中详细讨论。接下来，我们将举例说明如何将不同格式的文档转换为结构化数据。

将 **PDF** 文件转换为表格的示例

建筑项目中最常见的任务之一就是处理 PDF 格式的规范。为了演示从非结构化数据到结构化格式的过渡，让我们考虑一个实际例子：从 PDF 文档中提取表格并将其转换为 CSV 或 Excel 格式 （图 4.1-2）。



	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

图 4.1-2 与 PDF 不同，CSV 和 XLSX 格式非常普遍，很容易集成到各种数据管理系统中。

LLM 语言模型，如 ChatGPT、LlaMa、Mistral、DeepSeek、Grok、Claude、QWEN 大大简化了数据科学家处理数据的方式，减少了对编程语言深度学习的需求，使许多任务可以通过文本查询来解决。

因此，与其花时间在互联网（通常是 StackOverFlow 网站或专题论坛和聊天）上搜索解决方案或联系数据处理专家，我们可以使用现代在线或本地 LLM 的功能。只需提出查询，模型就会提供将 PDF 文档转换为表格格式的现成代码。

- 向任何 LLM - 模型（CHATGRT、LlaMa、Mistral、DeepSeek、Grok、Claude、QWEN 或任何其他）发送以下文本请求：

> 请编写一段代码，从包含表格的 PDF 文件中提取文本。代码应将文件路径作为参数，并将提取的表格作为 DataFrame 返回⏎

在大多数情况下，LLM - 模型响应将以 Python 的代码形式出现，因为这种语言被广泛用于数据处理、自动化和处理各种文件格式：

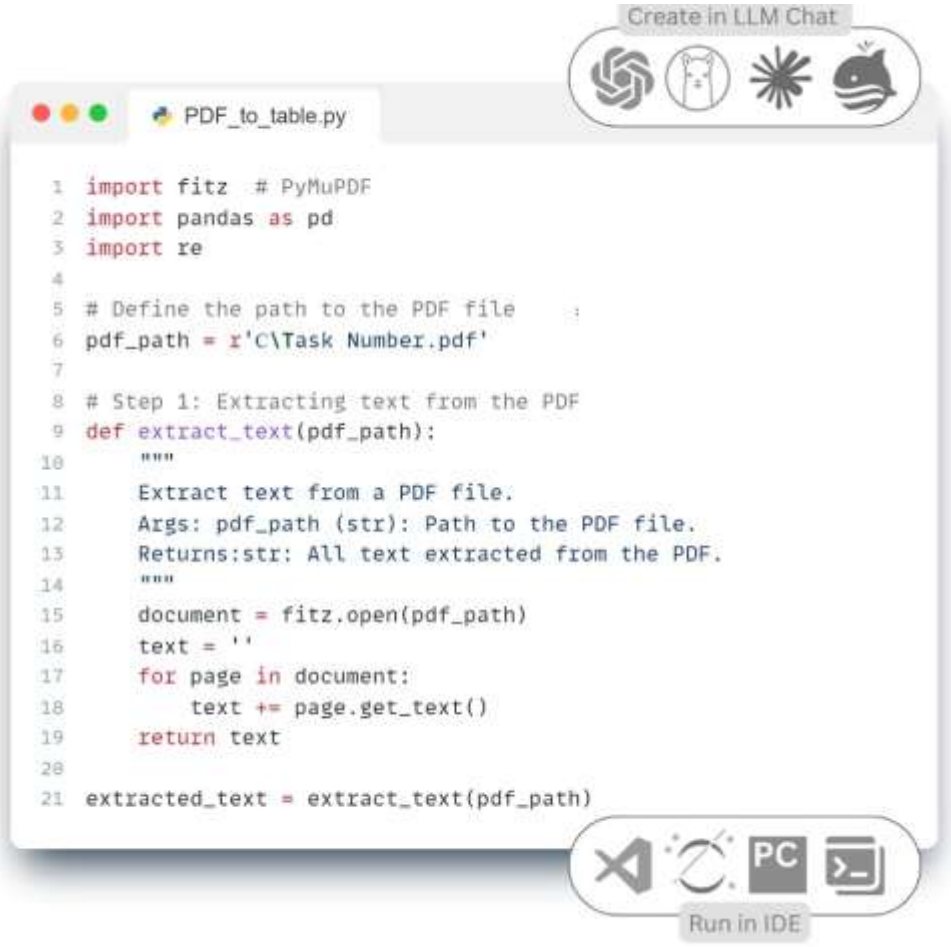


图 4.1-3 以 Python 代码 及其库和软件包（Pandas、Fitz）为形式的 LLM 响应从 PDF 文件中提取文本。

这段代码（图 4.1-3）可以在我们上面提到的流行集成开发环境之一中离线运行：PyCharm、Visual Studio Code (VS Code)、Jupyter Notebook 、Spyder、Atom、Sublime Text、带有 PyDev 插件的 Eclipse、Thonny、Wing IDE、带有 Python 插件的 IntelliJ IDEA 、JupyterLab 或流行的在线工具：Kaggle。com、Google Collab、Microsoft Azure Notebooks、Amazon SageMaker。

在 "转换 "步骤中，我们使用流行的 Pandas 库 （我们在 "Python Pandas：处理数据不可或缺的

工具 "一章中详细讨论过）将提取的文本读入 DataFrame ，并将 DataFrame 保存为 CSV 表格文件 或 XLXS：

> 我需要将 PDF 文件中的表格转换为 DataFrame 的代码。还要添加将 DataFrame 保存为 CSV 文件的代码。 ⏎

LLM 的答复：

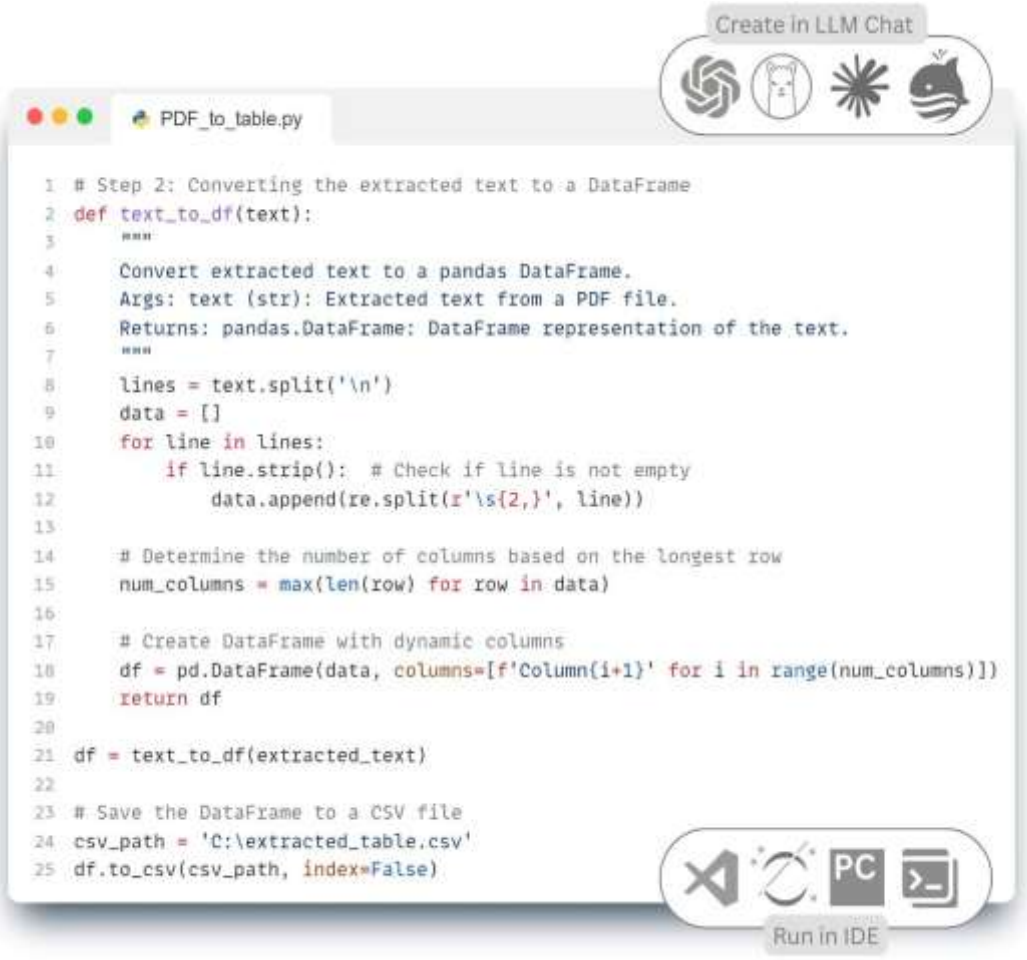


图 4.1-4 将提取的表格从 PDF 转换为 DataFrame 并将表格保存为 CSV 文件。

如果在代码执行过程中出现错误（图 4.1-3，图 4.1-4），例如由于缺少库或文件路径错误，只需将错误文本与源代码一起复制，然后重新提交给 LLM -模型即可。该模型将分析错误信息，解释问题所在，并提出修复

建议或其他步骤。

> 这样，与人工智能 LLM 的互动就变成了一个完整的循→→测试→ 反馈→ 修正 - 无需深厚的技术知识。

使用 LLM chat 中的纯文本查询和十几行 Python（我们可以在任何集成开发环境 中本地运行），我们将 PDF 文档转换为表格 CSV 格式，与 PDF 文档不同，CSV 格式易于机器阅读，并可快速集成到任何数据管理系统中。

我们可以将这段代码（图 4.1-3，图 4.1-4）从任何 LLM 聊天室服务器上的数以万计的新 PDF 文档中，从而自动完成将非结构化文档流转换为结构化 CSV 表格格式的过程。

但 PDF 文档 并不总是包含文本，更多的时候是需要作为图像处理的扫描文档。虽然图像本身是非结构化的，但识别库的开发和应用使我们能够提取、处理和分析图像内容，从而在业务流程中充分利用这些数据。

将 **JPEG**、**PNG** 图像转换为结构化形式

图像是非结构化数据最常见的形式之一。在建筑业和许多其他行业，大量的信息以扫描文件、示意图、照片和图纸的形式存储。这些数据包含有价值的信息，但无法直接进行处理，如 Excel 电子表格 或数据库。图像包含大量复杂信息，因为其内容、颜色和纹理各不相同，需要进行特殊处理才能提取 有用信息。

使用图像作为数据源的困难在于缺乏结构。图像不能像 Excel 电子表格 或数据库表格那样，以计算机可以立即理解或处理的直接、易于量化的方式传递意义。要将非结构化的图像数据转换成结构化的形式，必须使用能够解读图像信息的特殊库（图 4.1-5）。

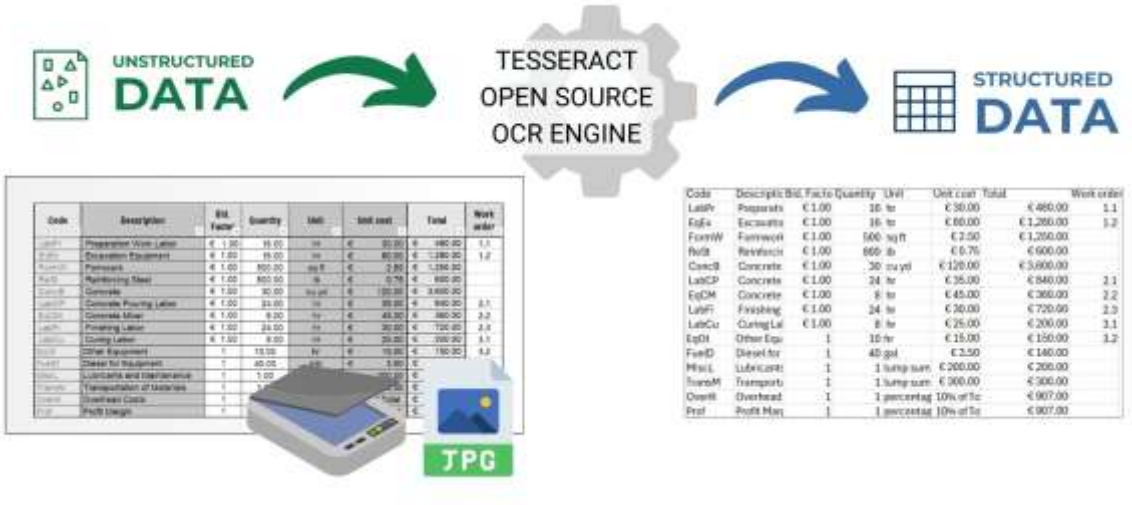


图 4.1-5 使用专门的 OCR 工具，可以将扫描文件和图像转换成结构化格式。

OCR （光学字符识别）技术用于从图像中提取文本。它可以识别文档扫描件、照片和 PDF 文件中的字母和数字，将其转化为可编辑和机器可读的文本。OCR 技术在文档自动化领域应用已久，如今已被轻松集成到任何业务流程和 Python 应用程序中。最流行的 OCR 工具之一是 Tesseract，它是一个开源 ，最初由 HP™ 开发，现在由 Google™ 支持。它支持 100 多种语言，识别准确率很高。

让我们请 LLM 聊天编写示例代码，以结构化的方式从扫描或拍摄的表格中检索数据。

- 向 LLM 聊天（ChatGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN 或其他）发送文本请求：

> 编写将包含表格的 JPEG 图像转换为 DataFrame 表格的代码 ⏎

在大多数情况下，LLM 的回复会建议使用 Pytesseract 库来识别图像中的文本：

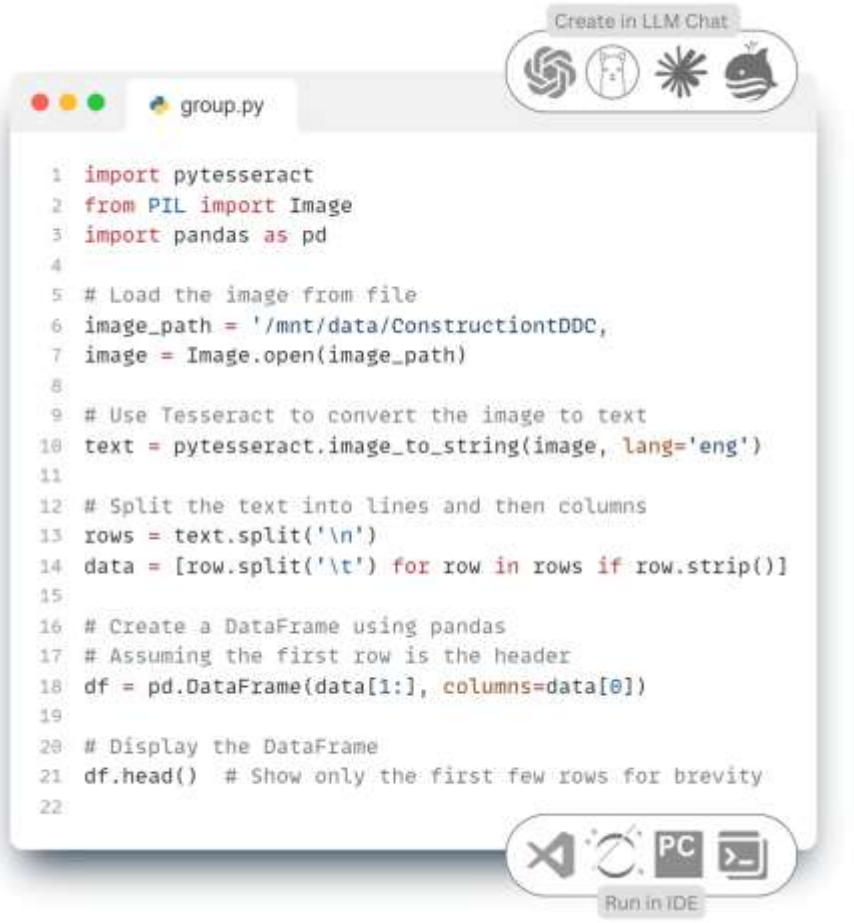


图 4.1-6 将从图像或照片表格中提取的文本转换为结构化表格。

在本例中，源于 LLM 的代码（图 4.1-6）使用 pytesseract 库（Tesseract for Python ）通过 OCR （光学字符识别）将图像转换为文本，并使用 Pandas 库 将文本转换为结构化形式，即 DataFrame。

> 转换过程通常包括预处理以提高图像质量，然后应用各种算法进行模式检测、特征提取或对象识别。这样，非结构化的视觉信息就被转换成了结构化数据。

虽然 PDF 和图像是非结构化信息的主要来源，但就数量而言，真正的 冠军是电子邮件、聊天、会议和信使中产生的文本。这些数据不仅数量多，而且分散、不正规、结构性极差。

将文本数据转换为结构化形式

除了带表格的 PDF 文件（图 4.1-2）和表格的扫描版本（图 4.1-5），项目文件中还有很大一部分信息是以文本形式呈现的。它既可以是文本文件中的连贯句子，也可以是散落在图纸和方案中的零散记录。在现代数据处理条件下，最常见的任务之一就是将这些文本转换成适合分析、可视化和决策 的结构化格式。

分类法是这一过程的核心，它是一种分类系统，可根据共同特征将信息组织成类别和子类别。

> 分类法是一种用于分组和组织对象的分层分类结构。在文本处理中，它是将项目系统地归入语义类别的基础，从而简化分析并提高数据处理质量。

分类法的创建需要经过实体提取、分类和上下文关联等步骤。为了模拟从文本数据中提取信息的过程，以下步骤与我们从 PDF 文档中构建数据的步骤类似：

- 提取)：您需要分析文本数据，提取有关项目进度延迟和变更的信息。
- 分类和分级 (Transofrm)：对收到的信息进行分类，如延误原因和计划变更。
- 整合（Load ）：最后，我们将结构化数据整合到外部数据管理系统中。

考虑这样一种情况：项目经理和工程师就进度问题进行对话。我们的目标是提取关键要素（延误原因、进度调整），并以结构化的方式呈现出来（图 4.1-7）。

让我们根据预期的关键字进行提取，创建一个 DataFrame 来模拟数据提取，并在转换后创建一个新的 DataFrame 表，该表将包含日期、事件（如延迟原因）和操作（如计划变更）列。

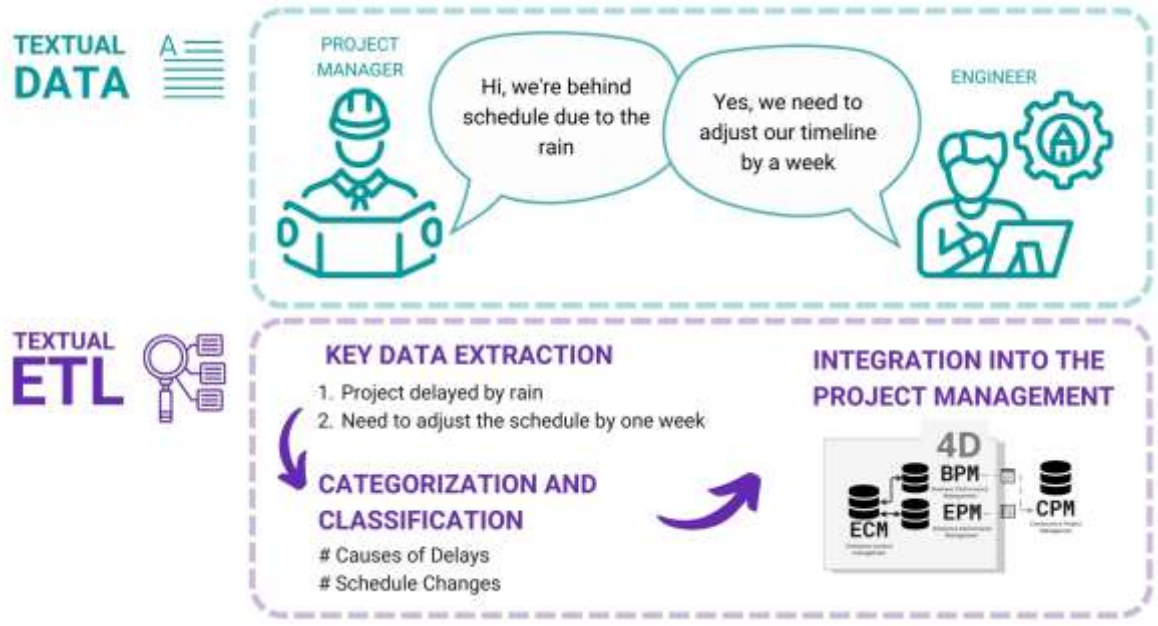


图 4.1-7 突出显示文本中关于需要调整时间表和将变更纳入项目管理系统的关键信息。

下面是使用一个语言模型中的文本查询来解决问题的代码，与前面的示例相同。

向任何 LLM 聊天室发送文本请求：

> 我有这样一段对话：一位经理说 "你好，由于下雨，我们的进度落后了"，一位工程师说 " 是的，我们需要将截止日期调整一周"。我需要一个脚本来分析未来类似的文本对话，从中提取延误的原因和必要的截止日期调整，然后根据这些数据生成一个 DataFrame。然后将 DataFrame 保存到 CSV 文件中。 ⏎

LLM 的回复通常包括使用正则表达式 (re - Regex) 和 Pandas (pd) 库 的 Python - 代码：

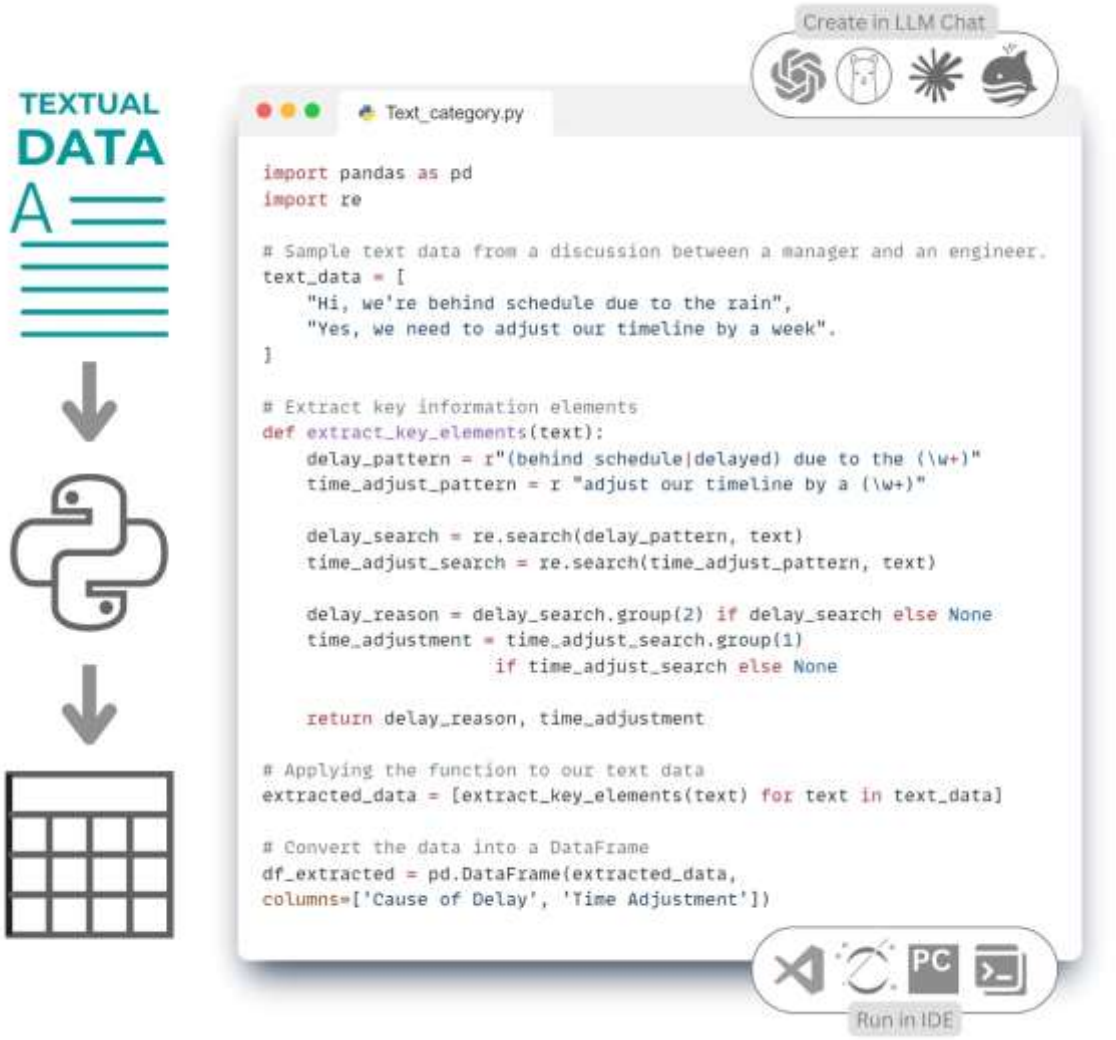
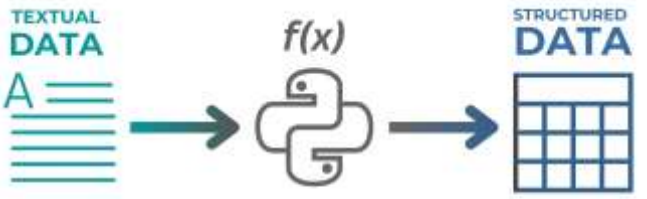


图 4.1-8 在表格中突出显示文本中关于需要调整截止日期的关键信息。

在这个示例中（图 4.1-7），对包含项目经理和工程师之间通信的文本数据进行了分析，以识别和提取可能会影响未来项目管理的类似对话的特定信息。使用正则表达式 （我们将在 "结构化需求 和 RegEx 正则表达式 "一章中讨论更多有关正则表达式的内容），通过模式识别出项目延误的原因和对时间进度的必要调整。本例中编写的函数根据模式从字符串中提取延迟原因或时间调整：高亮显示 "because of "后面的单词为延迟原因，或 "by "后面的单词为时间调整。

如果某行提到因天气原因造成的延误，"雨 "就会被识别为原因；如果某行提到某段时间的计划调整，该段时间就会被提取为时间调整（图 4.1-9）。如果某一行中没有这些词语，则相应属性列的值为 "无"。



Text string	Cause of Delay	Time Adjustment
0	rain	None
1	None	week

图 4.1-9 代码执行后得到的 DataFrame 汇总表包含有关存在延迟和必要时间调整的信息。

从文本（对话、信函、文件）中对条件进行结构化和参数化，可以快速消除施工中的延误：例如，在恶劣天气下，工人的缺乏会影响工作进度，因此，公司可以提前从施工现场工头和项目经理之间的对话（图 4.1-9）中了解延误参数，以便在出现不利预报时加强施工人员的力量。

使用相对简单、开放和免费的分类工具，可以将文件和图像转换为结构化格式。

对元素进行分类也是处理项目数据的一个关键部分，尤其是在使用 CAD 软件(BIM)的情况下。

将 **CAD** 数据 **(BIM)** 转换为结构化形式

对 CAD 数据进行结构化和分类 (BIM) 是一项更为复杂的任务，因为 CAD (BIM) 数据库中存储的数据几乎总是采用封闭或复杂的参数格式，通常同时结合了几何数据元素（半结构化）和元信息元素（半结构化或结构化数据）。

CAD (BIM) 系统中的原生数据格式通常受到保护，无法直接使用，除非有专门的软件 或 API - 开发人员自己的接口（图 4.1-10）。这种数据隔离形成了封闭的存储孤岛，限制了信息的自由交流，阻碍了公司端到端数字流程的创建。

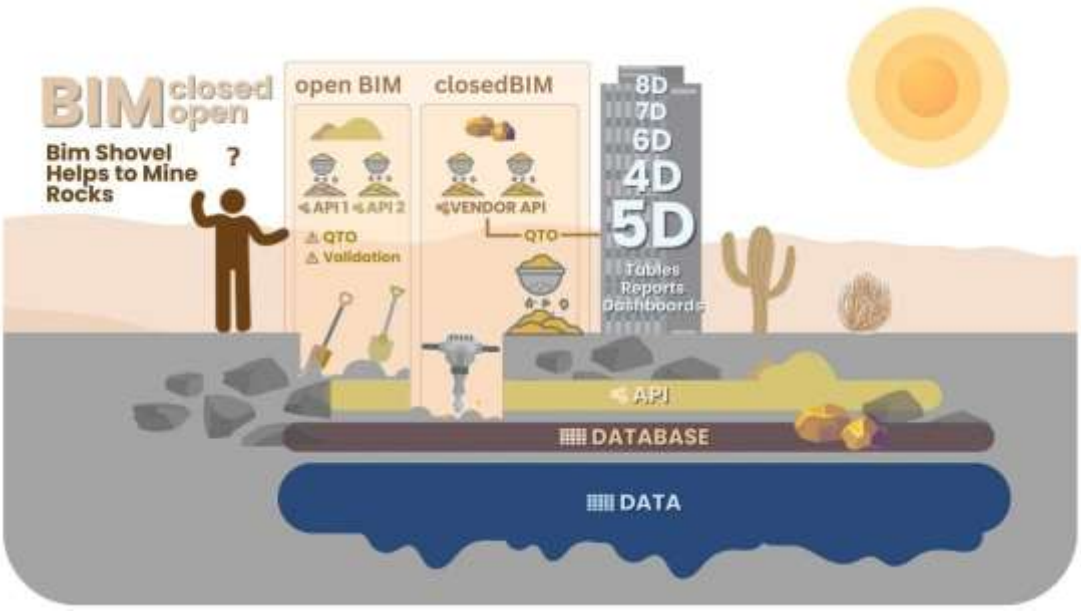


图 4.1-10 CAD 专业人员 (BIM) 可以通过 API - 连接或供应商工具访问本地数据。

在特殊的 CAD (BIM) 格式中，项目元素的特征和属性信息被收集在一个分层分类系统 中，具有相应属性的实体就像果树的果实一样，位于 数据分类分支的最新节点（图 4.1-11）。

从这种层次结构中提取数据有两种方法：一种是手动操作，点击每个节点，就像处理一棵树一样，用斧头砍下类别和类型的选定分支。另一种方法是使用应用编程接口（API），以更高效、自动化的方式检索和分组数据，最终将数据转换成结构化表格，供其他系统使用。

不同的工具，如 Dynamo、pyRvt、Pandamo（Pandas + Dynamo）、ACC 或开源解决方案，如用于 IFC 格式的 IfcOpSh 或 IFCjs，可用于从 CAD (BIM) 项目中提取结构化数据表。

为了简化数据处理和准备工作，现代数据导出和转换工具可以将 CAD 模型的内容分 为两个关键部分：几何信息和属性数据（图 4.1-13）--描述设计元素属性的元信息（图 3.1-16）。这两个数据层通过唯一标识符保持联系，因此可以将每个元素的几何描述（通过参数或多边形）精确映射到其属性：名称、材料、完成阶段、成本等。这种方法可确保模型的完整性，并允许灵活使用数据，既可用于可视化（几何模型数据），也可用于分析或管理任务（结构化或松散结构化），可分别或并行处理这两类数据。

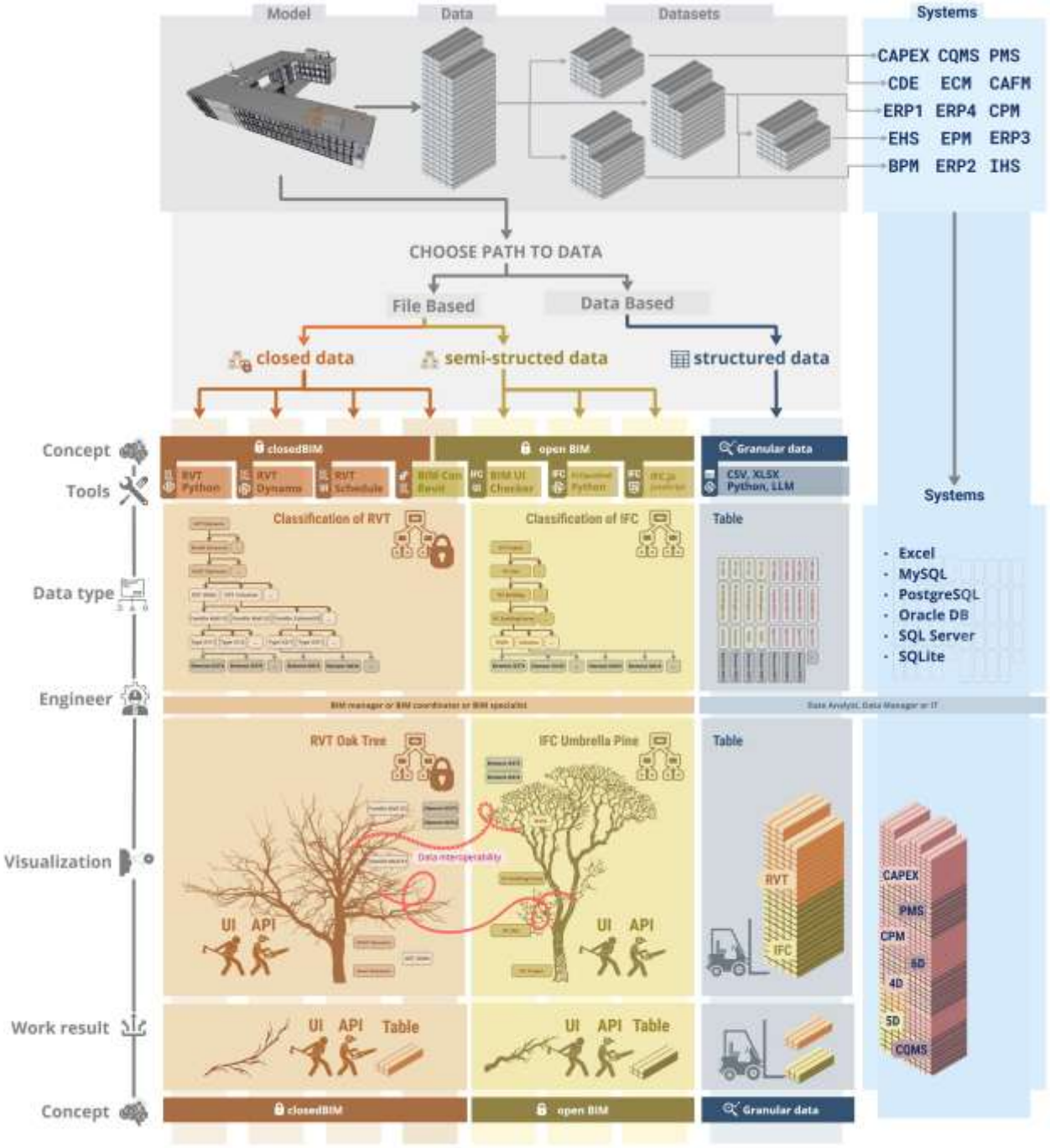


图 4.1-11 CAD 数据库的信息视图 (BIM) 以分类树的形式呈现给用户。

随着逆向工程技术 的发展和用于 CAD 数据转换的 SDK （软件开发工具包）的出现，从封闭的 CAD 软件格式 (BIM) 中获取和转换数据变得更加容易。现在，可以合法、安全地将数据从封闭格式 转换为通用格式，适合在其他系统中分析和使用。在 "结构化数据： 数字化转型的基础 "一章中讨论了第一批逆向工程工具（"

开放式 DWG"）的历史 以及争夺 CAD 供应商格式主导权的斗争。

> 逆向工程工具允许从封闭的专有格式中合法检索数据，将混合 CAD (BIM) 格式的信息分解为用户所需的数据类型和格式，使其更易于处理和分析。

利用逆向工程和直接从 CAD 数据库获取信息的方式，使信息可被访问，从而允许开放数据和开 放工具，以及使用标准工具进行数据分析、构建报告、可视化和与其他数字系统集成 （图 4.1-12）。

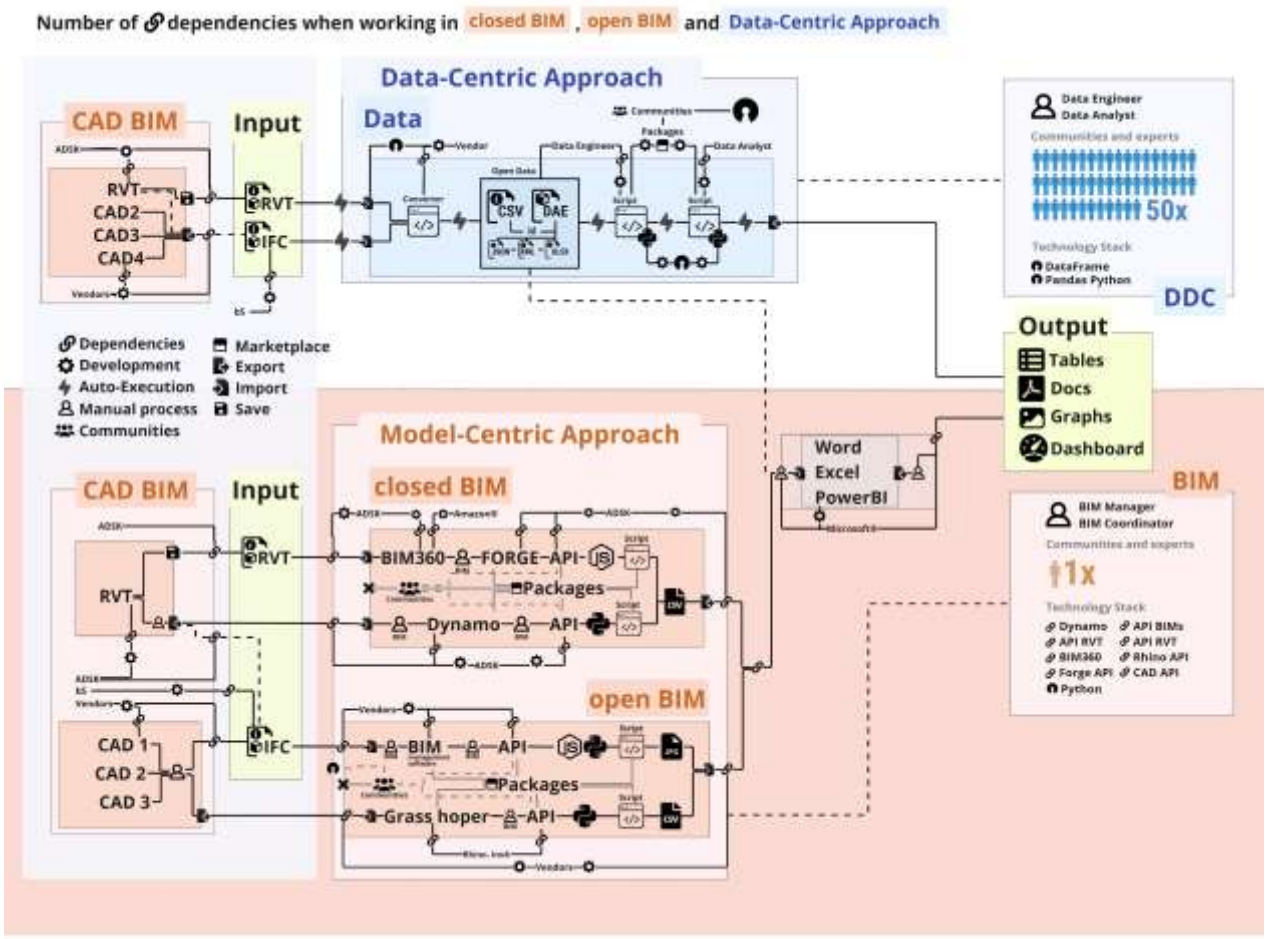


图 4.1-12 直接访问 CAD 数据可最大限度地减少对软件平台的依赖，并转向以数据为中心的方法。

自 1996 年以来，DWG 格式 ，2008 年以来，DGN 格式 ，2018 年以来，RVT ，借助逆向工程工具（图 4.1-13），可以方便高效地将最初封闭的 CAD 数据格式转换为任何 其他格式，包括结构化格式。如今，世界上几乎所有主要的 CAD (BIM) 和大型工程公司都使用 SDK - 逆向工程工具从封闭的 CAD (BIM) 供应商格

式中提取数据[92]。



图 4.1-13 使用逆向工程工具可以将 CAD (BIM) 程序数据库转换成任何方便的数据模型。

将数据从封闭的专有格式转换为开放格式，并将混合 CAD (BIM)格式分离为几何和元信息属性数据，可简化数据的处理过程，使其可用于分析、操作和与其他系统集成（图 4.1-14）。

> 在当今使用 CAD 数据 (BIM)的工作中，我们已经达到了无需征得 CAD（BIM）供应商许可即可访问 CAD 格式信息的程度。

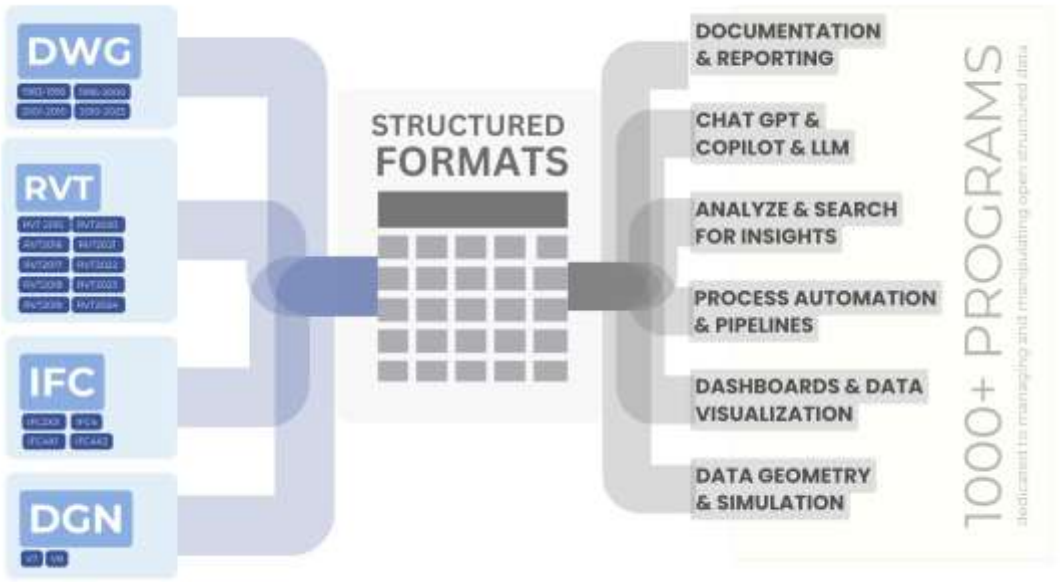


图 4.1-14 现代 SDK 工具 允许从专有 CAD 数据库格式合法转换数据 (BIM)。

目前，CAD 设计数据处理的趋势仍由主要市场参与者（CAD ）决定，这些供应商正在努力加强其在数据领域的地位，并创造新的格式和概念。

CAD 解决方案供应商正在向结构化数据方向发展

从 2024 年起，设计和建筑行业在数据使用和处理方面将发生重大技术变革。计算机辅助设计 - 系统供应商不再免费获取设计数据，而是专注于推广下一个新概念。BIM （创建于 2002 年）和开放式 BIM （创建于 2012 年）等方法正逐渐让位于 CAD 供应商开始推广的现代技术解决方案[93]：

- 逐步使用 "细粒度 "数据，以实现高效的信息管理和数据分析
- 美元 格式的出现和实体-组件-系统方法（ECS ）的实施，实现了灵活的数据组织
- 在数据处理、流程自动化和数据分析中积极使用人工智能
- 开发互操作性 - 改进不同计划、系统和数据库之间的互动

本书第六部分 "CAD 和 BIM ： 建筑设计数据的营销、现实和未来 "将更详细地讨论上述各个方面。在本章中，我们将简要概述变革的总体方向：主要的 CAD 供应商目前正在寻求重新思考设计信息的结构方式。其中一个关键的转变是放弃传统的基于文件的存储模式，转而采用细粒度的、以分析为导向的数据架构，从而提供对单个模型组件的连续访问[93]。

> 正在发生的事情的本质是，该行业正在逐步淘汰需要几何内核的繁琐、专业和参数化格式，转而采用更通用、机器可读和灵活的解决方案。

USD（通用场景描述）格式就是这种变革的驱动力之一，它最初是在计算机制图行业开发的，但由于英伟达公司开发了用于模拟和可视化的 Omniverse（和 Isaac Sim）平台，因此已经在工程应用中得到认可[93]。与参数化的 IFC 不同，USD 的结构更简单，可以用 JSON 格式描述几何体和对象属性 （图 4.1-15），从而方便了信息处理，加快了与数字流程的整合。新格式允许以 MESH 多边形的形式存储几何图形（除 BREP - NURBS 外--更多详情请参见本书第 6 部分 ），并以 JSON 格式存储对象属性，从而更便于自动化流程和云生态系统中的工作[94]。

一些 CAD 和 ERP 供应商已经使用了类似的格式（如 NWD、SVF 、CP2、 CPIXML） ，但其中大多数仍然是封闭的，不能供外部使用，这限制了数据集成和重用的可能性。在这种情况下，USD 可以扮演与 DXF 相同的角色，即 DWG 等专有格式的开放式替代格式。

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

图 4.1-15 美元 格式是 CAD 供应商为满足互操作性 和设计数据独立于几何内核 的需求而做出的尝试。

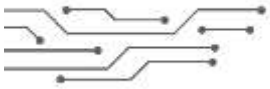
主要开发商向开放和简化的 USD 、GLTF 、OBJ 、XML（封闭的 NWD、CP2、SVF、SVF2、CPIXML） 和类似格式过渡（图 3.1-17），反映了全球趋势和行业对简化数据和提高可访问性的需求。在未来几年中，我们可以预见，依赖于几何内核 的复杂参数标准和格式将逐渐被淘汰，转而采用更轻便、更结构化的解决方案。这一转变将加速建筑行业的数字化进程，促进流程自动化并简化数据交换。

尽管计算机辅助设计 - 供应商制定了推广新开放格式的战略计划，但建筑行业的专业人员也可以通过使用逆

向工程工具，在不需要计算机辅助设计(BIM)工具的情况下，全面获取封闭式计算机辅助设计系统的数据。

> 所有这些趋势都不可避免地导致从笨重、单一的三维模型向通用、结构化数据转变，并使用在其他行业早已得到验证的格式。一旦项目团队开始将 CAD 模型视为不仅仅是视觉对象或一组文件，而是包含知识和信息的数据库，设计和管理方法就会发生巨大变化。

一旦团队学会了如何从文件、文本、图纸和 CAD 模型中提取结构化数据，并能够访问数据库，下一个关键步骤就是数据建模和质量保证。正是这一步在很大程度上决定了信息的处理和转换速度，而这些信息最终将用于在具体应用任务中做出决策。



章 4.2.

分类与整合：建筑数据的通用语言

决策速度取决于数据质量

当今的设计数据架构正在发生根本性的变化。行业正在从笨重、孤立的模型和封闭的格式向更加灵活、机器可读的结构转变，重点关注分析、集成和流程自动化。然而，仅仅向新格式过渡并不能保证效率，数据本身的质量必然是关注的焦点。

> 在本书中，我们谈了很多关于格式、系统和流程的内容。但是，如果没有一个关键要素：可信赖的数据，所有这些努力都毫无意义。数据质量是数字化的基石，我们将在下文中继续探讨。

现代建筑公司，尤其是大型建筑公司，使用数十个，有时甚至数千个不同的系统和数据库（图 4.2-1）。这些系统不仅必须定期填充新信息，还必须有效地相互交互。在处理输入信息时产生的所有新数据都会整合到这些环境中，并用于解决特定的业务任务。

过去，关于具体业务任务的决策是由高层管理人员--即所谓的 HiPPO（图 2.1-9）--根据经验和直觉做出的，而如今，随着信息量的急剧增加，这种方法正变得备受争议。使用实时数据的自动分析 正在取而代之。

> 传统的 "手工 "执行层面的业务流程讨论将转向运营分析 ，这就要求快速响应业务查询。

会计师、领班和估算师 手工生成报告、汇总表和项目数据展示，历时数天或数周的时代已经成为过去。如今，决策的速度和及时性已成为竞争优势的关键因素。



图 4.2-1 建筑业的计算和决策需要数天时间，不像其他行业只需数小时或数分钟。

建筑行业与数字化程度较高的行业（图 4.2-1）之间的主要区别在于数据质量和标准化程度较低。过时的信息生成、传输和处理方法会减慢流程并造成混乱。缺乏统一的数据质量标准 会阻碍端到端自动化的实施

> 主要挑战之一仍然是输入数据质量差，以及缺乏正规的数据准备和验证程序。没有可靠一致的数据，系统之间就无法有效整合。这将导致项目生命周期各阶段的延误、错误和成本增加。

在本书的以下章节中，我们将详细介绍如何提高数据质量、规范流程并缩短从信息到优质、有效和一致数据的过程。

数据标准化和集成

有效的数据管理需要明确的标准化战略。只有对数据结构和质量有明确的要求，才能在项目的各个阶段实现数据验证自动化，减少人工操作，加快知情决策。

在日常工作中，建筑公司每天需要处理数以百计的文件：电子邮件、PDF 文档、CAD 设计文件、 物联网传感器的数据，这些都需要集成到公司的业务流程中。

公司数据库和工具生态系统的森林（图 4.2-2）必须学会从输入的多格式数据中获取养分，以实现公司想要的结果。

> 要有效地处理数据流，不一定需要雇佣一大批管理人员，首先需要为数据制定严格的 要求和标准，并使用适当的工具来自动验证、协调和处理数据。

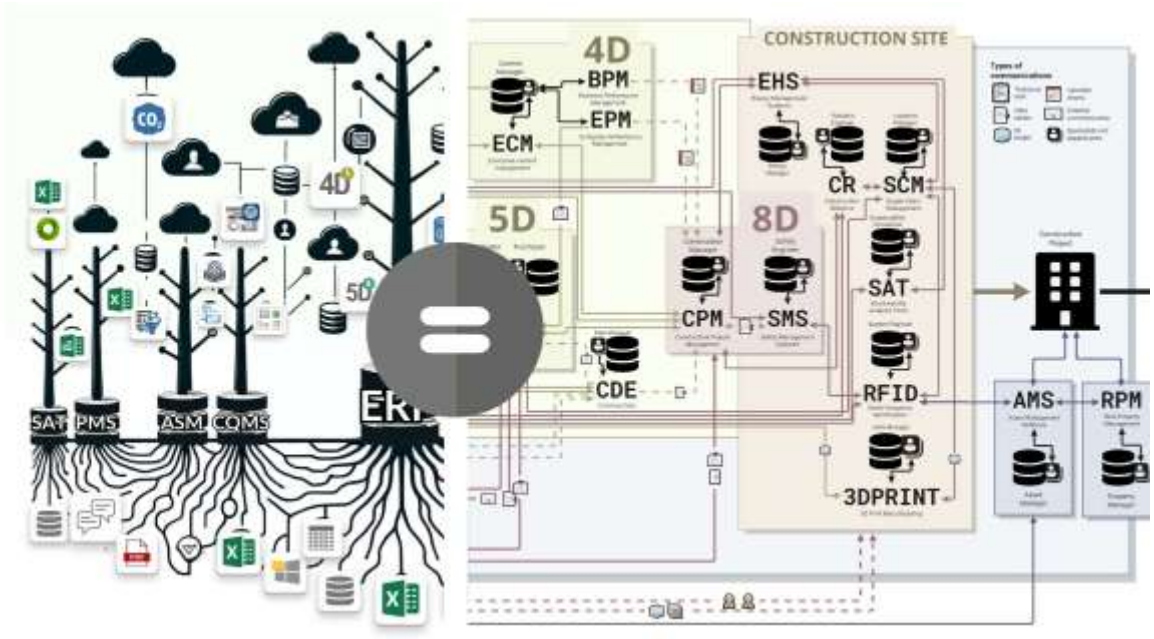


图 4.2-2 确保健康的公司生态系统需要高质量和及时的系统资源。

为了实现数据验证和统一过程的自动化（以便随后进行自动集成），应首先说明每个特定系统的最低必要数据要求。这些要求包括：

- 你到底需要什么?
- 采用何种形式（结构、格式）?
- 哪些属性是强制性的?
- 可接受的准确性和完整性公差是多少?

数据要求描述了接收和处理信息的质量、结构和完整性标准。例如，对于 PDF 文档中的文本，必须确保格式准确，符合行业标准（图 7.2-14 - 图 7.2-16）。CAD -模型中的对象必须有正确的属性（尺寸、代码、分

类器链接）（图 7.3-9，图 7.3-10）。对于合同扫描，明确的日期以及自动提取金额和关键条款的能力非常重要（图 4.1-7 - 图 4.1-10）。

> 制定数据要求并自动检查其合规性是最耗时但最关键的步骤之一。它是业务流程中最耗时的步骤。

正如本书第三部分所述，商业智能（BI）专业人员 50%到 90%的时间都花在了数据准备而非分析上（图 3.2-5）。这一过程包括数据收集、验证、确认、协调和结构化。

根据 2016 年的一项调查[95]，各行各业的数据科学家都表示，他们的大部分工作时间（约 80%）都花在了他们最不喜欢做的事情上（图 4.2-3）：收集现有数据集并对其进行组织（统一、结构化）。因此，只有不到 20% 的时间留给了创造性的工作，如寻找能带来新见解和新发现的模式和规律性。



图 4.2-3 验证 并确保数据质量 是准备将数据集成到其他系统的过程中成本最高、最耗时、最复杂的一步。

建筑公司要想成功地进行数据管理，就必须采取全面的方法，包括为任务设置参数、制定数据质量要求以及使用合适的工具进行自动验证。

数字互操作性始于需求

随着公司内部数字系统数量的增加，它们之间数据一致性的需求也在增加。负责不同 IT 系统的管理人员经常发现自己无法跟上日益增长的信息量和各种格式。在这种情况下，他们不得不请专家以适合其他应用程序和平台使用的形式创建数据。

这反过来又要求工程师和数据生成人员适应多种要求，而且往往不透明，也不清楚数据将来会在哪里以及如何使用。由于缺乏处理信息的标准化方法，导致验证阶段效率低下、成本增加，而由于数据的复杂性和非标准化性质，验证阶段往往需要人工操作。

> 数据标准化问题 不仅仅是一个方便或自动化的问题。它是一种直接的经济损失。根据 2016 年 IBM 的一份报告，美国每年因 数据质量差造成的损失高达 3.1 万亿美元[96]。此外，麻省理工学院和其他分析咨询公司的研究表明，糟糕的 数据质量造成的损失可高达公司收入的 15-25%[97]。

在这种情况下，明确定义数据要求和说明创建的对象中应包含哪些参数、采用何种格式和详细程度就变得至关重要。如果不将这些要求正规化，就不可能保证数据质量以及系统和项目阶段之间的兼容性（图 4.2-4）。

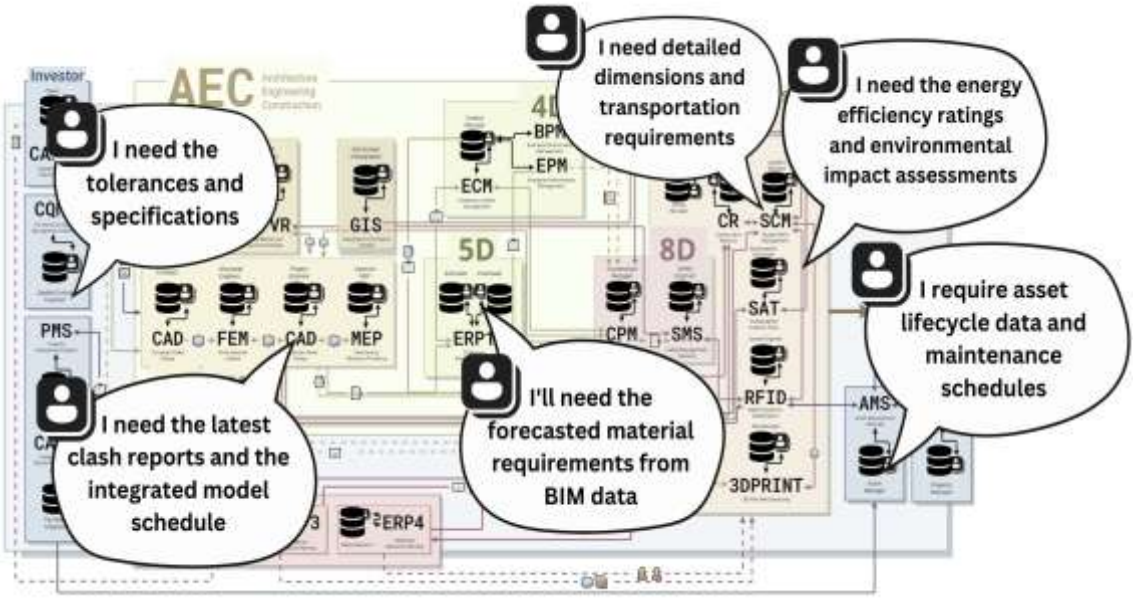


图 4.2-4 企业以不同角色的互动为基础，每个角色都需要某些对实现业务目标至关重要的参数和值。

为了制定正确的数据要求，您需要了解数据层面的业务流程。建筑项目的类型、规模和参与人数各不相同，而且每个系统，无论是建模（CAD (BIM)）、调度（ERP 4D）、成本计算（ERP 5D）还是物流（SCM），都需要各自独特的输入参数（输入实体-元素）。

根据这些需求，业务经理必须设计新的数据结构来满足要求，或者调整现有的表格和数据库。所创建数据的质量将直接取决于需求制定的准确性和正确性（图 4.2-5）。

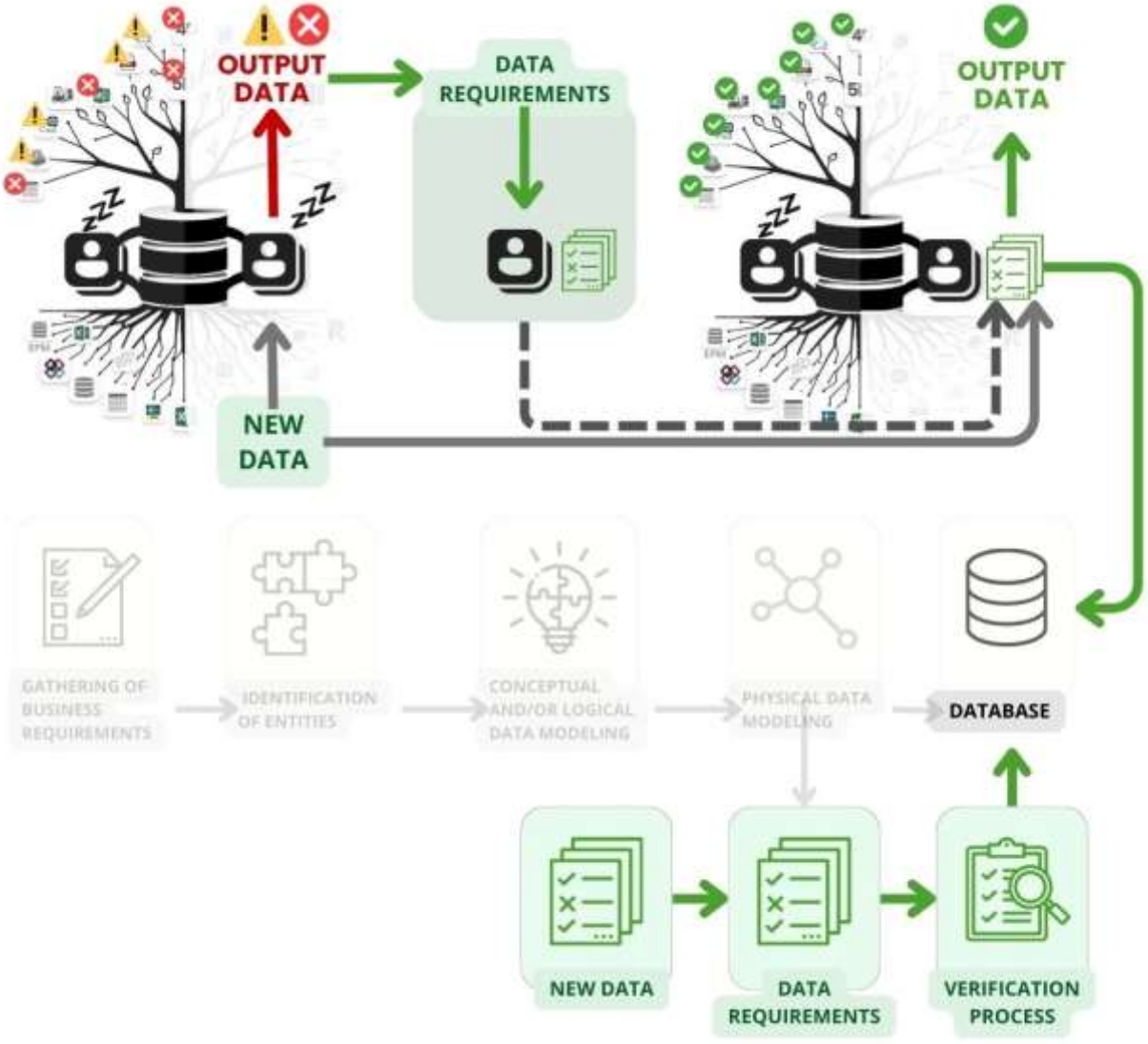


图 4.2-5 数据质量取决于为特定数据用例创建的需求的质量。

由于每个系统都有自己特定的数据要求，因此制定一般要求的第一步是对业务流程 中涉及的所有元素进行分类。这意味着必须将对象划分为与特定系统或应用任务相对应的类别和类别组。对于每个类组，都要制定单独的数据结构、属性和质量要求。

然而，在实践中，这种方法的实施面临着一个重大挑战：缺乏一种分组数据的通用语言。不同的分类、重复的标识符和不兼容的格式导致每家公司、每个软件甚至每个项目都形成了自己的、孤立的数据模型和类别。结果就形成了一个数字 "巴别塔"，在系统间传输信息需要多次转换到正确的数据模型和类，而且往往是手动完成。只有采用通用分类器和标准化需求集，才能克服这一障碍。

共同的建筑语言：分类器在数字化转型中的作用

在检测和处理过程数字化和自动化的背景下，分类系统 发挥着特殊的作用--一种 "数字字典"，确保了对象描述和参数化的统一性。分类器构成了一种 "通用语言"，可以根据意义对数据进行分组，并在项目生命周期的不同系统、管理级别和阶段之间对数据进行整合。

分级机最明显的影响在于建筑生命周期的经济性，其中最重要的方面是优化长期运营成本。研究表明，运营成本占建筑物总拥有成本的 80%，是初始建筑成本的三倍（图 4.2-6）[98]。这就意味着未来成本的决策主要是在设计阶段形成的

> 这就是为什么运营工程师的需求 （CAFM, AMS, PMS, RPM ）应成为设计阶段生成数据需求的起点（图 1.2-4）。这些系统不应被视为项目的最后阶段，而应是项目从概念到拆除的整个数字生态系统的组成部分

现代分类器 不仅仅是一个分组编码系统。它是建筑师、工程师、估算师、物流师、维护人员和 IT 系统之间相互理解的机制。就像汽车的自动驾驶仪必须准确无误地识别道路上的物体一样，数字建筑系统及其用户也必须通过要素分类，为不同系统准确无误地解释同一项目要素。

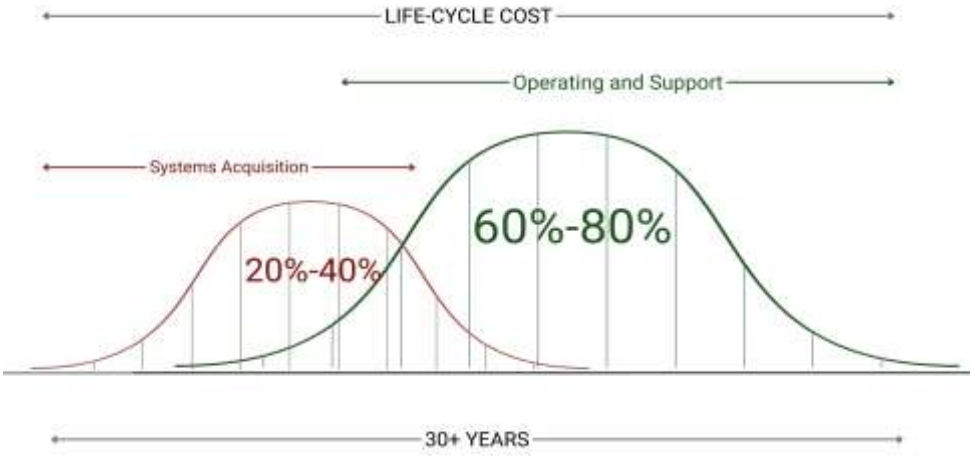


图 4.2-6 运行和维护成本是建筑成本的三倍，占建筑生命周期总成本的 60-80%（根据 [99]）。

分类器的发展水平直接关系到企业数字化的深度和数字化的成熟度。数字化成熟度低的企业面临着数据分散、信息系统不兼容的问题，因此分类器也不兼容且效率低下。在这些企业中，同一个元素在不同的系统中往往有不同的分组标识符，这严重阻碍了最终的整合，使流程自动化成为不可能。

例如，在 CAD 模型、估算和维护系统（图 4.2-7）中，一个项目中的同一个窗户可能会被贴上不同的标签，这是因为流程中不同的参与者对元素有多维的感知。对于窗户类元素的估算者来说，体积和成本很重要，对于维护服务来说，可用性和可维护性很重要，对于建筑师来说，美学和功能特性很重要。因此，同一要素可能需要不同的参数。

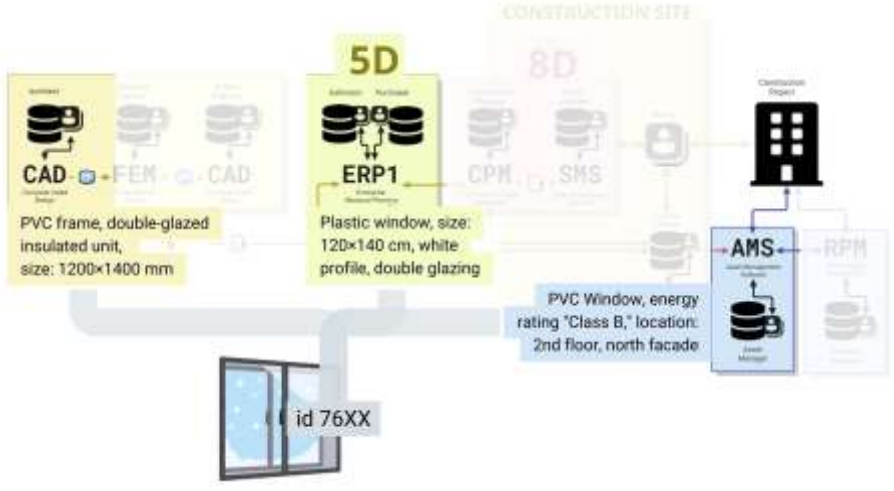


图 4.2-7 由于系统间的分类不一致，元素在过渡到另一个系统的每一步都会丢失一些属性信息。

由于难以明确界定建筑构件的分类，来自不同领域的专家往往会对同一构件划分不相容的类别。这就导致无法对物体形成统一的看法，这就需要随后进行人工干预，以协调不同的分类系统，并在不同专家定义的类型和类别之间建立一致性。

由于这种不一致性，采购部门（ERP）从制造商处购买建筑项目时收到的业务文件往往无法与建筑工地上的项目分类正确挂钩（PMIS ，SCM）。因此，关键信息很可能无法纳入基础设施和资产管理系统（CAFM、AMS），从而在试运行期间以及在随后的维护（AMS、RPM）或元件更换期间造成严重问题。

> 在数字化成熟度较高的企业中，分类器扮演着神经系统的角色，将所有信息流整合在一起。同一元素会收到一个唯一的标识符，使其可以在 CAD 、ERP 、AMS 和 CAFM 系统及其分类器之间传输，而不会失真或丢失。

要建立有效的分类器，需要了解数据的使用方式。同一个工程师在不同的项目中可能会对某个元素进行不同的命名和分类。只有收集多年来的使用统计数据，才能开发出稳定的分类系统。机器学习有助于实现这一点：算法分析成千上万的项目（图 9.1-10），通过机器学习确定可能的类别和参数（图 10.1-6）。在因数据量大而无法进行人工分类的环境中，自动分类尤其有价值。自动分类系统将能够根据最少填充的项目参数区分基本类别（更多详情将在本书第九和第十部分中介绍）。

已开发的分级机系统成为进一步数字化的催化剂，为分级机技术的发展奠定了基础：

- 自动估算项目成本和时间。
- 潜在风险和冲突的预测分析
- 优化采购流程和物流链
- 创建建筑物和结构的数字双胞胎
- 与智慧城市和物联网系统集成

> 转型的时间有限--随着机器学习 和计算机视觉技术的发展，几十年来无法解决的自动分类问题，，将在未来几年内得到解决。未能及时适应的建筑和设计公司有可能重蹈被数字平台取代的出租车队的命运。

本书的第五部分和第九部分将介绍更多有关成本计算和调度自动化以及大数据和 机器学习的内容。本书第十部分将详细讨论出租车队命运重演的风险以及建筑行业的优步化。

要了解分类器在建筑业数字化转型中的关键作用，就有必要回顾一下分类器的发展历史。正是历史背景让我们认识到分类方法是如何演变的，以及哪些趋势决定了分类方法的现状。

Masterformat, OmniClass, Uniclass 和 **CoClass**：分类系统的演变

从历史上看，建筑元素和工作分类器已经发展了三代，每一代都反映了特定时期的可用技术水平和当前的行业需求（图 4.2-8）：

- 第一代（20 世纪 50 年代初至 80 年代末）--纸质目录，本地使用的分层分类器（如 Masterformat 、SfB）。
- 第二代（20 世纪 90 年代末至 2010 年代中期）是在 Excel 和 Access 中实施的电子表格和结构化数据库（ASTM E 1557、OmniClass 、Uniclass 1997）。
- 第三代（2010 年代至今）--数字服务和应用程序接口 - 接口，与 CAD（BIM）集成 ，自动化（Uniclass 2015，CoClass）。

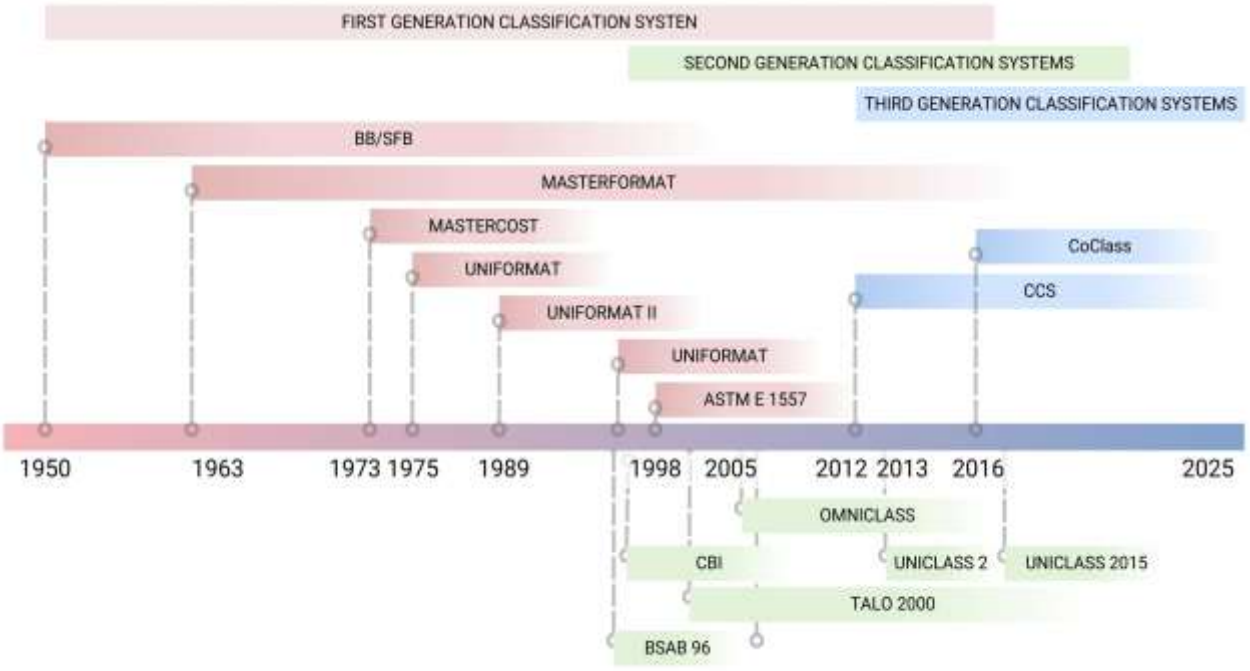


图 4.2-8 三代建筑行业分类器。

过去几十年来，分类器的分层复杂度不断降低（图 4.2-9）：早期的系统（如 OmniClass ）使用多达 7 层嵌套来描述 6887 个类别，而现代的解决方案（如 CoClass）则限制在 3 层，共 750 个类别。这使得在保持必要粒度的同时，更容易处理数据。Uniclass 2015 在英国经常被用作标准，它将 7210 个类合并为 4 个级别，方便了 CAD 项目和公共采购。

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

图 4.2-9 每一代新的分类器都会显著降低分类的复杂性。

在不同国家的建筑估算系统中，由于分类的不同，即使是一个典型的构件，如混凝土基础墙，也可以用非常不同的方式来描述（图 4.2-10）。这些差异反映了各国的施工实践、使用的测量系统、材料分类方法以及现行的法规和技术要求。

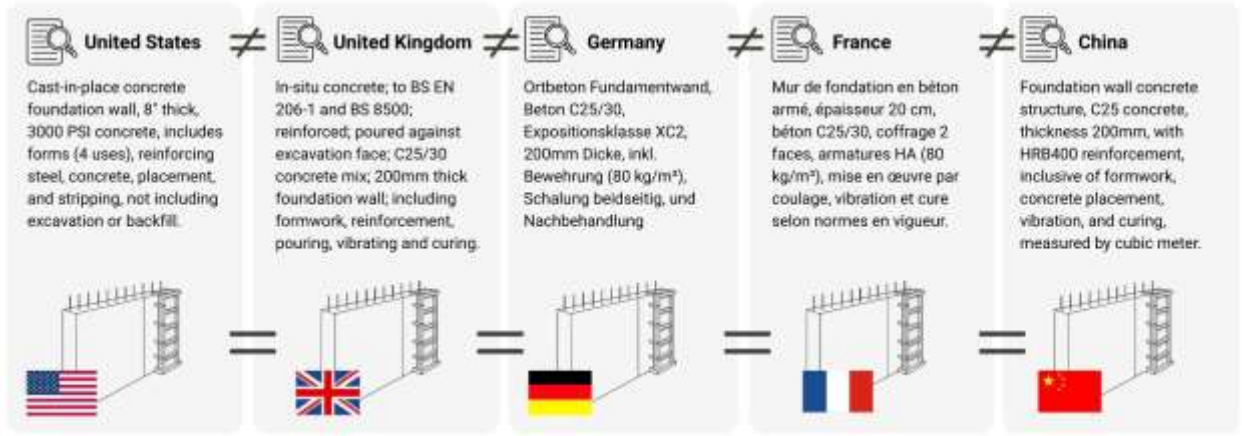


图 4.2-10 在不同国家的项目中，通过不同的描述和分类使用相同的元素。

对相同要素的不同分类使国际合作变得复杂，也使国际项目中的成本和工程范围的比较变得费时费力，有时甚至几乎不可能。目前，在全球范围内还没有一个通用的分类标准，每个国家或地区都根据当地的规范、语言和商业文化制定了自己的系统：

- **CCS**（丹麦）：成本分类系统--对 设施整个生命周期（设计、建造、运行）的成本进行分类的系统。重点是运行和维护逻辑，但也包括预算和资源管理。
- **NS 3451**（挪威）：按功能、设计要素和生命周期阶段对设施进行分类。用于项目管理、成本估算和长期规划。
- **MasterFormat**（美国）：一种将施工规范分为若干部分（如混凝土、电气、饰面）的系统。侧重于学科和工程类型，而不是功能要素（与 UniFormat 不同）。
- **Uniclass 2**（英国）：最详细的分类器之一，用于公共采购和 BIM 项目。将物体、工程、材料和空间的数据统一到一个系统中。
- **OmniClass**：国际标准（由美国 CSI 制定），用于管理从构件库到电子规格的对象信息。适用于长期数据存储，与 CAD (BIM) 和其他数字工具兼容。
- **COBie**：施工-运营建筑信息交换是设计、施工和运营阶段数据交换的国际标准。作为 "BIM - 可随时使用的模型 "概念的一部分，已纳入 BS 1192-4:2014。侧重于信息传输（如设备规格、保修、承包商联系方式）。

建筑业的全球化很可能导致建筑元素分类系统的逐步统一，这将大大减少对地方国家标准的依赖。这一过程可能会像互联网通信的发展一样，通用数据传输协议最终会取代不同的地方格式，确保全球系统的互操作性

。

另一条发展道路可能是直接过渡到基于机器学习技术的自动分类系统。这些技术目前主要是在自主运输领域开发的，在应用于大型 CAD 设计数据集方面具有巨大潜力（图 10.1-6）。

> 如今，这种情况已不仅仅局限于全国性的分级机集群。由于国家层面没有考虑到许多特殊性，每家公司都必须统一和规范其所使用的要素和资源类别。

通常情况下，这一过程从小规模开始--使用本地对象表或内部标签系统。然而，战略目标是转而使用一种通用语言来描述所有元素，这种语言不仅在公司内部可以理解，而且在公司外部也可以理解--理想情况下，与国际或行业分类标准相统一（图 4.2-8）。这种方法有利于与外部合作伙伴和数字系统集成，并促进在物品生命周期内形成统一的端到端流程。

在转向自动化和可扩展的 IT 系统之前，有必要使用国家级分类器或建立自己的、合乎逻辑的、无歧义的元素识别结构。每个对象，无论是窗户（图 4.2-11）、门还是 工程系统，都必须以这样一种方式加以描述，以便在任何公司的数字系统中都能明确无误地识别出来。这对于从平面图纸到数字模型的过渡至关重要，包括建筑物的设计阶段和运行阶段。

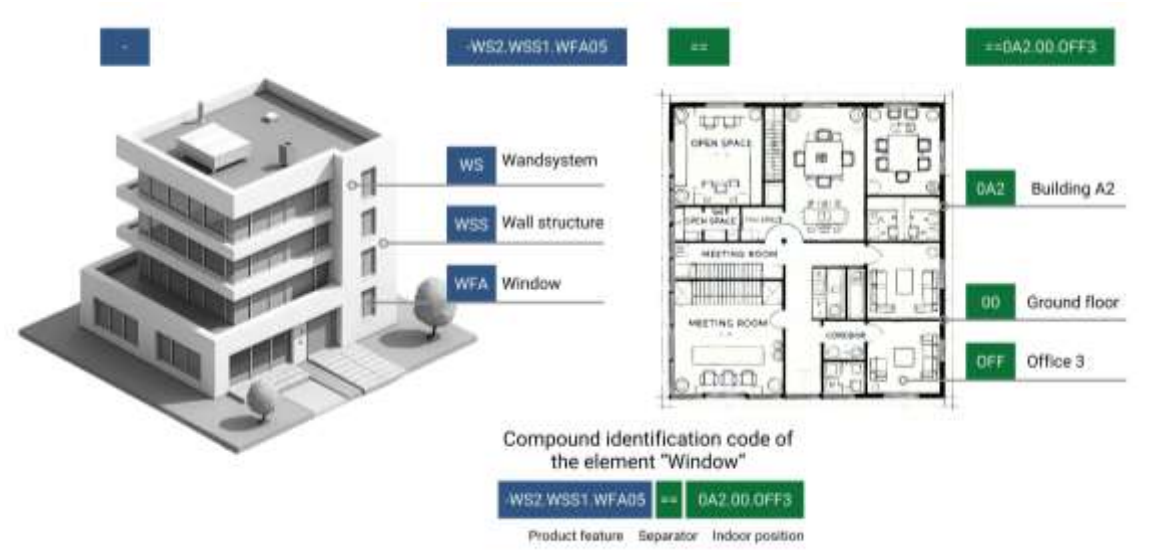


图 4.2-11 基于 分类和建筑物中位置的复合窗建筑元素标识符示例。

内部分类器的一个例子是开发一个复合识别代码（图 4.2-11）。这种代码结合了几个层次的信息：元素的

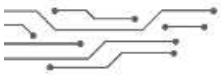
功能用途（如 "墙上的窗户"）、类型和确切的空间参照--A2 号楼，0 层，3 号房间。这种多层次结构可以创建一个统一的数字模型和文档导航系统，尤其是在数据验证和转换阶段，因为在这些阶段需要对元素进行明确的分组。明确的元素识别可确保各部门之间的一致性，减少重复、错误和信息丢失的风险。

一个完善的分类器 不仅仅是一份技术文档，它还是公司数字生态系统的基础：

- 确保系统之间的数据兼容性；
- 降低搜索和处理信息的成本；
- 提高透明度和可管理性；
- 为扩展和自动化奠定了基础。

通过使用国家分类器或自己的综合识别码对物体进行标准化描述，可成为数据一致性、 可靠的信息交换以及随后实施智能服务（从自动采购到数字双胞胎）的基础。

在完成多格式数据的结构化阶段并选择用于识别和分组元素的分类器后，下一步就是对数据进行正确建模。这一过程包括确定关键参数、建立逻辑数据结构和描述元素之间的关系。



章 4.3.

数据建模和示范中心

数据建模：概念模型、逻辑模型和物理模型

如果没有经过深思熟虑的存储和处理结构，就不可能对数据（由我们在前面进行结构化和分类）进行有效管理。为了确保信息在存储和处理阶段的可访问性和一致性，企业采用数据建模的方法，根据业务需求设计表格、数据库和它们之间的链接。

> 数据建模是建立任何数字生态系统的基础。如果没有对系统、需求和数据建模的描述，创建数据的工程师和专业人员就不知道或不了解他们创建的数据将用于何处。

创建数据仓库系统就像盖楼，没有计划就无法开始铺砖一样，需要清楚地了解将使用哪些数据、如何连接数据以及谁将使用这些数据。如果没有对流程和要求的描述，创建数据的工程师和专业人员就会忽略数据将来会在哪里以及如何使用。

数据模型是业务与 IT 之间的桥梁。数据模型可以使需求正规化 ，使信息结构化，并促进利益相关者之间的沟通。从这个意义上说，数据建模类似于建筑师的工作，建筑师根据客户的规划制定建筑计划，然后将其交给建设者--数据库管理员和开发人员--实施（创建数据库）。

因此，每家建筑公司除了对要素和资源进行结构化和分类之外（图 4.2-11），还必须掌握 "构建 "数据库（表格）的艺术，并学会如何在它们之间建立联系，就像将公司数据中的砖块连接成一堵可靠而坚固的知识墙一样。数据建模的关键概念（图 4.3-1）包括

- 实体是必须收集数据的对象。在早期设计阶段，实体可以是一个单独的元素（如 "门"），而在估算模型中，实体可以是一组分类元素（如 "内门"）。
- 属性是实体的特征，描述了重要的细节：尺寸、属性、组装成本、物流和其他参数。
- 关系（链接）--显示实体之间如何相互作用。它们可以是以下类型之一"一对一"、"多对一"、"多对多"。

- **ER** 图（实体-关系图）是显示 实体、属性和它们之间关系的可视化图表。ER 图可以是概念图、逻辑图和物理图--每种图都反映了不同的详细程度。

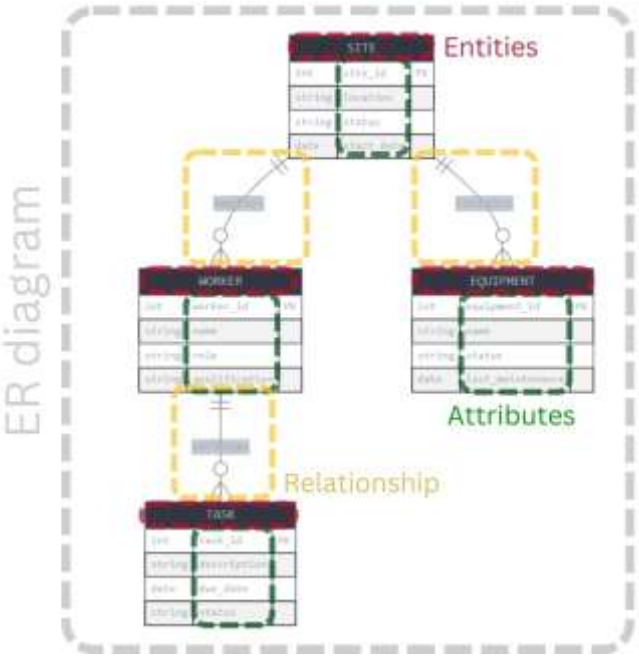


图 4.3-1 包含实体、属性和关系的概念数据库结构 ER 图。

设计数据和定义数据间关系的过程传统上分为三种主要模式。每种模型都能实现一定的功能，但在表示数据结构的详细程度和抽象程度上有所不同：

- 概念数据模型 ： 该模型描述主要实体及其关系，但不涉及属性细节。它通常用于规划的初始阶段。在这一阶段，我们可以从数据库和系统中勾勒出不同部门和专家之间的关系。

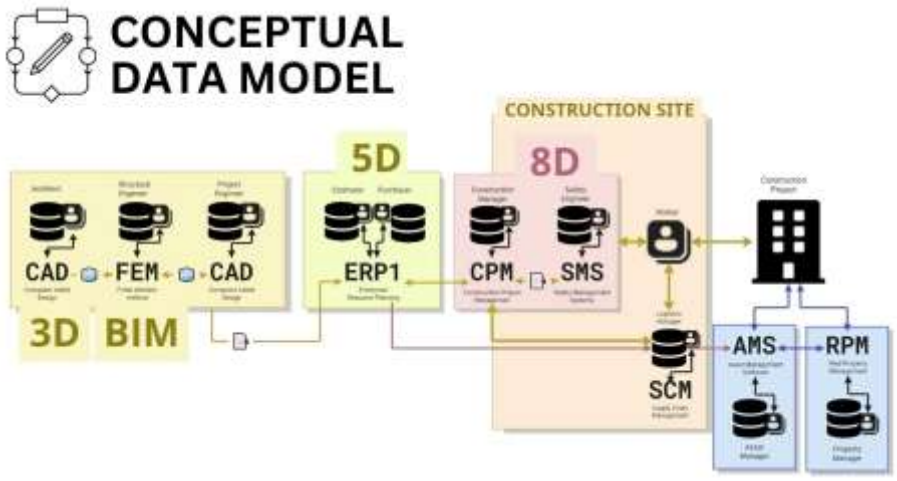


图 4.3-2 概念图描述了系统的内容：高层次的关系表述，没有技术细节。

- 逻辑数据模型 ： 逻辑数据模型以概念模型为基础，包括实体、属性、键和关系的详细描述，以及业务信息和规则的映射。

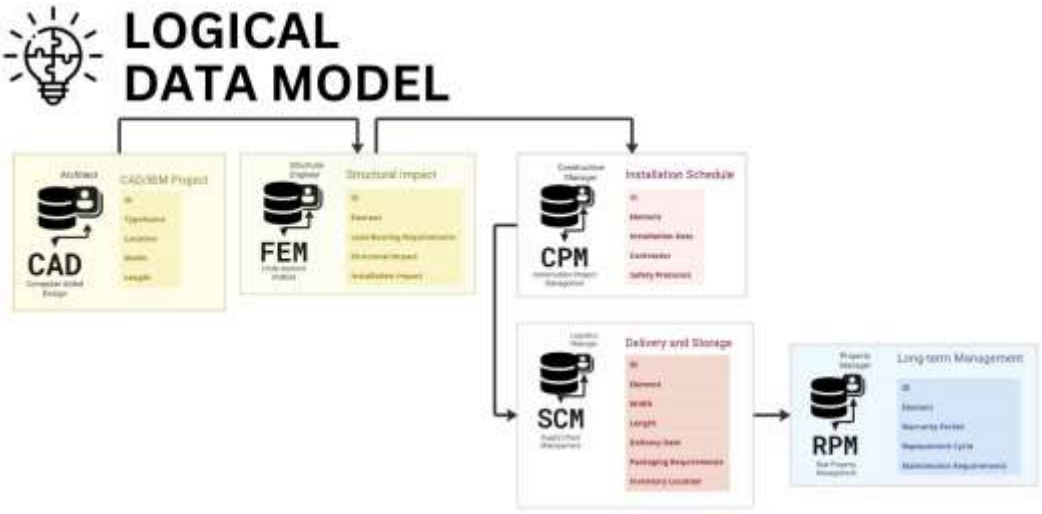


图 4.3-3 逻辑数据模型 详细说明了数据类型、关系和键，但没有系统实现。

- 物理数据模型 ： 该模型描述了实施数据库的必要结构，包括表、列和关系。它侧重于数据库性能、索引策略和物理存储，以优化数据库的物理部署。

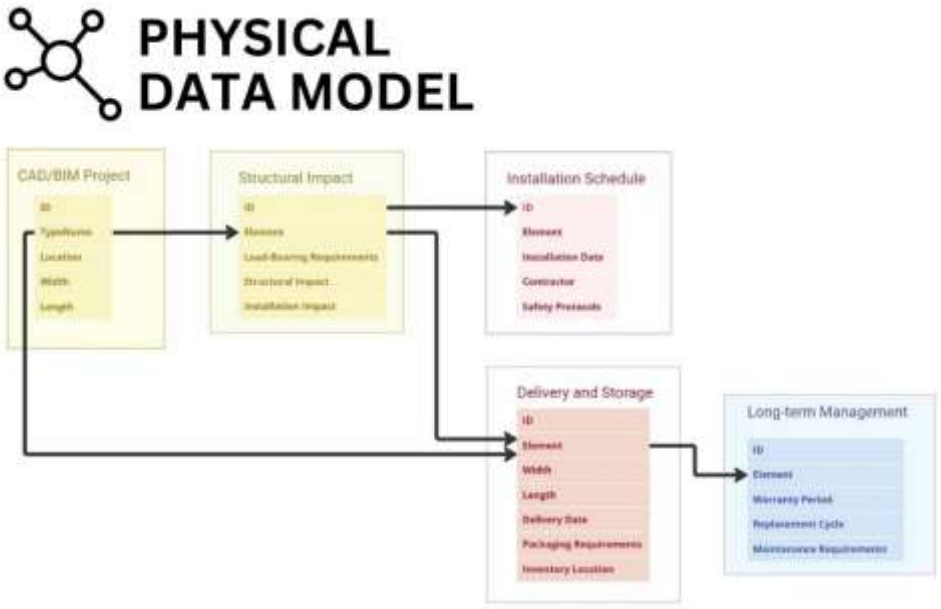


图 4.3-4 物理数据模型 定义了系统的实现方式，包括数据库的表和具体细节。

> 在设计数据库和设计表格关系时，了解抽象层次对建立有效的系统架构起着关键作用。

有效的数据建模方法可以将业务目标与技术实施相结合，使整个流程链更加透明和易于管理。数据建模不是一次性任务，而是一个涉及连续步骤的过程（图 4.3-5）：

- 收集业务需求：确定关键任务、目标和信息流。这是与专家和用户积极互动的阶段。
- 实体识别：强调未来系统中需要考虑的主要对象、类别和数据类型。
- 建立概念和逻辑模型：首先确定关键实体及其关系，然后确定属性、规则和详细结构。
- 物理建模：设计模型的技术实现：表、字段、关系、约束、索引。
- 创建数据库：最后一步是在选定的数据库管理系统中实施物理模型，进行测试并准备运行。

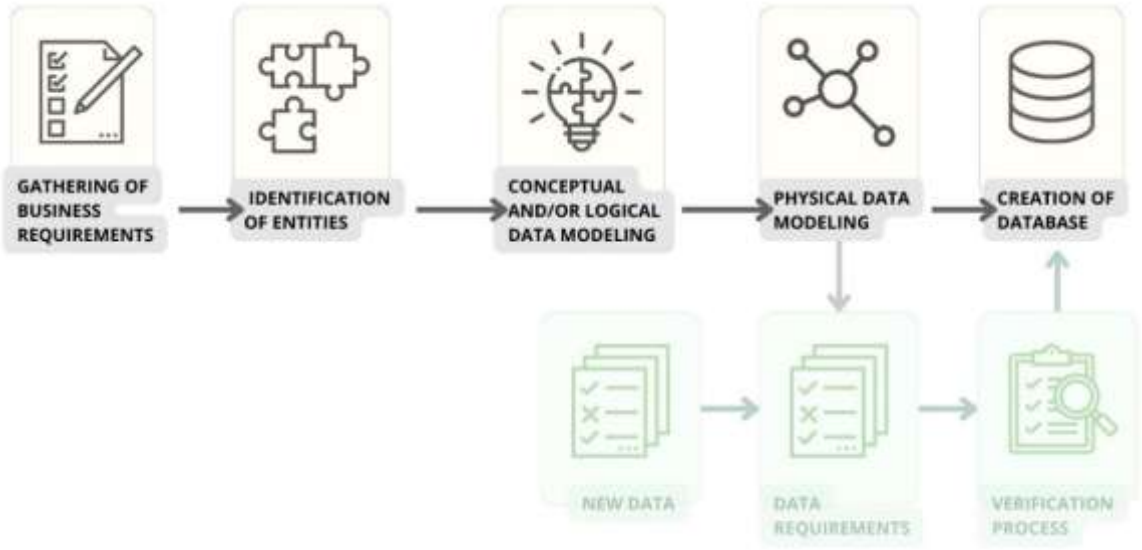


图 4.3-5 为业务流程创建数据库和数据管理系统始于需求生成和数据建模。

设计合理的数据建模流程可以实现透明的信息流，这对于建筑项目或现场管理等复杂项目尤为重要。让我们来看看从概念模型到逻辑模型，再到物理模型如何帮助简化流程。

建筑工程中的实用数据建模

让我们以建筑工地管理任务为例进行数据建模，将 领班的需求转换为结构化逻辑模型。根据工地管理的基本需求，我们定义了以下关键实体：工地 (SITE)、工人 (WORKER)、设备 (EQUIPMENT)、任务 (TASK) 和设备使用情况 (EQUIPMENT_USAGE)。每个实体都包含一组反映重要特征的属性。例如，对于任务（TASK）来说，属性可以是任务描述、到期日期、状态、优先级；对于工人来说，属性可以是姓名、在现场的角色、当前就业情况等。

逻辑模型建立了这些实体之间的关系，显示了它们在实际工作流程中如何相互影响（图 4.3-6）。例如，场地和工人之间的关系表明，许多工人可以在一个场地工作，而工人和任务之间的关系则表明，一个工人可以执行多项任务。

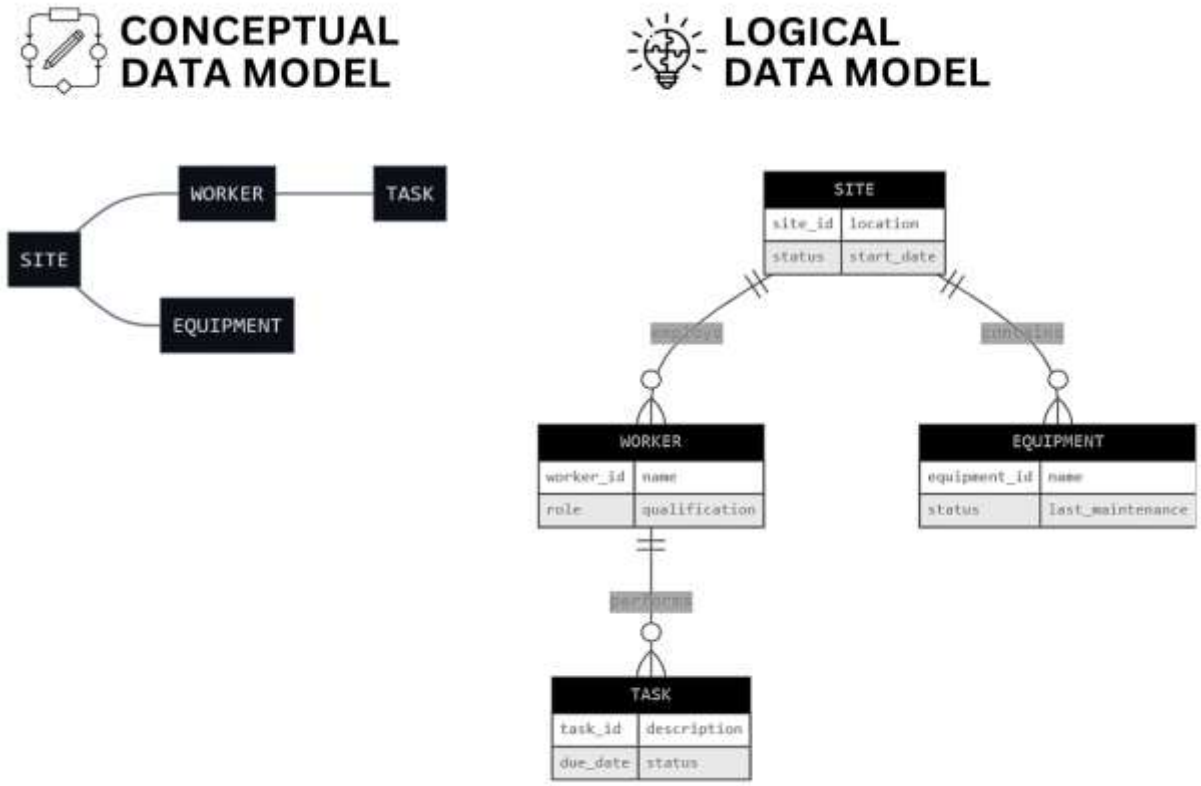


图 4.3-6 根据工头的要求生成的用于描述施工现场流程的概念和逻辑数据模型。

在转入物理模型时，需要添加技术实现细节：特定数据类型（VARCHAR、INT、DATE）、表间关系的主键和外键，以及优化数据库性能的索引（图 4.3-7）。

例如，应为状态定义具有可能值的特定类型，并为状态和 Worker_id 等关键字段添加索引，以提高搜索性能。这样，系统的逻辑描述就变成了具体的数据库实施计划，可以随时创建和实施。

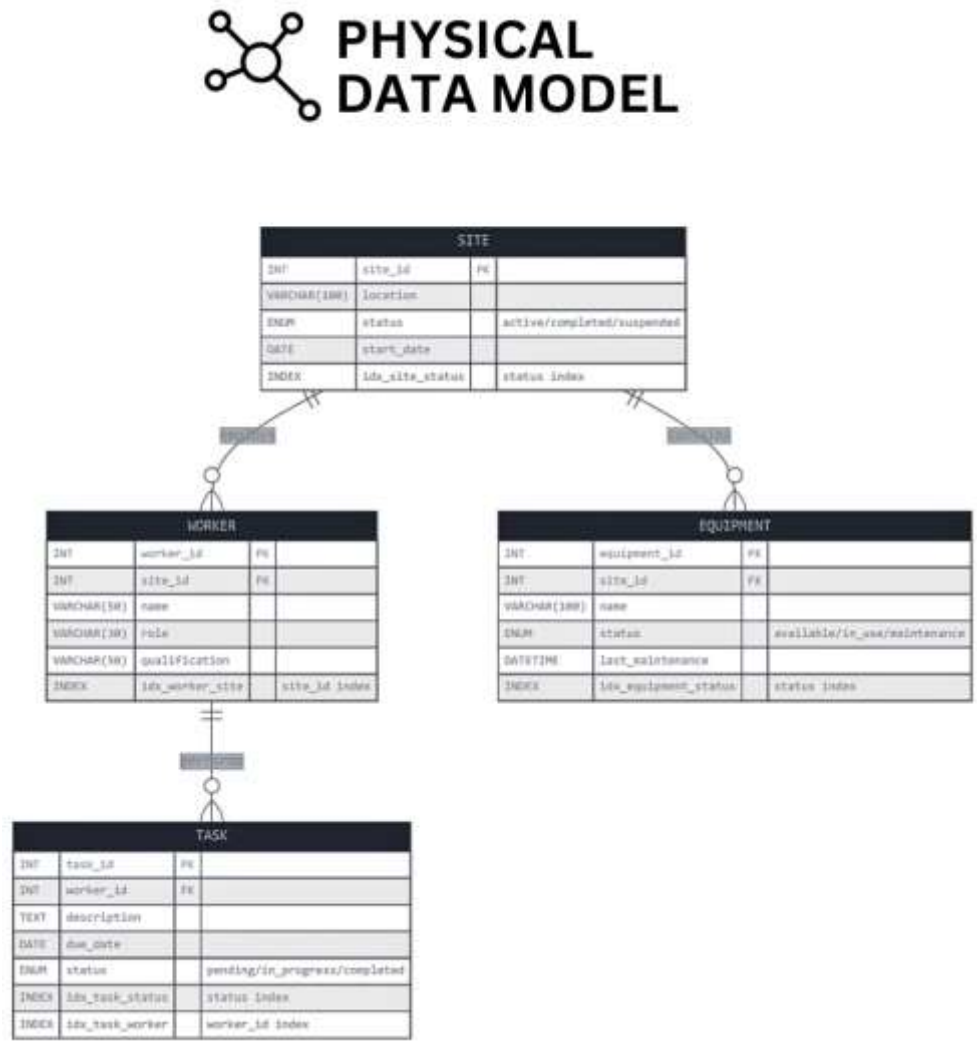


图 4.3-7 物理数据模型 通过最低要求的参数来描述建筑工地的实体。

> 物理模型通常不同于逻辑模型。建模时间的平均分配如下：约 50%用于概念模型（收集需求、讨论流程、确定实体），10%用于逻辑模型（指定属性和关系），40%用于物理模型（实施、测试、适应 DBMS）。

造成这种平衡的原因是，概念阶段为数据结构奠定了基础，而逻辑模型只规定了关系和属性。物理模型需要的资源最多，因为正是在这一阶段，数据要被实施到特定的平台和工具中

使用 **LLM** 创建数据库

有了数据模型和通过参数对实体的描述，我们就可以创建数据库--存储库，在这里我们将存储在特定流程的结构化阶段之后获得的信息。

让我们以 Python 编程语言 为例，尝试使用 SQLite 以最少的代码创建一个简单但功能强大的数据库。关系数据库 已在 "结构化关系数据库和 SQL 查询语言 "一章中详细讨论过。

> SQLite 是一个轻量级、开源、可嵌入、跨平台的数据库 ，支持核心 SQL 功能。SQLite 广泛应用于 CAD (BIM)、ERP 、EPM ，几乎存在于所有智能手机中。

让我们要求 LLM 模型根据上述创建的物理模型（图 4.3-4），为 FEM（结构影响）和 CPM （装配进度）系统创建一个小型 SQLlite 数据库。

- 向 LLM 聊天（ChatGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN 或其他）发送文本请求：

  > 根据附件中的图为 "建设性影响 "和 "安装时间表 "系统创建 SQLlite 数据库（将模式作为图片粘贴到聊天中（图 4.3-4）。 ⏎

LLM 的答复：

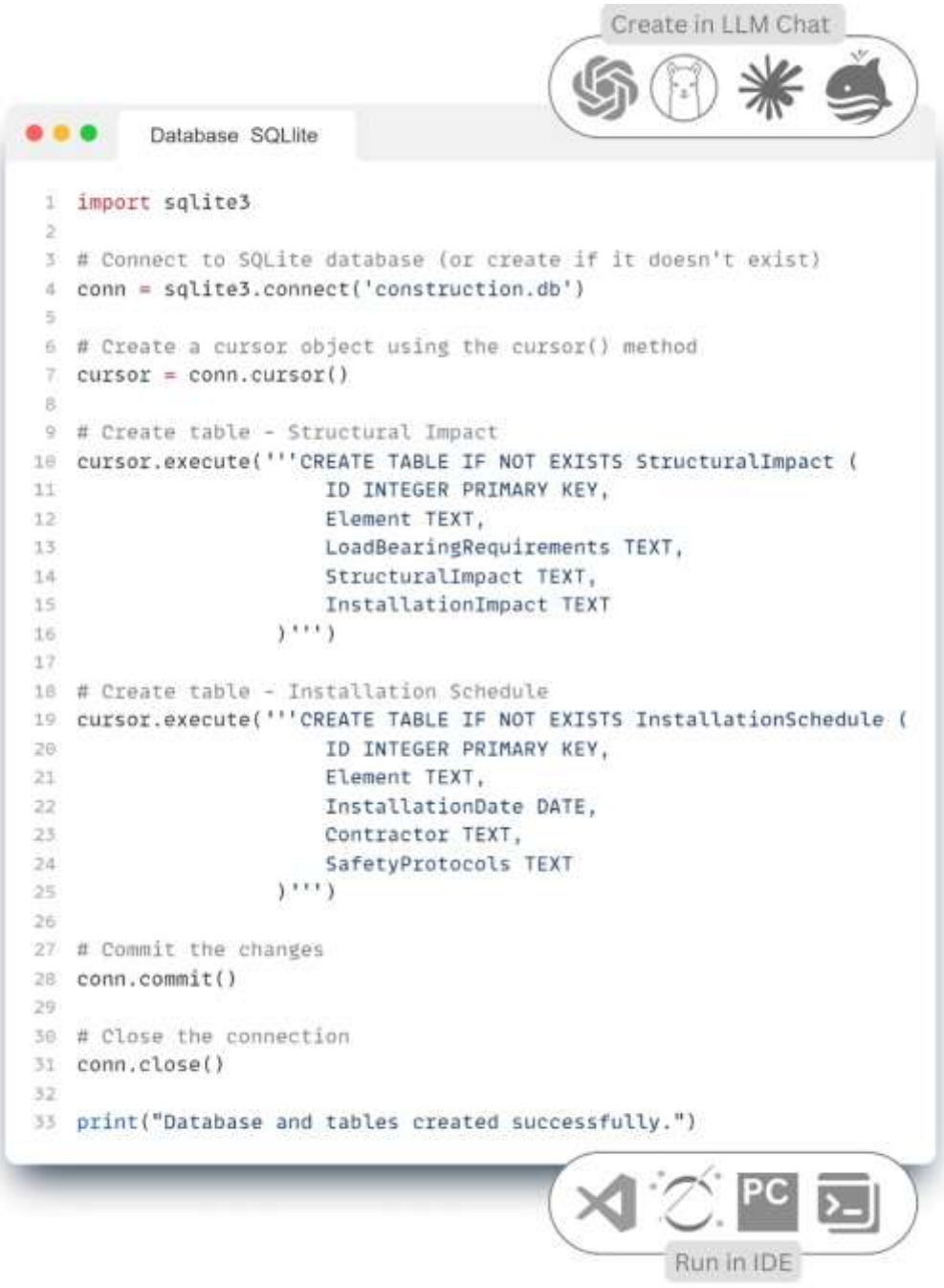


图 4.3-8 生成的代码会根据上传的图像创建数据库和表格，我们现在可以将数据加载到这些数据库和表格中。

该数据库代码可在支持 SQLite 的 Python 环境 中运行 ，该环境通常包含在标准 Python 环境中。它可以在任何能够运行脚本和创建文件的本地 Python 机器上运行。运行并执行该代码后，将在硬盘上创建一个文件--一个名为 construction.db 的 SQLite 数据库 （类似于带有工作表的 Excel 文件 ），其中包含 "结构影响 "和 "安装进度 "表。

一旦在这些表中创建了新的 **construction.db** 数据库，就可以通过 SQL _查询或导入将数据添加到这些表中，这样就可以在以后开始创建自动数据处理。数据可以从 CSV 文件 、Excel 电子表格导入 SQLite 数据库 ，也可以通过 API 从其他数据库和资料库导出。

要建立可持续的数据建模和有效的数据库管理流程，公司需要制定明确的战略，并在技术和业务团队之间进行协调。面对不同的项目和多个数据源，通常很难确保各个层面的一致性、标准化和质量控制。一个关键的解决方案可能是在公司内部创建一个专门的数据建模卓越中心（COE）。

数据建模卓越中心（**CoE**）

随着数据成为重要的战略资产之一，企业需要做的不仅仅是正确地收集和存储信息，更重要的是学会如何系统地管理数据。分类 和数据建模卓越中心（CoE ）是一个结构单位，负责确保组织内所有数据处理的一致性、质量和效率。

> 卓越中心（CoE）是专家支持的核心，也是企业数字化转型的方法论基础。它建立了一种数据驱动型文化，使企业能够建立基于结构化、经过验证和具有代表性的数据而非直觉或本地信息做出决策的流程。

卓越数据中心通常由跨职能团队组成，这些团队按照 "两个披萨 "原则开展工作。该原则由杰夫-贝索斯提出，意思是团队的规模应能满足两个披萨的需求，即不超过 6-10 人。这种方法有助于避免过度的官僚主义，提高工作的灵活性。CoE 团队应包括具备各种技术技能的员工，从数据分析 和机器学习 到特定业务领域的专业知识。数据工程师应具备深厚的技术知识，不仅能优化流程和建立数据模型，还能为同事提供支持，减少日常工作的时间（图 4.3-9）。

> 在自然界，生态系统的复原力是通过生物多样性来保证的，在数字世界，灵活性和适应性也是通过处理数据的方法多样性来实现的。然而，这种多样性必须以共同的规则和概念为基础。

卓越中心（CoE）可以比作森林生态系统的 "气候条件"，它决定了哪些类型的数据会蓬勃发展，哪些会被自动丢弃。通过为高质量数据创造有利的 "气候条件"，卓越中心促进了最佳实践和方法的自然选择，这些最佳实践和方法随后成为组织的标准。

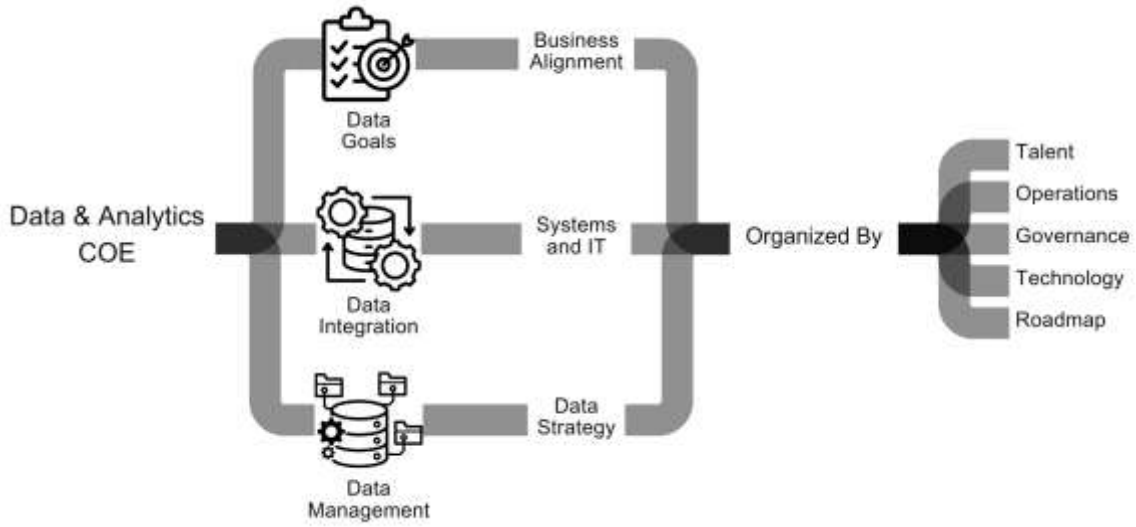


图 4.3-9 数据和分析卓越中心（CoE）汇集了数据管理、整合和战略等关键方面的专业知识。

为加快集成周期并取得更好的成果，CoE 应为其成员提供足够的决策自主权。这一点在动态环境中尤为重要，因为在这种环境中，不断试错、持续反馈和频繁发布可带来显著效益。然而，这种自主权只有在高层管理人员明确沟通和支持的情况下才会有效。如果没有战略眼光和高层协调，即使是最有能力的团队也会在实施计划时遇到障碍。

> 公司的首席运营官或高级管理层有责任确保数据建模方法不局限于一两个项目，而是要将其纳入整个信息管理和业务流程管理系统。

专业技能中心（CoE）除了与数据建模和数据管理有关的任务外，还负责为数据基础设施的部署和运行制定通用标准和方法。此外，它还在组织内培养一种持续改进、优化流程和有效使用数据的文化（图 4.3-10）。

CoE 内数据和模型管理的系统方法可大致分为几个关键部分：

- 流程和模型生命周期管理的标准化：CoE 制定并实施了统一创建和管理数据模型的方法。这包括：建立结构模板、质量控制方法和版本控制系统，以确保所有工作阶段的数据连续性。
- 角色管理和责任分配：特遣队所属装备定义了数据建模过程中的关键角色。为每个项目参与者分配明确界定的角色和责任范围，以促进团队合作并降低数据不一致的风险。
- 质量控制和审计：有效管理建筑数据需要对其质量进行持续监测。目前正在实施检查数据、识别错误和缺失属性的自动化机制。
- 元数据和信息结构管理：CoE 负责创建一个统一的分类系统 和标识符、命名和实体描述标准，这对系统之间的整合至关重要。

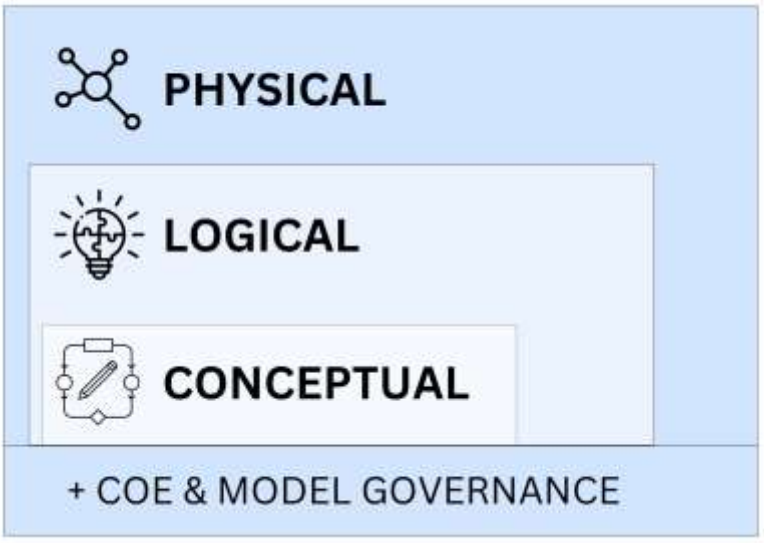


图 4.3-10 数据建模和数据质量管理是 CoE 的主要挑战之一

数据卓越中心（CoE ）不仅仅是一个专家小组，而是一个系统机制，它创建了一种新的数据驱动的 文化，并确保整个公司采用统一的方法处理数据。通过将建模流程有效整合到整体信息管理系统、标准化、分类和数据质量控制 ， CoE 帮助企业不断改进产品和业务流程，更快地应对市场变化，并根据可靠的分析结果做出明智的决策。

当与现代 DataOps 原则相结合时，此类中心尤其有效 - 根据 ，此举可确保数据的持续交付、自动化和质量控制。我们将在第 8 部分 "建筑行业的现代数据技术 "一章中详细介绍 DataOps。

在接下来的章节中，我们将从战略转向实践--让我们有条件地 "转型 "为数据中心：我们将举例说明任务参数

化、需求收集 和自动验证过程是如何进行的。



章 4.4.

要求系统化和信息验证

需求收集和分析： 将通信转换为结构化数据

收集和管理需求是确保数据质量的第一步。尽管开发了数字化工具，但大多数需求仍是通过非结构化方式制定的：信件、会议记录、电话和口头讨论。这种沟通方式难以实现信息的自动化、验证和重用。在本章中，我们将探讨如何将文本需求转化为正式结构，确保业务需求的透明性和系统性。

> Gartner 的研究 "数据质量：准确洞察的最佳实践 "强调了数据质量 对于成功的数据和分析计划至关重要 [100]。他们指出，糟糕的数据质量平均每年给企业造成至少 1290 万美元的损失，而可靠、高质量的数据对于创建数据驱动型公司至关重要。

由于缺乏结构化的要求，同一元素（实体）及其参数可能以不同的形式存储在不同的系统中。这不仅会降低流程效率，还会浪费时间、重复信息，并需要在使用数据前重新验证。因此，哪怕是一个遗漏--一个丢失的参数或一个描述错误的元素--都会导致决策速度减慢和资源使用效率低下。

> 因为缺少钉子，马蹄铁丢失了。
>
> 没有马蹄铁，就没有马。
>
> 因为没有马，失去了骑手。
>
> 没有骑手，就失去了信息。
>
> 没有信息，就失去了战斗。
>
> 没有战斗，就失去了王国。
>
> 这一切都是因为马蹄铁上缺少一枚钉子。
>
> - 谚语 [101］

分析和收集数据填充和存储过程的要求，首先要确定所有利益相关者。正如谚语所说的失去一颗钉子会导致一连串严重后果一样，在业务中，失去一个利益相关者、忽略一个需求或丢失哪怕是一个参数，不仅会严重

影响单个业务流程，还会影响项目的整个生态系统和整个组织。因此，即使是那些乍看之下无关紧要的要素、参数和角色，，但日后可能会被证明对企业的可持续发展至关重要。

假设某公司有一个项目，客户提出了一项新要求--"在大楼北侧增加一扇窗户"。在 "客户要求在当前项目中增加一个新窗口 "这个小流程中，涉及到建筑师、客户、CAD 专家 (BIM)、施工经理、物流经理、ERP - 分析师、质量控制工程师、安全工程师、控制经理和物业经理。

> 即使是一个小流程，也可能涉及几十个不同的专家。每个流程参与者都必须了解 专家的要求，因为他们在数据层面上与这些专家有联系。

在文本层面（图 4.4-1），客户与流程链中的专家之间的沟通如下：

- 客户："我们决定在北侧增加一扇窗户，以获得更好的采光。能实现吗？
- 建筑师: "当然，我会修改项目，将新窗户包括在内，并发送更新的 CAD 图纸 (BIM)"。
- **CAD** 专家 （**BIM** ）："接到一个新项目。我更新了 CAD（BIM）模型，增加了窗户，并在与 FEM 工程师协调后，提供了新窗户的确切位置和尺寸"。
- 施工经理："收到一个新项目。我们正在调整 4D 的安装日期，并通知所有相关分包商。"
- 设施工程师 **(CAFM)**："我将把新窗口上的 6D 数据输入 CAFM 系统，用于未来的设施管理和维护规划"。
- 物流经理："我需要新窗户的尺寸和重量，以便组织将窗户运送到工地"。
- **ERP** -analyst: "I need the scope tables and exact window type for the 5D budget update in our ERP system to reflect the cost of the new window in the overall project estimate."
- 质量控制工程师："一旦窗户规格准备就绪，我将确保它们符合我们的质量和材料标准"。
- 安全工程师："我将对新窗口的安全方面进行评估，重点是 8D 计划下的合规性和疏散问题 "。
- 控制经理："根据 ERP 的确切工作范围，我们将更新 4D 时间线 ，以反映新窗口的安装情况，并将新数据存储在项目内容管理系统中"。
- 工人（安装人员）："您需要关于安装、装配和工作时间安排的说明。此外，还有什么特别的安全规定需要我遵守吗？
- 物业经理："一旦安装完毕，我将记录保修和维护信息，以便进行长期管理"。
- 资产经理"设备工程师，请发送资产跟踪和生命周期管理的最终数据。

客户："等等，也许我赶时间，不需要窗户。也许我应该做一个阳台"。

在这种经常发生的情况下，即使是很小的变化也会在多个系统和角色之间引起连锁反应。在这种情况下，最初 阶段几乎所有的通信都是文本形式的：电子邮件、聊天记录、会议记录（图 4.4-1）。

在这种基于文本的建筑项目通信系统中，对所有数据交换操作和所有决定进行法律确认和记录的系统非常重要。这是为了确保所做的每项决定、指令或更改都具有法律效力并可追溯，从而降低将来发生 "误解 "的风险

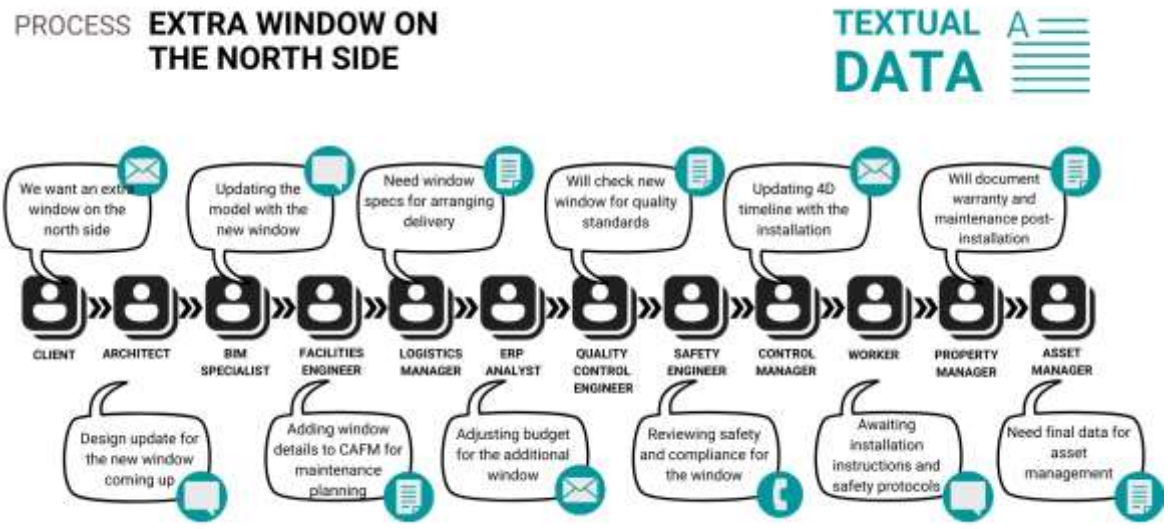


图 4.4-1 在项目初始阶段，客户与承包商之间的通信通常包含多种格式的文本数据。

> 在建筑项目的相关系统中，如果缺乏对决策的法律控制和确认，就会给所有相关人员带来严重问题。在没有适当文件记录和验证的情况下做出的每项决定、命令或变更都可能导致纠纷（和诉讼）。

只有通过大量的签名文件才能确保所有决定在文本交流中得到法律上的巩固，而这就落 在了有义务记录所有交易的管理层的肩上。因此，如果每个参与者都必须为每项行动签署文件，系统就会失去灵活性，成为官僚主义的迷宫。缺乏交易确认不仅会延误项目的实施，还可能导致财务损失和参与方之间关系的恶化， 甚至引发法律问题。

这种交易审批流程通常以文本讨论开始，在接下来的阶段逐渐演变成多格式文件交换（图 4.4-2），使过去

仅通过文本进行的交流变得更加复杂。如果没有明确界定的要求，要将这些充满多格式数据和大量文本要求的流程自动化几乎是不可能的。

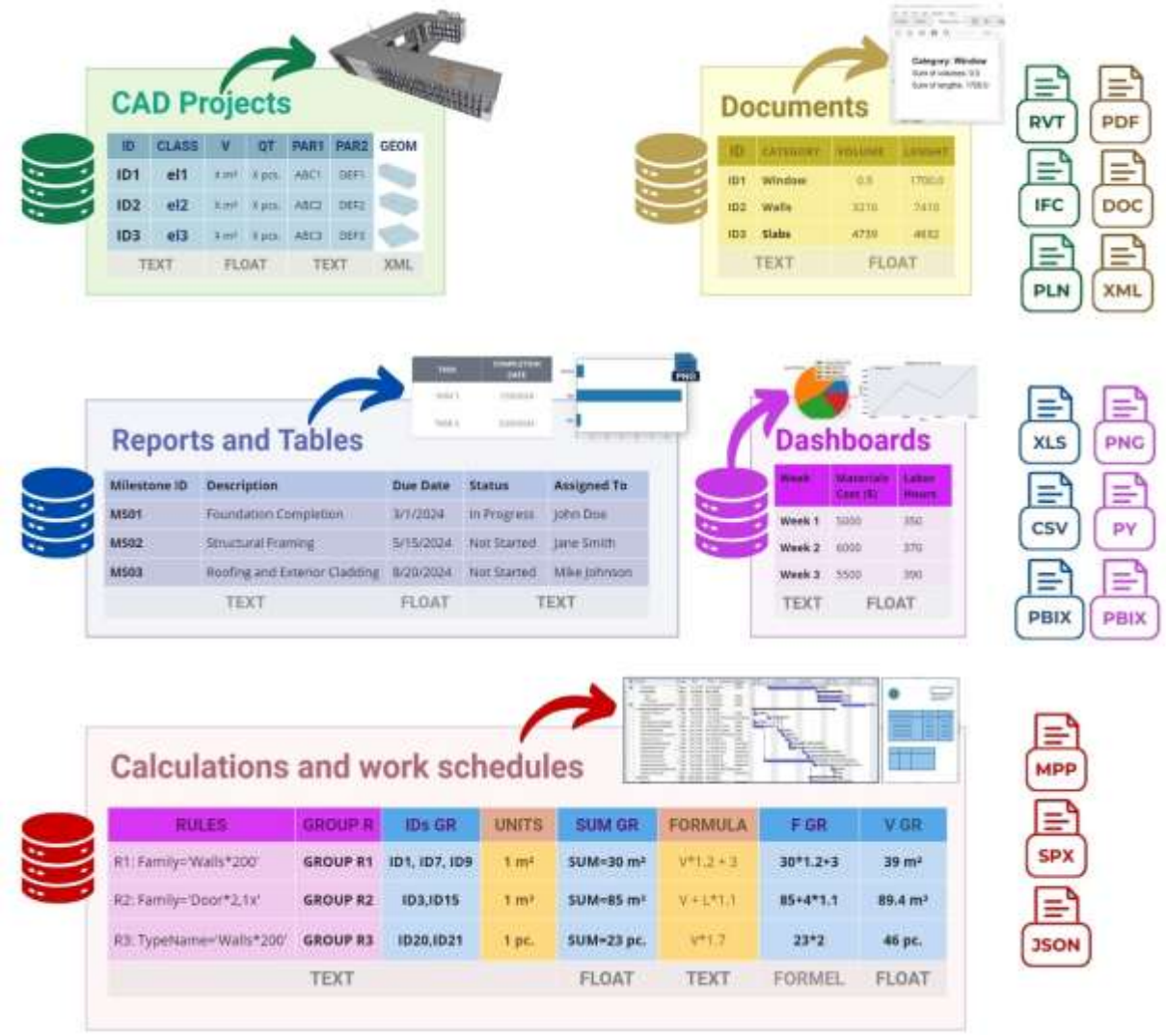


图 4.4-2 建筑公司的每个系统都是各种格式的法律相关文件的来源。

文本通信要求每个专业人员要么熟悉全部通信内容，要么定期参加所有会议，以了解项目的当前状况。

要克服这一局限，就必须从文本交流过渡到结构化的需求模型。这只有通过系统分析、流程可视化以及以流程图和数据模型的形式描述交互作用才能实现（图 4.4-3）。就像数据建模一样（图 4.3-7），我们通过添加参与者使用的系统和工具以及它们之间的联系，从上下文-想法层面进入概念层面。

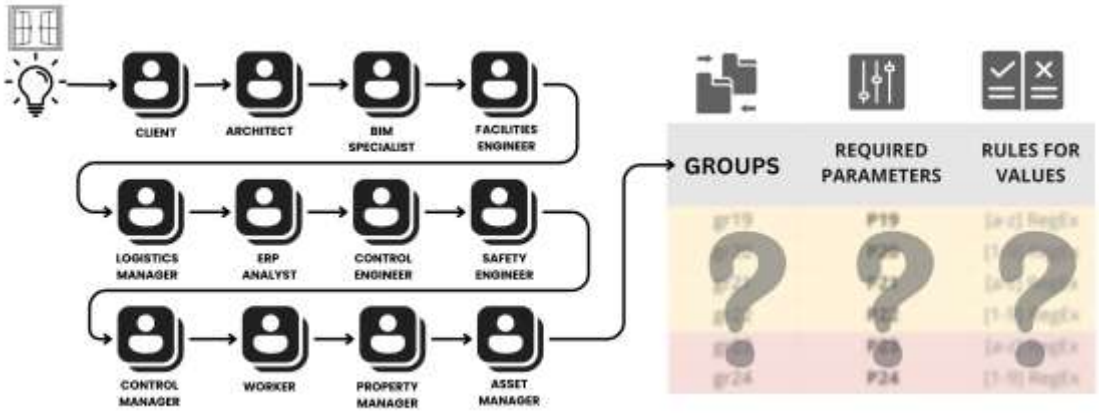


图 4.4-3 要了解如何管理和自动化验证流程，就必须将流程可视化，并将需求结构化。

将需求和关系系统化的第一步是使用概念流程图将所有链接和关系可视化。概念层面的 不仅能让所有流程参与者更容易理解整个流程链，还能清楚地显示每个流程步骤需要数据（和需求）的原因和对象。

流程图和概念框架的有效性

为了缩小传统数据管理方法与现代数据管理方法之间的差距，企业需要有意识地从零散的文本描述转变为结构化的流程表示。数据的演变--从泥板到数字生态系统--需要新的思维工具。使用流程图进行概念建模就是这样一种工具。创建可视化图表--流程图、过程图、交互图--可以让项目参与者意识到他们的行动和决策是如何影响整个决策系统的。

> 如果流程不仅需要存储数据，还需要分析数据或将数据自动化，那么你就需要开始处理创建概念化和可视化需求层的问题。

在我们的示例中（图 4.4-1），每个专家可以是一个小团队的成员，也可以是一个更大的部门的成员，该部门在总经理的控制下有多达十几个专家。每个部门都使用一个专门的应用数据库（图 1.2-4，如 ERP, CAD, MEP, CDE, ECM, CPM 等），定期更新创建文件、记录决策的法律状态和管理流程所需的输入信息。

> 交易过程类似于 4000 年前古代管理者的工作，当时人们使用泥板和纸莎草纸来合法确认决策。现代系统与其前身泥板和纸莎草纸的区别在于，现代方法还包括将文本信息转换为数字形式的过程，以便在其他系统和工具中进一步自动处理。

以概念流程图的形式将流程可视化，有助于描述每个步骤和不同角色之间的互动，使复杂的工作流程变得简单明了。

> 将 流程可视化，可确保流程逻辑透明，便于所有团队成员使用。

以文本、信息（图 4.4-1）和框图形式描述的为项目添加窗口的交流过程与我们在数据建模一章中讨论的概念模型（图 4.4-4）相似。

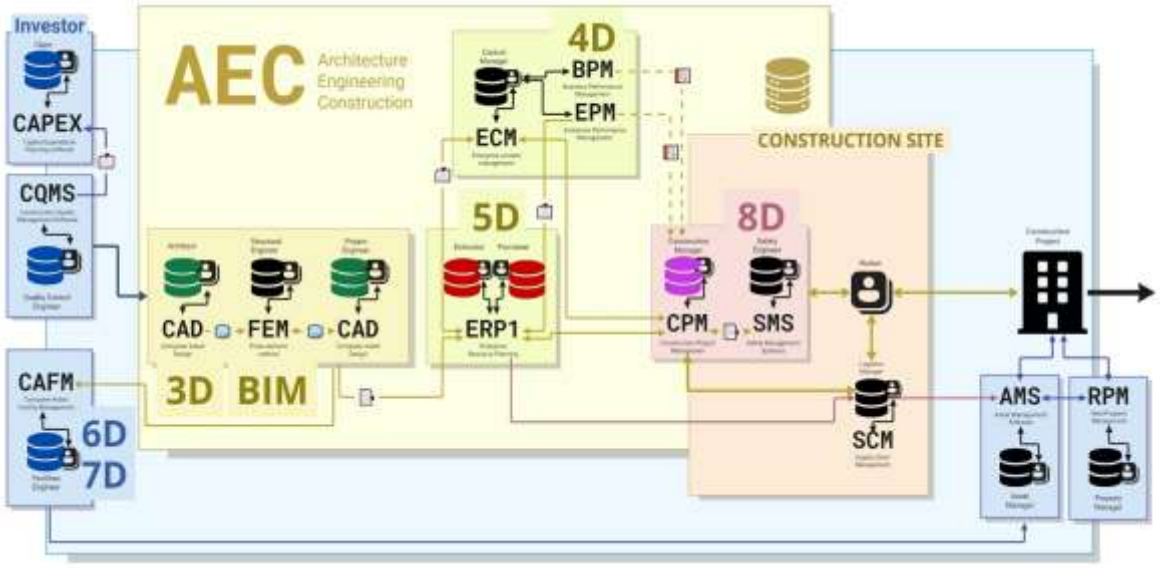


图 4.4-4 概念图显示，项目参与者是数据库的用户，他们的查询将不同的系统连接起来。

尽管概念图是一个重要步骤，但许多公司将自己局限在这一层面，认为直观图就足以理解流程。这就造成了一种可管理性的假象：管理人员更容易从这种流程图中了解全局，看到参与者和各阶段之间的联系。然而，这样的流程图并不能让人清楚地了解每个参与者需要哪些数据，应以何种格式传输数据，以及哪些参数和属性是实现自动化的必备条件。概念流程图更像是一张路线图：它指出了谁与谁之间的互动，但并没有揭示这些互动中传输的内容。

> 即使使用流程图在概念层面对流程进行了详细描述，也不能保证其有效性。可视化 通常会简化管理人员的工作，使他们能够通过逐步报告更轻松地跟踪流程。然而，对于数据库工程师来说，概念表述可能不够清晰，也可能无法清楚地了解如何在参数和需求层面实施流程。

随着我们向更复杂的数据生态系统迈进，概念和可视化工具的初步实施变得至关重要，以确保数据流程不仅高效，而且符合组织的战略目标。要将此窗口添加流程（图 4.4-1）完全转换到数据需求层面，我们需要更深入一层，将流程的概念可视化转换到数据、所需属性及其边界值的逻辑和物理层面。

结构化需求 和 **RegEx** 正则表达式

公司创建的数据中高达 80% 是非结构化或半结构化格式的数据[52]--文本、文档、信件、PDF -文件、对话。这些数据（图 4.4-1）难以分析、验证、在系统间传输和在自动化中使用。

为了确保可管理性、透明度和自动验证，有必要将文本和半结构化的需求 转换为定义明确的结构化格式。结构化过程不仅涉及数据（我们已在本书前几章详细讨论过），还涉及需求本身，项目参与者通常在整个项目生命周期中以自由文本形式提出需求，但往往没有想到这些过程可以自动化。

> 正如我们已经将数据从非结构化的文本形式转换为结构化形式一样，在需求工作流中，我们将把文本需求 转换为结构化的 "逻辑和物理层 "格式。

作为添加窗口示例的一部分（图 4.4-1），下一步是以表格形式描述数据要求。我们将通过指定关键属性及其边界值来组织项目参与者使用的每个系统的信息

举例来说，其中一个系统（图 4.4-5）--施工质量管理系统（CQMS ）--由客户方的质量控制工程师使用。在该系统的帮助下，他可以检查项目的新要素（在本例中为 "新窗口"）是否符合既定的标准和要求。

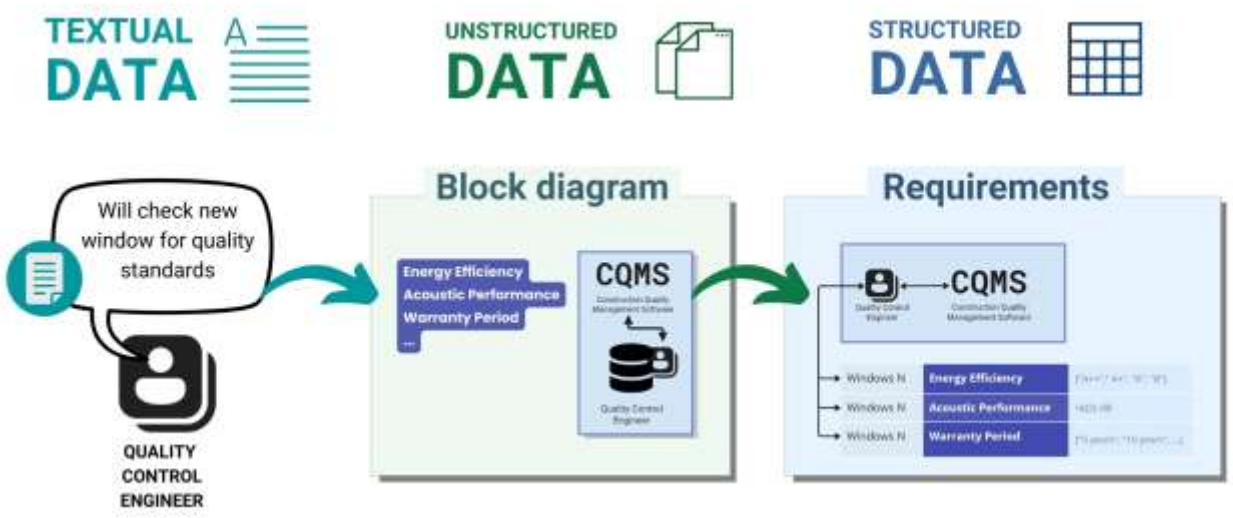


图 4.4-5 将文本需求转换成包含实体属性描述的表格格式，可简化其他专家的理解。

例如，在 CQMS -system（图 4.4-6）中，对 "窗户系统 "类型实体属性的一些重要要求：能效、声学性能和保修期。每个类别都包括在设计和安装窗户系统时需要考虑的某些标准和规范。



图 4.4-6 质量控制工程师应检查新窗型部件的能效、隔音和保修标准。

例如，质量控制工程师以表格形式指定的数据要求有以下边界值：

- 窗户的能效等级从 "A++"到 "B "不等，**"A++"**表示最高能效，"B "表示最低可接受水平，这些等级由可接受值列表**["A++"**、**"A+"**、**"A"**、**"A"**、**"B"]**表示。
- 窗户的隔音性能以分贝为单位，显示其降低街道噪音的能力，用正则表达式 \d{2}dB 来定义。

- 窗口类型 "实体的 "保修期 **"**属性以 5 年为起点，将此期限设定为选择产品时允许的最 短期限；也可指定保修期值，如**["5** 年**"**、**"10** 年 **"**等**]**或逻辑条件**">5**（年）**"**。

根据收集到的要求，在既定属性范围内，等级低于 "B"（如 "C "或 "D"）的新窗口类别或等级元素将无法通过能效测试。提交给质量保证工程师的数据或文件中的窗户隔声量应标注为两位数，后跟后缀 "dB"，如 "35 dB "或 "40 dB"，不在此格式内的数值，如 "9 D B "或 "100 分贝 "将不被接受（因为它们将无法通过 RegEx 字符串的模式）。质保期必须至少从 "5 年 "开始，质保期较短的窗口，如 "3 年 "或 "4 年"，将不符合质量工程师在表格格式中描述的要求。

在验证过程中，为了对照要求中的边界值检查这些属性参数值，我们会使用允许值列表（["A"、"B"、"C"]）、字典（["A"：["H1"、"H2"；"B"：W1"、"W2"]）、逻辑运算（如数字值的">"、"<"、"<="、">="、"=="）和正则表达式（如 "声学性能 "属性中的字符串和文本值）。正则表达式 是处理字符串值时一个极其重要的工具。

> 正则表达式 (RegEx) 在编程语言（包括 Python (Re 库)）中用于查找和修改字符串。Regex 就像字符串世界中的侦探，能够精确地识别文本中的文本模式。

在正则表达式中，字母直接使用相应的字母字符来描述，而数字则可以使用特殊字符 \d 来表示，它对应于 0 到 9 之间的任何数字。方括号用于表示字母或数字的范围，例如，[a-z] 表示拉丁字母表中的任何小写字母，[0-9] 相当于 \d。对于非数字和非字母字符，分别使用 \D 和 \W。

常用 RegEx 用例 （图 4.4-7）：

- 验证 电子邮件地址：要检查一个字符串是否是有效的电子邮件地址，可以使用 "^ **[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$"** 模板。
- 日期提取："\b\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}**.\d{4}\b"** 模板可用于以 DD.MM.YYYYYY 格式从文本中提取日期。
- 验证 电话号码：要验证格式为 +49(000)000-0000 的电话号码，模式应为**"\+\d{2}\(\d{3}\)\d{3}-\d{4}"**。

通过将质量保证工程师的要求 转换为属性格式及其边界值（图 4.4-6），我们将其从原始文本格式（对话、

信件和规范性文件）转换为有组织的结构化表格，从而可以自动检查和分析任何传入的数据（例如，窗口类别的新 元素）。有了这些要求，未经检查的数据就会被自动丢弃，而经过检查的数据则会被自动传输到系统中进行进一步处理。

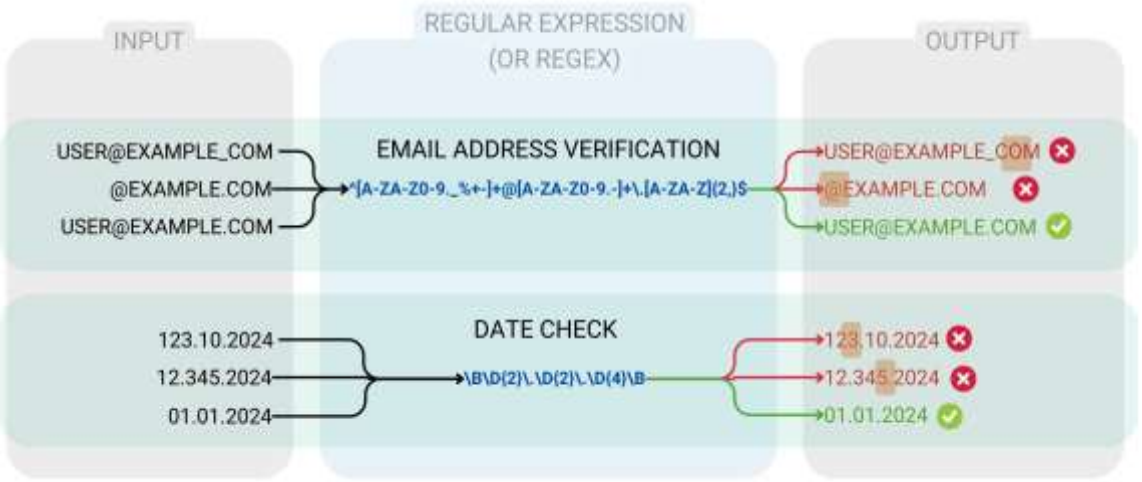


图 4.4-7 使用正则表达式 是文本数据验证过程中极为重要的工具。

现在，我们将从概念层面转向逻辑层面来处理需求，将安装新窗口过程中所有专家的需求 （图 4.4-4）转换成属性格式的有序列表，并将这些列表和必要的属性添加到每个专家的流程图中（图 4.4-8）。

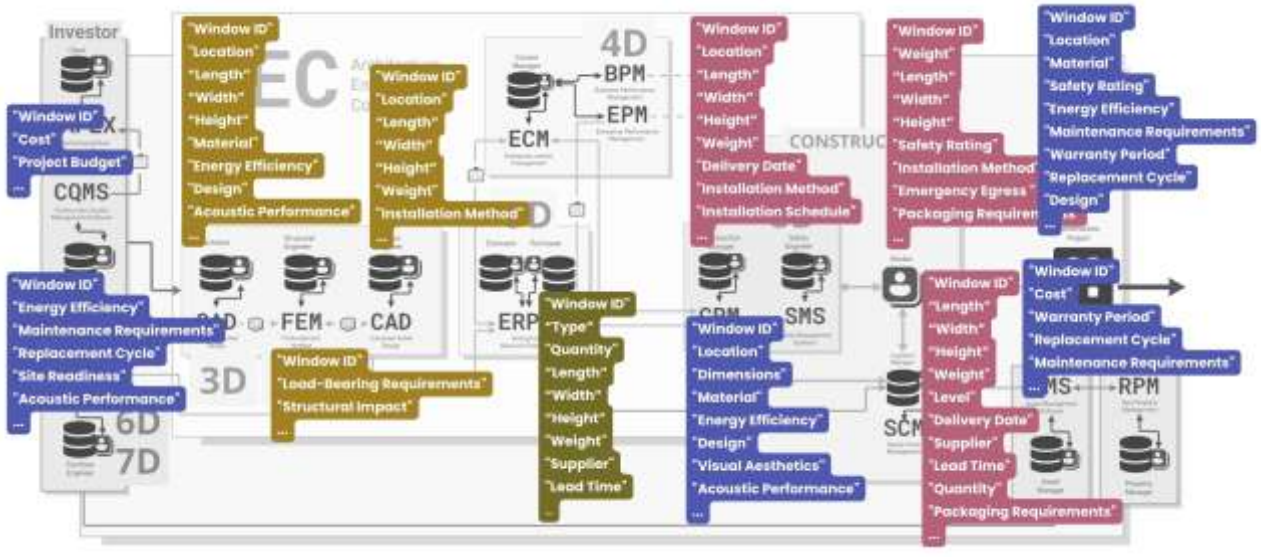


图 4.4-8 在逻辑流程层面，每个专家处理的属性都会添加到各自的系统中。

通过将所有属性添加到一个通用流程表中，我们将以前在概念层面以文本和对话形式呈现的信息（图 4.4-1

）转换为结构化和系统化的物理层面表格形式（图 4.4-9）。

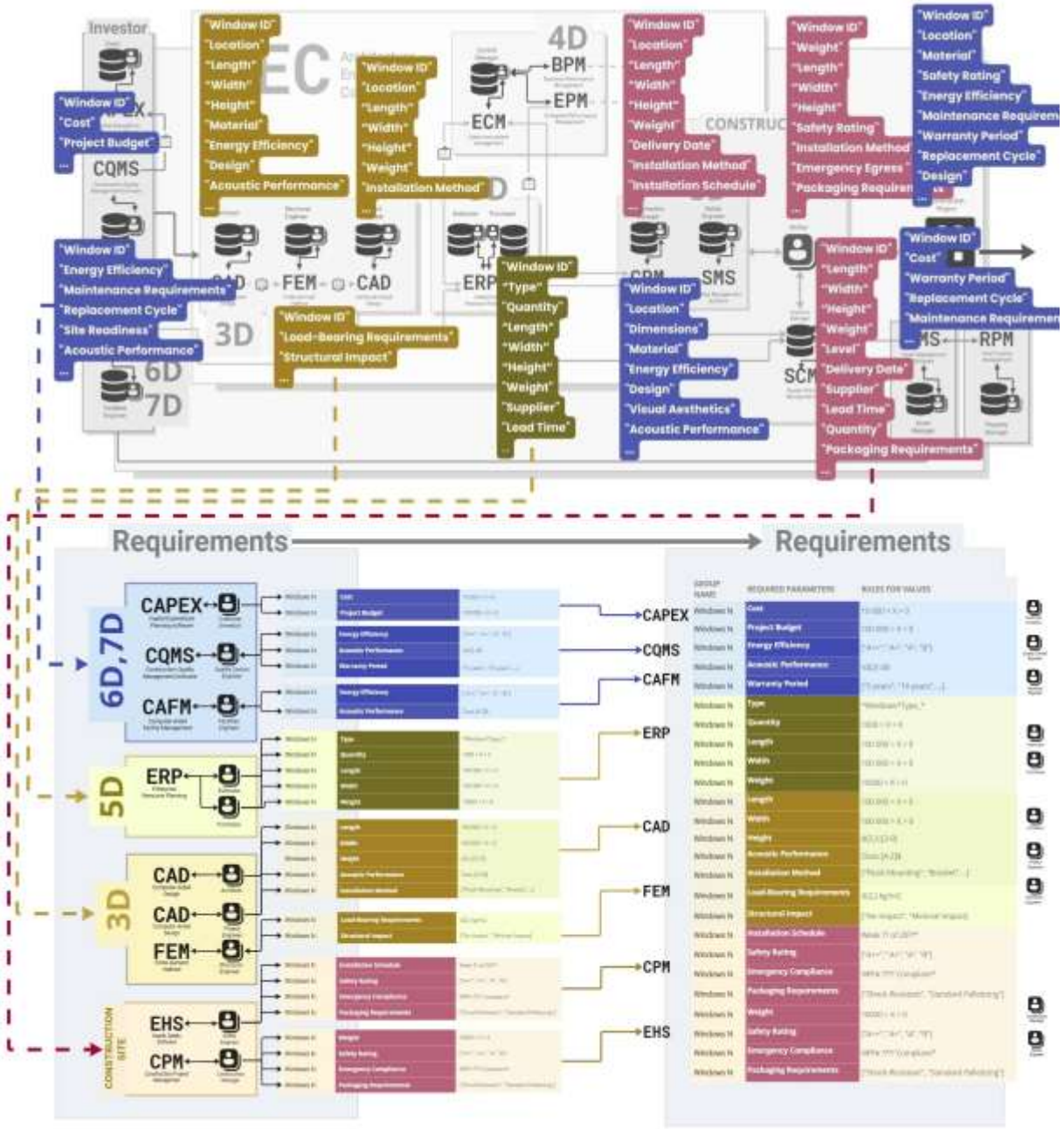


图 4.4-9 将非结构化专家对话转换为结构化表格有助于了解 在物理层面的要求。

现在，需要将数据要求传达给为特定系统创建信息的专家。例如，如果您正在使用 CAD 数据库，那么在开始对元素进行建模之前，您应该根据数据的最终使用场景收集所有必要的参数。这通常从操作阶段开始，然后是施工现场、物流部门、估算部门、结构计算部门等。只有在考虑了所有这些方面的要求后，才能开始根

据收集到的参数创建数据。这样，您就可以在未来自动验证和传输整个链条上的数据。

当新数据符合要求时，就会自动集成到公司的数据生态系统中，直接进入用户和系统。根据属性及其值对数据进行验证，可确保信息符合规定的质量标准，并可在公司场景中使用。

数据要求已经确定，现在，在开始核查之前，必须创建、获取或收集要核查的数据，或者记录数据库中信息的当前状态，以便在核查过程中使用。

数据收集

在开始验证之前，必须确保以适合验证过程的形式提供数据。这并不意味着简单地提供信息，而是要做好准备工作：必须收集数据，并将其从非结构化、松散结构化、文本和几何格式转换为结构化形式。前几章详细介绍了这一过程，并讨论了转换不同类型数据的方法。经过所有转换后，输入数据将以开放式结构表的形式出现（图 4.1-2、图 4.1-9、图 4.1-13）。

有了要求和包含必要参数和边界值的结构化表格（图 4.4-9），我们就可以开始验证数据--既可以是一个单一的自动流程（管道），也可以是对每个传入文档的逐步验证。

为了启动检查，需要接收一个新文件作为输入，或者固定数据的当前状态--创建快照或导出当前和输入的数据，或者建立与外部或内部数据库的连接。在本示例中，这种快照是通过自动将 CAD 数据从 转换为结构化格式来创建的，例如在 2024 年 3 月 29 日（星期五）23:00:00，即所有设计人员都回家之后记录的数据。



图 4.4-10 CAD 数据库快照 (BIM) 显示当前版本项目模型中 "窗口 "类新实体的当前属性信息。

借助 "将 CAD 数据 (BIM) 转换为结构化形式 "一章中讨论的逆向工程工具，可以将来自不同 CAD (BIM) 工具和编辑器的信息组织到单独的表格中（图 4.4-11），或将其合并到一个包含项目不同部分的通用表格中（图 9.1-10）。

这种表格--数据库显示门窗的唯一标识符（ID 属性）、类型名称（TypeName）、尺寸（宽度、长度）、材料（材料）以及能耗和声学性能指标和其他特征。在 CAD 程序 (BIM) 中填写的此类表格由设计工程师从各部门和文件中收集，形成项目的信息模型。



图 4.4-11 来自 CAD 系统的结构化数据 可以是一个二维表格，列表示元素的属性。

真正的 CAD (BIM) 项目包括数万或数十万个元素（图 9.1-10）。CAD 格式中的元素按类型和类别自动分类，从门窗到楼板、板和墙。独特的标识符（如 CAD 解决方案自动设置的本地 ID）或类型属性（类型名称、类型、族）允许在不同系统中跟踪同一对象。例如，一栋建筑北墙上的一扇新窗户可以在该组织的所有相关系统中通过单一标识符 "W-NEW "进行唯一标识。

> 虽然实体名称和标识符在所有系统中都应保持一致，但与这些实体相关的属性和值却会因使用环境的不同而大相径庭。建筑师、结构工程师、建筑、物流和物业维护专业人员都以不同的方式看待相同的元素。他们每个人都有自己的分类标准和目标：有些人纯粹从美学角度看待窗户，评估其形状和比例，而另一些人则从工程或操作角度看待窗户，分析其导热性、安装方法、重量或维护要求。因此，在建立数据模型和描述元素时，必须考虑到其用途的多样性，并确保数据的一致性，同时考虑到行业的特殊性。

公司流程中的每个角色都有专门的数据库，从设计和计算到物流、安装和楼宇运行，都有自己的用户界面（图 4.4-12）。每个系统都由一个专业的专家团队通过专门的用户界面或数据库查询进行管理，在这个链条的末端，系统管理员或部门经理负责对输入数据的法律效力和质量负责，并为其他系统提供服务。

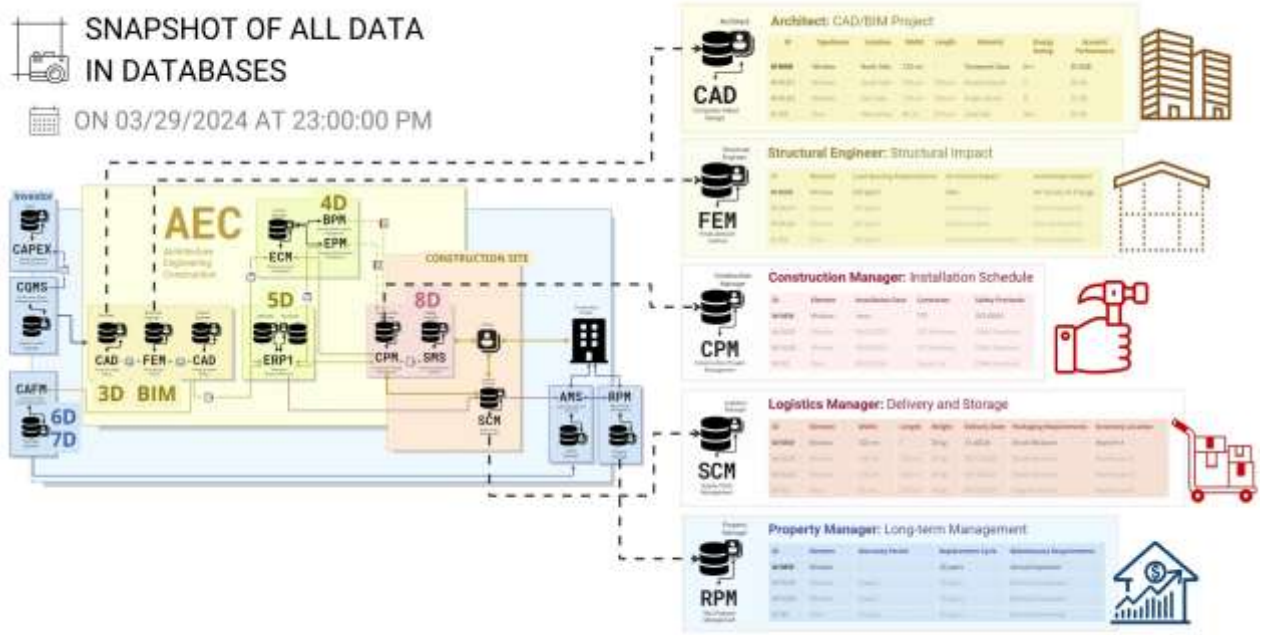


图 4.4-12 同一实体在不同系统中具有相同的标识符，但只有在该系统中才具有重要的不同属性。

一旦我们在逻辑和物理层面上组织了结构化需求和数据的收集工作，我们就需要建立一个程序，根据之前收集的需求，自动检查来自不同传入文件和不同系统的数据。

核实 数据和核实结果

所有进入系统的新数据，无论是来自客户、建筑师、工程师、工头、物流师或物业经理的文件、表格或数据库条目，都必须根据之前制定的要求进行验证（图 4.4-9）。验证过程至关重要：数据中的任何错误都可能导致计算错误、工期延误甚至经济损失。为了最大限度地降低这种风险，应该组织一个系统的、可重复的、反复的数据验证程序。

> 要验证进入系统的新数据（非结构化数据、文本数据或几何数据），必须将其转换为松散的结构化或结构化格式。然后，验证过程必须根据所需属性及其允许值的完整列表对数据进行检查。

将不同类型的数据：文本、图像、PDF -文档和混合 CAD (BIM) 数据转换为结构化形式，在 "将数据转换为结构化形式 "一章中进行了详细讨论。

例如，从 CAD (BIM) 项目中获得的表格（图 4.4-11）。它包括半结构化几何数据和项目实体的结构化属性信息（图 3.1-14）--例如 "视窗 "类中的一个元素。

为了进行验证，我们将属性值（图 4.4-11）与专家以需求形式定义的参考边界值（图 4.4-9）进行比较。最终的比较表（图 4.4-13）将明确哪些值是可接受的，哪些值需要修正，然后才能将数据用于 CAD 应用程序之外(BIM)。

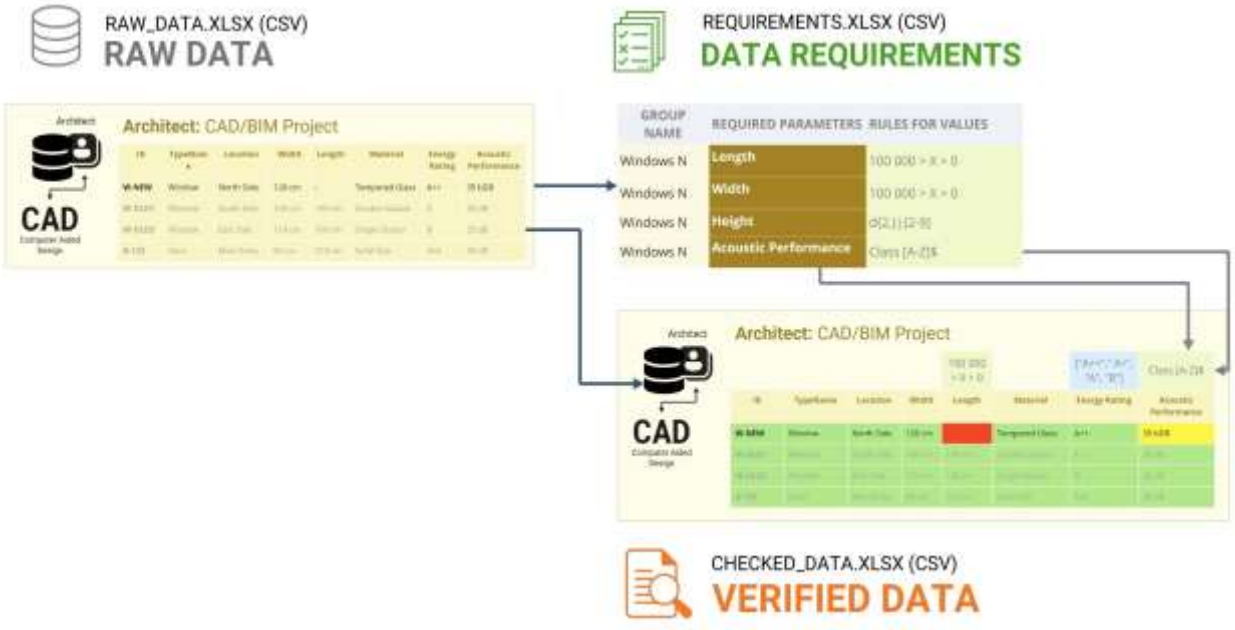


图 4.4-13 最后的验证表突出显示了 "Windows "类新实体应注意的属性值。

在 "Pandas：数据分析不可或缺的工具 "一章中，我们介绍了 Pandas 库 ，使用 Pandas 库实施类似的解决方案，我们将使用从另一个表格需求文件 中提取的数据（图 4.4-9）来验证从 CAD 文件 (BIM) (RVT, IFC, DWG, NWS, DGN) 中提取的表格文件数据（图 4.4-11）。

要获取代码，我们需要在 LLM 的提示符中说明，我们需要从 **raw_data.xlsx** 文件（CAD 数据库 (BIM) 中的一套完整数据）中加载数据 ，对其进行检查，并将结果保存到一个新文件 **checked_data.xlsx** 中（图 4.4-13）。

让我们在不提及 Pandas 库的情况下使用 LLM 获取代码：

> 编写代码来检查 raw_data.xlsx 文件中的表格，并使用以下验证规则对其进行验证："宽度 "和 "长度 "列的值大于零，"能效等级 "包含在['A++'、'A+'、'A'、'B']列表中，"声学性能 " 作为我们稍后指定的变量--添加了最后的验证列，并将最终表格保存到一个新的 Excel 文件中 checked_data.xlsx ⏎

LLM 的回复将描述一个 Python 代码的简短示例，该示例可通过后续提示进行完善和扩充：

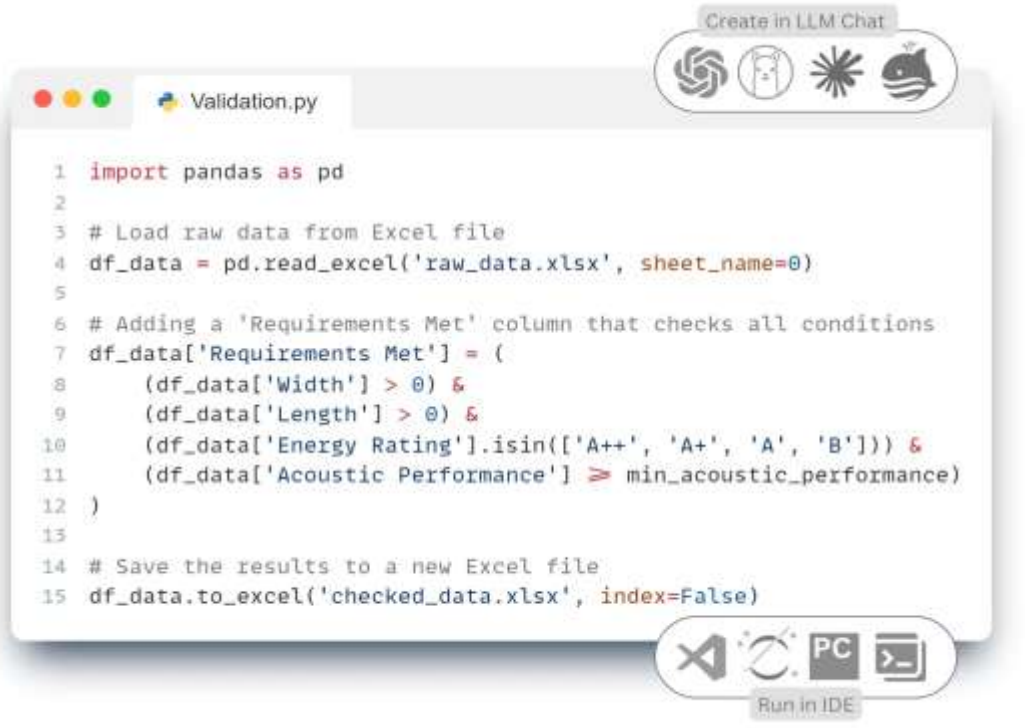


图 4.4-14 LLM- 模型生成的代码根据边界值形式的属性要求检查转换后的 CAD (BIM) 设计。

LLM 语言模型生成的代码 ，可用于任何流行的集成开发环境或在线工具：PyCharm、Visual Studio Code

(VS Code)、Jupyter Notebook 、Spyder、Atom、Sublime Text、带有 PyDev 插件的 Eclipse、Thonny、Wing IDE、带有 Python 插件的 IntelliJ IDEA 、JupyterLab 或流行的在线工具 Kaggle。com、Google Collab、Microsoft Azure Notebooks、Amazon SageMaker。

执行代码 (图 4.4-14)将显示 CAD 数据库 (BIM) 中的 "实体元素 "W-OLD1、W-OLD2 和 D-122（以及其他 元素）符合属性要求：宽度和长度大于零，能效等级为列表值 "A++"、"A"、"B"、"C "之一 (图 4.4-15)。

我们需要并在最近添加的 W-NEW 元素负责北侧的新元素类别 "窗户"，但它不符合要求（属性 "满足要求"），因为它的长度为零（根据我们的 "宽度">0 规则，"0.0 "的值是不可接受的），而且它没有指定能效等级。



	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

图 4.4-15 检查 识别未通过验证流程的实体，并在结果中添加值为 "假 "或 "真 "的新属性。

同样，在为项目添加窗口的过程中，我们会检查所有项目元素（实体）的一致性，以 及从不同专家处收到的所有数据中每个系统、表格或数据库所需的属性（图 4.4-1）。

在最后的表格中，可以用颜色突出显示检查结果，以便直观显示：用绿色标出成功通 过检查的属性，黄色标出非关键偏差值，红色标出关键偏差值（图 4.4-16）。

验证的结果是（图 4.4-16），我们会得到一个可信和已验证元素的列表，其 ID 经过验证符合属性要求。经过验证的元素为所有参与添加 Window 类或任何其他类元素过程的系统提供了保证，即这些元素符合既定的标准和规范（我们将在 "自动 ETL 和数据验证 "一章中讨论更多有关自动数据验证和创建自动 ETL 流程 的内容）。

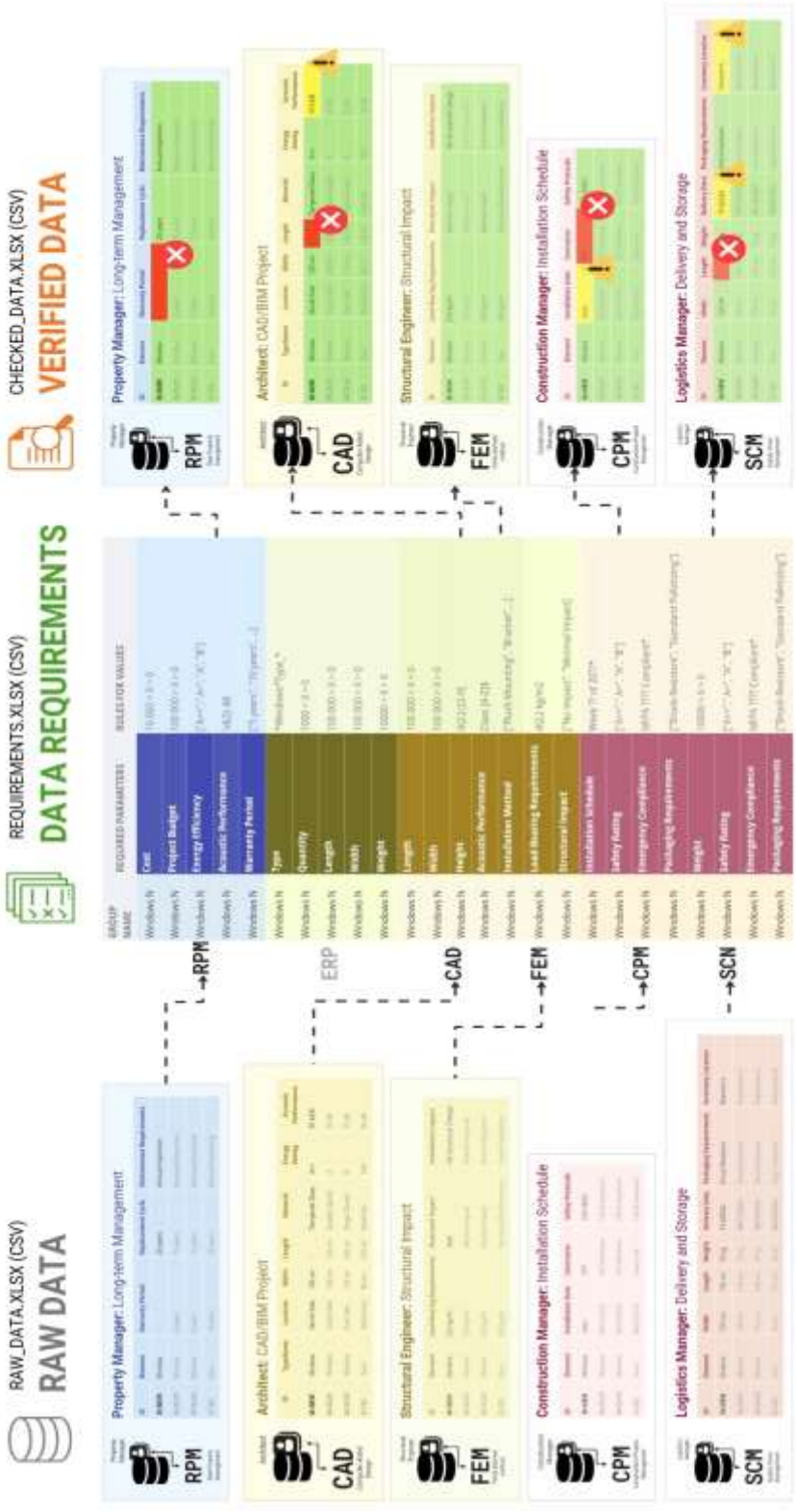


图 4.4-16 对所有系统进行检查的结果可以确定哪些数据不符合公司要求。

成功通过验证的实体通常不需要太多关注。它们可以顺利进入下一阶段的处理和集成到其他系统中。与 "优质 "项目相比，未能通过验证的项目才是最值得关注的。有关这些偏差的信息至关重要：不仅要以表格报告的形式传达，还要使用各种可视化工具。验证结果的图形显示有助于快速评估数据质量的整体状况 ，找出问题所在，并及时采取纠正或纠正措施。

验证结果的可视化

可视化 是解释检查结果的重要工具。除了常规的汇总表外，它还包括信息面板、图表和自动生成的 PDF 文档，这些文档按检查状态对项目元素进行分组。在这里，颜色编码可以起到辅助作用：绿色表示已成功验证的项目，黄色表示需要额外注意的项目，红色表示存在严重错误或缺少关键数据的项目。

在我们的示例中（图 4.4-1），我们逐步分析每个系统的数据：从 CAD (BIM) 和财产管理到物流和安装计划（图 4.4-16）。审核结束后，会自动为每位专家生成个人警报或报告文件，如 PDF 格式的 （图 4.4-17）。如果数据无误，专家会收到一条短信："感谢您的合作"。如果发现差异，则会发送一份详细报告，措辞如下："本文件列出了未经合规检查的元素、其标识符、属性和值"。



图 4.4-17 验证和自动生成报告加快了创建数据的专业人员发现和了解数据缺陷的过程。

由于采用了自动验证程序--一旦发现错误或数据缺口，就会以聊天信息、电子邮件或 PDF 文档的形式向负责创建或处理相关实体及其属性的人员发送即时通知（图 4.4-18），并列出尚未验证的元素和属性说明。



图 4.4-18 自动检查报告更容易理解错误并加快完成项目数据的工作。

例如，如果物业管理系统收到的文件（结构化后）中 "保修期 "属性填写有误，物业经理就会收到提示，列出需要检查和更正的属性。

同样，安装计划或物流数据中的任何缺陷都会导致自动报告的生成，例如，向相关专家发送包含检查结果的聊天通知或电子邮件。

除了 PDF -文档和结果图表外，还可以创建仪表盘和交互式 3D -模型（图 7.1-6，图 7.2-12），突出显示属性缺失的元素，使用户可以直观地使用元素的 3D 几何图形来筛选和评估项目中这些元素的质量和完整性。

> 以自动生成文件、图表或仪表盘的形式将检查结果可视化，大大简化了数据解释，并促进了项目参与者之间的有效沟通。

自动检查来自各种系统和信息源的数据的过程可以比作日常生活中的知情决策。正如建筑行业的公司会考虑许多变量--从输入数据的可靠性到它们对项目实现的时间、成本和质量的影响--同样，在做出重要决策时，例如在选择居住地时，人们会权衡一系列因素：交通便利性、基础设施、成本、安全性和生活质量。所有这些考虑因素构成了一个标准体系，成为我们生活中最终决策的基础。

数据质量检查与人类生活需求的比较

尽管数据质量控制方法和工具在不断发展，但信息合规性的基本原则却始终未变。无论是在业务中还是在日常生活中，这一原则都是成熟管理系统的基础。

> 迭代数据验证过程很像每个人每天都要面对的决策过程。在这两种情况下，我们都要依靠经验、数据和新获得的信息。越来越多的生活和职业决策，从战略决策到日常决策，都是基于数据做出的。

例如，在选择居住地或生活伴侣时，我们会凭直觉在头脑中形成一个标准和特征表，并据此对备选方案进行比较（图 4.4-19）。这些特征--无论是人的个人品质还是房产的参数--都是影响最终决定的属性。

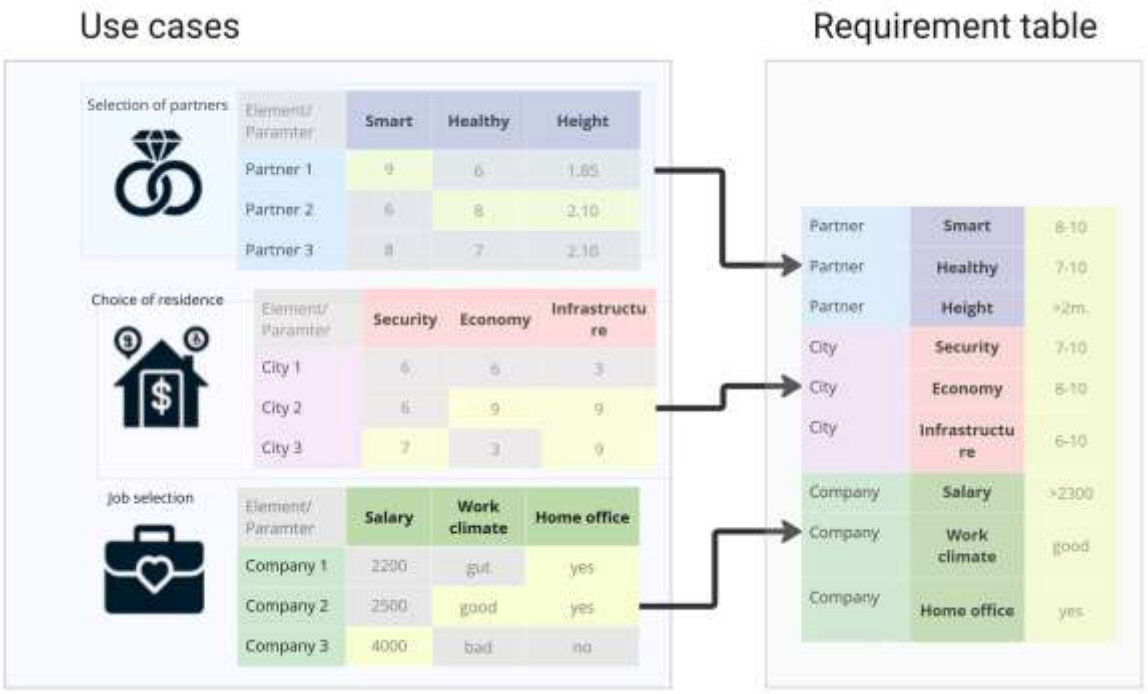


图 4.4-19 根据个人属性要求选择居住、就业或合伙。

使用结构化数据和正规化方法来描述需求（图 4.4-20）有助于在职业和个人生活中做出更加明智和知情的选择。

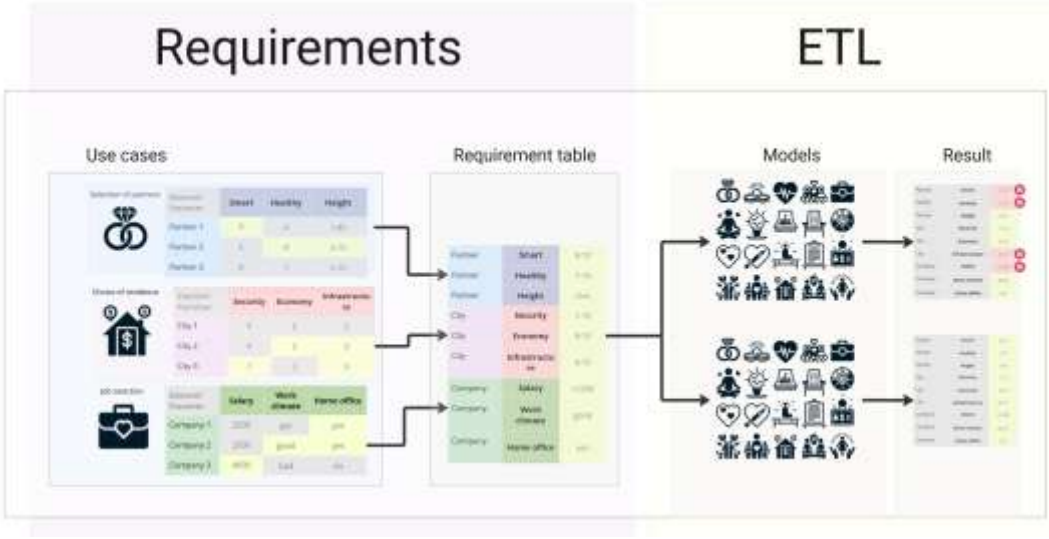


图 4.4-20 需求的形式化可以使生活感知和业务决策系统化。

数据驱动决策方法并不只是一种商业工具。它也可以无缝集成到日常生活中，遵循常见的数据处理步骤（图 4.4-21），类似于 ETL 流程 （提取、转换、加载），我们在本部分开始时已经讨论过数据结构化的问题，在本书第七部分的任务自动化中我们将详细讨论这一问题：

- 数据是基础（摘录 ）：在任何领域，无论是工作还是个人生活，我们都会收集信息。在企业中，它可以是报告、数字、市场数据；在个人生活中，它可以是个人经历、亲人的建议、反馈、观察。
- 评价标准（**Transform** ）：根据预先确定的标准来解释收集到的信息。在工作中，这些标准是绩效指标（KPI）、预算限制和规范；在个人生活中，这些标准是价格、便利性、可靠性、魅力等参数。
- 预测和风险分析（**Load** ）：最后阶段是在分析转换数据和比较可能后果的基础上做出决策。这与业务流程类似，数据要经过业务逻辑和风险过滤器。

> 我们所做的决定，从早餐吃什么这样的琐碎喜好，到选择职业或生活伴侣这样的人生大事，本质上都是处理和评估数据的结果。

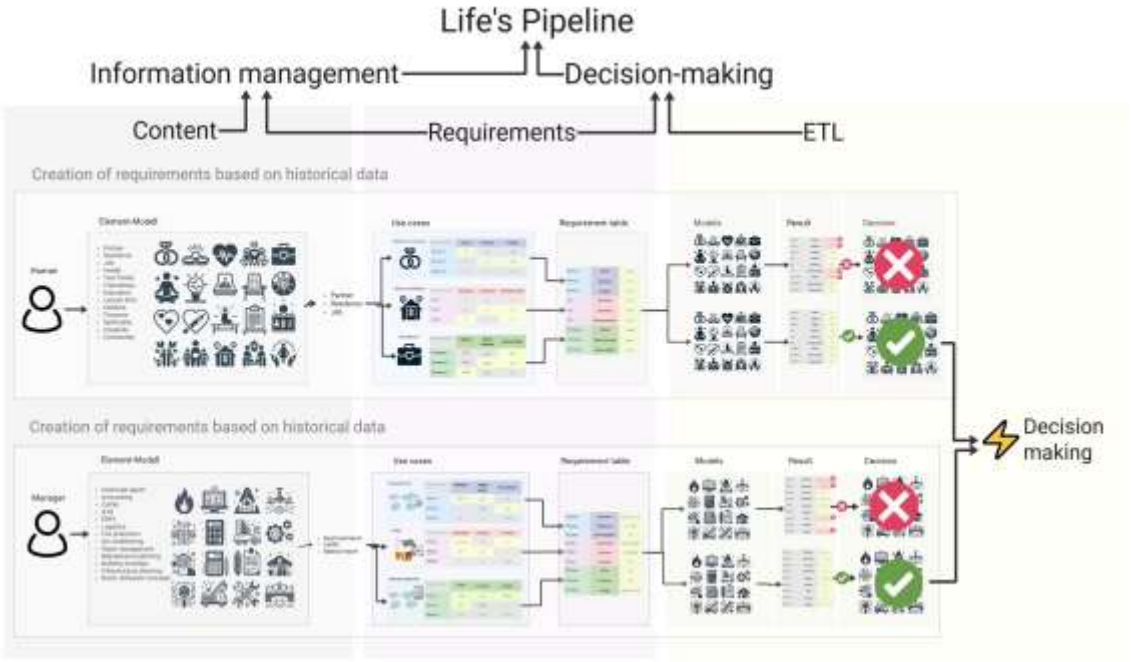


图 4.4-21 商业和生活一般都是一系列数据驱动的决策，而决策所用数据的质量是关键因素。

我们生活中的一切都相互关联，正如包括人类在内的生物体遵循自然规律，不断进化并适应不断变化的条件一样，人类的工作流程，包括我们如何收集和分析数据，也反映了这些自然原则。大自然与人类活动之间的密切关系不仅证明了我们对大自然的依赖，也证明了我们希望应用经过数百万年进化磨练的规律来创建数据架构、流程和系统，以便进行决策。

新技术，尤其是建筑领域的新技术，是人类一次又一次从大自然中获得灵感，创造出更好、更可持续和更高效解决方案的最好例证。

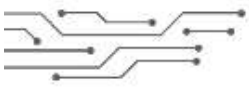
下一步：将数据转化为精确的计算和计划

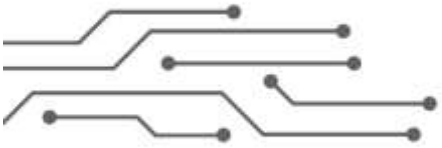
在这一部分中，我们探讨了如何将非结构化数据转换为结构化格式、开发数据模型以及组织流程以检查施工项目中的信息质量。数据管理、标准化和分类是一个基本过程，需要系统的方法和对业务需求的清晰理解。本部分讨论的技术和工具可在对象的整个生命周期内实现不同系统之间的可靠集成。

在总结这部分内容时，让我们强调一下主要的实际步骤，这些步骤将有助于您在日常工作中应用所讨论的方法：

- 从系统化需求开始
  - [ ] 为项目和 流程的关键要素创建属性和参数注册表
  - [ ] 记录每个属性的边界值
  - [ ] 使用流程图（如在 Miro、Canva、Visio 中）可视化流程以及类、系统和属性之间的关系
- 自动进行数据转换
  - [ ] 检查哪些流程中经常使用的文件可以使用 OCR 库进行数字化，并转换成表格形式
  - [ ] 查看逆向工程工具 ，从 CAD 提取数据 (BIM)
  - [ ] 尝试将工作中经常使用的文档或格式中的数据自动检索到电子表格中
  - [ ] 设置不同数据格式之间的自动转换
- 创建分类知识库
  - [ ] 开发符合行业标准的内部分类器或使用现有的 元素分类器
  - [ ] 记录不同分类系统之间的相互关系
  - [ ] 与你的团队讨论使用统一的元素识别和明确分类系统的问题
  - [ ] 开始建立自动验证数据的流程，包括团队内部使用的数据和传递到外部系统的数据

通过使用这些方法，可以显著提高数据质量，简化后续处理和转换。在本书的后续部分，我们将探讨如何将已经结构化和准备好的数据用于自动计算、成本估算、进度安排和施工项目管理。





V 部分

成本和时间计算：将数据纳入施工流程

第五部分专门讨论利用数据优化建筑项目成本计算和规划的实际问题。详细分析了基于资源的成本估算方法 和估算过程自动化。考虑了从 CAD (BIM)模型自动获取工程量清单)的方法及其与计算系统的整合。探讨了用于时间安排和施工成本管理的 4D 和 5D 建模技术，并提供了具体的应用实例。对扩展信息层 6D -8D 进行了分析，为评估房地产对象的可持续性、运行和安全性提供了综合方法。详细介绍了在现代环境要求和标准背景下计算建筑项目碳足迹和 ESG 指标的方法。对传统 ERP 和 PMIS 系统在施工过程管理中的可能性和局限性进行了批判性评估，并分析了它们对价格透明度的影响。预测了从封闭式解决方案向开放式标准和灵活的数据分析工具过渡的前景，这些工具能够确保提高建筑流程的效率。

章 5.1.

建筑项目的成本计算和估算

建筑基础知识：估算数量、成本和时间

在决定建筑业公司可持续性的众多业务流程中，尤其重要的是准确估算要素数量、项目成本和交付时间的流程（图 5.1-1），几千年前也是如此。

文字的发展是多种因素共同作用的结果，其中包括早期社会记录经济交易、贸易和资源管理的需要。第一批具有重要法律意义的文件，即计算材料成本和人工费用的泥板，被用于贸易和建筑领域。这些石板记录了各方在建筑结构建造中的义务，并作为协议、货币和贸易关系的证据保存起来。

> 数千年来，估算方法基本保持不变：依靠估算工程师的经验和直觉进行手工计算。然而，随着模块化 ERP 系统和 CAD 工具的出现，估算数量、成本和时间的传统方法开始迅速转变。如今，数字化技术使时间和成本的关键计算得以完全自动化，从而提高了建筑项目资源规划的准确性、速度和透明度。

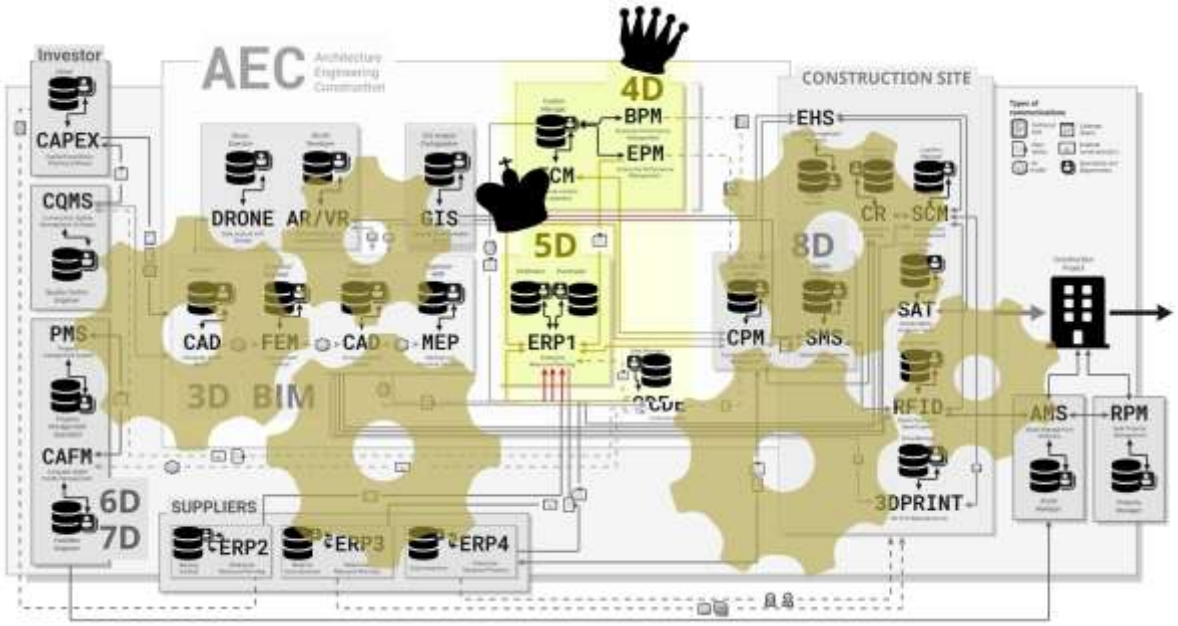


图 5.1-1 在众多不同的系统中，负责数量、成本和时间指标的工具是业务中最重要的。

建筑公司的重点是准确的时间和成本数据。这些数据反过来又取决于材料和劳动力的使用量，其透明度会影响盈利能力。然而，计算过程的复杂性和缺乏透明度往往会导致项目成本上升、误期甚至破产。

> 根据毕马威会计师事务所（KPMG）的报告《熟悉的问题-新方法》（2023 年），只有 50% 的建筑项目能够按时完工，87% 的公司表示对基本建设项目经济性的控制有所加强。主要问题与缺乏合格人员和难以进行风险预测有关[2]。

在建筑公司的整个生命周期中，在过去的项目施工过程中收集历史成本计算和流程时间数据，并将其输入各种系统的数据库（ERP 、PMIS BPM、EPM 等）。

> 拥有高质量的历史成本计算数据是建筑企业的一大竞争优势，直接影响到企业的生存能力。

建筑和工程公司设立估算和成本计算部门的目的是收集、存储和更新项目计算的历史数据。其主要功能是积累公司的经验并使之系统化，从而随着时间的推移提高对新项目的范围、时间安排和成本估算的准确性。这种方法有助于根据已实施项目的实践和结果，最大限度地减少未来计算的误差。

项目估计费用的计算方法

造价专家使用多种估算方法，每种方法都侧重于特定类型的数据、可用信息和项目的详细程度。最常见的方法包括

- 基于资源的方法 ：根据对材料、设备和劳动力等所有所需资源的详细分析，估算项目的估计费用。这种方法要求详细列出所有任务和执行每项任务所需的资源，然后计算其成本。这种方法精确度高，在成本估算中得到广泛应用。
- 参数法：使用统计模型根据项目参数估算费用。这可能涉及分析每个计量单位（如建筑面积或工程范围）的成本，并根据具体项目条件调整这些值。这种方法在早期阶段还没有详细资料时特别有效。
- 单位法（单位造价法）：根据计量单位（如每平方米或每立方米）的造价计算项目的估算造价。这为比较和分析不同项目或部分项目的费用提供了一种快速简便的方法。
- 专家判断法（德尔菲法）：基于专家的意见，他们利用 自己的经验和知识来估算项目的价值。当没有准确的基准数据或项目具有独特性时，这种方法非常有用。

值得注意的是，参数法和专家判断可适用于机器学习模型。这样就可以根据训练样本自动生成项目成本和进度预测。在 "使用机器学习 查找项目成本和进度的实例 "一章中，将详细讨论此类模型的应用实例（图 9.3-5）。

尽管如此，基于资源的方法仍然是世界上最流行和广泛使用的方法。它不仅能准确评估估算成本，还能计算施工现场各个工序和整个项目的工期（更多详情，请参阅 "施工进度和 4D - 项目数据 "一章）。

基于资源的方法 建筑成本计算和估算

以资源为基础的成本计算是一种管理会计方法，项目成本基于对所有相关资源的直接核算。在建筑工程中，这种方法涉及对完成工作所需的所有材料、劳动力和技术资源进行详细分析和评估。

> 基于资源的方法，在预算规划方面具有很高的透明度和准确性，因为它注重估算时资源的实际价格。这一点在不稳定的经济环境中尤为重要，因为价格波动会对项目的总体成本产生重大影响。

在接下来的章节中，我们将详细介绍基于资源的成本计算过程。为了更好地理解其构造原理，我们将以餐厅晚餐的成本计算来进行类比。餐厅经理在计划晚餐时，会列出必要的产品清单，考虑每道菜的烹饪时间，然后将成本乘以客人数量。在建筑工程中，过程也是类似的：对每一类项目要素（对象）进行逐项估算 Recipes，将所有费用相加得出总账单，即按类别进行的最终估算，从而确定项目的总成本。

以资源为基础的方法的关键和初始阶段是创建公司的初始数据库。在成本核算的第一阶段，要编制一份结构化清单，列出公司在建筑项目中可支配的所有项目、材料、工种和资源--从仓库中的一颗钉子到通过资格和小时工资对人员的描述。这些信息被系统化地纳入一个单一的 "建筑资源和材料数据库"--一个包含名称、特征、计量单位和当前价格等数据的表格登记册。该数据库是所有后续资源计算（包括工程成本和时间安排）的主要信息来源。

建筑资源数据库：建筑材料和工程目录

建筑资源和材料数据库或表格--包括建筑项目中可使用的每个要素--产品、项目、材料或服务的详细信息，包括其名称、描述、计量单位和单位成本，并以结构化的形式记录下来。在该表中，你可以找到从项目中使用的不同类型的燃料和材料到以不同类别形式列出的专家详细名单，以及每小时的费率说明（图 5.1-2）。

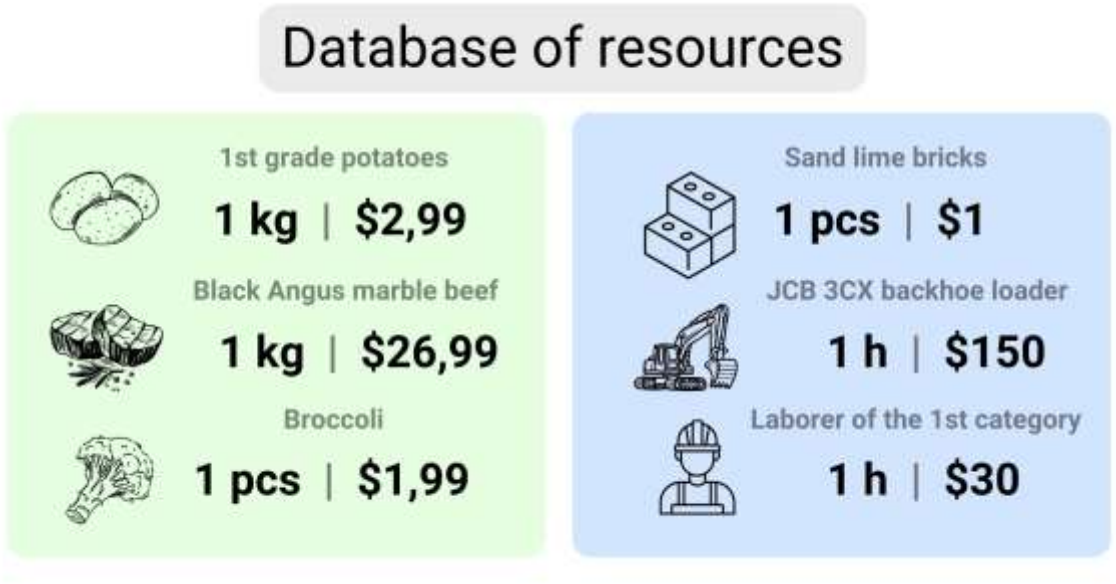


图 5.1-2 资源表是描述材料和服务的配料表，并附有单位成本。

> "资源数据库 "类似于网上商店的产品目录，每个商品都有详细的属性说明。这使得成本估算人员更容易选择正确的资源（就像在购物篮中添加商品时选择产品一样），以计算（网上商店中的最终订单）的形式计算特定施工过程所需的资源。

资源数据库也可以看作是餐厅食谱中所有配料的清单。每种建筑材料、设备和服务都类似于食谱中使用的配料。"资源数据库 "是所有配料--建筑材料和服务--的详细清单，包括其单位成本：件、米、小时、升等。

新的实体元素可以通过两种方式添加到 "建筑资源数据库 "表中，一种是手动添加（图 5.1-3），另一种是通过与公司的库存管理系统或供应商数据库集成自动添加。

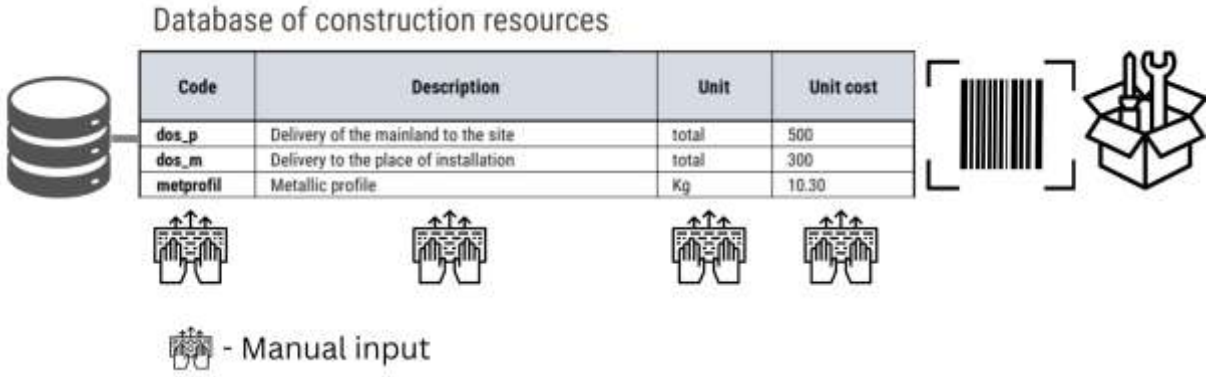


图 5.1-3 资源数据库由人工填写或自动采用其他数据库的数据。

一个典型的中型建筑公司使用一个数据库，其中包含数以千计，有时甚至数以万计的项目，并附有可用于建筑项目的详细说明。这些数据会自动用于合同和项目文件中，以准确描述工作和流程的构成

> 为了跟上通货膨胀等不断变化的市场条件，资源数据库（图 5.1-3）中每个产品（货物或服务）的 "单位成本 "属性都会定期手动更新，或从其他系统或在线平台自动下载当前价格。

根据资源的性质、通货膨胀和外部经济环境，可以每月、每季度或每年更新资源的单位成本。这种更新对于保持计算和估算的准确性是必要的，因为这些是作为成本估算人员工作起点的基本要素。最新数据用于生成估算、预算和进度表，以反映真实的市场情况，并降低后续项目计算出现错误的风险。

计算汇编 并根据资源基础计算工程成本

在 "建筑资源数据库"（图 5.1-3）中填入最小单位实体后，您就可以开始创建计算 ，这些计算是针对施工现场的每道工序或每项工作的特定计量单位进行的：例如，一立方米混凝土、一平方米石膏板墙、每米路缘石或每扇窗户的安装。

例如，根据以往项目的经验，要建造 1 平方米的砖墙（图 5.1-4），大约需要 65 块砖（实体 "硅酸盐砖"），每块成本 1 美元（属性 "每块成本"），共计 65 美元。此外，根据我的经验，需要使用建筑设备（实体 "JCB 3CX 装载机"）10 分钟，将砖块放置在工作区附近。由于租用设备的费用为每 小时 150 美元，使用 6 分钟的费用约为 15 美元。此外，还需要一名砌砖承包商 2 小时，每小时 30 美元，共计 60 美元。

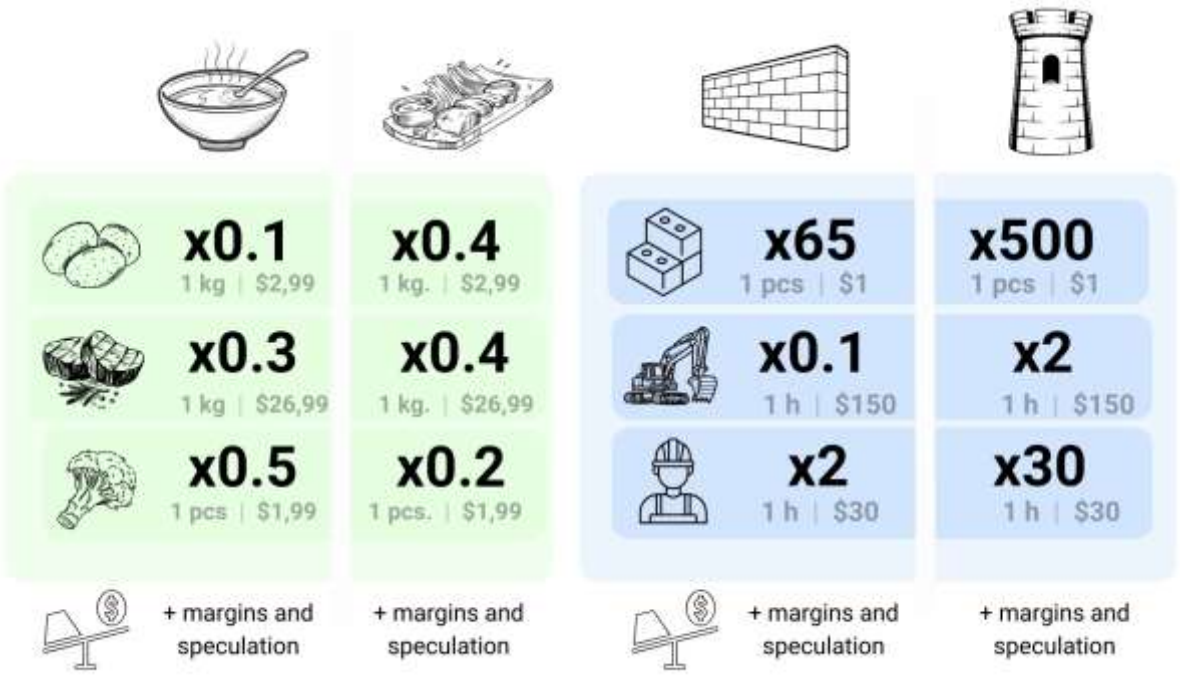


图 5.1-4 成本计算提供了施工和工序所需的建筑材料和服务的详细清单。

计算的组成（所谓的 "配方"）是根据公司在完成大量类似工程过程中积累的历史经验形成的。这些实践经验通常是通过施工现场的反馈积累起来的。特别是，工头直接在施工现场收集信息，记录实际的人工成本、材料消耗和技术操作的细微差别。然后与估算部门合作，反复完善这些信息：完善流程描述，调整资源组合，更新成本计算，以反映近期项目的实际数据。

> 就像菜谱描述了烹制菜肴所需的配料和数量一样，估算表也提供了完成特定工作或流程所需的所有建筑材料、资源和服务的详细清单。

定期进行的工作使工人、领班和估算人员能够确定所需资源的数量：材料、燃料、劳动时间和完成单位工作所需的其他参数（图 5.1-5）。这些数据以表格的形式输入估算系统，其中每项任务和操作都通过资源基础的最低要素（价格不断更新）进行描述，从而确保计算的准确性。

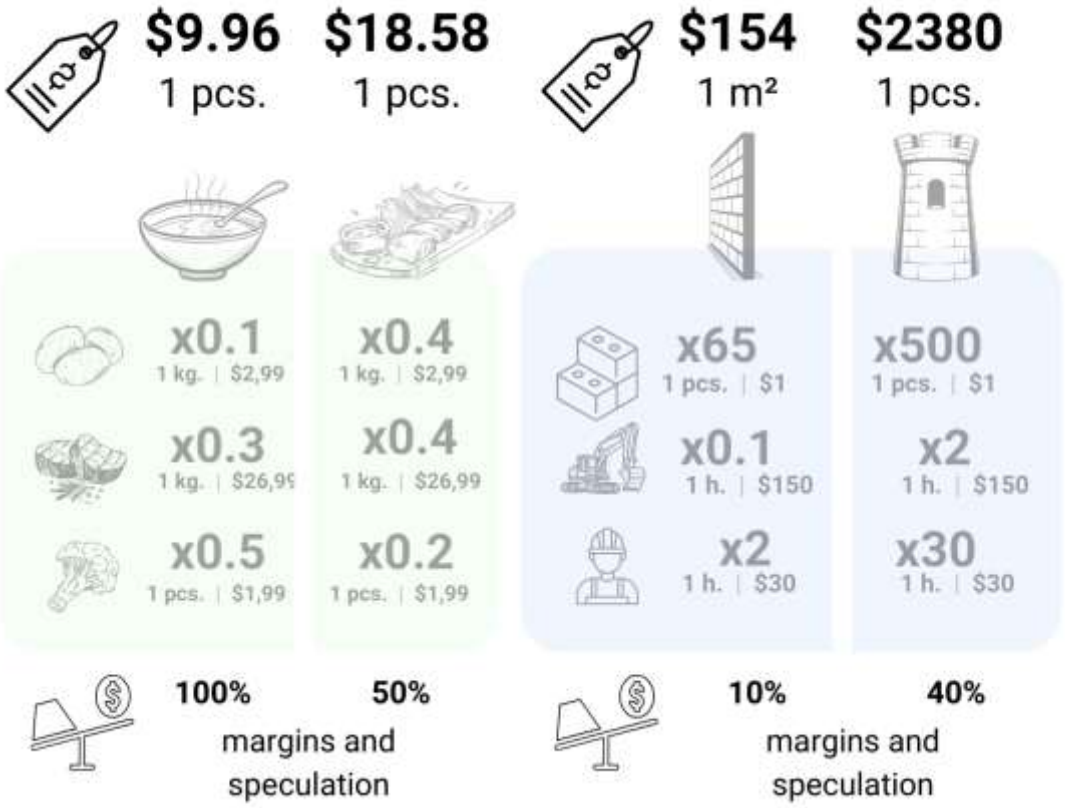


图 5.1-5 为每项工作收集单位费率，即实体的体积属性乘以数量，再加上利润百分比。

为了得出每个流程或活动（成本计算对象）的总成本，成本属性要乘以其数量和系数。系数可以考虑各种因素，如工作的复杂性、地区特点、通货膨胀率、潜在风险（预期间接费用百分比）或投机（额外利润系数）。

> 估算师作为分析师，将工头的经验和建议转化为标准化估算，通过表格形式的资源实体描述施工过程。从本质上讲，估算师的任务是通过参数和系数收集和整理来自施工现场的信息。

因此，单位工程（如平方米或立方米，或单个单元的一次安装）的最终成本不仅包括材料和人工的直接成本，还包括公司加价、间接费用、保险和其他因素（图 5.1-6）

同时，在（配方）计算中，我们不必再担心实际价格，因为实际价格总是反映在 "资源库"（配料表）中。在计算层面资源库中的数据会自动加载到表格中（例如，通过项目代码或其唯一标识符），加载说明和实际单

价，而实际单价又可以从在线平台或在线建材商店中自动加载。 工程计算层面的估算人员只需通过 "资源数量 "属性和附加因素来描述工程或流程。

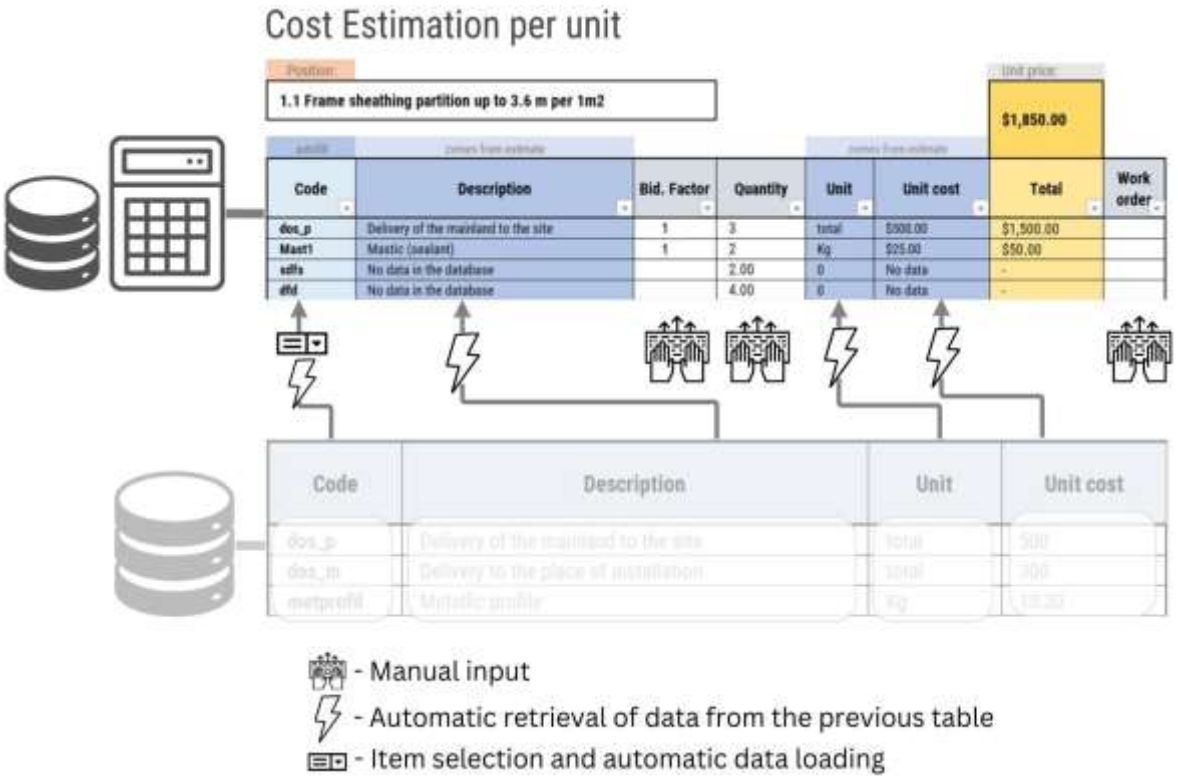


图 5.1-6 在计算单位工时成本阶段，只需填写所需资源数量的属性，其他内容均从资源数据库 自动加载。

创建的工作成本计算以典型项目模板表的形式存储，并与建筑资源和材料数据库直接相连。这些模板代表了未来项目中重复工种的标准化配方，确保了整个公司计算的统一性。

当数据库中任何资源的成本发生变化时（图 5.1-3）--无论是手动还是通过下载当前市场价格（如在通胀条件下）自动发生变化--更新都会立即反映在所有关联的成本计算中（图 5.1-6）。这意味着只需更改资源基础，而成本计算模板和估算则保持不变。这种方法确保了价格波动计算的稳定性和可重复性，而价格波动只在一个相对简单的资源表中进行计算（图 5.1-3）。

对于每个新项目，都会创建一个标准成本计算模板的副本，这样就可以在不改变公司采用的原始模板的情况下，根据具体要求对活动进行更改和调整。这种方法为调整计算提供了灵活性：您可以考虑施工现场的具体情况、客户的意愿、引入风险或盈利（投机）系数--所有这一切都不会违反公司的标准。这有助于公司在利润最大化、客户满意度和保持竞争力之间找到平衡。

在一些国家，经过几十年的积累，这些成本计算模板已在国家层面实现标准化，并成为国家建筑成本计算系统标准的一部分（图 5.1-7）。

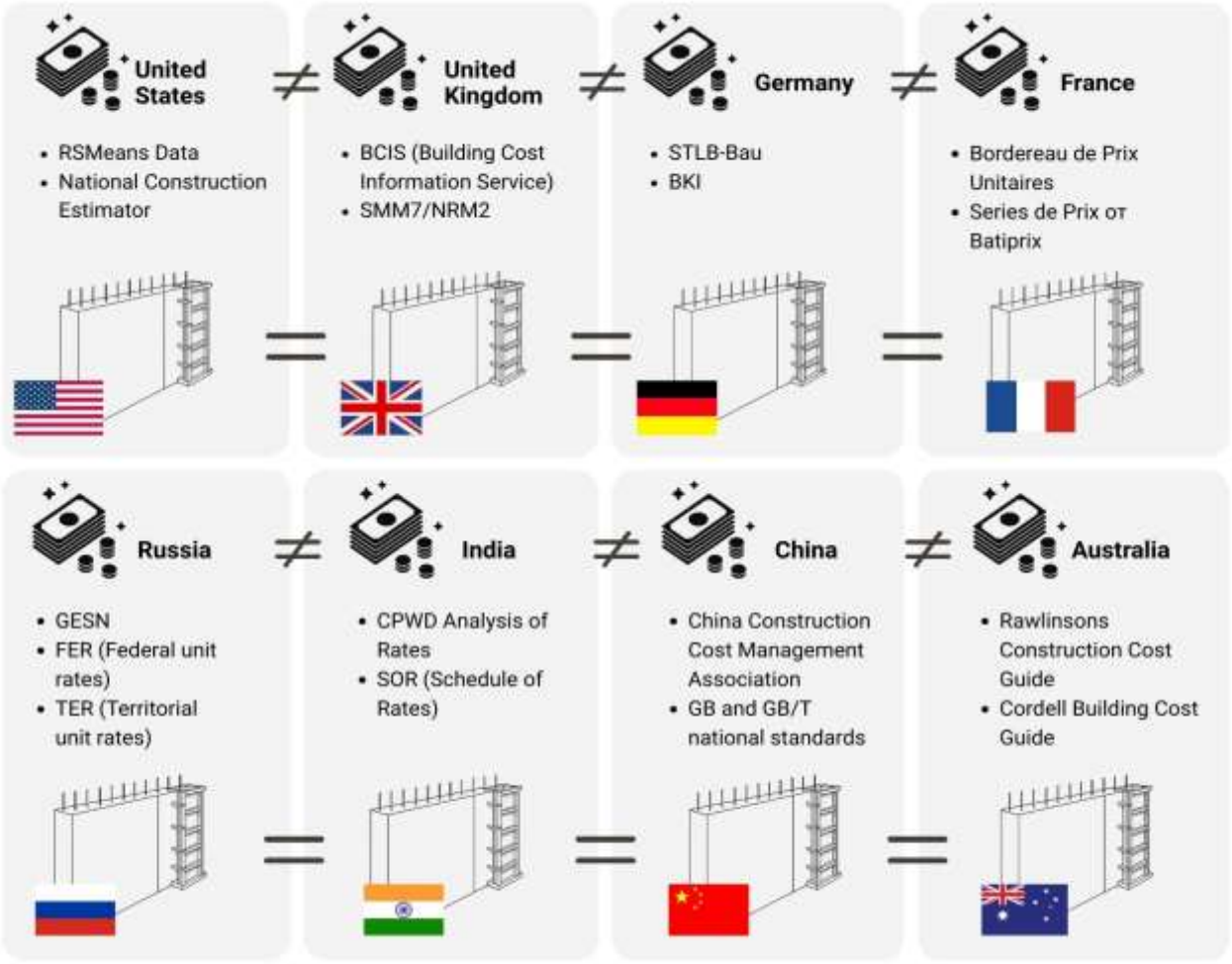


图 5.1-7 世界上不同的国家都有自己的成本计算规则，对同一要素的成本计算都有自己的（规范性）建筑工程汇编和标准。

这种标准化的资源估算基础（图 5.1-7）是所有市场参与者必须使用，尤其是公共融资项目。这种标准化可确保价格制定和客户合同义务的透明度、可比性和公平性

最终项目成本计算：从估算到预算

在不同国家的建筑实践中，国家和特定行业的估算标准发挥着不同的作用。一些国家要求严格遵守单一标准，而大多数发达经济体则采用更为灵活的方法。在市场经济国家，政府的建筑标准通常只是一个基准。建筑公司会根据自己的运营模式对这些标准进行调整或完全修改，并辅以自己的定制因素。这些调整反映了公司的经验、资源管理效率，通常还包括公司的投机利润等因素。

> 因此，竞争程度、市场需求、目标利润率，甚至与特定客户的关系，都可能导致与标准化规范的重大偏差。这种做法为市场提供了灵活性，但也使不同承包商的投标难以进行透明的比较，在计算过程的这一阶段为建筑业引入了投机定价的因素。

一旦准备好单项活动和工序的计算模板--或者更常见的情况是，直接从标准的政府估 算中复制（图 5.1-7），并添加系数以反映特定公司的 "特殊性"--最后一步是将每个项 目的成本乘以新项目中工作范围或工序的相应属性。

> 在计算新建筑项目的总成本时，关键步骤是汇总所有成本计算项目的成本，再乘以这些项目在项目中的工程量。

为了计算工程的总成本，在我们简化的例子中，我们首先计算建造一平方米墙壁的成本， 然后用计算成本（如 "1m² 墙壁构件标准安装 "工程）乘以工程中墙壁的总平方米数（如工程 CAD 中 "墙壁构件 "类型实体的 "面积 "或 "数量 "属性（图 5.1-8）或工长的计算结果）。

同样，我们计算项目所有要素的成本（图 5.1-8）：我们用单位工作成本乘以项目中某一要素或某一组要素的数量。估算人员只需将项目中这些要素、活动或工序的数量以体积或数量的形式输入。这样就可以自动生成完整的施工估算。

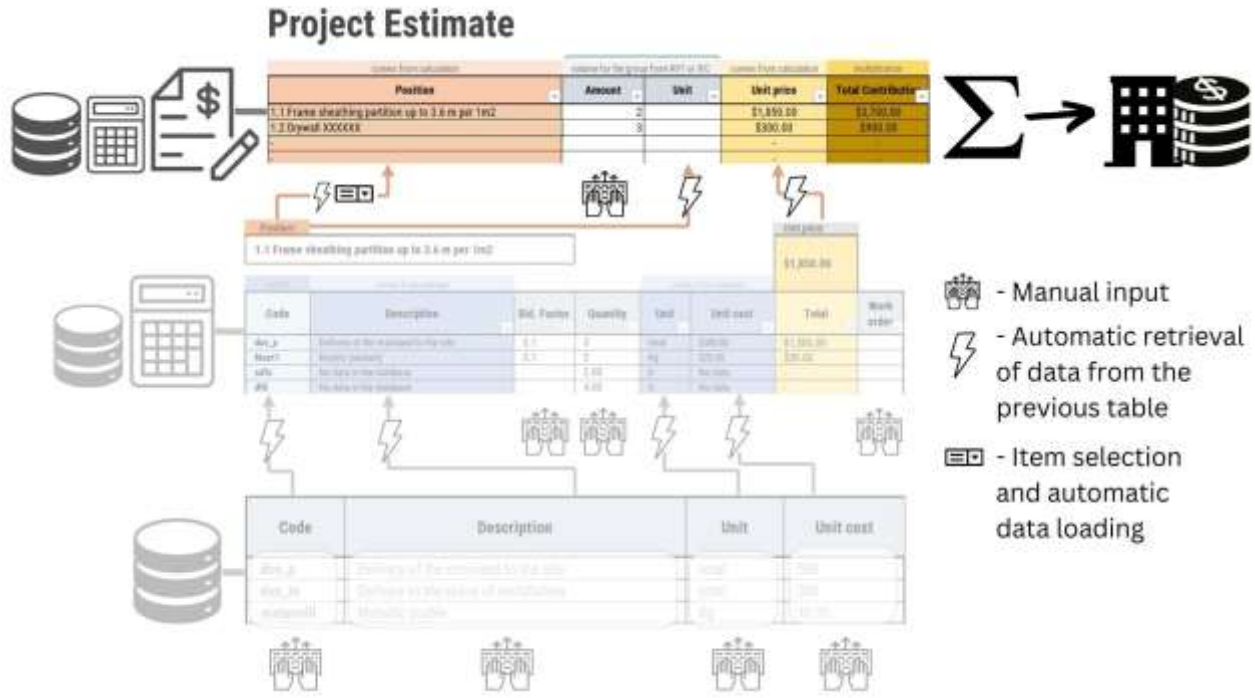


图 5.1-8 在创建估算阶段，我们只输入工作范围。

与计算一样，在这一级我们会自动上传准备就绪的计算项目（来自计算模板或从模板中复制并 编辑的新项目），这些项目会自动带来当前的单位工作成本（从资源数据库中自动更新（图 5.1-8 下表））。因此，如果资源数据库或成本计算表中的数据有任何变化，估算中的数据将自动更新为当 天的数据，而无需更改成本计算或估算本身。

在餐厅中，活动的最终成本也是以类似的方式计算的，等于整个晚宴的最终成本，即每道菜的成本乘以宾客人数加起来的总成本（图 5.1-9）。与建筑业一样，餐厅的烹饪食谱可能几十年都不会改变。与价格不同的是，配料的成本每小时都可能发生变化。

> 正如餐馆老板将每顿饭的成本乘以份数和人数来确定活动的总成本一样，成本估算经理也要将所有项目组成部分的成本相加，以得出完整的施工估算。

这样，项目中的每项活动的最终成本就确定了（图 5.1-9），再乘以与该活动相对应的实体的属性量，就得出了活动组的成本，并由此得出整个项目的最终成本。

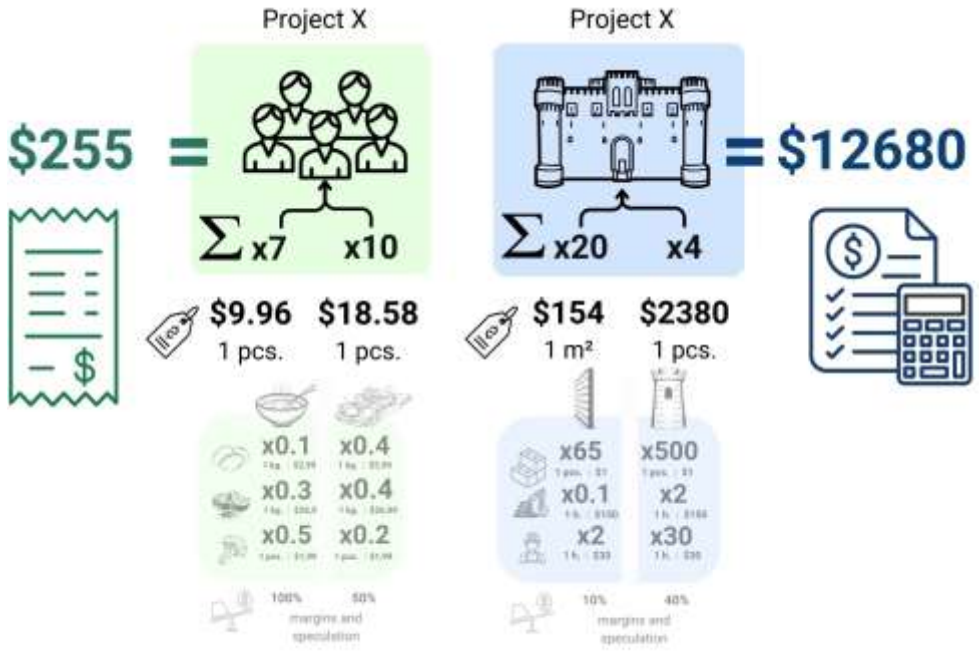


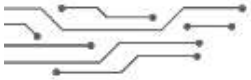
图 5.1-9 将每个元素的工作成本属性与其范围属性相加，计算出最终估算值。

项目总成本（图 5.1-8）提供了项目的财务状况，使客户、投资者或金融组织能够了解在任何给定的一天，考虑到当前的价格，交付项目所需的总预算和财务资源。

如果说编制资源库、计算和估算（工艺配方）的过程已经经过了数万年的半自动化和磨练，并在国家层面上记录在案，那么自动获取有关最终估算最后阶段的要素数量和质量的信息 ，如今仍然是所有项目成本和时间属性计算过程中的一个瓶颈，一般来说，也是项目总体预算的一个瓶颈。

> 几千年来，计算体积的传统方法一直是使用平面图纸手工测量体积和数量。随着数字化时代的到来，企业发现现在可以从 CAD 模型中包含的几何数据中自动提取体积和数量信息 ，彻底改变了几千年来获取定量数据的方式。

现代工艺估算和估价方法包括从 CAD 数据库中自动提取体积和数量属性，这些属性可以上载到 ，并与成本计算过程相连接，以便在从设计到运行的任何阶段获得最新的项目组体积。



章 5.2.

工程量核算以及自动创建估算和进度表

从三维到四维和五维：使用体积和定量参数

有了通过资源描述过程的成本计算表（图 5.1-8），下一步就是自动获取计算和最终估算所需的一组元素的体积或数量参数。

可从 CAD 数据库中自动提取工程构件（如墙壁或楼板）的体积特征。在 CAD 程序中创建的参数对象通过几何内核转换为长度、宽度、面积、体积和其他参数的数值。根据 3D 几何图形获取体积的过程将在下一部分，即第六部分（图 6.3-3）中详细讨论，该部分专门讨论如何使用 CAD (BIM)。除体积外，还可以通过按类别和属性对对象进行筛选和分组，从 CAD 模型数据库中获得相似元素的数量。这些可以进行分组的参数成为通过资源计算将项目元素与整个项目的计算、最终估算和 预算联系起来的基础。

因此，从三维（CAD）模型中提取的数据模型增加了新的参数层，即 4D 和 5D。在新的实体属性层 4D（时间）和 5D（成本）中，三维几何数据被用作实体体积属性值的来源。

- **4D** 是一个参数信息层，可在元素的 3D 参数中添加有关施工作业持续时间的信息。这些数据对于规划进度和管理项目时间表至关重要
- **5D** 是数据模型的下一级扩展，其中各要素都补充了成本特征。这为几何信息增加了财务方面的内容：材料、工程和设备的成本，从而可以在施工过程中进行预算计算、盈利分析和成本管理。

项目实体组的成本和 3D 、4D 和 5D 属性数据的描述方式与模块化 ERP、PIMS 系统（或类似 Excel 的工具）中的计算方式类似，用于单个组和整个项目预算的自动成本计算和预算规划。

5D 属性 ， 并从 **CAD** 获取属性体积

在编制建筑项目的最终估算 时（我们在前几章中已经讨论过该估算的编制工作）（图 5.1-8），每一类项目要素的体积属性 要么由人工收集，要么从 CAD 软件提供的体积属性规范中提取。

> 传统的人工计算工程量的方法包括工头和估算人员分析图纸，这些图纸几千年来一直以线条的形式出现在纸上，近 30 年来则以 PDF (PLT) 或 DWG 等数字格式出现。根据专业经验，他们通常使用直尺和量角器测量所需的工程量和材料。这种方法需要花费大量的精力和时间，还需要特别注意细节。

以这种方式确定范围属性可能需要几天到几个月的时间，具体取决于项目的范围。此外，由于所有的测量和计算都是人工完成的，因此存在人为失误的风险，可能导致数据不准确，进而影响项目时间和成本估算的错误，整个公司都将为此负责。

基于使用 CAD 数据库 的现代方法大大简化了体积的计算。在 CAD 模型中，元素的几何形状已经包含了体积属性，这些属性可以自动计算（通过几何形状内核（图 6.3-3）），并以表格形式显示或导出。

在这种情况下，估算部门会要求 CAD 设计人员提供项目元素的数量和体积特征数据。这些数据会以电子表格的形式导出，或直接集成到成本计算数据库（Excel、ERP 或 PMIS ）系统中。这一过程通常不是从正式请求开始，而是从客户（发起人）与建筑师施工或设计公司的估算人员之间的简短对话开始。下面是一个简化的示例，展示了如何通过日常交流形成自动计算结构表（QTO）：

- 客户--"我想在大楼上再加一层，结构与第一层相同"。
- 建筑师（CAD）--"增加第三层，配置与第二层相同"。然后将新的项目 CAD 版本发送给估算人员。
- 估算器自动进行分组和计算（ERP、PMIS 、Excel）--"我将通过带有 QTO 规则（ERP、PMIS）的 Excel 电子表格运行项目，按类别获得新楼层的工程量并生成估算"

因此，文本对话被转换为带有分组规则的表格结构：

要素	类别	楼层
重叠	OST_Floors	3
专栏	OST_StructuralColumns 结构列	3
楼梯	OST_Stairs	3

根据估算器的 QTO 规则对设计者提供的 CAD 模型进行自动分组，并通过资源计算自动乘以体积（图 5.1-8）后，我们得到以下结果，并将其发送给客户：

要素	卷数	楼层	单位价格	总成本
重叠	420 м²	3	150 €/м²	**63 000 €**
专栏	4 件	3	2450 欧元/件。	**9 800 €**
楼梯	2 件	3	4 300 欧元/件。	**8 600 €**
共计	-	-	-	**81 400 €**

- 客户--"谢谢你，数量太多了，我们需要减少几个房间"。如此循环往复。

这种情况可能会重复多次，尤其是在审批阶段，因为客户希望得到即时反馈。但实际上，这样的流程可能会拖上几天甚至几周。如今，由于引入了自动分组和计算规则，过去需要花费大量时间的工作现在只需几分钟即可完成。通过分组规则自动获取工程量，不仅加快了计算和估算速度，而且通过最大限度地减少人为因素，降低了出错的可能性，从而对项目成本进行透明、准确的评估。

如果最初在 CAD 系统中创建 3D 模型时就考虑到了估算部门的要求（这种情况在实践中还很少见），并且元素组的名称、标识符及其分类属性是以参数的形式设置的，与估算组和类的结构相吻合，那么体积属性就可以自动转移到估算系统中，而无需进行额外的转换。

从 CAD 中以规格表的形式自动提取体积属性，可以快速获得单项工程和整个项目的最新造价数据（图 5.2-1）。只需在计算过程或计算系统中更新包含工程量的 CAD 文件，公司就能根据最新变化快速重新计算估算，确保所有后续计算的高准确性和一致性。

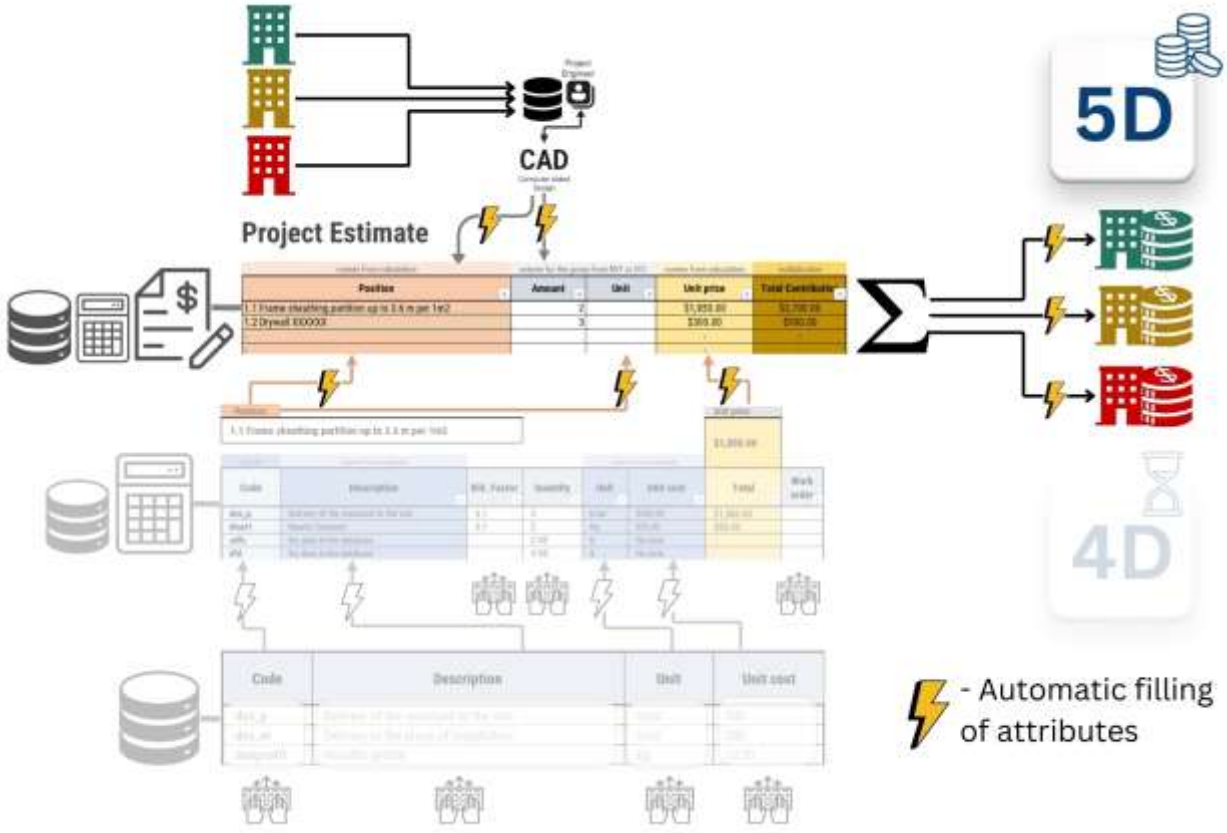


图 5.2-1 CAD 表格或数据库中的工程量属性 会自动输入到估算中，使您可以立即计算出项目总成本。

随着基本建设项目日益复杂，计算全额预算和分析 这种情况下的项目总成本（图 5.2-1）--已成为知情决策的重要工具。

> 根据埃森哲的《通过资本项目创造更多价值（2024 年）》研究[20]，领先企业正积极将数据分析融入数字化计划，利用历史信息预测和优化结果。研究表明，越来越多的业主-运营商正在应用大数据分析来预测市场趋势，并在设计开始前评估商业可行性。这是通过分析现有项目组合的数据仓库 实现的。此外，79% 的业主-运营商正在实施 "强大的 "预测分析，以评估项目绩效并支持实时运营决策。

现代建筑项目的有效管理与在设计的各个阶段以及设计之前的过程中处理 和分析大量信息密不可分。利用数据仓库、资源计算 、预测模型和机器学习 ，不仅可以最大限度地降低计算风险，还可以在设计的早期阶段做出项目融资的战略决策。我们将在本书的第九部分详细介绍数据仓库 和作为计算补充的预测模型。

从 CAD 项目 中自动获取估算所需的要素体积参数，是借助分组工具 QTO （工程量清单 ）进行的。QTO 工具利用 CAD 数据库中创建的规格和表格，通过特殊元素标识符或元素属性参 数对所有项目对象进行分组。

QTO 工程量清单： 按属性对项目数据进行分组

建筑工程中的 QTO（Quantity Take-Off，工程量计算）是提取实现项目所需要素数量的过程。在实践中，QTO 通常是一个半人工的过程，涉及从各种来源收集数据：PDF 文档、DWG 图纸和数字 CAD 模型。

在处理从 CAD 数据库提取的数据时，QTO 流程是通过一系列筛选、排序、分组和汇总操作来实现的。根据类、类别和类型参数选择模型元素，然后根据计算逻辑汇总其定量属性（如体积、面积、长度或数量）（图

5.2-2）。

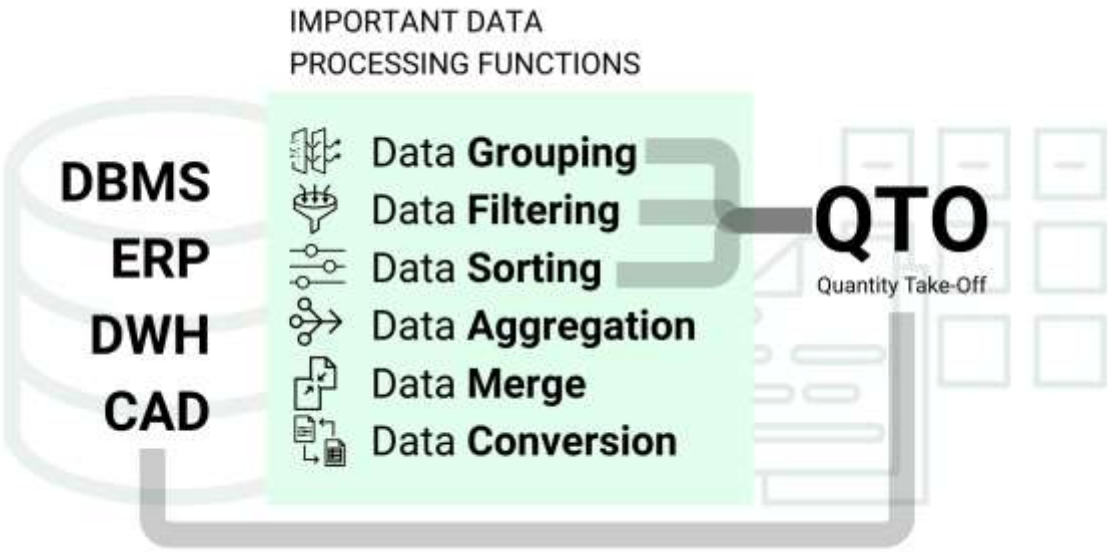


图 5.2-2 数据分组和过滤是数据库和数据仓库最常用的功能。

通过 QTO（过滤和分组）流程，可以将数据系统化，生成规格，并为计算估算、采购和工作计划准备输入信息。QTO 的基础是按测量属性类型对要素进行分类。对于每个要素或每组要素，都要选择相应的定量测量参数。例如

- 长度属性（路缘石 - 米）
- 面积属性（石膏板工程 - 以平方米为单位）
- 体积属性（混凝土工程 - 立方米）
- 数量属性（窗口 - 每件）

除了根据几何形状用数学方法生成的体积特性外，超限系数（图 5.2-12，例如：1.1，用于考虑 10%的物流和安装）通常应用于 QTO 分组后的计算中--考虑损耗、安装、储存或运输的修正值。这样就可以更准确地预测材料的实际消耗量，避免施工现场出现材料短缺或库存过多的情况。

> 自动数量核算流程 (QTO) 对于进行准确的 计算和估算、减少数量规格流程中的人为错误以及防止材料订购过多或过少至关重要。

作为 QTO 流程的一个示例，让我们考虑一个常见的情况，即需要从 CAD 数据库中显示某类、 某类元素的元素类型体积规格表。让我们将 CAD 项目墙类别 中的所有项目元素按类型分组，并汇总每种类型的体积属性，将结果作为 QTO 体积表 （图 5.2-3）。

以一个典型的 CAD 项目为例（图 5.2-3），CAD 数据库 中的所有墙体类别元素都按墙体类型分组，如 "Lamelle 11.5"、"MW 11.5 "和 "STB 20.0"，并有明确的体积属性，以公制立方体表示。

作为设计人员和计算专家之间的桥梁，管理人员的目标是自动获取所选类别中各元素类型的体积表。该表不仅适用于某一特定项目，还以通用形式适用于具有类似模型结构的其他项目。这使得该方法具有可扩展性，并可重复使用数据，避免重复劳动。

> 有经验的设计师 和估算师 拿着尺子，仔细测量纸张或 PDF -平面图上的每一条线的时代已经一去不复返了--这一传统在过去的几千年里从未改变过。随着三维 -建模技术的发展，现在每个元素的几何形状都与自动计算的体积属性直接相关，确定体积和 QTO 量的过程已经实现了自动化。

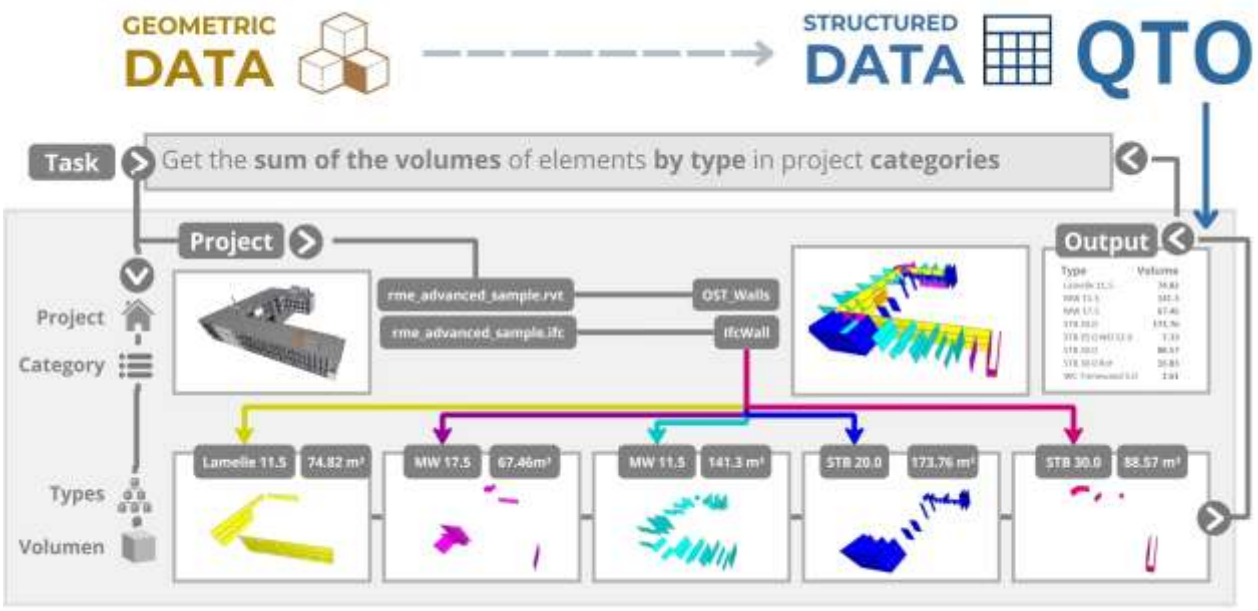


图 5.2-3 从项目中获取 QTO 体积和数量属性涉及对项目元素进行分组和过滤。

在我们的示例中，任务是 "选择项目中的一类墙壁，按类型对所有元素进行分组，并以结构化的表格形式呈现有关体积属性的信息"，以便数十名其他专家使用该表格计算成本、物流、工作进度和其他业务案例（图 6.1-3）。

由于 CAD 数据 的封闭性，如今并非每位专家都能直接访问 CAD 数据库（访问问题的原因和解决方案详见本书第六部分）。因此，许多人不得不求助于基于开放式 BIM 和封闭式 BIM [63]概念的专业 BIM 工具。在使用专门的 BIM 工具或直接在 CAD 软件环境中工作时，可以通过不同的方式生成带有 QTO（工程量清单）结果的表格，这取决于使用的是手动界面还是软件自动化。

例如，使用 CAD（BIM）软件的用户界面，只需执行约 17 个操作（点击按钮）就可以得到一个准备就绪的体量表（图 5.2-4）。不过，用户必须充分了解 CAD (BIM) 软件的模型结构和功能。

如果通过程序代码或 CAD 程序中的插件和 API 工具实现自动化，则可减少手动获取体积表的步骤，但需要编写 40 至 150 行代码，具体取决于所使用的库或工具：

- **IfcOpSh (**开放式 **BIM)** 或 **Dynamo IronPython (**封闭式 **BIM)** - 只需 ~40 行代码就能从 CAD 格式或 CAD 程序中获取 QTO 表。

- **IFC_js (**开放式 **BIM)** - 需要大约 150 行代码才能从 IFC 模型中提取大量属性。
- 界面 **CAD** 工具**(BIM)** - 让您只需点击 17 次鼠标，就能手动获得相同的效果。

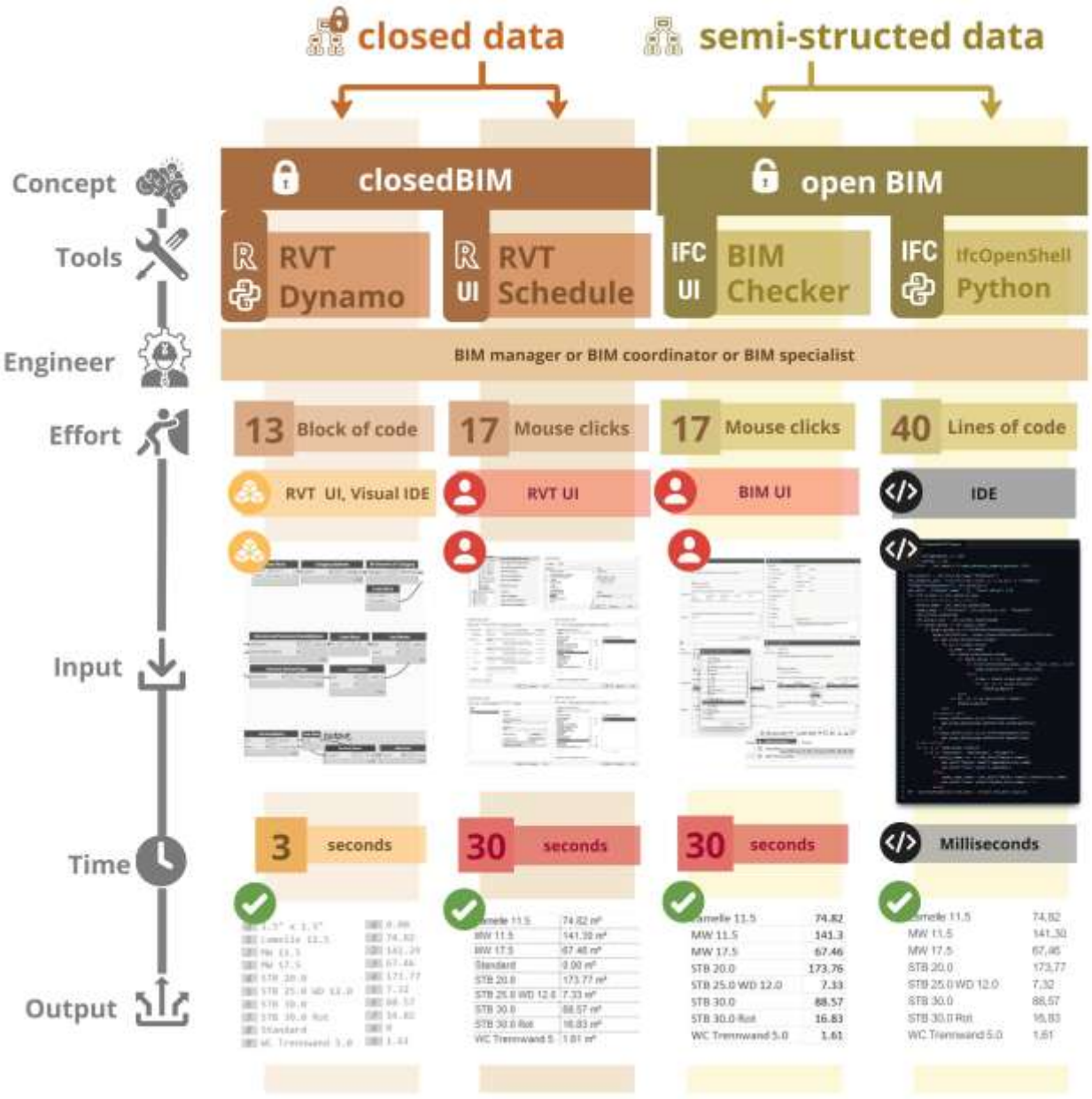


图 5.2-4 CAD (BIM)设计人员和管理人员使用 40 至 150 行代码或十几次按键来创建 QTO 表

结果是一样的--一个结构化表格，其中包含一组元素的数量属性。唯一不同的是人力成本和用户所需的专业

技术水平（图 5.2-4）。与人工收集数据量相比，现代工具大大加快了 QTO 流程，降低了出错的可能性。这些工具可直接从项目模型中提取数据，而无需像过去那样从图纸中手动重新计算工程量。

无论使用哪种方法--开放式 BIM 或封闭式 BIM--都有可能获得与项目元素体积相同的 QTO - 表（图 5.2-4）。然而，在使用 CAD - (BIM-) 概念中的项目数据时，用户依赖于供应商提供的专业工具和应用程序接口 （图 3.2-13）。这就产生了额外的依赖层，需要学习独特的数据模式，同时限制了对数据的直接访问。

由于 CAD 数据的封闭性，QTO 表和其他参数的获取使计算的自动化和与外部系统的集成变得 更加复杂。通过使用直接访问 数据库的工具，并通过逆向工程工具将项目的 CAD 数据转换为开放的结构化数据帧格式（图 4.1-13），只需一行代码即可获得相同的 QTO 表（图 5.2-5 - 带有粒度数据的变体）。

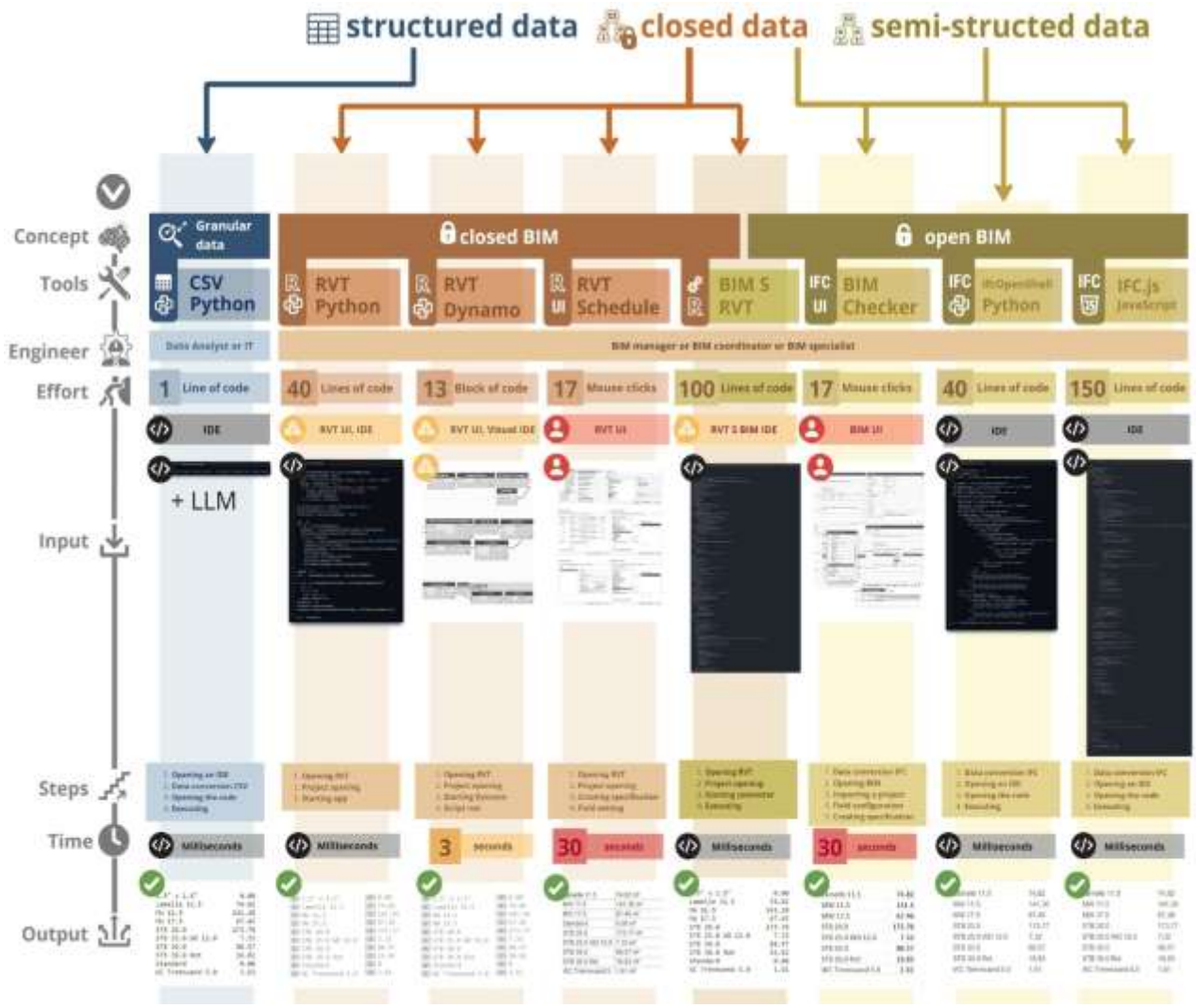


图 5.2-5 不同的工具以项目实体属性表的形式产生相同的结果，但劳动成本不同。

如 "将 CAD (BIM) 数据转换为结构化形式 "一章所述，在使用来自 CAD 项目的开放式结构化数据时，可大大简化分组过程 QTO。

基于使用开放式结构化数据或直接访问 CAD 模型数据库的方法摆脱了与 BIM 首字母缩写词 相关的营销限制。它们依赖于其他行业长期使用的成熟工具（图 7.3-10 ETL 流程）。

> 麦肯锡研究报告《开放数据：麦肯锡 2013 年进行的研究 "开放数据：利用流式信息释放创新和生产力"[102]显示[102]，在发电设施的设计、工程、采购和建设方面，使用开放数据每年可创造节省 300 亿至 500 亿美元的机会。这相当于节省 15% 的建设资本成本。

使用开放的结构化（粒度）数据可简化信息检索和处理，减少对专业 BIM 平台的依赖，并为自动化开辟道路，而无需使用专有系统或 CAD 格式的参数化复杂数据模型。

使用 **LLM** 和结构化数据的 **QTO** 自动化

将非结构化数据转换为结构化形式可显著提高各种流程的效率：简化数据处理（图 4.1-1，图 4.1-2），并通过使 要求清晰透明来加快验证流程，这一点我们在前几章中已经讨论过。同样，将 CAD 数据 (BIM) 转换为结构化的开放形式（图 4.1-12，图 4.1-13）也有助于属性分组过程和 QTO 过程。

QTO 属性表具有结构化形式，因此在使用结构化 CAD 数据时，我们使用的是单一数据模型（图 5.2-5），这样就不需要转换和翻译项目数据模型和分组规则，使之成为一个公分母。这样，我们只需一行代码就能按一个或多个属性对数据进行分组。相比之下，在开放式 BIM 和封闭式 BIM 中，数据以半结构化、参数化或封闭式格式存储，处理过程需要数十行甚至数百行代码，并使用 API 与几何和属性信息交互。

- 按一个属性对 QTO 结构项目进行分组的示例。在任何 LLM 聊天室（ChatGP、LlaMa 、Mistral 、DeepSeek 、Grok 、Claude 、QWEN 或其他）中进行文本查询：

> 我有 CAD -project 作为 DataFrame - 请过滤项目数据，以获取 "类型 "参数仅包含 "类型 1 值的项目。 ⏎

- LLM 的答案很有可能是使用 Pandas 的 Python 代码：



图 5.2-6 使用 LLM 编写的一行代码，可以按 "类型 "属性对整个 CAD 项目进行分组，并获得所需的元素组。

由于二维 DataFrame 结构简单，我们无需解释 LLM 模式和数据模型，这就缩短了解释步骤，加快了最终解决方案的创建。以前，即使是编写简单的代码也需要学习编程语言，但现在，现代语言模型（LLM）可以让我们在使用文本查询处理结构化数据时自动将流程逻辑转化为代码。

> LLM 自动化和语言模型通过提供使用文本查询解决问题的能力，使从事 CAD (BIM) 数据分组和处理的专业人员完全无需学习编程语言或 BIM 工具。

同样的查询--从 "墙壁 "类别中将所有项目元素分组并计算每种类型的体积（图 5.2-5）--在 CAD 环境(BIM)中需要在界面上点击 17 次或编写 40 行代码，而在开放式数据处理工具（如 SQL 或 Pandas）中看起来是一个简单直观的查询：

- 在 Pandas 中只需一行：

```
df[df['Category'].isin(['OST_Walls'])].groupby('Type')['Volume'].sum()
```

代码解码：从 df (DataFrame) 中提取属性列 Category "值为 "OST_Walls "的元素，按属性列 "Type "对所有获得的元素进行分组，并对获得的元素组的属性 "Volume "进行求和。

- 使用 SQL 对从 CAD 检索到的结构化项目进行分组：

```
SELECT Type, SUM(Volume) AS TotalVolume
FROM 元素
WHERE Category = 'OST_Walls' (类别 = 'OST_Walls')
GROUP BY Type;
```

- 在 LLM 的帮助下，我们可以将项目数据库的分组请求作为一个简单的文本参考--提示（图 5.2-7）：

> 对于项目数据框，按 "类型 "参数对项目进行分组，但仅限于 "类别 "参数等于 "OST_墙壁 "或 "OST_柱子 "的项目，并请对后的列参数 "体积 "进行汇总 ⏎

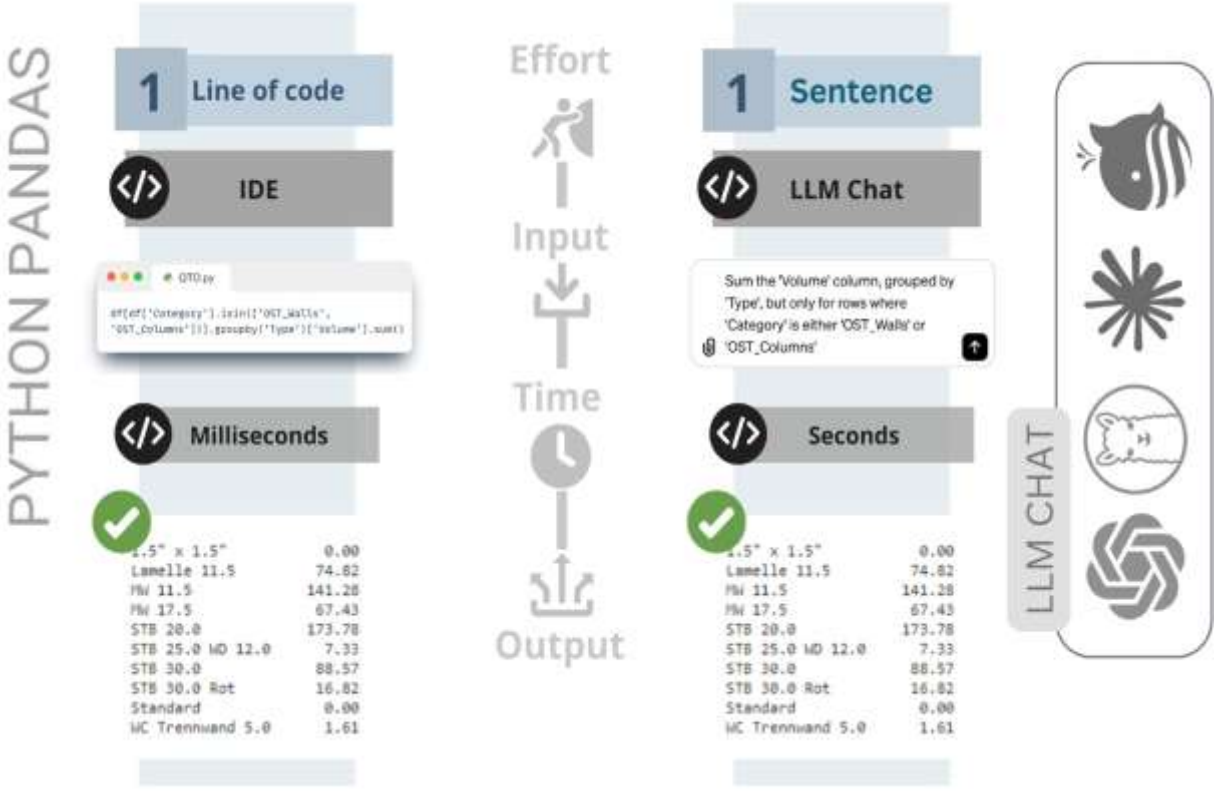


图 5.2-7 使用 SQL、Pandas 和 LLM 现在只需几行代码和文本查询就能实现数据处理自动化。

利用 LLM 工具（ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok ）从 CAD 数据 获取 QTO，极大地改变了提取单个物体和物体组的属性信息、定量和体积数据的传统方法。

> 现在，即使项目经理、成本核算或物流专家不具备深厚的设计知识，也没有专业的 CAD 软件 -(BIM-) 供应商，只要能访问 CAD 数据库，只需编写或口述一个查询，就能在几秒钟内获得墙壁或其他物体类别元素的总量。

在文本查询中（图 5.2-8），模型的 LLM 代理 处理用户的请求，将某个函数应用于一个或多个参数--表中的列。因此，与 LLM 通信的用户要么会收到带有新值的新列参数，要么会收到分组后的特定值

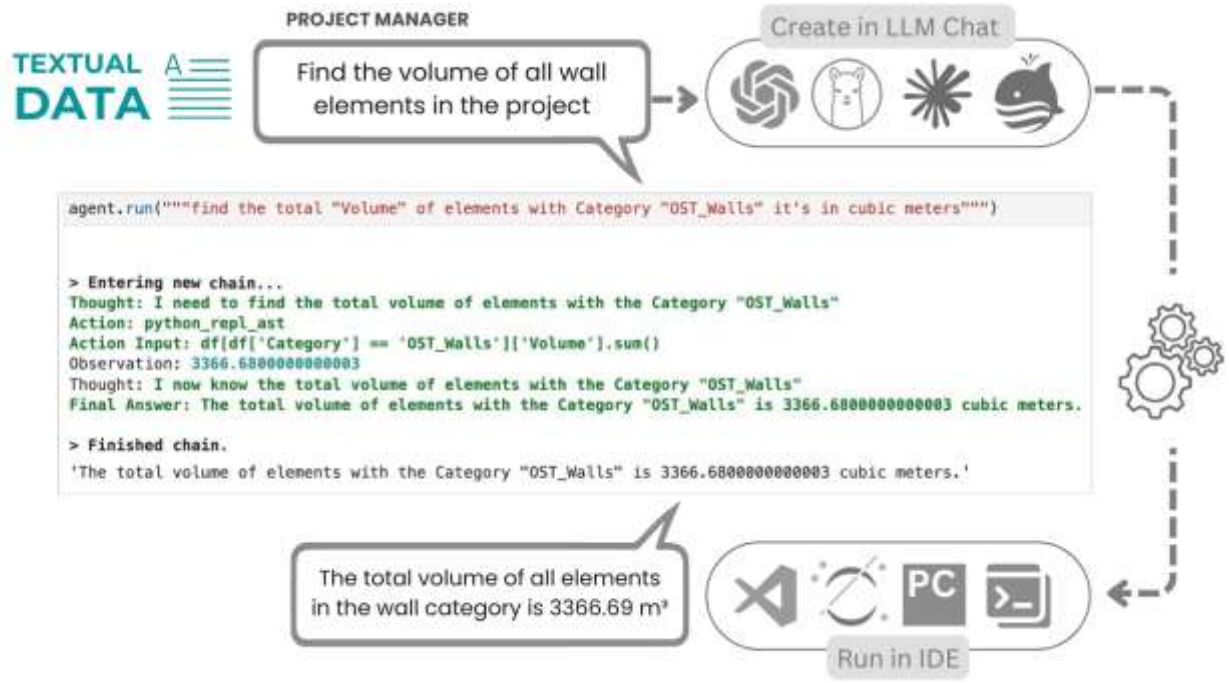


图 5.2-8 LLM 模型在处理结构化数据时，会根据文本查询的上下文理解用户询问的分组和属性。

如果只需获取一组要素的工程量，只需对 CAD 模型数据进行简单的 QTO 查询（图 5.2-7）即可。但是，在计算由多组元素组成的整个项目的预算或估算时，通常需要提取所有类型元素（类别 0）的工程量，其中每类元素都要单独处理--按相关属性分组。

在估算人员和评估人员的实践中，不同类型的对象使用不同的分组和计算规则。例如，窗户通常按楼层或区域分组（分组参数 - 属性 "楼层"、"房间"），墙壁按材料或建筑类型分组（参数 "材料"、"类型"）。为了实现分组过程的自动化，这些规则事先以分组规则表的形式在 中进行了描述。这些表格可作为配置模板，定义项目中每组元素在计算时应使用的属性。

QTO 利用 **Excel** 电子表格中的分组规则对整个项目进行计算

在实际建筑项目中，经常需要在一组元素中同时按多个属性进行聚合。例如，在处理 "Windows "类别（其中类别属性包含 OST_Windows 或 IfcWindows 等值）时，元素不仅可以按类型分组（例如，按类型名称或类型字段中的值分组），还可以按其他特征分组，例如按相应属性中指定的导热等级分组。这种多维分组方式可以为特定组别提供更准确的结果。同样，在计算墙壁或楼板类别时，属性的任意组合--如材料、层高、楼层、耐火等级和其他参数--可用作筛选器或分组标准（图 5.2-9）。

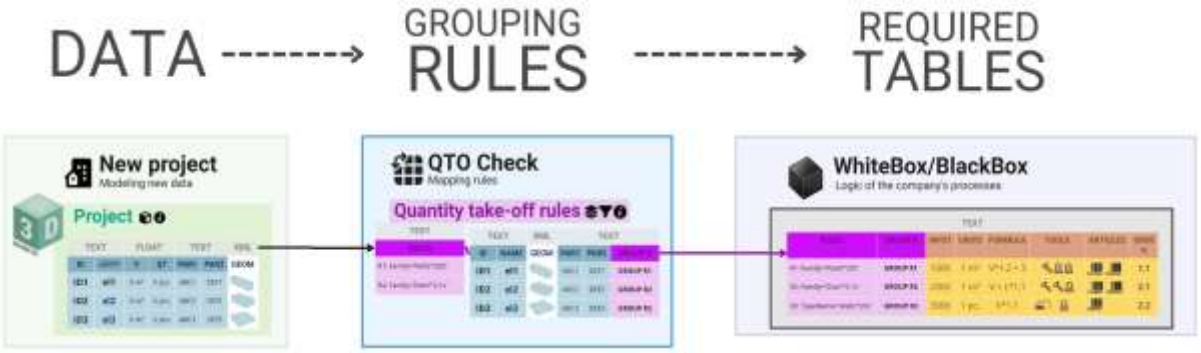


图 5.2-9 对于项目中的每一组或每一类实体，都有一个由一个或多个标准组成的不同分组公式。

定义这种分组规则的过程类似于 "创建需求 和质量检查 数据 "一章中描述的创建数据需求的过程（图 4.4-5），在该章中我们详细讨论了数据模型的工作。这种分组和计算规则可确保结果的准确性和相关性，以自动计算实体类别的数量或体积的总 属性，同时考虑到计算和计算中必须考虑的所有必要条件。

- 下面的代码示例过滤了项目表，使生成的数据集只包含 "类别 "属性列包含 "OST_Windows" 或 "IfcWindows "值，同时 "类型 "属性列包含 "类型 1 "值的实体：

> 我有一个 DataFrame 项目--过滤数据，使数据集中只保留属性 "Category "包含 "OST_Windows "或 "IfcWindows"，同时属性 "Type "包含 "Type 1 的项目。 ⏎

LLM 的答复：



图 5.2-10 类似于 Excel 公式 的一行代码，可让您按多个属性对所有项目实体进行分组。

将 CAD 数据 转换为结构化开放格式后生成的代码（图 5.2-10）（图 4.1-13）可以在我们上面提到的流行 IDE（集成开发环境）之一中以离线模式运行：PyCharm、Visual Studio Code (VS Code)、Jupyter Notebook 、Spyder、Atom、Sublime Text、带有 PyDev 插件的 Eclipse、Thonny、Wing IDE、带有 Python 插件的 IntelliJ IDEA 、JupyterLab 或流行的在线工具：Kaggle.com、Google Collab、Microsoft Azure Notebooks、Amazon SageMaker。

要检索 QTO DataFrame 表格 类别 "Windows "下仅具有特定热导率值的项目实体，我们可以使用下面的 LLM 查询：

> 我有一个 DataFrame 项目--过滤数据，使数据集中只保留 "Category "包含 "OST_Windows "或 "IfcWindows "值的记录，同时热导列的值应为 0.。 ⏎

LLM 的答复：



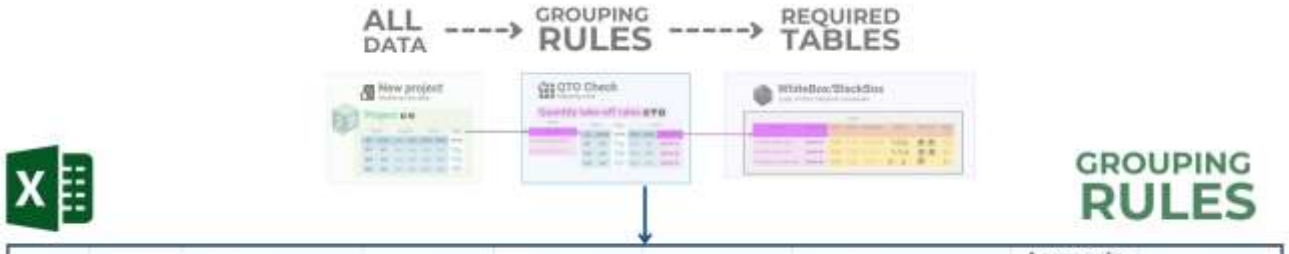
图 5.2-11 极其简单的 Pandas 查询语言 Python 允许您同时为任意数量的项目运行 QTO。

在从 LLM （图 5.2-11）收到的回复中，逻辑条件"&"用于合并两个标准：热导率值和属于两个类别之一。isin "方法检查属性列 "类别 "的值是否包含在所提供的列表中。

在有大量元素组的项目中，应为每一类项目实体（如：门窗、楼板）制定不同的组 合逻辑，其中可能包括附加系数或总属性计算公式。这些公式（图 5.2-12 属性 "formel"，如 x-数量值和 y-组群体积）和系数考虑了每个组群的独特性，如

- 考虑到超支，材料量增加百分比
- 定额补充材料
- 与可能的风险和公式计算错误有关的调整

一旦以参数公式的形式为每个项目类别制定了筛选和分组规则，就可以将其作为逐行表 格存储起来，例如 Excel 格式（图 5.2-12）。通过以结构化的形式存储这些规则，提取、过滤和分组项目数据的过程可以完全自 动化。系统只需读取参数表，并将相关规则应用于模型（整体项目数据框（图 4.1-13））， 为每类项目要素生成最终的 QTO 表格，而无需手动编写许多单独的查询。



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

图 5.2-12 QTO 属性分组表 建立了项目要素分组规则，为每个类别提供了准确的总数和范围。

通过收集的规则，您可以对整个项目进行分组，并执行所有必要的计算，包括调整体 积属性。结果是，体积被带入 "真实体积"，用于计算和计算，而不是最初在 CAD 模型 中设计阶段的体积。

> 在为整个项目自动创建 QTO 工程量表的过程中，应用程序必须查看分组规则表中的所有类别，提取分组属性，根据这些属性对所有项目元素进行分组，并汇总该组的工程量属性，再乘以一个精炼因子或系数。

让我们请 LLM 为我们编写这样一个解决方案的代码，代码必须加载两个表--分组规则表（图 5.2-12）和项目本身的数据表（图 4.1-13），然后应用分组规则，根据给定的规则对项目进行分组，计算汇总值，并将结果保存到一个新的 Excel 文件 中。

向法律硕士 聊天室发送文本请求：

> 我需要代码从文件 "basic_sample_project.xlsx "中读取项目数据，然后从 "Grouping_rules_QTO.xlsx "中读取规则，据此按 "参数 1 "和 "参数 2 "对所有数据进行分组，按 "聚合参数 "进行聚合，按 "表达式 2 "进行过滤，按 "表格 1 "进行计算，并将 QTO 表格保存到 "QTQ_table2.xlsx 中。 ⏎

LLM 的答复：

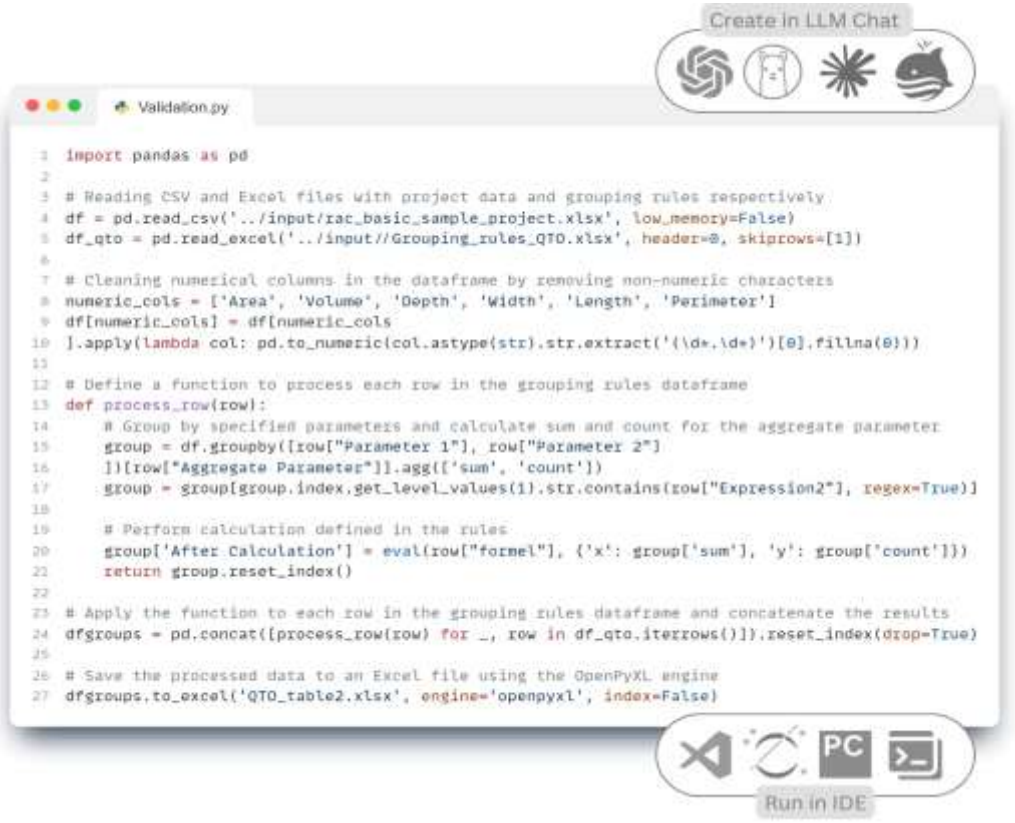


图 5.2-13 Python 代码 Pandas 处理 CAD 项目中的设计数据，并根据 Excel 中定义的规则进行分组。

代码执行的最终结果（图 5.2-13）将是一个实体组表，该表不仅包含原始 CAD- (BIM-) 模型中的通用体积属性，还包含一个新的实际体积属性，该属性考虑了 的所有要求，以便正确创建 计算和估算（示例图 5.2-14）。

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

图 5.2-14 执行自动计算实际体积的代码后，"计算后 "属性会添加到汇总表中。

生成的代码（图 5.2-13）可在流行集成开发环境中运行（如上所述），并应用于任何数量的现有项目或新项目（RVT 、IFC 、DWG 、NWS、DGN 等），无论是几个项目还是数百个不同格式的结构化项目（图 5.2-15）。

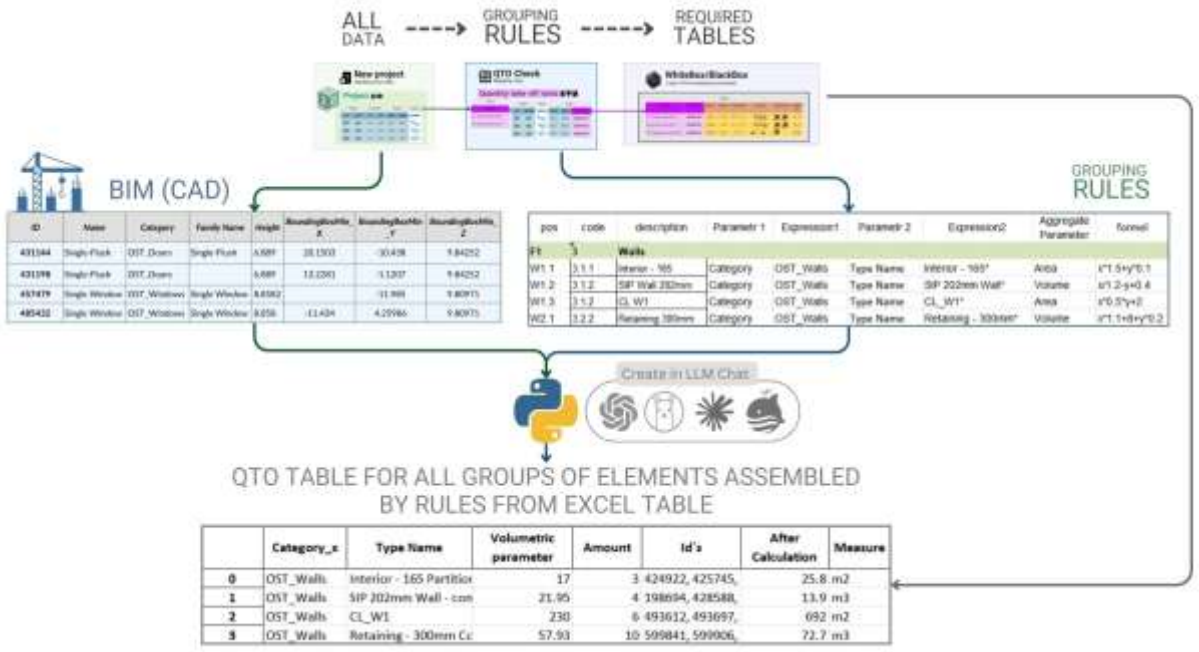
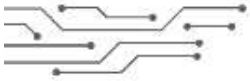


图 5.2-15 建筑数据自动分组过程通过 Excel 电子表格 中的规则将 BIM 数据 (CAD) 与 QTO 表格链接起来。

定制和参数化的体积数据收集流程（图 5.2-15）可全自动收集项目要素的定量属性和体积数据，以便进一步开展工作，包括成本估算、物流、工作进度、碳足迹计算和其他分析任务。

在掌握了可以根据特定属性轻松组织和分组项目元素的工具之后，我们现在就可以将分组和筛选后的项目与公司的各种计算和业务方案进行整合。



章 5.3.

4d, 6d -8d 和二氧化碳排放量的计算 **co_2**

4D 模型：将时间纳入施工估算

除成本计算外，设计数据在施工中的一个重要应用是确定时间参数，包括单个施工作业 和整个项目的时间参数。基于资源的估算方法和相关的计算数据库（在上一章 "建筑项目的计算和估算 "中进行了详细讨论）经常被用作自动计算时间和创建施工进度计划的基础。

基于资源的方法不仅考虑材料成本，还考虑时间资源。在成本计算中，可以为每个工序分配一个工单属性（图 5.3-1 - 工单参数）以及与执行该工序相关的时间和成本。这些参数对于描述那些没有固定市场价格且无法直接购买的作业尤为重要，例如建筑设备的使用、劳动力的雇佣或物流过程（通常以小时为单位）。在这种情况下，成本不是由采购部门决定，而是由实施公司根据内部规范或生产率直接决定 （图 5.3-1）。

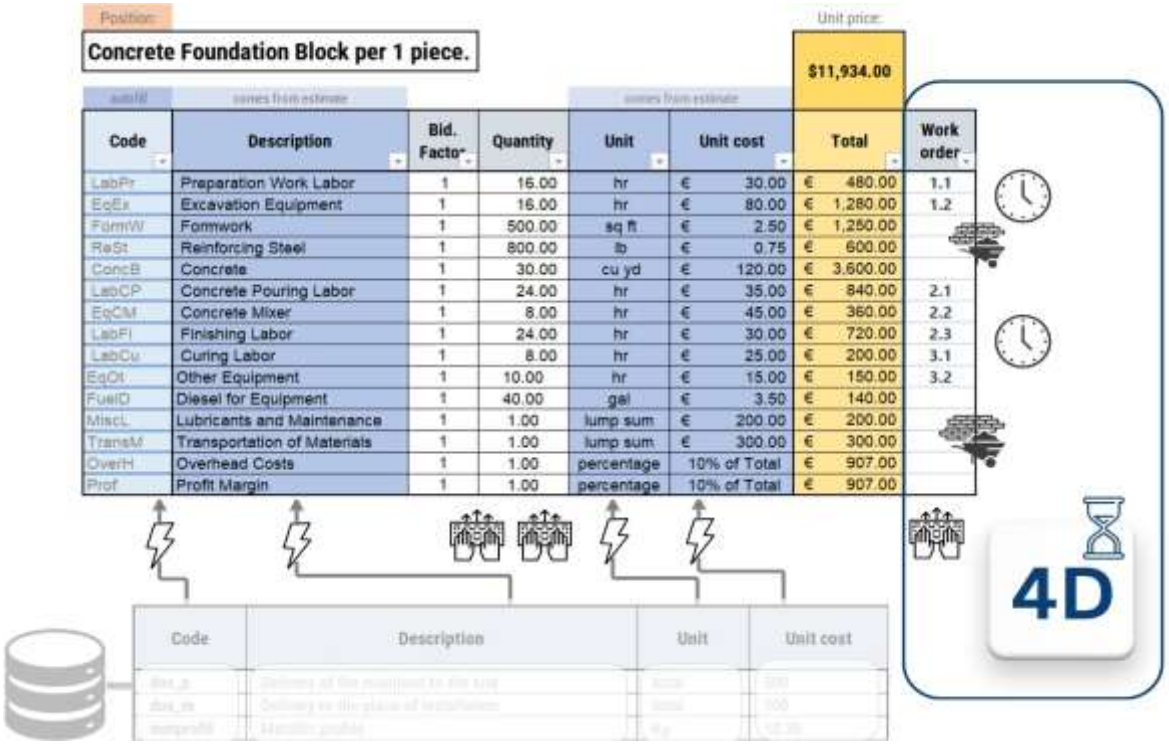


图 5.3-1 基于资源的估算方法中的工时计算包括工时计时。

因此，成本计算不仅包括燃料和材料成本（采购成本），还包括司机、技术人员和辅助工人在现场花费的时间。在所举的例子中（图 5.3-1），成本表是安装地基块的成本计算，包括工程的组成阶段，如准备工作、框架安装和混凝土浇筑，以及所需的材料和人工。不过，个别作业，如准备工作，可能没有材料成本，但可能包含大量以工时表示的临时人工成本。

为了规划施工现场的工作顺序（工作进度），需要在计算表中手动添加 "工作顺序 "属性 （图 5.3-1）。只有以时间（小时、天）为计量单位的项目才会在附加列中指定该属性。该属性是工作代码、描述、数量、计量单位（参数 "单位"）和成本之外的属性。通过活动的数字顺序（参数 "工作顺序"），可以确定施工现场的任务执行顺序，并用于排 程。

施工进度及其基于成本计算数据的自动化

施工进度计划是项目实施过程中工作计划和流程的直观体现。它是在详细的资源计算基础上创建的（图 5.3-1），其中除了资源成本外，还按时间和顺序安排了每项任务-工作。

在平均法中，时间估算是基于材料或设备安装的典型工时，而在基于资源的方法中，规划是基于成本计算中的实际数据。与人工成本有关的每个估算项目都是基于应用的日历，其中考虑到了工作期间资源使用的实际情况。通过成本计算中的系数（图 5.3-1 参数 "Bid.Factor"）对生产时间进行调整，可以考虑到生产率的差异和影响工作时间的季 节性特点。

为了确定甘特图中施工进度计划的开始和结束日期，我们从地基砌块成本计算中提取每个项目的时间属性值，然后乘以砌块数量（在本例中为混凝土地基砌块数量）。计算得出每项任务的工期。然后，我们将这些工期绘制在一个时间轴上，从项目开始日期开始，创建一个图表，其结果是一个直观的表示法，显示每项任务的开始和结束时间。此外，工序的 "工序 "参数还能让我们了解工作流程是并行运行（"工序"，如 1.1-1.1）还是顺序运行（1.1-1.2）。

> 甘特图 是一种用于规划和项目管理的图形工具，在时间轴上以水平条形图表示任务。每个横条代表任务的持续时间、开始和结束。

进度表或甘特图可帮助项目经理和工人清楚地了解施工各阶段的时间和顺序，从而确保有效利用资源和遵守

最后期限。

让我们利用上表中的计算结果，提出一个安装三块混凝土地基的时间表。利用上例中的成本表（图 5.3-1），请 LLM 安排这 3 块地基砖的安装时间，如 2024 年 5 月 1 日。

要向 LLM 发送成本计算，我们可以上传 XLSX 格式的成本计算表，或者直接将成本计算的 JPEG 图像截图插入 LLM 聊天室（图 5.3-1）。LLM 会独立找到一个库，将表格图像可视化，并通过将表格中活动的时间属性乘以活动量，将所有数据添加到图表中。

发送文本请求至 LLM：

> 使用时间相关表（附录中的 JPEG 图 5.3-1）中的值，以甘特图的形式创建一个安装 3 块地基的施工进度表。每个单元的工作应按顺序进行。工程开始时间应为 2024 年 5 月 1 日 ⏎

LLM 的答复

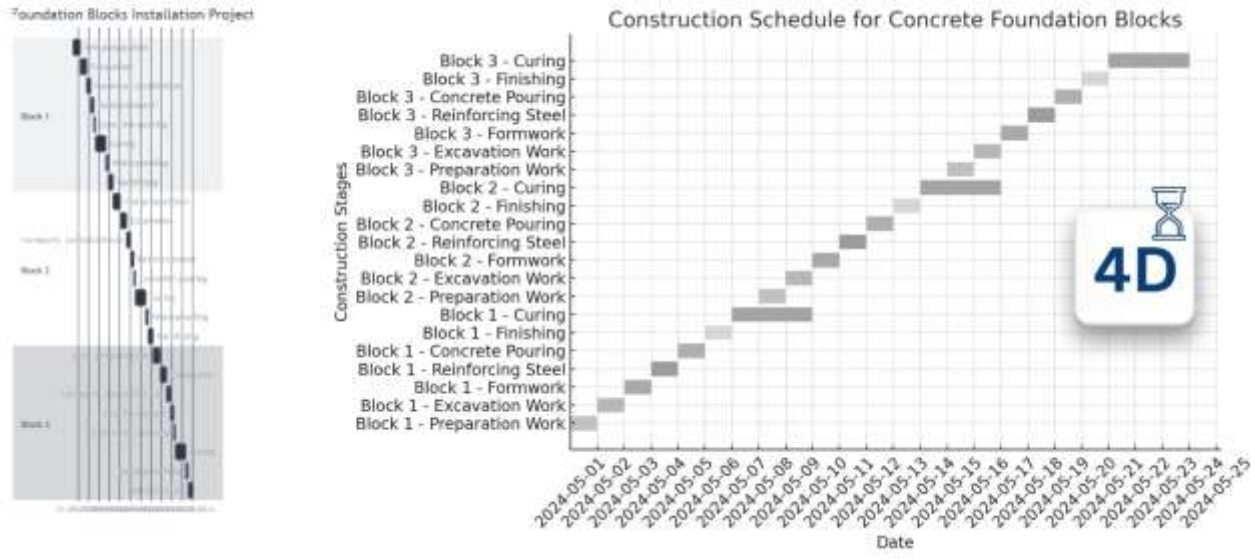


图 5.3-2 由多个 LLM 自动生成 甘特图显示了根据提示条件建造三个混凝土块的各个阶段。

由此得出的图表（图 5.3-2）是一个时间图，其中每个横条对应于地基工程的某一阶段，并显示了作业顺序（参数 "工序"），如准备、开挖、模板安装、加固、混凝土浇筑和收尾，，即在计算中填写了时间参数和

顺序的工序。

这样的时间表（图 5.3-2）并不考虑与工作日、轮班或工作时间标准相关的限制，而只是用于概念性的流程可视化。可以在聊天室中使用适当的提示或附加说明来补充反映工作并行性的精确时间表

使用单一成本计算（图 5.3-1），得益于三维 -geometry 的体积属性，可以通过自动 估算项目成本，同时以表格或图表的形式计算不同项目变体的组的时间特征（图 5.3-3）。

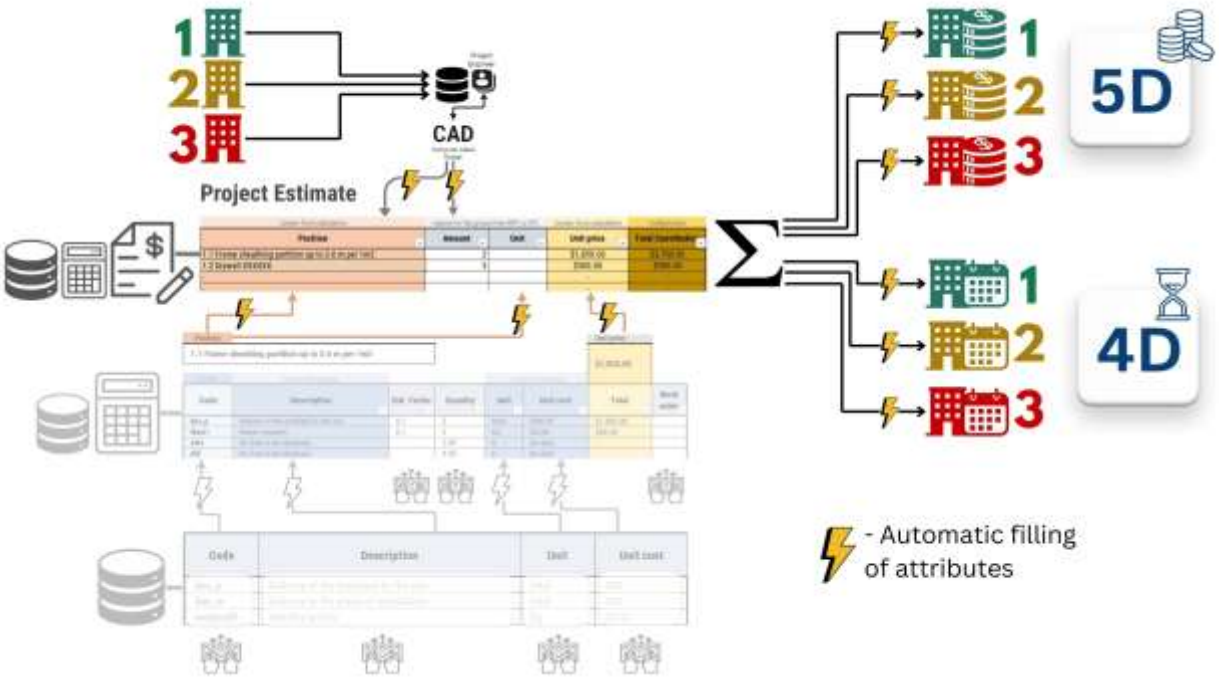


图 5.3-3 自动计算，可即时自动预测不同项目选项的成本和时间。

现代模块化 ERP 系统（图 5.4-4）从 CAD 模型加载数据，使用类似的自动时间计算方法，大大减少了决策过程。这样就可以即时准确地规划工作进度，并计算出完成项目实现过程中所有任务所需的总时间，同时考虑到实际价格。

扩展属性层 **6D -8D**：从能源效率到安全保证

6D 、7D 和 8D 是信息建模的扩展层次，每一层都为综合项目信息模型提供了额外的属性层，其基础 是三维模型的属性及其数量和范围。每个附加层都提供了在其他系统中进一步分组或进一步识别所需的特定参数，

例如财产管 理系统 (PMS)、计算机辅助设施管理 (CAFM)、施工项目管理 (CPM) 和安全管理系统 (SMS)。



图 5.3-4 数据信息模型中的属性 6D 、7D 和 8D 扩展了对项目各个方面的考虑，从能源效率到安全。

- 在 **6D** 中除了包含要素几何和体积属性的项目数据库（或数据框（图 4.1-13））外，还增加了环境可持续性信息（属性列）。这包括与能源效率、碳足迹、材料的可回收性以及环保技术的使用有关的信息。通过这些数据可以评估项目对环境的影响，优化项目决策，实现可持续发展目标（ESG）。
- **7D** 属性是对楼宇维护管理所需属性的补充。这些属性包括有关维护计划、组件生命周期、技术文档和维修历史的数据。这一系列信息可确保模型与维护系统（CAFM、AMS）集成，有效规划维护和设备更换，并在设施的整个生命周期内提供支持。
- **8D** 附加属性层--包括与安全有关的信息--无论是在施工阶段还是在随后的运行过程中。该模型包括人员安全措施、紧急指示、疏散系统和防火要求。将这些数据整合到数字模型中有助于提前考虑风险，并制定考虑到健康和安全要求的建筑、工程和组织解决方案。

在结构化表格形式中，4D 至 8D 层以列的形式表示附加属性，并在已填充的 3D -模型属性（如名称、类别、类型和体积特征）上添加已填充的值（图 5.3-5）。属性层 6D 、7D 和 8D 中的值包含附加的文字和数字数据，如回收百分比、碳足迹、保修期、更换周期、安装日期、安全协议等。



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

图 5.3-5 6D -8D 将属性层添加到数据信息模型中，该模型已包含来自三维 - 模型的几何和体积属性。

对于我们的新窗口（图 4.4-1），标识符为 W-NEW 的元素（图 5.3-5）可以具有以下 3D -8D 属性：

3D -attributes - 从 CAD 系统获取的几何信息：

- "类型名称"--元素 "窗口"
- "宽度" - 120 厘米
- 此外，您还可以将元素的 "边框 "点或其 "几何体 BREP / MESH "作为单独属性添加进来

6D - 环境的可持续性：

- 回收率达 90
- "碳足迹 - 1,622 千克 CO_2

属性 **7D** - 物件管理数据：

- "保修期" - 8 年
- "替换周期 "已有 20 年历史
- "维护"--每年需要

8D - 确保建筑物的安全使用和运行：

- 安装 "窗口 - 由 "XYZ Windows "公司提供
- "安全标准"--符合 ISO 45001 标准

数据库或数据集（图 5.3-5）中记录的所有参数都是不同部门的专家在分组、搜索或计算时所需要的。这种基于属性的多维度项目对象描述提供了项目对象生命周期、运行要求以及项目设计、建设和运行所需的许多其他方面的全貌。

估算二氧化碳排放量 并计算建筑项目的二氧化碳排放量

在第 6D 阶段 （图 5.3-5），除了建筑项目的可持续性这一主题外，现代建筑业还关注项目的环境可持续性，其中一个关键方面是评估和最大限度地减少项目生命周期各阶段（如制造和安装）产生的二氧化碳 CO_2 排放。

> 估算和计算建筑材料的碳排放量是一个过程，通过将项目中使用的一个或一组元素的体积属性乘以该类别的合适碳排放系数来确定碳排放总量。

在评估建筑项目时，将碳排放作为更广泛的 ESG 标准（环境、社会和治理）的一部分，为分析增加了新的复杂性。这对于客户-投资者获得相关认证尤为重要，如 LEED® （能源与环境设计先锋）、BREEAM® （建筑研究机构环境评估方法）或 DGNB® （德国建筑协会）。获得这些认证之一，可以大大提高房地产的市场竞争力，简化调试工作，并确保符合可持续发展导向型租户（ESG）的要求。根据项目要求，还可以使用 HQE（Haute Qualité Environnementale，法国绿色建筑标准）、WELL（WELL 建筑标准，关注用户健康和舒适度）和 GRESB（全球房地产可持续发展基准）

> 环境、社会和治理 **ESG** （环境、社会和治理）是一套广泛的原则，可用于评估企业对内对外的公司治理、社会和环境影响。

环境、社会和治理（ESG） ，最初是由金融基金在 2000 年代初开发的，旨在为投资者提供有关环境、社会和治理标准的广泛信息，现已发展成为评估公司和项目（包括建筑项目）的关键指标。根据主要咨询公司的

研究，环境、社会和治理（ESG）因素正在成为建筑行业不可分割的一部分。

> 根据安永会计师事务所（2023 年）的 "通往碳中和之路"，积极实施 ESG - 原则的公司不仅能降低长期风险，还能提高业务模式的效率，这在全球市场转型中尤为重要[103]。普华永道的《环境、社会和公司治理意识报告》指出，公司对环境、社会和公司治理 -因素重要性的认识从 67% 到 97% 不等，大多数组织认为这些趋势是未来可持续发展的关键[104]，而且大多数企业都看到了利益相关者要求整合环境、社会和公司治理原则的巨大压力。

> 因此，将 ESG 原则融入建筑项目，不仅有助于获得 LEED 、BREEAM、DGNB 等国际可持续发展认证，还能确保行业内企业的长期可持续性和竞争力。

影响建筑项目总体碳足迹的最重要因素之一是建筑材料和构件的生产和物流阶段。现场使用的材料通常会对二氧化碳的总排放量产生决定性影响 ，尤其是在项目生命周期的早期阶段--从原材料开采到运送到施工现场。

按建筑元素的类别或类型计算排放量需要使用参考碳排放系数，以反映不同材料生产过程中产生的二氧化碳量。这些材料包括混凝土、砖、再生钢、铝等。这些数值一般从知名来源和国际数据库中提取，如英国 ICE 2015（碳和能源清单）和美国 EPA 2006（美国环境保护局）[105]。下表（图 5.3-6）总结了一系列常见建筑材料的基准排放因子。为每种材料提供了两个关键参数：具体的二氧 化碳排放量（单位：千克/每千克材料）和体积重量 转换系数（单位：千克/每立方米），这对于将计算结 合到设计模型中并与 QTO 数据分组相联系是必 要的。



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

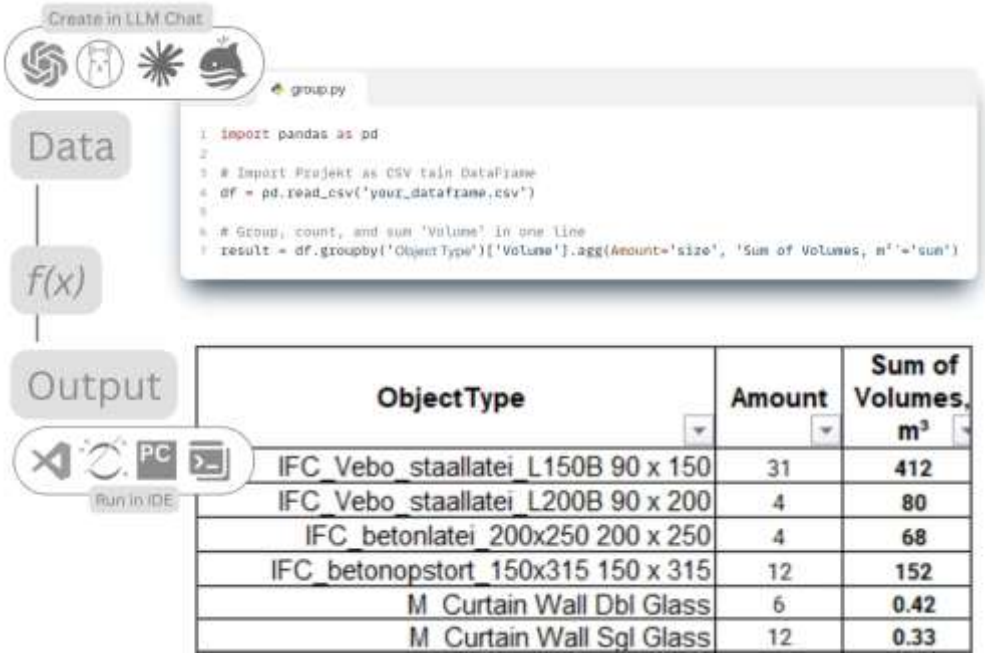
图 5.3-6 根据英国 ICE 和美国 EPA 数据库，不同建筑材料生产过程中的碳排放量。

与 4D 和 5D 计算一样，要计算一个项目的二氧化碳总排放量，必须确定每组设施的属性体积。这可以使用定量分析工具 (QTO) 来完成，方法是获得以立方米为单位的属性体积，详见 "数量推算 "一节。然后将这些体积乘以每个材料组的 "二 氧化碳过程排放 "属性的适当系数。

- 让我们按照前几章的方法，通过对所有项目数据进行分组，从 CAD (BIM) 项目中按元素类型自动提取体积表。要执行此任务，请参考 LLM。

> 请按列参数 "对象名称"（或 "类型"）对数据框表 from CAD (BIM) 项目进行分组，并显示每组中的项目数量，同时汇总该类型中所有项目的 "体积 "参数。 ⏎

LLM 的答复：



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

图 5.3-7 LLM 中生成的代码已按类型（ObjectType）为我们将项目实体分组，并汇总了属性 "体积"。

要自动计算整个项目的 CO_2 排放总量，只需在表格中设置自动数据映射，或手动将元素类型（图 5.3-7）与排放系数表中相应的材料类型（图 5.3-6）相连接。在 GitHub 上搜索 "CO_2_calculating-the-embodied-carbon.DataDrivenConstruction." [106]。[106].

因此，在将 CAD 数据库中的 QTO 元素分组后进行数据整合 ，可自动计算不同设计方案的二氧化碳排放量（图 5.3-8）。这样就可以分析不同材料在不同变体中的影响，并只选择符合客户二氧化碳排放要求解决方案 ，以便在建筑投入使用时获得特定证书

> 在建筑公司为设施获得 ESG 评级 （如 LEED 认证 ）的过程中，通过将系数乘以分组项目元素的体积来估算二氧化碳排放量 是一项典型的任务。

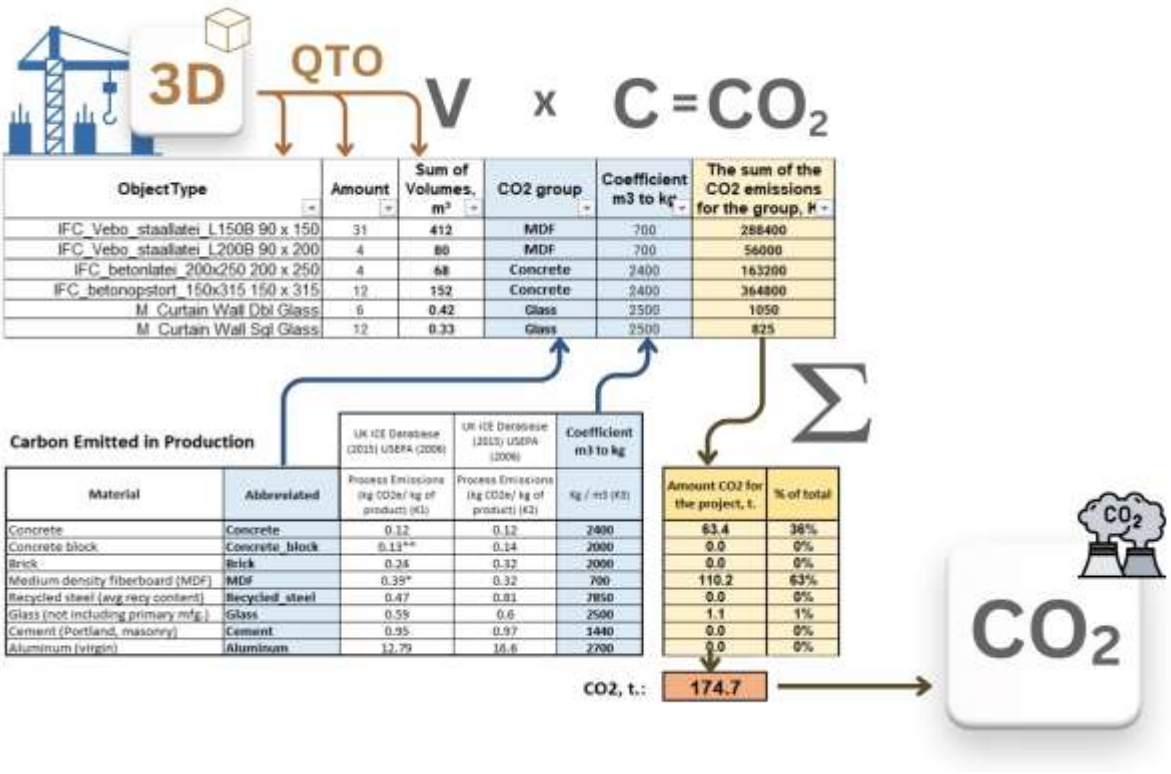


图 5.3-8 整合 CAD 数据库中的 QTO 组 在得出最终 CO_2 排放量估计值方面提供了准确性和自动化。

同样，通过定义元素组体积，我们可以进行材料控制和物流、质量监控和管理、能源建模和分析等一系列计算，从而获得单个元素组和整个项目的新属性状态（表中参数）。

如果公司中此类计算流程的数量开始增加，就会产生是否需要将此类计算自动化并将计算结果纳入公司流程和数据管理系统的问题。

由于完整解决方案的复杂性，建筑行业的大中型公司将这种自动化外包给 ERP （或 PMIS ）系统开发公司。开发公司为大型客户创建一个单一的综合模块化系统，以管理许多不同的信息层，包括材料和资源计算。

章 5.4.
建筑 **erp** 和 **pmis** 系统

建筑企业资源规划系统 **-** 计算和估算实例系统

模块化 ERP 系统将各种属性（信息）层和数据流整合到一个综合系统中，使项目经理能够在一个平台上同步管理项目的资源、财务、物流和其他方面。建筑企业资源规划系统可作为建筑项目的 "大脑"，通过自动化简化重复流程，在整个建筑过程中提供透明度和控制。

> 建筑企业资源规划 系统（企业资源规划）是一种综合性软件解决方案，旨在管理和优化建筑流程的各个方面。建筑企业资源规划系统的核心是管理成本计算和进度安排的模块，使其成为高效资源规划的重要工具。

企业资源规划模块 - 系统允许用户以结构化的方式输入、处理和分析数据，涵盖项目的各个方面，其中可能包括材料和人工成本核算、设备利用、物流管理、人力资源、联系人和其他建筑活动。

系统的功能模块之一是业务逻辑自动化模块--BlackBox/WhiteBox，它扮演着流程控制中心的角色。

BlackBox /WhiteBox 允许使用企业资源规划 系统的专家通过访问权限灵活管理其他用户或管理员已预先配置的业务的各个方面。在 ERP 系统中，BlackBox 和 WhiteBox 指的是系统内部逻辑的透明度和可控性：

- **BlackBox** （"黑盒子"）--用户通过界面与系统交互，无法进入流程执行的内部逻辑。系统根据预定义的规则自行执行计算，对最终用户保密。用户输入数据并得到结果，而不知道系统内部使用了哪些属性或系数。
- **WhiteBox** （"白盒"）--流程逻辑可供查看、定制和修改。高级用户、管理员或集成商可以手动定义项目实体之间的数据处理算法、计算规则和交互方案。

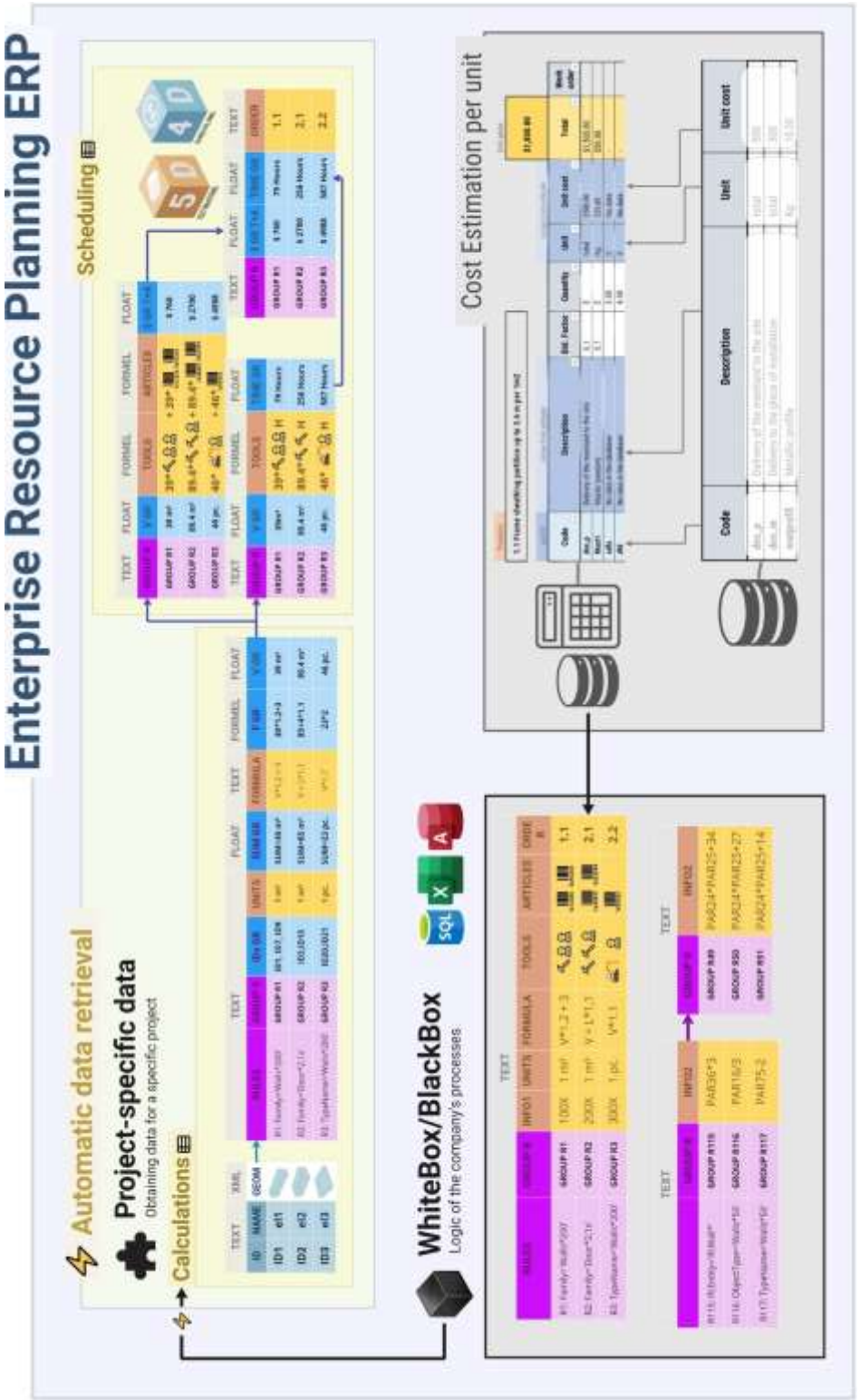

Enterprise Resource Planning ERP
Automatic data retrieval
Project-specific data
Obtaining data for a specific project
Calculations
Scheduling
WhiteBox/BlackBox
Logic of the company's processes
Cost Estimation per unit
Code
Description
Unit
Unit cost

图 5.4-1 建筑业企业资源规划 系统的结构，通过手动填写工程量属性来获取估算和工程进度计划

。

例如，经验丰富的用户或管理员可设定规则：估算中的哪些属性应相互相乘，或按某一特征分组，以及最终结果应记录在何处。之后，估算工程师等训练有素较低的专业人员只需通过用户界面将新数据载入企业资源规划系统，即可获得估算、进度表或规格，而无需编写代码或了解逻辑的技术细节。

> 在前几章中，我们结合 LLM 交互讨论了计算和逻辑模块。在企业资源规划环境中，此类计算和转换是在隐藏在按钮和表单界面后面的模块中进行的。

在以下示例中（图 5.4-1），ERP 系统管理员在 BlackBox /WhiteBox 模块中定义了估算实体属性与 QTO 分组属性的匹配规则。由于（由经理或管理员）配置了这一 BlackBox/WhiteBox 模块，用户（估算师或工程师）通过 ERP 用户界面手动添加数量或体积属性后，会自动收到完成的 估算和工作计划。这样，前几章中讨论的计算和估算生成过程在企业资源规划系统内部代码的帮助下，变成了半自动化的输送机。

例如，通过将 CAD 项目加载到为此目的预先配置的 ERP 模块中 ，将这一半自动化流程与 CAD (BIM) 模型（图 4.1-13）的体积属性连接起来，将数据流转化为一种同步机制，能够在将 CAD 模型加载到 ERP 时，根据设计阶段的任何变化，自主、即时地更新单个元素组或整个项目的值。

为了在 CAD (BIM) 和 ERP 系统之间创建自动数据流（图 5.4-2），必须以结构化的方式定义 CAD (BIM) 模型数据库数据的基本流程和要求 ，这在上一章 "要求和数据质量保证 "中已经讨论过。企业资源规划中的这一过程也分为类似的步骤：

- 创建验证规则 **(1)**，在确保进入企业资源规划 系统的数据准确性方面发挥重要作用。验证规则可作为过滤器，对实体及其属性进行验证，只允许通过要求的项目进入系统。有关验证和确认的更多信息，请参阅 "创建 要求并确认 数据质量 "一章。
- 然后，ERP 内部会进行验证流程 **(2)**，确认所有项目实体元素及其属性和值都已正确创建，并为下一步处理步骤做好准备。
- 如果存在属性数据不完整的问题，就会生成一份报告 **(3)**，并将项目连同更正说明一并送交修改，直至准备好进行下一次迭代。

- 项目数据经过验证和核实后，将用于另一个企业资源规划模块 **(4)**，以根据先前生成的规则（WhiteBox/BlackBox）为实体组、材料和资源创建数量属性的数量起飞表（**QTO** ）。
- 通过匹配规则或 QTO 将分组数据与计算（如成本和时间）自动整合 **(5)**。
- 在 ERP 系统的最后一步，用户通过将 QTO 表 中的范围属性与流程表（如估算项目） 的属性相乘，自动生成每个实体组和整个项目的计算结果 **(6)**（如成本 估算、工作时间表或二氧化碳排放量）。

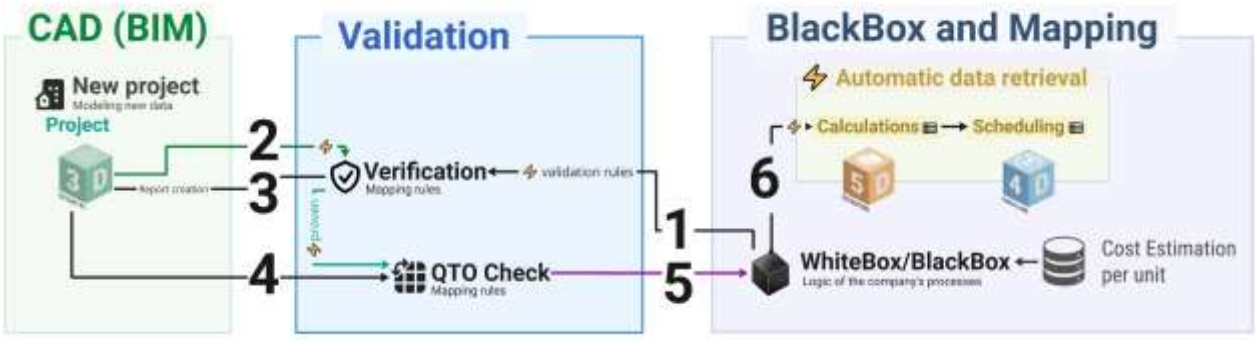


图 5.4-2 带有 CAD (BIM) 的建筑企业资源规划 系统结构，从创建验证规则 (1) 到自动计算成本和工作计划 (5-6)。

在模块化企业资源规划 系统中，流程是通过包括用户界面的软件进行整合的。界面后面是后端，结构化表格通过执行经理或管理员预先配置的各种操作来处理数据。因此，用户通过预先定义和定制的自动化逻辑（在 BlackBox /WhiteBox 模块中），可以收到半自动编制的文件，以完成其任务

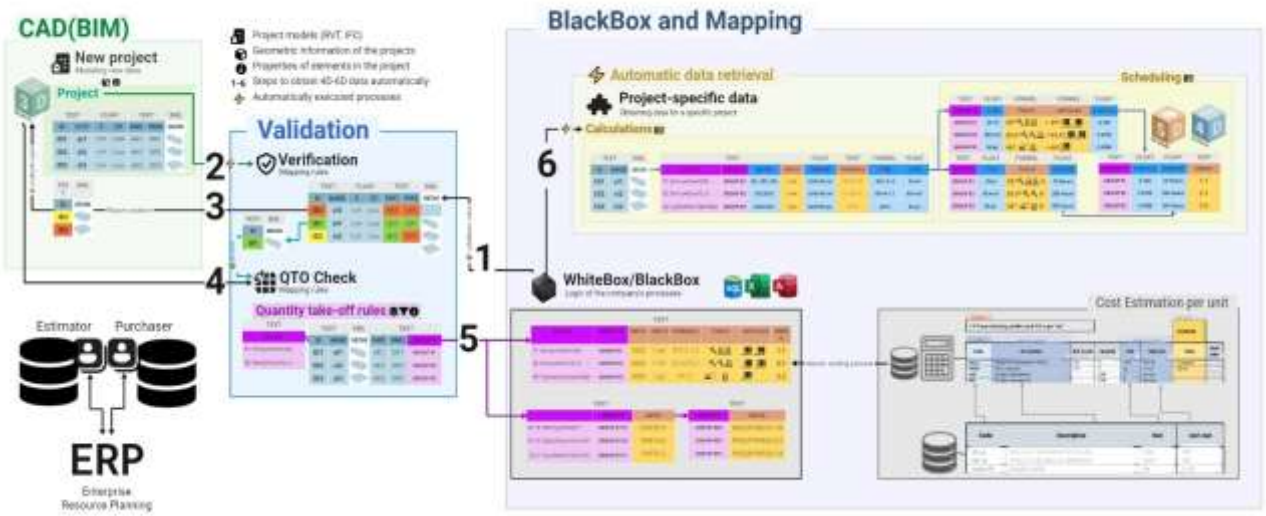


图 5.4-3 ERP - 系统帮助管理人员和用户在专业表之间移动，生成新数据。

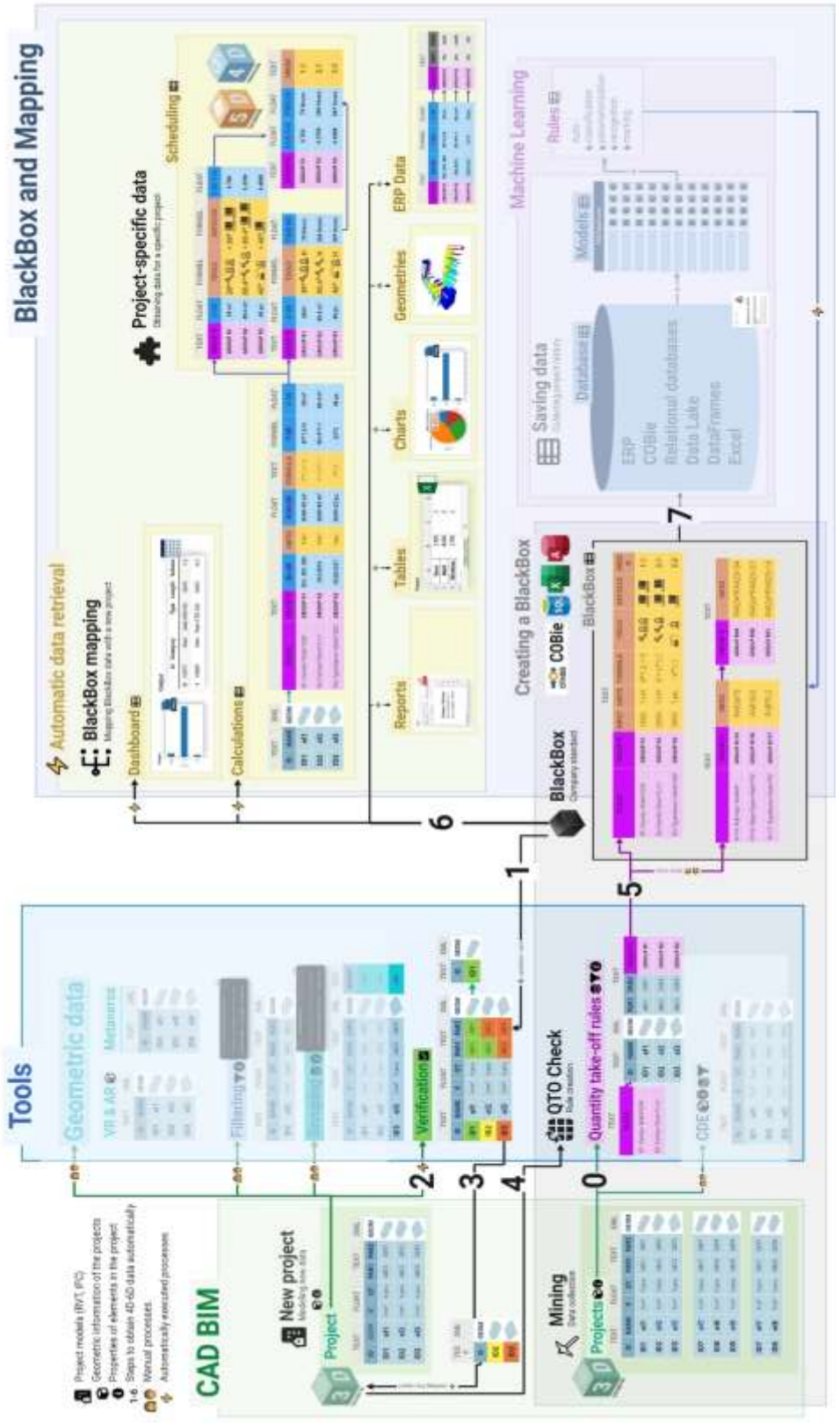

BlackBox and Mapping
Automatic data retrieval
BlackBox mapping
Project-specific data
Scheduling
Dashboard
Calculations
Tables
Charts
Reports
Geometries
ERP Data
Machine Learning
Rules
Models
Saving data
Database
ERP
COBie
Relational databases
Data Lake
DataFrames
Excel
Creating a BlackBox
BlackBox
COBie
BlackBox
Tools
Geometric data
Filtering
Verification
QTO Check
Quantity take-off rules
CAD BIM
New project
Project
Mining
Projects
Project models (RVT, IFC)
Geometric information of the projects
Properties of elements in the project
Steps to obtain 4D-6D data automatically
Manual processes
Automatically executed processes
0
1
2
3
4
5
6
7

图 5.4-4 ERP - 系统与分析工具集成，使公司决策过程自动化。

同样，ERP 系统中的流程，从开始到最终计算（步骤 1-6，图 5.4-3）是一连串相互关联的步骤，最终提供计划的透明度、效率和准确性。

现代建筑企业资源规划 系统不仅包括成本和进度计算模块，还包括数十个其他预配置模块，通常涵盖文件管理、项目进度跟踪、合同管理、供应链和物流，以及与其他业务系统和平台的集成。ERP 的集成分析工具允许用户自动创建仪表盘，以监控项目 KPI（KPI - 关键绩效指标）。这为建筑项目的各个方面提供了集中一致的管理，并尝试将大量应用程序和系统整合到一个平台上。

未来，ERP -analytics 将与机器学习结合使用，以提高准确性并优化计算未来项目属性的过程。在大数据 （图 5.4-4）中从企业资源规划系统中分析和收集的数据和属性今后将成为创建预测模型的基础，该模型可准确预测潜在的延误、风险或材料成本的可能变化等。

作为企业资源规划系统 的替代方案，建筑行业通常使用 PMIS （项目管理信息系统），这是一种项目管理系统，旨在对单个建筑项目的任务进行详细控制。

PMIS： **ERP** 与施工现场之间的中介

与涵盖公司业务流程整个链条的 ERP 不同，PMIS 侧重于管理具体项目、监控时间安排、预算、资源和文档。

> PMIS （项目管理信息系统）是一款建筑项目管理软件，旨在对项目的各个方面进行规划、跟踪、分析和报告。

PMIS 可让您管理文件、计划、预算，乍一看，PMIS 似乎是与 ERP 重复的解决方案，但关键区别在于管理水平：

- 企业资源规划系统 侧重于公司整体的业务流程：成本、合同、采购、人力资源和公司层面的资源管理。
- **PMIS** 专注于管理单个项目，提供详细规划、变更控制、报告和参与方协调。

在许多情况下，ERP 系统已经具备了足够的功能，实施 PMIS 更多的是为了方便和公司的偏好。许多承包商和客户使用 PMIS 并不是因为有必要，而是因为供应商或大客户希望在特定平台上汇总数据而强加的。

值得一提的是，在建筑项目管理的国际术语中，还有其他独立的流行概念，如 PLM （产品生命周期管理）和 EPC 和 EPC-M（工程、采购和施工管理）--建筑行业的承包方法。

如果公司已经使用带有项目管理模块的企业资源规划系统，那么引入 PMIS 可能是一个不必要的环节，造成功能重复。但是，如果流程没有自动化，数据分散，PMIS 可能是一个更方便、更易于维护的工具。

投机、利润、孤立和缺乏透明度

企业资源规划系统 和项目管理信息系统缺乏透明度

尽管界面和程序外部简单，但建筑企业资源规划系统和项目管理信息系 统在大多数情况下都是封闭和不灵活的解决方案。这些系统通常是由单一供应商提供的预配置软件包，对内部数据库和流程逻辑的访问有限。

CAD-(BIM-) 供应商正越来越多地接管此类系统的开发和控制工作，因为他们的数据库包含企业资源规划系统所需的信息：项目元素的数量和体积属性。然而，供应商并没有以开放或机器可读的格式提供对这些数据的访问，而是只提供有限的用户场景和封闭的处理逻辑--在 BlackBox 模块中预先定义。这就降低了系统的灵活性，使其无法适应特定的项目条件。

> 数据透明度有限仍然是建筑业数字化进程面临的主要挑战之一。封闭的数据库架构、无法获取建筑元素的全套属性、专注于 BlackBox 自动化模块以及缺乏开放式接口，都大大增加了文件官僚化的风险。这些限制造成了决策过程中的瓶颈，使信息难以核实，并为 ERP/PMIS 系统中的数据隐藏或投机打开了大门。用户通常只能获得有限的访问权限--无论是精简界面还是部分应用程序接口--而无法直接与主要数据源进行交互。当涉及到 CAD 项目自动生成的参数（如用于 QTO 计算的体积、面积和数量）时，这一点尤为重要。

因此，许多建筑公司不是通过流程自动化、开放数据、降低交易成本和创建新的商业模式来提高效率，而是专注于管理外部参数--在封闭的 ERP/PMIS 平台 中操纵影响项目成本的因素、调整因素和计算方法。这就为

投机创造了空间，扭曲了实际生产成本，降低了施工过程中所有参与者之间的信任度

> 在建筑业中，利润是指已完工项目的收入与可变成本之间的差额，可变成本包括设计、材料、人工和其他与项目实施直接相关的直接成本。然而，影响这些成本价值的关键因素不仅是技术或物流，还有计算的速度和准确性以及公司内部管理决策的质量。

在大多数建筑公司中，成本计算过程不仅对客户不透明，而且对不属于估算或财务部门的员工本身也不透明，这使得问题更加严重。这种封闭性有利于在公司内部形成一个特权专家群体--"财务专家 "的载体，他们拥有在 ERP/PMIS 系统中编辑属性和修正系数的专属权利。这些员工与公司领导一起，可以实际控制项目的财务逻辑。

> 在这种情况下，估算人员就成了 "财务魔术师"，既要实现公司利润最大化，又要为客户维持有竞争力的价格。同时，他们还必须避免明目张胆的粗暴操纵，以免损害公司的声誉。在这一阶段，需要制定系数来掩盖高估的数量或材料和工程成本。

因此，提高建筑业企业效率和盈利能力的主要办法不是自动化和加快决策进程， 而是对材料和工程价格进行投机（图 5.4-5）。在封闭的 ERP /PMIS - 系统中，通过系数夸大材料或工程量的市场平均价格的百分比（图 5.1-6），多报工程和材料成本，这在 "计算汇编 和根据资源基础计算工程成本 "一章中讨论过。

> 因此，客户收到的计算结果并不能反映真实的成本或工程范围，而是许多隐藏的内部系数的衍生物。同时，分包商为了达到总承包商低估的费率，往往被迫购买更便宜、更劣质的材料，从而导致最终的施工质量下降。

这种凭空寻找利润的投机过程最终伤害了客户，因为他们得到的是不可靠的数据，而执行人员则被迫寻找越来越多的投机模式

因此，项目越大，数据和流程管理的官僚主义程度就越高。每个步骤和每个模块往往都隐藏着计算算法和内部程序中不透明的系数和附加费用。这不仅给审计工作带来困难，还严重扭曲了项目的财务状况。在大型建筑 项目中，这种做法往往导致最终成本成倍增加（有时甚至高达十倍），而真正的工程量和 成本却仍然在

客户的有效控制之外（图 2.1-3 德国大型基础设施项目计划成本与实际成本的比较。

> 根据麦肯锡公司的报告《想象建筑业的数字化未来》（2016 年），大型建筑项目的竣工时间平均比计划晚 20%，超出预算高达 80%[107]。

估算和预算编制部门正在成为公司内部最受保护的环节。即使是内部专家进入这些部门也受到严格限制，而且由于逻辑和数据库结构的封闭性，不可能客观地评估项目决策的效率。由于缺乏透明度，公司被迫不去优化流程，而是通过 "创造性 "的数字和因素管理来为生存而战（图 5.3-1，图 5.1-6 - 例如，参数 "投标。因素 "）。

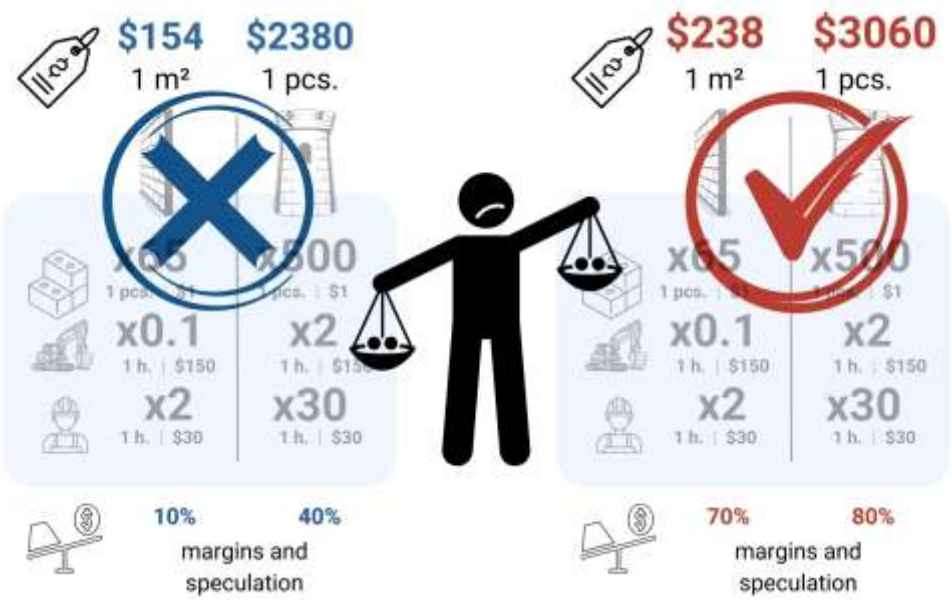


图 5.4-5 结算级投机比率是公司的主要利润，也是在工程质量和声誉之间玩弄的艺术。

所有这些都对建筑业继续使用封闭式 ERP/PMIS 系统提出了质疑。在数字化转型和客户对透明度要求不断提高的背景下（图 10.2-3），项目的实现不太可能长期依赖于限制灵活性、妨碍集成和阻碍业务发展的专有解决方案。

无论建筑公司在封闭的数据库中使用数据孤岛和不透明的数据是多么有利--不可避免的是，建筑行业的未来将涉及向开放平台、机器可读和透明的数据结构以及基于信任的自动化转型。这种转型将由高层推动--在客户、监管机构和社会的压力下，他们对问责制、可持续性、透明度和经济可行性的要求越来越高。

封闭式企业资源规划系统 /PMIS 时代终结：建筑业需要新方法

使用由数千万行代码组成的庞大模块化 ERP/PMIS 系统，对其进行任何更改都极为困难。在这种情况下，在已经为公司预先配置好的模块、资源数据库中数以万计的文章（图 5.1-3）和数以千计的现成计算（图 5.1-6）的存在下，向新平台的过渡变成了一个昂贵而漫长的过程。代码和传统架构越多，内部效率就越低，而每个新项目只会让情况变得更糟。在许多公司，数据迁移和新解决方案的集成成为多年的史诗，伴随着不断的返工和无休止的妥协。结果往往是回到熟悉的旧平台，尽管这些平台有其局限性。

正如关于建筑数据管理系统性失误的德国黑皮书报告[108]所强调的那样，信息分散和缺乏集中管理是导致效率低下的关键原因。没有标准化和整合，数据就会失去价值，成为档案而非管理工具。

> 数据质量损失 的一个主要原因是对施工项目的规划和控制不当，这往往会导致成本大幅增加。黑皮书的 "焦点：成本爆炸 "部分分析了造成这些不良后果的关键因素。这些因素包括需求分析不充分、缺乏可行性研究以及规划不协调，从而导致本可避免的额外成本。

在公司成熟的 IT 生态系统中，更换过时的系统就好比更换已建成大楼的承重柱。仅仅拆除旧的系统并安装新的系统是不够的，重要的是要确保建筑物保持稳定，天花板不会坍塌，所有通信都能继续工作。这就是困难所在：任何失误都可能对整个公司系统造成严重后果。

尽管如此，建筑行业大型企业资源规划系统的开发者们仍然将代码编写量作为支持其平台的论据。在专业会议上，我们仍然可以听到这样的说法："重新创建这样一个系统需要 150 人年的时间"，尽管事实上，这些系统的大部分功能都隐藏在数据库和相当简单的表格处理功能中，并封装在一个特殊的固定用户界面中。在实践中，"150 人年 "的代码量变成了一种负担，而不是一种竞争优势。代码越多，支持成本就越高，适应新情况就越困难，新开发人员和客户的进入门槛就越高。

如今，许多模块化建筑系统就像笨重过时的 "弗兰肯斯坦建筑"，任何不经意的改变都可能导致故障。每个新模块都会增加本已超负荷的系统的复杂性，使其变成只有少数专家才能理解的迷宫，从而使维护和现代化变得更加困难。

开发人员自己也意识到了复杂性，他们会定期暂停重构工作，即根据新技术的出现修改架构。然而，即使定期进行重构，复杂性还是会不可避免地增加。此类系统的架构师已经习惯了日益增长的复杂性，但对于新用户和专家来说，复杂性却成了难以逾越的障碍。因此，所有的专业知识都集中在少数开发人员手中，系统不再具有可扩展性。从短期来看，这些专家是有用的，但从长期来看，他们会成为问题的一部分。

> 企业将继续把 "小 "数据与大数据整合在一起，如果有人认为一个应用程序--无论多么昂贵或强大--就能处理所有问题，那就太愚蠢了[109]。
>
> - 关于协作的对话》播客主持人菲尔-西蒙

这就产生了一个合理的问题：如果其他行业早已能够使用数据开放、逻辑透明的分析工具来完成类似的任务，我们是否真的需要这种繁琐而封闭的系统，以表格的形式来计算工程成本和时间安排？

目前，建筑行业仍然需要封闭式模块平台，这主要是由于成本核算的特殊性（图 5.1-7）。此类系统通常用于运行灰色或不透明的计划，从而使实际成本无法被客户发现。然而，随着以客户为主的行业数字化的成熟，并进入所谓的 "Uberised 时代"，中介机构，即拥有企业资源规划系统的建筑公司，将失去其在时间和成本计算中的重要性。这将彻底改变建筑业的面貌。更多信息，请参阅本书最后一部分，以及 "建筑 5.0：如何在无法再躲藏的情况下赚钱 "一章。

> 在过去 30 年中积累起来的数千套传统解决方案，以及在开发过程中投入的数千个人工年，将开始迅速消失。向开放、透明和灵活的数据管理转变是不可避免的。唯一的问题是，哪些公司能够适应这些变化，哪些公司仍将受制于旧模式。

计算机辅助设计（BIM -）工具领域也存在类似情况，如今其数据充实了 ERP/PMIS -系统中设计实体的体积参数。BIM 的概念（早在 2002 年就已提出[110]）最初是基于单一集成数据库的概念，但实际上，如今使用 BIM 需要一整套专业软件和 格式。本应简化设计和施工管理，却变成了另一层专有解决方案，使集成变得复杂，降低了业务灵活性。

下一步：有效利用项目数据

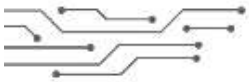
在本部分中，我们展示了结构化数据如何成为准确计算建筑项目成本和进度的基础。QTO、进度计划和估算过程的自动化可降低人工成本，并显著提高结果的准确性。

在总结这部分内容时，值得强调的是主要的实际步骤，它们将帮助您在日常工作中应用所讨论的方法。这些方法是通用的，既适用于公司的数字化转型，也适用于参与计算的专业人员的日常工作：

- 日常计算自动化
  - ☐ 尝试找到与工作相关的标准工作成本计算方法
  - ☐ 分析贵国建筑工地采用哪些方法来计算工程或工序的成本（图 5.1-7）。
  - ☐ 如果您正在使用 CAD 系统 - 学习在 CAD (BIM-) 软件中自动提取规格和 QTO 数据的功能。
  - ☐ 使用 LLM 编写代码草案，实现自动计算
- 为 QTO 开发自己的工具
  - ☐ 创建脚本或表格来自动计算容量
  - ☐ 将类别和要素组标准化，以便采用一致的评估方法
  - ☐ 记录计算方法，确保新项目结果的可重复性
- 将项目的不同方面融入你的工作中
  - ☐ 如果您使用的是模块化系统，那么不仅要将流程可视化为图表，还要将流程可视化为数据，尤其是表格形式的数据。
  - ☐ 使用 Python 代码，利用分组、过滤和汇总功能，掌握自动合并从 CAD 数据库中提取的数据并进行计算的方法
  - ☐ 为 QTO 小组创建清晰的可视化图表，向同事和客户展示复杂的信息

这些步骤将有助于在数据自动化和标准化的基础上建立一个可持续的计算系统。这种方法将提高准确性，减少日常计算问题。

以下各章主要介绍 CAD - (BIM-) 产品的技术方面，以及 CAD 数据库仍然难以集成到公司业务流程中的原因。如果您现在对建筑业实施 BIM 的历史、CAD 工具的演变以及使用这些技术的技术细节不感兴趣，您可以直接阅读本书的第七部分 "数据驱动的决策制定"。



印刷版带来的最大便利

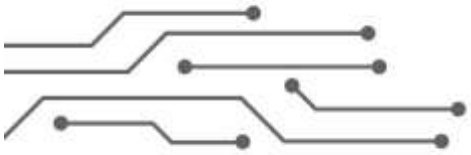
您现在持有的是《数据驱动的建筑》的免费电子版。为了更方便地工作和快速查阅资料，我们建议您关注印刷版：



- 随手可得：印刷版书籍将成为可靠的工作工具，让您在任何工作环境中都能快速找到并使用正确的可视化图示和图表
- 高质量插图：印刷版中的所有图像和图形均以最高质量呈现
- 快速获取信息：导航简便，可做笔记、书签，随时随地使用图书。

通过购买本书的完整印刷版，您将获得一个舒适高效地处理信息的便捷工具：能够在日常工作中快速使用视觉材料，快速查找必要的方案并做笔记。此外，您的购买还支持了开放知识的传播。

订购本书印刷版，请访问：datadrivenconstruction.io/books



VI 部分

cad 和 **bim**： 市场、现实和建筑设计数据的未来

本书第六部分对 CAD 和 BIM 技术的演变及其对建筑数据管理流程的影响进行了批判性分析。它追溯了 BIM 概念的历史变迁，从最初的集成数据库构想到目前软件供应商推广的营销构想。评估了专有格式和封闭系统对项目数据处理效率和建筑行业整体绩效的影响。详细分析了不同 CAD 系统的兼容性问题及其与建筑公司业务流程整合的困难。讨论了简化开放数据格式（如 USD ）的当前趋势及其对行业的潜在影响。介绍了从封闭系统中提取信息的其他方法，包括逆向工程技术。分析了应用人工智能和机器学习 实现建筑设计和数据分析过程自动化的前景。预测了以用户实际需求而非软件供应商利益为导向的设计技术发展前景。

章 6.1.

BIM- 概念在建筑业的兴起

本书第一版原本不包括第六部分，即 CAD (BIM)。专有格式、几何内核 和封闭系统等主题过于专业，细节过多，对于那些只想了解如何使用数据的人来说似乎毫无用处。然而，对本书第一版的反馈意见和增加说明的要求表明，如果不了解 CAD 系统内部运作的复杂性、几何内核 、格式的多样性以及相同数据不兼容的存储方案，就不可能真正理解为什么供应商推广的概念往往会使信息处理变得困难，并阻碍向开放参数化设计的过渡。这就是为什么这部分内容在本书结构中占据了自己的位置。如果 CAD（BIM）对您来说不是优先事项，您可以直接跳到下一部分--"第七部分：数据驱动决策、分析 、自动化和机器学习"。

BIM 和开放式 BIM 作为 CAD- 供应商营销概念出现的历史

随着 20 世纪 90 年代数字数据的出现，计算机技术不仅被引入业务流程，也被引入设计流程，从而产生了 CAD （计算机辅助设计系统）和后来的 BIM （建筑信息模型）等概念

然而，与任何创新一样，它们并不是发展的终点。BIM 等概念已成为建筑行业发展史上的一个重要里程碑，但它们迟早会被更好的工具和方法所取代，从而更好地应对未来的挑战。

> 由于受到 CAD 厂商的影响和自身实施的复杂性的困惑，2002 年出现的 BIM 概念很可能活不过它的三十岁生日，就像一颗耀眼夺目的摇滚明星，但很快就会陨落。原因很简单：数据科学家的需求变化速度比 CAD 供应商更快。

面对缺乏高质量数据的问题，当今建筑行业的专业人士要求跨平台互操作性和从 CAD 项目中获取开放数据，以简化分析和处理过程。复杂的 CAD 数据和混乱的 CAD 数据处理对参与建筑过程的每个人都有负面影响：设计师、项目经理、现场建筑工人以及最终的客户。

如今，客户和投资者收到的不是完整的操作数据集，而是 CAD- 格式的容器，需要复杂的几何内核 、对数据模式的理解、每年更新的 API -文档和专业的 CAD 软件（BIM ）来处理数据。同时，许多设计数据仍未使用。

在当今的设计和建筑领域，访问 CAD 数据的复杂性导致项目管理过度工程化。使用 CAD 数据或开发 BIM -

解决方案的大中型公司要么被迫与 CAD 供应商保持密切关系 解决方案，通过 API 访问数据，要么绕过 CAD 供应商的限制，使用昂贵的 SDK 转换器 进行逆向工程，以获取开放数据 [75]。

> 专有数据方法已经过时，不再符合当今数字环境的要求。未来的公司将分为两类：一类是有效使用开放数据的公司，另一类是退出市场的公司。

BIM （建筑信息模型）的概念是随着一家主要 CAD 供应商于 2002 年发布《BIM 白皮书》[54]而出现在建筑行业的，并由机械工程概念 BOM （材料清单）作为补充，源于创建和处理项目数据的参数化方法（图 6.1-1）。创建和处理设计数据的参数化方法是最早在机械工程设计 Pro-E 系统 (MCAD) 中实施的方法之一。该系统成为许多现代 CAD 原型[111]，包括当今建筑行业使用的解决方案。

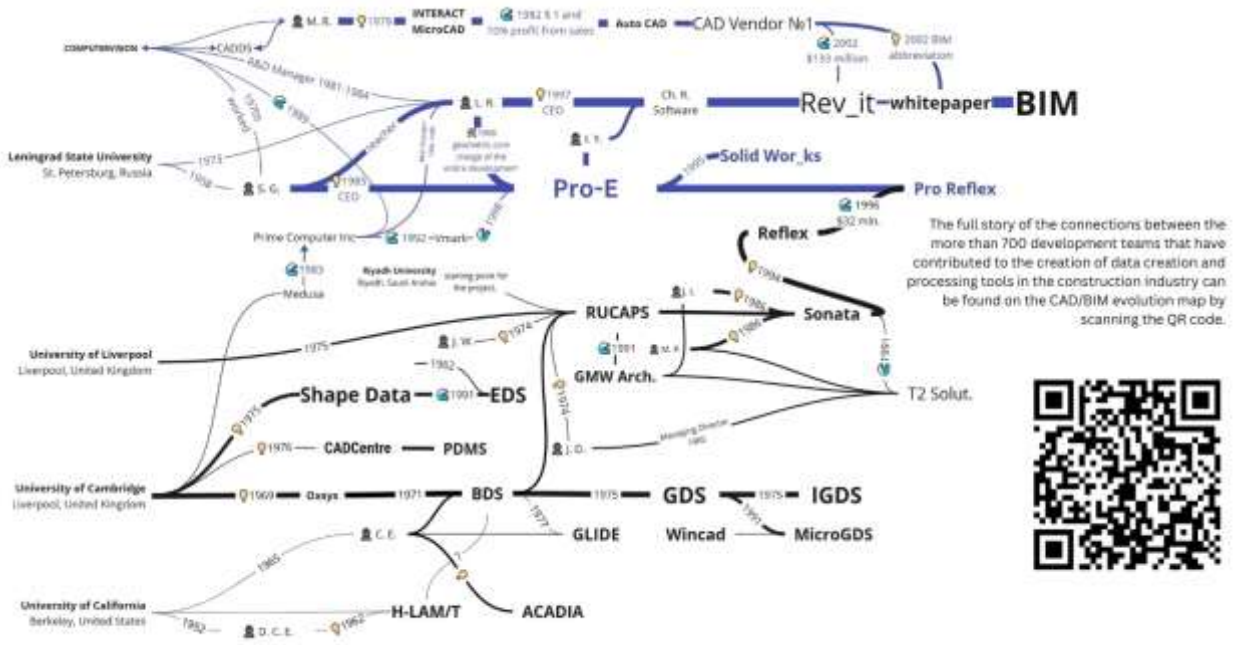


图 6.1-1 BIM 概念的历史图 和类似概念。

在本世纪初之前，记者和 AEC 顾问一直在推广 CAD 工具 - 供应商，而从 2002 年起，他们的注意力转向了白皮书 BIM。正是《2002-2004 年 BIM 白皮书》以及 2002、2003、2005 和 2007 年发表的文章在建筑行业普及 BIM 概念方面发挥了关键作用[112]。

> 建筑信息模型是.....[CAD 供应商公司名称]将信息技术应用于建筑业的战略。
>
> - 建筑信息模型白皮书, 2002 [60]

到 2000 年代中期，"研究人员 "开始将 CAD- 供应商于 2002 年发布的 BIM- 概念与查尔斯-伊士曼的 BDS 等早期科学著作联系起来，后者成为 GLIDE、GBM、BPM、RUCAPS 等系统的基础。Charles Eastman 在其开创性著作《建筑描述系统》（1974 年）中奠定了现代信息建模的理论基础。在他的著作中，"数据库 "一词出现了 43 次（图 6.1-2）--除了 "建筑 "一词之外，这个词出现的次数比其他任何词都多。

> 伊士曼的主要想法是，建筑物的所有信息--从几何图形到构件属性及其相互关系--都应存储在一个单一的结构化数据库中。从这个数据库中可以自动生成和分析图纸、规格、计算和规范合规性。伊士曼明确批评图纸是过时和多余的交流方式，指出存在信息重复、更新问题以及更改时需要手动更新等问题。相反，他建议在数据库中建立一个单一的数字模型，任何更改只需进行一次，并自动反映在所有视图中。

值得注意的是，伊士曼在他的概念中并没有把可视化放在中心位置。信息是他的系统的核心：参数、关系、属性、分析和自动化功能。在他的理解中，图纸只是显示数据库数据的形式之一，而不是设计信息的主要来源。

在领先的 CAD 供应商关于 BIM 的第一份白皮书中，"数据库 "一词的使用频率与查尔斯-伊斯特曼的《BDS》在七页纸上出现了 23 次[60]，是该文件中继 "建筑"、"信息"、"建模 "和 "设计 "之后最常用的词汇之一。然而，到 2003 年，"数据库 "一词在类似文件中只出现了两次[61]，到 2000 年代末，数据库话题几乎从设计数据的讨论中消失了。因此，"用于可视化和定量分析的单一集成数据库 "这一概念从未完全实现。

因此，建筑行业从查尔斯-伊士曼（Charles Eastman）强调数据库的渐进式 BDS 概念，到塞缪尔-盖斯伯格（Samuel Geisberg）关于在机械工程产品 Pro-E（当今建筑业使用的流行 CAD 解决方案的前身）中从数据库自动更新设计数据的想法，再到目前市场上的 BIM，其中几乎没有提及通过数据库进行数据管理，尽管这是最初理论背后的概念。

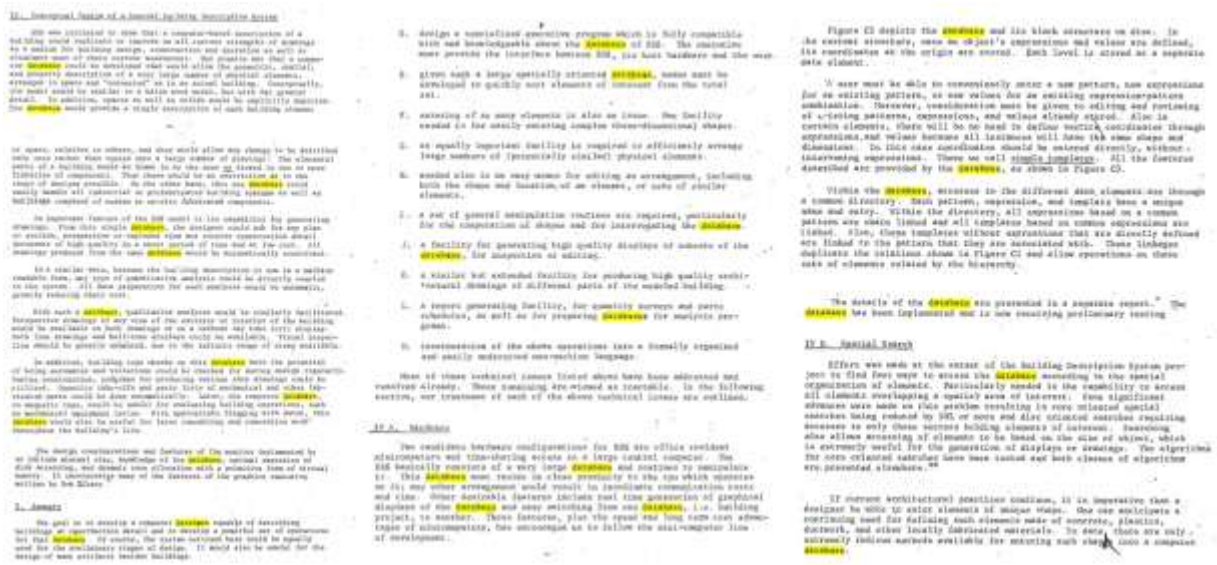


图 6.1-2 在 Charles Eastman 于 1974 年描述的 BDS 概念 中，"数据库"（用黄色标出）一词被使用了 43 次。

BDS 和 2000 年代之前的类似概念都是作为建筑物的数字数据库 而非可视化工具开发的。2002 年，BIM 成为一种设计工具，数据库退居其次。从 20 世纪 90 年代的 BDS 和类似概念到 2010 年代中期的 BIM ，我们失去了什么？

- 开放式数据库：BDS 和其他类似概念强调分析，BIM 强调设计。
- 灵活运用数据：BDS 强调数据分析，BIM 强调基于模糊数据的流程。
- 透明度：BDS 原本是一个开放的集成数据库 ，而 BIM 领域的 CAD 供应商 则将其数据库完全封闭，20 年来一直与开放专有格式的逆向工程工具 作斗争，但未取得成功。

在过去的 30 年中，设计师从未使用过 "集成数据库"，而在围绕 BIM - 工具进行了长达 20 年的狂热营销之后，建筑行业开始意识到这一流行趋势的后果。

BIM 的现实：取代集成数据库 **-** 封闭式模块系统

计算机辅助设计 - (BIM-) 系统的用户并不关注数据、数据结构以及将数据整合到统一的流程中，而是被迫使用一套支离破碎的专有解决方案，每个解决方案都有自己的游戏规则：

- 在第一份 BIM 白皮书 中讨论过的统一数据库 仍然是一个神话。尽管人们大声疾呼，但获取数据的途径仍然有限，而且分布在各个封闭系统中。

- **BIM -**模型已成为一个封闭的生态系统，而不是一种工具。用户不得不付费订阅并使用专有的应用程序接口，而不是进行透明的信息交换。
- 数据属于供应商，而非用户。项目信息被锁定在专有格式或云服务中，而不是以开放和独立的格式提供。

设计工程师和项目经理通常无法访问 CAD 数据库 系统，也无法访问他们自己的项目数据存储格式。这样就无法快速核实信息或制定数据结构和质量要求（图 6.1-3）。访问这些数据需要一整套通过应用程序接口（API） 和插件连接的专业程序，这导致了建筑行业流程的过度官僚化。同时，这些数据被数十个信息系统和数百名专家同时使用。

> 我们需要能够管理所有这些数据[CAD (BIM)]，将其数字化存储，并销售生命周期和流程管理软件，因为每一位工程师[设计师][在 CAD 软件中]创建一个东西，就有十个人在使用这些数据"[41]。
>
> - CAD 的首席执行官，该公司是创建 BIM 概念的供应商, 2005 年。

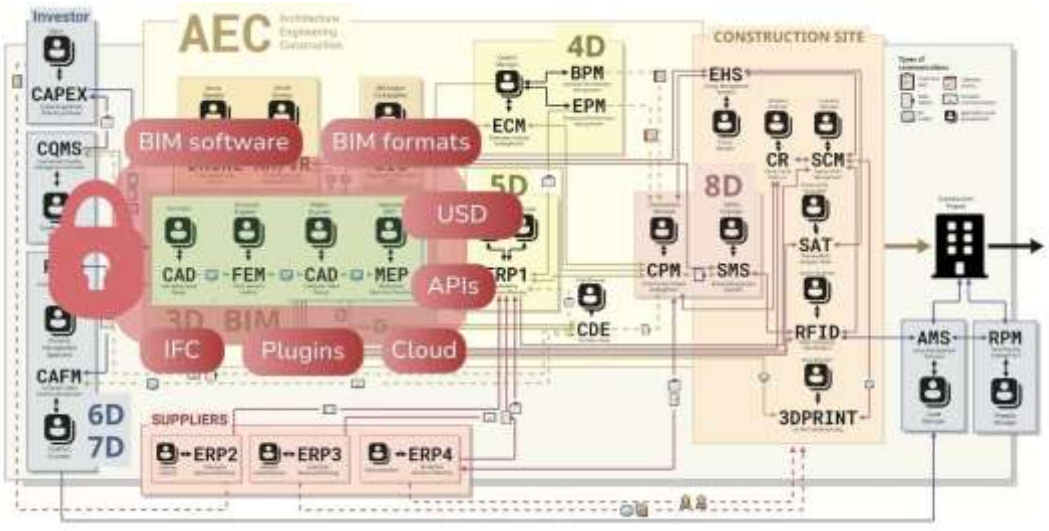


图 6.1-3 CAD- (BIM-) 数据库仍然是建筑业务生态系统中 IT 部门和数据管理人员最后的封闭系统之一。

> 当人们发现 BIM 更像是数据库商业化的一种手段，而不是一个成熟的数据库管理工具时，一个合乎逻辑的问题就出现了：我们如何才能重新获得对数据的控制权？答案就是使用开放式数据结构，让用户而不是软件供应商成为信息的所有者。

与其他行业的用户和解决方案开发商一样，建筑行业的用户和解决方案开发商将不可避免地摒弃过去 30 年中占主导地位的模糊的软件供应商术语，将重点放在数字化的关键方面--"数据 "和 "流程 "上。

20 世纪 80 年代末，建筑业数字化发展的关键领域是数据访问和项目信息管理。然而，随着时间的推移，重点发生了转移。IFC 格式 和开放式 BIM 概念 被积极推广，试图转移专家们对项目数据库管理主题的注意力，而不是开发透明、易用的数据工作方法。

开放格式 **IFC** 在建筑行业的出现

所谓的开放格式 IFC （工业基础类）被定位为确保不同 CAD(建筑信息模型 -)系统之间互操作性的标准。其开发是在主要 CAD 供应商创建和控制的组织框架内进行的。基于 IFC 格式，两家 CAD 公司于 2012 年开发了 OPEN BIM 营销概念[63]。

> **IFC** （工业基础类）是建筑行业数据交换的开放标准，旨在确保不同 CAD - (BIM-) 系统之间的互操作性。

开放式 BIM - 概念涉及使用 CAD 数据库中的信息，并通过用于交换 CAD 数据的开放格式 - IFC 在系统之间交换信息。

> 开放 BIM 计划是由.....[1 家 CAD 供应商]、.....[2 家 CAD 供应商]和其他公司发起的一项营销活动，旨在鼓励和促进在整个 AEC 行业全球协调推广开放 BIM 概念，并为计划参与者提供一致的沟通和通用品牌。
>
> - 摘自 CAD 供应商网站，开放式 BIM 计划，2012 [113]

IFC 由慕尼黑工业大学于 20 世纪 80 年代末从机械工程格式 STEP 改编而来，后来由一家大型设计公司和一

家大型 CAD 供应商于 1994 年注册成立 IAI（互操作性工业联盟）[114]（图 6.1-4）。IFC 格式的开发是为了确保不同 CAD 系统之间的互操作性，其基础是机械工程格式 STEP 中规定的原则，而 STEP 又是由一群 CAD 用户和供应商在 NIST（美国国家标准与技术研究院）和美国国防部的支持下于 1979 年创建的 IGES 格式发展而来[115]。

然而，IFC 的复杂结构、其对几何核心 的紧密依赖性以及不同软件解决方案在执行该格式时的差异导致了其实际应用中的许多问题。机械工程专家以前在使用 IGES 和 STEP 格式时也遇到过类似的困难--细节丢失、精度受限以及需要使用中间格式，而 IFC 正是在这两种格式的基础上产生的。



图 6.1-4 开发团队与 CAD 产品之间的联系图 (BIM) [116]。

> 2000 年，注册了 IFC 格式并创建了 IAI（后来的 bS）组织的同一家 CAD 供应商发布了白皮书 "集成设计与制造：优势与理由"[65]。[65].白皮书强调，在同一系统内的程序之间交换数据时，必须保持完整的数据粒度，而不能使用 IGES 、STEP [与 IFC 相同] 等中性格式。相反，文件建议应用程序应能直接访问底层 CAD 数据库，以防止信息准确性的损失。

2002 年，同一家 CAD 供应商购买了参数化 BOM 产品（图 3.1-18，更多详情见第三部分 ），并在此基础上形成了 BIM 概念。因此，只有封闭的 CAD 格式或 IFC 格式 (STEP) 才能用于交换建筑项目数据，而 CAD 供应商自己在 2000 年就已写明了这种格式的局限性，并将其引入了建筑行业。

在 "CAD（BIM）的演变"[116] 一图中，详细介绍了参与创建和处理建筑数据工具的 700 多个开发团队的互动历史。[116].

开放式 IFC 包括设计元素的几何描述和元信息描述。在 IFC 格式中，使用了多种方法来表示几何图形，如 CSG 和 Swept Solids：不过，参数表示法 BREP 已成为以 IFC 格式传输元素几何图形的主要标准，因为这种格式在从 CAD- (BIM-) 程序导出时受到支持，并允许在将 IFC 导入 CAD 程序时对元素进行潜在编辑。

IFC 格式问题取决于几何核心

在大多数情况下，当 IFC 中的几何体是按参数定义的（BREP ），仅凭 IFC 文件就不可能可视化或检索项目实体的体积或面积等几何属性，因为在这种 情况下要处理几何体并使其可视化，需要一个几何体内核（图 6.1-5），而这个内核最初是缺失的。

> 几何内核是一个软件组件，提供在 CAD (CAD)、BIM 和其他工程应用程序中创建、编辑和分析几何对象的基本算法。它负责构建二维和三维 几何图形，并对其进行布尔运算、平滑、交叉、变换和可视化等操作。

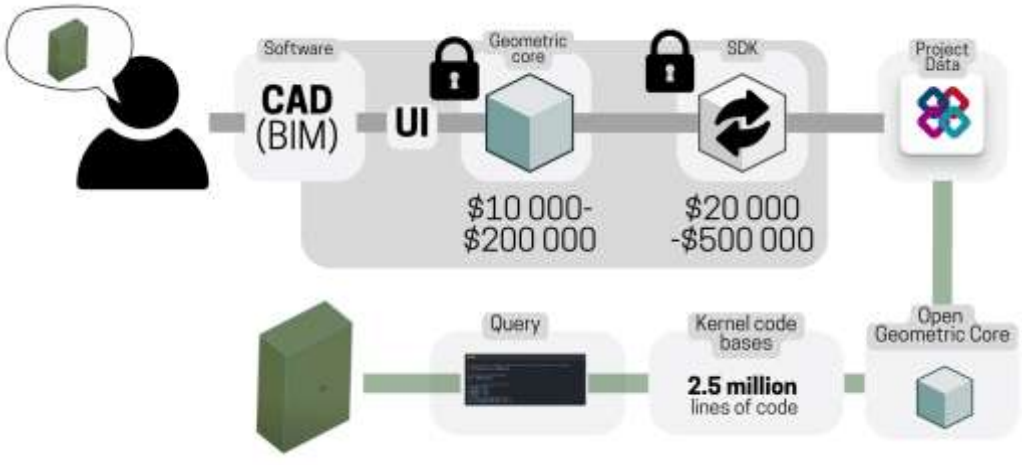


图 6.1-5 如今，通过 CAD- 软件创建几何图形需要使用专有几何图形内核和 SDK ，而 CAD 供应商通常并不拥有这些内核和 SDK。

每个 CAD 程序和任何使用参数或 IFC 格式的程序都有自己的或购买的几何内核。如果使用 IFC -BREP 格式的原始元素不会有任何问题，而且在使用不同几何内核的程序中，这些元素也能以类似方式显示，但除了使用不同几何内核引擎 的问题外，还有足够多的元素有其自身的特殊性，无法正确显示。2019 年出版的国际研究报告 "IFC 软件支持参考研究"[117]详细讨论了这一问题。

> 同样的标准化数据集会产生相互矛盾的结果，几乎找不到共同的模式，而且在支持标准方面也发现了严重的问题[IFC]，这可能是由于标准数据模型非常复杂。标准本身也有部分责任，因为它们往往对一些细节未加定义，自由度很高，可能会有各种不同的解释。它们允许对象的组织和存储具有很高的复杂性，这不利于有效的普遍理解、独特的实施和一致的数据建模[117]。
>
> - IFC 软件支持参考研究，2021 年

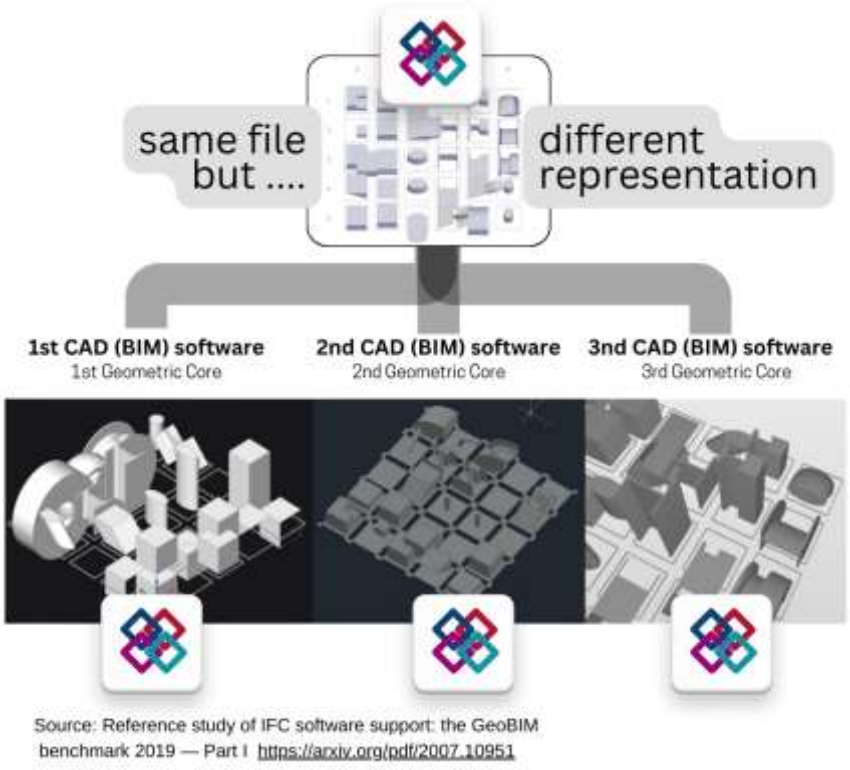


图 6.1-6 不同的几何核对参数描述的相同几何体有不同的表示方法（基于 [117]）。

对 "某些规定 "的正确理解只有开发 IFC 的特殊组织的付费成员才能获得。因此，谁想获得有关 IFC 某些特性的重要知识，就必须尝试与大型 CAD 供应商合作，或通过自己的研究对这些特性进行定性分析

> 您偶然发现了一个关于通过 IFC 格式导入和导出数据的问题，并询问其他供应商："为什么 IFC 文件中会有关于参数化前提传输的信息？开放式规范中并没有提及"。来自 "更有见识 "的欧洲供应商的回答："是的，没说，但允许"。
>
> - 摘自 CAD 2021 开发人员的访谈 [118]

IFC 通过参数基元描述几何图形，但不包含嵌入式内核 - 其作用由 CAD 程序完成，该程序通过几何内核编译几何图形。几何内核执行数学计算并定义交叉点，而 IFC 仅提供数据用于解释。如果 IFC 中包含不正确的面，不同几何内核的不同程序会根据内核的不同忽略这些面或产生错误。

> 因此，要使用 IFC 格式，就必须回答一个很难找到明确答案的主要问题--应使用哪种工具和哪种几何内核来获得项目最初在 CAD 程序中获得的数据质量，而 IFC 就是从 CAD 程序中获得的？

数据质量问题 和 IFC 格式的复杂性不允许直接使用设计数据进行流程自动化、分析和数据处理，这往往导致开发人员不可避免地需要使用封闭的 CAD --可 "高质量 "访问数据的解决方案[63]，1994 年注册 IFC 的供应商自己也写到了这一点[65]。

在几何内核中映射和生成 IFC 参数的所有特殊性只能由具有几何内核工作经验的大型开发团队来实现。因此，当前 IFC 格式的特殊性和复杂性主要有利于 CAD 供应商，与大型软件供应商 "采用、扩展、破坏 "的策略有很多共同之处，而标准的日益复杂性实际上为小型市场参与者制造了障碍[94]。

在这种战略中，大型供应商的策略可能是调整开放标准，增加专有扩展和功能，使用户对其产品产生依赖性，然后挤垮竞争对手。

> IFC 格式 旨在成为不同 CAD- (BIM-) 系统之间的通用桥梁，但在现实中却扮演着不同 CAD 平台几何核心之间兼容性问题指示器的角色，类似于最初产生该格式的 STEP 格式。

因此，如今对于大型 CAD 供应商 来说，全面、高质量地实施 IFC 本体是可行的，他们可以投入大量资源来

支持所有实体及其与自己内部几何核心 的映射，而作为标准的 IFC 并不存在这种映射。大型供应商也有能力在它们之间协调特征的技术细节，而即使是 IFC 格式开发组织中最活跃的参与者也可能无法获得这些技术细节。

> 对于致力于支持互操作格式开发的小型独立团队和开源项目 来说，缺乏内部几何内核是一个严重的问题。没有内核，几乎不可能考虑到与跨平台数据交换相关的所有微妙之处和细微差别。

随着 IFC 参数格式 和开放式 BIM 概念 的发展，建筑行业对本体 和语义 在数据和流程管理中的作用的讨论愈演愈烈。

出现在语义 和本体 的主题构建中

得益于 20 世纪 90 末语义互联网的理念以及参与开发 IFC 格式 的组织的努力，语义和本体 已成为 2020 年代中期建筑行业讨论的标准化关键要素之一。

> 语义技术是对大量异构数据进行统一、标准化和修改，并实现复杂搜索的技术。

OWL （网络本体语言）表示为 RDF 图 -triplets（资源描述框架）（图 6.1-7），用于存储语义数据。OWL 指的是图数据模型，我们将在 "数据模型：数据关系和元素之间的关系 "一章中详细讨论其类型。

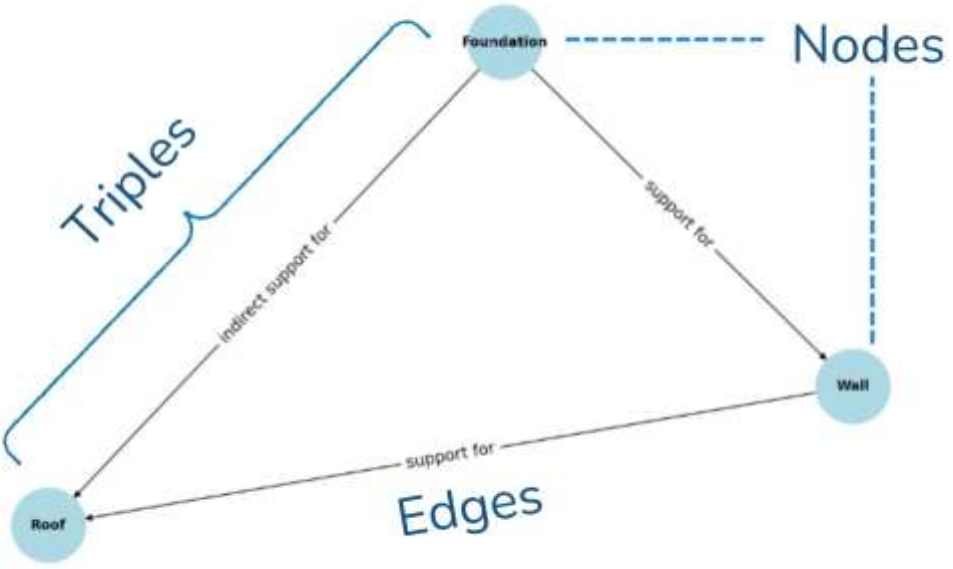


图 6.1-7 RDF 数据模型：节点、边和三元组说明了构件之间的关系。

从理论上讲，逻辑推理器（自动逻辑推理软件）可以从本体中推导出新的语句。例如，如果建筑本体 记录了 "地基是墙的支撑 "和 "墙是屋顶的支撑"（图 6.1-7），那么逻辑推理器就能自动推断出 "地基是屋顶的支撑"。

这种机制可以避免明确规定所有的依赖关系，因此有助于优化数据分析。不过，它并不创造新的知识，而只是识别和构建已知的事实。

> 语义学本身并不创造新的意义或知识，在这方面也不优于其他数据存储和处理技术。将关系数据库中的数据表示为三元组并不能使其更有意义。用图形结构取代表格可能有助于统一数据模型、方便检索和安全编辑，但并不会让数据变得更 "聪明"--计算机并不会开始更好地理解其内容。

数据中的逻辑关系无需复杂的语义技术即可组织（图 6.1-8）。传统的关系数据库（SQL ）以及 CSV 或 XLSX 格式都可以建立类似的依赖关系。例如，在列式数据库中，您可以添加一个 "屋顶支撑 "字段，并在创建墙体时自动将屋顶与地基关联起来。这种方法无需使用 RDF 、OWL 、图形或栅格，仍然是一种简单高效的数据存储和分析解决方案。

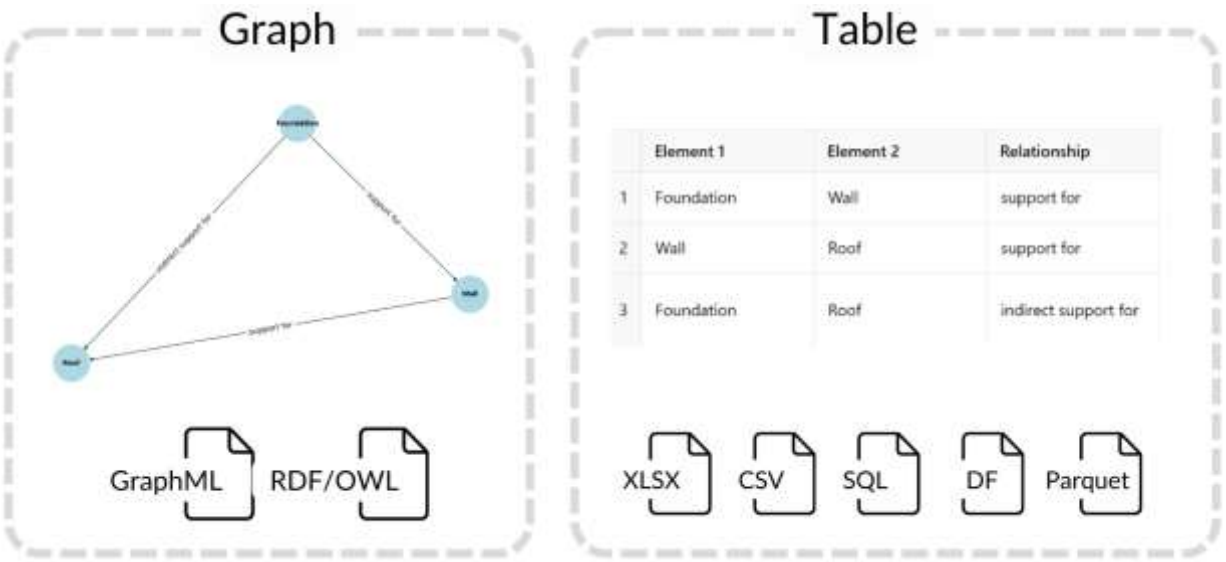


图 6.1-8 表示相同逻辑关系的图形和表格数据模型比较。

一些大型建筑公司和 IFC 格式开发组织 [94] 决定遵循语义网概念，这在 20 世纪 90 年代末似乎很有前途，但对建筑行业标准的制定产生了重大影响。

然而，矛盾的是，语义网 这一概念本身原本是为互联网而设的，但即使在其原生环境中也没有被广泛采用。尽管开发出了 RDF 和 OWL，但完整的语义网 并没有按照其最初的构想出现，而且已经不太可能诞生。

语义技术为何辜负了建筑行业的期望

其他行业也面临着利用语义 技术的限制。在游戏行业，由于变化的动态性很强，试图通过本体 来描述游戏对象及其交互的尝试被证明是无效的。因此，XML 和 JSON 等更简单的数据格式以及算法解决方案受到青睐。房地产行业的情况也类似：由于术语的地区差异和频繁的市场变化，本体的使用被证明过于复杂，而简单的数据库和标准（如 RETS [119]）则能更好地应对数据交换的挑战。

> 技术上的困难，如标记的复杂性、高劳动密集型支持和开发人员积极性不高，减缓了语义网 和其他经济部门采用语义网的速度。RDF （资源描述框架）没有成为大众标准，而本体 被证明过于复杂，在经济上也不合理。

因此，创建全球语义网 的宏伟构想未能实现。尽管本体 和 SPARQL 等一些技术元素已进入企业解决方案，但创建单一综合数据结构的最初目标并未实现。

在互联网上，计算机能够理解内容，这一概念已被证明在技术上是困难的，在商业上也是无利可图的。这就是为什么支持这一理念的公司最终将其应用减少到个别有用的工具上，让 RDF 和 OWL 满足高度专业化的企业需求，而不是整个互联网的需求。对过去 20 年谷歌趋势（图 6.1-9）的分析表明，语义网 的前景可能已经渺茫。

> 没有必要不必要地多重实体。如果对一种现象有几种逻辑上一致的解释，而且解释得同样好，那么在其他条件相同的情况下，人们应该选择其中最简单的解释。
>
> 剃刀

这里出现了一个逻辑问题：既然可以使用流行的结构化查询（SQL, Pandas, Apache® ）处理数据，为什么还要在构造中使用三元组、立方体和 SPARQL 呢？在企业应用中，SQL 是处理数据库的标准。相反，SPARQL 需要复杂的图形结构和专业软件，而且根据 Google 的发展趋势，并不吸引开发人员的兴趣。

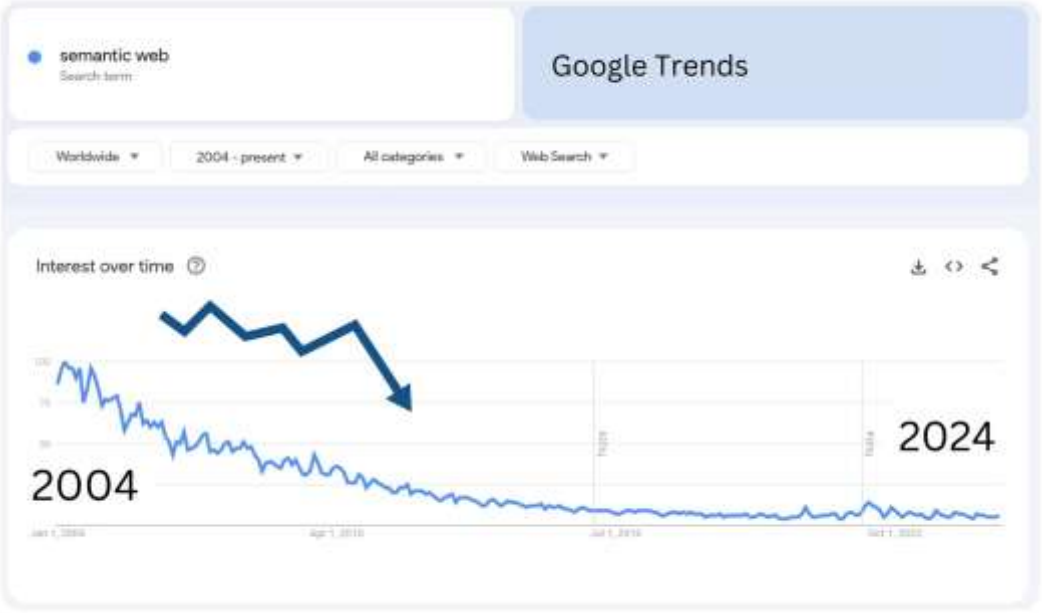


图 6.1-9 根据谷歌统计数据显示的 "语义互联网 "查询兴趣。

图形数据库 和分类树在某些情况下是有用的，但在大多数日常任务中，它们的应用并不总是合理的。因此，只有在需要统一不同来源的数据或实现复杂的逻辑结论时，创建知识图谱和使用语义网络技术 才有意义。

> 从表格到图形数据模型可以改进搜索并统一信息流，但并不能使数据对机器更有意义。问题不在于是否应该使用语义技术，而在于它们在哪些方面能真正发挥作用。在贵公司实施本体 、语义 和图形数据库 之前，先了解一下哪些公司已经成功使用了这些技术，以及在哪些方面失败了。

尽管人们抱有宏大的期望，但语义技术从未成为建筑行业数据结构化的通用解决方案。在实践中，这些技术并没有带来通用的解决方案，而只是增加了新的复杂性，这些努力与语义互联网概念未实现的雄心壮志不谋而合，后者的期望远远超过了现实。

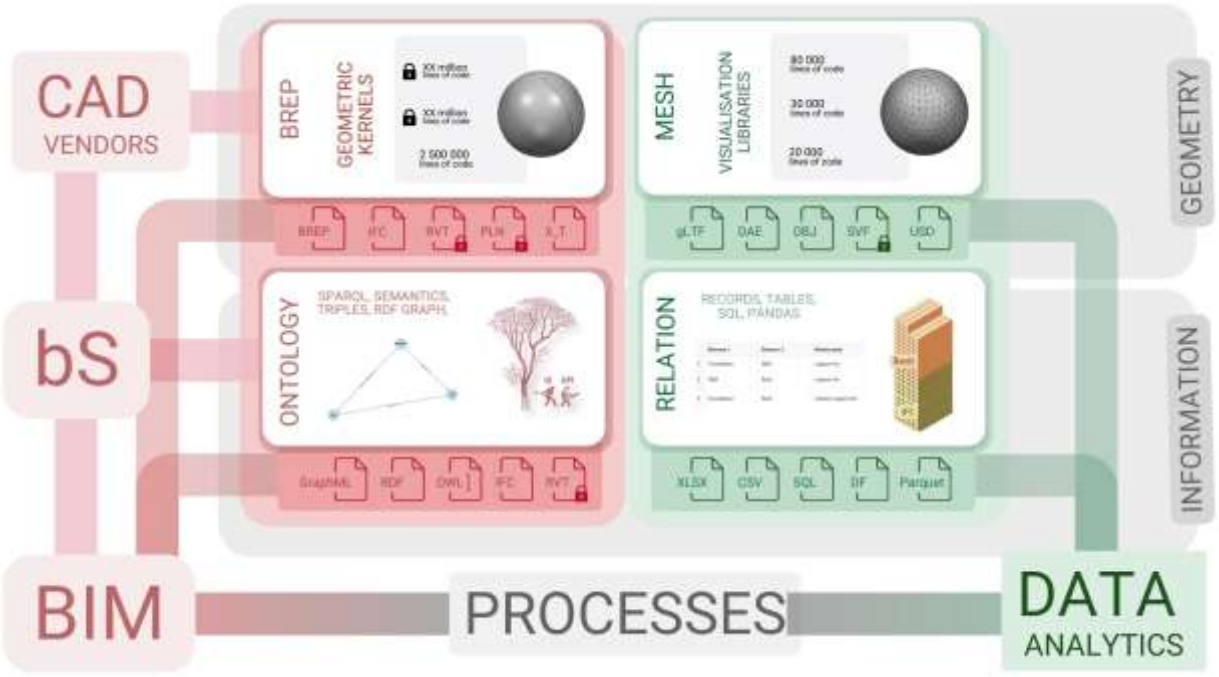


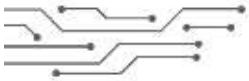
图 6.1-10 施工过程中的几何与信息：从复杂的 CAD 和 BIM- 系统到用于分析的简化数据。

在信息技术领域，语义网 的失败已被新技术（大数据、物联网 、机器学习 、AR/VR ）的出现所抵消，但建筑行业却没有这样的机会。

除了使用概念来沟通项目元素之间的数据关系所面临的挑战之外，还有一个根本性的问题--数据的可用性。建筑行业仍然以封闭系统为主，因此很难使用数据、共享信息和提高流程效率。

数据的封闭性正在成为阻碍建筑业数字化解决方案发展的主要障碍之一。在 IT 行业，开放的 和统一的数据格式已成为标准，但在 CAD 行业 (BIM)，每个软件都使用自己的格式，形成了封闭的生态系统，人为地限

制了用户。



章 6.2.

封闭项目格式和互操作性问题

封闭数据和生产力下降：**CAD** 行业的死胡同 **(BIM)**

CAD 系统的专有性质导致每个程序都有自己独特的数据格式，这些格式要么是封闭的，从 外部无法访问 - RVT 、PLN 、DWG 、NDW、NWD、SKP，或者通过相当复杂的转换过程以半结构化形式提供--JSON 、XML (CPIXML) 、IFC 、STEP 和 ifcXML、IfcJSON、BIMJSON、IfcSQL、CSV 等。.

存储相同项目的相同数据的不同数据格式不仅在结构上存在差异，还包括不同版本的内 部标记，开发人员需要考虑到这一点，以确保应用程序的兼容性。例如，2025 年的 CAD 格式可以在 2026 年的 CAD 程序中打开，但同一项目永远无法在 2025 年之前的所有 CAD 程序版本中打开。

> 由于不提供直接访问数据库的途径 ，建筑行业的软件供应商通常会创建自己独特的格式和工具 ，专业人员（设计工程师或数据管理员）必须使用这些格式和工具才能访问、导入和导出数据 。

因此，基本 CAD (BIM) 和相关解决方案（例如，ERP/PMIS) ）的供应商不断提高产品使用价格，普通用户被迫在数据传输的每个阶段按格式支付 "佣金"[63]：连接、导入、导出和使用用户自己创建的数据。

2025 年，从流行的 CAD - (BIM-) 产品的云存储中访问数据的成本将达到每笔交易 1 美元[120]，中型企业订购建筑企业资源规划 - 产品的费用每年高达五位数和六位数[121]。

> 现代建筑软件的本质不是自动化或提高效率，而是工程师对某一高度专业化软件的理解能力，它影响着建筑项目数据处理的质量和成本，以及承担建筑项目的公司的利润和长期生存。

由于无法访问数据库 CAD --这些系统被用于数十个其他系统和数百个流程[63]，因此各个专家之间缺乏高质量的交流，导致建筑业成为生产效率最低下的经济部门之一[44]。

在过去 20 年中，随着 CAD- （BIM- ）设计应用、新系统（ERP ）、新建筑技术和材料的出现，整个 建筑行业的生产率下降了 20%（图 2.2-1），而在访问数据库 和类似 BIM 营销概念方面没有重大问题的所有经

济部门的整体生产率提高了 70%（制造业提高了 96%）[122]。



图 6.2-1 由于建筑行业数十个部门和数百个流程所依赖的项目数据的孤立性和复杂性，决策速度比其他行业慢数倍。

不过，也有个别例子表明，可以采用其他方法在 CAD 解决方案之间创建互操作性。欧洲最大的建筑公司于 2018 年启动的 SCOPE 项目 [123]，展示了如何超越 CAD- （BIM- ）系统的经典逻辑。SCOPE 开发人员并没有试图征服 IFC 或依赖专有几何内核，而是使用 API 和 SDK 逆向工程 从各种 CAD 程序中提取数据，将其转换为中性格式，如 OBJ 或 CPIXML （基于唯一的开源 几何内核 OCCT ），并进一步将其应用到建筑和设计公司的数百个业务流程中。然而，尽管想法很先进，但这些项目面临着免费几何内核的局限性和复杂性，它们仍然是一家公司封闭生态系统的一部分，再现了单一供应商解决方案的逻辑。

由于封闭系统的限制和数据格式的差异，以及缺乏统一数据格式的有效工具，必须使用 CAD 格式的公司面临着大量结构和封闭程度不同的数据的积累。这些数据无法正常使用，只能消失在档案中，永远被遗忘和闲置。

> 在设计阶段花费大量精力获得的数据，由于其复杂性和封闭性，无法进一步使用。

因此，在过去的 30 年中，建筑行业的开发人员不得不一次又一次地面对同样的问题：每一种新的封闭格式或专有解决方案都会产生与现有开放式和封闭式 CAD 系统集成的需求。不断尝试确保不同 CAD 和 BIM 解

决方案之间的互操作性只会使数据生态系统更加复杂，而不是促进其简化和标准化。

CAD 系统之间互操作性的神话

20 世纪 90 年代中期，CAD 环境互操作性发展的主要方向是打破专有的 DWG 格式--最终，开放 DWG 联盟[75]取得了胜利，整个建筑行业真正开放了最流行的绘图格式--而到了 20 世纪 20 年代中期，重点发生了转移。在建筑行业，一种新的趋势正在蓄势待发：众多开发团队正致力于在封闭式 CAD 系统（封闭式 BIM）、IFC 格式和开放式解决方案（开放式 BIM）之间建立所谓的 "桥梁"。这些举措大多基于 IFC 格式和 OCCT 几何内核的使用，为不同平台之间架起了一座技术桥梁。这种方法被认为是一个很有前途的方向，可以显著改善数据交换和软件工具的互操作性。

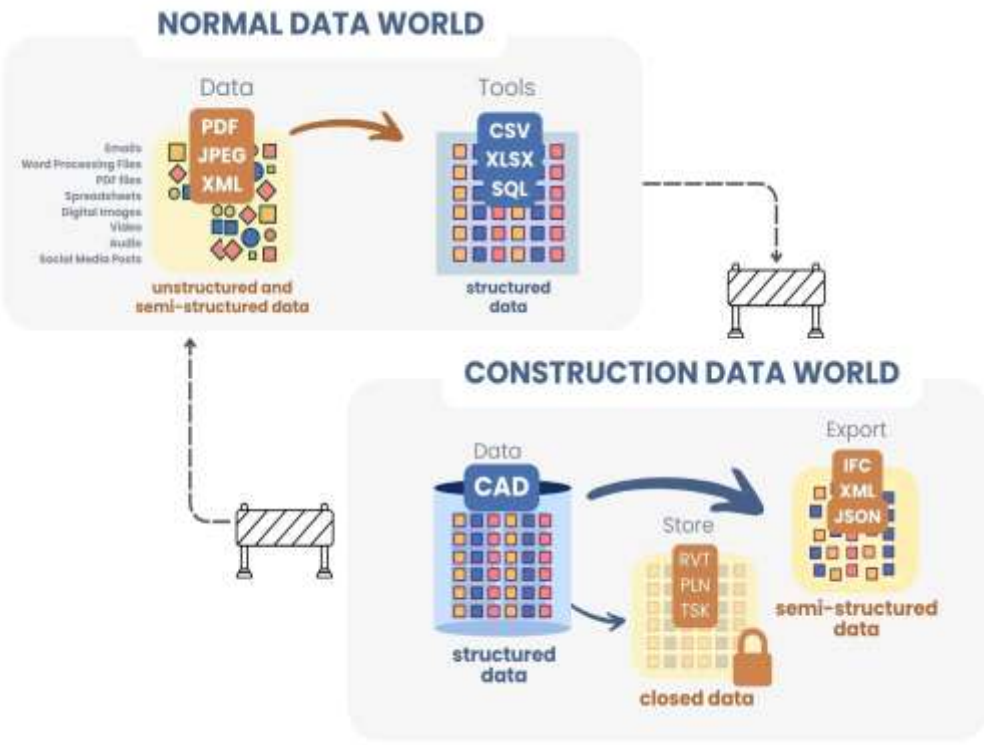


图 6.2-2 当其他行业使用开放数据时，建筑行业却不得不使用封闭或结构松散的 CAD 格式 (BIM)。

这种方法在历史上有相似之处。2000 年代，开发人员为了克服最大的图形编辑器供应商（二维世界）的主导地位，试图在其专有解决方案和免费开源 （GIMP 的替代品）之间建立无缝集成（图 6.2-3）。当时，就像今天的建筑行业一样，在保留复杂参数、图层和内部软件逻辑的同时，试图在封闭系统和开放系统之间架

起一座桥梁。

然而，用户实际上寻求的是简单的解决方案--扁平、开放的数据 ，没有过多复杂的层和程序参数（类似于 CAD 中的几何核心）。用户寻求的是简单、开放的数据格式，没有过多的逻辑。JPEG、PNG 和 GIF 就是这样的图形格式。如今，它们被广泛应用于社交网络、网站和应用程序中--无论平台或软件供应商如何，它们都易于处理和解释。

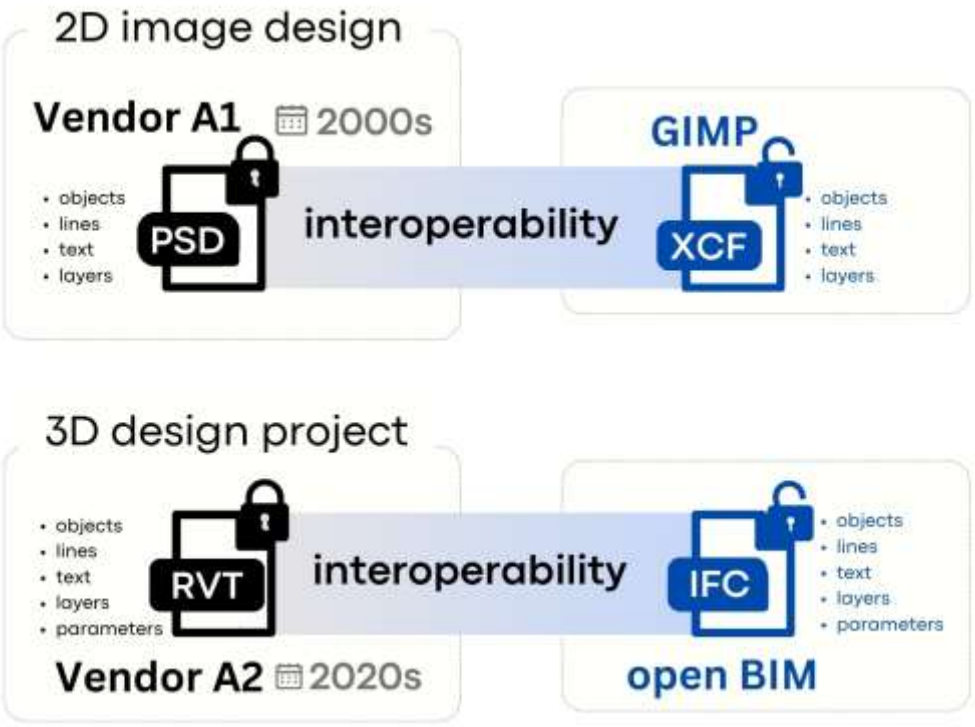


图 6.2-3 建筑数据格式的互操作性与 2000 年代尝试合并流行供应商的专有产品和开源 GIMP 的路径类似。

因此，如今图像行业几乎没有人使用 PSD 或开放式 XCF 等封闭格式来制作应用程序、Facebook 和 Instagram 等社交网络或网站内容。取而代之的是，大多数任务都使用扁平开放的 JPEG、PNG 和 GIF 格式，以便于使用和广泛兼容。JPEG 和 PNG 等开放格式因其通用性和广泛的支持，已成为图像共享的标准，使其易于在各种平台上使用。视频和音频等其他交换格式也出现了类似的转变，MPEG 和 MP3 等通用格式因其压缩效率高和兼容性广而受到重视。这种向标准化的转变简化了内容和信息的共享和播放，使所有用户都能在多个平台上使用它们（图 6.2-4）。

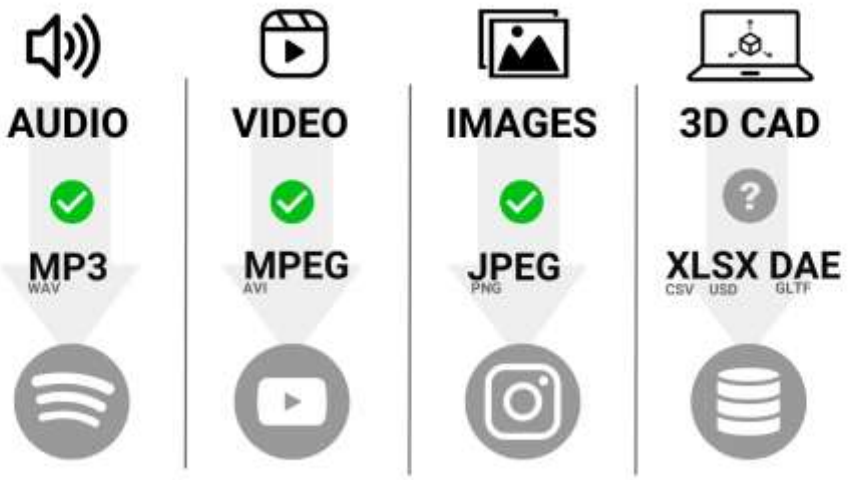


图 6.2-4 没有复杂编辑功能的简化格式已成为共享和使用数据的常用格式。

类似的过程也发生在三维 建模中。简单而开放的格式，如 USD, OBJ, glTF, DAE, DXF, SQL 和 XLSX 在项目中越来越多地用于 CAD 环境之外的数据交换 (BIM)。这些格式存储了所有必要的信息，包括几何图形和元数据，无需使用复杂的 BREP 结构、几何图形内核 或供应商特定的内部分类器。领先软件供应商提供的 NWC 、SVF、SVF2、CPIXML 和 CP2 等专有格式也具有类似功能，但与开放标准不同，它们仍然是封闭的。

> 值得注意的是（值得再次回顾的是，上一章已经提到过），这一观点--拒绝使用 IGES 、STEP 和 IFC 等中间中立和参数化格式--早 2000 年就得到了主要 CAD 供应商的支持，该供应商创建了 BIM 白皮书 ，并于 1994 年注册了 IFC 格式。在 2000 年 "集成设计与制造 "白皮书中[65] 中强调 ，为了保持信息的完整性和准确性，必须在软件环境中对 CAD 数据库进行本地访问，而无需使用中间翻译器和参数格式。

建筑行业尚未就访问 CAD 数据库的工具或其强制逆向工程 达成一致，也未就采用通用简化数据格式在 CAD 平台之外使用达成一致 (BIM)。例如，中欧和德语地区许多建筑行业的大公司在其 ERP 系统中使用 CPIXML 格式[121]。这种专有格式是 一种 XML ，它将 CAD（BIM）项目数据（包括几何数据和元数据）整合到一个有组织的简化结构中。大型建筑公司也在创建自己的新格式和系统，如 SCOPE 项目 ，我们在上一章中讨论过该项目

参数 CAD 格式或复杂参数文件 IFC (STEP) 的封闭逻辑在大多数业务流程中都是多余的。用户正在寻找简化

的平面格式，如 USD, CPIXML, XML &OBJ, DXF, glTF, SQLlite, DAE &XLSX，这些格式包含所有必要的元素信息，但没有冗余的 BREP 几何逻辑、对几何内核的依赖以及特定 CAD 和 BIM 产品的内部分类（图 6.2-5）。6.2-5).

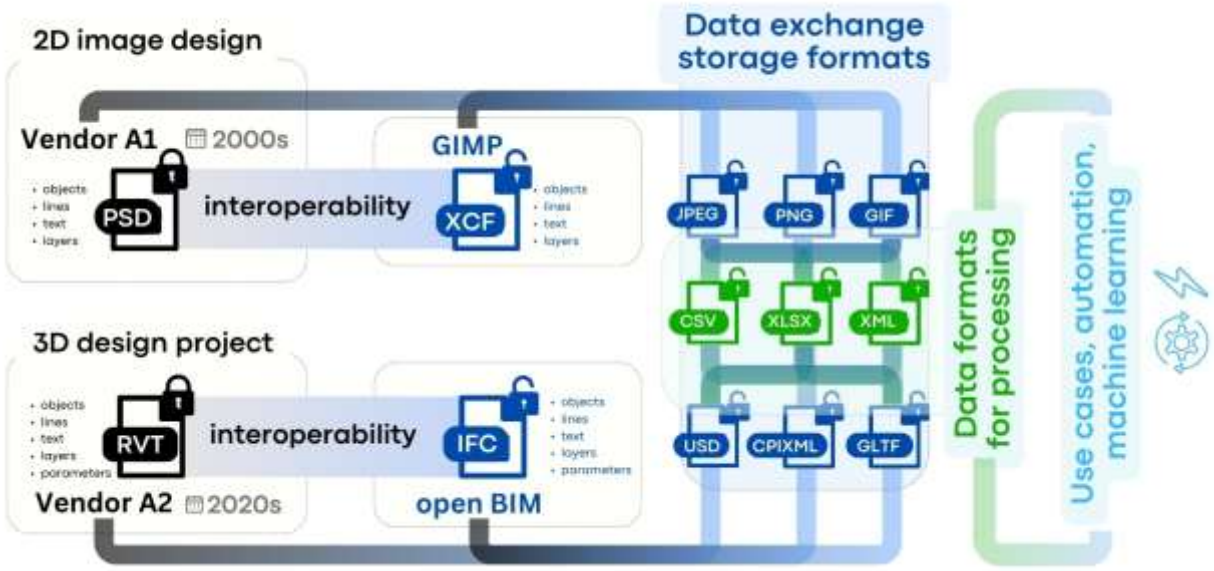


图 6.2-5 对于大多数用例，用户会选择独立于供应商程序的最简单格式。

JPEG、PNG 和 GIF 等平面图像格式的出现，摆脱了供应商内部引擎的冗余逻辑，推动了数以千计可互操作的图形处理和利用解决方案的发展。这导致了从润饰和过滤工具到 Instagram、Snapchat 和 Canva 等社交网络等各种应用的出现，在这些应用中，这些简化的数据无需与特定的软件开发商绑定即可使用。

> 设计 CAD 格式的标准化和简化将促进许多新的用户友好型独立工具的出现，用于施工项目。

从与封闭的几何内核相联系的供应商应用程序的复杂逻辑，到基于简化元素库的通用开放格式，为更加灵活、透明和高效的数据处理创造了先决条件。这也为参与施工过程的各方--从设计师到客户和维护服务人员--提供了获取信息的途径。

尽管如此，在未来几年内，CAD 供应商极有可能试图再次转移有关 互操作性和访问 CAD 数据库的争论。这已经涉及到 "新 "概念--如粒度数据、智能图、"联合模型"、云存储库中的数字孪生--以及建立行业联盟和标准，继续 BIM 和开放 BIM 的道路。尽管这些术语很有吸引力，但这些倡议可能会再次成为将用户留在专有生态系统中的工具。其中一个例子是，USD （通用场景描述）格式被积极推广为 2023 年起跨平台 CAD

（BIM）协作的 "新标准"。

转至美元 和粒度数据 **m**

2023 年 AOUSD 联盟[124]的出现标志着建筑行业的一个重要转折。我们正在目睹一个新的现实的开始，它由 CAD 厂商通过几个重大变化来塑造，以处理建筑数据。第一个重大变化涉及对 CAD -数据的认识。参与概念设计早期阶段的专业人员越来越认识到，在 CAD 环境中创建设计仅仅是一个起点。设计过程中生成的数据最终将成为分析、操作和管理对象的基础。这意味着这些数据必须能够在传统 CAD 工具之外的系统中访问和使用。

与此同时，领先开发商的方法也在发生变革。业界领先的 CAD 供应商，BIM 概念 和 IFC 格式 的创造者，其战略正在发生意想不到的转变。从 2023 年起，该公司将摒弃以独立文件存储数据的传统做法，专注于处理细粒度（规范化和结构化）数据，并转向以数据为中心的方法[125]。

> 供应商正在追随其他行业的历史趋势：大多数用户并不需要封闭的 CAD 格式（类似于 PSD）或复杂的参数化 IFC 文件（类似于带有图层逻辑的 GIMP）。他们需要的是可以用于 CAFM（建筑 Instagram）、ERP（Facebook） 以及其他成千上万个充满 Excel 电子表格和 PDF 文档的流程的简单对象图像。

目前建筑行业的发展趋势可能是逐渐放弃参数化和复杂的格式，转而采用通用性更强的独立格式 USD, GLTF, DAE, OBJ （元信息既包含在混合格式中，也包含在单独的结构化或松散结构化格式中）。历史上的领导者，包括曾在 20 世纪 90 年代中期积极推广 IFC 的主要设计公司，现在都在公开推广新的 USD 格式[93]，强调其简洁性和通用性（图 6.2-6）。USD 在产品中的大规模应用、GLTF 兼容性以及与 Blender、虚幻引擎和 Omniverse 等工具的积极整合，都显示出一种新的数据工作范式正在形成。随着欧洲通用 ERP 中使用的欧洲统一美元格式 CPIXML 等本地化解决方案的普及，有可能加强美元在中欧的地位。参与开发 IFC 格式 的机构已经开始将其战略调整为美元格式[126]，这也证实了这一转变的必然性。

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

图 6.2-6 IFC 和 USD 格式的技术规格比较。

> 在此背景下，USD 有可能成为事实上的标准，有望克服当前的许多局限性，主要涉及现有 CAD - (BIM-) 格式的复杂性及其解释对几何核的依赖性。

取代参数化和复杂的 CAD -formats 和 IFC - 简化的数据格式 USD, gLTF, DAE, OBJ 以及元素元信息 CSV, XLSX, JSON, XML 将因其简单性和灵活性而在建筑行业占据一席之地。

建筑行业当前的变化乍一看像是一种技术突破，与从过时的 IFC 过渡到更现代的 USD 有关。然而，值得考

虑的是，早在 2000 年，开发 IFC 的同一家 CAD 供应商就曾撰文指出 IFC 存在的问题以及访问数据库的必要性[65]，而现在则积极推动向新标准 USD 过渡。

> 在 "开放数据 "USD 和细粒度数据管理 "新 "概念的另一个表象背后，CAD 供应商开始推广的基于云的应用程序可能隐藏着供应商垄断项目数据管理的意图，用户会发现自己在格式选择上更多地与企业利益而非实际需求相关。

对关键事实的分析[93]表明，这些变化的主要目的与其说是为了方便用户，不如说主要是为了保持对生态系统和数据流的控制，以维护供应商的利益，因为 40 年来，他们从未能够提供 CAD 数据库的访问。

也许，现在是企业停止等待软件供应商提供新概念，专注于以数据为中心的自我发展的时候了。在通过逆向工程工具摆脱了数据访问问题之后，该行业将能够独立转向现代、免费和便捷的数据处理和分析工具，而无需强加新概念。

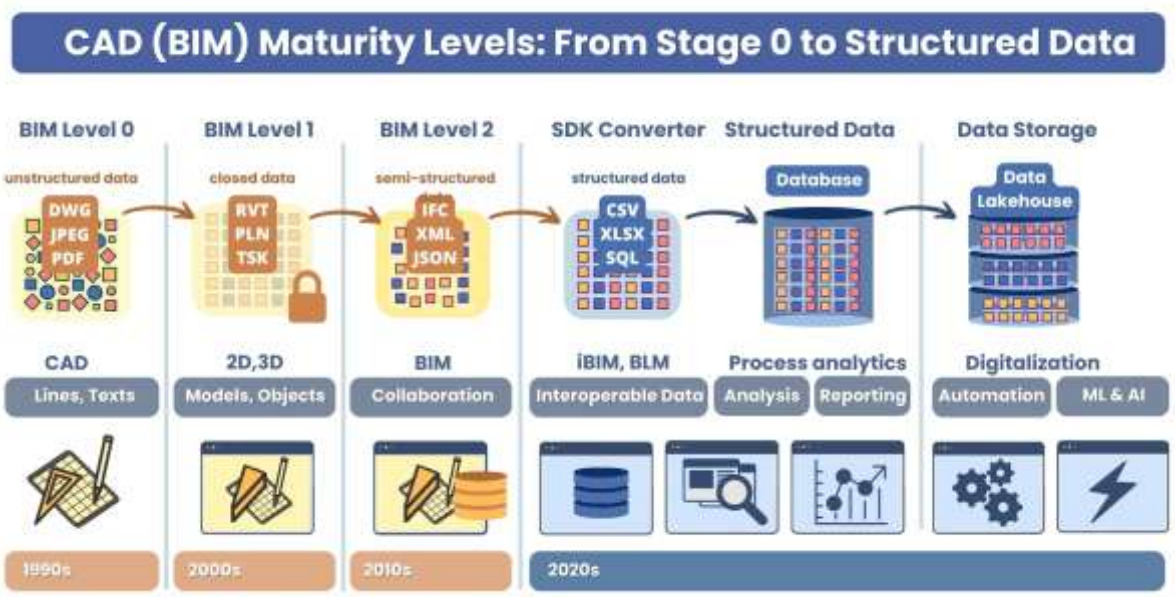


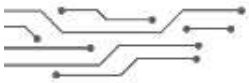
图 6.2-7 CAD 成熟度 (BIM): 从非结构化数据到结构化数据和存储库。

访问数据库 、开放数据和格式将不可避免地成为建筑行业的标准，无论供应商如何阻挠这一进程--这只是时间问题（图 6.2-7）。如果越来越多的专业人员开始熟悉开放格式、数据库工具和可用的逆向工程 SDK，从而可以直接访问 CAD 数据 - 系统，那么这一转变的步伐就会大大加快[92]。

> 未来在于开放、统一和可分析访问的数据。为了避免依赖供应商的解决方案，避免受制于封闭的生态系统，建筑和工程公司迟早要依靠开放性和独立性，选择能够完全控制数据的格式和解决方案。

当今建筑业所创建的数据将成为未来业务决策的关键资源。它将成为推动建筑公司发展和提高效率的战略 "燃料"。建筑业的未来在于利用数据的能力，而不是数据格式或模型的选择。

要了解开放格式 USD 、glTF、DAE 、OBJ 和专有参数化 CAD 格式之间的区别，就必须考虑可视化和设计计算中最复杂、最关键的数据元素之一--几何图形及其生成过程。要了解几何数据如何成为建筑分析 和计算的基础，有必要进一步探讨几何生成、转换和存储的机制。



章 6.3.

建筑几何：从线条到立方米

当线条变成金钱或建筑商为何需要几何图形

建筑中的几何图形不仅是一种可视化，也是精确定量计算的基础。在项目模型中，几何图形通过长度、面积和体积等重要的体积特征对元素参数列表（图 3.1-16）进行了补充。这些体积参数值通过几何核 自动计算，是估算、进度计划和资源模型的起点。正如我们在本书第 5 部分和 " 建筑项目的成本计算和估算 "一章中所讨论的那样，正是 CAD -模型中对象组的体积参数构成了现代 ERP、PMIS -系统/几何的基础，不仅在设计阶段，而且在项目实施管理、进度控制、预算编制和运行中都发挥着重要作用。就像几千年前建造埃及金字塔时，项目的准确性取决于立方和立方厘米等长度度量一样，今天 CAD - 程序中几何解释的准确性直接影响到结果：从预算和期限到承包商选择和交付物流

> 在竞争激烈、预算紧张的环境中，体积计算的准确性直接取决于几何形状，成为生存的关键因素。现代 ERP - 系统直接依赖于从 CAD - 和 BIM - 模型中获得的正确体积特征。因此，精确的构件几何描述不仅是一种可视化，更是管理成本和施工时间的关键工具。

从历史上看，几何一直是工程交流的主要语言。从纸莎草纸上的线条到数字模型，图纸和几何图形一直是设计人员、工头和估算人员之间交流信息的手段。在计算机出现之前，计算都是使用直尺和量角器手工完成的。如今，由于采用了体积建模技术，这项工作实现了自动化：CAD 软件的几何内核将线和点转换为三维体，并自动从中提取所有必要的特征。

在 CAD 程序中，通过 CAD-(BIM-) 程序的用户界面创建用于计算的几何元素。为了将点和线转化为体积体，需要使用几何内核，它的主要任务是将几何体转化为体积模型，并在近似后自动计算出元素的体积特征。

从线条到体积：面积和体积如何成为数据

在工程实践中，体积和面积是通过分析描述的几何曲面 或参数模型（如 BREP（边界元素表示法）框架内的 NURBS （非均匀有理 B-样条曲线））计算得出的。

> NURBS （Non-Uniform Rational B-Splines，非均匀有理 B-Splines）是一种描述曲线和曲面的数学方法，而 BREP 则是一种描述物体完整三维几何形状的框架，包括其边界，可以使用 NURBS 进行定义。

尽管 BREP 和 NURBS 非常精确，但它们需要强大的计算资源和复杂的算法。然而，根据这种数学上精确的描述直接进行计算通常在计算上比较困难，因此在实际应用中，几乎总是使用曲面细分（将曲面转换为三角形网格），从而简化了后续计算。曲面细分 是将复杂曲面划分为三角形或多边形。在 CAD /CAE 环境中，这种方法用于可视化、体积计算、碰撞搜索 、导出为 MESH 等格式以及碰撞分析。自然界中的一个例子是蜜蜂蜂巢，复杂的形状被分解成规则的网格（图 6.3-1）。

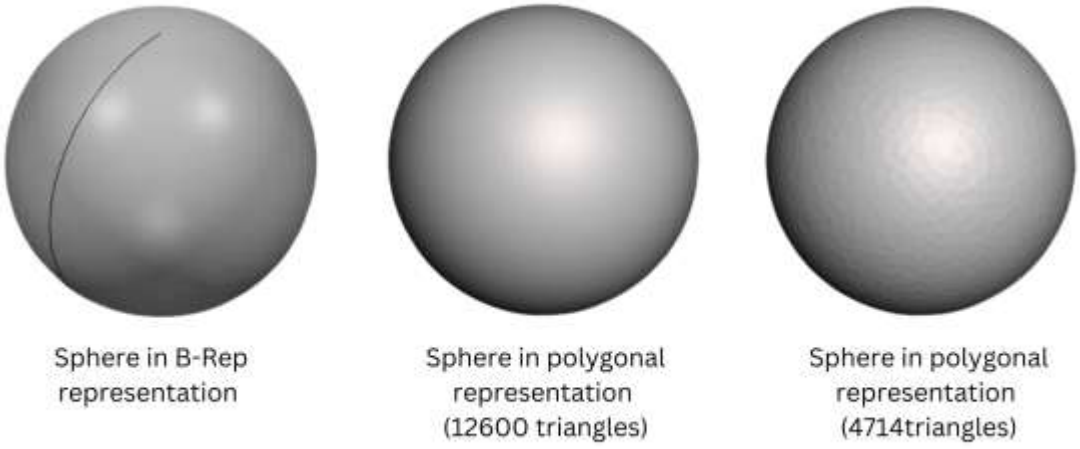


图 6.3-1 同一球体在参数描述 BREP 和多边形表示法中三角形数量不同。

BREP （NURBS ）用于 CAD ，并非几何的基本模型。它是一种方便的工具，用于表示圆和有理花键，并最大限度地减少几何数据的存储量。然而，它也有局限性，例如无法准确描述螺旋线和曲面的正弦曲线，而且需要使用复杂的几何核。

相比之下，三角形网格和参数图形的细分具有简单、内存使用效率高和能够处理大量数据的特点（图 6.3-2 ）。这些优点使得在计算几何图形时，无需使用复杂昂贵的几何内核 ，以及内嵌在几何内核中的数千万行代码。

> 在大多数建筑案例中，如何精确定义体积特征并不重要--通过参数模型（BREP、IFC）或多边形（USD, glTF, DAE, OBJ）。几何体仍然是近似的形式：无论是通过 NURBS 还是 MESH，它始终是对形状的近似描述。

定义为多边形或 BREP (NURBS) 的几何图形在某种程度上仍然只是对连续形状进行近似描述的一种方法。正如菲涅尔积分没有精确的解析表达式一样，通过多边形或 NURBS 进行几何离散化始终是一种近似方法，正如三角形 MESH 一样。

BREP 格式的参数化几何图形主要在以下情况下是必要的：对数据量的最小化要求很高，并且有 可能使用资源密集型和昂贵的几何图形内核进行处理和显示。这通常是 CAD 程序开发人员的特点，他们为此在其产品中应用 MCAD 供应商的几何内核。在这种情况下，即使在这些程序中，用于可视化和计算的 BREP 模型在分割 的过程中也经常被转换为三角形（类似于 PSD 文件被简化为 JPEG 文件的情况）。

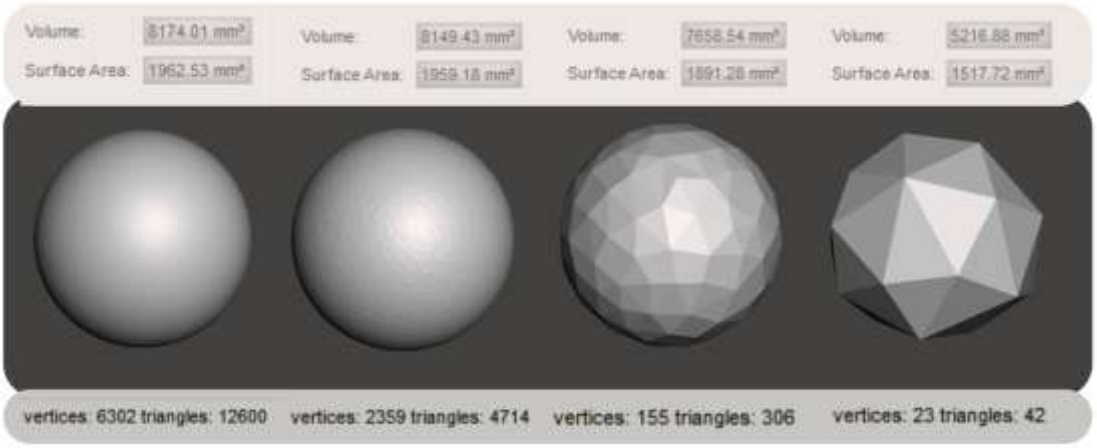


图 6.3-2 不同多边形数量的图形的体积特征差异。

多边形 MESH 和参数化 BREP 都有各自的优势和局限性，但目标是一致的--在考虑用户任务的情况下描述几何体。最终，几何模型的准确性不仅取决于其表示方法，还取决于特定任务的要求。

> 在大多数建筑问题中，对参数几何和复杂几何核的需求可能是多余的。

在每个特定的计算自动化任务中，值得考虑的是，参数几何的重要性是否被 CAD 开发人员夸大了，因为他们想推广和销售自己的软件产品。

转到 **MESH**、**USD** 和多边形：使用细分 绘制几何图形

在建筑行业中，在对设计信息和特征几何图形进行流式处理、开发系统、数据库或自动流程时，必须努力实现与特定 CAD 编辑器和几何图形内核的独立性。

> 计算部门 和施工现场使用的交换格式不应基于特定的 CAD (BIM) 程序。格式中的几何信息应直接通过细分表示，而无需参考几何核心 或 CAD 架构。

来自 CAD 的参数几何图形可视为中间来源，但不能作为通用格式的基础。无论如何，大多数参数描述（包括 BREP 和 NURBS ）都会转换为多边形 MESH ，以便进一步处理。如果结果相同（细分曲面和多边形）且处理过程更简单，那么选择是显而易见的。这类似于图形本体和结构化表格之间的选择（我们在第四部分讨论过）：过多的复杂性 很少是合理的（图 3.2-10，图 6.1-8）。

开放式格式如：OBJ 、STL、glTF、SVF 、CPIXML 、USD 和 DAE ，使用通用的三角形网格结构，具有显著的优势。这些格式具有极佳的互操作性--使用现有的开放源代码库就可以轻松读取和可视化，而不需要包含数百万行代码的复杂专业几何内核（图 6.3-3）。这些通用几何格式的应用范围很广，从宜家™中相对简单的厨房设计工具，到电影和 VR 应用中复杂的对象可视化系统，不一而足。这些格式的一个重要优势是有大量免费的开放源代码库，可用于大多数平台和编程语言。

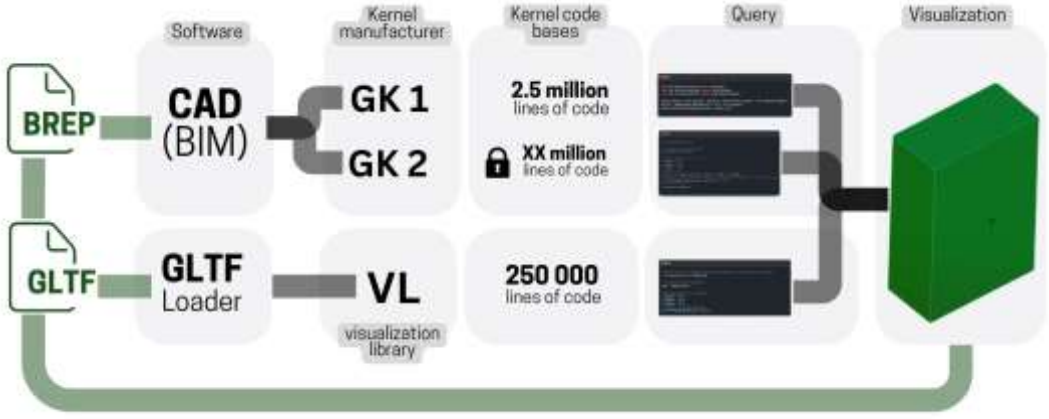


图 6.3-3 通过使用参数格式和几何内核 ，或通过使用三角形格式和开源可视化库，可以实现相同的几何表示。

由于几何内核不同，CAD - 供应商和用户本身都面临着解释外国参数 CAD 格式或开放 IFC 的问题。在实践中，所有 CAD -供应商无一例外地使用 逆向工程 SDK 在系统间传输数据，他们都不依赖 IFC 或 USD [93] 等格式来实现互操作性。

> 对于 CAD 解决方案的开发人员和用户来说，与其使用 CAD 厂商联盟推广的概念，而自己并不使用这些概念，不如集中精力了解每种方法在特定情况下的优势，并根据使用情况选择一种或另一种几何类型，这样会更有成效。在不同的几何表示法之间做出选择，需要在精度、计算效率和特定任务的实际需求之间做出权衡。

在处理设计数据时，传统上由大型供应商强加给建筑行业的与使用几何内核相关的复杂性往往是多余的。基于 MESH 几何图形的 USD 格式可以成为建筑行业的 "潘多拉盒子"，为开发人员在 IFC 和 CAD 供应商典型的参数 BREP 结构框架之外组织数据交换提供新的可能性。

在仔细研究了 USD、DAE、gLTF、OBJ 等格式的结构之后，我们会发现有一些更简单的开放式格式可以有效地组织几何信息的传输和使用，而无需依赖复杂的参数和封闭的几何内核。这种方法不仅降低了开发人员的技术门槛，而且有利于开发灵活、可扩展和真正开放的数字建筑解决方案。

LOD, LOI, LOMD - CAD 中 细部的唯一分类 (BIM)

除了几何表示格式外，在不同行业使用不同详细程度和深度数据的世界中，CAD - (BIM-) 方法提供了自己独特的分类系统 ，这些系统构建了为建筑模型提供信息的方法。

标准化新方法的实例之一是引入模型开发等级，反映图形和信息组件的准备程度和可靠性。为了区分 CAD - (BIM-) 数据工作中的信息内容，出现了 LOD （详细程度）- 模型图形部分的详细程度，以及 LOI （信息程度）- 数据详细程度。此外，在综合方法中还引入了 LOA（精度等级）和 LOG（几何等级）的概念，前者指的是所展示元素的精度，后者指的是图形表示的精度。

详细程度（LOD ）用从 100 到 500 的数字表示，反映了模型开发的程度。LOD 100 是具有一般形状和尺寸的概念模型。LOD 200 包括更精确的尺寸和形状，但有条件的细节。LOD 300 是具有精确尺寸、形状和元素位置的详细模型。LOD 400 包含元素制作和安装所需的详细信息。LOD 500 反映了设施建成后的实际状况，用于运行和维护。这些级别描述了 CAD（BIM）模型在生命周期不同阶段的信息饱和结构，包括 3D 、

4D 、5D 以及更多。

在实际项目中，高层次的细节（LOD400）往往过多，使用 LOD100 几何图形甚至平面图纸就足够了，而其余数据可以通过计算或从可能没有明显几何图形的相关元素中获取。例如，空间和房间元素（房间元素类别）可能没有可视几何图形，但包含大量信息和数据库，许多业务流程都是围绕这些信息和数据库建立的。

> 因此，在开始设计之前，必须明确定义所需的详细程度。对于 4D -7D 用例，即使是 DWG 图纸和最小 LOD100 几何图形通常也已足够。需求流程的关键任务是在模型的丰富性和实用性之间找到平衡。

从本质上讲，如果我们将 CAD（BIM）数据视为数据库（确实如此），那么通过新的缩略语对模型饱和的描述不过是信息系统的逐步数据建模，从概念层开始，到物理层结束（图 6.3-4），这在本书的第三和第四部分中有详细讨论。LOD 和 LOI 的每一次增加都意味着增加了新任务所需的信息：计算、施工管理、运行，其特点是以各种参 数的形式用附加信息层（3D -8D ）连续丰富模型，我们将在本书的第五部分进行讨论。

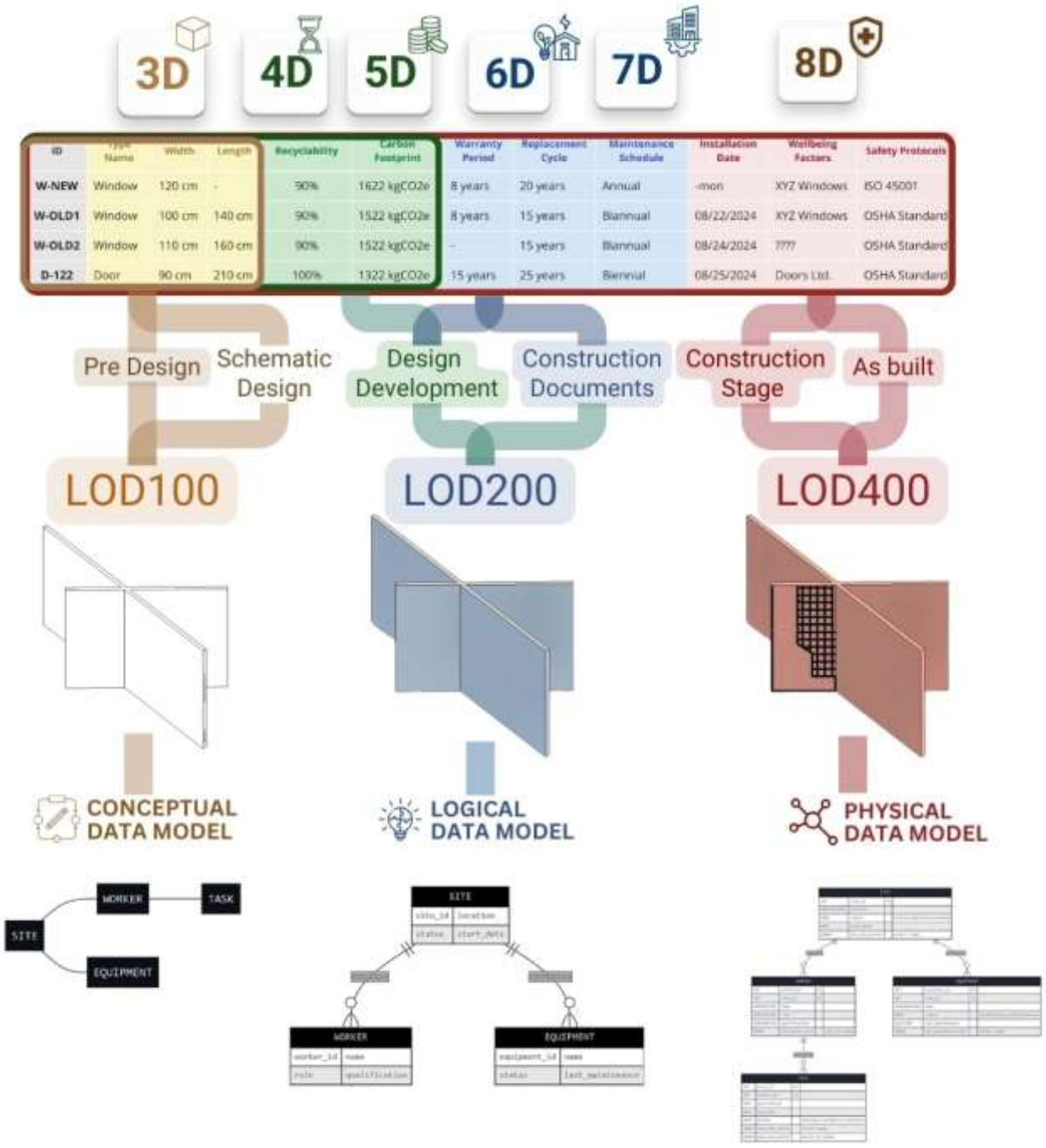


图 6.3-4 充实项目细节的过程与从概念到物理数据模型的数据建模过程相同。

几何图形只是设计数据的一部分，在建筑项目中并不一定需要这些数据，使用 CAD - 数据的关键问题并不在于如何将模型可视化，而在于如何在 CAD (BIM -) 程序之外使用这些模型的数据。

到 2000 年代中期，建筑行业面临着前所未有的挑战，管理和数据处理系统中的数据量迅速增加，尤其是来自 CAD 部门的数据 (BIM)。数据量的急剧增加让公司管理人员措手不及，他们对数据质量和管理方面日益

增长的需求毫无准备。

新的 **CAD** 标准 **(BIM) - AIA, BEP, IDS, LOD, COBie**

利用 CAD 数据库 的不开放访问和数据处理市场的有限竞争，并利用与新的首字母缩略词 BIM 相关的营销活动，参与开发 CAD 数据处理方法的组织已开始创建新的标准和概念，这些标准和概念在法律上应旨在改进数据管理实践。

> 尽管计算机辅助设计（CAD）供应商和开发商 (BIM) 直接或间接支持的几乎所有倡议都以优化工作流程为目标，但这些倡议导致各利益相关方游说制定了大量标准，从而导致建筑行业在数据流程方面出现一些模糊和混乱。

除了 LOD、LOI、LOA、LOG 之外，让我们列举一些近年来在建筑行业出现的新数据标准：

- **BEP**（BIM 执行计划）--说明如何在项目中集成和利用 CAD （BIM），定义数据处理方法和流程。
- **EIR** 文件 **/AIA** （客户的信息要求 ）--由客户在招标前准备，包含 对承包商准备和提供信息的要求。它是相关项目中 BEP 的基础。
- **AIM** （资产信息模型）是 BIM 流程 的一部分。项目交付和完成后，数据模型被称为资产信息模型或 AIM。AIM 的目的是管理、维护和运营已实现的资产。
- **IDS** （信息交付规范）--规定了 的要求，以及建筑项目不同阶段所需的数据和格式。
- **iLOD** 是指 BIM - 模型中表示信息的详细程度 LOD。它定义了模型中信息的详细和完整程度，从基本的几何表示到详细的规格和数据。
- **eLOD** - LOD CAD 模型中单个元素的详细程度 (BIM)。它定义了每个元素的建模程度以及相关信息，如尺寸、材料、性能特征和其他相关属性。
- **APS**（平台服务）和主要 CAD 供应商的其他产品 (BIM) - 描述 创建链接和开放数据模型所需的工具和基础设施。

虽然实施 CAD（BIM）标准（如 LOD、LOI、LOA、LOG、BEP、EIR、AIA、AIM、IDS、iLOD、eLOD）的公开目的是提高数据管理质量和扩大自动化能力，但在实际使用中，这些标准往往会导致流程过于复杂和

分散。如果我们将 CAD（BIM）模型视为一种数据库，就会发现其中许多标准重复了其他行业在信息系统工作中长期使用的有效方法。这些举措非但没有简化和统一，反而经常造成额外的术语负担，阻碍真正开放和灵活的解决方案的实施。

值得注意的是，其中许多新概念实际上正在取代本书前几部分详细讨论过的建模和数据验证过程，而这些过程在其他经济部门早已使用。另一方面，在建筑领域，标准化进程往往朝着相反的方向发展--新的数据描述格式、新的标准和新的数据验证概念应运而生，但它们并不总能带来真正的统一性和实际适用性。因此，该行业非但没有实现处理过程的简化和自动化，反而面临着更多的监管和官僚主义（图 6.3-1），这并不总是有利于提高效率。

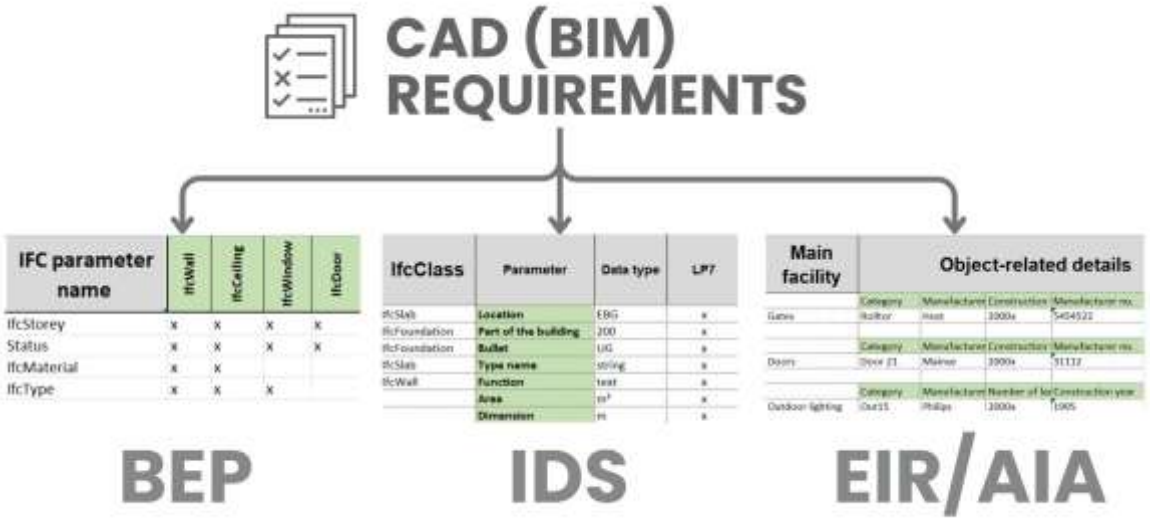


图 6.3-1 数据和信息内容要求简化为属性及其边界值的描述，使用表格进行描述。

> 与计算机辅助设计（BIM）数据有关的新概念不仅没有简化数据处理，反而往往在解释和基本定义阶段就产生更多的复杂性和争议。

新概念的最新实例之一是 IDS 格式 （2020 年推出），它可以描述开放式 BIM 概念中信息模型属性组成的要求 。IDS 要求以结构化表格（Excel 或 MySQL ）的形式描述属性信息及其边界值，然后转化为半结构化 XML 格式的标记 ，从 XML 更名为特殊缩写 IDS。

> 与供应商宣传并得到 BIM 和开放式 BIM 支持的观点相反，即由于使用 CAD 和 BIM 等专业工具，建筑业的数据处理是独一无二，该行业的数据格式和数据管理实践与其他行业并无不同

项目和 CAD (BIM) 格式的需求数量可以通过使用一个带有属性列的单一需求表来简化，详见 "将需求转化为结构化形式 "一章，而无需将原本结构化的 需求转化为非表格格式（IDS 最初是通过表格来描述的）。

简化方法（图 6.3-2）包括实体标识符列、属性列和边界值列，这些在前几章（图 4.4-9、图 4.4-16、图 7.3-10）中已详细讨论过，无需将需求转换为 IDS-XML 格式。这种方法为数据质量控制提供了一种直接、简便、透明的机制。它依赖于广泛使用的工具，从正则表达式（RegEx）到数据框、Pandas 和标准 ETL - payplanes - 与其他经济部门的专业人员处理数据时使用的工具如出一辙。

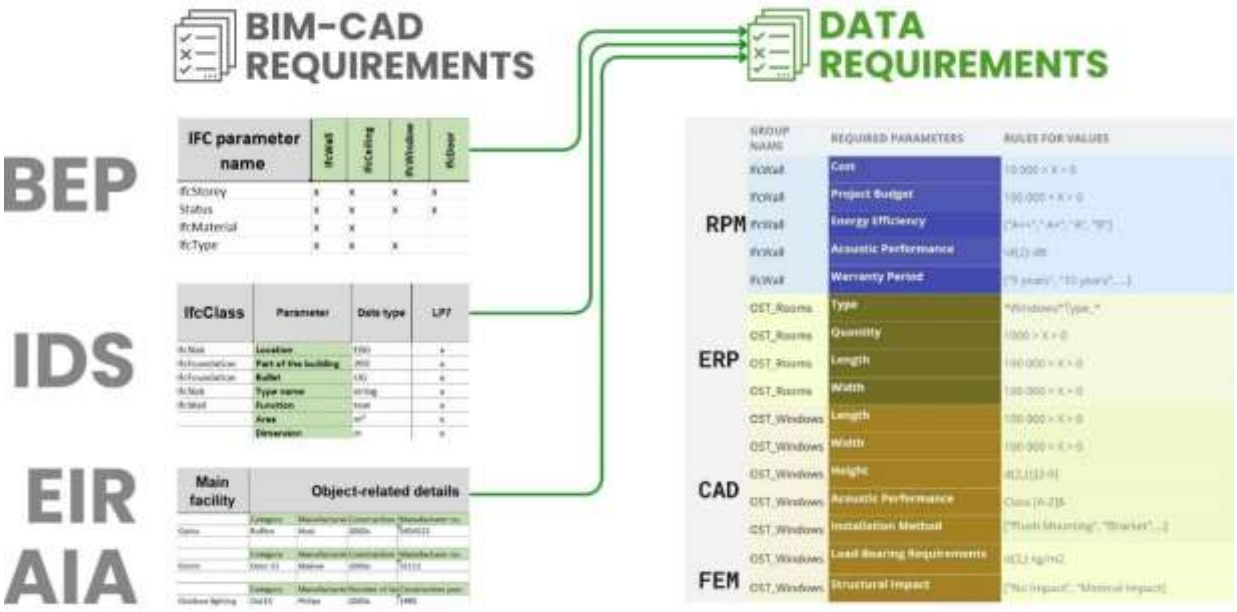


图 6.3-2 其他行业的数据要求简化为属性及其边界值的结构化描述。

随着时间的推移，由于数据的封闭性，建筑行业出现了越来越多的新方法和新技术来控制和管理这些多样化的数据，尽管建筑项目中的数据 与其他领域的数据基本相同。其他行业成功地采用了标准化的数据处理方法，而建筑业则继续开发新的、独特的数据格式、 要求和验证概念。

> 用于收集、准备和分析建筑数据的方法和工具不应与其他经济部门的专家所使用的方法和工具有本质区别。

该行业已经形成了一个独特的术语生态系统，需要进行批判性反思和重新评估：

- STEP 格式 被定位在新名称 IFC 下，并辅以建筑分类，但不考虑 STEP 格式本身的局限性。
- 尽管缺乏可视化和计算所需的统一几何核心，但在数据通信过程中使用了参数格式 IFC。
- 对数据库 CAD 系统的访问是在 "BIM "的术语下进行宣传的，但没有讨论这些数据库的具体内容和对它们的访问。
- 供应商通过 IFC 和 USD 格式来推广互操作性 ，但往往没有付诸实践，而是使用昂贵的逆向工程，而他们自己也很难做到这一点。
- LOD、LOI、LOA、LOG、BEP、EIR、AIA、AIM、IDS、iLOD、eLOD 等术语被普遍用于描述相同的实体参数，而不涉及其他行业长期使用的建模和验证工具。

建筑业的情况表明，尽管听起来很奇怪，但上述所有情况在建筑业都是可能发生的--尤其是如果主要目标是通过销售专业服务和软件来实现数据处理各阶段的货币化。从商业角度看，这样做并没有错。然而，这些与计算机辅助设计（CAD）（BIM）相关的缩写和方法是否真的能增加价值并简化专业流程，仍然是一个未决问题。

> 在建筑行业，这种制度之所以行之有效，是因为建筑行业本身的大部分投机利润都来自于这种迷宫般的制度和缩略语。对透明流程和开放数据感兴趣的公司很少。这种复杂的局面很可能会无限期地持续下去--直到客户、顾客、投资者、银行和私募基金开始要求更清晰、更明智的信息管理方法。

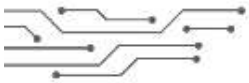
业界积累了过多的缩略语，但它们都在不同程度上描述了相同的流程和数据要求。这些缩略语在简化工作流程方面的实际作用仍然值得怀疑。

虽然概念和营销缩略语层出不穷，但数据要求验证流程本身将永远是业务流程不可分割的一部分。建筑行业应借鉴在金融、工业和 IT 等其他领域已被证明行之有效的工具，而不是在 上创建越来越多的专门格式和规定。

大量的术语、缩略语和格式给人造成一种数字建设流程十分详尽的错觉。然而，市场营销概念和复杂的术语往往掩盖了一个简单却不方便的事实：数据仍然难以获取、记录不全，而且与特定的软件解决方案紧密相连。

要摆脱这种为格式而格式的缩略语和格式的恶性循环，就必须把 CAD (BIM) 系统看作是专业数据库，而不

是神奇的信息管理工具。正是通过这面棱镜，我们才能理解市场营销的终点和真正的信息工作的起点。



章 6.4.

设计参数化和使用 **LML** 进行 **cad** 操作

CAD 数据唯一性的假象 **(BIM):** 通往分析和开放格式之路

现代 CAD (BIM) 平台极大地改变了设计和施工信息管理方法。以前，这些工具主要用于创建图纸和三维模型，而如今，它们已成为全面的项目数据存储库。在 "单一真实来源 "的概念下，参数化模型正日益成为项目信息的主要来源，而且往往是唯一的来源，以确保其在项目整个生命周期内的完整性和相关性。

> CAD - (BIM -) 平台与其他建筑数据管理系统的主要区别在于需要使用专门的工具和应用程序接口 来访问信息（唯一的真实来源）。这些数据库不是传统意义上的通用数据库：它们不是开放式结构和灵活的集成，而是一个封闭的环境，与特定的平台和格式硬连接。

尽管计算机辅助设计 - 数据的工作非常复杂，但有一个更重要的问题却超越了技术层面：计算机辅助设计数据库（BIM ）究竟是什么？要回答这个问题，有必要超越软件开发人员强加的通常缩略语和概念。相反，值得关注的是项目信息工作的本质：数据及其处理。

> 建筑业的业务流程不是从 CAD 或 BIM 工具开始的，而是从项目需求的形成和数据建模开始的。首先要定义任务参数：实体清单、实体的初始特征以及在解决特定任务时需要考虑的边界值。然后才根据指定的参数在 CAD（BIM）系统中创建模型和元素

在 CAD - (BIM-) 数据库中创建信息之前的过程与本书第四部分 "数据建模：概念模型、逻辑模型和物理模型 "一章中详细讨论的数据建模过程完全相同（图 4.3-1）。

正如在数据建模中，我们为以后要在数据库中处理的数据创建需求 一样，对于 CAD 数据库，管理人员以若干表列或键值对列表的形式创建设计需求 （图 6.4-1，步骤 1-2）。只有在这些初始参数的基础上，设计人员才能使用 API 自动或手动在 CAD- (BIM) 数据库中创建（或改进）对象（步骤 3-4），然后再次检查它们是否符合初始要求（步骤 5-6）。这个过程--定义→ 创建→ 验证→ 调整（步骤 2-6）--反复进行，直到数据质量与数据建模一样，达到目标系统--文档、表格或仪表盘--所需的水平（步骤 7）。

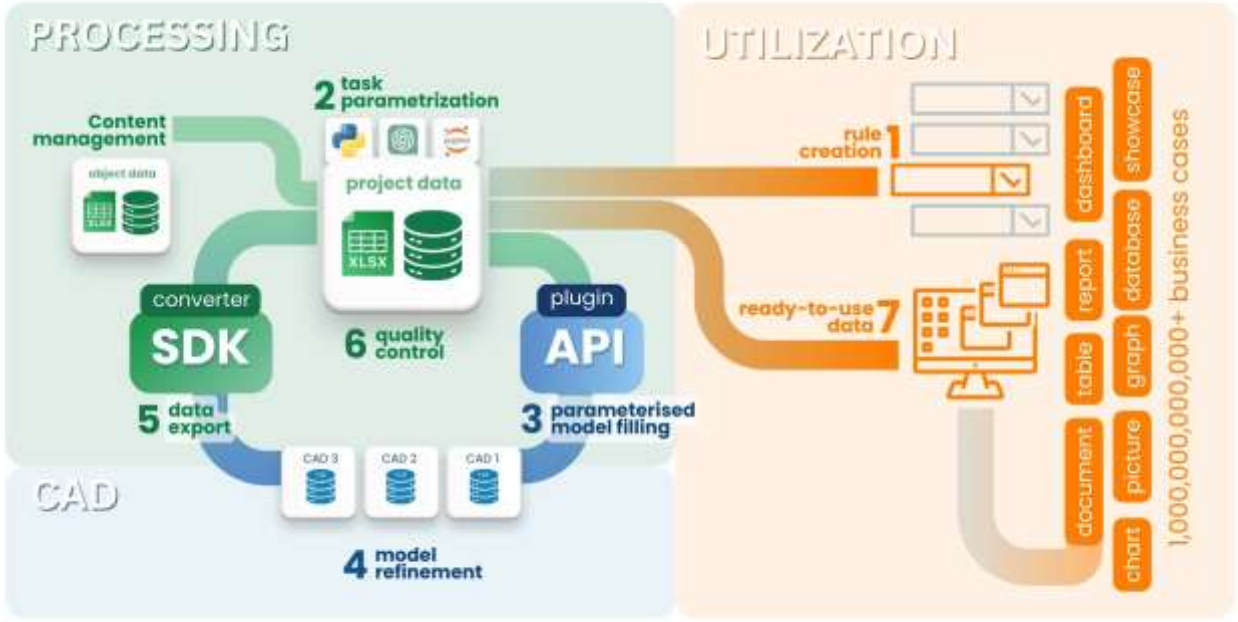


图 6.4-1 建筑项目实施过程中业务流程数据库信息饱和的周期。

如果我们将 CAD (BIM) 视为一种参数传输机制，其形式是根据在设计环境之外定义的要求生成的一组键值对（图 6.4-1，步骤 1-2），那么讨论的重点就会从具体的软件解决方案及其局限性转移到更基本的方面--数据结构、数据模型和数据要求。实质上，我们讨论的是数据库的参数饱和度和经典的数据建模过程（步骤 2-3 和 5-6）。唯一不同的是，由于 CAD 数据库的封闭性和所使用格式的特殊性，这一过程需要使用专门的 BIM 工具。问题来了：如果其他行业没有类似的方法，那么 BIM 的独特性何在？

过去 20 年来，BIM 的定位已不仅仅是一个单一的数据源。CAD -BIM 捆绑软件通常被推销为一种参数化工具，具有内在的集成数据库[64]，能够实现建筑项目设计、建模和生命周期管理过程的自动化。然而，在现实中，BIM 更多地是一种让用户留在供应商平台上的工具，而不是一种方便的数据和流程管理方法。

因此，CAD- (BIM-) 数据被隔离在其平台之外，将项目信息隐藏在专有的 API 和几何内核之后。这使得用户无法独立访问 数据库，也无法绕过供应商生态系统，提取、分析、自动处理数据并将数据传输到其他系统。

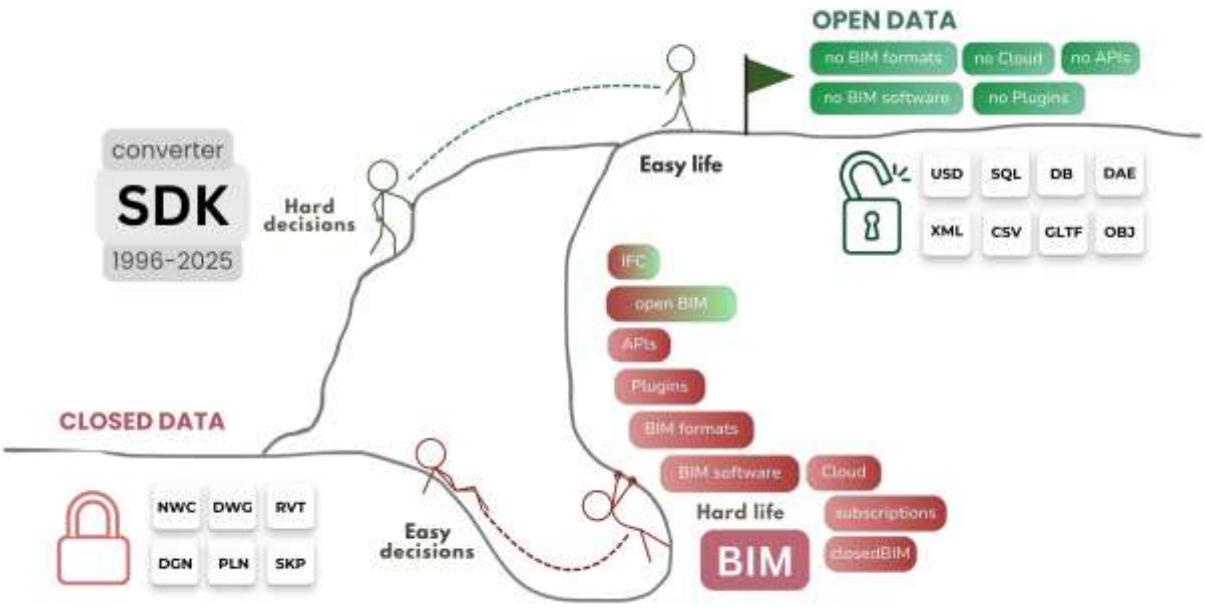


图 6.4-2 在建筑领域，现代格式需要复杂的几何内核 、每年更新的 API 以及 CAD -(BIM -) 软件的特殊许可证。

使用现代 CAD 工具的公司在处理数据时应采用所有 CAD 厂商自己在实践中无一例外使用的方法：使用 SDK - 逆向工程工具 进行数据转换，CAD 厂商自 1995 年以来一直在与这种方法作斗争 [75]。完全访问 CAD 数据库 并使用逆向工程工具，我们可以获得 [127] 一组带有属性的平面实体，并将其导出为任何方便的开放格式（图 6.4-2），包括设计元素的几何形状和参数。这种方法从根本上改变了处理信息的模式--从面向文件的架构转变为以数据为中心的架构：

- RVT 、IFC 、PLN 、DB1、CP2、CPIXML 、USD 、SQLite 、XLSX 、PARQUET 等数据格式都包含相同项目元素的相同信息。这意味着对特定格式及其模式的了解不应成为处理数据本身的障碍。
- 来自任何格式的数据都可以组合成一个单一的开放式结构和颗粒结构（图 9.1-10），其中包含 MESH 三角几何图形 和所有对象实体的属性，而不受几何内核的限制。
- 数据分析 努力实现通用性：使用开放数据，无论使用何种格式，您都可以 项目数据。
- 最大限度地减少对 API 和供应商插件的依赖：处理数据不再依赖于 API 技能。

当 和 CAD - 数据要求转化为易于分析的结构化表示格式时开发人员就不再依赖于特定的数据模式和封闭的生态系统。

通过参数进行设计：**CAD** 和 **BIM** 的未来

世界上没有一个建筑项目是在 CAD 程序中开始的。在 CAD 中形成图纸或模型之前，需要经过概念化阶段（图 6.4-1，阶段 1-2），重点是确定未来对象的基本思想和逻辑的参数。这一阶段相当于数据建模中的概念层（图 4.3-6）。参数可能只存在于设计者的头脑中，但理想的情况是以结构化列表、表格的形式排列或存储在数据库中（图 6.4-3），这样可以实现设计过程的透明度、可重复性和进一步自动化。

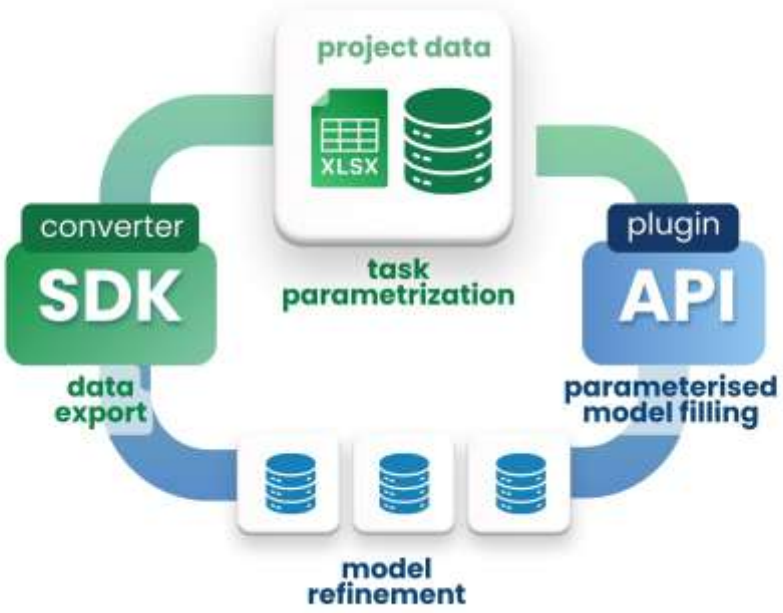


图 6.4-3 设计过程是一个迭代过程，利用价值链中的需求，将外部信息填充到 CAD 数据库。

在开始 CAD 建模之前（数据建模的逻辑和物理阶段（图 4.3-7）），必须定义作为项目基础的边界参数。与其他要求一样，这些属性是从数据利用链（如系统）的最末端收集的，通过它们，项目中未来对象的约束条件、目标和关键特征已经确定。

如果需求定义明确，借助参数化建模工具（图 6.4-3），建模本身可以完全自动化 60-100%。一旦项目以参数的形式描述出来，其形成在技术上就变得可行，例如，借助可视化编程语言，如 Grasshopper Dynamo，嵌入现代 CAD 环境或 Blender、UE、Omniverse 等免费解决方案。

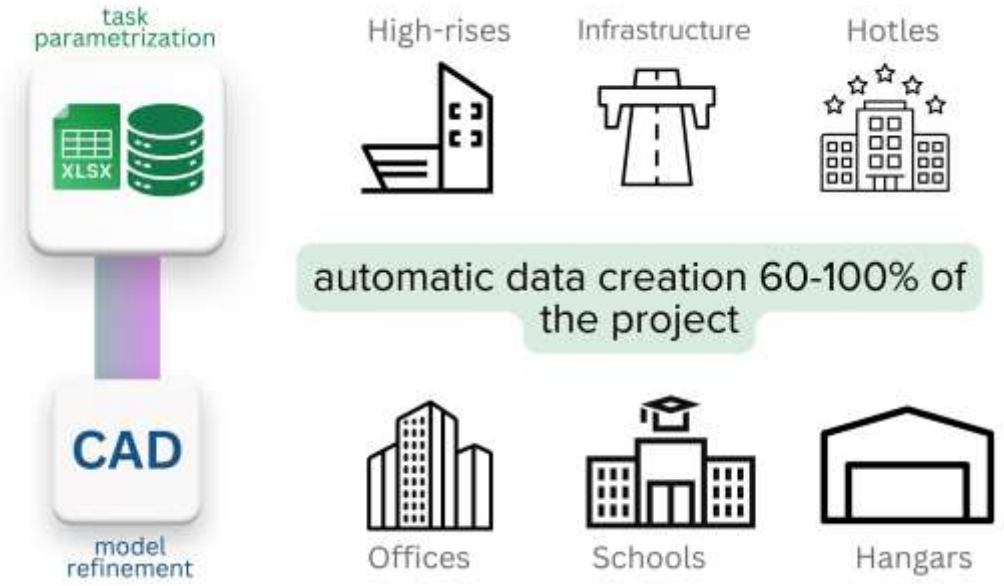


图 6.4-4 借助参数化编程工具，如今大多数类型化项目都已实现全自动创建。

如今，大型工业项目和类型化项目已不再由设计部门亲手制作，而是通过参数化工具和可视化编程来完成。这样就可以根据数据而不是某个设计师或经理的主观决定来建立模型。

> 内容先于设计。没有内容的设计不是设计，而是装饰 [128]。
>
> - Jeffrey Zeldman，网络设计师和企业家

这一过程不是从绘图或 3D -建模开始，而是从形成需求开始。正是需求 决定了项目中将使用哪些元素，哪些数据需要传输到其他部门和系统。只有存在结构化的需求，才有可能定期（例如，甚至每 10 分钟）自动检查模型，而不会分散设计人员的工作精力。

> 也许在未来，CAD (BIM-) 系统将仅仅成为一个填充数据库界面，而在哪种 CAD 工具中完成建模（物理层面）将不再重要。

同样，在机械工程领域，三维建模也经常被使用，但并不是项目的必要或强制要素。在大多数情况下，传统的二维文件就足够了，必要的信息模型是在其基础上创建的。该模型由根据行业标准构建的组件组装而成，包含了解设计和生产组织所需的所有信息。然后，利用工厂信息模型创建工厂信息模型，并在此基础上添加

特定产品和流程图，这些产品和流程图已经以技术专家的需求为导向。整个流程的组织不会产生不必要的复杂性，也不会让系统过多地使用没有实际优势的 3D 图形。

> 重要的是要明白，三维模型本身和 CAD 系统不应扮演主要角色--它只是定量和几何分析的工具。除了描述实体的几何参数外，其他所有参数都应尽可能在 CAD 环境之外进行存储和处理(BIM)。

通过参数进行设计不仅是一种趋势，也是建筑业不可避免的未来。设计人员将不再手动创建复杂的三维 模型，而是利用数据、验证数据并实现流程自动化，从而使建筑业更接近编程世界。随着时间的推移，设计流程将建立在软件开发的原则之上：

- 创建需求 → 创建模型→ 上传到服务器→ 验证 更改→ 拉取请求
- 作为拉动请求的一部分，拉动请求会根据设计前或设计过程中创建的需求自动运行模型检查
- 在数据质量检查 并获得批准后，将更改落实到项目、通用数据库或自动转移到其他系统中

目前，在机械工程领域，此类设计变更都是以创建变更通知开始的。建筑行业也将采用类似的方案：设计将是一个迭代的过程，每一步都有参数要求的支持。这样的系统将允许设计人员针对特定要求创建自动检查和自动拉动请求。

> 未来的设计师首先是数据操作员，而不是手工建模者。他的任务是在项目中填充参数化实体，几何形状只是其中的一个属性。

本书前几章详细讨论了数据建模、分类 和标准化的重要性，正是对这些重要性的理解将在转型中发挥重要作用。未来的设计规范将以 XLSX 或 XML -schemas 的形式作为键值参数对正式确定下来。

> 建筑业的未来是收集数据、分析数据、验证数据并利用分析工具实现流程自动化。BIM （或 CAD ）并非最终目标，而只是发展的一个阶段。当专业人员意识到他们可以绕过传统的 CAD 工具直接使用数据工作时，"BIM "一词本身将逐渐让位于使用结构化和细粒度建筑项目数据的概念。

加速转型的关键因素之一是大型语言模型（LLM ）和基于这些模型的工具的出现。这些技术正在改变处理

设计数据的方式，使人们无需深入了解 API 或供应商解决方案就能获取信息。有了 LLM，创建需求 和与 CAD 数据交互的过程变得直观易用。

设计 **CAD** 数据处理过程中出现的 **LLM**

除了开发 CAD 数据库访问工具以及开放和简化的 CAD -format 之外，LLM -tools（大型语言模型）的出现正在彻底改变设计数据的处理方式。过去获取信息主要通过复杂的界面，需要编程技能和 API 知识 ，而现在则可以使用自然语言与数据交互。

没有技术背景的工程师、管理人员和规划人员可以用普通语言进行查询，从项目数据中获取必要的信息。只要数据是结构化和可访问的（图 4.1-13），只需在 LLM 聊天工具中提出一个问题，如 "在按类型分组的表格中显示所有体积超过 10 立方米的墙体"，模型就会自动将该查询转换为 SQL 或 Pandas 中的代码，生成汇总表、图表甚至成品文档。

下面是一些实际例子，说明 LLM - 模型如何与以不同 CAD- (BIM-) 格式表示的设计数据进行交互。

- 将 RVT 格式（图 4.1-13）的 CAD 项目转换为表格数据帧（CHATGP、LlaMa 、Mistral 、DeepSeek 、Grok 、Claude 、QWEN 或任何其他格式）后的 LLM 聊天查询示例：

> 在对 "体积 "参数求和时，按 "类型名称 "对从 RVT 文件 中获得的数据帧中的数据进行分组，并显示分组中的元素数量。请以横向直方图的形式显示所有这些数据，且不显示零值。

LLM 响应 为水平条形图（PNG 格式）：

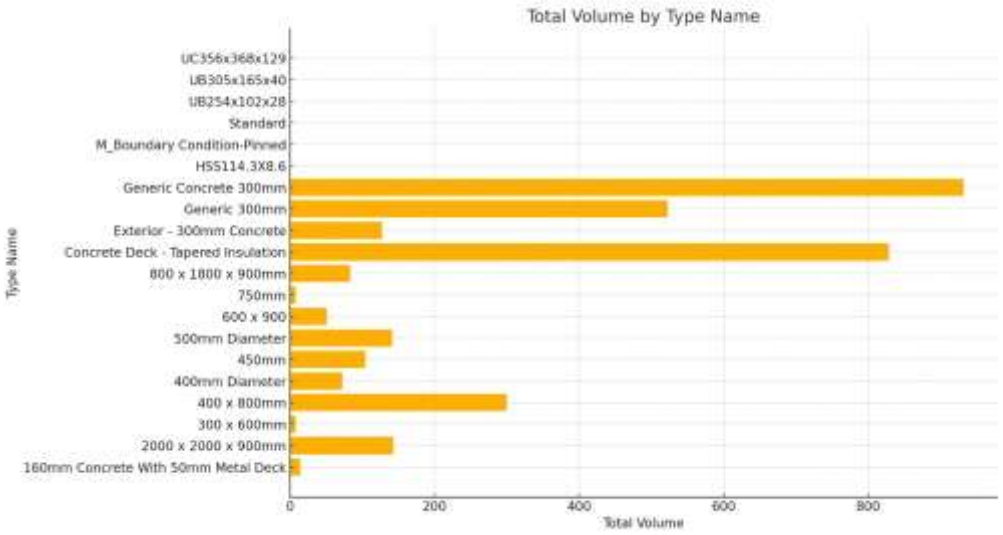


图 6.4-5 在 LLM 中，我们通过文本查询立即检索 QTO 表，而不是通过使用插件点击 17 次鼠标或编写 40 行代码。

为了从 "墙壁 "类别中生成一个包含总面积和数量的墙壁类型 QTO 表，让我们为 LLM -chat 生成一个文本查询：

> 从项目数据框中仅提取 "类别 "参数中包含 "OST_Walls "的项目，按 "类型名称 "对其分组，求 "面积 "列的值之和，加上数量，去除空值后将其显示在表格中。

LLM 以完成的 QTO 表格的形式回复：

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

图 6.4-6 使用自然语言创建 QTO 表，可获得与使用 CAD - (BIM-) 工具相同的质量结果。

让我们将 IFC 格式的项目转换为表格数据框后进行查询，并在任何 LLM 聊天工具中输入类似的文本查询：

> 从项目中仅提取 "父级 "参数中具有 1 级和 2 级值的项目，并提取 "类别 "参数中具有 IfcSlab 值的项目，然后按 "对象类型 "参数对这些项目进行分组，将 "PSet_RVT _Dimensions Area "参数中的值相加，并以饼图的形式显示出来。

- LLM 响应，作为来自 IFC 数据 的元素组别饼图成品：

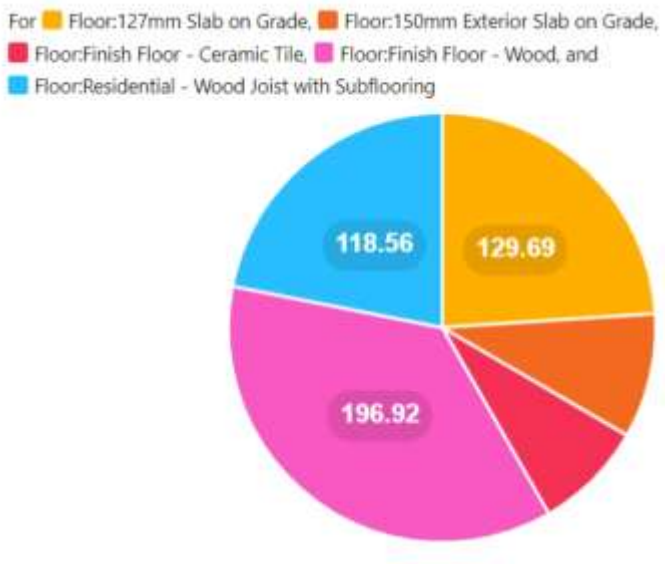


图 6.4-7 以结构化格式查询 IFC 数据的结果可以是任何类型的图形，以便于理解数据。

每个现成解决方案（图 6.4-5 - 图 6.4-7）的背后都是十几行使用 Pandas 库的 Python 代码。这些代码可以从 LLM 聊天室复制，并在任何本地或在线集成开发环境中使用，从而在 LLM 聊天室之外获得相同的结果。

在同一个 LLM 聊天室中，我们不仅可以处理从 3D CAD (BIM) 格式获得的项目，还可以处理 DWG 格式的平面图纸，我们可以通过 LLM 聊天室查询这些图纸，以显示以线条或 3D 几何图形形式转换为结构化形式后的元素组数据。

利用 **LLM** 和 **Pandas** 自动分析 **DWG -** 文件

由于信息的非结构化性质，从 DWG - 文件中进行数据处理的过程一直是一项复杂的任务，需要专门的软件和经常的人工分析。不过，随着人工智能 和 LLM 工具 的发展，许多步骤都可以实现自动化，而如今，这一过程大多是人工操作。让我们考虑一个真实的 Pipeline，请求 LLM（本例中为 ChatGPT）处理 DWG 图纸，这样就可以处理项目：

- 按图层、ID 和坐标过滤 DWG 数据
- 可视化元素的几何形状

- 根据参数自动注释图纸
- 将墙面折线扩展到水平面
- 创建交互式 3D - 平面数据可视化
- 无需复杂的 CAD - 工具即可构建和分析建筑数据

在我们的案例中，Pipeline 的构建过程是从通过 LLM 生成顺序代码开始的。首先，生成一个描述任务的查询。ChatGPT 生成 Python - 代码，执行并分析代码，在聊天室内显示结果。如果结果与预期不符，则更正请求并重复该过程

> 管道 是处理和分析数据的一系列自动化步骤。在这样的流程中，每一步都将数据作为输入，执行转换并将结果传递给下一步。

得到所需的结果后，从 LLM 复制代码，并以块的形式粘贴到任何一个方便的集成开发环境中的代码中，在我们的例子中就是 Kaggle 平台.com。由此产生的代码片段被组合成一个单一的 Pipeline ，实现从数据加载到最终分析的全过程自动化。这种方法可以快速开发和扩展分析流程，而无需深厚的编程专业知识。在 Kaggle.com 平台上搜索 "DWG Analyse with ChatGPT | DataDrivenConstruction"，可以找到下面所有片段的完整代码以及示例查询。[129].

在将 DWG 数据转换为结构化形式（图 4.1-13） 之后，让我们开始处理 DWG 数据的过程，其中最重要的一步是对所有绘图数据进行分组和过滤，对于我们的任务来说，墙壁元素是必需的，特别是折线（参数 "ParentID "允许将折线分组），在参数（数据框列）"Layer "中，有一个包含以下字母组合的字符串值（RegEx）--"wall"。

- 要获取类似任务的代码以及图片形式的结果，您应该在 LLM 中编写以下查询：

> 首先，检查从 DWG 获取的数据帧是否包含已定义的列：层"、"ID"、"ParentID "和 "点"。然后从 "图层 "列中筛选出包含 "墙 "字符串的 ID。查找 "ParentID "列中与这些标识符匹配的项目。定义一个函数来清理和分割 "点 "列中的数据。这包括删除括号并将值分割为 "x"、"y "和 "z "坐标。使用 matplotlib 绘制数据图。针对每个唯一的 "ParentID"，绘制连接 "点 "坐标的单独折线。尽可能确保第一个点和最后一个点相连。设置适当的标签和标题，确保 x 轴和 y 轴的比例相同。

答案 LLM 将为您提供一张现成的图片，图片背后隐藏着生成图片的 Python 代码：

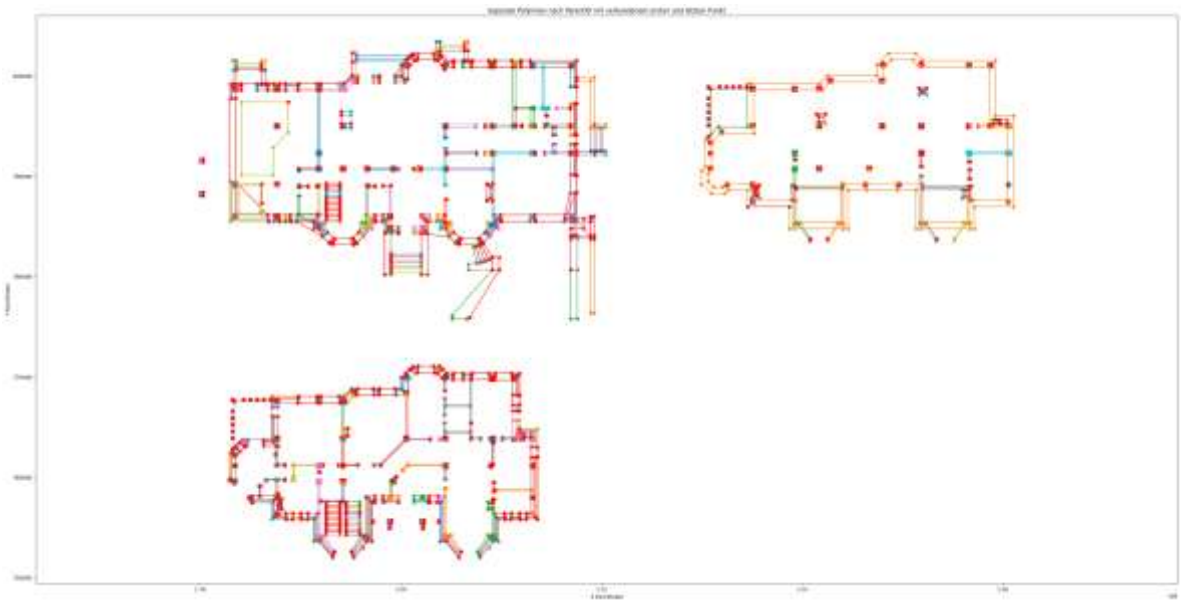


图 6.4-8 LLM 代码从 DWG 文件中提取了 "墙 "图层的所有线条，清除了它们的坐标，并使用一个 Python 库构建了折线。

现在，让我们在线条中添加每条多段线属性中的面积参数（在数据帧的一列中）：

> 现在，只需从每条折线中获取一个 "ParentID"（父 ID）--在 "ID "列中找到该 ID，求出 "Area"（面积）值，除以 1,000,000 并将该值添加到图表中。

LLM 回复 将显示一个新图表，其中每条折线都有一个标题，标明其面积：

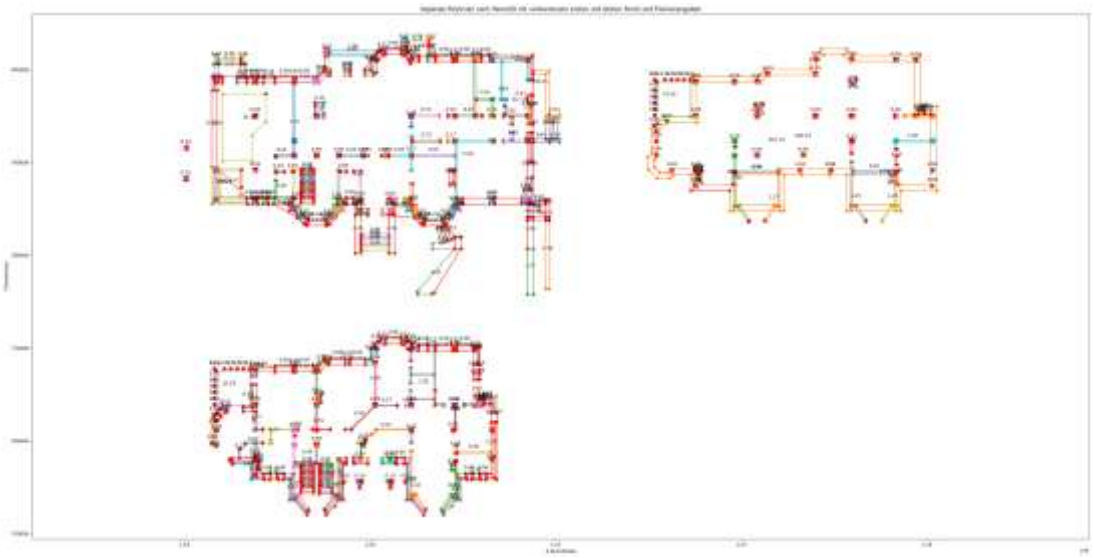


图 6.4-9 LLM 添加了代码，用于获取每条折线的面积值，并将其添加到图像中，使线条可视化。

然后，我们将把每条折线转换成一条水平线，再添加一条高度为 3000 毫米的平行线，并将它们连接成一个平面，以这种方式显示墙体构件表面的布局：

> 您需要从 "Layer"（层）列中提取所有值为 "wall "的元素。从 "ID "列中提取这些 ID 作为列表，然后从 "ParentID "列中的整个数据帧中查找这些 ID。所有元素都是组合成一条折线的线条。每条折线与 "点 "列中第一个点的 x、y 几何形状不同。您必须依次绘制每条折线，并从 0,0 点水平绘制折线上每段折线的长度。然后画出完全相同的线，只是高出 3000，将所有点连成一个平面。

LLM 响应 将输出允许您在平面上绘制墙面图纸的代码：

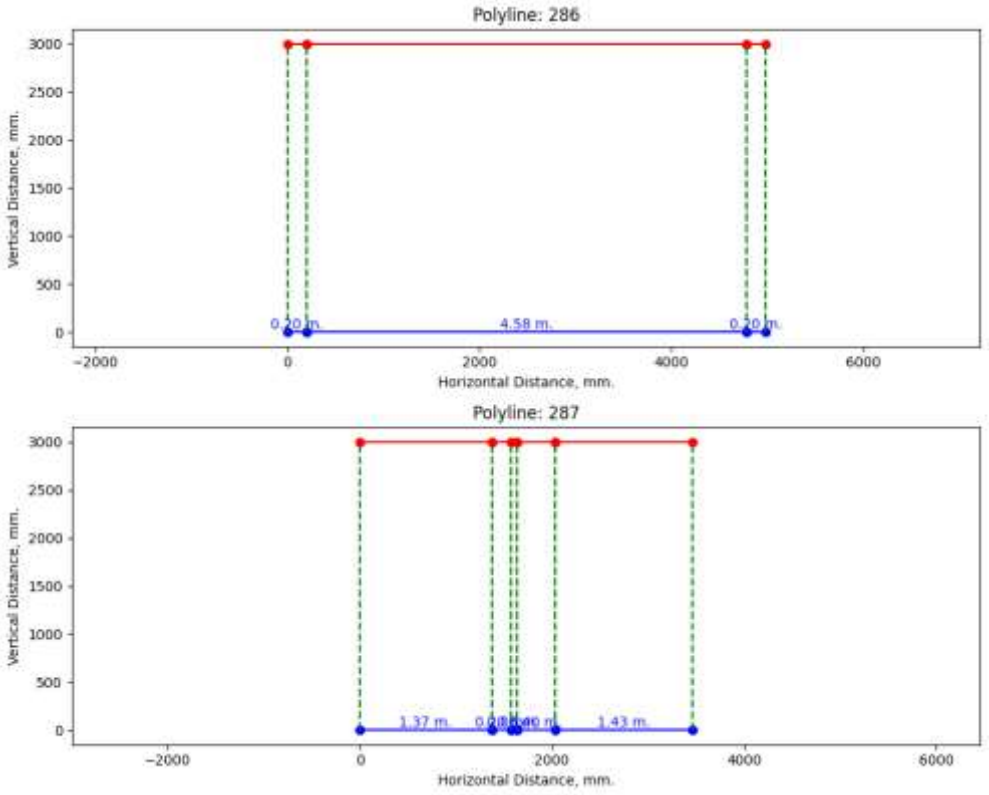


图 6.4-10 我们使用提示将每条折线转化为布局，直接在 LLM 聊天室中将墙面平面可视化。

现在，让我们从二维投影转到三维 - 通过连接顶层和底层的多段线，用平面线对墙壁进行建模：

> 在 3D 中可视化墙体元素，在高度 z = 0 和 z = 3000 毫米处连接多段线。创建表示建筑物墙壁的封闭几何体。使用 Matplotlib 三维图形工具。

- LLM 将生成一个交互式三维 -graph，其中每条折线将表示为一组平面。用户可以使用计算机鼠标在各元素之间自由移动，通过将代码从聊天工具复制到集成开发环境，以三维模式探索模型：

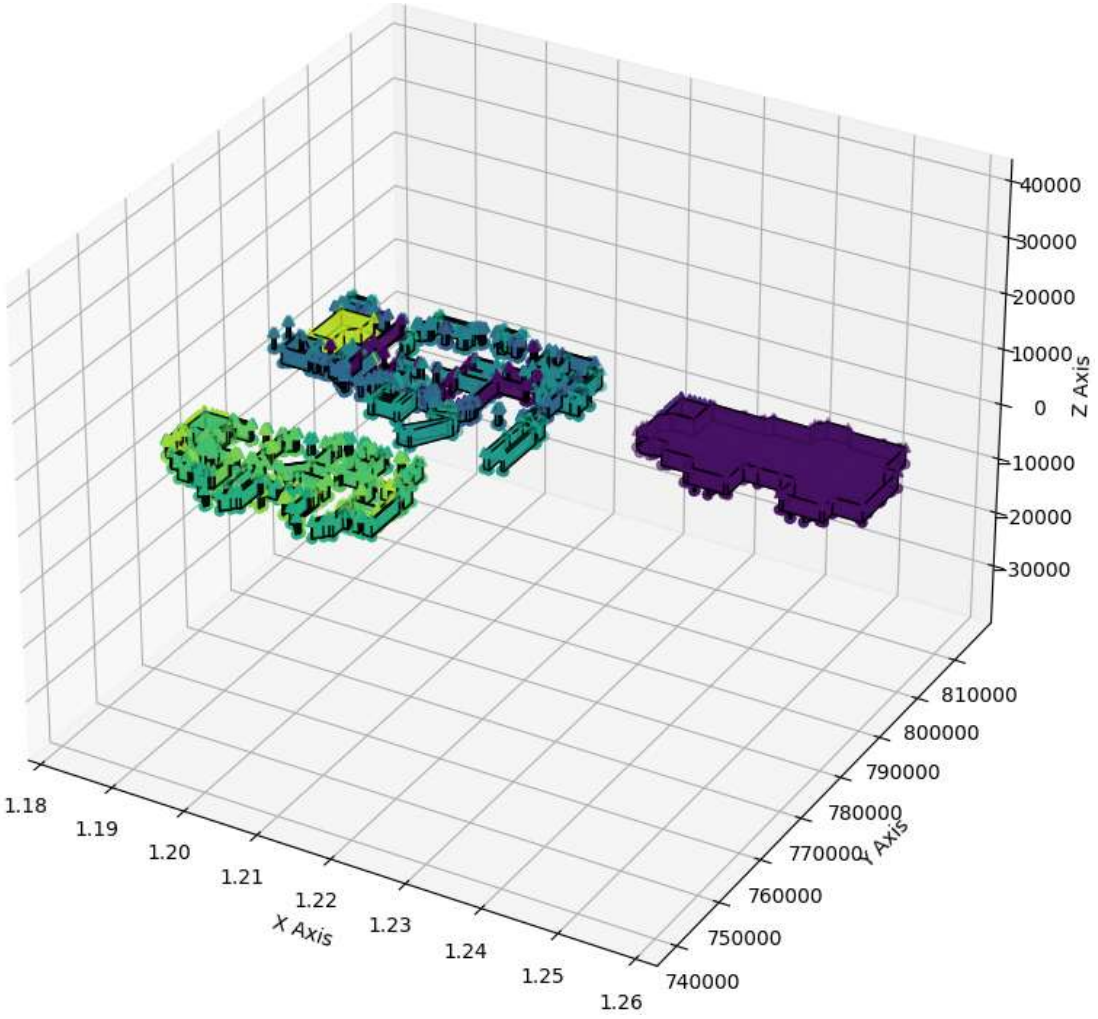


图 6.4-11 LLM 帮助构建代码 [129]，将平面绘图线条可视化为三维 视图，可在集成开发环境的三维查看器中进行探索。

为了建立一个合理且可重复的 Pipeline - 从最初的转换和加载 DWG - 文件到最终结果 - 建议在每个步骤后将生成的 LLM - 代码块复制到集成开发环境中。这样，您不仅可以在聊天中检查结果，还可以立即在开发环境中运行。这样，您就可以按顺序构建流程，并根据需要进行调试和调整。

您可以在 Kaggle 平台.com 上搜索 "DWG Analyse with ChatGPT | DataDrivenConstruction"，找到所有片段的完整 Pipeline 代码（图 6.4-8 至图 6.4-11）以及示例查询。[129].在 Kaggle 上，您不仅可以查看代码和使用的提示，还可以在云中使用原始 DWG 数据帧免费复制和测试整个 Pipeline，而无需安装任何其他

软件或集成开发环境本身。

通过本章介绍的方法，您可以完全自动化地检查、处理和生成基于 DWG 项目的文档。所开发的管道既适用于处理单个图纸，也适用于批量处理数十、数百或数千个 DWG 文件，并为每个项目自动生成必要的报告和可视化效果。

该流程可以按顺序以透明的方式构建：首先，CAD 文件中的数据会自动转换为 XLSX 格式的 ，然后加载到数据框中，接着是分组、检查和结果生成 - 所有这些都可以在任何流行的集成开发环境中通过一个 Jupyter 笔记本或 Python 脚本实现。如有必要，还可通过与项目文件管理系统集成轻松扩展该流程：可根据指定标准自动检索 CAD 文件，将结果返回存储系统，并在结果准备就绪时通过电子邮件或信使通知用户。

> 使用 LLM 聊天室和代理处理设计数据，可减少对专业 CAD - 程序的依赖，使您无需手动与界面交互即可执行建筑设计分析和可视化，无需点击鼠标和记忆复杂的菜单导航。

随着时间的推移，建筑行业将越来越多地听到 LLM 、粒状结构化数据、数据框 和列式数据库。由各种数据库和 CAD 格式形成的统一二维 DataFrames ，将成为其他行业专家积极使用的现代分析工具的理想燃料。

自动化过程本身将大大简化--不再需要研究封闭式利基产品的 API ，也不再需要编写复杂的脚本来分析或转换参数，现在只需以一组单独文本命令的形式制定任务即可，这些命令将被折叠到所需编程语言的所需管道或工作流过程中，几乎可在任何设备上免费运行。不再需要等待 CAD- (BIM-) 工具供应商提供新产品、新格式、新插件或更新。工程师和建筑商将能够在 LLM 聊天工具和代理的协助下，使用简单、免费和易于理解的工具独立处理数据。

下一步：从封闭格式转向开放数据

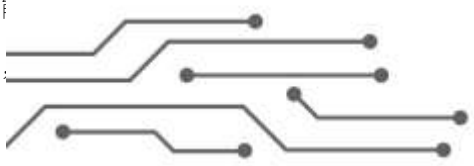
在处理未来的设计数据时，任何人都不太可能真正需要了解专有工具的几何内核，或学习包含相同信息的数百种不兼容格式。然而，如果不了解为什么转向开放式结构数据非常重要，就很难支持使用新的免费工具、开放式数据和软件供应商不太可能推广的方法。

在本章中，我们讨论了 CAD (BIM) 数据的主要特点、其局限性和机遇，并指出，尽管供应商在市场营销方面做出了承诺，但工程师和设计师在提取、传输和分析设计信息方面每天。了解这些系统的架构并学习其他方法--基于开放格式并通过 LLM 实现自动化--可以让单个专业人员的生活变得更加轻松，更不用说公司了。

在总结这部分内容时，值得强调的是主要的实际步骤，这些步骤将帮助您在日常工作中应用本节讨论的方法：

- 扩展您处理项目数据的工具包
  - 探索可用的插件和实用程序，以便从您使用的 CAD - (BIM-) 系统中提取数据
  - 探索可用的 SDK 和 API，让您无需手动打开专用软件即可自动从封闭格式中提取数据
  - 掌握使用开放式非参数几何格式（OBJ、glTF、USD、DAE）和相应开源库的基本技能
  - 尝试在 CAD (BIM) 解决方案之外建立一个与几何图形分离的项目元数据存储系统，以简化分析和与其他系统的集成。
  - 使用 LLM 自动处理格式之间的数据转换问题
- 创建自己的项目信息处理流程
  - 开始以简单和结构化的格式，通过参数及其值来描述任务和建模要求
  - 为经常执行的操作创建个人脚本库或代码块
- 在工作中推广使用开放标准
  - 邀请同事和合作伙伴以开放格式共享不受软件供应商生态系统限制的数据
  - 通过具体实例展示使用结构化数据的好处
  - 就封闭格式存在的问题和可能的解决方案展开讨论

，但如果您能亲自了解以开放格式处理项目数据的原则，就



法来提取和转换不同格式的数据，您不仅可以优化工作流程

VII 部分

数据驱动决策、分析、自动化和机器学习

第七部分侧重于建筑行业的数据分析和流程自动化。该部分讨论了数据如何成为决策的基础，并解释了可视化信息以进行有效分析的原则。详细介绍了关键绩效指标（KPI ）、投资回报评估方法（ROI ）以及创建项目监控仪表板。特别关注 ETL 流程 （提取 ，转换 ，加载 ）及其自动化，使用管道 (Pipeline) 将不同的数据转化为结构化信息以供分析。讨论了工作流协调工具，如 Apache Airflow 、Apache NiFi 和 n8n ，这些工具允许在没有深厚编程知识的情况下构建自动化数据管道。大型语言模型（LLM）及其在简化数据分析和自动化例行任务方面的应用正发挥着重要作用

章 7.1.

数据分析和数据驱动决策

经过信息的收集、结构化、清理和验证等步骤后，一个连贯的、可分析的数据集出现了。本书的前几部分介绍了从 PDF 文档和会议文本记录到 CAD 模型和几何数据等不同来源的系统化和结构化。书中详细描述了按照各种系统和分类器的要求检查和调整信息、消除重复和不一致的过程。

对这些数据进行的所有计算（本书第三、第四部分）--从简单的转换到时间、成本和环境、社会和治理指标的计算（第五部分）--都是综合分析任务。它们是了解项目现状、评估项目参数并做出决策的基础。因此，经过计算，数据从一组不同的记录转变为能够回答关键业务问题的可管理资源。

前几章详细介绍了用于典型业务案例的数据收集和质量控制流程，以及建筑行业特有的流程。这种情况下的分析在很多方面与其他行业的应用类似，但也有一些特殊之处。

以下各章将详细介绍数据分析过程，包括从最初获取信息、信息转换到随后将信息传输到目标系统和文件的自动化步骤。首先，将介绍理论部分，重点是数据驱动决策的某些方面。然后，在接下来的章节中，将介绍与 ETL 自动化和构建相关的实践部分 - Pipeline。

数据作为决策资源

数据驱动型决策通常是一个迭代过程，首先要从各种信息源系统地收集信息。就像自然循环一样，单个数据元素和整个信息系统逐渐落入土壤--在公司的信息库中积累起来（图 1.3-2）。随着时间的推移，这些数据就像落叶和树枝一样，被转化为有价值的材料。数据工程师和分析师的菌丝体组织和准备信息，以供未来使用，并将掉落的数据和系统转化为有价值的堆肥，从而长出新的嫩芽和新的系统（图 1.2-5）。

各行各业广泛使用分析 的趋势标志着一个新时代的开始，在这个时代，数据工作成为专业活动的基础（图 7.1-1）。建筑业专业人士必须适应这些变化，做好进入新时代的准备--数据和分析时代

> 在表格之间手动移动数据和手动执行计算已逐渐成为过去，取而代之的是自动化、数据流分析、分析和机器学习。这些工具正在成为现代决策支持系统的关键要素。

> 麦肯锡在《重启。McKinsey's Guide to Overcoming Competition in the Age of Digital Technology and Artificial Intelligence》[130]一书中引用了 2022 年对来自不同地区、行业和职能领域的 1330 名高级管理人员进行的一项研究。[该报告引用了 2022 年对来自不同地区、行业和职能领域的 1330 名高级管理人员进行的一项研究。研究结果显示，70% 的领导者使用高级分析技术来产生自己的想法，50% 的领导者使用人工智能来改进决策过程并使之自动化。



图 7.1-1 数据分析 和分析 是提高公司决策速度的主要工具。

数据分析 ，就像菌丝的传播一样，渗透到过去决策的腐殖质中，帮助连接各个系统，引导管理者获得有价值的见解。这些知识就像从腐烂的数据系统树木中汲取的养分一样，为公司的新决策提供养分，从而带来有效的变革和高质量的信息增长，就像从肥沃健康的土壤中长出的新芽和新苗一样（图 1.2-5）。

> 数字有一个重要的故事要讲述。它们寄希望于你为它们发出清晰而有说服力的声音[131]。
>
> - 数据可视化专家 Stephen Few

在中小型公司，为进一步分析而提取和准备信息的工作如今是一项极其劳动密集型的过程（图 7.1-2），堪比十八世纪的煤矿开采。直到最近，数据挖掘和准备工作仍是高度专业化的冒险家们的专利，他们只能使用一套小而有限的工具来处理来自非结构化、松散结构化、混合和封闭来源的不同类型数据。

决策者和管理者往往对异构数据和系统缺乏经验，但又需要根据这些数据和系统做出决策。因此，在过去几十年中，现代建筑行业的数据驱动决策不像一个自动化流程，而更像早期煤矿中矿工的多日手工劳动。

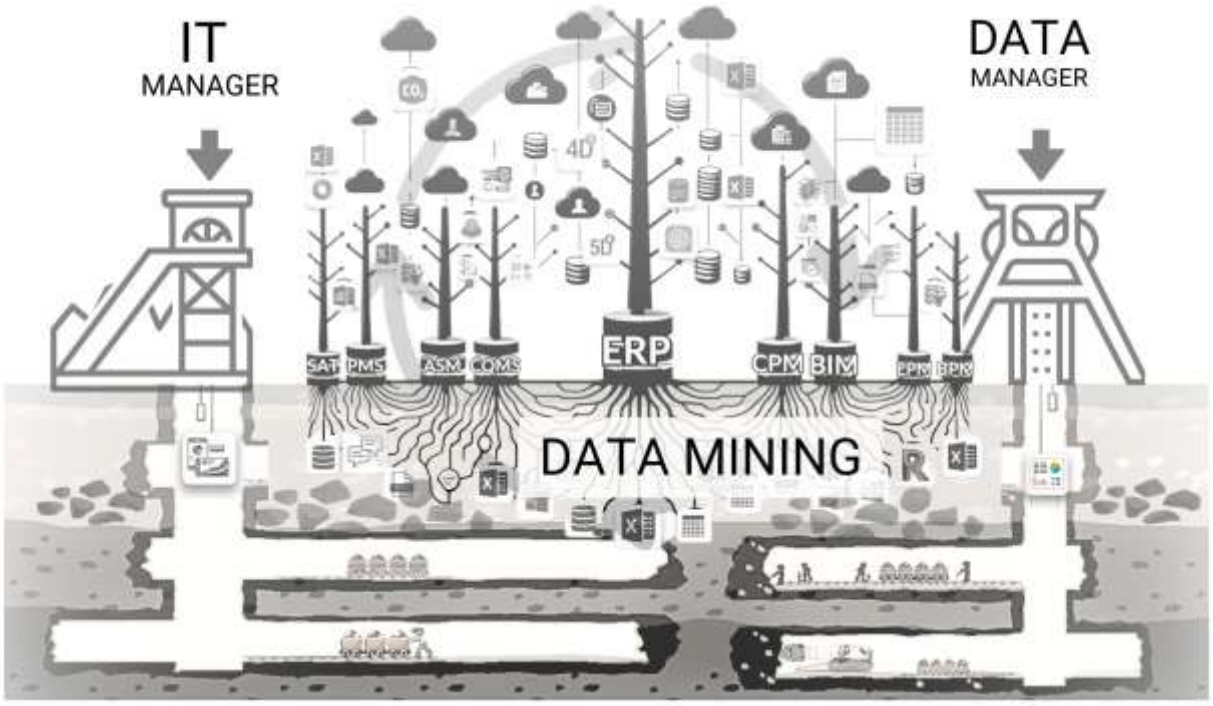


图 7.1-2 在数据挖掘过程中，专家们要经历复杂的数据准备过程--从清理到结构化，以便进行后续分析。

与 12 世纪矿工的原始技术相比，建筑行业的现代数据提取方法无疑更加先进，但这仍然是一项复杂且高风险的任务，需要大量的资源和专业知识，只有大公司才能负担得起。从过去项目积累的 数据中提取和分析数据的过程，直到最近仍主要由技术先进的大公司承担，这些公司已持续收集和存储数据数十年

以前，在分析领域发挥主导作用的是技术成熟的公司，它们已经积累了几十年的数据。如今，情况正在发生变化：数据和数据处理工具的获取正变得民主化--以前复杂的解决方案现在人人都能免费获得。

应用分析技术可以让公司实时做出更准确、更明智的决策。下面的案例研究说明了历史数据如何帮助做出合

理的财务决策：

- 项目经理 - "现在全市混凝土的平均价格是 82€ /m³，我们的估算是 95 欧元/m³"。
- 估算师--"在以前的项目中，超支率约为 15%，所以我打了退堂鼓"。
- 数据经理或客户方控制工程师--"让我们看看最近三次招标的分析结果"。

对过去项目的 DataFrame 进行分析后，我们得到了以下结果：

- 平均实际购买价格：84.80 /m³€
- 平均超支率： +4.7
- 建议估算率： ~ 85 /m³€

这种决定将不再基于主观感觉，而是基于具体的历史统计数据，这有助于降低风险，提高投标的有效性。对过去项目数据的分析成为一种 "有机肥料"，新的、更准确的解决方案就是从这种 "有机肥料 "中发芽的。

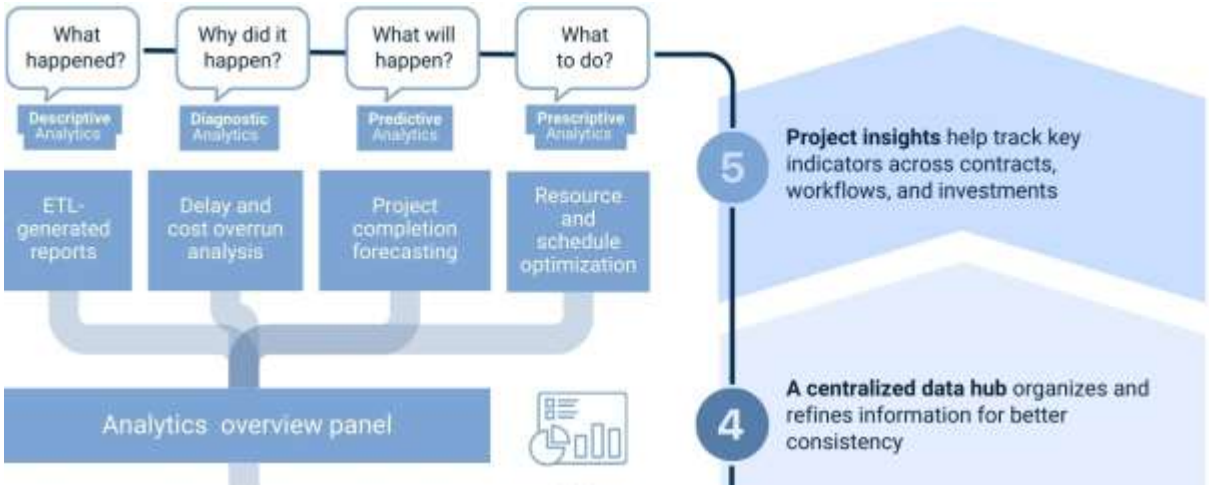


图 7.1-3 数据分析回答了三个关键问题：发生了什么、为什么会发生以及下一步应该做什么。

决策者和管理者经常需要处理各种数据和系统，却没有足够的专业技术知识。在这种情况下，可视化作为分析过程中最重要的第一步，是理解数据的关键辅助工具。它允许以可视和可理解的形式呈现信息。

可视化 数据：理解和决策的关键

在当今的建筑行业中，项目数据具有复杂性和多层次结构的特点，可视化在其中发挥着关键作用。 数据的可视化可让项目经理和工程师直观地看到隐藏在大量异构数据中的复杂模式和趋势。

> 将 数据可视化可以更容易地了解项目的状况：资源分配、成本趋势或材料使用情况。图形和图表使复杂枯燥的信息变得通俗易懂，让您能够快速识别需要关注的关键领域并发现潜在问题。

数据的可视化不仅有助于解读信息，还是分析过程和知情管理决策的关��步骤，有助��回答 "发生了什么？"和 "如�����生的？"等问题（图 2.2-5）。(图 2.2-5）。

> 图形是解决逻辑问题的可视化工具[132]。
>
> - Jacques Bertin，"图形和图形信息处理[133]。

在做出关键决策之前，项目经理更倾向于使用可视化的数据表示，而不是电子表格或文本信息中枯燥难解的数字。

> 没有可视化的数据就像建筑工地上胡乱散落的建筑材料：它们的潜力并不明确。只有当它们被清晰地可视化，就像用砖块和混凝土砌成的房子一样，它们的价值才会变得一目了然。在房子建成之前，我们无法断定这堆材料会变成一间小屋、一栋豪华别墅还是一座摩天大楼。

公司拥有来自各种系统的数据(图 1.2-4 至图 2.1-10)、财务交易和大量文本数据。然而，如何利用这些数据为企业带来效益往往具有挑战性。在这种情况下，可视化成为传达数据含义的重要工具，有助于以任何专家都能理解的格式（如仪表盘、图形和图表）呈现信息。

> 普华永道的研究报告《学生需要什么才能在快速变化的商业世界中取得成功》（2015 年）强调[9]，成功的公司不仅会进行数据分析，还会积极使用图表、信息图和分析仪表盘等交互式可视化工具来支持决策制定。报告指出--数据可视化通过图表、仪表盘和交互式数据模型帮助客户了解数据所讲述的故事。

将信息转换为可视化图形形式（如图表、图形和示意图）的过程可提高人脑对数据的理解和解释能力（图 7.1-4）。这使项目经理和分析人员能够更快地评估复杂的情况，并根据可视化的趋势和模式而不是直觉做出明智的决策。

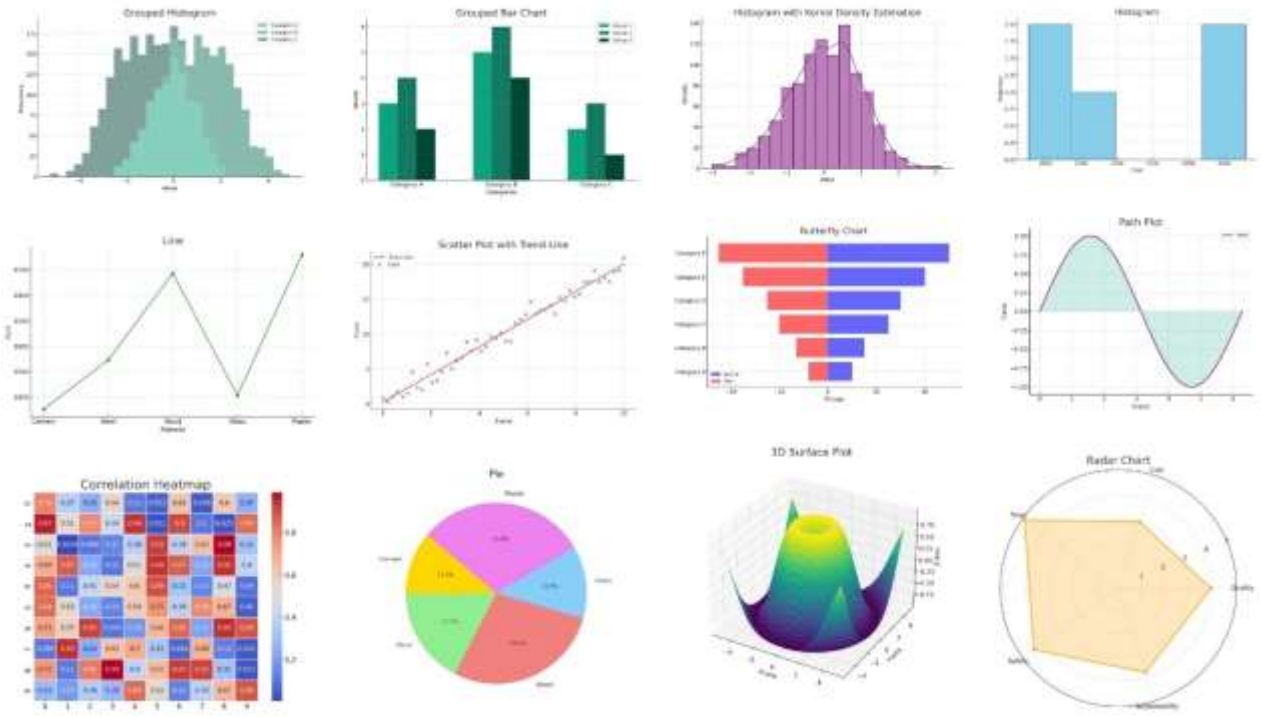


图 7.1-4 各种类型的可视化旨在帮助人脑更好地理解和感知枯燥的数字信息。

下一章 ETL 流程中将详细讨论从数据创建可视化以及使用各种免费可视化库的问题。

可视化正在成为建筑行业数据工作中不可或缺的元素--它不仅有助于 "看到 "数据，还有助于理解数据在管理任务中的意义。然而，为了使可视化真正发挥作用，有必要事先确定哪些内容需要可视化，哪些指标对评估项目绩效真正重要。这就是关键绩效指标（KPI）和投资回报率（ROI）等绩效指标发挥作用的地方。没有这些指标，即使是最漂亮的仪表盘也有可能只是 "信息噪音"。

关键绩效指标 和投资回报率

在当今的建筑行业中，绩效指标（关键绩效指标 和投资回报率 ）的管理以及通过报告和仪表盘将其可视化，在提高生产率和项目管理效率方面发挥着关键作用。

> 与任何企业一样，建筑业也必须明确界定衡量成功、投资回报和绩效的标准。在获取各种流程的数据时，数据驱动型组织必须首先学会确定关键绩效指标 （关键绩效指标）--反映战略和运营目标实现程度的量化指标。

要计算关键绩效指标 ，通常要使用一个包括实际指标和计划指标的公式（图 7.1-5）。例如，要计算某个项目、员工或流程的单项关键绩效指标，可以用实际绩效除以计划绩效，再乘以 100%。

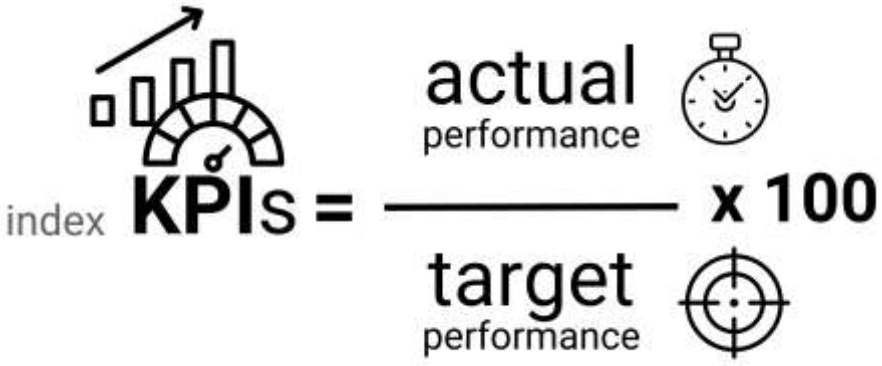


图 7.1-5 关键绩效指标 用于衡量项目或流程是否成功实现了关键目标。

在网站层面，可以使用更详细的 KPI 指标：

- 关键里程碑的时间安排（地基、安装、收尾）--以便控制工作计划的执行情况。
- 材料超支的百分比--有助于管理采购和尽量减少浪费。
- 计划外停机次数 - 影响生产率和成本。

选择错误的衡量标准会导致错误的 "做什么？"决策（图 2.2-5）。例如，如果一家公司只关注每平方米的成本，却不考虑改建成本，那么节省下来的材料可能会导致未来项目的质量下降和成本上升。

在制定目标时，必须明确衡量的内容。含糊的措辞会导致错误的结论，并使控制变得复杂。让我们来看看建筑业中成功和不成功的关键绩效指标。

良好的关键绩效指标：

- "到今年年底，将重新装修的比例降低 10%"。

- "到下一季度，在不影响质量的前提下，将外墙安装速度提高 15
- "通过优化工作计划，在年底前将机器停机时间减少 20%"。

这些衡量标准可以明确衡量，有具体的数值和时间范围。

糟糕的关键绩效指标：

- "我们将加快建设速度"（快多少？"快 "是什么意思？）
- "我们将提高混凝土工程的质量"（质量究竟如何衡量？）
- "我们将改进承包商在现场的互动"（改进的标准是什么？）

> 一个好的关键绩效指标 是可以衡量和客观评估的。在建筑业，这一点尤为重要，因为如果没有明确的指标，就不可能监控绩效并取得稳定的结果。

除了关键绩效指标 之外，还有一个用于评估投资效果的额外指标：投资回报率 （**ROI**）--投资回报率指标反映了利润与投入资金之间的比率。投资回报率有助于评估引进新方法、新技术或新工具是否合理：从数字化解决方案和自动化（如图 7.3-2）到新建筑材料的使用。这一指标有助于根据投资对企业盈利能力的实际影响，做出进一步投资的明智决策

> 在建筑项目管理方面，如果公司的目标是衡量一个项目、技术或流程改进的投资回报，那么投资回报率 （投资回报率）就可以作为关键绩效指标（KPI ）之一。例如，如果正在实施一项新的施工管理技术，那么投资回报率就可以显示该技术对盈利能力的改善程度。

根据从各种来源收集的数据（如材料消耗、工时和成本）定期衡量关键绩效指标和投资回报率，可使项目管理人员有效管理资源并迅速做出决策。长期储存这些数据可以分析未来趋势，优化流程。

各种图表和图形用于可视化关键绩效指标、投资回报率和其他指标 ，这些指标通常被组合成仪表盘。

仪表板和仪表盘：将指标可视化，实现有效管理

各种图表和图形用于将指标和衡量标准可视化，通常被组合成数据展示和仪表盘。这些仪表盘可集中显示项

目或项目部分的状态，显示关键指标（最好是实时显示）。最新和持续更新的仪表盘可以让团队对变化做出快速反应。

> 仪表板 是将定量评估可视化的工具，便于所有项目参与者使用和理解。

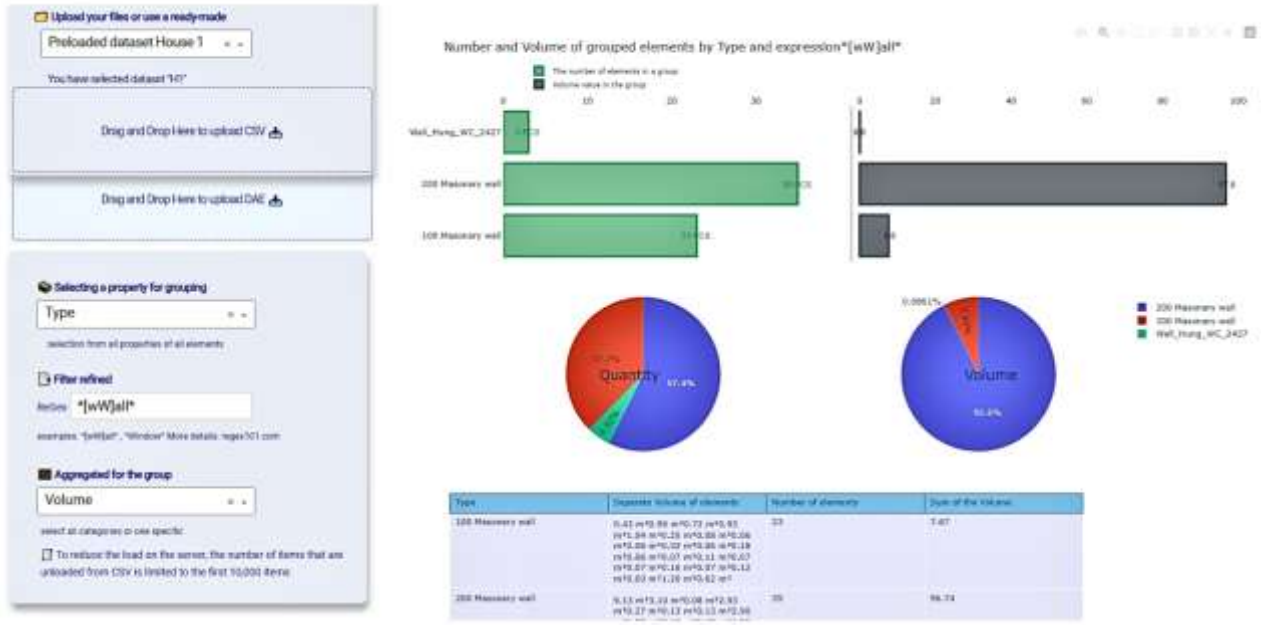


图 7.1-6 管理关键绩效指标 并通过仪表盘将其可视化是提高项目生产力和效率的关键。

下面是一些可以创建仪表盘的常用工具示例：

- **Power BI** 是微软推出的一款用于创建交互式报告和仪表盘的工具。
- **Tableau** 和 **Google Data Studio** 是无需编写代码即可实现数据可视化和创建仪表盘的强大工具。
- **Plotly** （图 7.1-6，图 7.2-12）是一个用于创建交互式图表的库，而 Dash 是一个用于创建分析数据的网络应用程序的框架。两者结合使用可创建交互式仪表盘。
- 众多 **Python** 库 （图 7.2-9 - 图 7.2-11）--Python 有许多 开源和免费的数据可视化库，如 Matplotlib 、Seaborn 、Plotly 、Bokeh 等。这些库可用于创建图表，并使用 Flask 或 Django 等框架将其集成到网络应用程序中。
- **JavaScript** 库： 允许您使用 D3.js 或 Chart.js 等开源 JavaScript 库创建交互式仪表盘，并将其集

成到网络应用程序中。

> 要评估关键绩效指标 并创建仪表板，您需要最新的数据以及收集和分析信息的明确时间表。

一般来说，建筑行业中的关键绩效指标 、投资回报率 和仪表板是项目管理分析方法的基础。它们不仅有助于监控和评估当前状态，还能为未来规划和优化过程提供有价值的见解--这些过程直接取决于对数据的解读以及及时提出正确的问题。

数据分析 和提问的艺术

数据解读是分析的最后阶段，在这一阶段，信息变得有意义并开始 "说话"。关键问题的答案就在这里形成："做什么？"和 "怎么做？"（图 2.2-5）。在这一阶段，可以总结结果、识别模式、建立因果关系，并根据可视化和统计分析得出结论。

> 也许不久的将来，人们就会意识到，要想完全成为一个正在发展中的新的复杂的世界大国的有效公民，就必须会计算，会从平均数、最大值和最小值的角度思考问题，就像现在必须会读会写一样[133]。
>
> - 塞缪尔-威尔克斯（Samuel S. Wilkes）在 1951 年向美国统计协会发表的主席致辞中引述

> 根据英国政府发布的《政府项目实施中的数据分析和人工智能》（2024 年）报告[83]，实施分析数据和人工智能 （AI ）可显著改善项目管理流程，提高时间和成本预测的准确性，并降低风险和不确定性。该文件强调，使用先进分析工具的公共组织在基础设施项目中取得了更高的绩效。

在第四次工业革命的激烈竞争和低利润环境中，现代建筑企业的运营可以比作军事行动。在这里，公司的生存和成功取决于获取资源和优质信息的速度，因此也取决于及时和知情的决策（图 7.1-7）。

如果说数据可视化是提供概览的 "情报"，那么数据分析 就是采取行动所需的 "弹药"。它回答了 "做什么 "和 "怎么做 "的问题，为赢得市场竞争优势奠定了基础。

> 分析将不同的数据转化为结构化和有意义的信息，并以此为基础做出决策。

分析师和管理人员的任务不仅仅是解读信息，还要根据项目的目标和具体情况做出明智的决策、识别趋势、确定不同类型数据之间的关系并对其进行分类。他们利用可视化工具和统计分析方法，将数据转化为公司的战略资产。

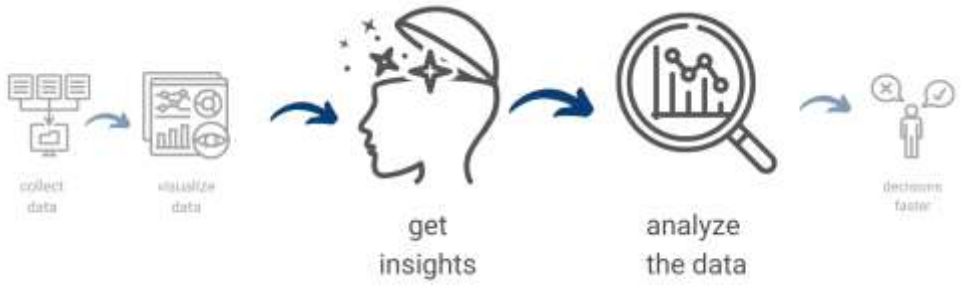


图 7.1-7 正是数据分析 最终将收集到的信息转化为决策依据。

为了在分析过程中做出真正明智的决策 ，有必要学习如何正确提出针对数据的问题。这些问题的质量直接影响所获见解的深度，进而影响管理决策的质量。

> 过去只存在于今天的记录中。而这些记录所代表的是什么，取决于我们提出的问题。除了这段历史，没有其他历史 [134]。
>
> - 约翰-阿奇博尔德-惠勒，物理学家 1982 年

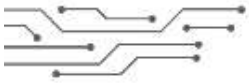
提出深刻问题和进行批判性思考的艺术是处理数据的基本技能。大多数人倾向于提出简单、肤浅的问题，不费吹灰之力就能回答。然而，真正的分析始于有意义的深思熟虑的问题， ，这些问题可以揭示隐藏在多层推理背后的信息中的隐藏关系和因果关系。

> 根据 "数据驱动转型：立即大规模加速 "研究（BCG，2017 年）[135]，成功的数字化转型需要对分析能力、变革管理计划进行投资，并使业务目标与 IT 计划保持一致。创建数据驱动文化的公司应投资于数据分析能力，并启动变革管理计划，以灌输新的思维、行为和工作方式。

如果不在发展分析文化、改进数据工具和培训专家方面进行投资，公司将继续面临根据过时或不完整的信息做出决策的风险--或依赖 HiPPO 管理人员的主观意见（图 2.1-9）。

> 意识到分析 和仪表板的相关性和不断更新的必要性，必然会使管理层认识到分析流程自动化的重要性。自动化提高了决策速度，减少了人为因素的影响，并确保了数据的相关性。随着信息量的指数级增长，速度不仅是一种竞争优势，也是持续成功的关键因素。

一般来说，数据分析和处理流程的自动化与 ETL （提取 ，转换 ，加载 ）的主题密不可分。正如在自动化流程中我们需要转换数据一样，在 ETL 流程中，数据从各种来源提取，根据必要的要求进行转换，然后加载到目标系统供进一步使用。



章 7.2.
无需人工的数据流：为什么需要 **etl**

ETL 自动化：降低成本，加快数据处理

当关键绩效指标（KPIs ）在数据量和团队规模增加的情况下停止增长时，公司管理层不可避免地会意识到流程自动化的必要性。这种意识迟早会成为启动复杂自动化的动力，其主要目标是降低流程的复杂性，加快处理，减少对人为因素的依赖。

> 麦肯锡的研究报告《如何建立数据架构来推动创新--今天和明天》（2022 年）[136]指出，使用流数据架构的公司可以获得显著优势，因为它们可以实时分析信息。流技术可以直接分析实时信息，并通过分析实时传感器数据在制造业中应用预测性维护。

> 流程简化就是自动化，传统的人工功能被算法和系统所取代。

自动化问题，或者说 "最大限度地减少人类在数据处理中的作用"，对每家公司来说都是一个不可逆转且高度敏感的过程。任何专业领域的专家往往都不愿向其他优化人员完全透露自己的工作方法和微妙之处，因为他们意识到，在快速发展的技术环境中，他们有可能失去工作。

> 如果你想树敌，就试着改变现状 [137]。
>
> - 伍德罗-威尔逊，在推销员大会上的演讲，底特律，1916 年

尽管自动化带来的好处显而易见，但许多公司在日常工作中，特别是在工程数据领域，仍有很高比例的人工劳动。为了说明目前的情况，让我们来看一个在此类流程中进行顺序数据处理的典型例子。

手工数据处理可以通过与 CAD 数据库中的信息交互为例进行说明。CAD (BIM)部门根据设计数据创建属性表或创建文档的传统数据处理（"手动 "ETL -process）顺序如下（图 7.2-1）：

1. 手动提取（提取）：用户手动打开项目--通过启动 CAD 应用程序 （BIM ）（图 7.2-1 步骤 1）。

2. 验证：下一步通常涉及手动运行几个插件或 辅助应用程序，以准备数据并评估其质量（图 7.2-1 步骤 2-3）。
3. 人工转换（**Transform**）：准备工作完成后，开始进行数据处理，这需要人工操作各种软件工具准备上传数据（图 7.2-1 第 4 步）。
4. 手动上传（**Load**）：将转换后的数据手动上传至外部系统、数据格式和文件（图 7.2-1 第 5 步）。

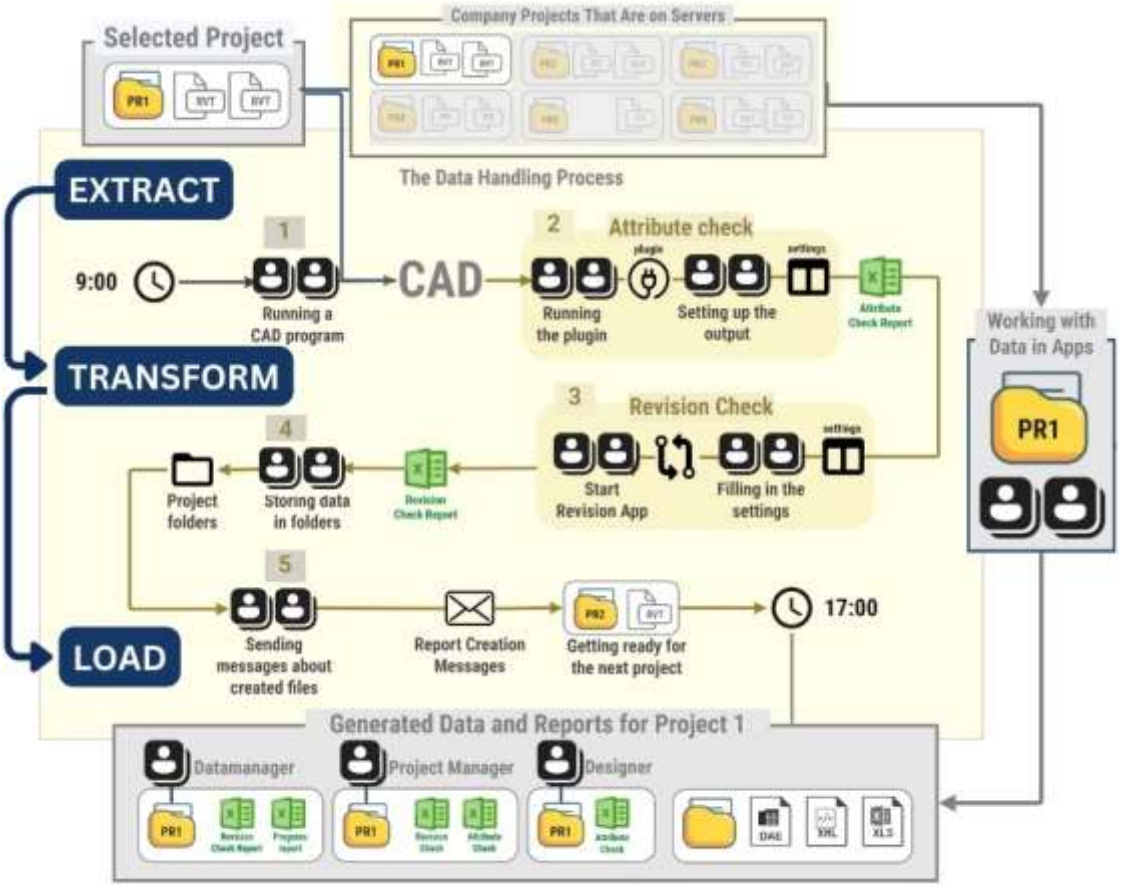


图 7.2-1 传统的手动 ETL 处理受限于技术人员个人的愿望和体能。

这样的工作流程就是一个典型的 ETL 流程--提取、转换和加载（ETL）的例子。与自动 ETL 管道早已成为标准的其他行业相比，建筑行业仍以手工劳动为主，这不仅减慢了流程速度，还增加了成本

> **ETL** （**Extract, Transform, Load** ）是从各种来源提取数据，将其转换为所需格式，并加载到目标系统供进一步分析和使用的过程。

ETL 是一个流程，表示数据处理的三个关键组成部分：提取 、转换 和加载 （图 7.2-2）：

- 提取 - 从不同来源（文件、数据库、API ）提取数据。
- 转换 - 数据清理、汇总、规范化和逻辑处理。
- 加载 - 将结构化信息加载到数据仓库、报表或商业智能系统中。

在本书的前半部分，ETL 的概念只是偶尔被提及：将非结构化扫描文档转换为结构化表格格式（图 4.1-1）、将需求形式化以系统化生活和业务流程的感知（图 4.4-20），以及从 CAD 解决方案中实现数据验证和数据处理的自动化。现在，让我们结合典型的工作流程更详细地了解一下 ETL。

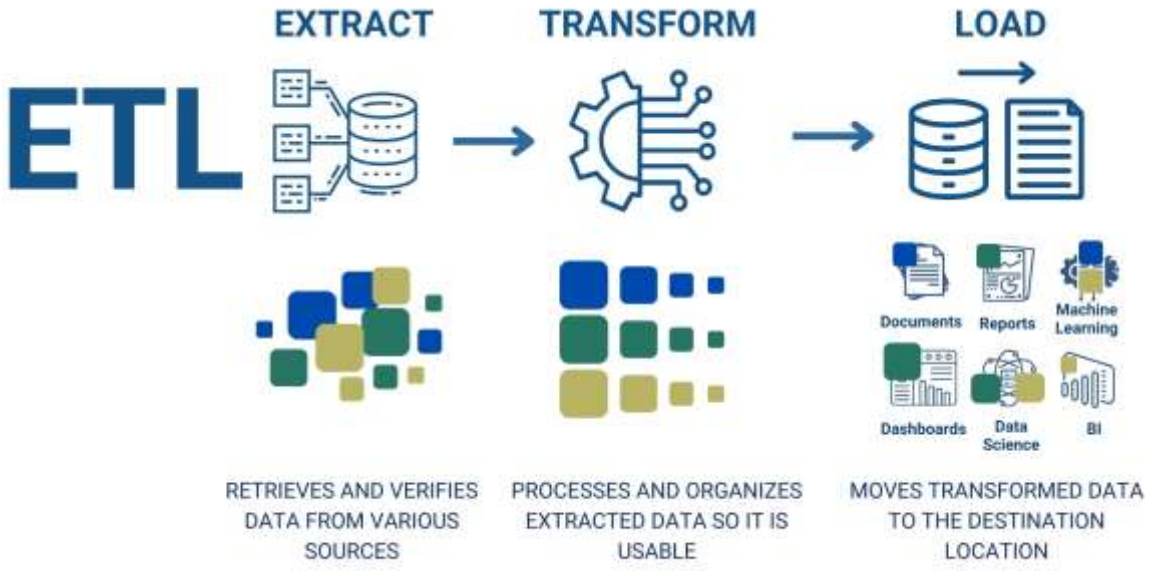


图 7.2-2 ETL 自动执行重复性数据处理任务。

手动或半自动化 ETL --流程意味着从数据收集到报告生成的所有步骤都由管理人员或技术人员手动管理。这样的流程需要花费大量时间，尤其是在工作时间有限的情况下（如 9:00 至 17:00）。

企业往往通过购买模块化集成解决方案（ERP 、PMIS 、CPM 、CAFM 等）来解决效率低、速度慢的问题，

然后再由外部供应商和顾问进一步开发。但是，这些供应商和第三方开发商往往成为一个关键的依赖点：他们的技术局限性直接影响到整个系统和整个业务的性能，这在前面关于专有系统和格式的章节中有详细介绍。碎片化和依赖性造成的问题在 "建筑企业如何陷入数据混乱 "一章中进行了详细讨论。

> 如果一家公司还没有准备好实施某个供应商提供的大型模块化平台，它就会开始寻找其他自动化方法。其中一种方法是开发自己的模块化开放式 ETL - 输送机，其中每个阶段（提取、转换、验证、加载）都是作为按计划执行的脚本来实现的。

在同一 ETL 工作流程的自动化版本 （图 7.2-1）中，工作流程看起来像一个模块化代码，首先处理数据并将其转换为开放的结构化形式。收到结构化数据后，各种脚本或模块会按计划自动运行，以检查变化、转换和发送信息（图 7.2-3）。

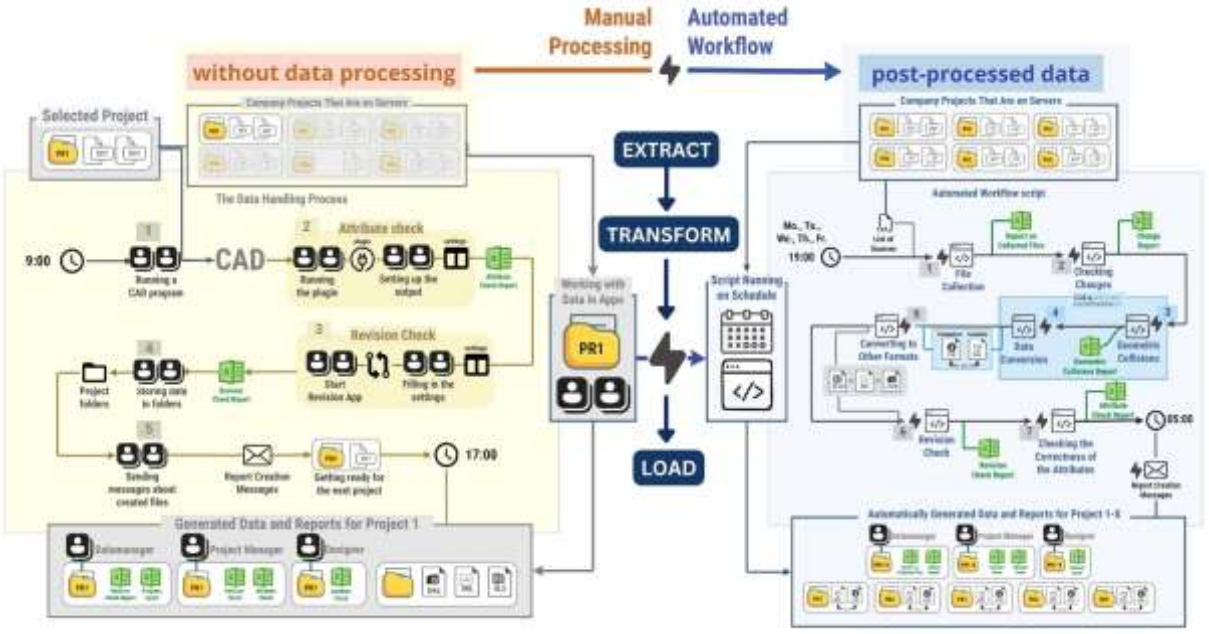


图 7.2-3 左边是手动加工，右边是自动加工，与传统的手动加工不同，自动加工不受用户能力的限制。

在自动工作流程中，通过 ET(L) 数据预处理：结构化和统一，简化了数据处理。

> 在传统处理方法中，专家们处理的是 "原样 "数据--从系统或软件中获取的数据。相比之下，在自动化流程中，数据通常首先要经过 ETL -payplane，在这里，数据被转换成适合进一步使用和分析的一致结构和格式。

让我们以一个实用的 ETL 示例 为例，演示 "验证 数据和验证结果 "一章中描述的数据表验证过程（图 4.4-13）。为此，我们将 Pandas 库 与 LLM 结合使用，以实现自动数据分析和处理流程。

ETL 提取： 数据收集

ETL 流程的第一阶段--提取) --从编写代码开始，收集数据集，以便进一步检查和处理。为此，我们将扫描生产服务器的所有文件夹，收集具有特定格式和内容的文件，然后将其转换为结构化形式。这一过程将在 "将非结构化数据和文本数据转换为结构化数据 "和 "将 CAD 数据 (BIM) 转换为结构化数据 "两章中详细讨论（图 4.1-1 - 图 4.1-12）。

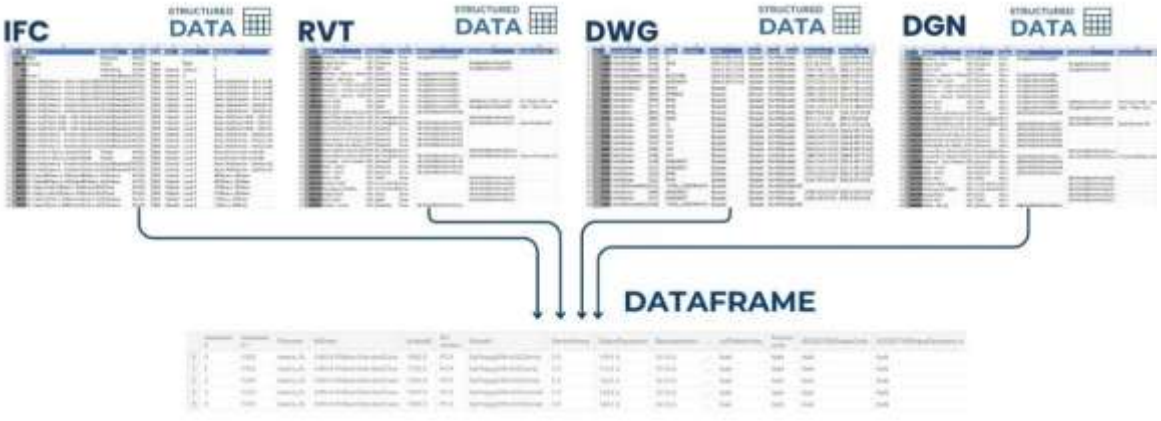


图 7.2-4 将 CAD 数据 (BIM) 转换为一个包含所有项目部分的大型数据框。

作为一个示例，我们使用提取 数据加载步骤，并使用支持逆向工程的 RVT 和 IFC 格式转换器 [138] 获得所有 CAD- (BIM-) 项目表格（图 7.2-4），以获得所有项目的结构化表格，并将其合并为一个大型 DataFrame 表。



图 7.2-5 使用 Python 代码和 SDK 逆向工程工具将 RVT 和 IFC 文件转换为一个大型结构化 (df) DataFrame。

Pandas DataFrame 可从多种来源加载数据，包括 CSV 文本文件 、Excel 电子表格 、JSON - 和 XML - 文件 、大数据存储格式（如 Parquet 和 HDF5 ）以及 MySQL 、PostgreSQL 、SQLite 、Microsoft SQL Server 、Oracle 和其他数据库。此外，Pandas 还支持从 API 、网页、云服务和存储系统（如 Google BigQuery 、Amazon Redshift 和 Snowflake）加载数据。

- 要编写连接数据库并从数据库中收集信息的代码，请向 LLM 聊天室（CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN 或其他）发送类似的文本请求：

> 请编写一个连接到 MySQL 并将数据转换为 的示例 ⏎

LLM 的答复：

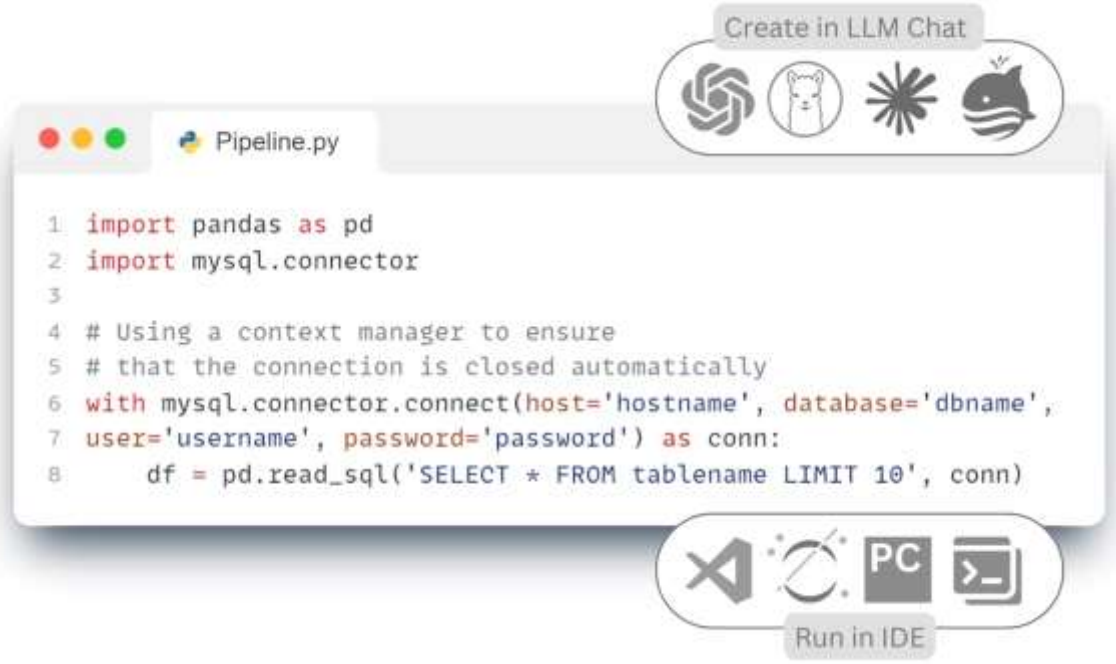


图 7.2-6 通过 Python 连接到 MySQL 数据库 并将数据从 MySQL 数据库导入 DataFrame 的示例。

生成的代码（图 7.2-5、图 7.2-6）可以在离线模式下在我们上面提到的流行 IDE（集成开发环境）之一中运行：PyCharm、Visual Studio Code (VS Code)、Jupyter Notebook 、Spyder、Atom、Sublime Text、带 PyDev 插件的 Eclipse、Thonny、Wing IDE、带 Python 插件的 IntelliJ IDEA 、JupyterLab 或流行的在线工具：Kaggle.com、Google Collab、Microsoft Azure Notebooks、Amazon SageMaker。

通过将多格式数据加载到变量 "df "中（图 7.2-5 - 第 25 行；图 7.2-6 - 第 8 行），我们将数据转换为 Pandas DataFrame 格式，这是数据处理中最常用的结构之一，是一个包含行和列的二维表格。我们将在 "建筑行业的数据存储和管理 "一章中详细介绍 ETL -Pipelines 中使用的其他存储格式，如 Parquet 、Apache ORC 、JSON 、Feather 、HDF5 以及现代数据仓库（图 8.1-2）。

在数据提取和结构化阶段（提取 ）之后，就形成了一个单一的信息阵列（图 7.2-5，图 7.2-6），可用于进一步处理。不过，在将这些数据加载到目标系统或用于分析之前，有必要确保其质量、完整性和符合特定要

求。数据转换（Transform ）就是在这一阶段进行的，这是确保后续结论和决策可靠性的关键步骤。

ETL Transform：应用验证和转换规则

转换 步骤是对数据进行处理和转换。这一过程可能包括正确性检查、规范化、填补缺失值以及使用自动化工具进行验证

> 普华永道的研究报告《数据驱动。学生需要什么才能在瞬息万变的商业世界中取得成功》（2015 年）[9]，现代审计公司正在从随机数据核查转向使用自动化工具分析海量信息。这种方法不仅能发现报告中的差异，还能提出优化业务流程的建议。

在建筑领域，类似的方法也可用于设计数据的自动验证、建筑质量控制和承包商业绩评估等。在 ETL 流程的数据转换阶段（Transform ）使用正则表达式 (RegEx) 是实现数据处理自动化和加速的工具之一。RegEx 可以让您有效地验证数据字符串，检测不一致之处，并以最少的资源消耗确保信息的完整性。有关 RegEx（图 4.4-7）的更多详情，请参阅 "将需求转化为结构化形式 "一章。

让我们举一个实际例子：在属性对象管理系统（RPM ）中，管理者为对象的关键属性设置要求 （图 7.2-7 ）。在转换阶段，需要验证以下参数：

- 验证对象标识符格式（属性 "ID）
- 控制替换保修期值（属性 "保修期）
- 验证元件更换周期（属性 "维护要求）



图 7.2-7 验证 质量首先要设定属性要求及其边界值。

例如，要为参数验证设置边界值，假设我们从 的经验中知道，"ID "属性的有效值只能包括字符串值 "W-NEW"、"W-OLD1 "或 "D-122 "或类似值，其中第一个字符是字母，后面是破折号，然后是三个字母字符 "NEW"、"OLD "或任何三位数字（图 7.2-7）。以下正则表达式 (RegEx) 可用来验证这些标识符：

```
^W-NEW$|^W-OLD[0-9]+$|^D-1[0-9]{2}$
```

该模板确保数据中的所有标识符都符合指定的标准。如果有任何值没有通过检查，系统就会记录错误。要创建 Python 代码来转换数据，并使用生成的数据创建结果表，只需在 LLM 聊天室中制定一个查询即可。

文本请求发送至 LLM：

> 编写代码，使用正则表达式 验证 DataFrame 的列，通过 RegEx 检查格式为 "W-NEW "或 "W-OLD "的标识符、以字母 "A "至 "G "表示的能源效率、以年为单位的数字值表示的保修期和更换周期 ⏎

LLM 的答复：

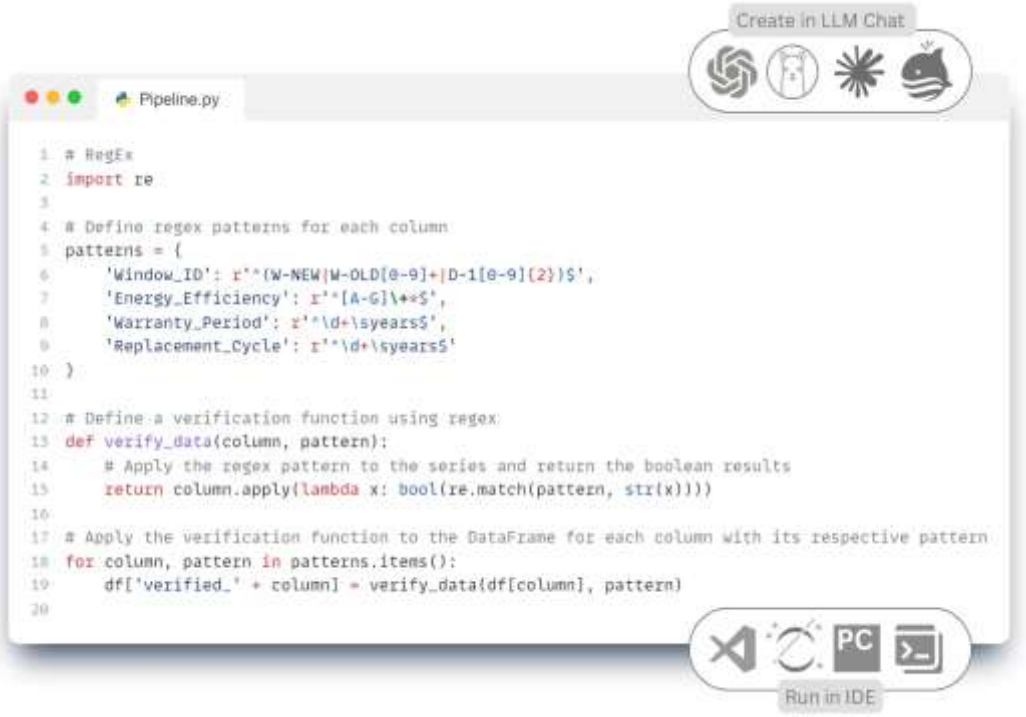


图 7.2-8 代码通过将 RegEx 模板应用于数据帧参数列，实现验证过程自动化。

上述自动生成的 Python 代码（图 7.2-8）使用 "re "库（RegEx 正则表达式 ）定义了一个函数，用于检查 DataFrame 中数据项的每个属性。对于指定的每一列（属性），函数应用 RegEx 模式检查每条记录是否符合预期格式，并将结果作为新值 (False/True) 添加到 DataFrame 的新属性列中。

> 这种自动验证可确保正式数据的合规性，并可作为转换阶段质量控制系统的一部分。

成功完成转换 步骤和质量检查后，数据就可以上传到目标系统。经过转换和验证的数据可以上传到 CSV 、JSON 、Excel 、数据库和其他格式供进一步使用。根据任务的不同，还可以通过报告、图表或分析仪表板来展示结果。

ETL Load： 通过图表和图形可视化 结果

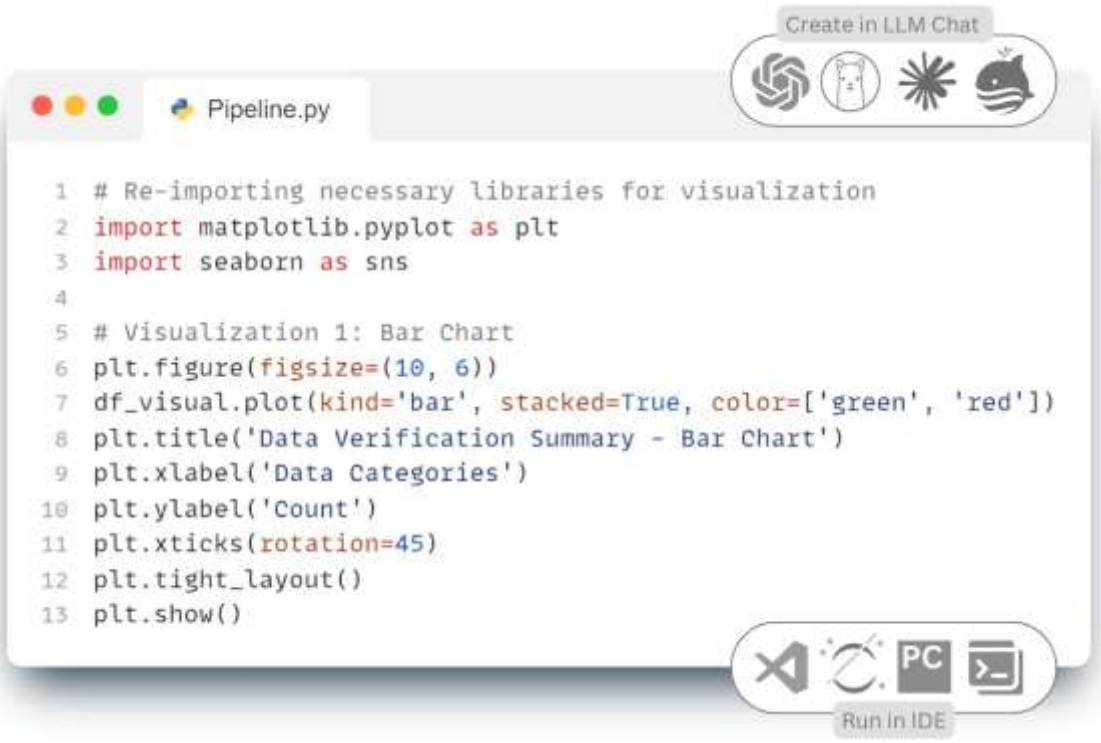
在转换 阶段完成后，数据已经结构化并经过验证，最后一个阶段是加载 ，在此阶段可以将数据加载到 目标系统并进行可视化分析。数据的可视化展示可以快速识别偏差、分析分布并向所有项目参与者（包括没有技术背景 的人员）传达关键结论。

我们可以使用信息图表、图形和仪表盘（dashboards）来代替表格和数字来展示信息。在 Python 中可视化结构化数据的最常用、最灵活的工具之一是 Matplotlib 库 （图 7.2-9，图 7.2-10）。它允许您创建静态、动画和交互式图表，并支持多种图表类型。

- 要可视化 RPM 系统的属性检查结果（图 7.2-7） ，可以使用下面的语言模型查询：

> 编写可视化 DataFrame 数据 的代码（图 7.2-7），用直方图显示属性的错误率。 ⏎

LLM 以代码形式做出回应，并直接在 LLM 聊天室可视化代码执行结果：



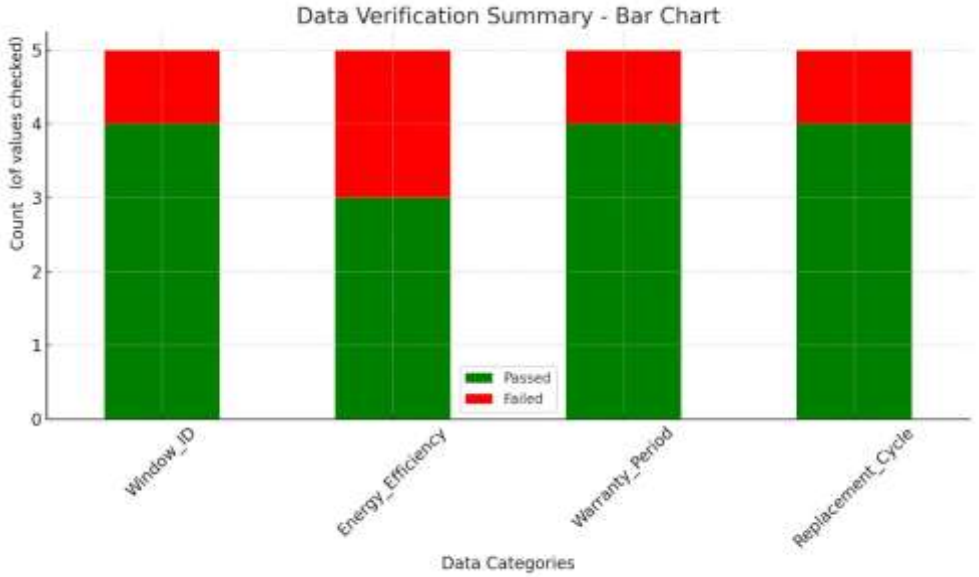


图 7.2-9 可视化 从 RPM 系统 （图 7.2-7）检查属性值的 Transform 步骤的结果，作为负载 步骤中的直方图

。

有许多开源和免费的可视化库可以让您以不同格式展示结构化数据。让我们继续用另一种类型的图表可视化结果，聊天 中的提示如下：

> 将相同数据绘制成折线图 ⏎

LLM 的答复：



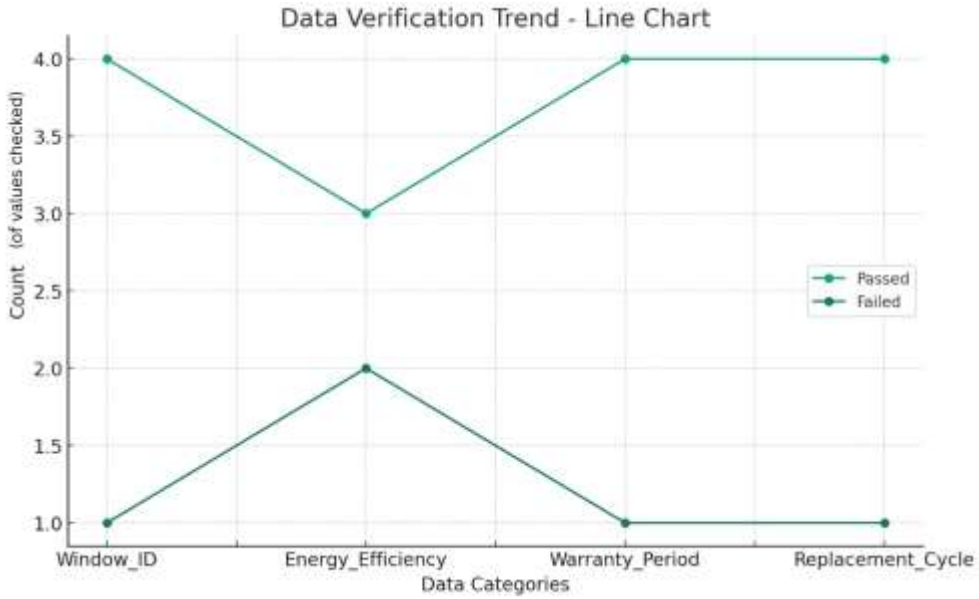


图 7.2-10 验证数据（图 7.2-8）的可视化线图，使用 Matplotlib 库。

有许多开源和免费的可视化库，例如

- Seaborn - 用于统计图表（图 7.2-11）
- Plotly - 用于交互式网络可视化（图 7.2-12，图 7.1-6）
- Altair - 用于声明式可视化
- Dash 或 Streamlit - 创建完整的仪表盘

您无需掌握可视化特定库的知识--包括 LLM 在内的现代工具允许您根据任务描述自动生成图形和整个应用程序的代码。

工具的选择取决于项目的目标：是报告、演示还是在线仪表盘。例如，开放源代码库 Seaborn 特别适合处理分类数据，有助于识别模式和趋势。

- 要查看运行中的 Seaborn 图书馆 ，您可以直接向图书馆管理员提出使用所需图书馆的要求，或者向图书馆管理员发送类似的文本请求，网址是：

  显示热图 ⏎

- LLM 以代码和完成图形的形式给出了答案，其中的绘图代码现在可以复制到集成开发环境中，图形本身也可以复制或保存，以便粘贴到文档中：

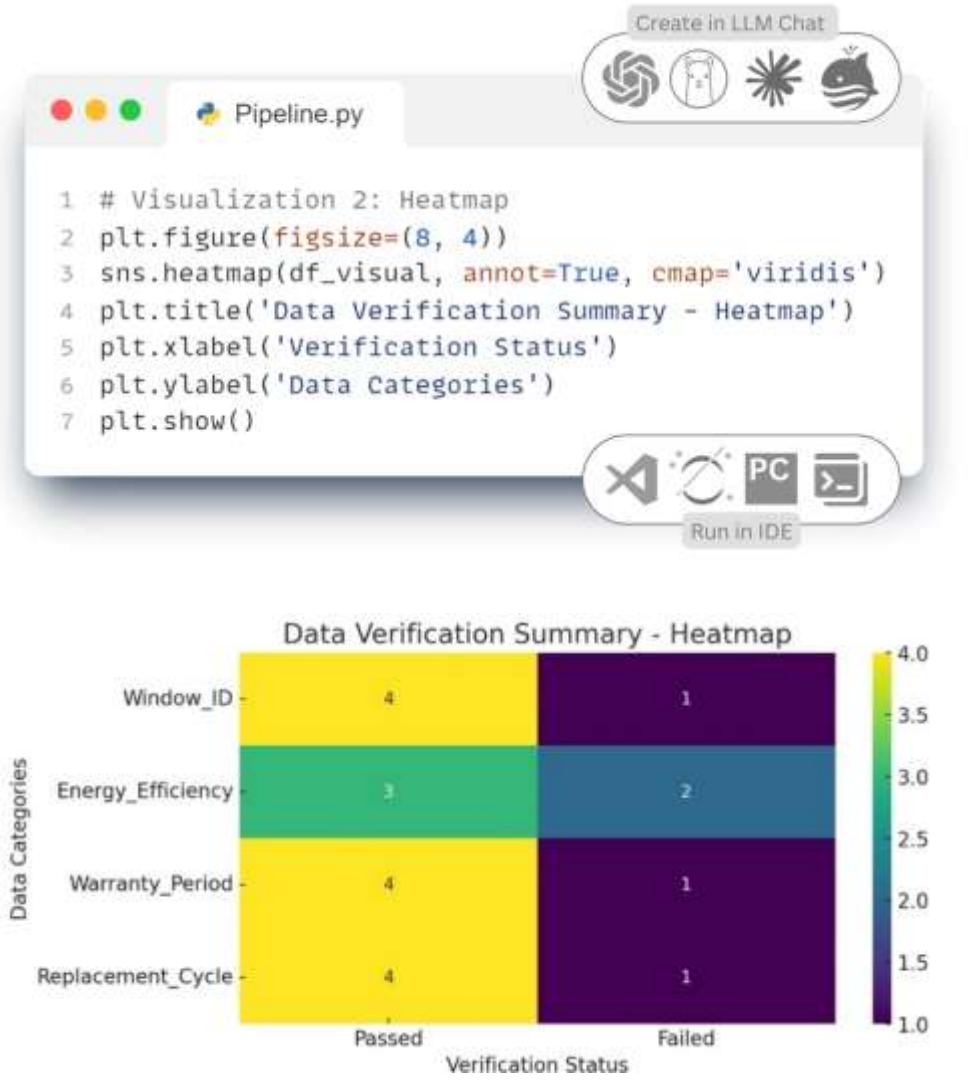


图 7.2-11 使用 Seaborn 库验证数据（图 7.2-8）的结果可视化。

对于喜欢交互式方法的用户，有一些工具可以创建具有交互功能的动态图表和面板。Plotly 库 （图 7.1-6，图 7.2-12）提供了创建高度交互式图表和面板的功能，这些图表和面板可以嵌入到网页中，并允许用户与数据进行实时交互。

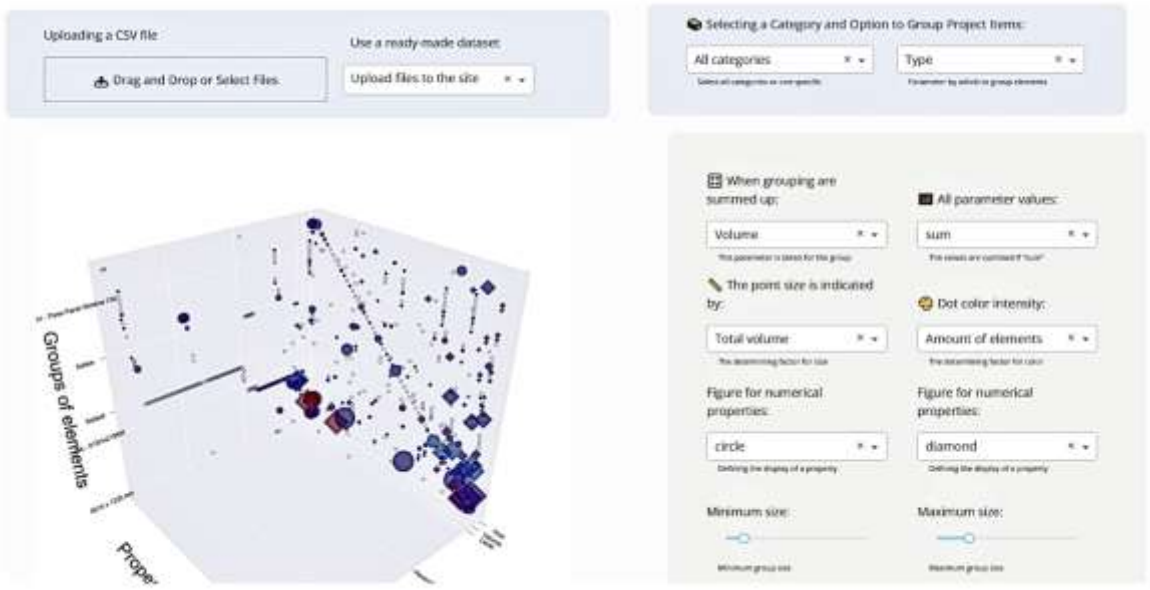


图 7.2-12 使用 Plotly 库 对 CAD- (BIM-) 项目中的元素属性进行交互式三维 可视化。

专门的开源库 Bokeh 、Dash 和 Streamlit 提供了一种便捷的数据展示方式，无需深厚的网络开发知识。Bokeh 适用于复杂的交互式图表，Dash 用于构建全面的分析仪表盘，而 Streamlit 则可让您快速创建用于数据分析的网络应用程序。

有了这些可视化工具，开发人员和分析人员 可以有效地向同事和利益相关者传播结果，实现与数据的直观互动，简化决策。

ETL Load: 自动创建 **PDF** 文档

在数据加载阶段，您不仅可以将数据可视化、上传到表格或数据库，还可以自动生成报告，包括必要的图表和关键分析指标，供等待审计结果的经理或专家使用。自动报告既可以包含评论，也可以包含对数据和可视化元素（表格、图表）的文字解释。

- 要创建带有直方图（图 7.2-9）和基于前几章验证的分析说明的 PDF 报告，只需在与 LLM 的继续对话中提出请求即可，例如：

> 编写代码，创建 PDF 文件，其中包含直方图和上述数据验证结果的说明（在聊天中），并编写文本警告，说明某些类别尚未验证，需要填写缺失的⏎

LLM 的回复作为代码和完成的 PDF，结果：

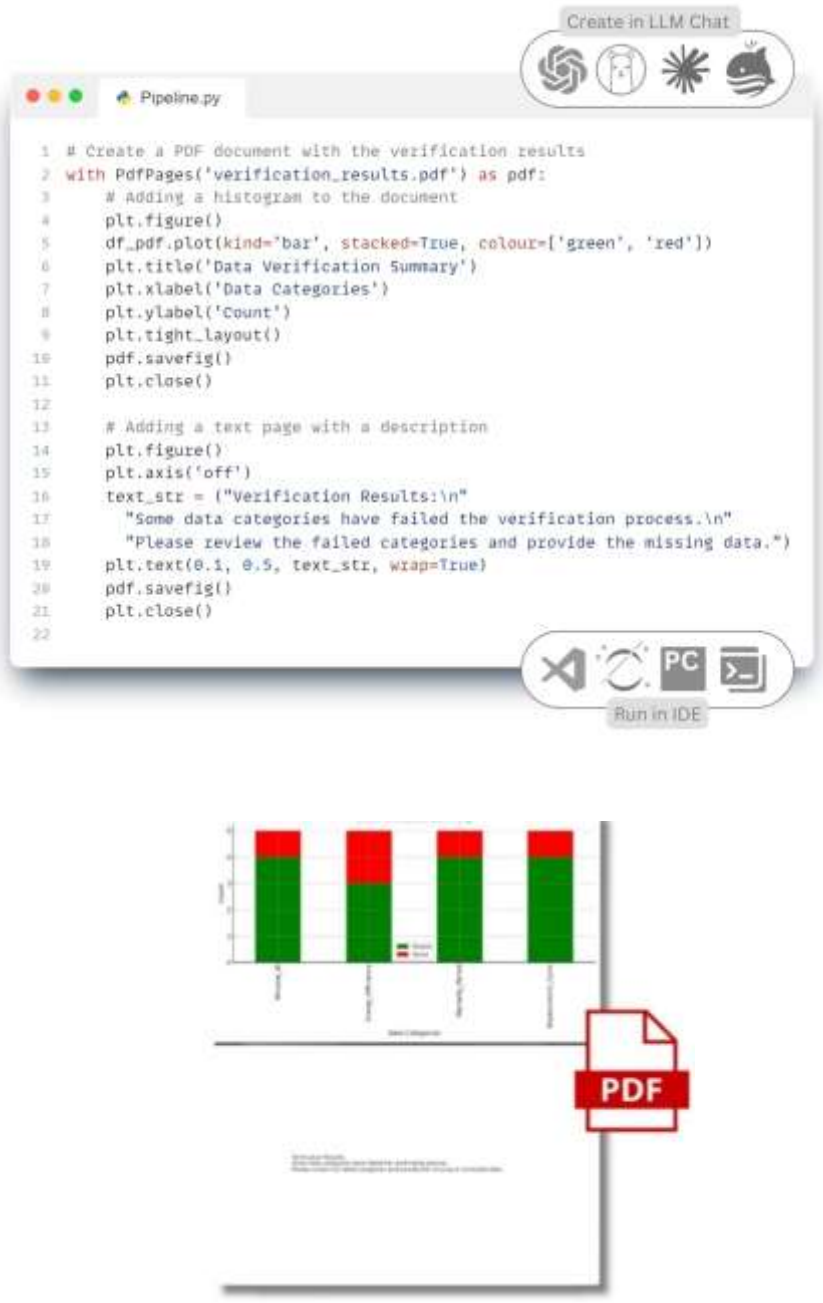


图 7.2-13 自动代码会创建一个 PDF 文档，其中包含测试数据直方图和测试结果文本。

使用 LLM 自动编写的解决方案只需 20 行代码，就能立即创建所需的 PDF （或 DOC）文档，并以属性直方图的形式（图 7.2-13）显示通过和未通过验证的数据数量，同时添加一个文本块，总结结果并进一步行动的建议。

自动生成文件是负载阶段 的关键要素，尤其是在对报告速度和准确性要求极高的项目环境中。

ETL Load：从 **FPDF** 自动生成文件

在 ETL 阶段自动生成报告 Load 是数据处理的重要步骤，尤其是当分析结果需要以易于沟通和理解的格式呈现时。在建筑行业，这通常与进度报告、项目数据统计、质量保证报告或财务文件有关。

此类任务最方便的工具之一是开源库 、FPDF ，可用于 Python 和 PHP。

> 开源 **FPDF** 库 提供了一种通过代码生成文档的灵活方式，允许您添加页眉、文本、表格和图像 。使用代码代替手工编辑可以减少错误，加快 PDF 格式报告的编制过程。

创建 PDF 文档的关键步骤之一是添加标题和注释或说明形式的主要文本。但是，在创建报告时，不仅要添加文本，还要合理安排文本结构。标题、缩进、行间距 - 所有这些都会影响文档的可读性。使用 FPDF ，您可以设置格式化参数，控制元素的排列，并自定义文档的样式。

FPDF 在原理上与 HTML 非常相似。已经熟悉 HTML 的用户可以使用 FPDF 轻松生成任何复杂程度的 PDF 文档，因为其代码结构与 HTML 标记非常相似：标题、文本、图像和表格都是以类似的方式添加的。不熟悉 HTML 的人也不必担心，您可以使用 LLM ，它可以立即帮助您编写代码，生成所需的文档布局。

- 下面的示例演示了如何生成带有页眉和正文的报告。在任何支持 Python 的集成开发环境 中执行此代码，都会生成一个包含所需页眉和正文的 PDF 文件：

```
from fpdf import FPDF  # 导入 FPDF 库
pdf = FPDF()  # 创建 PDF -document
pdf.add_page()  # 添加页面
```

```
pdf.set_font("Arial", style='B', size=16)  # 设置字体：Arial，粗体，大小 16
pdf.cell(200, 10, "Project Report", ln=True, align='C')  # 创建标题并居中显示
pdf.set_font("Arial", size=12)  # 将字体更改为 12 号普通 Arial 字体
pdf.multi_cell(0, 10, "本文档包含项目文件核查结果数据.....")# 添加多行文本
pdf.output(r "C:\reports\report.pdf") # 保存 PDF -file
```



图 7.2-14 只需几行 Python 代码 ，我们就能自动生成所需的 PDF 文本文档。

在编制报告时，必须考虑到形成文档的数据很少是静态的。标题、文本块（图 7.2-14）通常是动态形成的，在 ETL 流程中的转换阶段 接收值。

> 使用该代码可以创建包含最新信息的文档：项目名称、报告生成日期、参与者信息或当前状态。通过在代码中使用变量，您可以在报告的必要位置自动插入这些数据，完全无需在发送前进行手动编辑。

除了简单的文字和标题外，表格在项目文件中也占有特殊的地位。几乎每份文档都包含结构化数据：从对象描述到检查结果。根据 Transform 阶段的数据自动生成表格，不仅可以加快文件编制过程，还可以最大限度地减少信息传输时的错误。FPDF 允许在 PDF 文件中插入表格（作为文本或图片），设置单元格边框、列尺寸和字体（图 7.2-15）。在处理动态数据时，行数和列数会根据文档任务的不同而变化，这时使用表格就特别方便。

- 下面的示例展示了如何自动创建表格，例如包含材料清单、估算或参数测试结果的表格：

```
data = [
["Item", "Quantity", "Price"], # 列标题
["Concrete", "10 m³", "$ 500."], # 第一行数据
["Rebar", "2 tonnes", "$ 600"], # 第二行数据。
["Brick," "5,000 pieces," "$ 750."], # 第三行数据。
]

pdf = FPDF () # 创建 PDF -document
pdf.add_page() # 添加页面
pdf.set_font("Arial", size=12) # 设置字体

for row in data: # 搜索表格行
for item in row: # 搜索行中的单元格
pdf.cell(60, 10, item, border=1) # 创建一个有边框、宽 60、高 10 的单元格
pdf.ln() # 移动到下一行
pdf.output(r "C:\reports\table.pdf") # 保存 PDF -file
```



图 7.2-15 您不仅可以自动生成文本，还可以在 PDF 中通过转换步骤 自动生成任何表格信息。

在实际报告场景中，表格通常是在数据转换阶段动态生成的信息。在给定的示例中（图 7.2-15），表格是以静态形式插入 PDF 文档的：示例的数据放在数据字典中（代码的第一行），而在实际情况中，这种数据变量是在数据框分组后自动填入的。

在实践中，这类表格通常是根据来自 各种动态源的结构化数据建立的：数据库、Excel - 文件、API - 接口或分析计算的结果。通常在转换 (ETL) 阶段，数据会被聚合、分组或过滤，然后才会以图表或二维表格的形式转换成报告中显示的总数。这意味着表格内容会根据所选参数、分析周期、项目筛选器或用户设置而发生变化。

> 在转换 阶段使用动态数据框和数据集，可使加载 阶段的报告流程尽可能灵活、可扩展和易于重复，而无需人工干预。

除了表格和文本外，FPDF 还支持添加表格数据的图表，这样您就可以在报告中嵌入用 Matplotlib 或上文讨论过的其他可视化库生成的图像。任何图形、图表和示意图都可以通过代码添加到文档中。

- 使用 Python 库 FPDF ，让我们在 PDF 文档中添加用 Matplotlib 预先生成的图形：

```
import matplotlib.pyplot as plt  # 导入 Matplotlib 以创建图表

fig, ax = plt.subplots()  # 创建图表的图形和坐标轴
categories = ["Concrete", "Rebar", "Brick"]  # 类别名称
values = [50000, 60000, 75000]  # 类别值
ax.bar(categories, values)  # 创建条形图
plt.ylabel("Value,$.")  # 在 Y 轴上签名
plt.title("Cost Distribution")  # 添加标题
plt.savefig(r "C:\reports\chart\chart.png")  # 将图表保存为图像

pdf = FPDF ()  # 创建 PDF -document
pdf.add_page() # 添加页面
pdf.set_font("Arial", size=12)  # 设置字体
pdf.cell(200, 10, "Cost Chart", ln=True, align='C')  # 添加页眉

pdf.image(r "C:\reports\chart\chart.png", x=10, y=30, w=100)  # 将图像插入 PDF（x, y - 坐标，w - 宽度）
pdf.output(r "C:\reports\chart_report.pdf")  # 保存 PDF 文件
```

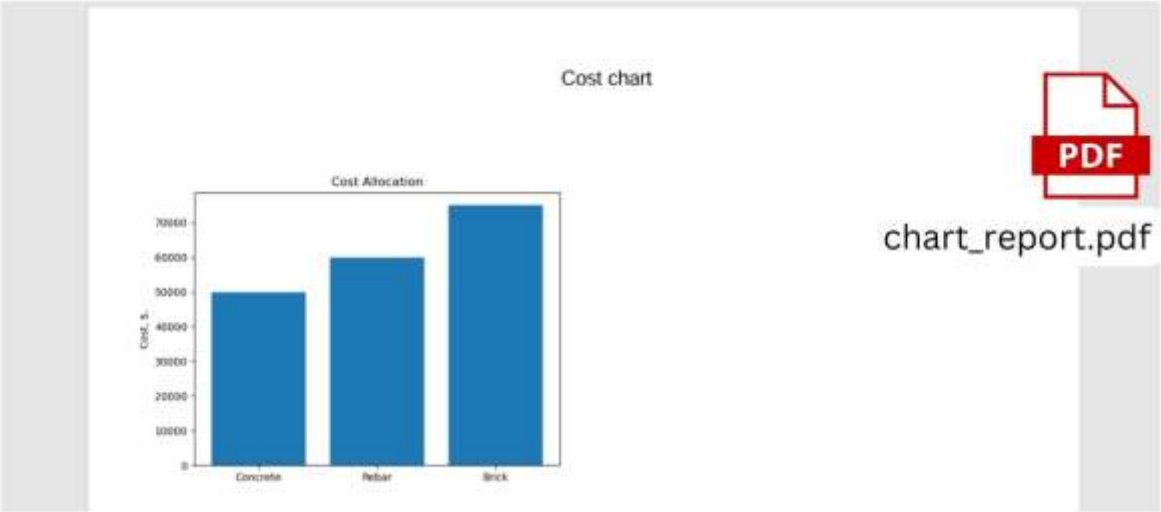


图 7.2-16 只需十几行代码，您就可以生成图形、保存图形，然后将其粘贴到 PDF 文档中。

FPDF 使文件准备和逻辑过程透明、快速、便捷。代码中内置的模板允许生成带有最新数据的文件，无需手动填写。

> 使用 ETL 自动化--而不是耗时的手动报告，专业人员可以专注于分析数据和做出决策，而不是选择合适的工具来处理具有清晰用户界面的特定数据孤岛。

因此，FPDF 库 为自动创建任何复杂程度的文档提供了灵活的工具--从简短的技术报告到复杂的带表格和图表的分析总结，它不仅能加快文档流程，还能大大降低与手动数据输入和格式化相关的错误概率。

ETL Load：报告和加载到其他系统

在加载阶段，结果以表格、图形和最终 PDF 报告的形式生成，这些都是按照既定要求编制的。此外，还可以将这些数据导出为机器可读格式（如 CSV），这对于与外部系统（如 ERP、CAFM、CPM、BI 平台和其他企业或行业解决方案）集成是必不可少的。除 CSV 外，还可上传至 XLSX、JSON、XML 或直接上传至支持自动信息交换的数据库。

- 要生成适当的代码来自动执行加载步骤，只需查询 LLM 接口即可，例如 ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude 或 QWEN：

> 编写代码，在 DataFrame 中生成数据验证结果报告，对以 "verified_"为前缀的列进行计数，将其重命名为 "通过 "和 "未通过"，缺失值用 0 代替，然后只将通过所有验证的行导出到 CSV 文件。

LLM 的答复：



图 7.2-17 在转换步骤中获得的经过验证的数据 从最终数据帧导出到 CSV 文件，以便与其他系统集成。

上述代码（图 7.2-17）实现了 ETL -process 的最后阶段--加载，在此期间，经过验证的数据以 CSV 格式保存 ，与大多数外部系统和数据库兼容。这样，我们就完成了 ETL 流程的整个周期，包括数据的提取、转换、可视化、记录和导出到我们所需的系统和格式，从而确保了信息工作的可重复性、透明度和自动化。

ETL 管道既可用于处理单个项目，也可用于大规模应用--当分析成百上千个以文档、图像、扫描、CAD 项目、点云、PDF 文件或来自分布式系统的其他来源为形式的传入数据时。ETL 能够完全自动化处理过程，这使得它不仅仅是一个技术处理工具，更是数字化建筑信息基础设施的基础。

ETL with LLM: 可视化 数据来自 **PDF -**文档

现在是时候建立一个全面的 ETL 流程了，它涵盖了单一方案中数据处理的所有关键阶段--提取、转换和加载

。让我们建立一个自动化的 ETL 管道，让您无需手动操作即可处理 PDF 文档--从文档中提取数据、可视化、分析并传输到其他系统。

我们示例中的 ETL 流程将通过提示来描述，提示必须向语言模型 (LLM) 解释所有 ETL 流程，并说明要获得的最终结果。在本例中，任务是查找指定文件夹及其子文件夹中的所有 PDF 文件，从中提取相关信息（如材料名称、数量和成本），并将结果以结构化表格（DataFrame）的形式呈现，以便进一步分析

- 在 LLM 中进行首次文本查询，自动从多个 PDF 文档中提取数据，并为提取步骤创建数据帧：

> 编写代码，从给定文件夹及其子文件夹中的 PDF 文件中提取材料信息。PDF 文件中的数据包括材料名称、数量和成本。提取结果应保存在 DataFrame 中⏎

LLM 的答复：

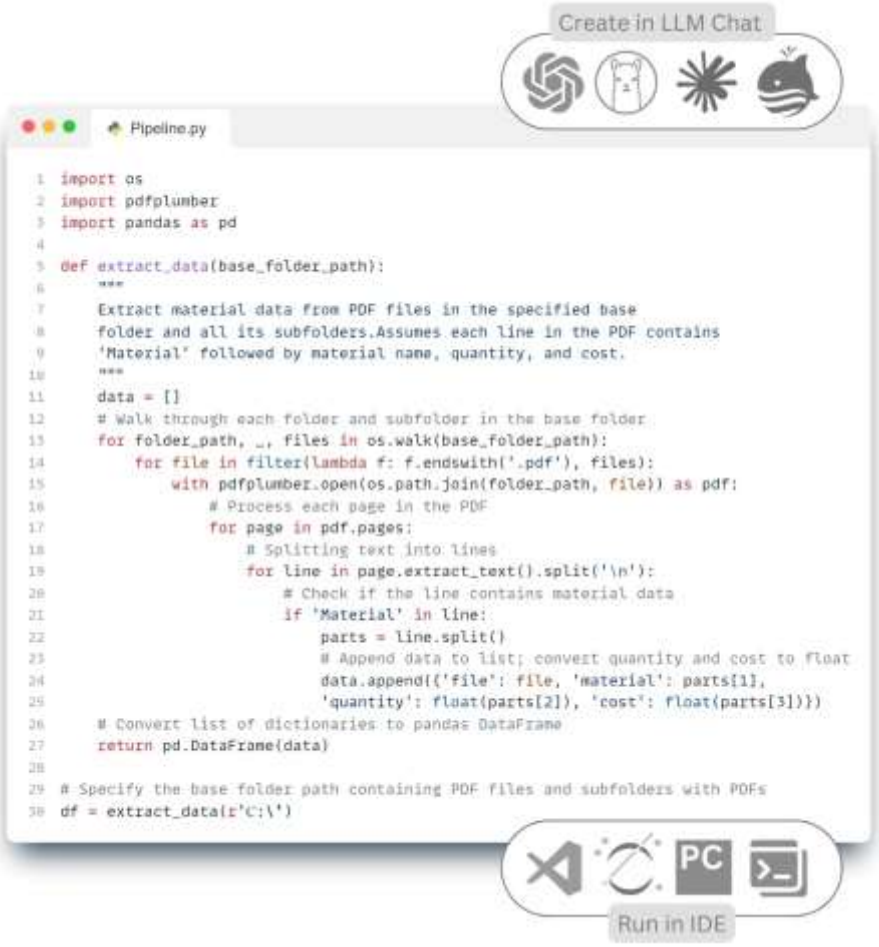


图 7.2-18 LLM 创建 Python - 代码，从特定文件夹及其所有子文件夹中的 PDF - 文件中提取数据。

LLM 响应 （图 7.2-18）是一个现成的 Python -脚本，它可以自动遍历所有文件夹，打开找到的 PDF -文件，从中提取文本信息并转换成表格。生成的聊天代码可以在以下流行的集成开发环境中运行：PyCharm、Visual Studio Code (VS Code)、Jupyter Notebook 、Spyder、Atom、Sublime Text、带有 PyDev 插件的 Eclipse、 Thonny、 Wing IDE、带有 Python 插件的 IntelliJ IDEA、JupyterLab 或流行的在线工具 Kaggle.com、Google Collab、Microsoft Azure Notebooks、Amazon SageMaker。

另外，为了简化收集过程，我们也可以将十几个 PDF - 文件直接上传到 LLM 聊天室（图 7.2-19），而不是从 LLM 复制代码并在集成开发环境中使用这些代码，并得到一个表格作为输出，而无需查看代码或运行代

码。执行此代码的结果将是一个包含我们所选属性的表格。

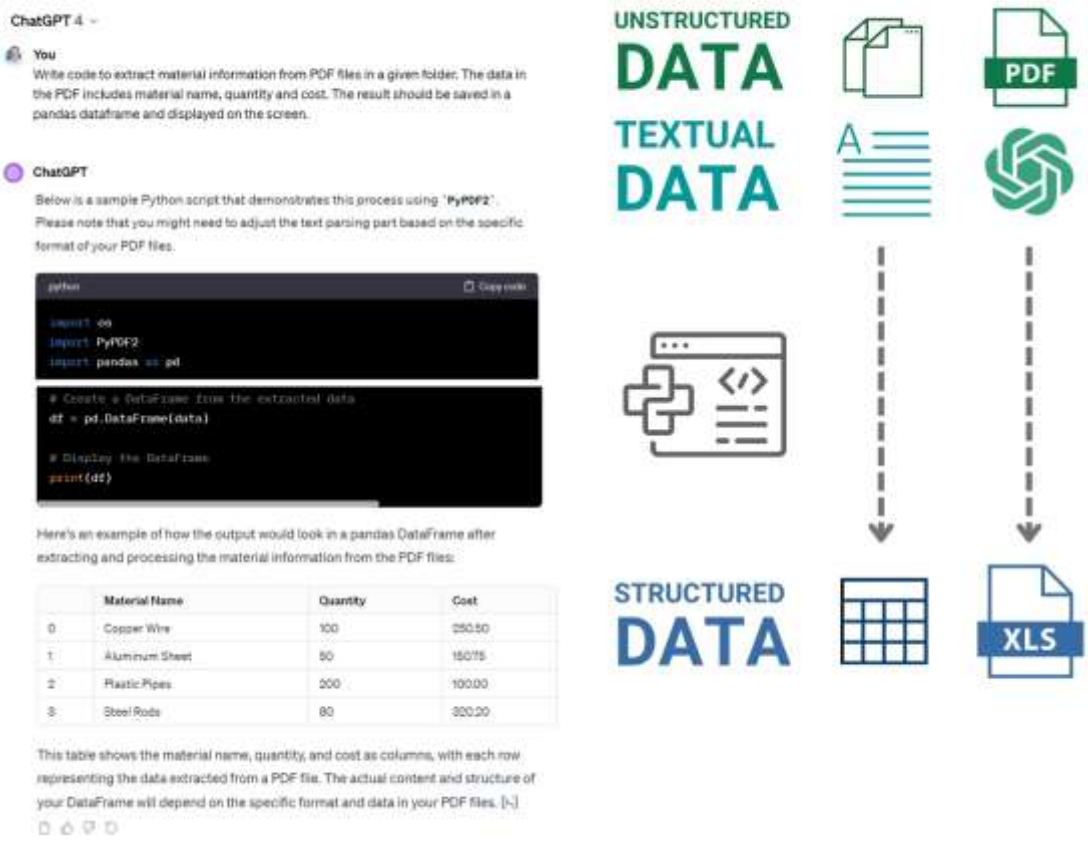


图 7.2-19 在 LLM 中执行代码的结果，该代码以结构化数据帧视图从 PDF - 文件中提取带有选定属性的数据。

下一步，我们要求对获得的数据建立语言模型，例如，比较材料使用的成本和数量，并创建一些可视化示例，作为进一步分析的基础。

- 继续与 LLM 聊天，要求根据 "转换 "步骤中制作的表格绘制一些图表（图 7.2-18）：

> 从 DataFrame 直观了解每种材料的总成本和数量（图 7.2-18↵

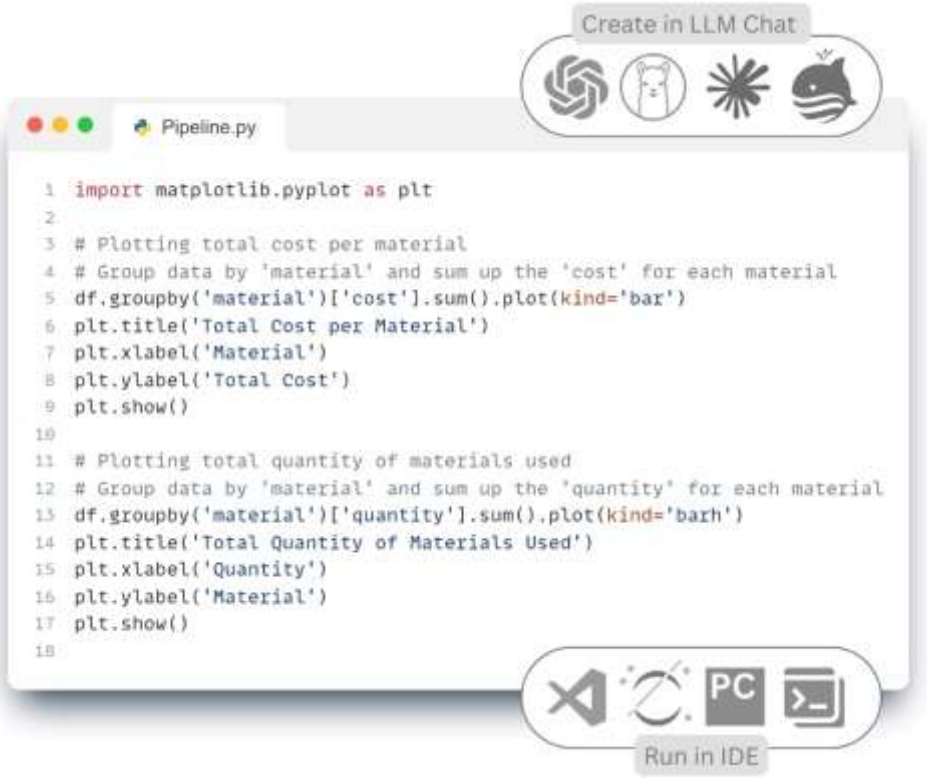


图 7.2-20 LLM 响应 - 模型作为 Python 代码 使用 matplotlib 库可视化数据帧中的数据。

LLM 使用 matplotlib 库自动生成并执行 Python - 代码（图 7.2-20）。执行该代码后，我们可以直接在 chat 中获得建筑项目的成本和材料用量图表（图 7.2-21），这大大简化了分析工作。

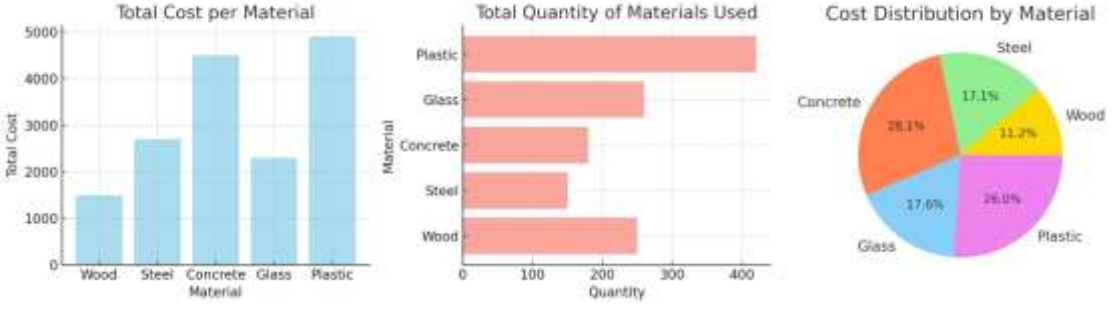


图 7.2-21 可视化 基于数据框中收集的数据，以图表形式显示 LLM 响应 。

通过 LLM 中的简单文本查询，无需学习编程基础知识，即可为编写 ETL 代码、分析和执行代码、可视化结果提供支持。人工智能 工具（如 LLM）的出现无疑正在改变编程和数据处理自动化的方法（图 7.2-22）。

> 根据普华永道的报告《人工智能对企业的真正价值是什么，如何利用它？(2017）[139]，流程自动化和生产率提高将成为经济增长的主要驱动力。而生产率的提高预计将占 2017 年至 2030 年人工智能驱动的所有 GDP 增长的 55% 以上"。

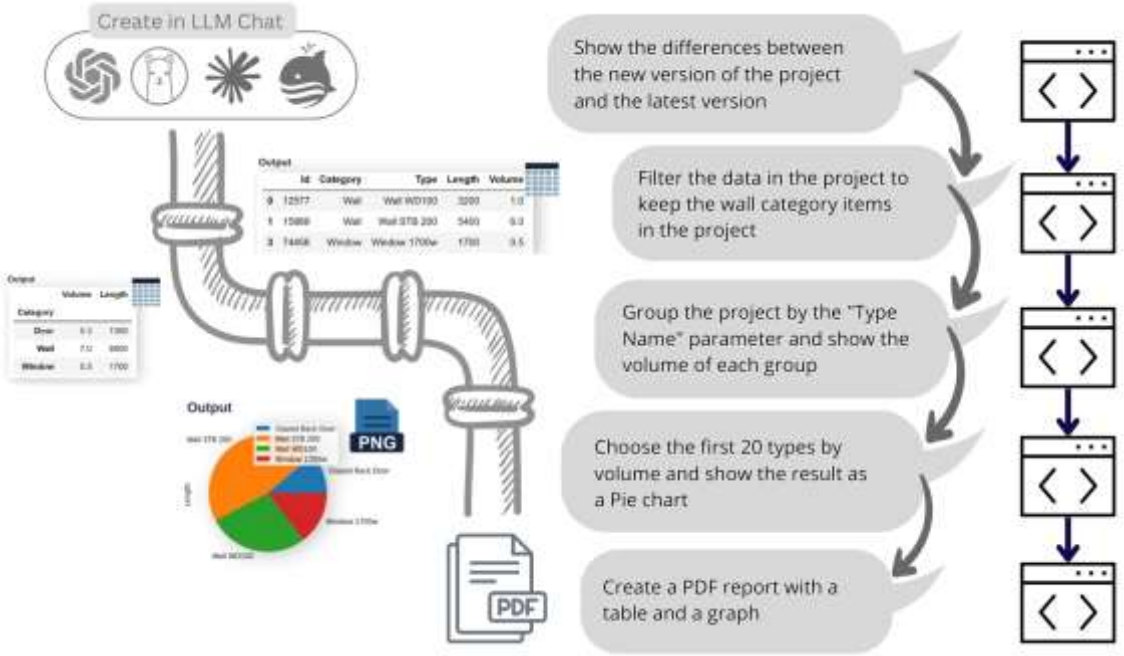


图 7.2-22 AI LLM 帮助生成可应用于未来项目的代码草案，而无需 LLM。

> 利用 ChatGPT 、LlaMa 、Mistral 、Claude 、DeepSeek 、QWEN 、Grok 等工具，以及开放数据 和开放源代码软件 ，我们可以将以前只能通过专业化、高成本和难以维护的模块化专有系统完成的流程自动化。

就建筑业而言，这意味着率先实施自动化管道 - 数据流程的公司将获得巨大收益，从提高项目管理效率到减少财务损失，再到消除分散的应用程序和孤立的数据仓库。

ETL 流程 中描述的业务任务执行逻辑是分析 和数据处理流程自动化的重要组成部分，是更广泛概念--管道（Pipelines）--的具体变体。

章 7.3.

自动 **etl** 输送机（管道）

管道： 自动 **ETL** 输送机 数据

ETL 流程 传统上用于处理分析系统中的数据，包括结构化和非结构化数据源。不过，在当今的数字化环境中，人们越来越多地使用一个更宽泛的术语--管道 （输送机 ），它描述了任何连续的处理链，其中一个阶段的输出成为下一个阶段的输入。

这种方法不仅适用于数据，也适用于其他类型的自动化：任务处理、建立报告、与软件和数字工作流 （图 7.3-1）。

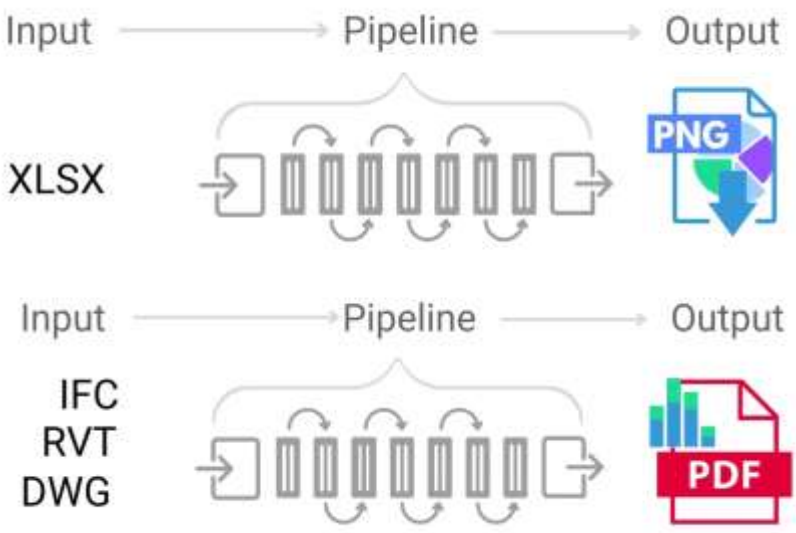


图 7.3-1 管道 是一个处理序列，其中一个阶段的输出成为下一阶段的输入。

使用管道 是自动化的主要元素之一，尤其是在处理大量异构数据时。管道架构允许以模块化、一致和可管理的格式组织复杂的处理步骤，从而提高了可读性，简化了代码维护，并实现了增量调试和可扩展测试。

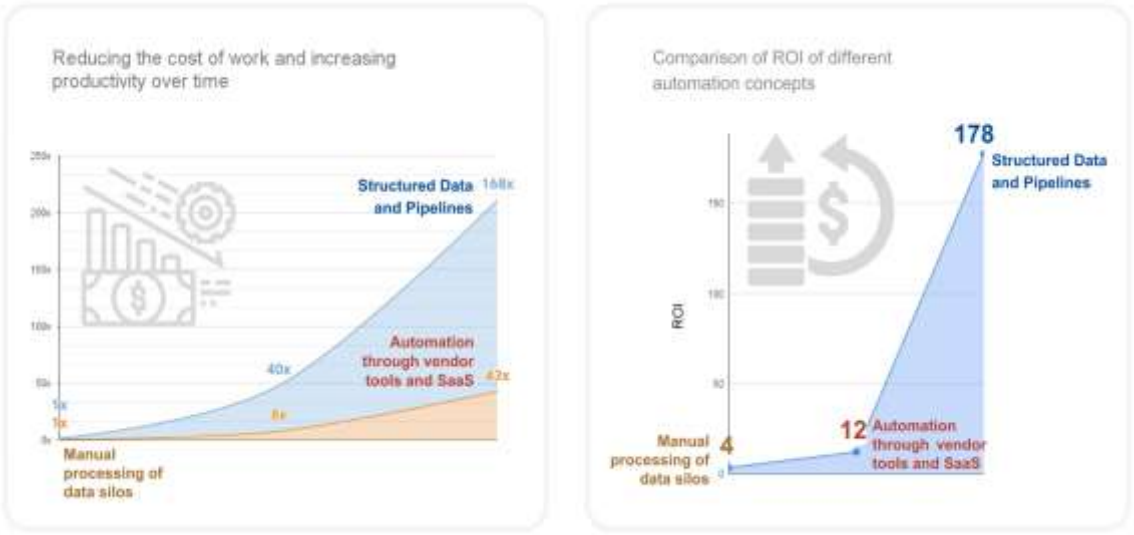


图 7.3-2 ROI Pipeline 数据验证过程与使用传统工具 [74] 处理相比，执行时间缩短了几十甚至几百倍。

与专有系统（ERP 、PMIS 、CAD 等）中的手工作业不同，流水线作业可以显著（图 7.3-2）提高任务速度，避免重复劳动，并在适当的时间自动启动流程（图 7.3-3）。

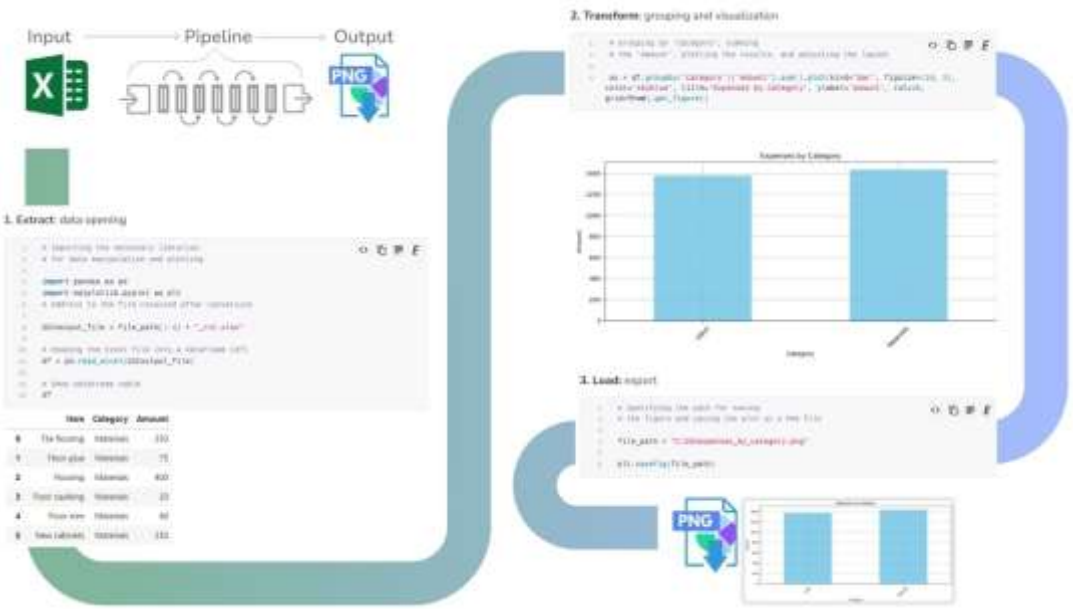


图 7.3-3 ETL 示例 关于在不打开 Excel 的情况下从 XLSX 文件中的表格数据自动检索图表的管道。

要处理流式数据并建立自动化管道（类似于 ETL 流程），您需要提前确定数据来源以及收集数据的时间框架--可以是特定业务流程的数据，也可以是整个公司的数据。

> 在建筑项目中，数据来自许多不同来源，更新时间间隔也各不相同。为了创建可靠的数据展示，记录信息的检索和更新时间至关重要。这样才能及时做出决策，提高项目管理效率。

一种方法是在固定的时间启动装配过程，例如在工作日结束时的 19:00 点。此时，负责汇总各系统和存储数据的第一个脚本被激活（图 7.3-4 第 1 步）。随后是自动处理数据并将其转换为适合分析的结构化格式（图 7.3-4 第 2-4 步）。在最后阶段，利用准备好的数据，自动生成报告、仪表板和前几章所述的其他产品（图 7.3-4 第 6-7 步）。因此，到早上 5:00 时，管理人员已经可以获得所需格式的最新项目状态报告（图 7.3-5）。

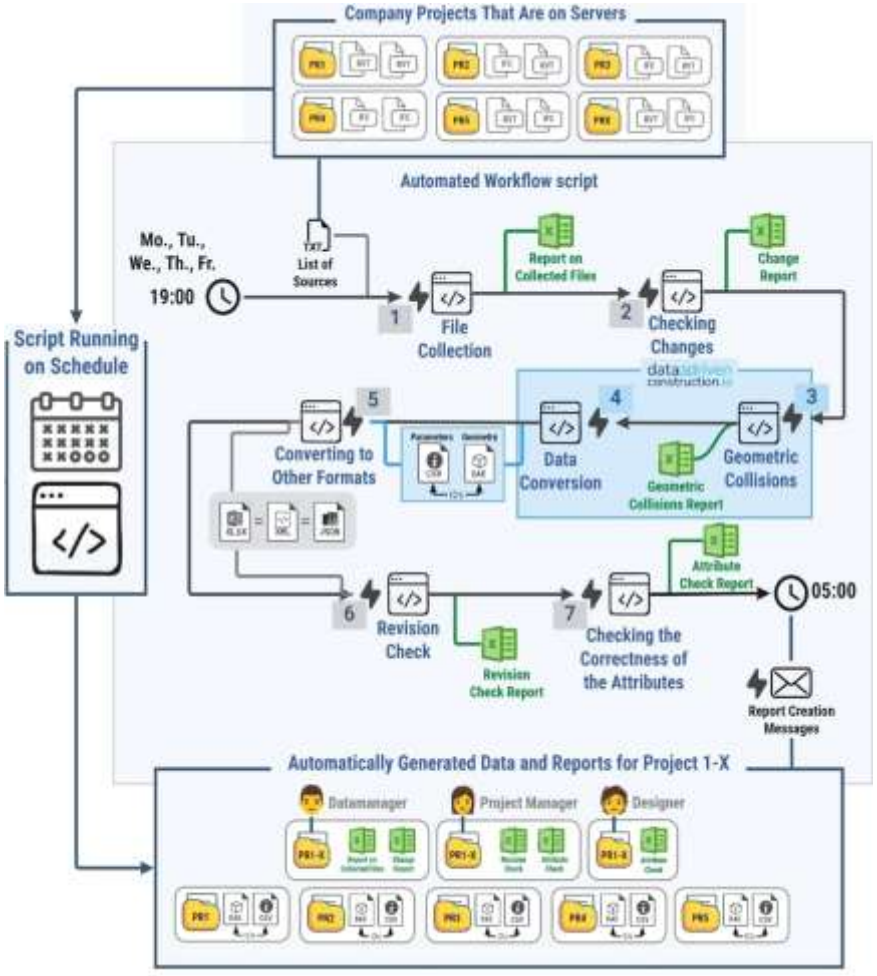


图 7.3-4 管道 中的数据，晚上自动收集，通宵处理，以便管理人员在早上收到最新报告和新报告

。

> 及时的数据收集、关键绩效指标定义、转型流程自动化 以及通过仪表盘实现可视化，都是成功的数据驱动决策的关键要素。

这种自动化流程（图 7.3-4）可以完全自主执行：它们按计划运行，在无人值守的情况下处理数据，既可部署在云中，也可部署在公司自己的服务器上（图 7.3-5）。这样，此类 ETL 管道就可以集成到现有的 IT 基础设施中，保持对数据的控制，并提供扩展的灵活性。

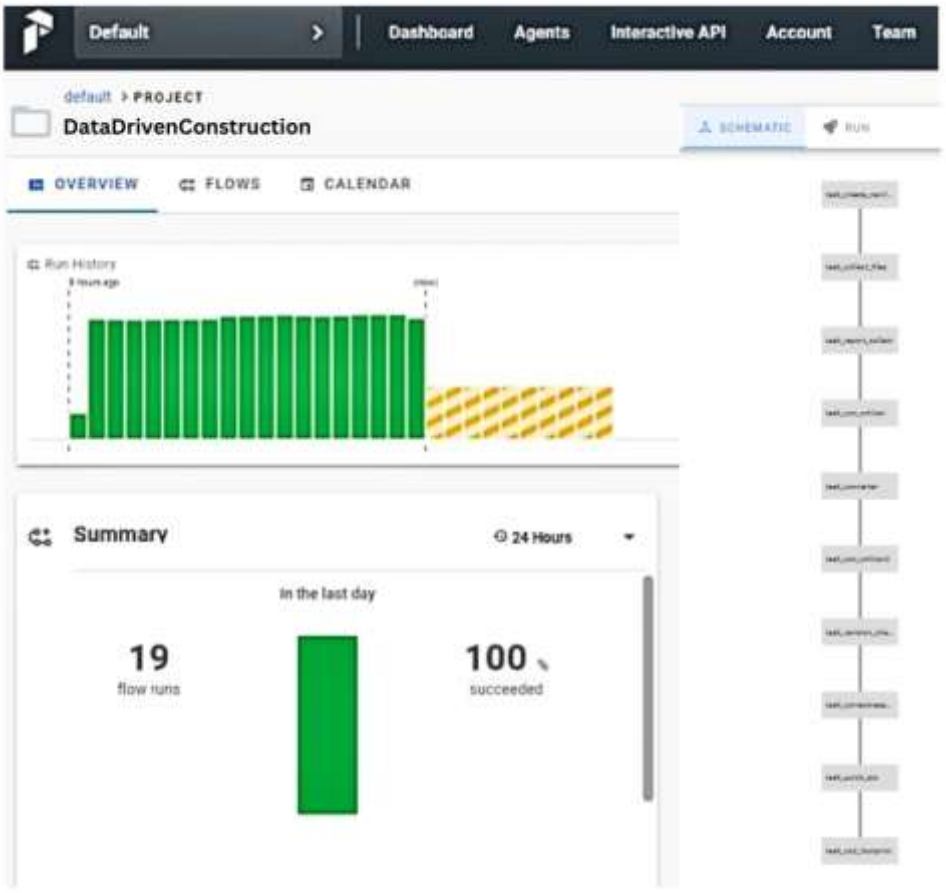


图 7.3-5 Prefect 平台上的自动 ETL - 输送机 流程（图 7.3-4），其中 10 个 python 脚本在每个工作日 19:00 后交替运行。

将工作流程自动化不仅可以腾出时间来完成更有意义、不那么常规的任务，从而提高团队的工作效率，而且还是将人工智能 (AI) 技术融入业务流程的重要第一步，我们将在预测分析和机器学习一章中对此进行更详细的讨论。

管道 -ETL 使用 LLM 进行数据验证过程

在前面关于创建数据需求和 ETL 自动化的章节 中，我们逐步分解了数据准备、转换、验证和可视化的过程。这些活动以独立代码块的形式实现（图 7.2-18 - 图 7.2-20），每个代码块执行一项特定任务。

现在，我们的下一个目标是将这些元素整合到一个单一、连贯和自动化的数据处理管道 - 一个管道，即 ETL

- 管道 - 其中所有阶段（加载、验证、可视化、导出）都在一个可自动执行的脚本中按顺序执行。

在下面的示例中，将实现一个完整的数据处理周期：从加载源 CSV 文件 →到使用正则表达式检查结构和值 →计算结果→ 生成 PDF 格式的可视化报告。

- 您可以在 LLM 中使用以下文本查询来获取相应的代码：

> 请编写一个代码示例，从 CSV 中加载数据，使用正则表达式 验证 DataFrame ，检查格式为 "W-NEW "或 "W-OLD "的标识符、从字母 "A "到 "G "的能效、以年为单位的保修期和更换周期，最后创建一份报告，其中包含通过和未通过值的计数，生成一份包含结果直方图的 PDF ，并添加文字说明。 ⏎

LLM 的答复：




```python
# Importing necessary libraries
import pandas as pd
import re
import matplotlib.pyplot as plt
from matplotlib.backends.backend_pdf import PdfPages


# Load data
df = pd.read_csv('path_to_csv_file.csv')

# Define patterns and verify data
patterns = {
    'Window_ID': r'^(W-NEW|W-OLD[0-9]+|D-1[0-9]{2})$',
    'Energy_Efficiency': r'^[A-G]\+*$',
    'Warranty_Period': r'^\d+\syears$',
    'Replacement_Cycle': r'^\d+\syears$'
}

for column, pattern in patterns.items():
    df['verified_' + column] = df[column].apply(lambda x: bool(re.match(pattern, str(x))))

# Create report
report = df.filter(like='verified_').apply(lambda x: x.value_counts()).T
report.columns = ['Passed', 'Failed']

# Generate PDF
pdf_path = '/mnt/data/verification_results_full_pipeline.pdf'
with PdfPages(pdf_path) as pdf:
    # Plot and save bar chart
    report.plot(kind='bar', stacked=True, color=['green', 'red'])
    plt.title('Data Verification Summary')
    plt.xlabel('Data Categories')
    plt.ylabel('Count')
    plt.tight_layout()
    pdf.savefig()
    plt.close()

    # Add text page
    plt.figure()
    plt.axis('off')
    text_str = "Verification Results Summary"
    plt.text(0.1, 0.5, text_str, wrap=True)
    pdf.savefig()
    plt.close()

pdf_path

```

图 7.3-6 Pipeline (ETL) 实现数据处理整个周期的自动化：从下载和验证到创建 PDF 格式的结构

化报告。

LLM 聊天室或 DIE 中的自动代码（图 7.3-6）在复制代码后，将使用指定的正则表达式 验证 CSV 文件中的数据，创建一份关于通过和未通过记录数的报告，然后将验证结果保存为 PDF 文件。

这种 ETL - 输送机结构（从数据加载到报告生成的每个步骤都作为独立模块实现）确保了透明度、可扩展性和可重复性。将验证逻辑以易于阅读的 Python 代码呈现，不仅使开发人员，而且使数据管理、质量和分析方面的专家都能透明地理解流程。

> Pipeline 的 自动化数据处理方法可实现流程标准化，提高流程的可重复性，并简化对新项目的适应性。这样，无论数据来源或任务类型如何，无论是合规性测试、生成报告还是将数据传输到外部系统，都能创建统一的数据分析方法。

这种自动化减少了人为错误，降低了对专有解决方案的依赖，提高了结果的准确性和可靠性，使其既适用于项目层面的运营分析，也适用于公司层面的战略分析。

管道 **-ETL:** 核查 **CAD** 中项目元素的数据和信息 **(BIM)**

来自 CAD 系统和数据库 (BIM) 的数据是建筑公司业务中最复杂和动态更新的数据源。这些应用程序不仅使用几何图形描述项目，而且还辅以多层文字信息：体积、材料属性、房间分配、能效等级、公差、预期寿命和其他属性。

> 在 CAD - 模型中分配给实体的属性是在设计阶段形成的，并成为进一步业务流程的基础，包括成本计算、进度安排、生命周期评估以及与 ERP 和 CAFM - 系统的集成，其中流程的效率在很大程度上取决于来自设计部门的数据 的质量。

CAD- (BIM-) 模型中属性验证的传统方法包括人工验证（图 7.2-1），当模型量较大时，这个过程会变得漫长而昂贵。考虑到现代建筑项目的规模和数量及其定期更新，数据验证和转换过程变得不可持续且难以承受。

> 总承包商和项目经理需要处理大量的项目数据，包括同一模型的多个版本和片段。这些数据来自 RVT、DWG、DGN、IFC、NWD 和其他格式的设计机构（图 3.1-14），需要定期检查是否符合行业和公司标准

对人工操作和专业软件的依赖使数据验证过程成为与全公司模型数据相关的工作流的瓶颈。自动化和使用结构化需求可以消除这种依赖性，显著提高数据验证的速度和可靠性（图 7.3-7）。

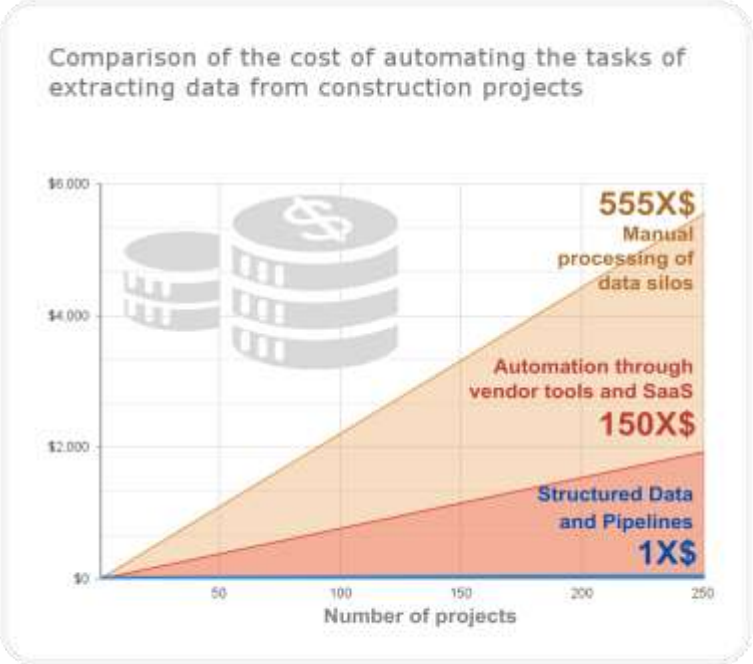


图 7.3-7 自动化提高了数据核查和处理的速度，使工作成本降低了数十倍[140]。

CAD 数据验证过程包括从各种封闭式（RVT, DWG, DGN, NWS 等）或开放式半结构化和参数化格式（IFC, CPXML, USD）中提取数据（ETL 阶段 提取）。在此阶段，可以使用正则表达式 RegEx （图 7.3-8）将规则表应用于每个属性及其值（转换阶段），我们将在本书第四部分详细讨论这一过程。

在创建 PDF 错误报告（和成功验证记录时，应以结构化格式完成输出（加载步骤），只考虑可用于进一步流程的已验证实体。

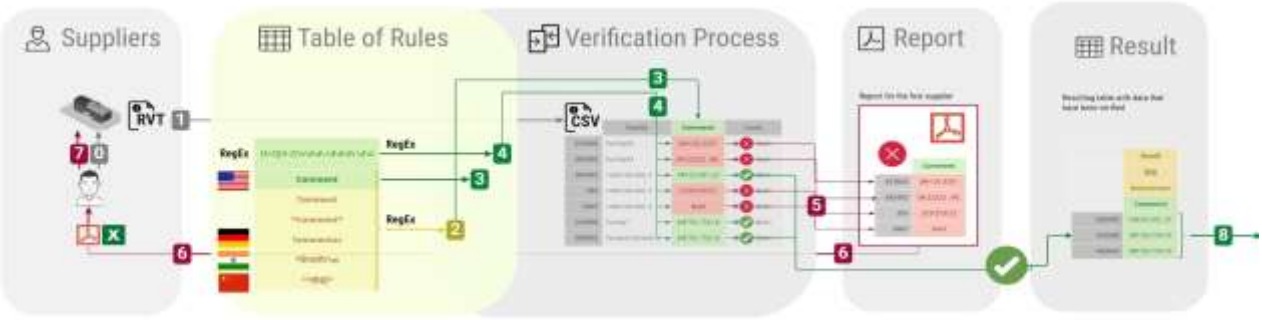


图 7.3-8 使用正则表达式验证从项目数据提供者到最终报告的数据验证过程。

通过 ETL 管道（图 7.3-9）自动验证来自 CAD 系统 (BIM)、结构化需求和新数据流的数据，可减少验证过程中的人工参与（前几章已讨论过每个验证和数据需求过程）。

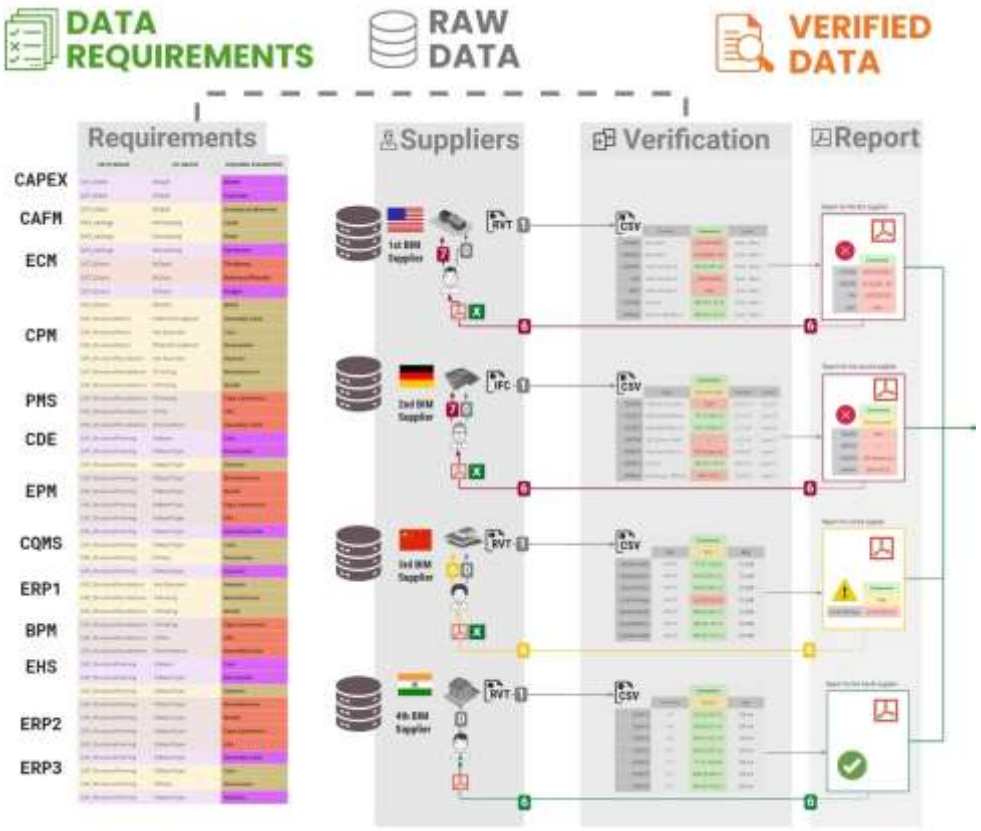


图 7.3-9 通过 ETL 自动进行数据验证可加快流程，从而简化施工项目管理。

传统上，对承包商和 CAD（BIM）专家提供的模型进行验证可能需要几天到几周的时间。然而，随着自动

化 ETL 流程的引入，验证时间可以缩短到几分钟。在一种典型的情况下，承包商会说"模型已验证且符合要求"。这句话开启了承包商数据质量声明的验证链：

- 项目经理 - "承包商说，'模型已经过测试，一切正常'"。
- 数据管理器 - 加载验证：
  - Pandas 中的一个简单脚本就能在几秒钟内检测到违规行为。自动化消除了争议：
    - 类别: OST_StructuralColumns, 参数: FireRating IS NULL。
    - 生成违规 ID 列表→ 导出到 Excel/PDF。

  Pandas 中的一个简单脚本就能在几秒钟内检测到漏洞：

```
df = model_data[model_data["类别"] == "OST_StructuralColumns"]  # 筛选
issues = df[df["FireRating"].isnull()] # 空值  # 空值
issues[["ElementID"]].to_excel("fire_rating_issues.xlsx") # 导出 ID
```

- 数据管理员致项目管理员 - " 的检查结果显示，有 18 列没有填充 FireRating 参数"。
- 项目经理致承包商--"模型被退回修改：FireRating 参数是强制性的，没有它就无法验收"。

因此，CAD 模型无需经过验证，自动化消除了争议，承包商几乎可以立即收到一份结构化报告，其中包含有问题元素的 ID 列表。这样，验证过程就变得透明、可重复并可避免人为错误（图 7.3-10）。

> 这种方法将数据验证流程转变为工程功能，而非人工质量控制。这不仅能提高生产率，还能将相同的逻辑应用于公司的所有项目，实现从设计到运营的端到端流程数字化转型。

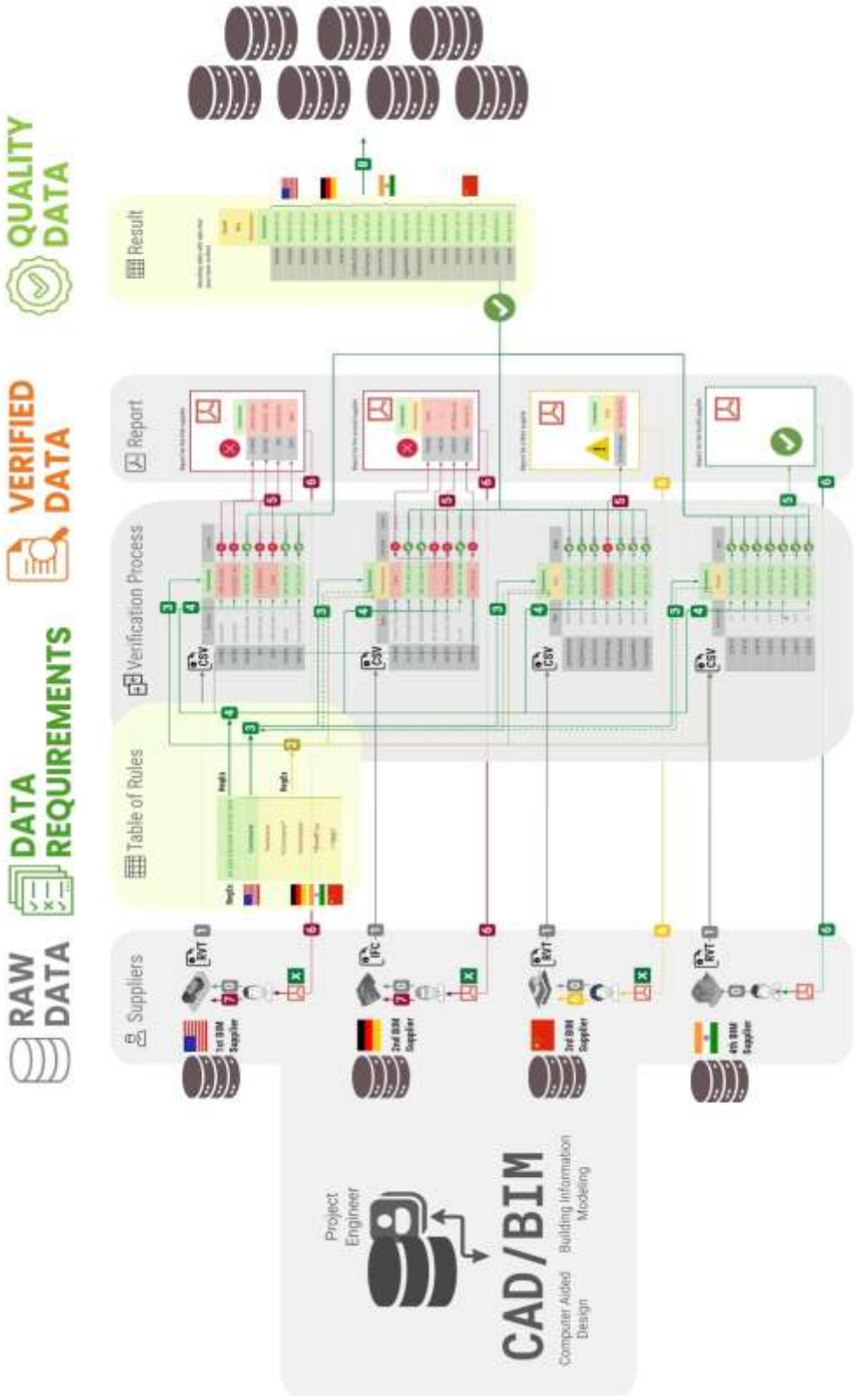

QUALITY DATA
VERIFIED DATA
DATA REQUIREMENTS
RAW DATA
Result
Report
Verification Process
Table of Rules
Suppliers
CAD/BIM
Computer Aided Design
Building Information Modeling
Project Engineer

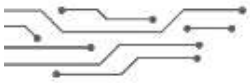
图 7.3-10 自动检查元素属性可消除人为错误并降低出错概率。

通过使用自动化管道 (图 7.3-10)，期待从 CAD- (BIM-) 系统获得高质量数据的系统用户可以立即获得所需的输出数据--表格、文档、图像--并将其快速集成到工作任务中。

> 控制、处理和分析的自动化正在推动建筑项目管理方式的变革，特别是不同系统的互操作性，而无需使用复杂而昂贵的模块化专有系统或封闭的供应商解决方案。

虽然概念和营销缩略语层出不穷，但数据要求验证流程本身将永远是业务流程不可分割的一部分。建筑行业与其创建越来越多的专用格式和标准，不如借鉴已在其他行业证明有效的工具。如今，有一些功能强大的平台可以实现数据处理和流程整合的自动化，使企业能够大大减少日常操作所花费的时间，并最大限度地减少提取 、转换 和加载 中的错误。

Apache Airflow 是 ETL 流程自动化和协调解决方案的流行范例之一，可用于组织复杂的计算流程和管理 ETL 管道。除 Airflow 外，其他类似解决方案也得到了广泛应用，如用于数据路由和流的 Apache NiFi 和用于业务流程自动化的 n8n。



章 7.4.

协调 **etl** 和工作流程：实用解决方案

DAG 和 **Apache Airflow**： 工作流自动化和协调

Apache Airflow 是一个免费的开源平台，旨在自动化、协调和监控工作流（ETL -conveyors）。

每天都需要处理大量数据：

- 从不同来源下载文件 - 提取 （例如，从供应商或客户处）。
- 将这些数据转换为所需格式 - 转换 （结构、清理和验证）
- 发送结果以进行验证并创建报告 - 加载 （上载到所需系统、文件、数据库或仪表板）。

手动执行此类 ETL 流程需要花费大量时间，并存在人为错误的风险。数据源的变化或某个步骤的失败都可能导致延迟和错误结果。

自动化工具，如 Apache Airflow，可让您构建可靠的 ETL - 输送机，最大限度地减少错误，缩短处理时间，并确保数据在每个阶段都正确无误。Apache Airflow 的核心是 DAG （有向无环图）的概念--在有向无环图中，每个任务（操作符）都与其他依赖关系相连，并严格按照指定顺序执行。DAG 消除了循环，为任务执行提供了一个逻辑和可预测的结构。

Airflow 负责协调工作--管理任务之间的依赖关系、控制执行时间表、跟踪状态并自动应对故障。这种方法最大限度地减少了人工干预，确保了整个流程的可靠性。

> **Task Orchestrator** 是一种工具或系统，旨在管理和控制复杂计算和信息环境中的任务执行。它可促进任务执行的部署、自动化和管理过程，从而提高性能并优化资源。

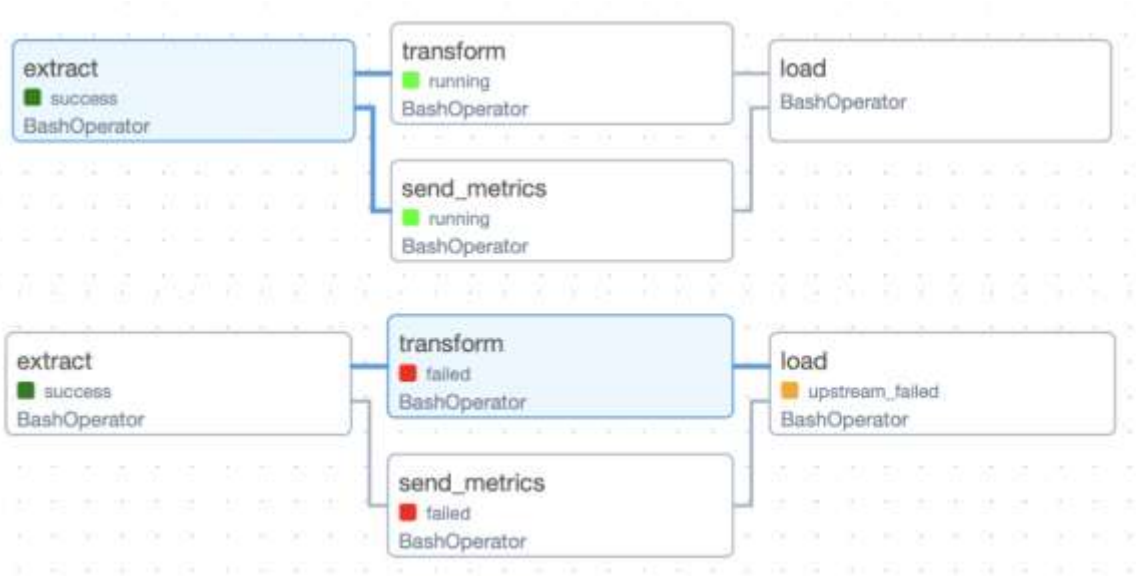


图 7.4-1 Apache Airflow 提供了一个用户友好型界面，您可以在此可视化 DAG -ETL 、查看执行日志、任务启动状态等。

Airflow 广泛应用于分布式计算的协调和自动化、数据处理、ETL （Extract, Transform, Load ）流程管理、任务调度和其他数据场景。默认情况下，Apache Airflow 使用 SQLite 作为数据库。

简单 DAG 的一个示例，类似于 ETL ，由任务组成--提取 、转换 和加载。在通过用户界面管理的图中（图 7.4-1），定义了任务（代码片段）的执行顺序：例如，先执行提取，然后转换（和发送度量），最后加载任务完成工作。当所有任务都完成后，数据加载过程就算成功。

Apache Airflow： ETL 自动化的实际应用

Apache Airflow 广泛用于组织复杂的数据处理流程，允许构建灵活的 ETL -输送机。Apache Airflow 既可通过网络界面运行，也可通过 Python 代码编程运行（图 7.4-2）。在 Web 界面（图 7.4-3）上，管理员和开发人员可以直观地监控 DAG、运行任务并分析执行结果。

使用 DAG ，您可以设置明确的任务顺序，管理任务之间的依赖关系，并自动对源数据的变化做出反应。让我们以使用 Airflow 自动处理报告为例（图 7.4-2）。

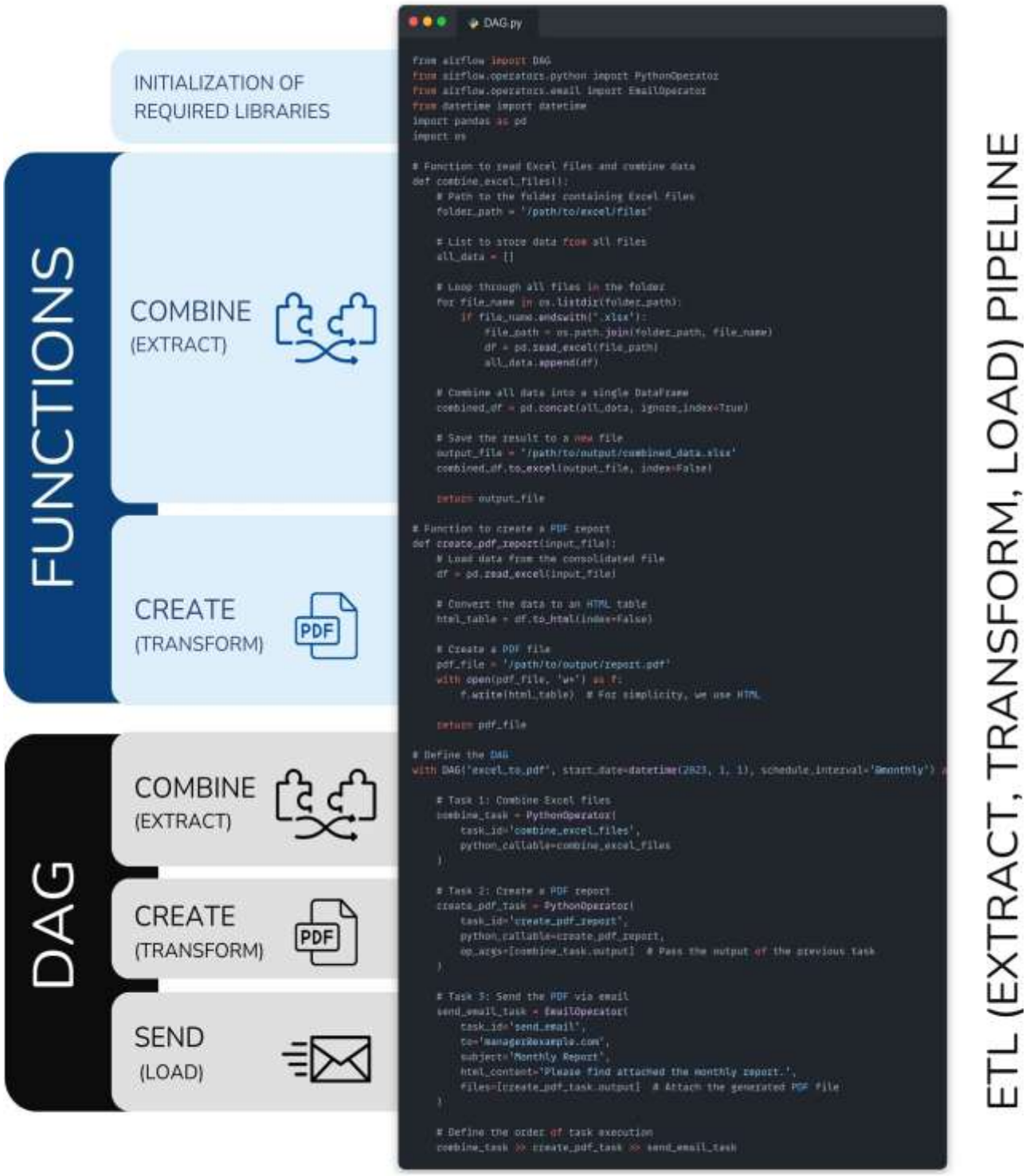


图 7.4-2 ETL 概念 - 使用 Apache Airflow 处理数据的传送带。

本示例（图 7.4-2）考虑了 DAG，它在 ETL - 输送机中执行关键任务：

- 读取 **Excel -files**（提取）：
  - 顺序遍历指定目录下的所有文件。
  - 使用 pandas 库从每个文件中读取数据。
  - 将所有数据合并为一个 DataFrame。
- 创建 **PDF -**文档（**Transform**）：
  - 将合并后的 DataFrame 转换为 HTML -表格。
  - 将表格保存为 PDF 格式（在演示版本中--通过 HTML）。
- 通过电子邮件发送报告（加载）：
  - 应用 EmailOperator 通过电子邮件发送 PDF 文档。
- 配置 **DAG**：
  - 定义任务顺序：提取数据→ 生成报告→ 发送。
  - 指定启动时间表（@monthly - 每月第一天）。

自动 ETL - 示例（图 7.4-2）显示了如何从 Excel - 文件中收集数据、创建 PDF - 文档并通过电子邮件发送。这只是 Airflow 的众多可能用例之一。此示例可适用于任何特定任务，以简化数据处理并使其自动化。

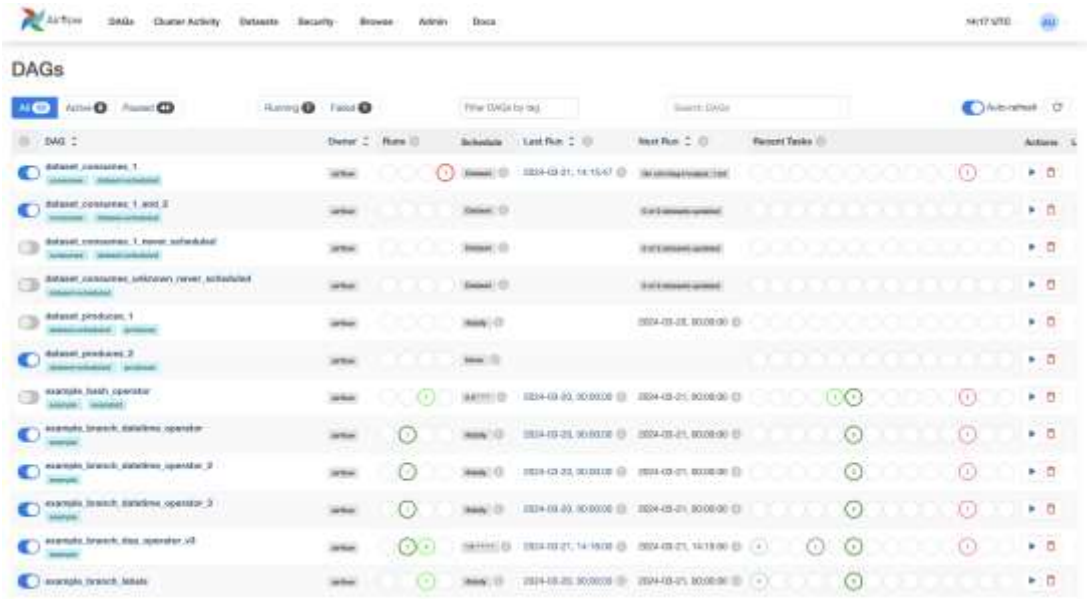


图 7.4-3 环境中所有 DAG 的概览，以及最近运行的信息。

Apache Airflow 网页界面（图 7.4-3）为管理数据工作流提供了一个全面的可视化环境。它将 DAG 显示为交互式图形，其中节点代表任务，边代表任务之间的依赖关系，从而便于跟踪复杂的数据工作流。界面包括一个仪表盘，其中包含任务执行状态、运行历史、详细日志和性能指标等信息。管理员可以通过直观的用户界面手动启动任务、重启失败的操作、暂停 DAG 和配置环境变量。

这种架构可以通过数据验证、执行状态通知、与外部应用程序接口 或数据库集成来补充。Airflow 允许您灵活调整 DAG ： 添加新任务、更改任务顺序、合并任务链--这使其成为复杂数据处理流程自动化的有效工具。在 Airflow 网页界面（图 7.4-3，图 7.4-4）上运行 DAG 时，您可以监控任务的状态。系统采用彩色指示：

- 绿色 - 任务已成功完成。
- 黄色 - 流程正在进行中。
- 红色 - 执行任务时出错。

如果出现故障（如文件丢失或数据结构损坏），系统会自动发送通知。

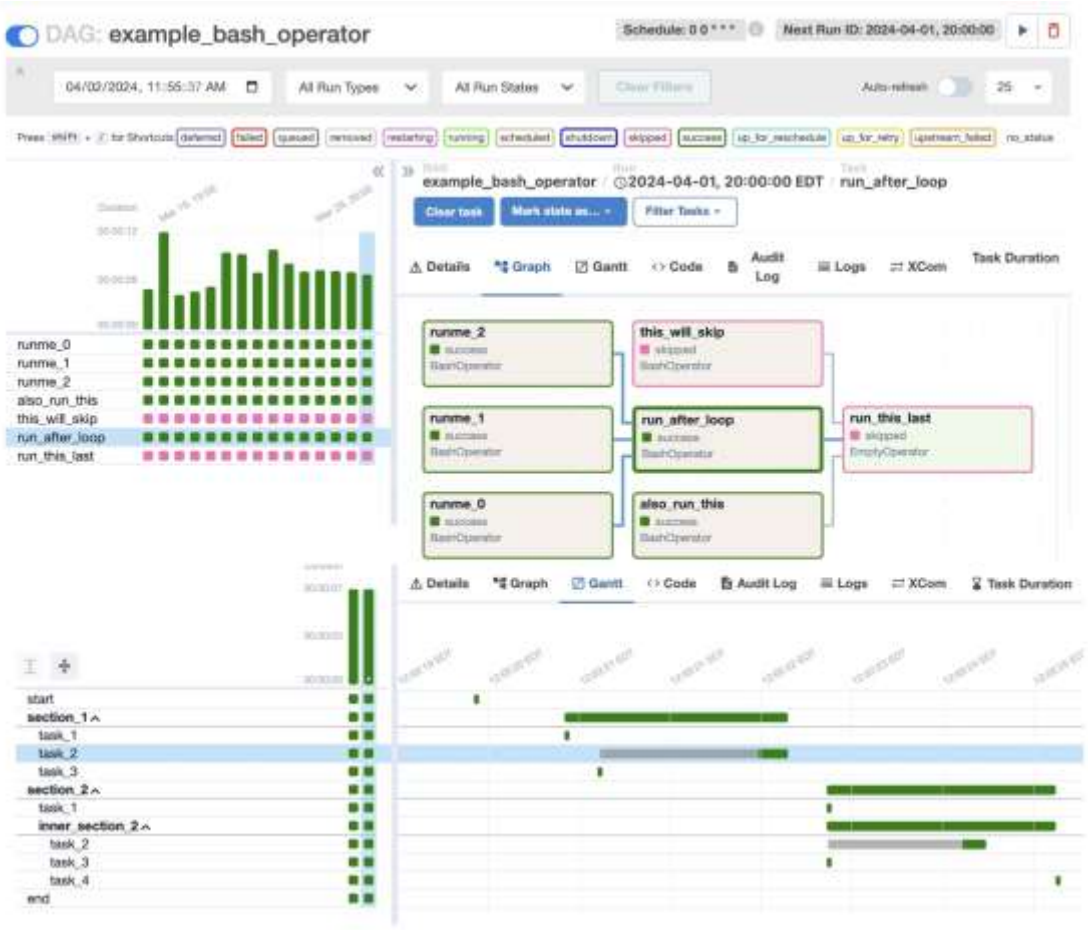


图 7.4-4 Apache Airflow 极大地简化了复杂数据处理管道的问题诊断、流程优化和团队协作。

> Apache Airflow 可自动执行日常任务，无需手动操作，因此非常方便。它通过监控流程执行和提供即时错误通知来确保可靠性。系统的灵活性使其可以轻松添加新任务或修改现有任务，调整工作流程以满足不断变化的要求。

除了 Apache Airflow 之外，还有一些类似的工作流协调工具。例如，开源免费的 Prefect（图 7.3-5）提供了更简单的语法，并能更好地与 Python 集成，Spotify 开发的 Luigi 提供了类似的功能，并能很好地处理大数据。值得注意的还有 Kronos 和 Dagster，它们提供了构建管道 的现代方法，注重模块化和可扩展性。任务协调工具的选择取决于项目的具体需求，但它们都有助于自动化复杂的 ETL 数据流程

特别值得注意的是 Apache NiFi，这是一个开源平台，专为数据流和路由设计。与侧重于批处理和依赖性管理的 Airflow 不同，NiFi 侧重于实时、即时的数据转换和系统间的灵活路由。

Apache NiFi 用于路由选择和数据转换

Apache NiFi 是一个功能强大的开源平台，旨在自动处理不同系统之间的数据流。最初由美国国家安全局（NSA）于 2006 年开发，名为 "Niagara Files"，供内部使用。2014 年，该项目开源并转让给阿帕奇软件基金会，成为其技术转让计划的一部分 [141]。

Apache NiFi 设计用于实时收集、处理和传输数据。与使用批处理任务并需要明确时间表的 Airflow 不同，NiFi 采用流处理模式，允许数据在不同服务之间连续传输。

> Apache NiFi 非常适合与物联网 设备、建筑传感器、监控系统集成，例如在服务器上对 CAD 格式进行流验证，因为在这种情况下可能需要对数据变化做出即时响应。

通过内置的过滤、转换和路由工具，NiFi 可以在将数据传输（加载 ）到存储或分析系统之前对数据进行标准化处理（转换 ）。其主要优势之一是内置的安全支持和访问控制，使其成为处理敏感信息的可靠解决方案。

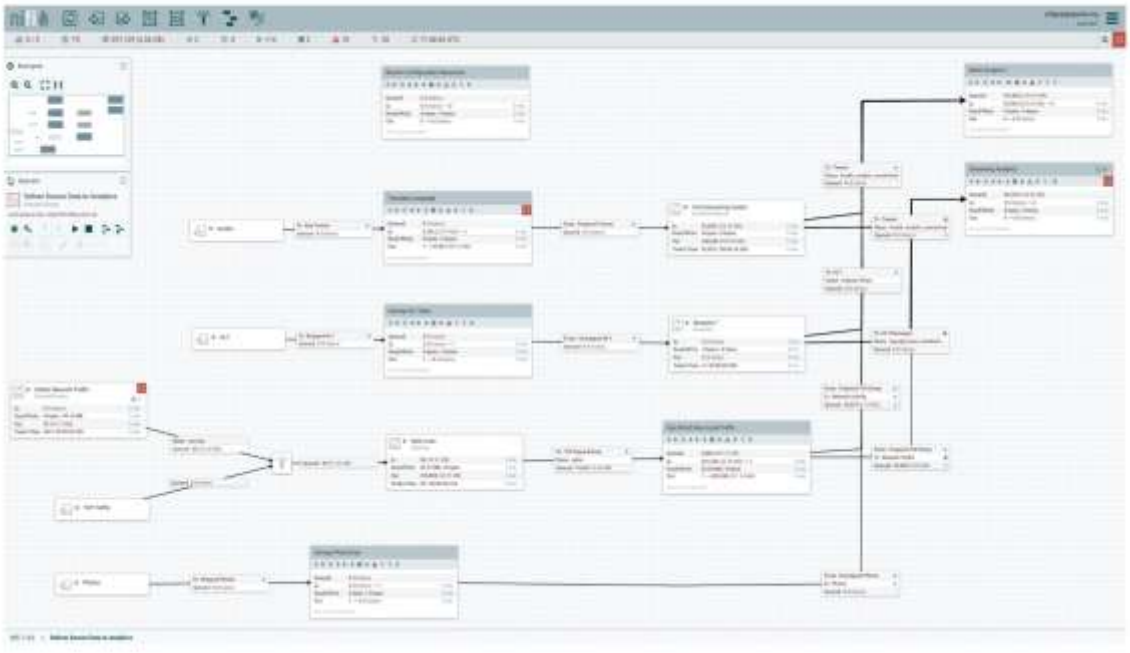


图 7.4-5 Apache NiFi 接口中数据流的图形表示。

Apache NiFi 可高效处理实时数据流、过滤和路由任务。它非常适合技术密集型场景，在这些场景中，系统

间的稳定信息传输和高吞吐量非常重要。

然而，当主要目标是整合各种服务、实现日常操作自动化以及在不需要深厚编程知识的情况下快速建立工作流程时，就需要入门门槛低且具有最大灵活性的解决方案。n8n 就是这样一个工具，它是一个低代码 / 无代码 类平台，专注于业务自动化和流程的可视化协调。

n8n 低代码**,** 无代码 流程协调

n8n 是一个开源的低代码 / 无代码 平台，用于构建自动化工作流程，具有易用性、灵活性和与各种外部服务快速集成的能力。

> 无代码 是一种无需编写代码即可创建数字产品的方法。创建过程中的所有元素--从逻辑到界面--都完全借助可视化工具来实现。无代码平台面向没有技术背景的用户，可以快速创建自动化、表单、集成和网络应用程序。例如：用户无需编程知识，即可通过拖放界面设置自动通知发送或 Google Sheets 集成。

n8n 具备开源和本地部署能力，在自动化和 ETL Pipelines 创建过程中，公司可完全控制其数据，同时确保安全并独立于云提供商。

Apache Airflow 面向具有严格协调性的计算任务，需要 Python 知识 ，而 n8n 则不同，它提供了一个可视化编辑器，无需编程语言知识即可编写脚本（图 7.4-6）。虽然其界面允许在不编写代码的情况下创建自动化流程（无代码 ），但在更复杂的情况下，用户可以添加自己的 JavaScript 和 Python 函数来扩展功能（低代码 ）。

> 低代码 是一种软件开发方法，在这种方法中，应用程序或流程的基本逻辑是通过图形界面和可视化元素创建的，程序代码仅用于定制或扩展功能。低代码平台不仅能让程序员参与进来，还能让具备基本技术技能的企业用户参与进来，从而大大加快解决方案的开发速度。例如：用户可以利用现成的模块构建业务流程，必要时还可以在 JavaScript 或 Python 中添加自己的脚本。

尽管 n8n 被定位为一个入门门槛较低的平台，但它需要具备基本的编程知识、对网络技术的了解以及使用 API 的技能。该系统的灵活性使其能够适应各种任务--从自动数据处理到与信使、物联网 - 设备和云服务的集成。

使用 n8n 的主要功能和优势：

- 开放源码和本地部署选项可确保完全的数据控制、安全合规性以及独立于云提供商。
- 与超过 **330** 种服务集成，包括 CRM 、ERP 、电子商务、云平台、信使和数据库。
- 场景灵活性：从简单的通知到复杂的链条，包括对 - 请求、决策逻辑和人工智能 - 服务的 API 处理。
- 支持 **JavaScript** 和 **Python**：用户可根据需要嵌入自定义代码，扩展自动化功能。
- 直观的可视化界面：可以快速配置和可视化所有流程步骤。

低代码类平台 提供了用最少代码创建数字解决方案的工具，非常适合不具备深厚技术专业知识但需要实现流程自动化的团队。

在建筑领域，n8n 可用于自动化各种流程，如与项目管理系统集成、流检查、编写开箱即用的 报告和信函、自动更新材料库存数据、向团队发送任务状态通知等。n8n 中定制的管道 可成倍减少人工操作，降低出错的可能性，加快项目执行的决策速度。

> 您可以从 n8n.io/workflows 提供的近两千个现成、免费和开源的 n8n Pipeline 中进行选择，实现建筑工作流程和个人任务的自动化，减少日常操作。

以 n8n.io [142] 上免费提供的现成管道模板 为例，它可以在 Gmail 中自动创建回复草稿（图 7.4-6），帮助那些收到大量电子邮件或难以撰写回复的用户。

此 n8n "Gmail AI 自动回复：创建传入邮件的回复草稿 "模板（图 7.4-6）使用 ChatGPT 的 LLM 分析传入邮件，确定是否需要回复，从 ChatGPT 生成草稿并将文本转换为 HTML ，然后将其添加到 Gmail 的消息链中。这不会自动发送电子邮件，您可以手动编辑和批准回复。设置大约需要 10 分钟，包括 Gmail API 的 OAuth 配置和 OpenAI API 集成。因此，这是一个方便、免费的解决方案，可在不失去对电子邮件内容控制的情况下实现日常电子邮件通信的自动化。

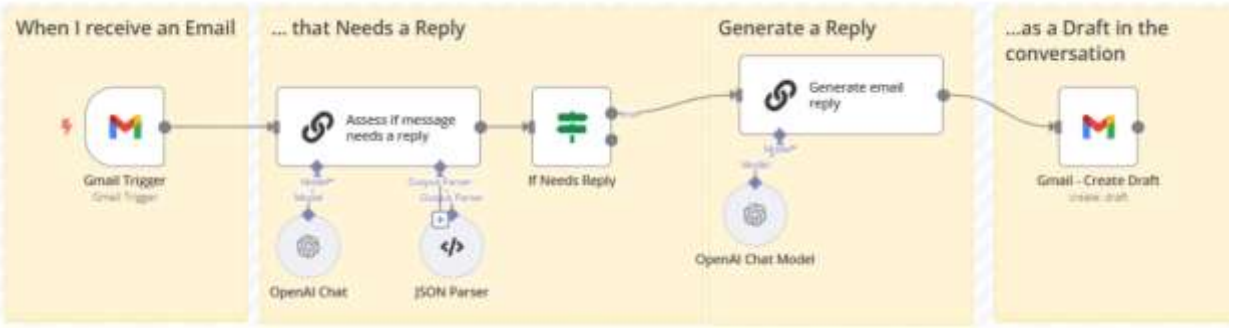


图 7.4-6 使用 n8n 自动生成电子邮件回复的过程。

使用 n8n 实现自动化的另一个例子是在房地产市场上寻找优惠 [143]。N8n Pipeline "利用 Zillow API、Google Sheets 和 Gmail 实现每日房产交易自动化"，利用 Zillow API 收集符合给定条件的每日相关报价。它自动计算关键投资指标（现金投资回报率 、每月现金流、首付），更新 Google Sheets ，并通过电子邮件发送总结报告（图 7.4-7），让投资者节省时间，快速响应最佳报价。



图 7.4-7 评估房地产投资吸引力的自动化流程。

它的灵活性和可扩展性使 n8n 成为公司寻求数字化转型的重要工具，并通过相对简单和免费的开源工具 提高市场竞争力。

Apache NiFi 、Airflow 和 n8n 等工具可视为数据处理的三个层次（图 7.4-8）。NiFi 管理数据流，确保数据的交付和转换；Airflow 通过将数据汇聚到处理管道来协调任务的执行；n8n 自动集成外部服务并管理业务逻辑。





	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

图 7.4-8 Apache Airflow 、Apache NiFi 和 n8n 可视为现代数据管理架构的三个互补层。

这些免费的开源工具共同构成了建筑行业数据和流程管理的有效生态系统，使公司能够利用信息进行决策和实现流程自动化。

下一步：从人工操作转向基于分析的解决方案

当今的建筑公司在高度不确定的环境中运营：材料价格不断变化、交货延迟、劳动力短缺和项目期限紧迫。使用分析仪表板、ETL - 输送机和商业智能系统可帮助企业快速识别问题领域、评估资源效率并预测变化 ，以免造成经济损失。

在总结这部分内容时，值得强调的是有助于您在日常工作中应用所讨论技术的主要实用步骤：

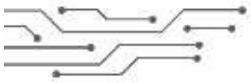
- 实施数据可视化和分析仪表板
  - ☐ 掌握创建仪表盘的流程，以监控关键绩效指标 (KPIs)
  - ☐ 使用数据可视化工具（Power BI, Tableau, Matplotlib, Plotly)
- 通过 ETL 实现数据处理自动化
  - ☐ 通过 ETL 流程从各种来源（文档、表格、CAD ）设置自动数据收集功能
  - ☐ 使用 Python 脚本组织数据转换（如正则表达式检查或计算
  - ☐ 尝试使用 FPDF 库设置自动 PDF（或 DOC）报告，使用 Excel 文件中的数据或从其他 PDF

文档中提取信息

- 使用语言模型 (LLM) 实现自动化
  - 使用大型语言模型 (LLM) 生成代码，帮助从非结构化文档中提取和分析数据
  - 熟悉 n8n 的自动化工具，探索其网站上的现成模板和案例研究。使用无代码/低代码方法确定哪些流程可以完全自动化

数据分析方法和流程自动化不仅能减少日常操作所花费的时间，还能提高决策质量。实施可视化分析工具和 ETL - 输送机的公司有机会对变化做出快速反应

使用 n8n 、Airflow 和 NiFi 等工具实现业务流程自动化只是迈向数字化成熟的第一步。下一步是高质量地存储和管理作为自动化基础的数据。在第 8 部分中，我们将深入探讨建筑公司如何建立可持续的数据存储架构，从混乱的文档和多格式文件转向集中存储和分析平台。





VIII 部分

建筑数据的存储和管理

第 8 部分探讨了建筑行业的现代数据存储和管理技术。它分析了处理大量信息的有效格式--从简单的 CSV 和 XLSX 到性能更高的 Apache Parquet 和 ORC，并详细比较了它们的功能和局限性。讨论了数据仓库 (DWH)、数据湖) 及其混合解决方案 (Data Lakehouse) 的概念，以及数据治理) 和数据最小化) 的原则。详细介绍了数据沼泽（ Data Swamp ）的问题和防止信息系统混乱的策略。本部分介绍了处理数据的新方法，包括矢量数据库 以及通过边界框（Bounding Box）概念在构建中的应用。本部分还介绍了作为组织数据工作流新标准的 DataOps 和 VectorOps 方法。

章 8.1.

数据基础设施：从存储格式到数字资料库

数据原子：有效信息管理的基础

宇宙万物都是由原子和分子这些最小的构件组成，随着时间的推移，所有生物和非生物都不可避免地回归到这种初始状态。在自然界中，这一过程以惊人的速度发生，我们试图将其移植到人类控制的过程中。

> 在森林中，任何生物最终都会转化为一种营养物质，成为新植物的基础。反过来，这些植物又会成为新生物的食物，而新生物是由数百万年前创造宇宙的相同原子构成的。

在商业世界中，同样重要的是将复杂的多层结构分解为最基本、最少量的处理单元--就像自然界中的原子和分子一样。这样就可以有效地存储和管理数据原子，将其转化为丰富、肥沃的基础，成为提高分析 和决策质量的关键资源。



图 8.1-1 分析和决策基于曾经处理和存储的重复使用数据。

音乐作品由音符组合而成，创造出复杂的乐曲，而文字则是由最原始的单位--字母--声音创造出来的。无论是自然、科学、经济、艺术还是技术，在追求破坏、结构、循环和创造的过程中，世界都表现出惊人的统一与和谐。同样，成本核算系统中的流程在成本核算 和进度表的层面上被分解为结构化的微小单位--资源项目。这些单位，就像笔记一样，然后被用来形成更复杂的计算和计划。计算机辅助设计系统也采用同样的原理，复杂的建筑和工程项目是由基本要素--单个要素和库组件--构建而成的，而复杂建筑或结构项目的完整三维。

> 自然和科学中固有的周期性和结构性概念也反映在现代数据世界中。正如在自然界中，所有生物都会还原为原子和分子一样，在现代数据处理工具的世界中，信息也往往会还原为最原始的形式。

具有有限不可分割性的最小元素是业务流程的基本构件。必须从一开始就仔细考虑如何从各种来源收集、组织（分解为原子）和存储这些微小的构件。组织和存储数据不仅仅是将其分解为各个组成部分的问题。同样重要的是，要确保以结构化的方式对它们进行整合和存储，以便在需要时可以轻松检索、分析和使用数据进行决策。

为了高效处理信息，必须精心选择数据存储的格式和方法，就像为树木生长准备土壤一样。必须对数据存储进行组织，以确保信息的高质量和相关性，消除冗余或不相关的数据。这种 "信息土壤 "的结构越合理，用户就能越快越准确地找到正确的数据并解决分析问题。

信息存储：文件或数据

数据仓库使公司能够收集和合并来自不同系统的信息，创建一个单一的中心，用于进一步分析。收集到的历史数据不仅能对流程进行更深入的分析，还能识别可能影响业务绩效的模式。

假设一家公司正在同时进行几个项目。工程师想了解已经浇筑了多少混凝土，还需要购买多少混凝土。如果采用传统方法，他必须手动搜索服务器并打开多个估算表，将其与已完成工程的证书进行比较，并检查当前的库存余额。这需要几个小时甚至几天的时间。即使有 ETL 流程和自动脚本，这项任务仍然是半手工的：工程师仍然需要手动指定服务器上文件夹或特定文件的路径。这就降低了自动化的整体效果，因为它会继续占用宝贵的工作时间。

改用数据管理后，工程师不再使用服务器文件系统，而是可以访问统一的存储结构，并在其中实时更新信息。通过代码、SQL 查询，甚至是 LLM 代理调用的形式进行一次查询，就能立即提供有关当前余额、已完成工作量和即将交付的货物的准确数据，前提是这些数据已提前准备好并合并到数据仓库中，而无需在文件夹中漫步、打开数十个文件和手动比较数值。

长期以来，建筑公司使用 PDF -文档、DWG -图纸、RVT -模型以及成百上千的 Excel -表格和其他不同格式的文件，这些文件存储在公司服务器上的特定文件夹中，给信息搜索、检查和分析带来困难。因此，项目完成后留下的文件通常会被移回服务器的档案存储文件夹中，而这些文件夹在未来实际上是不会被使用的。随着数据流的增加，这种传统的基于文件的数据存储因容易出现人为错误而失去意义。

> 文件只是一个存储数据的独立容器。文件是为人而不是为系统创建的，因此需要手动打开、阅读和解释。例如，Excel 表格、PDF 文档或 CAD 图纸，需要在特定工具中专门打开才能获取所需信息。如果不进行结构化检索和处理，其中的信息就会被闲置。
>
> 数据则是机器可读的信息，可自动链接、更新和分析。在单个数据仓库（如数据库、DWH 或数据湖）中，信息以表格、记录和关系的形式表示。这样就能实现统一存储、自动查询、价值分析和实时报告。

使用数据而不是文件（图 8.1-1）消除了人工搜索过程并统一了处理过程。已经采用这种方法的公司可以快速获取信息并将其快速整合到业务流程中，从而获得竞争优势。

> 从使用文件到使用数据的转变是不可避免的，它将塑造建筑业的未来。

建筑行业的每家公司都将面临一个关键的选择：继续将信息存储在不同的文件和孤岛中，必须由人工使用专门的程序进行读取；或者在处理的最初阶段将其转换为结构化数据，为自动化项目管理创建一个单一的集成数字基础。

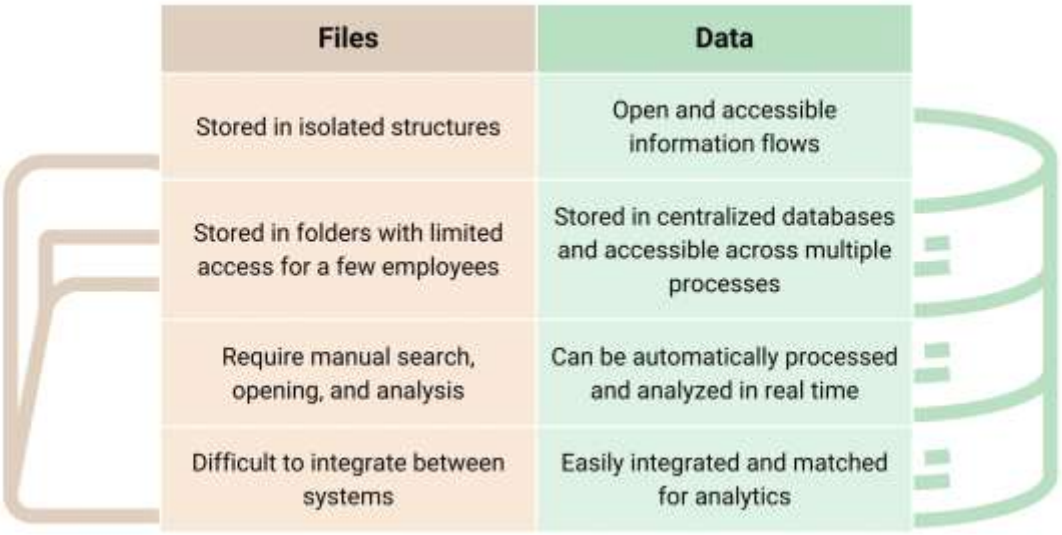


图 8.1-1 信息流的演变：从孤立文件到综合数据。

随着信息的爆炸式增长，传统的文件存储和处理方法的效率越来越低。在建筑行业，与其他行业一样，依靠不同文件格式的文件夹或互不关联的数据库已不再足够。

企业要想在数字时代保持竞争力，必然会转向综合数字平台，利用大数据技术和自动分析系统。

从基于文件的存储转向数据驱动的工作流程，需要重新思考信息管理方法，并有意识地选择适合进一步整合到集中存储库的格式。这种选择将决定数据处理的效率、数据访问的速度以及数据与公司数字流程集成的便捷程度。

大数据存储：分析流行格式及其有效性

存储格式对分析基础设施的可扩展性、可靠性和性能起着关键作用。为了分析和处理数据（如过滤、分组和聚合），我们的示例使用了 Pandas DataFrame，这是一种在 RAM 中处理数据的常用结构。

然而，Pandas DataFrame 并没有自己的存储格式，因此一旦处理完成，数据就会被导出为外部格式之一--最常见的是 CSV 或 XLSX。这些表格格式易于交换，并与大多数外部系统兼容，但有 一系列限制：存储效率低、缺乏压缩和版本支持差：

- **CSV**（逗号分隔值）：一种简单的文本格式，各种平台和工具都广泛支持。它易于使用，但不支持

复杂的数据类型和压缩。

- **XLSX**（Excel Open XML Spreadsheet）：Microsoft Excel 文件格式，支持公式、图表和样式等复杂功能。虽然它对手动数据分析和可视化很有用，但并不适合大规模数据处理。

除了常用的表格 XLSX 和 CSV 之外，还有几种有效存储结构化数据的常用格式（图 8.1-2），根据具体的数据存储和分析要求，每种格式都有独特的优势：

- **Apache Parquet**：一种专为数据分析系统优化使用的列式数据存储文件格式。它提供高效的数据压缩和编码方案，是复杂数据结构和大数据处理的理想选择。
- **Apache ORC**（Optimised Row Columnar）：与 Parquet 相似，ORC 可提供高压缩和高效的数据存储。它针对重读取操作进行了优化，非常适合存储数据湖。
- **JSON**（JavaScript Object Notation）：虽然与 Parquet 或 ORC 等二进制格式相比，JSON 的数据存储效率不高，但它非常容易访问和使用，因此非常适合对可读性和网络兼容性要求较高的脚本。
- **Feather**：一种快速、轻量、易用的面向分析的二进制列式数据存储格式。它设计用于在 Python (Pandas) 和 R 之间高效传输数据，因此是涉及这些编程环境的项目的绝佳选择。
- **HDF5**（层次数据格式第 5 版）：专为存储和组织大量数据而设计。它支持多种数据类型，非常适合处理复杂的数据集合。由于 HDF5 能够高效地存储和访问大型数据集，因此在科学计算领域特别受欢迎。



	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)

图 8.1-2 数据格式比较，显示存储和处理方面的主要差异。

为了对 ETL 流程加载阶段使用的格式进行比较分析，我们编制了一个显示文件大小 和读取时间的表格（图 8.1-3）。研究中使用了数据完全相同的文件：表格包含 10,000 行和 10 列随机值。

研究包括以下存储格式：CSV、Parquet、XLSX 和 HDF5，以及它们在 ZIP 档案中的压缩版本。原始数据使用 NumPy 库生成，并以 Pandas DataFrame 结构表示。测试过程包括以下步骤：

- 文件保存：数据帧以四种不同的格式保存：CSV 、Parquet 、XLSX 和 HDF5。每种格式在数据存储方式上都有其独特之处，会影响文件大小和读取速度。
- ZIP 文件压缩：为了分析标准压缩的效果，每个文件都被进一步压缩成 ZIP 档案。
- 文件读取（ETL - 加载）：从 ZIP 解压缩后，测量每个文件的读取时间。这样就能估算出从压缩包中提取后的数据访问速度。

值得注意的是，Pandas DataFrame 并不直接用于分析大小或读取时间，因为它本身并不代表一种存储格式。它只是作为 的中间结构，用于生成数据并随后将其保存为不同格式。



	Original Size (MB)	ZIP Size (MB)	Read Time (s)
CSV	18.82	8.73	0.48
Parquet	9.67	9.22	0.17
XLSX	12.59	12.59	41.73
HDF5	8.50	7.62	0.13

图 8.1-3 按大小和读取速度对存储格式进行比较。

CSV 和 HDF5 文件具有（图 8.1-3）很高的压缩效率，在 ZIP 中打包时可显著减小文件大小，这在需要优化存储的情况下特别有用。另一方面，XLSX 文件几乎不可压缩，其在 ZIP 中的大小仍与原始文件相当，因此不太适合用于大数据量或对数据访问速度要求较高的环境。此外，与其他格式相比，XLSX 的读取时间明显较长，因此不太适合快速数据读取操作。Apache Parquet 由于采用列式结构，在分析任务和大数据量方面表现出很高的性能。

使用 **Apache Parquet** 优化存储

用于存储和处理大数据的流行格式之一是 Apache Parquet。这种格式专为列式存储而设计（类似于 Pandas），可大幅减少内存占用，提高分析查询的速度。与 CSV 和 XLSX 等传统格式不同，Parquet 支持本机压缩，并针对 Spark、Hadoop 和云存储等大数据系统进行了优化。

> Parquet 的主要功能包括支持数据压缩和编码，通过直接对所需列而不是所有数据行进行处理，极大地减少了存储空间，加快了数据读取操作。

为了举例说明将数据转换为 Apache Parquet 所需的代码是多么容易，让我们使用 LLM。

- 向 LLM 聊天（CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN ）发送文本请求：

> 编写代码，将数据从 Pandas DataFrame 保存到 Apache Parquet。 ⏎

- LLM 的答复：

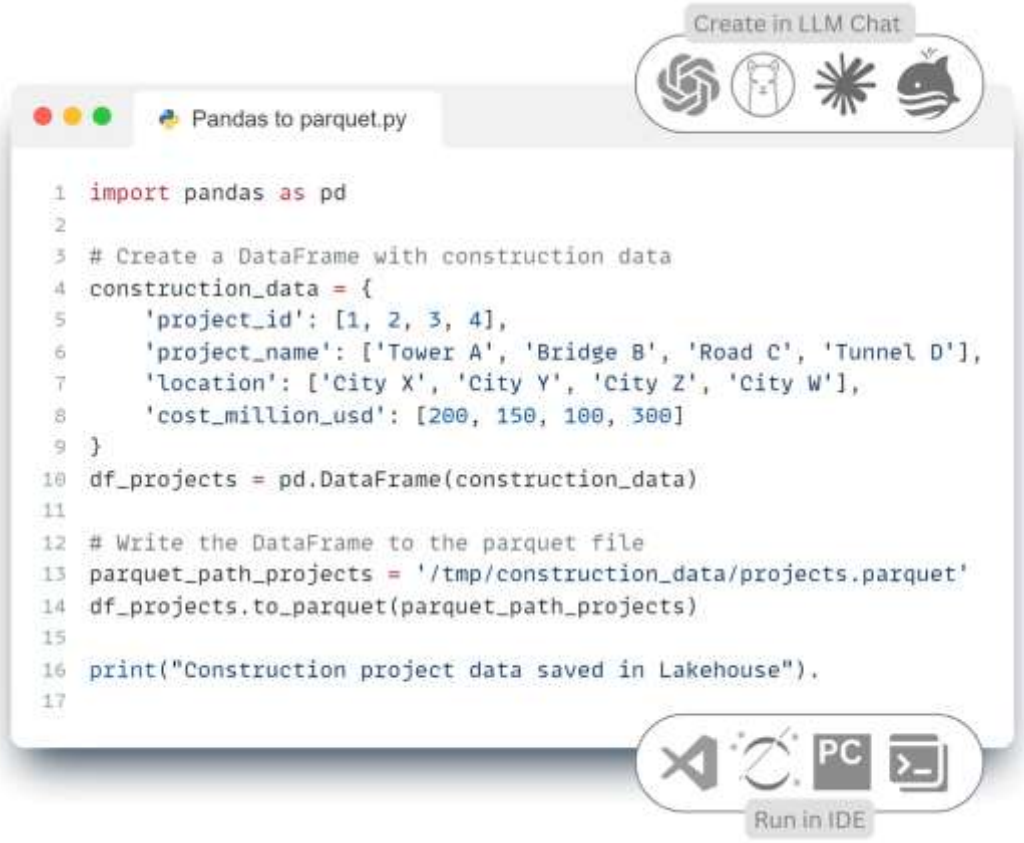


图 8.1-4 使用几行 Python 将数据帧数据从 RAM 传输到存储效率高的 Apache Parquet 格式 。

下一个示例：让我们模拟 ETL 流程 ， 使用以 Parquet 格式存储的数据，根据 "cost_million_usd "属性之一

的特定值过滤项目（图 8.1-4）。

聊天继续进行时，请向 LLM 发送文本请求：

> 编写代码，过滤表格中的数据，只保存 Apache Parquet 数据 中成本（参数 cost_million_usd）超过 1.5 亿美元的项目（表格行）。 ⏎

LLM 的答复：

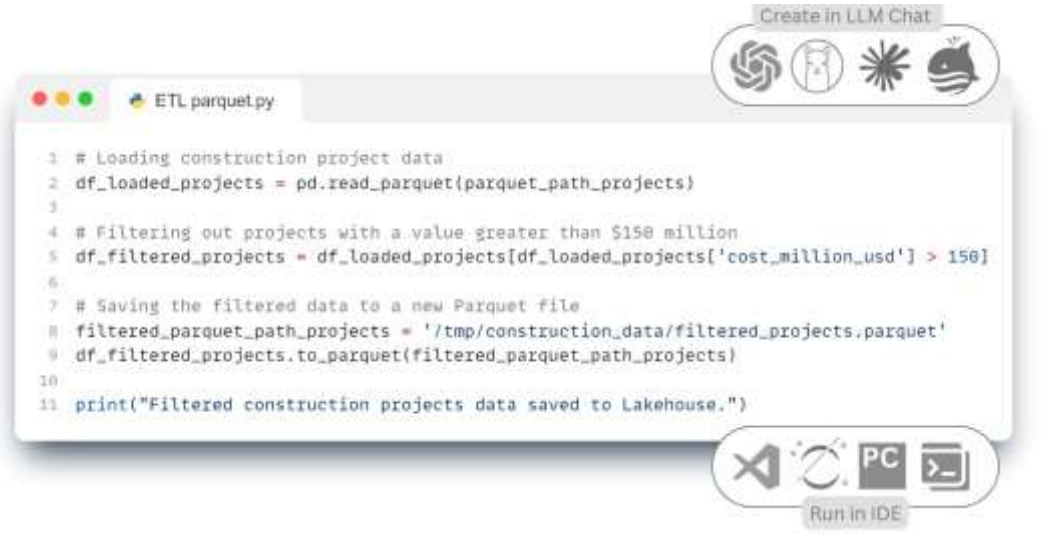


图 8.1-5 处理 Apache Parquet 格式数据时的 ETL 流程 与处理其他结构化格式的流程相同。

使用 Parquet 格式（相对于 XLSX、CSV 等格式）， 可显著减少存储的信息量并加快搜索操作。这使它成为存储和分析数据的绝佳选择。Parquet 可与各种处理系统集成，在混合架构中提供高效访问。

然而，高效的存储格式只是完整数据体验的一个要素。要创建一个可持续和可扩展的环境，需要一个精心设计的数据管理架构。DWH（数据仓库）类系统就能实现这一功能。这些系统可以汇总来自不同来源的数据，提高业务流程的透明度，并可使用商业智能工具和机器学习算法进行复杂的分析。

DWH: 数据仓库 数据仓库

正如 Parquet 格式是为高效存储大量信息而优化的一样，数据仓库 也是为整合和构建数据而优化的，以便

支持分析 、预测和管理决策。

在当今的企业中，数据来自许多不同的来源：ERP、CAFM、CPM、CRM 系统、会计和仓库管理、建筑物的数字 CAD 模型、物联网传感器和其他解决方案。要想获得全貌，仅仅收集数据是不够的，还需要对数据进行组织、标准化并集中到一个单一的存储库中。这正是 DWH 的作用所在--它是一个集中存储系统，可以汇总各种来源的信息，对其进行结构化处理，并将其用于分析和战略管理。

> **DWH** （数据仓库 ）是一个集中式数据仓库系统，可汇集来自多个来源的信息，构建信息结构，并将其提供给分析 和报告。

在许多公司中，数据分散在不同的系统中，这一点我们在本书的前半部分已经讨论过（图 1.2-4）。DWH 整合这些数据源，确保信息完全透明可靠。DWH 数据仓库是一个专门的数据库（大型数据库 ），用于收集、处理和存储来自多个来源的数据。DWH 的主要特点是

- 使用 **ETL --**流程（提取, 转换, 加载 ）--从数据源提取数据、清理数据、转换数据、将数据加载到存储库，并将这些流程自动化，本书第七部分对此进行了讨论。
- 数据粒度 - DWH 中的数据既可以汇总形式（汇总报告）存储，也可以粒度形式（原始数据）存储。从 2024 年起，CAD- 供应商开始谈论粒度数据[125]，这或许表明该行业正准备过渡到专门的云存储，以处理数字建筑模型数据。
- 支持分析 和预测 - 数据仓库 为 BI 工具、大数据 - 分析和机器学习 提供基础。

DWH 可作为商业智能 的基础，实现关键绩效指标分析、销售、采购和成本预测以及自动报告和数据可视化（图 8.1-6）。

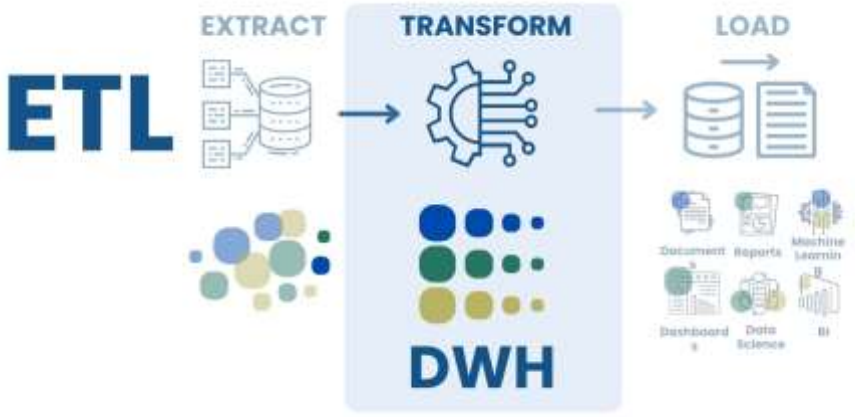


图 8.1-6 在 ETL 流程中，DWH 可充当中央存储库，从不同系统提取的数据在此进行转换和卸载。

DWH 在整合、清理和结构化信息方面发挥着关键作用，为商业智能 和决策过程奠定了坚实的基础。然而，在当今数据量快速增长、数据源日益多样化的环境中，传统的 DWH 信息存储方法往往需要以 ELT 和数据湖形式进行扩展。

数据湖 **-** 从 **ETL** 到 **ELT** 的演变：从传统清理到灵活处理

传统的 DWH - 数据仓库 ，旨在以分析查询的最佳格式存储结构化数据，但在处理非结构化数据和可扩展性方面受到限制。为应对这些挑战，数据湖**)** 应运而生，为大量异构数据提供了灵活的存储空间。

> 数据湖 提供了一种可供选择的 DWH -方法，允许您处理非结构化、半结构化和原始数据，而无需事先制定严格的模式。这种存储方法通常适用于实时数据处理、机器学习 和高级分析。数据湖与 DWH 不同，DWH 在加载前对数据进行结构化和聚合，而数据湖允许以原始形式存储信息，因此具有灵活性和可扩展性

正是对传统数据仓库（RDBMS、DWH）的失望和对 "大数据 "的兴趣导致了数据湖的出现，在数据湖中，数据不再是复杂的 ETL ，而是被简单地加载到一个结构松散的存储库中，并在分析阶段对其进行处理：

- 在传统数据仓库中，数据在加载到仓库之前通常要经过预处理、转换和清理（ETL - Extract, Transform, Load ）（图 8.1-6）。这意味着数据是有结构的，并针对未来特定的分析 和报告任务

进行了优化。重点是保持高查询性能和数据完整性。不过，这种方法成本较高，在整合新数据类型和快速变化的数据模式方面灵活性较差。

- 另一方面，数据湖旨在以原始格式存储大量原始数据（图 8.1-7）。ETL（提取 ， 转换，加载 ）、流程正被 ELT （提取 ，加载 ，转换 ）所取代，在这种流程中，数据首先 "按原样 "加载到仓库，然后才能根据需要进行转换和分析。这提供了更大的灵活性和存储异构数据的能力，包括文本、图像和日志等非结构化数据。

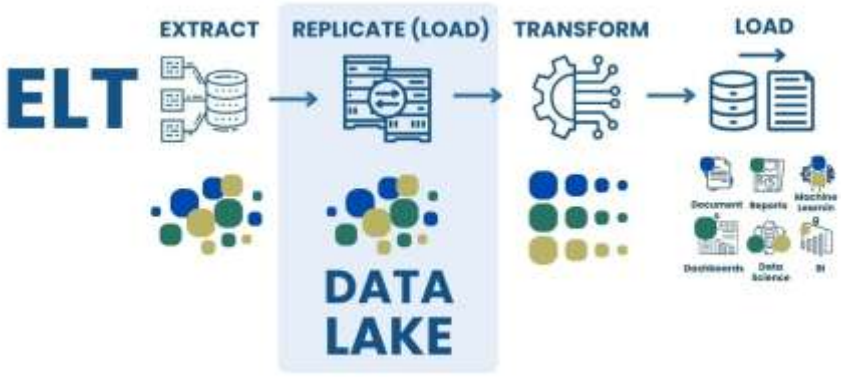


图 8.1-7 与 ETL 不同，数据湖 使用 ELT ，其中信息首先以 "原始 "形式上传，然后在上传阶段进行转换。

传统的数据仓库 侧重于预处理数据，以确保高查询性能，而数据湖则优先考虑灵活性：它们存储原始数据并根据需要进行转换（图 8.1-8）。

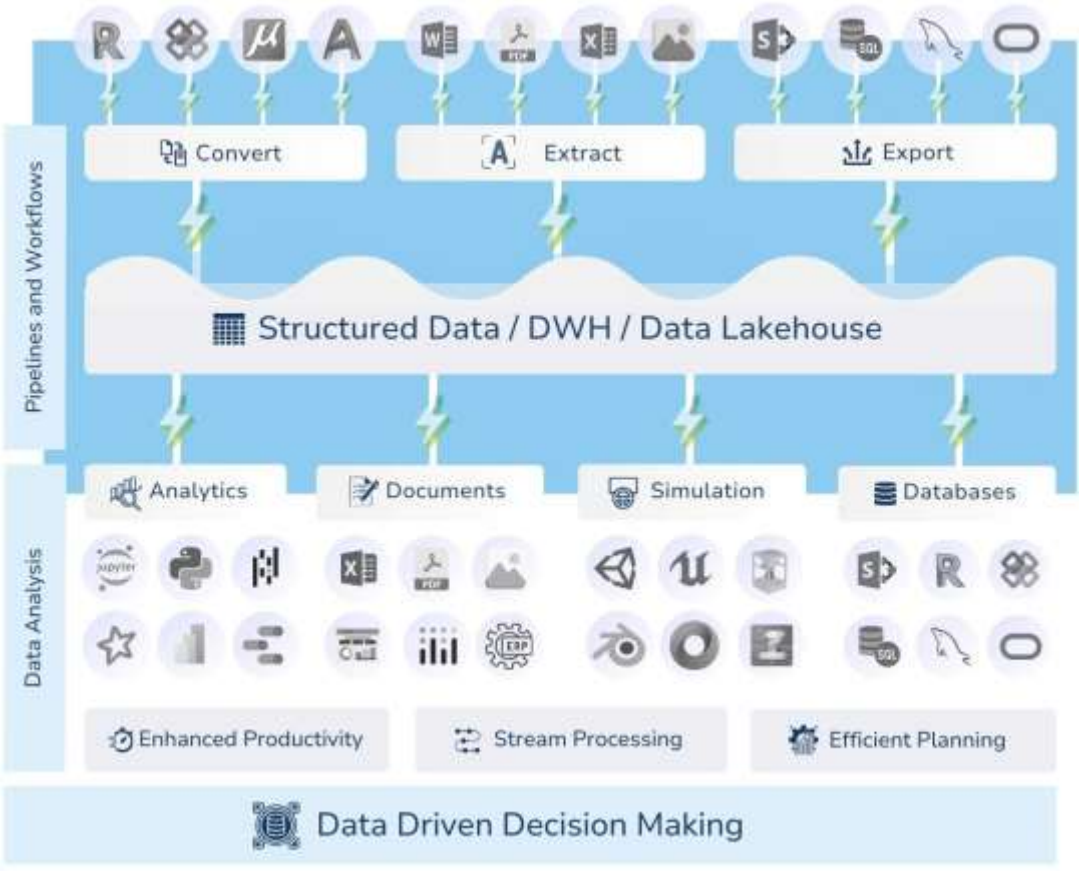


图 8.1-8 现代存储概念旨在为决策目的存储和处理所有类型的数据。

然而，尽管数据湖具有种种优势，但也并非没有缺点。由于缺乏严格的结构和信息管理的复杂性，可能会导致数据重复、相互矛盾或不相关的混乱局面。此外，在这样的存储库中搜索和分析数据需要花费大量精力，尤其是在处理异构信息时。为了克服这些限制，并结合传统数据仓库和数据湖的最佳功能，。

数据湖架构： 仓库和数据湖的协同作用

为了结合 DWH （结构化、可管理、高性能分析 ）和数据湖 （可扩展性、处理异构数据）的最佳特性，开发了数据湖方法。这种架构将数据湖的灵活性与传统仓库典型的强大处理和管理工具相结合，在存储、分析和机器学习之间取得了平衡。Data Lakehouse 是数据湖和数据仓库的综合体 ，将前者的灵活性和可扩展性与后者的可管理性和查询优化相结合。

> 数据湖 是一种架构方法，旨在将数据湖的灵活性和可扩展性与数据仓库的可管理性和查询性能结合起来（图 8.1-9）。

Data Lakehouse 的主要功能包括

- 开放式数据存储格式：使用开放式数据存储格式（如 Apache Parquet ）可提高效率并优化查询。
- 只读模式：与 DWH 中只写模式的传统方法不同，Lakehouse 支持只读模式，从而可以更灵活地管理数据结构。
- 灵活、可扩展：支持结构化和非结构化数据的存储和分析，通过存储级优化提供高查询性能。

Data Lakehouse 提供了一种折衷解决方案，结合了两种方法的优势，非常适合需要灵活处理数据的现代分析工作负载。

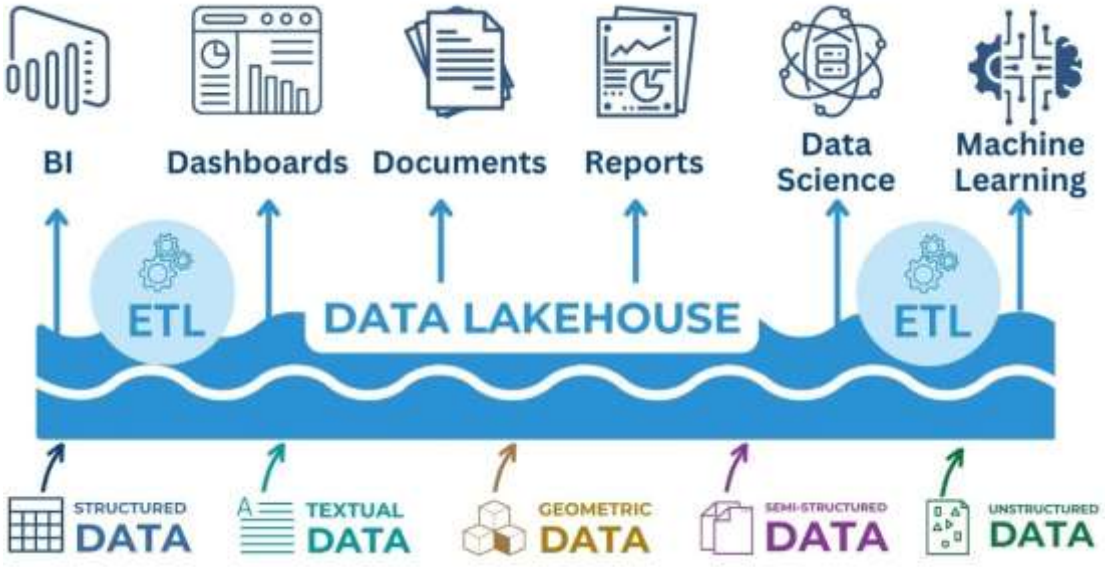


图 8.1-9 Data Lakehouse 是新一代存储系统，旨在满足复杂多变的需求。

现代数据仓库的理念似乎很简单：如果所有数据都在一个地方，就更容易分析。然而，在实践中，一切并不那么顺利。想象一下，一家公司决定完全放弃常用的会计和管理系统（ERP 、PMIS 、CAFM 或其他），代之以一个人人都能访问的巨大数据湖。会发生什么情况呢？很可能会出现混乱：数据重复、相互矛盾，关键信息丢失或损坏。即使数据湖仅用于分析 ，如果没有适当的管理，也会受到严重影响：

- 数据难以理解：在传统系统中，数据具有清晰的结构，但在湖泊中，数据只是大量文件和表格的集合。要找到某样东西，专家必须弄清每一行和每一列的作用。
- 数据可能不准确：如果同一信息的多个版本存储在一个地方，就很难知道哪个版本是最新的。因此，决策会基于过时或错误的数据做出。
- 为工作准备数据很困难：数据不仅要存储，还要以方便的形式呈现--以报告、图形、表格的形式。在传统系统中，这些都是自动完成的，但在数据湖中则需要额外的处理。

因此，每种数据仓库概念都有自己的特点、处理方法和业务应用。传统数据库侧重于事务性操作，数据仓库 (DWH) 为分析提供结构，数据湖 (Data Lake) 以原始形式存储信息，而混合仓库 (Data Lakehouse) 则结合了 DWH 和数据湖的优势（图 8.1-10）。

	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

图 8.1-10 DWH, Data Lake 和 Data Lakehouse：数据类型、使用场景、处理方法和存储方式的主要区别。

选择存储架构是一个复杂的过程，取决于业务需求、信息量和分析要求。每种解决方案都有其优缺点：DWH 提供结构性，数据湖 提供灵活性，而 Lakehouse 则在两者之间取得平衡。企业很少局限于单一的数据架构。

> 无论选择何种架构，自动化数据管理系统都明显优于人工方法。它们能最大限度地减少人为错误，加快信息处理速度，并确保业务流程各阶段数据的透明度和可追溯性。

虽然集中式数据仓库已成为许多经济领域的行业标准，但建筑业的情况仍然很分散。 这里的数据分布在不同的平台上（CDE、PMIS、ERP 等），因此很难创建一个统一的情况，这就要求架构能够将这些数据源整合到一个整体的、可用于分析的数字环境中。

CDE, PMIS, ERP 或 **DWH** 和数据湖

一些建筑和工程公司已经根据 ISO 19650 使用了通用数据环境（CDE ）的概念。从本质上讲，CDE 的功能与其他行业的数据仓库 (DWH) 相同：集中信息、提供版本控制、提供对有效信息的访问。

> 通用数据环境（**CDE** ）是一个集中的数字空间，用于在设施生命周期的各个阶段管理、存储、共享和协作项目信息。CDE 通常采用云技术，并与 CAD (BIM) 系统集成。

几十年来，金融、零售、物流和工业部门一直在使用集中式数据管理系统，将不同来源的信息整合在一起，控制其相关性并提供分析。CDE 将这些原则进一步加以改进，以应对建筑设计和生命周期管理方面的挑战。

与 DWH 一样，CDE 构建数据、捕捉变化并提供单点访问验证信息。随着向云计算的迁移以及与分析工具的集成，两者之间的差异正变得越来越不明显。在 CDE 的基础上，增加了细粒度数据（CAD - 供应商自 2023 年以来一直在讨论这一概念）[93, 125] ，我们可以看到与经典 DWH 的更多相似之处。

在前面的 "建筑企业资源规划 和 PMIS 系统 "一章中，我们已经了解了 PMIS（项目管理信息系统）和企业资源规划（ERP）。在建筑项目中，CDE 和 PMIS 协同工作：CDE 充当图纸、模型和项目文件等数据的存储库，而 PMIS 则管理截止日期、任务、资源和预算等流程。

负责管理整体业务（财务、采购、人事、生产）的 ERP 可与 PMIS 集成，在公司层面提供成本和预算控制。在分析 和报告方面，DWH 可用于收集、构建和汇总来自 CDE 、PMIS 和 ERP 的数据，以评估财务 KPI（投资回报率 ）并识别模式。反过来，数据湖 (DL) 可以通过存储原始和非结构化数据（如日志、传感器数据、图像）来补充 DWH。这些数据经过处理后，可加载到 DWH 中进行进一步分析。

> 因此，CDE 和 PMIS 侧重于项目管理，ERP 侧重于业务流程，DWH 和 Data Lake 侧重于分析和数据操作。

在将 CDE 、PMIS 和 ERP 系统与 DWH 和 Data Lake 进行比较时，可以发现 在供应商独立性、成本、集成灵活性、数据独立性、适应变化的速度和分析能力方面存在显著差异（图 8.1-11）。传统的系统，如 CDE 、PMIS 和 ERP 通常与特定的供应商解决方案和标准绑定，使其灵活性降低，并因许可证和支持而增加了成本。此外，这些系统中的数据通常以专有的封闭格式封装，限制了其使用和分析。

	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)

图 8.1-11 DWH 和数据湖 与 CDE 、PMIS 和 ERP 等系统相比，具有更大的灵活性和数据独立性。

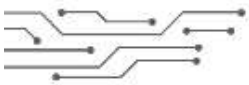
相比之下，DWH 和 Data Lake 在与不同数据源集成方面具有更大的灵活性，而且它们使用开放式技术和平台，有助于降低总体拥有成本。此外，DWH 和数据湖支持多种分析工具，从而增强了分析和管理能力。

> 随着 CAD 格式逆向工程工具和 CAD 应用数据库访问的发展，问题变得越来越尖锐：如果设计数据必须提供给在数十个承包商和设计机构工作的众多专家，那么继续使用封闭、孤立的平台有多大的合理性？

这种对特定供应商技术的依赖会极大地限制数据管理的灵活性，减缓对项目变化的响应速度，并阻碍参与者之间的有效协作。

传统的数据管理方法--包括 DWH 、Data Lake 、CDE 和 PMIS --主要侧重于信息的存储、结构化和处理。然而，随着人工智能 和机器学习 的发展，人们越来越需要新的数据组织方式，这种方式不仅能汇总数据，还能揭示复杂的关系、发现隐藏的模式并提供对最相关信息的即时访问。

矢量数据库 - 一种针对高维嵌入进行了优化的新型存储 - 正开始在这一方向发挥特殊作用。



章 8.2.

数据仓库管理和混乱预防

矢量数据库和边界框

矢量数据库是一种新型的存储库，它不仅可以存储数据，还可以根据意义进行搜索，根据语义的接近程度对对象进行比较，并创建智能系统：从推荐到自动分析和上下文生成。与注重精确匹配的传统数据库不同，矢量数据库根据属性查找相似对象--即使没有精确匹配

> 矢量数据库 是一种专门的数据库类型，它将数据存储为多维矢量，每个矢量代表一定的特征或质量。这些向量可以有不同的维数，具体取决于数据的复杂程度（在某种情况下可能只有几个维数，而在另一种情况下则可能有数千个维数）。

矢量数据库的主要优势是通过语义相关性而不是值的精确匹配进行搜索。与使用 "等于 "或 "包含 "过滤器的 SQL 和 Pandas 查询不同，我们使用的是在特征空间中搜索近邻（k-NN ）（我们将在本书下一部分详细讨论 k-NN）。

随着 LLM （大型语言模型）和生成模型的发展，与数据库的交互开始发生变化。现在，我们可以用自然语言查询数据，对文档进行语义搜索，自动提取关键术语，并在对象之间建立上下文关系，而这一切都不需要熟练掌握 SQL 或表格结构知识。这一点在 "LLM 及其在数据处理和业务流程中的作用 "一节中有更详细的讨论。

不过，我们必须认识到，LLM 不会自动构建信息结构并将其排列有序。模型只是在数据中漂浮，并根据查询的上下文找到最相关的数据。如果数据没有经过预先清理或转换，深度搜索就像试图在数字 "垃圾 "中寻找答案一样--它可能会起作用，但结果的质量会比较低。理想的情况是，数据可以结构化（例如，将文档翻译成 Markdown）并加载到矢量数据库中。这将大大提高输出结果的准确性和相关性。

> 最初，矢量数据库用于机器学习，但如今它们在机器学习之外的应用也越来越多，如搜索引擎、内容个性化和智能分析等。

边框 （边界平行四边形）是矢量法在构造中最明显的例子之一。它是一种描述三维空间中物体边界的几何结构。边界框由最小和最大 X、Y 和 Z 坐标定义，在物体周围形成一个 "框"。通过这种方法，您可以估算元素的大小和位置，而无需分析整个几何体。

每个边框 都可以表示为多维空间中的一个向量：例如 [x、y、z、宽、高、深] - 已经有 6 个维度（图 8.2-1）。



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

图 8.2-1 边框 - 元素坐标信息及其在项目模型中的位置类似于矢量数据库。

这种数据表示方式为许多任务提供了便利，包括检查对象之间的交叉、规划建筑元素的空间分布以及执行自动计算。边框 可以作为复杂三维模型和传统矢量数据库之间的桥梁，让您在建筑和工程建模中有效利用两种方法的优势

> 边框 是 "几何矢量化"，而嵌入 （一种转换抽象事物的方法）是 "意义矢量化"。无论是三维 - 项目模型中的物体还是文本中的概念，这两种方法都能让您从手动搜索转向智能搜索。

项目中的对象搜索（例如，"查找所有宽度大于 1.5 米的窗户"）类似于向量数据库中的近邻搜索（k-NN ），其中标准定义了特征空间中的一个 "区域"。(我们将在下一部分机器学习中详细讨论 k-NN 近邻搜索）（图 8.2-2）。如果我们在边界框属性中添加额外的参数（材料、重量、生产时间），表格就会变成一个高维

向量，其中每个属性都是一个新的维度。这更接近于现代向量库，其中的维数以百或千计（例如，神经网络的嵌入）。

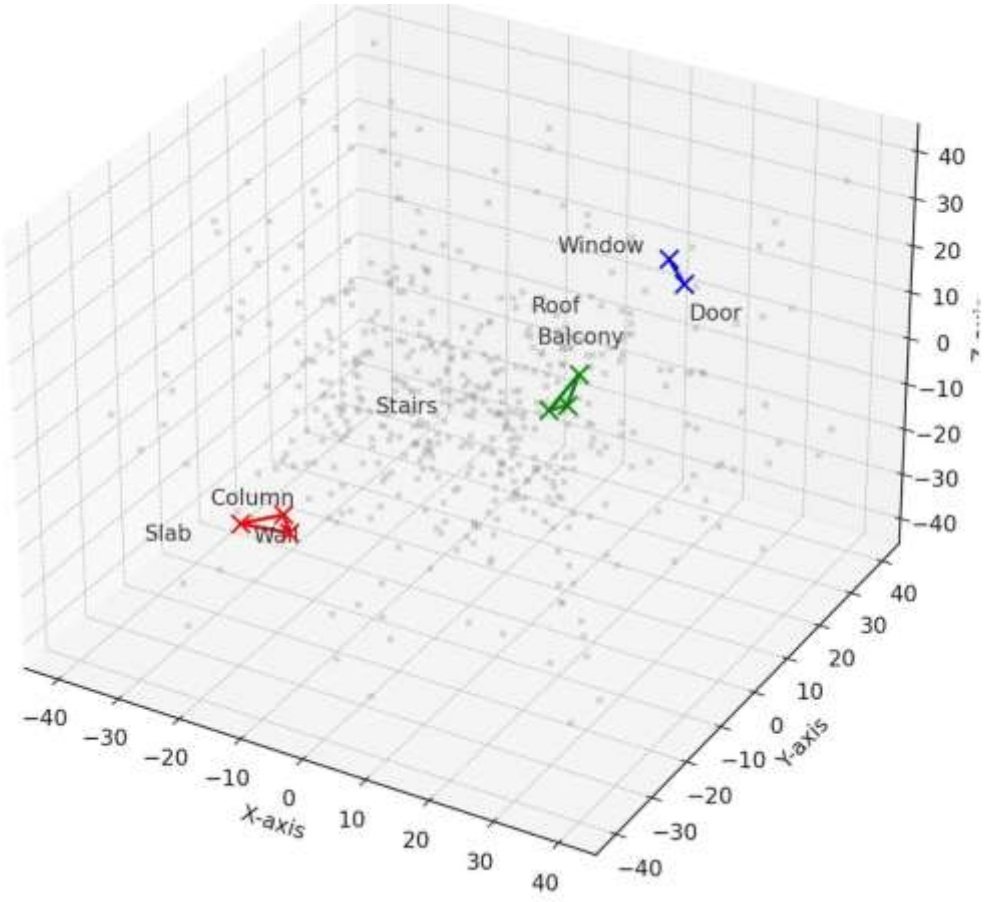


图 8.2-2 使用矢量数据库搜索项目中的对象。

边界框 中使用的方法不仅适用于几何对象，也适用于文本和语言分析。数据的矢量表示法已被广泛应用于自然语言处理（NLP ）。正如建筑项目中的物体可以根据空间距离的远近进行分组一样（图 8.2-2），文本中的词语也可以根据语义和上下文的远近进行分析。

例如，"建筑师"、"建筑"、"设计 "等词在矢量空间中相邻，因为它们具有相似的含义。在 LLM 中，这种机制可以实现自动分类，无需人工：

- 确定文本主题
- 对文件内容进行语义搜索

- 生成自动注释和文本摘要
- 查找同义词和相关术语

矢量数据库可以让您分析文本并找到其中的相关术语，就像边界框 可以帮助您分析三维 - 模型中的空间对象一样。项目元素的边界框示例有助于理解，矢量表示法并非来自 ML 的纯粹 "人工 "概念，而是一种自然的数据结构方式，可用于解决应用问题，无论是搜索 CAD 项目中的列，还是数据库中语义相近的图像。

> 数据库专家应关注向量存储。它们的扩散表明数据库发展进入了一个新阶段，传统的关系系统和面向人工智能 的技术开始交织在一起，形成未来的混合解决方案。

开发复杂和大规模人工智能 应用程序的用户将使用专业数据库进行矢量搜索。与此同时，那些只需要将单独的人工智能功能集成到现有应用程序中的用户则更倾向于选择他们已经使用的数据库（PostgreSQL 、Redis）中的内置矢量搜索功能。

尽管 DWH、数据湖、CDE、PMIS、矢量数据库等系统提供了不同的数据存储和管理方法，但它们的有效性不仅取决于其架构，还取决于数据本身的组织和管理情况。即使使用现代解决方案（无论是矢量数据库、传统的关系数据库管理系统还是数据湖类型的仓库），如果缺乏管理、构建和更新数据的明确规则，也会导致用户在处理不同文件和多格式数据时遇到同样的困难。

如果没有数据治理)，即使是最强大的解决方案也会变得混乱和无序，使数据湖变成数据沼泽)。为避免这种情况，企业不仅要选择正确的存储架构，还要实施数据最小化)、访问管理和质量控制策略，将数据转化为有效的决策工具。

数据治理)、数据最小化) 和数据沼泽)

理解和实施数据治理)、数据最小化)和防止数据沼泽)等概念是成功管理数据仓库和实现业务价值的关键（图 8.2-3）。

> 根据 Gartner（2017 年）的一项研究，85% 的大数据项目都以失败告终，其中一个重要原因就是数据质量和数据治理不足[144]。

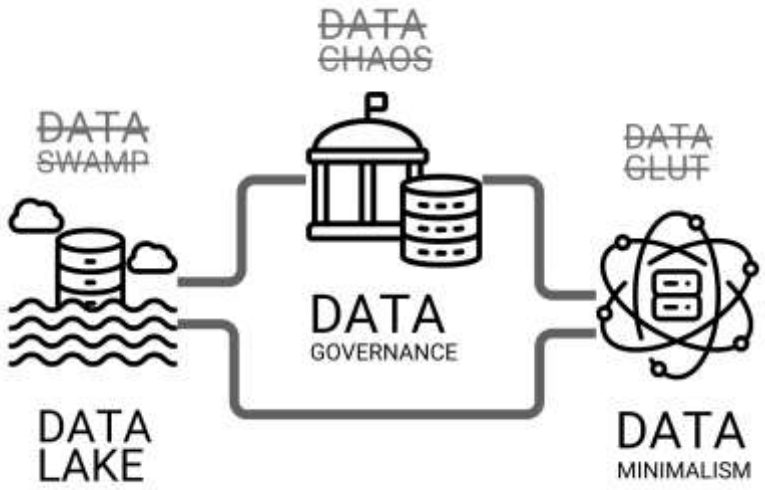


图 8.2-3 数据治理的一些关键方面是数据治理 和数据最小化。

数据治理（Data Governance ）是数据管理的基本组成部分，可确保数据在所有业务流程中得到合理有效的使用。它不仅涉及制定规则和程序，还涉及确保数据的可用性、可靠性和安全性：

- 对 数据进行定义和分类：对实体进行明确的定义和分类，可使组织了解公司需要哪些实体，并确定如何使用这些实体。
- 访问权限和管理：制定访问和管理数据的政策和程序，确保只有经授权的用户才能 访问某些数据。
- 保护数据免受外部威胁：保护数据免受外部威胁是数据管理的一个重要方面。这不仅包括技术措施，还包括对员工进行信息安全基础知识培训。

数据最小化（Data Minimalism ）是一种将数据减少到形成中最有价值和意义的属性和实体的方法（图 8.2-4），从而降低成本并提高数据利用率：

- 简化决策：通过减少分析和处理数据所需的时间和资源，将对象及其属性的数量减少到最相关的属性，从而简化决策。
- 关注重要信息：选择最相关的实体和属性可让您关注对业务真正重要的信息，消除噪音和不必要的数据。
- 高效的资源分配：数据最小化可实现更高效的资源分配，降低数据存储和处理成本，提高数据质量和安全性。

处理数据的逻辑不应该从创建数据开始（图 8.2-4），而应该在数据生成过程开始之前就了解未来使用这些数据的情景。这种方法可以提前定义对属性、属性类型和边界值的最低必要要求。这些要求构成了在信息模型中创建正确和稳定实体的基础。对数据目的和用途的初步了解有助于形成适合分析的结构。数据建模：概念、逻辑和物理模型 "一章讨论了概念、逻辑和物理层面数据建模方法的更多细节。

> 在建筑公司的传统业务流程中，数据处理更像是将数据倒入沼泽，先创建数据，然后由专家尝试将其整合到其他系统和工具中。

数据沼泽（Data Swamp ）是指在没有适当组织、结构化和管理的情况下，无节制地收集和存储数据，导致数据无序、难以使用且价值不高。

如何防止信息流通陷入泥潭：

- 数据结构管理：确保 数据的结构化和分类，使数据井然有序、易于访问，从而防止数据沼泽化。
- 理解和解释数据：明确说明数据的来源、修改和含义，确保正确理解和解释数据。
- 维护数据质量 ： 定期维护和清理数据有助于维护数据质量、相关性以及对分析和业务流程的价值。

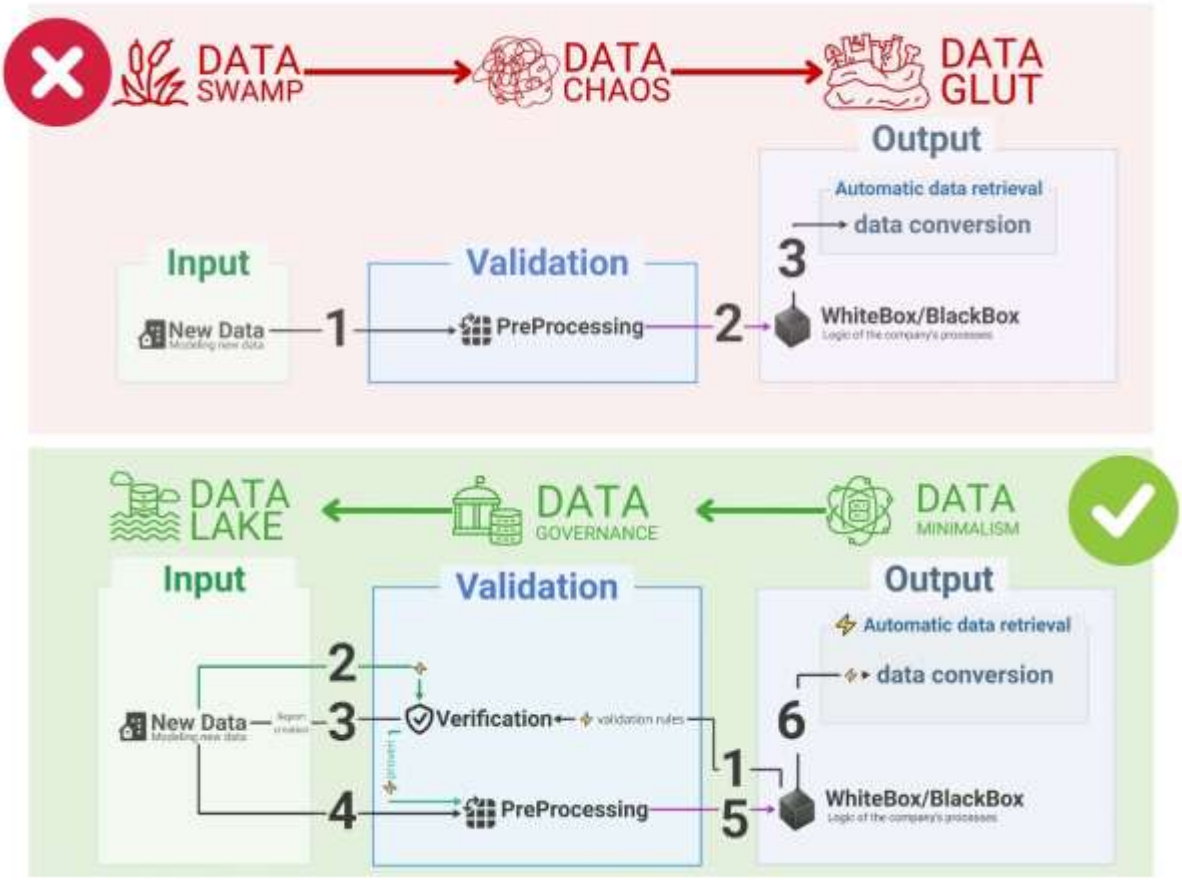


图 8.2-4 为避免数据仓库杂乱无章，应从收集属性需求开始数据创建过程。

> 通过将数据治理和数据最小化原则纳入数据管理流程，并积极防止数据仓库 成为数据沼泽，企业可以最大限度地发挥数据的潜力。

在解决了管理和最小化问题之后，数据工作发展的下一个阶段是自动处理的标准化、质量保证以及使数据可用于分析、转换和决策的方法的实施。这就是 DataOps 和 VectorOps 方法论正在做的事情，它们正在成为处理大数据和机器学习的公司的重要工具。

DataOps 和 **VectorOps**： 新数据标准

数据治理 负责控制和组织数据，而数据运营 则帮助确保数据的准确性、一致性和在公司内部的顺畅流动。这对于建筑业的许多业务案例尤为重要，因为在这些案例中，数据不断产生，需要及时处理。例如，在建筑

信息模型、项目 要求和分析报告需要在一个工作日内在不同系统之间同步的情况下，数据运营的作用就非常关键。它使您能够建立稳定、可重复的数据处理流程，降低延迟和信息相关性丧失的风险。

仅有数据管理是不够的--至关重要的是，数据不仅要存储，还要在日常运营中积极使用。这就是 DataOps（一种侧重于自动化、集成和持续数据流的方法）发挥作用的地方。

> **DataOps** 专注于改善组织内数据流的协作、集成和自动化。采用 DataOps 实践可提高数据的准确性、一致性和可用性，这对以数据为中心的应用至关重要。

DataOps 生态系统中的关键工具是用于工作流协调的 Apache Airflow （图 7.4-4）和用于路由和转换数据流的 Apache NiFi （图 7.4-5）。这些技术共同实现了灵活、可靠和可扩展的数据管道，以自动处理、控制和集成系统间的信息（更多详情，请参阅 "自动 ETL - 输送机 "一章）。在建筑流程中实施 DataOps 方法时，必须考虑四个基本方面：

1. 人和工具比数据更重要：孤立的数据存储库 可能被视为一个主要问题，但实际情况更为复杂。除了数据分散之外，团队的孤立和他们使用的不同工具也起着重要作用。在建筑业中，来自不同学科的专家都要处理数据：数据工程师和分析师 、商业智能和可视化团队，以及项目管理和质量专家。他们每个人都有不同的工作方式，因此创建一个生态系统，让数据在参与者之间自由流动，提供单一、一致的信息版本，就变得非常重要。
2. 自动测试和错误检测：建筑数据总是包含错误，无论是模型不准确、计算错误还是规格过时。定期测试数据并消除重复出现的错误可以显著提高数据质量。作为 DataOps 的一部分，您需要实施自动控制和验证机制，以监控数据的正确性、分析错误和识别模式，并捕捉和解决每个工作流程中的系统故障。自动化验证程度越高，整体数据质量就越高，最后阶段出现错误的可能性就越低。
3. 测试数据的方式应与测试程序代码的方式相同：大多数楼宇应用都基于数据处理，但对数据的控制往往由次要角色负责。如果机器学习模型 在不准确的数据上进行训练，就会导致错误的预测和经济损失。在数据运营部门，数据应像软件代码一样接受严格审查：逻辑检查、压力测试、输入值变化时模型行为评估。只有经过验证的可靠数据才能作为管理决策的依据。
4. 不影响性能的数据可观察性：数据监控不仅仅是 一系列指标，而是一种战略性的质量管理工具。要使 DataOps 有效发挥作用，必须在数据处理的各个阶段（从设计到运行）都内置可观察性。与此同时，重要的是，监控工作不能拖慢系统的运行速度。在建筑项目中，至关重要的是不仅要收集数

据，还要保证创建数据的专业人员（如设计师）的工作不受任何干扰。这种平衡既能控制数据质量，又不会影响工作效率。

DataOps 不是数据科学家的额外负担，而是他们工作的支柱。通过实施 DataOps，建筑公司可以从混乱的数据管理转变为高效的生态系统，让数据为业务服务。

反过来，VectorOps 代表了 DataOps 演进的下一阶段，重点是处理、存储和分析多维矢量数据（上一章已讨论过）。这与数字双胞胎、神经网络模型和语义搜索等领域尤为相关，这些领域正开始进入建筑行业。VectorOps 依靠矢量数据库 高效地存储、索引和搜索对象的多维表示。

> **VectorOps** 是继 DataOps 之后的又一发展方向，侧重于在构建过程中处理、分析和使用矢量数据。DataOps 侧重于数据流、一致性和质量，而 VectorOps 则不同，它侧重于管理机器学习所需的多维对象表示。

与传统方法不同，VectorOps 可让您实现更精确的对象描述，这对于数字孪生、生成设计系统和转换为矢量格式的 CAD 数据的自动错误检测至关重要。DataOps 和 VectorOps 的组合实施为可扩展、自动处理大量信息（从传统表格到语义丰富的空间模型）奠定了坚实的基础

下一步：从混乱存储到结构化存储

传统的数据存储方法往往会造成 "信息孤岛 "的形成，无法获得重要的分析和决策见解。现代存储概念，如数据仓库 、数据湖 以及它们的混合体，可以统一不同的信息，并以集中的方式提供给数据流和商业智能。重要的是，不仅要选择正确的存储架构，还要实施数据治理) 和数据最小化) ，以防止存储设施成为无法控制的数据沼泽)。

在总结这部分内容时，值得强调的是主要的实践步骤，这些步骤将帮助您把讨论过的概念应用到日常工作中：

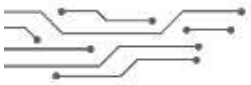
- 选择高效的数据存储格式
  - [ ] 从 CSV 和 XLSX 转向更高效的格式（Apache Parquet ，ORC），以存储大量数据
  - [ ] 实施数据版本系统以跟踪变更

- [ ] 使用元数据描述信息的结构和来源

- 创建统一的公司数据架构
  - [ ] 比较不同的存储架构：RDBMS、DWH 和数据湖。选择最能满足您的可扩展性、源集成和分析处理需求的架构
  - [ ] 为您的任务设计一个流程图，用于从各种来源提取、加载和转换数据（ETL）。使用可视化工具（如 Miro、Lucidchart 或 Draw.io）将关键步骤和集成点可视化
- 实施数据治理实践 和数据最小化
  - [ ] 遵循数据极简主义方法 - 只存储和处理真正有价值的内容
  - [ ] 实施数据管理原则 - 界定数据责任，确保质量和透明度
  - [ ] 了解有关数据管理政策和 DataOps 概念的更多信息, VectorOps
  - [ ] 在 DataOps 中定义数据质量标准 和数据验证程序

井然有序的数据存储为公司分析流程的集中化奠定了基础。从杂乱无章的文件积累过渡到结构化存储，可以将信息转化为战略资产，有助于做出明智的决策，提高业务流程的效率。

一旦数据收集、转换、分析和结构化存储过程实现自动化和标准化，数字化转型的下一个阶段就是全面处理大数据。





IX 部分

大数据、机器学习和预测

第九部分侧重于建筑行业的大数据、机器学习和预测分析。它探讨了从直观决策到基于历史数据的客观分析的转变。通过实际案例展示了建筑业的大数据分析--从解析旧金山建筑许可数据集到处理 CAD --拥有数百万元素的项目。特别关注用于预测建筑项目成本和进度的机器学习方法 ，并详细讨论了线性回归和 k 近邻算法。该部分展示了结构化数据如何成为预测模型的基础，以评估风险、优化资源和提高项目管理效率。该部分还提供了选择代表性数据样本的建议，并解释了为什么有效分析并不一定需要大型数据集。

章 9.1.
大数据及其分析

建筑业大数据：从直觉到可预测性

大数据 "一词并没有严格的定义。这一概念最初是在信息量开始超出传统处理方法的能力时出现的。如今，包括建筑业在内的许多行业的数据量和复杂性都已大大增加，以至于无法容纳在计算机的本地内存中，需要使用新技术进行处理。

大数据工作的精髓不仅在于存储和处理，还在于预测能力。在建筑行业，大数据开辟了从基于对表格和可视化效果的主观解读（如前所述）的直观决策到基于实际观察和统计数据的明智预测的道路。

与人们普遍认为的相反，大数据工作的目标不是 "让机器像人一样思考"，而是应用数学模型和算法分析海量数据，以识别模式、预测事件和优化流程。

> 大数据并不是一个冰冷的、不受人类影响的算法世界。相反，大数据与我们的本能、错误和创造力共同发挥作用。正是人类思维的不完美，让我们能够找到非标准的解决方案并取得突破。

随着数字技术的发展，建筑行业开始积极利用 IT 行业的数据处理技术。借助 Pandas 和 Apache Parquet 等工具，可以将结构化数据和非结构化数据结合起来，简化了信息的获取，减少了分析的损失，而来自文件或 CAD 项目的大型数据集（图 9.2-10 - 图 9.2-12），可以在项目生命周期的各个阶段收集、分析和预测数据。

大数据正在对建筑业产生变革性影响，并可能以各种方式对其产生影响。大数据技术的应用正在一些关键领域取得成果，例如包括以下方面：

- 投资潜力分析--根据以往设施的数据预测项目的盈利能力和投资回收期。
- 预测性维护--在设备故障真正发生之前就识别出可能发生的故障，从而减少停机时间。
- 供应链优化--预测中断并提高物流效率。
- 能源效率分析--协助设计低能耗建筑。

- 安全监测--使用传感器和可穿戴设备监测现场情况。
- 质量控制 - 实时监控流程标准的合规性。
- 人力管理--绩效分析和人员需求预测。

在建筑领域，很难找到一个不需要数据分析 和预测的领域。预测算法的主要优势在于它们能够自我学习，并随着数据的积累而不断改进。

> 在不久的将来，人工智能将不仅仅为建筑商提供帮助，还将做出从设计流程到建筑运营问题的关键决策。

本书的下一部分 "机器学习与预测 "将进一步讨论如何生成预测和使用学习模型。

向全面的大数据工作过渡需要改变分析方法本身。迄今为止，我们所考虑的经典系统侧重于因果关系，而大数据分析正在将重点转向寻找统计模式和相关性，从而使我们能够识别隐藏的关系，并在不完全了解所有因素的情况下预测对象的行为。

质疑大数据的可行性：相关性、统计和数据抽样

传统上，建筑是基于主观假设和个人经验。工程师以一定的概率假设材料的性能、结构能承受的荷载以及项目能持续多久。这些假设在实践中得到验证，但往往要付出时间、资源和未来风险的代价。

> 随着大数据时代的到来，这种方法正在发生巨大的变化：决策不再是凭直觉，而是分析大规模数据集的结果。建筑逐渐不再是一门凭直觉的艺术，而是一门精确预测的科学。

在向使用大数据的理念过渡时，不可避免地会提出一个重要问题：数据量到底有多重要，可靠的预测分析到底需要多少信息？人们普遍认为 "数据越多，准确性越高"，但在实践中并不总能证明这一点。

> 早在 1934 年，统计学家耶日-诺依曼（Jerzy Neumann）就证明 [145]，统计推断准确性的关键不在于数据量，而在于数据的代表性和抽样的随机性。

这一点在建筑行业尤为明显，因为该行业使用物联网 - 传感器、扫描仪、监控摄像头、无人机甚至多格式

CAD - 模型收集大量数据，增加了盲点、异常值和数据失真的风险。

让我们以路面状况监测为例。一个包含所有路段的完整数据集可能需要 X GB，处理时间约为一天。与此同时，如果随机取样，只包括每 50 个路段，则只需 X/50 GB，半小时即可处理完毕，而某些计算的估计精度却相差无几（图 9.1-1）。

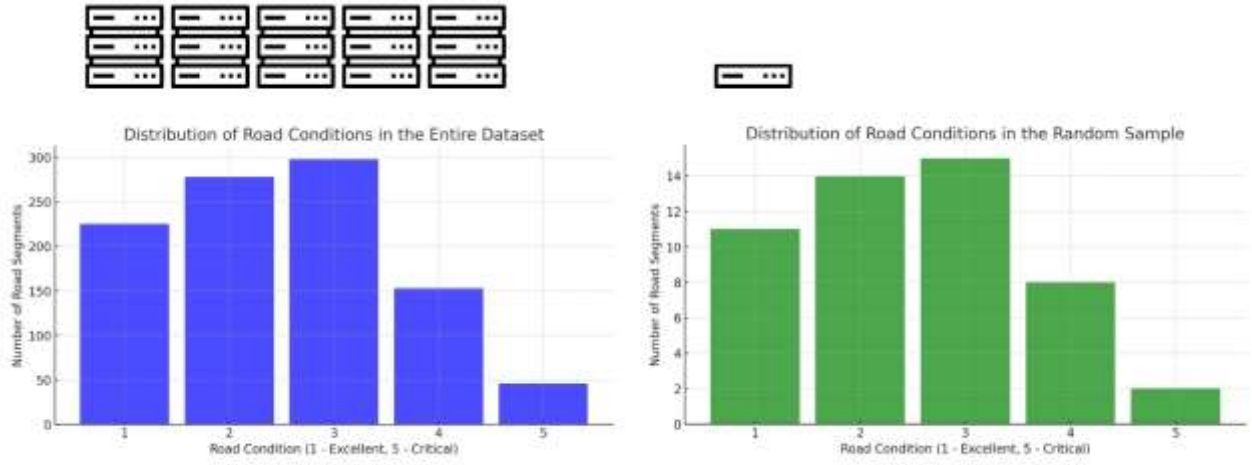


图 9.1-1 路面状况柱状图：全部数据集和随机取样显示的结果相同。

因此，数据分析成功与否的关键往往不在于数据量的多少，而在于样本的代表性和所用处理方法的质量。要改用随机抽样和更具选择性的方法，建筑行业就必须转变思维。从历史上看，公司一直遵循 "数据越多越好"的逻辑，认为涵盖所有可能的指标将最大限度地提高准确性。

这种做法让人想起项目管理中的一个流行误区："我吸引的专家越多，工作就越有效"。然而，与人力资源一样，质量和工具比数量更重要。如果不考虑数据或项目参与者之间的相互关系（相关性），增加数量只会导致噪音、失真、重复和不必要的浪费。

最后的结果往往是，拥有一个规模较小、但在质量上准备充分的数据集，能够做出稳定合理的预测，要比依赖包含许多相互矛盾信号的海量但混乱的信息更有成效。

> 数据量过大不仅不能保证更高的准确性，还可能因为存在噪声、冗余特征、隐藏的相关性和无关信息而导致结论失真。在这种情况下，过度拟合模型的风险会增加，分析结果的可靠性也会降低。

在建筑领域，处理大数据的一个主要挑战是确定数据的最佳数量和质量。例如，在监测混凝土结构状况时，使用成千上万个传感器并每分钟收集信息，会使存储和分析系统不堪重负。但是，如果进行相关性分析并选择 10% 信息量最大的传感器，就可以获得几乎相同的预测精度，而所花费的资源却要少很多倍，有时甚至是几十倍甚至几百倍。

使用较小的数据子集可以减少所需的存储量和处理时间，从而显著降低存储和分析数据的成本，这通常使随机抽样成为预测分析的理想解决方案，尤其是在大型基础设施项目或实时工作中。归根结底，施工过程的效率并不是由收集的数据量决定的，而是由数据分析的质量决定的。如果没有批判性的方法和仔细的分析，数据可能会导致错误的结论。

> 数据量达到一定程度后，每个新的信息单位产生的有用结果就会越来越少。与其无休止地收集信息，不如关注信息的代表性和分析方法（图 9.2-2）。

Allen Wallis [146]对这一现象进行了很好的描述，他以测试美国海军两种可供选择的弹丸设计为例，说明了统计方法的使用。

> 海军通过进行一系列配对发射，测试了两种可供选择的弹丸设计（A 和 B）。在每一轮试验中，A 的性能比 B 好或差，A 得 1 或 0，反之亦然。标准的统计方法包括进行固定数量的试验（如 1000 次），并根据百分比分布来确定胜负（例如，如果 A 得到 1 的次数超过 53%，则认为它是最好的）。当艾伦-沃利斯与（海军）加勒特-L-斯凯勒（Garrett L. Schuyler）上尉讨论这个问题时，上尉反对说，引用艾伦的故事，这样的测试可能毫无用处。如果当时有一位像斯凯勒这样睿智而经验丰富的军械军官在现场，他一定会在最初的几百次[射击]后发现，试验不需要终止，因为新方法明显不如预期，或者因为它明显优于预期[146]。
>
> - 哥伦比亚大学美国政府统计研究小组，二战时期

这一原理被广泛应用于各行各业。例如，在医学领域，新药的临床试验是在随机抽样的病人身上进行的，这样就可以获得具有统计学意义的结果，而无需在地球上生活的所有人身上进行试验。在经济学和社会学领域，通过代表性调查来反映社会意见，而无需对全国所有人进行访谈。

正如政府和研究机构通过调查小规模人群来了解社会总体趋势一样，建筑行业的公司也可以利用随机数据样本来有效监测和创建项目管理预测（图 9.1-1）。

> 大数据可能会改变社会科学的方法，但不会取代统计常识 [147]。
>
> - Thomas Landsall-Welfair, "Predicting the nation's current mood," Significance v.9(4), 2012 r.

从节约资源的角度来看，在为未来预测和决策收集数据时，必须回答这样一个问题：如果可以使用规模更小、成本更低且可以逐步扩展的测试数据集，那么花费大量资源收集和处理庞大的数据集是否有意义？随机抽样的有效性表明，企业可以通过选择不要求全面覆盖但仍能提供足够准确性和代表性的数据收集方法，将收集和训练模型的成本降低数倍甚至数千倍。这种方法甚至可以让小公司利用更少的资源和数据量取得与大公司同等的结果，这对于希望利用少量资源优化成本和加快知情决策的公司来说非常重要。在接下来的章节中，我们将探讨基于公共数据集、使用大数据工具进行 分析和预测分析的实例。

大数据：分析旧金山百万建筑许可数据集的数据

使用开放数据集为实践前几章讨论的原则提供了独特的机会：明智的特征选择、代表性取样、可视化和批判性分析。在本章中，我们将探讨如何利用开放数据（特别是旧金山的 100 多万条建筑许可记录）来研究复杂的现象，如大城市的建筑活动。

来自 "旧金山建筑部 "的超过 100 万份建筑��可的公开数据（图 9.1-2）（记录为 CSV 格式��两个数据集）[148]，使我们���仅可以使用原始 CSV 表分析该市的建筑活动，还可以批判性地分析旧金山建筑业在过去 40 年中的最新趋势和历史。[148]使我们能够使用原始 CSV 表不仅分析该市的建筑活动，，还能批判性地分析旧金山建筑业在过去 40 年（从 1980 年到 2019 年）中的最新趋势和历史。

用于创建数据集可视化的代码示例（图 9.1-3- 图 9.1-8），以及带有代码、解释和注释的可视化图表，可在 Kaggle 平台 上搜索 "旧金山 "找到。建筑业 1980-2019 年"。[149].



count 1.137695e+06

Building Permits on or after January 1, 2013

June 13, 2020 2,237 Views

Building Permits before January 1, 2013

June 13, 2020 860 Views

permit_creation_date	description	current_status	current_status_date	filed_date	issued_date	completed_date
07/01/1998	repair stucco	complete	07/07/1998	07/01/1998	07/01/1998	07/07/1998
12/13/2004	reroofing	expired	01/24/2006	12/13/2004	12/13/2004	NaN
02/18/1992	install auto fire spks.	complete	06/29/1992	02/18/1992	03/18/1992	06/29/1992

	permit_number	permit_expiration_date	estimated_cost	revised_cost	existing_use	Zipcode	Location
362780	9812394	11/01/1998	780.0	NaN	1 family dwelling	94123.0	(37.7963468760498, -122.432264144357 4)
570817	200412131233	06/13/2005	9000.0	9000.0	apartments	94127.0	(37.7292585160083 86, -122.464424566 57462)
198411	9202396	09/18/1992	9000.0	NaN	apartments	94111.0	(37.7950600255297 4, -122.395932244 61805)

图 9.1-2 数据集包含不同对象属性的已签发建筑许可证信息。

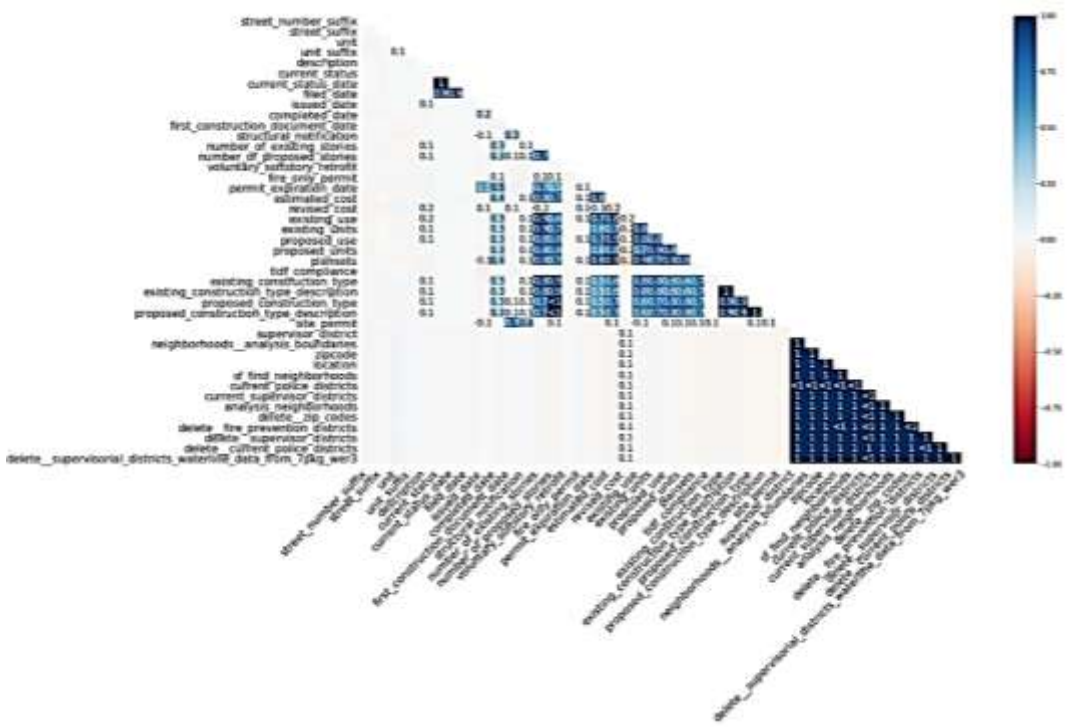


图 9.1-3 热图（Pandas 和 Seaborn ）可将数据集的所有属性可视化，并有助于识别属性对之间的关系。

从旧金山建筑部提供的表格（图 9.1-2）中看不出任何趋势或结论。表格形式的枯燥数字不能作为决策依据。为了使数据在视觉上易于理解，正如有关数据可视化的章节所详细讨论的那样，应使用本书第七部分 "ETL 和将结果可视化为图表 "中讨论的各种库将数据可视化。

通过使用 Pandas DataFrame 和 Python 可视化库 对 1,137,695 份许可证的价值进行数据分析[148]，我们可以得出结论：旧金山的建筑活动与经济周期密切相关，尤其是在蓬勃发展的硅谷科技产业（图 9.1-4）。

经济的繁荣和萧条对建筑项目的数量和价值有重大影响。例如，建筑活动的第一个高峰与 20 世纪 80 年代中期的电子产品热潮相吻合（使用 Pandas 和 Matplotlib ），随后的高峰和衰退则与互联网泡沫和近年来的技术热潮有关。

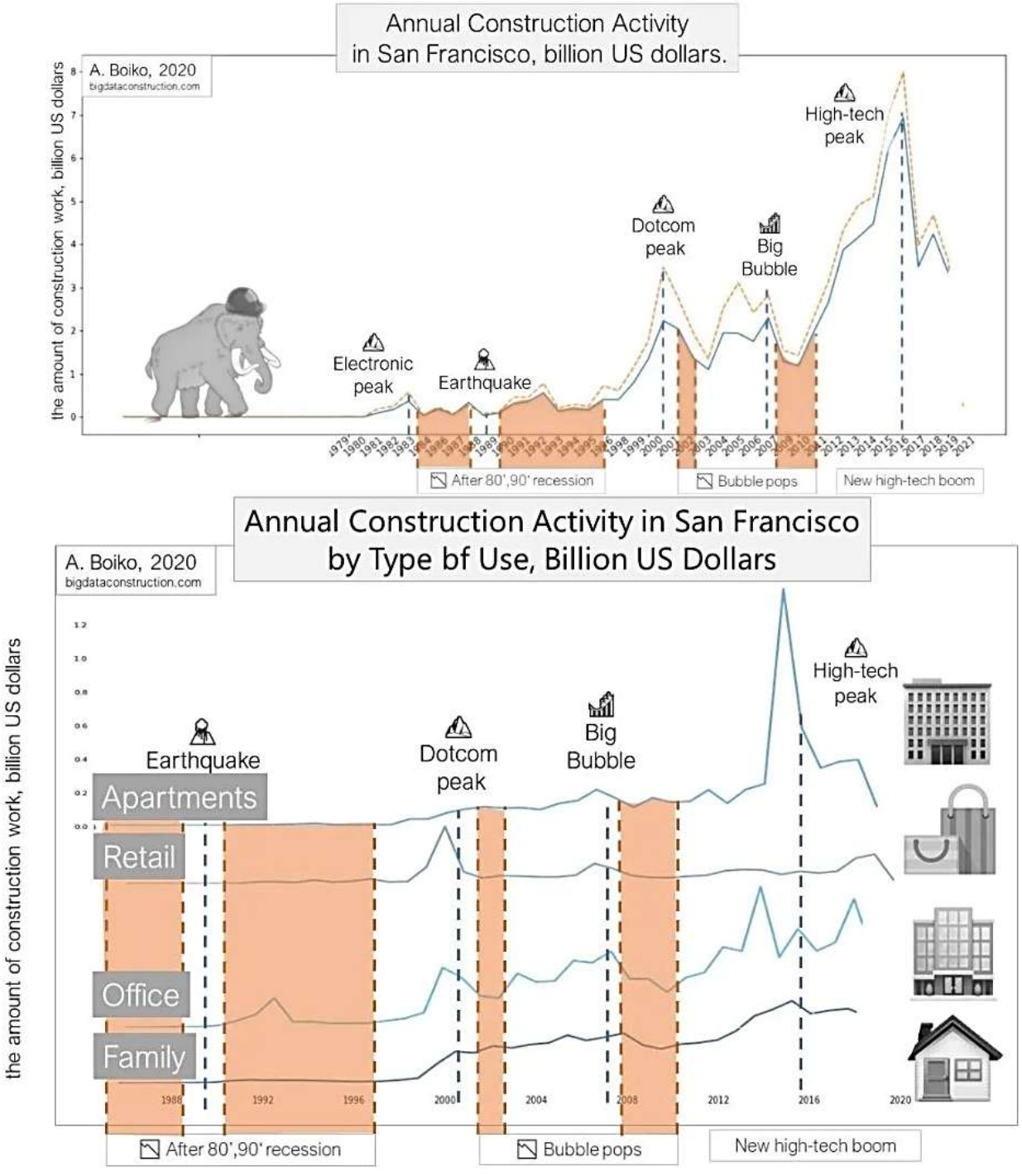


图 9.1-4 在旧金山的房地产领域，投资与硅谷的技术发展相关。

数据分析显示，在旧金山，过去十年中用于建设和重建的 915 亿美元投资中的大部分（近 75%）集中在市中心（图 9.1-5 - 使用 Pandas 和 Folium 可视化库）和市中心 2 公里半径范围内，反映出这些中心区域的投资密度较高。

建筑许可证的平均费用因街区不同而有很大差异，市中心的申请费用是市中心以外地区的三倍，原因是土地、劳动力、材料成本较高，以及严格的建筑法规要求使用更昂贵的材料来提高能源效率。

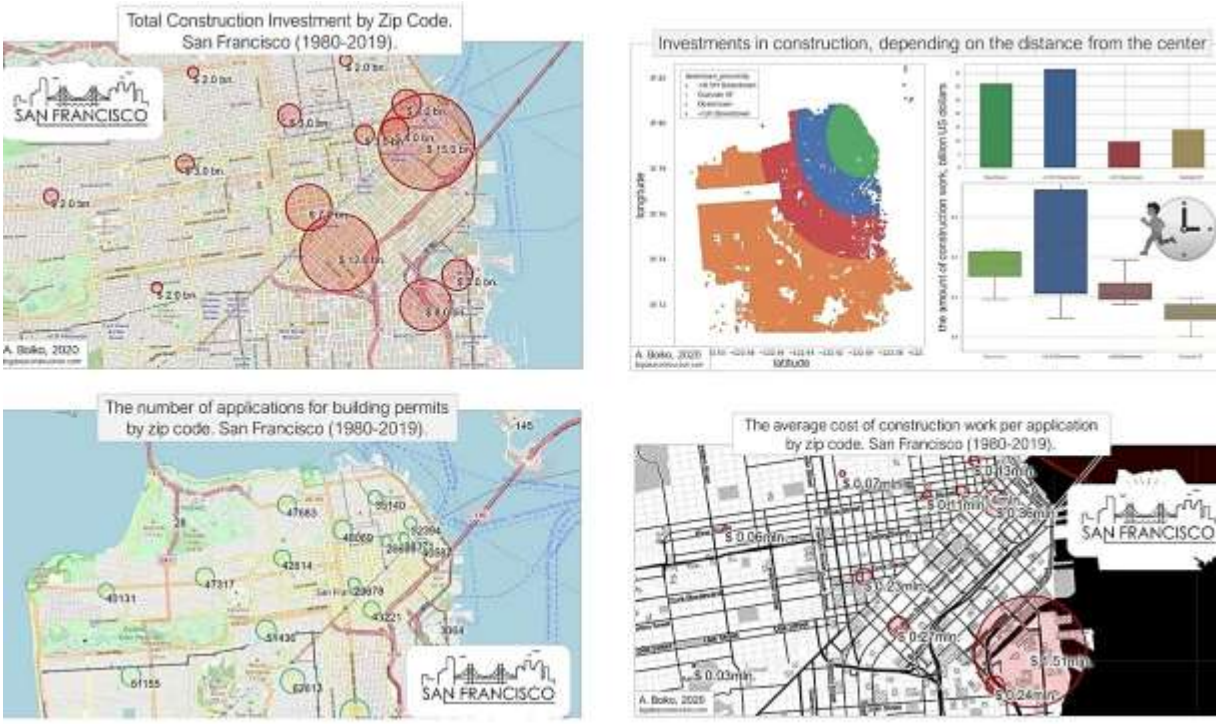


图 9.1-5 在旧金山，75% 的建筑投资（915 亿美元）集中在市中心。

该数据集不仅可以按房屋类型计算平均维修价格，还可以按城市街区和单个地址（邮政编码）计算平均维修价格。在旧金山，不同类型的翻修和住房的房屋翻修成本动态呈现出不同的趋势（图 9.1-6 - 使用 Pandas 和 Matplotlib ）。厨房翻新的费用明显高于浴室翻新：单户型住宅的厨房翻新费用平均约为 28,000 美元，而双户型住宅的厨房翻新费用平均为 25,000 美元。

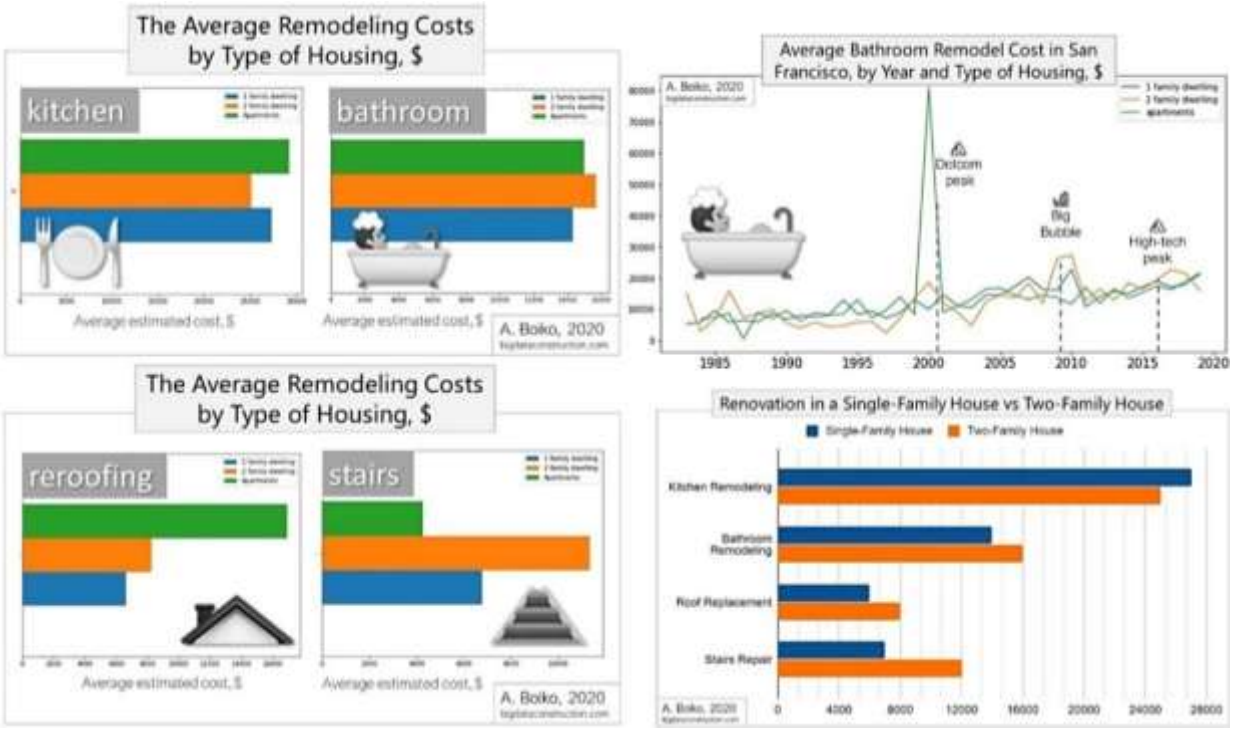


图 9.1-6 在旧金山，厨房翻新的费用几乎是浴室翻新费用的两倍，房主需要在 15 年内每月预留$ 350，以支付房屋大修的费用。

通过分析按房屋类型和年份分组的数据，可追溯旧金山多年来的建筑成本通胀情况（图 9.1-7 - 由 Pandas 和 Seaborn 使用），该图显示自 1990 年以来平均维修成本稳步上升，并揭示了多户建筑维修成本的短期三年周期。

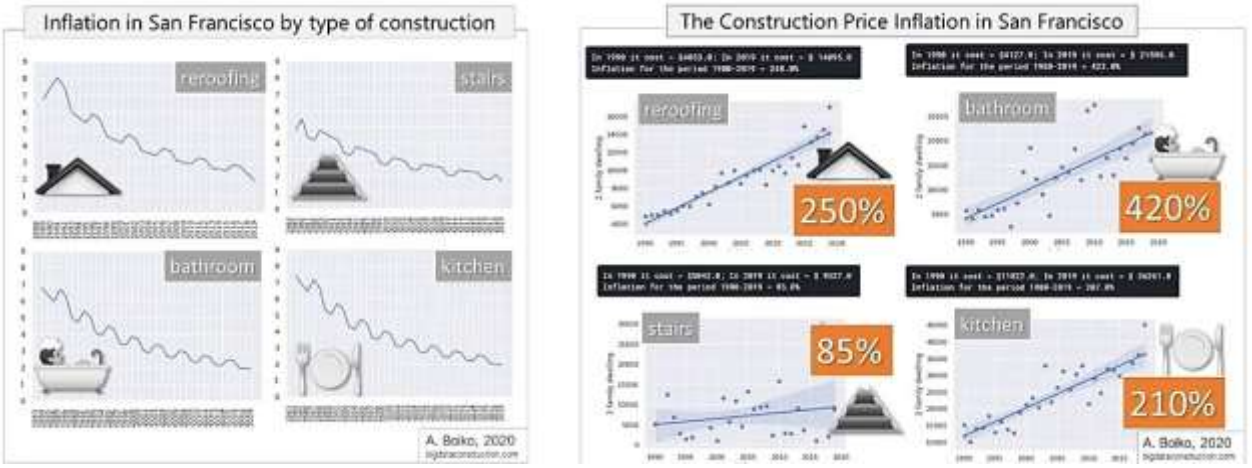


图 9.1-7 从 1980 年到 2019 年，旧金山的浴室翻新费用增长了五倍，而屋顶和厨房翻新的价格增长了三倍，楼梯翻新只增长了 85%。

对旧金山建筑部公开数据的研究（图 9.1-3）显示，该市的建筑成本受多种因素影响，变化极大，且往往难以预测。这些因素包括经济增长、技术创新以及不同类型房屋的独特要求。

过去，此类分析需要深入的编程和分析知识。然而，随着 LLM- 工具的出现，从设计部门的工程师到高级管理人员，建筑行业的各类专业人员都可以使用和理解这一过程。

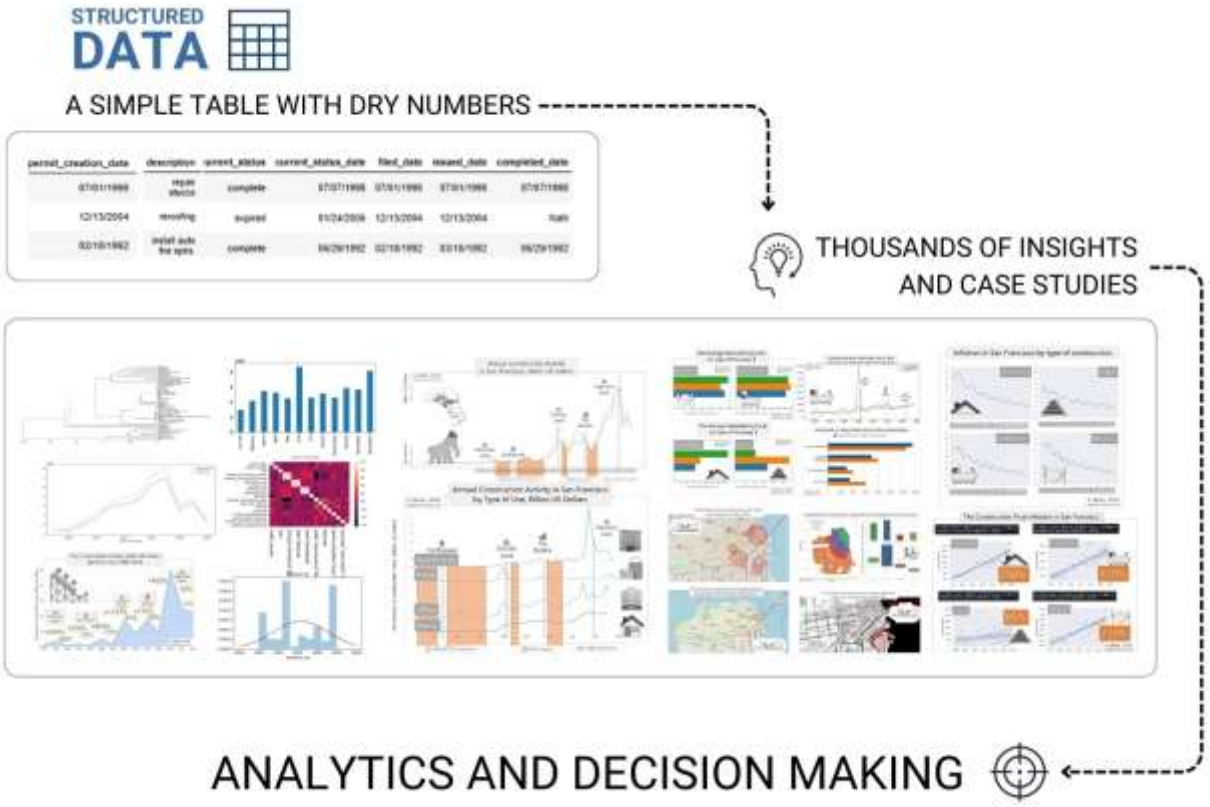


图 9.1-8 向可视化理解数据的转变可通过识别隐藏模式实现自动决策。

正如我们分析 "旧金山建筑管理局 "表格数据集的数据一样，我们可以对任何数据集进行可视化分析--从图像和文档到物联网数据，或从衍生的 CAD 数据库。

基于 **CAD** 数据的大数据示例 **(BIM)**

在下面的示例中，我们将使用来自不同 CAD 工具 (BIM) 的数据分析一个大型数据集。为了收集和创建这个大型数据集，我们使用了一个专门的自动网络爬虫（脚本），该爬虫被 配置为从提供免费建筑模型的网站上自动搜索和收集设计文件，这些模型的格式为 RVT 和 IFC。在几天内，爬虫成功找到并下载了 4,596 个 IFC 文件、6,471 个 RVT 文件和 156,024 个 DWG 文件[149]。

在收集了不同版本的 RVT 和 IFC 格式的项目并使用免费的 逆向工程 SDK 将其转换为结构化的 CSV 格式 之后，近 1 万个 RVT 和 IFC 项目被收集到一个大型 Apache Parquet 表文件 中，并上传到 Pandas DataFrame 中进行分析（图 9.1-9）。

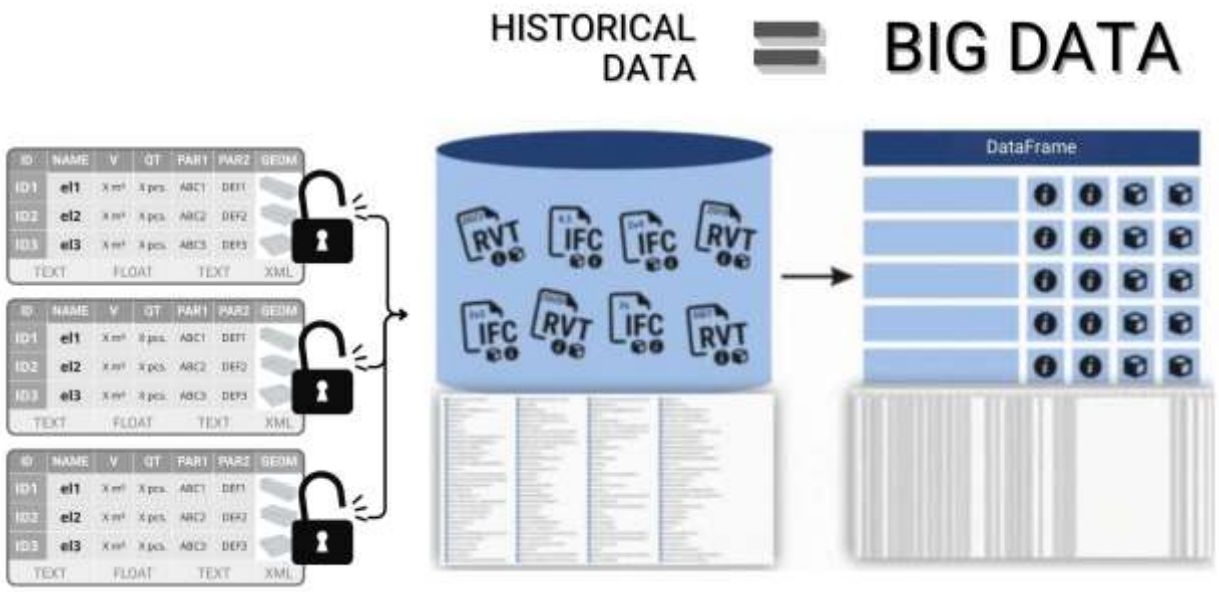


图 9.1-9 结构化数据 项目数据允许您将任意数量的项目合并到一个二维表中。

这一大规模收集的数据包含以下信息：IFC 文件集 包含约 400 万个实体（行）和 24,962 个属性（列），RVT 文件集 包含约 600 万个实体（行）和 27,025 个不同属性（列）。

这些信息集（图 9.1-10）涵盖了数百万个元素，其中每个元素的边界框 几何图形（定义项目中对象边界的矩形）的坐标都被额外获取并添加到一个通用表中--边界框 几何图形（定义项目中对象边界的矩形）的坐标以及每个元素的 PNG 格式图像和开放 XML 格式的几何图形 - DAE (Collada)。

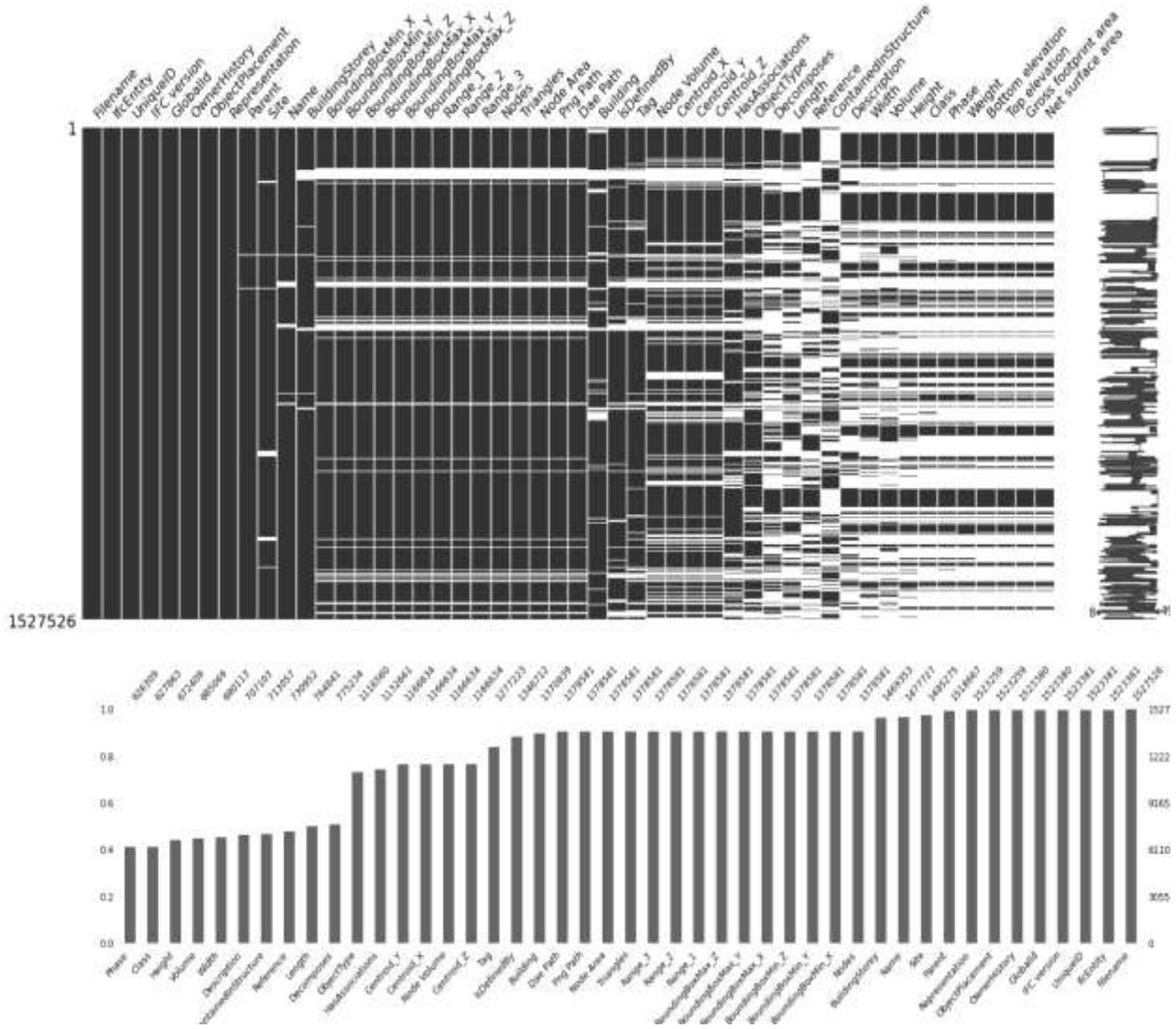


图 9.1-10 150 万个元素的 Sabset 和直方图（missingno 库）前 100 个属性的占用率。

因此，我们从 4,596 个 IFC 项目 和 6,471 个 RVT 项目 中获得了有关数千万个元素的所有信息，其中所有实体元素的所有属性-属性及其几何形状 （边框 ）都被转换成了单一表格的结构形式（数据框 ）（图 9.1-10 - 数据框中的数据以柱状图显示）。

> 在分析过程中绘制的直方图（图 9.1-10、图 9.2-6、图 9.2-7）可以快速评估列中值的数据密度和出现频率。这样就可以初步了解特征的分布情况、是否存在异常值以及各个属性在分析和建立机器学习模型时的潜在作用。

该数据集（图 9.1-10）的一个实际应用实例是 "5000 个 IFC 和 RVT "项目。[149] ，可在 Kaggle 平台 上查

看。它为 Jupyter Notebook 提供了完整的管道解决方案：从数据预处理和分析到使用 Python 库（pandas、matplotlib、seaborn、folium 等）实现结果可视化（图 9.1-11）。

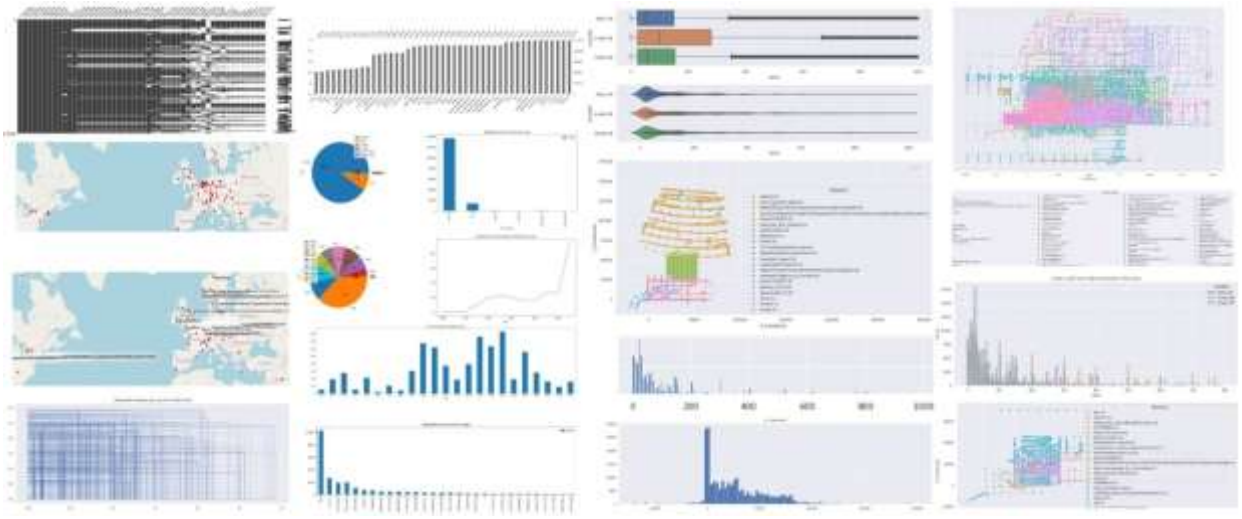


图 9.1-11 使用 Python 可视化库和 pandas 库分析 CAD 格式数据的示例 (BIM)。

根据元信息，可以确定某些项目是在哪些城市开发的，并将其显示在地图上（如使用褶皱库）。此外，通过数据中的时间戳，还可以探索文件保存或编辑的时间模式：按星期、时间和月份划分。

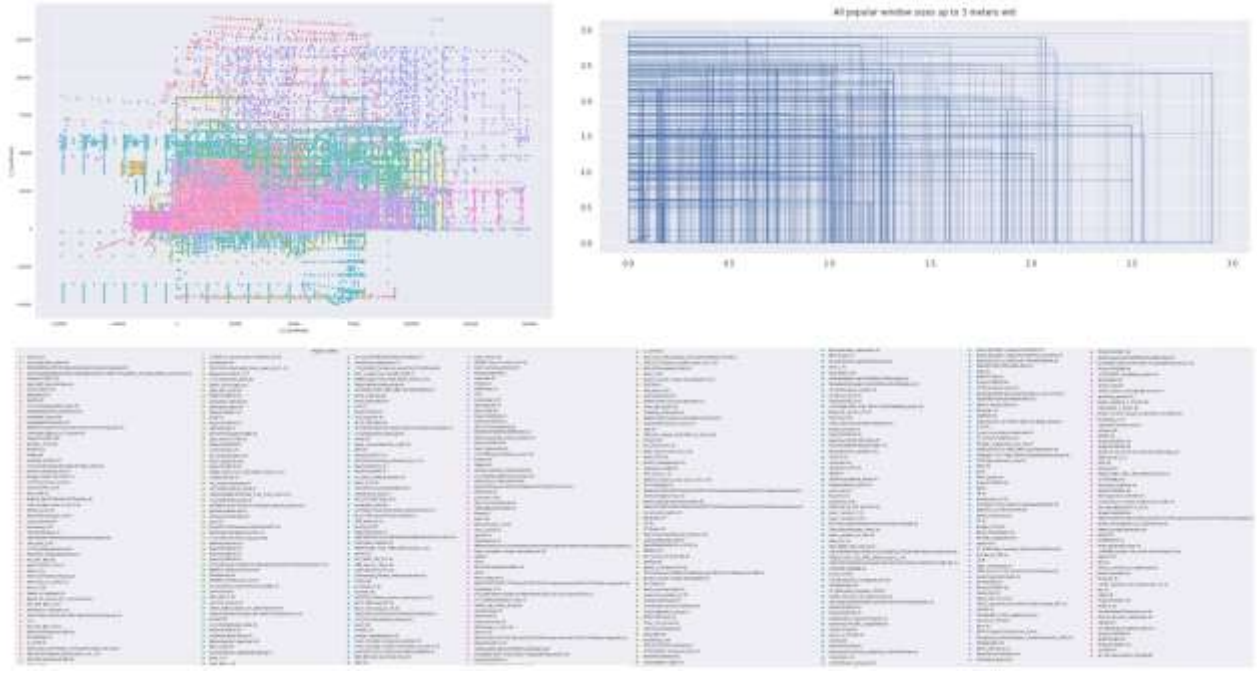


图 9.1-12 可视化 从图表底部的列表中查看所有立柱的几何位置和所有窗口的尺寸，最长可达 3 米。

从模型中提取的边界框形式的几何参数也可用于总体分析。例如，图 9.1-12 显示了两幅图：左边的图显示了所有项目相对于零点的列间距离分布，右边的图显示了在数万个窗元素样本中高度不超过 3 米的所有窗户的尺寸（将整个数据集按参数 "类别 "分组后，值为 "OST_Windows"、"IfcWindows"）。

本例的 Pipeline 分析代码 和数据集本身可在 Kaggle 网站 上获取，标题为 "5000 IFC and RVT | DataDrivenCo-nstruction.io projects" [149]。完成的管道和数据集可以在 Kaggle 上免费在线复制和运行，也可以在以下流行的集成开发环境中离线运行：PyCharm、Visual Studio Code (VS Code)、Jupyter Notebook 、Spyder、Atom、Sublime Text、带有 PyDev 插件的 Eclipse、Thonny、Wing IDE、带有 Python 插件的 IntelliJ IDEA 、JupyterLab 或流行的在线工具 Kaggle.com、Google Collab、Microsoft Azure Notebooks、Amazon SageMaker。

> 从处理和研究海量结构化数据中获得的分析见解将在建筑业的决策过程中发挥至关重要的作用。

通过这种基于以往项目的信息分析，专家们可以有效地预测材料和劳动力需求，并在施工开始前优化设计方案

然而，虽然设计数据或建筑许可证是相对静态的信息，变化相对缓慢，但施工过程本身正迅速充斥着各种传感器和物联网 - 设备：摄像头、实时传输数据的自动监控系统 - 所有这些都将施工现场变成一个需要实时分析数据的动态数字环境。

IoT 物联网和智能合约

物联网 物联网代表了数字化转型的新浪潮，每个设备都有自己的 IP 地址，并成为全球网络的一部分。物联网的概念是将物理对象连接到互联网，以收集、处理和传输数据。在建筑领域，这意味着能够实时控制施工过程，最大限度地减少材料浪费，预测设备磨损情况，并实现决策自动化。

> 根据 CFMA 的文章 "用互联建筑为未来做好准备"[150]，建筑行业将在未来十年经历重大的数字化转型，最终形成互联建筑的概念--一个完全集成的自动化建筑工地。[150]，建筑业将在未来十年经历一场重大的数字化转型，最终形成互联建筑的概念--一个完全集成和自动化的建筑工地。

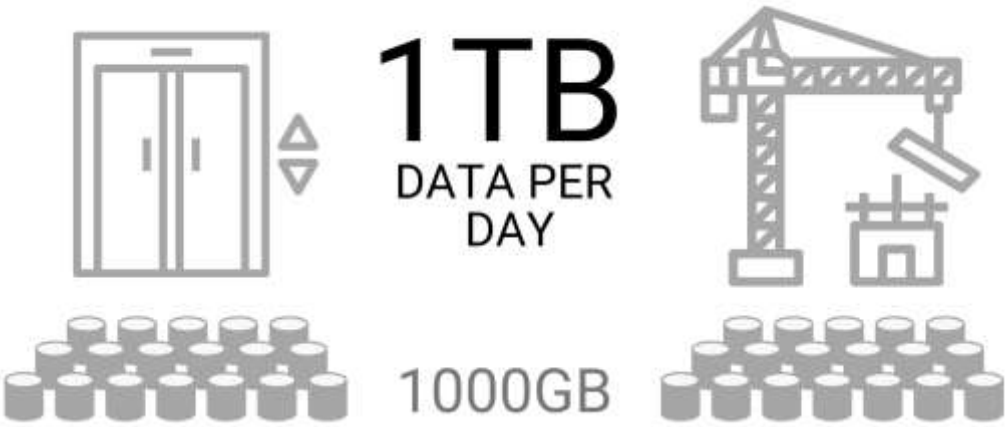


图 9.1-13 物联网或建筑工地数据设备每天可产生和传输 TB 级数据。

数字施工现场意味着施工的所有要素--从规划和物流到施工现场的工作执行和质量控制（使用固定摄像机和四旋翼飞行器）--都将被整合到一个单一的动态数字生态系统中。在本书第 7 部分的前面部分，我们已经介绍了 Apache NiFi 的功能（图 7.4-5），这是一款免费的开源工具，可实现实时数据流（从各种来源收集到传输到存储或分析平台）。

有关施工进度、材料消耗、设备状态和安全的数据将实时传送到分析系统（图 9.1-13）。这样就可以预测潜在风险，及时应对偏差，优化现场流程。数字化施工现场的主要组成部分包括

- 物联网 - 传感器 - 跟踪环境参数、监控建筑设备和控制劳动条件。
- 数字双胞胎--建筑物和基础设施的虚拟模型，用于预测可能出现的偏差并防止错误。
- 自动化物流系统 - 实时供应链管理，减少停机时间和成本。
- 机器人建筑系统--使用自主机器执行常规和危险任务。

机器人化物联网 以及 "互联工地（建筑）"的数字工地概念，不仅将提高效率、降低成本，还将开创一个安全、可持续建筑和预测性项目管理的新时代。

> RFID （射频识别）标签也是物联网 最重要的组成部分之一。它们用于识别和跟踪建筑工地上的材料、机械甚至人员，从而提高项目资源的透明度和控制力。

RFID - 该技术利用无线电信号自动识别物体。它由三个关键要素组成：

- RFID - 标签（无源或有源） - 包含一个唯一的标识符，并附在材料、工具或机械上。

- 扫描仪是从标签中读取信息并将其传输到系统的设备。
- 中央数据库 - 存储有关物体位置、状态和移动的信息。

RFID 在建筑业中的应用：

- 自动材料核算--预拌混凝土产品、钢筋或夹芯板包装上的标签可实现库存控制和防盗。
- 人员工作控制 - RFID - 员工徽章记录轮班开始和结束时间，提供工作时间记录。
- 设备监控 - RFID - 系统跟踪设备的移动，防止停机并提高物流效率。

与该技术套件相辅相成的是基于区块链的智能合约，可自动完成支付、交付控制和合同合规，无需中间人，从而降低了欺诈和延误的风险。

> 如今，在缺乏通用数据模型的情况下，智能合约只是参与者商定的代码。然而，有了以数据为中心的方法，就有可能创建一个通用的合约参数模型，将其编码到区块链中，并自动履行条款。

例如，在供应链管理系统中，智能合约可以通过物联网 - 传感器和 RFID - 标签跟踪货物的交付情况，并在货物到达时自动转账付款。同样，在建筑工地上，智能合约可以根据无人机或建筑传感器提供的数据，记录工作阶段（如安装钢筋或浇筑地基）的完成情况，并自动启动对承包商的下一次付款，而无需人工检查和纸质证明。

但是，尽管有新技术和国际标准化组织的努力，大量相互竞争的标准还是让物联网的前景变得更加复杂。

> 根据思科 2017 年发布的研究报告[151]，近 60% 的物联网计划（物联网 ）止步于概念验证阶段，只有 26% 的公司认为其物联网项目完全成功。此外，在已完成的项目中，有三分之一没有实现既定目标，甚至在实施后也不被认为是成功的。

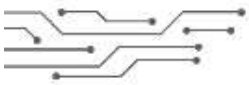
其中一个关键原因是处理不同传感器数据的平台之间缺乏互操作性。因此，数据仍然被隔离在不同的解决方案中。与其他类似情况（我们在本书中已经介绍过）一样，这种方法的替代方案是围绕数据本身这一主要资产建立的架构。

> 物联网传感器不仅在监控设备技术状况方面发挥着关键作用，而且在预测分析方面也发挥着重要作用，通过预测故障和偏差，降低施工现场的风险，提高整体流程性能。

物联网传感器和 RFID 标签收集的数据可通过机器学习算法进行实时处理，从而检测出异常情况，并提前向工程师发出潜在故障警报。这可能包括混凝土结构中的微小裂缝，也可能包括塔式起重机运行中的异常停顿，表明存在技术故障或违反法规。此外，先进的行为分析算法还能捕捉可能表明人员身体疲劳等情况的行为模式，从而加强对现场安全和员工福利的主动管理。

在建筑行业，无论是机械还是人员，事故和故障都很少突然发生。在它们发生之前，通常会出现一些微小的偏差，而这些偏差却没有被注意到。预测分析 和机器学习可以在早期阶段检测到这些信号，甚至是在出现严重后果之前。

文档、项目文件以及来自物联网设备和 RFID 标签的数据构成了建筑项目的数字足迹，而机器学习则有助于从中提取有用的见解。随着数据的增长和数据访问的民主化，建筑行业正在分析、预测分析和人工智能应用方面获得新的机遇。



章 9.2.

机器学习和预测

机器学习 和人工智能将改变我们的构建方式

建筑企业中各种系统的数据库--其不可避免的衰变和日益复杂的基础设施--正在成为未来解决方案的温床。公司服务器就像一片森林，蕴藏着大量的重要信息，这些信息往往隐藏在文件夹和服务器的地下。今天从各种系统中产生的大量数据--在使用之后、掉落到服务器底部之后以及经过多年的僵化之后--将为未来的机器学习 和语言模型提供动力。公司内部聊天室（例如，本地配置的 ChatGPT 、LlaMa 、Mistral 、DeepSeek 的独立实例）将建立在这些内部模型的基础上，利用集中存储快速方便地检索信息并生成必要的图表、仪表盘和文档。

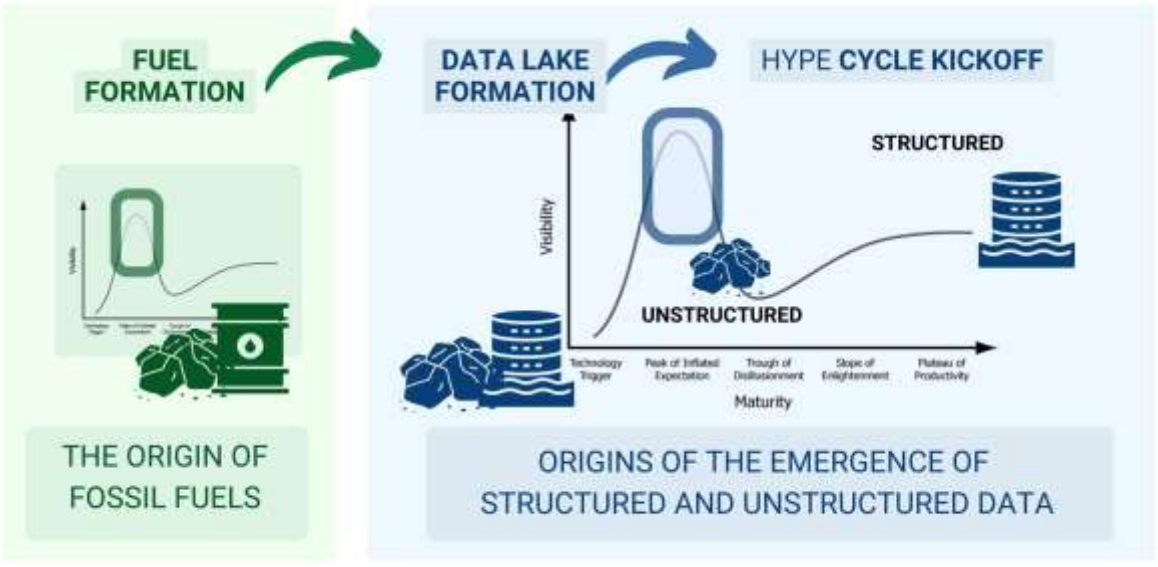


图 9.2-1 正如树木会变成煤炭一样，在时间和分析的压力下，信息也会随着时间的推移变成宝贵的商业能源。

植物体化石与压力和温度相结合，形成了由生活在不同时期的不同物种的树木组成的结构独特的同质体--木炭[152]。同样，在分析部门的压力和质量管理的温度作用下，不同格式、不同时间记录在硬盘上的信息最终会形成一个结构均匀的有价值的信息块（图 9.2-1）。

这些信息层（更常见的是孤立的信息块）是由经验丰富的分析师通过艰苦的数据组织工作创建的，他们开始

从看似长期无关的数据中逐步提取 有价值的信息。

当这些成熟的数据层不再只是 "刻录 "在报告中，而是开始在业务流程中流通，丰富决策和改进流程时，公司就为下一步--向机器学习和人工智能过渡--做好了准备（图 9.2-2）。

> 机器学习 （ML - Machine learning）是一类解决人工智能问题的方法。机器学习算法 可识别大型数据集中的模式，并利用这些模式进行自我学习。每个新的数据集都能让数学算法根据获得的信息进行改进和调整，从而不断提高建议和预测的准确性。

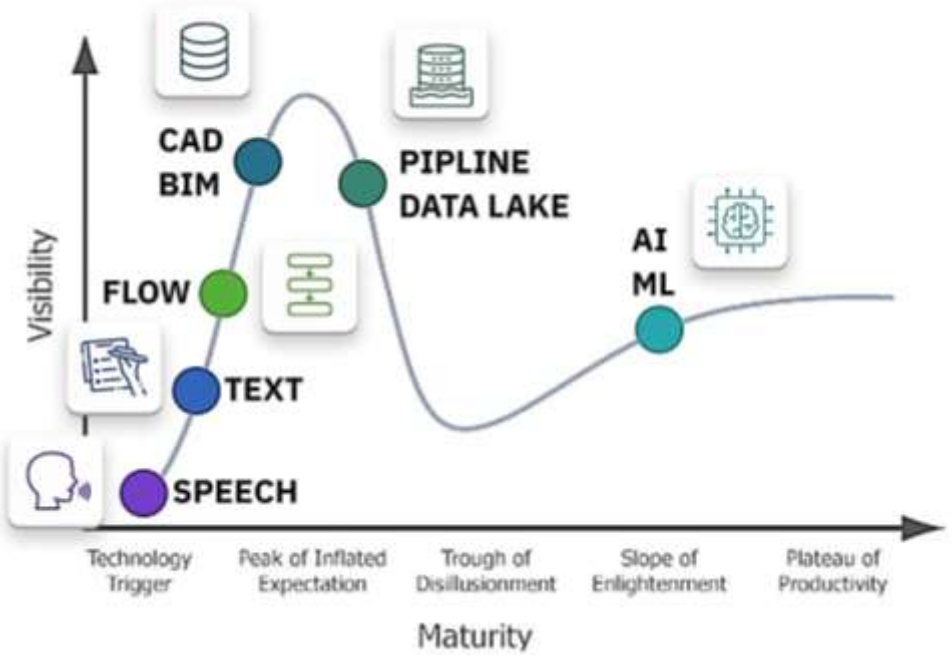


图 9.2-2 数据创建技术的消退和分析工具的应用 为机器学习主题打开了大门。

世界上最大的投资基金（该基金拥有几乎所有最大建筑软件公司的重要股份，也拥有世界上最大的房地产公司[55]）的首席执行官在 2023 年的一次采访中说道- 机器学习将改变建筑世界。

> 人工智能 潜力巨大。它将改变我们的工作方式和生活方式。人工智能和机器人技术将改变我们的工作方式和建设方式，我们将能够利用人工智能和机器人技术作为一种手段，创造更高的生产力[153]。
>
> - 全球最大投资基金首席执行官，采访，2023 年 9 月。

机器学习 (ML) 的工作原理是处理大量数据，利用统计技术模仿人类思维的各个方面。然而，大多数公司都没有这样的数据集，即使有，也往往没有充分标注。这正是语义技术和迁移学习可以提供帮助的地方，迁移学习是一种可以让人工智能在处理少量数据时更加有效的技术，其可行性已在本部分前几章中讨论过。

迁移学习的精髓在于，与其从头开始学习每项任务，不如利用在相关领域获得的知识。有必要认识到，其他行业的模式和发现可以在建筑业中得到调整和应用。例如，零售业开发的优化物流流程的方法有助于提高建筑供应链管理的效率。在金融业得到积极应用的大数据分析可以应用于建筑项目的成本预测和风险管理。而在工业领域开发的计算机视觉和机器人技术已经在自动质量控制、安全监控和建筑工地设施管理中得到应用。

> 转移学习不仅可以加快创新成果的引进，还可以利用其他行业已经积累的经验降低创新成果的开发成本。



图 9.2-3 人工智能 技术和机器人技术将成为未来提高建筑业生产率的主要推动力。

人类思维的组织原理与此类似：我们以先前获得的知识为基础来解决新问题（图 4.4-19，图 4.4-20，图 4.4-21）。在机器学习中，这种方法也同样有效--通过简化数据模型并使其更加优雅，我们可以降低 ML 算法的问题复杂度。这反过来又减少了对大量数据的需求，降低了计算成本。

从主观评估到统计预测

战略决策依赖管理者个人直觉的时代（图 9.2-4）已经成为过去。在竞争日益激烈和充满挑战的经济环境中，主观臆断的方法风险太大，效率太低。如果公司继续依赖个人观点，而不是客观分析 数据 ，就会丧失快速应对变化的能力。

> 竞争环境要求基于数据、统计模式和可计算概率的准确性和可重复性。决策不能再凭感觉，而必须基于相关性、趋势以及从分析 和机器学习 中得出的预测模型。这不仅仅是工具的变化，更是思维逻辑的变化：从假设到证据，从主观概率到统计计算的偏差，从感觉到事实。



图 9.2-4 随着大数据和机器学习的出现，由 HiPPO （薪酬最高的员工的意见）做出决策的时代将成为过去。

过去只凭感觉行事的管理者将不可避免地面临新的现实：权威不再决定选择。现在，管理的核心是分析数以百万计的参数和矢量、识别隐藏模式并提出最佳战略建议的系统。

如今，企业仍然避免实施 ML 的主要原因是其缺乏透明度。对于管理者来说，大多数模型就像 "黑盒子 "一样，无法解释它们是如何。这就导致了一些问题：算法可能会强化刻板印象甚至造成幽默的局面，比如微软

的聊天机器人就很快变成了一个有毒的交流工具[154]。

> 前国际象棋世界冠军加里-卡斯帕罗夫（Garry Kasparov）在《深度思考》（Deep Thinking）一书中回顾了他被 IBM Big Blue 计算机击败的经历[155]。他认为，人工智能 的真正价值不在于复制人类智能，而在于补充我们的能力。人工智能应该执行人类薄弱的任务，而人类则应发挥创造力。计算机改变了分析国际象棋的传统方法。计算机国际象棋程序不再创造引人入胜的对局故事，而是公正地评估每一步棋，只看它的实际强弱。卡斯帕罗夫指出，人类倾向于将事件视为连贯的故事，而不是单个行动，这往往会导致错误的结论--不仅在国际象棋中如此，在一般生活中也是如此。

因此，如果您计划使用机器学习进行预测和分析，那么在开始在工作中使用机器学习工具 和人工智能 之前，了解其基本原理--算法如何工作以及数据如何处理--是非常重要的。最好的入门方法就是亲身体验。

Jupyter Notebook 和广受欢迎的经典数据集是初步介绍机器学习 和预测主题的最方便工具之一，它将为分析数据和构建 ML 模型的关键方法提供直观的介绍。

泰坦尼克数据集：分析世界中的 "你好世界" 数据和大数据

在数据分析中使用 ML 的一个最著名的例子就是对泰坦尼克号数据集的分析，该数据集通常用于研究乘客的生存概率。学习这个表格类似于学习编程语言时的 "Hello World "程序。

1912 年泰坦尼克号沉没导致 2224 人中的 1502 人死亡。泰坦尼克号数据集不仅包含乘客是否幸存的信息，还包含年龄、性别、船票等级和其他参数等属性。该数据集可在各种离线和在线平台上免费打开和分析。

> 泰坦尼克号数据集链接：
> https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv

在前面的 "支持 LLM 的集成开发环境 和未来的编程变化 "一章中，我们已经讨论了 Jupyter Notebook - 用于数据分析和机器学习的最流行的开发环境之一。与 Jupyter Notebook 类似的免费云平台有 Kaggle 和 Google Collab 平台，它们允许您在不安装软件的情况下运行 Python 代码，并提供对计算资源的免费访问。

> Kaggle 是最大的数据分析、机器学习竞赛平台，拥有集成的代码执行环境。截至 2023 年 10 月，Kaggle 已拥有来自 194 个国家/地区的 1500 多万用户 [156]。

在 Kaggle 平台上下载并使用 Titanic 数据集（图 9.2-5）来存储数据集（数据集的副本），并直接在浏览器中运行带有预装库的 Python 代码，而无需安装专用的集成开发环境。

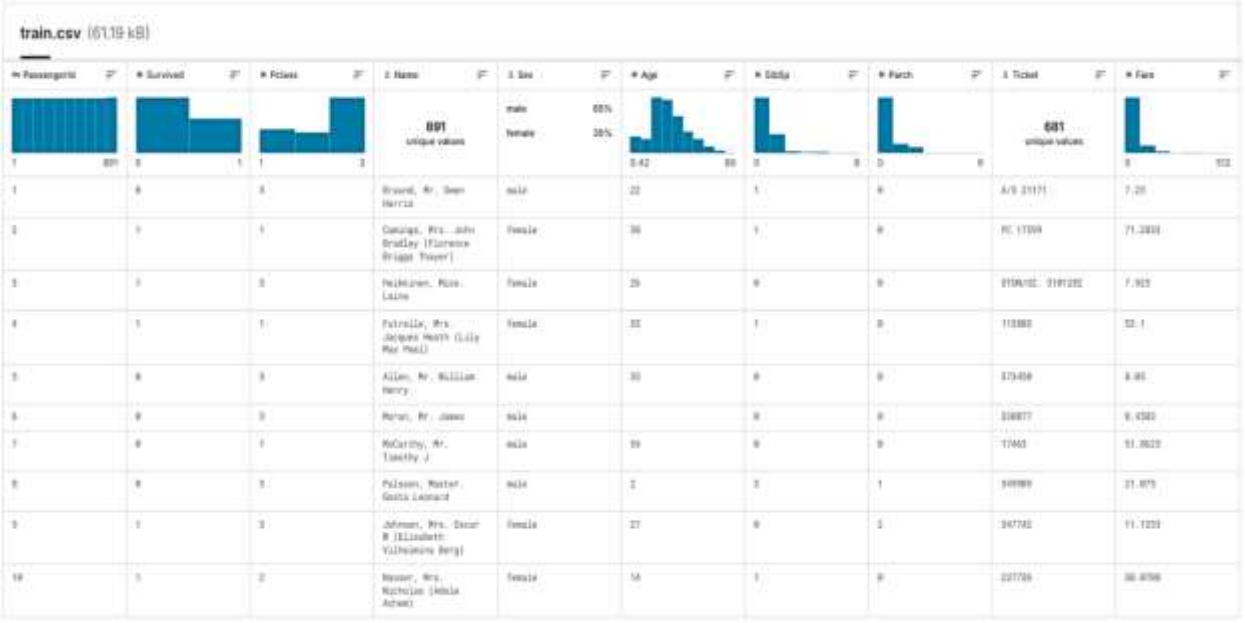


图 9.2-5 泰坦尼克号表格统计--学习数据分析和机器学习最常用的训练数据集。

泰坦尼克号数据集包括 1912 年泰坦尼克号沉船时船上 2224 名乘客的数据。该数据集以两个独立的表格（训练 (train.csv) 和测试 (test.csv) 样本）呈现，因此既可用于训练模型，也可用于评估新数据的准确性。

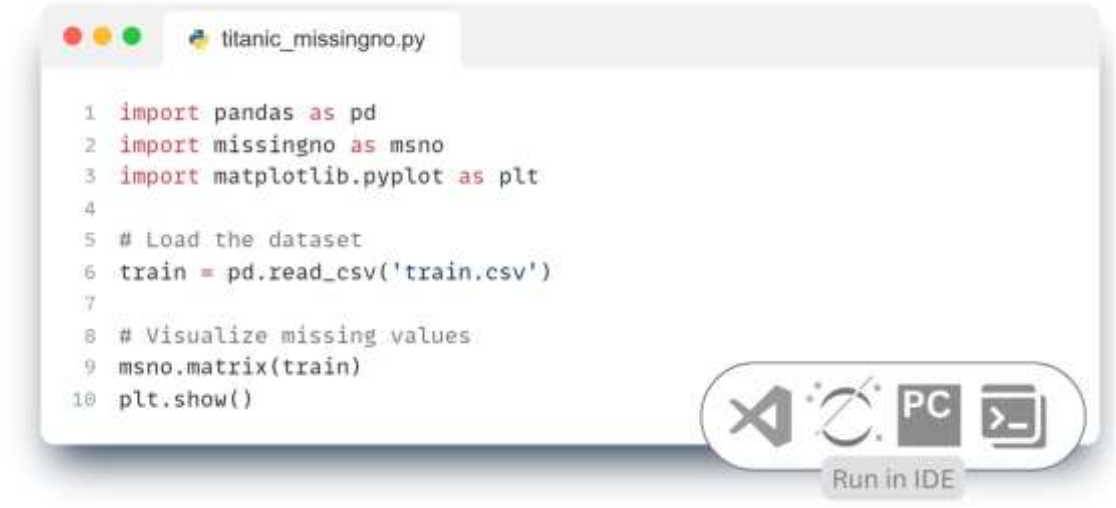
训练数据集包含两个属性--乘客属性（年龄、性别、机票等级及其他）和生还者信息（带二进制值 "生还者 "列）。训练数据集（图 9.2-6 - 文件 train.csv）用于训练模型。测试数据集（图 9.2-7 - 文件 test.csv）只包含乘客属性，不包含幸存者信息（不包含单列 "幸存者"）。测试数据集用于在新数据上测试模型并评估其准确性。

因此，训练数据集和测试数据集中的乘客属性几乎完全相同。唯一的关键区别在于，在测试数据集中，我们有一个没有 "幸存者 "列的乘客列表，而 "幸存者 "列正是我们希望通过各种数学算法学习预测的目标变量。建立模型后，我们就可以将模型输出与测试数据集中的真实参数 "Survivor "进行比较，从而评估结果。

表中主要栏目，训练和测试数据集中的乘客参数：

- **PassengerId** - 唯一的乘客标识符
- 幸存 - 1 表示乘客幸存，0 表示乘客死亡（测试集中无此数据）。
- **Pclass** - 机票等级（1、2 或 3）
- 姓名 - 乘客姓名
- 性别 - 乘客的性别（男/女）
- 年龄
- **SibSp** - 船上兄弟姐妹或配偶人数
- **Parch** - 船上父母或子女人数
- 机票 - 机票号码
- 票价
- 机舱 - 机舱编号（许多数据缺失）
- **Embarked** 是登船港口（C = 瑟堡，Q = 皇后镇，S = 南安普顿）。

要可视化两个表中的缺失数据，可以使用 missingno 库（图 9.2-6，图 9.2-7），它以直方图的形式显示缺失值，其中白色区域表示缺失数据。这种可视化方式可以在处理前快速评估数据质量。



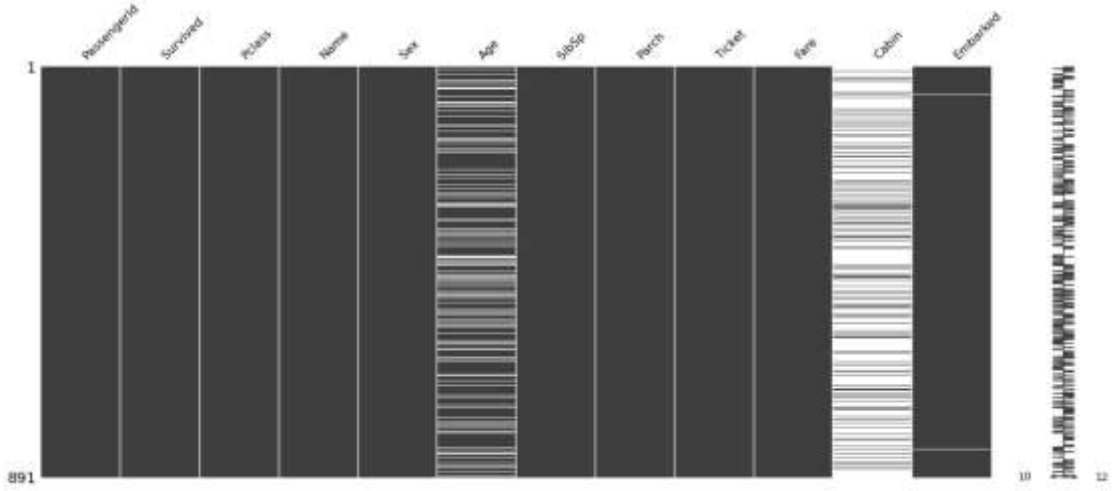


图 9.2-6 在泰坦尼克号训练数据集中，训练的关键参数是 "存活 "参数，我们使用几行代码对缺失数据进行可视化处理。

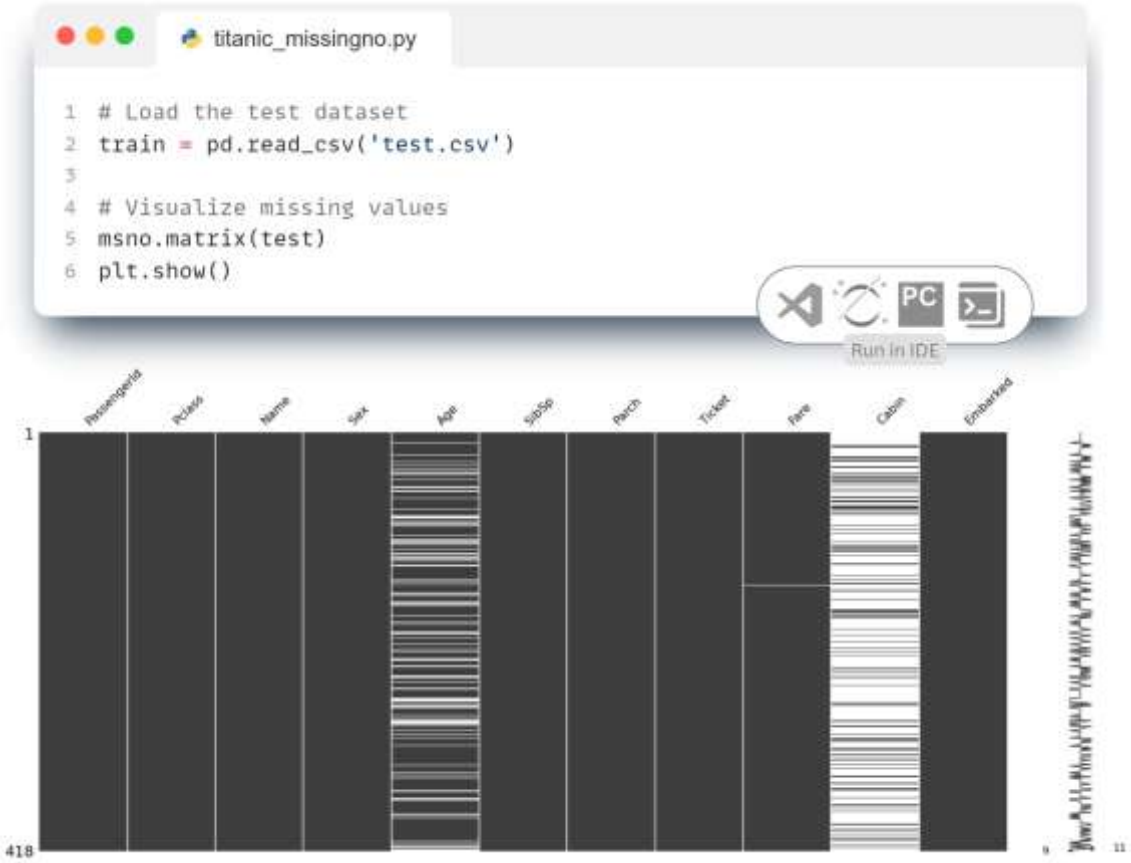


图 9.2-7 可视化 泰坦尼克号测试数据集中的缺失数据，该数据集仅包含乘客特征，没有任何信息

。

在根据数据集提出假设和进行预测之前，可视化分析有助于识别数据中的关键模式、评估数据质量并确定可能的依赖关系。有许多可视化技术可以帮助您更好地理解泰坦尼克号数据集。您可以使用分布图来分析乘客年龄组、按性别和阶级分类的生存图以及缺失数据矩阵来评估信息质量和理解数据。

- 让我们向任何 LLM 模型（CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN 或任何其他模型）发送以下文本请求，请 LLM 帮助我们将泰坦尼克号数据集的数据可视化：

> 请展示一些泰坦尼克数据集的简单图表。自行下载数据集并展示↵

- LLM 响应 以现成代码和图表的形式将数据集参数可视化

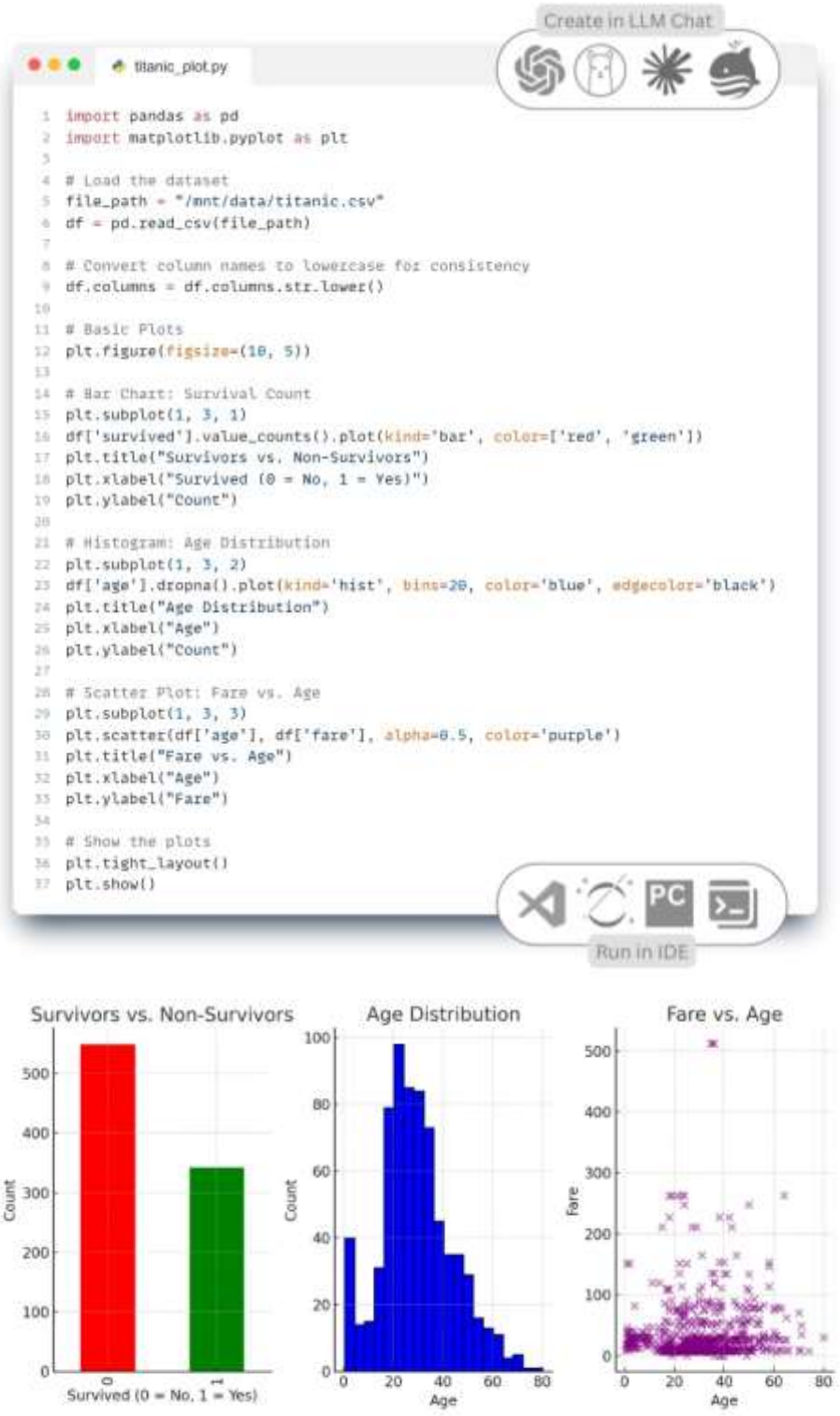


图 9.2-8 LLM 帮助您获得数据集数据的即时可视化。

数据可视化是为随后构建机器学习模型做准备的重要步骤，只有了解数据才能获取机器学习模型。

机器学习 在行动：从泰坦尼克号乘客到项目管理

在泰坦尼克号数据集的基础上，用于探索机器学习框架 的主要假设是，某些乘客群体的存活几率更高。

> 泰坦尼克号乘客的小桌子已经风靡全球，数百万人使用它进行训练、实验和模型测试，以便根据泰坦尼克号乘客的训练数据集找出哪种算法和假设能建立最准确的生存预测模型。

泰坦尼克号数据集的吸引力在于它的紧凑性：几百行和十二列（图 9.2-6），为分析提供了充分的机会。相对而言，该数据集是二元分类解决方案的一个典型示例，其中问题的目标--生存--以方便的 0 或 1 格式表示。

> 约翰-惠勒在《来自比特的它[7]中认为，宇宙是建立在二进制选择的基础上的。同样，由分子组成的人所经营的企业实际上也是建立在一系列二进制的二进制选择之上的。

此外，这些数据基于真实的历史事件，因此与人为创建的示例不同，具有很高的研究价值。仅在最大的数据管道 和 ETL 之一的 Kaggle 平台 上，就有 1,355,998 人参与了基于泰坦尼克号数据集的挑战，开发了 53,963 个独特的数据管道解决方案[157]（图 9.2-9）。

> 这看似不可思议，但泰坦尼克号乘客的 1000 行数据和 12 个参数，已成为数以百万计的假设、逻辑链和独特数据管道的领域。从一个小数据集中诞生了无穷无尽的见解、假设和解释--从简单的生存模型到考虑到隐藏模式和复杂推理迷宫的复杂组合。

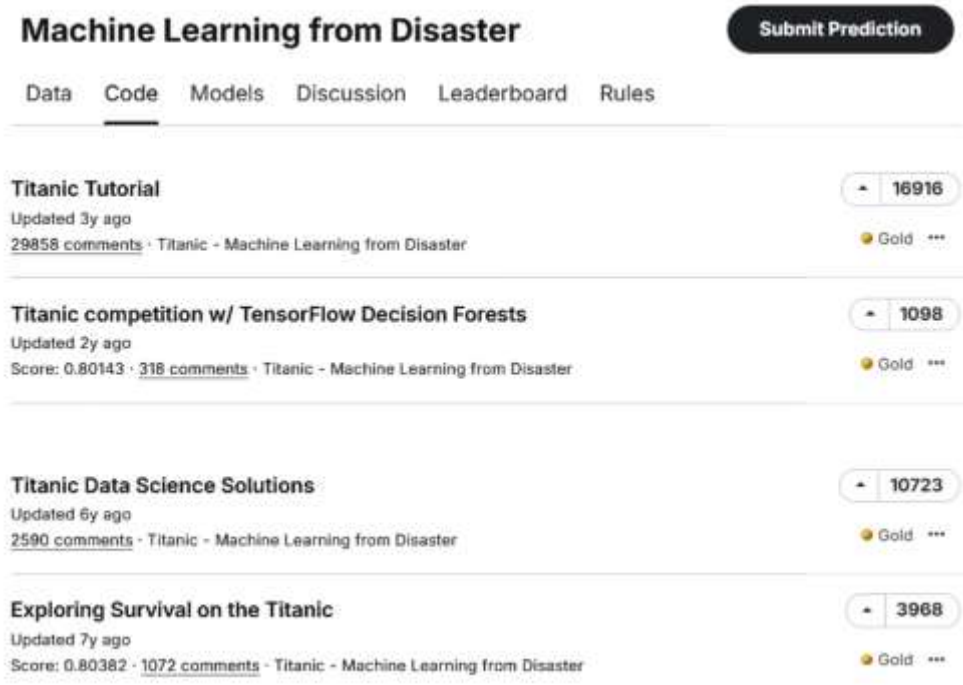


图 9.2-9 在总共 53,963 个现成的开源 Pipeline 解决方案中的前五个解决方案。仅在 Kaggle 上就有近 150 万人尝试解决这一问题 [157]。

如果连这么小的表格都能产生数百万个唯一解（图 9.2-9），那么对于参数以万为单位的实际工业建筑数据集，我们又能说什么呢?

> 一个相对较小的建筑的标准 CAD 设计包含数以万计的实体和数千个参数--从几何特征到成本和时间属性。试想一下，在贵公司多年来收集的所有项目数据中，隐藏着多少潜在的见解、关系、预测和管理假设。历史项目数据不仅仅是一个档案，它还是一个组织鲜活的记忆，是其数字足迹，可以通过分析建立大量独特的假设。

最重要的是，您不必等待 Kaggle 社区对您的公司或数据产生兴趣。您现在就可以开始使用自己的数据：在自己的数据上运行分析，在自己的数据上训练模型，识别重复、异常和 模式。过去需要多年的实验和昂贵的咨询，而现在只要有主动性、法学硕士学位、开放的数据方法和学习意愿就足够了。

- 为了建立一个机器学习算法 ，它将根据 train.csv 乘客训练数据集预测乘客存活率，让我们请 LLM 为我们解决这个问题：

> 基于泰坦尼克号乘客训练数据集，建立机器学习模型 以预测存活率 ⏎

LLM 的答复：

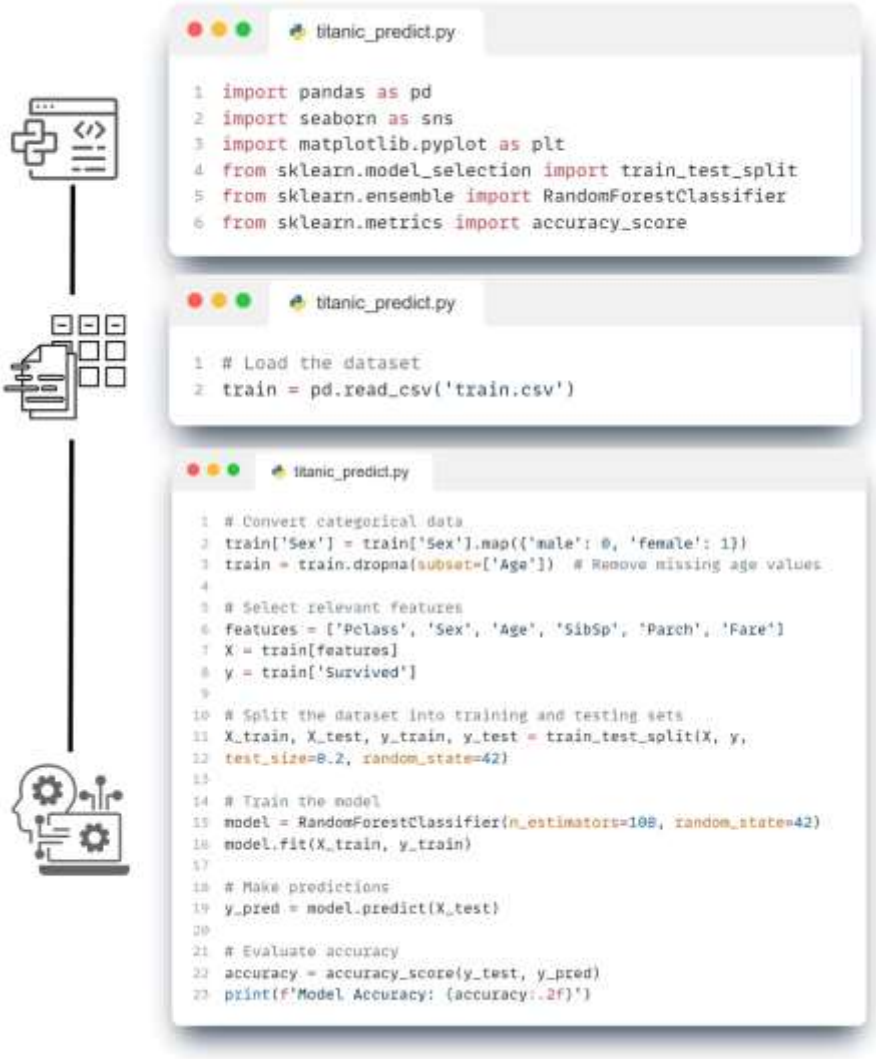


图 9.2-10 LLM 使用机器学习算法 Random Forest 建立了泰坦尼克号幸存者预测模型。

LLM 生成的代码（图 9.2-10）加载泰坦尼克号乘客数据、清理数据、转换分类变量（如将性别转换为数字格式），并通过 RandomForestClassifier 算法训练模型，以预测乘客是否幸存（我们将在后续章节中详细

介绍常用算法）。

代码将训练数据分为训练集和测试集（Kaggle 网站 已创建了现成的 test.csv（图 9.2-7）和 train.csv（图 9.2-6）用于训练，然后在训练数据上训练模型，并在测试数据上测试，以了解特定预测模型的性能如何。训练完成后，test.csv 中的测试数据（包含存活或未存活者的真实数据）将被输入模型，并由模型预测谁存活了，谁未存活。在我们的案例中，我们获得的机器学习模型 的准确率约为 80%，这表明它能很好地捕捉模式。

> 机器学习 可以比作一个孩子试图把长方形的积木塞进一个圆形的洞里。在初始阶段，算法会尝试许多方法，遇到错误和不一致。这个过程看似效率低下，但却提供了重要的学习机会：通过分析每个错误，模型可以改进其预测，并做出越来越准确的决策。

现在，这个模型（图 9.2-10）可以用来预测新乘客的存活率，例如，如果您使用参数的 model.predict 函数将乘客信息输入该模型：例如，如果使用参数为 "男性"、"三等舱"、"25 岁"、"无亲属在船上 "的 model.predict 函数向其输入乘客信息，该模型将生成一个预测结果--如果该乘客在 1912 年乘坐泰坦尼克号，那么他将有 80% 的概率无法在灾难中幸存（图 9.2-11）。

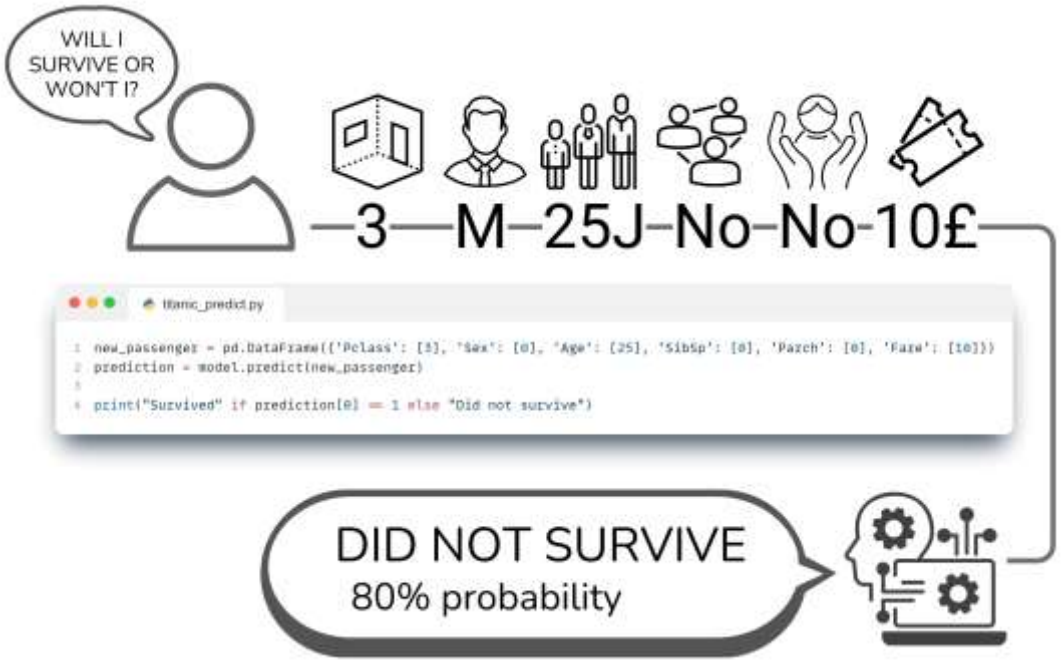


图 9.2-11 我们上面创建的模型现在可以以 80% 的概率预测泰坦尼克号新乘客是否会幸存。

泰坦尼克号乘客存活率预测模型说明了一个更为广泛的概念：每天，建筑行业成千上万的专业人员都在做出类似的 "双重 "决定--决定一个决策、一个项目、一个估算 、一个工具、利润或损失、安全或风险的生死。在泰坦尼克号的例子中，结果取决于各种因素（性别、年龄、阶级），而在建筑业中，决策的每个方面都受到许多自身因素和变量（表格中的一列）的影响：材料成本、工人资质、截止日期、天气、物流、技术风险、评论以及成千上万的其他参数。

> 在建筑行业，机器学习遵循与其他领域相同的原则：根据项目、合同、估算等历史数据对模型进行训练，以测试各种假设并找到最有效的解决方案。这个过程就像通过试验和错误来教育孩子一样：在每一个循环中，模型都会适应并变得更加准确。

累积数据的使用为建筑业开辟了新天地。通过训练模型，可以高度准确地预测未来项目的关键特征，而不是耗时的手工计算。通过这种方式，预测分析 将建筑行业变成了一个不仅可以规划，还可以自信地预测发展的空间。

基于历史数据的预测和预报

公司收集的项目数据为建立模型提供了可能性，这些模型能够预测未来尚未实现的目标的成本和时间特征，而无需耗时的人工计算和比较。这样就可以大大加快和简化估价过程，不依赖主观臆断，而是依赖可靠的数学预测。

在本书的第四部分，我们详细讨论了项目成本估算的传统方法，包括基于资源的方法 ，还提到了参数法和专家法。这些方法仍然具有现实意义，但在现代实践中，它们已开始被统计分析和机器学习工具所充实 ，这可以显著提高估算的准确性和可重复性。

人工和半自动计算价格和时间属性的过程，今后将由能够分析历史数据、发现隐藏模式并提出明智决策的人工智能模型的意见和预测加以补充。

新的数据和方案将从已有信息中自动生成--类似于语言模型（LLM ）根据多年来从开放源收集的数据创建文本、图像和代码[158]。

正如今天人类依靠经验、直觉和内部统计数据来评估未来事件一样，未来几年，建筑项目的未来将越来越多地由积累的知识和数学机器学习模型相结合来决定。

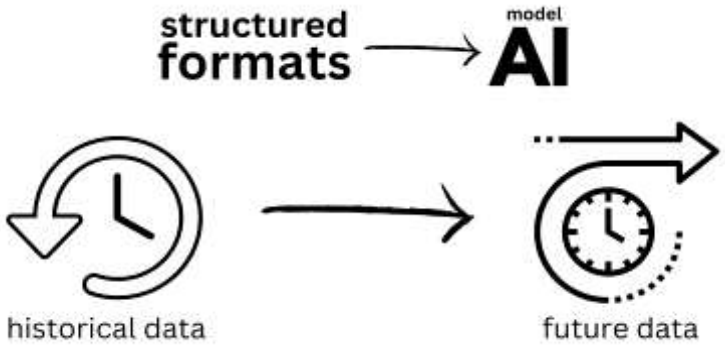


图 9.2-12 定性和结构化的公司历史数据是机器学习模型 和预测的基础材料。

举个简单的例子：根据房屋面积、地块大小、房间数量和地理位置预测房屋价格。一种方法是建立一个经典模型，分析这些参数并计算预期价格（图 9.2-13）。这种方法要求事先有一个精确的已知公式，而这在实际操作中实际上是不可能的。

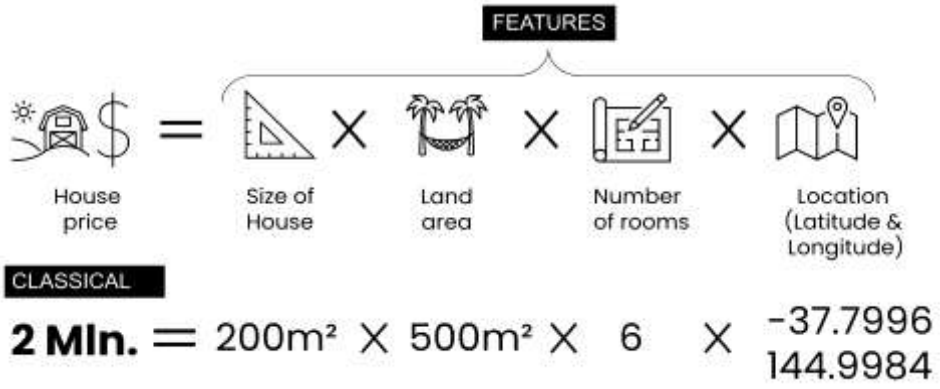


图 9.2-13 可采用经典算法估算房屋价值，并求得固定公式。

机器学习 消除了人工寻找公式的过程，取而代之的是经过训练的算法，这些算法可以独立识别依赖关系，其准确性比任何预定方程都要高出许多倍。或者，让我们创建一个机器学习算法 ，它将根据对问题的预先理解和可能不完整的历史数据生成一个模型（图 9.2-14）。

以定价为例，机器学习可以创建不同类型的数学模型，而无需了解成本形成的确切机制。该模型从以往项目的数据中 "学习"，根据建筑参数、成本和期限之间的实际模式进行调整。

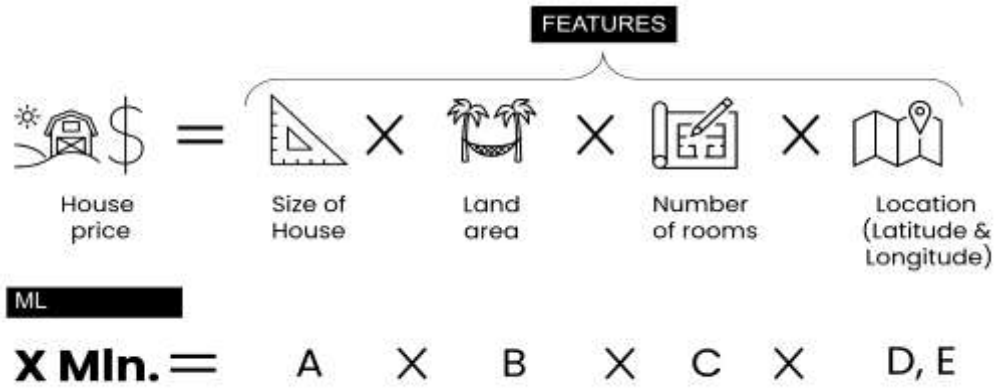


图 9.2-14 与传统的基于公式的估算不同，机器学习算法 是根据历史数据进行训练的。

在有监督机器学习 中，训练数据集中的每个项目都包含输入属性（如类似建筑的成本和时间）和预期输出值 （如成本或时间）。类似的数据集可用于创建和定制机器学习模型 （图 9.2-15）。数据集越大，数据质量越高，模型就越精确，预测结果也就越准确。

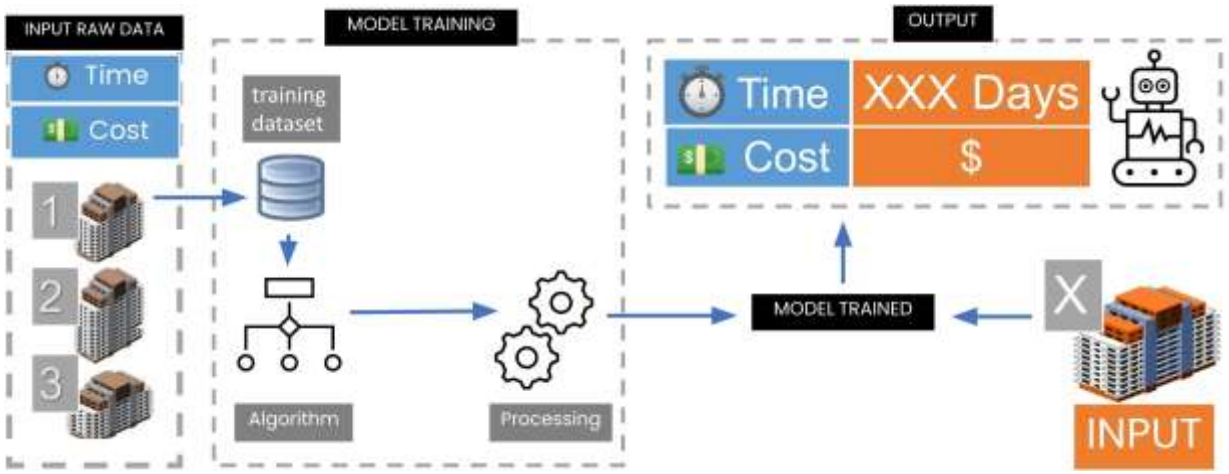


图 9.2-15 根据过去项目的成本和进度数据训练的 ML 模型将以一定的概率确定新项目的成本和进度。

一旦创建了模型并对其进行了估算新项目建设的训练，只需向模型提供新项目的新属性，模型就会根据先前

学习的模式以一定的概率提供估算结果。

机器学习的关键概念

机器学习 不是魔术，它只是数学、数据和寻找模式。它没有真正的智能，只是一个根据数据训练出来的程序，能够识别模式并做出决策，而不需要人类的不断参与。

机器学习使用一些关键概念来描述其结构（图 9.2-16）：

- 标签是模型应该预测的目标变量或属性（泰坦尼克号数据集中的 "幸存者 "参数）。例如：建筑成本（如美元）、建筑工期（如月）。
- 特征是作为模型输入的独立变量或属性。在预测模型中，它们用于预测标签。例如：地块大小（平方米）、建筑层数、建筑总面积（平方米）、地理位置（经纬度）、建筑材料类型。特征的数量也决定了数据的维度。
- 模型是一组不同的假设，其中一个近似要预测或近似的目标函数。例如：机器学习模型 ，它使用回归分析技术来预测建筑成本和时间。
- 学习算法 学习算法是利用一组训练数据在模型中找到与目标函数完全匹配的最佳假设的过程。例如：线性回归、KNN 或随机森林算法，用于分析成本和施工时间数据，以确定关系和模式。
- 训练 - 在训练过程中，算法会分析训练数据，找到与输入属性和目标标签之间的关系相对应的模式。这一过程的结果就是一个经过训练的机器学习模型 ，可用于预测。例如：算法分析历史建筑数据（成本、时间、设施特征）以创建预测模型的过程。

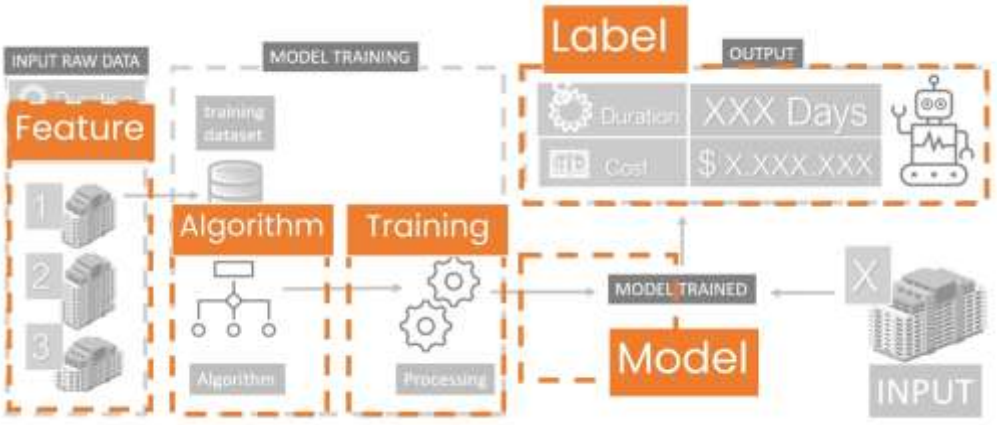


图 9.2-16 ML 使用标签和属性创建模型，这些模型通过算法对数据进行训练，以预测结果。

机器学习 并非孤立存在，而是包括统计学、数据库、数据挖掘、模式识别、大数据分析和人工智能在内的更广泛分析学科生态系统的一部分。图 9.2-17 展示了这些领域如何相互重叠和互补，为现代决策和自动化系统提供了一个全面的框架。

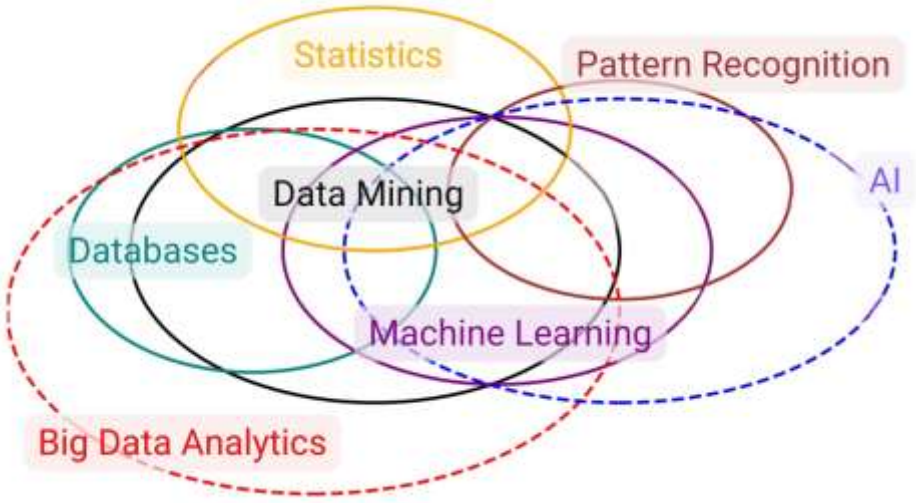
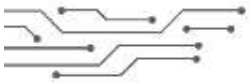


图 9.2-17 数据分析不同领域之间的关系：统计学、机器学习、人工智能、大数据、模式识别和数据挖掘。

> 机器学习 的主要目标是赋予计算机在没有人类干预或帮助的情况下自动学习知识的能力，并据此调整自己的行动 [159]。

因此，在未来，人类的角色将仅仅是为机器提供认知能力--他们将设定条件、权重和参数，而机器学习模型将完成剩下的工作。

下一章我们将讨论算法应用的具体实例。我们将使用真实表格和简化模型来展示如何逐步建立预测。



章 9.3.

利用机器学习预测成本和期限

使用机器学习 查找项目成本和进度的示例

施工时间和成本估算是建筑公司活动的关键流程之一。传统上，这种估算是由专家根据经验、参考书和监管数据库做出的。然而，随着数字化转型和数据可用性的不断提高，现在可以使用机器学习（ML）模型 来提高此类估算的准确性和自动化程度。

将机器学习 引入建筑成本和时间的计算过程，不仅可以提高规划效率，还可以成为将智能模型融入其他业务流程（从风险管理到优化物流和采购）的起点。

> 能够快速确定一个项目的建设时间和总成本是非常重要的。自建筑业诞生以来，这些有关项目时间和成本的问题一直是客户和建筑公司最关心的问题。

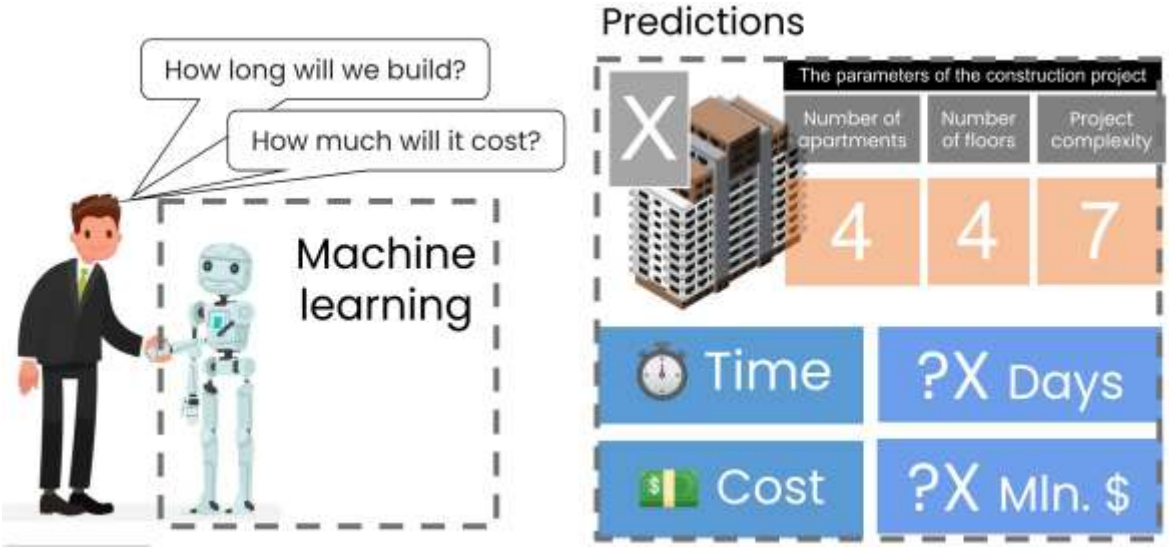


图 9.3-1 在建筑项目中，估算施工时间和成本的速度和质量是成功的关键因素。

在下面的示例中，我们将从过去的项目中提取关键数据并用于开发机器学习模型 ，这样我们就可以使用该模型来估算带有新参数的新建筑项目的成本和时间安排 （图 9.3-1）。

考虑具有三个关键属性的三个项目：单位数（为便于可视化，100 个单位等同于 10 个单位）、楼层数和建

筑复杂性的条件度量（1 到 10 分），其中 10 分是复杂性的最高分。在机器学习中，将 100 转换为 10 或将 50 转换为 5 等数值的转换和简化过程称为 "归一化"。

> 机器学习中的归一化是将不同的数字数据归入一个共同的尺度，以方便处理和分析的过程。当数据具有不同的尺度和测量单位时，这一过程尤为重要。

假设第一个项目（图 9.3-2）有 50 个单位（标准化后为 5 个），7 层楼，复杂度得分为 2，这意味着施工相对简单。第二个项目已有 80 个单位，9 层，设计相对复杂。在这种情况下，第一栋和第二栋公寓楼的施工时间分别为 270 天和 330 天，项目总成本分别为 450 万美元和 580 万美元。



图 9.3-2 一组过去的项目实例，用于估算未来 X 项目的时间和成本。

在为此类数据建立机器学习模型 时，主要任务是确定预测的关键属性（或标签），在本例中就是建筑时间和成本。通过一个小型数据集，我们将使用以前的建筑项目信息来规划新项目：使用机器学习算法 ，我们必须根据新项目的给定属性（如 40 个单位、4 层楼和相对较高的项目复杂度 7）来预测新项目 X 的成本和施工工期（图 9.3-2）。在现实世界中，输入参数的数量可能会更多，从几十个到几百个不等。这些因素可能包括：建筑材料类型、气候区、承包商的资质等级、公用设施的可用性、地基类型、开工季节、工头的意见等。

> 要创建一个预测性机器学习模型，我们需要选择一种算法来创建它。在机器学习中，算法 就像一个数学配方，教计算机如何根据数据进行预测（混合正确的参数顺序）或做出决策。

要分析过去建筑项目的数据并预测未来项目的时间和成本（图 9.3-2），可以使用一种流行的机器学习算法：

- 线性回归 （线性回归）：该算法试图找到属性之间的直接关系，例如楼层数与建筑成本之间的关系。该算法的目的是找到最能描述这种关系的线性方程，从而进行预测。
- 算法 **k** 近邻算法（**k-NN**）：该算法将新项目与过去规模或复杂度相似的项目进行比较。k-NN 根据 k 个（数量）训练实例中最接近的实例对数据进行分类。在回归中，结果是 k 个近邻的平均值或中位数。
- 决策树：是一种预测建模模型，它使用树形结构根据不同的条件将数据划分为不同的子集。树的每个节点代表一个条件或问题，导致数据的进一步划分，每片叶子代表最终的预测或结果。该算法 根据不同的特征将数据分成更小的组，如首先根据故事数量，然后根据复杂程度等进行预测。

下面，让我们以线性回归和 K 近邻算法这两种常用算法为例，了解一下用于估算新项目成本的机器学习算法。

利用线性回归预测项目成本和时间

线性回归 是一种基本的数据分析算法，它根据一个变量与一个或多个其他变量的线性关系来预测该变量的值。该模型假定因变量与一个或多个自变量之间存在直接的线性关系，算法的目标就是找到这种关系。

> 线性回归简单明了，是各领域常用的工具。在处理单一变量时，线性回归就是通过数据点找到最佳拟合线。

线性回归 找到近似输入变量 X 和输出变量 Y 之间依存关系的最佳直线（红线）。通过这条直线，可以根据确定的线性关系预测新 X 值的 Y 值（图 9.3-3）。

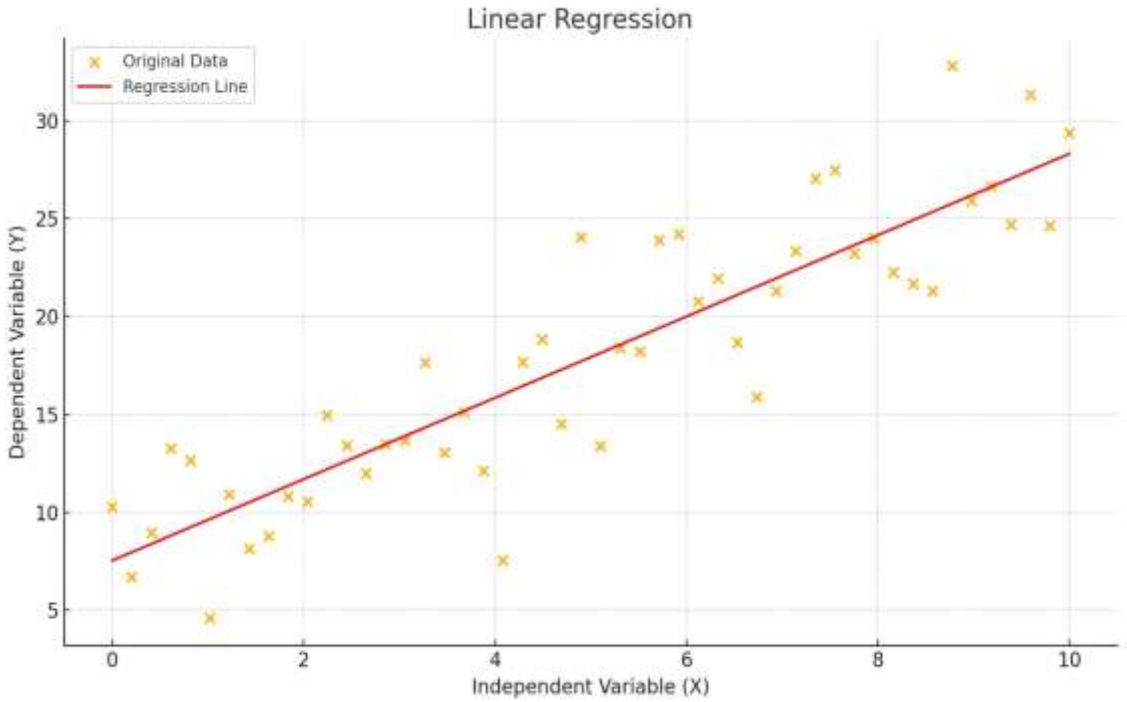


图 9.3-3 线性回归的原理是通过训练值找到最佳直线。

这条线由一个等式表示，在等式中输入一个自变量（X）的值，就会产生一个因变量（Y）的预测值。通过这一过程，可以根据已知的 X 值，利用它们之间的线性关系有效地预测 Y 值。在旧金山建筑许可数据评估（图 9.1-7）中可以看到找到这种统计平均线的例子，在该评估中，利用线性回归计算了不同类型设施的通货膨胀率。

让我们将项目数据表（上一章中的图 9.3-2）直接加载到 LLM ，并要求它为我们构建一个简单的机器学习模型。

向 LLM 聊天（CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN ）发送文本请求：

> 我们需要展示一个简单机器学习模型的构建 ，以预测实现新项目 X 的成本和时间（如附图 9.3-2 所示） ⏎

LLM 的答复：

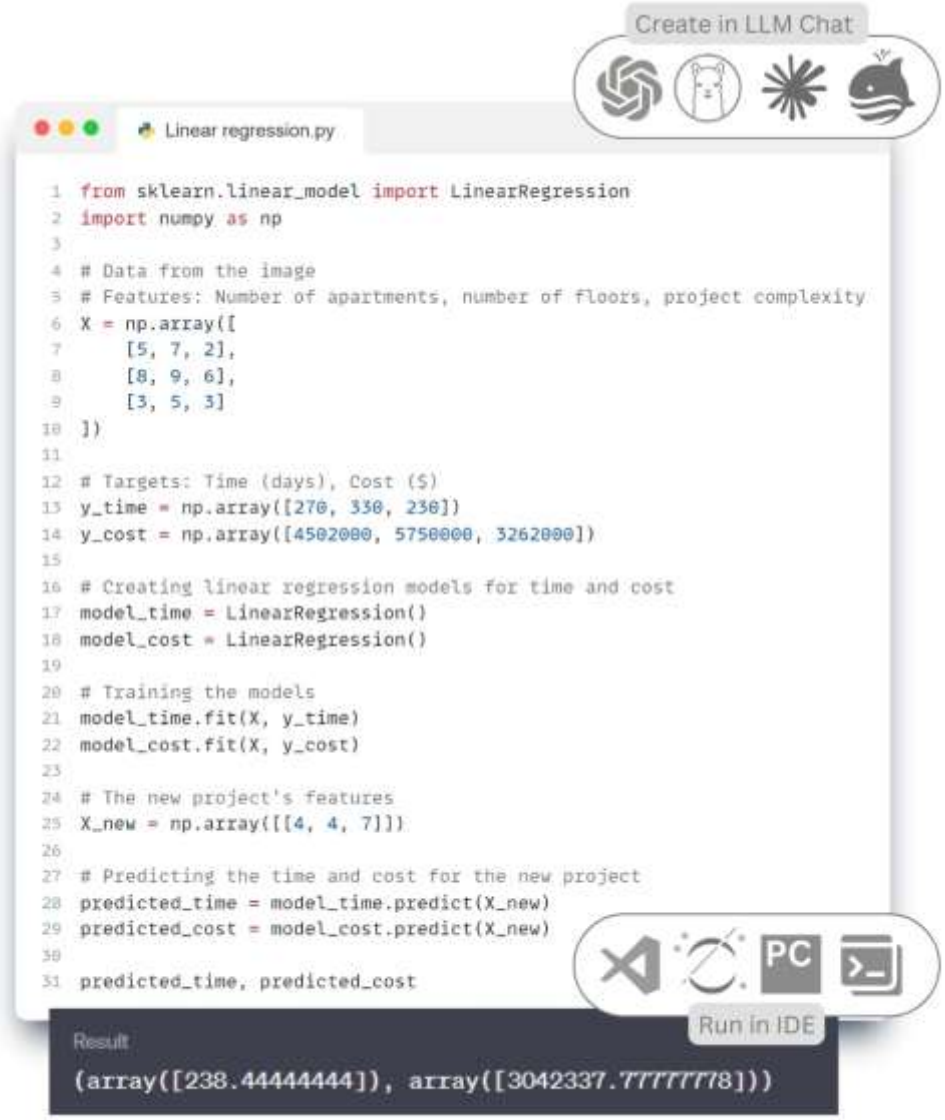


图 9.3-4 LLM 选择线性回归来创建机器学习模型，以预测项目成本和时间。

LLM 可自动识别所附图像中的表格，并将数据从可视格式转换为表格数组（图 9.3-4 - 第 6 行）。该数组被用作创建特征和标签的基础，根据这些特征和标签创建机器学习模型（图 9.3-4 - 第 17-22 行），该模型使用线性回归。

使用在 "极小 "数据集上训练的基本线性回归模型，对一个新的假定建筑项目 进行了预测，该项目被标记为

X 项目。在我们的问题中，该项目有 40 个单位，4 层楼，复杂程度为 7（图 9.3-2）。

正如基于新项目 X 的有限和少量数据集的线性回归模型所预测的那样（图 9.3-4 - 第 24-29 行）：

- 施工期约为 238 天 (238.4444444)
- 总支出约为$ 3,042,338 (3042337.777)

为了进一步探索项目成本假设，不妨尝试使用不同的机器学习算法和方法。因此，让我们使用 K-Nearest Neighbours 算法（k-NN ），基于一小组历史数据，预测新项目 X 的相同成本和时间值。

使用 **K** 近邻算法预测项目成本和时间（**k-NN)**

我们使用 K-Nearest Neighbours（k-NN ）算法作为额外的预测工具来估算新项目的成本和工期。 K-Nearest Neighbours (k-NN) 算法是一种监督机器学习（supervised machine learning）方法，可用于分类 和回归。我们之前还在向量数据库搜索中讨论过 k-NN 算法（图 8.2-2），该算法用于查找最接近的向量（如文本、图像或技术描述）。在这种方法中，每个项目都表示为多维空间中的一个点，其中每个维度都对应项目的一个 特定属性。

在我们的例子中，考虑到每个项目的三个属性，我们将它们表示为三维空间中的点（图 9.3-5）。因此，我们即将开展的项目 X 将以坐标（x=4、y=4、z=7）在该空间中定位。需要注意的是，在现实世界中，点的数量和空间的维度可能要大得多。

> K-NN（k-近邻）算法的工作原理是测量所需的项目 X 与训练数据库中的项目之间的距离。通过比较这些距离，该算法确定了最接近新项目 X 点的项目。

例如，如果原始数据集中的第二个项目（x=8，y=9，z=6）比其他项目离 X 更远（图 9.3-5），则可将其排除在进一步分析之外。因此，只能使用两个（k=2）最接近的项目进行计算，并在此基础上确定平均值。

这种方法通过邻域搜索，可以对项目之间的相似性进行评估，进而有助于根据以前实施的类似项目，对新项目的可能成本和时间安排得出结论。

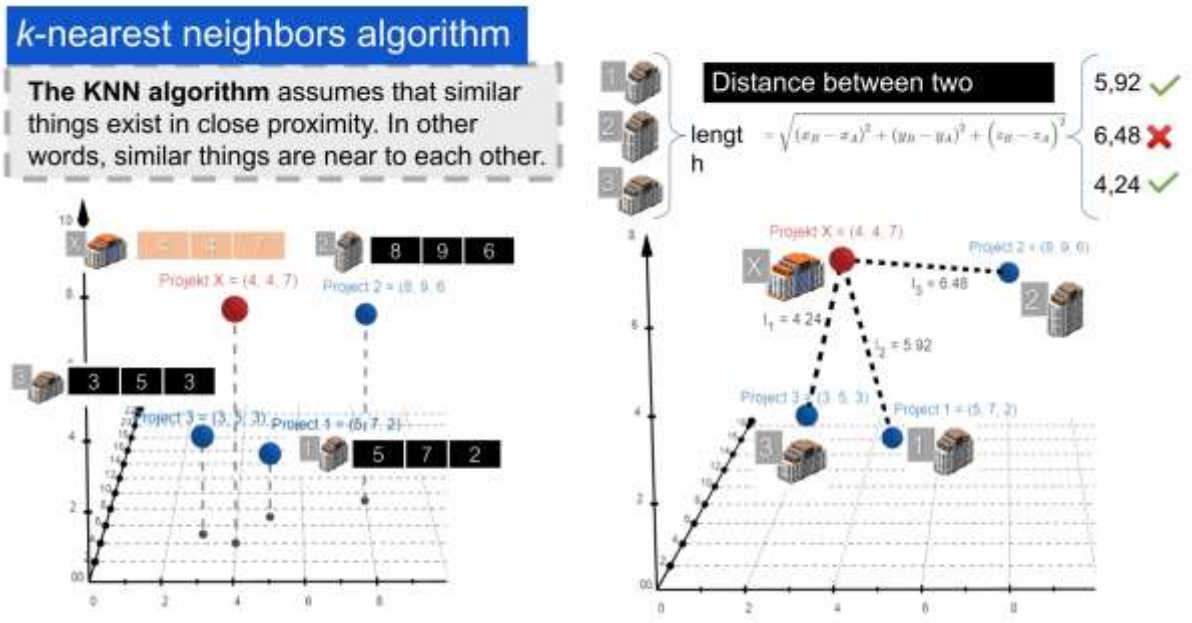


图 9.3-5 在 K-NN 算法中，项目表示为多维空间中的点，根据距离选择最近的项目进行相似性评估和预测。

k-NN 的工作涉及几个关键步骤：

- 数据准备：首先加载训练数据集和测试数据集。训练数据用于 "训练 "算法，测试数据用于检验算法的效率。
- 选择参数 **K**：选择一个数字 K，表示算法中应考虑多少个近邻（数据点）。K "的值非常重要，因为它会影响结果。
- 分类过程 和测试数据的回归：
  - 计算距离：计算测试数据中每个元素与训练数据中每个元素的距离（图 9.3-5）。可以使用不同的距离测量方法，如欧氏距离（最常用的方法）、曼哈顿距离或汉明距离。
  - 排序并选出 **K** 个近邻点：计算距离后，对距离进行排序，选出距离测试点最近的 K 个点。
  - 确定测试点的类别或值：如果是分类任务 ，则根据所选 K 个邻域中最常出现的类别确定测试点的类别。如果是回归任务，则计算 K 个邻域值的平均值（或其他中心倾向度量）。
- 完成流程：一旦所有测试数据都已分类或预测，流程即告完成。

k 近邻算法（k-NN ）在许多实际应用中都很有效，是机器学习专家的主要工具之一。这种算法因其简单高效而广受欢迎，尤其是在数据之间的关系易于解释的任务中。

在我们的例子中，应用 K 近邻算法后，（从我们的小样本中）找出了与项目 X 距离最短的两个项目（图 9.3-5）。在这些项目的基础上，算法确定了它们的价格和施工工期的平均值。经过分析（图 9.3-6），该算法通过最近邻的平均值得出结论，项目 X 的成本约为$ 380 万美元，完工时间约为 250 天。

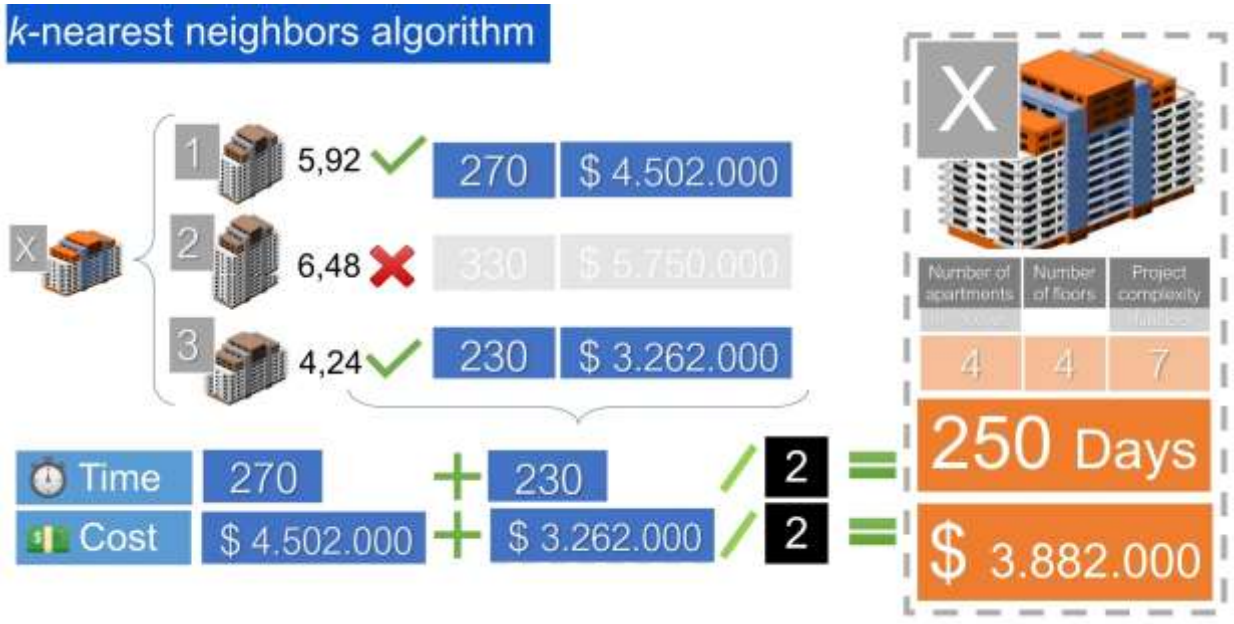


图 9.3-6 K 近邻算法通过分析样本中最接近的两个项目来确定项目 X 的成本和进度。

k-Nearest Neighbors（k-NN ）算法在 分类和回归任务（如推荐系统）中尤其流行，该算法用于根据与特定用户兴趣相似的偏好推荐产品或内容。此外，k-NN 还广泛应用于医疗诊断，根据病人症状对 疾病类型进行分类，模式识别，以及金融领域对客户信用度的评估。

即使数据有限，机器学习模型 也能提供有用的预测，并显著提高施工项目管理的分析能力。随着历史数据的扩展和清理，有可能转向更复杂的模型--例如，将建筑类型、地点、开工季节和其他因素考虑在内。

我们的简化任务使用三个属性在三维空间中进行可视化，但实际项目平均包括数百或数千个属性（参见 "基

于 CAD 的大数据 (BIM) "一章中的数据集），这大大增加了空间维度和以向量表示项目的复杂性（图 9.3-7）。

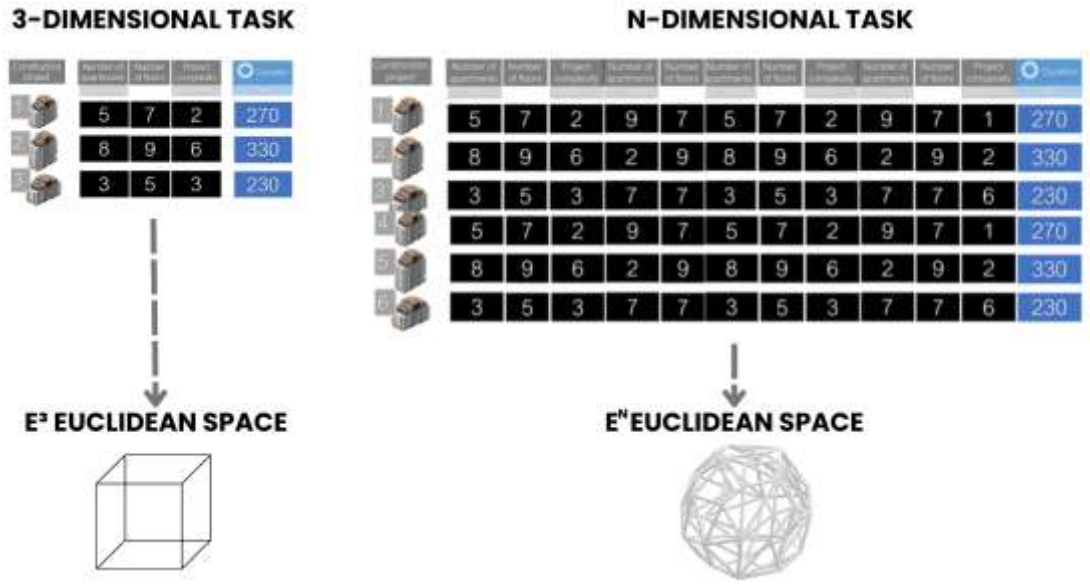


图 9.3-7 在简化的示例中，三维 可视化使用了三个属性，而实际项目则有更多属性。

项目 X 有 40 个单位，4 层楼，复杂程度为 7 级，对同一数据集采用不同算法得出的预测值各不相同。 线性回归算法预测完工时间为 238 天，成本为$ 3,042,338 （图 9.3-4），而 k-NN 算法 预测完工时间为 250 天，成本为$ 3,882,000 （图 9.3-6）。

使用机器学习模型 ，预测的准确性直接取决于输入数据的数量和质量。参与训练的项目越多，其特征（属性）和结果（标签）越完整、越准确，就越有可能获得误差值最小的可靠预测。

数据预处理技术在这一过程中发挥着重要作用，其中包括

- 标准化，使特征达到统一的比例；
- 异常值检测和消除，消除模型失真；
- 对分类属性进行编码，以便操作文本数据；
- 填补缺失值，增强模型的稳健性。

此外，还使用交叉验证方法来评估模型的通用性及其对新数据集的稳健性，以检测过度拟合并提高预测的可

靠性。

> 混沌是一种有待破译的秩序 [160]。
>
> - 何塞-萨拉马戈，"替身"

即使在你看来，你的任务的混乱无法用形式来描述，但你应该知道，世界上的任何事件，尤其是建筑过程都受数学规律的制约，这可能需要计算值的支持，而不是通过严格的公式，而是借助统计数据和历史数据。

传统的成本核算部门估算和机器学习模型 都不可避免地面临不确定性和潜在的误差来源。不过，如果有足够的高质量数据，机器学习模型 可以显示出与专家判断相当，有时甚至更高的预测准确性。

> 机器学习很可能成为一种可靠的辅助分析工具，它可以：完善计算、提出替代方案、识别项目参数之间隐藏的依赖关系。这些模型虽然不能说是万能的，但很快就会在计算和决策过程中占据重要地位。机器学习技术不会排斥工程师、估算师和分析师的参与，相反，还将通过提供基于历史数据的额外观点来扩展他们的能力。

如果适当地融入建筑公司的业务流程，机器学习有可能成为管理决策支持系统中的一个重要元素--不是取代人类，而是人类专业直觉和工程逻辑的延伸。

下一步：从存储到分析和预测

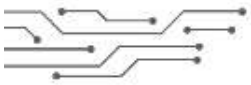
现代数据工作方法正在开始改变建筑行业的决策方式。从直观评估转向客观数据分析，不仅提高了准确性，还为优化流程带来了新的机遇。在总结这部分内容时，值得强调的是有助于您在日常工作中应用所讨论方法的主要实际步骤：

- 建立可持续的储存基础设施
  - 尝试将不同的文件和项目数据合并成单一的表格模型，将关键信息汇总到单一的数据框架中，以便进一步分析

- ☐ 使用高效的数据存储格式，例如 Apache Parquet 等列式格式，而不是 CSV 或 XLSX 格式，尤其是那些将来可能用于训练机器学习模型的数据集。
- ☐ 建立数据版本系统，以跟踪整个项目中的变化

- 实施分析和自动化工具
  - ☐ 开始分析历史项目数据（通过文档、模型、估算），以确定模式、趋势和异常情况
  - ☐ 掌握 ETL 流程（提取, 转换, 加载），自动加载和准备数据
  - ☐ 学习如何使用各种免费的 Python 可视化库将关键指标可视化
  - ☐ 开始应用统计方法和随机抽样，以得出具有代表性和可重复的分析结果
- 数据处理日益成熟
  - ☐ 通过简单明了的示例（如泰坦尼克号数据集）学习一些基本的机器学习算法
  - ☐ 分析当前流程，确定在哪些方面可以从僵化的因果逻辑转向预测和估算的统计方法
  - ☐ 开始将数据视为战略资产而非副产品：围绕数据模型而非特定软件解决方案建立决策流程

已经意识到数据价值的建筑公司正在进入一个新的发展阶段，在这个阶段，决定竞争优势的不是资源数量，而是基于分析的决策速度。



印刷版带来的最大便利

您现在持有的是《数据驱动的建筑》的免费电子版。为了更方便地工作和快速查阅资料，我们建议您关注印刷版：



- 随手可得：印刷版书籍将成为可靠的工作工具，让您在任何工作环境中都能快速找到并使用正确的可视化图示和图表
- 高质量插图：印刷版中的所有图像和图形均以最高质量呈现
- 快速获取信息：导航简便，可做笔记、书签，随时随地使用图书。

通过购买本书的完整印刷版，您将获得一个舒适、高效地处理信息的便捷工具：能够在日常工作中快速使用视觉材料，快速查找必要的方案并做笔记。此外，您的购买还支持了开放知识的传播。

订购本书印刷版，请访问：datadrivenconstruction.io/books



X 部分

数字时代的建筑业。机遇与挑战

最后第十部分是对数字化转型时代建筑业未来的全面展望。该部分分析了从因果分析到大数据关联工作的转变。该部分将美术的发展与建筑业数据工作的发展相提并论，展示了建筑业如何从细节控制转向对流程的整体理解。探讨了建筑行业的 "uberisation "概念，即数据透明度和自动计算可从根本上改变传统的商业模式，消除对中介的需求，减少投机机会。详细讨论了尚未解决的问题，如 元素的通用分类，以便建筑公司有时间适应新环境。本部分最后提出了制定数字化转型战略的具体建议，包括分析薄弱环节和扩展服务，以便在不断变化的行业中保持竞争力。

章 10.1.

生存战略：建立竞争优势

关联而非计算：建筑分析的未来

由于信息的快速数字化（图 1.1-5），现代建筑业正在经历一场根本性的变革，数据不仅成为一种工具，而且成为一种战略资产，可以从根本上改变传统的项目和业务管理方法。

数千年来，建筑活动一直依赖于确定性方法--精确计算、细部设计和对参数的严格控制。公元一世纪，罗马工程师将数学原理应用于水渠和桥梁的建造。在中世纪，建筑师们努力追求哥特式大教堂的理想比例，而在 20 世纪的工业化进程中，标准化的规范和规则体系形成，并成为大规模建筑的基础

如今，发展的矢量正在从寻找严格的因果关系转向概率分析、寻找相关性和隐藏模式。该行业正在进入一个新阶段--数据正在成为一种关键资源，基于数据的分析 正在取代直观和局部优化的方法。



图 10.1-1 建筑数据的隐藏潜力：公司现有的计算结果只是可供管理层分析的冰山一角。

> 公司的信息系统就像一座冰山（图 10.1-1）：公司管理层只能看到数据潜力的一小部分，而主要价值却隐藏在深处。重要的是，不仅要评估数据的当前用途，还要评估数据在未来释放的机会。只有那些学会如何从数据中提取隐藏模式和创造新见解的公司，才能创造可持续的竞争优势

寻找隐藏的模式和理解数据并不仅仅是处理数字，而是一个创造性的过程，需要抽象思维和洞察不同元素背后的全貌的能力。从这个意义上说，数据工作的演变可以比作视觉艺术的演变（图 10.1-2）。

建筑业的发展与美术的进步极为相似。在这两种情况下，人类都从原始方法发展到了复杂的可视化和分析技术。史前时代，人们使用洞穴图和原始工具来解决日常工作。在中世纪和文艺复兴时期，建筑和艺术的复杂程度显著提高。到了中世纪早期，建筑工具已经从简单的斧头发展到广泛的工具包，象征着技术知识的增长。

写实主义时代是视觉艺术的第一次革命：艺术家们学会了如何再现最微小的细节，达到最大的可信度。在建筑领域，与这一时期相对应的是精确的工程技术、详细的图纸和严格规范的计算，它们成为几个世纪以来设计实践的基础。

> 后来，印象派改变了人们对艺术现实的看法：艺术家们开始捕捉情绪、光线和动态，而不是照本宣科地渲染形式，旨在反映整体印象而非绝对准确。同样，建筑分析领域的机器学习也正在从僵化的逻辑模型转向模式识别和概率模式，从而使您能够 "看到 "数据中隐藏的依赖关系，而这在传统分析中是无法实现的。这种方法呼应了包豪斯的极简主义和功能性思想，即意义（功能）比形式更重要。包豪斯试图去除多余的东西，为了清晰、实用和大众化而放弃装饰。事物必须易懂、有用，没有多余的东西--美学诞生于设计和目的的逻辑之中。

随着 19 世纪末摄影术的出现，艺术获得了一种新的工具，能够以前所未有的精确度捕捉现实，颠覆了人们对视觉艺术的态度。同样，在建筑领域，21 世纪的工业革命正在引领机器人技术、激光、物联网 、RFID 以及互联建筑等概念的应用，其中，单个参数的收集已经发展成为对建筑工地的可扩展智能捕捉。

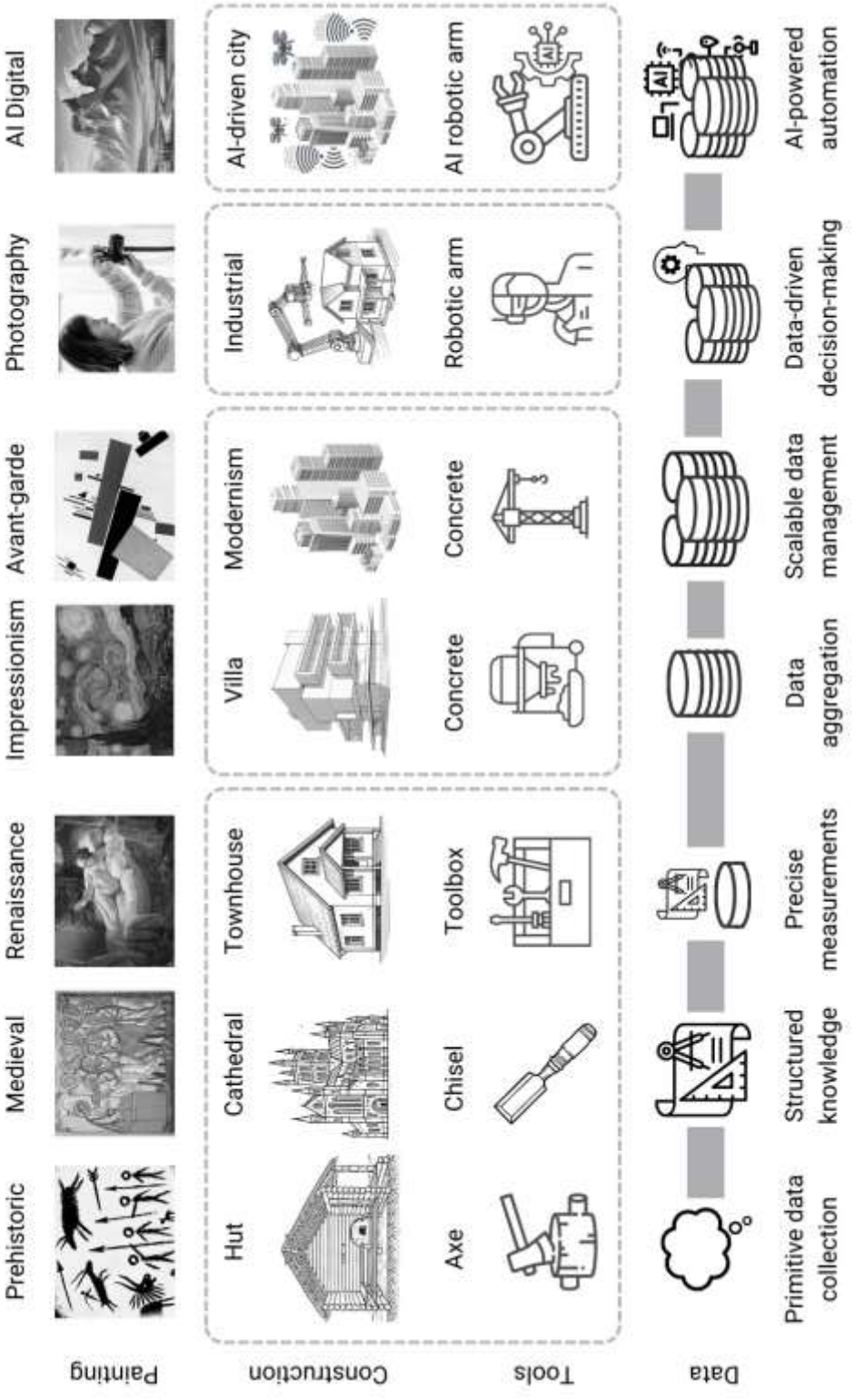


图 10.1-2 美术演变时代与建筑行业数据工作方法的发展相一致。

如今，随着人工智能工具 和 LLM 的出现，视觉艺术正经历着一场重新思考，建筑行业也正经历着另一次飞跃：由人工智能（AI ）驱动的智能系统、LLM 聊天工具允许在最少人工干预的情况下预测、优化和生成解决方案。

数据在设计和管理中的作用发生了翻天覆地的变化。过去，知识是通过口头和经验传播的，就像 19 世纪以前通过手绘图片捕捉现实一样，而今天的重点则是对建筑 "图片 "进行完整的数字化捕捉。在机器学习算法的帮助下，这种数字图像被转化为对施工现实的印象式再现--不是精确的复制品，而是对施工过程的概括性、概率性理解。

> 我们正在迅速步入这样一个时代：建筑的设计、建造和运营过程将不仅仅是增强型的，而且在很大程度上将由人工智能系统。正如现代数字艺术的创作无需画笔--使用文字提示和生成模型--未来的建筑和工程解决方案将由用户设置的关键询问和参数来塑造。

在 21 世纪，数据的获取、解读和分析质量对于项目的成功变得不可或缺。数据的价值不是由其数量决定的，而是由专家分析、验证并将其转化为行动的能力决定的。

数据驱动的建筑方法：基础设施的新水平

在人类历史上，每一次这样的技术飞跃都会给经济和社会带来根本性的变革。今天，我们正在目睹一场新的变革浪潮，其规模堪比 19 世纪的工业革命。然而，一百年前，变革的主要驱动力是机械力量和能源技术，而现在则是数据和人工智能。

机器学习 、LLM 和人工智能 代理改变了应用程序的本质，使传统的软件堆栈（将在本书第二部分讨论）变得没有必要（图 2.2-3）。所有数据逻辑都以人工智能代理为中心，而不是硬编码的业务规则（图 2.2-4）。

> 在数据时代，传统的应用观点正在发生根本性的转变。我们正在走向这样一种模式：庞大的模块化企业系统将不可避免地让位于开放、轻量级的定制解决方案。

> 未来，将只保留底层数据结构，所有与之的交互都将通过直接与数据库工作的代理进行。我真的相信，整个应用程序栈将会消失，因为当人工智能直接与底层数据库交互时，根本就不需要应用程序栈。我的整个职业生涯都是在 SaaS 工作--建立公司，在其中工作，说实话，我现在可能不会推出新的 SaaS 业务。我现在可能也不会投资 SaaS 公司。形势太不确定了。这并不是说未来不会有软件公司，只是它们的面貌会大不一样。未来的系统将是数据库，其业务逻辑将被引入[人工智能] 代理。这些代理将同时处理多个数据存储库，而不局限于单一数据库。所有逻辑都将进入人工智能层 [46]。
>
> - 马修-伯曼（Matthew Berman），Forward Future 首席执行官

新模式的关键区别在于最大限度地减少技术压舱物。取而代之的将是灵活、开放和可快速定制的模块，这些模块可以 "活 "在数据流中（图 7.4-1 - Apache Airflow，NiFi）。未来流程管理的架构设想使用微型应用程序--小巧、专用的工具，从根本上有别于大规模和封闭的 ERP、PMIS 、CDE 、CAFM 系统。新的代理将尽可能具有适应性、集成性和业务特定性（如低代码/无代码 图 7.4-6）。

> 所有业务逻辑都将交给这些[AI] 代理，这些代理将在多个存储库上执行 CRUD [创建、读取、更新和删除]操作，这意味着它们不会区分使用的是哪个后端。它们将更新多个数据库，而所有逻辑都将在所谓的人工智能 层完成。一旦所有逻辑都在人工智能层，人们就会开始更换后端。我们已经看到 Dynamics 后端和代理的使用占据了相当高的市场份额，我们将积极地朝这个方向发展，努力将这一切结合在一起。无论是在客户服务还是其他领域，比如不仅是 CRM ，还有我们的财务和运营解决方案。因为人们需要更多人工智能驱动的业务应用，在这些应用中，逻辑层可以由人工智能和人工智能代理驱动。[...].最让我兴奋的事情之一是 Excel 与 Python ，这相当于 GitHub 与 Copilot。也就是说，我们所做的是：现在你有了 Excel，就应该打开它，运行 Copilot，然后开始使用它。它不再只是了解可用的数字--它会自己制定计划。就像 GitHub Copilot 工作区创建计划然后执行一样，它就像数据分析师 ，使用 Excel 作为工具，将行和列可视化以进行分析。因此，Copilot 将 Excel 作为一种工具来使用，因为它可以生成数据并拥有 Python 解释器，所以它拥有 Excel 的所有功能。
>
> - 微软首席执行官萨提亚-纳德拉（Satya Nadella），2024 年 12 月接受 BG2 频道采访。[28]

我们正在目睹的办公应用逻辑的转变--从模块化的封闭系统转变为直接与开放数据打交道的人工智能代理--只是一个更大进程的一部分。这不仅仅是界面或软件架构的变化：这些变化将影响工作组织、决策和业务管理的基本原则。在建筑领域，这将导致一种数据驱动的逻辑，其中数据将成为从设计到资源管理和施工监控等流程的核心。

下一代数字办公室：人工智能 如何改变工作空间

近一个世纪前，人类已经经历了类似的技术革命。从蒸汽机到电动机的转变历时四十多年，但最终催化了前所未有的生产力增长--这主要得益于能源能力的分散化和新解决方案的灵活性。这一转变不仅改变了历史进程，使大部分人口从农村转移到城市，还奠定了现代经济的基础。技术的历史就是从体力劳动到自动化和智能系统的历程。正如拖拉机取代了大量的耕作机一样，现代数字技术也正在取代传统的以办公室为基础的建筑管理方法（图 10.1-3）。早在 20 世纪初，世界上大多数人都是靠手工耕作土地，直到 20 世纪 30 年代才开始使用机器和拖拉机进行机械化劳动。

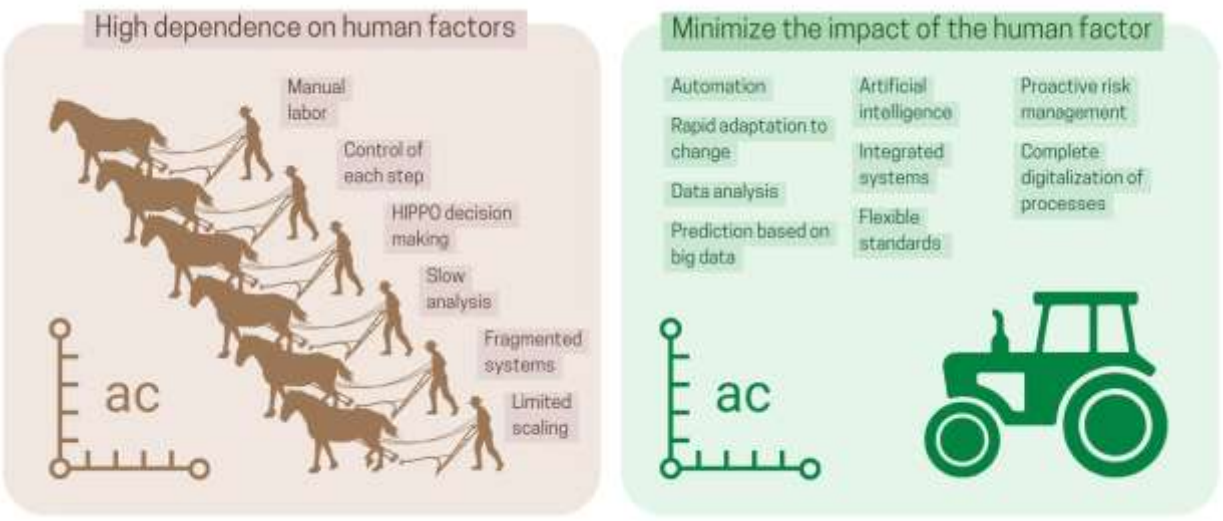


图 10.1-3 正如 20 世纪初拖拉机取代了几十个人一样，机器学习也将在 21 世纪取代传统的业务和项目管理方法。

正如一百年前人类从使用原始工具耕种单块土地到使用机械进行大规模耕作一样，今天我们正在从处理分散的信息 "孤岛 "过渡到在强大的 "拖拉机"--ETL--管道和人工智能算法的帮助下处理数据阵列。

我们正处于类似飞跃的边缘--但却是在数字层面：从传统的手工业务管理到数据驱动模式。

要实现全面的数据驱动型架构，需要时间、投资和组织方面的努力。但这条道路不仅能逐步改善，还能实现质的飞跃，提高建筑流程的效率、透明度和可管理性。所有这一切都需要系统地引入数字化工具，并摒弃过时的业务实践。

> 任务参数化、ETL、LLM 、物联网组件 、RFID 、标记化、大数据和机器学习将把传统建筑转变为数据驱动的 建筑，项目和建筑业务的每个细节都将由数据控制和优化。

过去，分析信息需要花费数千个工时。现在，这些任务由算法和龙 8 国际娱乐城来完成，它们通过提示将不同的数据集转化为战略资源。在技术领域，农业也正在发生同样的事情：我们正在从锄头转向自动化农业综合企业。同样，建筑业的办公室工作也从 Excel 文件和人工汇总转变为智能系统，在这个系统中，数据被收集、清理、结构化并转化为洞察力。

今天，企业就应该开始通过高质量的数据收集和信息结构构建来 "耕耘 "信息领域，并通过清理和规范化工具来 "施肥"，然后以预测分析和自动化解决方案的形式来 "收获"。如果说一个现代农民用一台机器就能取

代一百个耕作者，那么智能算法就能让员工摆脱日常工作，转而扮演信息流战略管理者的角色。

然而，重要的是要明白，创建一个真正以数据为导向的 组织并不是一个一蹴而就的过程。这是一个长期的战略方向，类似于创建一个新的场地，种植一片新的系统森林（图 1.2-5），在这个生态系统中，每一棵 "树 "都是一个独立的过程、能力或工具，需要时间来成长和发展。与真正的森林一样，成功不仅取决于种植材料的质量（技术），还取决于土壤（企业文化）、气候（商业环境）和护理（系统方法）。

企业将不再能够完全依赖封闭的、开箱即用的解决方案。与以往的技术发展阶段不同，目前向开放数据访问、人工智能和开源技术的过渡不太可能得到大型供应商的支持，因为这直接威胁到他们既有的商业模式和核心收入来源。

哈佛商学院的研究报告[40]（在第四和第五次技术革命章节中已经讨论过）显示，为所有公司从头开始创建最常用的开源解决方案的成本约为 41.5 亿美元。然而，如果我们设想每家公司在无法使用现有开源工具的情况下开发自己的替代方案，而这正是过去几十年一直在发生的事情，那么总的商业成本可能会达到 8.8 万亿美元之巨--这是 软件市场可以估价的非理性需求的价格。

技术进步将不可避免地导致对既有商业模式的重新思考。过去，企业可以从复杂、不透明的流程和封闭的数据中赚钱，而随着人工智能 和分析 的发展，这种方法越来越不可行。

> 由于数据和工具获取的民主化，传统的软件销售市场可能会大幅萎缩。但与此同时，一个新的市场也将成长起来，那就是数字专业技术、定制、集成和解决方案设计市场。在这里，价值将不是来自许可证的销售，而是来自建立灵活、开放和适应性强的数字化流程的能力。正如电气化和拖拉机的出现催生了新的行业一样，大数据、人工智能 和 LLM 的应用也将为建筑行业的业务开辟全新的前景，这不仅需要技术投资，还需要思维方式、流程和组织结构的深刻变革。那些意识到这一点并从今天开始行动的公司和专业人士将成为未来的领导者。

在一个开放数据正在成为重要资产的世界里，信息的可用性将改变游戏规则。投资者、客户和监管机构对透明度的要求将越来越高，而机器学习算法 将能够自动识别估算、时间表和成本中的差异。这就为数字化转型的新阶段创造了条件，并逐渐将我们引向建筑行业的 "uberisation "。

开放数据和 **Uber** 化对现有建筑企业构成威胁

建筑工程正在成为一个信息管理过程。数据越准确、质量越高、越完整，建筑物的设计、计算、成本估算、建造和运营就越有效率。未来，关键资源将不再是起重机、混凝土和钢筋，而是收集、分析和使用信息的能力。

> 未来，建筑公司的客户--投资者和建筑融资客户--将不可避免地利用开放数据和历史数据分析 的价值。这将带来自动计算项目时间表和成本的机会，而无需让建筑公司参与成本核算问题 ，这将有助于控制成本并更快地识别多余成本。

想象一下，在一个建筑工地上，激光扫描仪、四旋翼飞行器和摄影测量系统会实时收集有关混凝土用量的精确数据。这些信息会自动转换成带有元数据的简单平面 MESH -模型，绕过繁琐的 CAD (BIM) 系统，不依赖于复杂的几何内核 、ERP 或 PMIS。从施工现场收集的这些数据被集中传输到一个单一的结构化存储库中，供客户进行独立分析，其中上载了来自不同建筑商店的真实价格和各种参数，从信贷融资率到动态变化的因素，如天气条件、建筑材料的证券交易所报价、物流关税和劳动力价格的季节性波动统计。在这种情况下，设计与实际材料量之间的任何差异都会立即显现出来，从而使设计阶段和项目移交时的估算无法被操纵。因此，施工过程的透明度不是通过一大批监理和管理人员来实现的，而是通过客观的数字数据来实现的，这将最大限度地减少人为因素和投机取巧的可能性。

未来，这种数据控制工作将由客户方的数据管理人员来完成（图 1.2-4 CQMS manager）。这一点在计算和项目估算方面尤为明显：过去需要整个部门的估算人员，而明天就会有机器学习 和预测工具，为建筑公司设定价格限制。

> 鉴于[建筑]行业的分散性，大部分系统和子系统都由中小企业提供，因此数字化战略必须来自客户。客户必须创造条件和机制，释放供应链的数字化能力[20]。
>
> - 埃森哲公司的安德鲁-戴维斯和朱利亚诺-德尼科尔 "通过资本项目创造更多价值"

数据的这种开放性和透明度对建筑公司构成了威胁，因为它们习惯于从不透明性的流程和混乱的报告中赚钱

，投机和隐性成本可能隐藏在复杂、封闭的格式和模块化的专有数据平台背后。因此，与推广开源解决方案的供应商一样，建筑公司不太可能有兴趣在其业务流程中全面实施开放数据。如果数据可用且便于客户处理，就可以自动检查，从而消除高估工程量和操纵估算的可能性。

> 根据世界经济论坛的报告《塑造建筑业的未来》（2016 年）[5]，该行业面临的主要挑战之一仍然是客户的被动角色。然而，从早期规划到选择可持续的互动模式，再到监督绩效，客户应该对项目的结果承担更大的责任。没有项目业主的积极参与，就不可能实现建筑业的系统转型。

在过去 20 年里，对数量和成本计算的失控已经改变了其他行业，使客户可以不经中介直接，从而阻碍了他们的目标。数字化和数据透明化改变了许多传统商业模式，如 Uber 的出现改变了出租车司机（图 10.1-4），Airbnb 的出现改变了酒店经营者，亚马逊的崛起改变了零售商和商店，新银行和去中心化金融科技生态系统的崛起改变了银行，在这些行业中，信息的直接获取以及时间和成本计算的自动化大大降低了中介的作用。

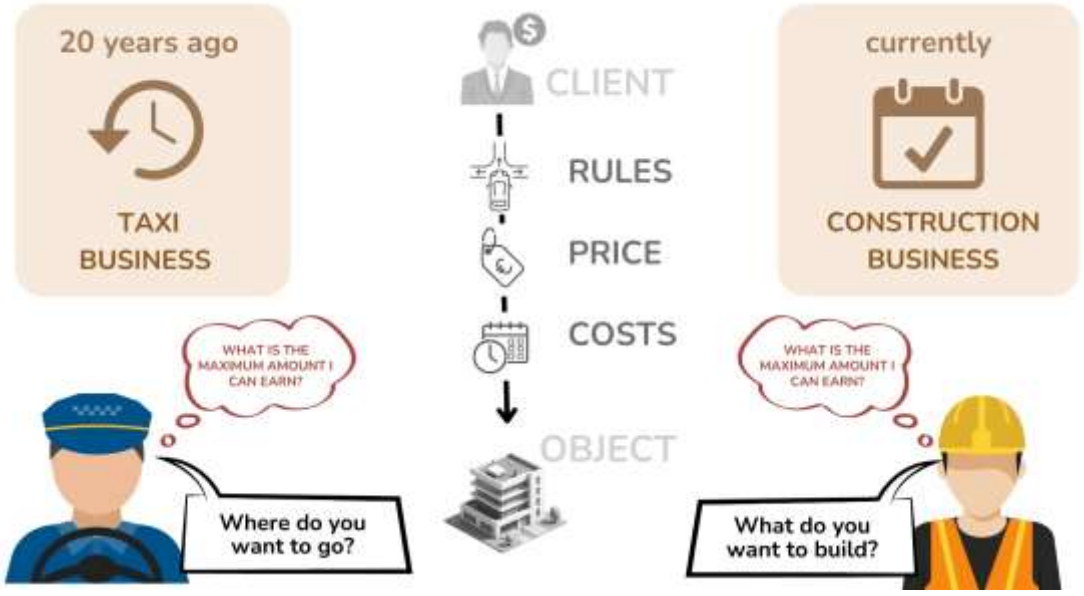


图 10.1-4 建筑业将面临出租车司机、酒店经营者和零售商 10 年前不得不面对的 uber 化。

数据访问和数据处理工具的民主化进程不可避免，随着时间的推移，所有项目组成部分的开放数据 将成为客户的要求和新标准。因此，引入开放格式和透明计算的问题将由投资者 、客户、银行和私募基金（私募基金 ）--这些人是所建对象的最终用户，随后将对对象进行数十年的运营。

> 主要投资者、客户和银行已经开始要求建筑行业保持透明。根据埃森哲公司的研究报告《通过资本项目创造更多价值》（2020 年）[20]，透明可靠的数据正成为建筑业投资决策的决定性因素。正如专家们所指出的，没有透明度，尤其是在危机时刻，就不可能进行可信和有效的项目管理。此外，资产所有者和承包商越来越倾向于签订鼓励数据共享和合作分析的合同，这反映出 、银行和监管机构对问责制和透明度的要求越来越高。

> 未来，投资者、客户从想法到建筑成品的过程，将类似于自动驾驶汽车的旅行--没有建筑公司形式的司机，有望摆脱投机和不确定性。

开放数据和自动化时代将不可避免地改变建筑业，正如它已经改变了银行业、商业、农业和物流业一样。在这些行业中，中介机构的作用和传统的经营方式正在让位于自动化和机器人化，没有了不合理加价和投机的空间。

人类所有经济活动中的数据和流程与建筑业专业人员所要处理的并无不同。从长远来看，如今通过制定价格和服务质量标准来主导市场的建筑公司，可能会失去其作为客户与建筑项目之间重要中介的角色。

uberisation 的未决问题 作为利用时间进行转型的最后机会

让我们回到建筑行业的现实。虽然自动驾驶汽车、分散式金融系统和基于人工智能的解决方案正在某些经济领域兴起 ，但相当一部分建筑公司仍然是纸质组织，其关键决策更多是依靠专家个人的直觉和经验。

在这种模式下，现代建筑公司可以比作一家有 20 年历史的出租车公司，它控制着资源、路线和交付时间，并对 "旅程 "的时间和成本负责--从项目创意（物流和安装过程）到项目交付。正如 GPS（在建筑物联网 、RFID 中）和时间/成本计算中的机器学习算法曾一度改变了运输领域一样，数据、算法和人工智能 代理也能改变建筑管理--从直观评估到预测、指导模型。在过去的 20 年里，金融、农业、零售和物流等许多行业都因数据不透明而逐渐丧失了投机能力。价格、交付成本或金融交易都是自动计算出来的，而且在数字平台上只需几秒钟就能统计出来。

展望未来，建筑公司需要认识到，数据获取的民主化以及分析数据的工具将打破估算项目成本和时间的传统方法，并消除对不透明的数量和价格数据进行推测的机会。

> 就像在没有司机干预的规范道路上驾驶一样，未来的施工过程将越来越像一个 "Uber 化 "的系统--自动估算时间和成本，透明地安排任务路线，尽量减少对人为因素的依赖。这将改变从构思到实现的 "旅程 "的本质，使其更具可预测性、可管理性和数据驱动性。

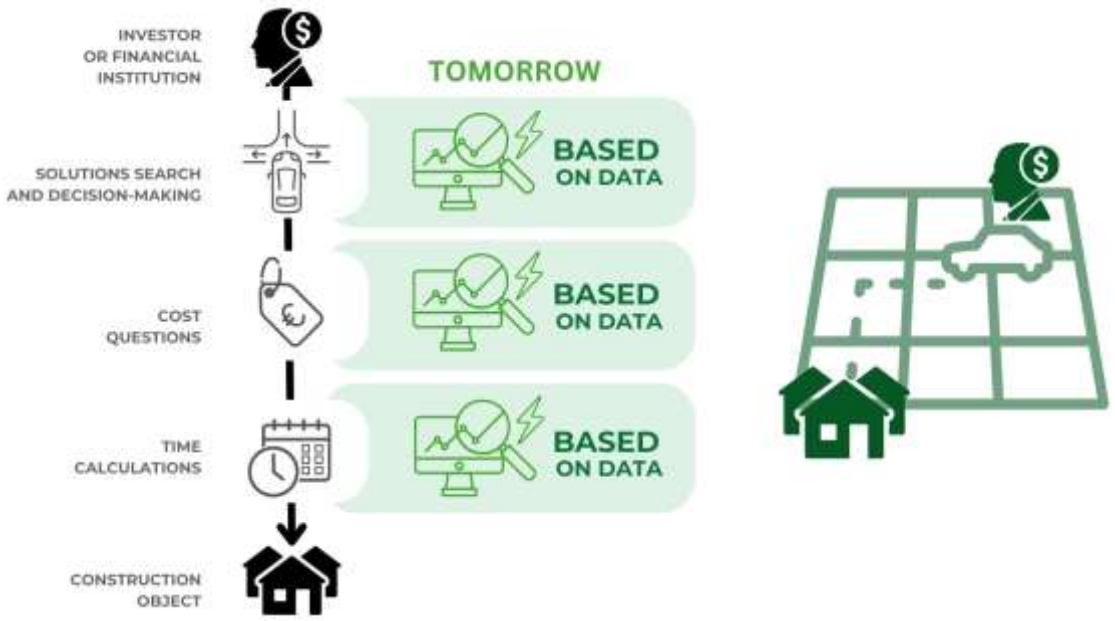


图 10.1-5 施工期间的成本和 "旅行 "时间将通过机器学习 和统计工具来确定。

世界上几乎所有国家都在逐步引入新的法规和要求，规定必须将 CAD- (BIM)-模型传输给客户或为建筑项目提供融资的银行，这样客户和顾客就有机会独立确保成本和范围计算的透明度。这对大客户和投资者尤其重要 ，他们有足够的能力和工具及时分析工程量并监控市场价格。对于实施大型标准项目（商店、办公楼、住宅小区）的公司来说，这种做法正在成为标准。

随着模型的信息内容越来越完整和标准化，操纵和投机的可能性几乎消失殆尽。数字化转型正在逐步改变建筑行业的游戏规则，不适应这些变化的企业可能会面临严峻挑战。

> 竞争加剧、技术颠覆和利润缩水有可能影响企业的可持续发展。在流动资金有限的情况下，越来越多的行业参与者开始转向自动化、分析和数据技术，以此来提高效率和流程透明度。这些工具正成为在不断变化的经济环境中保持竞争力的重要资源。

也许我们不应该等待外部环境迫使我们采取紧急行动--从今天开始，通过加强数字化能力、实施现代解决方

案和建立以数据为中心的文化来做好准备要有效得多。

建筑行业大规模数字化转型的最后一个关键技术障碍是建筑项目 要素的自动分类问题，这将在未来几年影响到每一家公司。

> 没有可靠、准确和可扩展的分类，就不可能利用人工智能和预测模型为全面分析、流程自动化和生命周期管理奠定基础。只要对象分类仍依赖于经验丰富的专家（工头、设计师、估算师）进行人工解读，建筑行业就仍有机会。我们可以利用这段时间为不可避免的变化做好准备：对透明度、工具和数据的民主化以及自动分类系统的出现提出了更高的要求，这些都将从根本上改变游戏规则。

对建筑世界中的元素进行自动分类的任务，其复杂程度堪比无人驾驶系统中的物体识别，这也是目前面临的主要挑战之一。让我们想象一下一辆从 A 点行驶到 B 点的无人驾驶汽车（图 10.1-5）。目前的自动驾驶系统被激光雷达和摄像头识别的物体分类问题所困扰。汽车仅仅 "看到 "障碍物或地标是不够的，还必须能够识别出是行人、路标还是垃圾桶。

整个建筑行业也面临着类似的基本挑战。项目元素（如门窗或柱子）可以记录在文件中、体现在 CAD 模型中、在施工现场拍摄或通过激光扫描在点云中识别。然而，要建立一个真正的自动化项目管理系统，仅靠目视或粗略的几何识别是不够的。有必要确保每个元素都被准确、一致地归类到一个类型中，该类型在所有后续流程中--从估算和规格到物流、库存控制以及最重要的运营--都是唯一可识别的（图 4.2-6）。

在这一阶段--从识别到有意义的分类--出现了一个关键的障碍。即使数字系统在技术上能够区分和识别模型中和施工现场的物体，但主要的困难在于如何在不同的软件环境中正确和稳定地定义元素类型。。例如，一扇门在 CAD 模型中可能被设计者标注为 "门 "类元素，但当传输到 ERP 或 PMIS 系统时，可能会由于设计者的错误或系统之间的不一致而被错误标注。此外，在数据导出和导入过程中，元素往往会丢失一些重要属性或从系统会计中消失。这就导致了数据流的缺口，破坏了建筑流程端到端数字化的原则。这在 "可见 "和 "可理解 "的语义之间造成了严重的差距，破坏了数据的完整性，并使建筑项目整个生命周期内的流程自动化变得更加复杂。

利用大数据和机器学习技术解决 建筑构件的通用分类问题 （图 10.1-6），将成为全行业转型的催化剂--或许也是许多建筑公司的意外发现。一个统一的、可学习的分类系统 将成为可扩展分析 、数字化管理和将人

工智能 应用于建筑组织日常实践的基础。

英伟达™（NVIDIA®）和其他技术领先企业已经在其他行业提供解决方案，能够自动对大量文本和视觉信息进行分类和结构化处理。

> 例如，英伟达公司的 NeMo Curator 模型[161]擅长将数据自动分类并归入预定义的类别，在优化信息处理管道以微调和预训练生成式人工智能模型方面发挥着关键作用。Cosmos 平台在真实世界的视频和三维 场景中进行训练 [162]，为英伟达生态系统中正在构建的自主系统和数字双胞胎奠定了基础。英伟达 Omniverse，到 2025 年已成为使用 USD 格式的领先工具，这是一种通用场景描述，最终可在设计信息传输过程中取代 IFC 格式。与机器人流程模拟器 Isaac Sim [163]一起，NeMo Curator、Cosmos 和 Omniverse 等解决方案代表了自动化的新水平：从数据清理和过滤到训练集生成、对象属性建模以及建筑工地上的机器人训练。而且所有这些工具都是免费开源的，大大降低了工程和施工实践中采用这些工具的障碍。

在结构化表格层面对数据进行自动分类并不像乍看起来那么困难。正如我们在上一章（图 9.1-10）中所展示的，如果有历史数据积累，就有可能根据其他要素的类似参数来弥补缺失或不正确的分类值。如果在几个已完成的项目中已经对具有类似特征的要素进行了正确分类，那么系统就可以很有 可能为新要素或未完成要素建议一个合适的数值（图 10.1-6）。这种基于平均值和上下文分析的逻辑在大量处理来自估算、规范或 CAD 模型的表格数据时尤为有效。

>

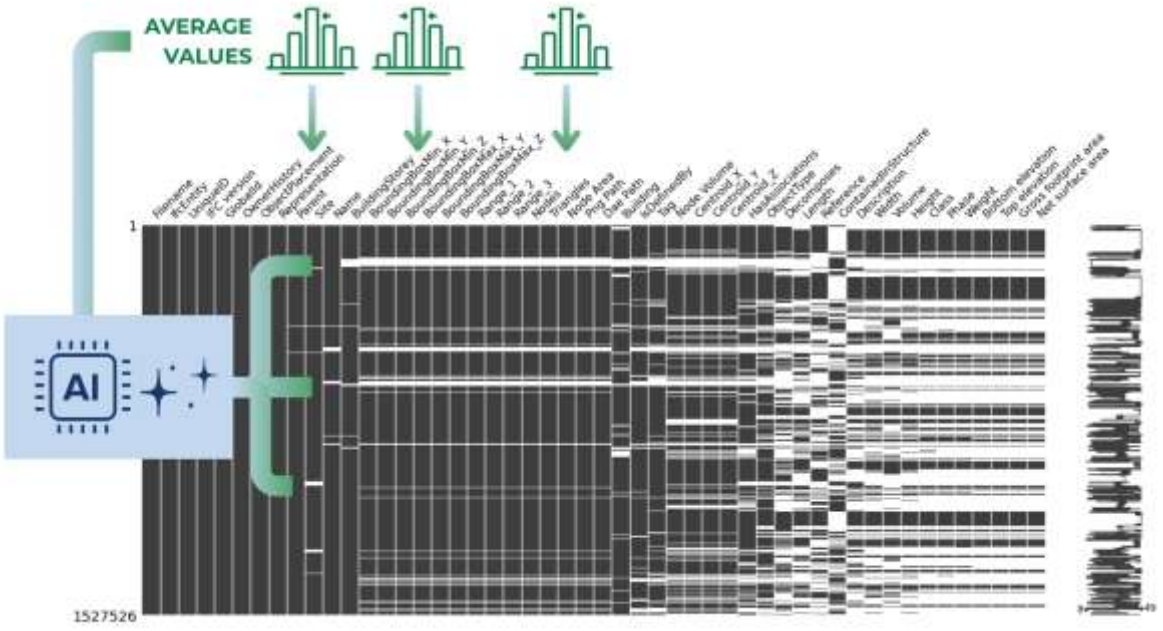
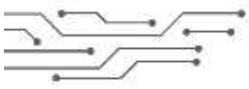


图 10.1-6 机器学习 可以帮助您根据过去的项目自动查找未填写（白框）表格参数的平均值。

> 在机器学习取得如此快速进步的背景下，我们可以清楚地看到：在 2025 年，认为建筑构件自动分类问题将长期得不到解决是天真的想法。是的，当前的算法尚未完全成熟，尤其是在数据不完整或异构的情况下，但适应的机会之窗正在迅速关闭。

已经投资于数据收集、清理和组织以及采用 ETL 自动化工具的公司将占据优势。其他公司则有可能落在后面--就像运输和金融行业的公司曾经无法应对数字化转型的挑战

那些继续依赖人工数据管理和传统成本与时间估算方法的公司，有可能会陷入 2000 年代出租车队的境地，到 2020 年代初无法适应移动应用程序和自动路线计算时代的到来。



章 10.2.

实施数据驱动方法的实用指南

从理论到实践：建筑业数字化转型路线图

建筑行业正逐步进入一个新的发展阶段，人们熟悉的流程越来越多地被数字平台和透明的互动模式所补充，有时甚至被取代。这不仅给企业带来了挑战，也带来了重大机遇。如今，那些已经制定了长期数字化战略的企业不仅能够保持其市场地位，还能通过向客户提供现代化的方法和可靠的技术解决方案来扩大市场。

> 必须认识到，概念和技术知识只是一个起点。管理人员和专家面临着一个实际问题：从哪里开始实施，如何将理论观点转化为实际价值。此外，问题也越来越多：如果客户可以随时修改传统的成本计算和时间安排方法，那么企业将以什么为基础。

答案可能并不在于技术，而在于形成一种新的职业文化，将数据工作视为日常实践不可分割的一部分。正是由于缺乏对数字技术和创新的关注，建筑行业在过去几十年中陷入了严重的落后状态[43]。

> 根据麦肯锡的数据，建筑业的研发支出不到收入的 1%，而在汽车和航空航天业，这一数字达到了 3.5%-4.5%。同样，建筑业的信息技术成本仍不到总收入的 1%[107]。

因此，不仅自动化水平在下降，建筑业的劳动生产率也在下降，到 2020 年，建筑工人的生产率将低于半个世纪前（图 10.2-1）

大多数发达国家和发展中国家都存在建筑业生产率低下的问题（29 个经合组织国家中有 16 个国家的建筑业生产率下降（图 2.2-1）），这不仅说明缺乏技术，而且说明需要对管理、培训和创新方法进行系统改革。

> 数字化转型的成功与否，与其说取决于工具的数量和可用性，不如说取决于组织是否有能力审查其流程，并形成一种乐于变革的文化。关键不在于技术本身，而在于确保有效使用技术、支持持续学习和鼓励接受新理念的人员和流程。

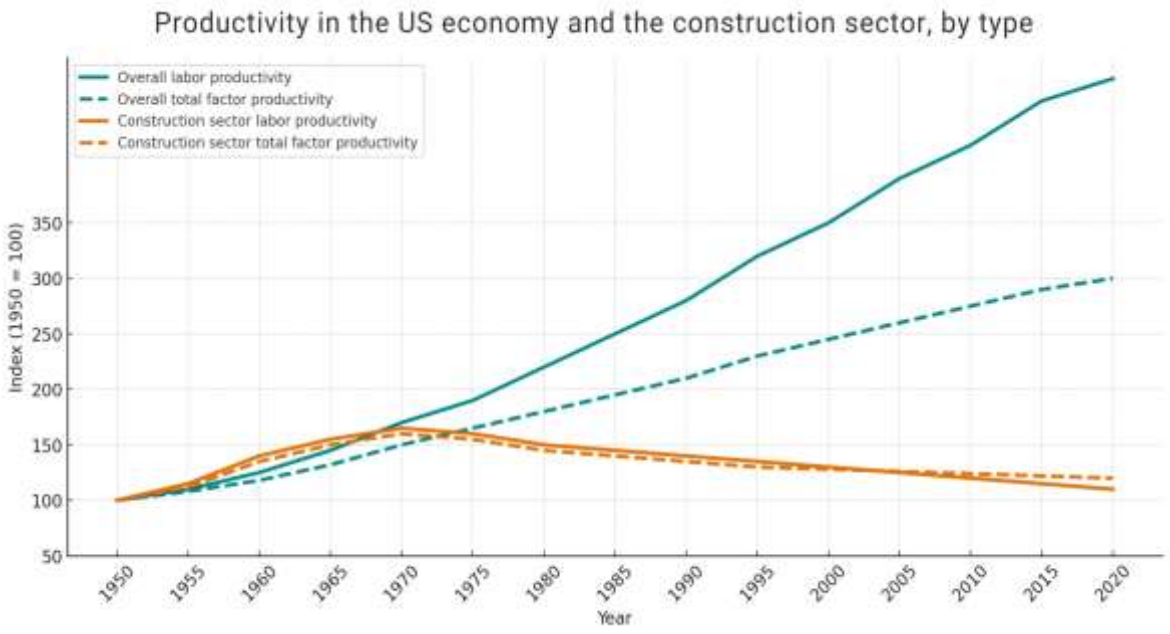


图 10.2-1 美国经济和建筑部门劳动生产率与总资源生产率的悖论（1950-2020 年）（基于 [43]）

。

在本书的早期部分，商业环境模式被比作森林生态系统（图 2.1-2，图 1.2-4，图 1.3-2）。在一个健康的森林中，周期性的火灾虽然破坏力巨大，但却在长期更新中发挥着关键作用。它们清除土壤中的老植被，恢复储存的养分，并为新生命创造空间。有些植物物种甚至进化到只有在火场温度较高时种子才会开放--这种自然机制为发芽提供了理想的时间。

同样，在企业中，危机可以作为一种 "有控制的倦怠"，鼓励新方法和新公司的出现，而不受制于过时的制度。这种时期迫使人们放弃低效的做法，为创新腾出资源。就像火灾后的森林会从先锋植物开始一样，危机后的企业也会形成新的、灵活的流程，成为成熟信息环境的基础。

那些能够正确解读这些 "信号火"，并将其破坏性能量转化为建设性变革的公司，其业绩将达到一个新的水平--数据流程更加透明、适应性更强，从而提高组织的自然更新和发展能力。

> 人工智能 和机器学习对商业环境的影响越来越大，这一点已毋庸置疑。这不仅是一种暂时的趋势，更是一种战略需要。在日益鼓励创新和灵活性的市场中，忽视人工智能 的公司有可能失去竞争力。

未来属于那些不仅将人工智能视为工具，而且将其视为重新思考业务各个方面（从优化流程到管理决策）的机会的人。

奠定数字化基础：迈向数字化成熟的 **1-5** 个步骤

在本章中，我们将探讨数字化转型路线图，并确定实施数据驱动方法所需的关键步骤，以帮助转变企业文化和公司的信息生态系统。

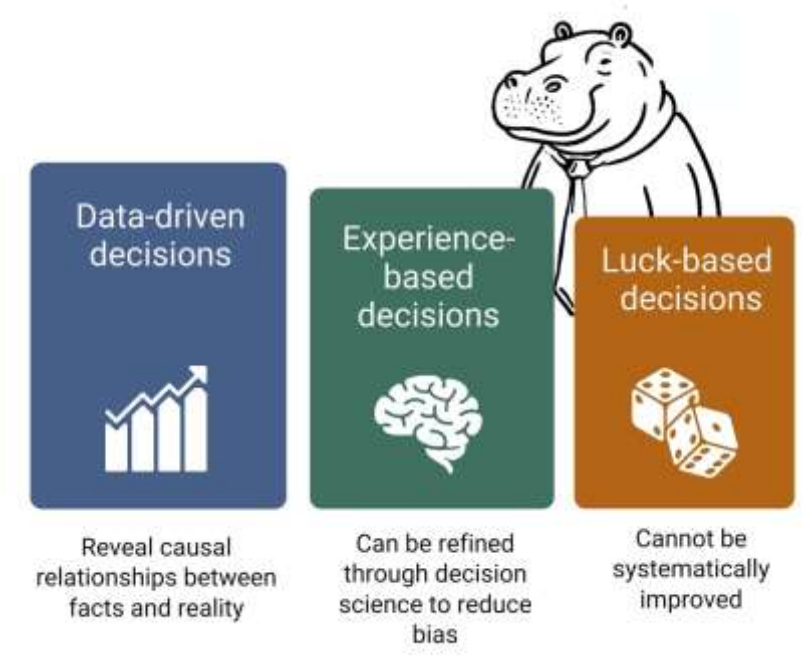


图 10.2-2 受控更新和战略选择：案例、经验或数据。

根据麦肯锡的研究报告《数字化战略为何失败》（2018 年），企业未能实现数字化转型目标至少有五个原因 [164]

- 定义模糊：高管和经理对 "数字化 "的含义有不同的理解，导致误解和不一致。
- 误解数字经济：许多公司低估了数字化给商业模式和行业动态带来的变革规模（图 10.1-6）。
- 忽视生态系统：公司只关注单个技术解决方案（数据孤岛），忽视了融入更广泛的数字生态系统的必要性（图 2.2-2，图 4.1-12）。

- 低估竞争对手的数字化：管理者没有考虑到竞争对手也在积极采用数字化技术，这可能导致竞争优势的丧失。
- 忽视数字化的双重性：首席执行官将数字化转型的责任下放给其他高管，从而使控制变得官僚化，并减缓了变革进程。

要应对这些挑战，就必须清楚地了解并协调组织各个层面的数字化战略。在制定数字化战略之前，了解起点非常重要。许多组织倾向于在不完全了解现状的情况下采用新的工具和平台。

步骤 **1**：对现有系统和数据进行审计。

在改变流程之前，了解已有的流程非常重要。通过审计，您可以发现数据管理中的薄弱环节，了解可以利用哪些资源。审计是业务流程的一种 "X 光"。它能让您识别风险领域，并确定哪些数据对您的项目或业务至关重要，哪些是次要的。

主要行动

- 绘制您的 IT 环境（使用 Draw.io、Lucidchart、Miro、Visio 或 Canva）。列出您的流程中所使用的系统（ERP 、CAD、CAFM 、CPM 、SCM 及其他），我们在 "现代建筑技术和管理系统 "一章中讨论过这些系统（图 1.2-4）
- 评估每个系统的数据质量问题，包括重复频率、可能的缺失值和格式不一致。
- 确定 "痛点"--流程可能崩溃或经常需要人工干预的地方--导入、导出和额外的验证流程。

> 如果想让团队信任报告，就必须从一开始就确保数据的正确性。

高质量的数据审计将向您展示哪些数据：

- 需要进一步开发（需要设置自动清洁程序或附加转换功能）
- 它们是只会堵塞系统的 "垃圾"，可以通过在流程中不再使用它们来清除。

自己进行审计是可能的。但有时聘请外部顾问（尤其是来自其他行业的顾问）也很有用：新的视角和独立于建筑 "特殊性 "的独立性将有助于冷静地评估现状，避免偏向某些解决方案和技术的典型陷阱。

步骤 **2**：确定数据统一的关键标准。

审计结束后，有必要制定处理数据的通用规则。正如我们在 "标准：从随机文件到智能数据模型 "一章中所讨论的，这将有助于消除孤立的数据流。

> 如果没有一个统一的标准，每个团队都将继续 "各自为政"，你将维持一个集成 "动物园"，每次转换都会丢失数据。

主要行动

- 选择数据标准，以便在系统间交换信息：
  - 对于表格数据，可以是 CSV 、XLSX 等结构化格式，也可以是 Parquet 等更高效的格式。
  - 用于交换结构松散的数据和文档：JSON 或 XML
- 掌握数据模型的工作方法：
  - 如 "数据建模：概念、逻辑和物理模型 "一章所述（图 4.3-2）
  - 随着对业务流程逻辑的深入研究，可以使用逻辑模型和物理模型中的参数将需求正规化（图 4.3-6）
  - 确定流程内的关键实体、其属性和关系，并将这些关系可视化 - 实体之间和参数之间的关系（图 4.3-7）
- 使用正则表达式（RegEx ）来验证和标准化数据（图 4.4-7） ，正如我们在 "结构化需求 和 RegEx 正则表达式 "一章中所讨论的。RegEx 并不复杂，但在创建物理数据模型需求的工作中是一个非常重要的课题。

如果没有数据级标准和流程可视化，就不可能提供一致且可扩展的数字环境。请记住，"糟糕的数据是昂贵的"。随着项目或组织变得越来越复杂，错误的代价也会越来越高。统一格式、定义命名、结构和验证规则是对未来解决方案稳定性和可扩展性的投资。

第 **3** 步：实施 **DataOps** 并实现流程自动化。

如果没有一个定义明确的架构，公司将不可避免地面临孤立的信息系统中包含的不同数据。数据将无法整合，在多个地点重复，而且维护成本高昂。

> 想象一下，数据是水，而数据架构则是将水从存储源头输送到使用点的复杂管道系统。数据架构决定了信息的收集、存储、转换、分析以及向终端用户或应用交付的方式。

DataOps （数据操作）是一种将数据的收集、清理、验证和使用整合到单一自动化流程中的方法，我们在本书第 8 部分中对此进行了详细讨论。

主要行动

- 创建和定制 ETL - 输送机，实现流程自动化：
  - 提取：从 PDF 文档（图 4.1-2、图 4.1-5、图 4.1-7）、Excel 电子表格、CAD -模型（图 7.2-4）、ERP -系统和您使用的其他来源组织自动数据收集
  - 转换：设置自动流程，将数据转换为单一的结构化格式，并自动进行在封闭应用程序之外进行的计算（图 7.2-8）
  - 加载：尝试将数据自动上传到汇总表、文件或集中存储库（图 7.2-9、图 7.2-13、图 7.2-16）。
- 自动计算和 QTO （工程量清单）流程，我们在 "QTO 工程量清单：按属性分组项目数据 "一章中讨论过：
  - 利用应用程序接口、插件或逆向工程工具（图 5.2-5），设置从 CAD - 模型中自动提取体积。
  - 以表格形式创建不同类别元素的属性分组规则（图 5.2-12）
  - 在模块化封闭系统之外，尝试将频繁重复的体积和成本计算自动化（图 5.2-15）
- 正如我们在 "Python Pandas：处理数据不可或缺的工具 "一章中所讨论的，开始使用 Python 和 Pandas 处理数据：

- 应用 DataFrame 处理 XLSX 文件并自动处理表格数据（图 3.4-6）
- 通过各种 Python 库自动汇总和转换信息
- 使用 LLM 可简化现成代码块和整个管道的编写工作（图 7.2-18）
- 尝试用 Python 构建一个管道 ，它能发现错误或异常并向负责人（如项目经理）发送通知（图 7.4-2）

基于 DataOps 原则的自动化可让您从手动和零散的数据处理转变为可持续和可重复的流程。这不仅能减轻每天处理相同转换的员工的负担，还能显著提高整个信息系统的可靠性、可扩展性和透明度。

步骤 **4**：创建开放式数据治理生态系统。

尽管开发了封闭式模块系统并将其与新工具集成，但企业仍面临一个严重问题--这些系统日益复杂，超出了其实用性。创建一个涵盖所有业务流程的单一专有平台的最初想法导致了过度的集中化，任何变化都需要大量的资源和时间来适应。

正如我们在 "企业菌丝：数据如何连接业务流程 "一章中所讨论的，有效的数据管理需要一个开放、统一的生态系统，将所有信息源连接起来。

生态系统的关键要素：

- 选择合适的数据存储：
  - 表格和计算使用数据库，例如 PostgreSQL 或 MySQL（图 3.1-7）。
  - 对于文档和报告，云存储（Google Drive、OneDrive）或支持 JSON 格式的系统可能比较合适
  - 查看数据仓库、数据湖和其他工具的功能，以集中存储和分析大量信息（图 8.1-8）。
- 实施获取专有数据的解决方案：
  - 如果使用专有系统，请通过 API 或 SDK 对其进行访问配置，以获取用于外部处理的数据（图 4.1-2）
  - 熟悉逆向工程工具的潜力 ，用于 CAD 格式（图 4.1-13）。

- 建立 ETL 管道，定期从应用程序或服务器收集数据，将其转换为开放的结构化格式，并保存到存储库中（图 7.2-3）
- 在团队内部讨论如何在不需要专有软件的情况下提供数据访问权限
- 记住：数据比界面更重要。提供长期价值的是信息的结构和可用性，而不是特定的用户界面工具

- 如我们在 "数据建模卓越中心（CoE）"一章中所讨论的，考虑创建一个数据卓越中心 （CoE ），或考虑如何以其他方式提供数据专业知识（图 4.3-9）。

数据管理生态系统创建了一个统一的信息空间，使所有项目参与者都能利用一致、最新和经过验证的信息开展工作。它是可扩展、灵活和可靠的数字化流程的基础

释放数据潜力：实现数字化成熟的 **5-10** 个步骤

除技术集成外，成功实施数字化解决方案的一个重要因素是最终用户的采用。让客户或用户参与绩效衡量既是改善用户体验的挑战，也是管理公司变革的挑战。如果一个解决方案无法融入熟悉的工作流程，或者无法解决用户或客户的实际问题，那么它就不会被使用，再多的额外措施和激励也无法解决这个问题。

转型是一个迭代过程，基于对用户与新流程交互数据的分析、频繁的测试周期、持续的反馈和改进。

步骤 **5**：建立数据文化，培训员工并收集反馈意见

没有员工的参与，再先进的系统也无法发挥作用。您需要创造一种环境，让数据每天都被使用，让团队了解数据的价值。

> 英国政府发布的 2024 年报告《政府项目交付中的数据分析和人工智能 》指出[83]，培训具备数据处理和解读必要能力的专业人员对于成功实施 数据分析和人工智能 至关重要。

缺乏数据分析专业知识是制约数字化转型的关键问题之一。高管们习惯于既定的例行公事：季度周期、确定优先级的计划和推进项目的传统方式。变革需要一位特殊的领导者--级别高到足以产生影响力，但级别还没有高到有时间和动力来领导一个长期的转型项目。

主要行动

- 认识到有必要从基于高薪员工意见的主观决策（HiPPO ）转向基于事实和数据的决策文化，正如 "HiPPO 或决策中意见的危险 "一章所讨论的那样（图 2.1-9）。
- 组织系统培训：
  - 举办关于如何使用结构化数据的培训班，并邀请其他行业的专家参加，这些专家不会偏向于当今建筑行业流行的产品和概念
  - 与同事讨论数据分析方法和工具，并独立学习 Python、pandas 和 LLM 等工具的实际操作（图 4.1-3，图 4.1-6）。
  - 创建一个以数据结构（图 3.2-15）和创建数据模型（图 4.3-6，图 4.3-7）为主题的教程库（最好有视频短片）。
- 利用现代学习技术：
  - 使用语言模型（LLM）支持代码和数据操作，包括代码生成、重构和分析，以及表格信息的处理和解释（图 3.4-1）
  - 探索在离线开发环境 (IDE) 中工作时，如何调整 LLM 生成的代码并将其集成到完整的管道解决方案中（图 4.4-14，图 5.2-13）。

> 如果管理者继续用 "老办法 "做决策，再多的培训也无法说服人们认真对待分析。

没有持续的反馈，就不可能建立数据文化。通过反馈，您可以发现流程、工具和战略中存在的差距，而这些差距是无法通过内部报告或正式的 KPI 指标发现的。用户对您的解决方案提出的赞扬性意见不会带来实际价值。重要的反馈意见才是有价值的，尤其是基于具体的观察和事实的反馈意见。但是，获取此类信息需要付出努力：您需要建立一套流程，让内部和外部参与者都能在不歪曲事实、不担心自己的意见会影响自身工作的情况下分享意见（匿名分享可能更有意义）。重要的是，他们在这样做时不能歪曲事实，也不能担心会给自己带来负面影响。

> 任何学习最终都是自学[165]。

> - 米尔顿-弗里德曼，美国经济学家和统计学家

在实施分析工具的同时，应定期验证其实际效果（投资回报率、关键业绩指标），这只能通过员工、客户和合作伙伴的结构化反馈来实现。这不仅能让企业避免重蹈覆辙，还能让企业更快地适应环境的变化。拥有收集和分析反馈的机制，是一个组织从偶尔的数字化举措走向可持续的持续改进模式的成熟标志之一（图 2.2-5）。

步骤 **6**：从试点项目到扩大规模

> 选择足够大而重要、足够小而能赢的战斗。
>
> - 乔纳森-科佐尔

"一次性、无处不在 "地启动数字化转型风险极大。更有效的方法是从试点项目开始，逐步推广成功的做法。

主要行动

- 为试点选择合适的项目：
  - 定义具有可衡量结果的具体业务目标或流程 (KPI, ROI) (图 7.1-5)
  - 选择 ETL 自动化流程，例如使用 Python 和 Pandas 进行自动数据验证或工作量计算 (QTO 图 5.2-10）
  - 制定明确的成功衡量标准（例如，将制作检验规范或数据验证报告的时间从一周缩短到一天）
- 采取迭代方法：
  - 从简单的数据转换流程开始，将多格式数据流转换为流程所需的格式（图 4.1-2，图 4.1-5）
  - 通过在集成开发环境中形成基于文档代码块的完整流水线，逐步提高任务的复杂性并扩大流程的自动化程度（图 4.1-7，图 7.2-18）。

- 记录成功的解决方案（最好有视频短片），并与同事或专业团体分享

- 开发模板和配套文件，以复制这些解决方案，从而使您的同事（或专业社区成员，包括社交媒体用户）能够有效地使用这些解决方案

逐步 "扩展 "使您能够保持高质量的变更，而不会陷入并行实施的混乱局面。从小到大 "的战略可最大限度地降低风险，让您从小错中吸取教训，而不会让小错演变成严重问题。

从员工仅部分参与的项目方法转变为组建永久性团队（如专业技术中心--CoEs），有助于确保产品的可持续发展，即使在其第一个版本发布之后也是如此。这些团队不仅支持现有的解决方案，还将继续对其进行改进。

这减少了对冗长审批的依赖：团队成员有权在其职责范围内做出决策。因此，管理人员可以从 微观管理的需要中解脱出来，而团队则可以专注于创造真正的价值。

> 开发新的解决方案不是短跑，而是马拉松。成功者都是那些最初以长期、持续的工作为目标的人。

必须认识到，技术需要不断发展。投资于技术解决方案的长期发展是成功运营的基础。

步骤 **7**：使用开放数据格式和解决方案

正如我们在模块化平台（ERP、PMIS、CAFM、CDE 等）章节中所讨论的，重要的是要注重开放和通用的数据格式，以确保独立于供应商的解决方案，并提高所有利益相关者的信息可用性。

主要行动

- 从封闭格式转向开放格式：

  - 使用开放格式而不是专有格式，或者找到一种方法，设置自动上传或将封闭格式转换为开放格式（图 3.2-15）。

  - 使用 Parquet 、CSV 、JSON 、XLSX 等工具，这些工具是大多数现代系统之间的交换标准（图 8.1-2）

- 如果三维几何图形在您的流程中起着重要作用，请考虑使用 USD、glTF、DAE 或 OBJ 等开放格式（图 3.1-14）。

- 使用矢量数据库 数据库有效地分析和搜索信息：
  - 使用边框 和其他方法简化三维 - 几何图形（图 8.2-1）
  - 考虑在哪些地方可以实现数据矢量化--将文本、对象或文档转换为数字表示（图 8.2-2）。
- 应用大数据分析工具：
  - 以适合分析的格式（Apache Parquet, CSV, ORC）组织存储积累的历史数据（如 PDF, XLSX, CAD ）（图 8.1-2）。
  - 开始应用基本统计方法并处理代表性样本，或至少熟悉统计的基本原理（图 9.2-5）
  - 实施并学习可视化数据和数据间关系的工具，使分析结果可视化。没有良好的可视化，就不可能完全理解数据本身或基于数据的流程（图 7.1-4）。

向开放数据格式的转变以及信息分析、存储和可视化工具的引入，为可持续和独立的数字治理奠定了基础。这不仅能减少对供应商的依赖，还能确保所有利益攸关方平等获取数据。

步骤 **8**：开始实施机器学习预测

许多公司积累了大量数据--一种尚未使用的 "信息喷泉"。这些数据是在成百上千个项目中收集的，但在后续流程中往往只使用过一次或根本没有使用过。以封闭格式和系统存储的文件和模型往往被视为过时和无用的压舱石。但实际上，它们是宝贵的资源，是分析所犯错误、实现日常操作自动化以及为未来项目的自动分类和特征识别开发创新解决方案的基础。

关键的挑战在于学习如何提取这些数据并将其转化为可操作的见解。正如 "机器学习和预测 "一章所述，机器学习技术有可能显著提高各种建筑相关流程中估算和预测的准确性。充分利用积累的数据为提高效率、降低风险和建立可持续的数字化流程开辟了道路。

主要行动

- 从简单的算法开始：

- 尝试使用线性回归--利用 LLM 的提示--来预测数据集的重复性能，在这些数据集中，大量因素的依赖性不存在或极小（图 9.3-4）。
- 考虑理论上 k 近邻（k-NN）算法可应用于流程的哪些阶段，例如分类任务、对象相似性评估或基于历史类比的预测（图 9.3-5）。

- 收集和构建数据以训练模型：
  - 在一个地方以单一格式收集项目历史数据（图 9.1-10）
  - 通过自动 ETL，努力提高训练样本的质量和代表性（图 9.2-8）
  - 学习将数据分为训练集和测试集，正如我们在泰坦尼克号数据集示例中所做的那样（图 9.2-6，图 9.2-7）
- 考虑将机器学习技术的应用扩展到广泛的任务中，从预测项目时间表到优化物流、资源管理和早期识别潜在问题。

机器学习是一种将存档数据转化为宝贵资产的工具，，用于预测、优化和明智决策。从小型数据集（图 9.2-5）和简单模型开始，逐渐增加复杂性。

步骤 **9**：整合物联网 和现代数据收集技术

建筑世界正在迅速数字化：每一张建筑照片、每一个团队帖子都已成为将现实参数化和标记化的更大进程的一部分。正如全球定位系统曾经改变了物流业一样，物联网、无线射频识别（RFID）和自动数据采集技术也正在改变建筑业。正如 "物联网和智能合约 "一章所述，带有传感器和自动监控功能的数字化施工现场是建筑业的未来。

主要行动

- 实施物联网 - 设备、RFID - 标签，并详细说明与之相关的流程：
  - 评估哪些领域或项目阶段可以获得最大的投资回报率（ROI），例如，用于监测温度、振动、湿度或的投资回报率
  - 考虑在整个供应链中使用 RFID 追踪材料、工具和设备

- 考虑如何将收集到的数据整合到一个单一的信息系统中，如 Apache NiFi，以进行自动实时处理和分析 （图 7.4-5）

- 建立实时监测系统：
  - 使用可视化工具（如 Streamlit、Flask 或 Power BI ）开发仪表板，跟踪关键流程或项目指标。
  - 设置自动通知，发出严重偏离计划或规范的信号（图 7.4-2）
  - 根据收集的数据和确定的模式，评估对设备进行预测性维护的潜力（图 9.3-6）
- 合并不同来源的数据：
  - 从物理层面上的数据模型可视化开始--反映来自 CAD 系统、物联网设备和 ERP 平台的信息流结构和关键参数（图 4.3-1）
  - 首先，创建数据分析和管理决策支持统一平台的大致轮廓。捕捉关键功能、数据源、用户和预期应用场景（图 4.3-7）。

越早开始将实际流程与数字世界连接起来，就能越早利用数据对其进行高效、透明和实时的管理。

步骤 **10.**为行业未来的变化做好准备

建筑公司经常受到外部环境的压力：经济危机、技术飞跃、法规变化。就像森林要经受雨雪、干旱和烈日的考验一样，企业也要不断适应环境。就像树木通过深根系统能够抵御霜冻和干旱一样，只有那些拥有坚实的自动化流程基础、能够预测变化并灵活调整战略的组织才能保持活力和竞争力。

正如 "生存策略：建立竞争优势 "一章所述，建筑行业正在进入一个彻底转型的阶段。客户与承包商之间的互动正在向说服模式转变 ，透明度、可预测性和数字化工具正在取代传统方法。在这一新的现实中，获胜的不是规模最大的企业，而是最灵活、技术最成熟的企业。

主要行动

- 分析开放数据背景下的业务漏洞：
  - 评估作为优步化一部分的数据访问民主化如何对您的竞争优势和业务产生破坏性影响（图

10.1-5）。

- 考虑制定一项战略，从不透明性和孤岛式流程转变为基于开放式解决方案、系统互操作性和数据透明度的业务模式（图 2.2-5）。

- 制定长期数字战略：
  - 确定您是希望成为创新领导者，还是更倾向于 "迎头赶上"，以节省资源
  - 描述各阶段：短期（流程自动化、数据结构化）、中期（实施 LLM 和 ETL ）、长期（数字生态系统、集中存储库）
- 考虑扩大服务范围：
  - 考虑提供新的服务（侧重于能源效率、环境、社会和治理、数据服务）。我们将在下一章详细讨论新的业务模式
  - 将自己定位为可靠的技术合作伙伴，为设施从设计到运营的整个生命周期提供支持。对您的信任应建立在系统化的方法、透明的流程以及提供可持续技术解决方案的能力之上

在转型环境中，获胜的不是那些简单应对变化的人，而是那些积极主动的人。灵活性、开放性和数字化成熟度是建设明天的可持续性基础。

转型路线图：从混乱到数据驱动型公司

以下计划可作为一个初始基准，也是您制定自己的数据驱动型数字化转型战略的起点：

- 审计和标准：分析现状，统一数据
- 数据结构化和分类：自动转换非结构化和松散结构化数据
- 自动分组、计算和计算 ： 使用开源工具和库实现自动化
- 生态系统和 **COE**：建立内部团队，在公司内部形成统一的数据生态系统
- 文化与学习：从 HiPPO - 解决方案转向数据驱动的解决方案
- 试点、反馈和推广：迭代行动：在有限的范围内测试新方法，收集有效反馈，并 逐步推广解决方案

。

- 开放格式：使用通用和开放格式，独立于软件供应商
- 机器学习 ：将 ML 算法嵌入流程，以进行预测和优化
- 物联网 和数字化建筑工地：将现代数据收集技术整合到流程中
- 战略调整：为未来的行业变化做好准备

> 最重要的是，要记住 "数据本身并不能改变公司：改变公司的是懂得如何使用数据的人"。强调文化、透明流程和持续改进的承诺

通过系统方法，可以从各自为政的数字计划转变为全面的数据驱动管理模式，在这种模式下，决策不是基于直觉或假设，而是基于数据、事实和经过数学计算的概率。建筑行业的数字化转型不仅仅是采用技术，而是要塑造一个业务生态系统，在这个生态系统中，项目信息可以在不同系统之间无缝、反复地传输。在此过程中，机器学习算法可提供自动、持续的分析、预测和流程优化。在这样的环境中，猜测和隐藏数据变得无关紧要，剩下的只有经过验证的模型、透明的计算和可预测的结果。

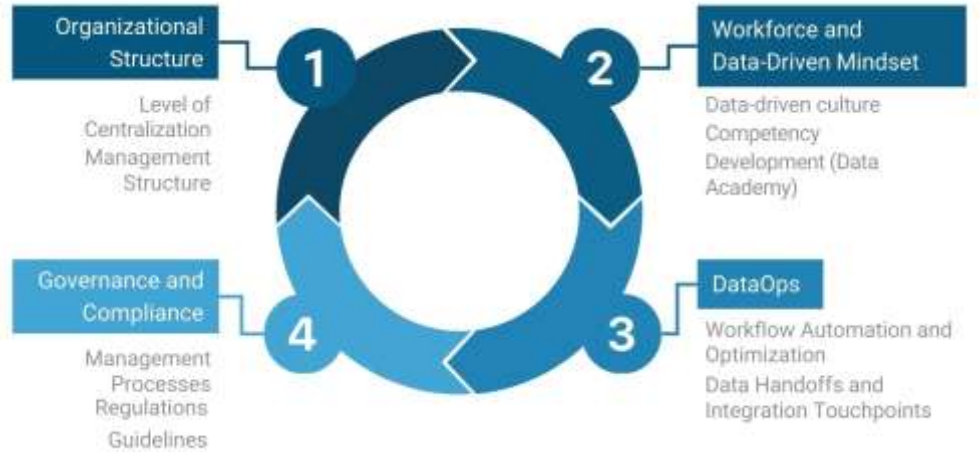


图 10.2-3 成功的公司级数据管理的关键要素。

本书的每一部分都与施工项目中数据处理和分析的特定阶段相对应（图 2.2-5）。如果您想回到前面讨论过的某个主题，并从数据利用流程的整体理解来看待它，可以参考图 10.2-4 中的部分标题。

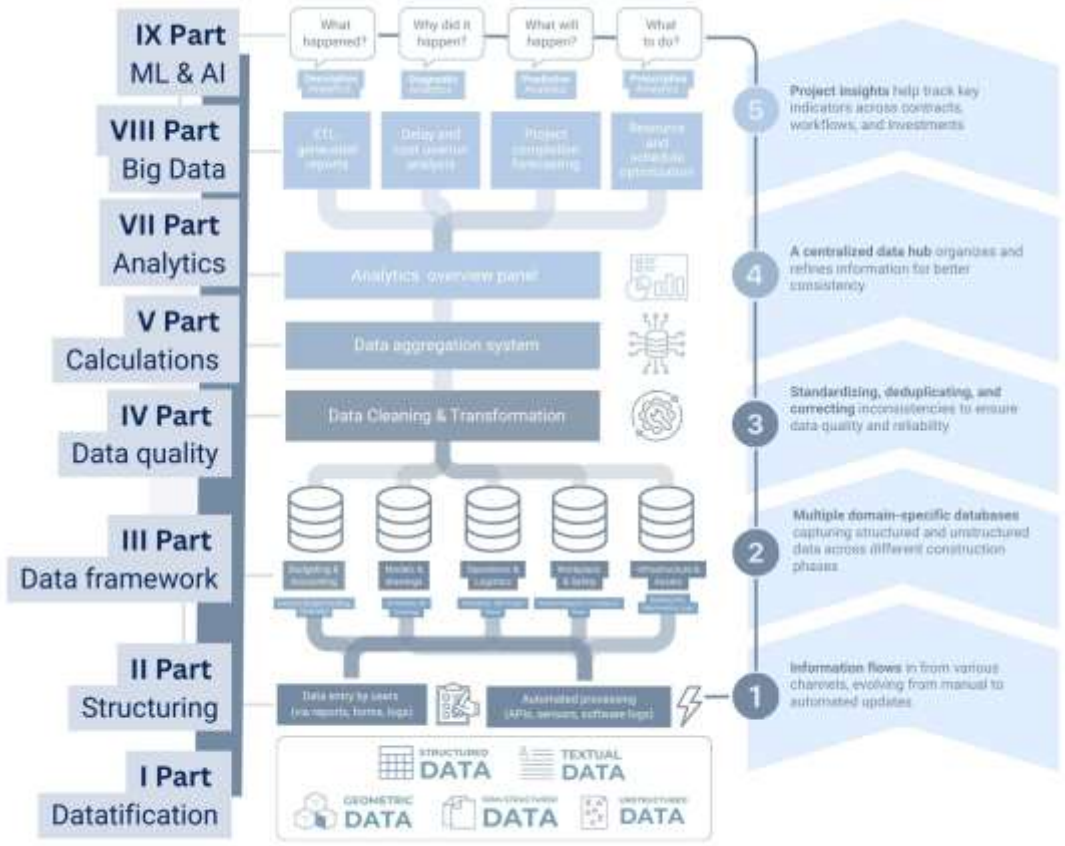


图 10.2-4 数据处理流水线（图 2.2-5）背景下的图书部分：从信息数字化到分析和人工智能。

无论贵组织的规模、技术成熟度或预算如何，您都可以从今天开始采用数据驱动的 方法。即使是朝着正确方向迈出的一小步，随着时间的推移也会产生效果。

> 数据驱动的转型不是一次性项目，而是一个持续、迭代的改进过程，包括引入新工具、重新设计流程和发展数据驱动决策的文化。

工业 **5.0** 中的建筑：无法再躲藏时如何赚钱

长期以来，建筑公司一直靠流程不透明赚钱。主要的商业模式是投机--在封闭的 ERP 和 PMIS 系统中高估材料成本、工程范围和加价百分比--外部审计无法进入这些系统。客户及其信任的人对原始项目数据的访问受到限制，这就为各种计划提供了土壤，几乎不可能对计算的可靠性进行验证。

然而，这种模式正在迅速失去意义。随着数据访问的民主化、LLM 的出现、开放数据的到来以及 ETL 自动化工具的出现，整个行业正在向新的工作标准迈进。

> 因此，不透明不再是一种竞争优势，很快就会成为一种难以割舍的负担。透明度从一种选择变成了立足市场的先决条件。

在新的数字化现实中，客户--银行、投资者、实体客户、私募基金 、政府客户--将与谁合作？答案显而易见：那些不仅能提供结果，还能提供实现结果过程中每一步的合理性的公司。随着开放数据量的增长，合作伙伴和客户将选择那些能保证结果的透明度、准确性和可预测性的公司。

在此背景下，基于数据管理和信任而非投机的新商业模式正在兴起：

- 销售流程而不是平方米：关键资产是信任和效率，而不是打折扣的具体协议。主要价值将是以可靠和经过验证的数据为基础的结果的可预测性。现代公司出售的不是建筑工程本身，而是：
  - 准确的最后期限和透明的工作时间表；
  - 合理的估算 ，并有计算支持；
  - 在项目的各个阶段实现全数字化可追溯性和控制。
- 工程和分析即服务："数据即服务 "模式（一种通过互联网向用户提供现成数据的服务方式），每个项目都成为数字数据链的一部分，商业价值在于管理这一数据链的能力。公司正在转型为智能平台，提供基于自动化和分析的解决方案：
  - 自动、透明地编制估算和计划；
  - 基于机器学习算法的风险和时机评估；
  - 计算环境指标（ESG、CO_2、能效）；

- 从经审计的公开来源生成报告。

- 工程专业知识的产品化：公司的研发成果可以在公司内部反复使用，并作为单独的产品进行销售--通过数字服务形成额外的收入来源。在新的环境中，企业不仅要创造项目，还要创造数字资产：
  - 组件库和估算模板；
  - 自动验证模块；
  - 用于处理数据的开源插件和脚本（销售咨询）。
- 新型公司：数据驱动型集成商：不依赖于特定软件供应商或模块化系统，也不 "锁定 "于单一软件界面的市场参与者。它可以自由使用数据，并在此基础上建立自己的竞争力。未来的建筑公司不仅仅是一个承包商，而是一个信息集成商，能够为客户履行以下职能：
  - 整合不同来源的数据并进行分析；
  - 确保程序的透明度和可信度；
  - 就优化业务流程提供建议；
  - 开发可在开放数据生态系统中使用的工具，LLM 、ETL 和 Pipelines。

工业 5.0（图 2.1-12）标志着 "手工平均比率时代 "和首席执行官与估算和会计部门之间的晚间会议的终结。以前隐藏的一切--计算、估算 、数量--都将变得公开、可验证，甚至非专业人士也能理解。最先调整自己的人将成为赢家。所有其他企业都将被排除在建筑业新的数字经济之外。

结 论

建筑业正在进入一个根本性变革的时代。从最早的泥板记录到项目服务器和施工现场流出的大量数字数据，建筑业信息管理的历史始终反映了当时技术的成熟程度。如今，随着自动化、开放格式和智能分析系统的出现，建筑业面临的不是渐进式发展，而是快速的数字化转型。

与其他经济部门一样，建筑业不仅要重新思考工作工具，还要重新思考工作原则。那些曾经主宰市场并充当客户与项目之间主要中介的公司正在失去其独特的地位。信任和数据工作能力--从收集和构建数据，到分析、预测和自动决策--正在凸显出来。

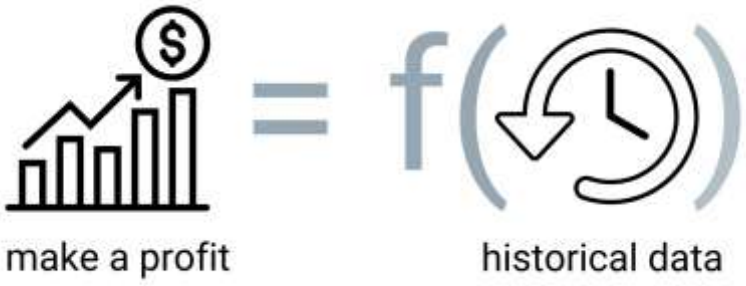


图 10.2-1 结构化历史数据是高效和可管理业务的燃料。

本书详细介绍了建筑行业数据管理的关键原则--从审计和标准化到流程自动化、可视化工具的使用以及智能算法的实施。我们探讨了如何在资源有限的情况下建立一个有效的数据架构，并开始根据可验证的事实而非直觉做出决策。数据工作不再仅仅是 IT 部门的任务，它已成为管理文化的基础，决定着公司的灵活性、适应性和长期可持续性。

如今，机器学习技术（）、自动处理系统、数字孪生和开放式格式的应用已使消除人为因素成为可能，而这在以前是至关重要的。建筑业正朝着自主化和可控化方向发展，从构思到项目实现的过程可以比作自动驾驶模式下的航行：不依赖主观决定，不需要在每个阶段进行人工干预，而是完全数字化的可追溯性和可控性（图 10.2-2）。

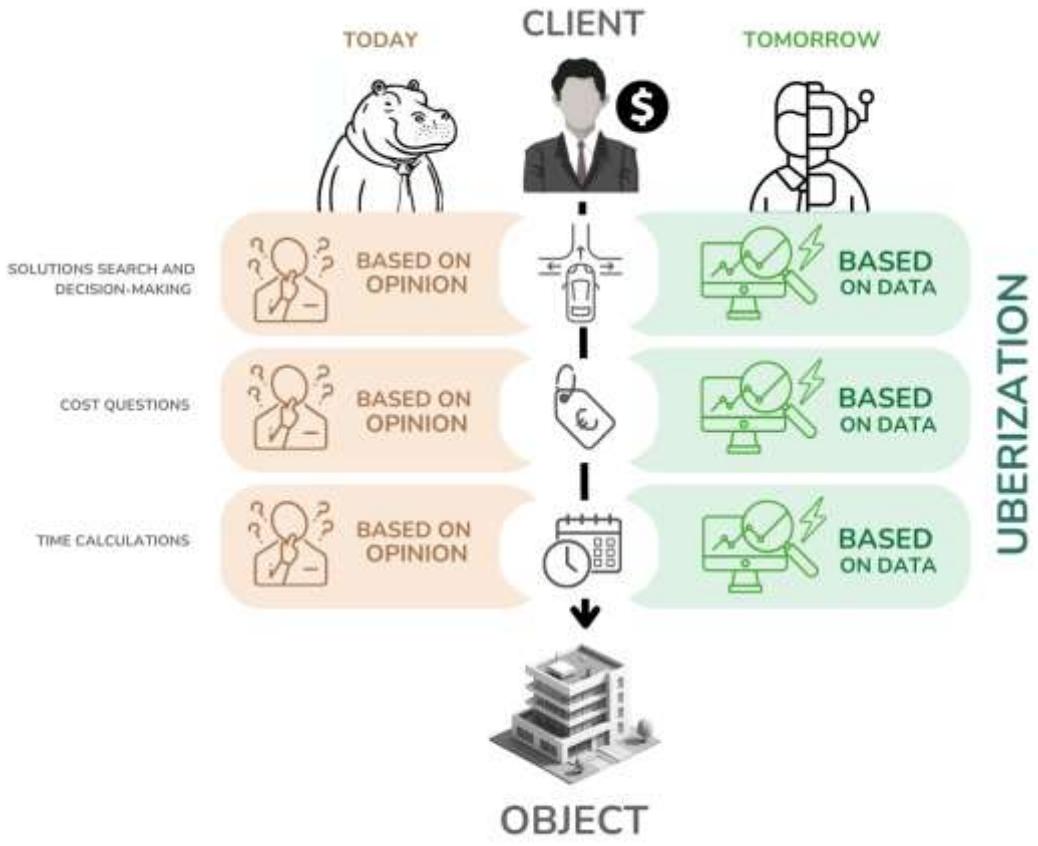


图 10.2-2 从基于重要专家意见（HiPPO ）的决策到数据分析的转变将主要由客户来推动。

通过学习本书介绍的方法、原理和工具，你将能够开始在公司中做出数据驱动而非直观的决策。您还将能够在 LLM 中运行模块链，将开箱即用的 ETL Pipelines 复制到您的开发环境 (IDE)，并自动处理数据，以您想要的形式获取所需的信息。在本书有关大数据和机器学习章节的基础上，您将能够实施更复杂的方案--从历史数据中提取新的见解，并应用机器学习算法来预测和优化您的流程。

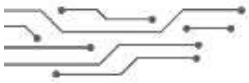
开放的数据和流程将为更准确地估算项目成本和时间提供基础，防止建筑公司对不透明的数据进行猜测。这既是挑战，也是机遇，建筑行业必须重新思考自己的角色，适应新的环境，使透明度和效率成为成功的关键因素。

> 愿意接受知识并将其付诸实践，是在数字化转型时代取得成功的关键。

首先意识到这一点的公司将在新的数字化竞争中占据优势。但重要的是要认识到，单靠数据并不能改变什么。许多人需要改变思维方式，而这需要激励。您的公司必须重新思考共享数据的方式

能够改变公司的人，是那些懂得如何使用这些数据、解读这些数据、利用这些数据进行优化并在此基础上创建新流程架构的人。

如果你正在阅读这些文字，那么你已经做好了改变的准备，你已经领先了一步。感谢您选择了这条路。欢迎来到数字化转型时代！



关于作者



我叫阿尔乔姆-博伊科。我的建筑工地之旅始于 2007 年--在圣彼得堡矿业大学学习矿山和地下建筑专业期间，我在家乡的一个油页岩矿当矿工。在这本书的封底，你可以看到一个爆破员在工作面进行爆破，我们在那里开采和爆破了数百立方的油页岩。我的职业生涯经历了许多不同的方向，从矿山和地下建筑工人到工业攀爬工、屋顶工和电梯设备安装工。我有幸参与过各种规模的项目，从私人住宅到世界各地大型工业设施的建设。

随着时间的推移，我的工作从实体建筑转向信息管理和数字化流程。自 2013 年以来，我在德国多个地区的大中小型建筑公司担任过不同职位，从设计师到数据管理经理。在数据管理方面，我的经验包括在各种 ERP 系统 、CAD (BIM)、MEP 、FEM、CMS 中处理数据。我曾在工业、住宅、基础设施和公用事业建筑公司的建筑工程规划、计算和执行阶段参与优化、流程自动化以及分析、机器学习和数据处理工作。

自 2003 年以来，我一直从事开源软件和开放数据方面的工作。在此期间，我实施了许多网络项目--从网站和在线商店到使用开源解决方案和开放式内容管理系统的成熟网络应用程序。这些平台在许多方面类似于现代建筑 ERP，具有模块化架构、高适应性和可访问性。这些经验为我的专业方法奠定了基础--注重开源技术和协作开发文化。尊重开源和自由交流知识是我在建筑行业努力推广的理念。我致力于提高建筑行业的数据可访问性，创建了多个社交媒体社区来讨论数据开放性和开源技术在建筑行业中的应用，还创办了多家初创公司，开发解决方案来提供各种封闭系统和平台的数据访问。

我为专业界做出的贡献包括：作为演讲者参加了涉及 CAD 互操作性 (BIM) 、ERP 、4D-5D、LLM 机器学习和人工智能 的会议，并在欧洲建筑行业出版物上发表了文章。我的显著成就之一是创建了 "BIM 历史"[111] 。[111]，这是一份关于建筑行业数据管理重要软件解决方案的综合地图。我的 7 篇系列文章 "BIM 发展与游说游戏 "被翻译成多种语言，作为揭示数字标准隐藏动态的尝试得到了广泛认可。

这就是我从开采岩石到开采建筑数据并将其系统化的过程。我随时欢迎专业对话、新想法和合作项目。我非常乐意接受任何反馈意见，并期待您的留言或在我的社交媒体粉丝中看到您的身影。非常感谢您将本书读完！如果本书能帮助您更好地理解建筑行业的数据主题，我将非常高兴。

进一步关系

读者的意见对出版物的进一步发展和优先主题的选择起着重要作用。对于哪些观点被证明是有用的，哪些观点引起了质疑，需要进一步澄清或引用资料来源，读者的意见尤其宝贵。本书包括大量材料和分析，其中一些可能会引起争议或带有主观性。如果您在阅读过程中发现不准确之处、错误引用资料来源、逻辑不一致或错别字，我将非常感谢您的意见、想法或批评，您可以将这些意见、想法或批评发送至 boikoartem@gmail.com。或在 linkedin.com/in/boikoartem.上留言.

如果有人在社交媒体上提及《数据驱动的建筑》一书，我将不胜感激--分享阅读体验有助于传播有关开放数据和工具的信息，并支持我的工作。

译注

本书的翻译使用了人工智能技术。这使得翻译过程更加快捷。然而，与任何技术操作一样，可能会出现错误或不准确之处。如果您发现任何不正确或错误的翻译，请给我发电子邮件。您的意见将有助于提高翻译质量。

社区

在这里，您可以自由地提出问题，分享您的问题和解决方案：

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

其他技能和概念

除了在建筑行业中使用数据的关键原则之外，《数据驱动建筑》还涉及了数据驱动专业人员所必需的大量其他概念、方案和技能。其中一些内容仅作概述，但对实践至关重要。

感兴趣的读者可以访问 DataDrivenConstruction.io 网站，获取有关关键技能的其他资料链接。这些资料包括使用 Python 和 Pandas 、构建 ETL -流程、建筑 CAD 项目中的数据处理示例 、大数据系统以及建筑数据可视化和分析的现代方法。

在编写《数据驱动建筑》一书和所有案例研究时，使用了许多开源工具和软件。作者在此对以下解决方案的开发者和合著者表示感谢：

- Python 和 Pandas - 数据操作和自动化的支柱
- Scipy、NumPy、Matplotlib 和 Scikit-Learn - 用于数据分析和机器学习的库
- SQL 和 Apache Parquet - 用于存储和处理大量建筑数据的工具
- 开放源 CAD (BIM) 开放格式的开放数据工具
- N8n、Apache Airflow 、Apache NiFi - 协调和工作流自动化系统
- DeepSeek, LlaMa, Mistral - 开放源 LLM

特别感谢所有参与专业社区和社交网络开放数据和工具主题讨论的人员，他们的批评、意见和想法有助于完善本书的内容和结构。

请在 DataDrivenConstruction.io 网站上关注该项目的发展该网站不仅会发布书籍的更新和更正，还会发布新的章节、教程和应用所述技术的实例。

印刷版带来的最大便利

您现在持有的是《数据驱动的建筑》的免费电子版。为了更方便地工作和快速查阅资料，我们建议您关注印刷版：



■随手可得：印刷版书籍将成为可靠的工作工具，让您在任何工作环境中都能快速找到并使用正确的可视化图示和图表

■高质量插图：印刷版中的所有图像和图形均以最高质量呈现

■快速获取信息：导航简便，可做笔记、书签，随时随地使用图书。

通过购买本书的完整印刷版，您将获得一个舒适高效地处理信息的便捷工具：能够在日常工作中快速使用视觉材料，快速查找必要的方案并做笔记。此外，您的购买还支持了开放知识的传播。

订购本书印刷版，请访问：datadrivenconstruction.io/books



独特机遇

战略定位

我们为您提供在免费版 DataDrivenConstruction 上刊登广告的机会。该刊物的付费版在出版后的第一年就吸引了来自世界 50 多个国家（从拉丁美洲到亚太地区）的

专家的关注。如需讨论个别合作条款并获得有关投放机会的详细信息，请在官方门户网站 datadrivenconstruction.io 上填写反馈表，或致函书末所列联系人。



本书各章节可在 **datadrivenconstruction.io** 网站查阅

您可以在 "数据驱动的建筑 "网站上阅读《数据驱动的建筑》的章节，该书的各个章节将陆续发布，以便您可以快速找到所需的信息并将其用于工作中。您还可以找到许多其他类似主题的出版物，以及应用实例和解决方案，帮助您提高技能并将数据应用于建筑业。



本书最新版本

从官方网站下载

可在 datadrivenconstruction.io 网站上下载当前和最新版本的《DataDrivenConstruction》一书。如果您希望收到新书章节更新、实用技巧或新应用评论，请订阅时事通讯：

- 您将率先了解本书的新章节
- 获得有关建筑业分析和自动化的实用案例研究和建议
- 了解最新趋势、出版物和应用实例

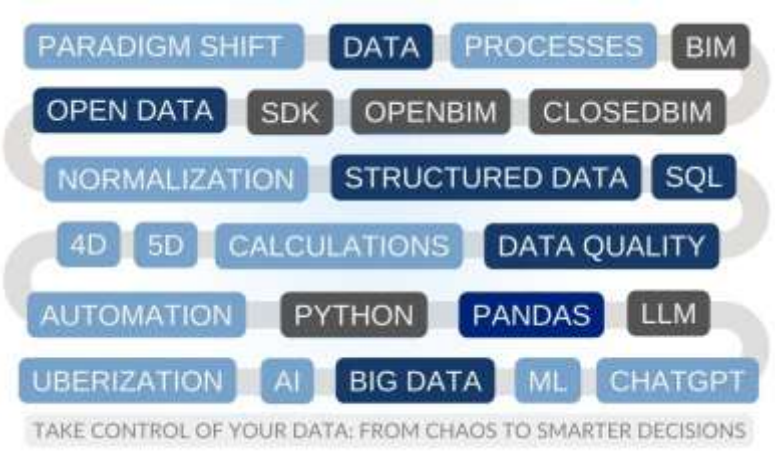
请访问 datadrivenconstruction.io 进行订阅！

数据驱动的建筑：

咨询、研讨会和培训

数据驱动建筑公司（DataDrivenConstruction）的培训计划和咨询已帮助全球数十家领先的建筑公司提高了效率、降低了成本并改善了解决方案的质量。DataDrivenConstruction 的客户包括建筑、咨询和 IT 公司

等数十亿欧元市场中的一些大型企业。



为什么选择我们?

- 相关性：谈论行业的主要趋势和见解
- 实用：通过 PoC 帮助专业人员高效解决日常任务。
- 个性化方法：为您的企业量身定制，使培训和建议的效益最大化

数据驱动施工团队的主要关注领域：

- 数据质量管理：帮助确定任务参数、收集需求 、验证和准备数据以进行自动处理。
- 数据挖掘**--**提取和构建数据：建立 ETL 流程，从电子邮件、PDF 、Excel 、图像和其他来源提取数据。
- **BIM** 和 **CAD** 分析 ： 从 RVT 文件 、IFC 、DWG 和其他 CAD (BIM) 格式中收集、构建和分析信息。
- 分析和数据转换：将分散的信息转化为结构化数据、分析、见解和决策。
- 数据集成和流程自动化：从自动创建文档到集成内部系统和外部数据库。

请联系 DataDrivenConstruction.io，了解使用自动化技术如何帮助贵公司取得切实的业务成果。

术语表

AI（人工智能）--人工智能；计算机系统执行通常需要人类智能的任务的能力，如模式识别、学习和决策。

Apache Airflow 是一个开源工作流协调平台，可让您使用 DAG（有向无环图）以编程方式创建、调度和跟踪工作流和 ETL。

Apache NiFi 是一款用于系统间数据流自动化的工具，专门用于数据路由和转换。

Apache Parquet 是一种高效的列式数据存储文件格式，专为在大数据分析系统中使用而优化。它具有显著的压缩和快速处理能力。

API（应用编程接口）--一种正式的接口，允许一个程序与另一个程序进行交互，而无需访问源代码，通过标准化的请求和响应来交换数据和功能。

属性 - 描述物体特征（如面积、体积、成本、材料）的特性或属性。

数据库是存储、管理和访问信息的有组织结构，用于高效的数据检索和处理。

BEP（**BIM** 执行计划）- 建筑信息模型实施计划，规定了在项目中实施 BIM 的目标、方法和流程。

海量数据--数量、种类和更新速度巨大的信息阵列，需要特殊技术进行处理和分析。

BI（商业智能）--商业智能；将数据转化为有意义的决策信息的流程、技术和工具。

BIM（**Building Information Modeling**，建筑信息模型）--建筑信息模型；创建和管理建筑项目物理和功能特征的数字表示过程，不仅包括三维模型，还包括性能、材料、时间和成本方面的信息。

黑盒**/**白盒--理解系统的方法：在第一种情况下，内部逻辑是隐藏的，只有输入和输出是可见的；在第二种情况下，处理过程是透明的，可供分析。

边框是一种几何结构，通过最小和最大 X、Y 和 Z 坐标来描述物体在三维空间中的边界，在物体周围形成一个 "框"。

BREP（边界表示法）是一种物体的几何表示法，它通过曲面的边界来定义物体。

CAD （计算机辅助设计）是一种计算机辅助设计系统，用于创建、编辑和分析建筑、施工、工程和其他行业的精确图纸和 3D 模型。

CAFM （计算机辅助设施管理）是一款财产和基础设施管理软件，包括空间规划、资产管理、维护和成本监控。

CDE（通用数据环境）--一个集中的数字空间，用于管理、存储、共享和协作设施生命周期各个阶段的项目信息。

卓越中心（**CoE**）是一个组织内的专门机构，负责开发某一特定领域的知识、制定标准和最佳做法、培训员工和支持引进创新。

CoClass 是一个现代化的第三代建筑元素分类系统。

概念数据模型是基本实体及其关系的高级表示，不包含属性细节，用于数据库设计的初始阶段。

CRM（客户关系管理）是一种客户关系管理系统，用于实现销售和服务流程的自动化。

DAG（有向无环图）是一种有向无环图，在数据协调系统（Airflow、NiFi）中用于确定任务序列和依赖关系。

Dash 是一个 Python 框架，用于创建基于网络的交互式数据可视化。

仪表盘 - 实时直观显示关键绩效指标和衡量标准的仪表盘。

以数据为中心的方法是一种优先考虑数据而不是应用程序或软件代码的方法，使数据成为组织的核心资产。

数据管理--一整套实践、流程和政策，确保数据在组织内的适当和有效使用，包括访问、质量和安全控制。

数据湖是一种存储设备，旨在以原始格式存储大量原始数据，直至使用。

Data Lakehouse 是一种将数据湖（Data **Lake**）的灵活性和可扩展性与数据仓库（DWH）的可管理性和性能相结合的架构方法。

数据驱动型施工是一种战略方法，在这种方法中，设施生命周期的每个阶段（从设计到运营）都由自动化的互联系统提供支持。这种方法提供了基于事实的持续学习，减少了不确定性，使公司能够实现可持续的行业领先地位。

数据驱动型集成商--一家专门从事将不同来源的数据结合起来并进行分析以做出管理决策的公司。

数据驱动法--一种将数据视为战略资产，根据对信息的客观分析而非主观意见做出决策的方法。

数据最小化--一种将数据减少到最有价值和最有意义的方法，从而简化信息的处理和分析。

数据沼泽（**Data Swamp**）--当信息在没有适当组织的情况下以不受控制的方式收集和存储时，就会出现大量分散的非结构化数据。

DataOps 是一种结合了 DevOps 原则、数据和分析的方法论，侧重于改进数据流的协作、集成和自动化。

信息数字化是将建筑活动的所有方面转化为适合分析、解释和自动化的数字形式的过程。

DataFrame - Pandas 库中的一种二维表格数据结构，其中行代表单个记录或对象，列代表其特征或属性。

描述性分析 - 分析历史数据，了解过去发生的事情。

诊断分析 - 分析数据以确定事情发生的原因。

甘特图是一种项目规划工具，在时间轴上用横条表示任务，让你直观地了解工作的顺序和持续时间。

DWH（数据仓库）是一个集中式数据仓库系统，它汇总来自多个来源的信息，对其进行结构化处理，并将其用于分析和报告。

ESG（环境、社会和治理）--一套评估公司或项目对环境、社会和治理影响的标准。

ELT（提取、加载、转换）是一个过程，首先从数据源提取数据并加载到存储库，然后为分析目的进行转换。

ETL（提取、转换、加载）是从各种来源提取数据，将其转换为所需格式，并加载到目标存储中进行分析的过程。

ER-图（实体**-**关系）--显示实体、实体属性和实体间关系的可视化图表，用于数据建模。

企业资源规划（**ERP**）是一个综合模块化的企业资源规划系统，用于管理和优化施工过程的各个方面。

特征 - 在机器学习中，用作模型输入的独立变量或属性。

物理数据模型 - 数据库结构的详细表示，包括表、列、数据类型、键和索引，针对特定 DBMS 进行了优化。

FPDF 是一个用于创建 PDF 文档的 Python 库。

Geometric Core 是一个软件组件，提供在 CAD、BIM 和其他工程应用中创建、编辑和分析几何对象的基本算法。

HiPPO（薪酬最高者的意见）--一种基于组织中薪酬最高者的意见而非客观数据的决策方法。

IDE（集成开发环境）--集成开发环境，用于编写、测试和调试代码的综合工具（如 PyCharm、VS Code、Jupyter Notebook）。

IDS（信息交付规范）是一种信息交付规范，它定义了项目不同阶段的数据要求。

IFC（工业基础类）是一种 BIM 数据交换格式，可确保不同软件解决方案之间的互操作性。

工业 **5.0** 是一种工业发展理念，它将数字化、自动化和人工智能的能力与人类潜能和环境可持续性相结合。

数据整合是将不同来源的数据整合到一个统一的系统中，以提供统一的信息视图。

信息孤岛是一种孤立的数据存储系统，不能与其他系统共享信息，从而阻碍了数据的有效利用。

物联网（**IoT**）的概念是将物理对象连接到互联网，以收集、处理和传输数据。

k-NN（**k-**最近邻）是一种机器学习算法，它根据与训练样本中最近邻的相似性对对象进行分类。

Kaggle 是一个数据分析和机器学习竞赛平台。

计算--计算一定计量单位（如 1 m^2 石膏板墙、1 m^3 混凝土）的建筑工程或工序费用。

KPI（关键绩效指标）是关键绩效指标，是用于评估公司或特定项目成功与否的可量化指标。

标签 - 在机器学习中，模型应预测的目标变量或属性。

学习算法 - 利用一组训练数据，在模型中找到与目标函数相对应的最佳假设的过程。

线性回归 - 一种模拟因变量与一个或多个自变量之间关系的统计方法。

LLM（大型语言模型）--大型语言模型是一种经过训练的人工智能，可以从海量数据中理解和生成文本，能够分析上下文并编写程序代码。

LOD（详细程度**/**开发程度）--模型的详细程度，决定了几何精度和信息内容。

逻辑数据模型是对实体、属性、键和关系的详细描述，反映了业务信息和规则，是概念模型和物理模型之间的中间阶段。

机器学习（**Machine Learning**）--一类人工智能技术，允许计算机系统在没有明确编程的情况下从数据中进行学习和预测。

Masterformat 是第一代分类系统，用于将施工规范划分为不同的章节和专业。

MEP（机械、电气、给排水）- 包括机械、电气和给排水组件的建筑工程系统。

网格是由顶点、边和面组成的三维物体的网格表示。

模型 - 在机器学习中，一组不同的假设，其中一个近似要预测或近似的目标函数。

数据建模是为在信息系统中实施数据及其关系创建结构化表示的过程，包括概念、逻辑和物理层面。

n8n 是一款开源工具，可通过低代码方法实现工作流程自动化和应用程序集成。

归一化 - 在机器学习中，将不同的数字数据统一到一个尺度上以方便处理和分析的过程。

逆向工程--通过分析物体的结构、功能和操作，研究其装置、功能和制造技术的过程。在数据方面--从专有格式中提取信息，供开放系统使用。

OCR（光学字符识别）是一种光学字符识别技术，可将文本图像（扫描文件、照片）转换为机器可读的文本格式。

OmniClass 是建筑信息管理的第二代国际分类标准。

本体论--将特定知识领域正规化的概念相互关系系统。

开放源代码 - 一种开发和发布开放源代码软件的模式，可供自由使用、研究和修改。

开放式 **BIM** 是开放式 BIM 的概念，涉及在不同软件解决方案之间使用开放式标准和格式进行数据交换。

开放标准--公开发布的实现特定任务的规范，允许不同系统交互和交换数据。

Pandas 是一个用于数据处理和分析的开源 Python 库，提供 DataFrame 和 Series 数据结构，可有效处理表格信息。

开放数据范式是一种数据处理方法，在这种方法中，任何人都可以免费使用、重复使用和传播信息。

参数法是一种建筑项目估算方法，使用统计模型根据项目参数估算价值。

PIMS（项目信息模型）是一个数字系统，旨在组织、存储和共享所有项目信息。

管道 - 一系列数据处理流程，从提取和转换到分析和可视化。

项目信息管理系统（**PMIS**）是一个项目管理系统，旨在对单个建筑项目进行详细的任务控制。

预测分析是分析的一个部分，它使用统计方法和机器学习来根据历史数据预测未来的结果。

规范性分析 - 分析的一个部分，它不仅能预测未来结果，还能提出实现预期结果的最佳行动建议。

专有格式是由特定公司控制的封闭数据格式，限制了共享信息的能力，增加了对特定软件的依赖。

QTO（工程量计算）是从设计文件中提取各要素的数量，以计算实现项目所需的材料数量。

质量管理系统 - 确保过程和结果符合既定要求的质量管理系统。

RAG（**Retrieval-Augmented Generation**，检索增强生成）是一种将语言模型的生成能力与从企业数据库中提取相关信息相结合的方法，可提高答案的准确性和相关性。

RDBMS（关系数据库管理系统）是一种关系数据库管理系统，它以相互关联的表格形式组织信息。

RegEx（正则表达式）是一种用于搜索和处理字符串的正式语言，允许您指定模板来检查文本数据是否符合某些标准。

回归是一种分析变量之间关系的统计方法。

CO_2 计算是一种估算与生产和使用建筑材料及工艺相关的二氧化碳排放量的方法。

资源估算法是在详细分析施工所需的所有资源（材料、人工、设备）的基础上进行估算的一种方法。

RFID（射频识别）是一种利用无线电信号自动识别物体的技术，用于追踪材料、机械和人员。

投资回报率（**ROI**）是反映利润与投入资金之间比率的指标，用于评估投资效果。

SaaS（软件即服务）是一种软件即服务模式，应用程序由提供商托管，并通过互联网提供给用户。

SCM（供应链管理）--供应链管理，包括协调和优化从材料采购到成品交付的所有流程。

数据孤岛是一个组织内部孤立的信息存储，没有与其他系统集成，因此难以共享数据，效率低下。

SQL（结构化查询语言）是一种结构化查询语言，用于处理关系数据库。

SQLite 是一种轻量级、可嵌入、跨平台的数据库管理系统，不需要单独的服务器，支持基本的 SQL 功能，广泛应用于移动应用程序和嵌入式系统。

结构化数据 - 以特定格式组织的信息，具有清晰的结构，如关系数据库或表格。

松散结构数据--具有部分组织和灵活结构的信息，如 JSON 或 XML，其中不同的元素可能包含不同的属性集。

实体是现实世界中的一个具体或抽象的物体，可以以数据的形式进行唯一识别、描述和表示。

监督学习 - 机器学习的一种类型，在这种类型中，算法在分区数据上进行训练，每个示例的预期结果都是已知的。

分类法是一种分层分类系统，用于根据共同特征对元素进行系统分类。

泰坦尼克数据集是用于训练和测试机器学习模型的常用数据集。

训练 - 机器学习算法分析数据以识别模式并形成模型的过程。

迁移学习是一种机器学习技术，它将为一项任务训练的模型作为另一项任务的起点。

转换 - 改变数据格式、结构或内容以供日后使用的过程。

数据要求 - 界定支持业务流程所需信息的结构、格式、完整性和质量的正式标准。

建筑业的 **Uber** 化是指在数字平台的影响下，传统的建筑业商业模式发生转变的过程，数字平台为客户和承包商之间提供了无中介的直接互动。

Uniclass 是英国广泛使用的第二代和第三代建筑元素分类系统。

USD（通用场景描述）是一种为计算机制图开发的数据格式，但由于其结构简单且独立于几何核心，已在工程系统中得到广泛应用。

数据验证是根据既定标准和要求检查信息的过程，以确保数据的准确性、完整性和一致性。

矢量数据库 - 一种专门的数据库类型，可将数据存储为多维矢量，以实现高效的语义搜索和对象比较。

矢量表示法（嵌入）是一种将数据转化为多维数字矢量的方法，可使机器算法高效地处理和分析信息。

VectorOps 是一种专注于处理、存储和分析多维矢量数据的方法，尤其适用于数字双胞胎和语义搜索等领域。

可视化 - 用图形表示数据，以便更好地感知和分析信息。

术语按英文名称的字母顺序分类。

参考文献和在线资料清单

[1] Gartner：《2017 年 IT 关键指标数据：已发布文件和指标索引》，2016 年 12 月 12 日。[Online].可查阅：https://www.gartner.com/en/documents/3530919。[访问日期：2025 年 3 月 1 日]。

[2] 毕马威，"熟悉的挑战 - 新的方法。2023 年全球建筑业调查"，2023 年 1 月 1 日。[Online].Available: https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2023/06/familiar-challenges-new-solutions-1.pdf.[Accessed: 5 March 2025].

[3] F.R.巴纳德，"一图胜千言"，10 Mapi 1927。[Online].可查阅：https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words。[参考日期：2025 年 3 月 15 日]。

[4] M.巴斯蒂安，"微软首席执行官萨蒂亚-纳德拉称自称 AGI 是'无稽的基准黑客'"，2025 年 2 月 21 日。[Online].Available: https://the-decoder.com/microsoft-ceo-satya-nadella-says-self-claiming-agi-is-nonsensical-benchmark-hacking/.[访问日期：2025 年 3 月 15 日]。

[5] W.E.论坛，"塑造建筑业未来的论坛--变革中的格局："，2016 年 1 月 1 日。[Online].Available: https://www3.weforum.org/docs/WEF_Shaping_the_Future_of_Construction.pdf.[访问日期：2025 年 3 月 2 日]。

[6] C.D. Gillespie，"粘土：地球在粘土时代的纠缠"，2024 年。[Online].Available: https://ufl.pb.unizin.org/imos/chapter/clay/.

[7] "公元前 3 世纪的纸莎草纸，语言为希腊语"，2024 年。[Online].Available: https://www.facebook.com/429710190886668/posts/595698270954525.

[8] "监测：利用现有工具"，1980 年。[Online].网址：https://pubmed.ncbi.nlm.nih.gov/10246720/。[发表日期：2025 年 3 月 15 日]。

[9] 普华永道，"数据驱动 学生需要什么才能在瞬息万变的商业世界中取得成功"，2015 年 2 月 15 日。

[Online].网址：https://www.pwc.com/us/en/faculty-resource/assets/PwC-Data-driven-paper-Feb2015.pdf。[访问日期：2025 年 3 月 15 日]。

[10] 斯堪斯卡美国公司，《秋季建筑市场趋势》，2023 年 11 月 2 日。[Online].Available: https://x.com/SkanskaUSA/status/1720167220817588714.

[11] "牛津精华语录（第 4 版）》，牛津大学出版社，2016 年。[Online].Available: https://www.oxfordreference.com/display/10.1093/acref/9780191826719.001.0001/q-oro-ed4-00006236.[访问日期：2025 年 3 月 1 日]。

[12] "名言：Sondergaard 论数据分析"。[Online].网址：https://www.causeweb.org/cause/resources/library/r2493。[访问日期：2025 年 3 月 15 日]。

[13] "全球对人工智能的兴趣如何推动数据管理市场的发展》，2024 年 5 月 28 日。[Online].网址：https://iot-analytics.com/how-global-ai-interest-is-boosting-data-management-market/。[访问日期：2025 年 3 月 15 日]。

[14] И.McCue，"ERP 历史"，2024 年。[Online].Available: https://www.netsuite.com/portal/resource/articles/erp/erp-history.shtml.

[15] erpscout，"ERP 价格：ERP 系统的成本是多少？[Online].Available: https://erpscout.de/en/erp-costs/.[访问日期：2025 年 3 月 15 日]。

[16] softwarepath, "What 1,384 ERP projects tell us about selecting ERP (2022 ERP report)"，2022 年 1 月 18 日。[Online].Available: https://softwarepath.com/guides/erp-report.[访问日期：2025 年 3 月 15 日]。

[17] 德勤，"数字资本项目中的数据驱动管理"，2016 年 12 月 16 日。[Online].网址：https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Real%20Estate/us-engineering-construction-data-driven-management-digital-capital-projects.pdf。[发表日期：2025 年 3 月 1 日]。

[18] 麦肯锡，《2025 年数据驱动型企业》，2022 年 1 月 28 日。[Online].Available:

https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025.[参考日期：2024 年 5 月 22 日]。

[19] 维基百科，"摩尔定律"。[Online].Available: https://en.wikipedia.org/wiki/Moore%27s_law.[访问日期：2025 年 3 月 15 日]。

[20] 埃森哲，"通过资本项目创造更多价值"，2020 年 1 月 1 日。[Online].Available: https://www.accenture.com/content/dam/accenture/final/a-com-migration/r3-3/pdf/pdf-143/accenture-industryx-building-value-capital-projects-highres.pdf.[访问日期：2024 年 3 月 3 日]。

[21] Б.马尔，《我们每天创造多少数据？每个人都应该阅读的令人震惊的数据"，2018 年。[Online].Available: 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] "每天产生多少数据？"，2024 年。[Online].Available: https://graduate.northeastern.edu/resources/how-much-data-produced-every-day/.

[23] T.沙利文，《人工智能与全球 "数据球"：到 2025 年人类将拥有多少信息？[Online].Available: https://www.datauniverseevent.com/en-us/blog/general/AI-and-the-Global-Datasphere-How-Much-Information-Will-Humanity-Have-By- 2025.html。

[24] Statista，"1454 至 1800 年间西欧各地区每半个世纪印刷书籍的总产量"，[互联网]。[Online].Available: https://www.statista.com/statistics/1396121/europe-book-production-half-century-region-historical/.[参考日期：2025 年 3 月 1 日]。

[25] "定价示例"，2024 年。[Online].Available: https://cloud.google.com/storage/pricing-examples.

[26] M.Ashare，"随着复杂性上升，企业外包数据存储"，2024 年 5 月 10 日。[在线]。可查阅：https://www.ciodive.com/news/enterprises-outsource-data-storage-complexity-rises/715854/。[访问日期：2025 年 3 月 15 日]。

[27] JETSOFTPRO, "SaaS 已死？微软首席执行官惊人预言解读》，2025 年 1 月 13 日。[在线]。Available: https://jetsoftpro.com/blog/saas-is-dead/.

[28] BG2 Pod，"Satya Nadella | BG2 w/ Bill Gurley & Brad Gerstner"，2024 年 12 月 12 日。[Online].Available: https://www.youtube.com/watch?v=9NtsnzRFJ_o.[访问日期：2025 年 3 月 15 日]。

[29] GoodReads，"蒂姆-伯纳斯-李"，[在线]。网址：https://www.goodreads.com/quotes/8644920-data-is-a-precious-thing-and-will-last-longer-than。[访问日期：2025 年 3 月 15 日]。

[30] 毕马威，"提示建筑 4.0：成败在此一举"，2023 年 1 月 1 日。[Online].Available: https://kpmg.com/ca/en/home/insights/2023/05/cue-construction-make-or-break-time.html.[访问日期：2025 年 3 月 5 日]。

[31] *И*.Deininger, B. Koch, R. Bauknecht, and M. Langhans, "Using digital models for decarbonising a production site: An example of connecting a building model, a production model and an energy model," 2024.[在线]。Available: 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] 麦肯锡，"重振建筑业：提高生产力的途径"，2017 年 2 月 1 日。[Online].Available: 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] 建筑工作组致副首相，"重新思考建筑业"，2014 年 10 月 1 日。[Online].可查阅：https://constructingexcellence.org.uk/wp-content/uploads/2014/10/rethinking_construction_report.pdf。

[34] 福布斯》，"没有观点，你只是另一个拥有数据的人"，2016 年 3 月 15 日。[Online].Available: https://www.forbes.com/sites/silberzahnjones/2016/03/15/without-an-opinion-youre-just-another-person-with-data/.[访问日期：2025 年 3 月 15 日]。

[35] Wikiquote，"查尔斯-巴贝奇"，[在线]。Available: https://en.wikiquote.org/wiki/Charles_Babbage.[访问日期：2025 年 3 月 15 日]。

[36] SAP，"最新研究发现，近半数高管相信人工智能而非自己"，2025 年 3 月 12 日。[Online].Available: https://news.sap.com/2025/03/new-research-executive-trust-ai/.[访问日期：2025 年 3 月 15 日]。

[37] 加拿大建筑协会和加拿大毕马威会计师事务所，2021 年，"数字世界中的建筑业"，2021 年 5 月 1 日。[Online].Available: https://assets.kpmg.com/content/dam/kpmg/ca/pdf/2021/05/construction-in-the-digital-age-report-en.pdf.[访问日期：2025 年 3 月 5 日]。

[38] ZCS，"解码第五次工业革命"。[Online].Available: https://www.pwc.in/decoding-the-fifth-industrial-revolution.html.[访问日期：2025 年 3 月 15 日]。

[39] M.K, Private Rights and Public Problems: The Global Economics of, 彼得森国际经济研究所，2012 年。经济学》，2012 年。

[40] F.N. a. Y. Z.哈佛商学院：Manuel Hoffmann，"开源软件的价值"，2024 年 1 月 24 日。[Online].Available: https://www.hbs.edu/faculty/Pages/item.aspx?num=65230.[访问日期：2025 年 3 月 15 日]。

[41] 海军成本分析中心空军成本分析机构，"软件开发成本估算手册"，2008 年 9 月 1 日。[Online].Available: https://www.dau.edu/sites/default/files/Migrated/CopDocuments/SW%20Cost%20Est%20Manual%20Vol%20I%20rev%2010.pdf.

[42] 麦肯锡，"提高建筑生产率"，[在线]。Available: https://www.mckinsey.com/capabilities/operations/our-insights/improving-construction-productivity.[发表日期：2025 年 3 月 15 日]。

[43] A.G. a. C.Syverson, "The Strange and Awful Path of Productivity in the US Construction Sector," 19 Jan. 2023.[Online].Available: https://bfi.uchicago.edu/insight/research-summary/the-strange-and-awful-path-of-productivity-in-the-us-construction-sector/.[访问日期：2025 年 3 月 1 日]。

[44] 麦肯锡，"实现建筑生产力不再是可有可无的"，2024 年 8 月 9 日。[Online].Available: https://www.mckinsey.com/capabilities/operations/our-insights/delivering-on-construction-productivity-is-no-longer-optional.[访问日期：2025 年 3 月 5 日]。

[45] ING 集团，"建筑业生产率落后导致建筑成本上升"，2022 年 12 月 12 日。[Online].Available: https://think.ing.com/articles/lagging-productivity-drives-up-building-costs-in-many-eu-countries/.[发表日期：2025 年 3 月 15 日]。

[46] M.Berman，"微软首席执行官的惊人预测：'代理将取代所有软件'"，2024 年 12 月 19 日。[Online].Available: https://www.youtube.com/watch?v=uGOLYz2pgr8.[访问日期：2025 年 3 月 15 日]。

[47] Business Insider，"Anthropic 首席执行官称，3 至 6 个月后，人工智能将编写软件开发人员负责的 90% 的代码"，2025 年 3 月 15 日。[Online].Available: https://www.businessinsider.com/anthropic-ceo-ai-90-percent-code-3-to-6-months-2025-3.[访问日期：2025 年 3 月 30 日]。

[48] Statista，"截至 2024 年 6 月全球数据库管理系统 (DBMS) 按类别的受欢迎程度比较"，2024 年 6 月。[Online].Available: https://www.statista.com/statistics/1131595/worldwide-popularity-database-management-systems-category/.[发布日期：2025 年 3 月 15 日]。

[49] DB-Engines, "DB-Engines Ranking," [Online].网址：https://db-engines.com/en/ranking。[参考日期：2025 年 3 月 15 日]。

[50] "Stack Overflow 开发人员调查 2023》，2024 年。[Online].Available: https://survey.stackoverflow.co/2023/.

[51] "SQL"，2024 年。[Online].Available: https://en.wikipedia.org/wiki/SQL.

[52] "结构化数据与非结构化数据：区别何在？"，2024 年。[在线]。Available: https://www.ibm.com/blog/structured-vs-unstructured-data/.

[53] DataDrivenConstruction, "PDF COMPARISON OF DATAFATS FOR CONSTRUCTION PROJECTS," 23 Apr 2024.[Online].Available: https://datadrivenconstruction.io/wp-content/uploads/2024/10/COMPARISON-OF-DATA-FORMATS-FOR-CONSTRUCTION-PROJECTS-1.pdf.

[54] "建筑信息模型白皮书网站"，2003 年。[Online].Available: https://web.archive.org/web/20030711125527/http:/usa.autodesk.com/adsk/servlet/item?id=2255342&siteID=123112.

[55] A.Boyko，"游说战争与 BIM 开发。第 5 部分：贝莱德是所有技术的主宰。企业如何控制开放源代码"，2024 年。[Online].Available: https://bigdataconstruction.com/autodesk-oracle-blackrock-open-source/.

[56] D.Ushakov, "Direct Modeling - Who and Why Needs It? A Review of Competitive Technologies," 14 11 2011.[Online].Available: https://isicad.net/articles.php?article_num=14805.[Date of address: 02 2025].

[57] C.伊士曼和 A. Cthers，《伊士曼、查尔斯和 Cthers》，1974 年 9 月。[Online].可查阅：https://files.eric.ed.gov/fulltext/ED113833.pdf。[访问日期：2025 年 3 月 15 日]。

[58] D.Ushakov, "Direct Modeling - Who and Why Needs It? A Review of Competitive Technologies," 11 November 2011.[Online].Available: https://isicad.net/articles.php?article_num=14805.[参考日期：2025 年 3 月 15 日]。

[59] D.魏斯伯格，"CAD 的历史"，2022 年 12 月 12 日。[Online].网址：https://www.shapr3d.com/blog/history-of-cad?utm_campaign=cadhistorynet.[参考日期：2025 年 3 月 15 日]。

[60] ADSK，"建筑信息模型白皮书"，2002 年。[Online].Available: https://web.archive.org/web/20060512180953/http:/images.adsk.com/apac_sapac_main/files/4525081_BIM_WP_Rev5.pdf#expand.[参考日期：2025 年 3 月 15 日]。

[61] ADSK，"建筑信息建模实践白皮书"，[在线]。Available: https://web.archive.org/web/20060512181000/http:/images.adsk.com/apac_sapac_main/files/

4525077_BIM_in_Practice.pdf.[访问日期：2025 年 3 月 15 日]。

[62] A.Boyko，"游说战争与 BIM 开发。第二部分：开放式 BIM VS 封闭式 BIM。欧洲 VS 世界其他地区"，2024 年。[Online].Available: 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, "Lobbykriege um Daten im Bauwesen | Techno-Feudalismus und die Geschichte von BIMs," 2024.[Online].Available: https://youtu.be/S-TNdUgfHxk?si=evM_v28KQbGOG0k&t=1360.

[64] ADSK，"BIM 白皮书"，2002 年。[Online].Available: https://web.archive.org/web/20060512180953/http:/images.autodesk.com/apac_sapac_main/files/4525081_BIM_WP_Rev5.pdf#expand.[发布日期：2025 年 3 月 15 日]。

[65] ADSK, "Integrated Design-Through-Manufacturing: Benefits and Rationale," [Online].Available: https://web.archive.org/web/20010615093351/http:/www3.adsk.com:80/adsk/files/734489_Benefits_of_MAI.pdf.[访问日期：2025 年 3 月 15 日]。

[66] M.Shacklett, "结构化数据和非结构化数据：主要区别"，2024 年。[在线]。Available: https://www.datamation.com/big-data/structured-vs-unstructured-data/.

[67] K.Woolard, "Making sense of the growth of unstructured data," 2024.[在线]。Available: 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，"互操作性不足的成本分析"，2004 年。[Online].Available: https://nvlpubs.nist.gov/nistpubs/gcr/2004/nist.gcr.04-867.pdf.[Date of access: 02 2025].

[69] CrowdFlower，《2016 年数据科学报告》，2016 年。[Online].Available: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf.[访问日期：2025 年 3 月 15 日]。

[70] Analyticsindiamag，《数据科学家最耗时的 6 项任务》，2019 年 5 月 15 日。[Online].Available: https://analyticsindiamag.com/ai-trends/6-tasks-data-scientists-spend-the-most-time-doing/.

[71] BizReport，"报告：数据科学家花费大量时间清理"，2015 年 7 月 6 日。[Online].网址：https://web.archive.org/web/20200824174530/http://www.bizreport.com/2015/07/report-data-scientists-spend-bulk-of-time-cleaning-up.html。[访问日期：2025 年 3 月 5 日]。

[72] S.霍金，"科学 AMA 系列：斯蒂芬-霍金 AMA 答案！"，2015 年 7 月 27 日。[Online].Available: https://www.reddit.com/r/science/comments/3nyn5i/science_ama_series_stephen_hawking_ama_answers/.[访问日期：2025 年 3 月 15 日]。

[73] Б.Cyphers and K. Doctorow, "Privacy without Monopoly: Data Protection and Interoperability," 2024.[Online].Available: https://www.eff.org/wp/interoperability-and-privacy.

[74] 麦肯锡全球研究院，"开放数据：利用流动信息释放创新和绩效"，2013 年 10 月 1 日。[Online].可查阅：https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information。[访问日期：2025 年 3 月 15 日]。

[75] A.Boyko, "建筑行业开放数据的斗争。AUTOLISP、intelliCAD、openDWG、ODA 和 openCASCADE 的历史"，2024 年 5 月 15 日。[Online].Available: https://boikoartem.medium.com/the-struggle-for-open-data-in-the-construction-industry-2b97200e6393.[Date of address: 16 02 2025].

[76] 维基百科，"微软与开放源码"，[在线]。Available: https://en.wikipedia.org/wiki/Microsoft_and_open_source.[访问日期：2025 年 3 月 15 日]。

[77] 时代》周刊，"开放式与封闭式人工智能模型之间的差距可能正在缩小。为何如此重要》，2024 年 11 月 5 日。[Online].Available: https://time.com/7171962/open-closed-ai-models-epoch/.[访问日期：2025 年 3 月 15 日]。

[78] The Verge，"谷歌超过四分之一的新代码由人工智能生成"，2024 年 10 月 29 日。[Online].Available: https://www.theverge.com/2024/10/29/24282757/google-new-code-generated-ai-q3-2024.[访问日期：2025 年 3 月 15 日]。

[79] 麦肯锡数字公司，"使用 GPU 加速分析处理的商业案例"，2020 年 12 月 15 日。[在线]。见 https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/the-

business-case-for-using-gpus-to-accelerate-analytics- processing。[访问日期：2025 年 3 月 15 日]。

[80] 普华永道，"普华永道开源监测 2019"，2019 年。[Online].Available: https://www.pwc.de/de/digitale-transformation/open-source-monitor-research-report-2019.pdf.[Accessed 15 Mar. 2025].

[81] 特拉弗斯-史密斯：《公开的秘密：开放源码软件》，2024 年。[Online].网址：https://www.traverssmith.com/knowledge/knowledge-container/the-open-secret-open-source-software/。[访问日期：2025 年 3 月 15 日]。

[82] 德勤，"企业转型中的数据传输过程"，2021 年。[Online].Available: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/finance/us-the-data-transfer-process-in-corporate-transformations.pdf.[参考日期：2025 年 3 月 15 日]。

[83] gov.uk，"政府项目交付中的数据分析和人工智能"，2024 年 3 月 20 日。[Online].Available: https://www.gov.uk/government/publications/data-analytics-and-ai-in-government-project-delivery/data-analytics-and-ai-in-government-project-delivery.[Date of access: 5 March 2025].

[84] "名言原点：一切都应尽可能简单，但不能更简单"，2011 年 5 月 13 日。[Online].网址：https://quoteinvestigator.com/2011/05/13/einstein-simple/.[访问日期：2025 年 3 月 15 日]。

[85] "Transformer（深度学习架构）"。[在线]。网址：https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)。[访问日期：2025 年 3 月 15 日]。

[86] "Python 软件包下载统计"，2024 年。[在线]。Available: https://www.pepy.tech/projects/pandas.

[87] Interview Bit，"十大 Python 库"，2023 年。[在线]。Available: https://www.interviewbit.com/blog/python-libraries/#:~:text=With%20more%20than%20137%2C000%20libraries,data%20manipulation%2C%20and%20many%20more.[访问日期：2025 年 3 月 30 日]。

[88] "英伟达™（NVIDIA®）和惠普™（HP®）为工作站上的数据科学和生成式人工智能增添动力》，

2025 年 3 月 7 日。[在线]。Available: https://nvidianews.nvidia.com/news/nvidia-hp-supercharge-data-science-generative-ai-workstations.[访问日期：2025 年 3 月 15 日]。

[89] P.Orac，"如何在数秒内处理数百万行的 DataFrame"，2024 年。[在线]。可登录：https://towardsdatascience.com/how-to-process-a-dataframe-with-millions-of-rows-in-seconds。

[90] Ç.Uslu，"什么是 Kaggle？"，2024 年。[Online].Available: https://www.datacamp.com/blog/what-is-kaggle.

[91] "英伟达™（NVIDIA®）首席执行官黄仁勋在 2024 年计算机展上发表主题演讲，"2024 年 6 月 2 日。[Online].可登录 https://www.youtube.com/live/pKXDVsWZmUU?si=Z3Rj1Las8wiPII2w。[Accessed 15 March 2025].

[92] "成员：创始人和公司成员"，2024 年。[Online].Available: https://www.opendesign.com/member-showcase.

[93] A.Boyko，"变革时代：IFC 已成为过去或 ADSK 和其他 CAD 供应商为何愿意放弃 IFC 换取 USD 的 14 个关键事实"，2024 年 11 月 24 日。[Online].Available: 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.[访问日期：2025 年 2 月 23 日]。

[94] A.Boyko，"后 BIM 世界。向数据和流程的过渡以及建筑业是否需要语义、格式和互操作性"，2024 年 12 月 20 日。[Online].Available: https://boikoartem.medium.com/the-post-bim-world-7e35b7271119.[发表日期：2025 年 2 月 23 日]。

[95] N.I.o.Health, "NIH STRATEGIC PLAN FOR DATA SCIENCE," 2016.[Online].Available: https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.[Date of access: 23 February 2025].

[96] 哈佛商业评论》，"糟糕的数据让美国每年损失 3 万亿美元"，2016 年 9 月 22 日。[Online].见 https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year。

[97] Delpha，"数据质量的影响"，2025 年 1 月 1 日。[Online].Available: https://delpha.io/impacts-of-

data-quality/.

[98] W.B.D.Guide, "Design for Maintainability: The Importance of Operations and Maintenance Considerations During the Design Phase of Construction Projects," [Online].[Online].Available: https://www.wbdg.org/resources/design-for-maintainability.[访问日期：2025 年 3 月 15 日]。

[99] O.o.D.C.P. a. Oversight，"军事系统和设备腐蚀防控规划指南"，2014 年 4 月。[Online].Available: https://www.dau.edu/sites/default/files/Migrated/CopDocuments/CPC%20Planning%20Guidebook%204%20Feb%2014.pdf.[访问日期：2025 年 3 月 15 日]。

[100] Gartner，《数据质量：准确洞察的最佳实践》，2025 年 1 月 1 日。[在线]。Available: https://www.gartner.com/en/data-analytics/topics/data-quality.

[101] "想要钉子"[Online].网址：https://en.wikipedia.org/wiki/For_Want_of_a_Nail。[访问日期：2025 年 3 月 15 日]。

[102] 麦肯锡全球研究院，《开放数据：利用流动信息释放创新和绩效》，2013 年 10 月。[Online].可查阅：
https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/open%20data%20unlocking%20innovation%20and%20performance%20with%20liquid%20i nformation/mgi_open_data_fullreport_oct2013.pdf。[访问日期：2025 年 3 月 15 日]。

[103] EY, "碳中和之路"，2023 年 3 月 10 日。[Online].Available: https://www.ey.com/ru_kz/services/consulting/the-path-to-carbon-neutrality.[访问日期：2025 年 3 月 15 日]。

[104] 普华永道，"ESG 意识"，2024 年 7 月 1 日。[Online].Available: https://www.pwc.com/kz/en/assets/esg-awareness/kz-esg-awareness-rus.pdf.[访问日期：2025 年 3 月 15 日]。

[105] G.Hammond, "Embodied Carbon - The Inventory of Carbon and Energy (ICE)," 2024.[Online].Available: https://greenbuildingencyclopaedia.uk/wp-content/uploads/2014/07/Full-BSRIA-ICE-guide.pdf.

[106] "CO2_计算所含碳量"，2024 年。[Online].Available: https://github.com/datadrivenconstruction/CO2_calculating-the-embodied-carbon.

[107] 麦肯锡，"想象建筑业的数字化未来"，2016 年 6 月 24 日。[Online].Available: https://www.mckinsey.com/capabilities/operations/our-insights/imagining-constructions-digital-future.[访问日期：2025 年 2 月 25 日]。

[108] Bund der Steuerzahler Deutschland e.V.，"Das Schwarzbuch"，2024 年 10 月 10 日。[Online].Available: https://steuerzahler.de/aktuelles/detail/das-schwarzbuch-202425/.[参考日期：2025 年 3 月 15 日]。

[109] SAS，"数据湖和数据仓库--了解两者的区别"，[在线]。可查阅：https://www.sas.com/en_is/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html。[参考日期：2025 年 3 月 15 日]。

[110] ADSK，《建筑信息建模》，2002 年。[Online].Available: https://www.laiserin.com/features/bim/autodesk_bim.pdf.[发表日期：2025 年 3 月 15 日]。

[111] A.Boyko，"BIM 历史地图"，2024 年。[Online].Available: https://bigdataconstruction.com/history-of-bim/.

[112] A.S.Borkowski, "各组织和标准对 BIM 的定义"，2023 年 12 月 27 日。[Online].Available: https://encyclopedia.pub/entry/53149.[发表日期：2025 年 3 月 5 日]。

[113] CAD Vendor，"OPEN BIM 计划"，2012 年。[在线]。Available: https://web.archive.org/web/20140611075601/http://www.graphisoft.com/archicad/open_bim/.[参考日期：2025 年 3 月 30 日]。

[114] 维基百科，"工业基础课"。[Online].Available: https://en.wikipedia.org/wiki/Industry_Foundation_Classes.[访问日期：2025 年 3 月 15 日]。

[115] 维基百科，"IGES"。[Online].Available: https://en.wikipedia.org/wiki/IGES.[参考日期：2025 年 3

月 30 日]。

[116] A.Boyko，"CAD（BIM）的历史"，2021 年 12 月 15 日。[Online].Available: https://miro.com/app/board/o9J_laML2cs=/.[发表日期：2025 年 2 月 24 日]。

[117] T.K.K.A.O.F.B.C.E.L. H. H. H. E. L. P. N. S. H. T. J. v. L. H. H. G. D. H. T. K. C.L.A.W.J. S. Francesca Noardo，"IFC 软件支持参考研究：GeoBIM 基准 2019 - 第一部分"，2021 年 1 月 8 日。[Online].Available: https://arxiv.org/pdf/2007.10951.[访问日期：2025 年 3 月 5 日]。

[118] И.Rogachev，"Let's Talk BIM：Maxim Nechiporenko | Renga | IFC | Domestic BIM"，2021 年 4 月 13 日。[Online].Available: https://www.youtube.com/watch?t=3000&v=VO3Y9uuzF9M&feature=youtu.be.[Date of address: 5 March 2025].

[119] D.Ares，"房地产中的 RETS：为什么它对效率和增长至关重要"，2024 年 12 月 17 日。[Online].Available: https://www.realpha.com/blog/rets-importance-in-real-estate-explained.[访问日期：2025 年 3 月 5 日]。

[120] "弹性令牌成本"，2024 年。[Online].Available: https://www.adsk.com/buying/flex?term=1-YEAR&tab=flex.

[121] A.Boyko, "Forget BIM and democratise access to data (17. Kolloquium Investor - Hochschule - Bauindustrie)," 2024.[Online].Available: https://www.bim.bayern.de/wp-content/uploads/2023/06/Kolloquium-17-TUM-Bauprozessmanagment-und-Bay-Bauindustrie.pdf.

[122] Д.Hill, D. Foldesi, S. Ferrer, M. Friedman, E. Loh, and F. Plaschke, "Solving the construction industry productivity puzzle," 2015.[Online].Available: https://www.bcg.com/publications/2015/engineered-products-project-business-solving-construction-industrys-productivity-puzzle.

[123] "SCOPE - Projektdatenumgebung und Modellierung multifunktionaler Bauprodukte mit Fokus auf die Gebäudehülle，"2018 年 1 月 1 日。[Online].Available: https://www.ise.fraunhofer.de/de/forschungsprojekte/scope.html.[访问日期：2025 年 3 月 2 日]。

[124] Apple.com，"Pixar、Adobe、Apple 和 NVIDIA 成立 OpenUSD 联盟，推动 3D 内容的开放标准"，2023 年 8 月 1 日。[Online].Available: https://www.apple.com/newsroom/2023/08/pixar-adobe-apple-adsk-and-nvidia-form-alliance-for-openusd/.[访问日期：2025 年 3 月 2 日]。

[125] AECmag，"ADSK 的粒度数据战略"，2024 年 7 月 25 日。[Online].Available: https://aecmag.com/technology/autodesks-granular-data-strategy/.[访问日期：2025 年 3 月 15 日]。

[126] A.Boyko，"变革时代：IFC 已成为过去或 ADSK 和其他 CAD 供应商为何愿意放弃 IFC 换取 USD 的 14 个关键事实"，2024 年 11 月 24 日。[Online].Available: 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.[Date of address: 23 fevryall 2025].

[127] A.Boyko, "ENG BIM Cluster 2024 | The Battle for Data and Application of LLM and ChatGPT in the Construction," 7 August 2024.[Online].Available: ENG BIM Cluster 2024 | The Battle for Data and Application of LLM and ChatGPT in the Construction.[Accessed: 15 March 2025].

[128] "杰弗里-扎尔德曼演讲"，2008 年 5 月 6 日。[Online].Available: https://zeldman.com/2008/05/06/content-precedes-design/.[访问日期：2025 年 3 月 15 日]。

[129] A.Boyko, "DWG Analyse with ChatGPT | DataDrivenConstruction," 5 Mar.[Online].Available: https://www.kaggle.com/code/artemboiko/dwg-analyse-with-chatgpt-datadrivenconstruction.[参考日期：2025 年 3 月 15 日]。

[130] 麦肯锡，《麦肯锡数字和人工智能时代的竞争指南》，2023 年。[Online].Available: https://www.mckinsey.com/featured-insights/mckinsey-on-books/rewired.[发表日期：2025 年 3 月 30 日]。

[131] 福布斯》，"讲数据故事：每个人都需要的数据科学基本技能"，2016 年 3 月 31 日。[Online].可登录：https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/。[访问日期：2025 年 3 月 15 日]。

[132] J.Bertin，"图形和图形信息处理"，2011 年 9 月 8 日。[Online].可查阅：

https://books.google.de/books/about/Graphics_and_Graphic_Information_Process.html?id=csqX_xnm4tcC&redir_esc=y。[访问日期：2025 年 3 月 15 日]。

[133] CauseWeb, "Wells/Wilks on Statistical Thinking," [Online].Available: https://www.causeweb.org/cause/resources/library/r1266.[访问日期：2025 年 3 月 15 日]。

[134] Ministrymagazine，"科学如何发现创造"，1986 年 1 月。[Online].网址：https://www.ministrymagazine.org/archive/1986/01/how-science-discovered-creation。[参考日期：2025 年 3 月 15 日]。

[135] BCG，《数据驱动的转型：立即大规模加速》，2017 年 5 月 23 日。[Online].网址：https://www.bcg.com/publications/2017/digital-transformation-transformation-data-driven-transformation。[2024 年 5 月 15 日访问]。

[136] "如何构建数据架构以推动创新--今天和明天"，2020 年 6 月 3 日。[Online].网址：https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/how-to-build-a-data-architecture-to-drive-innovation-today-and-tomorrow。[发表日期：2025 年 3 月 15 日]。

[137] 牛津大学，"伍德罗-威尔逊 1856-1924"。[Online].Available: https://www.oxfordreference.com/display/10.1093/acref/9780191866692.001.0001/q-oro-ed6-00011630.[演讲日期：2025 年 3 月 15 日]。

[138] "转换器"，2024 年。[Online].Available: https://datadrivenconstruction.io/index.php/convertors/.

[139] 普华永道，"Sizing the prize What's the real value of AI for your business and how can you capitalise?"，2017 年 1 月 1 日。[在线]。可查阅：https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf。[访问日期：2025 年 2 月 18 日]。

[140] "在建管道"，2024 年。[Online].Available: https://datadrivenconstruction.io/index.php/pipeline-in-construction/.

[141] 维基百科，"Apache NiFi"，2025 年 1 月 1 日。[Online].网址：

https://de.wikipedia.org/wiki/Apache_NiFi。[访问日期：2025 年 3 月 5 日]。

[142] n8n，"Gmail 人工智能自动回复：为收到的电子邮件创建回复草稿"，2024 年 5 月 1 日。[Online].Available: https://n8n.io/workflows/2271-gmail-ai-auto-responder-create-draft-replies-to-incoming-emails/.[Accessed 15 March 2025].

[143] n8n，"利用 Zillow API、Google Sheets 和 Gmail 实现房地产每日交易自动化"，2025 年 3 月 1 日。[在线]。可查阅：https://n8n.io/workflows/3030-real-estate-daily-deals-automation-with-zillow-api-google-sheets-and-gmail/。[访问日期：2025 年 3 月 15 日]。

[144] B.T.O'Neill，"分析、人工智能和大数据项目的失败率=85%--呀！"，2025 年 1 月 1 日。[Online].Available: https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/.

[145] J.内曼：《论代表性方法的两个不同方面：分层抽样法和目的性选择法》，牛津大学出版社，1934 年。

[146] T.J. S. a. J. S. Jesse Perla, "A Problem that Stumped Milton Friedman," Quantitative Economics with Julia, 1 Jan.[Online].Available: https://julia.quantecon.org/dynamic_programming/wald_friedman.html.[访问日期：2024 年 5 月 1 日]。

[147] T.Landsall-Welfair, Forecasting the nation's current mood, Significance, 2012.

[148] A.Boyko, "San Francisco.建筑业 1980-2019》，2024 年。[Online].Available: https://www.kaggle.com/search?q=San+Francisco.+Building+sector+1980-2019.

[149] A.Boyko, "Kaggle: RVT IFC Files 5000 Projects," 2024.[Online].Available: https://www.kaggle.com/datasets/artemboiko/rvtifc-projects.

[150] CFMA，"用互联建筑为未来做好准备"，[在线]。[Online].Available: https://cfma.org/articles/preparing-for-the-future-with-connected-construction.[访问日期：2025 年 3 月 15 日]。

[151] 思科，"思科调查显示近四分之三的物联网项目失败"，2017 年 5 月 22 日。[Online].Available:

https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2017/m05/cisco-survey-reveals-close-to-three-fourths-of-iot-projects-are-failing.html.

[152] "植物化石保存所需的条件"，2024 年。[Online].Available: https://ucmp.berkeley.edu/IB181/VPL/Pres/PresTitle.html.

[153] "贝莱德的芬克谈债券、并购、美国经济衰退和大选：访谈全文》，2023 年。[Online].Available: https://www.bloomberg.com/news/videos/2023-09-29/blackrock-s-fink-on-m-a-recession-election-full-intv-video.

[154] cio，"12 次著名的人工智能灾难"，2024 年 10 月 2 日。[Online].Available: https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html.[访问日期：2025 年 3 月 15 日]。

[155] G.卡斯帕罗夫，《深度思考》，PublicAffairs，2017。

[156] 维基百科，"Kaggle"，2025 年 1 月 1 日。[Online].Available: https://en.wikipedia.org/wiki/Kaggle.[访问日期：2025 年 3 月 15 日]。

[157] Kaggle，"泰坦尼克号--从灾难中学习机器"，2025 年 1 月 1 日。[Online].Available: https://www.kaggle.com/competitions/titanic/overview.[Accessed March 10, 2025].

[158] Ш.Johri，《创建 ChatGPT：从数据到对话》，2024 年。[Online].Available: https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/.

[159] П.多明戈斯，"关于机器学习的一些实用知识"，2024 年。[在线]。Available: https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.

[160] J.萨拉马戈，"名言警句"，[在线]。网址：https://www.goodreads.com/quotes/215253-chaos-is-merely-order-waiting-to-be-deciphered.[访问日期：2025 年 3 月 17 日]。

[161] 英伟达™（NVIDIA®），"利用全新英伟达™（NVIDIA®）NeMo Curator 分类器模型增强训练数据"，2024 年 12 月 19 日。[在线]。网址：https://developer.nvidia.com/blog/enhance-your-training-data-with-new-nvidia-nemo-curator-classifier-models/。[2025 年 3 月 25 日访问]。

[162] "英伟达™（NVIDIA®）宣布重大发布宇宙世界基础模型和物理人工智能数据工具，"2025 年 3 月 18 日。[Online].Available: https://nvidianews.nvidia.com/news/nvidia-announces-major-release-of-cosmos-world-foundation-models-and-physical-ai-data-tools.[访问日期：2025 年 3 月 25 日]。

[163] NVIDIA，"NVIDIA Isaac Sim"，[互联网上]。网址：https://developer.nvidia.com/isaac/sim。[2025 年 3 月 25 日访问]。

[164] M.季度》，"数字战略为何失败"，2018 年 1 月 25 日。[Online].Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-digital-strategies-fail.[访问日期：2025 年 3 月 15 日]。

[165] M.J.佩里，"我最喜欢的米尔顿-弗里德曼语录"，2006 年 11 月 17 日。[Online].Available: https://www.aei.org/carpe-diem/my-favorite-milton-friedman-quotes/.[访问日期：2025 年 3 月 1 日]。

[166] J.A.Wheeler, "Information, Physics, Quantum: the search for links," 1990.

[169] A.Boyko，"游说战争与 BIM 开发。第 5 部分：贝莱德是所有技术的主宰。企业如何控制开放源代码"，2024 年。[Online].Available: https://boikoartem.medium.com/lobbyist-wars-and-the-development-of-bim-d72ad0111a7d.

[170] T.Krijnen 和 J.Beetz, "A SPARQL query engine for binary-formatted IFC building models," Advanced Engineering Informatics, 2024.

[171] "2021 年按企业规模划分的英国建筑业企业数量"，2024 年。[Online].Available: https://www.statista.com/statistics/677151/uk-construction-businesses-by-size/.

[172] "5,000 个国际金融公司和区域技术合作项目"，2024 年。[Online].Available: https://www.kaggle.com/code/artemboiko/5000-projects-ifc-rvt-datadrivenconstruction-io.

[173] M.Popova, "It from Bit: Pioneering Physicist John Archibald Wheeler on Information, the Nature of Reality, and Why We Live in a Participatory Universe," 2008.[Online].Available:

https://www.themarginalian.org/2016/09/02/it-from-bit-wheeler/.[参考日期：2025 年 2 月]。

[174] 关于建筑数据的游说战争丨技术封建主义与 BIM 的隐秘历史。[电影]。德国：Artyom Boiko，2023 年。

[175] A.Boyko, "CHATGPT WITH REVIT AND IFC | Automatic retrieval of documents and data from projects," 16 Nov.[Online].Available: https://www.youtube.com/watch?v=ASXolti_YPs&t.[Date of access: 2 March 2025].

[176] M.&.公司，"获取数字化转型全部价值的三项新任务"，2022 年 1 月 22 日。[Online].Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/three-new-mandates-for-capturing-a-digital-transformations-full-value.[访问日期：2025 年 2 月 15 日]。

[177] 毕马威，"数字世界中的建筑业"，2021 年 5 月 1 日。[Online].Available: https://assets.kpmg.com/content/dam/kpmg/ca/pdf/2021/05/construction-in-the-digital-age-report-en.pdf。[访问日期：2024 年 4 月 5 日]。

[178] 毕马威会计师事务所，"Cue Construction 4.0: Make-or-Break Time"，2023 年 3 月 17 日。[Online].Available: https://kpmg.com/ca/en/home/insights/2023/05/cue-construction-make-or-break-time.html.[访问日期：2025 年 2 月 15 日]。

[179] O.Business, "Satya Nadella Reveals 'How AI Agents Will Disrupt SaaS Models'," 10 Jan. 2025.[Online].Available: https://www.outlookbusiness.com/artificial-intelligence/microsoft-ceo-satya-nadella-reveals-how-ai-agents-will-disrupt-saas-models.[访问日期：2025 年 3 月 15 日]。

[180] 福布斯》，"调查称，清理大数据：最耗时、最不愉快的数据科学任务"，2016 年 3 月 23 日。[在线]。可查阅：https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/。[访问日期：2025 年 3 月 15 日]。

[181] 英国外交、联邦和发展办公室，"2024 至 2030 年数字发展战略"，2024 年 3 月 18 日。[Online].Available: https://www.gov.uk/government/publications/digital-development-strategy-2024-to-2030/digital-development-strategy-2024-to-2030.[访问日期：2025 年 3 月 15 日]。

[182] "建筑设计行业的愿景与战略》，2003 年 11 月 7 日。[Online].网址：https://web.archive.org/web/20030711125527/http:/usa.adsk.com/adsk/servlet/item?id=2255342&siteID=123112。[发表日期：2025 年 3 月 5 日]。

[183] M.Bocharov，"信息建模"，2025 年 3 月。[Online].网址：https://www.litres.ru/book/mihail-evgenevich-bocharov/informacionnoe-modelirovanie-v-rossii-71780080/chitat-onlayn/?page=5.[发表日期：2025 年 3 月 15 日]。

[184] "集成设计-直通制造：效益与理由"，2000 年。[Online].可查阅：https://web.archive.org/web/20010615093351/http:/www3.autodesk.com:80/adsk/files/734489_Benefits_of_MAI.pdf。[参考日期：2025 年 3 月 25 日]。

[185] CAD Vendor，"开放 BIM 计划是一场营销活动"，2012 年 3 月 12 日。[Online].Available: https://web.archive.org/web/20120827193840/http:/www.graphisoft.com/openbim/.[发表日期：2025 年 3 月 30 日]。

主题索引

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
4IR, 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
友邦保险公司，289
AIM, 3, 289
上午，14，84，151，153
阿帕奇气流, 311, 361, 362, 363, 364, 366, 367, 369, 371, 399, 494
Apache NiFi、116、311、361、367、368、371、399、482、494
阿帕奇 ORC，62，330，378
阿帕奇镶木地板，62、67、380、381、494
国际出版商协会，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
大数据，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
虚化，320，337
炸弹、76、77、79、257、263
边界框，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, 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
咖啡管理条例》第 14 条、第 24 条、第 62 条、第 84 条、第 151 条、第 153 条、第 172 条、第 233 条、第 278 条、第 326 条、第 356 条、第 387 条、第 458 条、第 473 条
CAM, 16, 78
资本支出，14，82
CDE, 84, 175, 388, 389, 390
ChatGPT, 103, 104, 109, 110, 124, 129, 221, 303, 307, 343, 349, 421
克劳德、103、104、109、129、133、166、219、221、300、329、343、349、381、429、443
CO_2, 72
CO_2, 229, 234, 235, 236, 237, 238, 239, 243
COBie, 156, 289, 292.
欧洲委员会，56，168，169，170，477
副驾驶员，114，116，459
cpixml, 143, 272, 273, 276, 277, 279, 285, 296
CPM, 14, 17, 62, 166, 175, 233, 326, 473
传播和信息管理系统，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, 354, 356, 373, 377, 378, 379, 380, 401, 407, 414, 474, 480

D

裁军事务部，276，277，278，280，281，284，285，414
达格，362，363，365，366
Dash, 320, 336, 337.
数据治理，373、395、396、398、400、401
数据湖，214，373，376，384，385，386，387，388，389，390，400
数据湖房，67，373，386，387，388
数据极简主义，373、395、396、400、401
数据沼泽，373、395、397、401
数据仓库，382，383，400
数据即服务，487
数据驱动，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
数据收集网络，8，140，186，227，357
DWG, 8, 70, 71, 73, 97, 140, 186, 211, 227, 272, 287, 302, 303, 304, 307, 357, 376, 497
卫生司，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
环境管理计划，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, 123, 125, 132, 154, 167, 187, 210, 223, 224, 226, 227, 228, 278, 291, 329, 333, 342, 351, 365, 376, 378, 459, 475, 497
摘录，81，128，134，193，311，323，324，326，328，330，345，361，362，363，365，383，384，450，475

F

羽毛，62、123、330、378
FPDF, 339, 340, 341, 342, 343

G

GDPR, 109
地理信息系统，58
GLTF, 143, 278
谷歌工作表，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。
欧洲联盟委员会，8，73，138，142，186，227，261，262，263，264，265，266，267，268，272，273，273，276，277，278，279，280，284，286，292，296，302，329，357，414，415，417，497
iges、262、263、276。
iLOD, 289
物联网、10、18、67、271、367、369、405、413、417、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 笔记本，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, 166, 187, 219, 220, 220, 221, 222, 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
载荷、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
低代码，368，369

M

Matplotlib, 123, 306, 320, 333, 335, 342, 372, 409, 411, 494
监测和评估司，77、78、257、284
监测和评价计划，14，175，492
网目、234、283、284、285、296、462
微软 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 项目，70
MySQL, 63, 64, 65, 291, 329, 330

N

n8n, 116, 311, 361, 368, 369, 370, 371, 372
NLP, 69, 394
无代码，368，369
nurbs, 142, 282, 283, 284, 285
NWC, 8, 276

O

obj, 143, 273, 276, 277, 278, 280, 281, 284, 285
OCCT, 273
教育办事处，69，128，132，134
OmniClass, 154, 155, 156.
开放式 BIM，142、216、219、256、261、278、291
开放源码, 45, 55, 97, 98, 107, 108, 273, 275, 494
猫头鹰，267、268、269。

P

熊猫, 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
镶木地板，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, 342, 344, 345, 346, 347, 354, 355, 356, 357, 365, 376, 475, 480, 497
PDM, 16
PHP, 63, 339
管道, 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
私募股权，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, 349, 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
驻地协调员，311，317，319，321，351，370，389，479
转速、14、84、151、331、333、334
RVT, 8, 73, 77, 140, 186, 227, 272, 296, 300, 302, 322, 329, 357, 376, 414, 415, 417, 497

S

SaaS, 24, 50, 51, 52, 458
范围，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, 279, 276, 277, 300, 329, 376, 392, 494
SQLite, 63, 64, 65, 166, 167, 296, 329, 363
步骤，261，262，263，266，272，276，277，292
流光, 336, 337
SVF, 142, 276, 285

T

变革、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

统一分类，154、155、156。
美元，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, 280, 291, 299, 329, 378, 414, 474

探索数据如何重塑建筑行业

内容

- AEC 中超过 100 个关键数据相关主题
- 超过 300 种独特的可视化和图表
- 超过 50 个真实商业案例
- 实践法学硕士 (LLM) 与人工智能应用
- 代码示例与现成工作流程

主题

- 数据驱动的建设
- AEC的数字化转型
- 数据分析与自动化
- 数据质量与治理
- CAD、BIM与数据互操作性
- 建筑行业的法学硕士、人工智能与机器学习
- 项目成本与时间预测

观众群体

- 施工项目负责人
- 建筑师与土木工程师
- BIM 协调员与数据经理
- 数字化转型的引领
- AEC领域的IT与AI专家
- 城市规划师与可持续发展顾问
- 建筑与工程学专业的学生
- 教育者与研究人员

- 网站
- · www.datadrivenconstruction.io

对初版的评价：

“博伊科被誉为IT界的詹姆斯·卡维尔——后者那句广为引用的“笨蛋，问题在于经济”——只需用一个词来概括这部杰作即可。“笨蛋，问题在于数据。”（而非软件）为了在数据宇宙中找到方向，一句源自希腊语的古罗马谚语至今仍然适用：“Navigare necesse est”（航行是必要的）。作者以稳健的笔触和坚定的指南针，引导读者在数据海洋的深处与浅滩之间穿梭，更不用说全面的历史研究方法，以及同样重要的，极具原创性的图表……

— Burkhard Talebitari 博士

“Artem Boiko 的著作是建筑行业数字化民主化的重要里程碑，也是中小企业 (SME) 的真正变革者。这本书呼吁大家积极行动！对于任何希望深入了解建筑行业数字化转型，并以务实、高效和前瞻性的方式积极参与塑造这一进程的人来说，它都是一本极具价值的指南。现在正是共同分享这些知识、不断提升建筑行业生产力的最佳时机……

— 迈克尔·马克斯·布勒博士

这本实用指南将帮助专业人士和新手掌握快速发展的数据驱动建筑领域。从数据管理的基本概念到先进的数字化工作流程、人工智能工具及其实际应用，本书将为您描绘出更智能、更快速、更高效的建筑流程蓝图。

ISBN: 978-3-98262-557-7

9 783982 625577