

2ND EDITION
UPDATED & EXPANDED

DATA-DRIVEN CONSTRUCTION

NAVIGATING THE DATA AGE
IN THE CONSTRUCTION INDUSTRY

WITH AI & LLM USE CASES

Artem Boiko



100+

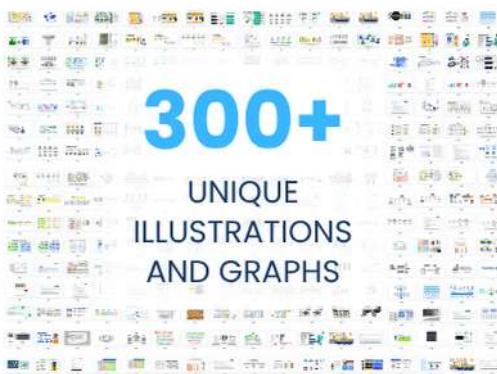
KEY DATA-
RELATED
TOPICS

MORE LANGUAGES ON THE SITE
DATADRIVENCONSTRUCTION.IO



50+

PRACTICAL
PROBLEMS
SOLVED



300+

UNIQUE
ILLUSTRATIONS
AND GRAPHS

DATA-DRIVEN CONSTRUCTION

Navigating the data age
in the construction industry

Second edition, corrected and supplemented

ARTEM BOIKO

“

"Boiko is the James Carville of IT - in the latter's much-quoted "It's the economy, stupid," only one word needs to be exchanged for this famous book. "It's the data, stupid." (And to find one's way in the data universe, a saying of the ancient Romans dating back to Greek is still valid today: "Navigare necesse est". The author navigates his readers through all the depths and shallows of the data ocean with a sure hand and an unwavering compass, not to mention a comprehensive historical approach and, last but not least, highly original graphics and a good sense of humor that is not only apparent at second glance. The international response to Boiko's book ranges from euphoric approval to rather bilious skepticism, which has done the second German edition of the book some good. Boiko is an original and undogmatic data thinker. He presents the reader with exciting insights and always courageous, even provocative theses that inspire further thought. Excellent medicine for the German disease of latent consensualism. Incidentally, the above Latin proverb has a complement: "vivere non est necesse." It does not apply to Boiko's approach to the world of data - data lives and its life is necessary, not to say crucial."

- **Dr. Burkhard Talebitari**, freelance editor - including for the Journal: BIM, published annually by Ernst & Sohn since 2013.

"Artem Boiko's book is a milestone for the democratization of digitization in the construction industry - and a real game changer for small and medium-sized enterprises (SMEs). Particularly groundbreaking: by using modern open-source low-code and no-code tools, companies can already efficiently integrate data into their business processes and analyze it profitably - without any in-depth programming knowledge. This makes the expensive use of cumbersome commercial software packages redundant. This book is a call to action! It is a valuable guide for anyone who not only wants to understand the digital transformation in the construction industry, but also wants to actively shape it - pragmatically, efficiently and in a forward-looking way. Now is the time to work together to share this knowledge and sustainably increase the productivity of the construction industry."

- **Dr. Michael Max Buehler**, Professor of Construction Management at HTWG Konstanz, Co-Owner at GemeinWerk Ventures, and Independent Director at DevvStream.

"DataDrivenConstruction book is one of the first steps beyond the boundaries of the usual world of builders, with their complex design and management systems, when, it would seem, the complexity and saturation of data does not even give a chance for radical simplification and increased transparency of work with construction data. In his book, Artem shows in simple language what opportunities modern technologies of working with data open before us, and literally gives concrete steps that you can immediately apply in your work. I urge everyone who wants to understand where automation systems will go in the construction industry to study this book carefully in order to realize that the data revolution in construction is already knocking on our door. It's only of interest to geeks now, but in a few years, like BIM, such approaches and software will be ubiquitous!"

- **Ihor Rogachew**, Head of IMT Competence Center, BIM & Digital Transformation at RGD, and Founder of InfraBIM.Pro.

"I highly recommend the DataDrivenConstruction book that addresses, as the title says, a data driven information management approach for AECO. I am currently using it to help initiate a number of discussions with various groups. I have found it a very accessible reference. As well as a thorough overview of the history context of tools in AECO, data and introducing several key technologies the book contains a number of very useful diagrams that outline the scope of data sources and end user artefacts with sample workflows. It strikes me that these are the types of diagrams we need more of when developing and monitoring information strategies and contribute to BEP's - defining the overall enterprise data model onto which the boundary for a PIM and AIM can be overlaid."

- **Paul Ransley**, Principal Consultant at Acmena, and Systems Integration Engineer at Transport for London.

"If "data is the new oil", we need to learn to define it, find it, mine it, refine it, to make it valuable. I've found the book DataDrivenConstruction very informative and insightful. The book provides a useful historical background and explains working with data in plain language. For those who are interested in digital transformation, it gives a good understanding of data - how it works, how it is structured and how it can be used."

- **Ralph Montague**, Director at ArcDox, Director of the BIM Coordinators Summit, and Chair of the BIM National Mirror Committee at the National Standards Authority of Ireland.

"As it was emphasized in the book, information is a crucial asset for the construction sector and having it in accessible formats greatly facilitates accurate decision-making and expedites project timelines. The book offers a neutral and efficient approach to accessing and taking advantage of this source in decision-making. The methodology presented in the book leverages a contemporary approach that combines artificial intelligence-driven programming with accessible open-source tools. By harnessing the power of AI and utilizing open-source software, the methodology aims to enhance automation, optimize processes, and promote accessibility and collaboration within the field. The language of the book is clear and easy to follow."

- **Dr. Salih Ofluoğlu**, Dean of the Faculty of Fine Arts and Architecture at Antalya Bilim University, and Organizer of the Eurasian BIM Forum.

"All I can say is, WOW! The way you incorporated history, the LLM, the graphics, and the overall ease of understanding your points is truly remarkable. The flow of the book is amazing. There are so many brilliant aspects to this book; it's genuinely a game-changer. It's a great source of information, and I commend you for the effort and passion you've put into it. Congratulations on creating such a remarkable work. I could go on, but suffice it to say, I'm incredibly impressed!"

- **Natasha Prinsloo**, Digital Practice Lead at energylab_.

"For anyone in the construction industry, from rookies to seasoned pros, this book is a game-changer! It's not your typical dusty read-it's packed with insights, strategies, and a touch of humor to keep you engaged. From ancient data recording methods to cutting-edge digital technologies, it covers the evolution of data usage in construction. It's like taking a time machine through the evolution of construction data. Whether you're an architect, engineer, project manager or data analyst, this comprehensive guide will revolutionize the way you approach projects. Get ready to optimize processes, enhance decision-making, and manage projects like never before!"

- **Pierpaolo Vergati**, Lecturer at Sapienza University of Rome, and Senior Construction Project Manager at Fintecna.

“

"I read the book in one breath, in less than 6 hours. The manufacturing quality of the book is excellent, dense glossy paper, color schemes, a pleasant font. The large number of practical examples on how to work with LLM specific to the construction industry will save you months, if not years, of self-study. Work examples are very diverse, ranging from simple to complex, without requiring you to purchase complex and expensive software. The book will allow owners of any business in the construction industry to take a fresh look at their business strategy, digitalization, and development prospects. And for smaller companies to increase efficiency with affordable and free tools."

- **Mikhail Kosarev**, Lecturer and Consultant on Digital Transformation in the Construction Industry at TIM-ASG.

“

"I should say that Data-Driven Construction is worthy of being taught as a textbook in universities and is a book that will make valuable contributions to the developing BIM field. Data-Driven Construction book contains a technical glossary that explains the concepts very well. Topics that are extremely difficult to explain are made simple and understandable with very beautiful visual language. I think that what is intended to be explained in the visuals should be expressed to the reader, even if briefly. The comprehensibility of some visuals, in other words, reading the visual requires separate information. I would also like to say that I am happy to introduce Artem Boiko's valuable work in my lectures and seminars at universities."

- **Dr. Ediz Yazicioglu**, Owner at ArchCube, and Lecturer in Construction Project Management in the Department of Architecture at Istanbul Technical University and at Medipol University.

“

"DATA DRIVEN CONSTRUCTION" book is a game-changer for anyone curious about where the construction industry is headed in the age of data. Artem doesn't just scratch the surface; he delves deep into the current developments, challenges, and promising opportunities in construction. What sets this book apart is its accessibility - Artem explains complex ideas using relatable analogies that make the content easy to grasp. I found the book to be incredibly informative yet engaging. In summary, Artem has crafted a valuable resource that not only informs but also inspires. Whether you're a seasoned professional or a newcomer to construction, this book will broaden your perspective and deepen your understanding of where the industry is heading. Highly recommended!"

- **Moayad Saleh**, Architect and BIM Implementation Manager at TMM GROUP Gesamtplanungs GmbH.

“

"Data Driven Construction vividly conveys the basics of information-based work with building data. A book that deals with information flows and fundamental economic concepts and thus sets itself apart from other BIM books, because it not only represents the perspective of a software manufacturer, but also tries to convey fundamental concepts. A book worth reading and seeing."

- **Jakob Hirn**, CEO and Co-Founder of Build Informed GmbH, and Initiator of the Innovation Forum "On Top With BIM".

“

"Data-Driven Construction" by Artem Boiko is an impressive work that offers a solid foundation for the construction industry in times of constantly growing technologies and information possibilities. Boiko manages to present complex topics in an understandable way while also introducing visionary ideas. The book is a well-thought-out compendium that not only highlights current developments but also provides an outlook on future innovations. It comes highly recommended for anyone who wants to get to grips with data-driven construction planning and execution."

- **Markus Eiberger**, Lecturer at Stuttgart University of Applied Sciences, Senior Project Manager and Deputy Branch Manager at Konstruktionsgruppe Bauen, Board Member of the BIM Cluster Baden-Württemberg Association.

“

"Data is the new oil" as they say, so its prospectors or miners should have the right tools and mindset to extract value from this 21st century resource. The construction industry has for too long been on the slippery slope of "3D information" based processes, whereby project delivery is based on someone else's baked information (e.g. they already plotted the pie or bar chart) whereas the underlying "data" (e.g. the raw spreadsheet) is capable of delivering much more, especially because multi-data fusion and AI bring unlimited potentials. If you are delivering (or teaching/researching) construction, this book is your best - and so far only - resource for navigating the data-driven world we have found ourselves."

- **Dr. Zulfikar Adamu**, Associate Professor of Strategic IT in Construction at LSBU, UK.



Second edition, April 2025.
© 2025 | Artem Boiko | Karlsruhe

ISBN 978-3-9827303-0-1



Artem Boiko Copyright

boikoartem@gmail.com
info@datadrivenconstruction.io

No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without the written permission of the author - except for non-commercial distribution in unaltered form. The book is distributed free of charge and may be freely shared with others for personal, educational, or research purposes, provided authorship and references to the original are retained. The author retains all non-proprietary rights to the text and makes no express or implied warranties. Companies, products, and names mentioned in the book may be fictitious or used as examples. The author is not responsible for any consequences of the use of the information given. The information contained in the book is presented "as is" without warranty of completeness or relevance. The author is not liable for incidental or consequential damages arising from the use of the information, code, or programs contained in this book. The code samples presented in this book are for educational purposes only. Readers use them at their own risk. The author recommends that all software solutions be tested before use in a production environment. All trademarks and product names mentioned in the text are trademarks, registered trademarks, or service marks of their respective companies and are the property of their respective owners. The use of these names in the book does not imply any relationship with or endorsement by their owners. Mention of third-party products or services does not constitute a recommendation or imply an endorsement. Company and product names used in the examples may be trademarks of their owners. Links to third-party websites are provided as a convenience and do not imply that the author endorses the information provided on those sites. All statistics, quotes, and studies cited were current at the time the book was written. Data may change over time.

This book is distributed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (CC BY-NC-ND 4.0). You may copy and distribute it for non-commercial purposes, provided you retain authorship and make no changes.



© 2024 Artem Boiko. First edition.
© 2025 Artem Boiko. Second edition, revised and supplemented.
All rights reserved

FOREWORD TO THE SECOND EDITION

This book is the result of a lively dialog with the professional community. It is based on numerous professional discussions on data management in the construction industry, which took place on various professional platforms and social media platforms. These discussions became the basis for articles, publications and visual materials that have generated a wide response in the professional community. The author's materials attract millions of views every year on various platforms and languages, bringing together professionals in the field of digitalization of construction.

Within a year of the publication of the first edition, the book was ordered by experts from more than 50 countries, from Brazil and Peru to Mauritius and Japan. The second edition of the book, which you are now holding in your hands, has been revised and expanded based on expert feedback, critical comments on the first edition and discussions in professional circles. Thanks to the feedback, the second edition has been significantly expanded: new chapters on CAD (BIM) technologies and creating effective ETL -processes have been added. The number of practical examples and case studies has also been significantly increased. Of particular value is the feedback from leaders of the construction industry, consulting companies and major IT-companies, who approached the author with questions of digitalization and interoperability both before and after the first version of the book was published. Many of them have already applied the approaches described in the book or plan to do so in the near future

You are holding in your hands a book created through discussion and active exchange of opinions. Progress is born in dialog, in the clash of views and openness to new approaches. Thank you for being part of this dialog. Your constructive criticism is the basis for future improvements. If there are errors in the text or if you would like to share ideas and suggestions, any feedback is welcome. Contact details are provided at the end of the book

WHY IS THE BOOK FREE?

This book was conceived as an open educational resource aimed at disseminating modern approaches to data management in the construction industry. The first version of the book served as a basis for collecting comments and suggestions from the professional community, which allowed improving the structure and content of the material. All comments, suggestions, and ideas were carefully analyzed and incorporated into this revised version. The aim of the book is to help construction professionals understand the importance of working with data: systematically, consciously and with an eye on the long-term value of the information. The author has collected examples, illustrations and practical observations from more than 10 years of working in the digitalization of construction. Most of this material was born out of real projects, discussions with engineers and developers, participation in international initiatives and training seminars. The book is an attempt to structure the accumulated experience and share it in an accessible way. If you want to support further dissemination of the book's ideas and get a convenient format for reading, working with examples and visual materials - you can purchase a [printed version](#).

RIGHTS OF USE

All materials, illustrations and fragments of this book may be reproduced, quoted or used in any format and on any media provided that the source is credited to Artem Boiko and the title of the book is "Data-Driven Construction". Thank you for respecting the work and spreading knowledge.

It is with sincere gratitude that I dedicate this book to my family,
who from an early age instilled in me a deep love of construction,
to my hometown of mining for lessons in resilience, and to my
surveyor wife, whose unwavering support has been my constant
inspiration.

FOR WHOM THIS BOOK IS FOR

Written in accessible language, this book is aimed at a wide range of readers in the construction industry, from students and novices who want to grasp the basics of modern construction processes to professionals who need an up-to-date methodology for managing data in construction. Whether you are an architect, engineer, foreman, construction manager, or data analyst, this comprehensive guide with many unique illustrations and charts offers valuable insights on how to use data in business to optimize and automate processes, improve decision-making, and manage construction projects at different levels using modern tools.

The book is a comprehensive guide that combines theoretical foundations and practical recommendations for integrating data management techniques into construction processes. The book focuses on the strategic use of information to optimize operations, automate processes, improve decision-making and effectively manage projects using modern digital tools.

This book covers theoretical and practical aspects of working with information in the construction industry. Through detailed examples, it explores the methodology of task parameterization, requirements gathering, processing unstructured and multiformat data and transforming it into effective solutions for construction companies.

The reader successively passes the way from the formation of requirements and development of basic data models to more complex processes of integration of heterogeneous sources of information, creation of ETL - processes, construction of information Pipeline and machine learning models. The sequential approach allows you to clearly demonstrate the mechanisms of organization and automation of business processes and decision support systems in the construction industry. Each part of the book concludes with a practical chapter containing step-by-step instructions that enable immediate application of the acquired knowledge in real projects.

SUMMARIZED DESCRIPTION OF PARTS OF THE BOOK

This book is organized around the concept of transforming data in the value chain: from data collection and quality assurance to analytical processing and extracting valuable practical solutions using modern tools and methodologies.

Part 1: Digital Evolution in Construction - Traces the historical transformation of data management from clay tablets to modern digital systems, analyzing the emergence of modular systems and the growing importance of information digitalization in the context of industrial revolutions.

Part 2: Information Challenges for the Construction Industry - explores the problems of data fragmentation, 'information silos', the impact of the HiPPO approach on decision making and the limitations of proprietary formats, suggesting consideration of the move to AI and LLM ecosystems.

Part 3: Data Systematization in Construction - forms a typology of construction data, describes methods of its organization, integration with corporate systems and discusses the creation of competence centers for standardization of information processes.

Part 4: Data Quality Assurance - reveals methodologies for turning disparate information into quality, structured data, including data extraction from various sources, validation, and modeling using LLM.

Part 5: Cost and Time Calculations - covers digitalization of cost and scheduling calculations, automation of obtaining volumes from CAD (BIM) models, 4D-8D modeling technologies, and ESG calculations for construction projects.

Part 6: CAD and BIM - Critically analyzes the evolution of design technologies, system interoperability issues, trends toward open data formats, and the prospects for applying artificial intelligence to design.

Part 7: Data Analytics and Automation - examines the principles of information visualization, key performance indicators, ETL processes, workflow orchestration tools, and the application of language models to automate routine tasks.

Part 8: Data Storage and Management - explores data storage formats, data warehouse and data lake concepts, data management principles, and new approaches including vector databases and the DataOps and VectorOps methodologies.

Part 9: Big Data and Machine Learning - focuses on the transition to objective analysis based on historical data, the Internet of Things on construction sites, and the application of machine learning algorithms to predict project costs and timelines.

Part 10: The Construction Industry in the Age of Digital Data - presents a look at the future of the construction industry, analyzing the shift from causal analysis to working with correlations, the concept of "Uberizing" construction, and strategies for digital transformation.

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



INTRODUCTION

How long can your company remain competitive in a world where technology is rapidly evolving and every aspect of business, from timing and costing to risk analysis, is being automated by machine learning models?

The construction industry, which has existed for as long as mankind itself, is on the threshold of revolutionary changes that promise to completely change the way we think about traditional construction. Already in other sectors of the economy, digitalization is not just changing the rules of thumb, but is ruthlessly driving out of the market companies that have failed to adapt to the new data processing environment and are unable to improve the speed of decision-making (Fig. 1).



Fig. 1 The speed of decision making in the construction industry depends on the human factor more often than in other industries.

Banking, retail, logistics and agribusiness are rapidly moving towards full digitalization, where inaccuracies and subjective opinions have no place anymore. Modern algorithms are able to analyze enormous amounts of data and provide customers with accurate predictions - whether it's the likelihood of loan repayment, optimal delivery routes or risk forecasting.

Construction is one of the last industries to make the inevitable transition from decisions based on the opinions of highly paid experts to data-driven solutions. This transition is driven not only by new technological capabilities, but also by increased market and customer demands for transparency, accuracy and speed.

Robotization, process automation, open data and forecasts based on them - all these are no longer just possibilities, but inevitability. Most companies in the construction industry, which were recently

responsible to the client for calculating the volume, cost, time of projects and quality control, now risk turning into mere executors of orders, not making key decisions (Fig. 2).

With the development of computing power, machine learning algorithms and democratized access to data, it is now possible to automatically combine data from different sources, enabling deeper process analysis, risk prediction and cost optimization at the stages of construction project discussions. These technologies create the potential for radical efficiency gains and cost reductions across the sector.

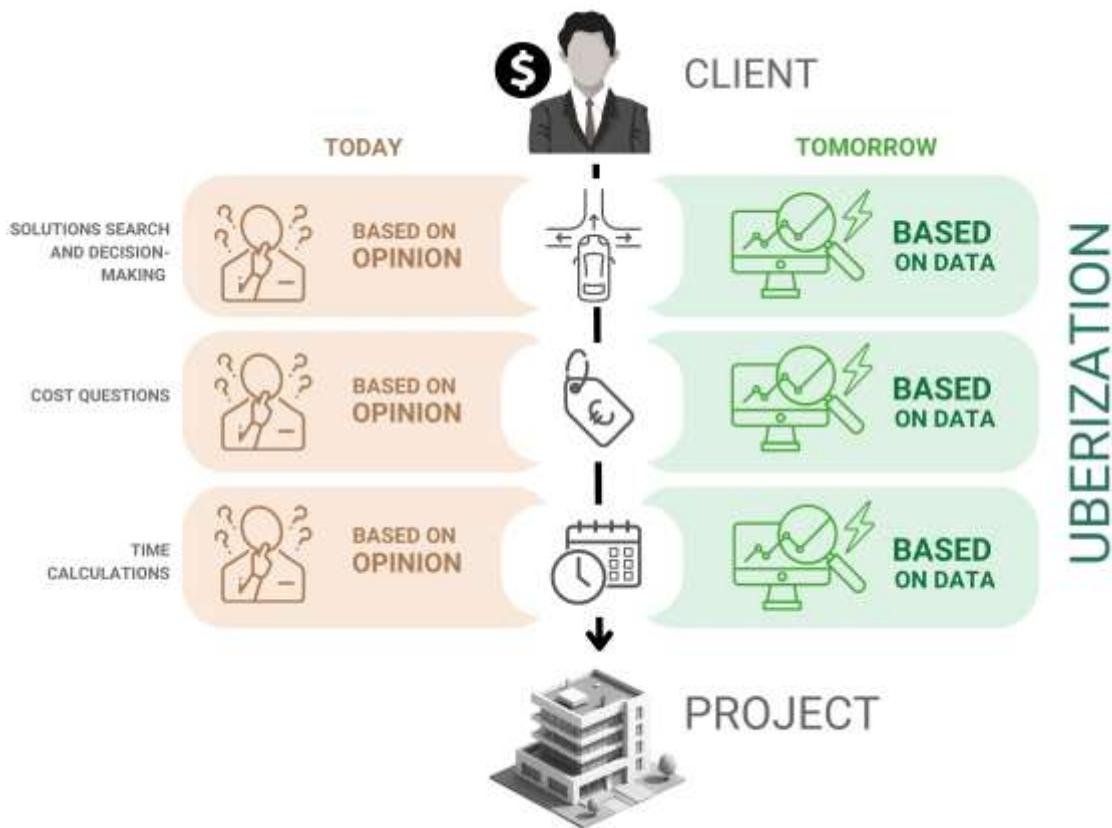


Fig. 2 The client is not interested in excessive human factors on the way to realizing their project.

Despite all the advantages of new tools and concepts, the construction industry lags far behind other sectors of the economy in the adoption of new technologies.

According to the IT Metrics Key Data 2017 report, the construction industry ranks last in IT spending among 19 other industries [1].

The rapid growth in data volume and process complexity is becoming a headache for company management, and the main problem in utilizing new technologies is that data, despite its abundance, remains fragmented, unstructured and often incompatible between different systems and software products. Therefore, many companies in the construction sector today are now primarily concerned about data quality issues, which can only be solved with the implementation of efficient, automated management and analytics systems.

According to a KPMG® survey of construction managers in 2023 [2], project management information systems (PMIS), advanced and basic data analytics and building information modeling (BIM) have the greatest potential to improve ROI on projects (Fig. 3).

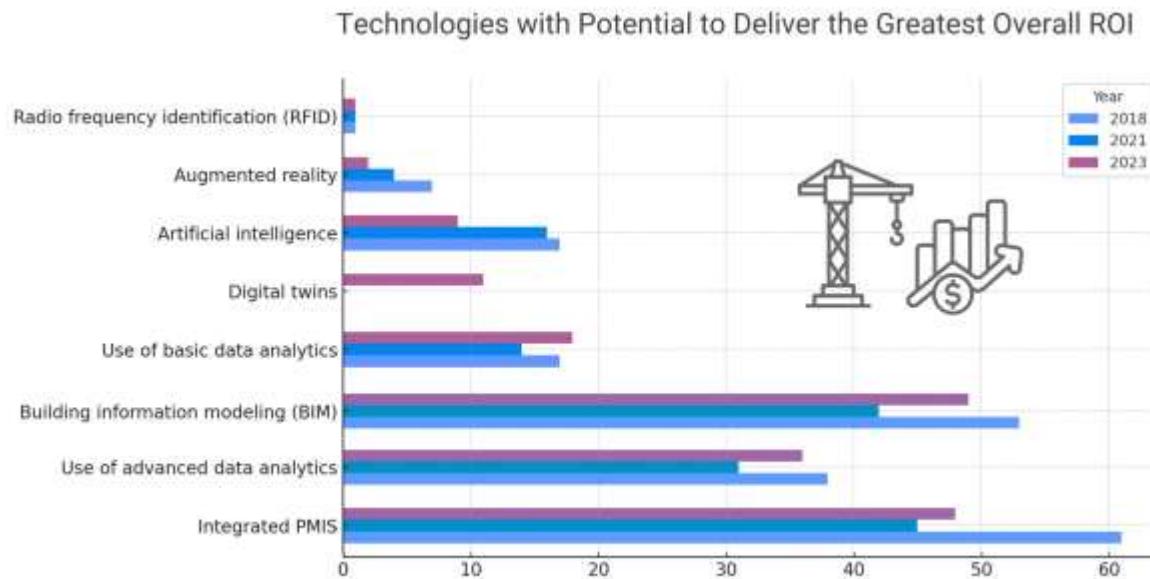


Fig. 3 Survey among construction company managers: which technologies will provide the highest return on investment (ROI) in capital projects? (based on materials [2]).

The solution to the challenges associated with integrating data into business processes is to ensure high quality information, use appropriate data formats, and apply effective methods for creating, storing, analyzing, and processing data.

Realizing the value of data is driving various industries away from siloed applications and complex bureaucratic management structures. Instead, the focus is shifting to creating new approaches to information architecture, transforming companies into modern data-driven enterprises. Sooner or later, the construction industry itself will take this step, moving from a gradual digital evolution to a true digital revolution affecting all companies.

The transition to data-driven business processes will not be easy. Many companies will face challenges because executives don't always understand how to use chaotic data sets to improve efficiency and business growth.

This book delves into the world of data, where information is becoming a key strategic resource that determines the efficiency and sustainability of business processes. With the rapid growth of information, companies are facing new challenges. Digital transformation is no longer just a buzzword - it is becoming a necessity.



Fig. 4 Data and processes are the foundation of construction.

To understand transformation means to be able to explain the complex in simple words. That is why the book is written in an accessible language and is accompanied by illustrations created specifically for the purpose of visualizing key concepts. These diagrams, charts and visualizations are designed to remove perception barriers and make the material understandable even to those who previously considered such topics too complex. All illustrations, charts and graphs in this book are created by the author and designed specifically to visualize the key concepts described in the text.

One picture is worth a thousand words [3].

— Fred R. Barnard, English illustrator, 1927.

To connect theory with practice, we will use artificial intelligence tools (in particular, language models) that allow you to develop solutions without the need for deep programming knowledge. If you are oriented towards practical material and are more interested in practical work with data, you can skip the first introductory part and go straight to the second part of the book, where the description of concrete examples and cases begins.

However, do not place excessive expectations on AI (Artificial Intelligence), machine learning and LLM (Large Language Models) tools in general. Without quality input data and a deep understanding of the subject matter, even the most advanced algorithms cannot provide reliable and meaningful results.

Microsoft CEO Satya Nadella warns of the risk of a bubble in artificial intelligence in early 2025 [4], comparing the current hype to the dot-com bubble. He emphasizes that claims of reaching AGI (Artificial General Intelligence) milestones without proper justification are "meaningless manipulation of metrics". Nadella believes that the real success of AI should be measured by its contribution to global GDP growth, not by an overemphasis on buzzwords.

Behind all the buzzwords about new technologies and concepts lies a complex and painstaking work to ensure data quality, parameterization of business processes and adaptation of tools to real tasks.

A data-driven approach is not a product you can just download or buy. It is a strategy that must be built. It starts with a fresh look at existing processes and problems, and then requires disciplined movement in the chosen direction.

Leading software developers and application vendors will not be the engine of change in the construction industry for many of them the data-driven approach is a threat to their established business model.

Other industries [unlike construction], such as automotive, have already undergone radical and disruptive change, and their digital transformation is well underway. Construction companies need to act quickly and decisively: nimble companies will reap huge rewards, while for those who hesitate, the risks will be severe. Consider the upheaval that digital photography has caused in the industry [5].

– World Economic Forum report Shaping the Future of Construction, 2016

Those companies that realize the opportunities and benefits of the new approach in a timely manner will gain a sustainable competitive advantage and will be able to develop and grow without dependence on solutions from large vendors.

This is your chance to not only weather the coming storm of information digitalization, but to take control of it. In this book you will find not just an analysis of the current state of the industry, but also concrete recommendations for rethinking and restructuring your processes and your business to become a leader in the new era of construction and enhance your professional experience.

The digital future of construction is not just about using new technologies and programs, but fundamentally rethinking data handling and business models.

Is your company ready for this strategic change?

TABLE OF CONTENTS

INTRODUCTION	1
TABLE OF CONTENTS	I
II PART FROM CLAY TABLETS TO THE DIGITAL REVOLUTION: HOW INFORMATION HAS EVOLVED IN CONSTRUCTION.....	2
CHAPTER 2.1. EVOLUTION OF DATA USE IN THE CONSTRUCTION INDUSTRY.....	3
The birth of the data era in construction.....	3
From clay and papyrus to digital technology	4
Process as a tool for data-driven experience.....	5
Digitalization of construction process information.....	7
CHAPTER 2.2. TECHNOLOGIES AND MANAGEMENT SYSTEMS IN MODERN CONSTRUCTION	11
Digital revolution and the emergence of modular MRP/ERP-systems	11
Data management systems: from data mining to business challenges	13
Corporate mycelium: how data connects to business processes	17
CHAPTER 2.3. THE DIGITAL REVOLUTION AND THE EXPLOSION OF DATA	20
The beginning of the data volume boom as an evolutionary wave.....	20
The amount of data generated in a modern company	22
Cost of data storage: the economic aspect.....	23
Frontiers of data accumulation: from mass to meaning	25
Next steps: from data theory to practical change	27
III PART HOW THE CONSTRUCTION BUSINESS IS DROWNING IN DATA CHAOS	29
CHAPTER 3.1. DATA FRAGMENTATION AND SILOS	30
The more tools, the more efficient the business?	30
Data silos and their impact on company performance	32
Duplication, and lack of data quality as a consequence of disunity	35
HiPPO or the danger of opinions in decision making.....	37
Continuous increase in the complexity and dynamism of business processes	40

The Fourth Industrial Revolution (Industry 4.0) and the Fifth Industrial Revolution (Industry 5.0) in the construction industry	43
CHAPTER 3.2. TURNING CHAOS INTO ORDER AND REDUCING COMPLEXITY	46
Redundant code and closed systems as a barrier to productivity improvement.....	46
From silos to a unified data warehouse	48
Integrated storage systems enable the transition to AI agents	49
From data collection to decision-making: the road to automation	52
Next steps: turning chaos into a manageable system	54
IV PART DATA FRAMEWORK IN CONSTRUCTION BUSINESS PROCESSES	56
CHAPTER 4.1. DATA TYPES IN CONSTRUCTION.....	57
The most important data types in the construction industry	57
Structured Data	61
Relational databases RDBMS and SQL query language	63
SQL-queries in databases and new trends.....	65
Unstructured data	67
Text data: between unstructured chaos and structure	68
Semi-structured and loosely structured data.....	69
Geometric data and its application.....	70
CAD data: from design to data storage.....	73
The emergence of the BIM (BOM) concept and the use of CAD in processes	76
CHAPTER 4.2. DATA UNIFICATION AND STRUCTURING	82
Filling systems with data in the construction industry.....	82
Data transformation: the critical foundation of modern business analysis	84
Data models: relationships in data and relationships between elements	88
Proprietary formats and their impact on digital processes	92
Open formats are changing the approach to digitalization	96
Paradigm Shift: Open Source as the End of the Era of Software Vendor Dominance	97
Structured open data: the foundation of digital transformation.....	100
CHAPTER 4.3. LLM AND THEIR ROLE IN DATA PROCESSING AND BUSINESS PROCESSES..	103
LLM chat rooms: ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok for automating	

data processing processes	103
Large Language Models LLM: how it works	104
Utilizing local LLMs for sensitive company data	107
Full control of AI in the company and how to deploy your own LLM	109
RAG: Intelligent LLM -assistants with access to corporate data.....	111
CHAPTER 4.4. IDE WITH LLM SUPPORT AND FUTURE PROGRAMMING CHANGES.....	113
Choosing an IDE: from LLM experiments to business solutions.....	113
IDE with LLM support and future programming changes	115
Python Pandas: an indispensable tool for working with data.....	116
DataFrame: universal tabular data format	120
Next steps: building a sustainable data framework	123
V PART DATA QUALITY: ORGANIZATION, STRUCTURING, MODELING	125
CHAPTER 5.1. DATA CONVERSION INTO A STRUCTURED FORM	126
Learning how to turn documents, PDF, pictures and texts into structured formats	126
Example of converting a PDF -document into a table	127
Convert JPEG, PNG image to structured form.....	131
Converting text data into a structured form.....	134
Translation of CAD data (BIM) into a structured form	137
CAD solution vendors move to structured data.....	142
CHAPTER 5.2. CLASSIFICATION AND INTEGRATION: A COMMON LANGUAGE FOR CONSTRUCTION DATA.....	145
Speed of decision making depends on data quality.....	145
Data standardization and integration	146
Digital interoperability starts with requirements.....	149
A common language of construction: the role of classifiers in digital transformation.....	151
Masterformat, OmniClass, Uniclass and CoClass: the evolution of classification systems	154
CHAPTER 5.3. DATA MODELING AND CENTER OF EXCELLENCE	159
Data modeling: conceptual, logical and physical model.....	159
Practical data modeling in the context of construction	163
Creating a database using LLM	165

Center of Excellence (CoE) for Data Modeling	167
CHAPTER 5.4. SYSTEMATIZATION OF REQUIREMENTS AND VALIDATION OF INFORMATION	170
Requirements gathering and analysis: transforming communications into structured data.....	170
Process flowcharts and the effectiveness of conceptual frameworks	174
Structured Requirements and RegEx regular expressions.....	176
Data collection for the verification process	181
Verification of data and results of verification	183
Visualization of verification results	188
Comparison of data quality checks with human life needs.....	190
Next steps: turning data into accurate calculations and plans	192
VI PART COST AND TIME CALCULATIONS: INCORPORATING DATA INTO CONSTRUCTION PROCESSES	194
CHAPTER 6.1. COST CALCULATIONS AND ESTIMATES FOR CONSTRUCTION PROJECTS ..	195
Construction basics: estimating quantity, cost and time.....	195
Methods of calculating the estimated cost of projects	196
Resource-based method estimates and costing in construction.....	197
Database of construction resources: catalog of construction materials and works.....	197
Calculation of calculations and costing of works on the basis of resource base	199
Final project costing: from estimates to budget.....	203
CHAPTER 6.2. QUANTITY TAKE-OFF AND AUTOMATIC CREATION OF ESTIMATES AND SCHEDULES.....	208
Moving from 3D to 4D and 5D: using volumetric and quantitative parameters	208
5D attributes and retrieving attribute volumes from CAD	208
QTO Quantity Take-Off: grouping project data by attributes	212
QTO automation using LLM and structured data	217
QTO calculation of the entire project using group rules from an Excel spreadsheet	221
CHAPTER 6.3. 4D, 6D -8D AND CALCULATION OF CARBON DIOXIDE EMISSIONS CO₂	227
4D model: integrating time into construction estimates.....	227
Construction schedule and its automation based on costing data	228

Extended attribute layers 6D -8D: from energy efficiency to safety assurance	230
Estimating CO ₂ and calculating carbon dioxide emissions from construction projects	233
CHAPTER 6.4. CONSTRUCTION ERP AND PMIS SYSTEMS	237
Construction ERP -systems on the example of calculations and estimates.....	237
PMIS: Intermediate between ERP and the construction site	242
Speculation, profit, insularity and lack of transparency in ERP and PMIS	243
The end of the era of closed ERP /PMIS: the construction industry needs new approaches	246
Next steps: efficient use of project data	248
VII PART CAD AND BIM: MARKETING, REALITY AND THE FUTURE OF DESIGN DATA IN CONSTRUCTION	251
CHAPTER 7.1. EMERGENCE OF BIM-CONCEPTS IN THE CONSTRUCTION INDUSTRY	252
History of the emergence of BIM and open BIM as marketing concepts of CAD- vendors .	252
The reality of BIM: instead of integrated databases - closed modular systems.....	255
The emergence of the open format IFC in the construction industry	257
IFC format problem depending on geometric core	259
Appearance in the construction of the topic of semantics and ontology.....	262
Why semantic technologies fail to live up to expectations in the construction industry	264
CHAPTER 7.2. CLOSED PROJECT FORMATS AND INTEROPERABILITY ISSUES.....	268
Closed data and falling productivity: the dead end of the CAD industry (BIM)	268
The myth of interoperability between CAD systems	270
Transition to USD and granular data.....	274
CHAPTER 7.3. GEOMETRY IN CONSTRUCTION: FROM LINES TO CUBIC METERS	278
When lines turn into money or why builders need geometry	278
From lines to volumes: How area and volume become data.....	279
Moving to MESH, USD and polygons: using tessellation for geometry	281
LOD, LOI, LOMD - unique classification of detailing in CAD (BIM)	282
New CAD standards (BIM) - AIA, BEP, IDS, LOD, COBie	285
CHAPTER 7.4. DESIGN PARAMETERIZATION AND USE OF LLM FOR CAD OPERATION	290
The illusion of uniqueness of CAD data (BIM): the path to analytics and open formats.....	290

Design through parameters: the future of CAD and BIM	293
Emergence of LLM in design CAD data processing processes	296
Automated analysis of DWG -files with LLM and Pandas.....	299
Next steps: moving from closed formats to open data	305
VIII PART DATA-DRIVEN DECISION-MAKING, ANALYTICS, AUTOMATION AND MACHINE LEARNING 307	
CHAPTER 8.1. DATA ANALYTICS AND DATA-DRIVEN DECISION-MAKING	308
Data as a resource in decision making	308
Visualizing data: the key to understanding and decision making	311
KPIs and ROI.....	313
Dashboards and dashboards: visualization of indicators for effective management.....	315
Analyzing data and the art of asking questions.....	317
CHAPTER 8.2. DATA FLOW WITHOUT MANUAL EFFORT: WHY ETL.....	320
ETL automation: lower costs and faster data handling	320
ETL Extract: data collection.....	324
ETL Transform: application of validation and transformation rules	327
ETL Load: Visualize results in charts and graphs.....	329
ETL Load: Automatic creation of PDF documents	335
ETL Load: automatic document generation from FPDF.....	336
ETL Load: Reporting and loading to other systems	340
ETL with LLM: Visualize data from PDF -documents	341
CHAPTER 8.3. AUTOMATIC ETL CONVEYOR (PIPELINE).....	347
Pipeline: Automatic ETL	347
Pipeline-ETL data validation process with LLM	351
Pipeline-ETL: verification of data and information of project elements in CAD (BIM).....	353
CHAPTER 8.4. ORCHESTRATION OF ETL AND WORKFLOWS: PRACTICAL SOLUTIONS.....	359
DAG and Apache Airflow: workflow automation and orchestration	359
Apache Airflow: practical application on ETL automation.....	360
Apache NiFi for routing and data conversion	364
n8n Low-Code, No-Code process orchestration	365

Next steps: moving from manual operations to analytics-based solutions	368
IX PART DATA STORAGE AND MANAGEMENT IN CONSTRUCTION.....	370
CHAPTER 9.1. DATA INFRASTRUCTURE: FROM STORAGE FORMATS TO DIGITAL REPOSITORIES	371
Data atoms: the foundation of effective information management	371
Information storage: files or data	372
Big Data Storage: Analyzing Popular Formats and Their Effectiveness	374
Optimize storage with Apache Parquet.....	377
DWH: Data Warehouse data warehouses	379
Data Lake - evolution of ETL to ELT: from traditional cleaning to flexible processing	381
Data Lakehouse architecture: synergy of warehouses and data lakes.....	383
CDE, PMIS, ERP or DWH and Data Lake	385
CHAPTER 9.2. DATA WAREHOUSE MANAGEMENT AND CHAOS PREVENTION.....	389
Vector Databases and the Bounding Box.....	389
Data Governance, Data Minimalism and Data Swamp.....	392
DataOps and VectorOps: new data standards.....	395
Next steps: from chaotic storage to structured storage.....	397
X PART BIG DATA, MACHINE LEARNING AND FORECASTS.....	399
CHAPTER 10.1. BIG DATA AND ITS ANALYSIS.....	400
Big data in construction: from intuition to predictability.....	400
Questioning the feasibility of big data: correlation, statistics and data sampling	401
Big data: analyzing data from San Francisco's million building permit dataset	404
Example of big data based on CAD data (BIM)	410
IoT Internet of Things and Smart Contracts	414
CHAPTER 10.2. MACHINE LEARNING AND PREDICTIONS	418
Machine learning and artificial intelligence will change the way we build	418
From subjective assessment to statistical forecast	420
Titanic dataset: Hello World in the world of analytics data and big data	422
Machine learning in action: from Titanic passengers to project management.....	428
Predictions and forecasts based on historical data.....	433

Key concepts of machine learning.....	435
CHAPTER 10.3. COST AND SCHEDULE FORECASTING USING MACHINE LEARNING	438
An example of using machine learning to find project cost and schedule.....	438
Project cost and time prediction using linear regression	440
Project cost and time predictions using the K-nearest neighbor algorithm (k-NN)	443
Next steps: from storage to analysis and forecasting	447
XI PART THE CONSTRUCTION INDUSTRY IN THE DIGITAL AGE. OPPORTUNITIES AND CHALLENGES	450
CHAPTER 11.1. SURVIVAL STRATEGIES: BUILDING COMPETITIVE ADVANTAGE.....	451
Correlations instead of calculations: the future of construction analytics.....	451
Data-driven approach in construction: a new level of infrastructure	454
The next-generation digital office: how AI is changing the workspace	456
Open data and Uberization is a threat to the existing construction business.....	458
Unresolved problems of uberization as last chance to use time for transformation	461
CHAPTER 11.2. A PRACTICAL GUIDE TO IMPLEMENTING A DATA-DRIVEN APPROACH	466
From Theory to Practice: A Roadmap for Digital Transformation in Construction.....	466
Laying the Digital Foundation: 1-5 Steps to Digital Maturity.....	468
Unlocking the potential of data: 5-10 steps to digital maturity.....	472
Transformation roadmap: from chaos to data-driven company	479
Building in Industry 5.0: How to make money when you can't hide anymore	482
CONCLUSION	484
ABOUT THE AUTHOR.....	487
FEEDBACK.....	488
TRANSLATION COMMENTARY	488
OTHER SKILLS AND CONCEPTS.....	489
GLOSSARY	493
LIST OF REFERENCES AND ONLINE MATERIALS	499
SUBJECT INDEX	516

MAXIMUM CONVENIENCE WITH THE PRINTED VERSION

You are holding a free digital version of **Data-Driven Construction**. For more convenient work and quick access to the materials, we recommend you to pay attention to the **printed edition**:



■ **Always at hand:** the book in printed format will become a reliable working tool, allowing you to quickly find and use the necessary visualizations and diagrams in any working situation

■ **High quality illustrations:** all images and graphics in the printed edition are presented in maximum quality

■ **Quick access to information:** easy navigation, the ability to make notes, bookmarks and work with the book anywhere.

By purchasing the full printed version of the book, you get a convenient tool for comfortable and efficient work with information: the ability to quickly use visual materials in everyday tasks, quickly find the necessary schemes and make notes. In addition, your purchase supports the dissemination of open knowledge.

You can order a print version of the book at: datadrivenconstruction.io/books



I PART

FROM CLAY TABLETS TO THE DIGITAL REVOLUTION: HOW INFORMATION HAS EVOLVED IN CONSTRUCTION

The first part of the book examines the historical evolution of data management in the construction industry, from primitive records on physical media to modern digital ecosystems. It analyzes the transformation of information management technologies, the emergence of ERP -systems and the impact of data fragmentation on the efficiency of business processes. Special attention is paid to the process of information digitalization and the growing importance of objective analysis instead of subjective expert assessments. The exponential growth of information volumes faced by the modern construction industry and the associated challenges for enterprise systems are examined in detail. The positioning of the construction industry in the context of the fourth and fifth industrial revolutions is explored, as well as the potential of using artificial intelligence and data-centric approaches to create sustainable competitive advantage.

CHAPTER 1.1.

EVOLUTION OF DATA USE IN THE CONSTRUCTION INDUSTRY

The birth of the data era in construction

About 10,000 years ago, in the Neolithic era, mankind made a revolutionary transition in its development, abandoning the nomadic lifestyle in favor of sedentary life, which led to the appearance of the first primitive buildings made of clay, wood and stone [6]. From this moment the history of the construction industry begins.

As civilizations developed, architecture became increasingly complex, leading to the first ritual temples and public buildings. The increasing complexity of architectural designs required the engineers and managers of antiquity to create the first records and calculations. The first records on clay tablets and papyri often included a description of the logic behind calculating the amount of building materials needed, their cost, and calculating payment for the work performed [7]. Thus began the era of data use in construction - long before the advent of modern digital technologies (Fig. 1.1-1).

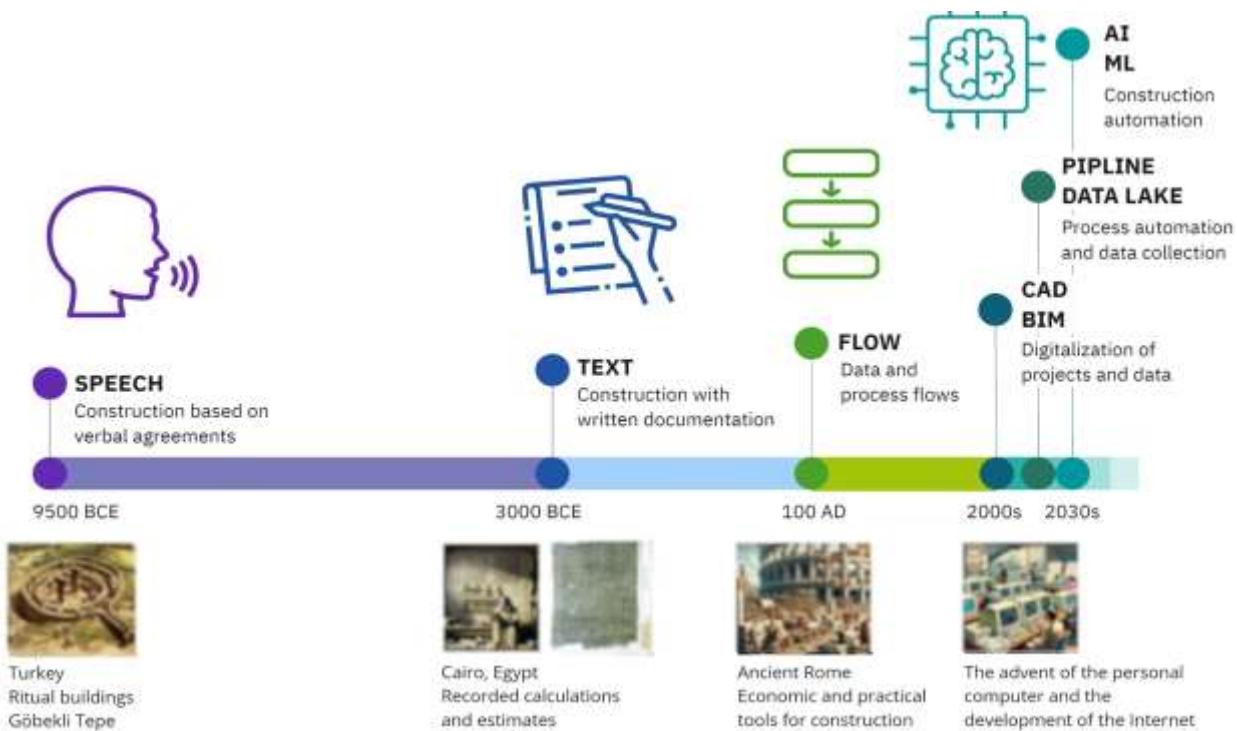


Fig. 1.1-1 Chronology of information technology development in construction: from verbal information to artificial intelligence.

From clay and papyrus to digital technology

The first documentary evidence in construction dates back to the period of pyramid building, around 3000-4000 BC[7]. Since then, the keeping of written records has facilitated and accompanied progress in the construction industry, allowing for the accumulation and systematization of knowledge that led to significant innovations in construction methods and architecture over the next 10,000 years.

The use of the first physical media in construction, such as clay tablets, papyrus from thousands of years ago (Fig. 1.1-2) or "A0" paper in the 1980s, to record data was not originally intended to apply this information to new projects. The main purpose of such records was to detail the current status of the project, including calculations of the materials required and the cost of the work. Similarly, in today's world, the availability of digital design data and models does not always guarantee their application in future projects and often serves mainly as information for current calculations of required materials and construction costs.

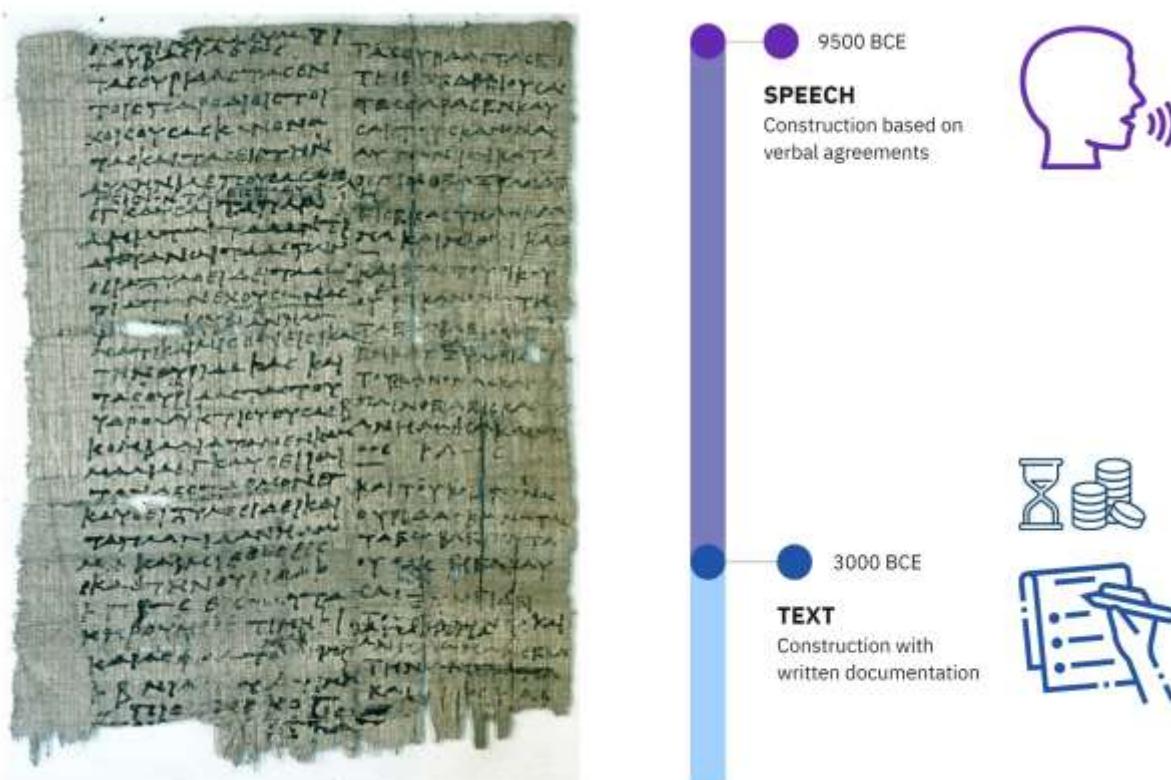


Fig. 1.1-2 A 3rd century BCE papyrus describing the cost of painting various types of windows in a royal palace using the encaustic technique.

It took humanity about 5,000 years to move from verbal conversations to written documents in construction project management, and the same amount of time to move from paper to digital data as the primary resource for planning and control.

Just as the development of trade and monetary relations stimulated the emergence of writing and the first lawyers to resolve disputes, so the first records of material costs and scopes of work in construction led to the emergence of the first managers in the construction industry, whose duties included documenting, monitoring, and being responsible for key information about project schedules and costs.

Today, data play a much more significant role: they not only record the decisions made, but also become a tool for predicting and modeling the future. This is the foundation on which the modern process approach in project management is built - turning accumulated experience into a decision-making system based on structured and verifiable data.

Process as a tool for data-driven experience

At the heart of any process is the transformation of past experience into a tool for planning the future. Experience in the modern sense is a structured set of data, the analysis of which allows making reasonable forecasts.

It is historical data that serves as the foundation of forecasting, as it clearly demonstrates the results of the work performed and provides insight into the factors affecting those results.

Let's take a concrete example from monolithic construction: usually when planning the timing of works, the volume of concrete, the complexity of the structure and weather conditions are taken into account. Suppose that a particular site foreman or the company's historical data for the last three years (2023-2025) show that pouring a 200 m² monolithic structure in rainy weather took between 4.5 and 6 days (Fig. 1.1-3). These accumulated statistics become the basis for predicting lead times and costing resources when planning similar work in future projects. Based on this historical data, the foreman or estimator can make an informed forecast, based on experience, of the time required to complete future similar work in 2026 under similar conditions.

In this case of time-analytic assessments, the analytical process acts as a mechanism for transforming disparate data into structured experience and then into a precise planning tool. Data and processes are a single ecosystem where one cannot exist without the other.

Count what can be counted, measure what can be measured, and make what cannot be measured measurable [8].

— Galileo Galilei

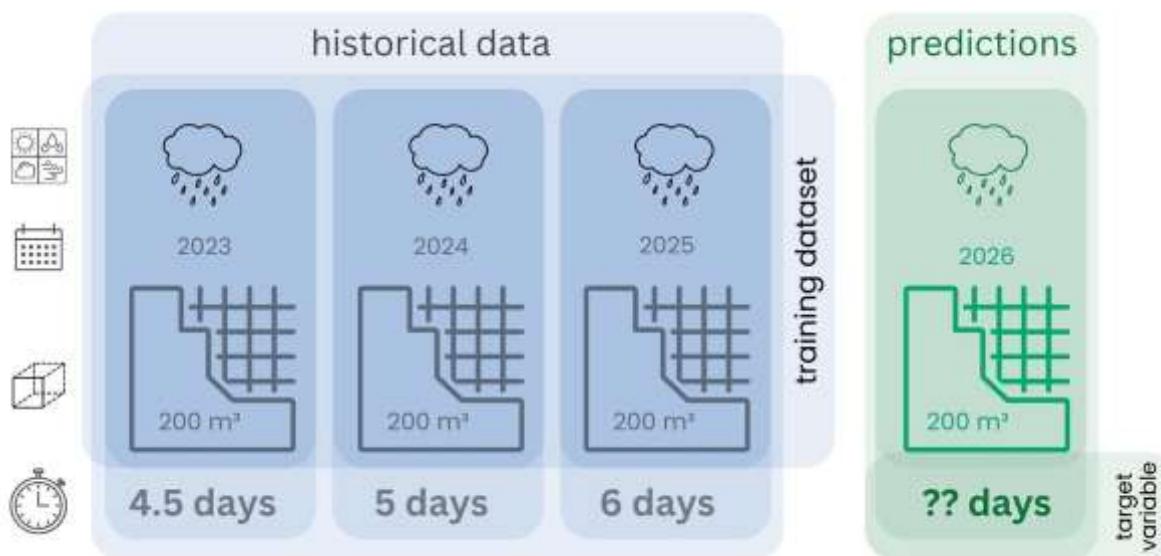


Fig. 1.1-3 Historical data acts as a training data set to predict one of the values in the future.

In today's business landscape, data analytics is becoming a critical component of effective project management, process optimization, and strategic decision making. The construction industry is gradually mastering four key levels of analytics, each of which answers a specific question and provides unique benefits (Fig. 1.1-4):

- **Descriptive analytics** - answers the question "*what happened?*" and provides historical data and reports on past events and results: over the past three years (2023-2025), it took 4.5 to 6 days to pour a 200 m² monolithic structure in rainy weather.
- **Diagnostic analytics** - answers the question "*why did this happen?*" by identifying the causes of the problems: the analysis shows that the pouring time of the monolithic structure increased due to rainy weather, which slowed down the concrete curing process.
- **Predictive analytics** - future-oriented, predicts possible risks and lead times by answering the question "*what will happen?*": based on historical data, it is predicted that pouring a similar 200 m² monolithic structure in rainy weather in 2026 will take approximately 5.5 days, taking into account all known factors and trends.
- **Prescriptive analytics** - provides automated recommendations and answers the question "*what to do?*", allowing companies to choose optimal actions: To optimize work, for example, it is recommended to: use special additives to accelerate concrete curing in high humidity.

conditions; plan pouring for periods with the lowest probability of precipitation; organize temporary shelters for the structure, which will reduce the work time to 4-4.5 days even in adverse weather conditions.

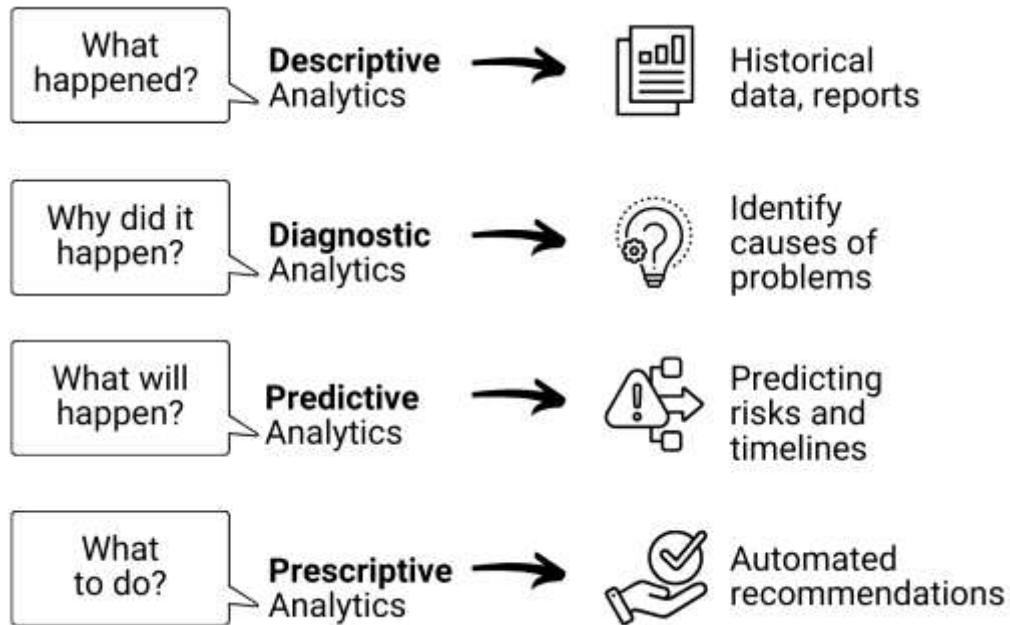


Fig. 1.1-4 Basic types of analytics: from past description to automated decision making.

Full-fledged digital transformation, which implies a transition to system analytics and data-driven management, requires not just outsourcing, but the formation of a competent internal team. The key members of such a team should be product managers, data engineers, analysts and developers, who will work in close collaboration with business units (Fig. 4.3-9). This collaboration is necessary to ask educated analytical questions and effectively parameterize business decision-making tasks. In an information society, data becomes not just an auxiliary tool, but the basis for forecasting and optimization.

In construction, digital transformation is fundamentally changing the way facilities are designed, managed and operated. This process is referred to as the digitalization of information - when all aspects of the construction process are digitized into a form suitable for analysis.

Digitalization of construction process information

For millennia, the amount of information recorded in construction has barely changed, but it has grown rapidly in recent decades (Fig. 1.1-5).

According to the PwC study® "Managed Data. What Students Need to Succeed in a Rapidly Changing Business World"(2015) [9], 90% of all data in the world has been created in the last two years (as of 2015). However, most companies are not making full use of this data as it either remains in siloed systems or is simply archived without real analysis.

The increase in data volume has only accelerated in recent years, doubling from 15 zettabytes in 2015 to 181 zettabytes in 2025 [10]. Every day the servers of construction and design companies are filled with project documentation, work schedules, calculations and calculations, financial reports. For 2D/3D -drawings DWG, DXF and DGN formats are used, and for 3D models - RVT, NWC, PLN and IFC™. Text documents, tables and presentations are saved in DOC, XLSX and PPT. In addition to video and images from the construction site - in MPG and JPEG, real-time data from IoT components, RFID® tags (identification and tracking) and BMS building management systems (monitoring and control)

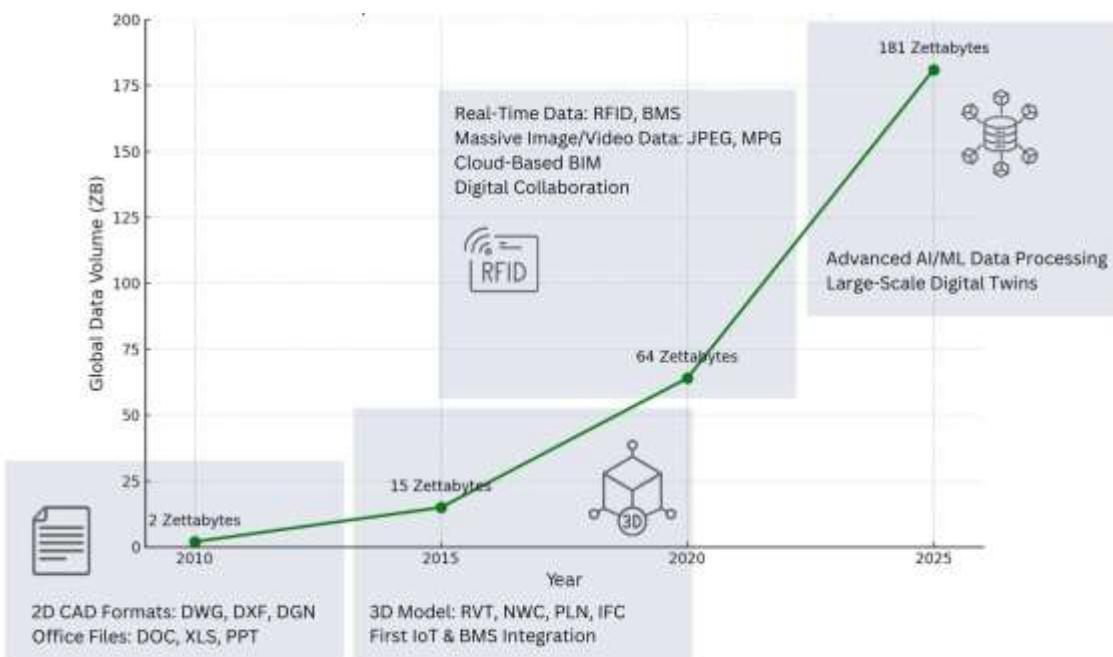


Fig. 1.1-5 Parabolic growth of data 2010-2025 (based on [10]).

With the rapid growth of information, the construction industry is faced with the need to not only collect and store data, but also to ensure its verification, validation, measurability and analytical processing. Today, the industry is going through an active phase of information digitalization - the systematic transformation of all aspects of construction activities into a digital form suitable for analysis, interpretation and automation.

Information digitalization means taking information about all entities and elements of a construction project and the construction process itself - including those we previously did not consider information at all - and converting it into a data format to make the information quantifiable and easy to analyze.

In the context of construction, this means capturing and digitizing information on all elements of projects and all processes - from the movement of machinery and people on the construction site to weather and climate conditions on the construction site, current material prices and central bank interest rates - in order to generate analytical models.

If you can measure what you are talking about and express it in numbers, then you know something about the subject. But if you cannot quantify it, your knowledge is extremely limited and unsatisfactory. It may be a starting point, but it is not the level of true scientific knowledge. [11].

– W. Thomson (Lord Kelvin), 1824-1907, British scientist

The digitalization of information goes far beyond the traditional approach to information collection, where only basic metrics such as man-hours or actual material costs were captured. Today, virtually any event can be transformed into a stream of data suitable for deep analysis using advanced analytics tools and machine learning techniques. The construction industry has undergone a fundamental shift from paper drawings, Excel spreadsheets and verbal instructions to digital systems (Fig. 1.2-4) in which every element of a project becomes a data source. Even employees - from engineers to construction workers on site - are now viewed as a collection of digital variables and data sets.

According to KPMG's "Familiar Challenges - New Approaches: Global Construction Outlook 2023", digital twins, AI (AI) and Big Data, are emerging as key drivers for improving project profitability [2].

Modern technologies not only simplify information collection, making it largely automatic, but also radically reduce the cost of data storage. As a result, companies are abandoning a selective approach and prefer to store the entire array of information for later analysis (Fig. 2.1-5), which opens up potential opportunities for optimizing processes in the future.

The digitalization of information and digitalization make it possible to reveal the hidden, previously untapped value of information. With proper organization, data can be reused, reinterpreted and integrated into new services and solutions.

In the future, the digitalization of information is likely to lead to the full automation of document management, the introduction of self-managed construction processes and the emergence of new professions - construction data analysts, AI project management experts and digital engineers. Construction projects will become dynamic sources of information, and decision-making will be based not on intuition or subjective experience, but on reliable and reproducible digital facts

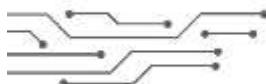
Information is the oil of the 21st century, and analytics is the internal combustion engine [12].

— Peter Sondergaard, Senior Vice President, Gartner®

According to IoT Analytics 2024 [13], global spending on data management and analytics is expected to grow dramatically from \$185.5 billion in 2023 to \$513.3 billion by 2030, with a compound annual growth rate of 16%. However, not all components are growing at the same rate: analytics is growing rapidly, while storage growth is slowing. Analytics will provide the fastest growth in the data management ecosystem: it is projected to grow from \$60.6 billion in 2023 to \$227.9 billion by 2030, a compound annual growth rate of 27%.

With the accelerating digitalization of information and the rapid growth of information volumes, construction project and company management is faced with the need to systematically store, analyze and process diverse, often heterogeneous data. In response to this challenge, starting in the mid-1990s, the industry began a massive shift to electronic creation, storage and management of documentation - from spreadsheets and design calculations to drawings and contracts.

Traditional paper documents that require signatures, physical storage, regular revision and archiving in cabinets are gradually being replaced by digital systems that store data in a structured way - in databases of specialized applications.



CHAPTER 1.2.

TECHNOLOGIES AND MANAGEMENT SYSTEMS IN MODERN CONSTRUCTION

Digital revolution and the emergence of modular MRP/ERP-systems

The era of modern digital data storage and processing began with the advent of magnetic tape in the 1950s, which opened up the possibility of storing and utilizing large amounts of information. The next breakthrough was the advent of disk drives, which radically changed the approach to data management in the construction industry.

With the development of data warehousing, a large number of companies have entered the solution market and started developing modular software to create, store, process data and automate routine tasks

The exponential growth of information and tools has led to the need for integrated, modular solutions that don't work with individual files, but help manage and control the flow of data across processes and projects.

The first comprehensive platform tools had to not only store documents, but also document all change requests and operations in processes: who initiated them, what was the scope of the request, and what was finally recorded as a value or attribute. For these purposes, a system was needed that could track accurate calculations and decisions made (Fig. 1.2-1). Such platforms were the first MRP (Material Requirements Planning) and ERP (Enterprise Resource Planning) systems that gained popularity since the early 1990s [14]

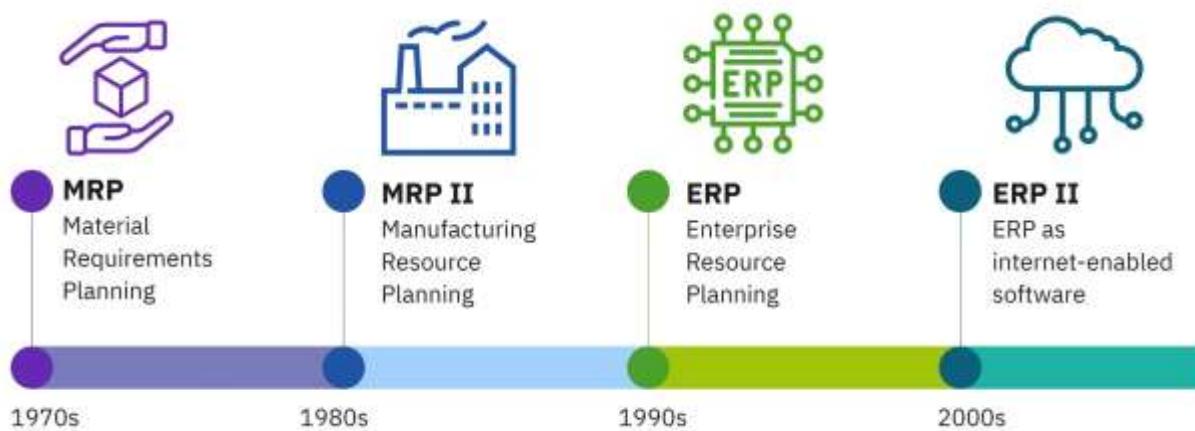


Fig. 1.2-1 Advances in data storage technology led to the emergence of ERP -systems in the 1980s.

The first MRP- and ERP- systems laid the foundation for the digitalization era in business process and construction project management. Modular systems, originally designed to automate key business processes, have over time become integrated with additional, more flexible and adaptive software solutions.

These additional solutions were designed for data processing and project content management (Fig. 1.2-2), they either replaced certain modules of large systems, or effectively complemented them, extending the functionality of the entire system.

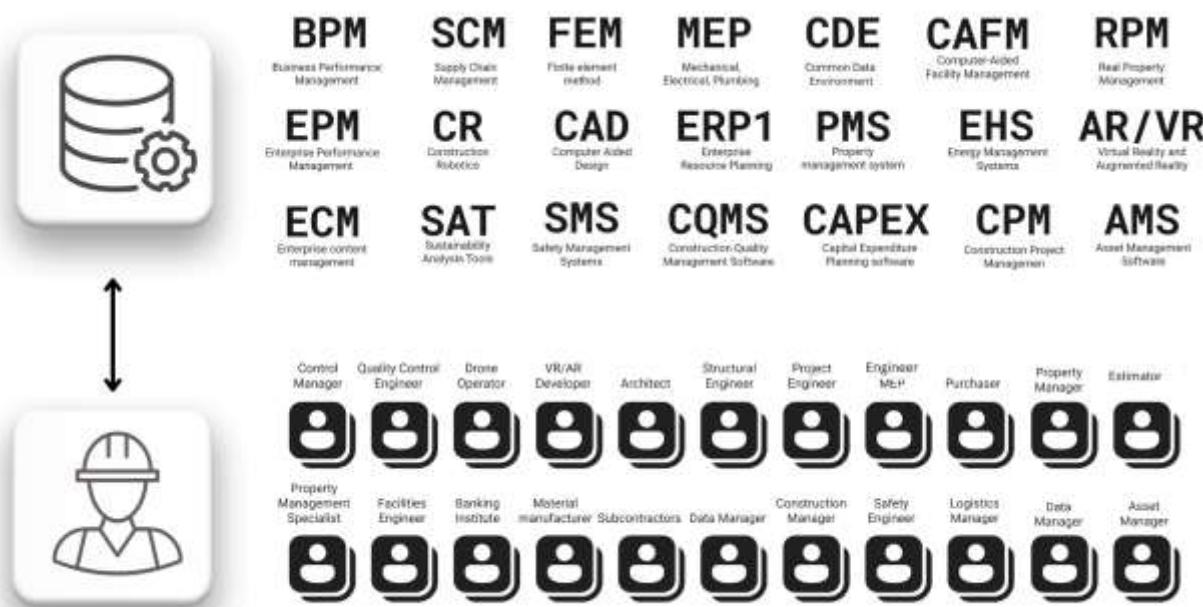


Fig. 1.2-2 New software solutions have attracted an army of managers into the business to manage data flows.

Over the past decades, companies have invested heavily in modular systems [15], perceiving them as long-term integrated solutions.

According to the Software Path report for 2022 [16], the average budget per user of an ERP - system is \$9,000. On average, about 26% of the company's employees use such systems. Thus, for an organization with 100 users, the total cost of ERP implementation reaches approximately \$900,000.

Investments in proprietary, closed, modular solutions are becoming less and less justified against the backdrop of the rapid development of modern, flexible and open technologies. If such investments have already been made, it is important to objectively reevaluate the role of existing systems: whether they remain necessary in the long term, or whether their functions can be revised and implemented more efficiently and transparently.

One of the key problems with today's modular data platforms is that they centralize data management within closed applications. As a result, data - a company's primary asset - becomes dependent on

specific software solutions, rather than the other way around. This limits information reuse, complicates migration and reduces business agility in a rapidly changing digital landscape.

If it is likely that the value or relevance of closed modular architecture will diminish in the future, it makes sense to recognize the costs incurred today as sunk costs and focus on a strategic shift to a more open, scalable and adaptive digital ecosystem.

Proprietary software is characterized by exclusive control by the development company over the source code and user data created as part of the use of such solutions. Unlike open source software, users do not have access to the internal structure of the application and cannot independently review, modify or adapt it to their needs. Instead, they are required to purchase licenses that grant the right to use the software within the limits set by the vendor.

A modern data-centric approach offers a different paradigm: data should be viewed as a core strategic asset - independent, durable and separate from specific software solutions. Applications, in turn, become mere data tools that can be freely replaced without the risk of losing critical information.

The development of ERP and MRP systems in the 1990s (Fig. 1.2-1) provided businesses with powerful tools for process management, but it also had the unintended consequence of significantly increasing the number of employees dedicated to maintaining information flows. Instead of automating and simplifying operational tasks, such systems often created new levels of complexity, bureaucracy and dependence on internal IT resources.

Data management systems: from data mining to business challenges

Today's companies are faced with the need to integrate multiple data management systems. Selecting data management systems, managing these systems well, and integrating disparate data sources is becoming critical to business performance.

In the mid-2020s, you can find hundreds (and thousands in large construction companies) of different systems (Fig. 1.2-3) that must work in harmony to make all aspects of the construction process run smoothly and cohesively.

According to Deloitte's 2016 study® "Data-Driven Management in Digital Capital Projects" - the average construction professional uses 3.3 software applications daily, but only 1.7 of them are integrated with each other [17].

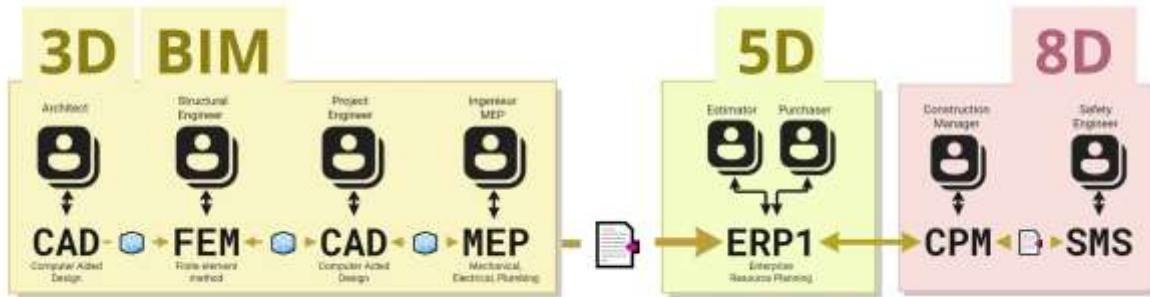


Fig. 1.2-3 Every business system requires a professional team and a responsible manager for quality data management.

The following is a list of popular systems for medium to large companies in the construction industry that are used in effective construction project management:

- **ERP (Enterprise Resource Planning)** - provides integration of business processes including accounting, procurement and project management.
- **CAPEX (Capital Expenditure Planning Software)** - used for budgeting and managing financial investments in construction projects, helps to determine the cost of fixed assets and investments in long-term assets.
- **CAD (Computer-Aided Design) and BIM (Building Information Modeling)** - are used to create detailed and accurate technical drawings and 3D -models of projects. The focus of these systems is on working with geometric information.
- **MEP (Mechanical, Electrical, Plumbing)** - Engineering systems that include mechanical, electrical, and plumbing components, and detail the internal "circulatory" system of a project.
- **GIS (Geographic Information Systems)** - used for terrain analysis and planning, including cartography and spatial analysis.
- **CQMS (Construction Quality Management Software)** - ensures that construction processes comply with established standards and regulations, helping to eliminate defects.
- **CPM (Construction Project Management)** - includes the planning, coordination and control of construction processes.
- **CAFM (Computer-Aided Facility Management)** - building management and maintenance systems.
- **SCM (Supply Chain Management)** is necessary to optimize the flow of materials and information between suppliers and the construction site.
- **EPM (Enterprise Performance Management)** - aimed at improving business processes and performance.
- **AMS (Asset Management Software)** - used to optimize the use, management and maintenance of equipment and infrastructure throughout the asset lifecycle.
- **RPM (Real Property Management)** - includes tasks and processes related to the management and operation of buildings and land, as well as related resources and assets.

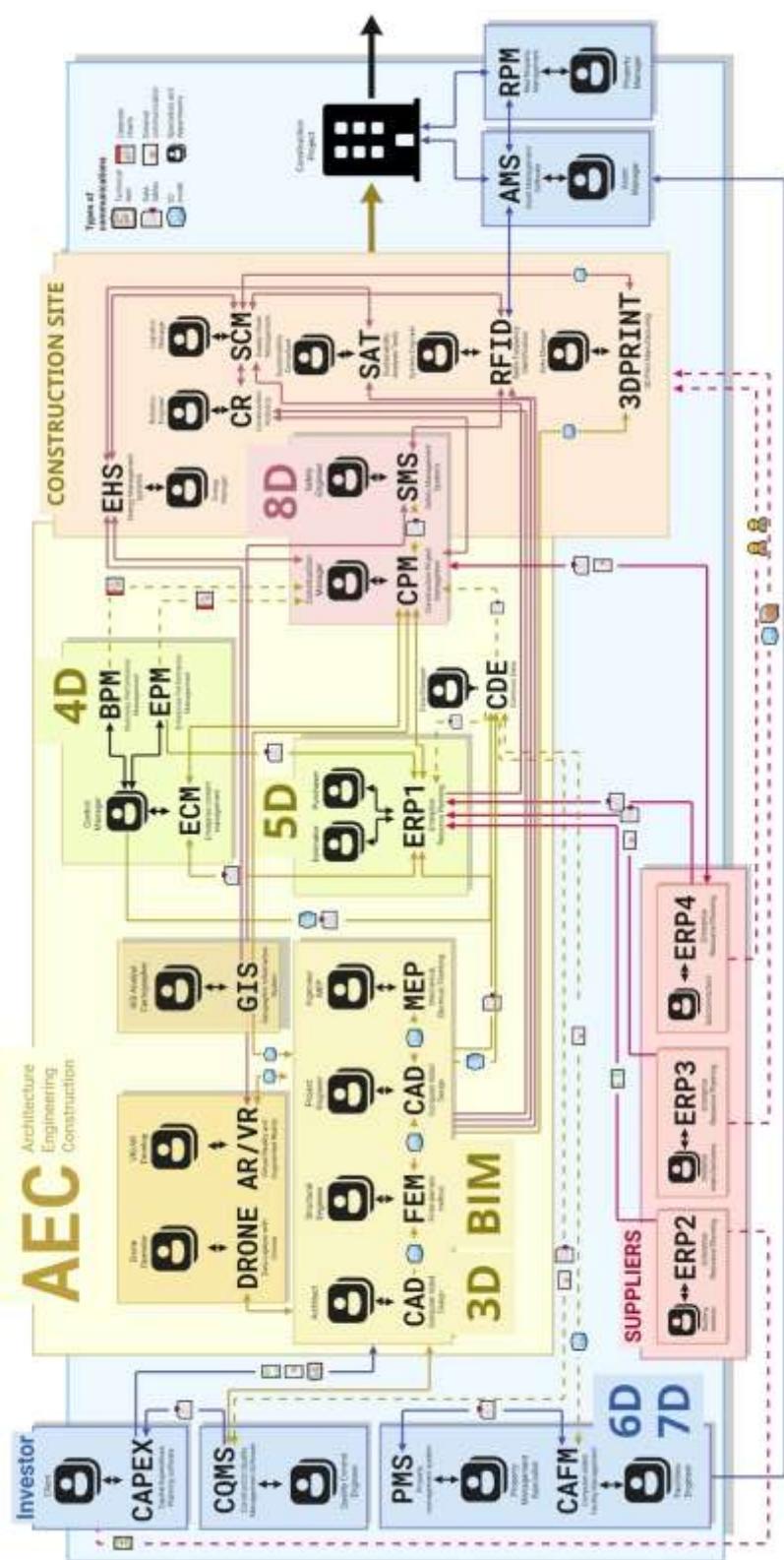


Fig. 1.2-4 Interconnectivity of systems that connects the company's processes with the flow of information between different departments.

- **CAE (Computer-Aided Engineering)** - Computer-aided engineering, includes computational and simulation systems such as finite element analysis (FEA) and computational fluid dynamics (CFD).
- **CFD (Computational Fluid Dynamics)** - Computational fluid dynamics, modeling of fluid and gas flows. CAE subcategory.
- **CAPP (Computer-Aided Process Planning)** - Computer-Aided Process Planning. It is used to create route and process maps.
- **CAM (Computer-Aided Manufacturing)** - computer-**aided** manufacturing, generation of control programs for CNC machines.
- **PDM (Product Data Management)** - Product Data Management, a system for storing and managing technical documentation.
- **MES (Manufacturing Execution System)** is a real-time manufacturing process control system.
- **PLM (Product Lifecycle Management)** - Life Cycle Management of a project element, integrates PDM, CAPP, CAM and other systems for complete control of a product from development to disposal.

These and many other systems incorporating a variety of software solutions have become an integral part of the modern construction industry (Fig. 1.2-4). In essence, such systems are specialized databases with intuitive interfaces that provide efficient input, processing and analysis of information at all stages of design and construction. The integration of digital tools with each other not only helps to optimize work processes, but also significantly improves the accuracy of decisions, which has a positive impact on the timing and quality of project implementation.

But there is no integration in half of the cases. According to statistics, only every second application or system is integrated with other solutions [17]. This indicates the continuing fragmentation of the digital environment and emphasizes the need to develop open standards and unified interfaces to ensure end-to-end information exchange within a construction project.

One of the main challenges in integration for modern companies remains the high complexity of digital systems and the requirements for user competence necessary for effective information retrieval and interpretation. A team of specialists, headed by a key manager, is formed to support each system implemented in the business (Fig. 1.2-2).

The key system manager plays a crucial role in directing the flow of data in the right direction and is responsible for the quality of the final information, just as the first managers thousands of years ago were responsible for the numbers written on papyrus or clay tablets.

To turn disparate information flows into a management tool, the ability to systematically integrate and manage data is essential. In this architecture, managers must act as elements of a unified network - like a mycelium that connects the individual parts of the company into a holistic living organism capable of adapting and evolving.

Corporate mycelium: how data connects to business processes

The process of integrating data into applications and databases relies on the aggregation of information from a variety of sources, including different departments and specialists (Fig. 1.2-4). Specialists search for relevant data, process it, and transfer it to their systems and applications for further use.

Each company system, consisting of a set of tools, technologies and databases, is a knowledge tree rooted in the soil of historical data and growing to bear new fruit in the form of ready-made solutions: documents, calculations, tables, graphs and dashboards (Fig. 1.2-5). The systems in a company, like trees in a particular patch of forest interact and communicate with each other, representing a complex and well-structured system, supported and managed by expert managers.

A company's information retrieval and transfer system works like a complex forest network consisting of trees (systems) and mycelium mushrooms (managers) that act as conductors and recyclers, ensuring that information is transferred and flows to the right systems. This helps to maintain a healthy and efficient flow and distribution of data within the company.

Experts, like roots, absorb raw data at the initial stages of a project, turning it into nutrients for the corporate ecosystem. Data and content management systems (Fig. 1.2-4 - ERP, CPM, BIM, etc.) act as powerful information highways through which this knowledge circulates through all levels of the company.

As in nature, where each element of the ecosystem plays its own role, in a company's business landscape, each process participant - from engineer to analyst - contributes to the growth and fertility of the information environment. These systemic "data trees" (Fig. 1.2-5) are not just mechanisms for gathering information, but a competitive advantage that ensures a company's sustainability.

Forest ecosystems are a surprisingly accurate reflection of how digital corporate structures are organized. Like the tiered structure of a forest - from the undergrowth to the treetops - corporate governance categorizes tasks by levels of responsibility and functional departments.

Deep and branching tree roots provide resilience and access to nutrients. Similarly, a strong organizational structure and stable processes for working with quality data support the entire information ecosystem of a company, contributing to its sustainable growth and development even during periods of (high wind) market instability and crises.

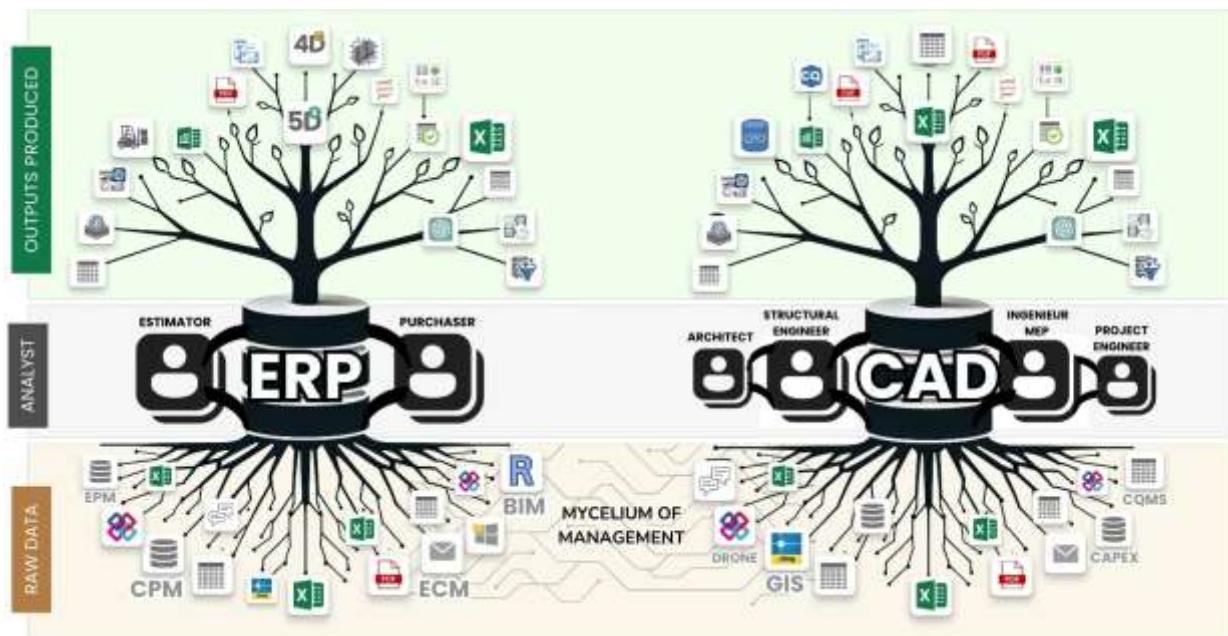


Fig. 1.2-5 Integrating data through different systems is like a mycelium that connects managers and specialists into a single information network.

The modern understanding of scale in business has evolved. Today, the value of a company is determined not only by its visible part - the "crowns" in the form of final documents and reports - but also by the depth of the "root system" of qualitatively collected and systematically processed data. The more information can be collected and processed, the higher the business value becomes. Companies that methodically accumulate a "compost" of already processed data and are able to extract useful insights from it gain a strategic advantage

Historical information is becoming a new kind of capital, enabling growth, process optimization and competitive advantage. In a data-driven world, it's not who has more, but who knows more that wins.

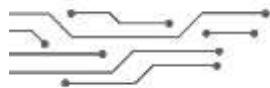
For the construction industry, this means moving to real-time project management, where all processes - from design and procurement to contractor coordination - will be based on relevant, daily updated data. The integration of information from various sources (ERP -systems, CAD -models, sensors IoT on construction sites, RFID) will make it possible to make more accurate forecasts, react quickly to changes and avoid delays caused by the lack of up-to-date data.

According to the Data-Driven Enterprise 2025 study (McKinsey & Company®, 2022 [18]), successful companies of the future will rely on data in all key aspects of their operations, from strategic decisions to operational interactions.

Data will cease to be just an analytics tool and will become an integral part of all business processes, providing transparency, control and automation of management. Data-driven under move will allow

organizations to minimize the influence of human factor, reduce operational risks and increase transparency and efficiency of decision-making.

The 21st century is turning the economic paradigm upside down: whereas oil used to be called "black gold" for its ability to power machinery and transportation, today, compressed under time pressure, historical data is becoming a new strategic resource, powering not machines but decision-making algorithms that will drive business.



CHAPTER 1.3.

THE DIGITAL REVOLUTION AND THE EXPLOSION OF DATA

The beginning of the data volume boom as an evolutionary wave

The construction industry is experiencing an unprecedented information explosion. If we think of business as a knowledge tree (Fig. 1.2-5) fed by data, the current stage of digitalization can be compared to the rapid growth of vegetation during the Carboniferous period, an era in which the Earth's biosphere was transformed by the rapid accumulation of biomass (Fig. 1.3-1).

With global digitalization, the amount of information in the construction industry is doubling every year. Modern technologies allow data to be collected in the background, analyzed in real time and used on a scale that seemed impossible just a short time ago.

According to Moore's Law, formulated by Gordon Moore (co-founder of Intel®), the density and complexity of integrated circuits and the amount of data processed and stored doubles approximately every two years [19].

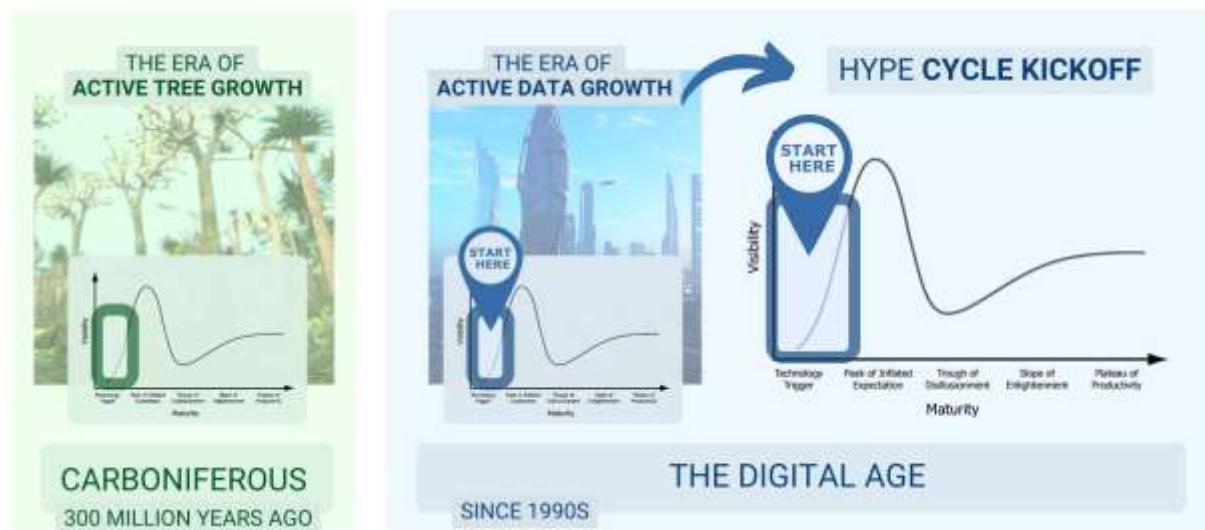


Fig. 1.3-1 The onset of digitalization has led to an exponential growth of data, much like the burst of vegetation in the coal age.

While ancient megalithic structures such as Göbekli Tepe (Turkey) did not leave behind documented knowledge suitable for reuse, today digital technologies make it possible to accumulate and reuse information. This can be compared to the evolutionary transition from spore plants to seed plants (angiosperms): the emergence of the seed gave rise to the widespread spread of life on the planet. (Fig. 1.3-2).

Similarly, data from past projects become a kind of "digital seeds" - DNA knowledge carriers that can be scaled and used in new projects and produces. The emergence of modern artificial intelligence

tools - machine learning and large language models (LLMs) such as ChatGPT, LlaMa, Mistral, Claude, DeepSeek, QWEN, Grok - allow data to be automatically extracted, interpreted and applied in new contexts

Just as seeds revolutionized the spread of life on an initially lifeless planet, "data seeds" are becoming the basis for the automatic emergence of new information structures and knowledge, allowing digital ecosystems to evolve independently and adapt to changing user requirements.

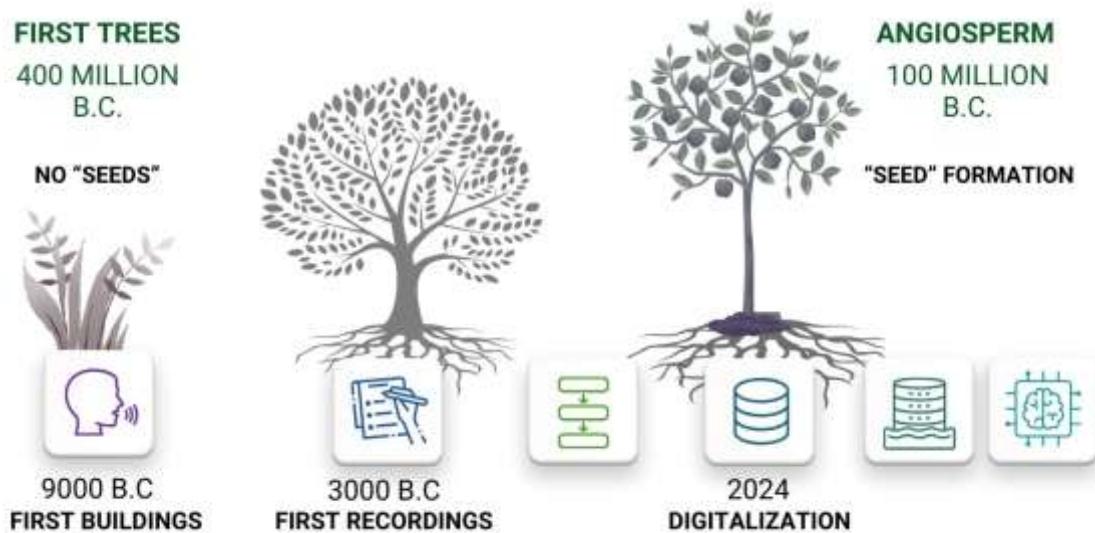


Fig. 1.3-2 Digital "data seeds" play the same evolutionary role as angiosperms, the flowering plants that transformed the Earth's ecosystem.

We stand at the threshold of a new era in construction, where the explosion of data and the active dissemination of "data seeds" - structured information from past and ongoing projects - are forming the foundation of the industry's digital future. Their "pollination" through big data language models (LLMs) allows us to not just observe digital change, but to actively participate in the creation of self-learning, adaptive ecosystems. This is not evolution - it's a digital revolution in which data is becoming the main building block of a new reality

The amount of data in the construction industry is increasing dramatically due to information from various disciplines throughout the life cycle of construction projects. This huge accumulation of data has pushed the construction industry towards the era of Big Data [20].

— Prof. Hang Yang, Department of Civil Engineering and Architecture, Wuhan University of Technology, Wuhan

The growth of data in the information age is reminiscent of evolutionary processes in nature: just as the development of forests changed the ancient landscape of the planet, the current information explosion is changing the landscape of the entire construction industry.

The amount of data generated in a modern company

In the last two years, 90% of all existing data in the world has been created [21]. As of 2023, each person, including construction industry professionals, generates about 1.7 megabytes of data per second [22], and the total amount of data in the world will reach 64 zettabytes in 2023 and is projected to exceed 180 zettabytes, or 180×10^{15} megabytes, by 2025 [23].

This information explosion has a historical precedent - the invention of the printing press by Johannes Gutenberg in the fifteenth century. Just fifty years after its introduction, the number of books in Europe doubled: in a few decades, as many books were printed as had been created by hand over the previous 1,200 years [24]. Today, we are seeing even more rapid growth: the amount of data in the world doubles every three years.

Given the current rate of data growth, the construction industry has the potential to generate as much information in the next few decades as it has accumulated in its entire previous history



Fig. 1.3-3 Each employee's daily storage of data on the company's servers contributes to an ever-increasing amount of data.

In today's world of construction business, even small companies generate a huge amount of multiformat information on a daily basis and the digital footprint of even a small construction company can

reach tens of gigabytes per day - from models and drawings to photographic records and sensors on site. If we assume that each technician generates on average about 1.7 MB of data per second, this is equivalent to about 146 GB per day, or 53 TB per year (Fig. 1.3-3).

When a team of 10 people work actively for only 3 hours daily, the cumulative amount of information generated per day reaches 180 gigabytes (Fig. 1.3-4).

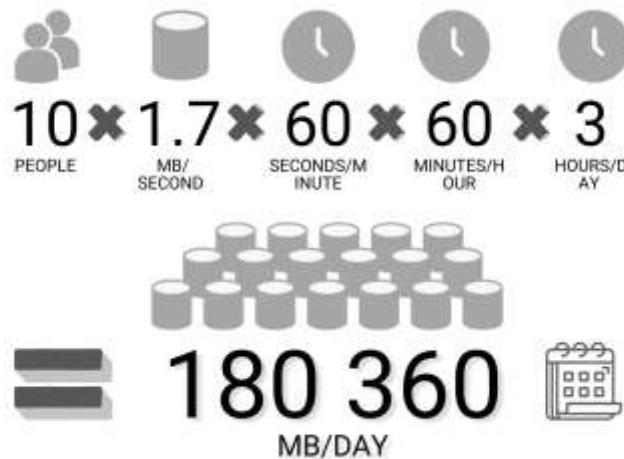


Fig. 1.3-4 A company of 10 people generates approximately 50-200 gigabytes of data per day.

Assuming that 30% of work data is new (the rest is overwritten or deleted), a 10-person firm can create on the order of several hundred gigabytes of new data per month (actual numbers depend on the type of business company does).

Thus, it becomes obvious: we are not just generating more and more data - we are facing a growing need for its efficient management, storage and long-term availability. And while previously data could "lie" on local servers at no cost, in the context of digital transformation, more and more companies are starting to use cloud solutions as the basis of their information infrastructure.

Cost of data storage: the economic aspect

In recent years, more and more companies are outsourcing data storage to cloud services. For example, if a company hosts half of its data in the cloud, at an average price of \$0.015 per gigabyte per month, its storage costs may increase by \$10-50 [25] each month.

For a small company with typical data generation patterns, cloud storage costs can range from hundreds to potentially over a thousand dollars per month (Fig. 1.3-5) in a few years, creating a potentially

significant financial burden.

According to Forrester's study "Enterprises Outsource Data Storage as Complexity Grows" [26], which surveyed 214 technology infrastructure decision makers [26], which surveyed 214 technology infrastructure decision makers, more than one-third of organizations are outsourcing storage to handle the growing volume and complexity of data operations, with nearly two-thirds of enterprises preferring a subscription-based model.



Fig. 1.3-5 Moving data to the cloud can increase monthly storage costs by up to \$2,000, even for a company with as few as 10 employees.

The situation is further complicated by the accelerated adoption of cloud-based technologies such as CAD (BIM), CAFM, PMIS and ERP -systems that further increase data storage and processing costs. As a result, companies are forced to look for ways to optimize costs and reduce dependence on cloud providers.

Since 2023, with the active development of large language models (LLM), approaches to data storage have started to change. More and more companies are thinking about taking back control of their data as it becomes safer and more profitable to process information on their own servers.

In this context, the trend away from cloud-based storage and processing of only the necessary data in favor of local deployment of enterprise LLM and AI -solutions comes to the fore. As Microsoft's CEO pointed out in one of his interviews [27], instead of relying on several separate applications or cloud-based SaaS solutions to perform different tasks, AI agents will manage processes in databases, automating the functions of different systems.

[...] the old approach to this [data processing] issue was: if you think back to how different business applications handled integration, they used connectors. Companies sold licenses for those connectors, and the business model was formed around that. SAP [ERP] is one of the classic examples: you could only access SAP data if you had the right connector. So it seems to me that something similar will emerge in the case of [AI] agent interaction [...]. The approach, at least that we take, is: I think that the concept of the existence of business applications will probably collapse in the era of [AI] agents. Because if you think about it, they are essentially databases with a bunch of business logic

— Satya Nadella, Microsoft CEO, interview with BG2 channel, 2024 [28]

In this paradigm, the data-driven LLM approach goes beyond classical systems. Artificial intelligence becomes an intermediary between the user and the data (Fig. 2.2-3, Fig. 2.2-4), eliminating the need for multiple intermediary interfaces and increasing the efficiency of business processes. We will talk more about this approach to working with data in the chapter "Turning Chaos into Order and Reducing Complexity".

While the architecture of the future is still taking shape, companies are already facing the consequences of past decisions. The massive digitalization of recent decades, accompanied by the introduction of disparate systems and uncontrolled accumulation of data, has led to a new problem - information overload.

Frontiers of data accumulation: from mass to meaning

Modern company systems successfully develop and function under managed growth, when the volume of data and the number of applications are in balance with the capabilities of IT departments and managers. However, in recent decades, digitalization has led to an uncontrollable increase in the volume and complexity of data, which has caused an oversaturation effect in companies' information ecosystem.

Today, servers and storage facilities are subjected to an unprecedented influx of unprocessed and multiformat information that has no time to turn into compost and is rapidly becoming irrelevant. Limited company resources cannot cope with this deluge, and data is accumulating in isolated silos (so-called "silos") that require manual processing to extract useful information.

As a result, like a forest overgrown with ivy and covered with mold, modern company management systems often suffer from information overload. Instead of nourishing information humus, isolated areas of different-format data form the basis of the corporate ecosystem, which inevitably leads to a decrease in the overall efficiency of business processes.

The long period of exponential growth of data volumes observed over the last 40 years will inevitably be followed by a phase of saturation and subsequent cooling. When storage reaches its limit, a qualitative shift will occur: data will no longer be just a storage object, but a strategic resource.

With the development of artificial intelligence and machine learning, companies have the opportunity to reduce information processing costs and move from quantitative growth to qualitative use of data. Over the next decade, the construction industry will have to shift its focus from creating more and more data to ensuring its structure, integrity and analytical value.

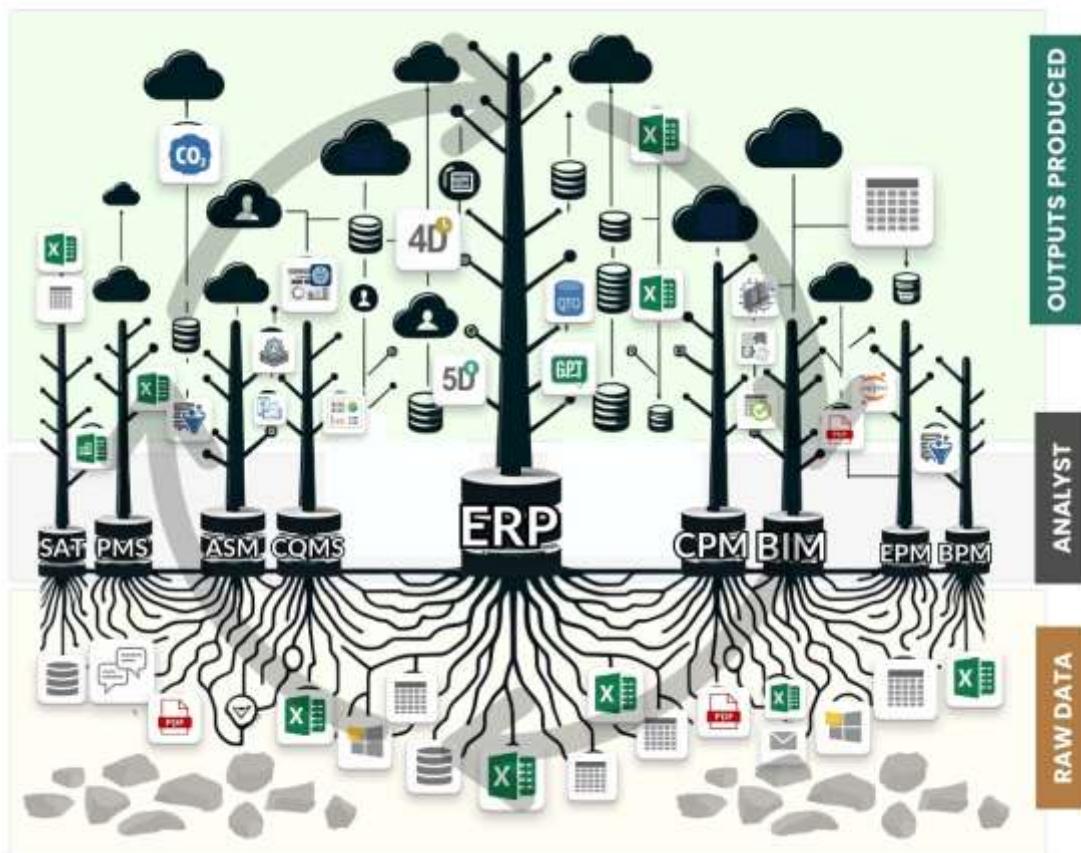


Fig. 1.3-6 Isolated data sources prevent information sharing between data systems.

The main value is no longer in the amount of information, but in the ability to automatically interpret it and turn it into applied knowledge useful for making management decisions. In order for data to become truly useful, it must be properly managed: collected, verified, structured, stored and analyzed in the context of specific business tasks.

The data analytics process in a company is similar to the cycle of life and decay of trees in a forest and the emergence of new young and strong trees: mature trees die off, decompose and become a breeding ground for new growth. Finished and completed processes at the end of their application are fed into the company's information ecosystem, eventually becoming information humus that feeds the future growth of new systems and data.

In practice, however, this cycle is often broken. Instead of organic renewal, a layered chaos is formed, similar to geological strata, where new systems are layered on top of old ones without deep integration and structuring. As a result, disparate information "silos" emerge, hindering the circulation of knowledge and complicating data management.

Next steps: from data theory to practical change

The evolution of data in construction is a journey from clay tablets to modern modular platforms. The challenge today is not to collect information, but to create a framework that turns disparate and diverse data into a strategic resource. Whether your role is that of a corporate executive or an engineer, understanding the value of data and how to work with it will be a key skill in the future.

To summarize this part, it is worth highlighting the main practical steps that will help you apply the discussed approaches in your daily tasks:

- Conduct a personal audit of information flows
 - Make a list of all the systems and applications you work with on a daily basis
 - Note where you spend the most time searching or double-checking data
 - Identify your key sources of information
 - Analyze your current application landscape for redundancy and duplication of functions
- Strive to move through the processes by levels of analytical maturity
 - Start your tasks with descriptive analytics (what happened?)
 - Gradually introduce diagnostic (why did this happen?)
 - Think about how in processes you can move to predictive (what will happen?) and prescriptive (what to do?) analytics
- Start structuring your work data
 - Implement a unified system for naming files and folders that you frequently use in

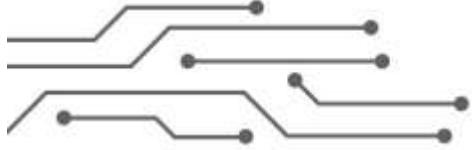
your work

- Create templates for frequently used documents and reports
- Regularly archive completed projects with a clear structure

Even if you can't change the entire information infrastructure in your team or company, start with your own processes and small improvements in your day-to-day work. Remember that the real value of data is not in its volume, but in the ability to extract actionable insights from it. Even small but properly structured and analyzed data sets can have a significant impact when integrated into decision-making processes.

In the next parts of the book, we'll move on to specific methods and tools for working with data, look at ways to transform unstructured information into structured sets, explore analytics automation technologies, and detail how to build an effective analytics ecosystem in a construction company.





II PART

HOW THE CONSTRUCTION BUSINESS IS DROWNING IN DATA CHAOS

The second part is devoted to critically analyzing the challenges faced by construction companies in dealing with increasing volumes of data. The consequences of information fragmentation and the phenomenon of "data in silos", which hinders effective decision-making, are discussed in detail. The problems of HiPPO -approach (Highest Paid Person's Opinion) and its impact on the quality of management decisions in construction projects are studied. The impact of dynamic business processes and their growing complexity on information flows and operational efficiency is assessed. Specific examples are given of how excessive system complexity increases costs and reduces the flexibility of organizations. Special attention is given to the limitations created by proprietary formats and the prospects for using open standards in the construction industry. The concept of moving towards AI and LLM-based software ecosystems is presented, which minimize excessive complexity and technical barriers.

CHAPTER 2.1.

DATA FRAGMENTATION AND SILOS

The more tools, the more efficient the business?

At first glance, it may seem that more digital tools lead to greater efficiency. In practice, however, this is not the case. With each new solution, whether it's a cloud service, a legacy system, or another Excel report, a company adds another layer to its digital landscape - a layer that is often not integrated with the rest (Fig. 2.1-1).

Data is like coal or oil: it takes years to build up, compacted under layers of chaos, errors, unstructured processes and forgotten formats. To extract truly useful information from it, companies must literally wade through layers of outdated solutions and digital noise.

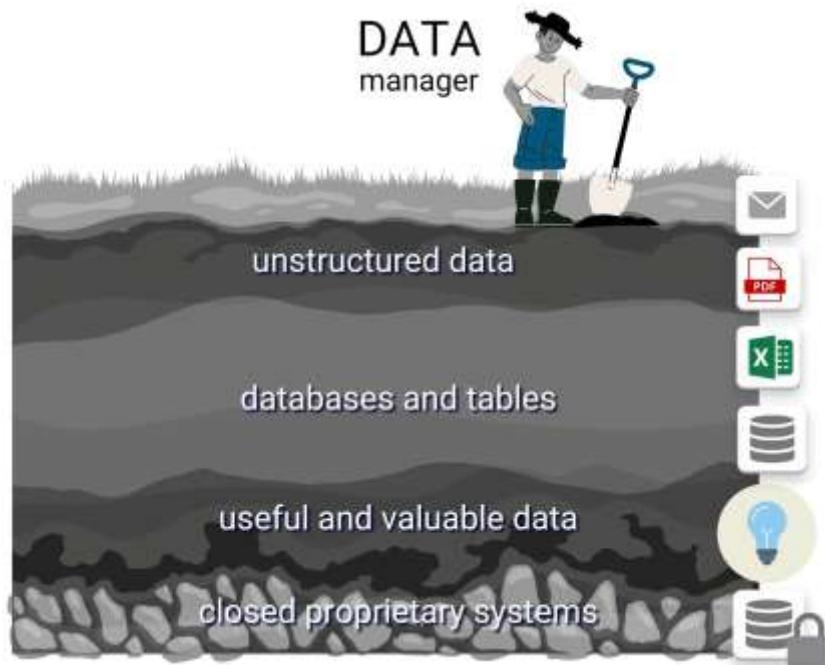


Fig. 2.1-1 Divergent data form compartmentalized layers - even "golden" insights are lost in geologic rocks of systemic complexity.

Every new application leaves behind a trail: a file, a table, or a whole isolated "silo" on the server. One layer is clay (outdated and forgotten data), another is sand (disparate tables and reports), and the third is granite (closed proprietary formats that cannot be integrated). Over time, a company's digital environment increasingly resembles a reservoir of uncontrolled information accumulation, where value is lost deep within the company's servers.

With each new project and each new system, not only the infrastructure becomes more complex, but also the path to useful quality data. Getting to the valuable "rock" requires deep cleansing, structuring

information, "chunking" it, grouping it into meaningful pieces, and extracting strategically important insights through analytics and data modeling.

Data is a valuable thing, and it will last longer than the systems [that process the data] themselves [29].

— Tim Berners-Lee, father of the World Wide Web and creator of the first Web site

Before data can become a "thing of value" and a reliable basis for decision-making, it must be thoroughly pre-processed. It is proper preprocessing that turns disparate data into structured experience, useful information humus, which then becomes a forecasting and optimization tool.

There is a misconception that you need perfectly clean data to start analyzing, but in practice, being able to work with dirty data is an essential part of the process.

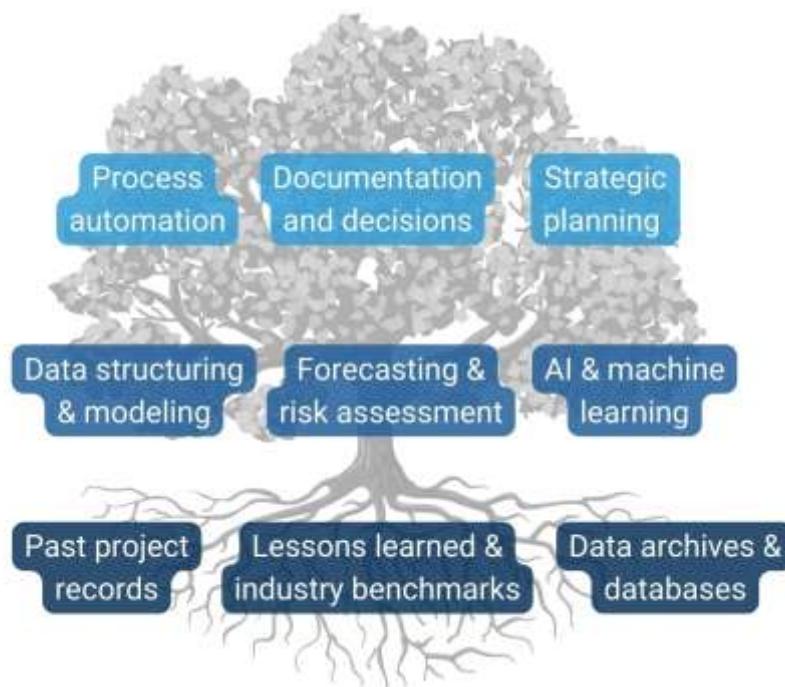


Fig. 2.1-2 Data is the root system and foundation of business, which in turn is based on decision-making processes.

As technology continues to advance, your business must also move forward and learn how to create value from data. Just as oil and coal companies are building the infrastructure to extract minerals, so too must businesses learn how to manage the flow of new information on their own servers and extract

valuable insights from unused, unformatted and outdated data, turning it into a strategic resource.

Creating fields (data warehouses) is the first step. Even the most powerful tools don't solve the problem of data isolation and multi-format data if companies continue to operate in siloed systems. When data exists separately from each other, without intersecting and sharing information, businesses face a "data silo" effect. Instead of a single, consistent infrastructure, companies are forced to spend resources on merging and synchronizing data.

Data silos and their impact on company performance

Imagine that you are building a residential complex, but each team has its own project. Some are building walls, others are laying communications, and others are paving roads without checking with each other. As a result, the pipes do not match the openings in the walls, the elevator shafts do not correspond to the storeys, and the roads have to be dismantled and re-laid.

This situation is not just a hypothetical scenario, but a reality of many modern construction projects. Due to the large number of general and subcontractors working with different systems and without a single coordinating center, the process turns into a series of endless approvals, rework and conflicts. All of this leads to significant delays and multiple project costs.

A classic situation on a construction site is a simple one: the formwork is ready, but the delivery of reinforcement has not arrived on time. When checking information in various systems, the communication is roughly as follows:

- ⌚ **The foreman** at the construction site on the 20th writes to the project manager, "*We finished setting the formwork, where are the rebar?*"
- ⌚ **Project Manager** (PMIS) to the Procurement Department: - "*The formwork is ready. In my [PMIS] system, the rebar was supposed to arrive on the 18th. Where are the rebar?*"
- ⌚ **Supply Chain Specialist** (ERP): - "*Our ERP says delivery will be on the 25th*"
- ⌚ **Data Engineer** or IT (responsible for integrations): - "*In PMIS the date is on the 18th, in ERP it is on the 25th. There is no OrderID link between ERP and PMIS, so the data is not synchronized. This is a typical example of an information gap*".
- ⌚ **Project manager** to general director - "*The delivery of fittings is delayed, the site is standing, and who is responsible is unclear*".

The cause of the incident was the isolation of data in disparate systems. By integrating and unifying data sources, creating a single repository of information, and automating through ETL-tools (Apache NiFi, Airflow, or n8n), the silos between systems can be eliminated. These and other methods and tools will be discussed in detail in later sections of the book.

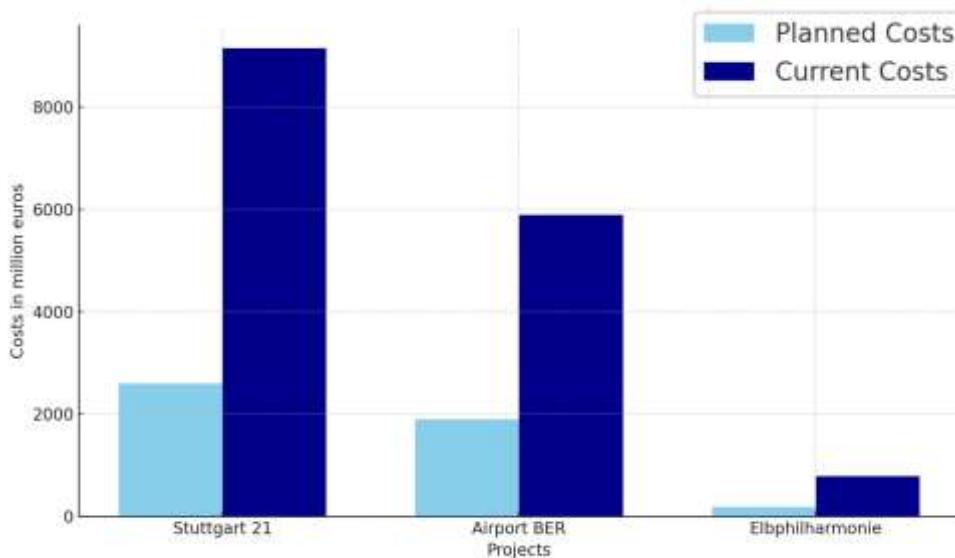


Fig. 2.1-3 Comparison of planned and actual costs of major infrastructure projects in Germany.

It's the same with enterprise systems: first you create isolated solutions, and then you have to spend huge budgets to integrate and harmonize them. If data and communication models had been thought through from the start, there would be no need for integration at all. Siloed data creates chaos in the digital world, like an uncoordinated construction process.

According to KPMG's 2023 study "Cue construction 4.0: Time to make or break", only 36% of companies share data effectively across departments, while 61% face serious problems due to isolated data "silos" [30].



Fig. 2.1-4 Years of hard-to-collect data accumulate in isolated storage "silos" at the risk of never being used.

Company data is stored in isolated systems, like individual trees scattered across the landscape. Each

contains valuable information, but the lack of connections between them prevents the creation of a single, interconnected ecosystem. This siloing hinders the flow of data and limits the organization's ability to see the full picture. Connecting these silos is an extremely long and complex process of growing mushroom mycelium at the management level to learn how to transfer individual pieces of information between systems.

According to a 2016 WEF study, one of the main barriers to digital transformation is the lack of common data standards and fragmentation.

The construction industry is one of the most fragmented in the world and depends on the smooth interaction of all participants in the value chain [5].

— World Economic Forum Report "Shaping the Future of Construction"

Designers, managers, coordinators and developers often prefer to work autonomously, avoiding the complexities of coordination. This natural inclination leads to the creation of information "silos" in which data is isolated within separate systems. The more such isolated systems there are, the more difficult it is to get them to work together. Over time, each system gets its own database and a specialized support department of managers (Fig. 1.2-4), further complicating integration.

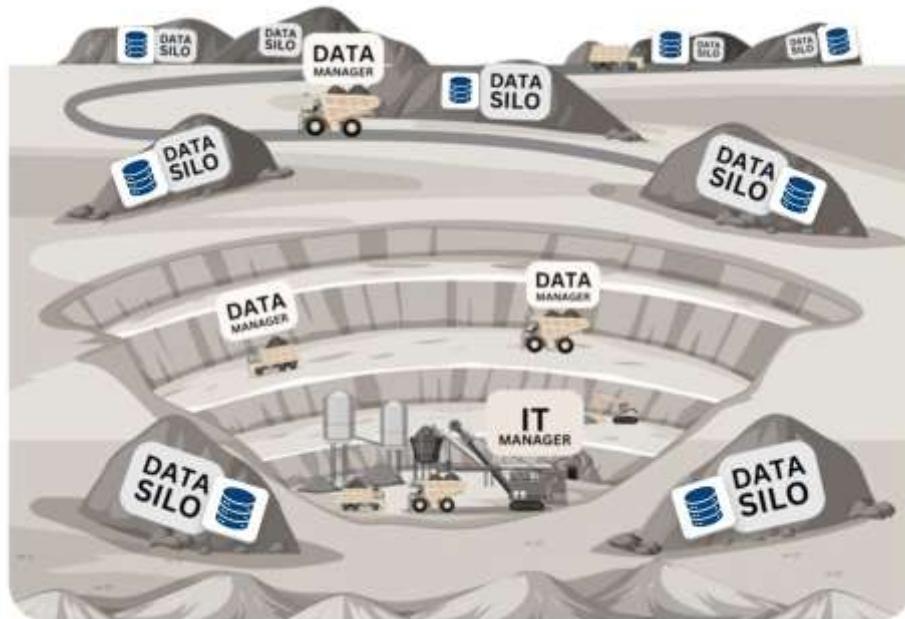


Fig. 2.1-5 Each system aims to create its own unique silo of data that needs to be processed by suitable tools [31].

The vicious circle in corporate systems looks like this: companies invest in complex iso customized solutions, then face high costs for their integration, and developers, realizing the complexity of combining systems, prefer to work in their closed ecosystems. All this increases the fragmentation of the

IT landscape and makes it more difficult to migrate to new solutions (Fig. 2.1-5). Managers end up criticizing data silos, but rarely analyze their causes and how to prevent them. Managers complain about outdated IT systems, but replacing them requires significant investment and rarely yields the expected results. As a result, even attempts to combat the problem often make matters worse.

The main reason for the disconnect is the prioritization of applications over data. Companies first develop separate systems or buy off-the-shelf solutions from vendors, and then try to unify them by creating duplicate and incompatible storage and databases.

Overcoming the problem of fragmentation requires a radical new approach - prioritizing data over applications. Companies must first develop data management strategies and data models, and then build systems or purchase solutions that work with a single set of information rather than creating new barriers.

We are entering a new world where data may be more important than software.

– Tim O'Reilly, CEO of O'Reilly Media, Inc.

McKinsey Global Institute's study "Rethinking Construction: the path to improved productivity" (2016) demonstrates that the construction industry lags behind other sectors in digital transformation [32]. According to the report, the adoption of automated data management and digital platforms can significantly improve productivity and reduce losses associated with process inconsistency. This need for digital transformation is also emphasized by Egan's (UK, 1998) report [33], which highlights the key role of integrated processes and a collaborative approach in construction.

As a result, while in the last 10,000 years the main problem for data managers has been a lack of data, with the avalanche of data and data management systems, users and managers are faced with a problem - an overabundance of data, making it difficult to find legally correct and quality information.

Disparate data silos inevitably lead to the serious problem of reduced data quality. With multiple independent systems, the same data can exist in different versions, often with conflicting values, creating additional complexity for users who need to determine which information is relevant and reliable.

Duplication, and lack of data quality as a consequence of disunity

Due to the problem of data silos, managers are forced to spend considerable time searching for and reconciling data. To hedge against quality problems, companies create complex information management structures in which a vertical of managers is responsible for searching, verifying, and reconciling data. However, this approach only increases bureaucracy and slows down decision-making. The more

data there is, the more difficult it is to analyze and interpret, especially if there is no uniform standard for its storage and processing.

With the plethora of software applications and systems that have been growing like mushrooms after rain in the last decade, the problem of silos and inappropriate data quality has become increasingly important to end users. The same data, but with different values, can now be found in different systems and applications (Fig. 2.1-6). This leads to difficulties for end users when trying to determine which version of data is relevant and correct among the many available. This leads to errors in analysis and ultimately decision making.

To insure against problems with finding the right data, company managers create a multi-level bureaucracy of verification managers. Their task is to be able to quickly find, check and send the necessary data in the form of tables and reports, navigating the maze of disparate systems.

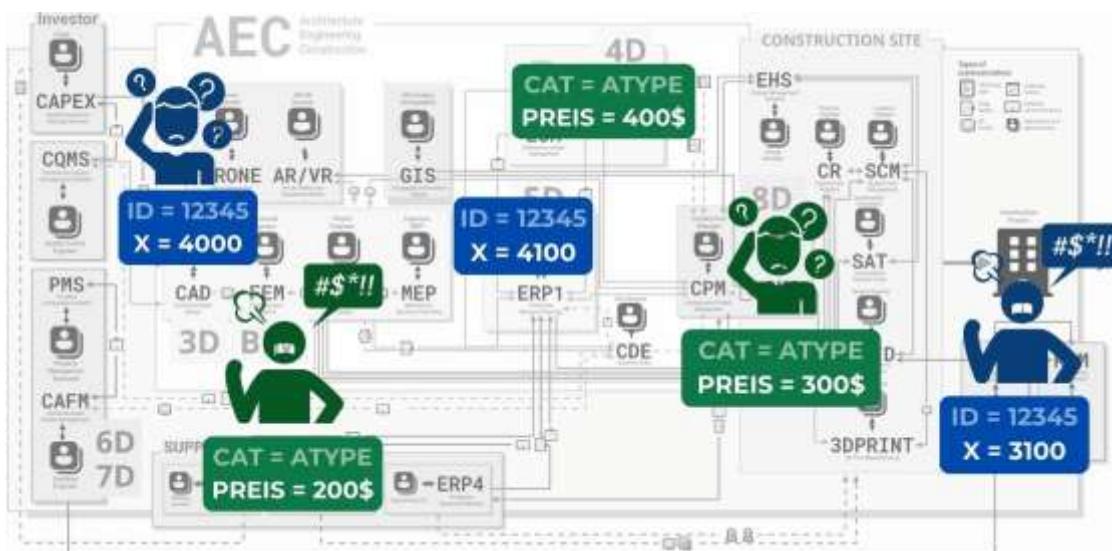


Fig. 2.1-6 In trying to find the right data, managers must ensure the quality and legal reliability of data between different systems.

In practice, however, this model creates new complexities. When data is managed manually and information is scattered across many unrelated decisions, every attempt to obtain accurate and up-to-date information through a pyramid of decision-makers (Fig. 2.1-7) becomes a bottleneck - time-consuming and error-prone.

The situation is exacerbated by the avalanche of digital solutions. The software market continues to be flooded with new tools that seem promising. But without a clear data management strategy, these solutions do not integrate into a unified system, but instead create additional layers of complexity and duplication. As a result, instead of simplifying processes, companies find themselves in an even more fragmented and chaotic information environment.

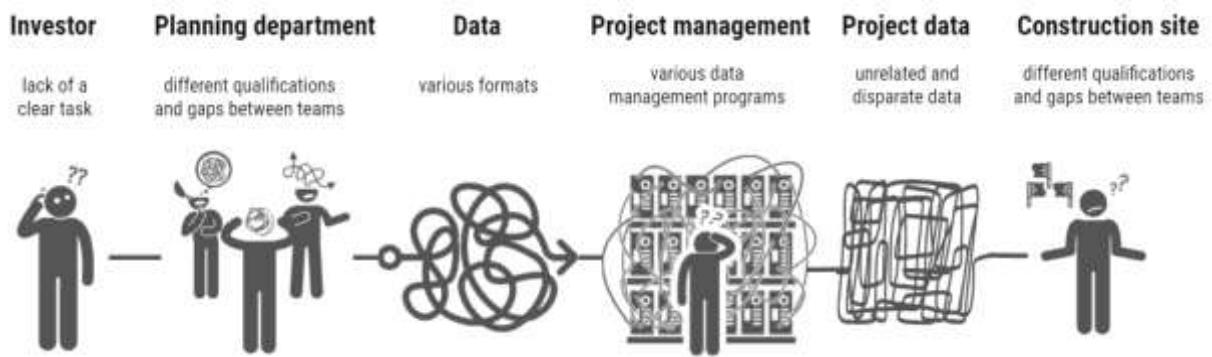


Fig. 2.1-7 The complexity of systems and the variety of data formats lead to a loss of consistency in the construction process.

All of these problems associated with managing a multitude of disparate solutions sooner or later bring company management to an important realization: it's not about the volume of data or finding the next "one-size-fits-all" tool to process it. The real reason lies in the quality of the data and how the organization creates, receives, stores and uses it.

The key to sustainable success is not in chasing new "magic" applications, but in building a data culture within the company. This means treating data as a strategic asset and making data quality, integrity and relevance a priority at all levels of the organization.

The solution to the quality vs. quantity dilemma lies in the creation of a unified data structure that eliminates duplication, eliminates inconsistencies and unifies information flows. This architecture provides a single, reliable source of data upon which to make informed, accurate and timely decisions.

Otherwise, as is still often the case, companies continue to rely on subjective opinions and intuitive assessments of HiPPO experts rather than on reliable facts. In the construction industry, where expertise traditionally plays a significant role, this is particularly noticeable.

HiPPO or the danger of opinions in decision making

Traditionally, in the construction industry, key decisions are made based on experience and subjective assessments. Without timely and reliable data, company managers have to act blindly, relying on the intuition of the highest paid employees (HiPPO - Highest Paid Person's Opinion) rather than on objective facts (Fig. 2.1-8).

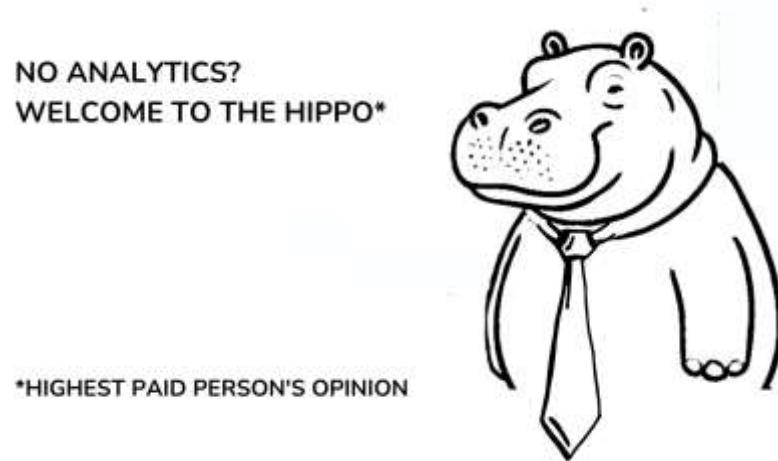


Fig. 2.1-8 In the absence of analytics business depends on the subjective opinion of experienced professionals.

This approach may be justified in a stable and slow-changing environment, but in an era of digital transformation, it becomes a serious risk. Decisions based on intuition and guesswork are prone to distortion, often based on unsupported hypotheses, and do not take into account the complex picture reflected in the data

What is passed off as intelligent debate at the decision-making level in a company is often not based on anything concrete. A company's success should not depend on the authority and salary level of experts, but on the ability to work effectively with data, identify patterns and make informed decisions.

It is important to abandon the notion that authority or experience automatically means the right decision. The data-driven approach is a game changer: data and analytics, not position and salary, are now the basis for decision-making. Big data, machine learning, and visual analytics make it possible to identify patterns and rely on facts rather than guesswork (Fig. 1.1-4).

Without data, you are just another person with an opinion [34].

– W. Edwards Deming, scholar and management consultant

Modern data management methods also ensure knowledge continuity in the company. Clearly described processes, automation and a systematic approach make it possible to transfer even key roles without losing efficiency.

However, blind trust in data can also lead to serious errors. Data itself is just a collection of numbers. Without proper analysis, context and the ability to identify patterns, they have no value and cannot drive

processes. The key to success lies not in choosing between HiPPO intuition and analytics, but in building intelligent tools that transform disparate information into manageable, informed decisions.

In a digital construction environment, it is not seniority and place in the hierarchy that become the decisive success factors, but responsiveness, decision accuracy and resource efficiency

Data are tools, not absolute truths. It should complement human thinking, not replace it. Despite the benefits of analytics, data cannot completely supplant human intuition and experience. Their role is to help make more accurate and informed decisions.

Competitive advantage will be achieved not just by meeting standards, but by being able to outperform competitors in the efficient use of resources that are the same for everyone. In the future, data skills will become as important as literacy or math skills once were. Professionals who can analyze and interpret data will be able to make more accurate decisions, displacing those who rely only on personal experience (Fig. 2.1-9).

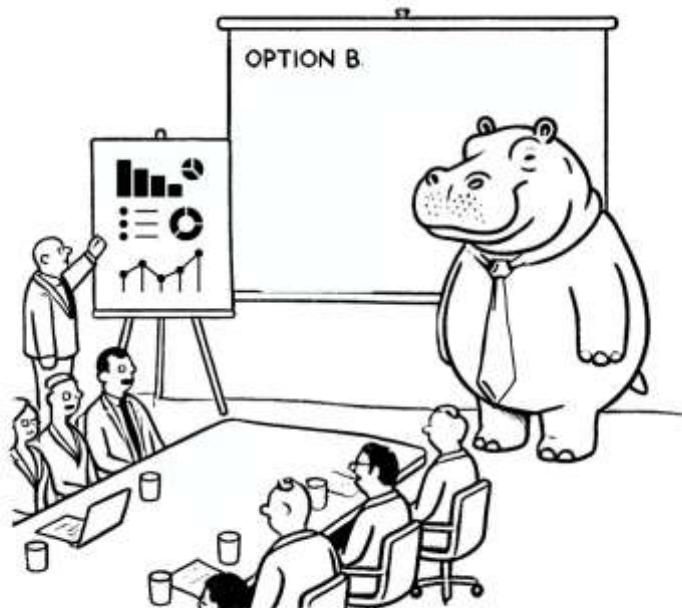


Fig. 2.1-9 Decisions should be based on objective analysis, not the opinion of the highest paid employee.

Managers, specialists and engineers will act as data analysts, studying the structure, dynamics and key indicators of projects. Human resources will become elements of the system, requiring flexible data-driven customization to maximize efficiency.

Errors using inadequate data are much smaller than those using no data [35].

— Charles Babbage, inventor of the first analytical calculating machine

The emergence of big data and the introduction of LLM (Large Language Models) have radically changed not only the ways of analysis, but also the very nature of decision making. Whereas previously the focus was on causality (why something happened - diagnostic analytics) (Fig. 1.1-4), today the ability to predict the future (predictive analytics) and, in the future, prescriptive analytics, where machine learning and AI suggest the best choice in the decision-making process, is at the forefront.

According to the new SAP™ study, "New Study Finds Nearly Half of Executives Trust Artificial Intelligence More Than Themselves" 2025 [36], 44% of senior executives would be willing to change their previous decision based on AI advice, and 38% would trust AI to make business decisions on their behalf. Meanwhile, 74% of executives said they trust AI advice more than their friends and family, and 55% work in companies where AI-derived insights replace or often bypass traditional decision-making methods - especially in organizations with annual revenues over \$5 billion. Additionally, 48% of respondents use generative AI tools on a daily basis, including 15% who use them multiple times a day.

With the development of LLM and automated data management systems, a new challenge arises: how to use information effectively without losing its value in the chaos of incompatible formats and heterogeneous sources, which is complemented by the growing complexity and dynamics of business processes.

Continuous increase in the complexity and dynamism of business processes

The construction industry today faces serious challenges in data and process management. The main challenges are siloed information systems, excessive bureaucracy and a lack of integration between digital tools. These challenges are intensifying as business processes themselves become more complex - driven by technology, changing customer requirements and evolving regulations.

The uniqueness of construction projects is due not only to their technical peculiarities, but also to differences in national standards and regulatory requirements in different countries (Fig. 4.2-10, Fig. 5.1-7). This requires a flexible, individualized approach to each project, which is difficult to implement within traditional modular control systems. Because of the complexity of processes and the large amount of data, many companies turn to vendors offering specialized solutions. But the market is overloaded - many startups offer similar products, focusing on narrow tasks. As a result, a holistic approach to data management is often lost.

Adaptation to the continuous flow of new technologies and market requirements is becoming a critical factor of competitiveness. However, existing proprietary applications and modular systems have low adaptability - any changes often require lengthy and costly revisions by developers who do not always understand the specifics of construction processes.

Companies find themselves hostage to technological lag, waiting for new updates instead of prompt implementation of innovative integrated approaches. As a result, the internal structure of construction organizations is often a complex ecosystem of interconnected hierarchical, and often closed, systems coordinated through a multi-level network of managers (Fig. 2.1-10).

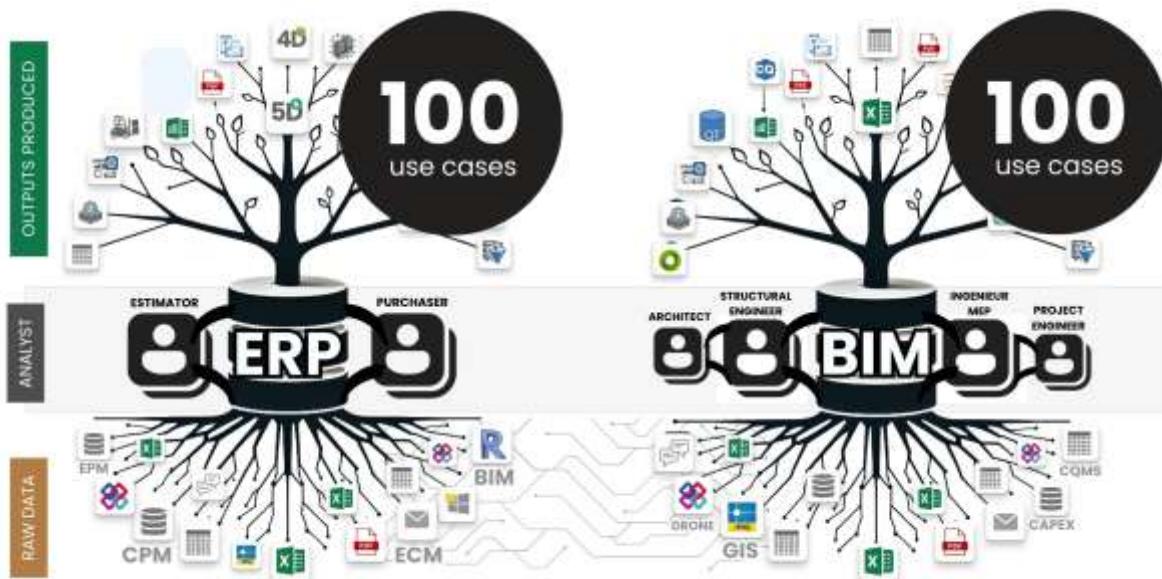


Fig. 2.1-10 Companies are made up of interconnected systems whose interconnection forms processes that require automation.

According to a study conducted by the Canadian Construction Association and KPMG Canada 2021 [37], only 25% of companies believe they are in a significant or different position compared to competitors in terms of technology adoption or digital solutions. Only 23% of respondents reported that their solutions are significantly or heavily data-driven. At the same time, the majority of survey respondents characterized their use of a number of other technologies as purely experimental or admitted to not using them at all.

This reluctance to participate in technological experiments is especially evident in large infrastructure projects, where mistakes can cost millions of dollars. Even the most advanced technologies - digital twins, predictive analytics - often meet resistance not because of their effectiveness, but because of the lack of proven reliability in real projects.

According to the World Economic Forum (WEF) report "Shaping the Future of Construction" [5], the introduction of new technologies in construction faces not only technical difficulties but also psychological barriers on the part of customers [5], the introduction of new technologies in construction faces not only technical difficulties, but also psychological barriers on the part of customers. Many clients fear that the use of advanced solutions will make their projects an experimental site and make them "guinea pigs", and unpredictable consequences may lead to additional costs and risks.

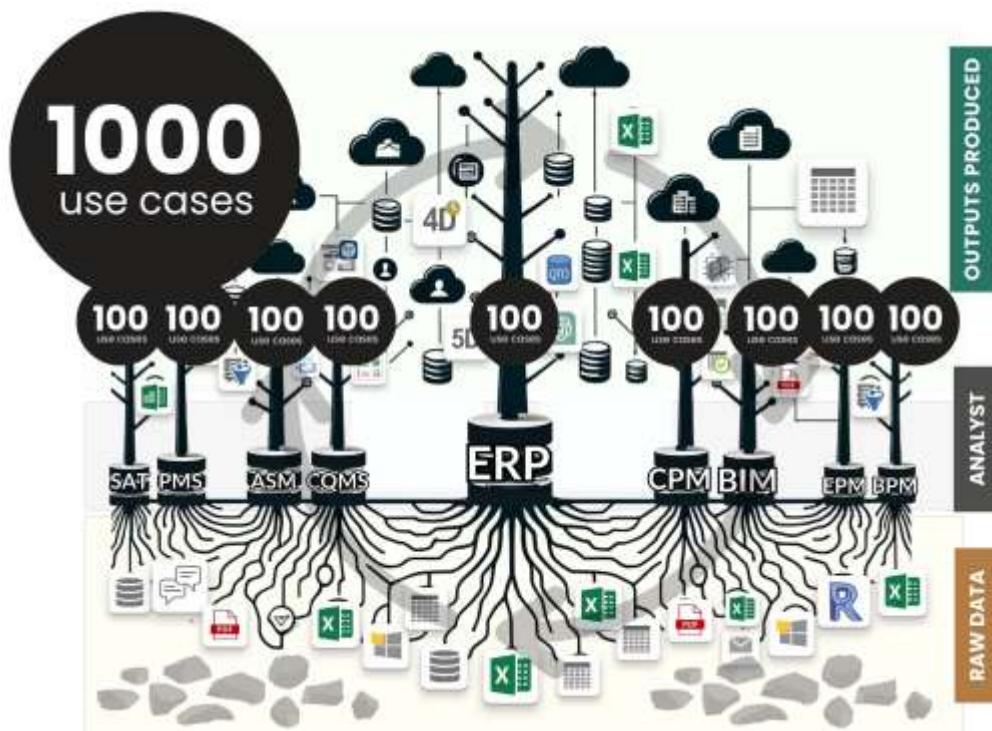


Fig. 2.1-11 For each data use case, the solutions market offers applications for process optimization and automation.

The construction industry is very diverse: different projects have different requirements, regional peculiarities, legal regulations of classifications (Fig. 4.2-10), calculation standards (Fig. 5.1-7), etc. Therefore, it is practically impossible to create a proprietary universal application or system that would perfectly fit all these requirements and project peculiarities.

Trying to cope with the growing complexity of systems and dependence on software vendors, more and more people are realizing that the key to effective data management is not only openness and standardization, but also simplification of the process architecture itself. The growing complexity and dynamism of business processes requires new approaches, where the priority shifts from accumulating data to structuring and organizing it. It is this shift that will be the next step in the development of

the construction industry, marking the end of the era of software vendor dominance and the beginning of the era of meaningful information organization.

The realization of the limitations of one-size-fits-all solutions and vulnerability to increasing complexity is leading to a shift in priorities from closed platforms and data hoarding to transparency, adaptability and structured information handling. This shift in thinking reflects broader changes in the global economy and technology, described through the lens of so-called "industrial revolutions." To understand where construction is heading and its future direction, it is necessary to consider the industry's place in the context of the Fourth and Fifth Industrial Revolutions - from automation and digitalization to personalization, open standards and the service-based data model.

The Fourth Industrial Revolution (Industry 4.0) and the Fifth Industrial Revolution (Industry 5.0) in the construction industry

Technological and economic stages are theoretical concepts used to describe and analyze the evolution of society and economy at different stages of development. At the same time, different researchers and experts may interpret them in different ways.

- **The fourth industrial revolution** (4IR or Industry 4.0) is related to information technology, automation, digitalization and globalization. One of its key elements is the creation of proprietary software solutions, i.e. specialized digital products designed for specific tasks and companies. These solutions often become an important part of the IT infrastructure, but they are poorly scalable without additional modifications.
- **The Fifth Industrial Revolution** (5IR) is now at an earlier stage of conceptualization and development than 4IR. Its core principles include increased personalization of products and services. 5IR is a movement towards more adaptable, flexible and personalized economic activity with a focus on personalization, consulting and service-oriented models. A key aspect of the fifth economic mode is the use of data for decision-making, which is virtually impossible without the use of open data and open tools (Fig. 2.1-12).

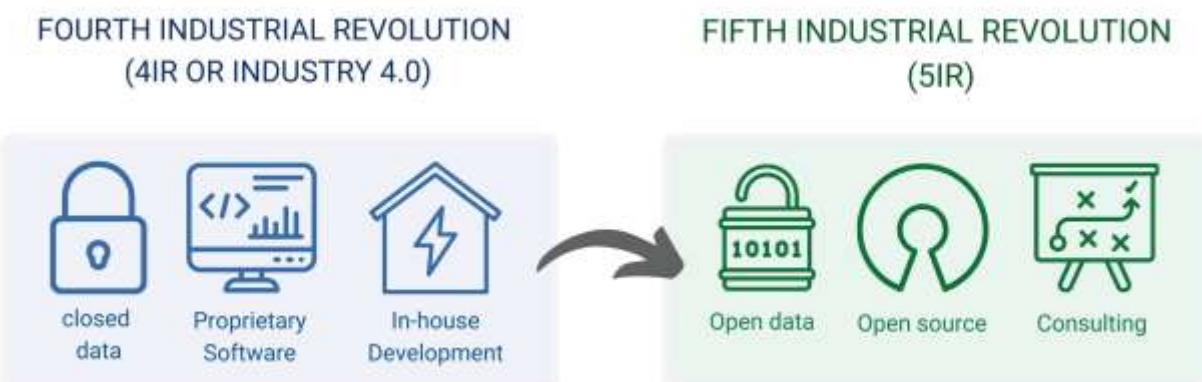


Fig. 2.1-12 The fourth pattern focuses on solutions, while the fifth pattern focuses on personalization and data.

Creating an application for companies in the construction industry for use in ten or one hundred organizations does not guarantee its successful scaling to other companies, regions or countries without significant modifications and enhancements. The likelihood of successfully scaling such solutions remains low, as each organization has unique processes, requirements and conditions that may require personalized adaptations.

It is important to understand that already today successful integration of technological solutions implies a deeply personalized approach to each process, project and company. This means that even after developing a universal framework, tool or program, it will require detailed adaptation and customization to meet the unique requirements and conditions of each specific company and project.

According to the PwC report "Decoding the Fifth Industrial Revolution" [38], about 50% of senior executives in various industries this year rely on the integration of advanced technologies and human expertise. This approach allows them to quickly adapt to changes in product design or customer requirements, creating personalized production.

Each process requires the development of a unique function or application, which, given the size of the global construction industry and the diversity of projects, leads to the existence of a huge number of business cases, representing each time a unique Pipeline logic (Fig. 2.1-13). Each such case has its own peculiarities and requires a customized approach. We will look in more detail at the variety of possible solutions to the same analytical problem in the context of different approaches in the chapter devoted to machine learning and the parsing of the Titanic dataset (Fig. 9.2-9).

Pipeline in the context of digital processes is a sequence of activities, processes and tools that enable an automated or structured flow of data and work through the different phases of the project lifecycle.

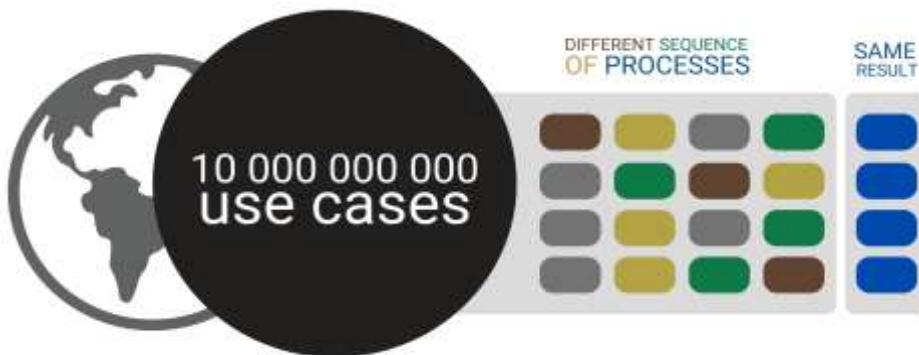


Fig. 2.1-13 The individuality and variability of business cases makes attempts to create scalable closed platforms and tools impossible.

Our lives have already changed in many ways under the influence of digital transformation, and today we can talk about the advent of a new stage in the economic development of the construction industry.

In this "new economy", competition will be organized according to different rules: the one who is able to efficiently turn public knowledge and open data into demanded products and services gains a key advantage in the conditions of the fifth industrial revolution.

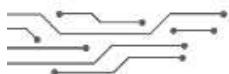
As noted by economist Kate Maskus in the book "Private Rights and Public Problems: The Global Intellectual Property Economy in the 21st Century" 2012 [39], "we live in a global knowledge economy, and the future belongs to those who know how to turn scientific discoveries into commodities".

The transition to the fifth economic mode implies a shift in focus from closed IT solutions to open standards and platforms. Companies will begin to move away from traditional software products in favor of service-oriented models, where data, rather than proprietary technologies, will become the main asset.

Harvard Business School 2024 study [40] shows the huge economic value of open source software (Open Source Software, OSS). According to the study, OSS is present in 96% of all software codes, and some commercial software consists of 99.9% OSS components. Without OSS, companies would spend 3.5 times more on software.

Building company ecosystems, following global trends, will gradually shift to the fifth economic paradigm, where data-centric analytics and consulting services will become a higher priority than isolated closed solutions with rigidly defined use cases.

The digitalization era will change the balance of power in the industry: instead of relying on vendor solutions, companies will base their competitiveness on their ability to leverage data. As a result, the construction industry will move from legacy rigid systems to flexible, adaptive ecosystems where open standards and interoperable tools will be the foundation of project management. The end of the era of application vendor dominance will create a new environment where value will not be defined by possession of closed source code and specialized connectors, but by the ability to turn data into a strategic advantage.



CHAPTER 2.2.

TURNING CHAOS INTO ORDER AND REDUCING COMPLEXITY

Redundant code and closed systems as a barrier to productivity improvement

Over the past decades, technological changes in the IT sphere have been driven primarily by software vendors. They set the course of development, determining which technologies companies should adopt and which ones should be left behind. In the era of transition from siloed solutions to centralized databases and integrated systems, vendors promoted licensed products, providing control over access and scaling. Later, with the advent of cloud technologies and Software as a Service (SaaS) models, this control evolved into a subscription model, cementing users as loyal customers of digital services.

This approach has given rise to a paradox: despite the unprecedented volumes of created program code, only a small part of it is actually used. Perhaps there are hundreds or thousands of times more code than necessary, because the same business processes are described and duplicated in dozens or hundreds of programs in different ways, even within one company. At the same time, development costs have already been paid for, and those costs are sunk. Nevertheless, the industry continues to reproduce this cycle, creating new products with minimal added value for the end user, more often under the pressure of market expectations than real needs.

According to the Defense Acquisition University (DAU) Software Development Cost Estimating Guide [41], the cost of software development can vary significantly depending on several factors, including the complexity of the system and the technology selected. Historically, development costs for 2008 have been about \$100 per line of source code (SLOC), while maintenance costs can rise to \$4,000 per SLOC.

Just one of the components of CAD applications - the geometry kernel - can have tens of millions of lines of code (Fig. 6.1-5). A similar situation is observed in ERP systems (Fig. 5.4-4), to the discussion of the complexity of which we will return in the fifth part of the book. However, a closer look reveals that much of this code does not add value, but merely acts as a "mailman" - mechanically moving data between the database, API, user interface, and other tables in the system. Despite the popular myth about the critical importance of the so-called business logic, the harsh reality is much more prosaic: modern code bases are full of outdated template blocks (legacy code), the only purpose of which is to ensure data transfer between tables and components without affecting decision making or business efficiency growth.

As a result, closed solutions that process data from various sources inevitably turn into confusing "spaghetti ecosystems". These complex, intertwined systems can only be handled by an army of managers working in a semi-routine mode. This organization of data management is not only inefficient in terms of resources, but also creates critical vulnerabilities in business processes, making the company dependent on a narrow circle of specialists who understand how this technological maze functions.

The continuous increase in the amount of code, the number of applications and the increasing complexity of concepts offered by vendors has led to a natural result - the growing complexity of the IT ecosystem in construction. This has made the practical implementation of digitalization through increasing the number of applications in the industry ineffective. Software products created without due attention to user needs often require significant resources for implementation and support, but do not bring the expected return.

According to McKinsey's study "Increasing Construction Productivity" [42], over the past two decades, global labor productivity growth in construction has averaged only 1% per year, compared to growth of 2.8% for the world economy as a whole and 3.6% for manufacturing. In the United States, construction labor productivity per worker has halved since the 1960s [43].

Increasing system complexity, isolation, and closed data have impaired communication among professionals, making the construction industry one of the least efficient (Fig. 2.2-1). to \$22 trillion by 2040, which will require significant efficiency gains.

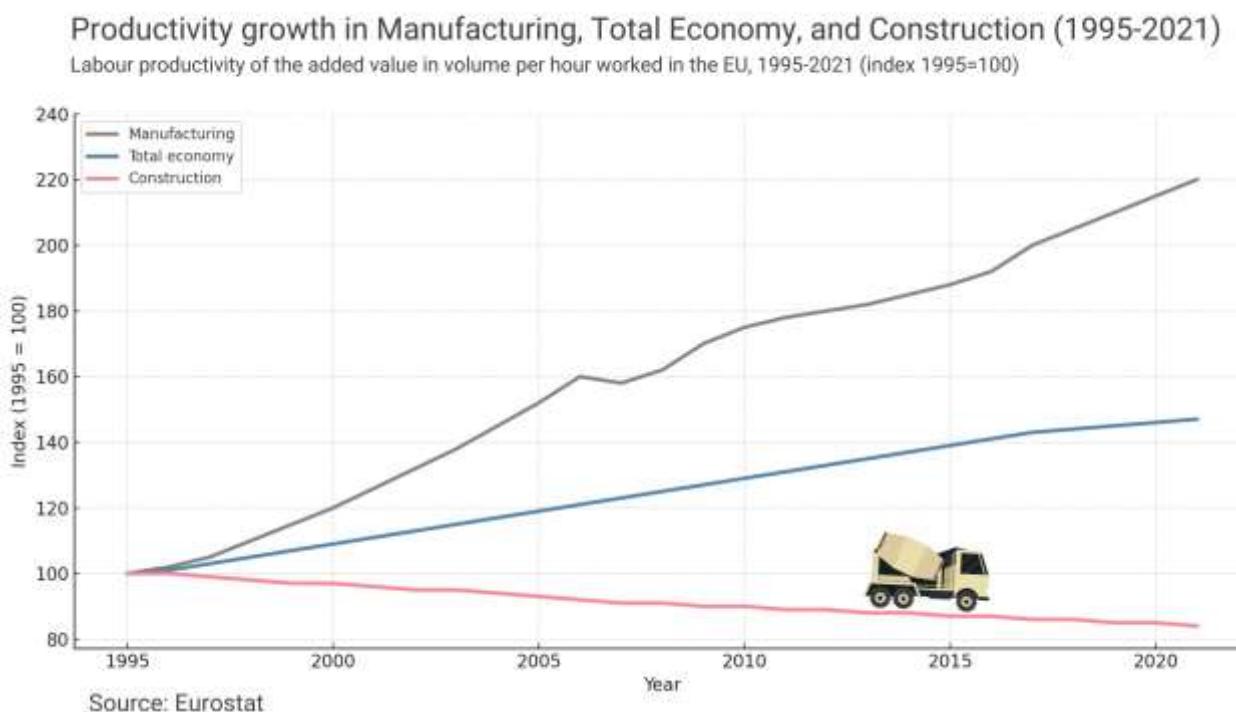


Fig. 2.2-1 Closed and complex data and as a consequence poor communication between specialists led the construction industry to one of the least efficient sectors of the economy (based on [44], [45]).

As emphasized in the McKinsey (2024) study "Ensuring construction productivity is no longer optional", with increasing resource scarcity and the industry's drive to double its growth rate, construction can no longer afford to remain at current productivity levels [44]. Global construction costs are projected to rise from \$13 trillion in 2023 to much higher levels by the end of the decade, making the issue of efficiency not just relevant, but critical.

One of the key ways to improve efficiency will be the inevitable unification and simplification of application structures and data ecosystem architectures. This approach to rationalization will eliminate excessive layers of abstraction and unnecessary complexity that have accumulated over the years in enterprise systems.

From silos to a unified data warehouse

The more data an organization accumulates, the harder it becomes to extract real value from it. Because of the fragmented nature of storing information in isolated silos, modern companies' business processes are like builders trying to construct a skyscraper out of materials stored in thousands of different warehouses. The excess of information not only makes it difficult to access legally relevant information, but also slows down decision-making: every step has to be repeatedly checked and confirmed.

Each task or process is hard-wired to a separate table or database, and data exchange between systems requires complex integrations. Errors and inconsistencies in one system can cause chain failures in others. Incorrect values, late updates, and duplicate information force employees to spend significant time manually reconciling and reconciling data. As a result, the organization spends more time dealing with the consequences of fragmentation than developing and optimizing processes

This problem is universal: some companies continue to struggle with chaos, while others find a solution in integration - moving information flows into a centralized storage system. Think of it as one big table where you can store any entities related to tasks, projects and objects. Instead of dozens of disparate tables and formats, a single cohesive repository appears (Fig. 2.2-2), allowing:

- minimize data loss;
- eliminate the need for constant harmonization of information;
- improve data availability and quality;
- simplify analytical processing and machine learning

Bringing data to a single standard means that regardless of the source, information is converted into a unified and machine-readable format. Such organization of data allows to check its integrity, analyze it in real time and promptly use it for making managerial decisions.

The concept of integrated storage systems and their application in analytics and machine learning will be discussed in more detail in the chapter "Big Data Storage and Machine Learning". The topics of data modeling and structuring will be covered in detail in the chapters "Transforming data into a structured

form" and "How standards change the game: from random files to an elaborate data model".

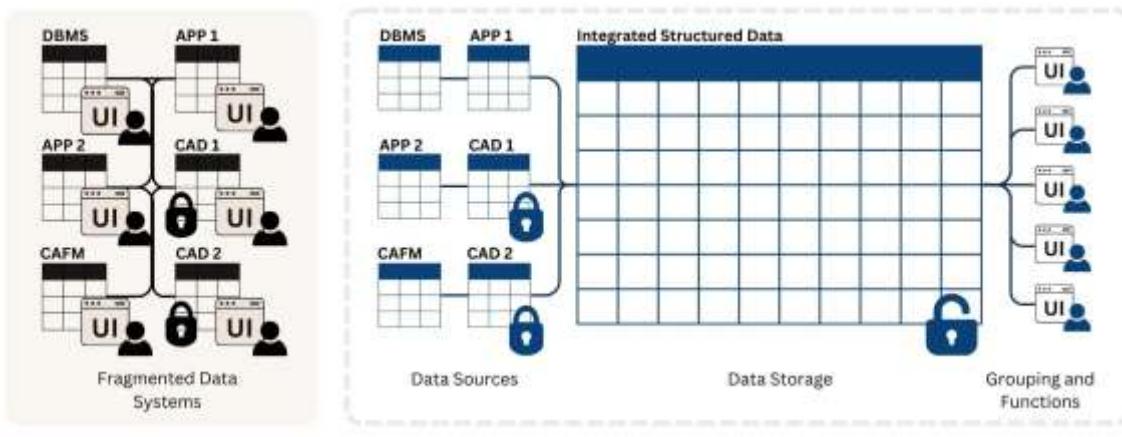


Fig. 2.2-2 Data integration eliminates silos, improves information availability, and optimizes business processes.

Once the data has been structured and merged, the next logical step is to validate it. With a single integrated repository, this process is greatly simplified: no more multiple inconsistent schemas, duplicate structures and complex relationships between tables. All information is aligned to a single data model, eliminating internal inconsistencies and speeding up the validation process. Validating and ensuring data quality are cornerstone aspects of all business processes, and we will explore them in more detail in the relevant chapters of the book.

At the final stage, the data are grouped, filtered and analyzed. Various functions are applied to them: aggregation (addition, multiplication), calculations between tables, columns or rows (Fig. 2.2-4). Working with data becomes a sequence of steps: collection, structuring, validation, transformation, analytical processing and offloading to final applications where the information is used to solve practical problems. We will talk more about building such scenarios, automating steps and building processing flows in the chapters on ETL-processes and data pipeline approach.

Thus, digital transformation is not just about simplifying the handling of information. It is about eliminating excessive complexity in data management, moving from chaos to predictability, from multiple systems to a manageable process. The lower the complexity of the architecture, the less code is required to support it. And in the future, code as such may disappear altogether, giving way to intelligent agents that independently analyze, systematize and transform data.

Integrated storage systems enable the transition to AI agents

The less complex the data and systems are, the less code you need to write and maintain. And the easiest way to save development is to get rid of code altogether, replacing it with data. When application code development moves from code to data models, there is inevitably a shift towards a data-centric (data-driven) approach, because there is a completely different way of thinking behind these concepts.

When one chooses to work with data at the center, one begins to see its role differently. Data is no longer just "raw material" for applications - it is now the foundation around which architecture, logic and interaction are built.

The traditional approach to data management usually starts at the application level and in construction resembles a cumbersome bureaucratic system: multi-level approvals, manual checks, endless versions of documents through the relevant software products. With the development of digital technologies, more and more companies will be forced to move to the principle of minimalism - to store and use only what is really necessary and will be used.

The logic of minimization has been taken up by vendors. To simplify data storage and processing, user work is being moved from offline applications and tools to cloud services and so-called SaaS solutions.

The SaaS concept (Software as a Service, or "software as a service") is one of the key trends in modern IT infrastructures, allowing users to access applications via the Internet without having to install and maintain software on their own computers.

On the one hand SaaS has facilitated scaling, version control and reduced support and maintenance costs, but on the other hand, in addition to dependence on the logic of a particular application, it has also made the user completely dependent on the provider's cloud infrastructure. If a service goes down, access to data and business processes can be temporarily or even permanently blocked. In addition, all user data when working with SaaS applications is stored on the provider's servers, which creates security and regulatory compliance risks. Changes in tariffs or terms of use may also result in increased costs or the need for urgent migration.

The development of AI, LLM -agents and data-centric approach has questioned the future of applications in their traditional form and SaaS execution. Whereas applications and services were previously required to manage business logic and process data, with the advent of AI agents, these functions may shift to intelligent systems that work directly with data.

This is why hybrid architectures are increasingly being discussed in IT departments and at the management level, where AI -agents and on-premises solutions complement cloud services, reducing dependency on SaaS -platforms.

The approach we take recognizes that traditional business applications or SaaS -applications may change dramatically in the agent era. These applications are essentially CRUD [create, read, update and delete] databases with business logic. But in the future, this logic will be taken over by AI agents [46].

– Satya Nadella, CEO of Microsoft, 2024.

A data-centric approach and the use of AI/LLM agents can reduce redundant processes and thus reduce the workload of employees. When data is organized properly, it becomes easier to analyze, visualize and apply it to decision-making. Instead of endless reports and checks, specialists get access to up-to-date information in a few clicks or with the help of LLM agents automatically in the form of ready documents and dashboards.

We will be assisted in data manipulation by artificial intelligence tools (AI) and LLM chats. In recent years, there has been a trend away from traditional CRUD operations (create, read, update, delete) towards the use of large language models (LLMs) for data management. LLMs are capable of interpreting natural language and automatically generating appropriate database queries, which simplifies interaction with data management systems (Fig. 2.2-3).

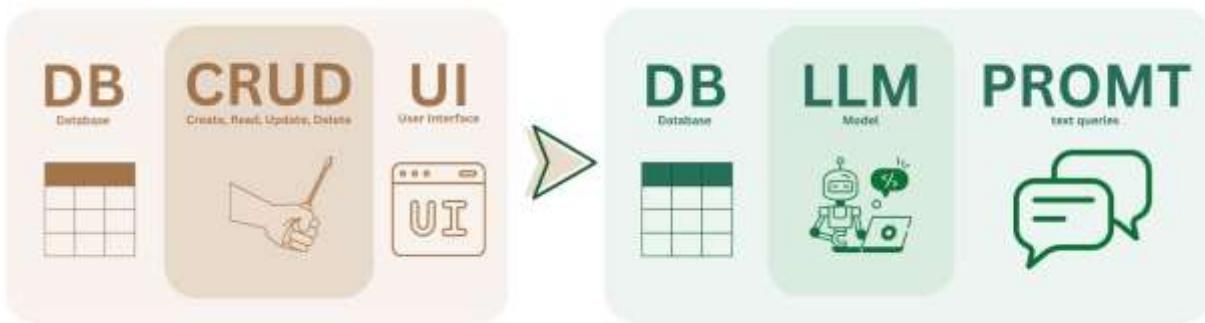


Fig. 2.2-3 AI will replace and integrate storage and database solutions, gradually displacing traditional applications and CRUD -operations.

In the next 3-6 months, AI will be writing 90% of the code, and in 12 months, almost all of the code could be generated by AI [47].

— Dario Amodei, CEO of LLM Anthropic, March 2025.

Despite the rapid development of AI development tools (e.g., GitHub Copilot), in 2025 developers still play a key role in this process. AI agents are becoming increasingly useful assistants: they automatically interpret user queries, generate SQL and Pandas queries (more on this in the following chapters), or write code to analyze data. Thus, artificial intelligence is gradually replacing traditional application user interfaces.

The proliferation of artificial intelligence models, such as language models, will drive the development of hybrid architectures. Instead of completely abandoning cloud solutions and SaaS products, we may see the integration of cloud services with local data management systems. For example, federated learning enables powerful AI models without having to move sensitive data to the cloud. In this way,

companies can maintain control of their data while gaining access to advanced technologies.

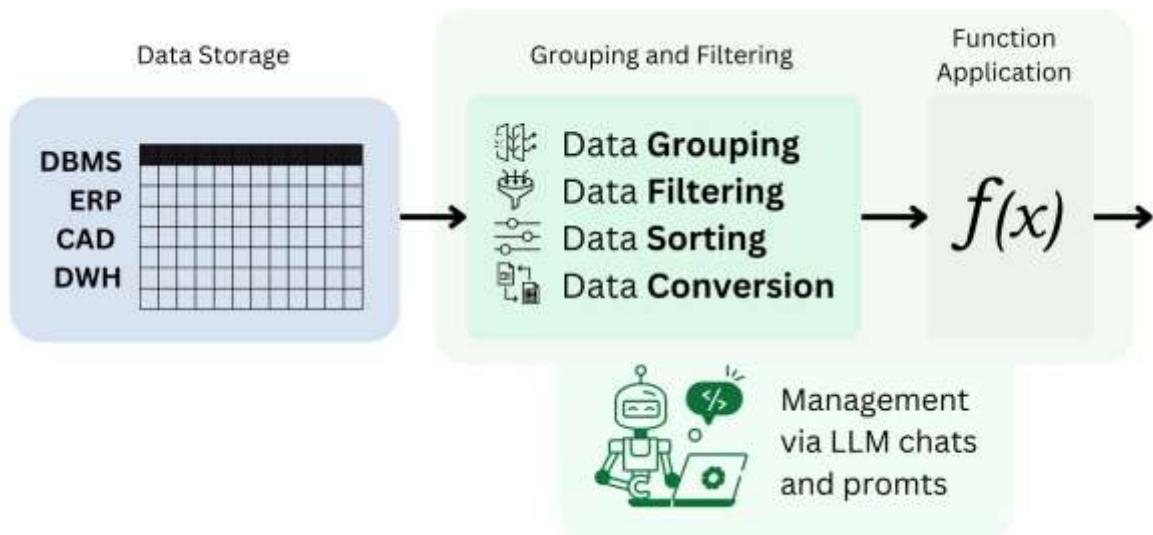


Fig. 2.2-4 The basic operations of grouping, filtering, and sorting followed by the application of functions will be handled by LLM chats.

The future of the construction industry will be based on a combination of on-premises solutions, cloud power and intelligent models working together to create efficient and secure data management systems. LLM will enable users without deep technical knowledge to interact with databases and data warehouses by formulating their queries in natural language. We will talk more about LLM and AI agents and how they work in the chapter "LLM agents and structured data formats".

Properly organized data and simple, easy-to-use LLM-enabled analytics tools will not only make it easier to work with information, but will also help minimize errors, increase efficiency and automate processes.

From data collection to decision-making: the road to automation

In later parts of the book, we will look in detail at how specialists interact with each other and how data becomes the basis for decision making, automation, and operational efficiency. Fig. 2.2-5 provides an example diagram showing the sequence of data processing steps in a data-centric approach. This diagram illustrates the Continuous Improvement Pipeline), parts of which will be discussed in detail later in the book.

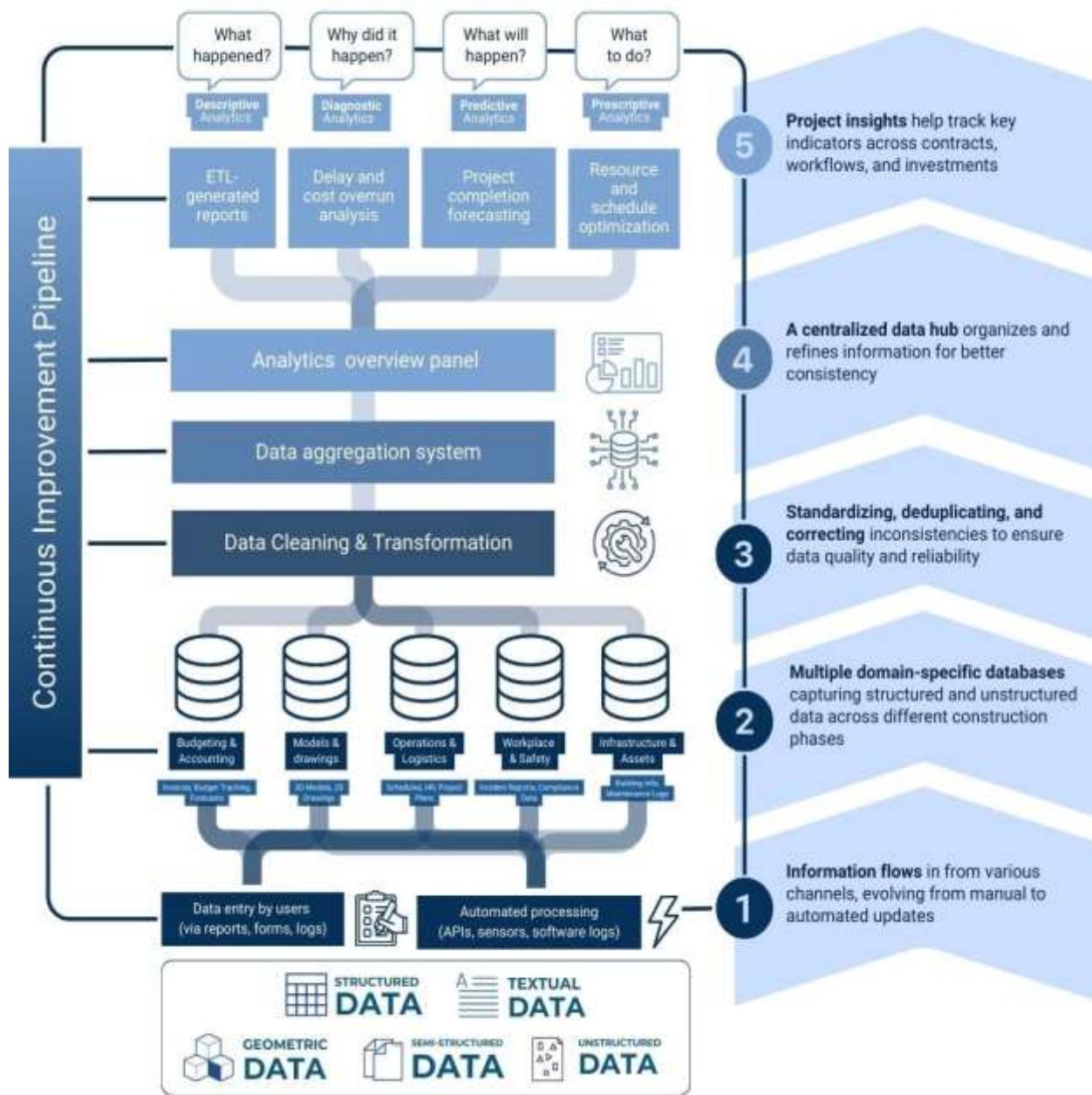


Fig. 2.2-5 An example of a continuous data improvement pipeline: the flow of data processing and analysis in construction projects.

The system describing the business processes of a medium-sized company is built on a multi-level principle. It includes: data collection, cleaning, aggregation, analytical processing, and decision-making based on the results. We will study all these stages later in the book - both in a theoretical context and through practical examples:

- At the first level, **data entry** takes place (Fig. 3.1-1). Information is received both manually (through reports, forms, logs) and in an automated form (from API, sensors, software systems). Data can be of different structure: geometric, text, unstructured. At this stage there is a

need for standardization, structuring and unification of information flows.

- The next level is **data processing and transformation**. It includes the processes of cleaning, removing duplicates, correcting errors, and preparing information for further analysis (Fig. 4.2-5). This stage is critical because the quality of analytics directly depends on the cleanliness and accuracy of the data.
- The data is then **entered into specialized tables, dataframes or databases** divided by functional areas: budgeting and accounting, models and drawings, logistics, security and infrastructure. This division allows for easy access and cross-analysis of information.
- The data is then **aggregated and displayed in an analytics dashboard** (showcase). Here, the methods of descriptive, diagnostic, predictive and prescriptive analytics are applied. This helps answer key questions (Fig. 1.1-4): what happened, why it happened, what will happen in the future, and what actions need to be taken. For example, the system can identify delays, predict project completion, or optimize resources.
- Finally, the last level generates **analytical conclusions and key indicators** that help to monitor contract fulfillment, manage investments and improve business processes (Fig. 7.4-2). This information becomes the basis for decision-making and the company's development strategy.

Similarly, data goes from being collected to being used in strategic management. In the following parts of the book, we will look at each stage in detail, focusing on data types, data processing techniques, analytics tools and real-life cases of how these approaches are used in the construction industry.

Next steps: turning chaos into a manageable system

In this part, we explored the challenges of information silos and looked at the impact of excessive system complexity on business performance, analyzing the transition from the fourth industrial revolution to the fifth, where data rather than applications are central. We saw how siloed information systems are creating barriers to knowledge sharing, and the continuing complexity of the IT landscape is reducing productivity and inhibiting innovation in the construction industry.

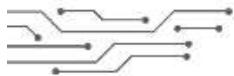
To summarize this part, it is worth highlighting the main practical steps that will help you apply the discussed approaches in your daily tasks:

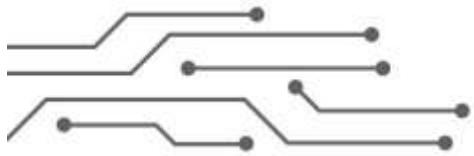
- Visualize your information landscape
 - Create a visual map of the data sources (Miro, Figma, Canva) you work with on a regular basis
 - Add the systems and applications you use in your work to this map
 - Identify potentially duplicate functionality and redundant solutions
 - Identify critical points where data loss or corruption can occur during transmission between systems
- Implement personalized data management practices
 - Shift focus from applications to data as a key asset in processes

- Document data sources and processing methodology to ensure transparency
 - Develop mechanisms to assess and improve data quality
 - Strive to ensure that data is entered once and used repeatedly - this is the basis of efficient process organization
- Promote a data-centric (data-driven) approach in your team
- Suggest the use of standardized and uniform formats for peer-to-peer data exchange
 - Regularly raise issues related to data quality and availability in team meetings
 - Get to know Open Source alternatives to the tools you use to solve your issues

Start small - pick one specific process or data set that is critical to your work and apply a data-centric approach to it, shifting the focus from tools to data. By succeeding in a single pilot, you'll gain not only hands-on experience, but also a clear demonstration of the benefits of the new methodology to your team. In completing most of these steps, if you have questions, you can seek clarification and assistance from any up-to-date LLM.

In the following parts of the book, we will move on to a more detailed look at data structuring and unification techniques and explore practical approaches to integrating heterogeneous information. Particular attention will be paid to the transition from disparate silos to unified data ecosystems, which play a key role in the digital transformation of the construction industry.





III PART

DATA FRAMEWORK IN CONSTRUCTION BUSINESS PROCESSES

In the third part, a comprehensive understanding of the typology of data in construction and methods of their effective organization is formed. The characteristics and specifics of working with structured, unstructured, semi-structured, textual and geometric data in the context of construction projects are analyzed. Modern storage formats and protocols of information exchange between different systems used in the industry are considered. Practical tools and techniques for transforming multi-format data into a single structured environment are described, including how to integrate CAD (BIM) data. Approaches to ensuring the quality of data through standardization and validation critical to the accuracy of construction calculations are proposed. Practical aspects of using modern technologies (Python Pandas, LLM -models) with code examples to solve typical problems in the construction industry are analyzed in detail. The value of creating a competence center (CoE) as an organizational structure for coordination and standardization of information management approaches is substantiated.

CHAPTER 3.1.

DATA TYPES IN CONSTRUCTION

The most important data types in the construction industry

In the modern construction industry, the systems, applications and data warehouses of companies are actively filled with information and data of various types and formats (Fig. 3.1-1). Let's take a closer look at the main types of data that form the information landscape of a modern company operating in the construction industry:

- **Structured data:** this data has a clear organizational structure, e.g. Excel Spreadsheets and Relational Databases.
- **Unstructured data:** this is information that is not organized according to strict rules. Examples of such data are text, video, photos, and audio recordings.
- **Loosely structured data:** these data occupy an intermediate position between structured and unstructured data. They contain elements of structure, but this structure is not always clear or often described through different schemas. Examples of semi-structured data in construction are: technical specifications, project documentation or progress reports.
- **Textual data:** includes anything derived from oral and written communications, such as emails, transcripts of meetings and appointments.
- **Geometric data:** this data comes from CAD programs in which experts create geometric data of project elements for visualization, confirmation of volume values or collision checking.

It is important to note that geometric and text (alphanumeric) data are not a separate category, but can be present in all three types of data. Geometric data, for example, can be part of both structured data (parametric CAD formats) and unstructured data (scanned drawings). Text data can similarly be both organized in databases (structured data) and exist as documents without a clear structure.

Each type of data in a construction company is a unique element in the mosaic of the company's information assets. From unstructured data, such as images from construction sites and audio recordings of meetings, to structured records, including tables and databases, each element plays an important role in shaping the company's information landscape.

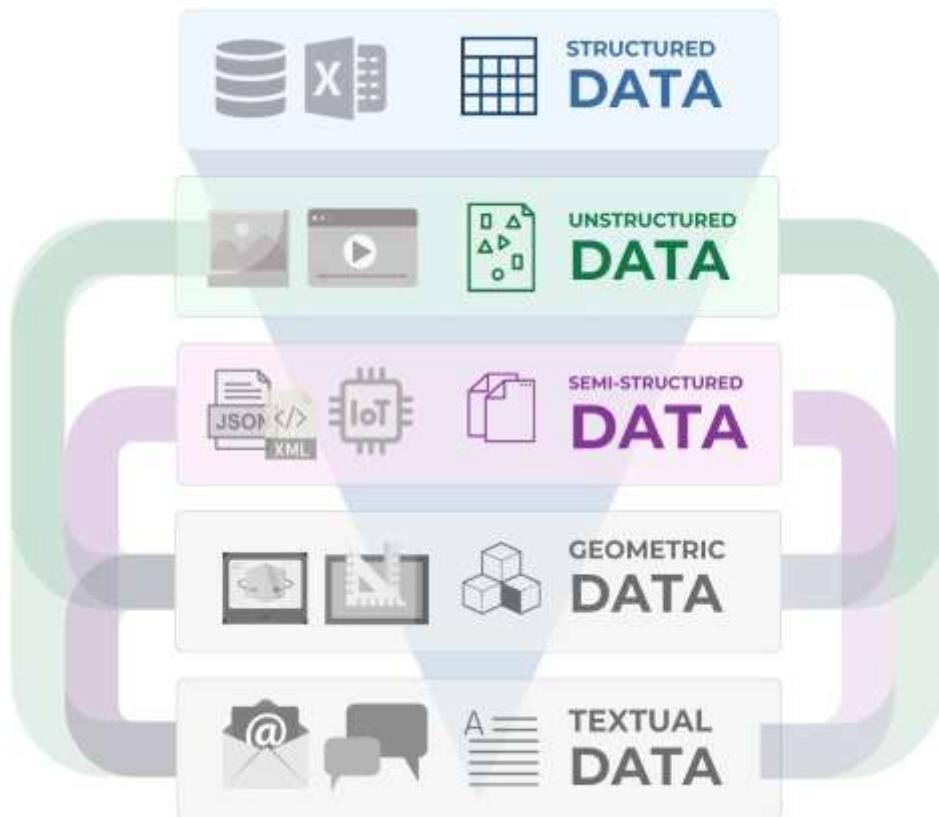


Fig. 3.1-1 Engineers and data managers must learn to work with all types of data used in the construction industry.

Here is a sample list of just some of the systems and associated data types (Fig. 3.1-2) used in construction:

- **ERP** (Enterprise Resource Planning) - handles generally structured data to help manage enterprise resources and integrate various business processes.
- **CAD** (Computer-Aided Design) combined with **BIM** (Building Information Modeling) - uses geometric and semi-structured data to design and model construction projects, ensuring accuracy and consistency of information during the design phase.
- **GIS** (Geographic Information Systems) - works with geometric and structured data to create and analyze map data and spatial relationships.
- **RFID** (Radio-Frequency Identification) - uses semi-structured data to efficiently track materials and equipment on a construction site using radio frequency identification.
- **ECM** (Engineering Content Management) is a system for managing engineering data and documentation, including semi-structured and unstructured data such as technical drawings and design documents.

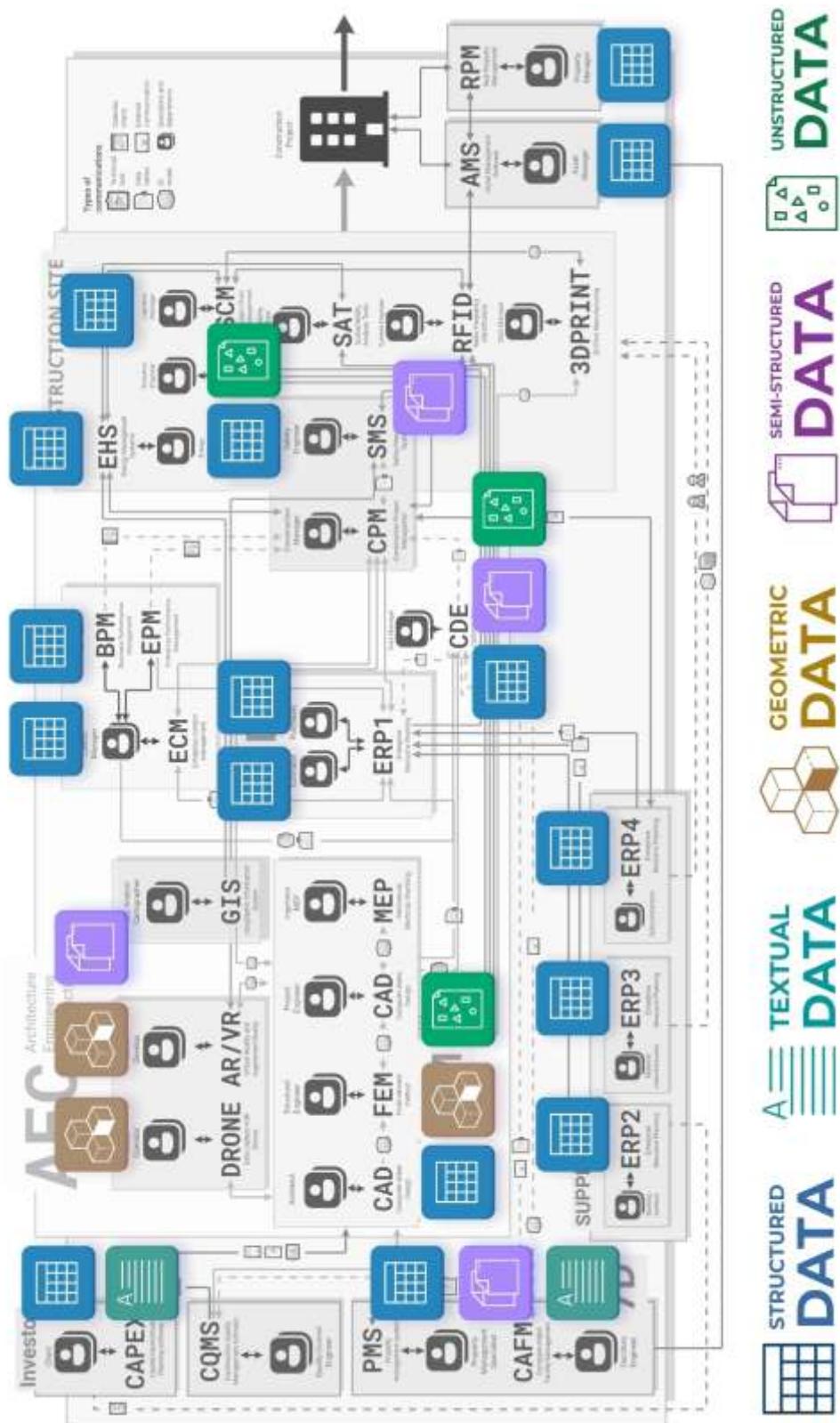


Fig. 3.1-2 Different formats and data populate different systems, requiring translation into a form suitable for complex integration.

These and many of the company's other systems manage a wide range of data, from structured tabular data to complex geometric models, providing integrated collaboration in the design, planning and construction management processes.

In the example of a simplified dialog (Fig. 3.1-3), different types of data are exchanged between construction project specialists:

- ⦿ **Architect:** "Taking into account the client's wishes, I have added a rooftop seating area. Please take a look at the new design" (geometric data - model).
- ⦿ **Structural engineer:** "The project has been received. I am calculating the load-bearing capacity of the roof for the new recreation area" (structured and semi-structured data - calculation tables).
- ⦿ **Purchasing manager:** "Need specifications and quantities of materials for the recreation area to organize the purchase" (textual and semi-structured data - lists and specifications).
- ⦿ **Health and Safety Engineer:** "Received data on new area. I am assessing the risks and updating the safety plan" (semi-structured data - documents and plans).
- ⦿ **Specialist in BIM -modeling:** "Making changes to the overall project model to adjust the working documentation" (geometric data and semi-structured data).
- ⦿ **Project Manager:** "I am incorporating the new rest area into the work schedule. I am updating schedules and resources in the project management system" (structured and semi-structured data - schedules and plans).
- ⦿ **Facility Maintenance Specialist (FM):** "I prepare data for future maintenance of the recreation area and input it into the property management system" (structured and semi-structured data - instructions and maintenance plans).

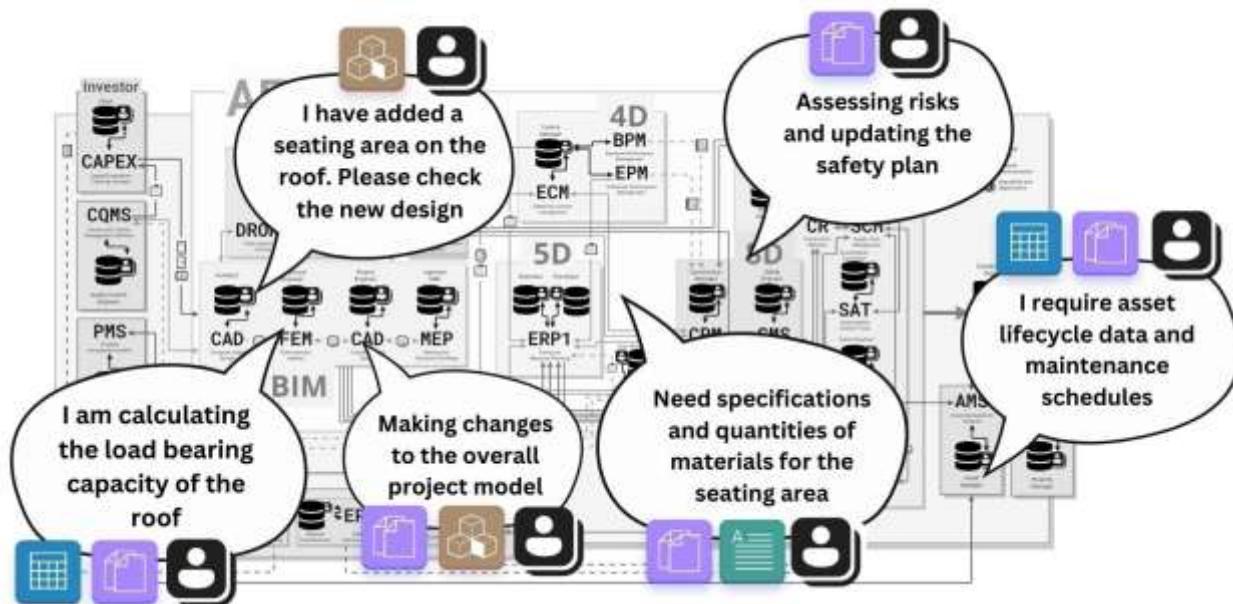


Fig. 3.1-3 Communication between specialists occurs at both text and data levels.

Each professional works with different types of data to ensure effective team collaboration and project success. Understanding the differences between structured, semi-structured and unstructured data

allows you to recognize the unique role each type plays in digital business processes. It is important to not only know that there are different forms of data, but also to understand how, where and why they are used.

Not so long ago, the idea of combining such diverse data seemed ambitious but difficult to realize. Today, it is already part of everyday practice. Integration of data of different schemas and structures has become an integral part of modern information systems architecture.

In the following chapters we will look in detail at the key standards and approaches that enable structured, semi-structured and unstructured data to be combined into a single coherent view. Special attention will be paid to structured data and relational databases - as the main mechanisms for storing, processing and analyzing information in the construction industry.

Structured Data

In the construction industry, information comes from many sources - drawings, specifications, schedules and reports. To effectively manage this flow of information, it needs to be structured. Structured data allows you to organize information in a convenient, readable and accessible form.

According to JB Knowledge's 5th Annual Construction Technology Report [17], 67% of construction project management professionals track and evaluate job performance manually or using spreadsheets.

Some of the most common structured data formats are XLSX and CSV. They are widely used for storing, processing and analyzing information in spreadsheets. In such spreadsheets, data is presented in the form of rows and columns, which makes them easy to read, edit, and analyze.

XLSX, a format created by Microsoft, is based on the use of XML -structures and is archived using the ZIP algorithm. The main features of the format:

- Support for complex formulas, charts and macros.
- Ability to store data in different sheets as well as format information.
- Optimized for Microsoft Excel, but compatible with other office suites as well.

CSV format is a plain text file in which values are separated by commas, semicolons, or other delimiter characters. Key benefits:

- Universal compatibility with various programs and operating systems.
- Easy import/export to databases and analytical systems.
- Easy processing even in text editors.

However, CSV does not support formulas and formatting, so its main application is data exchange between systems and mass updates of information. Due to its versatility and platform independence,

CSV has become a popular tool for data transfer in heterogeneous IT environments.

The two formats XLSX and CSV act as a link between different systems that work with structured data (Fig. 3.1-4). They are particularly useful in tasks where readability, manual editing and basic compatibility are important

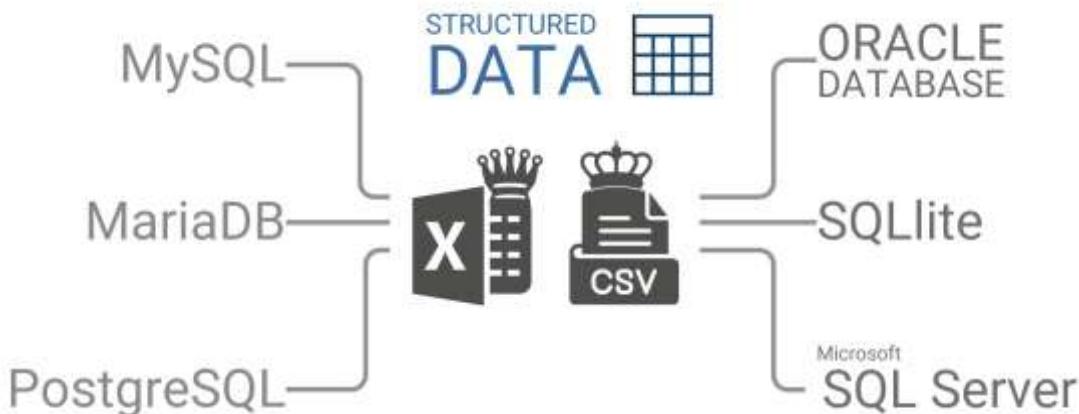


Fig. 3.1-4 The XLSX and CSV formats are the link between different systems that work with structured data.

Platform independence makes CSV the most popular format for data transfer in heterogeneous IT environments and systems.

However, XLSX and CSV are not designed for high-performance computing or long-term storage of large amounts of data. More modern structured formats such as Apache Parquet, Apache ORC, Feather, HDF5 are used for such purposes. These formats will be discussed in more detail in the chapter "Storing Big Data: Analyzing Popular Formats and Their Effectiveness" in Part 9 of this book.

In practice, Excel with XLSX format is more often used for small tasks and automation of routine processes. More complex scenarios require the use of data management systems such as ERP, PMIS CAFM, CPM, SCM and others (Fig. 3.2-1). These systems store structured data on which the organization and management of the company's information flows are based.

Modern data management information systems used in the construction industry rely on structured data organized as tables. For reliable, scalable and holistic management of large volumes of information, application and system developers are turning to relational database management systems (RDBMS).

Relational databases RDBMS and SQL query language

Relational databases (RDBMS) are data warehousing systems that organize information into tables with defined relationships between them to efficiently store, process, and analyze data.

Data organized in databases (RDBMS) are not just digital information; they are the basis for transactions and interactions between different systems.

Here are a few of the most common relational database management systems (RDBMS) (Fig. 3.1-5):

- **MySQL** (Open Source) is one of the most popular RDBMS, which is a part of LAMP stack (Linux, Apache, MySQL, PHP /Perl/Python). It is widely used in web development due to its simplicity and high performance.
- **PostgreSQL** (Open Source) is a powerful object-relational system known for its reliability and advanced features. It is suitable for complex enterprise solutions.
- **Microsoft SQL Server** is a commercial system from Microsoft that is widely used in corporate environments due to its integration with other company products and high level of security.
- **Oracle Database** is one of the most powerful and reliable RDBMS used in large enterprises and mission-critical applications.
- **IBM DB2** - targeted at large corporations, providing high performance and fault tolerance.
- **SQLite** (Open Source) is a lightweight embedded database, ideal for mobile applications and standalone systems such as CAD design programs (BIM).

Popular database management systems in the construction business - MySQL, PostgreSQL, Microsoft SQL Server, Oracle® Database, IBM® DB2 and SQLite - work with structured data. All these DBMSs are powerful and flexible solutions for managing a wide range of business processes and applications, from small Web sites to large-scale enterprise systems (Fig. 3.2-1).

According to Statista [48], relational database management systems (RDBMS) account for about 72% of the total DBMS in use in 2022.

	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

Fig. 3.1-5 Popularity of using structured databases (marked in blue) in DBMS ranking (based on [49]).

It is quite easy to install open source databases - even without extensive technical knowledge. Open source systems, such as PostgreSQL, MySQL or SQLite, are available for free and work on most operating systems: Windows, macOS and Linux. All you need is to go to the official website of the project, download the installer and follow the instructions. In most cases, installation takes no more than 10-15 minutes. We will model and create one of such databases in the fourth part of the book (Fig. 4.3-8).

If your company uses cloud services (for example, Amazon Web Services, Google Cloud or Microsoft Azure), you can deploy the database in a couple of clicks - the platform will offer you ready-made templates for installation. Due to the openness of the code, such databases are easy to customize for your tasks, and a huge community of users will always help you find a solution to any problem.

RDBMS remain the foundation for a multitude of business applications and analytics platforms (Fig. 3.1-6) that enable companies to efficiently store, process, and analyze data - and therefore make informed and timely decisions.

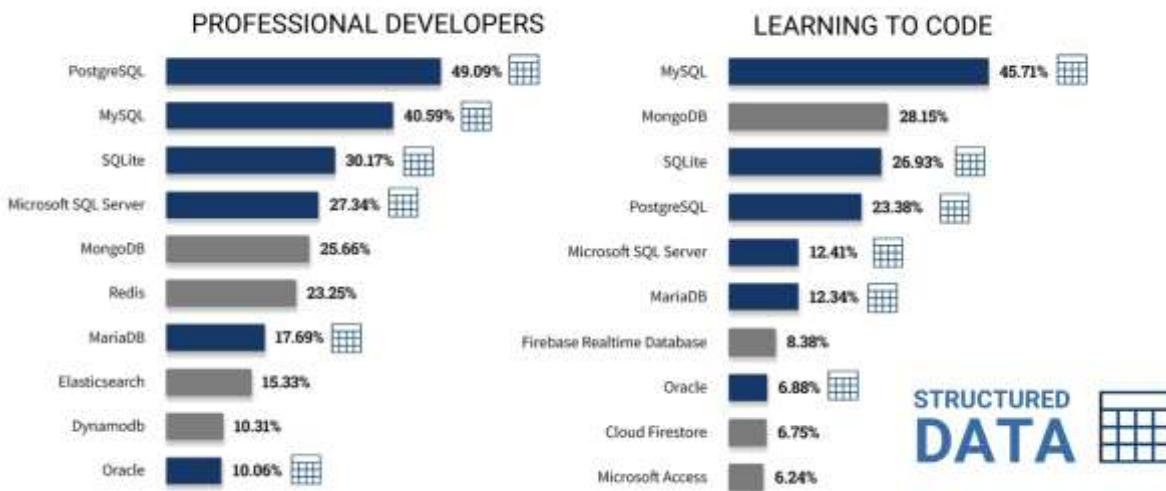


Fig. 3.1-6 A survey of developers at StackOverFlow (the largest IT forum) about which databases they used last year and which they want to use next year (RDBMSs are highlighted in blue) (based on [50]).

RDBMS provide reliability, data consistency, transaction support and use a powerful query language - SQL (Structured Query Language), which is often used in analytics and allows you to easily obtain, modify and analyze information stored in databases. SQL is the main tool for working with data in relational systems.

SQL-queries in databases and new trends

The main advantage of the SQL language, often used in relational databases, over other types of information management (for example, with the help of classic Excel spreadsheets) is the support of very large volumes of databases at high speed of query processing.

Structured Query Language (SQL) is a specialized programming language designed for storing, processing and analyzing information in relational databases. SQL is used to create, manage and access data, allowing you to efficiently find, filter, combine and aggregate information. It serves as a key tool for accessing data, providing a convenient and formalized way to interact with information stores.

The evolution of SEQUEL-SQL systems goes through significant products and companies such as Oracle, IBM DB2, Microsoft SQL Server, SAP, PostgreSQL and MySQL, and culminates in the emergence of SQLite and MariaDB [51]. SQL provides spreadsheet capabilities not found in Excel, making data manipulation more scalable, secure, and easy to automate:

- **Creating and managing data structures (DDL):** In SQL you can create, modify, and delete tables in a database, establish links between them, and define data storage structures. Excel,

on the other hand, works with fixed sheets and cells, without clearly defined relationships between sheets and data sets.

- **Data manipulation (DML):** SQL allows you to massively add, modify, delete and retrieve data at high speed, performing complex queries with filtering, sorting and table joins (Fig. 3.1-7). In Excel, processing large amounts of information requires manual actions or special macros, which slows down the process and increases the probability of errors.
- **Access control (DCL):** SQL allows you to differentiate access rights to data for different users, limiting the ability to edit or view information. In Excel, on the other hand, access is either shared (when transferring a file) or requires complex settings with rights sharing via cloud services.

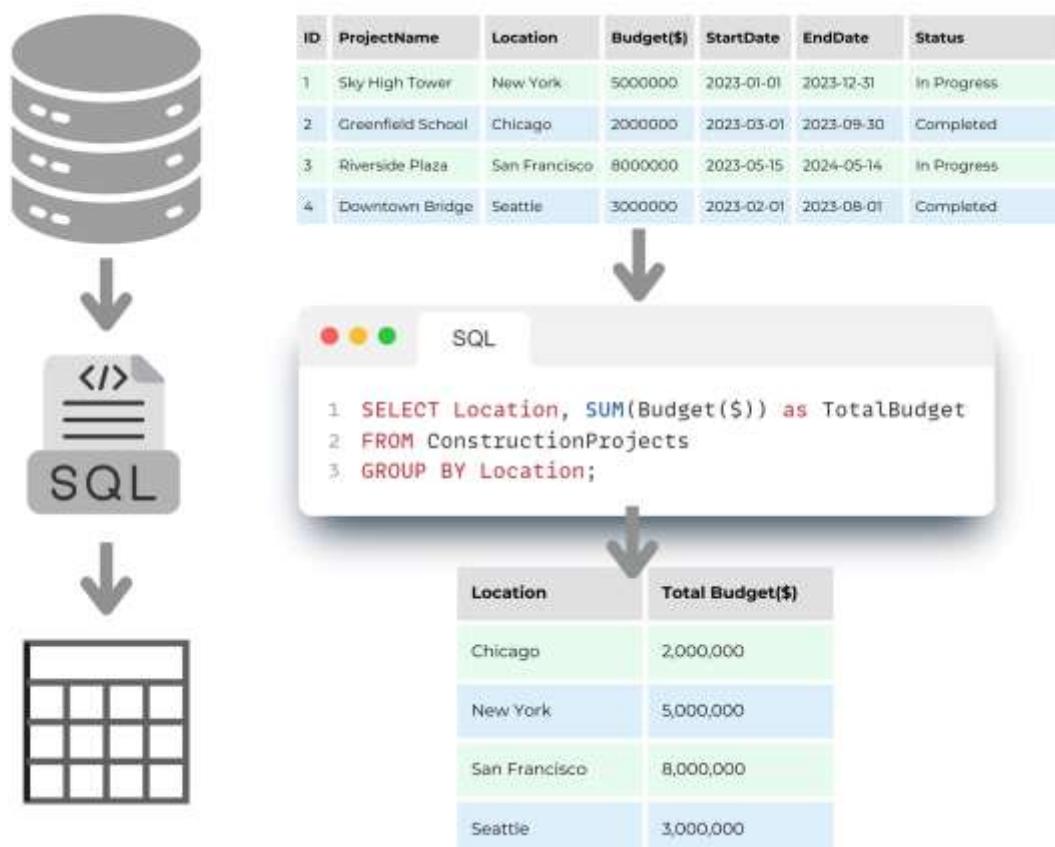


Fig. 3.1-7 Example of DML in SQL: fast processing, grouping, and aggregation with a few lines of code for automatic data processing.

Excel makes it easy to work with data because of its visual and intuitive structure. However, as the amount of data increases, Excel's performance decreases. Excel also faces limits on the amount of data it can store - a maximum of one million rows - and performance degrades long before this limit is reached. So while Excel looks preferable for visualization and manipulation of small amounts of data, SQL is better suited for handling large data sets.

The next stage in the development of structured data was the emergence of columnar databases (Columnar Databases), which are an alternative to traditional relational databases, especially when it comes to significantly larger data volumes and analytical calculations. Unlike row databases, where data is stored line by line, columnar databases record information by column. Compared to classical databases, this allows:

- Reduce storage space by efficiently compressing uniform data in columns.
- Speed up analytic queries, as only the required columns are read, not the entire table.
- Optimize Big Data and data warehouses, e.g. Data Lakehouse Architecture.

We will talk more about columnar databases, Pandas DataFrame, Apache Parquet, HDF5, as well as about creating Big Data -stores based on them for the purposes of data analysis and processing in the following chapters of this book - "DataFrame: a universal tabular data format" and "Data storage formats and working with Apache Parquet: DWH -data warehouses and Data Lakehouse architecture".

Unstructured data

Although most of the data used in applications and information systems is in structured form, most of the information generated in construction is in the form of unstructured data - images, videos, text documents, audio recordings and other forms of content. This is especially true at the construction, operation and technical supervision stage, where visual and textual information predominates.

Unstructured data is information that has no predetermined model or structure, not organized into traditional rows and columns as in databases or tables.

In general terms, unstructured data can be classified into two categories:

- Human-generated unstructured data, which includes various types of human-generated content: text documents, emails, images, videos, and so on.
- Machine-generated unstructured data is created by devices and sensors: these include log files, GPS data, Internet of Things (IoT) results, and other telemetry information from a construction site, for example.

Unlike structured data, which are conveniently organized in tables and databases, unstructured data require additional processing steps before their integration into information systems (Fig. 3.1-8). The use of technologies for automated collection, analysis and transformation of such data opens up new opportunities for improving construction efficiency, reducing errors and minimizing the influence of human factor.

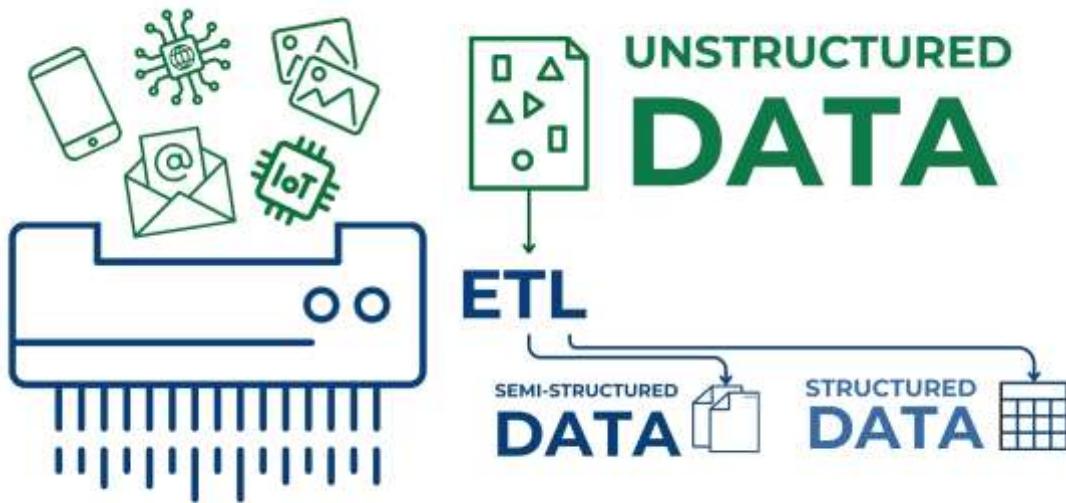


Fig. 3.1-8 The processing of unstructured data begins by transforming it into semi-structured and structured data.

Unstructured data account for up to 80% of all information [52] encountered by professionals in companies, so we will discuss their types and processing in detail with examples in the following chapters of the book.

For ease of discussion, we will separate textual data into a separate category. Although they are a type of rather unstructured data, their importance and prevalence in the construction industry require special attention.

Text data: between unstructured chaos and structure

Textual data in the construction industry covers a wide range of formats and types of information, from paper documents to informal methods of communication such as letters, conversations, work correspondence and verbal meetings at the construction site. All of this textual data carries important information for managing construction projects, from details of design decisions and changes in plans to discussions of safety issues and negotiations with contractors and clients (Fig. 3.1-9).



Fig. 3.1-9 Text data, one of the most popular types of information used in communication between project participants.

Textual information can be both formalized and unstructured. Formalized data includes Word documents (.doc,.docx), PDF, as well as text files of meeting minutes (.txt). Non-formalized data include messenger and mail correspondence, meeting transcripts (Teams, Zoom, Google Meet), and audio recordings of discussions (.mp3,.wav) that require conversion to text.

But while written documents such as formal requests, contract terms and conditions, and emails usually already have some structure, verbal communications and work correspondence often remain unstructured, making them difficult to analyze and integrate into project management systems.

The key to effective management of text data is to convert it into a structured format. This allows you to automatically integrate the processed information into existing systems that already work with structured data.



Fig. 3.1-10 Converting textual content into structured data.

To make effective use of textual information, it must be automatically converted into a structured form (Fig. 3.1-10). This process usually involves several steps:

- **Text Recognition (OCR)** - converting images of documents and drawings into a machine-readable format.
- **Text analysis (NLP)** - automatic identification of key parameters (dates, amounts and figures related to the project).
- **Data classification** - categorizing information (finance, logistics, risk management).

After recognition and classification the already structured data can be integrated into databases and used in automated reporting and management systems.

Semi-structured and loosely structured data

Semi-structured data contains some level of organization, but does not have a strict schema or structure. Although such information includes structured elements (e.g. dates, employee names and lists of tasks completed), the format of presentation may vary considerably from project to project or even from one employee to another. Examples of such data are time logs, progress reports and schedules, which can be presented in a variety of formats.

Semi-structured data is easier to analyze than unstructured data, but requires additional processing to integrate into standardized project management systems.

Working with semi-structured data, characterized by the presence of constantly changing structure, presents significant challenges. This is because the variability of the data structure requires separate individual approaches to processing and analyzing each source of semi-structured data.

But while dealing with unstructured data requires a lot of effort, processing semi-structured data can be done with relatively simple methods and tools.

Semi-structured data is a more general term that describes data with minimal or incomplete structure. It is most often text documents, chats, emails where some metadata (e.g. date, sender) is found, but most of the information is presented in a chaotic manner.

In construction, loosely structured data is found in a variety of processes. For example, they may include:

- Estimates and quotations - tables with material, volume and cost data, but without a uniform format.
- Drawings and engineering schematics - files in PDF or DWG, containing text annotations and metadata, but without a strictly fixed structure.
- Work schedules - data from MS Project, Primavera P6 or other systems, which may have different export structure.
- CAD (BIM -models) - contain elements of the structure, but data representation depends on the software and project standard.

Geometric data, produced by CAD systems, can be classified in the same way as semi-structured data. However, we will distinguish geometric CAD (BIM) data as a separate data type because it, like text data, can often be treated as a separate data type in company processes.

Geometric data and its application

If metadata about project elements are almost always stored in the form of tables, structured or weakly structured formats, then geometric data of project elements in most cases are created using special CAD tools (Fig. 3.1-11), allowing to visualize project elements in detail as a set of lines (2D) or geometric bodies (3D).

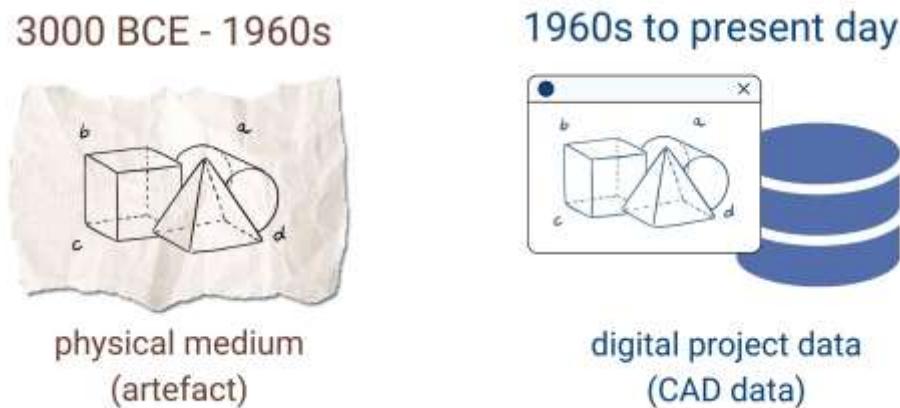


Fig. 3.1-11 CAD tools have helped move geometric information from physical media to database form.

When working with geometric data in construction and architecture, three main applications of geometric data can be distinguished (Fig. 3.1-12):

- **Confirmation of volumes:** geometric data, generated within CAD programs (BIM) using special geometric kernels, are required to automatically and accurately determine the volumes and dimensions of project elements. This data includes automatically calculated areas, volumes, lengths and other important attributes required for planning, budgeting and ordering of resources and materials
- **Visualization of the project:** in case of any changes in the project, visualization of elements allows to automatically generate updated drawings in different planes. Visualization of the project at the initial stages allows to accelerate the mutual understanding between all participants to save time and resources during the construction process.
- **Checking collisions:** In complex construction and engineering projects where the interaction of multiple categories of elements (e.g. pipes and walls) without "geometric conflicts" is critical, collision checking plays a key role. Utilizing collision detection software allows you to proactively identify potential geometric conflicts between project elements, preventing costly errors during the construction process.

From the very beginning of engineering and design bureaus, from the time of construction of the first complex structures, structural engineers have provided geometric information in the form of drawings, lines and flat geometric elements (on papyrus, "A0" paperboard or in DWG, PDF, PLT formats), on the basis of which foremen and estimators (Fig. 3.1-11), for the last millennia, with the help of rulers and transportation, collected attributive volumes or quantities of elements and groups of elements.

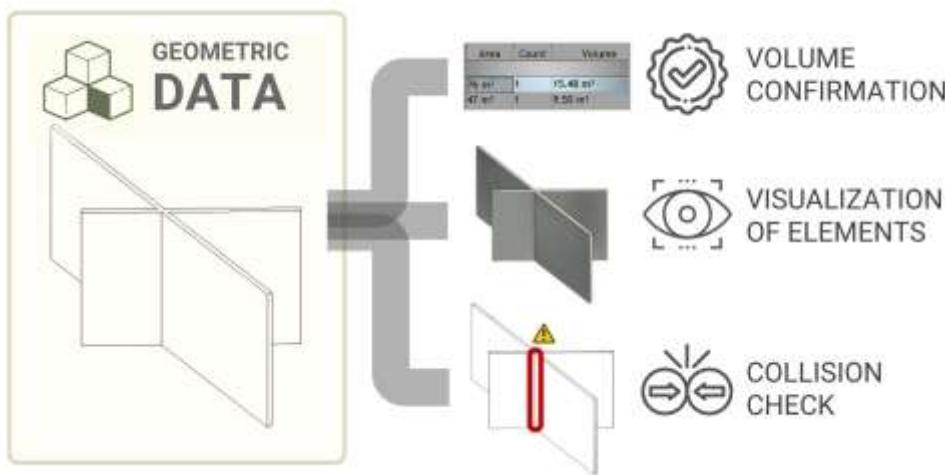


Fig. 3.1-12 The geometry is the basis for obtaining the volumetric parameters of the elements, which are then used to calculate the cost and schedule of the project.

Today, this manual and time-consuming task is solved by full automation thanks to the emergence of volumetric modeling in modern CAD tools (BIM), which allows automatically, with the help of a special geometric kernel, to obtain volumetric attributes of any element without the need to calculate volumetric parameters manually.

Modern CAD tools also allow you to classify and categorize project elements so that specification tables can be downloaded from the project database for use in various systems, such as cost estimating, scheduling, or CO₂ calculation (Fig. 3.1-13). We will discuss Obtaining specifications, QTO tables and quantities, and practical examples in the chapter "Obtaining quantities and quantification".

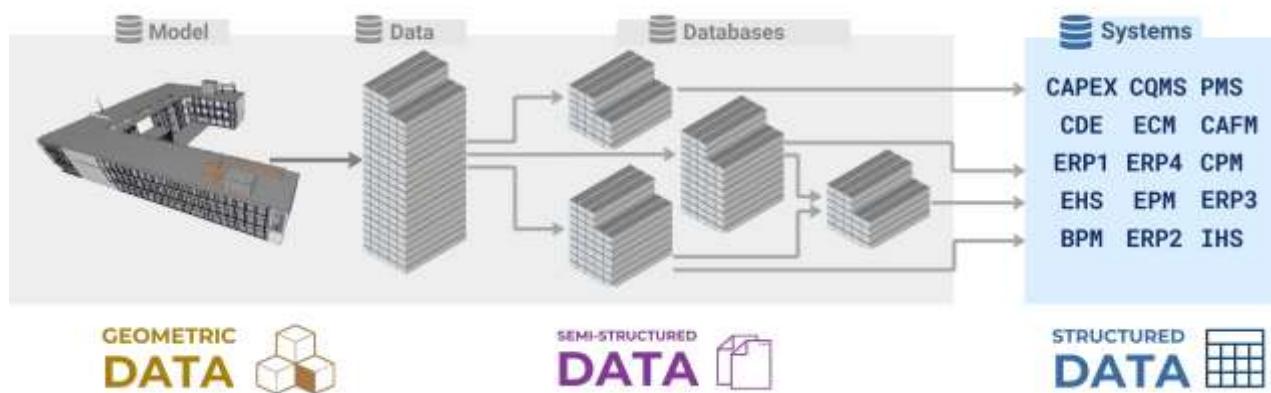


Fig. 3.1-13 CAD tools (BIM) store data in databases that are designed to integrate and interact with other systems.

Due to the closed nature of databases and formats used in CAD environment, geometric data created in CAD solutions have actually become a separate type of information. It combines both geometry of elements and meta-information (structured or semi-structured), enclosed in specialized files and formats.

CAD data: from design to data storage

Modern CAD and BIM systems store data in their own, often proprietary formats: DWG, DXF, RVT, DGN, PLN and others. These formats support both 2D and 3D representations of objects, preserving not only the geometry but also the attributes associated with the objects. Here are the most common ones:

- **DWG** is a binary file format used to store two-dimensional (and less commonly three-dimensional) design data and metadata.
- **DXF** is a text format for exchanging 2D and 3D -drawings between CAD -systems. It contains geometry, layers and attribute data, supports both ASCII and binary representation.
- **RVT** is a binary format for storing CAD models including 3D -geometry, element attributes, relationships, and design parameters.
- **IFC** is an open text format for exchanging construction data between CAD (BIM) systems. It includes geometry, object properties and information about their relationships.

In addition to these, other formats are used: PLN, DB1, SVF, NWC, CPIXML, BLEND, BX3, USD, XLSX, DAE. Although they differ in purpose and level of openness (Fig. 3.1-14), they can all represent the same project information model in different forms. In complex projects, these formats are often used in parallel, from drafting to project model coordination.

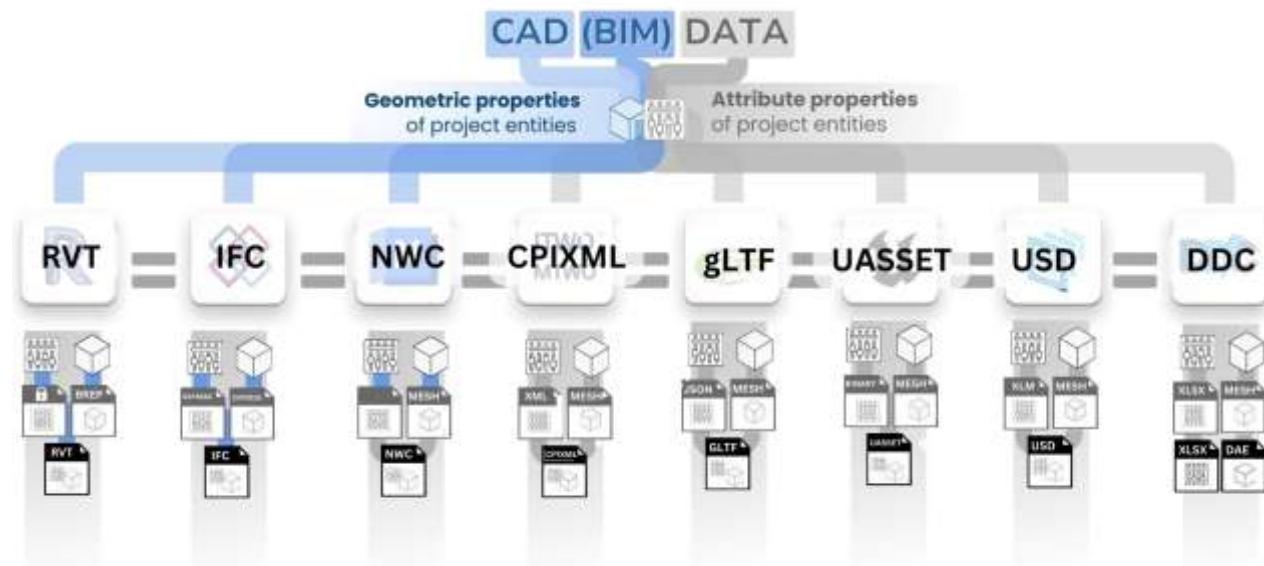


Fig. 3.1-14 Popular CAD storage formats describe geometry through BREP or MESH parameters, supplemented by attribute data.

All of the above formats allow you to store data about each element of a construction project and all of the above formats contain two key types of data:

- **Geometric parameters** - describe the shape, location and dimensions of an object. Geometry and its use will be discussed in detail in the sixth part of the book dedicated to CAD (BIM) solutions;

- **Attribute properties** - contain various information: materials, element types, technical characteristics, unique identifiers and other properties that project elements may have.

Attribute data are of particular importance in modern projects, as they determine the operational characteristics of objects, allow for engineering and costing calculations and provide end-to-end interaction between participants in design, construction and operation. For example:

- For windows and doors: type of construction, type of glazing, opening direction (Fig. 3.2-1).
- For walls, information on materials, thermal insulation and acoustic performance is recorded.
- For engineering systems the parameters of pipelines, ducts, cable routes and their connections are stored.

These parameters can be stored both within the CAD-(BIM-)files themselves and in external databases - as a result of export, conversion or direct access to internal CAD structures via reverse engineering tools. This approach facilitates the integration of design information with other corporate systems and platforms

Reverse engineering in the context of CAD (BIM) is the process of extracting and analyzing the internal structure of a digital model to recreate its logic, data structure and dependencies without access to the original algorithms or documentation.

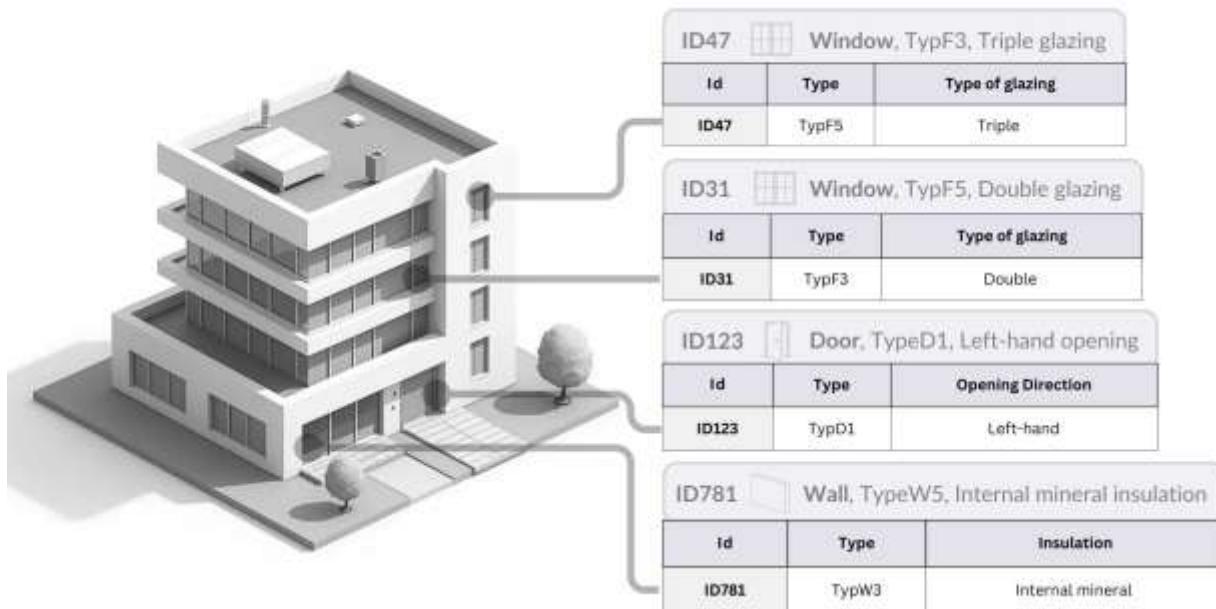


Fig. 3.1-15 A project element, in addition to describing parametric or polygonal geometry, contains information about the parameters and properties of elements.

As a result, a unique set of parameters and properties is formed around each element, including both unique characteristics of each object (e.g., identifier and dimensions) and common attributes for groups of elements. This allows not only to analyze individual elements-entities of the project, but also

to combine them into logical groups, which can then be used by other specialists for their tasks and calculations in systems and databases.

An entity is a concrete or abstract object of the real world that can be uniquely identified, described and represented in the form of data.

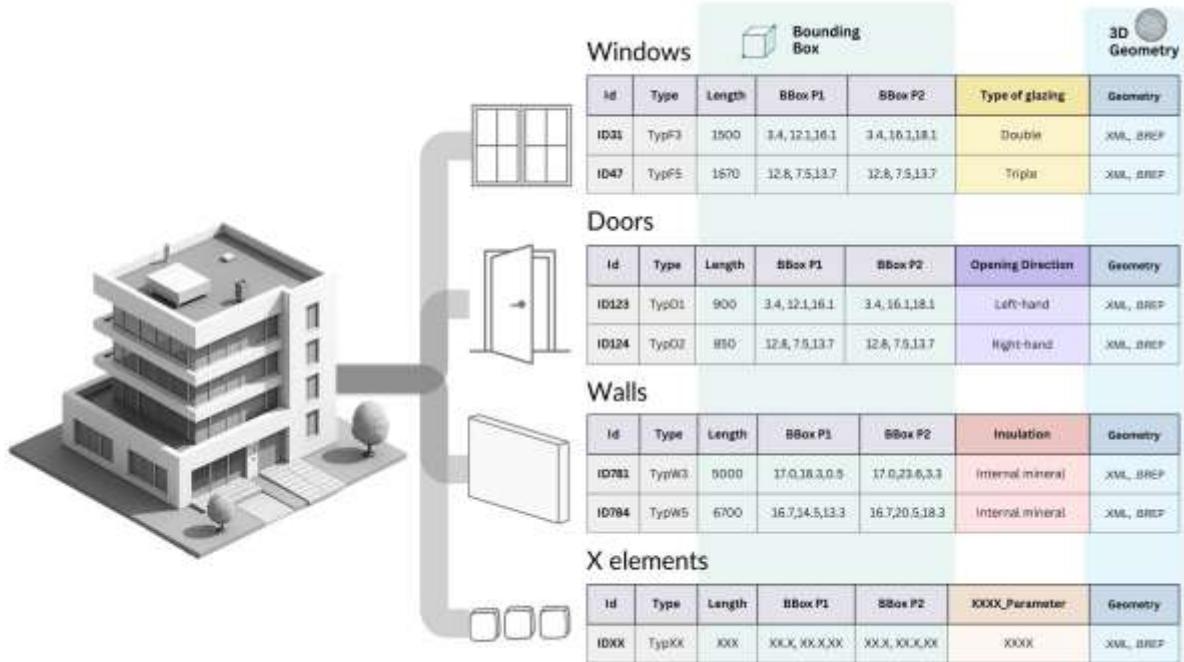


Fig. 3.1-16 Each design element contains attributes that are either entered by the designer or calculated within the CAD program.

Over the last decades, the construction industry has developed many new CAD (BIM) formats that simplify the creation, storage and transfer of data. These formats can be closed or open, tabular, parametric or graphical. However, their diversity and fragmentation significantly complicate data management at all stages of the project lifecycle. A table comparing the main formats used for information exchange in construction is presented in Fig. 3.1-17 (full version available by QR code).

To solve the problems of interoperability and access to CAD data, managers (BIM) and coordinators are included, whose task is to control exports, check data quality and integrate parts of CAD (BIM) data into other systems.

However, due to the closed nature and complexity of formats, it is difficult to automate this process, which forces specialists to perform many operations manually, without the ability to build full-fledged in-line data processing processes (pipeline).

The table is a comprehensive comparison of data formats for construction projects. It includes sections for 'Geometric properties of project entities' and 'Attribute properties of project entities'. The table is very wide and contains numerous colored cells representing different data types across various categories.

Fig. 3.1-17 Table comparing the main data formats in which project element information is stored [53].

To understand why there are so many different data formats, and why most of them are closed, it is important to delve into the processes that take place inside CAD (BIM) programs, which will be explored in detail in the sixth part of the book.

An additional information layer added to the geometry was introduced by CAD system developers in the form of the BIM concept (Building Information Modeling), a marketing term actively promoted in the construction industry since 2002 [54].

The emergence of the BIM (BOM) concept and the use of CAD in processes

The concept of Building Information Modeling (BIM), first outlined in the 2002 BIM Whitepaper [54], originated from the marketing initiatives of CAD software manufacturers. It emerged from the marketing initiatives of CAD software developers and was an attempt to adapt the principles already well established in mechanical engineering to the needs of the construction industry.

The inspiration for BIM came from the concept of BOM (Bill of Materials), a product composition specification that has been used extensively in industry since the late 1980s. In mechanical engineering, BOM allowed linking data from CAD systems with PDM (Product Data Management), PLM (Product Lifecycle Management) and ERP systems, providing holistic management of engineering information throughout the entire product lifecycle (Fig. 3.1-8).

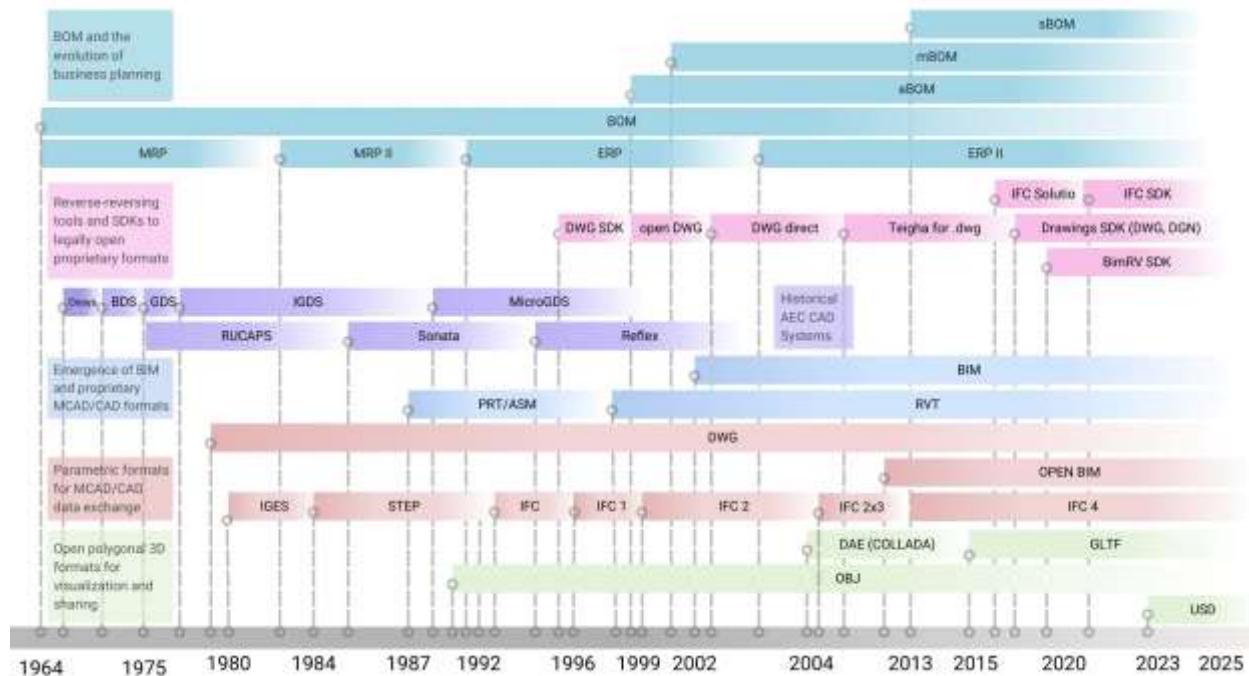


Fig. 3.1-18 Evolution of specifications (BOM), information modeling (BIM), and digital formats in the engineering construction industry.

The modern development of the BOM concept has led to the emergence of an extended framework - XBOM (Extended BOM), which includes not only product composition, but also behavioral scenarios, operational requirements, sustainability parameters and data for predictive analytics. XBOM essentially fulfills the same role as BIM in construction: both approaches strive to turn the digital model into a Single Source of Truth for all project participants throughout the entire lifecycle of an object.

A key milestone in the emergence of BOM in construction was the introduction of the first parametric CAD (MCAD) specifically adapted for the construction industry in 2002. It was developed by the team that had previously created Pro-E®, a revolutionary MCAD system for mechanical engineering that appeared in the late 1980s and became an industry standard [55].

Already in the late 1980s, the goal was to eliminate the limitations [56] of the then existing CAD -programs. The main objective was to reduce the labor required to make changes to the parameters of design elements and to make it possible to update the model based on data outside CAD programs via a database [57]. The most important role in this was to be played by parametrization: automatic retrieval of characteristics from the database and using them to update the model inside CAD-systems.

Pro-E and the concept of elemental parametric modeling c BOM underlying it have had a significant impact on the development of the CAD - and MCAD - market [58]. For 25 years this model has been in the industry and many modern systems have become its conceptual successors.

The goal is to create a system that is flexible enough to encourage the engineer to easily consider different designs. And the cost of making changes to the design should be as close to zero as possible. Traditional CAD / CAM software unrealistically restricts making inexpensive changes only at the very beginning of the design process [59].

— Samuel Geisberg, founder of Parametric Technology Corporation®, developer of MCAD - product Pro-E and teacher of the creator of a CAD product using the RVT format

In mechanical engineering, PDM, PLM, MRP and ERP systems have become key platforms. They play a central role in data and process management, gathering information from CAx systems (CAD, CAM, CAE) and organizing design activities based on the product structure (BOM: eBOM, pBOM, mBOM) (Fig. 3.1-18). This integration reduces errors, avoids data duplication and ensures end-to-end traceability from design to production.

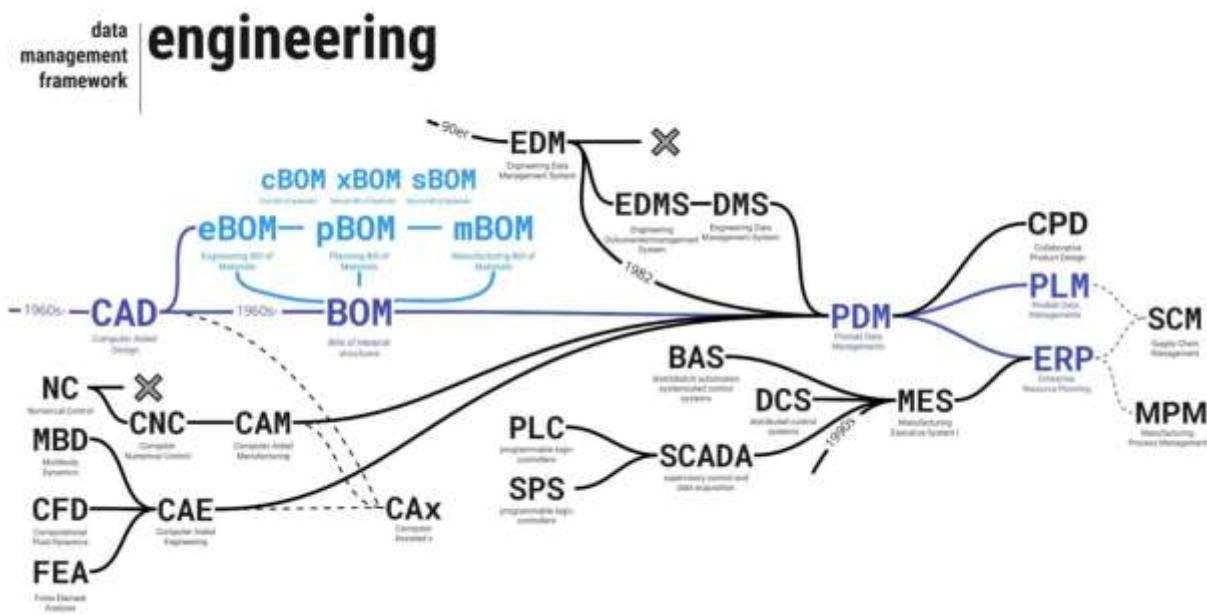


Fig. 3.1-19 Historically, BOM emerged in the 1960s as a way to structure data from CAx systems and pass it to control systems.

The purchase by one of the leading vendors of a CAD solution developed by the former Pro-E team and based on the BOM approach was marked by the almost immediate publication of the BIM Whitepaper series (2002-2003)[60][61]. As early as the mid-2000s, the BIM concept began to be actively promoted in the construction industry, which markedly increased interest in parametric software. The popularity grew so rapidly that the construction fork of mechanical engineering Pro-E - parametric CAD promoted by this vendor - has actually displaced competitors in the architectural and structural design segment

(Fig. 3.1-20). By the early 2020s, it has de facto consolidated global dominance in the BIM (CAD) market [62].



Fig. 3.1-20 Google search query popularity (RVT versus IFC): parametric CAD created by the former Pro-E team with BOM support -BIM has gained popularity in almost most countries of the world.

Over the past 20 years, the abbreviation BIM has acquired a multitude of interpretations, the polysemy of which has its roots in the initial marketing concepts that emerged in the early 2000s. The ISO 19650 standard, which played an important role in popularizing the term, actually secured the status of BIM as a "scientifically based" approach to information management. However, in the text of the standard itself, which is dedicated to data management throughout the life cycle of objects using BIM, the abbreviation BIM is mentioned, but never clearly defined.

The vendor's original website, which published a series of Whitepaper on BIM in 2002[60] and 2003[61], actually reproduced marketing materials on the BOM (Bills of Materials) and PLM (Product Lifecycle Management) concepts previously used in Pro-E mechanical engineering software back in the 1990s [63].

Building Information Modeling, an innovative new approach to building design, construction, and management introduced by..... [CAD vendor company name] in 2002, has changed the way industry professionals around the world think about how technology can be applied to the design, construction, and management of buildings.

— BIM Whitepaper, 2003 [61]

These early publications linked BIM directly to the concept of a centralized integrated database. As stated in the 2003 Whitepaper, BIM is building information management where all updates occur in a single repository, keeping all drawings, cuts and specifications (BOM - Bills of Materials) synchronized.

BIM is described as building information management, where all updates and all changes take place in a database. So whether you are dealing with schematics, sections or sheet drawings, everything is always coordinated, consistent and up to date.

— CAD company website vendor with BIM Whitepaper, 2003 [54]

The idea of managing design through a single integrated database has been widely discussed as early as in the studies of the 1980s. For example, Charles Eastman's BDS concept [57] included 43 references to the term "database" (Fig. 6.1-2). By 2004, this number had almost halved to 23 in the 2002 Whitepaper on BIM [64]. And by the mid-2000s, the topic of databases had virtually disappeared from vendors' marketing materials and the digitalization agenda in general.

Although it was the database and access to it that was originally conceived as the core of the BIM - system, over time the emphasis shifted to geometry, visualization and 3D. The very registrar of the IFC standard in 1994, who published the BIM Whitepaper in 2002 - the same vendor - in the Whitepaper of the early 2000s explicitly pointed out the limitations of neutral formats such as IGES, STEP and IFC and the need for direct access to CAD databases:

Different applications may be incompatible and re-entered data may be inaccurate [...]. The result of traditional computer-aided design [CAD]: higher costs, longer time-to-market, and lower product quality. Today, all major applications use industry standard interfaces for low-level data exchange. By using the old IGES standards or the new STEP [IFC is a de facto and de jure copy of the STEP/IGES format] to exchange data between applications from different vendors, users can achieve some data compatibility between best-of-breed products. But IGES and STEP only work at low levels, and they cannot exchange data as rich as the information generated by today's leading applications [...]. And while these and other standards are improving almost daily, they will always lag behind today's vendor products in terms of data richness. [...] programs within an application must be able to share and preserve data richness without resorting to neutral translators such as IGES, STEP [IFC] or PATRAN. Instead, framework applications should be able to directly access the underlying CAD database so that the detail and accuracy of the information is not lost.

— CAD vendor Whitepaper (IFC, BIM) "Integrated Design and Manufacturing: Benefits and Rationale," 2000 [65]

Thus, already in the 1980s and early 2000s, the key element of digital design in the CAD environment was considered to be the database rather than the format-file or the neutral IFC format. It was suggested that translators should be abandoned and applications should have direct access to the data. However, in reality, by the mid-2020s, the concept of BIM began to resemble a "divide and conquer"

strategy, where the interests of software vendors using closed geometric kernels are prioritized over the development of open information exchange.

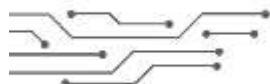
Today, BIM is perceived as an integral part of the construction industry. But over the past two decades, the promise of simplified collaboration and data integration has largely gone unrealized. Most solutions are still tied to closed formats or neutral formats and specialized tools. We will look in detail at the history of BIM, open BIM and IFC, as well as the issues of interoperability and geometric kernels in Part 6 of the book "CAD and BIM: Marketing, Reality and the Future of Design Data in Construction".

Today, the industry faces a key challenge to move from the traditional understanding of CAD (BIM) as a modeling tool to its use as a full-fledged database. This requires new approaches to working with information, abandoning the dependence on closed ecosystems and implementing open solutions.

With the development of reverse engineering tools that allow access to CAD databases and the proliferation of Open Source and LLM technologies, users and developers in the construction industry are increasingly moving away from the vague terms of software vendors. Instead, the focus is shifting to what really matters: data (databases) and processes.

Behind the trendy acronyms and visualizations are standard data management practices: storage, transfer and transformation - i.e. the classic ETL process (Extract, Transform, Load). As in other industries, the digitalization of construction requires not only exchange standards, but also clearly structured handling of heterogeneous information.

In order to fully utilize the potential of CAD (BIM) data, companies need to rethink their approach to information management. This will inevitably lead to a key element of digital transformation - unification, standardization and meaningful structuring of the data that construction professionals work with on a daily basis.



CHAPTER 3.2.

DATA UNIFICATION AND STRUCTURING

Filling systems with data in the construction industry

Whether it is large corporations or medium-sized companies, specialists are daily engaged in filling program systems and databases with various interfaces with multiformat information (Fig. 3.2-1), which, with the help of managers, must interact with each other in a coherent manner. It is this complex of interacting systems and processes that ultimately creates revenue and profit for the company.

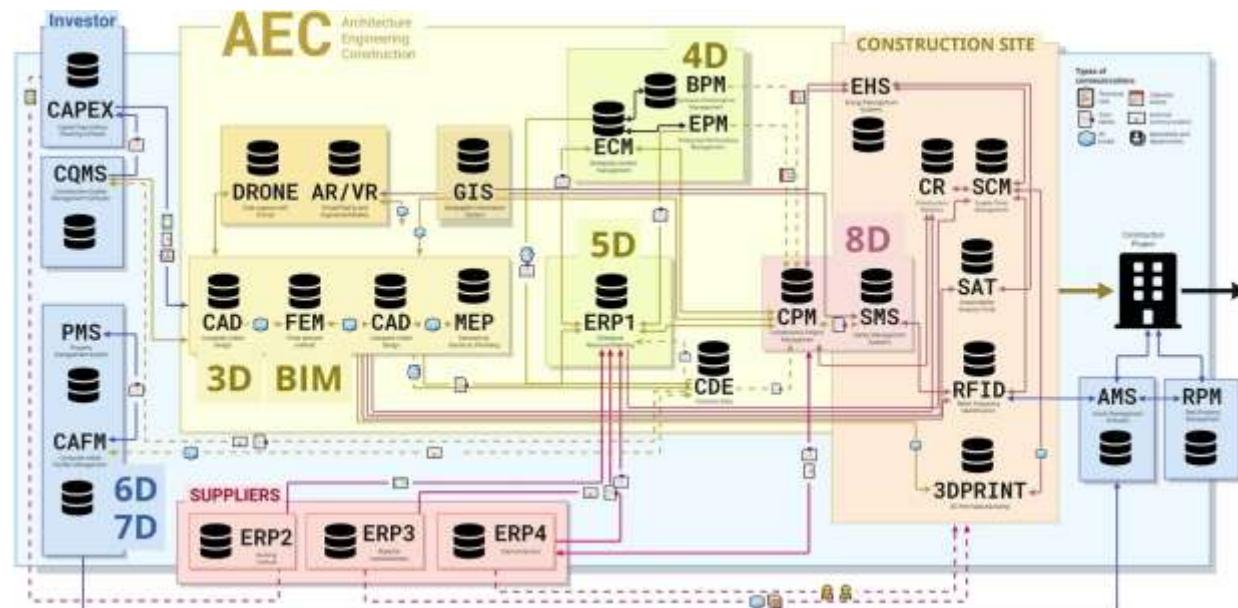


Fig. 3.2-1 Virtually every system or application used in the construction business has one of the popular RDBMS databases at its core.

Each of the categories of systems mentioned earlier and applied in the construction industry works with its own data types corresponding to the functional role of these systems. To move from the abstract level to the concrete, we move from data types to their representation as formats and documents.

To the previously presented list of systems (Fig. 1.2-4), we now add the specific types of formats and documents they often work with:

■ Investor (CAPEX)

- Financial data: budgets, expenditure forecasts (structured data).
- Market trend data: market analysis (structured and unstructured data).
- Legal and contractual data: contracts (text data).

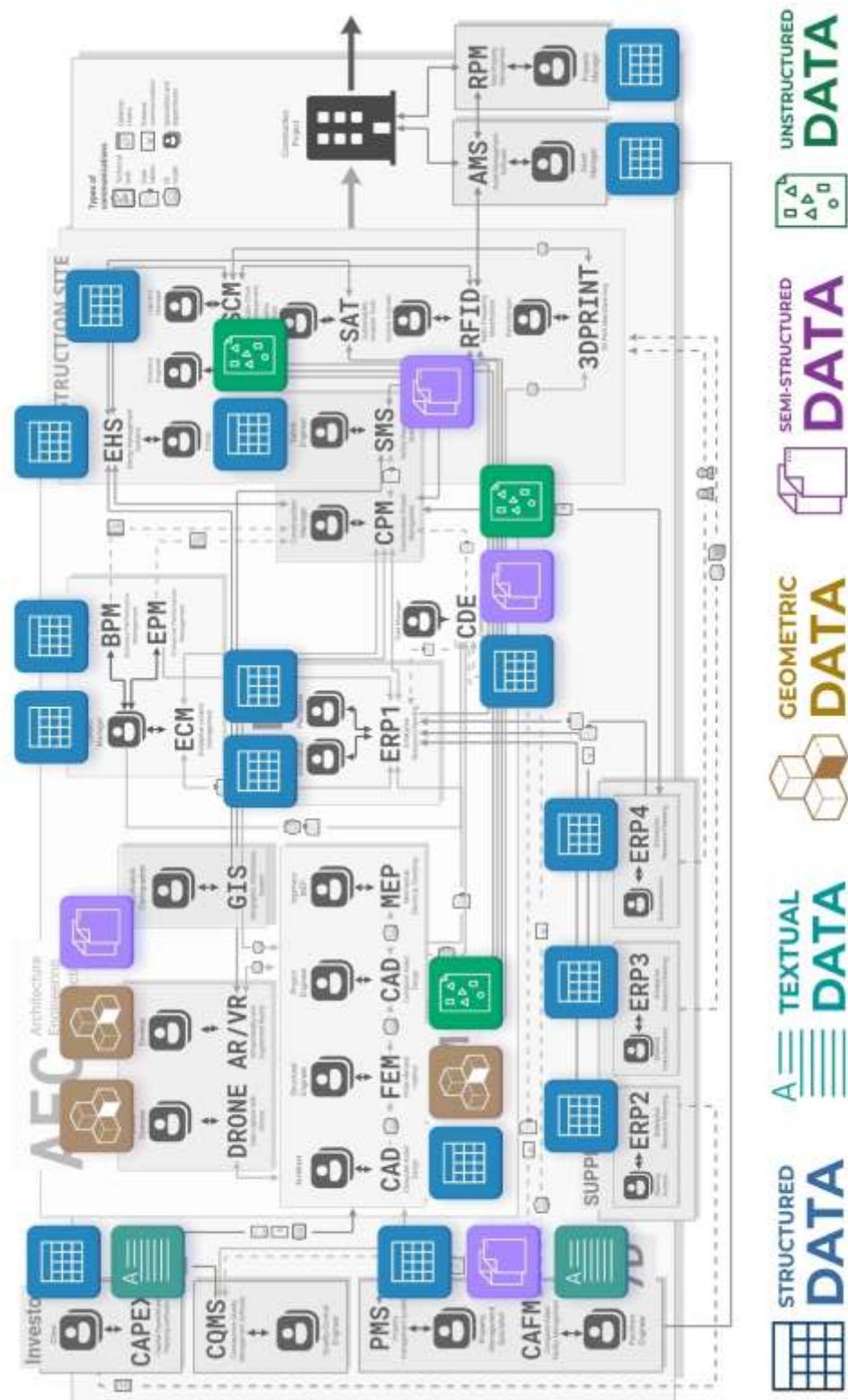


Fig. 3.2-2 The construction industry uses many systems with different interfaces that handle different types of data.

■ **Management systems (PMS, CAFM, CQMS)**

- Project data: graphs, tasks (structured data).
- Facility maintenance data: maintenance plans (text and semi-structured data).
- Quality control data: standards, inspection reports (textual and unstructured data).

■ **CAD, FEM and BIM**

- Technical drawings: architectural, structural plans (geometric data, unstructured data).
- Building models: 3D -models, material data (geometric and semi-structured data).
- Engineering calculations: load analysis (structured data).

■ **Construction site management systems (EHS, SCM)**

- Safety and health data: safety protocols (textual and structured data).
- Supply chain data: inventories, orders (structured data).
- Daily reports: working hours, productivity (structured data).

■ **Drones, AR/VR, GIS, 3D -printing**

- Geodata: topographic maps (geometric and structured data).
- Real-time data: video and photos (unstructured data).
- Models for 3D -printing: digital drawings (geometric data).

■ **Additional management systems (4D BPM, 5D ERP1)**

- Time and cost data: schedules, estimates (structured data).
- Change management: project change records (text and structured data).
- Performance reporting: indicators of success (structured data).

■ **Data integration and communication (CDE, RFID, AMS, RPM)**

- Data exchange: document exchange, data models (structured and textual data).
- RFID and tracking data: logistics, asset management (structured data).
- Monitoring and control: sensors on sites (structured and unstructured data).

Thus, each system in the construction industry - from site management systems to operational databases - operates with its own type of information: structured, textual, geometric, etc. The "data landscape" that professionals have to work with on a daily basis is extremely diverse. The "data landscape" with which specialists have to work on a daily basis is extremely diverse. However, a simple enumeration of formats does not reveal the complexity of real work with information.

In practice, companies are faced with the fact that data, even when retrieved from systems, is not ready to be used "as is". This is especially true for texts, images, PDFs, CAD files and other formats that are difficult to analyze with standard tools. That is why the next key step is data transformation - a process without which processing, analysis, visualization and decision-making cannot be effectively automated.

Data transformation: the critical foundation of modern business analysis

Today, most companies are facing a paradox: about 80% of their daily processes still rely on classic structured data - familiar Excel spreadsheets and relational databases (RDBMS) [66]. However, at the

same time, 80% of new information entering the digital ecosystem of companies is unstructured or loosely structured (Fig. 3.2-3) [52]. This includes text, graphics, geometry, images, CAD -models, documentation in PDF, audio and video recordings, electronic correspondence, and much more.

Moreover, the volume of unstructured data continues to grow rapidly - the annual growth rate is estimated at 55-65% [67]. Such dynamics creates serious difficulties in integrating new information into existing business processes. Ignoring this flow of multiformat data leads to the formation of information gaps and to a decrease in the manageability of the entire digital environment of the company.



Fig. 3.2-3 The annual growth in unstructured data creates challenges in integrating streaming information into business processes.

Ignoring complex unstructured and confusing loosely structured data in automation processes can lead to significant gaps in a company's information landscape. In today's world of uncontrolled and avalanche-like information movement, companies need to adopt a hybrid approach to data management that incorporates effective methods for handling all types of data.

The key to effective data management lies in organizing, structuring and classifying different types of data "Babel" (including unstructured, textual and geometric formats, into structured or loosely structured data). This process transforms chaotic data sets into organized structures for integration into systems, thereby enabling decision making based on them (Fig. 3.2-4).

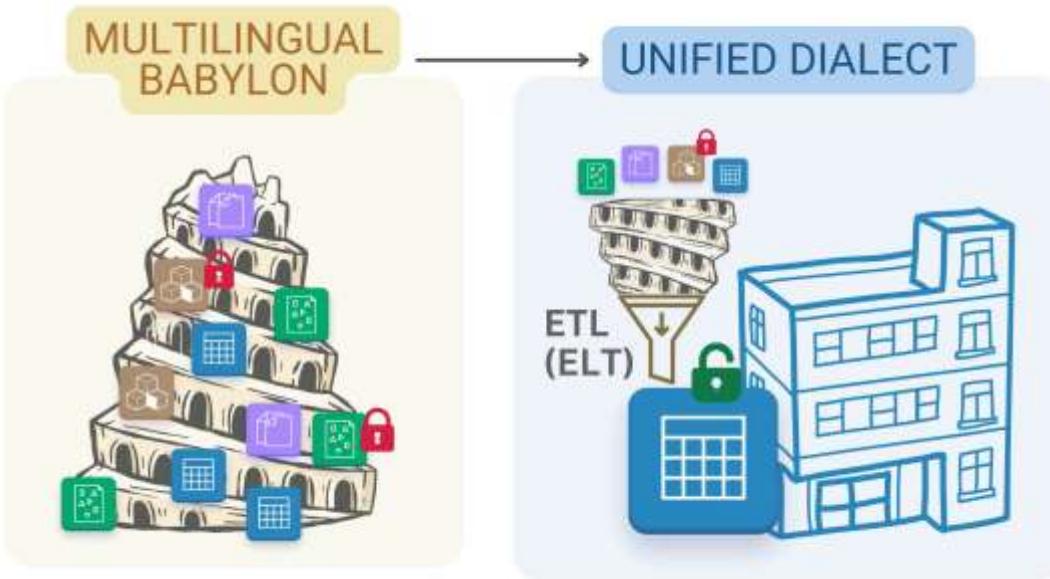


Fig. 3.2-4 The main task of data management departments is to translate the "Babylon" of diverse and multi-format data into a structured and categorized system.

One of the key obstacles to such unification remains the low level of interoperability between different digital platforms - the "silos" we discussed in the previous chapters.

According to the report, the National Institute of Standards and Technology (NIST, USA) emphasizes [68] that poor data compatibility between different building platforms leads to loss of information and significant additional costs. In 2002 alone, due to software interoperability problems, losses in US capital construction amounted to 15.8 billion dollars per year, where two-thirds of these losses are borne by building owners and operators, especially during operation and maintenance [68]. The study also notes that standardization of data formats can reduce these losses and improve efficiency throughout all phases of the facility life cycle.

According to the 2016 CrowdFlower study [69], which covered 16,000 data scientists around the world, the main problem remains "dirty" and multiformat data. According to this study, the most valuable resource is not the final databases or machine learning models, but the time spent on preparing information.

Cleaning, formatting, and organizing takes up to 60 percent of an analyst's and data manager's time. Nearly one-fifth is spent finding and collecting the right data sets, which are often hidden in closed storage ("silos") and inaccessible for analysis. And only about 9 percent of the time is spent directly on modeling, analytics, making predictions and testing hypotheses. The rest is spent communicating, visualizing, reporting, and researching supporting information sources

On average, a manager's data work is distributed as follows (Fig. 3.2-5):

- **Cleaning and organizing data (60%):** having clean and structured data can significantly reduce the analyst's work time and speed up the process of completing tasks.
- **Data collection (19%):** A major challenge for data science professionals is finding relevant datasets. Often, company data is stacked in chaotically organized "silos," making it difficult to access the information they need.
- **Modeling/Machine Learning (9%):** Often complicated by lack of clarity of business objectives on the part of customers. The lack of a clear mission statement can negate the potential of even the best model.
- **Other tasks (5%):** in addition to processing data, analysts have to deal with research, exploring data from different angles, communicating results through visualizations and reports, and recommending optimization processes and strategies.

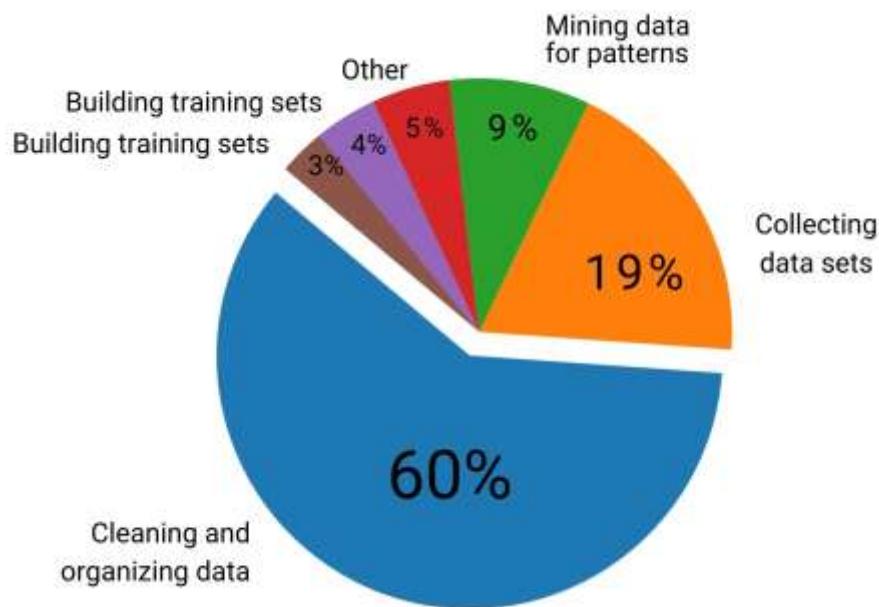


Fig. 3.2-5 What data managers working with data spend the most time on (based on [70]).

These estimates are supported by other studies as well. According to the Xplenty study published in BizReport in 2015 [71], between 50% and 90% of business intelligence (BI) professionals' time is spent on preparing data for analysis.

Cleaning, validating and organizing data represents a critical foundation for all downstream data and analytics processes, taking up to 90% of data scientists' time.

This painstaking work, invisible to the end user, is critical. Errors in raw data inevitably distort analysis results, are misleading and can lead to costly management errors. That's why data cleaning and standardization processes - from eliminating duplicates and filling in omissions to harmonizing units of measure and aligning to a common model - are becoming a cornerstone of today's digital strategy.

Thus, thorough transformation, cleaning and standardization of data not only occupy the majority of

specialists' time (up to 80% of work with data), but also determine the possibility of their effective use within the framework of modern business processes. However, data organization and cleaning alone do not exhaust the task of optimal management of the company's information flows. During the stage of organization and structuring becomes the choice of a suitable data model, which directly affects the convenience and efficiency of working with information in subsequent stages of processing.

Since data and business objectives are different, it is important to understand the characteristics of data models and be able to select or create the right structure. Depending on the degree of structuring and the way the relationships between elements are described, there are three main models: structured, loosely structured and graph models. Each is suitable for different tasks and has its own strengths and weaknesses.

Data models: relationships in data and relationships between elements

Data in information systems are organized in different ways - depending on the tasks and requirements for storing, processing and transmitting information. The key difference between the types of data models, the form in which information is stored, is the degree of structuring and the way in which the relationships between elements are described.

Structured data has a clear and repeatable schema: it is organized as tables with fixed columns. This format provides predictability, ease of processing and efficiency when performing SQL -queries, filtering and aggregation. Examples - databases (RDBMS), Excel, CSV.

Loosely structured data allows flexible structure: different elements can contain different set of attributes and be stored as hierarchies. Examples are JSON, XML or other document formats. These data are convenient when it is necessary to model nested objects and relations between them, but on the other hand, it complicates data analysis and standardization (Fig. 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

Fig. 3.2-6 A data model is a logical structure that describes how data is organized, stored, and processed in a system.

The choice of the appropriate format depends on the objectives:

- If the speed of filtering and analytics is important - relational tables (SQL, CSV, RDBMS, columnar databases) will do.
- If flexibility of structure is required - it is better to use JSON or XML.
- If the data has complex relationships - graph databases provide visibility and scalability.

In classical relational databases (RDBMS), each entity (e.g., a door) is represented by a row and its properties by table columns. For example, a table of items from the category "Doors" may contain the fields ID, Height, Width, Fire Resistance, and Room ID indicating the room (Fig. 3.2-7).

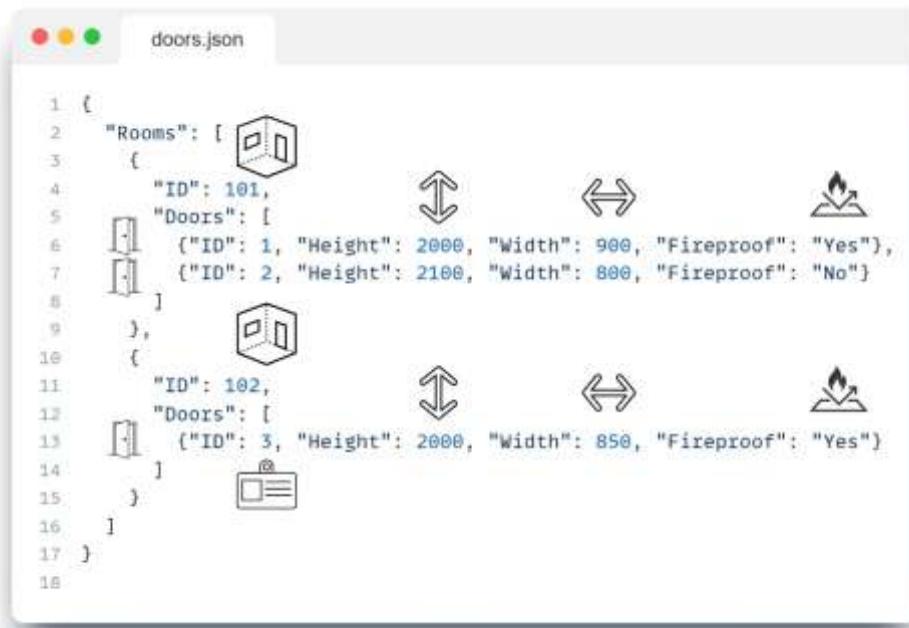
In classical relational databases (RDBMS) relations are formed in the form of tables, where each record represents an object and columns represent its parameters. In the tabular format the data about doors in the project looks like this, where each row represents a separate element - a door with its unique identifier and attributes, and the relationship with the room is realized through the parameter "Room ID".



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

Fig. 3.2-7 Information about the three elements of the "Doors" category of the project in tabular structured form.

In loosely structured formats such as JSON or XML, data is stored in a hierarchical or nested form, where elements may contain other objects and their structure may vary. This allows complex relationships between elements to be modeled. Similar information about doors in the project, which was recorded in structured form (Fig. 3.2-7), is represented in a loosely structured format (JSON) in such a way (Fig. 3.2-8) that they become nested objects within Rooms (Rooms - ID), which logically reflects the hierarchy.



```

1 {
2   "Rooms": [
3     {
4       "ID": 101,
5       "Doors": [
6         {"ID": 1, "Height": 2000, "Width": 900, "Fireproof": "Yes"},
7         {"ID": 2, "Height": 2100, "Width": 800, "Fireproof": "No"}
8       ]
9     },
10    {
11      "ID": 102,
12      "Doors": [
13        {"ID": 3, "Height": 2000, "Width": 850, "Fireproof": "Yes"}
14      ]
15    }
16  ]
17 }
18

```

The screenshot shows a Mac OS X window titled "doors.json". Inside the window, there is a JSON file containing information about rooms and doors. The JSON structure includes arrays for "Rooms" and "Doors", with various properties like ID, Height, Width, and Fireproof status.

Fig. 3.2-8 Information about elements of the "Doors" category of the project in JSON format.

In a graph model, data is represented as nodes (vertices) and links (edges) between them. This allows you to visualize the complex relationships between objects and their attributes. In the case of door and room data in the project, the graph representation of is as follows:

- **Nodes (nodes)** represent the main entities: rooms (Room 101, Room 102) and doors (ID1001, ID1002, ID1003)
- **Ribs (links)** show the relationships between these entities, e.g., the belonging of a door to a certain room
- **Attributes** are mapped to nodes and contain entity properties (height, width, fire resistance for doors)



Fig. 3.2-9 Project door entity information in graph view.

In the graph data model of door description, each room and each door are separate nodes. Doors are linked to rooms through edges that indicate whether the door belongs to a particular room. The attributes of the doors (height, width, fire resistance) are stored as properties of the corresponding nodes. More details about graph formats and how graph semantics appeared in the construction industry will be discussed in the chapter "The emergence of semantics and ontology in construction".

Graph databases are effective in cases where it is not so much the data itself that is important, but the relationships between them, for example, in recommender systems, routing systems, or when modeling complex relationships in facility management projects. The graph format simplifies the creation of new relationships by allowing new data types to be added to the graph without changing the storage structure. However, compared to relational tables and structured formats, there is no additional data connectivity in a graph - transferring two-dimensional database data into a graph does not increase the number of relationships and does not allow to obtain new information.

The form and schema of the data should be tailored to the specific use case and tasks to be solved. To work effectively in business processes, it is important to use those tools and those data models that help you get results as quickly and easily as possible.

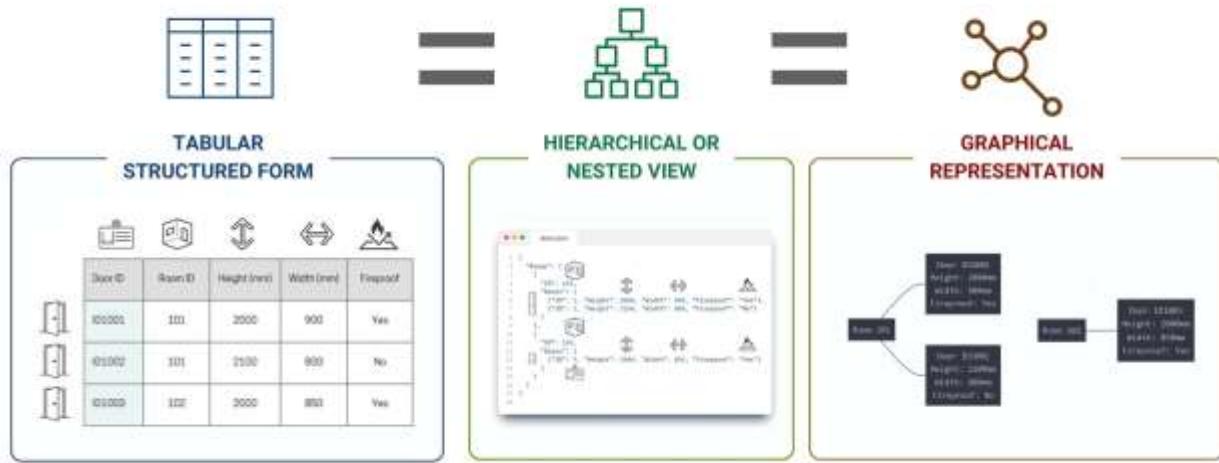


Fig. 3.2-10 The same information about project elements can be stored in different formats using different data models.

Today, most large companies face the problem of excessive data complexity. Each of hundreds or thousands of applications uses its own data model, which creates excessive complexity - an individual model is often dozens of times more complex than necessary, and the aggregate of all models is thousands of times more complex. This excessive complexity significantly hampers the work of both developers and end users.

Such complexity imposes serious limitations on the development and maintenance of the company's

systems. Each new element in the model requires additional code, implementation of new logic, thorough testing and adaptation to existing solutions. All this increases costs and slows down the work of the automation team in the company, turning even simple tasks into costly and time-consuming processes.

Complexity affects all levels of data architecture. In relational databases, it is expressed in the growing number of tables and columns, often redundant. In object-oriented systems, complexity is increased by the multiplicity of classes and interrelated properties. In formats like XML or JSON, complexity is manifested through confusing nested structures, unique keys, and inconsistent schemas.

The excessive complexity of data models makes systems not only less efficient, but also difficult to be understood by end users and in the future by large language models and LLM agents. It is the problem of understanding and complexity of data models and data processing that raises the question: how to make data easy enough to use that it actually starts to be useful quickly.

Even when data models are chosen wisely, their utility is dramatically reduced if access to the data is limited. Proprietary formats and closed platforms hinder integration, complicate automation, and take away companies' control over their own information, creating not just a silo of new data, but a locked silo that can only be accessed with the permission of the vendor. To understand the scale of the problem, it's important to consider exactly how closed systems affect digital processes in construction.

Proprietary formats and their impact on digital processes

One of the key challenges faced by construction companies during digitalization is limited access to data. This makes it difficult to integrate systems, reduces the quality of information and complicates the organization of efficient processes. The use of proprietary formats and closed software solutions is often at the root of these difficulties.

Unfortunately, until now, many programs used in the construction industry allow the user to save data exclusively in proprietary formats or cloud storage, which can only be accessed through strictly limited interfaces. And it's not uncommon for these solutions to be built in reliance on even more closed systems from larger vendors. As a result, even those developers who would like to offer more open architectures are forced to comply with the rules dictated by the large vendors.

While modern construction data management systems increasingly support open formats and standards (Fig. 3.1-5), CAD- (BIM)-based databases and related ERP and CAFM systems remain isolated proprietary "islands" in the industry's digital landscape (Fig. 3.2-11).



Fig. 3.2-11 The closed and proprietary nature of data creates barriers to data integration and access.

Closed and monopolized formats and protocols are not only a problem for the construction industry. In many sectors of the economy, the fight against closed standards and limited access to data started with slowing innovation (Fig. 3.2-12), the existence of artificial barriers to entry for new players and deepening dependence on large suppliers. With the rapid growth in the importance of data, competition authorities simply do not have time to respond to the challenges posed by new digital markets, and as a result, closed formats and restricted access to data essentially become digital "borders" that constrain the flow of information and growth [63].

If machines produce everything we need, then our situation will depend on how these goods are distributed. Everyone will only be able to enjoy a life of prosperity if the wealth produced by machines is shared. Or most people will end up living in abject poverty if car owners can successfully lobby against the redistribution of wealth. So far, things seem to be going the second way, with technology leading to ever greater inequality [72].

— Stephen Hawking, astrophysicist, 2015

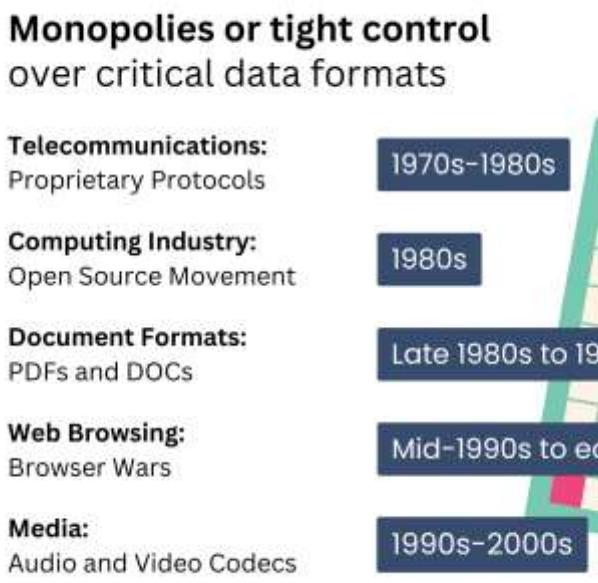


Fig. 3.2-12 Monopoly ownership over key data formats and protocols is not exclusive to the construction industry.

As a result, due to closed access to databases programs, data managers, analysts, IT specialists and developers creating applications for data access, processing and automation in the construction industry today face numerous dependencies on software vendors (Fig. 3.2-13). These dependencies in the form of additional access layers require the creation of solutions with specialized API -connections and special tools and software.

An API (Application Programming Interface) is a formalized interface through which one program can interact with another, exchanging data and functionality without having to access the source code. An API describes what requests an external system can make, what format they should be in, and what responses it will receive. It is a standardized "contract" between software modules.

The large number of dependencies on closed solutions causes the entire code architecture and business process logic in a company to become a "spaghetti architecture" of tools that depend on the software vendor's policy to provide quality access to data.

Dependence on closed solutions and platforms leads not only to loss of flexibility, but also to real business risks. Changing licensing terms, closing access to data, changing formats or API structure - all this can block critical processes. Suddenly it turns out that updating a single table requires reworking an entire block of integrations and connectors (Fig. 3.2-13), and any large-scale update to software or its API vendor becomes a potential threat to the stability of the entire company's system.

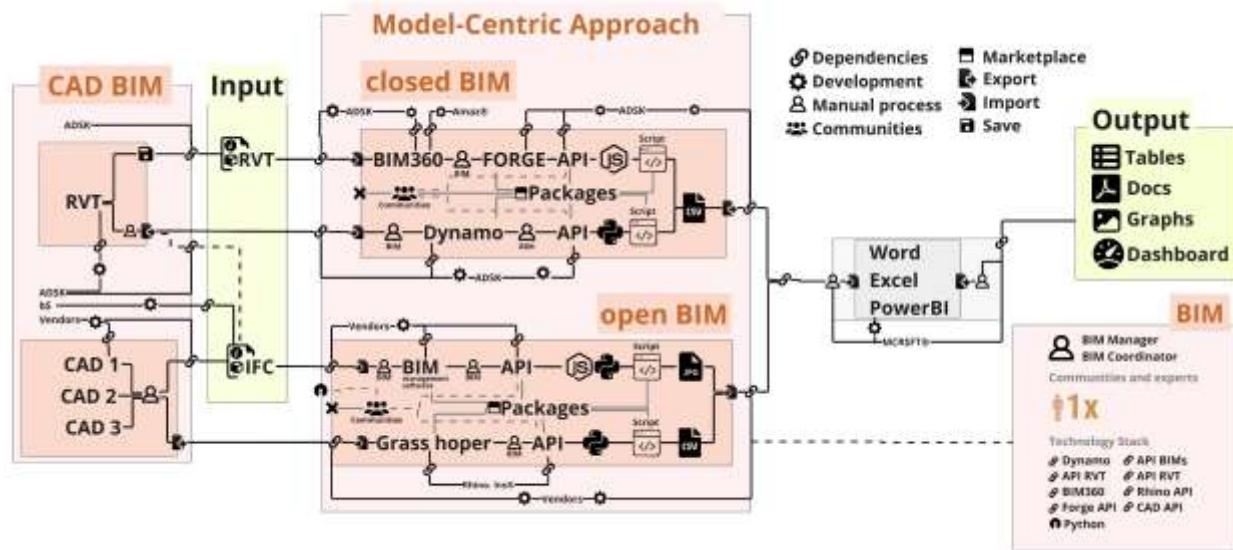


Fig. 3.2-13 An example of the large number of dependencies in CAD processing -data creates barriers to data integration in the construction company ecosystem.

Developers and system architects in such conditions are forced to work not for anticipation, but for survival. Instead of implementing new solutions, they adapt. Instead of developing, they try to maintain compatibility. Instead of automating and speeding up processes, they spend time on studying the next closed interfaces, API documentation and endless code rebuilding.

Working with closed formats and systems is not just a technical challenge - it is a strategic constraint. Despite the obvious opportunities offered by modern automation, AI, LLM and predictive analytics, many companies fail to realize their full potential. And the barriers erected by proprietary formats (Fig. 3.2-13) deny businesses access to their own data. This is perhaps the irony of digital transformation in construction.

Data transparency and open systems are not a luxury, but a prerequisite for speed and efficiency. Without openness, business processes are filled with unnecessary bureaucracy, multi-layered approval chains and a growing dependence on the HiPPO principle - making decisions based on the opinion of the highest paid person.

Nevertheless, a paradigm shift is forming on the horizon. Despite the dominance of proprietary solutions, more and more companies are realizing the limitations of Fourth Industrial Revolution-inspired architectures. Today, the vector is shifting toward the principles of the Fifth Revolution, which centers on data as a strategic asset, open interfaces (APIs), and true interoperability between systems.

This transition marks a shift away from closed ecosystems towards flexible, modular digital architectures where open formats, standards and transparent data exchange are key.

Open formats are changing the approach to digitalization

The construction industry was one of the last to address the problem of closed and proprietary data. Unlike other sectors of the economy, digitalization has been slow to develop here. The reasons for this include the traditional conservative nature of the industry, the prevalence of disparate local solutions, and the deep-rooted paper-based workflow. For decades, key construction processes relied on physical drawings, phone calls and unsynchronized databases. In this context, closed formats have long been perceived as the norm rather than an obstacle.

Experience from other industries shows that removing barriers to closed data leads to a surge in innovation, accelerated development and increased competition [73]. In science, the exchange of open data allows to accelerate discoveries and promote international cooperation. In medicine, it can improve the efficiency of diagnosis and treatment. In software engineering - to create ecosystems of co-creation and rapid product improvement.

According to the McKinsey report "Open Data: Unlock Innovation and Productivity with Information Flow" 2013. [74], open data has the potential to unlock \$3 to \$5 trillion annually across seven key industries, including construction, transportation, healthcare, and energy. According to the same study, decentralized data ecosystems enable large construction companies and contractors to reduce software development and maintenance costs, accelerating digital adoption.

The transition to open architectures, which has long started in other sectors of the economy, is gradually embracing the construction industry. Large companies and public clients, and especially financial organizations that control investments in construction projects, are increasingly demanding the use of open data and access to the source code of calculations, calculations and applications. Developers are no longer just expected to create digital solutions and show the final figures of a project - they are expected to be transparent, reproducible and independent of third-party application vendors.

Using open source solutions provides the customer with the assurance that even if external developers stop collaborating or leave the project, it will not affect the ability to further develop tools and systems. One of the main benefits of open data is its ability to eliminate the dependence of application developers on specific platforms to access data.

If a company cannot completely abandon proprietary solutions, a possible compromise is the use of reverse engineering techniques. These legal and technically sound methods allow closed formats to be transformed into more accessible, structured and suitable for integration. This is especially important when connecting to legacy systems or migrating information from one software landscape to another is required.

One of the brightest examples in the history of transition to open formats and the use of reverse engineering (legal hacking of proprietary systems) in construction is the history of the struggle to open the DWG format, widely used in computer-aided design systems (CAD). In 1998, in response to the monopoly of one software vendor, the other 15 CAD vendors formed a new alliance called "Open DWG" to provide developers with free and independent tools to work with the DWG format (the de facto standard for drawing transfer) without the need for proprietary software or closed APIs. This event was a turning point that allowed tens of thousands of companies to get free access to the closed format of a popular CAD solution from the late 1980s to today and create compatible solutions that fostered competition in the CAD market [75]. Today, the "Open DWG" SDK, which was first created back in 1996, is used in almost all solutions in which it is possible to import, edit and export DWG format, outside the official application of the DWG format developer.

Other technology giants are forcing similar transformations. Microsoft, once a symbol of proprietary approach, opened up the .NET Framework source code, started using Linux in the Azure cloud service infrastructure, and acquired GitHub to strengthen its position in the Open Source community. [76]. Meta (formerly Facebook) released open source AI models, such as the Llama series, to foster innovation and collaboration in AI agent development. CEO Mark Zuckerberg envisions that open source platforms will lead the way in technological advancements over the next decade [77].

Open Source is a software development and distribution model in which the source code is open for free use, study, modification, and distribution.

Open data and open source solutions are becoming not just a trend, but the foundation of digital sustainability. They give companies flexibility, resilience, control over their own decisions, and the ability to scale digital processes without depending on vendor policies. And, just as importantly, they give businesses back control over the most valuable resource of the 21st century - their data.

Paradigm Shift: Open Source as the End of the Era of Software Vendor Dominance

The construction industry is undergoing a shift that cannot be monetized in the usual way. The concept of data-driven, data-centric approach and the use of Open Source tools is leading to a rethinking of the rules of the game on which the software giants of the market stand.

Unlike previous technology transformations, this transition will not be actively promoted by vendors. The paradigm shift threatens their traditional business models based on licensing, subscriptions and consulting. The new reality does not involve an out-of-the-box product or a paid subscription - it requires a re-engineering of processes and thinking.

To manage and develop data center solutions based on open technologies, companies will need to

rethink internal processes. Specialists from different departments will have to not only collaborate, but also rethink how they work together.

The new paradigm implies the use of open data and Open Source solutions, where tools based on artificial intelligence and large language models (LLM) rather than programmers will play a special role in creating program code. Already by mid-2024, more than 25% of new code at Google is created with AI [78]. In the future, coding with LLMs will do 80% of the work in just 20% of the time (Fig. 3.2-14).

According to McKinsey's 2020 study [79], GPUs are increasingly replacing CPUs in analytics due to their high performance and support by modern Open Source tools. This allows companies to accelerate data processing without significant investments in expensive software or hiring scarce specialists.

Leading consulting firms such as McKinsey, PwC and Deloitte emphasize the growing importance of open standards, Open Source applications across industries.

According to the PwC Open Source Monitor 2019 report [80], 69% of companies with 100 or more employees consciously use Open Source solutions. OSS is especially actively used in large companies: 71% of companies with 200-499 employees, 78% in the 500-1999 employee category, and up to 86% among companies with more than 2000 employees. According to the Synopsys OSSRA 2023 report, 96% of the analyzed codebases contained open source components [81].

The future of the developer's role is not to manually write code, but to design data models, flow architectures, and manage AI agents that create the right calculations on demand. User interfaces will become minimalistic and interaction will become dialog-based. Classical programming will give way to high-level design and orchestration of digital solutions (Fig. 3.2-14). Current trends - such as low-code platforms (Fig. 7.4-6) and LLM-enabled ecosystems (Fig. 7.4-4) - will significantly reduce the cost of developing and maintaining IT systems.

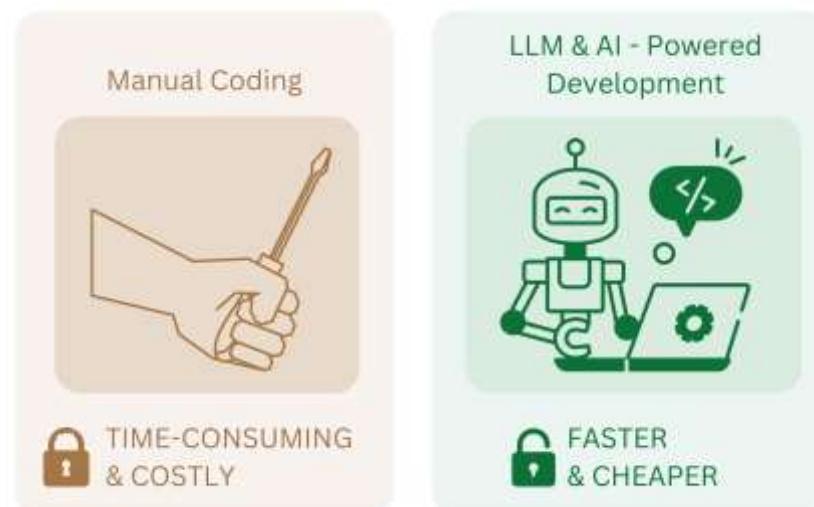


Fig. 3.2-14 While today applications are manually created by programmers, in the future a significant portion of code will be generated by AI and LLM-based solutions.

This transition will be unlike previous ones and the big software vendors will likely not catalyze it.

Harvard Business School study "The Value of Open Source Software" 2024 [40], the total value of open source software is estimated from two points of view. On the one hand, if we calculate how much it would take to build all existing Open Source solutions from scratch, this amount would be about 4.15 billion dollars. On the other hand, if we imagine that each company develops its own analogs of Open Source solutions on its own (which happens everywhere), without having access to existing tools, then the total cost of business would reach a colossal 8.8 trillion dollars - this is the cost of demand.

It's not hard to guess that no major software vendor is interested in shrinking a software market with a potential value of \$8.8 trillion to just \$4.15 billion. This would mean reducing the volume of demand by more than 2,000 times. Such a transformation is simply unprofitable for vendors whose business models are built on years of maintaining customer dependence on closed solutions. Therefore, companies expecting someone to offer them a convenient and open turnkey solution may be disappointed - such vendors simply won't show up.

The shift to an open digital architecture does not mean job or revenue losses. On the contrary, it creates the conditions for flexible and adaptive business models that may eventually displace the traditional license and boxed software market.

Instead of selling licenses - services, instead of closed formats - open platforms, instead of dependence on a vendor - independence and the ability to build solutions for real needs. Those who used to simply use tools will be able to become their co-authors. And those who can work with data, models, scenarios and logic will find themselves at the center of the industry's new digital economy. We will talk more about these changes and what new roles, business models, and collaboration formats are emerging around open data in the final, tenth part of the book.

Solutions based on open data and open code will allow companies to focus on the efficiency of business processes rather than on struggling with outdated APIs and integrating closed systems. A conscious transition to open architecture can significantly increase productivity and reduce dependence on vendors.

The transition to a new reality is not just a change in approaches to software development, but also a rethinking of the very principle of working with data. At the center of this transformation is not code, but information: its structure, accessibility and interpretability. And this is where open and structured data comes to the forefront, becoming an integral part of the new digital architecture.

Structured open data: the foundation of digital transformation

While in past decades business sustainability was largely determined by the choice of software solutions and dependence on specific vendors, in today's digital economy the key factor is data quality and the ability to work with it effectively. Open source code is an important part of the new technological paradigm, but its potential is only truly unlocked when data is understandable, organized and machine-readable. Among all types of data models, structured open data is becoming the cornerstone of sustainable digital transformation.

The main advantage of structured open data is unambiguous interpretation and the possibility of automated processing. This allows for significant efficiency gains both at the level of individual operations and across the organization.

According to Deloitte's report "The Data Transfer Process in Enterprise Transformation" [82], working with IT to manage the transfer of structured data is critical. According to the UK government report "Data Analytics and AI in Government Project Delivery" (2024) [83], removing barriers to data sharing between different projects and organizations is key to improving efficiency in project management. The document emphasizes that the standardization of data formats and the introduction of open data principles can avoid duplication of information, minimize time loss and improve the accuracy of forecasts.

For the construction industry, where traditionally a high degree of fragmentation and diversity of formats prevails, the structured-unification process and structured open data play a crucial role in shaping harmonized and manageable processes (Fig. 4.1-14). They allow project participants to focus on improving productivity rather than on solving technical problems related to incompatibilities between closed platforms, data models and formats.

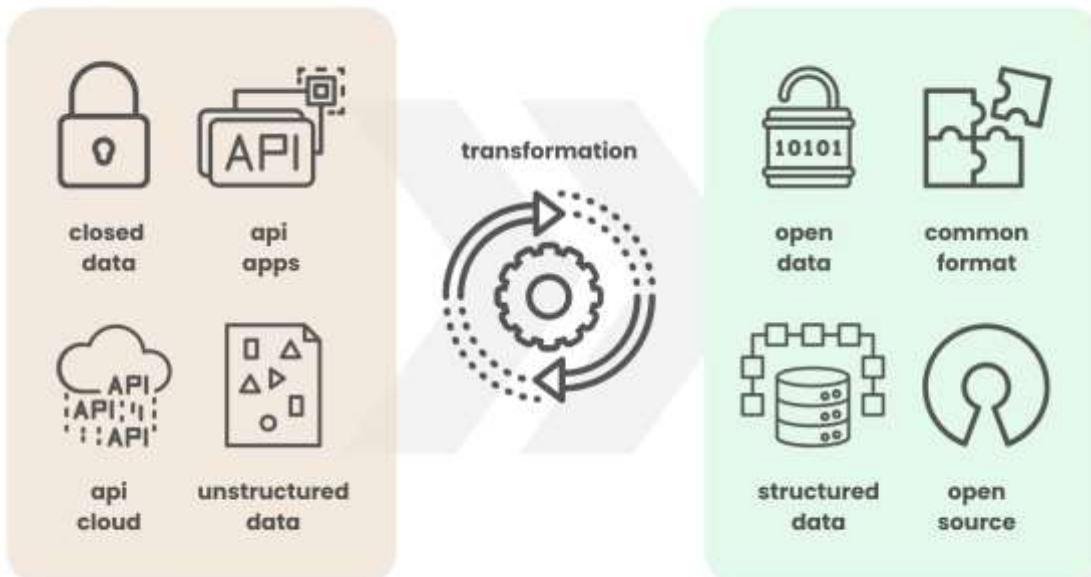


Fig. 3.2-15 Open structured data reduces dependence on software solutions and platforms and accelerates innovation.

Modern technology tools, which we will further discuss in detail in the book, allow not just collecting information, but also cleaning it automatically: eliminating duplications, correcting errors, and normalizing values. This means that analysts and engineers are not working with disparate documents, but with an organized knowledge base suitable for analysis, automation, and decision-making.

Make it as simple as you can, but no simpler.

– Albert Einstein, theoretical physicist (ownership of the quote is disputed [84])

Today, most user interfaces for working with data can be created automatically - without the need to manually write code for each business case. This requires an infrastructure layer that understands the data structure, model and logic without additional instructions (Fig. 4.1-15). It is structured data that makes this approach possible: forms, tables, filters and views can be automatically generated with minimal programming effort.

The most important user-critical interfaces may still require manual revision. But in most cases - and this is 50 to 90 percent of work scenarios - automatic generation of applications and calculations without the use of special applications for this purpose is sufficient (Fig. 3.2-16), which significantly reduces development and maintenance costs, reduces errors and speeds up the implementation of digital solutions.

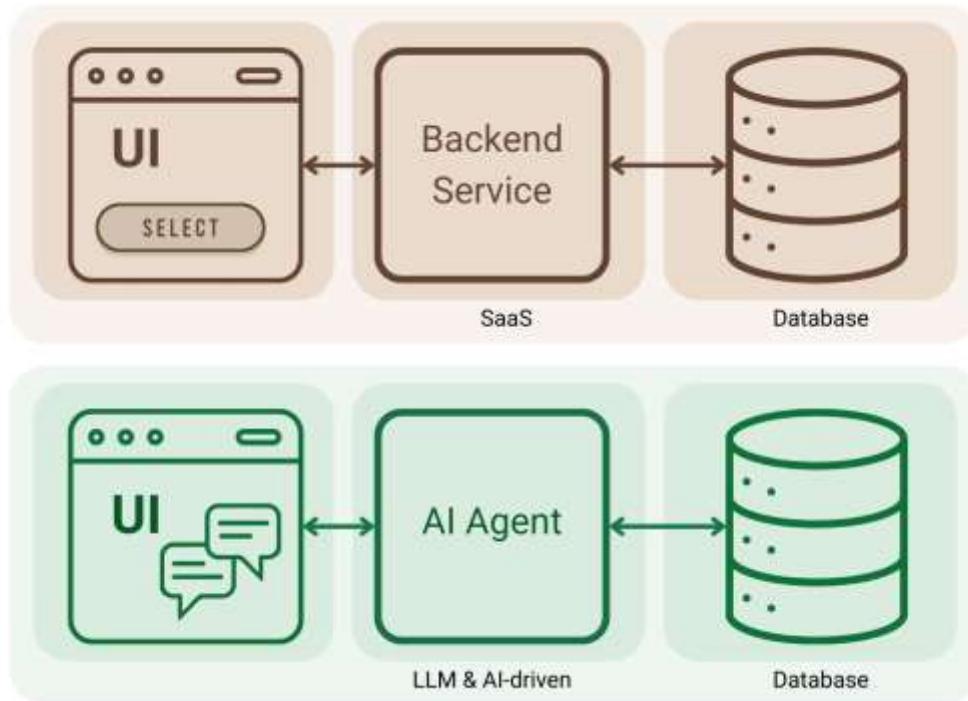
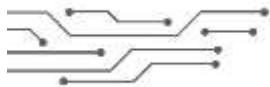


Fig. 3.2-16 Architectural models for working with data: traditional application architecture and AI-oriented model with LLM.

The transition from architectures built on individual applications to intelligently managed systems based on language models (LLMs) is the next step in digital evolution. In such an architecture, structured data becomes not only an object of storage, but also the basis for interaction with AI tools capable of analyzing, interpreting, and recommending actions based on context.

In the following chapters, we will look at real-life examples of implementing an architecture based on open structured data and show how language models are applied to automatically interpret, validate, and process data. These practical cases will help you better understand how the new digital logic works in action - and what benefits it brings to companies that are ready for transformation.



CHAPTER 3.3.

LLM AND THEIR ROLE IN DATA PROCESSING AND BUSINESS PROCESSES

LLM chat rooms: ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok for automating data processing processes

The emergence of Large Language Models (LLMs) was a natural extension of the movement towards structured open data and the Open Source philosophy. When data becomes organized, accessible and machine-readable, the next step is a tool that can interact with this information without the need to write complex code or possess specialized technical knowledge.

LLMs are a direct product of openness: large open datasets, publications, and the Open Source movement. Without open scholarly articles, publicly available textual data, and a culture of collaborative development, there would be no ChatGPT or other LLMs. The LLM is, in a sense, a "distillate" of humanity's accumulated digital knowledge, gathered and educated through the principles of openness.

Modern large language models (LLM - Large Language Models) such as ChatGPT ® (OpenAI), LLaMa™ (Meta AI), Mistral DeepSeek™, Grok™ (xAI), Claude™ (Anthropic), QWEN™ provide users with the ability to formulate queries to data in natural language. This makes working with information accessible not only to developers, but also to analysts, engineers, designers, managers, and other professionals previously distant from programming

LLM (Large Language Model) is an artificial intelligence that is trained to understand and generate text based on vast amounts of data collected from all over the internet. It is capable of analyzing context, answering questions, engaging in dialogue, writing text, and generating program code.

If earlier visualization, processing or analysis of data required knowledge of a special programming language: Python, SQL, R or Scala, as well as the ability to work with libraries like Pandas, Polars or DuckDB and many others, then starting from 2023 the situation has changed radically. Now the user can simply describe what he wants to get - and the model itself will generate the code, execute it, display a table or graph and explain the result. For the first time in decades, the development of technology has not followed the path of complication, but the path of radical simplification and accessibility.

This principle - "process data with words (prompts)" - marked a new stage in the evolution of working with information, effectively taking the creation of solutions to an even higher level of abstraction. Just as it was once no longer necessary for users to understand the technical underpinnings of the Internet

to run online stores or create websites using WordPress, Joomla, and other open source modular systems (as far as the book has been working with such systems since 2005, including educational and engineering online platforms). - This, in turn, has led to a boom in digital content and online business - today engineers, analysts, and managers can automate work processes without knowledge of programming languages. This is facilitated by powerful LLMs - both free and open source, such as LLaMA, Mistral, Qwen, DeepSeek and others - that make advanced technologies accessible to the widest possible audience.

Large Language Models LLM: how it works

Big language models (ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok) are neural networks trained on huge amounts of textual data from the Internet, books, articles and other sources. Their main task is to understand the context of human speech and generate meaningful responses.

Modern LLM is based on the Transformer architecture proposed by Google researchers in 2017 [85]. The key component of this architecture is the attention mechanism, which allows the model to consider relationships between words regardless of their position in the text.

The learning process of LLM is remotely similar to the way humans learn a language, only millions of times larger. The model analyzes billions of examples of words and expressions, identifying patterns in the structure of language and in the logic of semantic transitions. The entire text is divided into tokens - minimal semantic units (words or their parts), which are then transformed into vectors in a multidimensional space (Fig. 8.2-2). These vector representations allow the machine to "understand" the hidden relationships between concepts, rather than simply operating the text as a sequence of symbols.

Big Language Models are not just tools for generating text. They are able to recognize meaning, find connections between concepts, and work with data, even if it is presented in different formats. The main thing is that information should be broken down into understandable models and represented as tokens that the LLM can work with.

The same approach can be applied to construction projects. If we think of a project as a kind of text, where each building, element or construction is a token, we can start to process such information in a similar way. Construction projects can be compared to books that are organized into categories, chapters, and groups of paragraphs consisting of minimal tokens - elements of a construction project (Fig. 3.3-1). By translating data models into a structured format, we can also translate structured data into vector bases (Fig. 8.2-2), which are an ideal source for machine learning and technologies such as LLM.

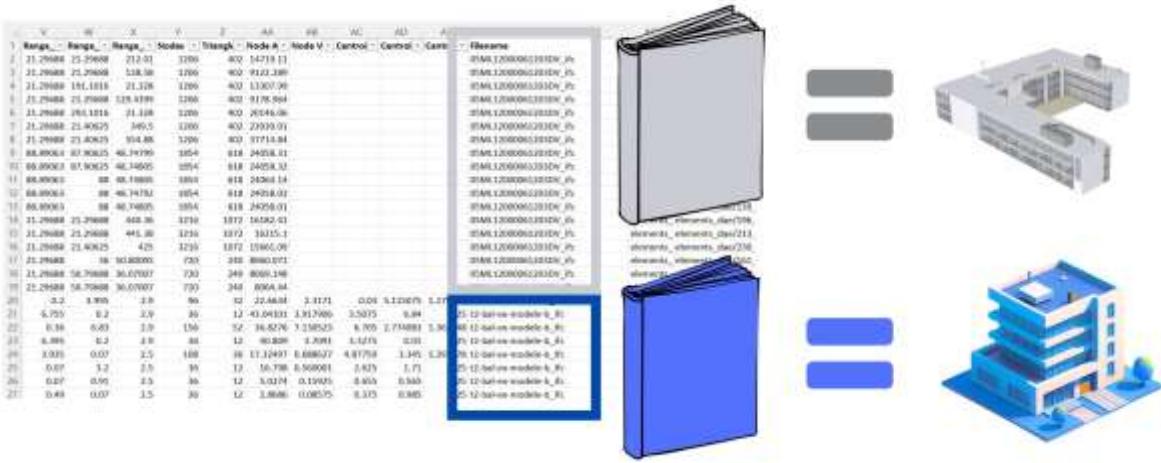


Fig. 3.3-1 A construction project element is like a token in a text: a minimum unit from which groups (paragraphs) sections (categories) of the entire project are formed.

If a construction project is digitized and its elements are represented as tokens or vectors, it becomes possible to access them not through rigid formal queries, but in natural language. This is where one of the key advantages of LLM comes into play - the ability to understand the meaning of a query and link it to the relevant data.

The engineer no longer has to write SQL -query or Python code to get the required data - he can simply, understanding the LLM and data structure, formulate the task in the usual way: *"Find all reinforced concrete structures with concrete class higher than B30 and calculate their total volume"*. The model will recognize the meaning of the query, turn it into a machine-readable form, find the data (group and transform) and return the final result.

Documents, tables, project models are converted into vector representations (embedding) and stored in the database. When a user asks a question, the query is also converted into a vector, and the system finds the closest meaningful data. This allows the LLM to rely not only on its trained knowledge, but also on actual corporate data, even if it has already appeared after the model has been trained.

One of the most important advantages of LLM in construction is the ability to generate program code. Instead of passing the technical task to a programmer, specialists can describe the task in natural language, and the model will create the necessary code, which can be used (by copying it from the chat) in the creation of process automation code. LLM -models allow specialists without deep programming knowledge to contribute to the automation and improvement of the company's business processes.

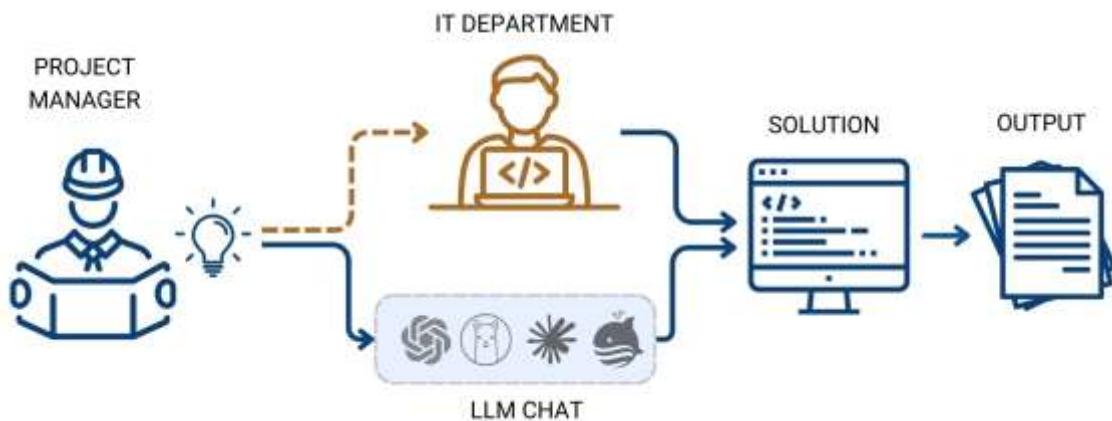


Fig. 3.3-2 LLMs provide the ability for users to write code and get results without the need for programming skills.

According to a study conducted by Wakefield Research and sponsored by SAP in 2024 [36], which surveyed 300 senior executives at companies with annual revenues of at least \$1 billion in the US: 52% of senior executives trust AI to analyze data and provide recommendations for decision making. Another 48% use AI to identify previously unaccounted-for risks, and 47% use AI to suggest alternative plans. In addition, 40% use AI for new product development, budget planning and market research. The study also showed the positive impact of AI on personal life: 39% of respondents reported improved work-life balance, 38% reported improved mental health, and 31% reported lower stress levels.

However, for all their power, LLMs remain a tool that is important to use consciously. Like any technology, they have limitations. One of the most well-known problems is so-called "hallucinations" - cases where the model confidently produces a plausible but factually incorrect answer. Therefore, it is critical to understand how the model works: what data and data models it can interpret without errors, how it interprets queries, and where it gets its information from. It is also worth remembering that the LLM's knowledge is limited to the date of its training, and without a connection to external data, the model may not take into account current norms, standards, prices, or technologies.

The solution to these problems is to regularly update vector databases, connect to relevant sources, and develop autonomous AI -agents that do not just answer questions, but proactively use data for training, manage tasks, identify risks, offer optimization options, and monitor project performance.

The transition to LLM -interfaces in construction is not just a technological novelty. It's a paradigm shift, removing barriers between people and data. It's the ability to work with information as easily as we talk to each other - and still get accurate, verified and actionable results.

Those companies that start using such tools earlier than others will gain a significant competitive advantage. This includes speeding up work, reducing costs, and improving the quality of design decisions due to quick access to data analysis and the ability to quickly find answers to complex questions.

But there are also security issues to consider. The use of cloud-based LLM -services can be associated with risks of data leakage. Therefore, organizations are increasingly looking for alternative solutions that allow them to deploy LLM tools in their own infrastructure - locally, with full protection and control over information.

Utilizing local LLMs for sensitive company data

The appearance of the first chat-LLMs in 2022 marked a new stage in the development of artificial intelligence. However, immediately after the widespread adoption of these models, a legitimate question arose: how secure is it to transfer company-related data and queries to the cloud? Most cloud-based language models stored communication history and uploaded documents on their servers and for companies dealing with sensitive information, this became a serious barrier to AI adoption.

One of the most sustainable and logical solutions to this problem has been the deployment of Open Source LLM locally, within the corporate IT infrastructure. Unlike cloud services, local models work without an Internet connection, do not transfer data to external servers and give companies full control over information

The best open model [Open Source LLM] is currently comparable in performance to closed models [such as ChatGPT, Claude], but with a lag of about one year [77].

— Ben Cottier, lead researcher at Epoch AI, a nonprofit research organization, 2024

Major technology companies have started to make their LLMs available for local use. Meta's open source LLaMA series and the rapidly growing DeepSeek project from China were examples of the move to open architecture. Alongside them, Mistral and Falcon have also released powerful models free from the constraints of proprietary platforms. These initiatives have not just accelerated the development of global AI, but have also given privacy-conscious companies real alternatives for independence, flexibility and security compliance.

In a corporate environment, especially in the construction industry, data protection is not just a matter of convenience, but of regulatory compliance. Working with tender documents, estimates, drawings and confidential correspondence requires strict controls. And this is where local LLM provides the necessary assurance that data stays inside the company's perimeter.

	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

Fig. 3.3-3 Local models provide complete control and security, while cloud-based solutions offer easy integration and automatic updates.

Key Benefits of Local Open Source LLM:

- Complete control over data. All information remains inside the company, which eliminates unauthorized access and data leakage.
- Standalone operation. No dependence on the Internet connection, which is especially important for work in isolated IT infrastructures. This also ensures uninterrupted operation in the face of sanctions or blocked cloud services.
- Application flexibility. The model can be used for text generation, data analysis, program code writing, design support and business process management.
- Adaptation to corporate objectives. LLM can be trained on internal documents, which allows

you to take into account the specifics of the company's work and its industry features. The local LLM can be connected to CRM, ERP or BI platforms, allowing you to automate the analysis of customer requests, report generation or even trend forecasting.

Deploying DeepSeek's free and open source model -R1-7B on a server, for access by an entire team of users, at a cost of \$1000 per month can potentially cost less than annual fees for cloud APIs, such as ChatGPT or Claude and allows companies to take full control of their data, eliminates its transfer to the internet and helps comply with regulatory requirements such as GDPR

In other industries, local LLMs are already changing their approach to automation. In support services, they respond to frequent customer requests, reducing the workload of operators. In HR departments, they analyze resumes and select relevant candidates. In e-commerce, they generate personalized offers without revealing user data.

A similar effect is expected in the construction industry. Thanks to the integration of LLM with project data and standards, it is possible to accelerate the preparation of documentation, automate the preparation of estimates and predictive cost analysis. The use of LLM in conjunction with structured tables and dataframes is becoming a particularly promising area.

Full control of AI in the company and how to deploy your own LLM

Modern tools allow companies to deploy a large language model (LLM) locally in just a few hours. This gives complete control over data and infrastructure, eliminating dependence on external cloud services and minimizing the risk of information leakage. This solution is especially relevant for organizations working with sensitive project documentation or confidential business data.

Depending on the tasks and resources, different deployment scenarios are available, from out-of-the-box solutions to more flexible and scalable architectures. One of the easiest tools is Ollama, which allows you to run language models literally in one click, without the need for deep technical knowledge. A quick start with Ollama:

1. Download the distribution for your operating system (Windows / Linux / macOS) from the official website: ollama.com
2. Install the model via the command line. For example, for the *Mistral* model:

```
ollama run mistral
```

3. After running the model is ready to work - you can send text queries through the terminal or integrate it into other tools. Run the model and execute a query:

```
ollama run mistral "How to create a calculation with all the resources for the work to install a 100mm wide plasterboard partition wall?"
```

For those who prefer to work in a familiar visual environment, there is LM Studio, a free application with an interface reminiscent of ChatGPT

- Install LM Studio by downloading the distribution kit from the official website - lmstudio.ai
- Through the built-in catalog, select a model (e.g. Falcon or GPT-Neo-X) and download it
- Work with the model through an intuitive interface reminiscent of ChatGPT, but completely localized

	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

Fig. 3.3-4 Comparison of popular local open source LLM -models.

The choice of model depends on the requirements for speed, accuracy and available hardware capabilities (Fig. 3.3-4). Small models such as Mistral 7B and Baichuan 7B are suitable for lightweight tasks and mobile devices, while powerful models such as DeepSeek -V3 require significant computational resources but provide high performance and support for multiple languages. In the coming years, the LLM market will grow rapidly - we will see more and more lightweight and specialized models. Instead of general-purpose LLMs covering all human content, models trained on narrow domain expertise will

emerge. For example, we can expect to see the emergence of models designed solely to work with engineering calculations, construction estimates, or CAD-formatted data. Such specialized models will be faster, more accurate and safer to use - especially in professional environments where high reliability and subject matter depth are important.

Once the local LLM has been launched, it can be adapted to the company's specific tasks. For this purpose, the fine-tuning technique is used, whereby the model is further trained on internal documents, technical instructions, contract templates or project documentation.

RAG: Intelligent LLM -assistants with access to corporate data

The next stage in the evolution of LLM application in business is the integration of models with actual real-time corporate data. This approach is called RAG (Retrieval-Augmented Generation) - Retrieval-Augmented Generation. In this architecture, the language model becomes not just a dialog interface, but a full-fledged intelligent assistant capable of navigating documents, drawings, databases and providing accurate, contextualized answers.

The main advantage of RAG is the ability to utilize internal company data without the need to pre-train the model, while maintaining high accuracy and flexibility in information handling.

RAG technology combines two main components:

- **Retrieval:** the model connects to data stores - documents, tables, PDF -files, drawings - and retrieves relevant information as requested by the user.
- **Augmented Generation:** based on the extracted data, the model generates an accurate, informed response, taking into account the context and specificity of the query.

In order to run LLM with RAG support, there are a few steps to follow:

- **Data preparation:** gather the necessary documents, drawings, specifications, tables. They can be in different formats and structures, from PDF to Excel.
- **Indexing and vectorization:** using tools such as LlamaIndex or LangChain, data is converted into vector representations that allow you to find semantic links between text fragments (more about vector databases and translating large arrays into vector representation, including CAD projects, in Part 8).
- **Query the assistant:** once the data has been uploaded, you can ask the model questions and it will search for answers within the corporate framework rather than in general knowledge gathered from the internet.

Suppose a company has a folder `constructionsite_docs`, where contracts, instructions, estimates and

tables are stored. Using a Python script (Fig. 3.3-5), we can scan this folder and build vector indexing: each document will be converted into a set of vectors reflecting the semantic content of the text. This turns the documents into a kind of "map of meanings" on which the model can efficiently navigate and find connections between terms and phrases.

For example, the model "remembers" that the words "return" and "complaint" are often found in the section of the contract concerning the shipment of materials to the construction site. Then, if a question is asked - for example, "What is our return period?" (Fig. 3.3-5 - line 11 of code) - the LLM will analyze internal documents and find accurate information, acting like an intelligent assistant capable of reading and understanding the contents of all corporate files.



```

1 from llama_index import SimpleDirectoryReader, VectorStoreIndex
2
3 # Load documents from the folder
4 documents = SimpleDirectoryReader("constructionsite_docs").load_data()
5
6 # Creating a vector index for semantic search
7 index = VectorStoreIndex.from_documents(documents)
8
9 # Integration with LLM (e.g. Llama 3)
10 query_engine = index.as_query_engine()
11 response = query_engine.query("What are the return terms in the contracts?")
12 print(response)

```

Fig. 3.3-5 LM reads a file folder - similar to the way a person opens it and searches for a desired document

The code can be run on any computer with Python installed. We'll talk more about using Python and IDEs to run the code in the next chapter.

Local deployment of LLM is not just a trend, but a strategic solution for companies that value security and flexibility. However, deploying LLM, whether on local company computers or using online solutions, is only the first step. In order to apply LLM capabilities to real-world tasks, companies must utilize tools that allow them to not only receive chat responses, but also store the logic created in the form of code that can be run outside of the context of using LLM. This is important for scaling solutions - properly organized processes make it possible to apply AI developments to several projects or even the entire company at once.

In this context, the choice of a suitable development environment (IDE) plays an important role. Modern programming tools allow not only to develop LLM-based solutions, but also to integrate them into existing business processes, turning them into automated ETL -Pipeline



CHAPTER 3.4.

IDE WITH LLM SUPPORT AND FUTURE PROGRAMMING CHANGES

Choosing an IDE: from LLM experiments to business solutions

When diving into the world of automation, data analysis, and artificial intelligence - especially when working with large language models (LLMs) - it is critical to choose the right integrated development environment (IDE). This IDE will be your main working tool: the place where the code generated by the LLM will be run, both on a local computer and within the corporate network. The choice of IDE determines not only the convenience of your work, but also how quickly you will be able to move from experimental LLM requests to full-fledged solutions embedded in real business processes.

An IDE (Integrated Development Environment) is a versatile building block on your computer for automating processes and processing data. Instead of keeping a saw, hammer, drill, and other tools separately, you have one device that can do it all - cut, fasten, drill, and even check the quality of materials. IDE for programmers is a single space where you can write code (in analogy with construction - create blueprints), test its work (building model assembly), find errors (like checking the strength of structures in construction) and run the finished project (commissioning the house).

An overview of popular IDEs:

- **PyCharm®** (JetBrains) is a powerful professional IDE for Python. It is well suited for serious projects due to the large number of built-in features. However, basic support for interactive Jupyter files (IPYNB) is only available in the paid version, and beginners may find the interface overwhelming.

A file with the IPYNB (Interactive Python Notebook) extension is a format for interactive Jupyter® Notebooks (Fig. 3.4-1) where code, visualizations, and explanations are combined in a single document. This format is ideal for building reports, analytics and training scenarios.

- **VS Code®** (Microsoft) is a fast, flexible and customizable tool with free IPYNB support and many plugins. Suitable for both beginners and professionals. Allows integration of GitHub Co-pilot and language model plugins, making it a great choice for AI and data science projects.
- **Jupyter Notebook** - A classic and popular choice for experimentation and learning. It allows you to write code, add explanations, and visualize results in a single interface (Fig. 3.4-1). Ideal for quickly testing hypotheses, working with LLM, and creating reproducible wild data analysis steps. To manage dependencies and libraries, we recommend using Anaconda Navigator, a visual interface for managing the Python environment.

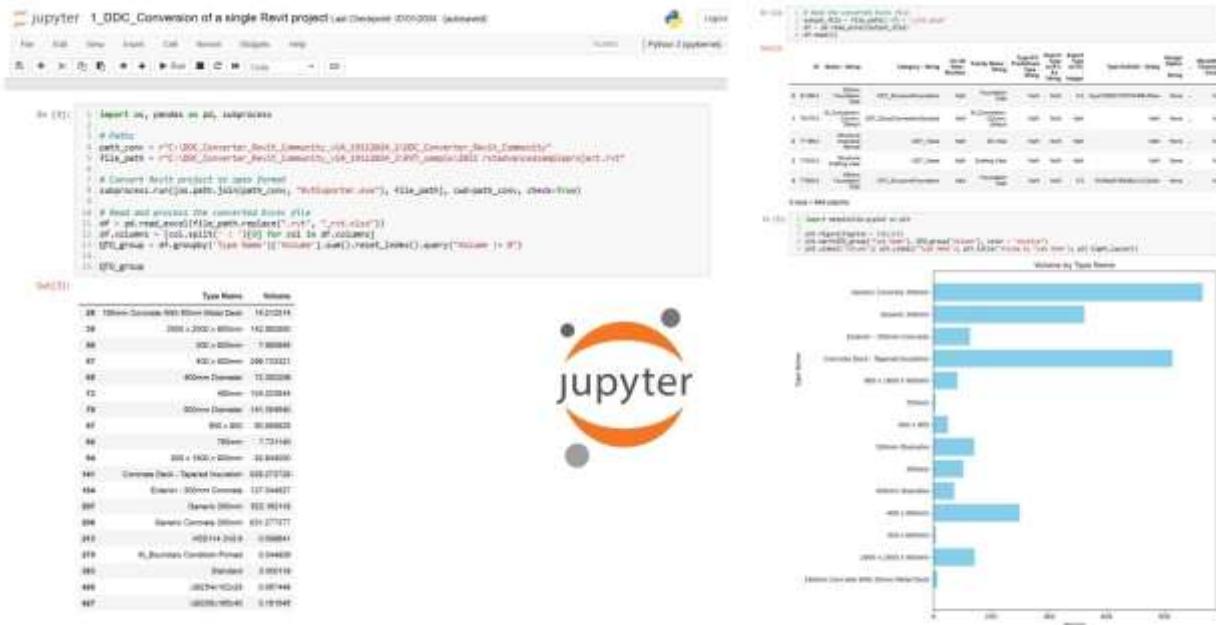


Fig. 3.4-1 Jupyter Notebook one of the most convenient and popular tools for creating Pipeline processes.

- **Google Collab™** (and the Kaggle platform (Fig. 9.2-5)) is a cloud-based alternative to Jupyter that provides free GPU/TPU access. It's a great solution for getting started - no local software installation and the ability to work directly from a browser. It supports integration with Google Drive and recently with Gemini (Google's LLM).

	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

Fig. 3.4-2 IDE Comparison: Jupyter Notebook one of the most convenient and easy tools for creating Pipeline processes.

The choice of IDE depends on your tasks. If you want to quickly start working with AI, try Jupyter Notebook or Google Collab. For serious projects it is better to use PyCharm or VS Code. The main thing is to get started. Modern tools allow you to quickly turn your experiments into working solutions.

All described IDEs allow you to create data processing pipelines - that is, chains of code block modules (which could be generated by LLM), each of which is responsible for a different stage, for example:

- analytical scenarios,
- chains of information extraction from documents,
- automatic responses based on RAG,
- generation of reports and visualizations.

Thanks to the modular structure, each step can be represented as a separate block: data loading→filtering→analysis→visualization→exporting results. These blocks can be reused, - adapted and assembled into new chains, like a constructor, just for data.

For engineers, managers and analysts, this opens up the possibility of documenting decision-making logic in the form of code that can be generated with LLM. This approach helps to speed up routine tasks, automate typical operations, and create repeatable processes where every step is clearly documented and transparent to all team members.

The automated ETL Pipelines (Fig. 7.2-3), Apache Airflow (Fig. 7.4-4), Apache NiFi (Fig. 7.4-5) and n8n (Fig. 7.4-6) tools for building blocks of logic for process automation will be discussed in more detail in Part 7 and Part 8 of the book.

IDE with LLM support and future programming changes

The integration of artificial intelligence into development processes is changing the programming landscape. Modern environments are no longer just text editors with syntax highlighting - they are turning into intelligent assistants capable of understanding project logic, completing code, and even explaining how a particular code fragment works. Products are appearing on the market that use AI to push the boundaries of conventional development:

- **GitHub Copilot** (integrates with VS Code, PyCharm): AI is an assistant that generates code based on comments or partial descriptions, turning textual hints into finished solutions.
- **Cursor** (a fork of VS Code with AI -kernel): allows not only to finish writing code, but also to ask questions to the project, look for dependencies and learn from the code base.
- **JetBrains AI Assistant**: a plugin for JetBrains IDE (including PyCharm) with the function of explaining complex code, optimization and test creation.
- **Amazon CodeWhisperer**: an analog of Copilot with a focus on security and support for Amazon's AWS services.

Programming will undergo a dramatic change in the coming years. The main focus will shift from routine code writing to model and data architecture design - developers will be more involved in system design, while AI will take over template tasks: code generation, tests, documentation and basic functions. The future of programming is a collaboration between humans and AI, where machines take over the technical routine and humans focus on creativity.

Natural language programming will become an everyday occurrence. IDE personalization will reach a new level - development environments will learn to adapt to the user's work style, and companies, anticipating patterns, offering contextual solutions and learning from previous projects

This does not abolish the developer's role, but radically transforms it: from writing code to managing knowledge, quality, and processes. This evolution will also affect business intelligence, where the creation of reports, visualizations, and decision support applications will increasingly take place through the generation of code and logic with the help of AI and LLM, chat and agent interfaces.

Once a company has set up LLM chats and selected a suitable development environment, the next important step is data organization. This process involves extracting information from disparate sources, cleaning it, transforming it into a structured form and integrating it into corporate systems.

In a modern Data-Centric approach to data management, a key goal is to bring data into a single universal form that is compatible with a large number of tools and applications. Specialized libraries are needed to handle structuring processes and structured data. One of the most powerful, flexible and popular is the Pandas library for Python. It allows you to conveniently process tabular data: filter, group, clean, append, perform aggregations and build reports.

Python Pandas: an indispensable tool for working with data

Pandas occupies a special place in the world of data analysis and automation. It is one of the most popular and widely used libraries of the Python programming language [86], designed to work with structured data.

A library is like a set of ready-made tools: functions, modules, classes. Just as on a construction site you don't need to invent a hammer or a level every time, so in programming libraries allow you to quickly solve problems without reinventing basic functions and solutions.

Pandas is an open source Python library , providing high-performance and intuitive data structures, in particular DataFrame, a universal format for working with tables. Pandas is a Swiss knife for data-driven analysts, engineers, and developers.

Python is a high-level programming language with a simple syntax that is actively used in analytics, automation, machine learning, and web development. Its popularity is due to its code readability, cross-platform nature, and rich ecosystem of libraries. To date, more than 137,000 open source packages have been created for Python [87], and this number continues to grow almost daily. Each such library is a kind of repository of ready-made functions: from simple mathematical operations to complex tools for image processing, big data analysis, neural networks, and integration with external services.

In other words, imagine that you have free and open access to hundreds of thousands of out-of-the-box software solutions - libraries and tools that you can directly embed into your business processes. It's like a huge catalog of applications for automation, analysis, visualization, integration, and more - and it's all available right after you install Python.

Pandas is one of the most popular packages in the Python ecosystem. In 2022, the average number of downloads of the Pandas library reached 4 million per day (Fig. 3.4-3), whereas by early 2025, this Fig. has increased to 12 million downloads per day, reflecting its growing popularity and widespread use in data analytics and LLM chat [86]

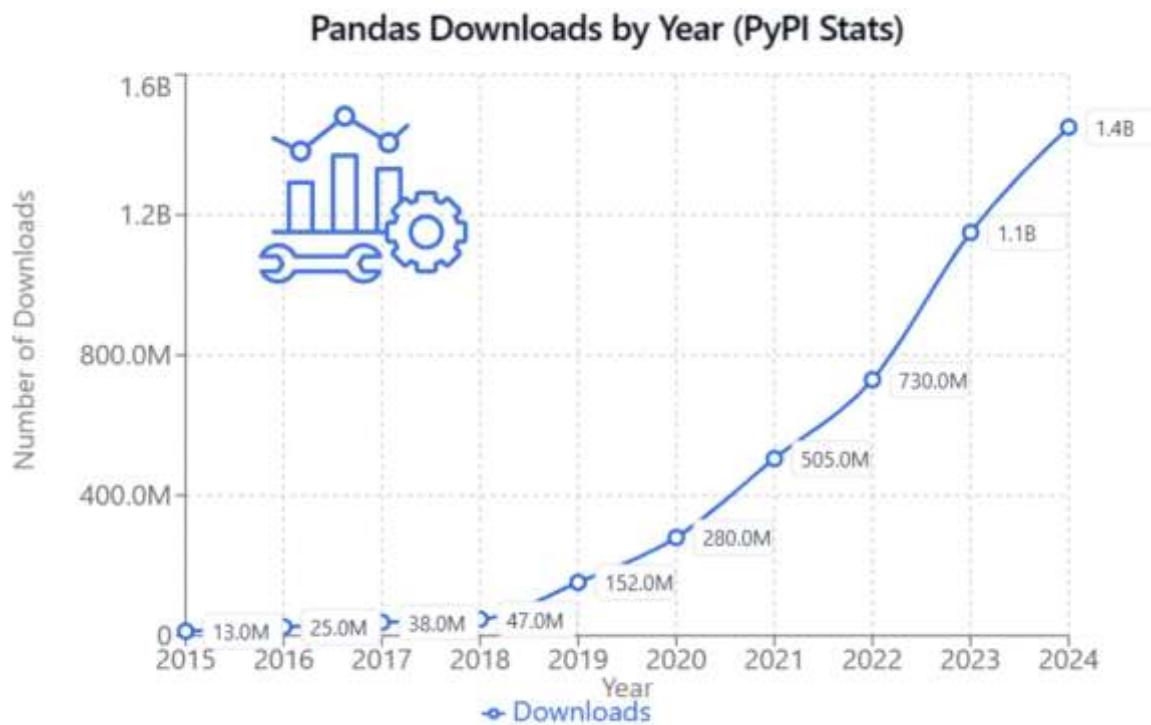


Fig. 3.4-3 Pandas is one of the most downloaded libraries. In 2024, its annual number of downloads exceeded 1.4 billion.

The query language in the Pandas library is similar in functionality to the SQL query language, which we discussed in the chapter "Relational Databases and SQL Query Language".

In the world of analytics and structured data management, Pandas stands out for its simplicity, speed and power, providing users with a wide range of tools to effectively analyze and process information.

Both tools - SQL and Pandas - provide powerful data manipulation capabilities, especially when compared to traditional Excel. They support operations such as sampling, filtering (Fig. 3.4-4), with the only difference being that SQL is optimized for working with relational databases, while Pandas processes data in RAM, which allows it to run on any computer, without the need to create databases and deploy a separate infrastructure.

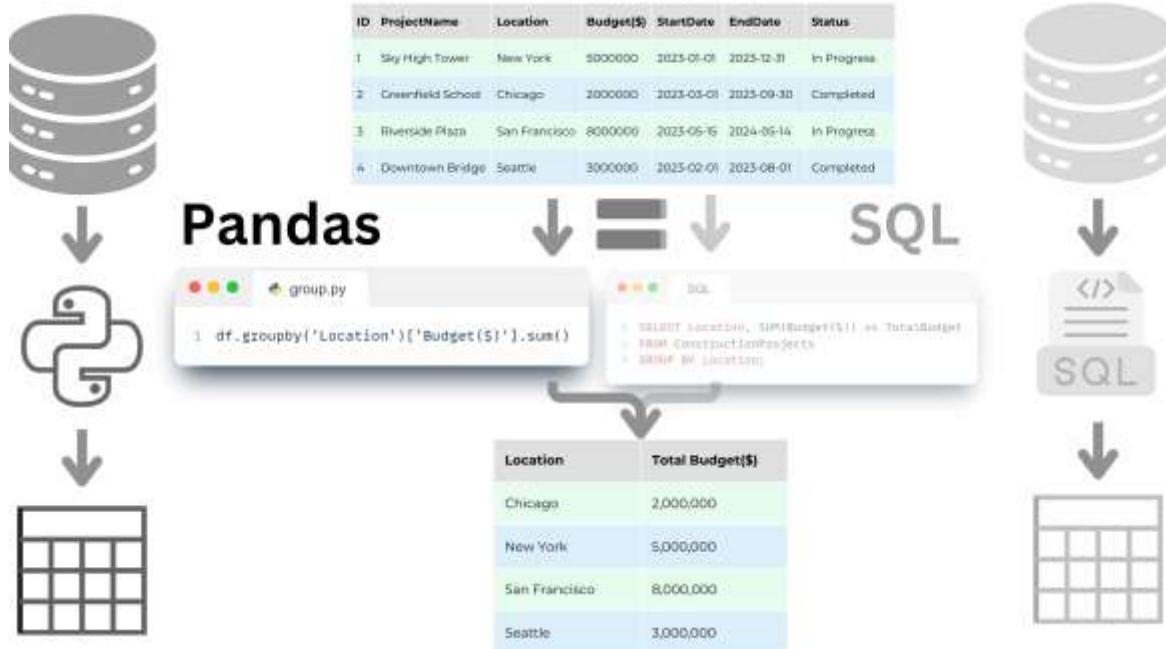


Fig. 3.4-4 Pandas, unlike SQL, has the flexibility to work with a variety of data formats, not limited to databases.

Pandas is often preferred for scientific research, process automation, pipeline creation (including ETL) and data manipulation in Python, while SQL is a database management standard and is often used in enterprise environments to handle large amounts of data.

The Pandas library of the Python programming language allows you to perform not only basic operations such as reading and writing tables, but also more complex tasks, including merging data, grouping data, and performing complex analytical calculations.

Today, the Pandas library is used not only in academic research and business analytics, but also in conjunction with LLM -models. For example, Meta® division (Facebook™), when publishing a new open source model LLaMa 3.1 in 2024, paid special attention to working with structured data, making one of the key and first cases in its release the processing of structured dataframes (Fig. 3.4-5) in CSV format and integration with Pandas library directly in chat.

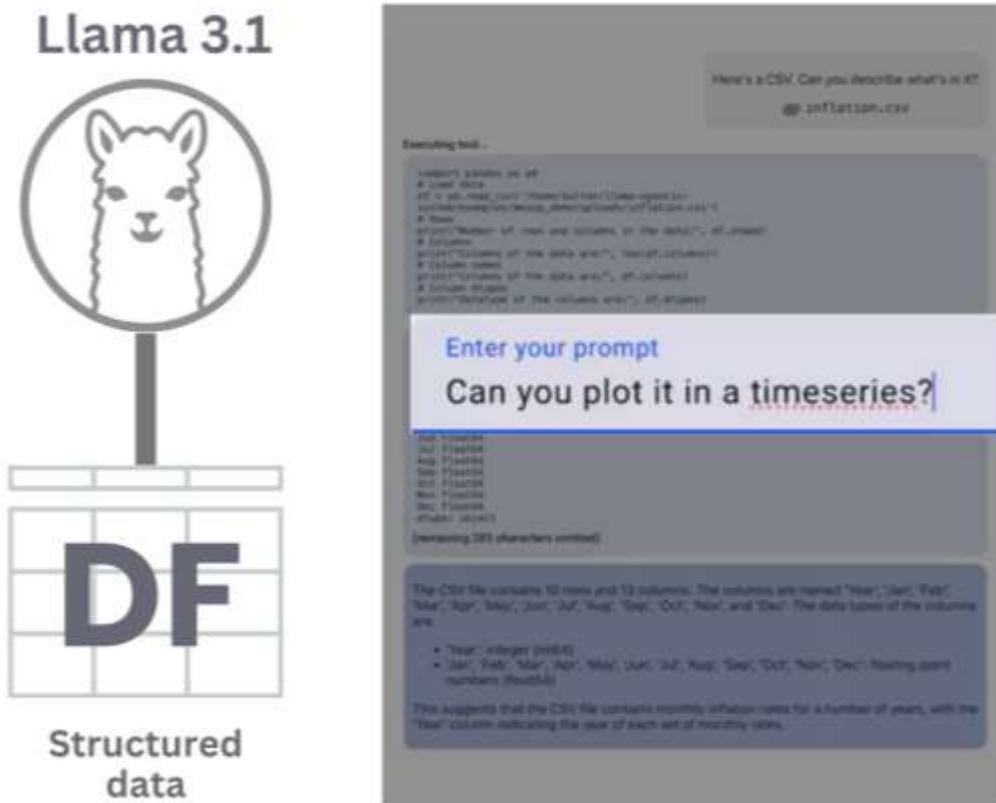


Fig. 3.4-5 One of the Meta team's first and main cases presented in LLaMa 3.1 in 2024 was building applications using Pandas.

Pandas is an essential tool for millions of data scientists processing and preparing data for generative AI. Accelerating Pandas with zero code changes will be a huge step forward. Data scientists will be able to process data in minutes instead of hours and get orders of magnitude more data to train generative AI models [88].

— Jensen Huang, founder and CEO of NVIDIA

Using Pandas, you can manage and analyze datasets far beyond the capabilities of Excel. While Excel is typically capable of handling up to 1 million rows of data, Pandas can easily handle datasets (Fig. 9.1-2, Fig. 9.1-10) containing tens of millions of rows [89]. This capability allows users to perform sophisticated data analysis and visualization on large datasets, providing deep insights and facilitating data-driven decision making. In addition, Pandas has strong community support [90]: hundreds of millions of developers and analysts worldwide (Kaggle.com, Google Collab, Microsoft® Azure™ Notebooks, Amazon SageMaker) use it online or offline every day, providing a large number of out-of-the-box solutions for any business problem.

At the heart of most Python analytic processes is a structured form of data called DataFrame, provided

by the Pandas library. It is a powerful and flexible tool for organizing, analyzing, and visualizing tabular data.

DataFrame: universal tabular data format

DataFrame is the central structure in the Pandas library, which is a two-dimensional table (Fig. 3.4-6) where rows correspond to individual objects or records and columns correspond to their characteristics, parameters, or categories. This structure visually resembles Excel spreadsheets, but is far superior in terms of flexibility, scalability, and functionality.

A **DataFrame** is a way to represent and process tabular data stored in the computer's RAM.

DataFrame is a way of representing and processing tabular data stored in the computer's RAM. In a table, rows can reflect, for example, elements of a construction project, and columns - their properties: categories, dimensions, coordinates, cost, terms and so on. Moreover, such a table can contain both information on one project (Fig. 4.1-13) and data on millions of objects from thousands of different projects (Fig. 9.1-10). Thanks to vectorized Pandas operations, it is easy to filter, group and aggregate such volumes of information at high speed.

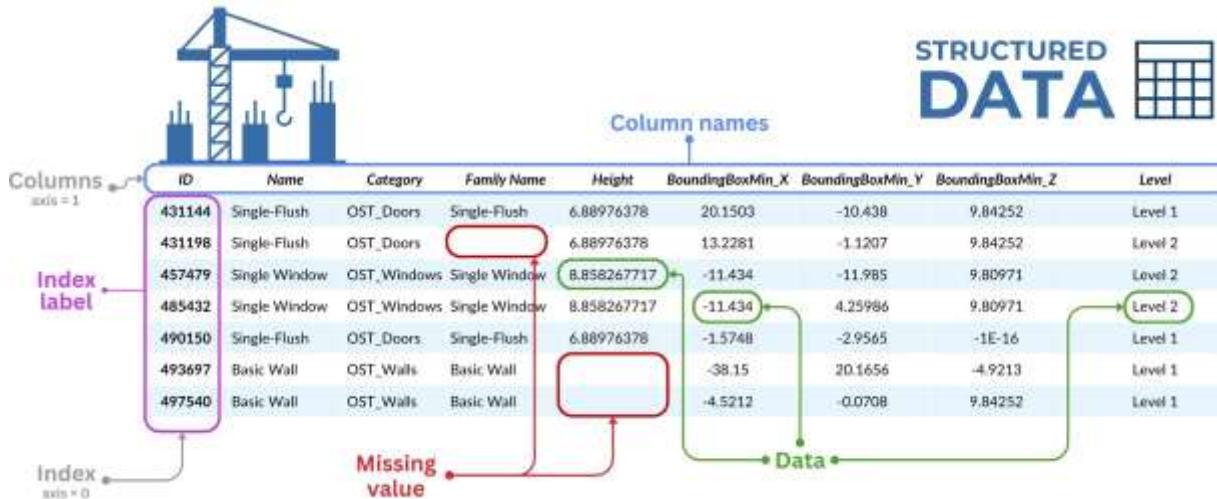


Fig. 3.4-6 Construction Project as a DataFrame is a two-dimensional table with elements in rows and attributes in columns.

Nvidia estimates that already today up to 30% of all computing resources are used to process structured data - dataframes, and this share continues to grow.

Data processing is what probably one third of the world's computing is done in every company. The data processing and data of most companies are in DataFrame, in table format

— Jensen Huang, CEO of Nvidia [91]

Let's list some key features of DataFrame in Pandas:

- **Columns:** in DataFrame, data is organized into columns, each with a unique name. Attribute columns can contain data of different types, similar to columns in databases or columns in tables.
- **Pandas Series** is a one-dimensional data structure in Pandas, similar to a list or column in a table, where each value corresponds to a different index

Pandas Series has over 400 attributes and methods, making working with data incredibly flexible. You can directly apply one of the four hundred available functions to a column, perform math operations, filter data, replace values, work with dates, strings, and more. In addition, Series supports vectorized operations, which greatly speeds up the processing of large datasets compared to cyclic calculations. For example, you can easily multiply all values by a number, replace missing data, or apply complex transformations without writing complicated loops.

- **Rows:** in DataFrame can be indexed with unique values. This index allows you to quickly change and adjust the data in specific rows.
- **Index:** By default, when you create a DataFrame Pandas assigns each row an index from 0 to N-1 (where N is the number of all rows in the DataFrame). However, the index can be changed to include special designations such as dates or unique characteristics.
- **Indexing** rows in a DataFrame means that each row is assigned a unique name or label, which is called the DataFrame index.
- **Data Types:** DataFrame supports a variety of data types, including: `int`, `float`, `bool`, `datetime64` and `object` for text data. Each DataFrame column has its own data type that determines what operations can be performed on its contents.
- **Data operations:** DataFrame supports a wide range of operations for data processing, including aggregation (`groupby`), merge (`merge` and `join`), concatenation (`concat`), split-apply-combine, and many other data transformation techniques.
- **Size Manipulation:** DataFrame allows you to add and remove columns and rows, making it a dynamic structure that can be modified according to your data analysis needs.
- **Visualizing data:** using built-in visualization techniques or interacting with popular data visualization libraries such as Matplotlib or Seaborn, DataFrame can be easily converted to graphs and charts to present data graphically.

- **Data input and output:** Pandas provides functions to read import and export data to various file formats such as CSV, Excel, JSON, HTML and SQL, potentially making DataFrame a central hub for data collection and distribution.

Unlike CSV and XLSX, Pandas DataFrame provides greater flexibility and performance when working with data: it can handle large amounts of information in RAM, supports extended data types (including dates, logical values, and time series), and provides extensive capabilities for filtering, aggregating, merging, and visualizing data. While CSV does not store information about data types and structure, and XLSX is often overloaded with formatting and has low scalability, DataFrame remains the optimal choice for rapid analytics, process automation, and integration with AI -models (Fig. 3.4-7). In the following chapters we will explore each of these aspects of data in detail, also in Part 8 of the book, similar formats such as Parquet, Apache Orc, JSON, Feather, HDF5 and data warehouses will be discussed in detail (Fig. 8.1-2).

	XLSX	CSV	Pandas DataFrame
	Storage	Tabular	Tabular
	Usage	Office tasks, data presentation	Simple data exchange
	Compression	Built-in	None
	Performance	Low	Medium
	Complexity	High (formatting, styles)	Low
	Data Type Support	Limited	Very limited
	Scalability	Low	Medium (memory limited)

Fig. 3.4-7 DataFrame is the optimal choice of data manipulation with high performance and advanced data type support.

Because of their flexibility, power, and ease of use, the Pandas library and DataFrame format have become the de facto standard in Python data analysis. They are ideal for both creating simple reports and building complex analytic pipelines, especially in conjunction with LLM models.

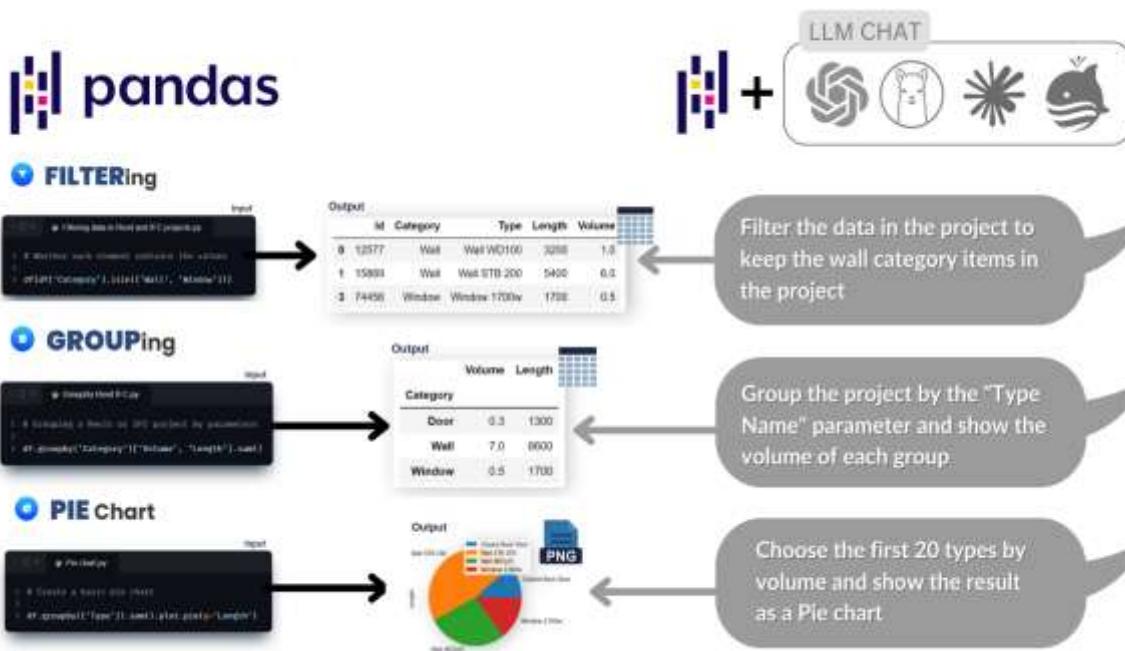


Fig. 3.4-8 LLMs simplify interaction with Pandas: a text query is sufficient instead of code.

Today Pandas is actively used in LLM-based chat rooms such as ChatGPT, LLaMa, DeepSeek, QWEN and others. In many cases, when a model receives a query related to table processing, data validation or analytics, it generates code exactly using the Pandas library. This makes DataFrame a natural "language" for representing data in AI dialogs (Fig. 3.4-8).

Modern data technologies such as Pandas make it easier to analyze, automate and integrate data into business processes. They deliver results quickly, reduce the workload of specialists, and ensure repeatable operations.

Next steps: building a sustainable data framework

In this part, we reviewed the key types of data used in the construction industry, got acquainted with different formats of their storage and analyzed the role of modern tools, including LLM and IDEs, in information processing. We have seen that effective data management is the foundation for informed decision making and business process automation. Organizations that are able to structure and organize their data gain a significant competitive advantage in the stages of data processing and transformation.

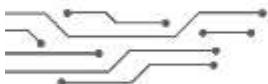
To summarize this part, it is worth highlighting the main practical steps that will help you apply the discussed approaches in your daily tasks:

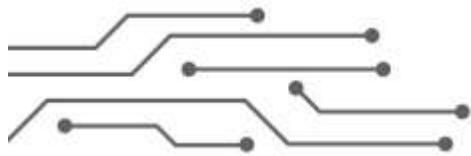
- Audit the data in your processes
- Make an inventory of all the data types you use in your projects

- Determine which data types and models are most critical to your business processes
- Identify problem areas where information remains often unstructured, poorly structured or inaccessible
- Start forming a data management strategy**
 - Raise policy issues and standards for dealing with different data types
 - Analyze which of your workflows can be improved by converting unstructured data into structured data
 - Create a data storage and access policy that takes security and privacy into account
- Install and master basic tools for working with data**
 - Choose a suitable IDE that matches your tasks (e.g. install VS Code or Jupyter Notebook)
 - Try setting up a local LLM to handle your personal data confidentially
 - Start experimenting with the Pandas library to process XLSX tabular data
 - Describe to the LLM the typical tasks you handle in spreadsheet tools or databases and ask the LLM to automate the work using Pandas

Applying these steps will allow you to gradually transform your approach to working with data, moving from disparate, unstructured data sets to a unified ecosystem where data becomes an accessible and understandable asset. Start small - create your first DataFrame in Pandas, start a local LLM, automate your first routine task using Python (e.g. Excel spreadsheets).

The fourth part of the book will focus on data quality, data organization, structuring, and modeling. We will focus on methodologies that transform disparate data sources - from PDFs and texts to images and CAD models - into structured data sets suitable for analysis and automation. We will also explore how data requirements are formalized, how conceptual and logical models are built in construction projects, and how modern language models (LLMs) can help in this process.





IV PART

DATA QUALITY: ORGANIZATION, STRUCTURING, MODELING

The fourth part focuses on methodologies and technologies for transforming disparate information into structured data sets of high quality. The processes of forming and documenting data requirements as a basis for effective information architecture in construction projects are discussed in detail. Practical methods of extracting structured information from various sources (PDF -documents, images, text files, CAD -models) with examples of implementation are presented. The use of regular expressions (RegEx) and other tools for automatic validation and verification of data is analyzed. The process of data modeling at conceptual, logical and physical levels is described step by step, taking into account the specifics of the construction industry. Specific examples of using language models (LLM) to automate the processes of structuring and validation of information are demonstrated. Effective approaches to the visualization of analysis results are proposed, increasing the availability of analytical information for all levels of construction project management

CHAPTER 4.1.

DATA CONVERSION INTO A STRUCTURED FORM

In the era of the data-driven economy, data is becoming the basis for decision-making rather than an obstacle. Instead of constantly adapting information to each new system and its formats, companies are increasingly striving to form a single structured data model that serves as a universal source of truth for all processes. Modern information systems are designed not around formats and interfaces, but around the meaning of data - because the structure may change, but the meaning of information remains the same for much longer.

The key to working effectively with data lies not in its endless conversion and transformation, but in organizing it correctly from the start: creating a universal structure capable of providing transparency, automation and integration at all stages of the project lifecycle.

The traditional approach forces manual adjustments with each new platform implementation: migrating data, changing attribute names, and adjusting formats. These steps do not improve the quality of the data themselves, but only mask problems, creating a vicious cycle of endless transformations. As a result, companies become dependent on specific software solutions, and digital transformation slows down.

In the following chapters, we will look at how to structure data properly and then how to create universal models, minimize platform dependency, and focus on what matters most - data as a strategic resource around which sustainable processes are built.

Learning how to turn documents, PDF, pictures and texts into structured formats

In construction projects, the vast majority of information exists in unstructured form: technical documents, statements of work, drawings, specifications, schedules, and protocols. Their diversity - both in format and content - complicates integration and automation.

The conversion process to structured or semi-structured formats may vary depending on the type of input data and the desired processing results.

Transforming data from unstructured to structured form is both an art and a science. This process varies depending on the type of input data and the purpose of the analysis and often takes up a significant portion of the work of the data engineer (Fig. 3.2-5) and analyst, with the goal of producing a clean, organized data set.



Fig. 4.1-1 Converting an unstructured scanned document into a structured tabular format.

Turning documents, PDF, pictures, and texts into a structured format (Fig. 4.1-1) is a step-by-step process that includes the following steps:

- Extract): In this step, a source document or image containing unstructured data is loaded. This can be, for example, a PDF -document, a photo, a drawing or a schematic.
- **Data conversion (Transform):** This is followed by the step of converting unstructured data into a structured format. For example, this may involve recognizing and interpreting text from images using optical character recognition (OCR) or other processing methods.
- **Loading and saving data (Load):** the last step involves saving the processed data in various formats such as CSV, XLSX, XML, JSON, for further work, where the choice of format depends on specific requirements and preferences.

This process, known as ETL (Extract, Transform, Load), plays a key role in automated data processing and will be discussed in more detail in the chapter "ETL and Pipeline: Extract, Transform, Load". Next, we will look at examples of how documents of different formats are transformed into structured data.

Example of converting a PDF -document into a table

One of the most common tasks in construction projects is to process specifications in PDF format. To demonstrate the transition from unstructured data to a structured format, let's consider a practical example: extracting a table from a PDF document and converting it to CSV or Excel format (Fig. 4.1-2).

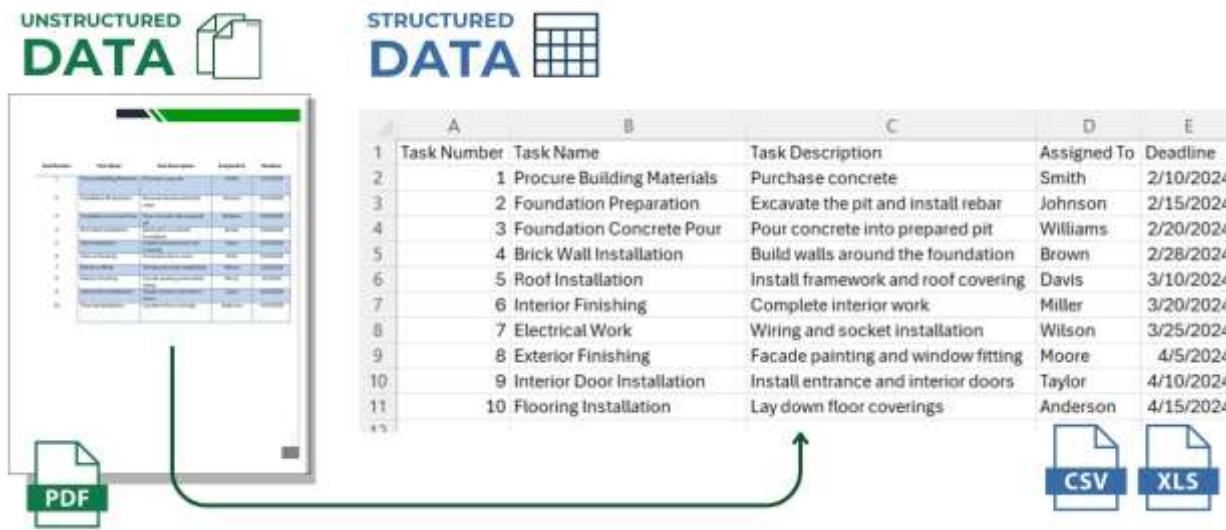


Fig. 4.1-2 Unlike PDF, the CSV and XLSX formats are widely available and easily integrated into various data management systems.

LLM language models, such as ChatGPT, LLaMa, Mistral DeepSeek, Grok, Claude, QWEN greatly simplify the way data scientists work with data, reducing the need for deep learning of programming languages and allowing many tasks to be solved with text queries.

Therefore, instead of spending time searching for solutions on the Internet (usually the StackOverFlow website or thematic forums and chats) or contacting data processing specialists, we can use the capabilities of modern online or local LLMs. It is enough to ask a query and the model will provide ready code for converting a PDF -document into a tabular format.

- Send the following text request to any LLM -model (CHATGRT, LLaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

Please write a code to extract text from a PDF -file that contains a table. The code should take the file path as an argument and return the extracted table as a DataFrame.

- The LLM -model response will in most cases be in the form of code in Python, as this language is widely used for data processing, automation and working with various file formats:

```

1 import fitz # PyMuPDF
2 import pandas as pd
3 import re
4
5 # Define the path to the PDF file
6 pdf_path = r'C\Task Number.pdf'
7
8 # Step 1: Extracting text from the PDF
9 def extract_text(pdf_path):
10     """
11         Extract text from a PDF file.
12         Args: pdf_path (str): Path to the PDF file.
13         Returns: str: All text extracted from the PDF.
14     """
15     document = fitz.open(pdf_path)
16     text = ''
17     for page in document:
18         text += page.get_text()
19     return text
20
21 extracted_text = extract_text(pdf_path)

```

Run in IDE

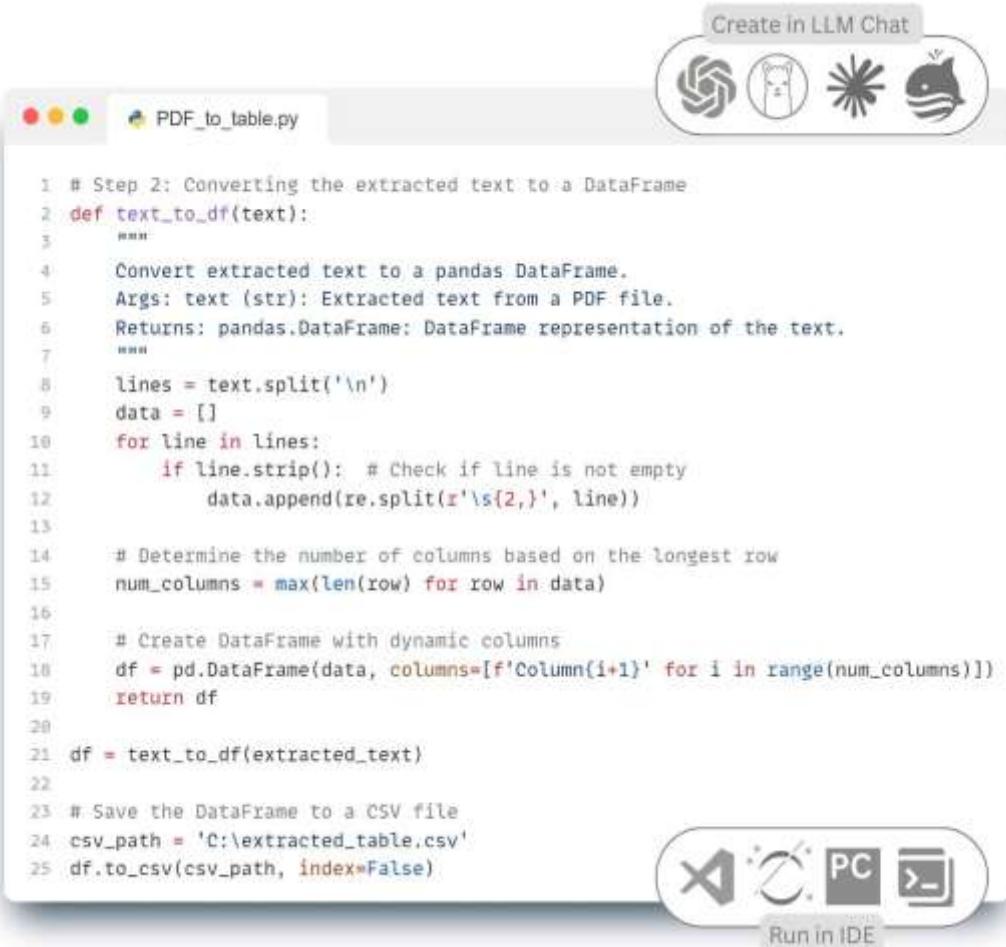
Fig. 4.1-3 The LLM response in the form of Python code and its libraries and packages (Pandas, Fitz) extracts text from a PDF -file.

This code (Fig. 4.1-3) can be run offline in one of the popular IDEs we mentioned above: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse with PyDev plugin, Thonny, Wing IDE, IntelliJ IDEA with Python plugin, JupyterLab or popular online tools: Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

- In the "Convert" step, we use the popular Pandas library (which we discussed in detail in the chapter "Python Pandas: an indispensable tool for working with data") to read the extracted text into the DataFrame and save the DataFrame to a CSV table file or XLSX:

I need code that will convert the resulting table from a PDF -file to a DataFrame. Also add code to save the DataFrame to a CSV file ↴

✍ LLM's response:



The screenshot shows a code editor window titled "PDF_to_table.py". The code is as follows:

```

1 # Step 2: Converting the extracted text to a DataFrame
2 def text_to_df(text):
3     """
4         Convert extracted text to a pandas DataFrame.
5         Args: text (str): Extracted text from a PDF file.
6         Returns: pandas.DataFrame: DataFrame representation of the text.
7     """
8     lines = text.split('\n')
9     data = []
10    for line in lines:
11        if line.strip(): # Check if line is not empty
12            data.append(re.split(r'\s{2,}', line))
13
14    # Determine the number of columns based on the longest row
15    num_columns = max(len(row) for row in data)
16
17    # Create DataFrame with dynamic columns
18    df = pd.DataFrame(data, columns=[f'Column{i+1}' for i in range(num_columns)])
19    return df
20
21 df = text_to_df(extracted_text)
22
23 # Save the DataFrame to a CSV file
24 csv_path = 'C:\extracted_table.csv'
25 df.to_csv(csv_path, index=False)

```

At the bottom right of the code editor, there are icons for "Run in IDE" and other toolbars.

Fig. 4.1-4 Converting the extracted table from PDF to DataFrame and saving the table to a CSV -file.

If an error occurs when executing the code (Fig. 4.1-3, Fig. 4.1-4) - for example, due to missing libraries or wrong file path - the error text can simply be copied together with the source code and resubmitted to the LLM -model. The model will analyze the error message, explain what the problem is and suggest fixes or additional steps.

Thus, interaction with the AI LLM becomes a complete cycle: request → response → test → feedback → correction - without the need for deep technical knowledge.

Using a plain text query in LLM chat and a dozen lines of Python that we can run locally in any IDE, we

converted a PDF -document into a tabular CSV format, which, unlike a PDF document, is easily machine readable and quickly integrated into any data management system.

We can apply this code (Fig. 4.1-3, Fig. 4.1-4), by copying it from any LLM chat room, to tens or thousands of new PDF documents on the server, thereby automating the process of converting a stream of unstructured documents into a structured CSV table format.

But PDF documents do not always contain text, more often than not they are scanned documents that need to be processed as images. Although images are inherently unstructured, the development and application of recognition libraries allow us to extract, process and analyze their content, enabling us to make full use of this data in business processes.

Convert JPEG, PNG image to structured form

Images are one of the most common forms of unstructured data. In construction and many other industries, a huge amount of information is stored in the form of scanned documents, schematics, photographs and drawings. Such data contains valuable information but cannot be directly processed, such as an Excel spreadsheet or a database. Images contain a lot of complex information because their content, colors, textures are varied and special processing is required to extract useful information.

The difficulty in using images as a data source is the lack of structure. Images do not convey meaning in a direct, easily quantifiable way that a computer can immediately understand or process, as an Excel spreadsheet or a database table does. To convert unstructured image data into a structured form, it is necessary to use special libraries that can interpret the visual information they contain (Fig. 4.1-5).

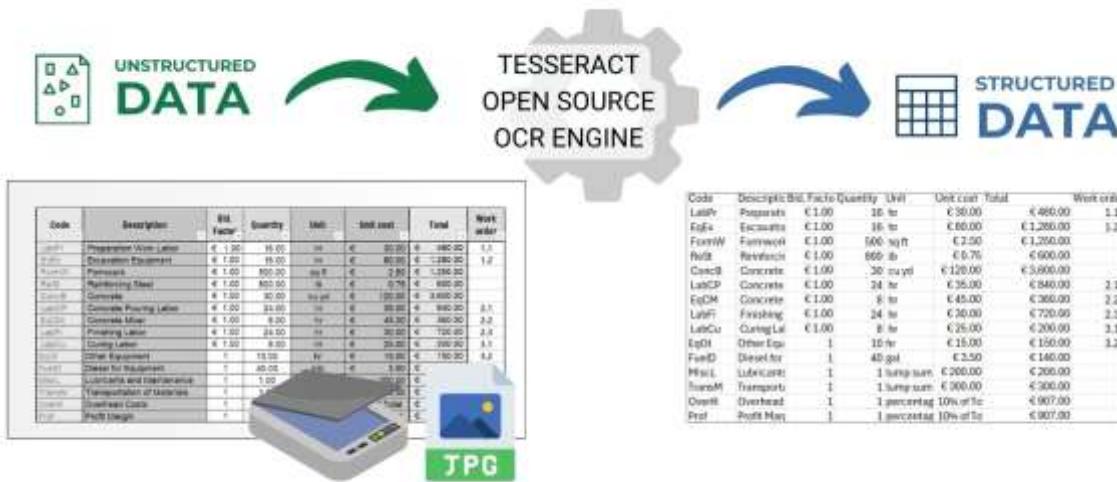


Fig. 4.1-5 Converting scanned documents and images into structured formats is possible using special OCR tools.

OCR (Optical Character Recognition) technology is used to extract text from images. It allows you to

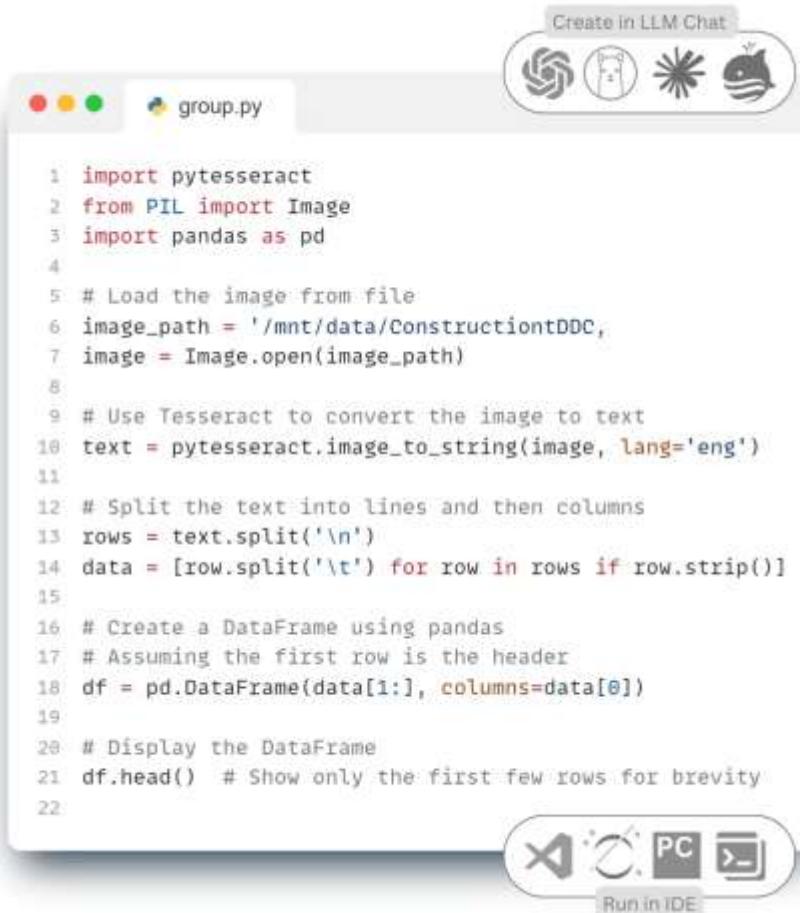
recognize letters and numbers in document scans, photos and PDF -files, turning them into editable and machine-readable text. OCR technologies have long been used in document automation and today they are easily integrated into any business processes and Python -applications. One of the most popular OCR tools is Tesseract, an open source, originally developed by HP™ and now supported by Google™. It supports over 100 languages and has high recognition accuracy.

Let's ask LLM chat to write sample code to retrieve data from a scanned or photographed table in a structured way.

- ❷ Send a text request to LLM chat (CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

```
Write code to convert a JPEG image containing a table to a DataFrame  
table ↴
```

- 2 LLM's response will in most cases suggest using the Pytesseract library to recognize text in images:



```

1 import pytesseract
2 from PIL import Image
3 import pandas as pd
4
5 # Load the image from file
6 image_path = '/mnt/data/ConstructionDDC'
7 image = Image.open(image_path)
8
9 # Use Tesseract to convert the image to text
10 text = pytesseract.image_to_string(image, lang='eng')
11
12 # Split the text into lines and then columns
13 rows = text.split('\n')
14 data = [row.split('\t') for row in rows if row.strip()]
15
16 # Create a DataFrame using pandas
17 # Assuming the first row is the header
18 df = pd.DataFrame(data[1:], columns=data[0])
19
20 # Display the DataFrame
21 df.head() # Show only the first few rows for brevity
22

```

Fig. 4.1-6 Convert text extracted from a table of images or photos into a structured tabular representation.

In this example - the code (Fig. 4.1-6) derived from the LLM, uses the pytesseract library (Tesseract for Python) to convert an image into text using OCR (optical character recognition) and the Pandas library to convert this text into a structured form, i.e., a DataFrame.

The conversion process usually involves preprocessing to improve image quality, after which various algorithms are applied for pattern detection, feature extraction or object recognition. As a result, unstructured visual information is converted into structured data.

While PDF and images are key sources of unstructured information, the real champion in terms of volume is text generated in emails, chats, meetings, messengers. This data is not just numerous - it is scattered, unformalized, and extremely poorly structured.

Converting text data into a structured form

In addition to PDF documents with tables (Fig. 4.1-2) and scanned versions of tabular forms (Fig. 4.1-5), a significant part of information in project documentation is presented in text form. It can be both coherent sentences in text documents and fragmentary records scattered over drawings and schemes. In modern conditions of data processing one of the most common tasks is to convert such text into a structured format suitable for analysis, visualization and decision making.

Central to this process is taxonomy, a classification system that organizes information into categories and subcategories based on common features.

A taxonomy is a hierarchical classification structure used to group and organize objects.

In the context of text processing, it serves as a basis for systematically categorizing items into semantic categories, thus simplifying analysis and improving the quality of data processing.

Taxonomy creation is accompanied by the steps of entity extraction, categorization and contextualization. To simulate the process of extracting information from textual data, we need to perform the following steps, similar to those we have already applied to structuring data from PDF documents:

- **Data extraction** (Extract): you need to analyze text data to extract information about delays and changes in the project schedule.
- **Categorization and Classification** (Transofrm): categorize the information received, e.g. Reasons for delays and schedule changes.
- **Integration** (Load): at the end we prepare structured data for integration into external data management systems.

Consider a situation: we have a dialog between a project manager and an engineer discussing problems with the schedule. Our goal is to extract the key elements (reasons for delay, schedule adjustments) and present them in a structured way (Fig. 4.1-7).

Let's perform the extraction based on the expected keywords, create a DataFrame to simulate data extraction and after transformation, a new DataFrame table that will contain columns for date, event (e.g., reason for delay) and action (e.g., schedule change).

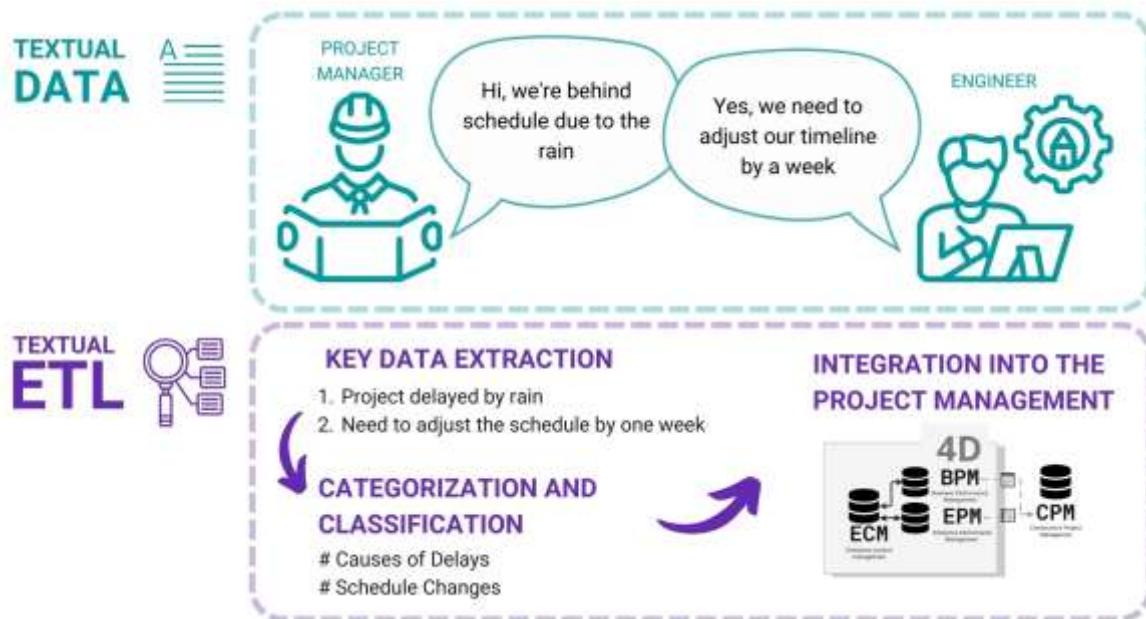


Fig. 4.1-7 Highlight key information from the text about the need to adjust timelines and integrate changes into the project management system.

Here is the code to solve the problem using a text query in one of the language models as in the previous examples.

- Send a text request to any LLM chat room:

```
I have a conversation between a manager, "Hello, we are behind schedule due to rain" and an engineer, "Yes, we need to adjust the deadline by a week". I need a script that will analyze future similar text conversations, extract the reasons for delays and necessary deadline adjustments, and then generate a DataFrame from that data. The DataFrame should then be saved to a CSV -file ↵
```

- 2 The response from LLM will typically include Python -code using regular expressions (re - Regex) and the Pandas (pd) library:

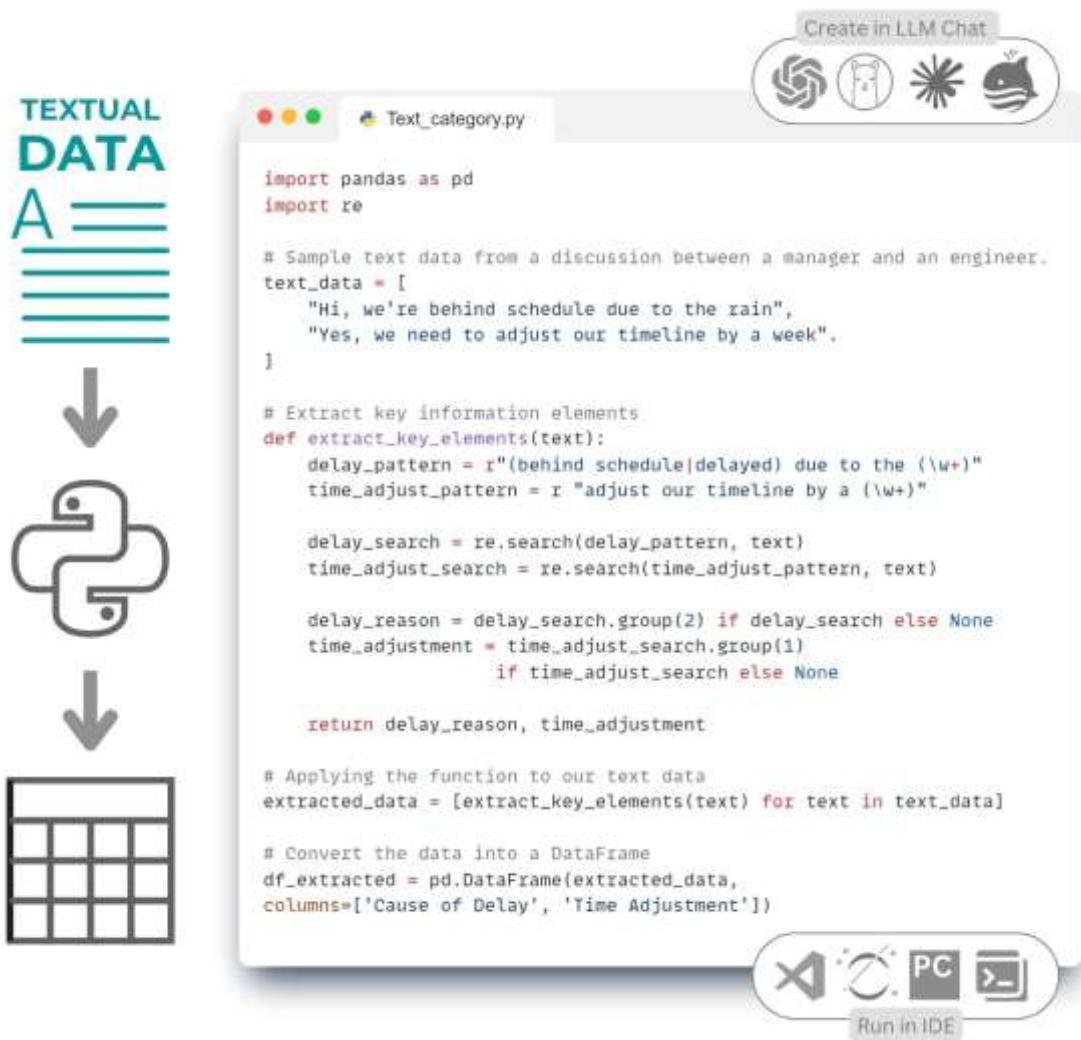


Fig. 4.1-8 Highlighting key information from the text about the need to adjust deadlines in a table.

In this example (Fig. 4.1-7), text data containing correspondence between a project manager and an engineer is analyzed to identify and extract specific information that may affect the management of future projects with similar dialogs. Using regular expressions (we'll talk more about regular expressions in the chapter "Structured Requirements and RegEx Regular Expressions"), the causes of project delays and necessary adjustments to the time schedule are identified through patterns. The function written in this example extracts either the cause of delay or the time adjustment from the strings based on the patterns: highlighting the word after "because of" as the cause of delay or the word after "by" as the time adjustment.

If a row mentions a delay due to weather, "rain" is identified as the cause; if a row mentions a schedule

adjustment for a specific period, that period is extracted as a time adjustment (Fig. 4.1-9). The absence of any of these words in a row results in a value of "None" for the corresponding attribute-column.

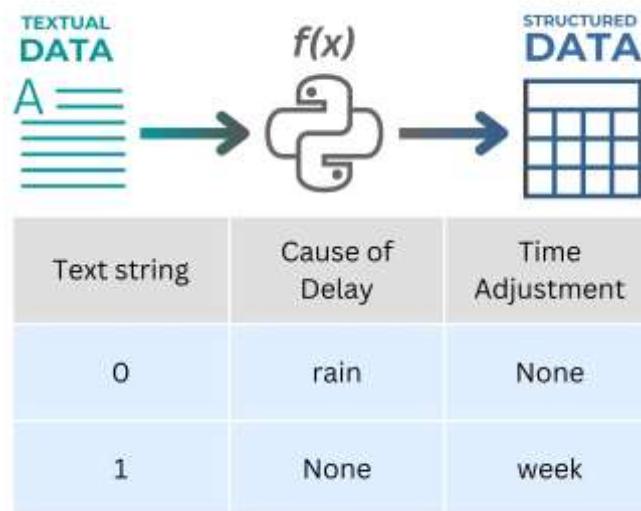


Fig. 4.1-9 The summary table obtained as DataFrame, after code execution, contains information about the existence of delays and necessary time adjustments.

Structuring and parameterization of conditions from the text (dialog, letter, document) allows to promptly eliminate delays in construction: for example, lack of workers can affect the pace of work in bad weather, so companies, knowing the delay parameters from dialogs (Fig. 4.1-9) between the foreman at the construction site and the project manager - in advance can strengthen the crew in case of an unfavorable forecast.

Converting documents and images into a structured format can be achieved with relatively simple, open and free categorization-based tools.

Element categorization is also a key part of working with project data, especially in the context of using CAD software (BIM).

Translation of CAD data (BIM) into a structured form

Structuring and categorizing CAD data (BIM) is more challenging because data stored from CAD (BIM) databases are almost always in closed or complex parametric formats, often combining geometric data elements (semi-structured) and metainformation elements (semi-structured or structured data) simultaneously.

Native data formats in CAD (BIM) systems are usually protected and inaccessible for direct use, unless specialized software or API - interfaces of the developer himself (Fig. 4.1-10). Such data isolation forms closed storage silos that limit the free exchange of information and inhibit the creation of end-to-end digital processes in the company.

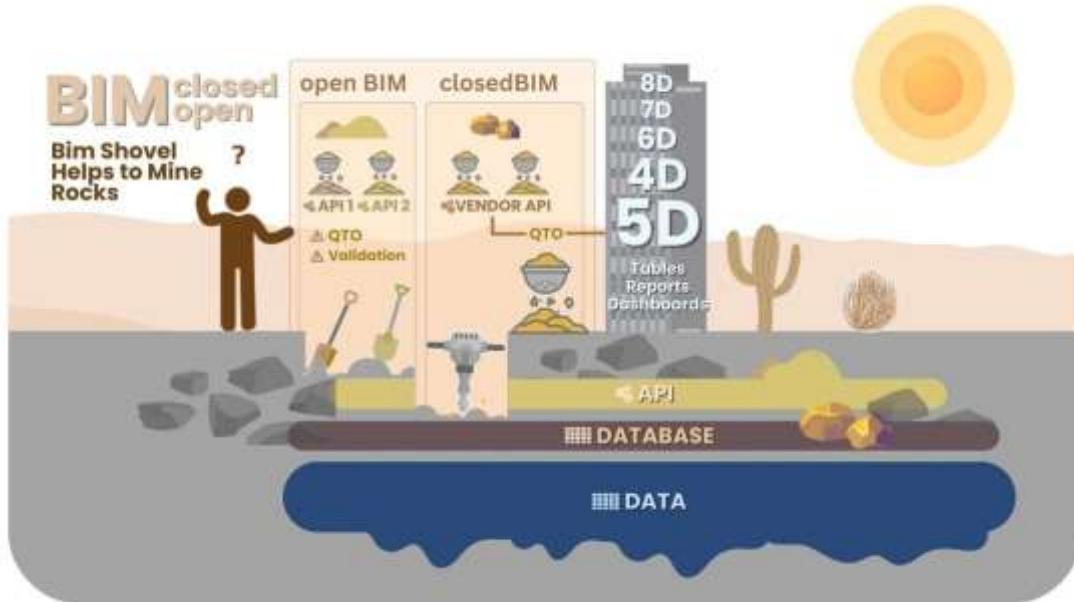


Fig. 4.1-10 CAD professionals (BIM) can access native data through API -connections or vendor tools.

In special CAD (BIM) formats, information about the characteristics and attributes of project elements is collected in a hierarchical classification system, where entities with corresponding properties are located, like the fruit of a fruit tree, in the most recent nodes of the data classification branches (Fig. 4.1-11).

Data extraction from such hierarchies is possible in two ways: either manually, by clicking on each node, as if processing a tree, cutting down selected branches of categories and types with an axe. An alternative option - the use of program interfaces (APIs) - implies a more efficient, automated approach to data retrieval and grouping, eventually transforming it into a structured table for use in other systems.

Different tools such as Dynamo, pyRvt, Pandamo (Pandas + Dynamo), ACC, or open source solutions, such as IfcOpSh or IFCjs for IFC format, can be used to extract structured data tables from CAD (BIM) projects.

Modern data export and conversion tools allow to simplify data processing and preparation by dividing the content of CAD models into two key components: geometry information and attribute data (Fig. 4.1-13) - meta-information describing the properties of design elements (Fig. 3.1-16). These two layers of data remain linked through unique identifiers, thanks to which it is possible to precisely map each element with geometry description (via parameters or polygons) to its attributes: name, material, stage of completion, cost, and so on. This approach ensures the integrity of the model and allows flexible use of data both for visualization (geometric model data) and for analytical or management tasks (structured or loosely structured), working with the two types of data separately or in parallel.

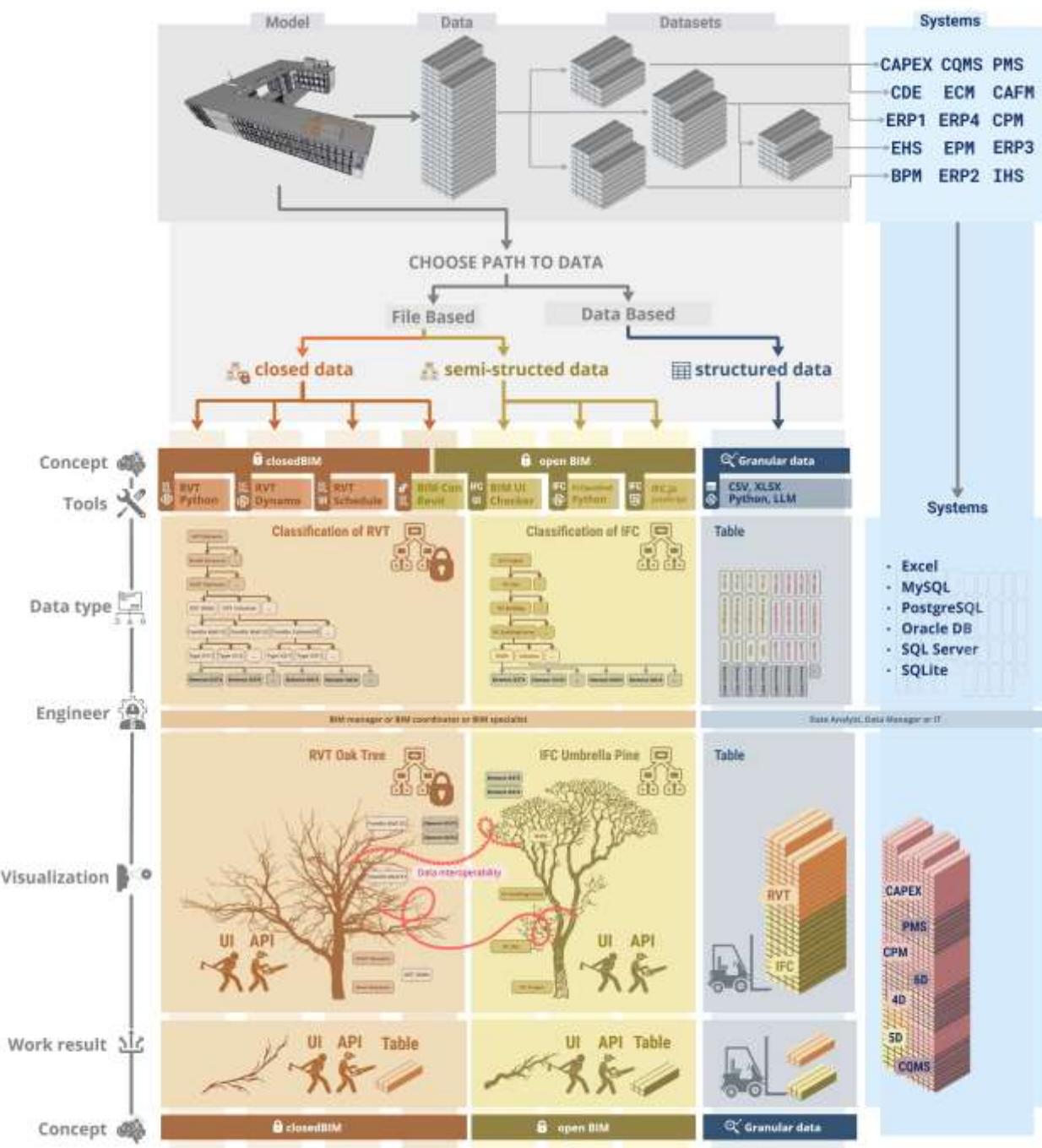


Fig. 4.1-11 The view of information from CAD databases (BIM) is presented to the user in the form of classification trees.

With the development of reverse engineering technologies and the advent of software development kits SDK (Software Development Kit) for CAD data conversion - availability and conversion of data from closed CAD program formats (BIM) has become much easier. It is now possible to legally and safely convert data from closed formats into universal formats suitable for analysis and use in other systems.

The history of the first reverse engineering tools ("Open DWG") and the struggle for dominance over CAD vendors' formats was discussed in the chapter "Structured data: the foundation of digital transformation".

Reverse engineering tools allow legitimate retrieval of data from closed proprietary formats, breaking down information from the mixed CAD (BIM) format into the data types and formats required by the user, making it easier to process and analyze.

Using reverse engineering and direct access to information from CAD databases makes information accessible, allowing open data and open tools, as well as analyzing data with standard tools, building reports, visualizations, and integrating with other digital systems (Fig. 4.1-12).

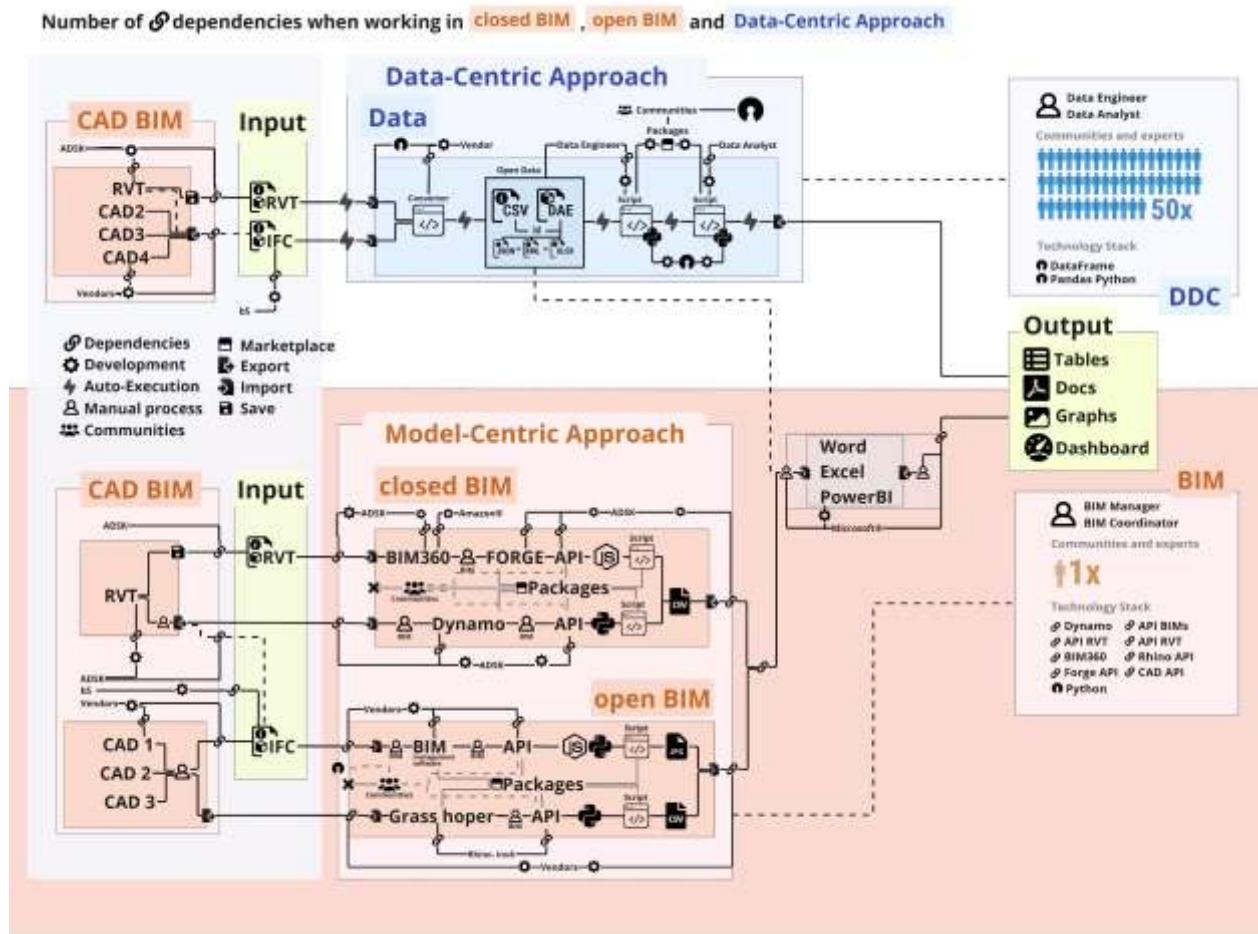


Fig. 4.1-12 Direct access to CAD data minimizes dependencies on software platforms and moves to a data-centric approach.

Since 1996 for DWG format, since 2008 for DGN format and since 2018 for RVT it has been possible to convert initially closed CAD data formats into any other formats, including structured formats, conveniently and efficiently with the help of reverse engineering tools (Fig. 4.1-13). Today, almost all major

CAD (BIM) and large engineering companies in the world use SDKs - reverse engineering tools to extract data from closed CAD (BIM) vendor formats [92].



Fig. 4.1-13 Using reverse engineering tools allows you to convert CAD (BIM) program databases into any convenient data model.

Converting data from closed, proprietary formats to open formats and separating mixed CAD (BIM) formats into geometric and meta-information attribute data simplifies the process of working with it, making it available for analysis, manipulation, and integration with other systems (Fig. 4.1-14).

In today's work with CAD data (BIM), we have reached the point where you don't need to request permission from CAD (BIM) vendors to access information from CAD formats.

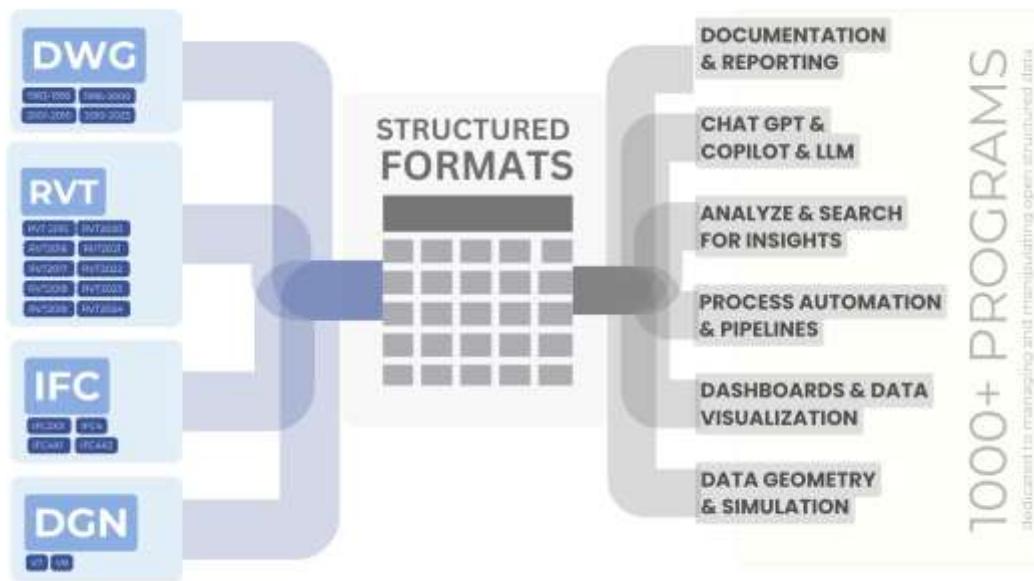


Fig. 4.1-14 Modern SDK tools allow legal conversion of data from proprietary CAD database formats (BIM).

Current trends in CAD design data processing continue to be shaped by key market players - CAD - vendors who are working to strengthen their position in the data world and create new formats and concepts.

CAD solution vendors move to structured data

From 2024, the design and construction industry is undergoing a significant technological shift in the use and processing of data. Instead of free access to design data, CAD -system vendors are focusing on promoting the next new concepts. Approaches such as BIM (created in 2002) and open BIM (created in 2012) are gradually giving way to modern technological solutions that CAD vendors are starting to promote [93]:

- Moving to the use of "granular" data that allows for efficient information management and a shift to data analytics
- Emergence of USD format and implementation of Entity-component-system approach (ECS) for flexible data organization
- Active use of artificial intelligence in data processing, process automation and data analytics
- Developing interoperability - improved interaction between different programs, systems and databases

Each of these aspects will be discussed in more detail in the sixth part of the book "CAD and BIM: marketing, reality and the future of design data in construction". Within this chapter we will only briefly outline the general vector of change: the largest CAD vendors are now striving to rethink the way design information is structured. One of the key shifts is the rejection of the classic file-based storage model in favor of a granular, analytics-oriented data architecture that provides continuous access to individual components of the model [93].

The essence of what is happening is that the industry is phasing out cumbersome, specialized and parametric formats that require geometric kernels in favor of more universal, machine-readable and flexible solutions.

One such driver of change is the USD (Universal Scene Description) format, originally developed in the computer graphics industry, but already recognized in engineering applications thanks to the development of the NVIDIA Omniverse (and Isaac Sim) platform for simulations and visualizations [93]. Unlike the parametric IFC, USD offers a simpler structure and allows describing geometry and object properties in JSON format (Fig. 4.1-15), which facilitates information processing and speeds up its integration into digital processes. The new format allows storing geometry (in addition to BREP -NURBS - more details in Part 6 of the book) in the form of MESH polygons, and object properties in JSON, which makes it more convenient for automated processes and work in cloud ecosystems [94].

Some CAD and ERP vendors already use similar formats (e.g. NWD, SVF, CP2, CPIXML), but most of them remain closed and unavailable for external use, which limits the possibilities of data integration and reuse. In this context, USD can play the same role as DXF did in its time as an open alternative to proprietary formats like DWG.

General Information				Comparison / Notes
Year of format creation	1991	2016		IFC focuses on construction data, USD on 3D graphics
Creator-developer	TU Munich	Pixar		IFC was founded in Germany, USD in America
Prototypes and predecessors	IGES, STEP	PTEX, DAE, GLTF		IFC evolved from IGES/STEP, USD from PTEX/DAE/GLTF
Initiator in Construction	ADSK	ADSK		ADSK initiated the adoption of both formats in construction
Organizer of the Alliance	ADSK	ADSK		ADSK organized both alliances
Name of the Alliance	bS (IAI)	AOUSD		Different alliances for each format
Year of Alliance Formation	1994	2023		The IFC alliance was formed in 1994, AOUSD for USD in 2023
Promoting in the construction	ADSK and Co	ADSK and Co		ADSK and Co actively promotes both formats in bS (IAI) since the introduction

Purpose and Usage				Comparison / Notes
Purpose	Semantic description and interoperability	Data simplification, visualization, unification		IFC for semantics and exchange; USD for simplification and visualization
Goals and Objectives	Interoperability and semantics	Unification for visualization and data processing		IFC focuses on semantics; USD on visualization
Use in Other Industries	Predominantly in construction	In film, games, VR/AR, and now in construction		USD is versatile and used in various fields
Supported Data Types	Geometry, object attributes, metadata	Geometry, shaders, animation, light, and camera		USD supports a wider range of data types suitable for complex visualizations; IFC focuses on construction-specific data

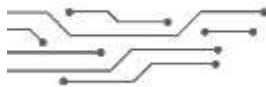
Fig. 4.1-15 USD format as an attempt by CAD vendors to meet the demand for interoperability and independence of design data from geometry kernels.

The transition of major developers to open and simplified USD, GLTF, OBJ, XML (closed NWD, CP2, SVF, SVF2, CPIXML) and similar formats (Fig. 3.1-17) reflects the global trend and industry demand for data simplification and increased accessibility. In the coming years, we can expect a gradual move away from complex parametric standards and formats with dependence on geometric kernels in favor of lighter and more structured solutions. This transition will accelerate the digitalization of the construction industry, facilitate process automation and simplify data exchange.

Despite the strategic plans of CAD -vendors to promote new open formats, construction industry professionals can also fully access data from closed CAD systems, without the need for CAD (BIM) tools, by using reverse engineering tools.

All these trends inevitably lead to a shift from bulky, monolithic 3D models to universal, structured data and to the use of formats that have long been proven in other industries. Once project teams begin to see CAD models not just as visual objects or a set of files, but as databases containing knowledge and information, the approach to design and management changes dramatically.

Once teams have learned how to extract structured data from documents, texts, drawings and CAD models, and have access to databases, the next key step is data modeling and quality assurance. It is this step that largely determines the speed of processing and transformation of information that will ultimately be used to make decisions in specific application tasks.



CHAPTER 4.2.

CLASSIFICATION AND INTEGRATION: A COMMON LANGUAGE FOR CONSTRUCTION DATA

Speed of decision making depends on data quality

Today's design data architecture is undergoing fundamental changes. The industry is moving away from bulky, isolated models and closed formats towards more flexible, machine-readable structures focused on analytics, integration and process automation. However, the transition to new formats alone does not guarantee efficiency - the quality of the data itself is inevitably at the center of attention.

In the pages of this book, we talk a lot about formats, systems, and processes. But all these efforts are meaningless without one key element: data that can be trusted. Data quality is a cornerstone of digitalization that we will return to throughout the parts that follow.

Modern construction companies - especially large ones - use dozens and sometimes thousands of different systems and databases (Fig. 4.2-1). These systems must not only be filled with new information on a regular basis, but also interact effectively with each other. All new data generated as a result of processing incoming information are integrated into these environments and serve to solve specific business tasks.

And if earlier decisions on specific business tasks were made by top managers - so-called HiPPOs (Fig. 2.1-9) - on the basis of experience and intuition, today, with the sharp increase in the volume of information, this approach is becoming controversial. Automated analytics, which works with real-time data, is replacing it.

"Traditional-manual" executive-level business process discussions will shift toward operational analytics, which requires quick responses to business queries.

The era when accountants, foremen and estimators manually generated reports and summary tables and project data showcases over days and weeks is a thing of the past. Today, speed and timeliness of decision-making are becoming a key factor in competitive advantage.

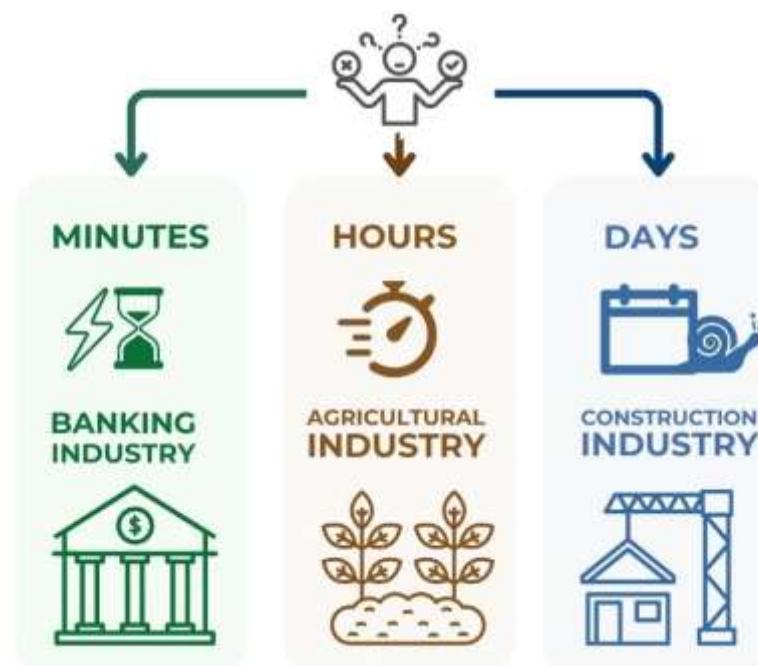


Fig. 4.2-1 The construction industry takes days to calculate and make decisions, unlike other industries where this happens in hours or minutes.

The main difference between the construction industry and more digitally developed industries (Fig. 4.2-1) is the low level of data quality and standardization. Outdated approaches to the generation, transmission and processing of information slow down processes and create chaos. The lack of uniform data quality standards hinders the implementation of end-to-end automation

One of the main challenges remains the poor quality of input data, as well as the lack of formalized processes for their preparation and verification. Without reliable and consistent data, effective integration between systems is impossible. This leads to delays, errors and increased costs at every stage of the project lifecycle.

In the following sections of the book, we detail how you can improve data quality, standardize processes, and shorten the path from information to quality, validated, and consistent data.

Data standardization and integration

Effective data management requires a clear standardization strategy. Only with clear requirements for data structure and quality can data validation be automated, manual operations reduced and informed decision making accelerated at all stages of a project.

In daily practice, a construction company has to process hundreds of files every day: e-mails, PDF - documents, CAD design files, data from IOT sensors, which need to be integrated into the company's business processes.

The forest of a company's ecosystem of databases and tools (Fig. 4.2-2) must learn how to derive nutrients from incoming multiformat data to produce the results the company needs.

To effectively deal with the flow of data, you don't necessarily need to hire an army of managers, you first need to develop strict requirements and standards for data and use appropriate tools to automatically validate, unify and process it.

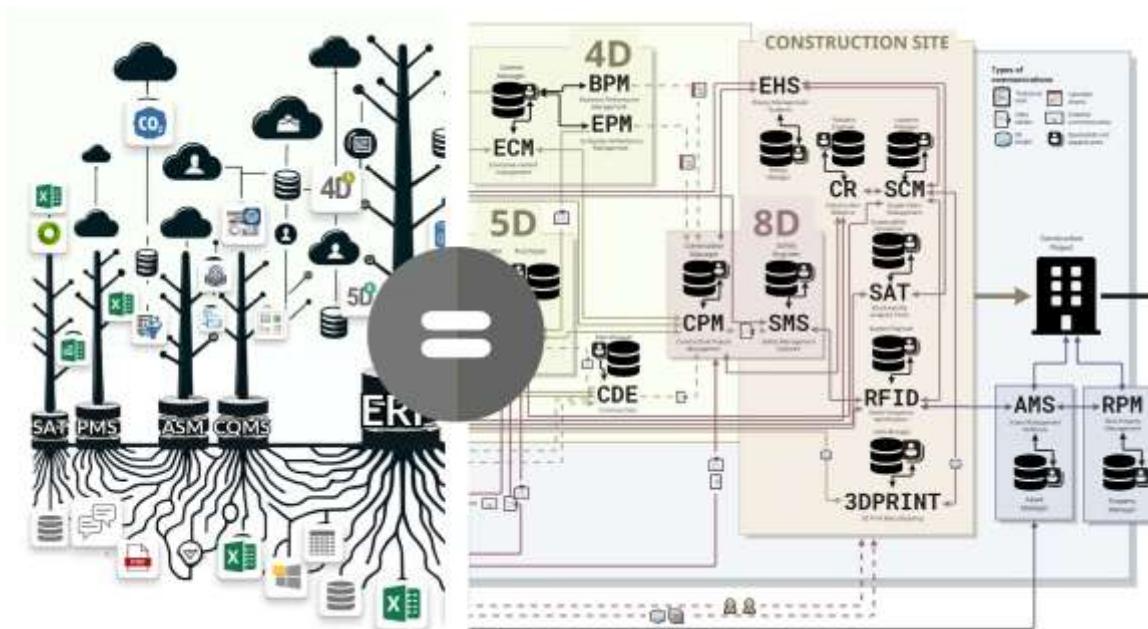


Fig. 4.2-2 Ensuring the healthy vitality of a company's ecosystem requires quality and timely resourcing of its systems.

To automate the process of data validation and unification (for subsequent automatic integration) you should start by describing the minimum necessary data requirements for each specific system. These requirements define:

- What exactly do you need to get?
- In what form (structure, format)?
- What attributes are mandatory?
- What tolerances in accuracy and completeness are acceptable?

Data requirements describe the criteria of quality, structure and completeness of the received and processed information. For example, for texts in PDF -documents it is important to ensure accurate formatting in accordance with industry standards (Fig. 7.2-14 - Fig. 7.2-16). Objects in CAD -models must have correct attributes (dimensions, codes, links to classifiers) (Fig. 7.3-9, Fig. 7.3-10). And for contract scans, clear dates and the ability to automatically extract the amount and key terms are important (Fig. 4.1-7 - Fig. 4.1-10).

Formulating data requirements and automatically checking their compliance is one of the most time-consuming but critical steps. It is the most time-consuming step in business processes.

As mentioned in Part 3 of this book, between 50% and 90% of business intelligence (BI) professionals' time is spent on data preparation rather than analysis (Fig. 3.2-5). This process includes data collection, verification, validation, harmonization, and structuring.

According to a 2016 survey [95], data scientists in a wide variety of broad-spectrum fields stated that they spend most of their work time (about 80%) doing what they least like to do (Fig. 4.2-3): collecting existing datasets and organizing (unifying, structuring) them. Thus, less than 20% of their time is left for creative tasks, such as finding patterns and regularities that will lead to new insights and discoveries.

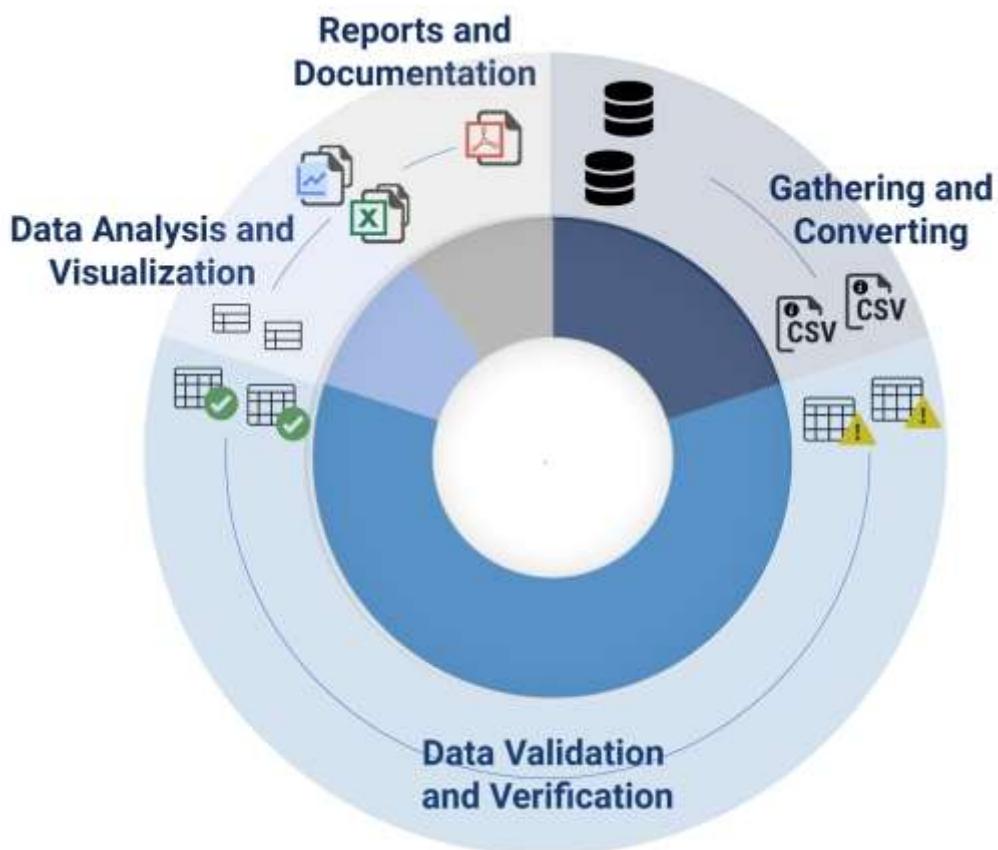


Fig. 4.2-3 Verifying and ensuring data quality is the most costly, time-consuming, and complex step in preparing data for integration into other systems.

Successful data management in a construction company requires a comprehensive approach that includes parameterization of tasks, formulation of data quality requirements, and use of suitable tools for their automated validation.

Digital interoperability starts with requirements

As the number of digital systems within companies grows, so does the need for data consistency between them. Managers responsible for different IT systems often find themselves unable to keep up with the increasing volume of information and the variety of formats. In such circumstances, they are forced to ask specialists to create data in a form suitable for use in other applications and platforms.

This, in turn, requires engineers and data generation staff to adapt to a multitude of requirements, often without transparency and a clear understanding of where and how the data will be applied in the future. The lack of standardized approaches to handling information leads to inefficiencies and increased costs during the verification phase, which is often manual due to the complexity and non-standardized nature of the data.

The issue of data standardization is not just a matter of convenience or automation. It is a direct financial loss. According to a 2016 IBM report, the annual loss from poor data quality in the US is \$3.1 trillion [96]. Additionally, studies by MIT and other analytical consulting firms show that the cost of poor data quality can be as high as 15-25% of a company's revenue [97].

Under these conditions, it becomes critical to have clearly defined data requirements and descriptions of what parameters, in what format and with what level of detail should be included in the created objects. Without formalizing these requirements, it is impossible to guarantee the quality and compatibility of data between systems and project stages (Fig. 4.2-4).

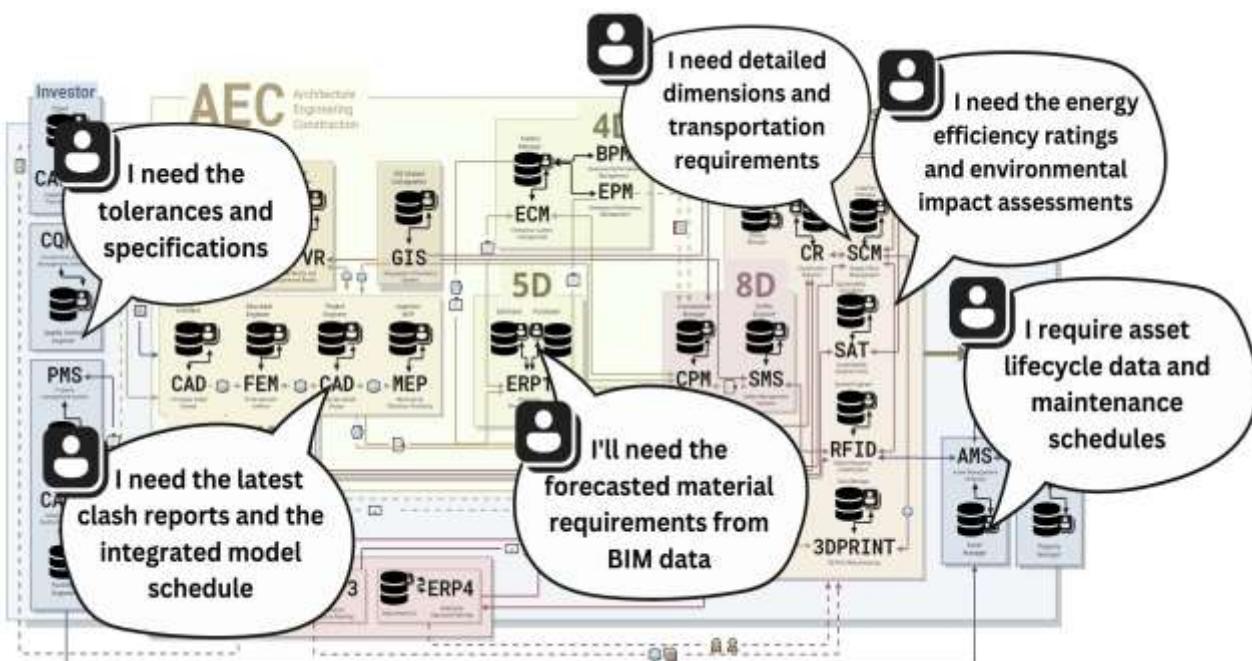


Fig. 4.2-4 Business is based on the interaction of different roles, each of which requires certain parameters and values that are critical to accomplishing business objectives.

In order to formulate the correct data requirements, you need to understand the business processes at the data level. Construction projects vary in type, scope, and number of participants, and each system - be it modeling (CAD (BIM)), scheduling (ERP 4D), costing (ERP 5D), or logistics (SCM) - requires its own unique parameters for inputs (input entity-elements).

Depending on these needs, business managers must either design new data structures to meet the requirements or adapt existing tables and databases. The quality of the data created will directly depend on how precisely and correctly the requirements are formulated (Fig. 4.2-5).

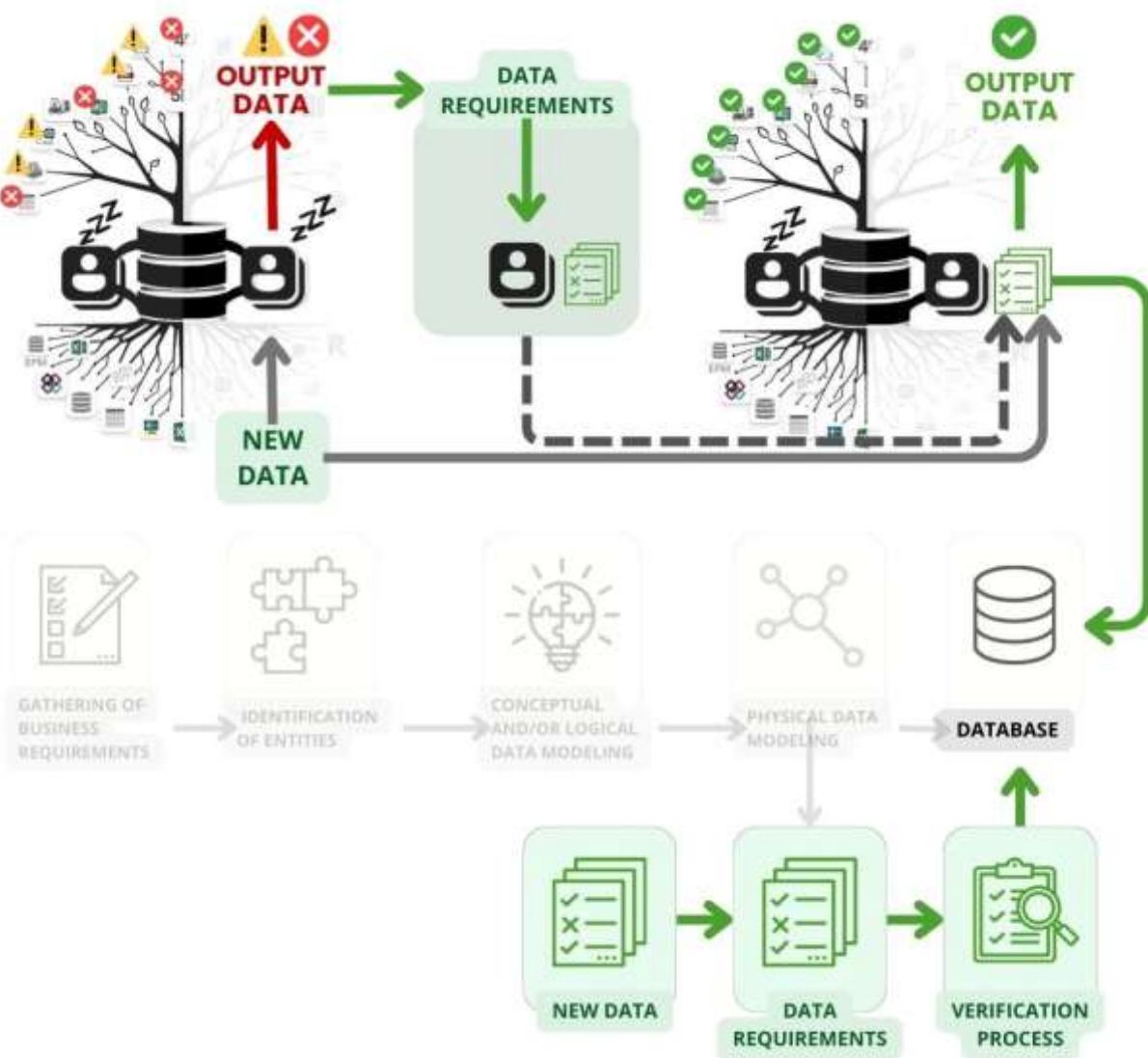


Fig. 4.2-5 Data quality depends on the quality of the requirements that are created for specific data use cases.

Since each system has its own specific data requirements, the first step in formulating general requirements should be to categorize all elements involved in business processes. This means the need to categorize objects into classes and groups of classes corresponding to specific systems or application tasks. For each such group, separate requirements for data structure, attributes and quality are developed.

In practice, however, the implementation of this approach faces a major challenge: the lack of a common language for grouping data. Disparate classifications, duplicate identifiers and incompatible formats result in each company, each software and even each project forming its own, isolated data models and classes. The result is a digital "Tower of Babel" where transferring information between systems requires multiple conversions to the right data models and classes, often done manually. This barrier can only be overcome by moving to universal classifiers and standardized sets of requirements.

A common language of construction: the role of classifiers in digital transformation

In the context of digitalization and automation of inspection and processing processes, a special role is played by classification systems elements - a kind of "digital dictionaries" that ensure uniformity in the description and parameterization of objects. Classifiers form the "common language" that allows data to be grouped by meaning and data to be integrated between different systems, management levels and phases of the project lifecycle.

The most tangible impact of classifiers is in the economics of the building life cycle, where the most important aspect is the optimization of long-term operating costs. Studies show that operating costs account for up to 80% of the total cost of building ownership, which is three times higher than the initial construction costs (Fig. 4.2-6) [98]. This means that the decision on future costs is largely formed at the design stage

This is why requirements from operations engineers (CAFM, AMS, PMS, RPM) should become the starting point for generating data requirements during the design phase (Fig. 1.2-4). These systems should not be viewed as the final stage of the project, but as an integral part of the entire digital ecosystem of the project, from concept to disassembly

A modern classifier is not just a system of codes for grouping. It is a mechanism for mutual understanding between architects, engineers, estimators, logisticians, maintenance and IT systems. Just as a car's autopilot must unambiguously recognize road objects with high precision, digital construction systems and their users must interpret the same project element unambiguously for different systems via the element class.



Fig. 4.2-6 Operating and maintenance costs exceed the cost of construction three times, accounting for 60-80% of all life cycle costs of a building (according to [99]).

The level of classifier development directly correlates with the depth of a company's digitalization and its digital maturity. Organizations with a low level of digital maturity are faced with fragmented data, incompatible information systems and, as a result, incompatible and inefficient classifiers. In such companies, the same element can often have different grouping identifiers in different systems, which critically complicates final integration and makes process automation impossible.

For example, the same window in a project can be labeled differently in CAD model, estimating and maintenance system (Fig. 4.2-7) because of the multidimensional perception of elements by different participants in the process. For the estimator in the windows category element, volume and cost are important, for the maintenance service - availability and maintainability, for the architect - aesthetic and functional characteristics. As a result, the same element may require different parameters.

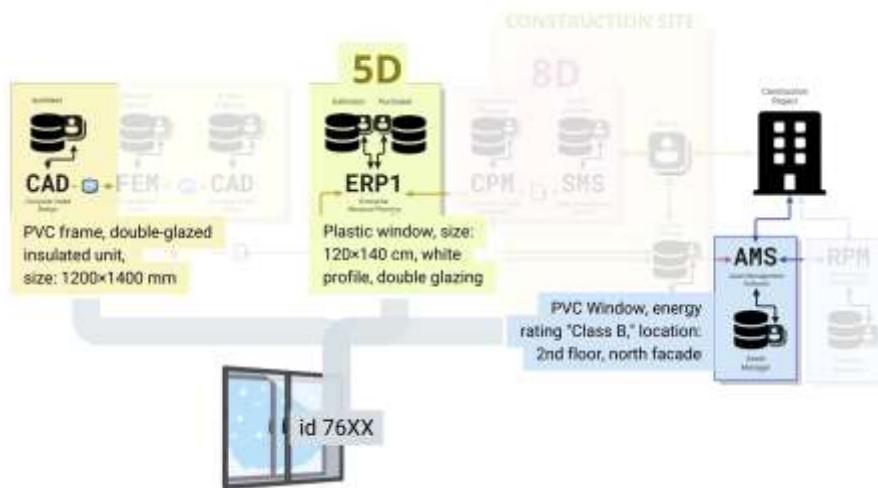


Fig. 4.2-7 With inconsistent classification between systems an element will lose some attribute information at each step of its transition to another system.

Due to the difficulty of unambiguously defining the classification of building elements, specialists from

different fields often assign incompatible classes to the same element. This leads to a loss of a unified view of the object, which requires subsequent manual intervention to harmonize the different classification systems and to establish correspondence between the types and classes defined by different specialists.

As a result of this inconsistency, the operational documentation received by the procurement department (ERP) when a construction item is purchased from a manufacturer often cannot be correctly linked to the classification of that item at the construction site (PMIS, SCM). As a consequence, critical information is not likely to be integrated into infrastructure and asset management systems (CAFM, AMS), which creates serious problems during commissioning, as well as during subsequent maintenance (AMS, RPM) or replacement of the element.

In companies with high digital maturity, classifiers play the role of a nervous system that integrates all information flows. The same item receives a unique identifier that allows it to be transferred between CAD, ERP, AMS and CAFM -systems and their classifiers without distortion or loss.

To build effective classifiers, you need to understand how the data is used. The same engineer may name and classify an element differently in different projects. Only by collecting usage statistics over the years can a stable classification system be developed. Machine learning helps with this: algorithms analyze thousands of projects (Fig. 9.1-10), identifying likely classes and parameters through machine learning (Fig. 10.1-6). Automatic classification is especially valuable in environments where manual classification is not possible due to the volume of data. Automatic classification systems will be able to distinguish basic categories based on minimally populated item parameters (more details in the ninth and tenth parts of the book).

Developed classifier systems become catalysts for further digitalization, creating the basis for:

- Automated estimation of project cost and schedule.
- Predictive analysis of potential risks and conflicts
- Optimization of procurement processes and logistics chains
- Creating digital doubles of buildings and structures
- Integrations with smart city and Internet of Things systems

The time for transformation is limited - with the development of machine learning and computer vision technologies, the problem of automatic classification, which has been unsolvable for decades, will be solved in the coming years, and construction and design companies that fail to adapt in time risk repeating the fate of taxicabs displaced by digital platforms.

More about the automation of costing and scheduling as well as big data and machine learning will be

covered in the fifth and ninth parts of the book. The risk of a repeat of the fate of taxi fleets and the Uberization of the construction industry are discussed in detail in the tenth part of the book.

Understanding the key role of classifiers in the digital transformation of the construction industry, it is necessary to turn to the history of their evolution. It is the historical context that allows us to realize how approaches to classification have evolved and what trends determine their current state.

Masterformat, OmniClass, Uniclass and CoClass: the evolution of classification systems

Historically, construction element and work classifiers have evolved in three generations, each reflecting the level of available technology and the current needs of the industry in a particular time period (Fig. 4.2-8):

- **First generation** (early 1950s to late 1980s) - paper-based directories, hierarchical classifiers used locally (e.g. Masterformat, SfB).
- **The second generation** (late 1990s to mid-2010s) are spreadsheets and structured databases implemented in Excel and Access (ASTM E 1557, OmniClass, Uniclass 1997).
- **Third generation** (2010s to present) - digital services and APIs -interfaces, integration with CAD (BIM), automation (Uniclass 2015, CoClass).

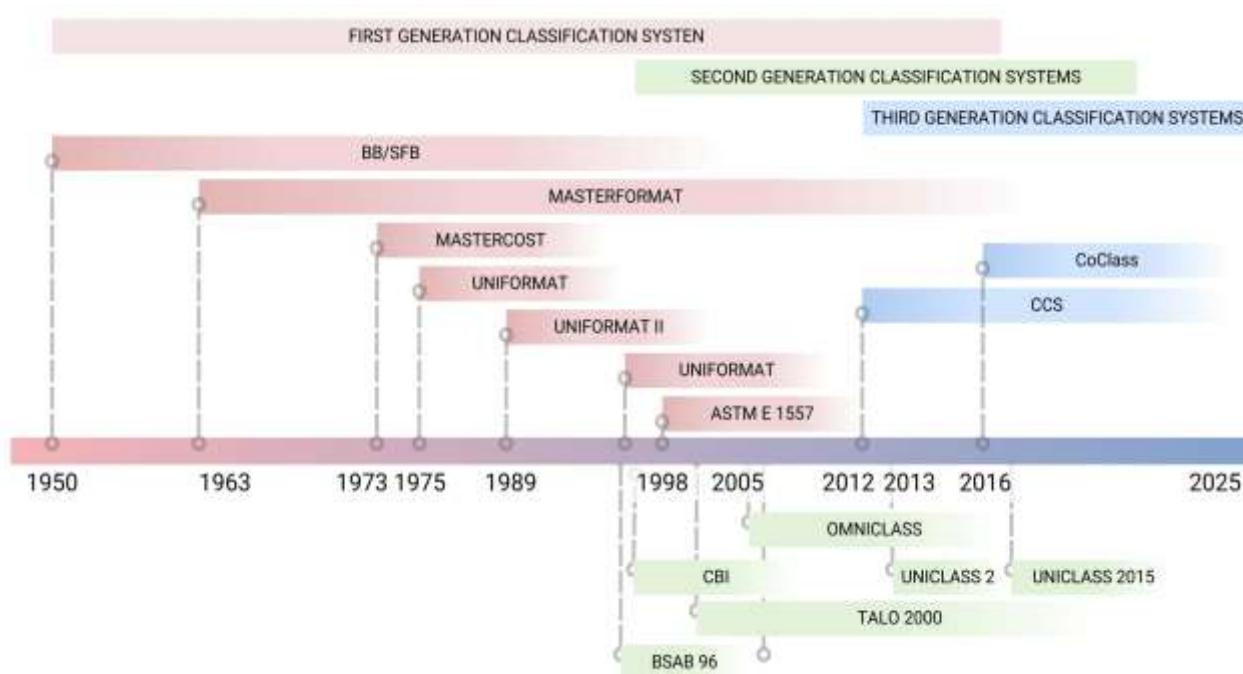


Fig. 4.2-8 Three generations of construction industry classifiers.

Over the last decades there has been a reduction in the hierarchical complexity (Fig. 4.2-9) of classifiers: while early systems such as OmniClass, used up to 7 nesting levels to describe 6887 classes, modern solutions such as CoClass are limited to 3 levels with 750 classes. This simplifies data handling while maintaining the necessary granularity. Uniclass 2015, often used as a standard in the UK, combines 7210 classes in just 4 levels, making it convenient for CAD projects and public procurement.

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

Fig. 4.2-9 With each new generation of classifiers, the complexity of categorization decreases by leaps and bounds.

In the construction estimating systems of different countries, even a typical element such as a concrete foundation wall can be described in very different ways due to different classifications (Fig. 4.2-10). These differences reflect national construction practices, measurement systems used, approaches to material classification, and the regulatory and technical requirements in force in each country.



Fig. 4.2-10 The same element is used in projects in different countries through different descriptions and classifications.

The diversity of classifications of the same elements complicates international cooperation, makes the comparison of cost and scope of work within international projects laborious and sometimes almost impossible. At the moment, there is no one universal classifier at the global level - each country or region develops its own systems based on local norms, language and business culture:

- **CCS (Denmark): Cost Classification System** - a classification system costs throughout the life

cycle of a facility (design, construction, operation). The focus is on the O&M logic, but also includes budget and resource management.

- **NS 3451** (Norway): categorizes facilities by function, design elements and life cycle stages. Used for project management, cost estimation and long-term planning.
- **MasterFormat** (USA): a system for structuring construction specifications into sections (e.g. concrete, electrical, finishing). Focus on disciplines and work types rather than functional elements (unlike UniFormat).
- **Uniclass 2** (UK): one of the most detailed classifiers, used in public procurement and BIM - projects. Unifies data on objects, works, materials and spaces into a single system.
- **OmniClass**: an international standard (developed by CSI in the USA) for managing object information, from component libraries to electronic specifications. Suitable for long-term data storage, compatible with CAD (BIM) and other digital tools.
- **COBie**: Construction-Operation Building information exchange is an international standard for the exchange of data between the design, construction and operation phases. Included in BS 1192-4:2014 as part of the concept of "BIM -model ready for operation". Focuses on information transfer (e.g. equipment specifications, warranties, contractor contacts).

The globalization of the construction industry is likely to lead to a gradual unification of building element classification systems, which will significantly reduce dependence on local national standards. This process may develop similarly to the evolution of Internet communications, where universal data transfer protocols eventually supplanted disparate local formats, ensuring global interoperability of systems.

An alternative development path may be a direct transition to automatic classification systems based on machine learning technologies. These technologies, which are being developed today mainly in the field of autonomous transportation, have significant potential for application to large CAD design data sets (Fig. 10.1-6).

Today the situation is not limited only to the national clustering of classifiers. Due to the many peculiarities not taken into account at the national level, each company has to independently engage in unification and standardization of the categories of elements and resources with which it works.

As a rule, this process starts small - with local tables of objects or internal designation systems. However, the strategic goal is to move to a common language for describing all elements, which would be understandable not only within the company, but also outside it - ideally, harmonized with international or industry classifiers (Fig. 4.2-8). This approach facilitates integration with external partners, digital systems and promotes the formation of unified end-to-end processes within the life cycle of objects.

Before moving to automation and scalable IT systems, it is necessary to either use national-level classifiers or build your own, logical and unambiguous element identification structure. Every object - whether it is a window (Fig. 4.2-11), a door, or an engineering system - must be described in such a way that it can be unmistakably recognized in any company's digital system. This is critical in the transition from flat drawings to digital models, covering both the design phase and building operation.

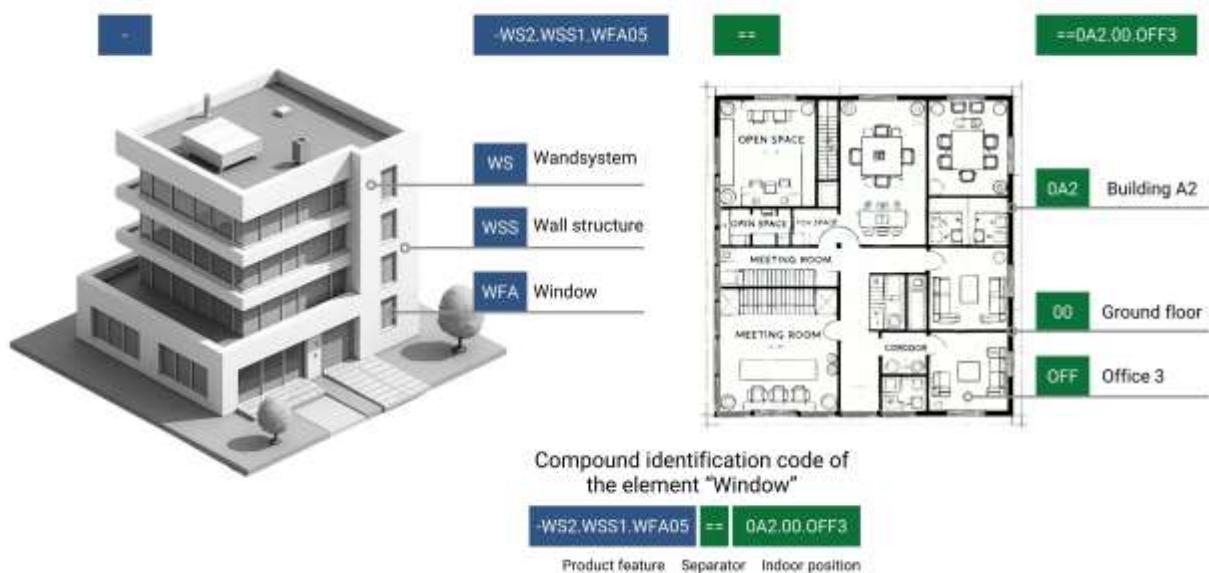


Fig. 4.2-11 Example of a composite window building element identifier based on classification and position in the building.

One example of internal classifiers could be the development of a composite identification code (Fig. 4.2-11). Such a code combines several levels of information: the functional purpose of the element (e.g., "window in the wall"), its type, and the exact spatial reference - building A2, floor 0, room 3. Such multilevel structure allows to create a unified system of navigation through digital models and documentation, especially at the stages of data verification and transformation, where unambiguous grouping of elements is required. Unambiguous element recognition ensures consistency between departments and reduces the risks of duplication, errors and loss of information.

A well-aligned classifier is not just a technical document, it is the foundation of a company's digital ecosystem:

- ensures data compatibility between systems;
- reduces the cost of searching and processing information;
- increases transparency and manageability;
- Creates a foundation for scaling and automation.

Standardized description of objects, through the use of national classifiers or own composite identification codes becomes the basis for consistent data, reliable information exchange and subsequent implementation of intelligent services - from automated procurement to digital twins.

After completing the structuring stage of multi-format data and selecting a classifier to be used for recognizing and grouping elements, the next step is correct data modeling. This process involves iden-

tifying key parameters, building a logical data structure and describing the relationships between elements.



CHAPTER 4.3.

DATA MODELING AND CENTER OF EXCELLENCE

Data modeling: conceptual, logical and physical model

Effective management of data (structured and categorized by us earlier) is impossible without a well thought-out storage and processing structure. To ensure access and consistency of information at the storage and processing stages, companies use data modeling - a methodology that allows designing tables, databases and links between them in accordance with business requirements.

Data modeling is the foundation on which any digital ecosystem is built. Without a description of systems, requirements, and data modeling, engineers and professionals creating data don't know or understand where the data they create will be used.

Like building a building, where you can't start laying bricks without a plan, creating a data warehouse system requires a clear understanding of what data will be used, how it will be linked, and who will work with it. Without a description of the processes and requirements, the engineers and professionals creating the data lose sight of where and how the data will be used in the future.

The data model serves as a bridge between business and IT. It allows formalizing requirements, structuring information and facilitating communication between stakeholders. In this sense, data modeling is similar to the work of an architect who, according to the customer's plan, develops a building plan and then passes it to the builders - database administrators and developers - for implementation (database creation).

Thus, every construction company, in addition to structuring and categorizing elements and resources (Fig. 4.2-11), must master the art of "building" databases (tables) and learn how to create links between them, as if connecting the bricks into a reliable and strong wall of knowledge from the company's data. Key concepts in data modeling (Fig. 4.3-1) include:

- **Entities** are objects about which data must be collected. In the early design phase, an entity can be a single element (e.g., "door"), and in the estimate model, it can be a group of elements organized by category (e.g., "interior doors").
- **Attributes** are characteristics of entities that describe important details: dimensions, properties, assembly costs, logistics, and other parameters.
- **Relationships (links)** - show how entities interact with each other. They can be of one of the following types: "one to one", "many to one", "many to many".
- **ER diagrams** (Entity-Relationship diagrams) are visual diagrams that show entities, attributes, and the relationships between them. ER-diagrams can be conceptual, logical, and physical - each reflecting a different level of detail.

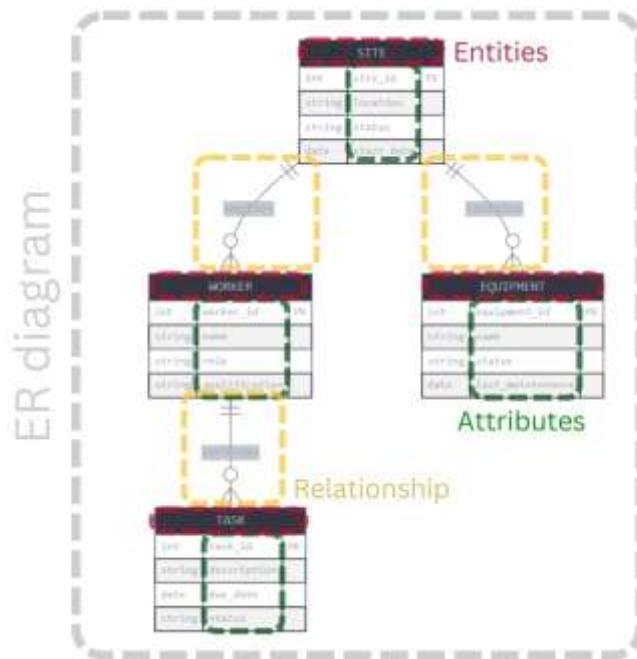


Fig. 4.3-1 ER diagram of a conceptual database structure with entities, attributes, and relationships.

The process of designing data and defining relationships between them is traditionally divided into three main models. Each of them performs certain functions, differing in the level of detail and degree of abstraction in representing the data structure:

- **Conceptual data model:** this model describes the basic entities and their relationships without going into attribute details. It is usually used in the initial stages of planning. At this stage we can sketch from databases and systems to show the relationship between different departments and specialists.

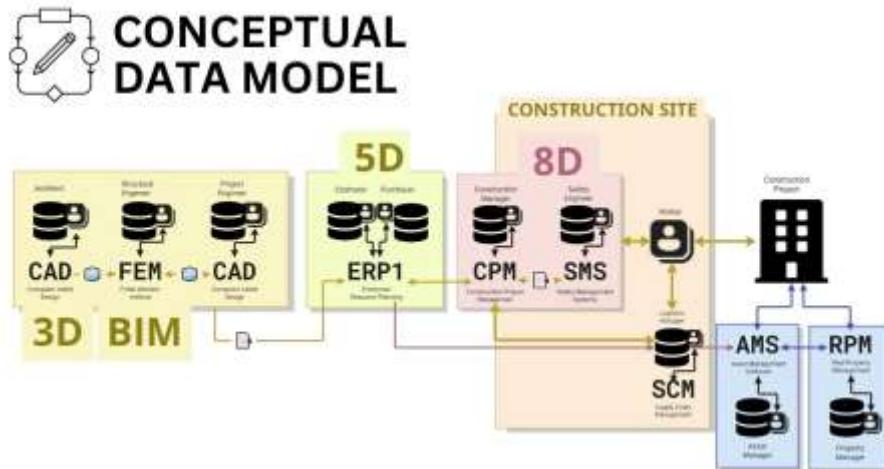


Fig. 4.3-2 The conceptual diagram describes the content of the system: a high-level representation of the relationships, without technical details.

- **Logical Data Model:** Based on the conceptual model, the logical data model includes detailed descriptions of entities, attributes, keys, and relationships, mapping business information and rules.

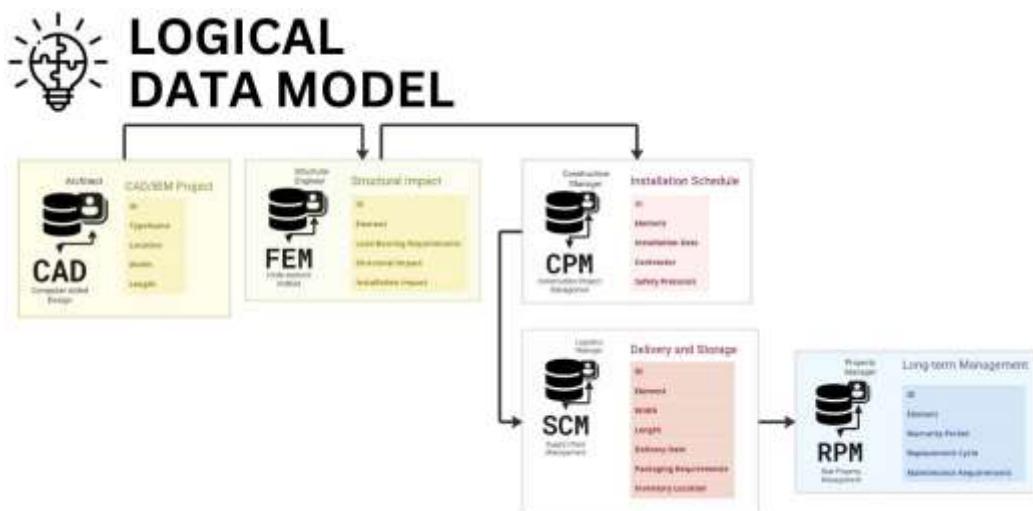


Fig. 4.3-3 Logical Data Model describes data types, relationships, and keys in detail, but without system implementation.

- **Physical Data Model:** This model describes the necessary structures for implementing a database, including tables, columns, and relationships. It focuses on database performance, indexing strategies, and physical storage to optimize physical database deployment.

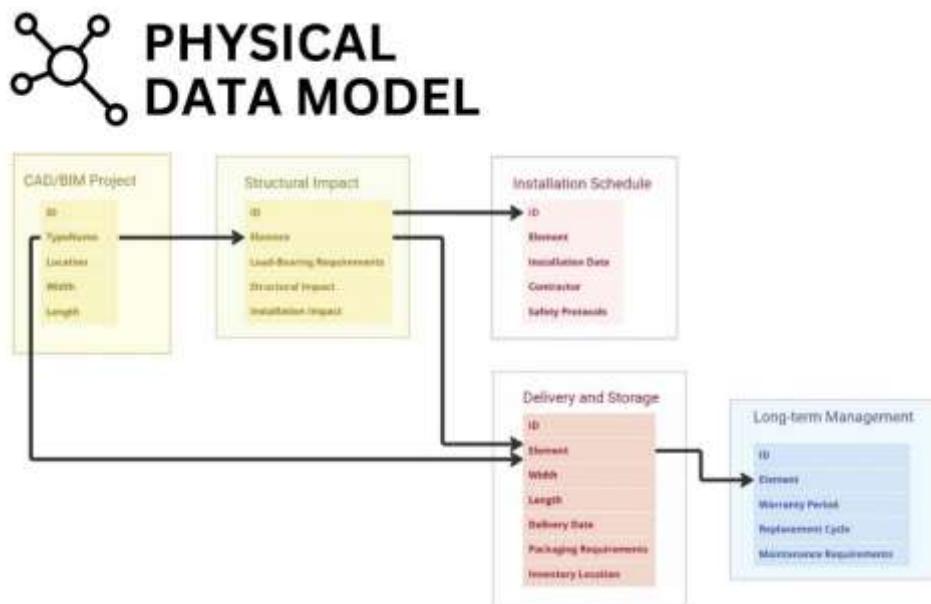


Fig. 4.3-4 The physical data model defines how the system will be implemented, including the tables and specific details of the database.

When designing databases and designing tabular relationships, understanding the levels of abstraction plays a key role in building an effective system architecture.

An effective data modeling methodology allows you to combine business objectives with technical implementation, making the entire process chain more transparent and manageable. Data modeling is not a one-time task, but a process involving sequential steps (Fig. 4.3-5):

- **Gathering business requirements:** key objectives, goals and information flows are defined. This is the stage of active interaction with experts and users.
- **Entity identification:** the main objects, categories and data types that are important to consider in the future system are highlighted.
- **Development of a conceptual and logical model:** first the key entities and their relationships are captured, then the attributes, rules and detailed structure.
- **Physical modeling:** the technical implementation of the model is designed: tables, fields, relationships, constraints, indexes.
- **Database creation:** the final step is to implement the physical model in the selected DBMS, conduct testing and prepare for operation.

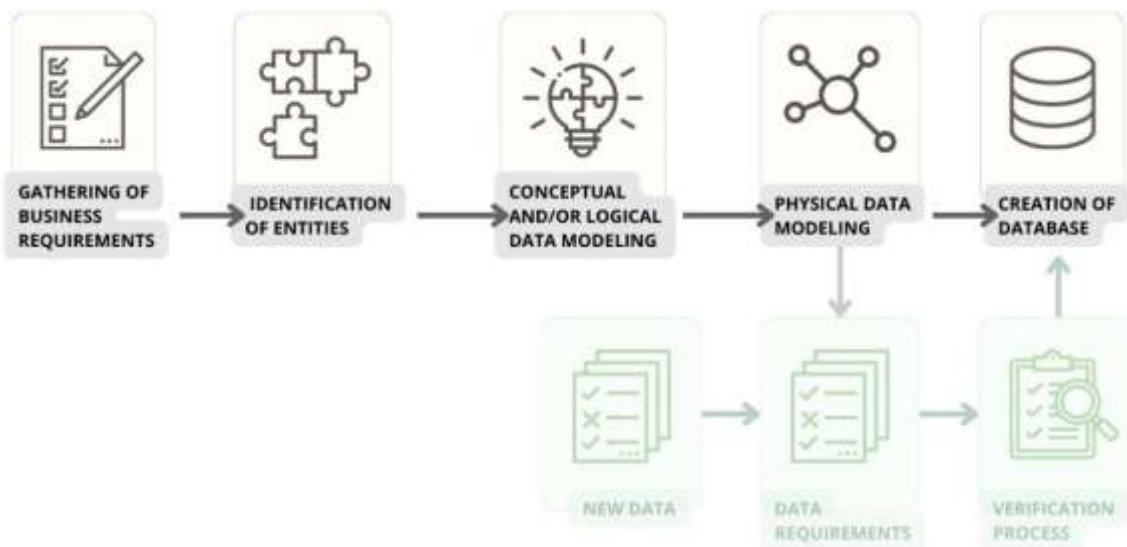


Fig. 4.3-5 Creating databases and data management systems for business processes begins with requirements generation and data modeling.

Properly designed data modeling processes allow for transparent information flows, which is especially important in complex projects such as construction project or site management. Let's look at how moving from a conceptual model to a logical model and then to a physical model can help streamline processes.

Practical data modeling in the context of construction

Let's take the construction site management task as an example of data modeling and transform the requirements of foreman into a structured logic model. Based on the basic needs of construction site management, we define key entities for: construction site (SITE), workers (WORKER), equipment (EQUIPMENT), tasks (TASK) and equipment utilization (EQUIPMENT_USAGE). Each entity contains a set of attributes that reflect important characteristics. For example, for a TASK, this could be a task description, due date, status, priority; for a WORKER, it could be the name, role on the site, current employment, etc.

The logic model establishes the relationships between these entities, showing how they interact with each other in actual work processes (Fig. 4.3-6). For example, the relationship between site and workers indicates that many workers can work on the same site, and the relationship between workers and tasks reflects that one worker can perform multiple tasks.

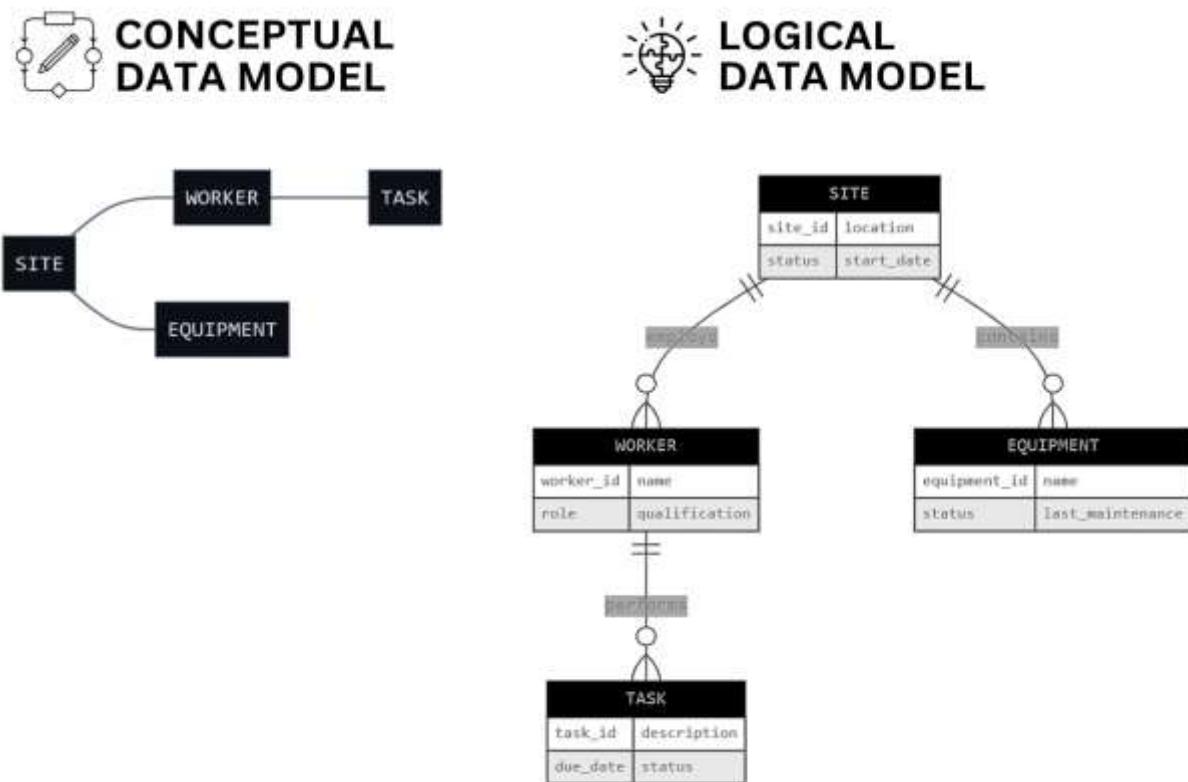


Fig. 4.3-6 Conceptual and logical data model generated by foreman requirements to describe construction site processes.

When moving to the physical model, technical implementation details are added: specific data types (VARCHAR, INT, DATE), primary and foreign keys for relationships between tables, and indexes to optimize database performance (Fig. 4.3-7).

For example, specific types with possible values should be defined for statuses, and indexes on key fields such as status and worker_id should be added to improve search performance. This turns a

logical description of the system into a concrete database implementation plan, ready to be created and implemented.

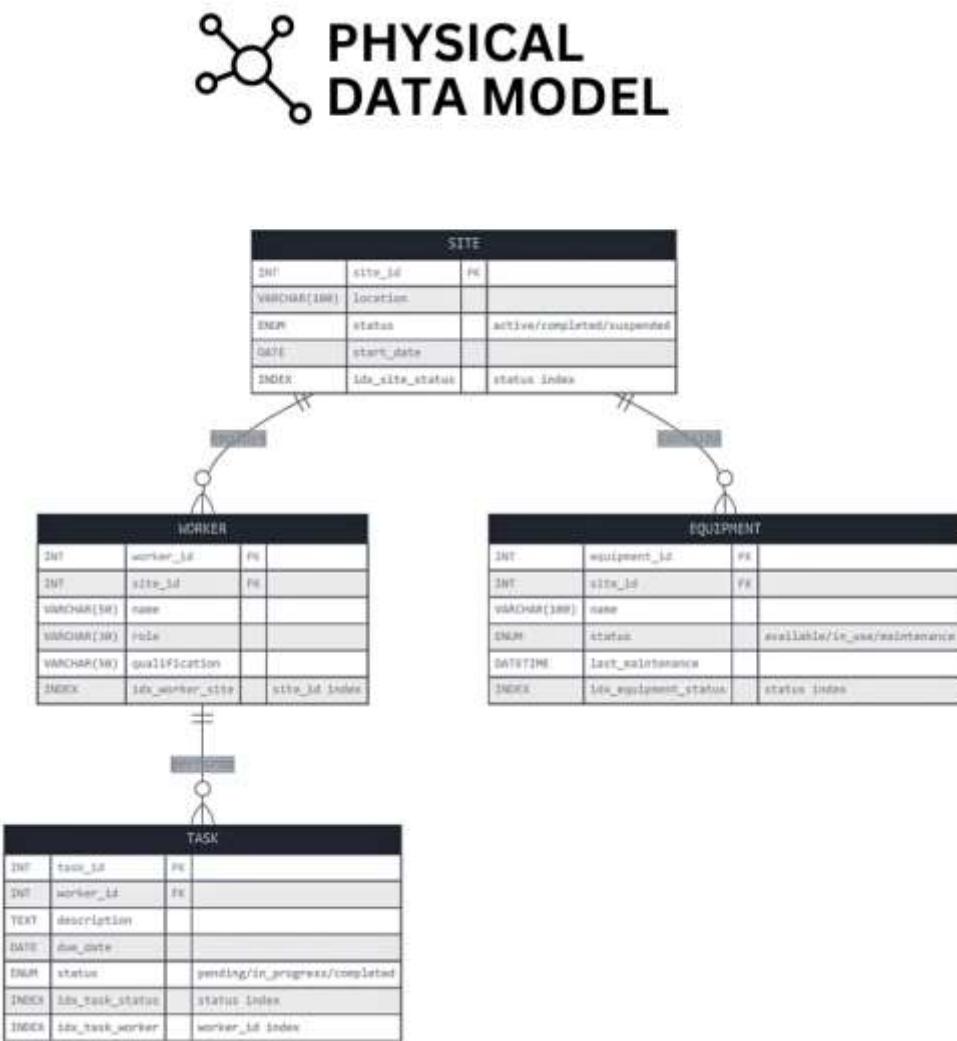


Fig. 4.3-7 The physical data model describes the entities of a construction site through the minimum required parameters.

The physical model often differs from the logical model. On average, the distribution of modeling time is as follows: about 50% is spent on the conceptual model (gathering requirements, discussing processes, identifying entities), 10% on the logical model (specifying attributes and relationships), and 40% on the physical model (implementation, testing, adapting to DBMS).

This balance is due to the fact that the conceptual stage lays the foundation for the data structure, while the logical model only specifies relationships and attributes. The physical model requires the most resources, as it is at this stage that data is implemented into specific platforms and tools.

Creating a database using LLM

Having a data model and description of entities through parameters, we are ready to create databases - storages, where we will store information coming after the structuring stage on specific processes.

Let's try to create an example of a simple but functional database with a minimum amount of code using SQLite using the Python programming language as an example. Relational databases were discussed in detail in the chapter "Structured relational databases and SQL query language".

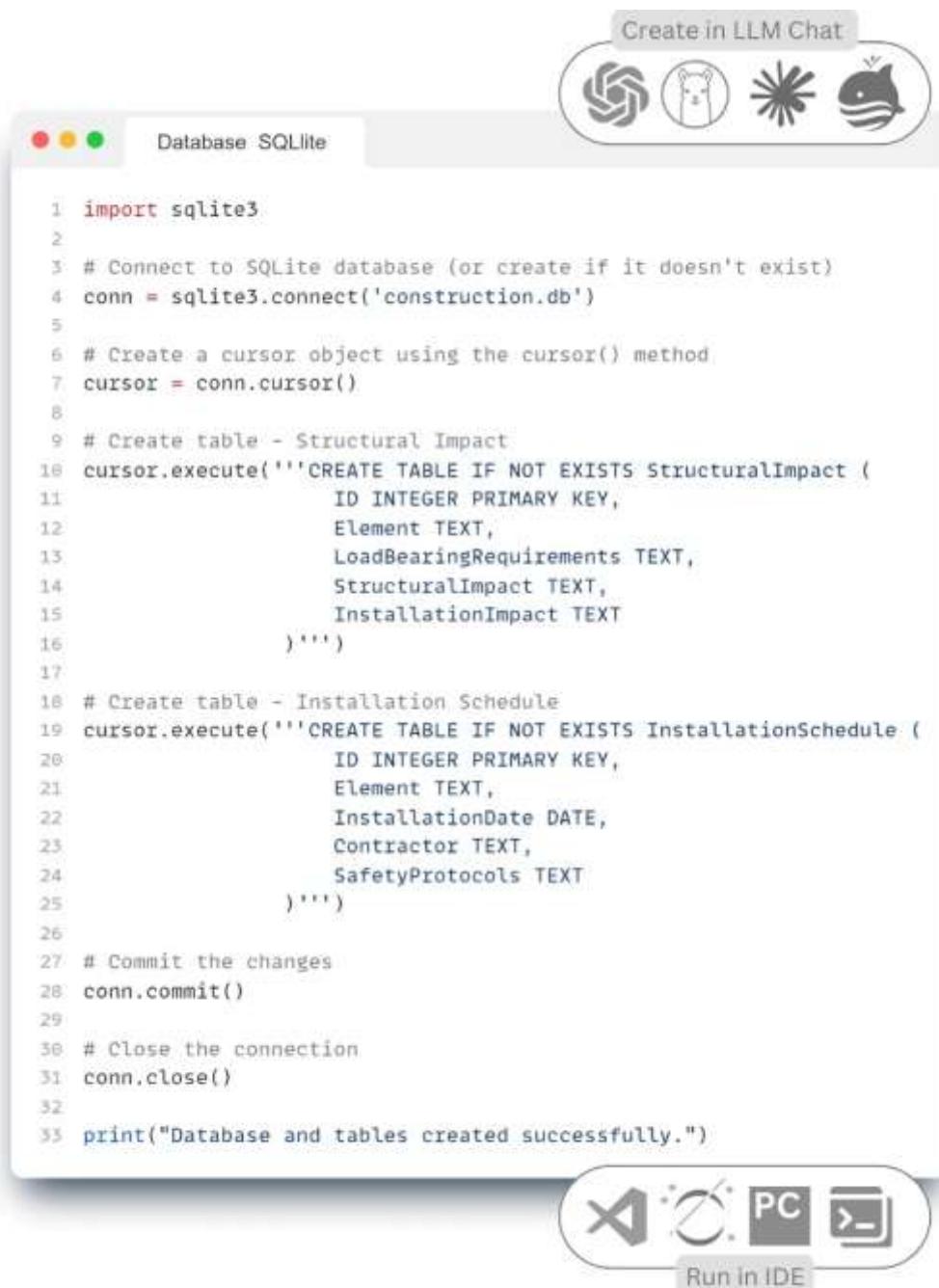
SQLite is a lightweight, open source, embeddable, cross-platform database, supporting core SQL features. SQLite is widely used in CAD (BIM), ERP, EPM and is present in almost every smartphone.

Let's ask the LLM model to create a small SQLlite database for us based on the physical model created above (Fig. 4.3-4) for the FEM (structural impact) and CPM (assembly schedule) systems.

- Send a text request to LLM chat (CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

Create SQLlite databases for the Constructive Impact and Installation Schedule systems from the Fig. in the attachment (paste the schema as an image in the chat (Fig. 4.3-4)) ↴

2 LLM's response:



```
1 import sqlite3
2
3 # Connect to SQLite database (or create if it doesn't exist)
4 conn = sqlite3.connect('construction.db')
5
6 # Create a cursor object using the cursor() method
7 cursor = conn.cursor()
8
9 # Create table - Structural Impact
10 cursor.execute('''CREATE TABLE IF NOT EXISTS StructuralImpact (
11         ID INTEGER PRIMARY KEY,
12         Element TEXT,
13         LoadbearingRequirements TEXT,
14         StructuralImpact TEXT,
15         InstallationImpact TEXT
16     )''' )
17
18 # Create table - Installation Schedule
19 cursor.execute('''CREATE TABLE IF NOT EXISTS InstallationSchedule (
20         ID INTEGER PRIMARY KEY,
21         Element TEXT,
22         InstallationDate DATE,
23         Contractor TEXT,
24         SafetyProtocols TEXT
25     )''' )
26
27 # Commit the changes
28 conn.commit()
29
30 # Close the connection
31 conn.close()
32
33 print("Database and tables created successfully.")
```



Fig. 4.3-8 The generated code creates databases and tables from the uploaded image, into which we can now load data.

This code for creating a database can be executed in the Python environment with support for SQLite, which is usually included in the standard Python environment. It can be run on any local Python machine capable of running scripts and creating files. Once run and executed, this code will create a file on your hard disk - a SQLite database named construction.db (similar to an Excel file with worksheets) that will contain the "Structural Impact" and "Installation Schedule" tables.

After creating a new **construction.db** database to these tables, data can be added to these tables via SQL _queries or import, which will allow you to start creating automatic data processing later. Data can be imported into the SQLite database from CSV files, Excel spreadsheets or exported from other databases and storages via API.

To establish sustainable data modeling and effective database management processes, a company needs a clearly defined strategy as well as coordination between technical and business teams. With disparate projects and multiple data sources, it is often difficult to ensure consistency, standardization and quality control at all levels. One key solution may be to create a dedicated Data Modeling Center of Excellence (CoE) within the company.

Center of Excellence (CoE) for Data Modeling

With data becoming one of the key strategic assets, companies need to do more than just collect and store information correctly - it is important to learn how to manage data systematically. The Center of Excellence for Classification and Data Modeling (CoE) is a structural unit that ensures consistency, quality and efficiency of all data handling in the organization.

The Center of Excellence (CoE) is the core of expert support and a methodological foundation for digital transformation in a company. It builds a data-driven culture and enables organizations to build processes that make decisions based on structured, validated and representative data rather than on intuition or local information.

A data center of excellence is usually formed from cross-functional teams that work according to the "two pizzas" principle. This principle, proposed by Jeff Bezos, means that the size of the team should be such that it can be fed with two pizzas, i.e. not to exceed 6-10 people. This approach helps to avoid excessive bureaucracy and increases the flexibility of work. The CoE team should include employees with a variety of technical skills, from data analytics and machine learning to expertise in specific business areas. With their deep technical knowledge, data engineers should not only optimize processes and model data, but also support colleagues by reducing time on routine tasks (Fig. 4.3-9).

Just as in nature ecosystem resilience is ensured by biodiversity, in the digital world flexibility and adaptability are achieved through a diversity of approaches to handling data. However, this diversity must be based on common rules and concepts.

A Center of Excellence (CoE) can be compared to the "climate conditions" of a forest ecosystem, which

determine which types of data will thrive and which will be automatically discarded. By creating a favorable "climate" for quality data, the CoE facilitates the natural selection of best practices and methodologies that later become organizational standards.

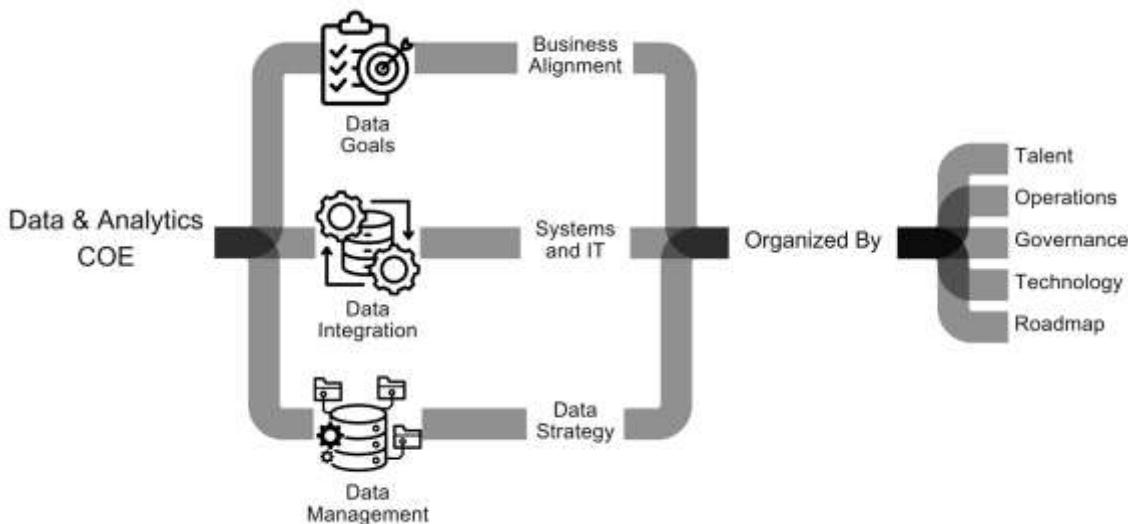


Fig. 4.3-9 The Center of Excellence (CoE) for Data and Analytics brings together expertise on key aspects of data management, integration, and strategy.

To accelerate integration cycles and achieve better results, CoE should provide its members with a sufficient degree of autonomy in decision-making. This is especially important in a dynamic environment where trial and error, constant feedback and frequent releases can bring significant benefits. However, this autonomy is only effective if there is clear communication and support from senior management. Without strategic vision and top-level coordination, even the most competent team can face barriers in implementing their initiatives.

It is the COE or senior management's responsibility to ensure that the data modeling approach is not limited to one or two projects, but is embedded in the overall information management and business process management system.

The Center of Expertise (CoE), in addition to tasks related to data modeling and Data Governance, is responsible for the development of common standards and approaches to the deployment and operation of the data infrastructure. It also creates a culture of continuous improvement, process optimization and efficient use of data in the organization (Fig. 4.3-10).

The systematic approach to data and model management within CoE can be roughly divided into several key blocks:

- **Process standardization and model lifecycle management:** CoE develops and implements methodologies to unify the creation and management of data models. This includes: establishing structural templates, quality control methods and version control systems to ensure data continuity across all phases of work.

- **Role management and responsibility assignment:** The COE defines key roles in the data modeling process. Each project participant is assigned clearly defined roles and areas of responsibility, which promotes teamwork and reduces the risk of data inconsistencies.
- **Quality control and auditing:** effective management of construction data requires continuous monitoring of its quality. Automated mechanisms are being introduced to check data, identify errors, missing attributes.
- **Metadata and Information Architecture Management:** CoE is responsible for creating a unified system of classification and identifiers, naming and entity description standards, which is critical for integration between systems.

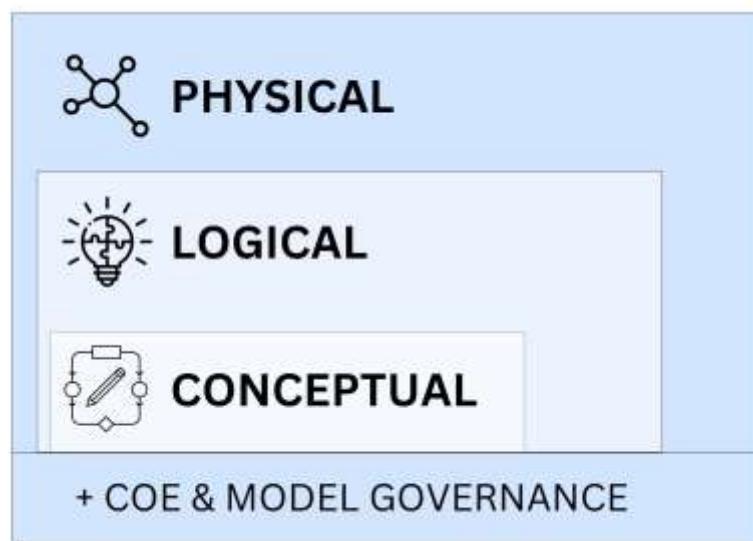


Fig. 4.3-10 Data modeling and data quality management is one of the main challenges of CoE

The Center of Excellence (CoE) for Data is not just a group of experts, but a systemic mechanism that creates a new data-driven culture and provides a unified approach to working with data throughout the company. Through competent integration of modeling processes into the overall information management system, standardization, classification and data quality control, CoE helps businesses to continuously improve their products and business processes, respond faster to market changes and make informed decisions based on reliable analytics.

Such centers are particularly effective when combined with modern DataOps principles - under a move that ensures continuous delivery, automation and quality control of data. We will talk more about DataOps in Part 8, in the chapter "Modern Data Technologies in the Construction Industry".

In the following chapters, we will move from strategy to practice - let's conditionally "transform" into a data center: we will look at several examples of how task parameterization, requirements gathering and the automatic validation process take place.



CHAPTER 4.4.

SYSTEMATIZATION OF REQUIREMENTS AND VALIDATION OF INFORMATION

Requirements gathering and analysis: transforming communications into structured data

Collecting and managing requirements is the first step to ensuring data quality. Despite the development of digital tools, most requirements are still formulated in an unstructured way: through letters, meeting minutes, phone calls and verbal discussions. This form of communication makes it difficult to automate, validate, and reuse information. In this chapter, we look at how to translate textual requirements into formal structures, ensuring that business requirements are transparent and systematic.

Gartner's research, "Data Quality: Best Practices for Accurate Insights," emphasizes the critical importance of data quality for successful data and analytics initiatives [100]. They note that poor data quality costs organizations an average of at least \$12.9 million annually and that reliable, high-quality data is essential to creating a data-driven company.

The lack of structured requirements leads to the fact that the same element (entity) and its parameters may be stored in different systems in different variations. This not only reduces the efficiency of processes, but also results in wasted time, duplication of information and the need to re-validate data before it can be used. As a result, even a single omission - a lost parameter or a single incorrectly described element - can slow down decision-making and cause inefficient use of resources.

*For want of a nail, the horseshoe was lost.
 For want of a horseshoe, the horse was lost.
 For want of a horse, the rider was lost.
 For want of a rider, a message was lost.
 For want of a message, the battle was lost.
 For want of a battle, a kingdom was lost.
 And all for the lack of a nail in the horseshoe.*

— Proverb [101]

Analyzing and gathering requirements for the process of filling and storing data starts with identifying all stakeholders. Just as the proverbial loss of a single nail leads to a chain of critical consequences, in business, the loss of a single stakeholder, an overlooked requirement or the loss of even a single parameter can significantly impact not only an individual business process, but the entire ecosystem of a project and the organization as a whole. Therefore, it is extremely important to identify even those

elements, parameters and roles that at first glance seem insignificant, but later may turn out to be critical to business sustainability.

Let's imagine that a company has a project in which the client puts forward a new request - "add an additional window on the north side of the building". The small process "*client's request to add a new window to the current project*" involves architect, client, CAD specialist (BIM), construction manager, logistics manager, ERP -analyst, quality control engineer, safety engineer, control manager and real estate manager.

Even a small process may involve dozens of different specialists. Each process participant must understand the requirements of the specialists with whom they are connected at the data level.

At the text level (Fig. 4.4-1), the communication between the client and the specialists in the process chain takes place as follows:

- ⦿ **Customer:** "We have decided to add an additional window on the north side for better lighting. Can this be realized?"
- ⦿ **Architect:** "Sure, I will revise the project to include the new window and send updated CAD plans (BIM)."
- ⦿ **CAD Specialist (BIM):** "Received a new project. I update the CAD (BIM) model with the additional window and after coordination with the FEM engineer provide the exact location and dimensions of the new window".
- ⦿ **Construction Manager:** "A new project has been received. We are adjusting the installation dates for 4D and informing all relevant subcontractors."
- ⦿ **Facilities Engineer (CAFM):** "I will enter the 6D data on the new window into the CAFM system for future facility management and maintenance planning."
- ⦿ **Logistics Manager:** "I need the dimensions and weight of the new window to organize the delivery of the window to the facility."
- ⦿ **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."
- ⦿ **Quality Control Engineer:** "Once the window specs are ready, I will make sure they meet our quality and material standards."
- ⦿ **Safety Engineer:** "I will be assessing the safety aspects of the new window, with a particular focus on compliance and evacuation scheme 8D".
- ⦿ **Controls Manager:** "Based on the exact scope of work from ERP, we will update our 4D timeline to reflect the installation of the new window, and store the new data in the project's content management system."
- ⦿ **Worker (installer):** "Need instructions on installation, assembly and timing of work. In addition, are there any special safety rules that I have to follow?"
- ⦿ **Property Manager:** "Once installed, I will document warranty and maintenance information for long-term management."

- ⌚ **Asset Manager:** "Equipment Engineer, please submit final data for asset tracking and lifecycle management."
- ⌚ **Client:** "Wait, maybe I'm in a hurry and the window won't be needed. Maybe I should make a balcony."

In such scenarios, which happen frequently, even a small change causes a chain reaction between multiple systems and roles. In this case, almost all communication at the initial stage is in text form: emails, chats, meeting minutes (Fig. 4.4-1).

In such a text-based communication system for a construction project, a system of legal confirmation and recording of all data exchange operations and all decisions made is very important. This is to ensure that every decision, instruction or change made is legally valid and traceable, reducing the risk of future "misunderstandings"

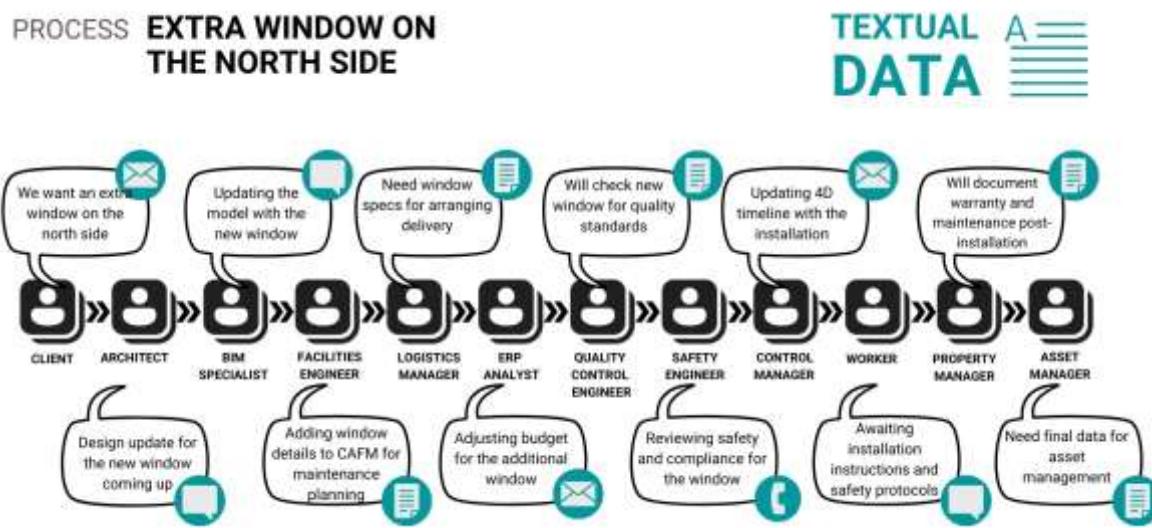


Fig. 4.4-1 Communication between client and contractor in the initial stages of a project often contains multi-format textual data.

The lack of legal control and validation of decisions in the relevant systems of a construction project can lead to serious problems for all involved. Every decision, order or change made without proper documentation and validation can lead to disputes (and litigation).

Legal fixation of all decisions in textual communication can only be ensured by a large number of signed documents, which will fall on the shoulders of the management, which is obliged to record all transactions. As a result, if every participant is required to sign documents for every action, the system loses flexibility and becomes a bureaucratic maze. Lack of transaction confirmations will not only delay project implementation, but may also lead to financial losses and deterioration of relations between the participants, up to and including legal problems.

Such a transaction approval process, which usually starts with text-based discussions, gradually evolves into a multiformat document exchange format in the following stages (Fig. 4.4-2), significantly

complicating the communication that used to take place only through text. Without well-defined requirements, automating such processes filled with multiformat data and a large number of textual requirements becomes almost impossible.

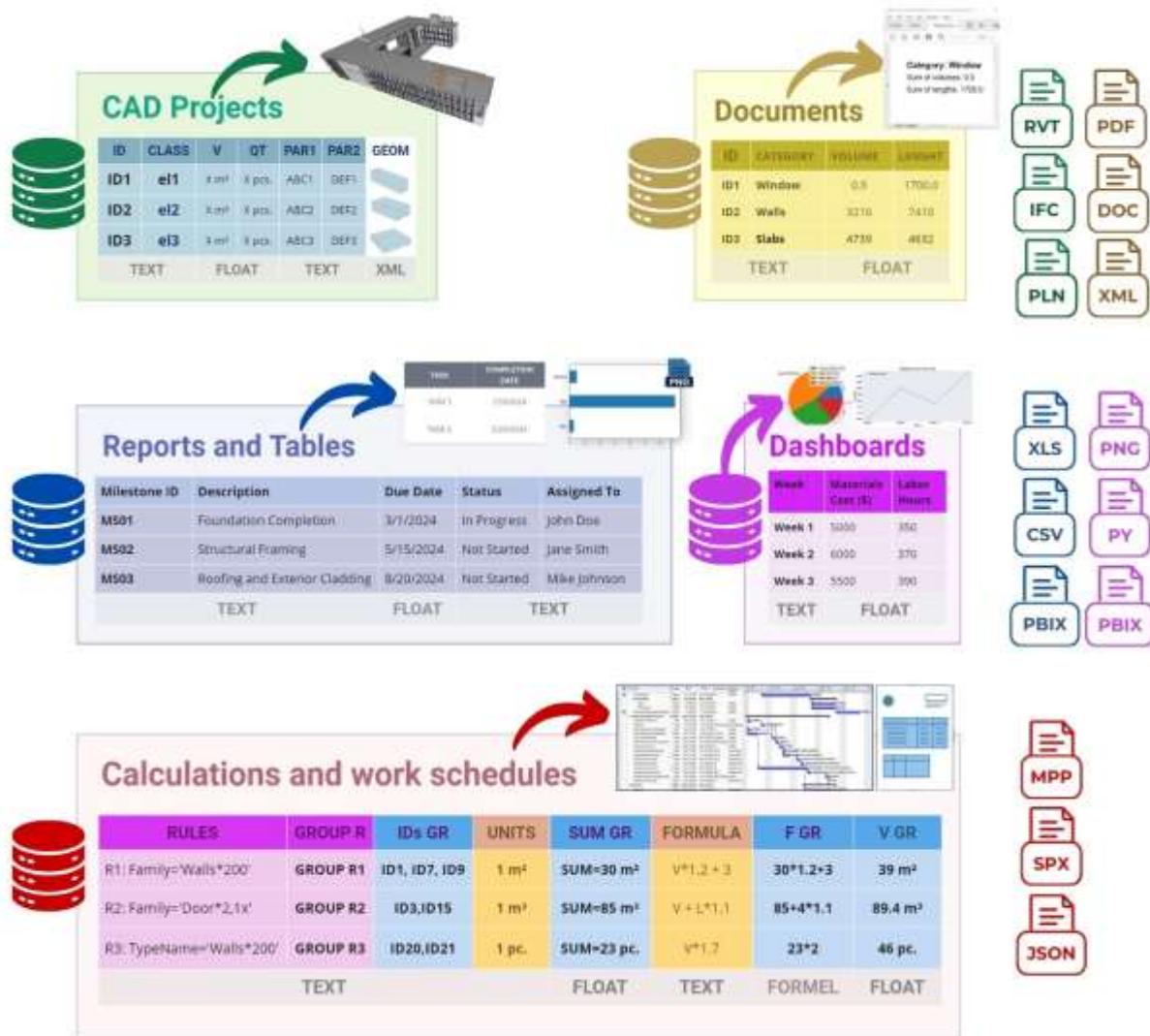


Fig. 4.4-2 Each system in the construction company landscape serves as a source of legally relevant documents in a variety of formats.

Text communications require each professional to either familiarize themselves with the full correspondence or attend all meetings on a regular basis to understand the current status of the project.

To overcome this limitation, a transition from textual communication to a structured requirements model is necessary. This is only possible through systematic analysis, process visualization, and description of interactions in the form of flowcharts and data models (Fig. 4.4-3). Just as in data modeling (Fig. 4.3-7), we moved from the contextual-idea level to the conceptual level by adding the systems and tools used by the participants and the links between them.

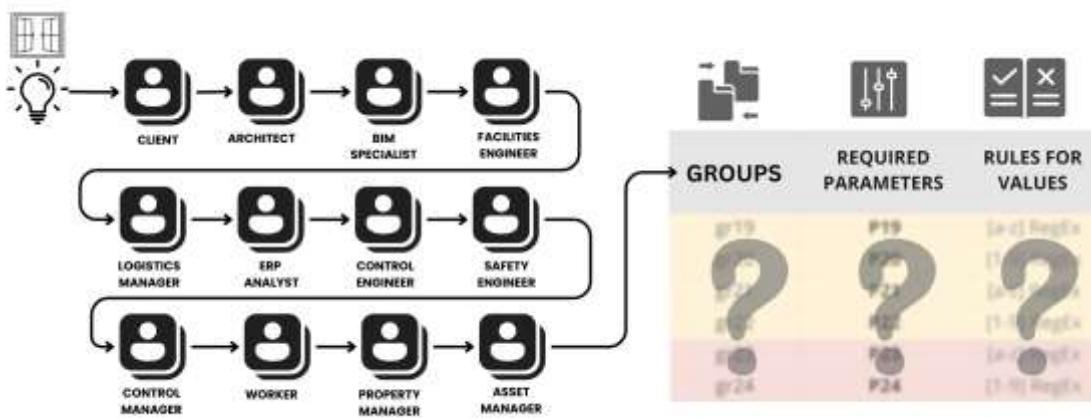


Fig. 4.4-3 To learn how to manage and automate the validation process, it is necessary to visualize processes and structure requirements.

The first step in systematizing requirements and relationships is to visualize all connections and relationships using conceptual flowcharts. The conceptual level will not only make it easier for all process participants to understand the entire process chain, but also visualize why and for whom the data (and requirements) are needed at each process step.

Process flowcharts and the effectiveness of conceptual frameworks

To bridge the gap between traditional and modern approaches to data management, companies need to consciously move from fragmented textual descriptions to structured process representations. The evolution of data - from clay tablets to digital ecosystems - requires new thinking tools. And one such tool is conceptual modeling using flowcharts. Creating visual diagrams - flowcharts, process diagrams, interaction diagrams - allows project participants to realize how their actions and decisions affect the entire decision-making system.

If processes require more than just storing data, but analyzing or automating it, then you need to start addressing the topic of creating a conceptual-visual level of requirements.

In our example (Fig. 4.4-1), each specialist can be part of a small team, but also of a larger department, including up to a dozen experts under the control of a general manager. Each department uses a specialized application database (Fig. 1.2-4 e.g. ERP, CAD, MEP, CDE, ECM, CPM etc.), which is regularly updated with incoming information needed to create documents, record the legal status of decisions and manage processes.

The transaction process is similar to the work of ancient managers 4,000 years ago, when clay tablets and papyrus were used to legally confirm decisions. The difference between modern systems and their clay and paper predecessors is that modern methods additionally include the process of converting textual information into digital form for further automatic processing in other systems and tools.

Creating a visualization of the process in the form of conceptual flowcharts will help describe each step and the interactions between different roles, making a complex workflow clear and simple.

Visualization of processes ensures that the process logic is transparent and accessible to all team members.

The same communicative process for adding a window to a project that was described in the form of text, messages (Fig. 4.4-1), and block diagram is similar to the conceptual model we discussed in the chapter on data modeling (Fig. 4.4-4).

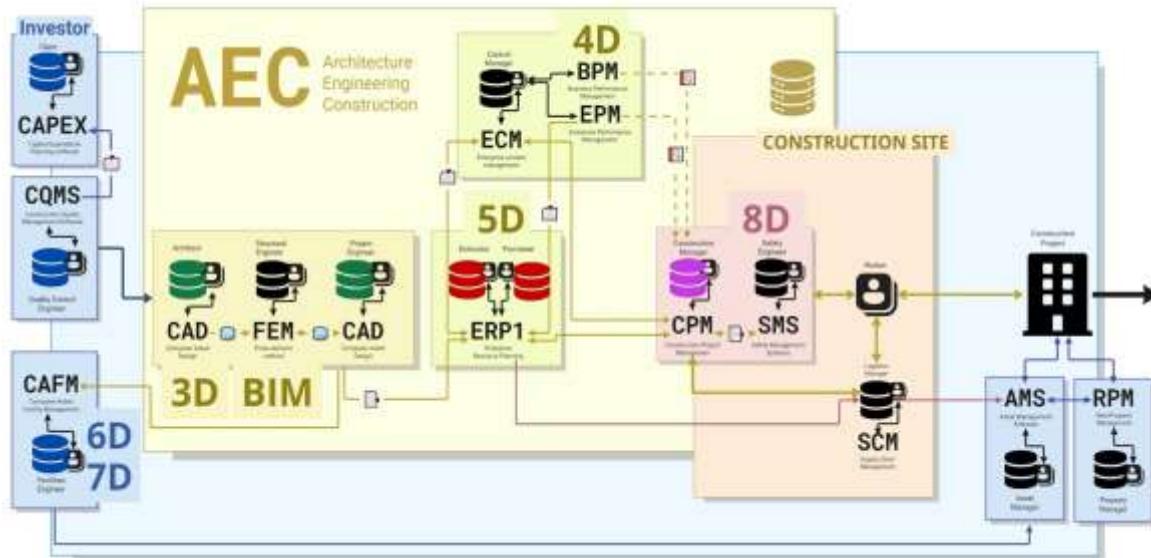


Fig. 4.4-4 The conceptual diagram shows the project participants as users of a database where their queries link the various systems.

Although conceptual diagrams are an important step, many companies limit themselves to this level, believing that a visual diagram is sufficient for understanding processes. This creates the illusion of manageability: it is easier for managers to perceive the big picture on such a flowchart, to see connections between participants and stages. However, such schemes do not give a clear idea of what data is necessary for each participant, in what format it should be transmitted and what parameters and attributes are mandatory for the realization of automation. A conceptual flowchart is more like a route map: it indicates who interacts with whom, but it does not reveal what is transferred in these interactions.

Even if a process is described in detail at the conceptual level using flowcharts, this does not guarantee its effectiveness. Visualization often simplifies the work of managers, allowing them to track the process more easily with step-by-step reporting. However, for database engineers, the conceptual representation may not be clear enough and may not provide a clear understanding of how to implement the process at the parameter and requirements level.

As we move toward more complex data ecosystems, the initial implementation of conceptual and visual tools becomes critical to ensure that data processes are not only efficient, but also aligned with the organization's strategic goals. To fully translate this window addition process (Fig. 4.4-1) to the data requirements level, we need to go a level deeper and translate the conceptual visualization of the process to the logical and physical level of data, required attributes, and their boundary values.

Structured Requirements and RegEx regular expressions

Up to 80% of data created in companies is in unstructured or semi-structured formats [52] - text, documents, letters, PDF -files, conversations. Such data (Fig. 4.4-1) is difficult to analyze, verify, transfer between systems and use in automation.

To ensure manageability, transparency, and automatic validation, it is necessary to translate textual and semi-structured requirements into well-defined, structured formats. The structuring process concerns not only the data (which we discussed in detail in the first chapters of this part of the book), but also the requirements themselves, which project participants usually formulate in free text form throughout the project lifecycle, often without thinking that these processes can be automated.

Just as we have already converted data from an unstructured textual form to a structured form, in the requirements workflow we will convert textual requirements to a structured "logical and physical layer" format.

Within the example of adding a window (Fig. 4.4-1), the next step is to describe the data requirements in tabular form. We will structure the information for each system used by the project participants by specifying key attributes and their boundary values

Consider, for example, one such system (Fig. 4.4-5) - Construction Quality Management System (CQMS), which is used by the quality control engineer on the client's side. With its help he checks whether a new element of the project - in this case "new window" - complies with the established standards and requirements.

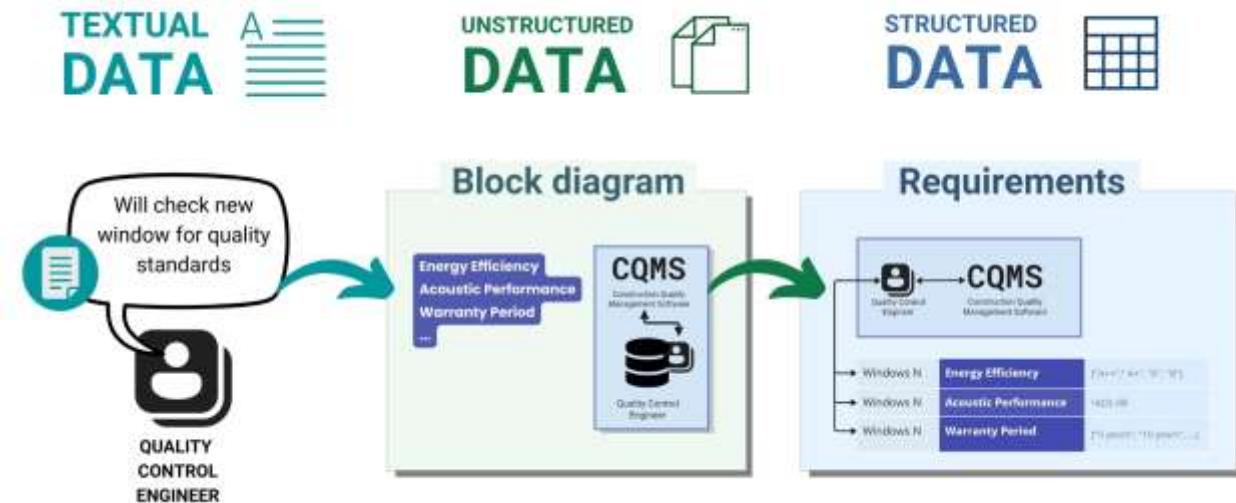


Fig. 4.4-5 Converting textual requirements into a table format with descriptions of entity attributes simplifies understanding for other specialists.

As an example, consider some important requirements for attributes of entities of type "window systems" in CQMS -system (Fig. 4.4-6): energy efficiency, acoustic performance and warranty period. Each category includes certain standards and specifications that need to be considered when designing and installing window systems.



Fig. 4.4-6 The Quality Control Engineer should inspect new Window Type elements for energy efficiency, sound insulation, and warranty standards.

The data requirements that a QA engineer specifies in a table have, for example, the following boundary values:

- **The energy efficiency class of windows** ranges from "A++", denoting the highest efficiency, to "B", considered the minimum acceptable level, and these classes are represented by a list of acceptable values ["A++", "A+", "A", "A", "A", "B"].

- **The acoustic insulation of windows**, measured in decibels and indicating their ability to reduce street noise, is defined by the regular expression `\d{2}dB`.
- **The "Warranty Period" attribute** for the Window Type entity starts at five years, setting this period as the minimum allowed when selecting a product; warranty period values such as `"5 years", "10 years", etc.` or the logical condition >5 (years) are also specified.

According to the collected requirements, within the established attributes, new window category or class elements with grades below "B" such as "C" or "D" will not pass the energy efficiency test. Acoustical insulation of windows in data or documents to the QA Engineer shall be labeled with a two-digit number followed by the postfix "dB", such as "35 dB" or "40 dB", and values outside this format such as "9 D B" or "100 decibels" will not be accepted (as they will not pass the pattern for RegEx strings). The warranty period must begin with a minimum of "5 years" and windows with shorter warranty periods such as "3 years" or "4 years" will not meet the requirements that the Quality Engineer has described in the table format.

To check such attribute-parameter values against boundary values from requirements in the validation process, we use either a list of allowed values (`["A", "B", "C"]`), dictionaries (`["A": "H1", "H2"; "B": W1, "W2"]`), logical operations (e.g., `>`, `<`, `<=`, `>=`, `==`) for numeric values) and regular expressions (for string and text values such as in the "Acoustic Performance" attribute). Regular expressions are an extremely important tool when working with string values.

Regular expressions (RegEx) are used in programming languages, including Python (Re library), to find and modify strings. Regex is like a detective in the string world, able to identify text patterns in text with precision.

In regular expressions, letters are described directly using the corresponding alphabet characters, while numbers can be represented using the special character `\d`, which corresponds to any digit from 0 to 9. Square brackets are used to indicate a range of letters or digits, e.g., `[a-z]` for any lowercase letter of the Latin alphabet or `[0-9]`, which is equivalent to `\d`. For non-numeric and non-letter characters, `\D` and `\W` are used, respectively.

Popular RegEx use cases (Fig. 4.4-7):

- **Verifying email address:** to check if a string is a valid email address, you can use the template `^ [a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}$`.
- **Date Extraction:** `\b\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\d{2}\d{4}\b` template can be used to extract date from text in DD.MM.YYYY format.
- **Verifying phone numbers:** to verify phone numbers in the format +49(000)000-0000, the pattern will look like `+\d{2}(\d{3})\d{3}-\d{4}`.

By translating the requirements of a QA engineer into the format of attributes and their boundary values (Fig. 4.4-6), we have transformed them from their original text format (conversations, letters, and regulatory documents) into an organized and structured table, thus making it possible to automati-

cally check and analyze any incoming data (e.g., new elements of the Window category). The presence of requirements allows for the automatic discarding of data that has not been checked, while the checked data is automatically transferred to the systems for further processing.

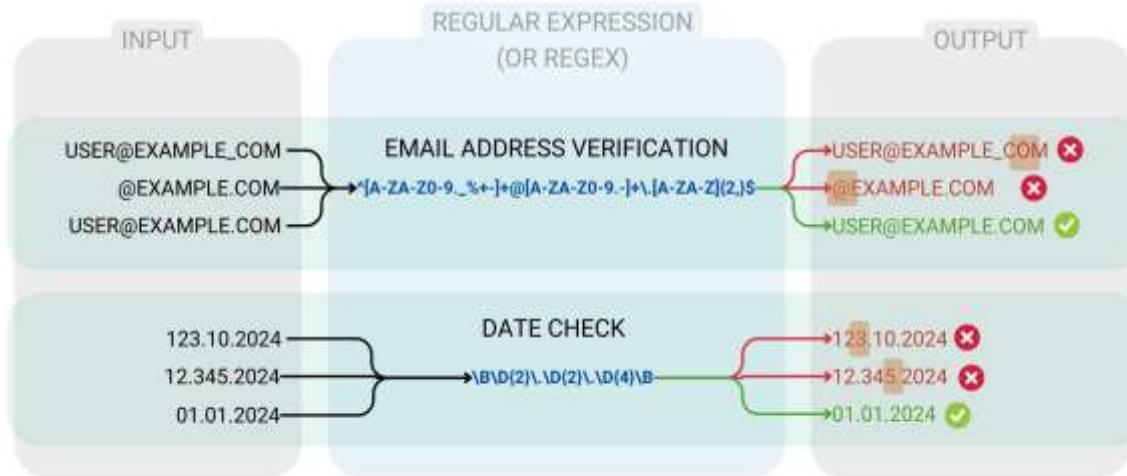


Fig. 4.4-7 The use of regular expressions is an extremely important tool in the text data validation process.

Now, moving from the conceptual to the logical level of working with requirements, we will convert all requirements of all specialists in our process of installing a new window (Fig. 4.4-4) into an organized list in attribute format and add these lists with the necessary attributes to our flowchart for each specialist (Fig. 4.4-8).

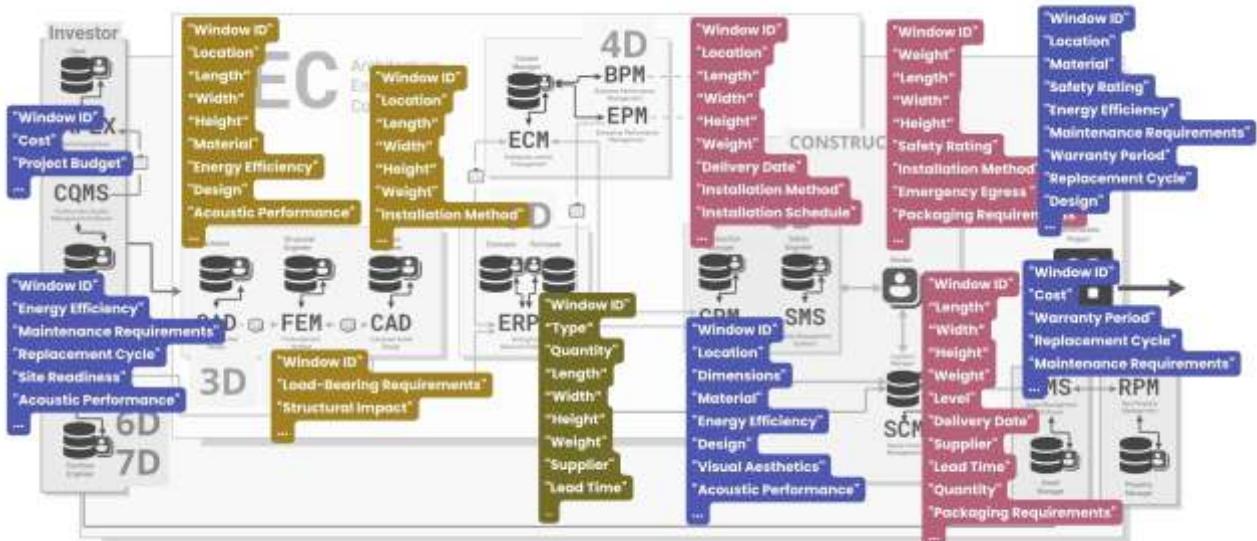


Fig. 4.4-8 At the logical process level, the attributes that each technician handles are added to their respective systems.

By adding all attributes to one common process table, we transform the information previously presented as text and dialogue at the conceptual level (Fig. 4.4-1) into the structured and systematized

form of physical-level tables (Fig. 4.4-9).

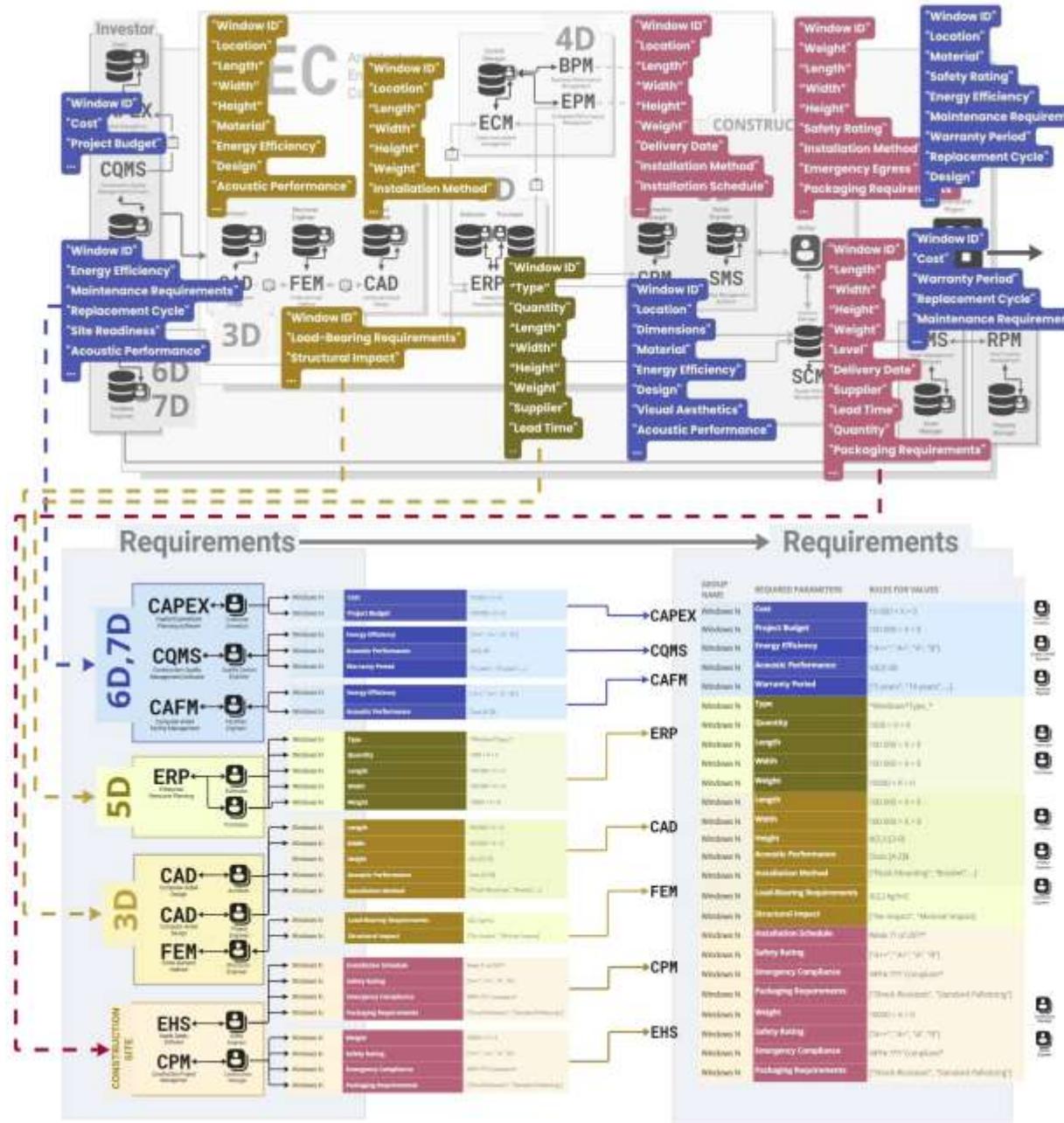


Fig. 4.4-9 Converting unstructured expert dialog into structured tables helps to understand requirements at the physical level.

Now the data requirements need to be communicated to the specialists creating information for specific systems. For example, if you are working in a CAD database, before you start modeling elements, you should collect all the necessary parameters based on the end use scenarios of this data. This usually starts with the operational phase, followed by the construction site, the logistics department, the estimating department, the structural calculations department, and so on. Only after you have

taken into account the requirements of all these links can you start creating data - based on the parameters collected. This will allow you to automate the verification and transfer of data along the chain.

When new data meets the requirements, it is automatically integrated into the company's data ecosystem, going directly to the users and systems for which it was intended. Verification of data against attributes and their values ensures that the information meets the required quality standards and is ready to be applied to company scenarios.

The data requirements have been defined and now, before verification can begin, the data to be verified must be created, obtained or collected, or the current state of information in databases must be recorded to be used in the verification process.

Data collection for the verification process

Before starting validation, it is important to make sure that the data are available in a form suitable for the validation process. This means not just having the information available, but preparing it: the data must be collected and transformed from unstructured, loosely structured, textual, and geometric formats into a structured form. This process is described in detail in the previous chapters, where methods for transforming different types of data were discussed. As a result of all transformations, the incoming data takes the form of open structured tables (Fig. 4.1-2, Fig. 4.1-9, Fig. 4.1-13).

With the requirements and structured tables with the necessary parameters and boundary values (Fig. 4.4-9), we can start validating the data - either as a single automated process (Pipeline) or as a step-by-step validation of each incoming document.

In order to start the check, it is required either to receive a new file as input or to fix the current state of the data - to create a snapshot or export current and incoming data, or to set up a connection to an external or internal database. In the example under consideration, such a snapshot is created by automatically converting CAD data from into a structured format recorded at, say, 23:00:00 on Friday, March 29, 2024, after all designers have gone home.



Fig. 4.4-10 CAD database snapshot (BIM) showing the current attribute information for a new entity of class "Window" in the current version of the project model.

Thanks to the reverse engineering tools discussed in the chapter "Translating CAD data (BIM) into a structured form", this information from different CAD (BIM) tools and editors can be organized into separate tables (Fig. 4.4-11) or combined into one common table connecting different sections of the project (Fig. 9.1-10).

Such table - database displays unique identifiers of windows and doors (ID attribute), type names (TypeName), dimensions (Width, Length), materials (Material), as well as indicators of energy and acoustic efficiency, and other characteristics. Such a table filled in CAD program (BIM) is collected by a design engineer from various departments and documents, forming an information model of the project.



Fig. 4.4-11 Structured data from CAD systems can be a two-dimensional table with columns denoting attributes of elements.

Real CAD (BIM) projects include tens or hundreds of thousands of elements (Fig. 9.1-10). Elements within CAD formats are automatically categorized by type and category, from windows and doors to slabs, floors and walls. Unique identifiers (e.g., native ID, which is set automatically by the CAD solution) or type attributes (Type Name, Type, Family) allow the same object to be tracked in different systems. For example, a new window on the north wall of a building can be uniquely identified through a single ID "W-NEW" in all relevant systems of the organization.

While entity names and identifiers should be consistent across all systems, the set of attributes and values associated with these entities can vary significantly depending on the context of use. Architects, structural engineers, construction, logistics, and real estate operations professionals perceive the same elements in different ways. Each of them relies on their own classifiers, standards and objectives: some consider the window from a purely aesthetic point of view, evaluating its shape and proportions, while others consider it from an engineering or operational point of view, analyzing thermal conductivity, installation method, weight or maintenance requirements. Therefore, when modeling data and describing elements, it is important to take into account the versatility of their use and ensure consistency of data while taking into account industry specifics.

For each role in the company's processes there are specialized databases with their own user interface - from design and calculations to logistics, installation and building operation (Fig. 4.4-12). Each such system is managed by a professional team of specialists through a special user interface or through database queries, where the sum of all decisions made on the entered values at the end of the chain is followed by the system manager or department manager, who is responsible for the legal validity and quality of the entered data before their counterparts serving other systems.

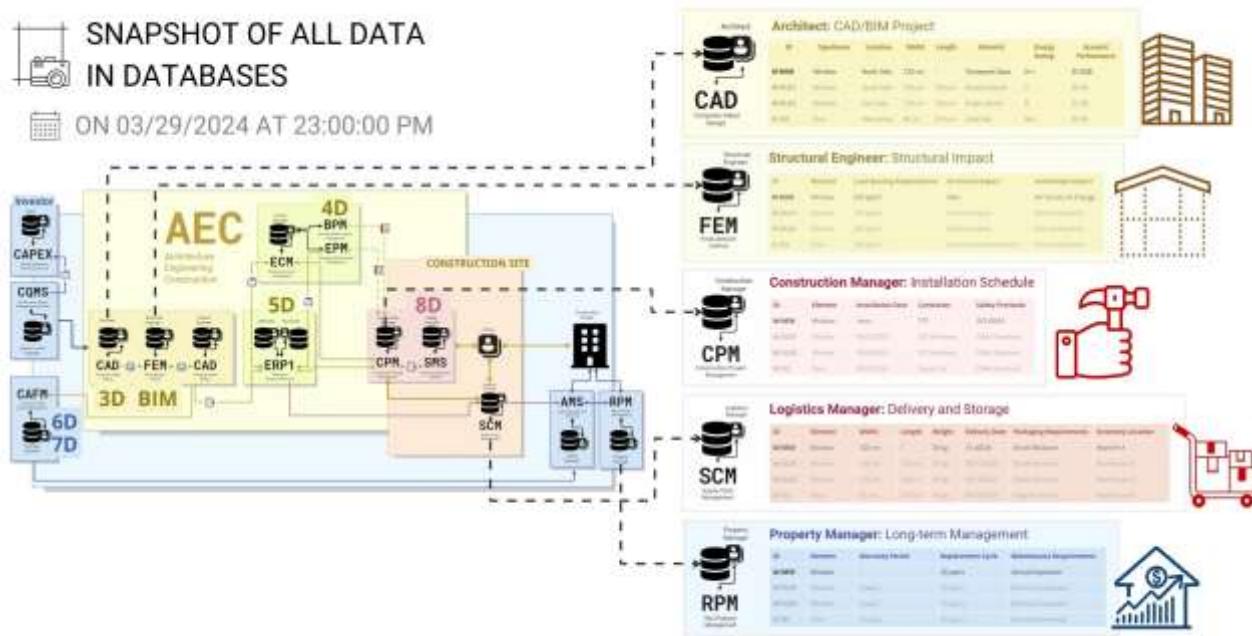


Fig. 4.4-12 The same entity has the same identifier in different systems, but different attributes that are important only in that system.

Once we have organized the collection of structured requirements and data at the logical and physical level, it remains for us to set up a process to automatically validate the data from different incoming documents and different systems against the previously collected requirements.

Verification of data and results of verification

All new data entering the system - be it documents, tables or database entries from the client, architect, engineer, foreman, logistician or property manager - must be validated against the requirements formulated earlier (Fig. 4.4-9). The validation process is critical: any errors in the data can lead to incorrect calculations, schedule delays, and even financial losses. To minimize such risks, it is necessary to organize a systematic and repeatable, iterative data validation procedure.

To validate new data entering the system - unstructured, textual or geometric - it must be converted into a loosely structured or structured format. The validation process must then check the data against the full list of required attributes and their allowed values.

Converting different types of data: text, images, PDF -documents and mixed CAD (BIM) data into a structured form was discussed in detail in the chapter "Converting data into a structured form".

An example is a table obtained from CAD (BIM) of a project (Fig. 4.4-11). It includes semi-structured geometric data and structured attribute information on project entities (Fig. 3.1-14) - for example an element from the class "Windows".

To perform the validation, we compare the attribute values (Fig. 4.4-11) with the reference boundary values that were defined by experts in the form of a requirement (Fig. 4.4-9). The final comparison table (Fig. 4.4-13) will allow us to understand which values are acceptable and which ones need to be corrected before the data can be used outside of CAD applications (BIM).

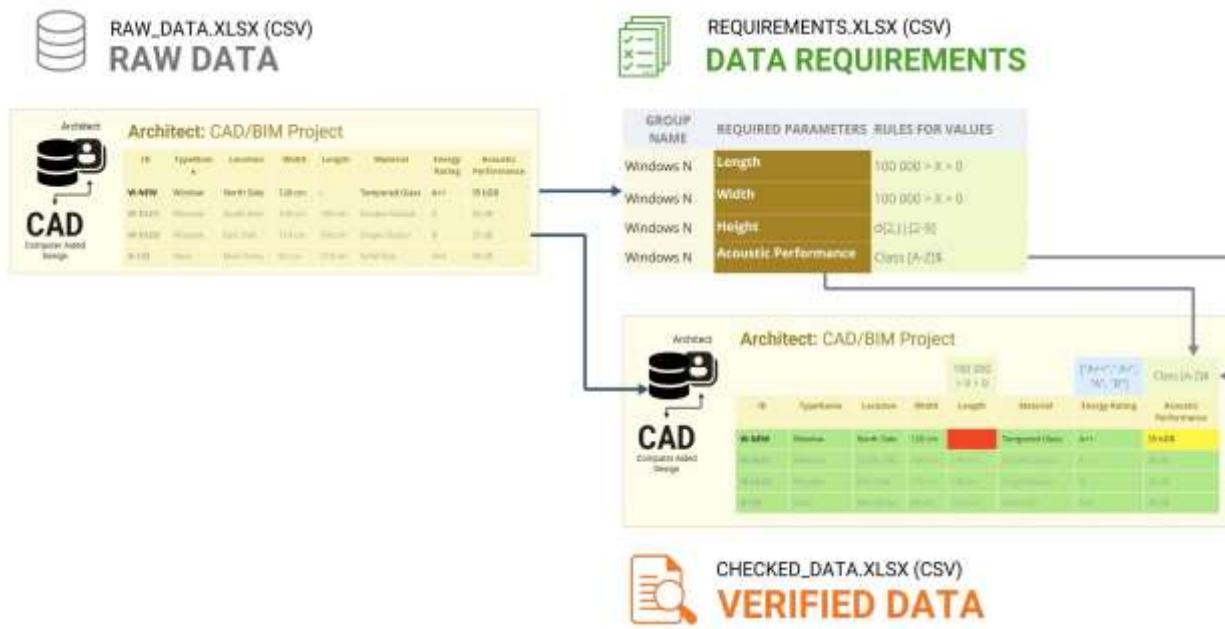


Fig. 4.4-13 The final validation table highlights those attribute values for the new entity of class "Windows" that you should pay attention to.

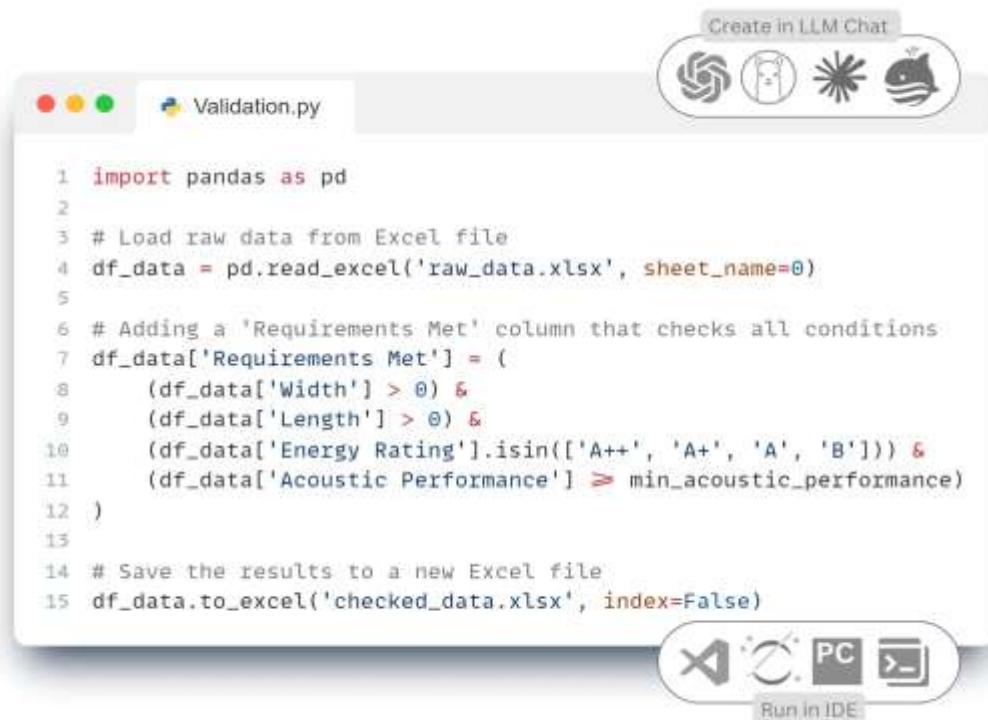
Implementing a similar solution using the Pandas library, which we described earlier in the chapter "Pandas: An Indispensable Tool for Data Analysis", we will validate data from a tabular file extracted from a CAD file (BIM) (RVT, IFC, DWG, NWS, DGN) (Fig. 4.4-11) using requirements from another tabular requirements file (Fig. 4.4-9).

To get the code, we need to describe in the prompt for LLM that we need to load the data from the file **raw_data.xlsx** (a complete set of data from the CAD database (BIM)), **check** them and save the result in a new file **checked_data.xlsx** (Fig. 4.4-13).

- 💡 Let's get the code using LLM without mentioning the Pandas library:

Write code to check the table from the raw_data.xlsx file and validate them using the following validation rules: the values of the 'Width' and 'Length' columns are greater than zero, 'Energy Rating' is included in the ['A++', 'A+', 'A', 'B'] list, and 'Acoustic Performance' as a variable that we'll specify later - with the final validation column added, and save the final table to a new Excel file checked_data.xlsx ↵

- 💡 LLM's response will describe a short example of Python code that can be refined and augmented by subsequent prompts:



```

1 import pandas as pd
2
3 # Load raw data from Excel file
4 df_data = pd.read_excel('raw_data.xlsx', sheet_name=0)
5
6 # Adding a 'Requirements Met' column that checks all conditions
7 df_data['Requirements Met'] = (
8     (df_data['Width'] > 0) &
9     (df_data['Length'] > 0) &
10    (df_data['Energy Rating'].isin(['A++', 'A+', 'A', 'B'])) &
11    (df_data['Acoustic Performance'] >= min_acoustic_performance)
12 )
13
14 # Save the results to a new Excel file
15 df_data.to_excel('checked_data.xlsx', index=False)

```

Fig. 4.4-14 The code generated by the LLM- model checks the converted CAD (BIM) design against the attribute requirements in the form of boundary values.

The code generated by the LLM language model, can be used in any popular IDE or online tool: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse with PyDev plugin, Thonny, Wing IDE, IntelliJ IDEA with Python plugin, JupyterLab or popular online tools Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

Execution of the code (Fig. 4.4-14) will show that the "entity elements" W-OLD1, W-OLD2, D-122 (and other elements) from the CAD database (BIM) meet the attribute requirements: width and length are

greater than zero, and the energy efficiency class is one of the list values 'A++', 'A', 'B', 'C' (Fig. 4.4-15).

The W-NEW element we need and recently added, which is responsible for the new element class "Window" on the north side, is not compliant (attribute "Requirements Met") because its length is zero (a value of "0.0" is considered unacceptable by our 'Width'>0 rule) and it does not specify an energy efficiency class.



CHECKED_DATA.XLSX (CSV)

VERIFIED DATA

ID	TypeName	Location	Width	Length	Material	Energy Rating	Acoustic Performance	Requirements Met	
0	W-NEW	Window	North Side	120	0.0	Tempered Glass		35	False
1	W-OLD1	Window	South Side	100	140.0	Double Glazed	A++	30	True
2	W-OLD2	Window	East Side	110	160.0	Single Glazed	B	25	True
3	D-122	Door	Main Entry	90	210.0	Solid Oak	B	30	True

Fig. 4.4-15 Checking identifies entities that have not gone through the verification process and adds a new attribute with values 'False' or 'True' to the results.

Similarly, we check the consistency of all project elements (entities) and required attributes for each of the systems, tables or databases in all the data we receive from different specialists (Fig. 4.4-1) during the process of adding a window to the project.

In the final table it is convenient to highlight the results of the check by color for visualization: attributes that successfully passed the check are marked in green, yellow - values with non-critical deviations, and red - critical inconsistencies (Fig. 4.4-16).

As a result of the validation (Fig. 4.4-16), we get a list of trusted and validated elements with their IDs that have been verified to meet attribute requirements. Validated elements provide assurance that these elements meet the stated standards and specifications for all systems involved in the process of adding elements of the Window class or any other class (we will talk more about automating data validation and creating an automated ETL process in the chapter "Automating ETL and Data Validation").

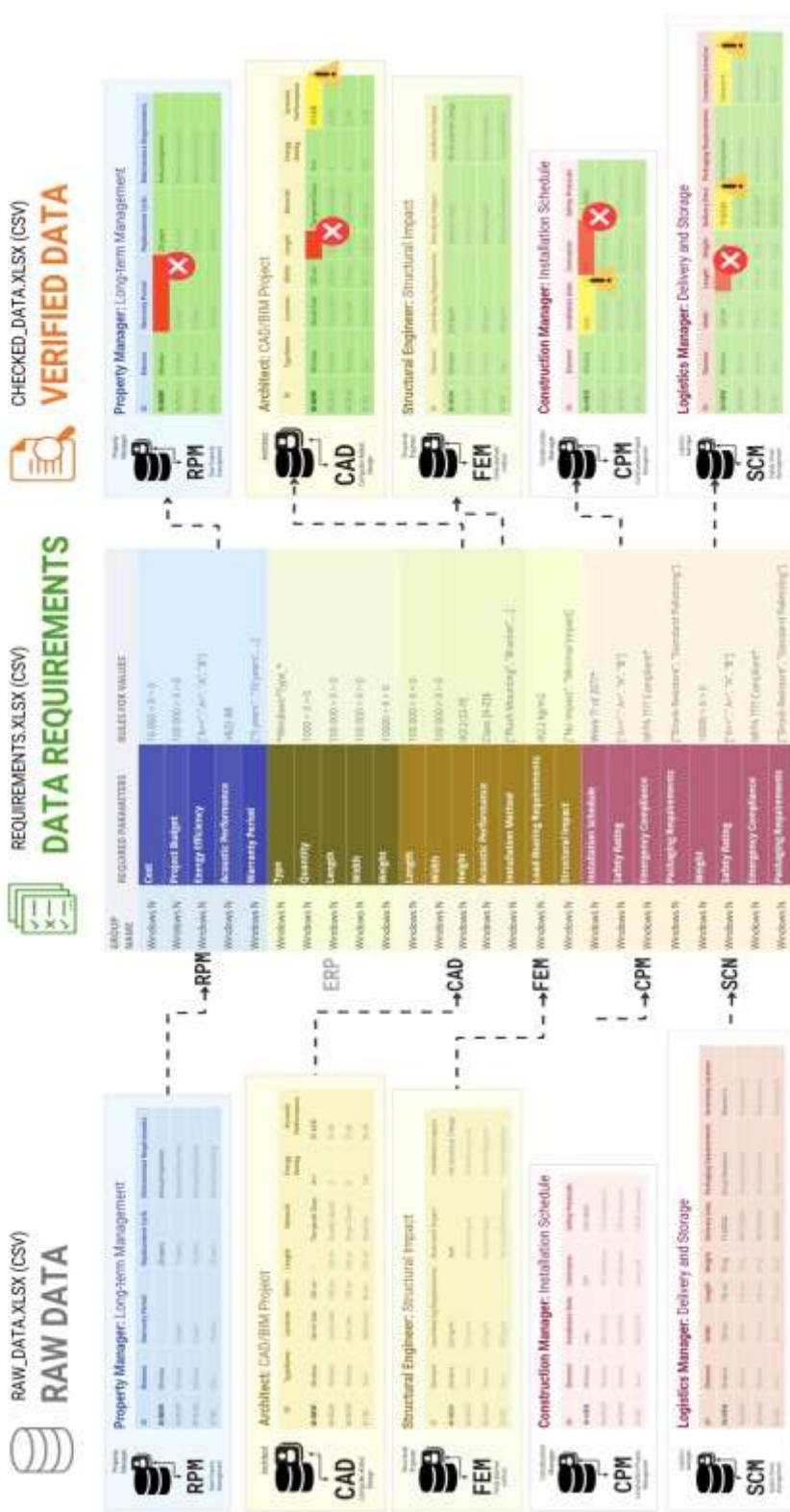


Fig. 4.4-16 The result of the check performed for all systems allows to determine which data does not meet the company's requirements.

Entities that have been successfully validated usually do not require much attention. They move on to the next stages of processing and integration into other systems without obstacles. In contrast to "quality" items, it is the items that fail validation that are of greatest interest. Information about such deviations is critical: it should be communicated not only in the form of tabular reports, but also using various visualization tools. Graphical representation of inspection results helps to quickly assess the overall state of data quality, identify problem areas and promptly take measures to correct or refine parameters.

Visualization of verification results

Visualization is an essential tool for interpreting inspection results. In addition to the usual summary tables, it can include dashboards, diagrams, and automatically generated PDF documents that group project elements by inspection status. Color coding can play a supporting role here: green can indicate items that have been successfully validated, yellow can indicate items that require additional attention, and red can indicate items that have critical errors or are missing key data.

In our example (Fig. 4.4-1), we analyze data from each system step by step: from CAD (BIM) and real estate management to logistics and installation schedules (Fig. 4.4-16). Based on the audit results, individual alerts or report documents are automatically generated for each specialist, e.g. in PDF format (Fig. 4.4-17). If the data are correct, the specialist receives a brief message: "Thank you for working together". If discrepancies are detected, a detailed report is sent with the wording: "This document lists elements, their identifiers, attributes and values that have not been checked for compliance".

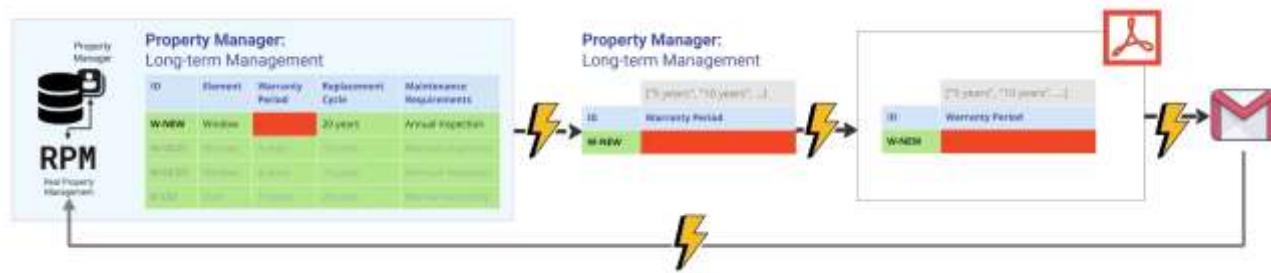


Fig. 4.4-17 Validation and automatic report generation speeds the process of finding and understanding data deficiencies for the professional who creates the data.

Thanks to the automated validation process - as soon as an error or data gap is detected, an instant notification is sent in the form of a chat message, e-mail or PDF -document to the person responsible for creating or processing the relevant entities and their attributes (Fig. 4.4-18), with a list of elements and attribute descriptions that failed the validation.



Fig. 4.4-18 Automated inspection reports make it easier to understand errors and speed up the work of completing project data.

For example, if a document arrives in the property management system (after structuring) with the Warranty Period attribute incorrectly filled in, the property manager receives an alert with a list of attributes that need to be checked and corrected.

Similarly, any deficiencies in the installation schedule or logistics data will result in an automatic report being generated and, for example, a chat notification or an e-mail with the results of the inspection being sent to the relevant specialist.

In addition to PDF -documents and graphs with results, it is possible to create dashboards and interactive 3D -models (Fig. 7.1-6, Fig. 7.2-12) highlighting elements with missing attributes, allowing users to visually utilize 3D element geometries to filter and evaluate the quality and completeness of these elements in the project.

Visualizing inspection results in the form of automatically generated documents, graphs or dashboards greatly simplifies data interpretation and facilitates effective communication between project participants.

The process of automatically checking data from various systems and information sources can be compared to informed decision-making in everyday life. Just as companies in the construction industry consider many variables - from the reliability of input data to their impact on the timing, cost, and quality of project implementation - so too, when making important decisions, for example, when choosing a place to live, a person weighs a number of factors: transportation accessibility, infrastructure, cost, safety, and quality of life. All of these considerations form a system of criteria, on the basis of

which the final decisions that make up our lives are made.

Comparison of data quality checks with human life needs

Despite the constant development of data quality control methods and tools, the fundamental principle of information compliance remains unchanged. This principle is built into the foundation of a mature management system, whether in business or in everyday life.

The process of iterative data validation is much like the decision-making process that everyone faces on a daily basis. In both cases, we rely on experience, data, and new information. In fact, more and more life and professional decisions - from strategic to mundane - are being made based on data.

For example, when choosing a place to live or a life partner, we intuitively form in our minds a table of criteria and characteristics by which we compare alternatives (Fig. 4.4-19). These characteristics - be it personal qualities of a person or parameters of a real estate object - represent attributes that influence the final decision.

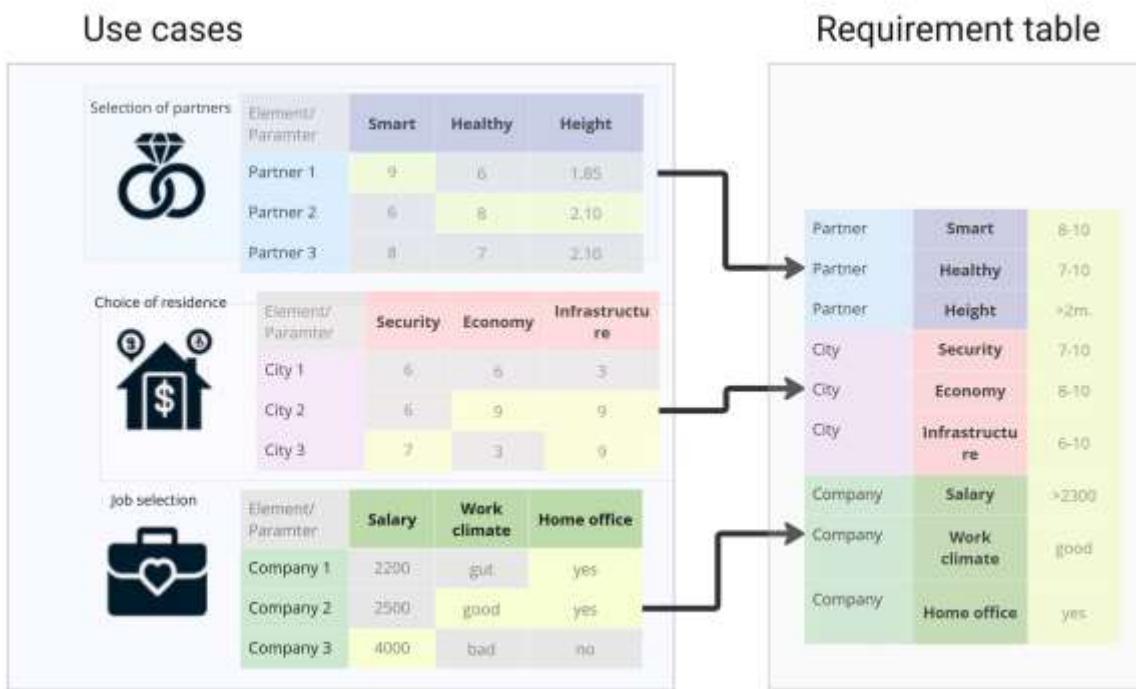


Fig. 4.4-19 The choice of residence, job, or partnership is based on individual attribute requirements.

The use of structured data and a formalized approach to describing requirements (Fig. 4.4-20) contributes to more informed and informed choices in both professional and personal life.

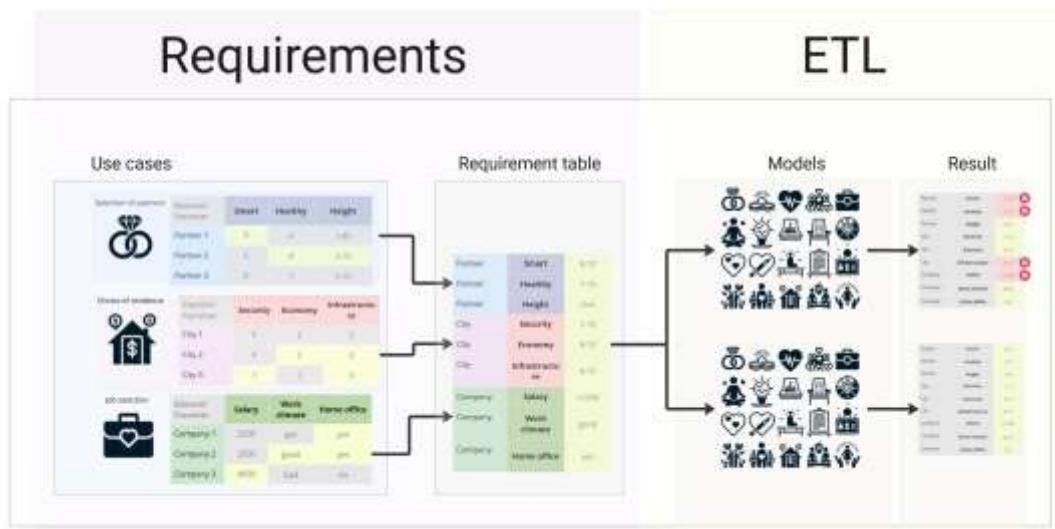


Fig. 4.4-20 Formalizing requirements helps to systematize the perception of life and business decisions.

The data-driven decision-making approach is not exclusively a business tool. It is seamlessly integrated into everyday life as well, following common data processing steps (Fig. 4.4-21) similar to the ETL process (Extract, Transform, Load) that we already covered at the beginning of this part when structuring data and that we will cover in detail in the Task Automation Contest in the seventh part of the book:

- **Data as the foundation (Extract):** In any field - whether it is work or personal life - we collect information. In business, it can be reports, figures, market data; in personal life, it can be personal experience, advice from loved ones, feedback, observations.
- **Evaluation criteria (Transform):** the information collected is interpreted based on predefined criteria. At work, these are performance indicators (KPIs), budget constraints and norms; in personal life, parameters such as price, convenience, reliability, charisma, etc.
- **Prediction and risk analysis (Load):** the final stage is decision making based on analyzing the transformed data and comparing possible consequences. This is similar to business processes, where data is passed through a business logic and risk filter.

The decisions we make - from trivial preferences like what to eat for breakfast to major life events like choosing a career or life partner - are inherently the result of processing and evaluating data.

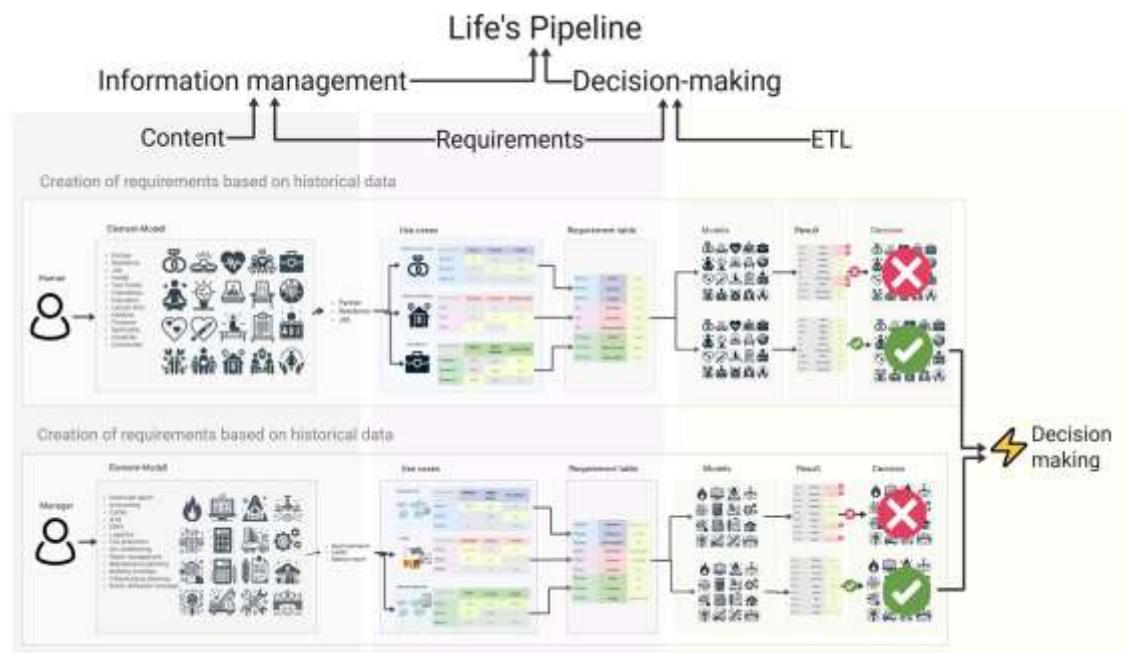


Fig. 4.4-21 Business and life in general is a series of data-driven decisions where the quality of the data used to make decisions is a key factor.

Everything in our lives is interconnected, and just as living organisms, including humans, follow the laws of nature, evolving and adapting to changing conditions, so human processes, including data collection and analysis methods, reflect these natural principles. The close relationship between nature and human activity confirms not only our dependence on nature, but also our desire to apply laws honed by millions of years of evolution to create data architectures, processes and systems for decision making.

New technologies, especially in construction, are a prime example of how humanity is inspired by nature time after time to create better, more sustainable and efficient solutions.

Next steps: turning data into accurate calculations and plans

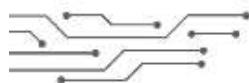
In this part we looked at how to convert unstructured data into a structured format, develop data models and organize processes for checking the quality of information in construction projects. Data management, standardization and classification is a fundamental process that requires a systematic approach and a clear understanding of business requirements. The techniques and tools discussed in this part allow for reliable integration between different systems throughout the entire lifecycle of an object.

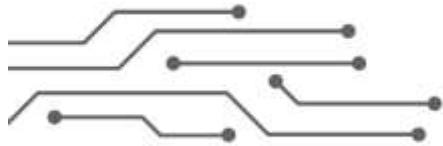
To summarize this part, let's highlight the main practical steps that will help you apply the discussed approaches in your daily tasks:

- Start by systematizing the requirements

- Create a registry of attributes and parameters for key elements of your projects and processes
 - Document the boundary values for each attribute
 - Visualize processes and relationships between classes, systems and attributes using flowcharts (e.g. in Miro, Canva, Visio)
- Automate data conversion
- Check which of your documents that are frequently used in processes can be digitized with OCR libraries and translated into tabular form
 - Check out reverse engineering tools to extract data from CAD (BIM)
 - Try setting up automatic data retrieval from documents or formats you frequently use in your work to a table form
 - Set up automatic conversions between different data formats
- Create a knowledge base for categorization
- Develop an internal or use an existing element classifier aligned with industry standards
 - Document the interrelationships between different classification systems
 - Discuss with your team the topic of using a unified system of identification and unambiguous classification of elements
 - Start building a process to automatically validate data - both that which you work with within the team and that which is passed to external systems

By using these approaches, you can significantly improve the quality of your data and simplify its subsequent processing and transformation. In the following parts of this book, we will look at how to apply already structured and prepared data for automated calculations, cost estimation, scheduling and construction project management.





V PART

COST AND TIME CALCULATIONS: INCORPORATING DATA INTO CONSTRUCTION PROCESSES

The fifth part is devoted to practical aspects of using data to optimize costing and planning of construction projects. The resource-based method of cost estimating and the automation of calculation processes are analyzed in detail. Methods of automated obtaining of quantities (Quantity Take-Off) from CAD (BIM)-models and their integration with calculation systems are considered. 4D and 5D modeling technologies for time scheduling and construction cost management are explored, with specific examples of their application. An analysis of extended information layers 6D -8D, providing an integrated approach to assessing the sustainability, operation and safety of real estate objects, is presented. The methods of calculating carbon footprint and ESG -indicators of construction projects in the context of modern environmental requirements and standards are discussed in detail. The possibilities and limitations of traditional ERP and PMIS systems in the management of construction processes are critically evaluated, analyzing their impact on pricing transparency. Prospects of transition from closed solutions to open standards and flexible data analysis tools capable of ensuring greater efficiency of construction processes are predicted.

CHAPTER 5.1.

COST CALCULATIONS AND ESTIMATES FOR CONSTRUCTION PROJECTS

Construction basics: estimating quantity, cost and time

Among the many business processes that determine the sustainability of a company in the construction industry, of particular importance - as they were thousands of years ago - are the processes for accurately estimating the number of elements, project cost and schedule (Fig. 5.1-1).

The development of writing was the result of a complex of factors, including the need to record economic transactions, the development of trade and resource management in early societies. The first legally significant documents, clay tablets with calculations of the cost of materials and labor, were used in the context of trade and construction. These tablets recorded the obligations of the parties in the construction of structures and were kept as evidence of bargaining and monetary relations.

For millennia, the approach to estimating has remained largely unchanged: calculations were performed manually, relying on the experience and intuition of the estimating engineer. However, with the advent of modular ERP systems and CAD tools, the traditional approach to estimating quantity, cost and time began to transform rapidly. Today's digital technologies allow for full automation of key time and cost calculations, enabling greater accuracy, speed and transparency in resource planning for construction projects.

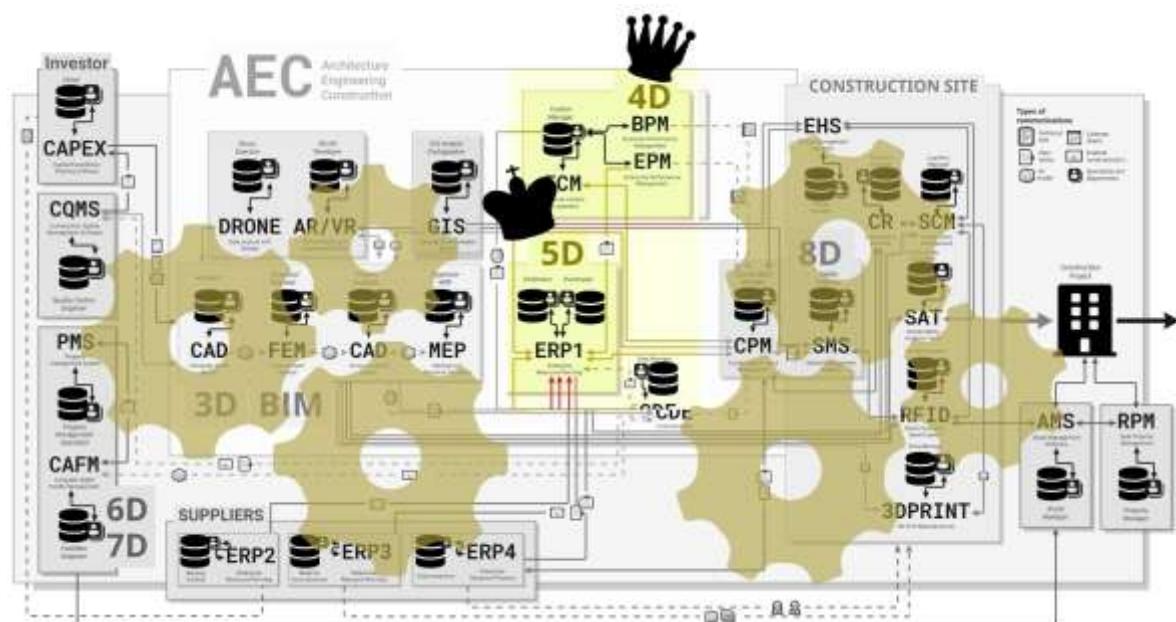


Fig. 5.1-1 Of the many different systems, the tools responsible for volume, cost, and time metrics are the most important in business.

The main focus of construction companies is on accurate time and cost data. These in turn depend on the amount of materials and labor used, and their transparency affects profitability. However, the complexity of calculation processes and their lack of transparency often lead to higher project costs, missed deadlines and even bankruptcy of companies.

According to the KPMG report "Familiar Problems - New Approaches" (2023), only 50% of construction projects are completed on time, and 87% of companies note increased control over the economics of capital projects. The main problems are related to the lack of qualified personnel and the difficulty of risk forecasting [2].

Historical costing and process time data is collected during the construction of past projects throughout the life of the construction company and entered into databases of various systems (ERP, PMIS BPM, EPM, etc.).

The availability of quality historical costing data is a major competitive advantage for a construction organization, directly affecting its survivability.

Estimating and costing departments in construction and engineering companies are created to collect, store and update historical data on project calculations. Their main function is to accumulate and systematize the company's experience, which allows the accuracy of estimating the scope, timing and cost of new projects to improve over time. This approach helps minimize errors in future calculations based on the practice and results of already implemented projects.

Methods of calculating the estimated cost of projects

A variety of estimating methods are used by costing professionals, each focusing on a specific type of data, availability of information, and level of project detail. The most common include:

- **Resource-based method:** estimating the estimated cost of a project based on a detailed analysis of all the resources required, such as materials, equipment and labor. This method requires a detailed list of all the tasks and resources required to accomplish each task, followed by a calculation of their cost. This method is highly accurate and is widely used in cost estimating.
- **Parametric method:** uses statistical models to estimate costs based on project parameters. This may involve analyzing the cost per unit of measure, such as building area or scope of work, and adapting these values to the specific project conditions. The method is particularly effective in the early stages when detailed information is not yet available.
- **Unit Method (Unit Cost Method):** calculates the estimated cost of a project based on a cost per unit of measurement (e.g. per square meter or cubic meter). This provides a quick and easy way to compare and analyze the cost of different projects or parts of projects.
- **Expert judgment (Delphi method):** based on the opinions of experts who use their experience and knowledge to estimate the value of the project. The approach is useful when accurate baseline data are not available or the project is unique.

It is worth noting separately that the parametric method and expert assessments can be adapted to

machine learning models. This allows you to automatically generate project cost and schedule forecasts based on training samples. Examples of using such models are discussed in more detail in the chapter "An example of using machine learning to find project cost and schedule" (Fig. 9.3-5).

Nevertheless, the resource-based method remains the most popular and widely used in the world practice. It provides not only an accurate assessment of the estimated cost, but also allows to calculate the duration of individual processes on the construction site and the entire project as a whole (more details in the chapter "Construction schedules and 4D -project data").

Resource-based method estimates and costing in construction

Resource-based costing is a method of management accounting, in which the cost of a project is formed on the basis of direct accounting of all resources involved. In construction, this approach involves a detailed analysis and evaluation of all material, labor and technical resources required to perform the work.

The resource-based method, provides a high degree of transparency and accuracy in budget planning, as it focuses on actual resource prices at the time of estimation. This is particularly important in an unstable economic environment where price fluctuations can significantly affect the overall cost of a project.

In the following chapters, we will look in detail at the resource-based costing process. In order to better understand its principles in construction, we will draw an analogy with the calculation of the cost of a dinner in a restaurant. The restaurant manager, planning the evening, makes a list of necessary products, takes into account the cooking time of each dish, and then multiplies the costs by the number of guests. In construction, the process is similar: for each category of project elements (objects), itemized estimates are generated recipes, and the total cost of the project is determined by summing up all the costs in the total bill - the final estimate by category.

The key and initial stage of the resource-based approach is the creation of the company's initial database. At the first stage of costing, a structured list of all items, materials, types of work and resources that the company has at its disposal within the framework of its construction projects - from a nail in the warehouse to the description of people through their qualifications and hourly rate - is compiled. This information is systematized into a single "Construction Resources and Materials Database" - a tabular register containing data on names, characteristics, units of measurement and current prices. It is this database that becomes the main and primary source of information for all subsequent resource calculations - both cost and timing of works.

Database of construction resources: catalog of construction materials and works

A database or table of construction resources and materials - includes detailed information about each

element that can be used in a construction project - a commodity, product, material or service, including its name, description, unit of measurement and unit cost, recorded in a structured form. In this table you can find everything from different types of fuels and materials used in projects to detailed lists of specialists in the form of different categories with descriptions of hourly rates (Fig. 5.1-2).

Database of resources			
	1st grade potatoes		Sand lime bricks
	1 kg \$2,99		1 pcs \$1
	Black Angus marble beef		JCB 3CX backhoe loader
	1 kg \$26,99		1 h \$150
	Broccoli		Laborer of the 1st category
	1 pcs \$1,99		1 h \$30

Fig. 5.1-2 A resource table is an ingredient list describing a material and service with a unit cost.

"Resource database" is akin to the product catalog of an online store, where each item has a detailed description of its attributes. This makes it easier for estimators to select the right resources (like selecting items when adding to a shopping cart) needed to calculate specific construction processes in the form of calculations (final order in the online store).

A resource database can also be thought of as a list of all the ingredients in a restaurant cookbook. Each building material, equipment and service is similar to the ingredients used in recipes. "Resource database" is a detailed list of all ingredients - building materials and services, including their cost per unit: piece, meter, hour, liter, etc.

New entity elements can be added to the "Construction Resource Databases" table in two ways - manually (Fig. 5.1-3) or automatically by integrating with the company's inventory management systems or supplier databases.

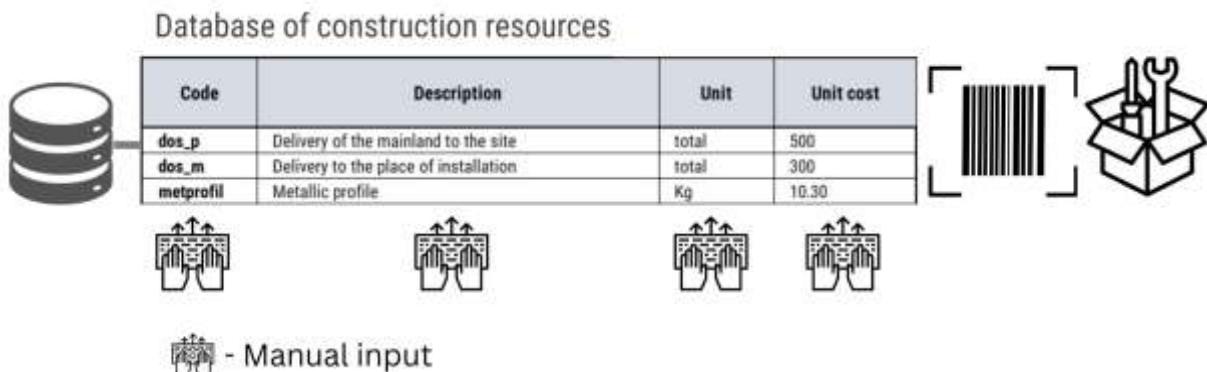


Fig. 5.1-3 The resource database is filled in manually or automatically adopts data from other databases.

A typical mid-sized construction company utilizes a database containing thousands, and sometimes tens of thousands, of items with detailed descriptions that can be used in construction projects. This data is then automatically used in contracts and project documents to accurately describe the work mix and processes

To keep up with changing market conditions such as inflation, the "unit cost" attribute for each product (good or service) in the resource database (Fig. 5.1-3) is regularly updated manually or by automatically downloading current prices from other systems or online platforms.

Updating the unit cost of a resource can be done monthly, quarterly or annually - depending on the nature of the resource, inflation and the external economic climate. Such updates are necessary to maintain the accuracy of calculations and estimates, as these basic elements are the starting point for the work of cost estimators. Up-to-date data is used to generate estimates, budgets and schedules that reflect real market conditions and reduce the risk of errors in subsequent project calculations.

Calculation of calculations and costing of works on the basis of resource base

Having filled the "Construction Resource Database" (Fig. 5.1-3) with entities-minimum units, you can start creating calculations, which are calculated for each process or work on the construction site for certain units of measurement: for example, for one cubic meter of concrete, one square meter of dry-wall, one meter of curb or one window installation.

For example, to build a 1 m² brick wall (Fig. 5.1-4), based on experience from previous projects, approximately 65 bricks (entity "Silicate Brick") are required at a cost of \$1 per piece (attribute "Cost per piece"), totaling \$65. Also, in my experience, it is required to use construction equipment (entity "JCB 3CX Loader") for 10 minutes that will place bricks near the work area. Since it costs \$150 per hour to rent the equipment, 6 minutes of use would cost approximately \$15. In addition, a brick-laying contractor will be needed for 2 hours, with an hourly rate of \$30 and a total of \$60.

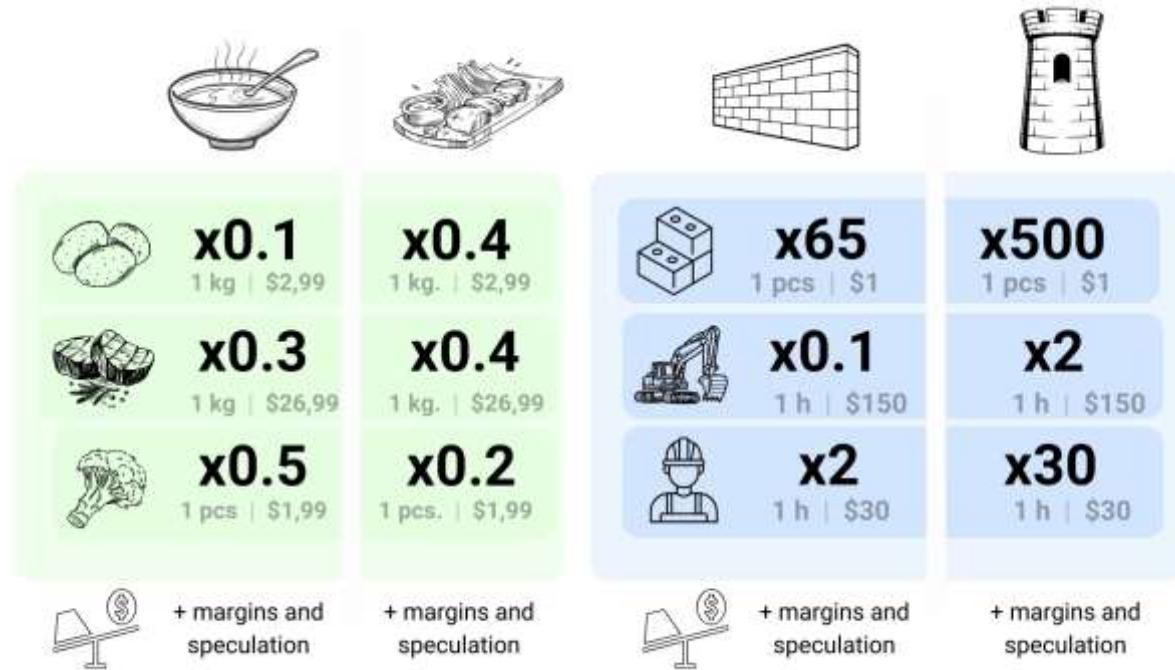


Fig. 5.1-4 Cost estimates provide a detailed list of construction materials and services required to perform the work and processes.

The composition of calculations (so-called "recipes") is formed on the basis of historical experience accumulated by the company in the process of performing a large volume of similar work. This practical experience is usually accumulated through feedback from the construction site. In particular, the foreman collects information directly at the construction site, recording actual labor costs, material consumption and nuances of technological operations. Then, in collaboration with the estimating department, this information is iteratively refined: process descriptions are refined, resource mix is adjusted, and cost estimates are updated to reflect actual data from recent projects.

Just as a recipe describes the necessary ingredients and quantities to prepare a dish, an estimate sheet provides a detailed list of all construction materials, resources, and services required to perform a particular job or process.

Regularly performed work allows workers, foremen and estimators to orient themselves in the necessary amount of resources: materials, fuel, working time and other parameters required to perform a unit of work (Fig. 5.1-5). These data are entered into the estimating systems in the form of tables, where each task and operation is described through the minimum elements of the resource base (with constantly updated prices), which ensures the accuracy of calculations.

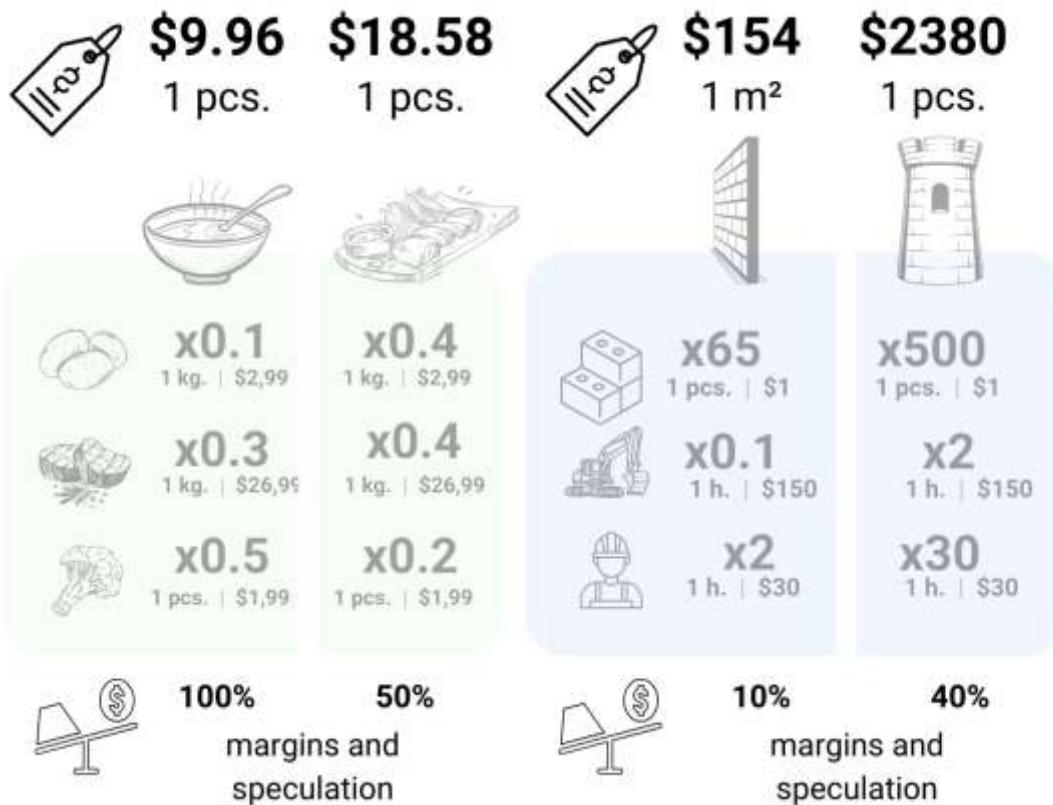


Fig. 5.1-5 Unit rates are collected for each job, where an entity's volume attribute is multiplied by its quantity with a profit percentage added.

To obtain the total cost of each process or work (costing object), the cost attribute is multiplied by its quantity and factors. The coefficients can take into account various factors such as complexity of the work, regional characteristics, inflation rate, potential risks (expected percentage of overhead) or speculation (additional profit factor).

The estimator, as an analyst, converts the experience and recommendations of the foreman into standardized estimates, describing construction processes through resource entities in tabular form. In fact, the estimator's task is to collect and structure through parameters and coefficients the information coming from the construction site.

Thus, the final cost per unit of work (e.g., square or cubic meter, or one installation of one unit) includes not only the direct costs of materials and labor, but also company markups, overhead, insurance, and other factors (Fig. 5.1-6)

At the same time, we no longer have to worry about the actual prices in (recipe) calculations, because the real prices are always reflected in the "resource base" (ingredient table). At the level of calculations in the table automatically (e.g. by item code or its unique identifier) are loaded from the resource base, which uploads the description, and the actual cost per unit, which in turn can be automatically loaded

from online platforms or online store of building materials. The estimator at the costing level of the work only has to describe the work or process through the attribute "quantity of resources" and additional factors.

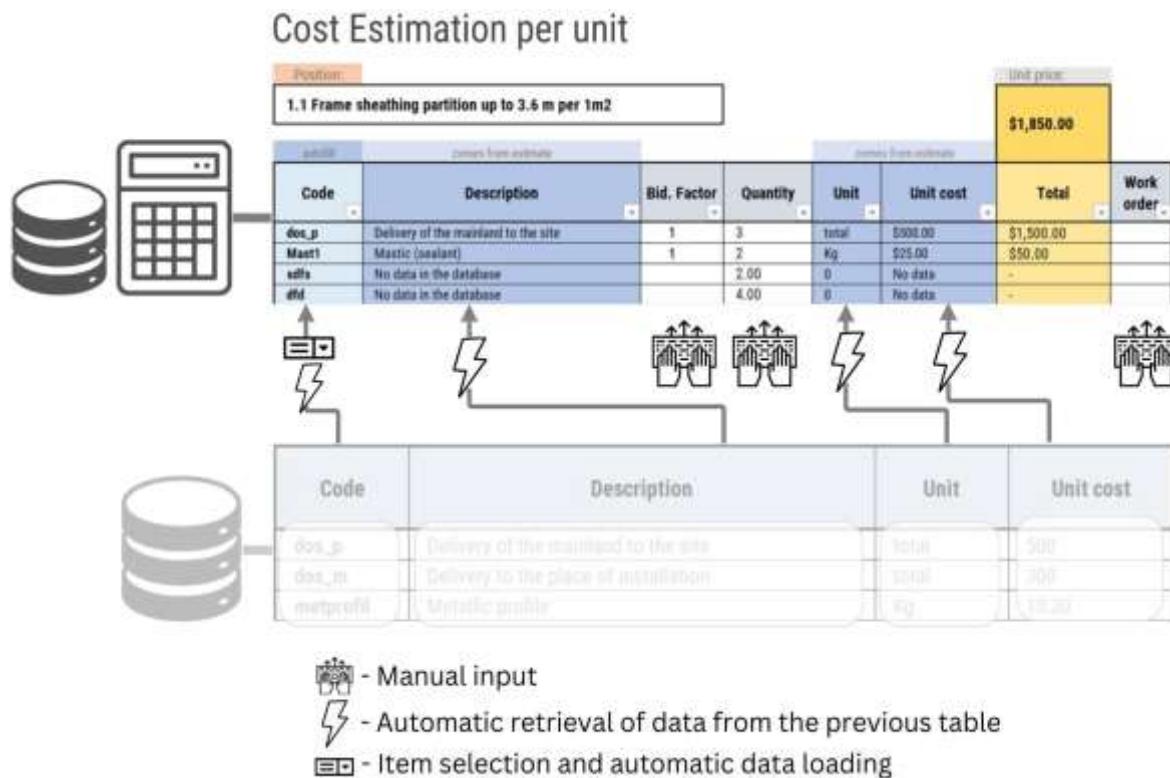


Fig. 5.1-6 At the stage of calculating the unit cost of work, only the attributes of the number of resources required are filled in, everything else is automatically loaded from the resource database.

The created cost estimates are stored in the form of template tables of standard projects, which are directly linked to the database of construction resources and materials. These templates represent standardized recipes for repetitive types of work for future projects, ensuring uniformity in calculations across the company.

When the cost of any resource changes in the database (Fig. 5.1-3) - whether manually or automatically via download of current market prices (e.g. in inflationary conditions) - the updates are immediately reflected in all linked costings (Fig. 5.1-6). This means that only the resource base needs to be changed, while the costing templates and estimates remain unchanged over time. This approach ensures the stability and reproducibility of the calculations for any price fluctuations, which are only accounted for in a relatively simple resource table (Fig. 5.1-3).

For each new project, a copy of the standard costing template is created, which makes it possible to make changes and adjust activities to meet specific requirements without changing the original template adopted by the company. This approach provides flexibility in adapting calculations: it is possible to take into account the specifics of the construction site, the customer's wishes, to introduce risk or profitability (speculation) coefficients - and all this without violating the company's standards. This helps the company find a balance between profit maximization, customer satisfaction and maintaining its competitiveness.

In some countries, such costing templates, accumulated over decades, are standardized at the national level and become part of the national standards of the construction costing system (Fig. 5.1-7).

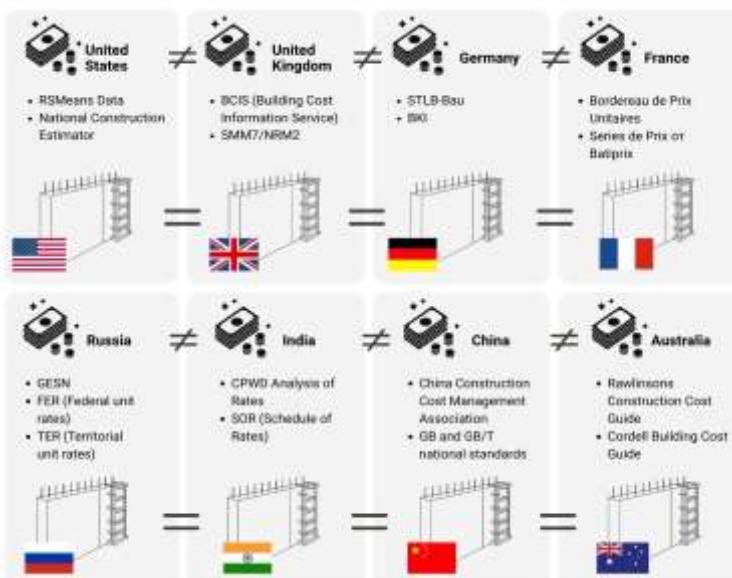


Fig. 5.1-7 Different countries around the world have their own costing rules with their own (prescription) compendia and standards for construction work for calculating the same element.

Such standardized resource bases of estimates (Fig. 5.1-7) are mandatory for use by all market participants, first of all, when implementing projects with public funding. Such standardization ensures transparency, comparability and fairness in the formation of prices and contractual obligations for the customer.

Final project costing: from estimates to budget

State and industry-specific estimating standards play different roles in construction practice in different countries. While some countries require strict adherence to a single standard, most developed economies adopt a more flexible approach. In market economies, government construction standards

usually serve only as a baseline. Construction companies adapt these standards to their operating models or completely revise them, supplementing them with their own customized factors. These adjustments reflect corporate experience, resource management efficiency and often factors in which, for example, a company's speculative profits may be factored in.

As a result, the level of competition, market demand, target margins and even relationships with specific customers can lead to significant deviations from standard norms. This practice provides market flexibility, but at the same time makes it difficult to transparently compare bids from different contractors, introducing an element of speculative pricing into the construction industry at this stage of the calculation process.

Once the calculation templates for individual types of work and processes have been prepared - or, more often, simply copied from standard government estimates (Fig. 5.1-7), with coefficients added to reflect the "peculiarities" of a particular company - the final step is to multiply the cost of each item by the corresponding attribute of the scope of work or processes in the new project.

When calculating the total cost of a new construction project, the key step is to summarize the costs of all costing items, multiplied by the volume of these items-work in the project.

To create the total cost of the project, in our simplified example, we start by calculating the cost of building one square meter of wall and multiply the cost of its calculation (e.g. the work "1m² standard installation of wall elements") by the total number of square meters of walls in the project (e.g. the "Area" or "Quantity" attribute (Fig. 5.1-8) of an entity of type "Wall elements" from CAD of the project or the foreman's calculations).

Similarly, we calculate the cost for all elements of the project (Fig. 5.1-8): we take the cost of a unit of work and multiply it by the volume of a particular element or its group in a given project. The estimator only has to enter the number of these elements, activities or processes in the project in the form of volume or quantity. This allows to automatically generate a complete construction estimate.

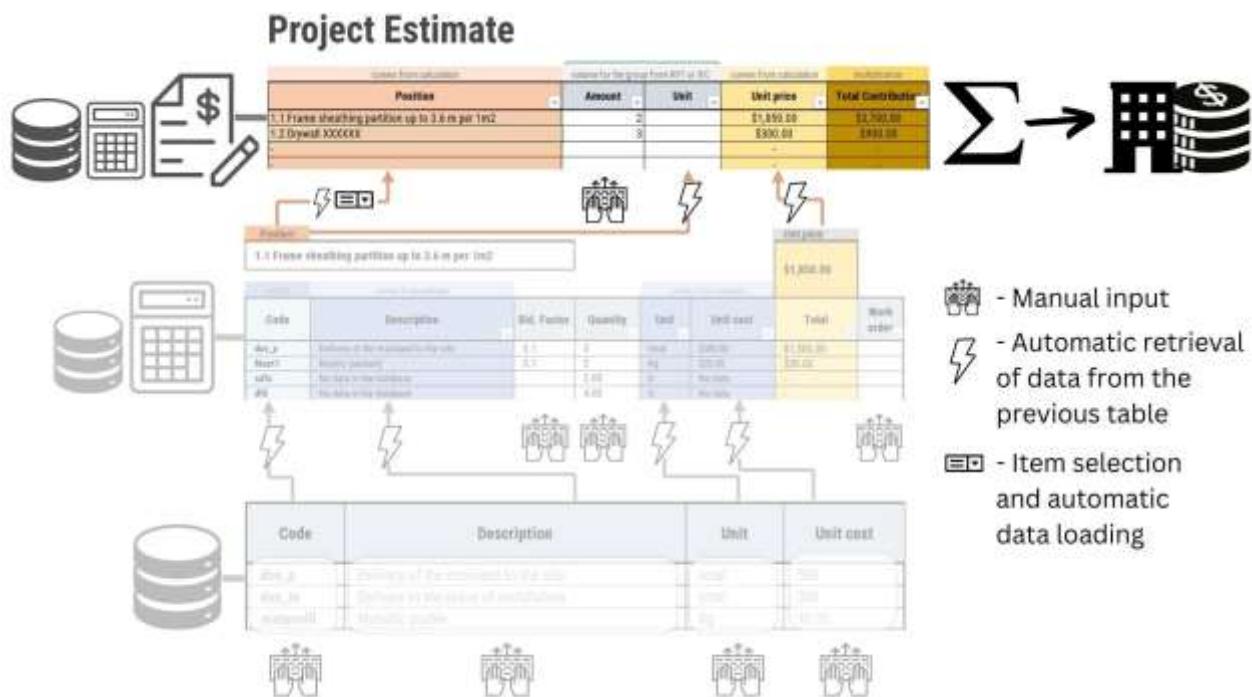


Fig. 5.1-8 At the estimate creation stage, we only enter the scope of work.

As in the case of calculations, at this level we load automatically ready calculated items (from the template of calculations or new ones copied from the template and edited), which automatically bring with them the current cost per unit of work (which is updated automatically from the resource database (Fig. 5.1-8 bottom table)). Accordingly, in case of any change of data in the resource database or costing tables - the data in the estimate will be automatically updated for the current day, without the need to change the costing or the estimate itself.

In the restaurant context, the final cost of an event is calculated in a similar manner and equals the final cost of the entire dinner, where the cost of each dish multiplied by the number of guests adds up to the total cost of the check (Fig. 5.1-9). And just as in construction, recipes for cooking in a restaurant may not change for decades. Unlike prices, where the cost of ingredients can change every hour.

Just as a restaurant owner multiplies the cost of each meal by the number of servings and people to determine the total cost of the event, the cost estimating manager adds up the cost of all project components to arrive at a complete construction estimate.

Thus, for each work in the project, its final cost is determined (Fig. 5.1-9), which, multiplied by the attribute volume of the entity corresponding to this work - gives the cost of groups of works, from which the final cost of the entire project is obtained.

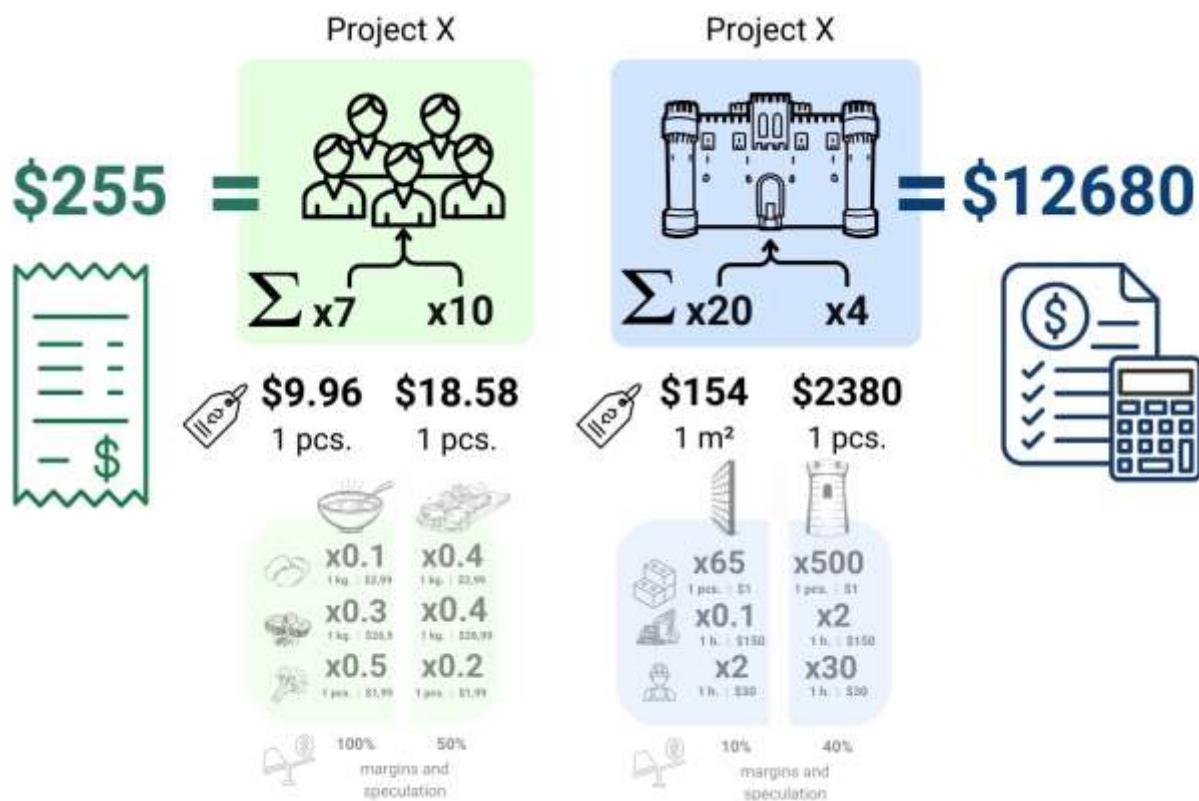


Fig. 5.1-9 The final estimate is calculated by summing each element's work cost attribute by its scope attribute.

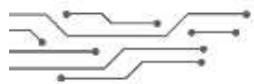
The Total Project Cost (Fig. 5.1-8) presents a financial picture of the project, allowing customers, investors, or financial institutions to understand the total budget and financial resources required for the project on any given day, taking into account current prices.

And if the processes of compiling resource bases, calculations and estimates (process recipes) have already been worked out, semi-automated and honed by tens of thousands of years and recorded at the state level, then the automatic obtaining of quality information about the volume and quantity of elements for the last stage of the final estimate - today remains a bottleneck in the processes of all calculations of cost and time attributes of the project, and in general the overall budget of the project.

For millennia, the traditional method of calculating volumes has been manual methods of measuring volume and quantity using flat drawings. With the advent of the digital age, companies have discovered that volume and quantity information can now be automatically extracted from the geometric data contained in CAD models, revolutionizing the millennia-old ways of obtaining quantitative data.

Modern approaches to process estimating and estimating involve automatically extracting volumetric and quantitative attributes from CAD databases, which can be uploaded and connected to the

costing process to get up-to-date project group volumes at any stage of design through to operation.



CHAPTER 5.2.

QUANTITY TAKE-OFF AND AUTOMATIC CREATION OF ESTIMATES AND SCHEDULES

Moving from 3D to 4D and 5D: using volumetric and quantitative parameters

With the costing tables with the described processes through the resources (Fig. 5.1-8) in hand, the next step is to automatically obtain the volume or quantity parameters for a group of elements that are needed for the calculations and for the final estimate.

Volumetric characteristics of project elements - e.g. walls or slabs - can be automatically extracted from CAD databases. Parametric objects created in CAD programs are converted by means of geometry kernel into numerical values of length, width, area, volume and other parameters. The process of obtaining volumes based on 3D geometry will be discussed in more detail in the next, sixth part (Fig. 6.3-3), dedicated to working with CAD (BIM). In addition to volumes, the number of similar elements can also be obtained from the CAD-model database by filtering and grouping objects by categories and properties. These grouping parameters become the basis for linking the project elements through resource calculations to the calculations, final estimate and budget of the entire project.

Thus, the data model extracted from the 3D (CAD) model is augmented with new parameter layers, denoted as 4D and 5D. In the new entity attribute layers, 4D (time) and 5D (cost), 3D geometric data are used as a source of entity volume attribute values.

- **4D** is a parameter information layer that adds information about the duration of construction operations to the 3D parameters of elements. This data is required for planning work schedules and managing project timelines
- **5D** is the next level of extension of the data model, in which the elements are supplemented with cost characteristics. Thus, the financial aspect is added to the geometric information: the cost of materials, works and equipment, which allows to perform budget calculations, analyze profitability and manage costs during the construction process.

Cost and 3D, 4D and 5D attribute data of groups of project entities are described in a similar way to costing in modular ERP, PIMS -systems (or Excel -like tools) and are used for automatic costing and budget planning of both individual groups and the full project budget.

5D attributes and retrieving attribute volumes from CAD

When preparing the final estimate of a construction project, the compilation of which we discussed in previous chapters (Fig. 5.1-8), the volume attributes for each category of project elements are either collected manually or extracted from the volume attribute specifications provided by CAD programs.

The traditional manual method of calculating quantities involves the foreman and estimator analyzing drawings that have been presented for thousands of years as lines on paper, and for the last 30 years in digital formats such as PDF (PLT) or DWG. Drawing on professional experience, they measure the quantities of work and materials required, often with a ruler and protractor. This method requires considerable effort and time, as well as special attention to detail.

Determining scope attributes in this manner can take anywhere from a few days to several months, depending on the scope of the project. In addition, since all measurements and calculations are done manually, there is a risk of human error that can lead to inaccurate data, which subsequently affects errors in estimating project time and cost, for which the entire company will be held responsible.

Modern methods based on the use of CAD databases greatly simplify the calculation of volumes. In CAD models, the geometry of elements already includes volume attributes that can be automatically calculated (via the geometry kernel (Fig. 6.3-3)) and presented or exported in tabular form.

In such a scenario, the estimating department asks the CAD designer for data on the quantity and volume characteristics of the project elements. This data is exported as spreadsheets or directly integrated into costing databases - whether Excel, ERP or PMIS - systems. This process often begins not with a formal request, but with a brief dialog between the client (initiator) and the architect estimator from the construction or design company. Below is a simplified example showing how a structured table for automatic calculations (QTO) is formed from everyday communication:

- ⌚ Customer - "*I want to add another floor to the building, in the same configuration as the second floor*"
- ⌚ Architect (CAD) - "*Adding a third floor, the configuration is the same as on the second floor*". And after this message sends a new CAD version of the project to the estimator.
- ⌚ Estimator automatically performs grouping and calculation (ERP, PMIS, Excel) - "*I will run the project through an Excel spreadsheet with QTO rules (ERP, PMIS), get volumes by category for the new floor and generate an estimate*"

As a result, the text dialog is transformed into a table structure with grouping rules:

Element	Category	Floor
Ceiling	OST_Floors	3
Column	OST_StructuralColumns	3
Staircase	OST_Stairs	3

After the process of automatic grouping of CAD model from the designer according to the rules of QTO estimator and automatic multiplication of volumes by resource calculations (Fig. 5.1-8) we get the following results, which are sent to the customer:

Element	Volume	Floor	Price per unit.	Total cost
Ceiling	420 m ²	3	150 €/m ²	63 000 €
Column	4 pcs.	3	2450 €/piece.	9 800 €
Staircase	2 pcs.	3	4,300 €/piece.	8 600 €
TOTAL:	-	-	-	81 400 €

- ⌚ The customer - "Thank you, it's quite a lot, we need to cut some rooms". And the cycle repeats itself many times.

This scenario can be repeated many times, especially in the approval phase, where the customer expects instant feedback. In practice, however, such processes can drag on for days or even weeks. Today, thanks to the introduction of automatic grouping and calculation rules, activities that used to take considerable time should be completed in minutes. Automated acquisition of volumes through grouping rules not only speeds up calculations and generation of estimates, but by minimizing the human factor reduces the probability of errors, providing a transparent and accurate assessment of the project cost.

If the requirements of the estimation department were initially taken into account when creating the 3D model in the CAD system (which is rare in practice), and the names, identifiers of element groups and their classification attributes are set in the form of parameters that coincide with the structures of estimation groups and classes, then volumetric attributes can be automatically transferred to the estimation systems without additional transformations.

Automatic extraction of volumetric attributes from CAD in the form of specification tables allows to quickly obtain up-to-date data on the cost of individual works and the project as a whole (Fig. 5.2-1). By updating only the CAD file with project volumes in the calculation process or the calculation system, the company can quickly recalculate the estimate taking into account the latest changes, ensuring high accuracy and consistency of all subsequent calculations.

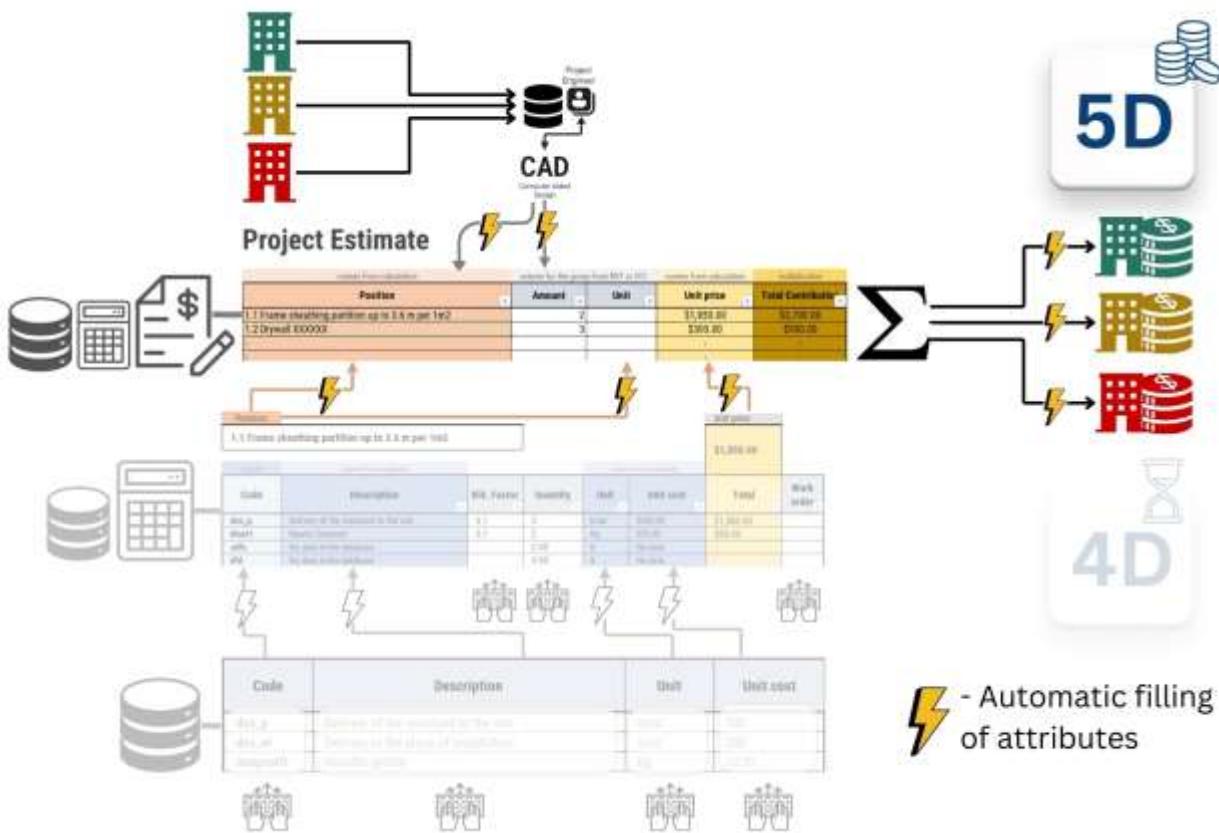


Fig. 5.2-1 Volume attributes from CAD tables or databases are automatically entered into the estimate, allowing you to instantly calculate the total project cost.

With the increasing complexity of capital projects, calculating the full budget and analyzing the total cost of projects under this type of scenario (Fig. 5.2-1) - becomes a key tool for informed decision making.

According to Accenture's Creating More Value with Capital Projects (2024) study [20], leading companies are actively integrating data analytics into digital initiatives, using historical information to predict and optimize outcomes. Research shows that more owner-operators are applying big data analytics to predict market trends and assess commercial viability before design begins. This is accomplished by analyzing data warehouses from an existing portfolio of projects. In addition, 79% of owner-operators are implementing 'robust' predictive analytics to assess project performance and support real-time operational decision-making.

Modern effective management of construction projects is inextricably linked to the processing and analysis of large amounts of information at all stages of design and those processes that precede the design. The use of data warehouses, resource calculations, predictive models and machine learning allows not only to minimize risks in calculations, but also to make strategic decisions on project financing at the early stages of design. We will talk more about data warehouses and predictive models that will complement the calculations in the ninth part of the book.

Automatic obtaining of volumetric parameters of elements from CAD projects, which are necessary for making estimates, is realized with the help of grouping tools QTO (Quantity Take-Off). QTO tools work by grouping all project objects by special element identifiers or element attribute parameters, using specifications and tables created in CAD database.

QTO Quantity Take-Off: grouping project data by attributes

QTO (Quantity Take-Off) in construction is the process of extracting the quantitative characteristics of the elements required to realize a project. In practice, QTO is often a semi-manual process involving data collection from various sources: PDF documents, DWG drawings and CAD models.

When working with data extracted from CAD databases, the QTO process is realized as a sequence of filtering, sorting, grouping and aggregation operations. Model elements are selected by class, category and type parameters, and then their quantitative attributes - such as volume, area, length or quantity - are summarized according to the calculation logic (Fig. 5.2-2).

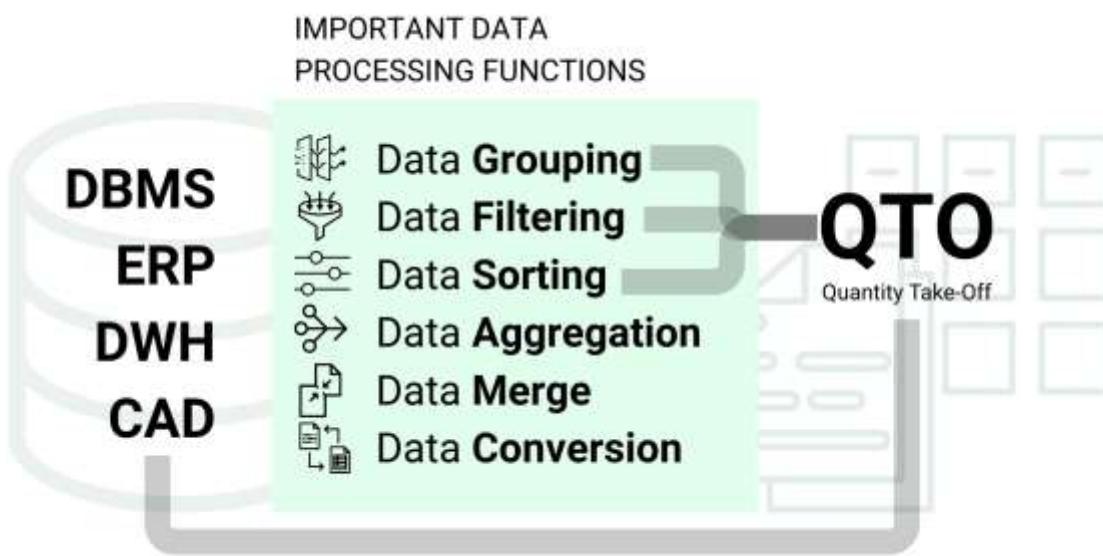


Fig. 5.2-2 Data grouping and filtering are the most popular functions applied to databases and data warehouses.

The QTO (filtering and grouping) process allows to systematize data, form specifications and prepare source information for calculating estimates, purchases and work schedules. The basis of QTO is the classification of elements according to the type of measured attributes. For each element or group of elements the corresponding quantitative measurement parameter is selected. For example:

- **Length attribute** (curbstone - in meters)
- **Area attribute** (drywall work - in square meters)
- **Volume attribute** (concrete works - in cubic meters)

■ **Quantity attribute** (windows - per piece)

In addition to the volumetric characteristics generated mathematically from the geometry, after grouping QTOs, overrun factors are often applied in calculations (Fig. 5.2-12 e.g. 1.1 to account for 10% in logistics and installation) - correction values that take into account losses, installation, storage or transportation features. This makes it possible to predict the actual consumption of materials more accurately and to avoid both shortages and overstocking on the construction site.

An automated quantity take-off process (QTO) is essential for producing accurate calculations and estimates, reducing human error in the processes of finding volume specifications and preventing over- or under-ordering of materials.

As an example of QTO process, let's consider a common case when it is necessary to show from CAD database a table-specification of volumes by element types for a certain category, classes of elements. Let's group all project elements by type from the CAD project wall category and summarize the volume attributes for each type to present the result as a QTO volume table (Fig. 5.2-3).

In the example of a typical CAD project (Fig. 5.2-3), all wall category elements within the CAD database are grouped by wall type, e.g. "Lamelle 11.5", "MW 11.5" and "STB 20.0", and have well-defined volume attributes represented in metric cubes.

The goal of the manager, who is at the interface between designers and calculation specialists, is to obtain an automated table of volumes by element type in the selected category. Not only for a specific project, but also in a universal form applicable to other projects with a similar model structure. This allows the approach to be scalable and allows data to be reused without duplication of effort.

Gone are the days when experienced designers and estimators used to arm themselves with a ruler, carefully measuring every line on paper or PDF -plans - a tradition that has not changed over the past millennia. With the development of 3D -modeling, where the geometry of each element is now directly linked to automatically calculated volumetric attributes, the process of determining volumes and QTO quantities has become automated.

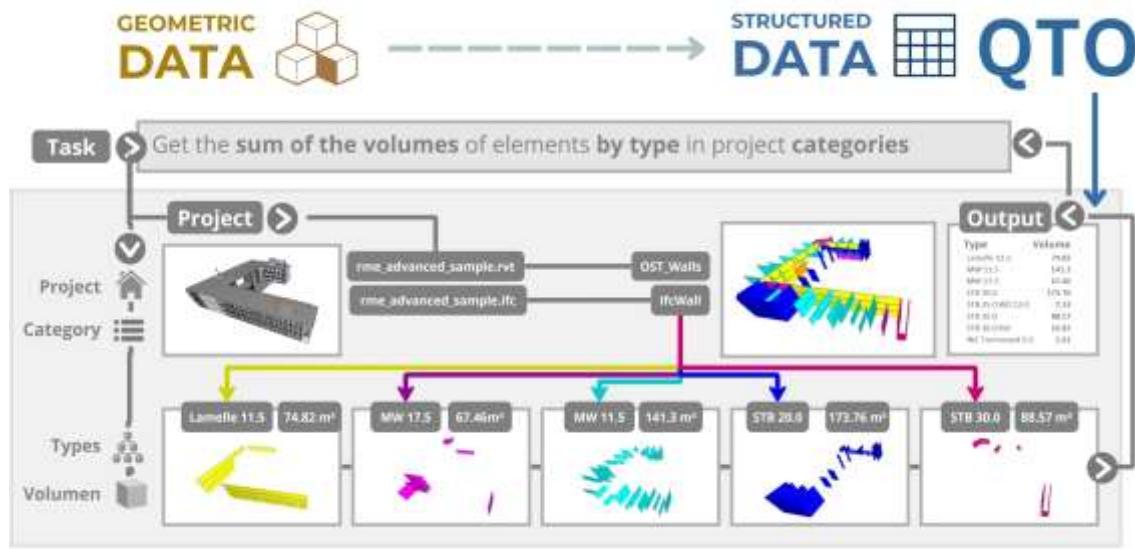


Fig. 5.2-3 Obtaining QTO scope and quantity attributes from a project involves grouping and filtering project elements.

In our example, the task is to "select a category of walls in a project, group all elements by type, and present the scope attribute information in a structured, tabular format" so that this table can be used by dozens of other professionals for costing calculations, logistics, schedules, and other business cases (Fig. 6.1-3).

Due to the closed nature of CAD data not every specialist today can use direct access to the CAD database (the reasons and solutions to the access problem are detailed in the sixth part of the book). Therefore, many are forced to turn to specialized BIM tools based on the concepts of open BIM and closed BIM [63]. When working with specialized BIM-tools or directly in the CAD program environment, the table with QTO (Quantity Take-Off) results can be generated in different ways - depending on whether manual interface or software automation is used.

For example, using the user interface of CAD (BIM) software, it is enough to perform about 17 actions (button clicks) to get a ready table of volumes (Fig. 5.2-4). However, the user must have a good understanding of the model structure and functions of the CAD (BIM) software.

If automation is applied through program code or through plug-ins and API tools within CAD programs, the number of manual steps to obtain the volume tables is reduced, but 40 to 150 lines of code will need to be written, depending on the library or tool used:

- **IfcOpSh (open BIM)** or **Dynamo IronPython (closed BIM)** - allow you to get a QTO table from a CAD format or CAD program in just ~40 lines of code.
- **IFC_js (open BIM)** - requires approximately 150 lines of code to extract voluminous attributes from the IFC model.
- **Interface CAD tools (BIM)** - allows you to get the same result manually, in 17 mouse clicks.

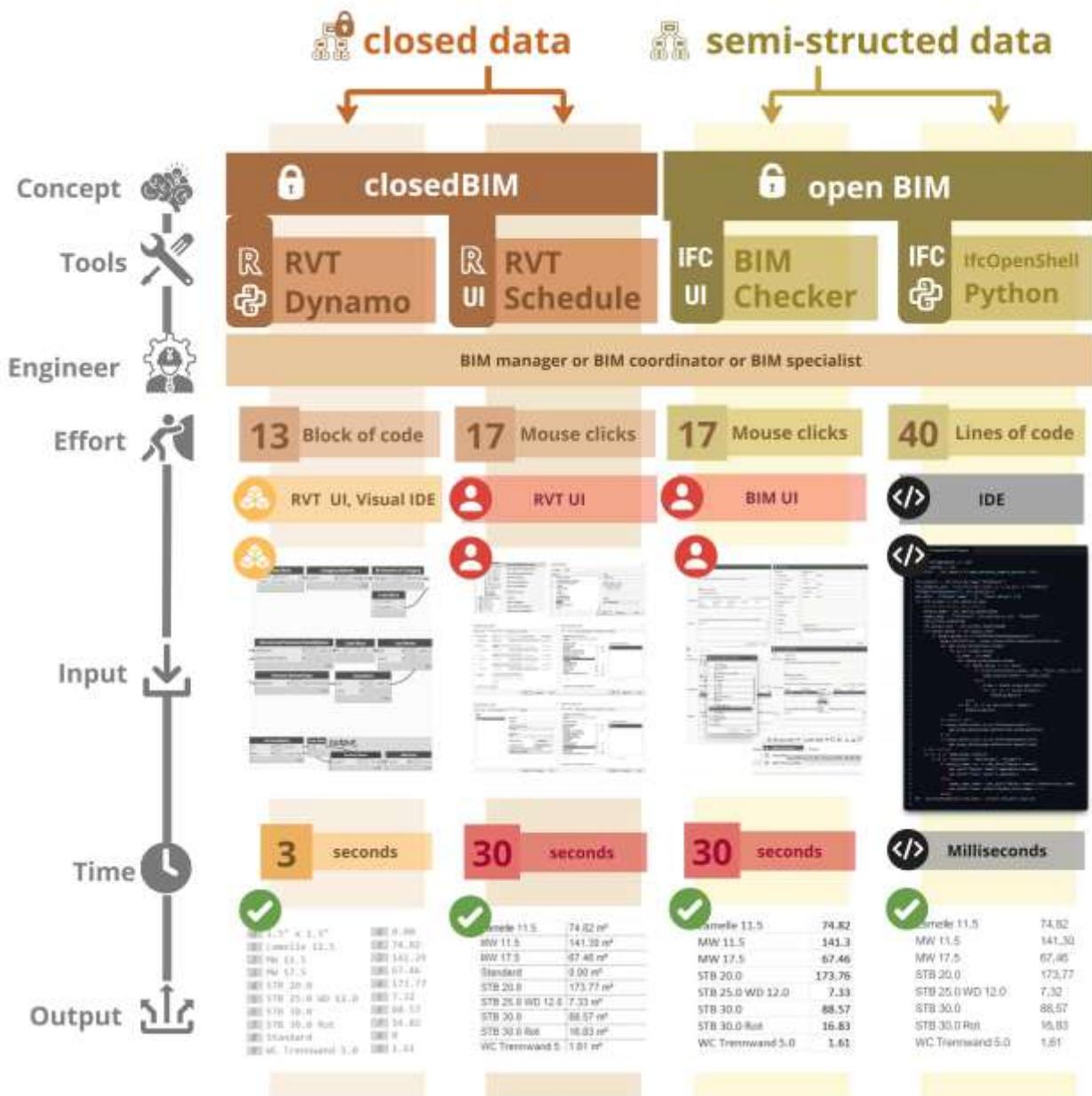


Fig. 5.2-4 CAD (BIM) designers and managers, use 40 to 150 lines of code or a dozen keystrokes to create QTO tables

The result is the same - a structured table with volume attributes for a group of elements. The only difference is the labor costs and the necessary level of technical training of the user (Fig. 5.2-4). Modern tools, in relation to manual collection of volumes, significantly speed up the QTO process and reduce the probability of errors. They allow data to be extracted directly from the project model, eliminating the need to manually recalculate volumes from drawings, as was done in the past.

Regardless of the method used - whether open BIM or closed BIM - it is possible to obtain an identical QTO - table with project element volumes (Fig. 5.2-4). However, when working with project data in CAD

- (BIM-) concepts, users depend on specialized tools and APIs, provided by vendors (Fig. 3.2-13). This creates additional layers of dependency and requires learning unique data schemas while limiting direct access to the data.

Due to the closed nature of CAD-data, obtaining QTO-tables and other parameters complicates the automation of calculations and integration with external systems. By using tools for direct access to databases and transferring CAD project data using reverse engineering tools into an open structured dataframe format (Fig. 4.1-13), an identical QTO table can be obtained with just one line of code (Fig. 5.2-5 - variant with granular data).

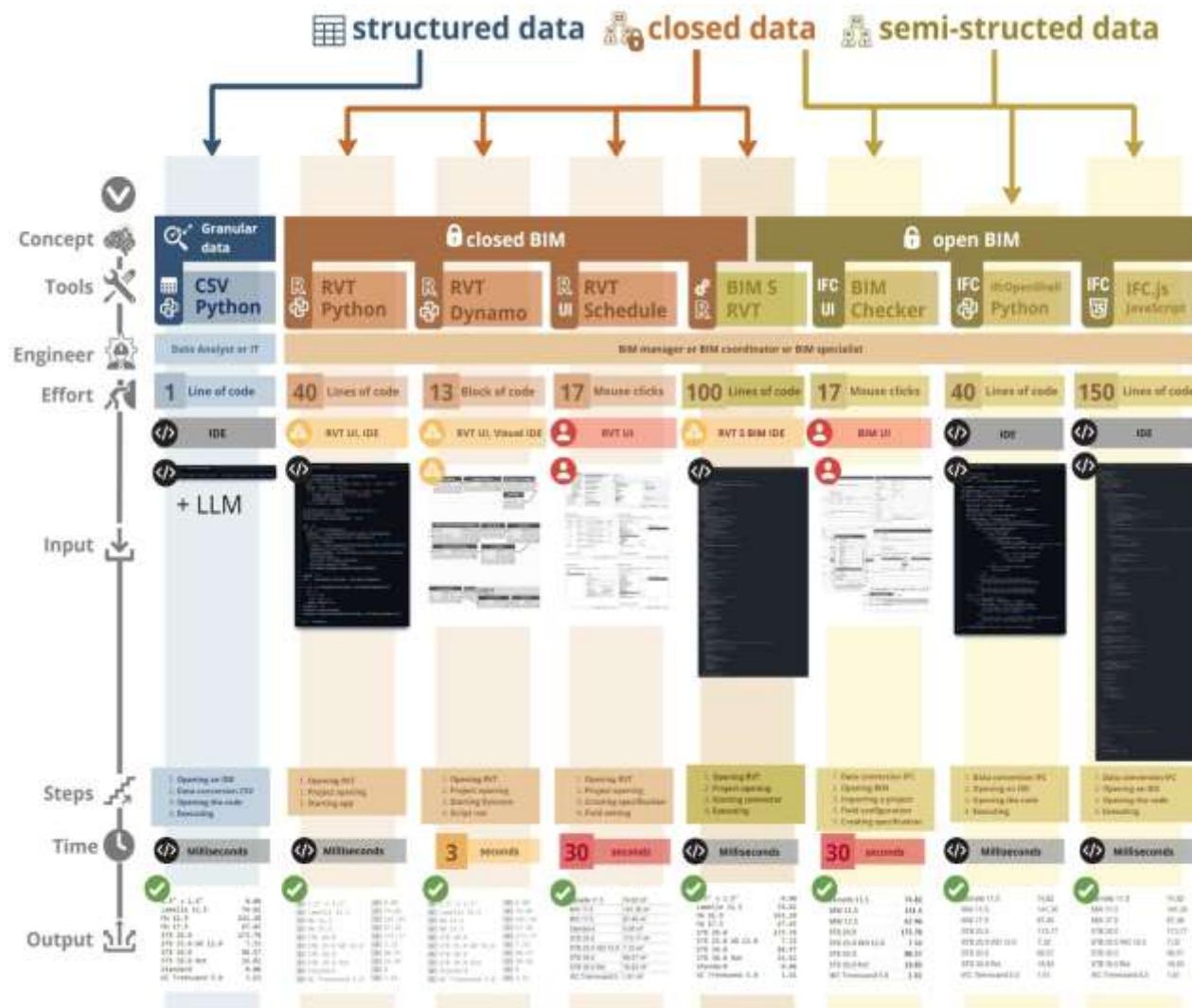


Fig. 5.2-5 Different tools produce the same results in the form of attribute tables of project entities, but with different labor costs.

When using open structured data from CAD projects, as mentioned in the chapter "Converting CAD (BIM) data into structured form", the grouping process, QTO, is greatly simplified.

Approaches based on the use of open structured data or direct access to CAD model databases are free from the marketing constraints associated with the acronym BIM. They rely on proven tools long

used in other industries (Fig. 7.3-10 ETL process).

According to the McKinsey study "Open Data: Unleash Innovation and Productivity with Streaming Information" [102] conducted in 2013, the use of open data can create opportunities for savings of \$30 to \$50 billion per year in the design, engineering, procurement, and construction of electric power facilities. This translates into a 15 percent savings in construction capital costs.

Working with open structured (granular) data simplifies information retrieval and processing, reduces dependence on specialized BIM platforms, and opens the door to automation without the need to use proprietary systems or parametric and complex data models from CAD formats.

QTO automation using LLM and structured data

Translating unstructured data into a structured form significantly improves the efficiency of various processes: it simplifies data processing (Fig. 4.1-1, Fig. 4.1-2) and speeds up the validation process by making the requirements clear and transparent, as we have already discussed in the previous chapters. Similarly, translating CAD data (BIM) into a structured open form (Fig. 4.1-12, Fig. 4.1-13) facilitates the attribute grouping process and the QTO process.

The QTO attribute table is structured, so when using structured CAD data, we work with a single data model (Fig. 5.2-5), which eliminates the need to convert and translate project data models and grouping rules to a common denominator. This allows us to group data by one or more attributes with just one line of code. In contrast, in open BIM and closed BIM, where data is stored in semi-structured, parametric or closed formats, processing requires dozens or even hundreds of lines of code, and the use of the API to interact with geometry and attribute information.

- ⌚ Example of grouping a QTO structured project by one attribute. Text query in any LLM chat room (ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

I have CAD -project as DataFrame - please filter the project data to get items with "Type" parameter containing only "Type 1 value. ↵

- LLM's answer is very likely to be in the form of Python code using Pandas:



Fig. 5.2-6 One line of code written with LLM, allows you to group an entire CAD project by the "Type" attribute and get the desired group of elements.

Thanks to the simple structure of the two-dimensional DataFrame we do not need to explain the LLM schema and data model, which shortens the interpretation steps and speeds up the creation of final solutions. Previously, writing even simple code required learning programming languages, but now modern language models (LLMs) allow us to automatically translate process logic into code when working with structured data using text queries.

LLM automation and language models can completely eliminate the need for professionals working with CAD (BIM) data grouping and processing from having to learn programming languages or BIM tools by providing the ability to solve problems with text-based queries.

The same query - grouping all project elements from the "walls" category and calculating volumes for each type (Fig. 5.2-5) - which in a CAD environment (BIM) requires 17 clicks in the interface or writing 40 lines of code, in open data processing tools (e.g. SQL or Pandas) looks like a simple and intuitive query:

- With a single line in Pandas:

```
df[df['Category'].isin(['OST_Walls'])].groupby('Type')['Volume'].sum()
```

Code decoding: take from df (DataFrame) the elements, which attribute-column "Category", has values "OST_Walls", group all obtained elements by attribute-column "Type" and summarize for the obtained group of elements attribute "Volume".

- Grouping a structured design retrieved from CAD using SQL:

```
SELECT Type, SUM(Volume) AS TotalVolume
FROM elements
WHERE Category = 'OST_Walls'
GROUP BY Type;
```

- With the help of LLM we can write a grouping request to the project database as a simple text reference - a prompt (Fig. 5.2-7):

For the project dataframe, group the items by the 'Type' parameter, but only for items with the 'Category' parameter equal to 'OST_Walls' or 'OST_Columns' and please summarize the column parameter 'Volume' for the resulting ↴

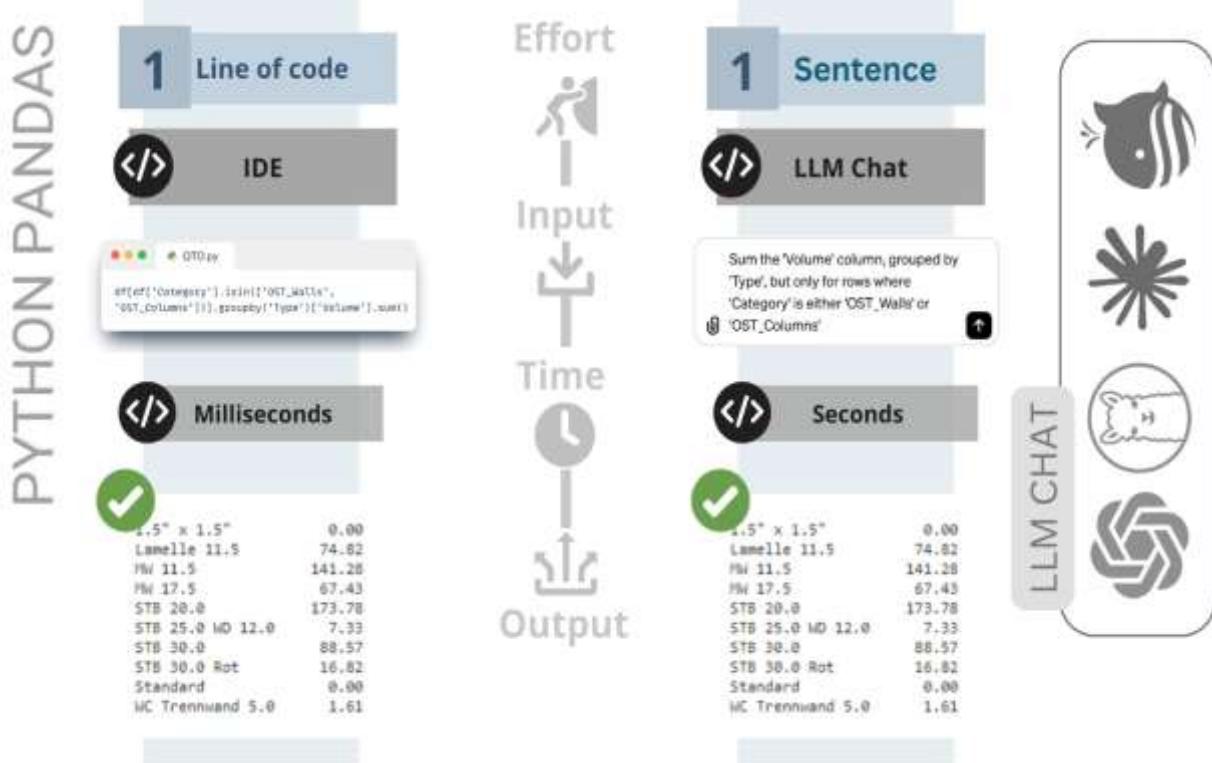


Fig. 5.2-7 Using SQL, Pandas and LLM, automating data processing is now possible with a few lines of code and text queries.

Obtaining QTO from CAD data using LLM tools (ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok), dramatically changes traditional methods of extracting attribute information, quantitative and volumetric data for individual objects and groups of objects.

Now even project managers, costing or logistics specialists who do not have a deep knowledge of design and do not have specialized CAD software - (BIM-) vendors, having access to the CAD database can get the total volume of elements of the category of walls or other objects in seconds, simply by writing or dictating a query.

In text queries (Fig. 5.2-8) the LLM agent of the model processes the user's request to apply a certain function to one or more parameters - columns of the table. As a result, the user in communication with the LLM receives either a new column-parameter with new values, or one specific value after grouping

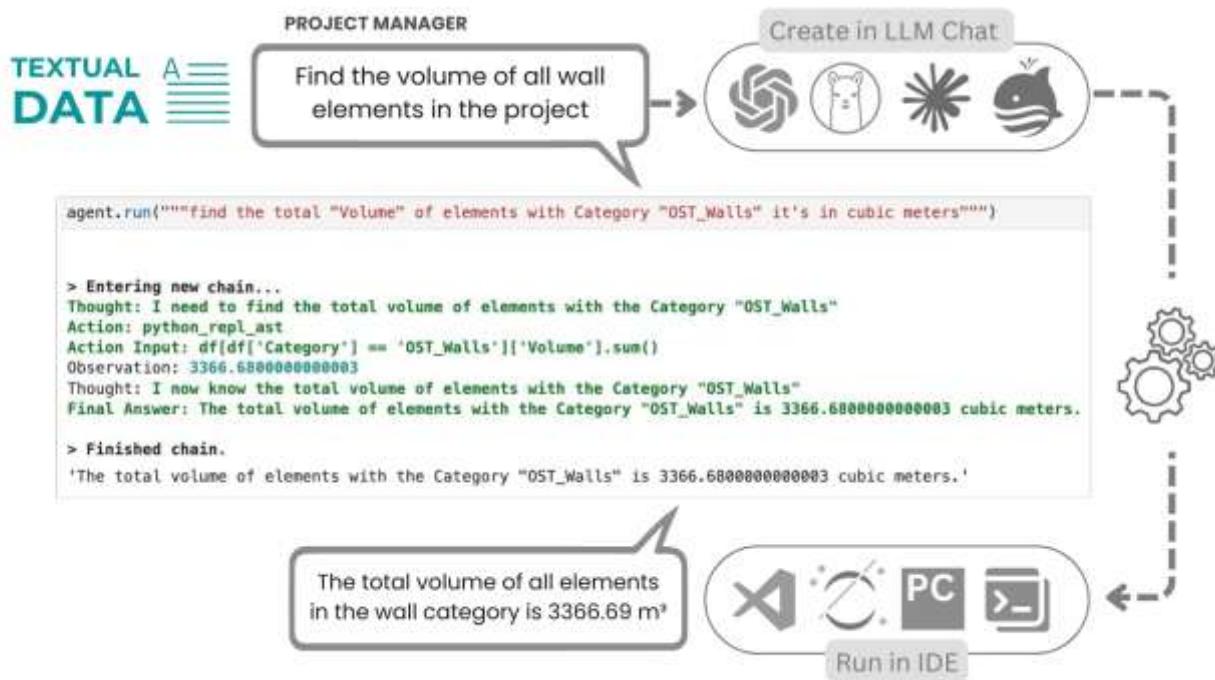


Fig. 5.2-8 LLM model, when working with structured data, understands from the context of a text query what grouping and attributes the user is asking about.

If it is necessary to retrieve quantities for only one group of elements, it is sufficient to perform a simple QTO query (Fig. 5.2-7) on the CAD model data. However, when calculating a budget or estimate for an entire project consisting of many groups of elements, it is often necessary to extract quantities for all types of elements (classes), where each category of elements is processed separately - with grouping by appropriate attributes.

In the practice of estimators and appraisers, individual grouping and calculation rules are used for different types of objects. For example, windows are usually grouped by floors or zones (grouping parameter - attribute Level, Rooms), and walls - by material or construction type (parameter Material, Type). To automate the grouping process, such rules are described in advance in the form of grouping rule tables. These tables act as configuration templates that define which attributes should be

used in calculations for each group of elements in the project.

QTO calculation of the entire project using group rules from an Excel spreadsheet

In real construction projects, it is often necessary to perform aggregation by several attributes simultaneously within one group of elements. For example, when working with the category "Windows" (where the Category attribute contains values like OST_Windows or IfcWindows), elements can be grouped not only by type - for example, by the value in the Type Name or Type field - but also by additional characteristics, such as the thermal conductivity level specified in the corresponding attribute. Such multidimensional grouping allows for more accurate results for a particular group. Similarly, when calculating wall or floor categories, arbitrary combinations of attributes - such as material, level, floor, fire resistance, and other parameters - can be used as filters or grouping criteria (Fig. 5.2-9).

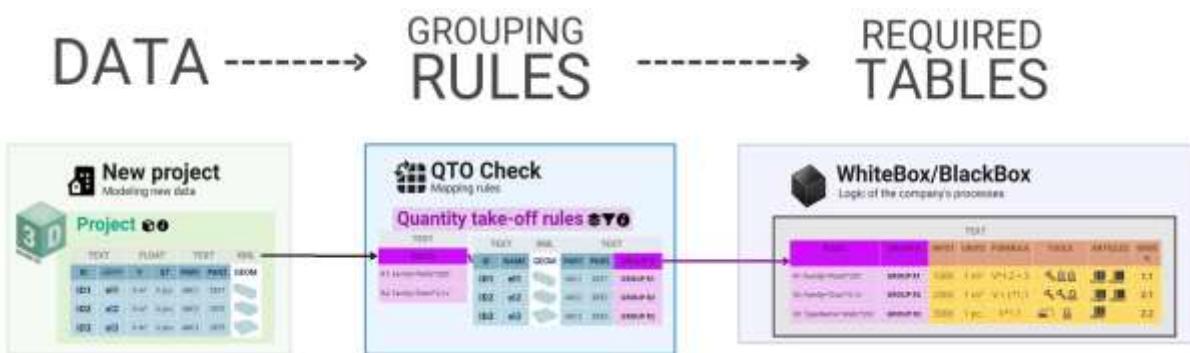


Fig. 5.2-9 For each group or category of entities in a project, there is a different grouping formula consisting of one or more criteria.

The process of defining such grouping rules is similar to the process of creating data requirements described in the chapter "Creating Requirements and Quality Checking data" (Fig. 4.4-5), where we discussed working with data models in detail. These grouping and calculation rules ensure the accuracy and relevance of the results to automatically calculate the total attributes of the quantity or volume of an entity category, taking into account all the necessary conditions that must be taken into account in calculations and calculations.

- ➲ The following code sample filters the projects table so that the resulting dataset contains only entities in which the "Category" attribute-column contains the values "OST_Windows" or "IfcWindows" and at the same time the "Type" attribute-column contains the value "Type 1":

I have a DataFrame project - filter the data so that only items that have the attribute "Category" containing the values "OST_Windows" or "IfcWindows" and at the same time the Type attribute contains the value "Type 1remain in the dataset ↴

❷ LLM's response:

```
group.py
df[(df['Category'].isin(['OST_Windows', 'IfcWindows'])) & (df['Type'].str.contains("Type 1"))]
```

Fig. 5.2-10 A single line of code, similar to the Excel formula, allows you to group all project entities by several attributes.

The resulting code (Fig. 5.2-10) after translation of CAD data in structured open formats (Fig. 4.1-13) can be run in one of the popular IDEs (integrated development environments) we mentioned above in offline mode: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse with PyDev plugin, Thonny, Wing IDE, IntelliJ IDEA with Python plugin, JupyterLab or popular online tools: Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

- ❸ To retrieve the project entities in the QTO DataFrame form under the category "Windows" only with a specific thermal conductivity value, we can use the following query to the LLM:

I have a DataFrame project - filter the data so that only records with "Category" containing "OST_Windows" or "IfcWindows" values remain in the dataset, and at the same time the ThermalConductivity column should have a value of 0.↳

2 LLM's response:

```
1 df[(df['Category'].isin(['OST_Windows', 'IfcWindows'])) & (df['ThermalConductivity'] == 0.5)]
```

Create in LLM Chat

Run in IDE

Fig. 5.2-11 The extremely simple Pandas query language Python allows you to run QTOs for any number of projects simultaneously.

In the response received from LLM (Fig. 5.2-11), the logical condition "&" is used to combine two criteria: thermal conductivity value and belonging to one of the two categories. The "isin" method checks whether the value of the attribute-column "Category" is contained in the provided list.

In projects with a large number of element groups, with different grouping logic - for each category of project entities (e.g.: windows, doors, slabs) it is necessary to set individual grouping rules, which may include additional coefficients or total attribute calculation formulas. These formulas (Fig. 5.2-12 attribute "formel", e.g. x-value of quantity and y-volume of group) and coefficients take into account the unique characteristics of each group, for example:

- add% to material volume to account for overruns
- fixed additional quantity of material
- adjustments related to possible risks and calculation errors in the form of formulas

Once the filtering and grouping rules are formulated in the form of parameter formulas for each item category, they can be saved as a line-by-line table - for example, in Excel format (Fig. 5.2-12). By storing these rules in a structured form, the process of extracting, filtering and grouping project data can be fully automated. Instead of manually writing many separate queries, the system simply reads the parameter table and applies the appropriate rules to the model (the overall project dataframe (Fig. 4.1-13)), generating final QTO tables for each category of project elements.

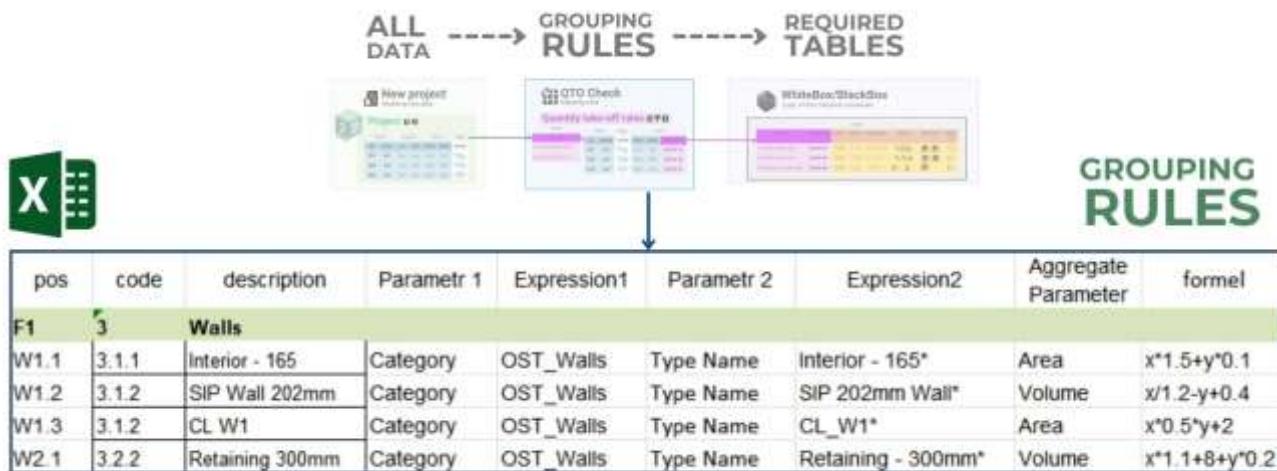


Fig. 5.2-12 QTO Attribute Grouping Table establishes rules for grouping project elements, providing an accurate total number and scope for each category.

The collected rules will allow you to group the entire project and perform all necessary calculations, including adjustments to volume attributes. As a result, the volumes are brought to the "real volume", which is used for calculations and calculations, and not those that were originally at the design stage in the CAD model.

In the process of automatically creating QTO volume tables for the entire project, the application must go through all the categories of the grouping rules table, take the grouping attributes, group all project elements by them, and aggregate the volume attribute for that group, additionally multiplying it by a refining factor or coefficient.

Let's ask LLM to write code for us for such a solution, where the code would have to load two tables, a table of grouping rules (Fig. 5.2-12) and a table of the project data itself (Fig. 4.1-13), and then apply the grouping rules, group the items according to the given rules, calculate the aggregated values and save the results to a new Excel file.

💡 Send a text request to the LLM chat:

```
I need code to read project data from file 'basic_sample_project.xlsx' and then rules from 'Grouping_rules_QT0.xlsx' by which to group all data by 'Parameter 1' and 'Parameter 2', aggregate 'Aggregate Parameter', filter by 'Expression2', perform calculations from 'Formel1' and save QTO table to 'QTQ_table2.xlsx'
```

💡 LLM's response:

```

1 import pandas as pd
2
3 # Reading CSV and Excel files with project data and grouping rules respectively
4 df = pd.read_csv('../input/rac_basic_sample_project.xlsx', low_memory=False)
5 df_qto = pd.read_excel('../input//Grouping_rules_QTO.xlsx', header=0, skiprows=[1])
6
7 # Cleaning numerical columns in the dataframe by removing non-numeric characters
8 numeric_cols = ['Area', 'Volume', 'Depth', 'Width', 'Length', 'Perimeter']
9 df[numeric_cols] = df[numeric_cols]
10 df.apply(lambda col: pd.to_numeric(col.astype(str).str.extract('(\d*\.\d*)')[0]).fillna(0))
11
12 # Define a function to process each row in the grouping rules dataframe
13 def process_row(row):
14     # Group by specified parameters and calculate sum and count for the aggregate parameter
15     group = df.groupby([row["Parameter 1"], row["Parameter 2"]])
16     [][row["Aggregate Parameter"]].agg(['sum', 'count'])
17     group = group[group.index.get_level_values(1).str.contains(row["Expression2"], regex=True)]
18
19     # Perform calculation defined in the rules
20     group['After Calculation'] = eval(row["formula"], {'x': group['sum'], 'y': group['count']})
21     return group.reset_index()
22
23 # Apply the function to each row in the grouping rules dataframe and concatenate the results
24 dfgroups = pd.concat([process_row(row) for _, row in df_qto.iterrows()]).reset_index(drop=True)
25
26 # Save the processed data to an Excel file using the OpenPyXL engine
27 dfgroups.to_excel('QTO_table2.xlsx', engine='openpyxl', index=False)

```

Fig. 5.2-13 Python code Pandas processes design data from a CAD project, grouping it according to rules defined in Excel.

The final result of the code execution (Fig. 5.2-13) will be an entity-group table that contains not only the generalized volume attributes from the original CAD- (BIM-) model, but also a new real volume attribute that takes into account all the requirements for correct creation of calculations and estimates (Example Fig. 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 Partition	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

Fig. 5.2-14 The "After Calculation" attribute is added to the summary table after code is executed that will automatically calculate the actual volume.

The resulting code (Fig. 5.2-13) can be run in one of the popular IDEs (which we mentioned above) and applied to any number of existing or new incoming projects (RVT, IFC, DWG, NWS, DGN etc.), be it a few projects or perhaps hundreds of projects in different formats in a structured form (Fig. 5.2-15).

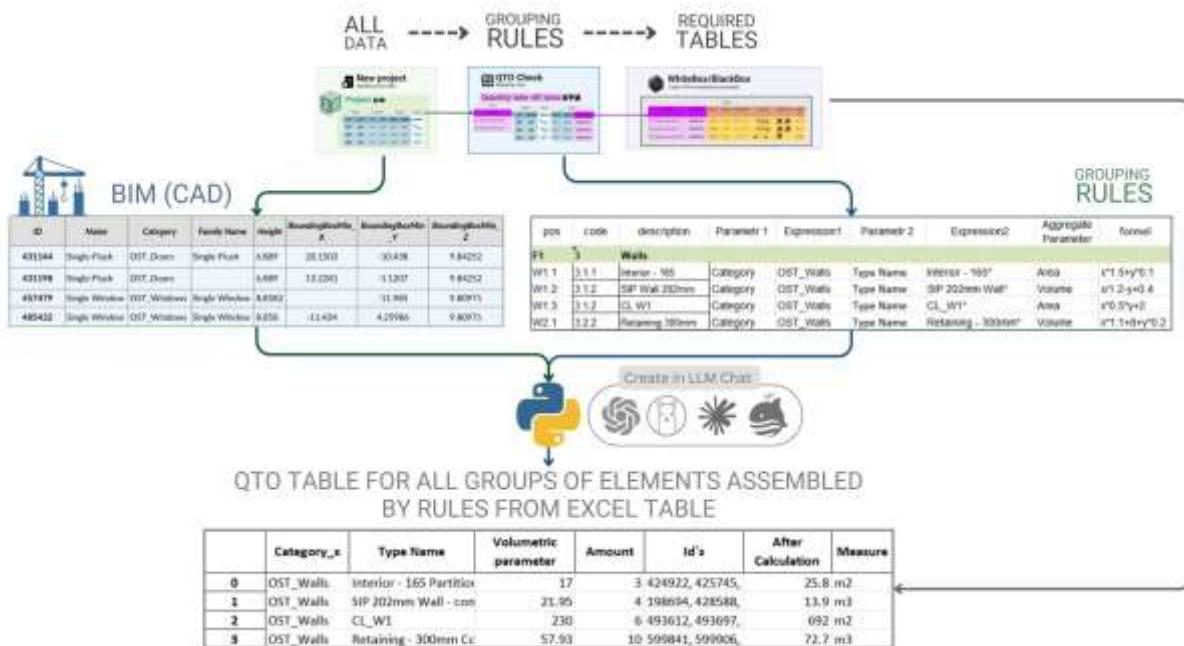
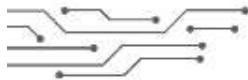


Fig. 5.2-15 The automatic construction data grouping process links BIM data (CAD) to QTO tables via rules from an Excel spreadsheet.

The customized and parameterized volumetric data collection process (Fig. 5.2-15) allows fully automated collection of data on quantitative attributes and volumes of project elements for further work with them, including cost estimation, logistics, work schedules and carbon footprint calculation and other analytical tasks.

Having learned tools that allow us to easily organize and group groups of project elements by specific attributes, we are now ready to integrate grouped and filtered projects with various company calculations and business scenarios.



CHAPTER 5.3.

4D, 6D -8D AND CALCULATION OF CARBON DIOXIDE EMISSIONS CO₂

4D model: integrating time into construction estimates

In addition to costing, one of the key applications of design data in construction is the determination of time parameters, both for individual construction operations and for the entire project. The resource-based estimating method and the associated calculation database, discussed in detail in the previous chapter "Estimates and Estimates for Construction Projects", are often used as a basis for automated calculation of time and the formation of a schedule for the execution of work.

The resource-based approach takes into account not only material costs but also time resources. In costing, each process can be assigned a work order attribute (Fig. 5.3-1 - Work order parameter) and the amount of time and cost associated with the execution of that process. These parameters are particularly important for describing operations that do not have a fixed market price and are not directly procurable - such as the use of construction equipment, worker employment, or logistical processes (which are usually expressed generally in hours). In such cases, the cost is not determined by the procurement department but directly by the implementing company based on internal norms or production rates (Fig. 5.3-1).

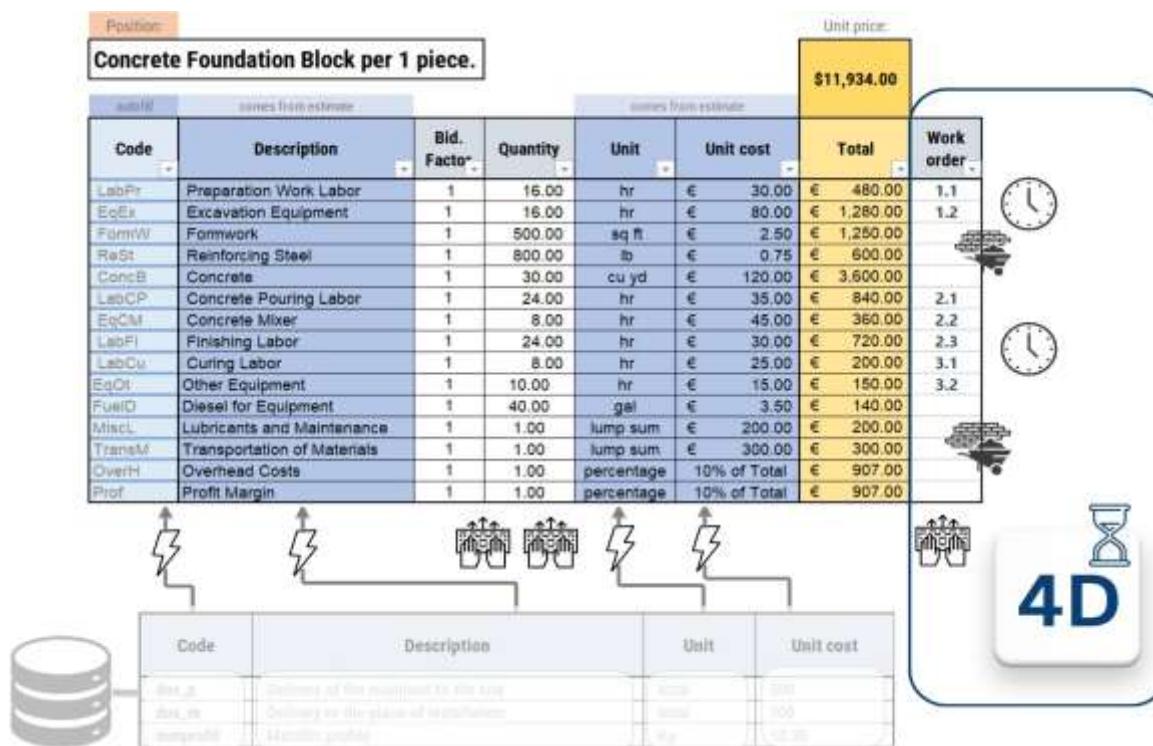


Fig. 5.3-1 The work calculations in the resource-based method of estimation include temporary labor hours.

Thus, calculations at the costing level include not only fuel and material costs (purchase cost), but also the time spent on site by drivers, technicians and auxiliary workers. In the example shown (Fig. 5.3-1), the cost table represents the cost of installing a foundation block, including the component stages of the work, such as preparation, frame installation and concrete pouring, as well as the materials and labor required. At the same time, some operations, e.g. preparation work, may have no material costs, but may contain significant temporary labor costs expressed in man-hours.

To plan the work sequence (for the work schedule) on the construction site, the attribute "Work order" is manually added to the calculation table (Fig. 5.3-1). This attribute is specified in an additional column only for items whose unit of measure is expressed in time (hour, day). This attribute is in addition to the work code, description, quantity, unit of measure (parameter "Unit") and costs. The numerical sequence (parameter "Work order") of activities allows you to set the order of tasks to be performed on the construction site and use it for scheduling.

Construction schedule and its automation based on costing data

The construction schedule is a visual representation of the plan of activities and processes to be performed as part of the project implementation. It is created on the basis of detailed resource calculations (Fig. 5.3-1), where each task-job is scheduled, in addition to the cost of resources, by time and sequence.

In contrast to averaging approaches, where time calculations are based on a typical number of hours for installing materials or equipment, in the resource-based method, planning is based on actual data in the costing. Each item of the estimate related to labor costs is based on the applied calendar, which takes into account the actual conditions of resource use during the working period. Adjustment of productive hours through coefficients at the level of calculations (Fig. 5.3-1 parameter "Bid. Factor"), allows to take into account differences in productivity and seasonal peculiarities affecting the timing of work.

To determine the process start and end dates for the construction schedule in the Gantt chart, we take the values of the time volume attribute for each item from the foundation block costing and multiply them by the number of blocks (in this case, the number of concrete foundation blocks). This calculation gives the duration of each task. We then plot these durations on a timeline, starting from the project start date, to plot a graph and the result is a visual representation showing when each task should start and finish. The "Work order" parameter for processes additionally allows us to understand whether the work process is running in parallel ("Work order" e.g. 1.1-1.1) or sequentially (1.1-1.2).

Gantt Chart is a graphical tool for planning and project management that represents tasks as horizontal bars on a timeline. Each bar represents the duration of the task, its start and end.

A schedule, or Gantt chart, helps project managers and workers clearly understand when and in what sequence the various phases of construction should be performed, ensuring efficient use of resources

and meeting deadlines.

Let's present a schedule for the installation of three concrete foundation blocks using the calculations from the table above. Using the cost table (Fig. 5.3-1) from the example above, let's ask LLM to schedule the installation of 3 foundation block elements, e.g. for the first of May 2024.

To send a costing to LLM, we can upload the costing table in XLSX format or simply insert a screenshot of a JPEG image of the costing directly into the LLM chat (Fig. 5.3-1). LLM will independently find a library to visualize the table image and, by multiplying the time attributes of the activities from the table by their volume, will add all the data into a graph.

- ❷ Send a text request to the LLM at:

Create a work schedule in the form of a Gantt chart using values from the time related table (Fig. 5.3-1 as JPEG in the appendix) for the installation of 3 foundation blocks. The work for each block shall be performed sequentially. Set the start of work to 01/05/2024 ↴

- ❸ LLM's response

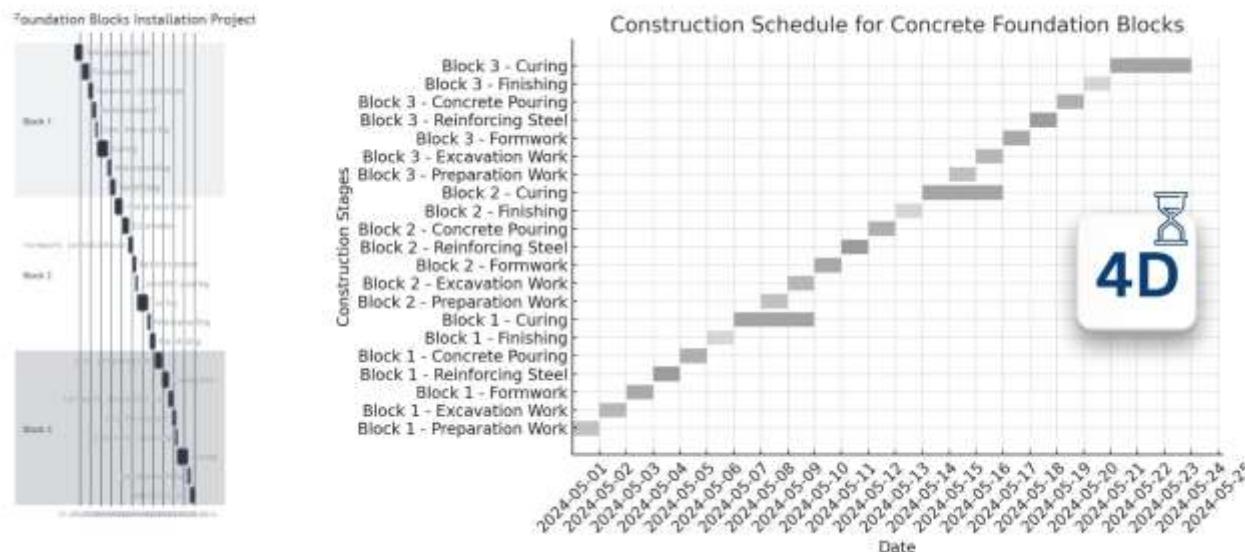


Fig. 5.3-2 Automatically generated by multiple LLMs Gantt chart shows the stages of construction of three concrete blocks, according to the conditions from the prompt.

The resulting graph (Fig. 5.3-2) is a time diagram in which each horizontal bar corresponds to a certain stage of work on the foundation block and shows the sequence of operations (parameter "Work order"), such as preparation, excavation, formwork installation, reinforcement, concrete pouring and finishing, i.e. those processes that have filled in time parameters and sequence in the calculations.

Such a schedule (Fig. 5.3-2) does not take into account constraints related to working days, shifts or working time standards, but is intended solely for conceptual visualization of the process. An accurate

schedule that will reflect the concurrency of work can be supplemented with appropriate prompts or additional instructions within the chat room

Using a single costing (Fig. 5.3-1), thanks to volume attributes from 3D -geometry, it is possible to automatically estimate both the project cost through automated estimates and at the same time to calculate the time characteristics of groups in the form of tables or graphs for different project variants (Fig. 5.3-3).

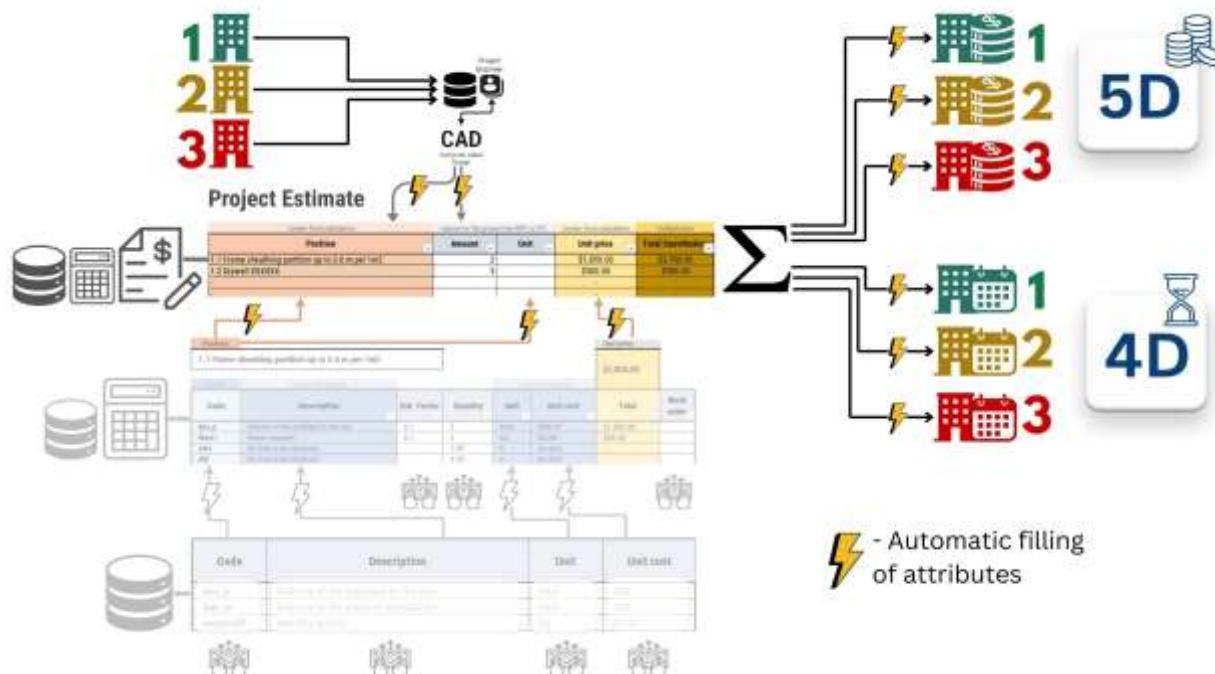


Fig. 5.3-3 Automatic Calculation, allows you to instantly and automatically forecast costs and time for various project options.

Modern modular ERP -systems (Fig. 5.4-4), loading data from CAD models use similar automated time calculation methods that significantly reduce the decision-making process. This allows you to instantly and accurately plan work schedules and calculate the total time required to complete all tasks in the project realization, taking into account real prices.

Extended attribute layers 6D -8D: from energy efficiency to safety assurance

6D, 7D and 8D are extended levels of information modeling, each of which contributes additional layers of attributes to the comprehensive project information model, the basis of which are the attributes of the 3D -model with their number and scope. Each additional layer contributes specific parameters that are required for further grouping or further identification in other systems, such as for example real estate management systems (PMS), computer-aided facilities management (CAFM), construction project management (CPM) and safety management systems (SMS).



Fig. 5.3-4 Attributes 6D, 7D and 8D in the data information model expand the consideration of various aspects of the project, from energy efficiency to safety.

- In **6D**, in addition to the project database (or data frame (Fig. 4.1-13)) with geometric and volumetric attributes of elements, information (attribute-columns) on environmental sustainability is added. This includes information related to energy efficiency, carbon footprint, recyclability of materials, and use of environmentally friendly technologies. This data allows the project's environmental impact to be assessed, project decisions to be optimized and Sustainable Development Goals (ESG) to be achieved.
- **7D** attributes supplement the attributes needed to manage building maintenance. These are data on maintenance schedules, component life cycles, technical documentation and repair history. This set of information ensures that the model can be integrated with maintenance systems (CAFM, AMS), allows for efficient scheduling of maintenance, replacement of equipment and provides support throughout the entire lifecycle of the facility.
- **8D** additional attribute layer, - includes information related to safety - both at the construction stage and during subsequent operation. The model includes measures to ensure the safety of personnel, emergency instructions, requirements for evacuation systems and fire protection. The integration of this data into the digital model helps to take risks into account in advance and develop architectural, engineering and organizational solutions that take into account health and safety requirements.

In structured tabular form, layers 4D to 8D represent additional attributes in the form of columns with populated values (Fig. 5.3-5) added to the already populated 3D -model attributes such as name, category, type and volumetric characteristics. The values in attribute layers 6D, 7D and 8D contain additional textual and numerical data such as recycling percentage, carbon footprint, warranty period, replacement cycle, installation date, safety protocols, etc.



Fig. 5.3-5 6D -8D add attribute layers to the data information model, which already contains geometric and volumetric attributes from the 3D -model.

For our new window (Fig. 4.4-1), the element with the identifier W-NEW (Fig. 5.3-5) can have the following 3D -8D attributes:

3D -attributes - geometric information obtained from CAD systems:

- "Type name" - element "Window"
- "Width" - 120 cm
- Additionally, you can add the "Bounding Box" points of an element or its "geometry BREP / MESH" as a separate attribute

Attributes of 6D - environmental sustainability:

- Recyclability rate of 90%
- "Carbon Footprint" - 1,622 kg CO₂

Attributes 7D - object management data:

- "Warranty period" - 8 years
- "Replacement Cycle" is 20 years old
- "Maintenance" - required annually

Attributes of 8D - ensuring the safe use and operation of buildings:

- "Installed" window - by "XYZ Windows" company
- "Safety Standard" - complies with ISO 45001

All parameters recorded in a database or dataset (Fig. 5.3-5) are needed by specialists in different departments for grouping, searching or calculations. This multidimensional attribute-based description of project objects provides a complete picture of their life cycle, operational requirements, and many other aspects necessary for project design, construction, and operation.

Estimating CO₂ and calculating carbon dioxide emissions from construction projects

In addition to the topic of sustainability of construction projects at stage 6D (Fig. 5.3-5), modern construction focuses on the environmental sustainability of projects, where one of the key aspects becomes the assessment and minimization of carbon dioxide CO₂ emissions, which occur during the stages of the project life cycle (e.g. manufacturing and installation).

Estimating and calculating the carbon emissions of building materials is the process by which total carbon emissions are determined by multiplying the volumetric attributes of an element or group of elements used in a project by a suitable carbon emission factor for the category.

Considering carbon emissions in the assessment of construction projects as part of the wider ESG criteria (environmental, social and governance) adds a new level of complexity to the analysis. This is particularly important for the client-investor in obtaining a relevant certification such as LEED® (Leadership in Energy and Environmental Design), BREEAM® (Building Research Establishment Environmental Assessment Method) or DGNB® (Deutsche Gesellschaft für Nachhaltiges Bauen). Obtaining one of these certifications can significantly increase the marketability of a facility, simplify commissioning and ensure compliance with sustainability-oriented tenants (ESG). Depending on the project requirements, HQE (Haute Qualité Environnementale, the French green building standard), WELL (WELL Building Standard, focused on user health and comfort) and GRESB (Global Real Estate Sustainability Benchmark) may also be used

Environmental, social and governance **ESG** (environmental, social and governance) is a broad set of principles that can be used to assess the corporate governance, social and environmental impact of a business both internally and externally.

ESG, originally developed in the early 2000s by financial funds to provide investors with information on broad environmental, social and governance criteria, has evolved into a key indicator for evaluating both companies and projects, including construction projects. According to research by major consulting firms, environmental, social and governance (ESG) considerations are becoming an integral part of the construction industry.

According to EY (2023) "The Path of Carbon Neutrality", companies that actively implement ESG -principles, not only reduce long-term risks, but also increase the efficiency of their business models, which is especially important in the global transformation of markets [103]. PwC's ESG Awareness report notes that companies' awareness of the importance of ESG -factors ranges from 67% to 97%, with most organizations seeing these trends as key to future sustainability [104] and that businesses for the most part are seeing significant pressure from stakeholders to integrate ESG principles.

Thus, the integration of ESG -principles into construction projects not only contributes to obtaining international sustainability certifications such as LEED, BREEAM, DGNB, but also ensures the long-term sustainability and competitiveness of companies in the industry.

One of the most significant factors affecting the overall carbon footprint of a construction project is the production and logistics stages of construction materials and elements. The materials used on site often have a decisive impact on total CO₂ emissions, especially in the early stages of the project life cycle - from extraction of raw materials to delivery to the construction site.

Calculating emissions by category or type of building element requires the use of reference carbon emission factors that reflect the amount of CO₂ generated from the production of different materials. These materials include concrete, bricks, recycled steel, aluminum, and others. These values are generally extracted from reputable sources and international databases such as UK ICE 2015 (Inventory of Carbon and Energy) and US EPA 2006 (U.S. Environmental Protection Agency) [105]. The following table (Fig. 5.3-6) summarizes the baseline emission factors for a range of common building materials. Two key parameters are shown for each material: specific CO₂ emissions (in kilograms per kilogram of material) and volume-to-weight conversion factors (in kilograms per cubic meter), which are necessary for integrating the calculations into the design model and linking to the QTO data grouping.



Carbon Emitted in Production		UK ICE Database (2015) USEPA (2006)	UK ICE Database (2015) USEPA (2006)	Coefficient m³ to kg
Material	Abbreviated	Process Emissions (kg CO2e/kg of product) (K1)	Process Emissions (kg CO2e/kg of product) (K2)	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

Fig. 5.3-6 Amount of carbon emitted during the production of different building materials, according to the UK ICE and US EPA database.

To calculate the total CO₂ emissions for a project, as with the 4D and 5D calculations, it is necessary to determine the attribute volumes of each group of facilities. This can be done using quantitative analysis tools (QTO) by obtaining the attribute volumes in cubic meters, as discussed in detail in the section on Quantity take-off. These volumes are then multiplied by the appropriate factors for the "CO₂ process emissions" attribute of each material group.

- ➲ Let's automatically extract the table of volumes by element type from the CAD (BIM) project by grouping all project data as already done in the previous chapters. To accomplish this task, let's refer to LLM.

Please group the DataFrame table from CAD (BIM) project by the column parameter "ObjectType" (or "Type") and show the number of items in each group, and summarize the "Volume" parameter for all items in the type. ↴

💡 LLM's response:

The screenshot shows a user interface for generating Python code to process a DataFrame. At the top, there are icons for creating a chat with an AI, and below them are four icons: a brain, a person, a sun, and a fish. The text 'group.py' is displayed next to a file icon. A large button labeled 'Create in LLM Chat' is at the top right.

Data Flow:

- A central box contains the generated Python code:

```

1 import pandas as pd
2
3 # Import Projekt as CSV train DataFrame
4 df = pd.read_csv('your_dataframe.csv')
5
6 # Group, count, and sum 'Volume' in one line
7 result = df.groupby('ObjectType')['Volume'].agg(Amount='size', Sum of Volumes='sum')

```

- To the left, a vertical flowchart shows the data processing steps: 'Data' leads to 'f(x)', which leads to 'Output'.
- Below the code box, there is a 'Run in IDE' button with icons for VS Code, PyCharm, and Jupyter Notebook.
- Output:** A table titled 'ObjectType' is shown, containing the grouped data:

ObjectType	Amount	Sum of Volumes, m ³
IFC_Vebo_staalplatei_L150B_90 x 150	31	412
IFC_Vebo_staalplatei_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

Fig. 5.3-7 The generated code in LLM has grouped the project entities for us by type (ObjectType) with the summarized attribute "Volume".

To automate the calculation of total CO₂ emissions for the entire project, simply set up automatic data mapping in the table or manually associate the element types (Fig. 5.3-7) with the corresponding material types (Fig. 5.3-6) from the emission factor table. The finished table with emission factors and formulas, as well as code to retrieve volumes from CAD formats (BIM) and automate CO₂ determination can be found on GitHub by searching for "CO₂_calculating-the-embodied-carbon. DataDrivenConstruction." [106].

Thus, data integration after grouping QTO elements from the database CAD allows automatic calculation of carbon dioxide emissions (Fig. 5.3-8) for different design options. This makes it possible to analyze the impact of different materials in different variants and to select only those solutions that meet customer's requirements for CO₂ emissions to obtain a particular certificate when the building is commissioned.

Estimating CO₂ emissions by multiplying factors by the volumes of grouped project elements is a typical example of a task in the process of a construction company obtaining an ESG rating (e.g., LEED certification) for a facility.

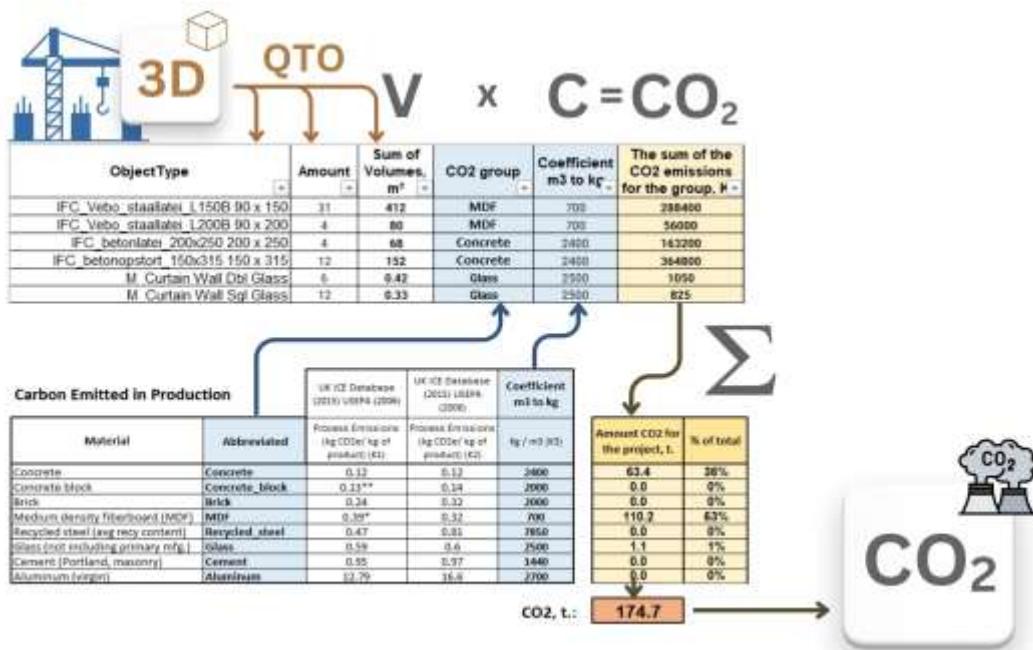


Fig. 5.3-8 Integration of QTO groups from CAD databases provides accuracy and automation in deriving estimates of final CO₂ emissions.

Similarly, by defining element group volumes, we can perform calculations for material control and logistics, quality monitoring and management, energy modeling and analysis, and a host of other tasks to obtain a new attribute status (parameter in the table) for both individual element groups and the entire project.

If the number of such calculation processes in the company starts to grow, the question arises about the need to automate such calculations and implement the results of calculations into the company's processes and data management systems.

Due to the complexity of a comprehensive solution, medium and large companies operating in the construction industry outsource such automation to ERP (or PMIS) system development companies. The development companies create a single comprehensive modular system for large clients to manage many different information layers, including material and resource calculations.



CHAPTER 5.4.

CONSTRUCTION ERP AND PMIS SYSTEMS

Construction ERP -systems on the example of calculations and estimates

Modular ERP systems integrate various attribute (information) layers and data flows into a single comprehensive system, allowing project managers to manage resources, finances, logistics and other aspects of a project in a synchronized manner within a single platform. A construction ERP system acts as the "brain" of construction projects, simplifying repetitive processes through automation, providing transparency and control throughout the construction process.

Construction ERP -systems (Enterprise Resource Planning) are comprehensive software solutions designed to manage and optimize various aspects of the construction process. At the core of construction ERP systems are modules for managing costing and scheduling, making them an important tool for efficient resource planning.

ERP modules -systems allow users to enter, process and analyze data in a structured manner covering various aspects of a project, which may include material and labor cost accounting, equipment utilization, logistics management, human resources, contacts and other construction activities.

One of the functional blocks of the system is the module of business logic automation - BlackBox/WhiteBox, which plays the role of the process control center.

BlackBox /WhiteBox allows professionals using an ERP -system to flexibly manage, via access rights, various aspects of the business that have already been pre-configured by other users or administrators. In the context of ERP systems, the terms *BlackBox* and *WhiteBox* refer to the levels of transparency and controllability of the system's internal logic:

- **BlackBox** ("black box") - the user interacts with the system through the interface, without access to the internal logic of process execution. The system performs calculations on its own, based on predefined rules hidden from the end user. He enters data and gets the result without knowing what attributes or coefficients were used inside.
- **WhiteBox** ("white box") - process logic is available for viewing, customization and modification. Advanced users, administrators or integrators can manually define data processing algorithms, calculation rules and interaction scenarios between project entities.

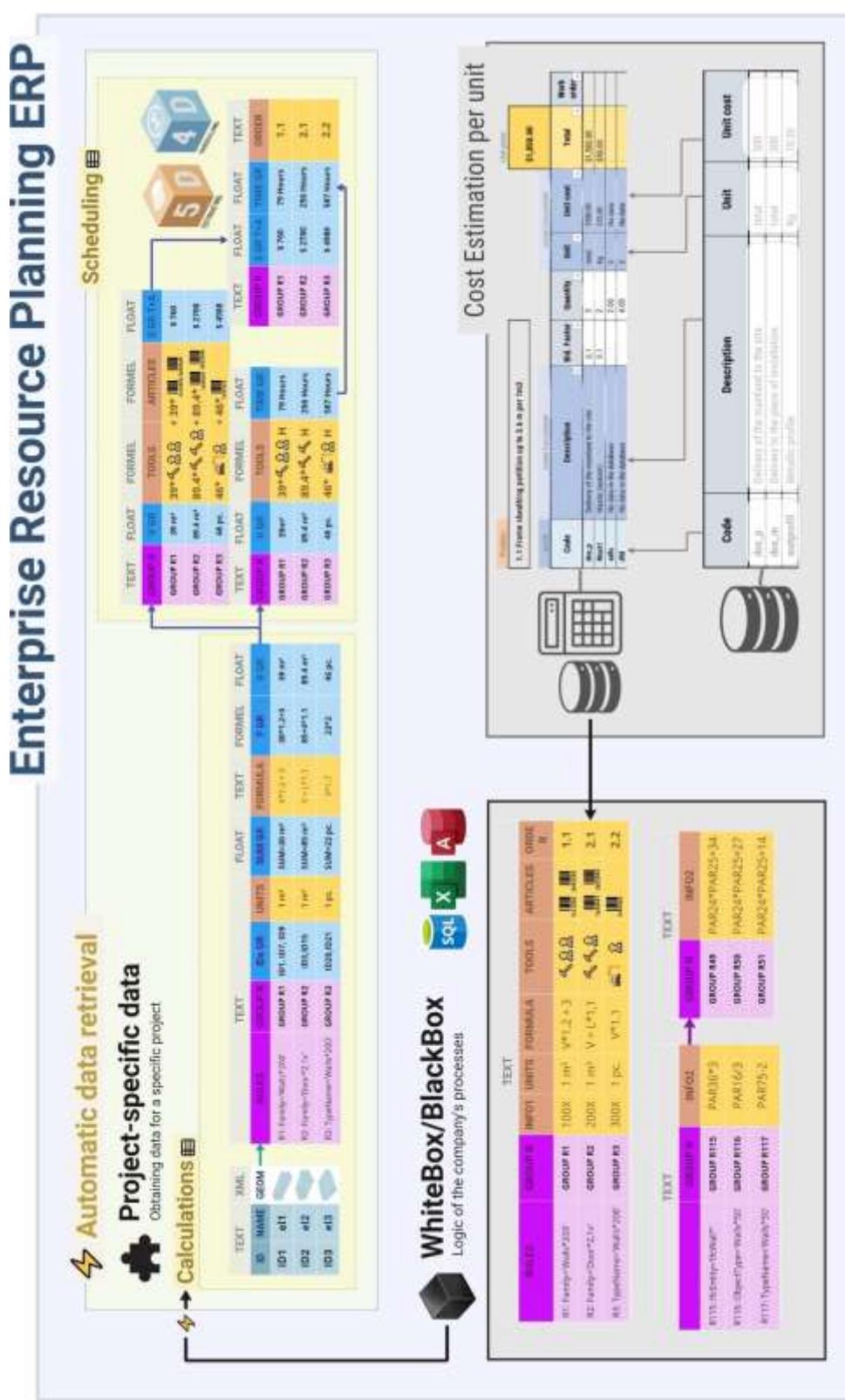


Fig. 5.4-1 Construction ERP Architecture -system, to obtain estimates and work schedules when manually filling in the scope attributes.

An example is where an experienced user or administrator sets a rule: which attributes in an estimate should be multiplied among themselves or grouped by a certain attribute, and where the final result should be recorded. Later, less trained professionals, such as estimating engineers, simply load new data into the ERP through the user interface - and get ready estimates, schedules or specifications without having to write code or understand the technical details of the logic.

In the previous chapters, the calculation and logic modules were discussed in the context of LLM interactions. In an ERP environment, such calculations and transformations take place inside modules hidden behind an interface of buttons and forms.

In the following example (Fig. 5.4-1), the administrator of the ERP -system in the BlackBox /WhiteBox module has defined rules for matching attributes of entities from estimates with attributes for grouping QTO. Thanks to this customized (by manager or administrator) BlackBox/WhiteBox module, the user (estimator or engineer), by manually adding a quantity or volume attribute through the ERP user interface, automatically receives the finished estimates and work schedules. In this way, the calculation and estimate generation processes discussed in the previous chapters with the help of code, inside the ERP, becoming a semi-automated conveyor.

Connecting this semi-automated process to the volumetric attributes from CAD (BIM) models (Fig. 4.1-13), through, for example, loading the CAD project into a preconfigured ERP module, turns the data flow into a synchronized mechanism capable of autonomously and instantly updating the value of individual groups of elements or the entire project in response to any changes in it during the design phase, when loading the CAD model into the ERP.

In order to create an automated data flow (Fig. 5.4-2) between CAD (BIM) and ERP systems, the basic processes and requirements for data from CAD (BIM) model databases must be defined in a structured way, as we have already discussed in the chapter above "Requirements and Data Quality Assurance ". This process in ERP is divided into similar steps:

- **Creating validation rules (1)**, which play an important role in ensuring the accuracy of the data entering the ERP -system. Validation rules serve as filters that validate entities and their attributes, allowing only those items that pass the requirements to enter the system. Learn more about verification and validation in the chapter "Creating requirements and validating data quality".
- **A verification process (2)** then takes place inside ERP, which confirms that all project entity elements with their attributes and values have been created correctly and are ready for the next processing steps.
- If there are problems with incomplete attribute data, **a report (3) is generated** and the project, along with instructions for correction, is sent for revision until ready for the next iteration.
- Once the project data has been validated and verified, it is used in another ERP module **(4) to create Quantity Take-Off tables (QTO)** that create quantity attributes for entity groups, materials and resources according to previously generated rules (WhiteBox/BlackBox).

- Grouped data by matching rules or QTO are automatically **integrated with calculations (e.g., cost and time) (5)**.
- In the last step of the ERP -system, the user, by multiplying the scope attributes from the QTO table with the attributes of the process tables (e.g., estimated items), **automatically generates calculation results (6)** (e.g., cost estimates, work schedules, or CO₂ emissions) for each entity group and for the project as a whole.

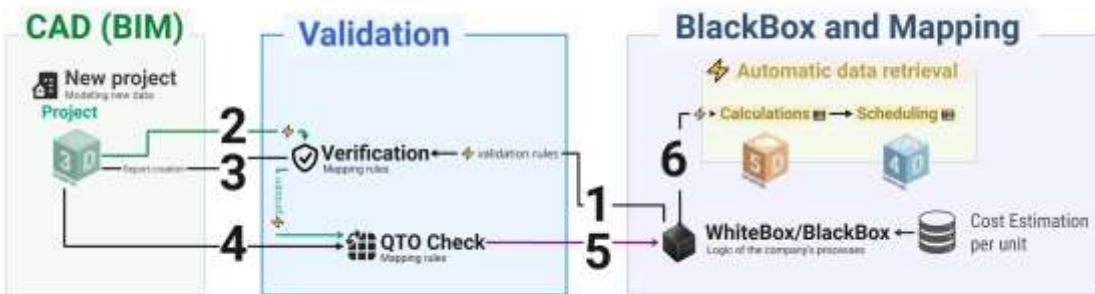


Fig. 5.4-2 Architecture of a construction ERP -system with CAD (BIM), from the creation of validation rules (1) to the automatic calculation of costs and work schedules (5-6).

In a modular ERP -system, processes are integrated using software that includes a user interface. Behind the interface is the back-end, where structured tables process data by performing various operations that the manager or administrator has pre-configured. As a result, the user, thanks to the pre-defined and customized automation logic (in the modules BlackBox /WhiteBox), receives semi-automatically prepared documents that meet his tasks

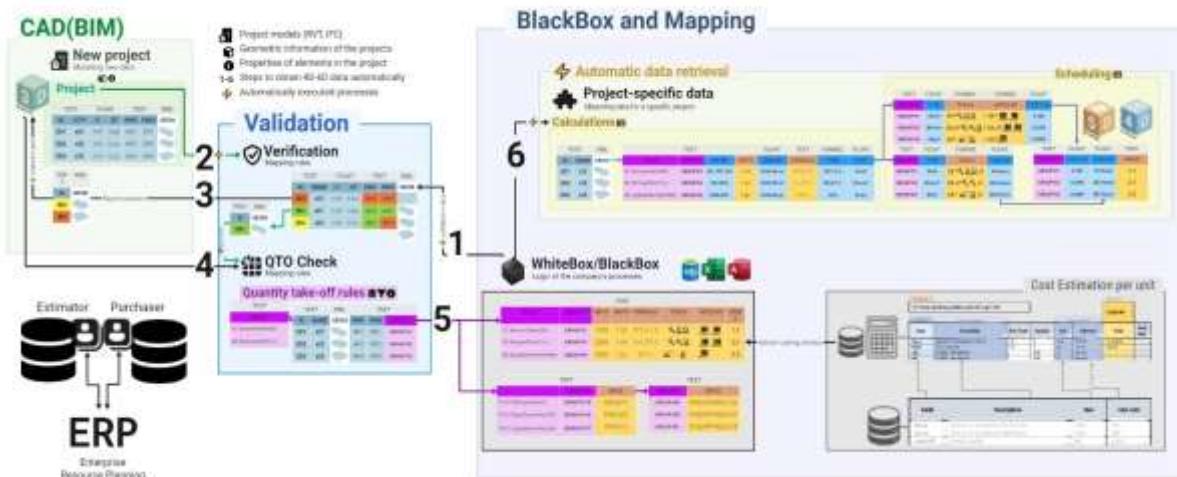


Fig. 5.4-3 ERP -system helps managers and users move between specialist tables to generate new data.

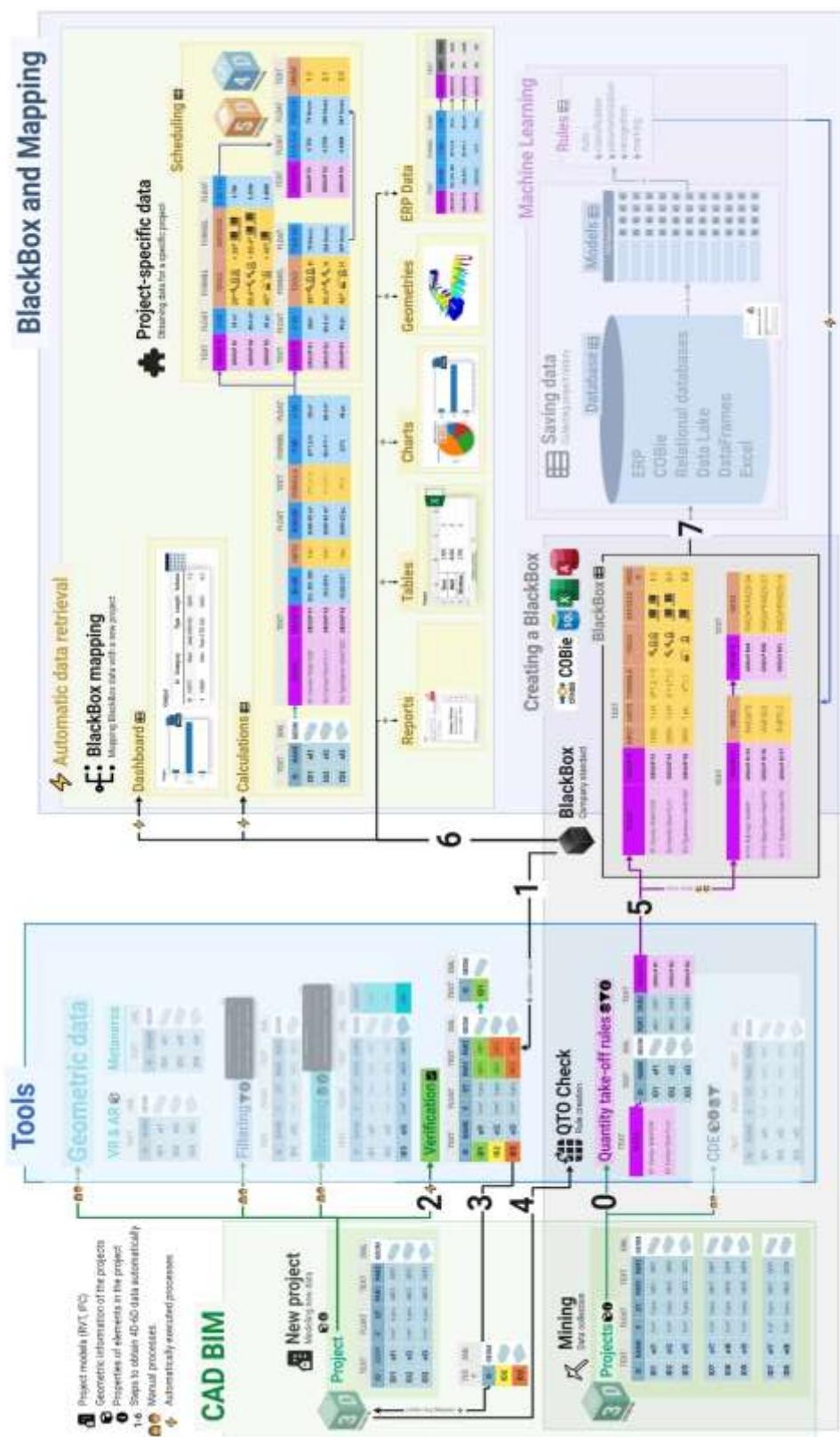


Fig. 5.4-4 ERP -system is integrated with analytical tools and automates the company's decision-making process.

Similarly, the processes in ERP -systems, from inception to final calculation (steps 1-6 Fig. 5.4-3) are a chain of interrelated steps that ultimately provide transparency, efficiency, and accuracy in planning.

Modern construction ERP -systems include not only cost and schedule calculation modules, but also dozens of other pre-configured modules, usually covering document management, project progress tracking, contract management, supply chain and logistics, as well as integration with other business systems and platforms. The integrated analytical tools of ERP allow users to automate the creation of dashboards to monitor project KPIs (KPI - key performance indicators). This provides centralized and consistent management of all aspects of a construction project, with an attempt to integrate a large number of applications and systems on a single platform.

In the future, ERP -analytics will be used in combination with machine learning to improve accuracy and optimize the process of calculating future project attributes. The data and attributes analyzed and collected from ERP systems in Big Data (Fig. 5.4-4) will in the future form the basis for creating predictive models that can accurately anticipate potential delays, risks or, for example, possible changes in material costs.

As an alternative to ERP, the construction industry often uses PMIS (Project Management Information System), a project management system designed for detailed control of tasks at the level of an individual construction project.

PMIS: Intermediate between ERP and the construction site

Unlike ERP, which covers the entire chain of a company's business processes, PMIS focuses on managing a specific project, monitoring timelines, budgets, resources and documentation.

PMIS (Project Management Information System) is construction project management software designed to plan, track, analyze and report on all aspects of a project.

PMIS allows you to manage documents, schedules, budgets and at first glance, PMIS may seem like a duplicate solution to ERP, but the key difference is the level of management:

- **ERP** is focused on the business processes of the company as a whole: managing costs, contracts, procurement, human resources and resources at the corporate level.
- **PMIS** focuses on managing individual projects, providing detailed planning, change control, reporting, and participant coordination.

In many cases, it is the ERP -systems that already have sufficient functionality, and the implementation of PMIS becomes more a matter of convenience and company preference. Many contractors and customers use PMIS not because it is necessary, but because it is imposed by the vendor or a large customer who wants to aggregate data on a particular platform.

It should be mentioned that in the international terminology for construction project management there are other separate popular concepts such as PLM (Product Lifecycle Management) and EPC and EPC-

M (Engineering, Procurement and Construction Management) - contracting methods in the construction industry.

If a company already uses ERP with project management modules, the implementation of PMIS may be an unnecessary link that duplicates functionality. However, if processes are not automated and data is fragmented, PMIS can be a more convenient and easy to maintain tool.

Speculation, profit, insularity and lack of transparency in ERP and PMIS

Despite the external simplicity of interfaces and procedures, construction ERP and PMIS -systems are in most cases closed and inflexible solutions. Such systems are usually delivered as a pre-configured software package from a single vendor, with limited access to internal databases and process logic.

CAD-(BIM-) vendors are increasingly taking over the development and control of such systems, as their databases contain the information required by ERP systems: quantitative and volumetric attributes of project elements. However, instead of providing access to this data in an open or machine-readable format, vendors offer only limited user scenarios and closed processing logic - predefined within Black-Box modules. This reduces the flexibility of the system and prevents it from being adapted to specific project conditions.

Limited data transparency remains one of the key challenges of digital processes in construction. Closed database architecture, lack of access to complete sets of attributes of construction elements, focus on *BlackBox* automation *modules* and lack of open interfaces significantly increase the risks of document bureaucracy. Such limitations create bottlenecks in the decision-making process, make it difficult to verify information, and open the door to data hiding or speculation within ERP/PMIS systems. Users typically receive only limited access - be it a stripped-down interface or a partial API - without the ability to interact with the primary data sources directly. This is especially critical when it comes to parameters automatically generated from CAD projects, such as volumes, areas and quantities used for QTO calculations.

As a consequence, instead of seeking efficiency through process automation, open data, reducing transaction costs and creating new business models, many construction companies focus on managing external parameters - manipulating factors, adjustment factors and calculation methods that affect project costs in closed ERP/PMIS platforms. This creates room for speculation, distorts real production costs and reduces trust between all participants in the construction process

In construction, profit is formed as the difference between the revenue from a completed project and variable costs, which include design, materials, labor and other direct costs directly related to the realization of the project. However, the key factor affecting the value of these costs is not only technology or logistics, but also the speed and accuracy of calculations and the quality of management decisions within the company.

The problem is aggravated by the fact that in most construction companies the cost calculation processes remain non-transparent not only for customers, but also for the employees themselves, who are not part of the estimating or financial departments. Such closedness contributes to the formation within the company of a privileged group of specialists - carriers of "financial expertise", who have the exclusive right to edit attributes and correction factors in ERP/PMIS -systems. These employees, together with the heads of companies, can actually control the financial logic of the project.

Estimators, in such conditions, turn into "financial jugglers", balancing between maximizing the company's profit and the need to maintain a competitive price for the client. At the same time, they have to avoid blatant and gross manipulations in order not to undermine the company's reputation. It is at this stage that coefficients are laid down to hide overestimated volumes or costs of materials and works.

As a result, the main scheme for increasing the efficiency and profitability of companies operating in the construction industry becomes speculation on the prices of materials and works rather than automation and acceleration of decision-making processes (Fig. 5.4-5). Overestimation of the cost of works and materials is carried out by "gray" accounting in closed ERP /PMIS - systems by inflating percentages over the average market prices for materials or volumes of works with the help of coefficients (Fig. 5.1-6), which were discussed in the chapter "Compilation of calculations and calculation of the cost of works on the basis of the resource base".

As a result, the customer receives a calculation that does not reflect the real cost or scope of work, but is a derivative of many hidden internal coefficients. At the same time, subcontractors, in an attempt to meet the underestimated rates set by the general contractor, are often forced to purchase cheaper and low-quality materials, which worsens the final quality of construction.

The speculative process of seeking profits out of thin air ends up hurting both clients who receive unreliable data and execs who are forced to find more and more speculation models

As a result, the larger the project, the higher the level of bureaucracy in data and process management. Every step and every module often hides opaque coefficients and surcharges embedded in calculation algorithms and internal procedures. This not only makes auditing difficult, but also significantly distorts the financial picture of the project. In large construction projects, such practices often lead to a multiple (sometimes up to tenfold) increase in the final cost, while the real volumes and costs remain

outside the effective control of the client (Fig. 2.1-3 Comparison of planned and actual costs of large infrastructure projects in Germany).

According to McKinsey & Company's report Imagining the Digital Future of Construction (2016), large construction projects are on average completed 20% later than planned and up to 80% over budget [107].

Estimating and budgeting departments are becoming the most guarded link within a company. Access to them is strictly limited even for internal specialists, and due to the closed logic and database structures, it is impossible to objectively assess the effectiveness of project decisions without distortions. The lack of transparency leads to the fact that companies are forced not to optimize processes, but to fight for survival by "creative" management of figures and coefficients (Fig. 5.3-1, Fig. 5.1-6 - for example, the parameter "Bid. Factor").

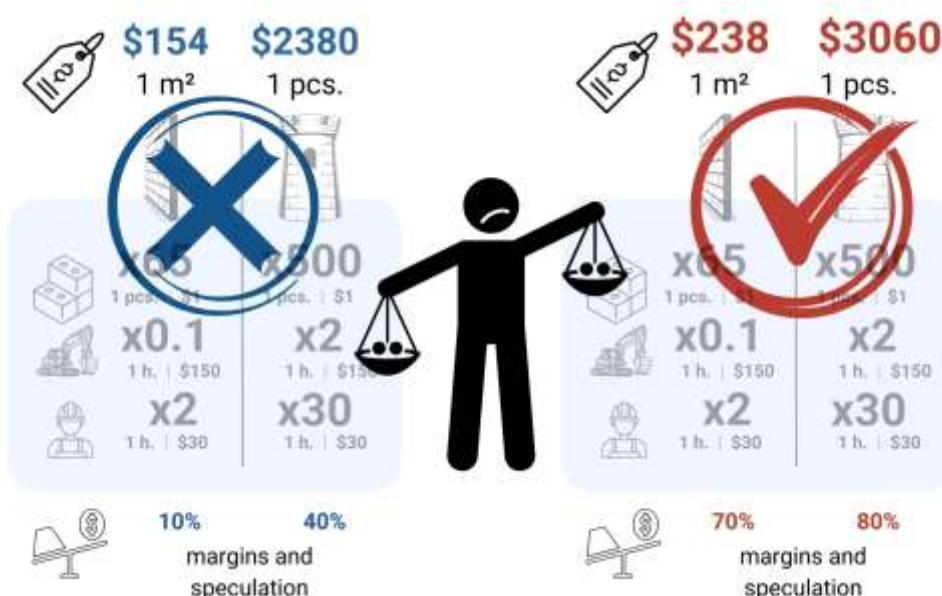


Fig. 5.4-5 Settlement-level speculation ratios are the main profit of companies and the art of juggling between quality of work and reputation.

All this casts doubt on the feasibility of further use of closed ERP/PMIS -systems in construction. In the context of digital transformation and increasing transparency requirements on the part of customers (Fig. 10.2-3), it is unlikely that project implementation in the long term will remain dependent on proprietary solutions that limit flexibility, hinder integration and hamper business development.

And no matter how beneficial it may be for construction companies to work with data silos and opaque data in closed databases - inevitably, the future of the construction industry will involve a transition to open platforms, machine-readable and transparent data structures, and trust-based automation. This transformation will be driven from the top - under pressure from customers, regulators and society, increasingly demanding accountability, sustainability, transparency and economic feasibility.

The end of the era of closed ERP /PMIS: the construction industry needs new approaches

The use of bulky modular ERP/PMIS -systems consisting of tens of millions of lines of code makes any changes in them extremely difficult. At the same time, the transition to a new platform with modules already customized for the company, tens of thousands of articles in resource databases (Fig. 5.1-3) and thousands of ready-made calculations (Fig. 5.1-6) turns into a costly and time-consuming process. The more code and legacy architecture - the higher the level of internal inefficiencies, and each new project will only make things worse. In many companies, data migration and integration of new solutions become multi-year epics accompanied by constant rework and endless search for compromises. The result is often a return to old, familiar platforms, despite their limitations.

As highlighted in the German Black Book report [108] on systemic failures in construction data management, fragmentation of information and lack of a centralized approach to its management is a key cause of inefficiency. Without standardization and integration, data loses its value, becoming an archive rather than a management tool.

A major cause of data quality loss is inadequate planning and control of construction projects, which often leads to significant cost increases. The Black Book's "Focus: The Cost Explosion" section analyzes the key factors contributing to these unintended consequences. These include inadequate needs analysis, lack of feasibility studies and uncoordinated planning leading to additional costs that could have been avoided.

In a mature IT ecosystem of a company, replacing an outdated system is comparable to replacing a load-bearing column in an already constructed building. It is not enough to simply remove the old one and install a new one - it is important to do it in such a way that the building remains stable, the ceilings do not collapse, and all communications continue to work. This is where the difficulty lies: any mistake can have serious consequences for the entire company's system.

Nevertheless, developers of large ERP products for the construction industry continue to use the amount of written code as an argument in favor of their platform. At specialized conferences one can still hear phrases like: "It will take 150 man-years to recreate such a system", despite the fact that most of the functionality of such systems hides databases and quite simple functions for working with tables, packed in a special fixed, user interface. In practice, the code volume of "150 man-years" turns into a burden rather than a competitive advantage. The more code - the higher the cost of support, the more difficult the adaptation to new conditions and the higher the entry threshold for new developers and clients.

Many modular building systems today resemble cumbersome and outdated "Frankenstein constructs" where any careless change can lead to failures. Each new module adds complexity to an already overloaded system, turning it into a labyrinth that only a few specialists can understand, making it even more difficult to maintain and modernize.

The complexity is also realized by the developers themselves, who periodically pause for refactoring - revising the architecture to take into account the emergence of new technologies. However, even if refactoring is done regularly, complexity inevitably grows. Architects of such systems get used to the growing complexity, but for new users and specialists it becomes an insurmountable barrier. As a result, all expertise is concentrated in the hands of a few developers, and the system ceases to be scalable. In the short term, such experts are useful, but in the long term, they become part of the problem.

Organizations will continue to integrate "small" data with their big data counterparts, and it is foolish for anyone to believe that one application – no matter how expensive or robust – can handle everything [109].

– Phil Simon, host of the Conversations About Collaboration podcast

A legitimate question arises: do we really need such cumbersome and closed systems for calculating the cost and timing of work in the form of tables, if other industries have long been able to handle similar tasks with analytical tools with open data and transparent logic?

Currently, closed modular platforms are still in demand in the construction industry, primarily due to the specifics of cost accounting (Fig. 5.1-7). Such systems are often used to run "gray" or opaque schemes, allowing real costs to be hidden from the customer. However, as the industry matures digitally, primarily customers, and moves into the so-called "Uberized Era", intermediaries, namely construction companies with their ERPs, will lose their importance in time and cost calculations. This will change the face of the construction industry forever. Read more in the last part of the book and in the chapter "Construction 5.0: How to make money when you can't hide anymore".

Thousands of legacy legacy solutions accumulated over the last 30 years with thousands of man-years invested in development will start to disappear rapidly. The transition to open, transparent and flexible data management is inevitable. The only question is which companies will be able to adapt to these changes and which will remain hostage to the old model.

A similar situation is observed in the field of CAD (BIM -) tools, whose data today fill the volumetric parameters of design entities in ERP/PMIS -systems. Initially the idea of BIM (developed back in 2002 [110] was based on the concept of a single integrated database, but in practice today work with BIM

requires a whole set of specialized programs and formats. What was supposed to simplify design and construction management has turned into another layer of proprietary solutions that complicate integration and reduce business flexibility.

Next steps: efficient use of project data

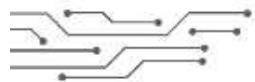
In this part, we have shown how structured data becomes the basis for accurate cost and schedule calculations for construction projects. Automating the QTO, scheduling and estimating processes reduces labor costs and significantly improves the accuracy of the results.

To summarize this part, it is worth highlighting the main practical steps that will help you apply the discussed approaches in your daily tasks. These approaches are universal - they are useful both for the digital transformation of a company and for the daily work of professionals involved in calculations:

- Automate routine calculations
 - Try to find standardized job costings that you can relate to in your work
 - Analyze which methods are used to cost or calculate works or processes at the construction site in your country (Fig. 5.1-7).
 - If you are working with a CAD system - explore the automatic extraction of specifications and QTO data in your CAD (BIM-) software.
 - Use the LLM to write draft code to automate calculations
- Develop your own tools for QTO
 - Create scripts or tables to automate volume counting
 - Standardize categories and groups of elements for a consistent approach to assessment
 - Document the calculation methodology to ensure reproducibility of results in new projects
- Integrate different aspects of the project into your work
 - If you're working with modular systems, try visualizing your processes not only as diagrams or charts, but also at the data level - especially in the form of tables
 - Master the automatic merging of data extracted from CAD databases with calculations - with Python code using grouping, filtering and aggregation
 - Create clear visualizations of QTO groups to present complex information to colleagues and customers

These steps will help build a sustainable calculation system based on automation and data standardization. This approach will improve accuracy and reduce the routine of day-to-day calculation issues.

The following chapters are dedicated to the technical aspects of CAD - (BIM-) products and the reasons why CAD databases are still difficult to integrate into companies' business processes. If you are not interested now in the history of BIM implementation in construction, the evolution of CAD tools and the technicalities of working with these technologies, you can go straight to the seventh part of the book "Data-Driven Decision Making".



MAXIMUM CONVENIENCE WITH THE PRINTED VERSION

You are holding a free digital version of **Data-Driven Construction**. For more convenient work and quick access to the materials, we recommend you to pay attention to the **printed edition**:



■ **Always at hand:** the book in printed format will become a reliable working tool, allowing you to quickly find and use the necessary visualizations and diagrams in any working situation

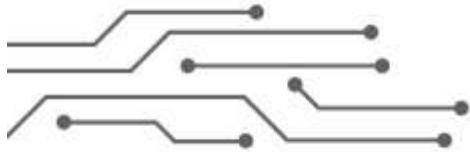
■ **High quality illustrations:** all images and graphics in the printed edition are presented in maximum quality

■ **Quick access to information:** easy navigation, the ability to make notes, bookmarks and work with the book anywhere.



By purchasing the full printed version of the book, you get a convenient tool for comfortable and efficient work with information: the ability to quickly use visual materials in everyday tasks, quickly find the necessary schemes and make notes. In addition, your purchase supports the dissemination of open knowledge.

Order the print version of the book at: datadrivenconstruction.io/books



VI PART

CAD AND BIM: MARKETING, REALITY AND THE FUTURE OF DESIGN DATA IN CONSTRUCTION

The sixth part of the book presents a critical analysis of the evolution of CAD and BIM -technologies and their impact on data management processes in construction. It traces the historical transformation of the BIM concept from the original idea of an integrated database to the current marketing constructs promoted by software vendors. The impact of proprietary formats and closed systems on the efficiency of project data handling and overall performance of the construction industry is evaluated. The problems of compatibility of various CAD-systems and difficulties of their integration with business processes of construction companies are analyzed in detail. Current trends in the transition to simplified open data formats, such as USD, and their potential impact on the industry are discussed. Alternative approaches to extracting information from closed systems are presented, including reverse engineering techniques. Prospects for the application of artificial intelligence and machine learning to automate design and data analysis processes in construction are analyzed. Forecasts for the development of design technologies focused on the real needs of users rather than on the interests of software vendors are formulated.

CHAPTER 6.1.

EMERGENCE OF BIM-CONCEPTS IN THE CONSTRUCTION INDUSTRY

Originally, this sixth part, dedicated to CAD (BIM), was not included in the first version of the book. The topics of proprietary formats, geometric kernels and closed systems are overly technical, overly detailed and seemingly useless for those who just want to understand how to work with data. However, feedback and requests to add clarification to the first version of the book showed that without understanding the complexities of the inner workings of CAD systems, geometry kernels, the variety of formats and incompatible storage schemes for the same data, it is impossible to truly understand why the concepts promoted by vendors often make it difficult to work with information and hinder the transition to open parametrized design. That is why this part has taken its own place in the structure of the book. If CAD (BIM) is not a priority for you, you can go straight to the next part - "PART VII: Data-Driven Decision Making, Analytics, Automation and Machine Learning".

History of the emergence of BIM and open BIM as marketing concepts of CAD-vendors

With the advent of digital data in the 1990s, computer technology was introduced not only in business processes but also in design processes, leading to concepts such as CAD (computer-aided design systems) and later, BIM (building information modeling)

However, like any innovation, they are not the end point of development. Concepts like BIM have become an important milestone in the history of the construction industry, but sooner or later they may give way to better tools and approaches that will better meet the challenges of the future.

Overwhelmed by the influence of CAD vendors and confused by the complexities of its own implementation, the concept of BIM, which appeared in 2002, may well not live to see its thirtieth anniversary, like a rock star that flashed brightly but quickly faded away. The reason is simple: the demands of data scientists are changing faster than CAD vendors can adapt to them.

Faced with a lack of quality data, today's construction industry professionals demand cross-platform interoperability and access to open data from CAD- projects to simplify their analysis and processing. The complexity of CAD data and the confusing processing of CAD data has a negative impact on everyone involved in the construction process: designers, project managers, construction workers on site and, ultimately, the client.

Instead of a full-fledged dataset for operation today, the customer and investor receive containers in CAD- formats that require complex geometric kernels, understanding of data schemas, annually updated API -documentation and specialized CAD software (BIM) to work with the data. At the same time, much of the design data remains unused.

In today's design and construction world, the complexity of accessing CAD data leads to over-engineered project management. Medium and large companies working with CAD data or developing BIM -solutions are either forced to maintain close relationships with CAD vendors solutions to access data via APIs, or bypass CAD vendor restrictions by using expensive SDK converters to reverse-engineer, to get open data [75].

The proprietary data approach is outdated and no longer meets the demands of today's digital environment. The future will divide companies into two types: those who use open data effectively, and those who will leave the market.

The concept of BIM (Building Information Modeling), appeared in the construction industry with the publication of one of the major CAD vendors - Whitepaper BIM [54] in 2002 and, supplemented by the mechanical engineering concept BOM (Bills of Materials), originated from the parametric approach to the creation and processing of project data (Fig. 6.1-1). The parametric approach to the creation and processing of design data was one of the first to be implemented in the Pro-E system for mechanical engineering design (MCAD). This system became a prototype [111] for many modern CAD -solutions, including those used today in the construction industry.

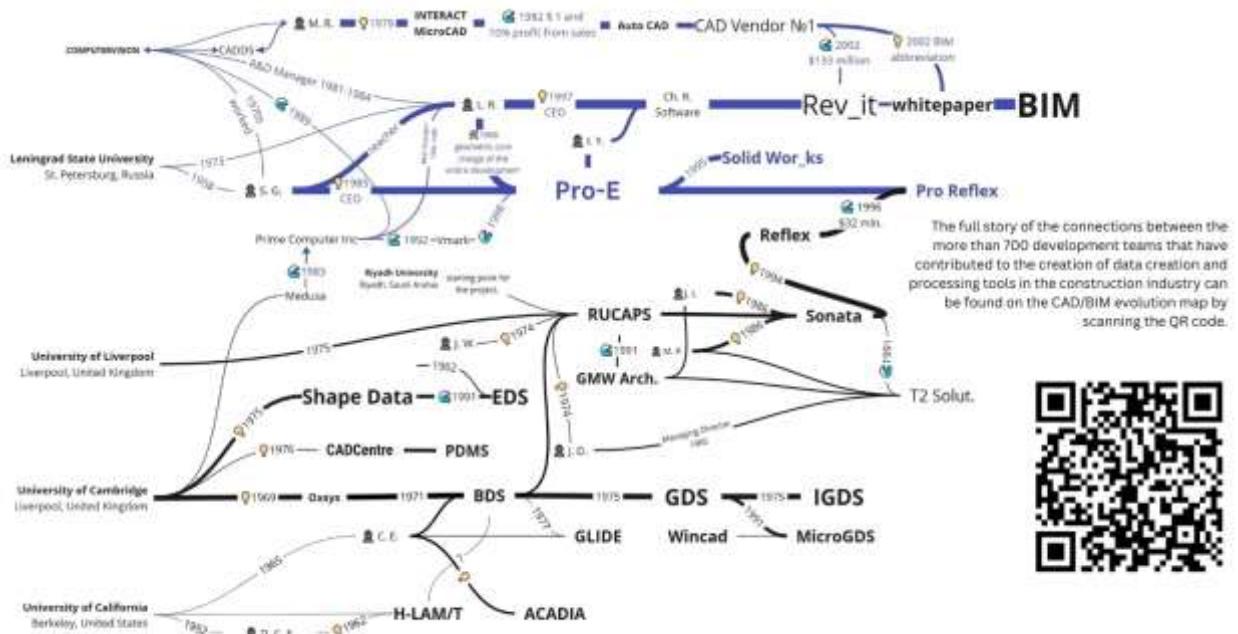


Fig. 6.1-1 Map of the history of the BIM concept and similar concepts.

Journalists and AEC consultants, who promoted CAD tools -vendors until the early 2000s, shifted their attention from 2002 to Whitepaper BIM. It was the BIM Whitepaper 2002-2004 and articles published in 2002, 2003, 2005 and 2007 that played a key role in popularizing the BIM concept in the construction industry [112].

Building Information Modeling is a strategy..... [CAD vendor company name] to apply information technology to the construction industry.

— BIM Whitepaper, 2002 [60]

By the mid-2000s, "researchers" began to link the BIM- concept published by CAD- vendor in 2002 with earlier scholarly works, such as Charles Eastman's BDS, which became the basis for systems such as GLIDE, GBM, BPM, and RUCAPS. In his groundbreaking work Building Description System (1974), Charles Eastman laid the theoretical foundations of modern information modeling. The term "database" appears 43 times in his work (Fig. 6.1-2) - more often than any other, except for the word "building".

Eastman's key idea was that all information about a building - from geometry to the properties of elements and their relationships - should be stored in a single structured database. It is from this database that drawings, specifications, calculations, and code compliance can be automatically generated and analyzed. Eastman explicitly criticized drawings as an outdated and redundant method of communication, pointing to duplication of information, problems with updating, and the need for manual updates when changes are made. Instead, he proposed a single digital model in a database where any change is made once and automatically reflected in all views.

It is noteworthy that in his concept Eastman did not put visualization at the forefront. The central place in his system was information: parameters, relationships, attributes, analysis and automation capabilities. Drawings in his understanding were only one of the forms of displaying data from the database, not the primary source of design information.

In the first Whitepaper on BIM from the leading CAD vendor, the phrase "database" was used as often as in Charles Eastman's BDS - 23 times [60] on seven pages and was one of the most popular words in the document after "Building", "Information", "Modeling" and "Design". However, by 2003, the term "database" appears only twice in similar documents [61], and by the late 2000s, the topic of databases had virtually disappeared from the discussion of design data. As a result, the concept of "a single integrated database for visual and quantitative analysis" was never fully realized.

Thus, the construction industry has gone from Charles Eastman's progressive BDS concept with its emphasis on databases and Samuel Geisberg's ideas about automatically updating design data from databases in the mechanical engineering product Pro-E (the predecessor of popular CAD -solutions used in construction today) to the current marketed BIM, where data management through databases is barely mentioned, despite the fact that this was the concept behind the original theoretical

IV - Integrated Design of a Building by Using Descriptive Systems

... was intended to show that a computer-based integration of all building models results in savings in all-around strengths of buildings and structures. This is the main reason for the emergence of the integrated concept of design and construction management. This means that a common and universal description of a very large number of physical elements, arranged in space and distributed, can be used in actual buildings. Consequently, detailed descriptions of individual parts of the building can be omitted. In addition, spaces as well as entire models can be simplified making the **database** more precise & highly concentrated on more interesting elements.

... spaces, volumes or volumes, and their world allow any change to be restricted only once when such changes cover a large number of details. The numerical description of a building model is based on the **database**, so that, in case of need, it can be easily modified. This is the main advantage of the integrated concept of design and construction management. On the other hand, this one **database** could cause difficulties in the use of different types of software. This is the reason why the **BDS** concept of design and construction management must be developed.

In practice, however, the **BDS** concept of design and construction management is not yet fully implemented. From this simple **database**, the designer would not be able to plan his work in a better way. He would not be able to make a better assessment of his work capacity in a more efficient way than at the present. All necessary products from this one **database** would be automatically generated.

It is similar, because the building management is now in a separate system. Because the building management is now in a separate system, all new requirements for each application must be manually greatly reducing their costs.

With such a **database**, visualization methods would be capable of being automatic and transparent, could be created for storing design information, and could be used on building sites. In addition, a very short construction time savings and efficiency should be available. Total implementation of the **BDS** concept of design and construction management will be the result of the following:

In addition, detailed form drawings were **database** were capable of being automatic and transparent, could be created for storing design information, and could be used on building sites. In addition, a very short construction time savings and efficiency should be available. Total implementation of the **BDS** concept of design and construction management will be the result of the following:

The design, construction and function of the models implemented by the **BDS** concept of design and construction management are now available to the **BDS** administrator, and demands more interaction with a previous form of drawing. It is important to pay attention to the function of the **BDS** system to the **BDS** administrator.

2. **Database**

The goal is to obtain a common **database** capable of describing building models. Of course, the status-oriented basic tools must be specified and the standard types of data. It would also be needed that the design of one architecture should be made.

3. design a **descriptive** program which is fully compatible with and bidirectional about the **database** of **BDS**. The descriptive user provides the interface between **BDS**, the built hardware and the user.
4. given such a large spatially oriented **database**, users must be enabled to quickly sort elements of interest from the total set.
5. existence of so many elements is also an issue. The **BDS** needs to be easily creating large three-dimensional shapes.
6. an rapidly increasing ability is required to efficiently arrange large numbers of potentially visible physical elements.
7. second also is an easy means for writing an arrangement, including both the shape and location of an element, or sets of similar elements.
8. a set of general manipulation routines are required, particularly for the composition of shapes and for interchanging the **database**.
9. a facility for generating high quality drawings of subsets of the **database**, for inspection of existing.
10. a similar but extended feature for producing high quality aesthetic renderings of different parts of the modelled building.
11. interconnectedness of the above operations into a logically organized and easily understood construction language.

One of course technical issues listed above have been addressed and resolved already. Those remaining are referred as **critical**. In the following sections, our treatment of most of the above technical issues are outlined.

IV A - **Database**

The specific hardware configurations for **BDS** are often selected administrators will time-sharing systems as a large central computer. The **BDS** basically consists of a very large **database** and requires no visualization. This **database** must reside in close proximity to the CPU which processes all the data. The **database** must be able to handle large amounts of data and time. Other auxiliary features include real time generation of graphical displays of the **database** and easy switching from one **database**, i.e., building projects. These features, plus the speed and low cost can advantages of **computer**, has encouraged us to follow the main computer line of development.

Figure 23 depicts the **database** and its block structure on disk. In the current structure, once an object's coordinates and colors are defined, its position on the origin are stored. Each level is stored as a separate data element.

"... user must be able to conveniently enter a new part, new expression for a building pattern, as new values for an existing representation are available. Therefore, consideration must be given to creating and maintaining of writing patterns, expressions, and values already stored. Also is certain elements, there will be need to define specific constraints through which the **database** can be used. These constraints are called **constraints**. To this end, a condition should be entered directly, without intervening expressions. There we will **single** language. All the features described are provided by the **database**, as shown in Figure 23."

Within the **database**, instead of the different data elements are through a common hierarchy. Each pattern, expression, and template have a unique name and entry. Within this hierarchy, all expressions based on a common template will have the same name. This allows the **database** to be used in a much more effective manner. The **database** is also designed to be used in a hierarchical manner, where all levels are linked together. These links depict the relation status in Figure 23 and allow operations on these sets of elements to be provided by the **database**, as shown in Figure 23.

The details of the **database** are presented in a separate section.⁹ The **database** has been implemented and is now resulting preliminary testing.

IV B - **Initial Design**

Effort was made at the start of the Building Description System project to find ways to ensure the **database** would be the core of organization of elements. Particularly needed is the capability to access all elements overlapping a spatial area of interest. This significantly reduces the need for data parallel rendering in very enlarged spatial areas. The **database** is also designed to be used in a much more effective manner. The **database** is also designed to be used in a hierarchical manner, where all levels are linked together. These links depict the relation status in Figure 23 and allow operations on these sets of elements to be provided by the **database**.

If current architectural conditions continue, it is imperative that a designer be able to enter elements of unique shape. One can anticipate a continuing need for refining such elements made of concrete, plastic, wood, and other locally distributed materials. In design, there are only currently rigidized methods available for entering such shapes in a unique **database**.

Fig. 6.1-2 In the BDS concept, described by Charles Eastman in 1974, the phrase "Database" (highlighted in yellow) was used 43 times.

BDS and similar concepts until the 2000s were developed as a digital database of buildings rather than as a visualization tool. BIM in 2002 became a design tool where the database took a back seat. What have we lost in the transition from BDS and similar concepts in the 1990s to BIM by the mid 2010s:

- Open databases: BDS and other similar concepts emphasized analytics, BIM emphasized design.
- Flexibility to work with data: BDS emphasized data analytics, BIM emphasized processes that must be based on obscure data.
- Transparency: the BDS was intended to be an open integrated database, while CAD vendors in BIM have made their databases completely closed and have fought unsuccessfully for 20 years against reverse engineering tools that open proprietary formats.

Over the past 30 years, designers have never had access to an "integrated database" and after twenty years of marketing euphoria around BIM -tools, the construction industry is beginning to realize the consequences of this fad.

The reality of BIM: instead of integrated databases - closed modular systems

Instead of focusing on data, structuring it and integrating it into unified processes, users of CAD - (BIM-) systems are forced to work with a fragmented set of proprietary solutions, each dictating its own rules of the game:

- **The unified database**, discussed in the first BIM Whitepaper, **has remained a myth**. Despite loud claims, access to data is still limited and distributed among closed systems.
- **BIM -models have become a closed ecosystem** rather than a tool. Instead of transparent information exchange, users are forced to pay for subscriptions and use proprietary APIs.

- **Data belongs to vendors, not users.** Project information is locked in proprietary formats or cloud services rather than available in open and independent formats.

Design engineers and project managers often do not have access to the database of CAD -systems, nor to the format in which their own project data is stored. This makes it impossible to quickly verify information or formulate requirements for data structure and quality (Fig. 6.1-3). Access to such data requires a whole set of specialized programs linked through APIs and plug-ins, which leads to excessive bureaucratization of processes in the construction industry. Meanwhile, these data are simultaneously used by dozens of information systems and hundreds of specialists.

*We need to be able to manage all this data [CAD (BIM)] store it digitally and sell lifecycle and process management software, because **for every engineer [designer]** who creates something [in a CAD program], **there are ten people** who work with that data" [41].*

— CEO of CAD – the vendor that created the BIM concept, 2005.

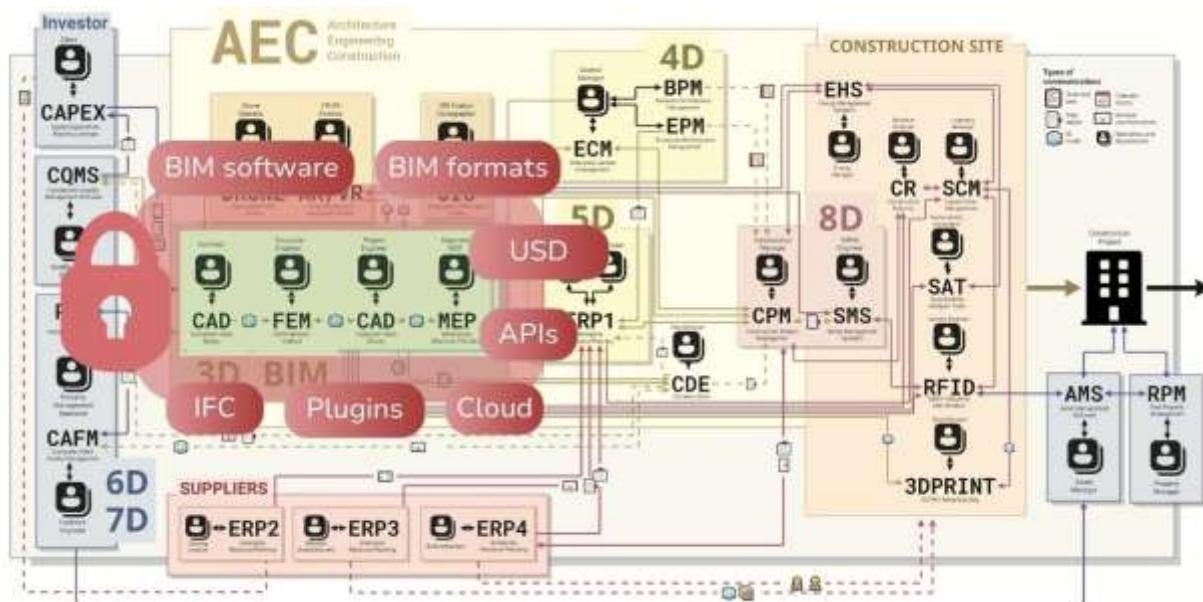


Fig. 6.1-3 CAD- (BIM-) databases remain one of the last closed systems for IT departments and data managers in the construction business ecosystem.

When it becomes obvious that BIM is more a means of commercializing databases rather than a full-fledged database management tool, a logical question arises: how to regain control over the data? The answer is to use open data structures where the user, not the software vendor, becomes the owner of the information.

Users and solution developers in the construction industry, like their counterparts in other industries, will inevitably move away from the vague software vendor terminology that has dominated the last 30 years, focusing on the key aspects of digitalization - "data" and "processes."

Back in the late 1980s, the key area of digital development in construction was envisioned as a matter of data access and project information management. Over time, however, the focus has shifted. Instead of developing transparent and accessible approaches to working with data, the IFC format and the open BIM concept were actively promoted as attempts to divert the attention of specialists from the topics of project database management.

The emergence of the open format IFC in the construction industry

The so-called open format IFC (Industry Foundation Classes) is positioned as a standard to ensure interoperability between different CAD (BIM-) systems. Its development was carried out within the framework of organizations that were created and controlled by major CAD vendors. Based on the IFC format, two CAD-companies in 2012 developed the marketing concept of OPEN BIM[63].

IFC (Industry Foundation Classes) is an open standard for data exchange in the construction industry, designed to ensure interoperability between different CAD - (BIM-) systems.

Open BIM - concept involves working with information from CAD databases and exchanging information between systems through an open format for exchanging CAD data - IFC.

The Open BIM Program is a marketing campaign initiated by... [1 CAD vendor],... [2 CAD vendor] and other companies to encourage and facilitate global coordinated promotion of the OPEN BIM concept throughout the AEC industry, with consistent communication and common branding available to program participants.

– From CAD vendor website, OPEN BIM Program, 2012 [113]

IFC was adapted by the Technical University of Munich from the mechanical engineering format STEP in the late 1980s, and later registered by a major design company and a major CAD-vendor to form the IAI (Industry Alliance for Interoperability) in 1994 [114] (Fig. 6.1-4). The IFC format was developed to provide interoperability between different CAD-systems and was based on the principles laid down in the mechanical engineering format STEP, which, in turn, emerged from the format IGES, created back in 1979 by a group of CAD users and vendors with the support of NIST (The National Institute of Standards and Technology) and the U.S. Department of Defense [115].

However, the complex structure of IFC, its close dependence on the geometric core, as well as differences in the implementation of the format by different software solutions have led to many problems in its practical application. Similar difficulties - loss of detail, limitation of accuracy and necessity to

use intermediate formats - were previously encountered by mechanical engineering specialists when working with IGES and STEP formats from which IFC was derived.

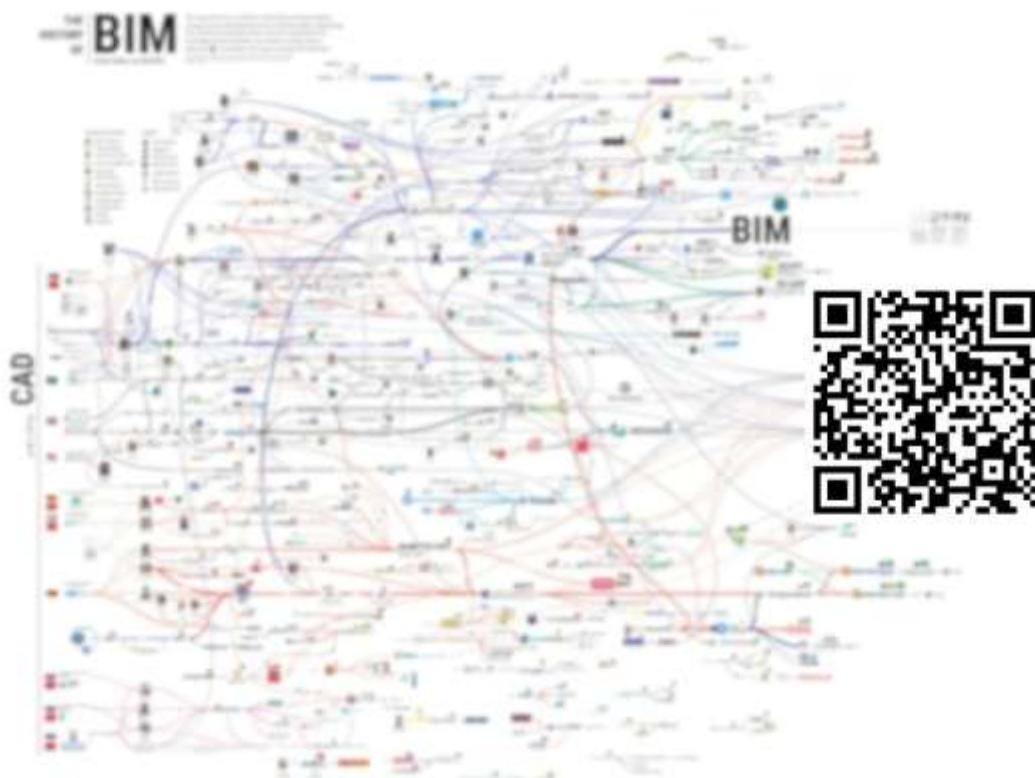


Fig. 6.1-4 Map of relationships between development teams and CAD products (BIM) [116].

In 2000, the same CAD -vendor that registered the IFC format and created the IAI (later bS) organization, publishes the Whitepaper "Integrated Design and Manufacturing: Benefits and Rationale" [65]. The paper stressed the importance of maintaining full data granularity when exchanging between programs within the same system, without using neutral formats such as IGES, STEP [identical to IFC]. Instead, it was proposed that applications should have direct access to the underlying CAD database to prevent loss of accuracy of information.

In 2002 the same CAD vendor buys parametric BOM product (Fig. 3.1-18, more details in of the third part) and on its basis forms the BIM concept. As a result, only closed CAD formats or the IFC format (STEP) are used in the exchange of construction project data, the limitations of which were written about by the CAD vendor himself in 2000, who brought this format to the construction industry.

A detailed history of the interaction of more than 700 development teams involved in building data creation and processing tools is presented in the map "The Evolution of CAD (BIM)" [116].

The open form IFC consists of a geometric description of the design elements and a description of the meta-information. Various methods are used to represent geometry in IFC format, such as CSG and

Swept Solids: however, the parametric representation BREP has become the leading standard for transferring element geometry in IFC format, as this format is supported when exporting from CAD-(BIM-) programs and allows for potential editing of elements when importing IFC back into CAD programs.

IFC format problem depending on geometric core

In most cases, when the geometry in IFC is defined parametrically (BREP), it becomes impossible to visualize or retrieve geometric properties, such as volume or area of project entities, with only an IFC file, because to work with and visualize the geometry in this case, a geometry kernel (Fig. 6.1-5) is required, which is initially missing.

Geometry kernel is a software component that provides basic algorithms for creating, editing and analyzing geometric objects in CAD (CAD), BIM and other engineering applications. It is responsible for building 2D and 3D -geometry, as well as for operations on it, such as: Boolean operations, smoothing, intersections, transformations and visualization.

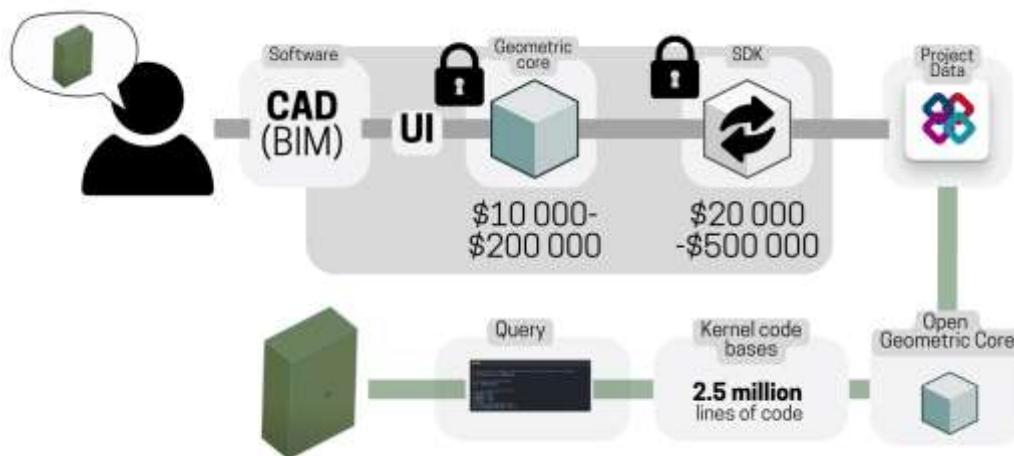


Fig. 6.1-5 Creating geometry through CAD- programs today goes through proprietary geometry kernels and SDKs, which are often not owned by CAD vendors.

Every CAD program and any program working with parametric or IFC formats has its own or purchased geometric kernel. And if with primitive elements in IFC -BREP format there can be no problems and in programs with different geometrical kernels these elements can be displayed similarly, but besides problems with different engines of geometrical kernels, there are enough elements which have their own peculiarities for correct displaying. This problem is discussed in detail in the international study "A reference study of IFC software support" published 2019 [117].

The same standardized datasets produce inconsistent results, with few common patterns found, and serious problems are found in supporting the standard [IFC], probably due to the very high complexity of the standard data model. The standards themselves are partly to blame here, as they often leave some details undefined, with high degrees of freedom and various possible interpretations. They allow high complexity in the organization and storage of objects, which is not conducive to effective universal understanding, unique implementations, and consistent data modeling [117].

— Reference study of IFC software support, 2021

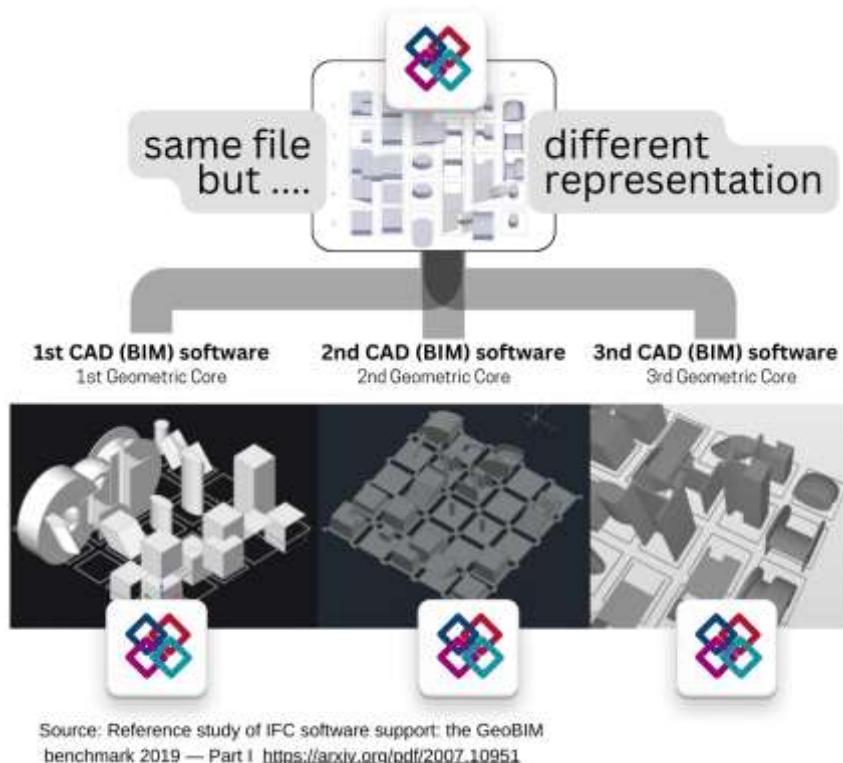


Fig. 6.1-6 Different geometric kernels give different representations of the same geometry described parametrically (based on [117]).

Correct understanding of "certain provisions" is available to paid members of special organizations that are engaged in IFC development. As a consequence, whoever wants to get access to important knowledge about certain features of IFC will try to cooperate with large CAD-vendors, or to reach a qualitative consideration of the features by his own research

You stumble upon a question about importing and exporting data via the IFC format and ask your fellow vendors: "Why is it so in the IFC file the information about parametric transfer of premises? The open specification does not say anything about it". Answer from "more knowledgeable" European vendors: "Yes, it is not said, but it is allowed".

— From the interview of CAD developer, 2021 [118]

IFC describes the geometry through parametric primitives, but does not contain a built-in kernel - its role is performed by the CAD program, which compiles the geometry through the geometry kernel. The geometry kernel performs mathematical calculations and defines intersections, and IFC only provides data for its interpretation. If the IFC contains incorrect faces, different programs with different geometry kernels can either ignore them or produce errors, depending on the kernel.

As a result, to work with the IFC format it is necessary to answer the main question, to which it is difficult to find an unambiguous answer - what tool, with what geometric kernel should be used to get the quality of data that the project originally had in the CAD program from which the IFC was obtained?

Data quality issues and the complexity of the IFC format do not allow direct use of project data for process automation, analysis and data processing, which often leads developers to the inevitable need to use closed CAD -solutions with "quality" access to data[63], which was written about by the vendor himself, who registered IFC in 1994 [65].

All peculiarities of display and generation of IFC parameters in geometry kernel can be realized only by large teams of developers who have experience in working with geometry kernels. Therefore, the current practice of IFC format peculiarities and complexity is beneficial primarily to CAD- vendors and has much in common with the strategy of large software vendors "adopt, extend, destroy", when the growing complexity of the standard actually creates barriers for small market players [94].

The strategy of large vendors in such a strategy may be to adapt open standards, add proprietary extensions and features to create user dependency on their products to then drive out competitors.

The IFC format, intended to be a universal bridge between different CAD- (BIM-) systems, in reality serves as an indicator of compatibility problems between the geometric cores of different CAD platforms, similar to the STEP format from which it originally emerged.

As a result, today a full and high-quality implementation of the IFC ontology is within the reach of large CAD vendors, who can invest significant resources to support all entities and their mapping to their own internal geometry core, which does not exist for IFC as a standard. Large vendors also have the ability to negotiate among themselves technical details of features that may not be available to even

the most active participant in IFC format development organizations.

For small independent teams and open-source projects, striving to support the development of interoperable formats, the lack of an in-house geometry kernel becomes a serious problem. Without it, it is virtually impossible to take into account all the various subtleties and nuances associated with cross-platform data exchange.

With the development of the IFC parametric format and the open BIM concept, discussions have intensified in the construction industry about the role of ontology and semantics in data and process management.

Appearance in the construction of the topic of semantics and ontology

Thanks to the ideas of the semantic web in the late 1990s and the efforts of organizations involved in the development of the IFC format, semantics and ontologies have become some of the key elements of the standardization being discussed in the construction industry by the mid-2020s.

Semantic technologies are unification, standardization and modification of large heterogeneous data sets and implementation of complex search.

To store semantic data we use OWL ontology language (Web Ontology Language), represented in the form of RDF graphs -triplets (Resource Description Framework) (Fig. 6.1-7). OWL refers to graph data models, the types of which we discussed in more detail in the chapter "Data models: data relations and relationships between elements".

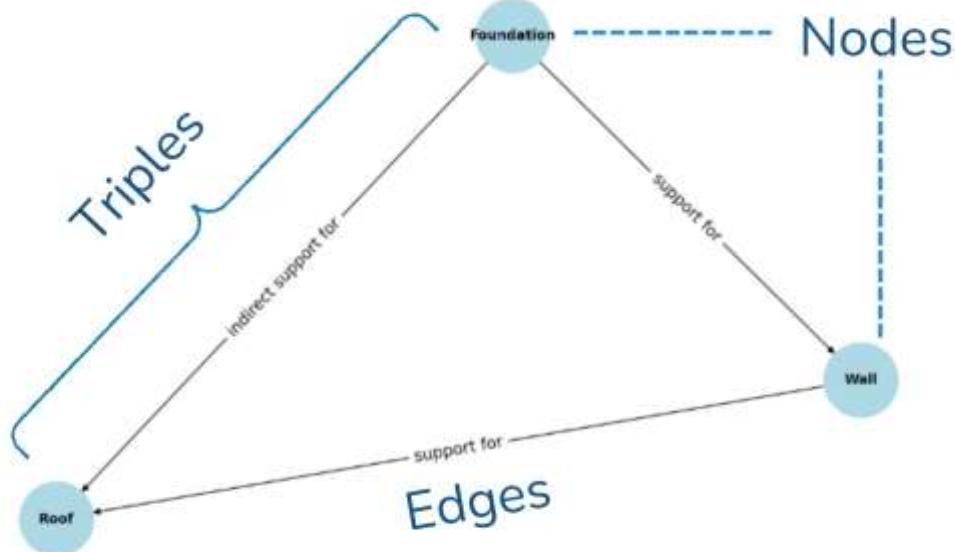


Fig. 6.1-7 RDF data model: Nodes, Edges and Triples illustrating the relationships between building blocks.

Theoretically, the logical inference of reasoners (programs for automatic logical inference) allows new statements to be derived from ontologies. For example, if the building ontology records that "a foundation is a support for a wall" and "a wall is a support for a roof" (Fig. 6.1-7), the reasoner is able to automatically infer that "a foundation is a support for a roof".

Such a mechanism is useful for optimizing data analysis because it avoids explicitly prescribing all dependencies. However, it does not create new knowledge, but only identifies and structures already known facts.

Semantics does not create new meaning or knowledge per se and is not superior to other data storage and processing technologies in this aspect. Representing data from relational databases as triplets does not make them more meaningful. Replacing tables with graph structures may be useful for unifying data models, easy searching and secure editing, but it does not make the data "smarter" - the computer does not begin to understand its content better.

Logical relationships in data can be organized without complex semantic technologies (Fig. 6.1-8). Traditional relational databases (SQL) as well as CSV or XLSX formats allow building similar dependencies. For example, in a columnar database, you can add a "roof support" field and automatically associate the roof with the foundation when creating a wall. This approach is implemented without the use of RDF, OWL, graphs or reasoners, remaining a simple and efficient solution for storing and analyzing data.

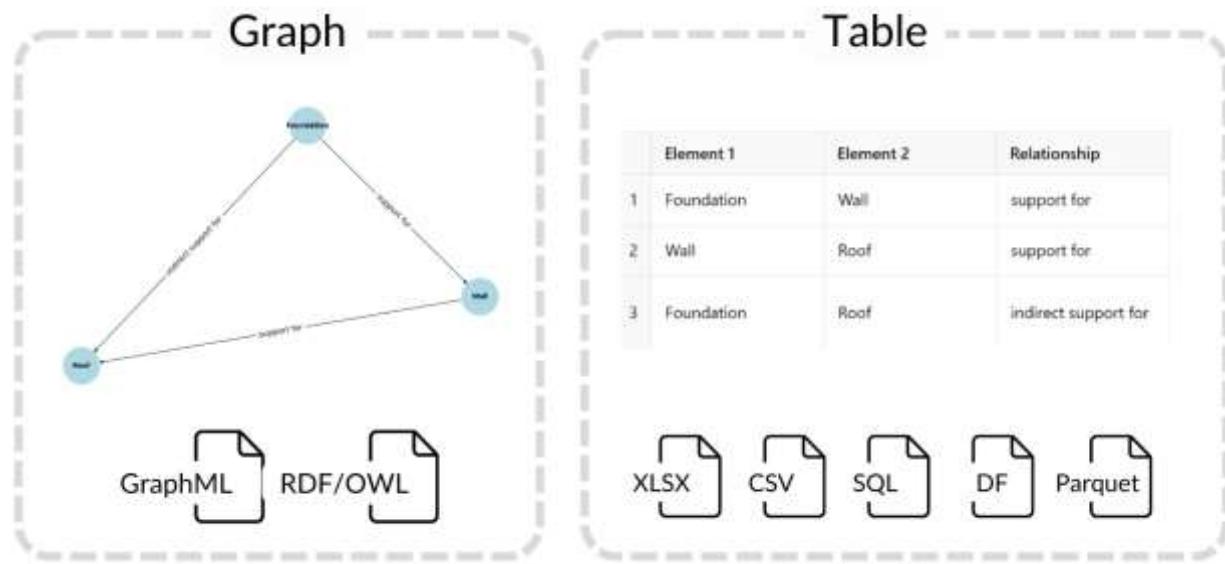


Fig. 6.1-8 Comparison of graph and table data models of representing the same logical relationships.

The decision of a number of large construction companies and the IFC format development organization [94] to follow the semantic web concept, which seemed promising in the late 1990s, has had a significant impact on the development of standards in the construction industry.

However, the paradox is that the very concept of the semantic web, originally intended for the Internet, has not been widely adopted even in its native environment. Despite the development of RDF and OWL, the full-fledged semantic web has not appeared in its original conception, and its creation is already unlikely.

Why semantic technologies fail to live up to expectations in the construction industry

Other industries have faced the limitations of technologies for using semantics. In the gaming industry, attempts to describe game objects and their interactions through ontologies have proven ineffective due to the high dynamics of change. As a result, simpler data formats such as XML and JSON, together with algorithmic solutions, have been preferred. The situation was similar in real estate: due to regional differences in terminology and frequent market changes, the use of ontologies proved to be overly complex, while simple databases and standards such as RETS [119] were better able to handle the data exchange tasks.

Technical difficulties, such as the complexity of markup, high labor-intensive support and low developer motivation, slowed down the adoption of semantic web and in other sectors of the economy. RDF (Resource Description Framework) did not become a mass standard, and ontologies proved to be too complex and economically unjustified.

As a result, the ambitious idea of creating a global semantic web failed to materialize. Although some elements of the technology, such as ontologies and SPARQL, have found their way into enterprise solutions, the original goal of creating a single comprehensive data structure has not been achieved.

The concept of an Internet in which computers are able to make sense of content has proven to be technically challenging and commercially unprofitable. This is why the companies that supported the idea eventually reduced its use to individual useful tools, leaving RDF and OWL for highly specialized corporate needs rather than for the Internet as a whole. An analysis of Google Trends (Fig. 6.1-9) over the last 20 years suggests that there may be no more prospects for the semantic web.

There is no need to multiply entities unnecessarily. If there are several logically consistent explanations of a phenomenon that explain it equally well, one should, all other things being equal, prefer the simplest of them.

— Occam's Razor

A logical question arises here: why use triplets, risoners and SPARQL in construction at all, if you can process data using popular structured queries (SQL, Pandas, Apache®)? In enterprise applications, SQL is the standard for working with databases. SPARQL, on the contrary, requires complex graph structures and specialized software and according to trends in Google does not attract the interest of developers.

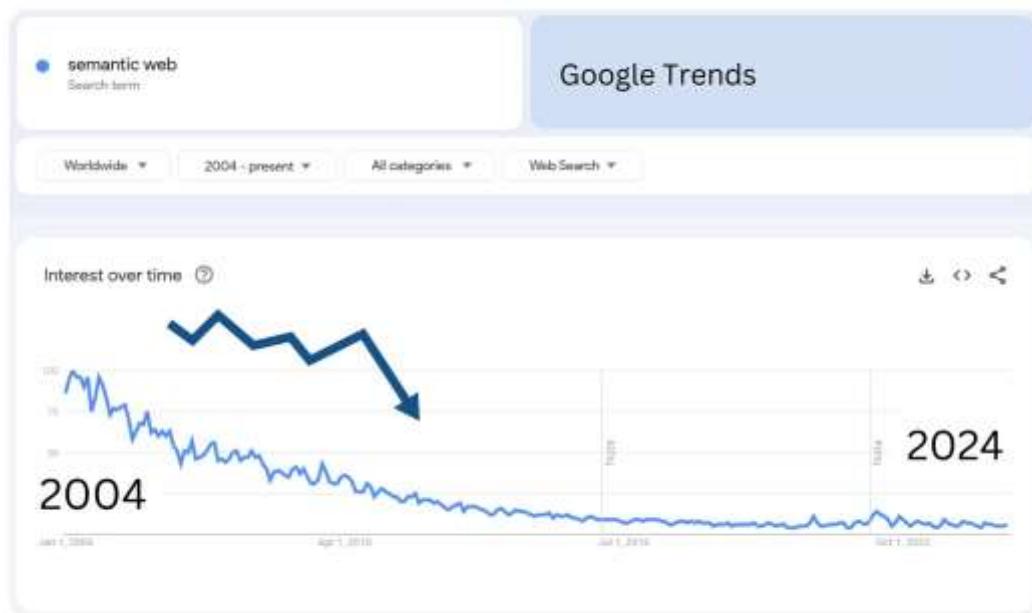


Fig. 6.1-9 Interest in "semantic internet" queries according to Google statistics.

Graph databases and classification trees can be useful in some cases, but their application is not always justified for most everyday tasks. As a result, creation of knowledge graphs and use of semantic web technologies makes sense only when it is necessary to unify data from different sources or to

realize complex logical conclusions.

Moving from tables to graph data models improves search and unifies the flow of information, but does not make the data more meaningful to machines. The question is not whether semantic technologies should be used, but where they really make a difference. Before implementing ontology, semantics and graph databases in your company, find out which companies are already successfully using these technologies and where they have failed.

Despite ambitious expectations, semantic technologies never became a universal solution for structuring data in the construction industry. In practice, these technologies have not led to a universal solution, but have only added new complexities, and these efforts echo the unrealized ambitions of the Semantic Internet concept, where expectations far exceeded reality.



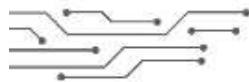
Fig. 6.1-10 Geometry and information in construction processes: from complex CAD and BIM-systems to simplified data for analytics.

While in IT the failures of the semantic web have been offset by the emergence of new technologies (big data, IoT, machine learning, AR/VR), the construction industry has no such occasions.

In addition to the challenges of using concepts to communicate data relationships between project elements, a fundamental problem remains - the very availability of that data. The construction industry is still dominated by closed systems, making it difficult to work with data, share information and improve process efficiency.

It is the closed nature of data that is becoming one of the key barriers that hinders the development of

digital solutions in construction. Unlike the IT industry, where open and unified data formats have become the standard, in the CAD sector (BIM) each software uses its own format, creating closed eco-systems and artificially limiting users.



CHAPTER 6.2.

CLOSED PROJECT FORMATS AND INTEROPERABILITY ISSUES

Closed data and falling productivity: the dead end of the CAD industry (BIM)

The proprietary nature of CAD -systems has led to the fact that each program has its own unique data format, which is either closed and inaccessible from the outside - RVT, PLN, DWG, NDW, NWD, SKP, or is available in semi-structured form through a rather complex conversion process - JSON, XML (CPIXML), IFC, STEP and ifcXML, IfcJSON, BIMJSON, IfcSQL, CSV etc..

Different data formats that can store the same data about the same projects not only differ in structure, but also include different versions of internal markup that developers need to consider to ensure application compatibility. For example, a CAD format from 2025 will open in a CAD program from 2026, but the same project will never open in all versions of the CAD program that may have been available before 2025.

By not providing direct access to databases, a software vendor in the construction industry often creates its own unique format and tools that a professional (design engineer or data manager) must use to access, import and export data.

As a consequence, vendors of basic CAD (BIM) and related solutions (e.g. ERP/PMIS) are constantly raising prices for using the products, and ordinary users are forced to pay a "commission" at each stage of data transfer by formats [63]: for connecting, importing, exporting and working with data that users have created themselves.

The cost of accessing data in cloud storage from popular CAD - (BIM-) products will reach \$1 per transaction in 2025 [120], and subscriptions to construction ERP -products for medium-sized companies reach five- and six-Fig. sums per year [121].

The essence of modern construction software is that it is not automation or increased efficiency, but the ability of engineers to understand a particular highly specialized software that affects the quality and cost of construction project data processing, as well as the profits and long-term survival of companies undertaking construction projects.

The lack of access to databases CAD -systems that are used in dozens of other systems and hundreds of processes [63], and the resulting lack of quality communication between individual professionals has led the construction industry to the status of one of the most inefficient sectors of the economy in terms of productivity [44].

Over the last 20 years of CAD- (BIM-) design applications, the emergence of new systems (ERP), new construction technologies and materials, the productivity of the entire construction industry has dropped by 20% (Fig. 2.2-1), while the overall productivity of all sectors of the economy that do not

have major problems accessing databases and marketing-like BIM concepts has increased by 70% (96% in the manufacturing industry) [122].

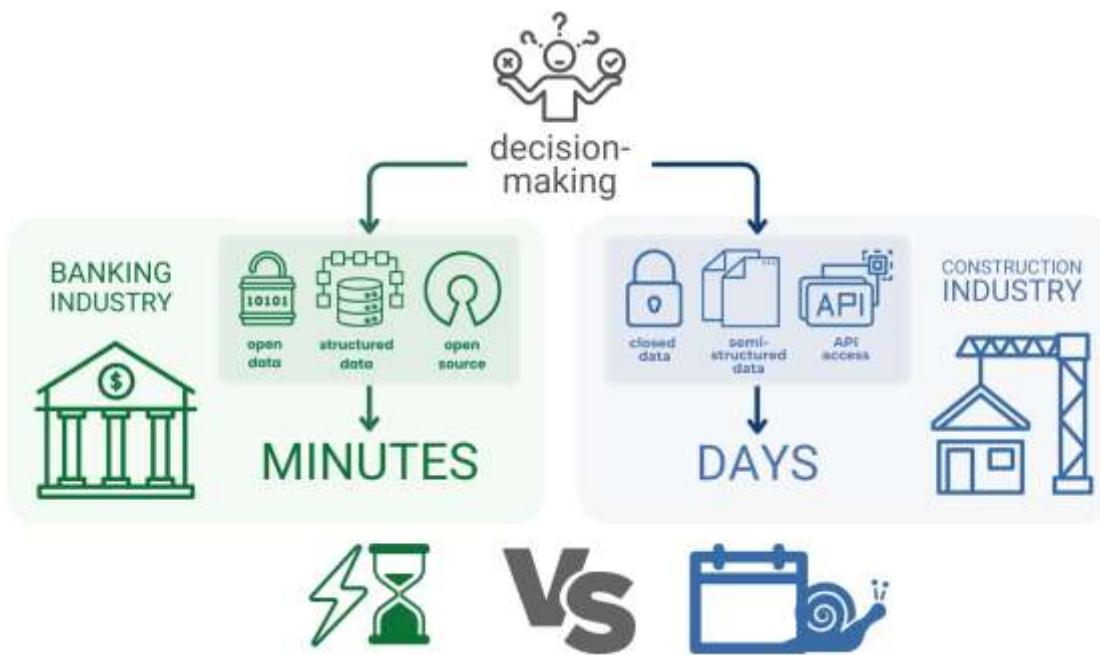


Fig. 6.2-1 Because of the isolation and complexity of project data on which dozens of departments and hundreds of processes depend in the construction industry, the speed of decision making is several times slower than in other industries.

However, there are also isolated examples of alternative approaches to creating interoperability between CAD solutions. Europe's largest construction company with the SCOPE project [123], started back in 2018, demonstrates how it is possible to go beyond the classical logic of CAD- (BIM-) systems. Instead of trying to subjugate IFC or relying on proprietary geometry kernels, SCOPE developers use APIs and SDKs reverse engineering to extract data from various CAD programs, convert them into neutral formats such as OBJ or CPIXML based on the only Open Source geometry kernel OCCT, and further apply them to hundreds of business processes of construction and design companies. However, despite the progressiveness of the idea, such projects face the limitations and complexity of free geometry kernels and they still remain part of closed ecosystems of one company that reproduce the logic of monovendor solutions.

Due to the limitations of closed systems and differences in data formats, as well as the lack of effective tools for their unification, companies that have to work with CAD formats are faced with the accumulation of significant amounts of data with varying degrees of structure and closedness. These data are not used properly and disappear in archives, where they remain forever forgotten and unused.

Data obtained through significant effort in the design phase becomes inaccessible for further use due to its complexity and closed nature.

As a result, over the past 30 years, developers in the construction industry have been forced to face the same problem over and over again: each new closed format or proprietary solution generates the need to integrate with existing open and closed CAD systems. These constant attempts to provide interoperability between different CAD and BIM solutions only serve to complicate the data ecosystem, instead of contributing to its simplification and standardization.

The myth of interoperability between CAD systems

If in the mid-1990s the key direction of interoperability development in the CAD environment was the breaking of the proprietary DWG format - culminating in the victory of the Open DWG alliance [75] and the actual opening of the most popular drawing format for the entire construction industry - then by the mid-2020s the focus has shifted. A new trend is gaining momentum in the construction industry: numerous development teams are focused on creating so-called "bridges" between closed CAD systems (closed BIM), IFC format and open solutions (open BIM). Most of these initiatives are based on the use of the IFC format and the OCCT geometry kernel, providing a technical bridge between disparate platforms. This approach is seen as a promising direction that can significantly improve data exchange and interoperability of software tools.

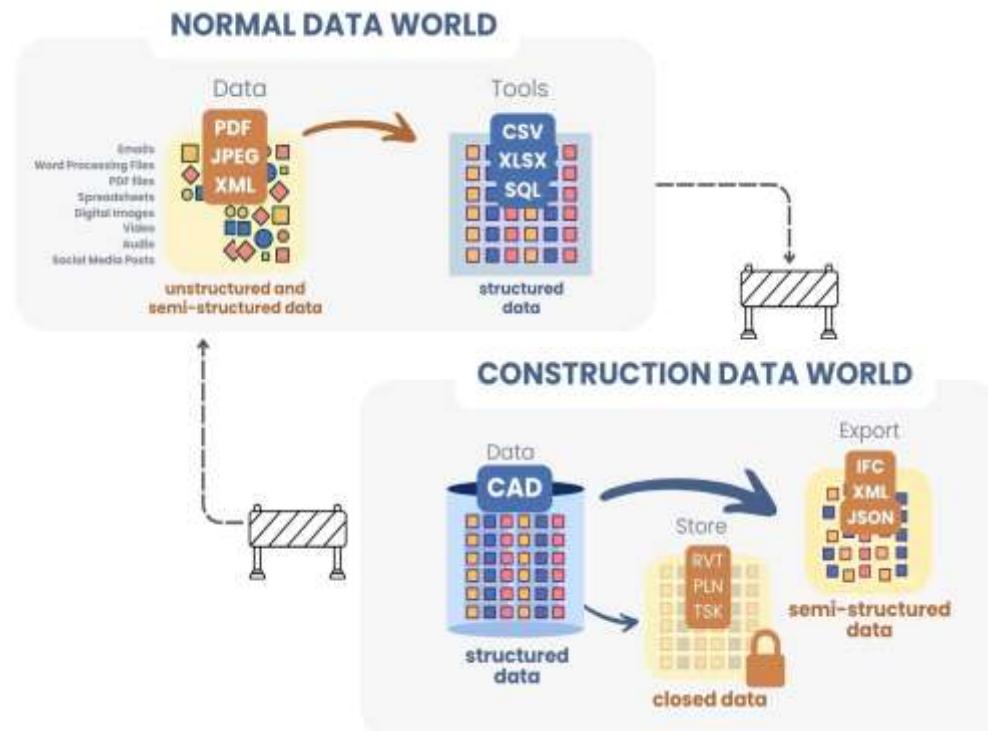


Fig. 6.2-2 While other industries work with open data, the construction industry has to work with closed or loosely structured CAD formats (BIM).

This approach has historical parallels. In the 2000s, developers, trying to overcome the dominance of the largest vendor of graphic editors (2D world), tried to create a seamless integration between its

proprietary solution and free Open Source - an alternative to GIMP (Fig. 6.2-3). Then, as today in construction, it was about trying to connect closed and open systems while preserving complex parameters, layers and internal logic of software operation.

However, users were actually looking for simple solutions - flat, open data without excessive complexity of layers and program parameters (analogous to the geometric core in CAD). Users sought simple and open data formats, free from excessive logic. JPEG, PNG and GIF became such formats in graphics. Today they are used in social networks, websites, applications - they are easy to process and interpret, regardless of the platform or software vendor.

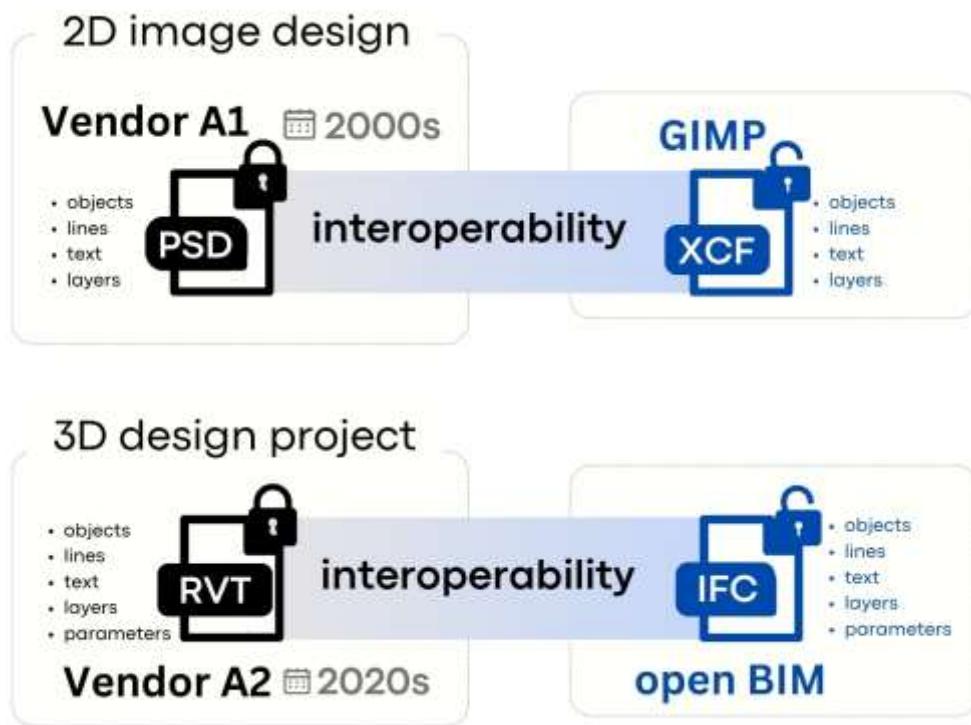


Fig. 6.2-3 The interoperability of data formats in construction is similar to the path from attempts to merge a popular vendor's proprietary product and Open Source GIMP in the 2000s.

As a result, almost no one in the imaging industry today uses closed formats like PSD or open XCF for applications, social networks like Facebook and Instagram, or as content on websites. Instead, most tasks utilize flat and open JPEG, PNG and GIF formats for ease of use and broad compatibility. Open formats such as JPEG and PNG have become the standard for image sharing due to their versatility and broad support, making them easy to use on a variety of platforms. A similar transition can be seen in other exchange formats, such as video and audio, where universal formats like MPEG and MP3 stand out for their compression efficiency and broad compatibility. Such a move towards standardization has simplified the exchange and playback of content and information, making them accessible to all users across platforms (Fig. 6.2-4).

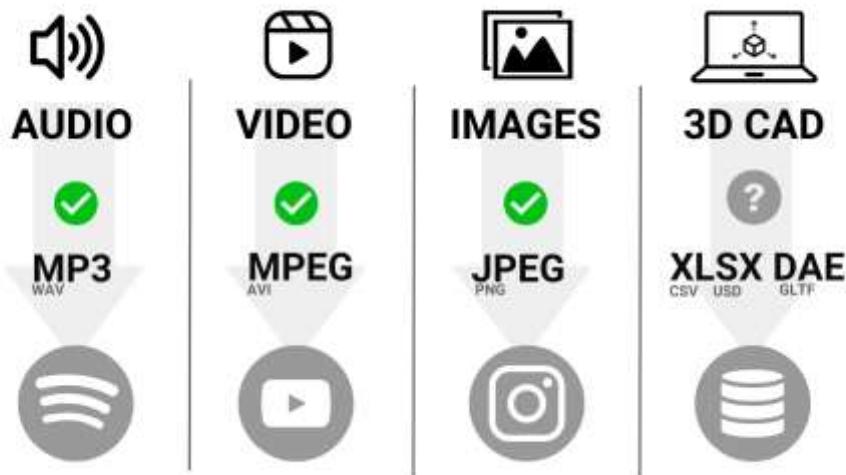


Fig. 6.2-4 Simplified formats without complex editing features, have become popular for sharing and using data.

Similar processes occur in 3D modeling. Simple and open formats like USD, OBJ, glTF, DAE, DXF, SQL and XLSX are increasingly used in projects for data exchange outside the CAD environment (BIM). These formats store all the necessary information, including geometry and metadata, without the need to operate a complex BREP structure, geometry kernels or vendor-specific internal classifiers. Proprietary formats such as NWC, SVF, SVF2, CPIXML and CP2 provided by leading software vendors also perform similar functions, but remain closed, unlike open standards.

It is noteworthy (and worth recalling again, as already mentioned in the previous chapter) that this idea - the rejection of intermediate neutral and parametric formats like IGES, STEP and IFC - was supported back in 2000 by the major CAD vendor that created the BIM Whitepaper and registered the IFC format in 1994. In the 2000 Whitepaper "Integrated Design and Manufacturing" [65] the CAD vendor emphasizes the importance of native access to the CAD database within the software environment, without the need to use intermediate translators and parametric formats, in order to maintain the completeness and accuracy of the information.

The construction industry has yet to agree either on tools to access CAD databases or their forced reverse engineering, or on the adoption of a common simplified data format for use outside CAD platforms (BIM). For example, many large companies in Central Europe and German-speaking regions operating in the construction sector use the CPIXML format in their ERP -systems [121]. This proprietary format, which is a kind of XML, combines CAD (BIM) project data, including geometric and metadata, into a single organized simplified structure. Large construction companies are also creating new formats and systems of their own, as in the SCOPE project, which we discussed in the previous chapter.

The closed logic of parametric CAD formats or complex parametric files IFC (STEP) are redundant in

most business processes. Users are looking for simplified and flat formats such as USD, CPIXML, XML &OBJ, DXF, glTF, SQLite, DAE &XLSX, which contain all the necessary element information, but are unencumbered by redundant BREP geometry construction logic, dependency on geometry kernels and internal classifications of specific CAD and BIM -products (Fig. 6.2-5).

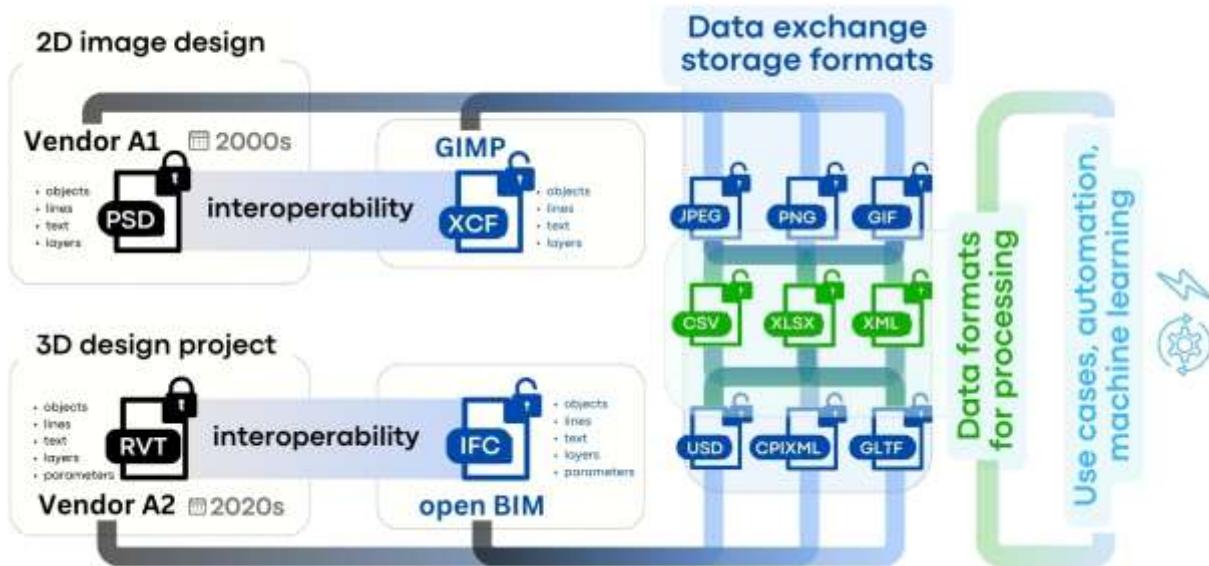


Fig. 6.2-5 For most use cases, users choose the simplest possible formats that are independent of vendor programs.

The advent of flat image formats such as JPEG, PNG and GIF, free from the redundant logic of vendors' internal engines, has fostered the development of thousands of interoperable solutions for processing and utilizing graphics. This has led to a variety of applications, from retouching and filtering tools to social media platforms such as Instagram, Snapchat and Canva, where this simplified data can be utilized without being tied to a specific software developer.

Standardization and simplification of design CAD-formats will stimulate the emergence of many new convenient and independent tools for working with construction projects.

Moving away from complex vendor application logic tied to closed geometry kernels and moving to universal open formats based on libraries of simplified elements creates the prerequisites for more flexible, transparent and efficient data handling. This also opens up access to information for all participants in the construction process - from designers to customers and maintenance services.

Nevertheless, it is highly likely that in the coming years CAD vendors will attempt to shift the debate about interoperability and access to CAD databases again. It will already be about "new" concepts - such as granular data, intelligent graphs, "federated models," digital twins in cloud repositories - as well as the creation of industry alliances and standards that continue the path of BIM and open BIM. Despite the attractive terminology, such initiatives may once again become tools to retain users.

within proprietary ecosystems. One example is the active promotion of the USD (Universal Scene Description) format as the "new standard" for cross-platform CAD (BIM) collaboration from 2023.

Transition to USD and granular data

The emergence of the AOUSD alliance [124] in 2023 marks an important turn in the construction industry. We are witnessing the beginning of a new reality shaped by CAD vendors in dealing with construction data through several significant changes. The first major change concerns the perception of CAD -data. Professionals involved in the early stages of conceptual design are increasingly realizing that creating a design in a CAD environment is only a starting point. The data generated during the design process eventually becomes the basis for analyzing, operating, and managing facilities. This means that they must be accessible and usable in systems beyond traditional CAD tools.

In parallel, a revolution in the approach of leading developers is taking place. The industry's leading CAD- vendor, which created the BIM concept and the IFC format, is making an unexpected turn in its strategy. From 2023, the company is moving away from the traditional storage of data in separate files, focusing on working with granular (normalized and structured) data with a transition to a data-centric approach [125].

Vendors are following the historical trends of other industries: most users don't need closed CAD formats (similar to PSD) or complex parametric IFC files (similar to GIMP with layer logic). They need simple object images that can be used in CAFM (construction In-stagram), ERP (Facebook) and thousands of other processes filled with Excel spreadsheets and PDF documents.

Current trends in the construction industry are potentially setting the stage for a gradual shift away from parametric and complex formats in favor of more universal and independent formats USD, GLTF, DAE, OBJ (with meta-information both within hybrid and in separate structured or loosely structured formats). Historical leaders, including major design companies that once actively promoted IFC in the mid-1990s, are now openly promoting the new USD format [93], emphasizing its simplicity and versatility (Fig. 6.2-6). The mass adoption of USD in products, GLTF compatibility, and active integration into tools such as Blender, Unreal Engine, and Omniverse indicate the potential for the beginning of a new paradigm for working with data. Along with the popularity of localized solutions such as the European flat USD format - CPIXML, used in popular European ERPs could potentially strengthen the USD position in Central Europe. Organizations involved in the development of the IFC format are already adapting their strategy to USD [126], which only confirms the inevitability of the shift.

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	

Fig. 6.2-6 Comparison of IFC and USD format technical specifications.

Against this background, USD has the potential to become the de facto standard, promising to overcome many current limitations, primarily related to the complexity of existing CAD - (BIM-) formats and the dependence of their interpretation on geometric kernels.

Instead of parametric and complex CAD -formats and IFC - simplified data formats USD, glTF, DAE, OBJ with element meta-information in CSV, XLSX, JSON, XML will gain a place in the construction industry due to their simplicity and flexibility.

Current changes in the construction industry at first glance look like a technological breakthrough associated with the transition from the aging IFC to the more modern USD. However, it should be taken

into account that back in 2000 the same CAD vendor, which developed IFC, wrote about its problems and the need for access to the database [65], and now actively promotes the transition to the new standard - USD.

Behind yet another facade of "open data" USD and "new" concepts for granular data management, through cloud-based applications that CAD vendors are starting to promote may hide the vendors' intention to monopolize project data management, where users find themselves in a position where the choice of format has more to do with corporate interests than real-world needs.

An analysis of key facts [93] shows that the main goal of these changes is rather less about user convenience than primarily about maintaining control over ecosystems and data flows for the benefit of vendors who, in 40 years, have never been able to provide access to CAD databases.

Perhaps now is the time for companies to stop waiting for new concepts from software vendors and focus on self-development in the data-centric direction. Having freed itself from data access problems through reverse engineering tools, the industry will be able to move independently to modern, free and convenient tools for working and analyzing data without imposing new concepts.

CAD (BIM) Maturity Levels: From Stage 0 to Structured Data

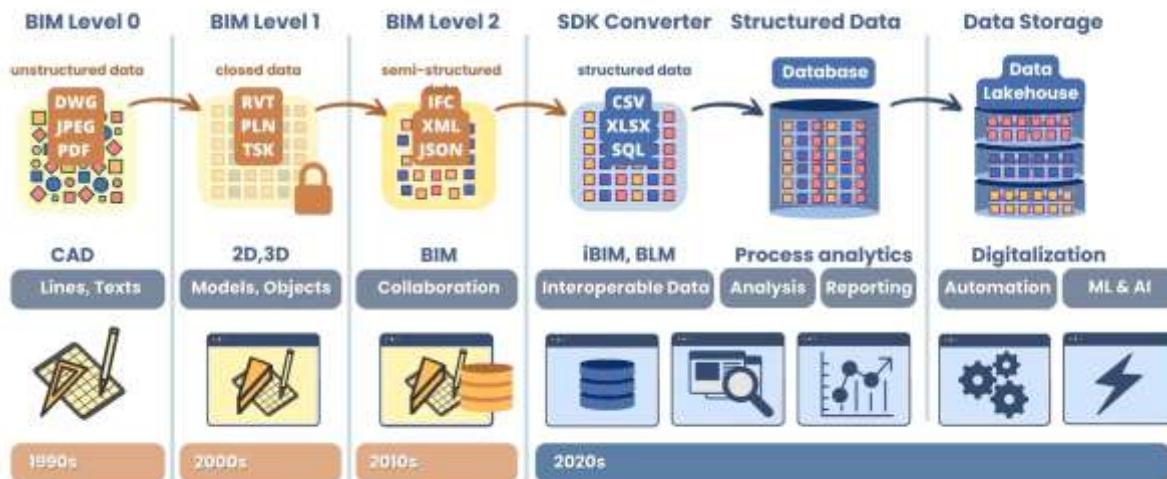


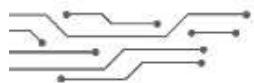
Fig. 6.2-7 CAD maturity level (BIM): from unstructured data to structured data and repositories.

Access to databases, open data and formats will inevitably become a standard in the construction industry, regardless of vendors' attempts to stall the process - it is only a matter of time (Fig. 6.2-7). The pace of this transition can increase significantly if more and more specialists become familiar with open formats, database tools and available reverse engineering SDKs, which allow organizing direct access to CAD data -systems [92].

The future lies in open, unified and analytically accessible data. To avoid dependence on vendor solutions and to avoid being held hostage to closed ecosystems, construction and engineering companies will sooner or later have to rely on openness and independence, choosing formats and solutions that provide full control over data.

The data that is being created in the construction industry today will be a key resource for business decisions in the future. It will act as the strategic "fuel" that fuels the development and efficiency of construction companies. The future of the construction industry lies in the ability to work with data, not in the choice of data formats or models.

To understand the difference between open formats USD, glTF, DAE, OBJ and proprietary parametric CAD formats, it is important to consider one of the most complex and key data elements in visualization and design calculations - geometry and its generation processes. And to understand how geometry data becomes the basis for analytics and calculations in construction, it is necessary to further explore the mechanisms of geometry generation, transformation and storage.



CHAPTER 6.3.

GEOMETRY IN CONSTRUCTION: FROM LINES TO CUBIC METERS

When lines turn into money or why builders need geometry

Geometry in construction is not only a visualization, but also the basis for accurate quantitative calculations. In the project model, geometry supplements the lists of element parameters (Fig. 3.1-16) with important volumetric characteristics such as length, area and volume. These volumetric parameter values are calculated automatically using the geometry kernels and are the starting point for estimates, schedules, and resource models. As we discussed in Part 5 of the book and in the chapter "Costing and Estimating construction projects", it is the volumetric parameters of object groups from CAD models that form the basis for modern ERP, PMIS systems. Geometry plays a fundamental role not only in the design phase, but also in project implementation management, schedule control, budgeting and operation. Just as thousands of years ago, when building Egyptian pyramids, the accuracy of a project depended on length measures like elbows and cubits, today the accuracy of geometry interpretation in CAD -programs directly affects the result: from budget and deadlines to contractor selection and delivery logistics

In a highly competitive and budget-constrained environment, the accuracy of volumetric calculations, which directly depends on geometry, becomes a survival factor. Modern ERP-systems directly depend on correct volumetric characteristics obtained from CAD - and BIM-models. That is why accurate geometric description of elements is not just a visualization, but a key tool for managing the cost and timing of construction.

Historically, geometry has been the primary language of engineering communication. From lines on papyrus to digital models, drawings and geometric representations have served as a means of exchanging information between designers, foremen and estimators. Before the advent of computers, calculations were done manually, using rulers and protractors. Today, this task is automated thanks to volumetric modeling: the geometric kernels of CAD programs transform lines and points into three-dimensional bodies, from which all the necessary characteristics are automatically extracted.

Working in CAD -programs, creation of geometric elements for calculations is performed through the user interface of CAD- (BIM-) programs. To transform points and lines into volumetric bodies, the geometric kernel is used, which performs the key task - transformation of geometry into volumetric models, from which the volumetric characteristics of the element are automatically calculated after approximation.

From lines to volumes: How area and volume become data

In engineering practice, volumes and areas are computed from geometric surfaces described analytically or through parametric models such as NURBS (nonuniform rational B-splines) within the BREP (boundary element representation) framework.

NURBS (Non-Uniform Rational B-Splines) is a mathematical way of describing curves and surfaces, whereas BREP is a framework for describing the complete three-dimensional geometry of an object, including its boundaries, which can be defined using NURBS.

Despite the accuracy of BREP and NURBS, they require powerful computational resources and complex algorithms. However, direct computation from such mathematically accurate descriptions is often computationally difficult, so in practice, tessellation - the transformation of surfaces into a grid of triangles - is almost always used, which simplifies subsequent computations. Tessellation is the partitioning of a complex surface into triangles or polygons. In CAD /CAE environments this method is used for visualization, volume calculations, collision search, export to formats like MESH and collision analysis. An example from nature is bee honeycomb, where a complex shape is broken down into a regular grid (Fig. 6.3-1).

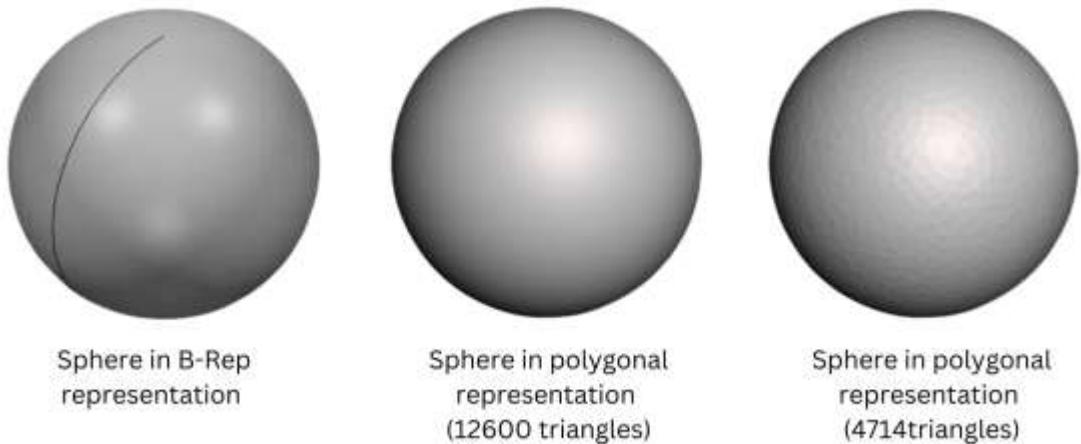


Fig. 6.3-1 The same sphere in parametric description BREP and polygonal representation with different number of triangles.

BREP (NURBS), used in CAD, is not a fundamental model of geometry. It was created as a convenient tool for representing circles and rational splines and for minimizing the storage of geometry data. However, it has limitations - for example, the inability to accurately describe the sinusoid that underlies helical lines and surfaces, and the need for complex geometry kernels.

In contrast, triangular meshes and tessellation of parametric shapes are characterized by simplicity, efficient use of memory and ability to process large amounts of data (Fig. 6.3-2). These advantages make it possible to do without complex and expensive geometry kernels, and the tens of millions of

lines of code embedded in them, when calculating geometric shapes.

In most building cases it does not matter how exactly the volumetric characteristics are defined - through parametric models (BREP, IFC) or through polygons (USD, glTF, DAE, OBJ). The geometry remains the form of approximation: whether through NURBS or MESH, it is always an approximate description of the shape.

Geometry defined as polygons or BREP (NURBS) remains to some extent only a way of approximating with an approximate description of a continuous form. Just as Fresnel integrals have no exact analytic expression, discretizing geometry through polygons or NURBS is always an approximation, just as triangular MESH.

Parametric geometry in BREP format is necessary mainly where minimal data size is important and it is possible to use resource-intensive and expensive geometry kernels for its processing and display. Most often it is characteristic for developers of CAD -programs, which for this purpose apply in their products geometrical kernels of MCAD -vendors. At the same time, even within these programs, BREP-models in the process of tessellation for visualization and calculations are often converted into triangles (similar to the way PSD-files are simplified into JPEG).

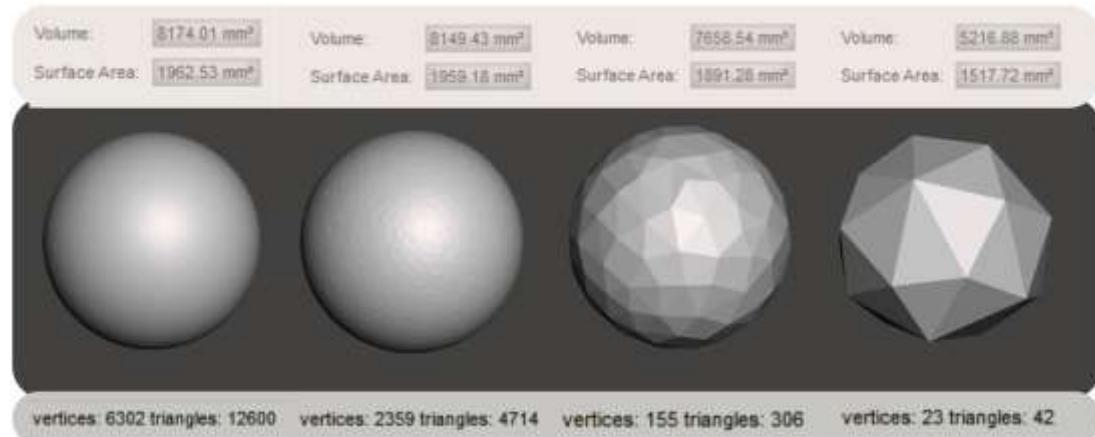


Fig. 6.3-2 Difference of volumetric characteristics in figures with different number of polygons.

Polygonal MESH, as well as parametric BREP, have their own advantages and limitations, but the goal is the same - to describe the geometry taking into account the user's tasks. Ultimately, the accuracy of a geometric model depends not only on the method of its representation, but also on the requirements of a particular task.

In most construction problems, the need for parametric geometry and complex geometric kernels may be redundant.

In each particular calculation automation task, it is worth considering whether the importance of parametric geometry is exaggerated by CAD developers who are interested in promoting and selling their own software products.

Moving to MESH, USD and polygons: using tessellation for geometry

In the construction industry, when streaming, developing systems, databases or automating processes that work with design information and feature geometry, it is important to strive for independence from specific CAD editors and geometry kernels.

The exchange format to be used both in the calculation departments and on the construction site should not be based on a specific CAD- (BIM-) program. Geometric information should be represented in the format directly through tessellation, without reference to the geometry core or CAD architecture.

Parametric geometry from CAD can be considered as an intermediate source, but not as the basis for a universal format. Most parametric descriptions (including BREP and NURBS) are in any case converted to polygonal MESH for further processing. If the result is the same (tessellation and polygons) and the process is simpler, the choice is obvious. This is analogous to the choice between graph ontologies and structured tables (which we discussed in part four): excessive complexity is rarely justified (Fig. 3.2-10, Fig. 6.1-8).

Open formats such as: OBJ, STL, glTF, SVF, CPIXML, USD and DAE, use a universal triangle mesh structure, which gives them significant advantages. These formats have excellent interoperability - they are easy to read and visualize using available open source libraries without the need for complex specialized geometry kernels containing millions of lines of code (Fig. 6.3-3). These versatile geometry formats are used in applications ranging from relatively simple kitchen design tools in IKEA™ to complex object visualization systems in movie and VR-applications. An important advantage is the availability of a large number of free and open source libraries for working with these formats, available for most platforms and programming languages.

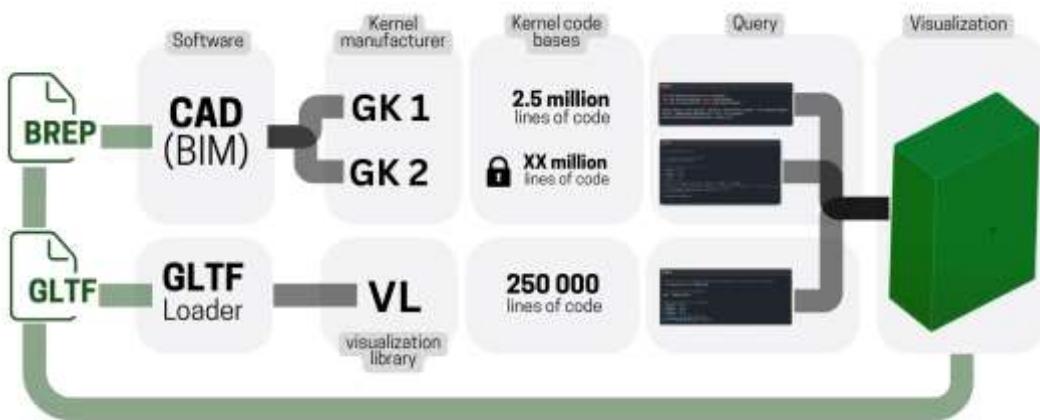


Fig. 6.3-3 The same geometry representation is achieved through the use of parametric formats and geometry kernels, or through the use of triangulated formats and open source visualization libraries.

As well as the users themselves, CAD -vendors face problems with interpreting foreign parametric CAD formats or open IFC because of different geometry kernels. In practice, all CAD -vendors, without exception, use the reverse engineering SDK to transfer data between systems, and none of them rely on formats like IFC or USD [93] for interoperability purposes.

Instead of using concepts promoted by alliances of CAD- vendors that they themselves do not use - it is more productive for developers and users of CAD solutions to focus on understanding the benefits of each approach in a specific context and to choose one or another type of geometry depending on the use case. Choosing between different geometric representations is a trade-off between accuracy, computational efficiency and the practical needs of a particular task.

The complexity associated with the use of geometric kernels, which is traditionally imposed on the construction industry by large vendors when processing design data, often turns out to be redundant. The USD format based on MESH geometry can become a kind of "Pandora's box" for the industry, opening new opportunities for developers to organize data exchange - outside the framework of IFC and parametric BREP structures typical for CAD vendors.

After a closer look at the structure of USD, DAE, gLTF, OBJ, etc., it becomes obvious that there are simpler, open formats that allow to efficiently organize the transfer and use of geometric information without the need to rely on complex parametrics and closed geometric kernels. This approach not only lowers the technical threshold of entry for developers, but also promotes the development of flexible, scalable and truly open solutions for digital construction.

LOD, LOI, LOMD - unique classification of detailing in CAD (BIM)

In addition to geometric representation formats, in a world where different industries use different

levels of detail and depth of data, CAD - (BIM-) methodologies offer their own unique classification systems, which structure the approach to informing building models.

One of the examples of new approaches to standardization is the introduction of levels of model development, reflecting the degree of readiness and reliability of both graphical and information components. For differentiation of information content in work with CAD - (BIM-) data there appeared LOD (Level Of Detail) - level of detail of the graphical part of the model, and LOI (Level Of Information) - level of data elaboration. In addition, for the integrated approach the concept of LOA (Level of Accuracy) was introduced - the accuracy of represented elements and LOG (Level of Geometry) to determine the accuracy of graphical representation.

Levels of detail (LOD) are indicated by numbers from 100 to 500, reflecting the degree of model development. LOD 100 is a conceptual model with general shapes and dimensions. LOD 200 includes more precise dimensions and shapes, but with conditional detail. LOD 300 is a detailed model with precise dimensions, shapes, and element locations. LOD 400 contains detailed information required for fabrication and installation of elements. LOD 500 reflects the actual condition of the facility after construction and is used for operation and maintenance. These levels describe the structure of CAD (BIM) model information saturation at different stages of the life cycle, including 3D, 4D, 5D and beyond.

In real projects, high level of detail (LOD400) is often excessive and it is sufficient to use LOD100 geometry or even flat drawings, while the rest of the data can be obtained either by calculation or from related elements that may not have a distinct geometry. For example, spaces and room elements (categories of "Premises" elements) may have no visual geometry, but still contain significant amounts of information and databases around which many business processes are built.

Therefore, it is important to clearly define the required level of detail before starting the design. For 4D -7D use cases, even DWG drawings and minimal LOD100 geometry are often sufficient. The key task in the requirements process is to find a balance between saturation and practical applicability of the model.

In essence, if we consider CAD (BIM) data as a database (which it is), the description of model saturation through new acronyms is nothing but a step-by-step modeling of data for information systems, starting from the conceptual level and ending with the physical one (Fig. 6.3-4), which was discussed in detail in the third and fourth parts of the book. Each increase of LOD and LOI means addition of information needed for new tasks: calculations, construction management, operation and is characterized by successive enrichment of the model with additional information layers (3D -8D) in the form of various parameters, which we discussed in the fifth part of the book.

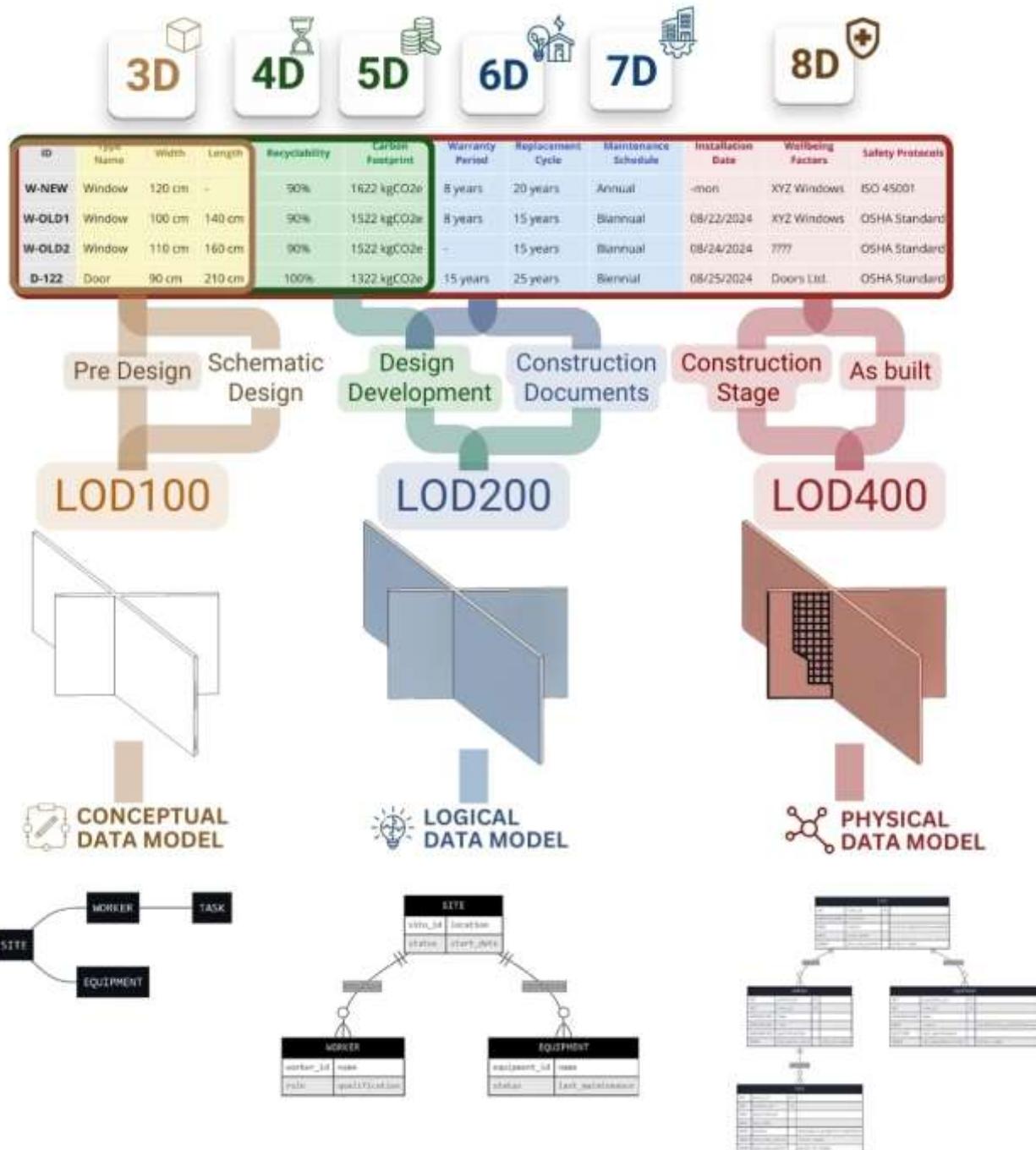


Fig. 6.3-4 The process of fleshing out the details of a project is identical to data modeling from a conceptual to a physical data model.

Geometry is only some of the design data, the need for which is not always justified in construction projects and the key issue of working with CAD -data is not so much how the models are visualized, but more how the data from these models can be used outside of CAD- (BIM-) programs.

By the mid-2000s, the construction industry faced an unprecedented challenge with the rapid increase

in the amount of data in management and data processing systems, especially those coming from CAD departments (BIM). This dramatic increase in data volume took company managers by surprise and they were unprepared for the growing demands on data quality and management.

New CAD standards (BIM) - AIA, BEP, IDS, LOD, COBie

Taking advantage of the lack of open access to CAD databases and limited competition in the data processing market, and using marketing campaigns related to the new acronym BIM, organizations involved in developing approaches to working with CAD data - have begun to create new standards and concepts that de jure should be aimed at improving data management practices.

While almost all initiatives directly or indirectly supported by CAD vendors and developers (BIM) have been aimed at optimizing workflows, they have resulted in a plethora of standards lobbied by various stakeholders, leading the construction industry to some ambiguity and confusion about data processes.

Let's list some of the new data standards, in addition to LOD, LOI, LOA, LOG, that have emerged in recent years in the construction industry:

- **BEP** (BIM Execution Plan) - describes how to integrate and utilize CAD (BIM) in a project by defining data processing methods and processes.
- **EIR document /AIA** (Information Requirements of the client) - prepared by the client prior to the call for tender and contains the requirements for the contractor to prepare and provide information. It serves as the basis for the BEP in the respective project.
- **AIM** (Asset Information Model) is part of the BIM process. Once the project is delivered and completed, the data model is called the Asset Information Model or AIM. The purpose of AIM is to manage, maintain and operate the realized asset.
- **IDS** (Information Delivery Specification) - defines the requirements of and what data and in what format is required at different stages of a construction project.
- **iLOD** is the level of detail LOD, with which information is represented in the BIM -model. It defines how detailed and complete the information in the model is, from basic geometric representations to detailed specifications and data.
- **eLOD** - LOD level of detail of individual elements in a CAD model (BIM). It defines the degree of modeling of each element and associated information such as dimensions, materials, performance characteristics and other relevant attributes.
- **APS** (Platform Services) and other products from major CAD vendors (BIM) - describe the tools and infrastructure needed to create linked and open data models.

Although the declared purpose of implementing CAD (BIM) standards - such as LOD, LOI, LOA, LOG,

BEP, EIR, AIA, AIM, IDS, iLOD, eLOD - is to improve the quality of data management and expand automation capabilities, in practice their use often leads to excessive complexity and fragmentation of processes. If we consider CAD (BIM) model as a kind of database, it becomes obvious that many of these standards duplicate long-established and effective approaches used in other industries when working with information systems. Instead of simplification and unification, such initiatives often create additional terminological burden and hinder the implementation of truly open and flexible solutions.

Notably, many of these new concepts are actually replacing the modeling and data validation processes that were discussed in detail in the first parts of the book and that have long been used in other sectors of the economy. In construction, however, the standardization process often moves in the opposite direction - new data description formats, new standards, and new concepts for data validation are created that do not always lead to real uniformity and practical applicability. As a result, instead of simplifying and automating processing, the industry faces additional levels of regulation and bureaucracy (Fig. 6.3-1), which is not always conducive to increased efficiency.

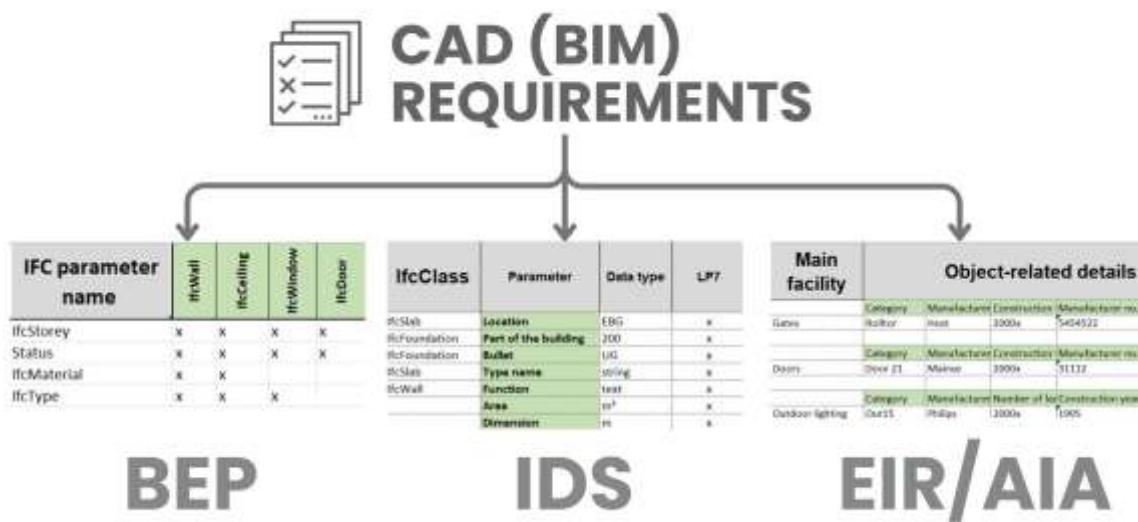


Fig. 6.3-1 Data and information content requirements are reduced to the description of attributes and their boundary values, described using tables.

Instead of simplifying data processing, new concepts related to CAD (BIM) data more often generate additional complexities and disputes already at the stage of interpretation and basic definitions.

One of the latest examples of new concepts is the IDS format (introduced in 2020) which allows to describe the requirements for the attribute composition of an information model in the open BIM concept. IDS requirements describe information about attributes and their boundary values in the form of a structured table (Excel or MySQL), which is then translated into the markup of a semi-structured XML format, renamed from XML to the special abbreviation IDS.

Contrary to the view promoted by vendors and supported by BIM and open BIM, that data handling in construction is unique because of the use of specialized tools such as CAD and BIM, data formats and data management practices in this industry are not different from data handling formats and concepts in other industries

The number of requirements for CAD projects and formats (BIM) can be simplified by using a single requirements table with attribute-columns, detailed in the chapter "Translating Requirements into Structured Form", without having to translate the originally structured requirements into non-table formats (IDS is initially described via a table).

The simplified approach (Fig. 6.3-2), which includes columns for entity identifiers, properties, and boundary values that were discussed in detail in previous chapters (Fig. 4.4-9, Fig. 4.4-16, Fig. 7.3-10), eliminates the need to convert requirements to IDS-XML format. This method provides a direct, less cumbersome, and more transparent mechanism for data quality control. It relies on widely used tools, from regular expressions (RegEx) to dataframes, Pandas, and standard ETL-Pipelines, just like those used by professionals in other sectors of the economy when working with data.

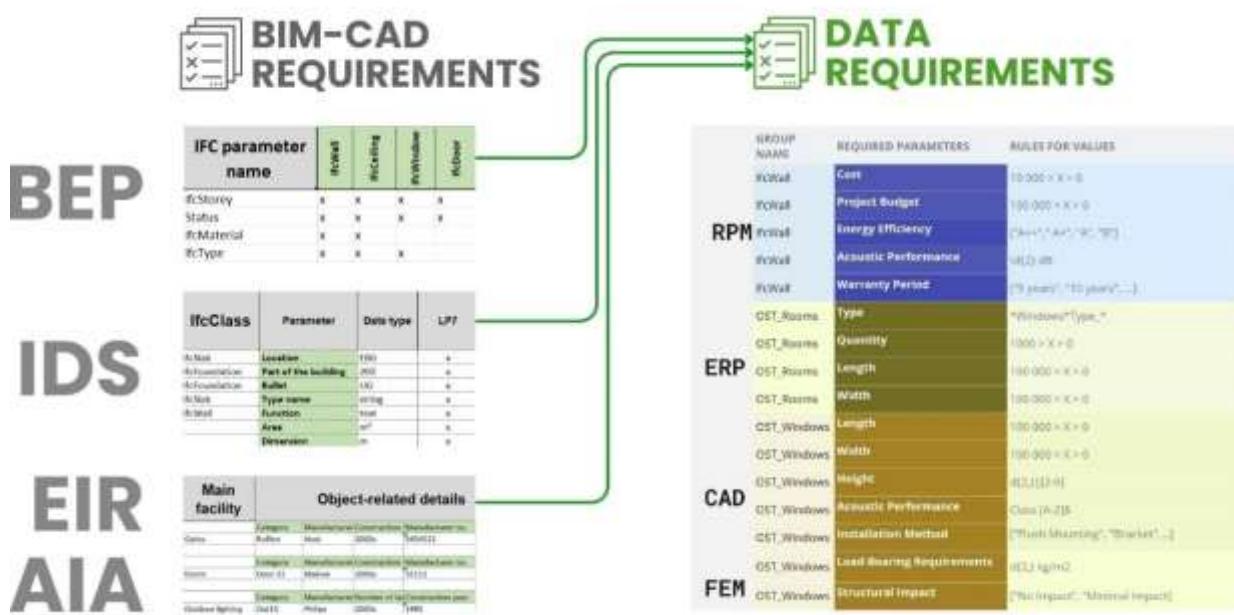


Fig. 6.3-2 Data requirements in other industries are simplified to a structured description of attributes and their boundary values.

Over time in the construction industry, due to the closed nature of the data, more and more new approaches and techniques are emerging to control and manage this diverse data, even though the data in construction projects is essentially the same as in other fields. While other industries have successfully made do with standardized approaches to data processing, construction continues to develop new and unique data formats, requirements and validation concepts.

The methods and tools used to collect, prepare and analyze data in construction should not be fundamentally different from those used by specialists in other sectors of the economy.

The industry has developed a distinct terminological ecosystem that requires critical reflection and re-evaluation:

- The STEP format is positioned under the new name IFC, supplemented by construction categorization, without taking into account the limitations of the STEP format itself.
- The parametric format IFC is used in data communication processes despite the lack of a unified geometric core needed for visualization and computation.
- Access to databases CAD -systems is promoted under the term "BIM ", with no discussion of the specifics of these databases and access to them.
- Vendors promote interoperability through the IFC and USD formats, often without practicing them, using costly reverse engineering that they themselves have struggled with.
- The terms LOD, LOI, LOA, LOG, BEP, EIR, AIA, AIM, IDS, iLOD, eLOD are used universally to describe the same entity parameters, without reference to modeling and verification tools long used in other industries.

The construction industry demonstrates that all of the above, although it sounds strange, is possible in the construction industry - especially if the main goal is to monetize each stage of data processing through the sale of specialized services and software. From a business point of view, there is nothing wrong with this. However, whether such acronyms and approaches related to CAD (BIM) really add value and simplify professional processes remains an open question.

In the construction industry, such a system works because the industry itself makes most of its speculative profits in these labyrinths of systems and acronyms. Companies interested in transparent processes and open data are rare. This complex situation will probably continue indefinitely - until customers, clients, investors, banks and private equity start demanding clearer and more informed approaches to information management.

The industry has accumulated an excessive number of acronyms, but they all describe the same processes and data requirements to varying degrees. Their real usefulness in simplifying workflows remains questionable.

While concepts and marketing acronyms come and go, the data requirements validation processes themselves will forever remain an integral part of business processes. Instead of creating more and more specialized formats and regulations, the construction industry should look to tools that have already proven effective in other areas such as finance, industry and IT.

The abundance of terms, acronyms and formats forms the illusion of deep elaboration of digital construction processes. However, marketing concepts and complex terminology often hide a simple but

inconvenient truth: data remains hard to access, poorly documented and rigidly tied to specific software solutions.

To get out of this vicious circle of acronyms and formats for the sake of formats, it is necessary to look at CAD (BIM) systems not as magical information management tools, but as what they really are - specialized databases. And it is through this prism that one can understand where marketing ends and real work with information begins.



CHAPTER 6.4.

DESIGN PARAMETERIZATION AND USE OF LLM FOR CAD OPERATION

The illusion of uniqueness of CAD data (BIM): the path to analytics and open formats

Modern CAD (BIM) platforms have significantly transformed the approach to design and construction information management. While previously these tools were mainly used to create drawings and 3D models, today they serve as full-fledged repositories of design data. Under the Single Source of Truth concept, the parametric model is increasingly becoming the main and often the only source of project information, ensuring its integrity and relevance throughout the entire lifecycle of an object.

The key difference between CAD - (BIM -) platforms and other construction data management systems is the need to use specialized tools and APIs to access the information (the only source of truth). These databases are not universal in the traditional sense: instead of an open structure and flexible integration, they are a closed environment, hardwired to a specific platform and format.

Despite the complexity of working with CAD -data there is a more important question that goes beyond the technical realization: what are CAD databases (BIM) really? To answer this question, it is necessary to go beyond the usual acronyms and concepts imposed by software developers. Instead, it is worth focusing on the essence of working with project information: data and its processing.

The business process in construction begins not with work in CAD - or BIM - tools, but with the formation of project requirements and data modeling. First, the task parameters are defined: the list of entities, their initial characteristics and boundary values that need to be considered when solving a particular task. Only after that models and elements are created in CAD (BIM) systems on the basis of the specified parameters

The process that precedes the creation of information in CAD - (BIM-) databases is completely the same as the data modeling process that was discussed in detail in the fourth part of the book and the chapter "Data Modeling: conceptual, logical and physical model" (Fig. 4.3-1).

Just as in data modeling we create requirements for the data we later want to process in the database, for CAD databases, managers create design requirements in the form of several table columns or lists of key-value pairs (Fig. 6.4-1, steps 1-2). And only on the basis of these initial parameters using the API automatically or manually, the designer creates (or rather refines) objects in CAD- (BIM) databases (steps 3-4), after which they are checked again for compliance with the initial requirements (steps 5-6). This process - definition→ creation→ validation→ adjustment (steps 2-6) - is re-

peated iteratively until the data quality, just as in data modeling, reaches the desired level for the target system - documents, tables or dashboards (step 7).

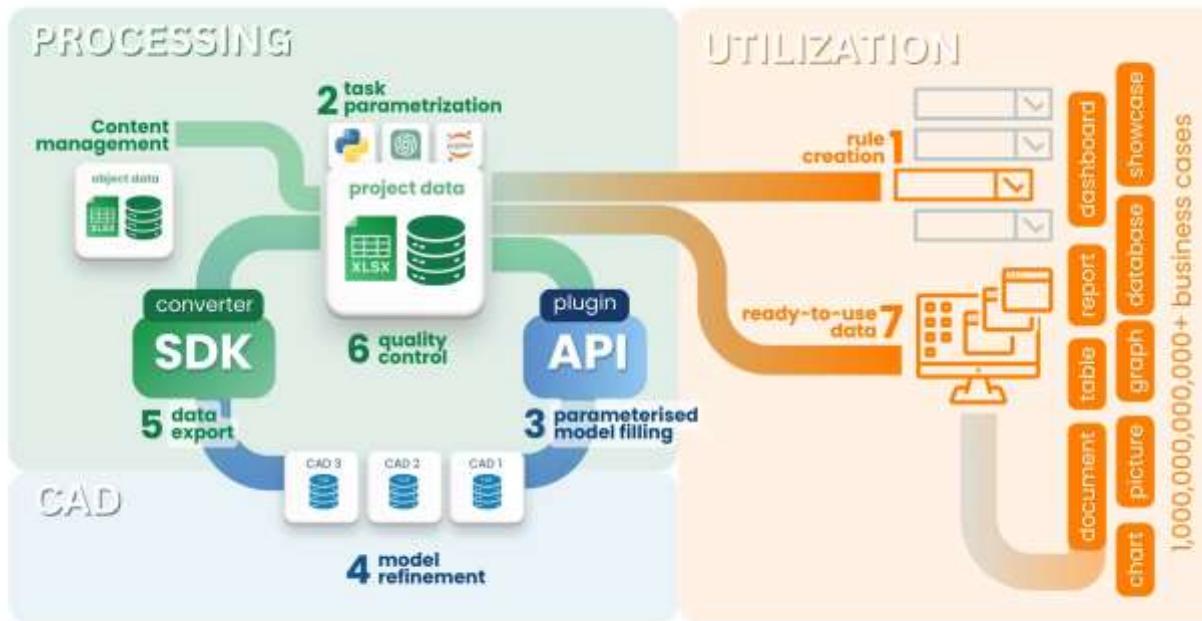


Fig. 6.4-1 The cycle of information saturation of databases for business processes in the realization of construction projects.

If we consider CAD (BIM) as a mechanism for parameter transfer in the form of a set of key-value pairs generated on the basis of requirements defined outside the design environment (Fig. 6.4-1, steps 1-2), the focus of the discussion shifts from specific software solutions and their limitations to more fundamental aspects - data structure, data models and data requirements. In essence, we are talking about parameter saturation of the database and the classical data modeling process (steps 2-3 and 5-6). The only difference is that due to the closed nature of CAD-databases and the peculiarities of the formats used, this process is accompanied by the use of specialized BIM-tools. The question arises: what is the uniqueness of BIM, if there are no similar approaches in other industries?

For the last 20 years, BIM has been positioned as more than just a single data source. The CAD -BIM marketing bundle is often sold as a parametric tool with an inherently integrated database [64], capable of automating the processes of design, modeling and life cycle management of construction objects. However, in reality, BIM has become more a tool to keep users on the vendors' platform than a convenient method of data and process management.

As a result, CAD- (BIM-) data is isolated within their platforms, hiding project information behind proprietary APIs and geometry kernels. This has deprived users of the ability to independently access databases and extract, analyze, automate, and transfer data to other systems, bypassing vendor ecosystems.

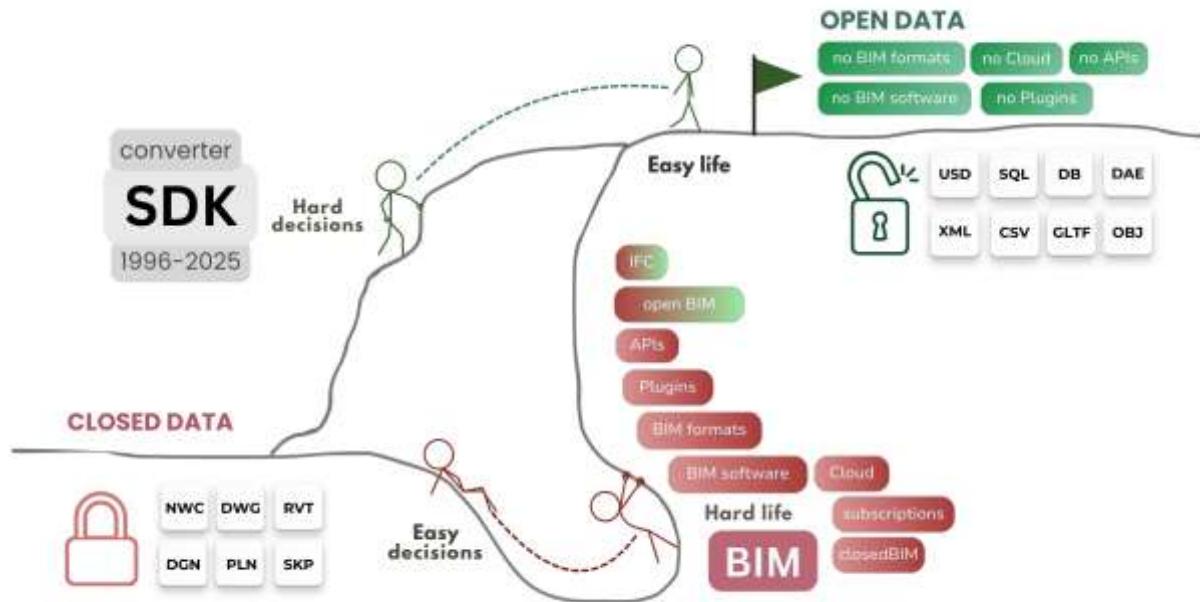


Fig. 6.4-2 In construction, modern formats require sophisticated geometry kernels, an annually updated API, and special licenses for CAD -(BIM -) programs.

Companies that work with modern CAD tools should use the same approach to working with data that all CAD vendors themselves without exception use in practice: data transformation using SDK - reverse engineering tools, which CAD vendors have been fighting against since 1995 [75]. Having full access to the CAD database and using reverse engineering tools, we can obtain [127] a flat set of entities with attributes and export them to any convenient open format (Fig. 6.4-2), including both geometry and parameters of design elements. This approach fundamentally changes the paradigm of working with information - from file-oriented to data-centered architecture:

- Data formats such as RVT, IFC, PLN, DB1, CP2, CPIXML, USD, SQLite, XLSX, PARQUET and others contain identical information about elements of the same project. This means that knowledge of a particular format and its schema should not be a barrier to working with the data itself.
- Data from any formats can be combined into a single open structured and granular structure (Fig. 9.1-10) containing the MESH triangular geometry and the properties of all object entities, without the constraints of geometric kernels.
- Data analytics strives for universality: using open data, you can work with project data regardless of the format used.
- Minimization as well as dependency on APIs and vendor plugins: working with data no longer depends on API skills.

When and CAD -data requirements are transformed into analytics-friendly structured representation formats - developers are no longer dependent on specific data schemas and closed ecosystems.

Design through parameters: the future of CAD and BIM

No construction project in the world has ever started in a CAD program. Before a drawing or model takes shape in CAD, it passes through the conceptualization stage (Fig. 6.4-1, stages 1-2), where the focus is on the parameters that define the basic idea and logic of the future object. This stage corresponds to the conceptual level in data modeling (Fig. 4.3-6). Parameters may exist solely in the designer's mind, but ideally they are formalized in the form of structured lists, tables or stored in databases (Fig. 6.4-3), which allows for transparency, reproducibility and further automation of the design process.

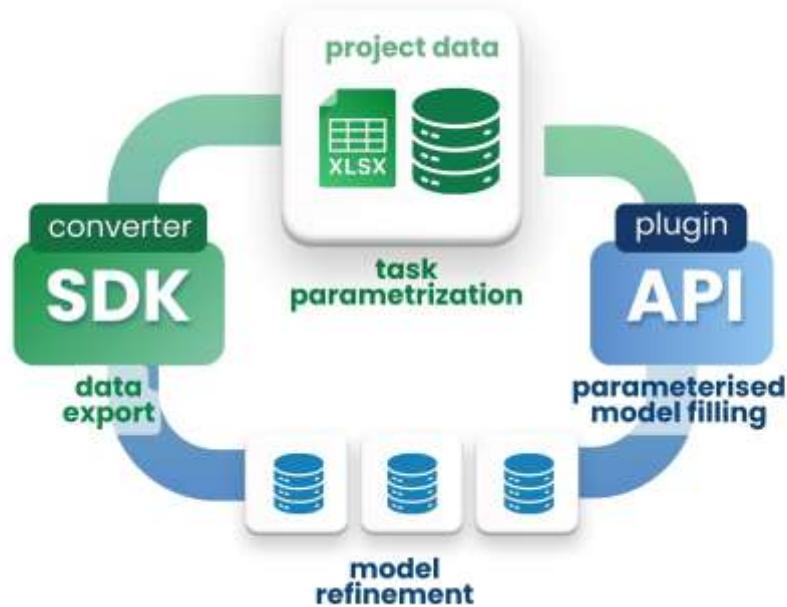


Fig. 6.4-3 The design process is an iterative process of populating the CAD database with information from the outside by means of requirements in the value chain.

Before starting the CAD modeling itself (the logical and physical phase of data modeling (Fig. 4.3-7)), it is important to define the boundary parameters that serve as the basis of the project. These attributes, as with other requirements, are collected from the very end of the data usage chain (e.g. systems) and through them the constraints, goals and key characteristics of future objects in the project are already defined.

The modeling itself can be fully automated by 60-100% with the help of parametric modeling tools (Fig. 6.4-3), if the requirements are well defined. As soon as the project is described in the form of parameters, its formation becomes technically feasible with the help of visual programming languages such as Grasshopper Dynamo, embedded in modern CAD-environments or free solutions in Blender, UE, Omniverse.

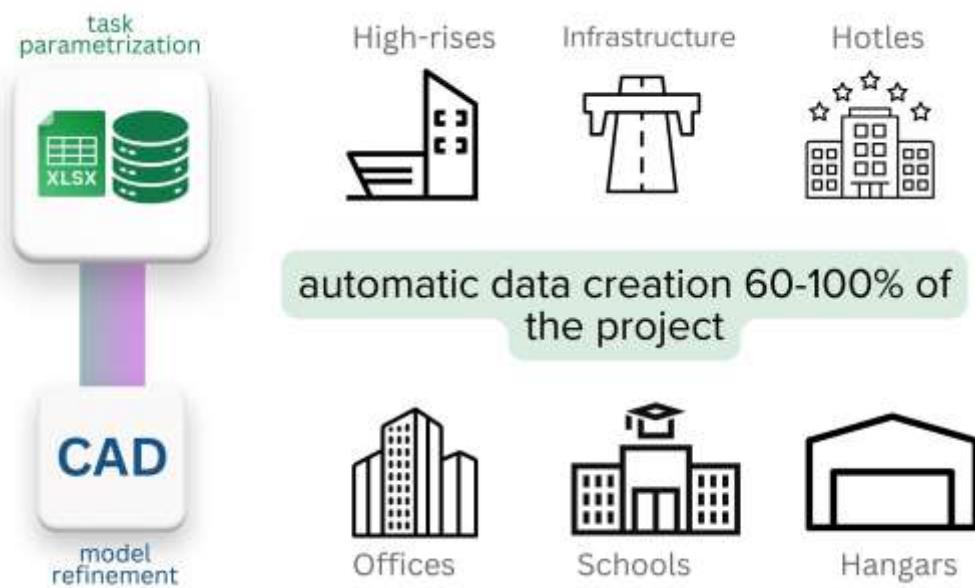


Fig. 6.4-4 The majority of typed projects are already created fully automatically today thanks to parametric programming tools.

Already today, large industrial and typified projects are not created by the hands of the design department, but through parametric tools and visual programming. This makes it possible to build a model based on data rather than on the subjective decisions of a particular designer or manager.

Content precedes design. Design without content is not design, but decoration [128].

— Jeffrey Zeldman, web designer and entrepreneur

The process does not start with drawing or 3D -modeling, but with the formation of requirements. It is the requirements that determine what elements will be used in the project, what data needs to be transferred to other departments and systems. Only the existence of structured requirements makes it possible to automatically check models on a regular basis (for example, even every 10 minutes without distracting the designer from his work).

Perhaps in the future CAD- (BIM-) system will become just an interface for filling the database, and in what CAD tool is modeling (physical level) - will not matter.

Similarly, in mechanical engineering, 3D modeling is often used but is not a necessary or mandatory element of the project. In most cases, the classic 2D documentation is sufficient and the necessary information model is created on its basis. This model is assembled from components structured according to industry standards and contains all the necessary information for understanding the design and organization of production. The factory information model is then used to create a factory infor-

mation model, to which specific products and flow charts, already oriented to the needs of technologists, are added. The whole process can be organized without unnecessary complexity, without overloading the system with 3D graphics where they do not provide real advantages.

It is important to realize that the 3D model itself and CAD -system should not play the main role - it is just a tool for quantitative and geometric analysis. All other parameters, except geometry, which describe the entity, should be stored and processed outside the CAD environment (BIM) if possible.

Design through parameters is not just a trend, but the inevitable future of the construction industry. Instead of creating complex 3D -models manually, designers will work with data, validate it and automate processes, bringing construction closer to the world of programming. Over time, design processes will be built on the principles of software development:

- Create requirements → Create model → Upload to server → Validate changes → Pull request
- The Pull request automatically runs model checks against requirements that were created before or during the design process
- After data quality checks and approval, changes are implemented into the project, the common database or transferred automatically to other systems

Already now in mechanical engineering, such design changes start with the formation of a change notice. A similar scheme awaits the construction industry: design will be an iterative process where each step is supported by parametric requirements. Such a system will allow designers to create automated checks and automated pull request for specific requirements.

The designer of the future is primarily a data operator, not a manual modeler. His task is to fill the project with parametric entities, where geometry is only one of the attributes.

It is the understanding of the importance of data modeling, classification and standardization, which have been discussed in detail in the previous chapters of the book, that will play an important role in the transformation. The design regulations of the future will be formalized as key-value parameter pairs in the form of XLSX or XML -schemas.

The future of the construction industry is about collecting data, analyzing it, validating it and automating processes with analytics tools. BIM (or CAD) is not the end goal, but only a stage of evolution. When professionals realize that they can work directly with data, bypassing traditional CAD tools, the term "BIM" itself will gradually give way to the concepts of using structured and granular construction project data.

One of the key factors accelerating the transformation has been the emergence of large language models (LLM) and the tools based on them. These technologies are changing the way design data is

handled, enabling access to information without the need for in-depth knowledge of APIs or vendor solutions. With LLMs, the process of creating a requirement and interacting with CAD data becomes intuitive and accessible.

Emergence of LLM in design CAD data processing processes

In addition to the development of CAD database access tools and open and simplified CAD -formats, the emergence of LLM -tools (Large Language Models) is revolutionizing the processing of design data. Whereas previously the access to information was mainly through complex interfaces and required programming and knowledge of API, now it is possible to interact with data using natural language.

Engineers, managers and planners with no technical background can obtain the necessary information from project data by formulating queries in ordinary language. Provided the data is structured and accessible (Fig. 4.1-13), it is enough to ask a question in LLM chat like: "*Show in a table with grouping by type all walls with a volume of more than 10 cubic meters*" - and the model will automatically convert this query into SQL or code in Pandas, generating a summary table, graph or even a ready document.

Below are some real-life examples of how LLM -models interact with design data represented in various CAD- (BIM-) formats.

- 💡 Example of a query in LLM chat to a CAD project in RVT format after conversion (Fig. 4.1-13) to a tabular dataframe (CHATGP, LLaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

Group the data in Dataframe obtained from RVT file by "Type name" when summarizing "Volume" parameter and show the number of elements in the group. And please show all this as a horizontal histogram without zero values.

- 💡 LLM response as a horizontal histogram (PNG format):

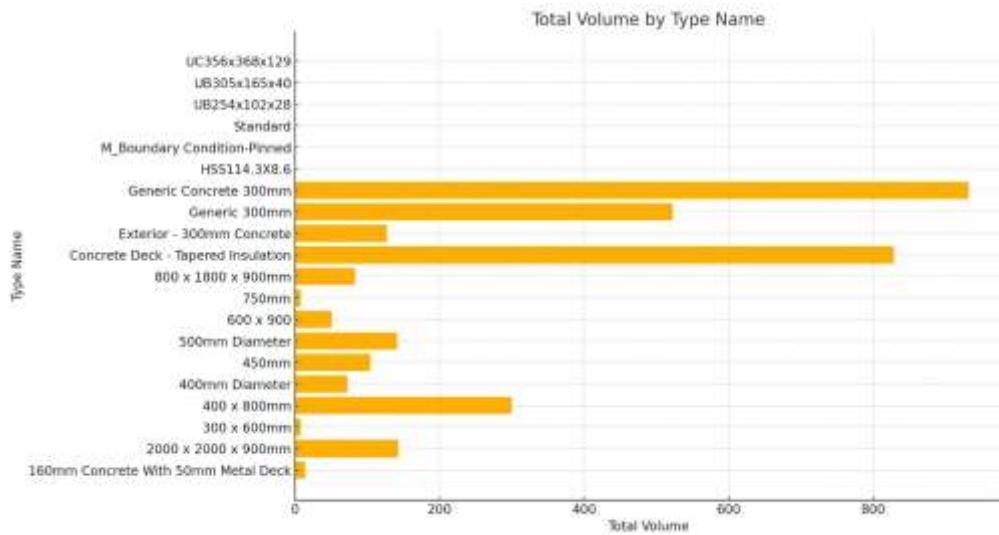


Fig. 6.4-5 Instead of 17 mouse clicks or 40 lines of code through the use of plug-ins, in LLM we instantly retrieve the QTO table through a text query.

- 💡 To generate a QTO table of wall types with total area and quantity from the category "Walls" let's formulate a text query for LLM -chat:

Take only those items from the project dataframe that have "OST_Walls" in the "Category" parameter, group them by "Type Name", sum the value of the "Area" column, add the quantity, and display them in the table by removing null values.

- ❷ LLM response in the form of a finished QTO table:

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

Fig. 6.4-6 Creating a QTO table in natural language provides the same quality results as when using CAD - (BIM-) tools.

- ❸ Let's query the project in IFC format after converting to a tabular dataframe and enter a similar text query into any LLM chat:

Take only those items from the project that have Level 1 and Level 2 values in the "Parent" parameter, and take the items that have IfcSlab values in the "Category" parameter, then group these items by the "ObjectType" parameter, sum the values in the "PSet_RVT_Dimensions_Area" parameter, and show them as a pie chart.

- LLM response as a finished Pie chart of item groups from IFC data:

For ■ Floor:127mm Slab on Grade, ■ Floor:150mm Exterior Slab on Grade, ■ Floor:Finish Floor - Ceramic Tile, ■ Floor:Finish Floor - Wood, and ■ Floor:Residential - Wood Joist with Subflooring



Fig. 6.4-7 The result of an IFC data query in a structured format can be any type of graph that is convenient for understanding the data.

Behind each of the resulting ready-made solutions (Fig. 6.4-5 - Fig. 6.4-7) is a dozen lines of Python code using the Pandas library. The resulting code can be copied from the LLM chat room and used in any local or online IDE to get identical results outside the LLM chat room.

In the same LLM chat we can work not only with projects obtained from 3D CAD (BIM) formats, but also with flat drawings in DWG format, to which after conversion to a structured form we can query the LLM chat in order to display, for example, data on element groups in the form of lines or 3D geometries.

Automated analysis of DWG -files with LLM and Pandas

The process of data processing from DWG -files due to the unstructured nature of the information - has always been a complex task, requiring specialized software and often manual analysis. However, with the development of artificial intelligence and LLM tools, it has become possible to automate many stages of this, today, mostly manual process. Consider a real Pipeline of requests to LLM (in this example ChatGPT) to work with DWG drawings, which allow you to work with the project:

- Filter DWG data by layer, ID and coordinates
- Visualize the geometry of the elements
- Automatically annotate drawings based on parameters
- Expand wall polylines to the horizontal plane

- Create interactive 3D -visualizations of planar data
- Structure and analyze construction data without complex CAD -tools

In our case, the process of building Pipeline starts with sequential code generation through the LLM. First, a query describing the task is generated. ChatGPT generates Python -code, which is executed and analyzed, showing the result inside the chat room. If the result is not as expected, the request is corrected and the process is repeated

Pipeline is a sequence of automated steps performed to process and analyze data. In such a process, each step takes data as input, performs transformations, and passes the result to the next step.

After obtaining the desired result, the code is copied from LLM and pasted into the code in the form of blocks in any of the convenient IDEs, in our case on the Kaggle platform.com. The resulting code fragments are combined into a single Pipeline, which automates the entire process - from data loading to its final analysis. This approach allows rapid development and scaling of analytical processes without deep programming expertise. The full code of all the fragments below, along with sample queries, can be found on the Kaggle.com platform by searching for "DWG Analyse with ChatGPT | DataDrivenConstruction" [129].

Let's start the process of working with DWG data, after conversion to structured view (Fig. 4.1-13), with a classical step - grouping and filtering of all the drawing data, necessary for our task wall elements, specifically polylines (parameter 'ParentID' allows to group lines into groups), which in the parameter (dataframe column) "Layer" has a string value containing the following combination of letters (RegEx) - "wall".

- ❑ To get the code for a similar task and the result in the form of a picture you should write the following query in LLM:

First, check if the dataframe obtained from DWG contains the defined columns: 'Layer', 'ID', 'ParentID' and 'Point'. Then filter out the IDs from the 'Layer' column that contain the string 'wall'. Find the items in the 'ParentID' column that match these identifiers. Define a function to clean and split the data in the 'Point' column. This includes removing brackets and splitting the values into 'x', 'y' and 'z' coordinates. Plot the data using matplotlib. For each unique 'ParentID', draw a separate polyline connecting the 'Point' coordinates. Make sure the first and last points are connected if possible. Set the appropriate labels and headers, ensuring that the x and y axes are equally scaled.

- ❑ The answer LLM will give you a ready-made picture behind which hides the Python code

that generated it:

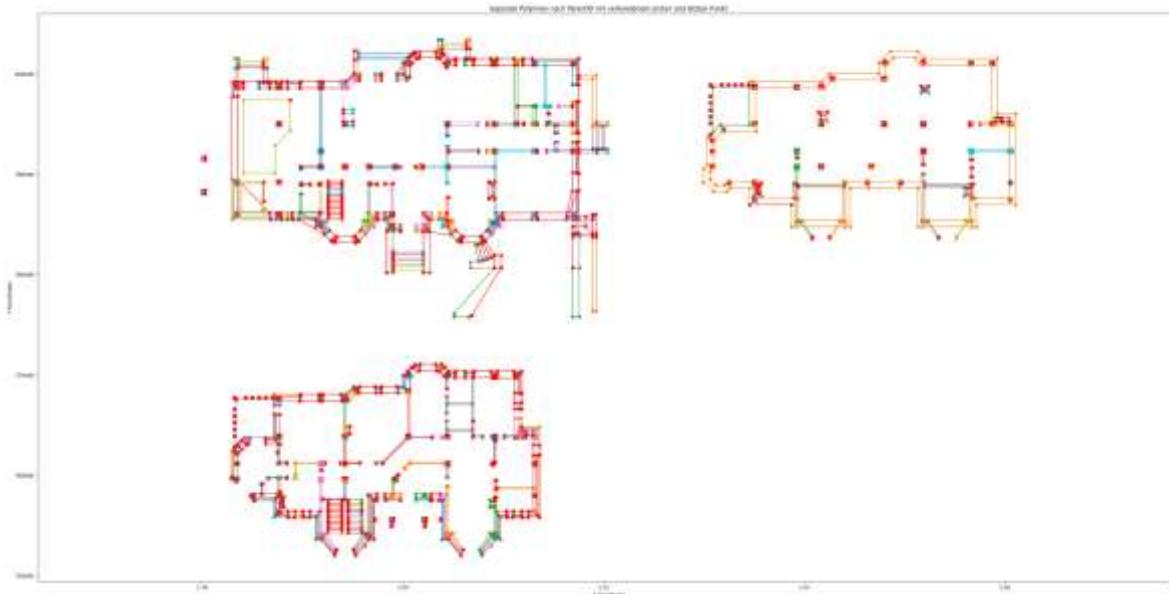


Fig. 6.4-8 LLM code extracted all the lines of the "wall" layer from the DWG -file, cleared their coordinates, and constructed the polylines using one of the Python libraries.

- Now let's add to the lines the area parameter that each polyline has in its properties (in one of the dataframe columns):

Now get just one "ParentID" from each polyline - find that ID in the "ID" column, take the "Area" value, divide by 1,000,000 and add that value to the graph

- The LLM response will show a new graph where each polyline will have a caption with its area:



Fig. 6.4-9 LLM has added code that takes the area values for each polyline and adds it to the image with line visualization.

- Then we will transform each polyline into a horizontal line, add a parallel line at a height of 3000 mm and connect them into a single plane, to show in this way the layout of the surfaces of the wall elements:

You need to take all the elements from the "Layer" column with the value "wall". Take these IDs as a list from the "ID" column and find these IDs from the whole dataframe in the "ParentID" column. All elements are lines that are combined into a single polyline. Each line has a different x, y geometry of the first point in the "Point" column. You must take each polyline in turn and from the point 0,0 horizontally draw the length of each segment from the polyline. the length of each segment of the polyline into one line. Then draw exactly the same lines only 3000 higher, connect all points into one plane.

- The LLM response will output code that allows you to plot wall drawings in the plane:

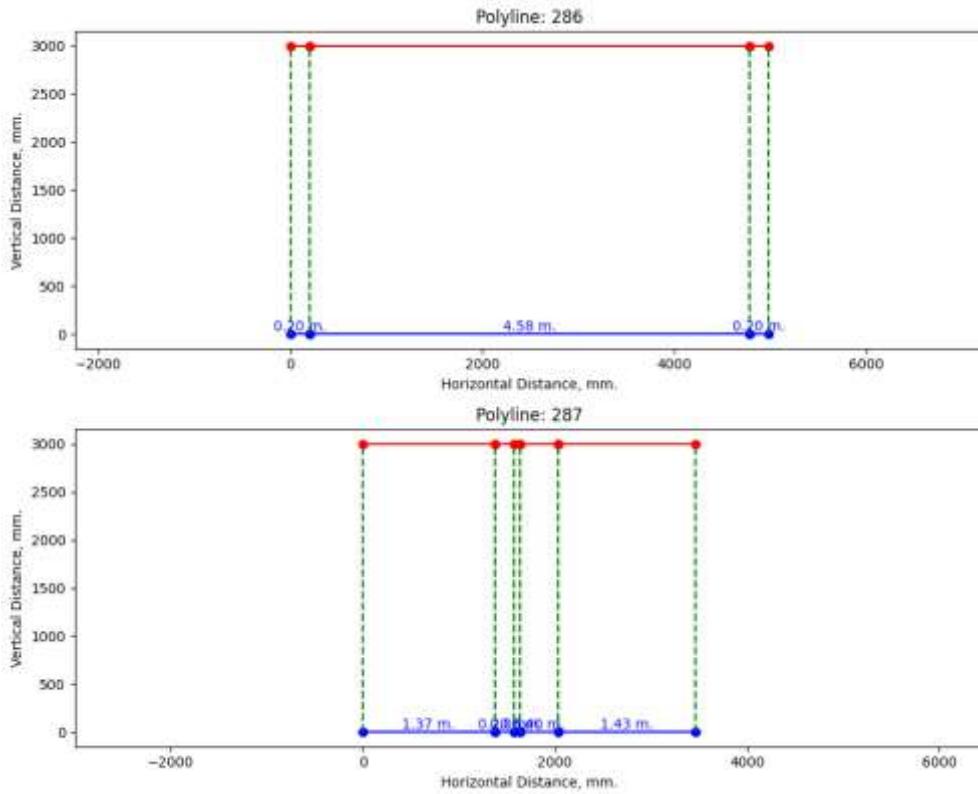


Fig. 6.4-10 We turn each polyline using prompts into a layout that visualizes the wall planes directly in the LLM chat.

- Now let's move from 2D projection to 3D -model of walls from flat lines by connecting upper and lower layers of polylines:

Visualize wall elements in 3D, connecting polylines at heights $z = 0$ and $z = 3000$ mm. To create a closed geometry representing the walls of the building. Use Matplotlib 3D graphing tool.

- LLM will generate an interactive 3D -graph in which each polyline will be represented as a set of planes. The user will be able to move freely between elements with a computer mouse, exploring the model in 3D mode by copying the code from the chat to the IDE:

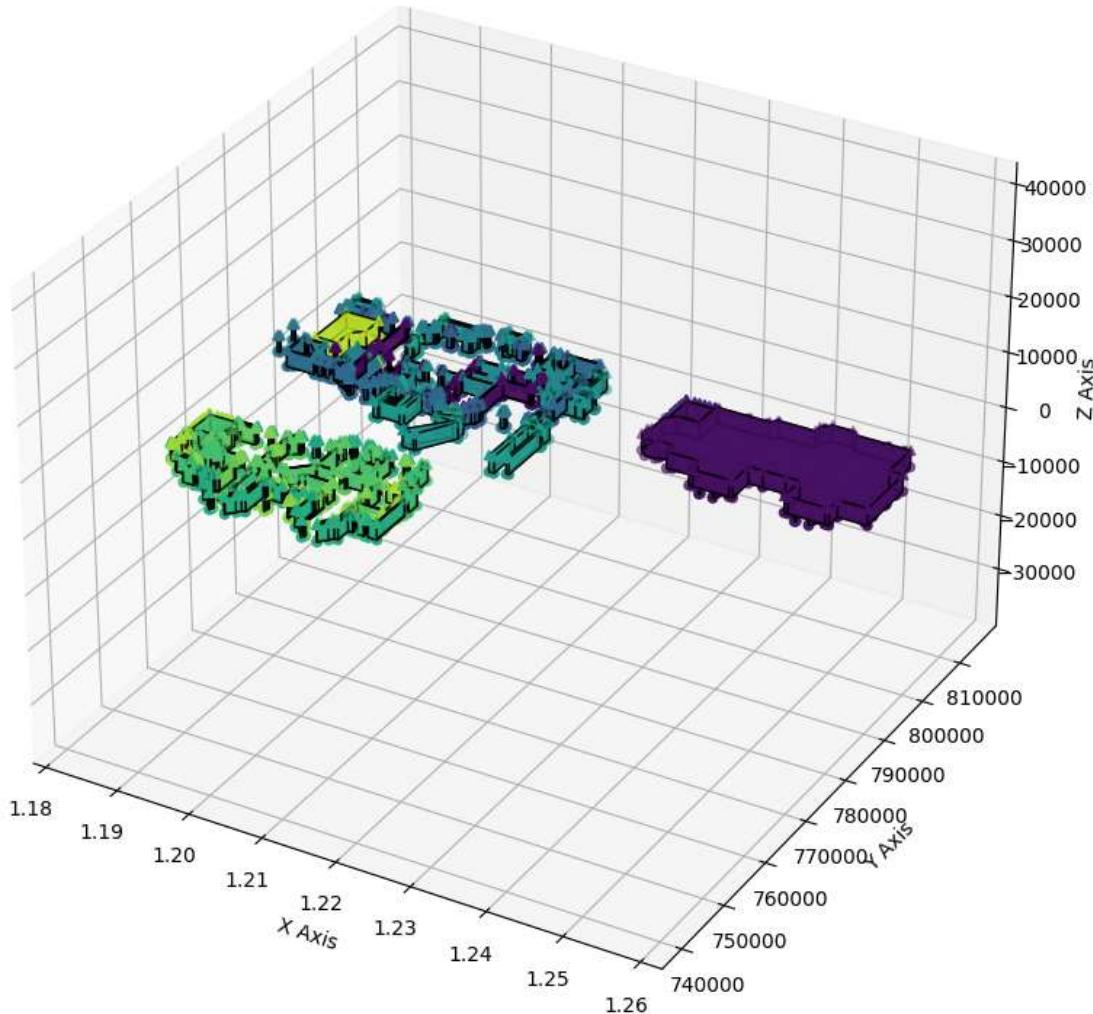


Fig. 6.4-11 LLM helped build code [129] to visualize flat drawing lines into a 3D view that can be explored in the 3D viewer inside the IDE.

To build a logical and reproducible Pipeline - from initial conversion and loading of DWG -file to the final result - it is recommended to copy the generated LLM -block of code to the IDE after each step. In this way, you not only check the result in chat, but also run it in your development environment immediately. This allows you to build the process sequentially, debugging and adapting it as needed.

You can find the complete Pipeline code of all fragments (Fig. 6.4-8 to Fig. 6.4-11) along with sample queries on the Kaggle platform.com by searching for "DWG Analyze with ChatGPT | DataDrivenConstruction" [129]. On Kaggle you can not only view the code and the prompts used, but also copy and

test the entire Pipeline with the original DWG dataframes in the cloud for free without having to install any additional software or the IDE itself.

The approach presented in this chapter allows you to fully automate the checking, processing and generation of documents based on DWG -projects. The developed Pipeline is suitable both for processing individual drawings and for batch processing of dozens, hundreds and thousands of DWG-files with automatic generation of necessary reports and visualizations for each project.

The process can be organized sequentially and transparently: first the data from CAD -file is automatically converted into XLSX format, then loaded into a dataframe, followed by grouping, checking and result generation - all this is realized in a single Jupyter notebook or Python -script, in any popular IDE. If necessary, the process can be easily extended through integration with project documentation management systems: CAD files can be automatically retrieved according to specified criteria, results can be returned back to the storage system and users can be notified when the results are ready - by email or messengers.

Using LLM chats and agents to work with design data reduces dependence on specialized CAD -programs and allows you to perform analysis and visualization of architectural designs without the need for manual interaction with the interface - without mouse clicks and remembering complex menu navigation.

With each passing day, the construction industry will hear more and more about LLM, granular structured data, DataFrames and columnar databases. Unified two-dimensional DataFrames formed from various databases and CAD formats, will be the ideal fuel for modern analytical tools that are actively handled by specialists in other industries.

The automation process itself will be significantly simplified - instead of studying API of closed niche products and writing complex scripts to analyze or transform parameters, now it will be enough to formulate a task in the form of a set of individual text commands, which will be folded into the required Pipeline or Workflow-process for the required programming language, which runs for free on almost any device. No more waiting for new products, formats, plug-ins or updates from CAD- (BIM-) tool vendors. Engineers and builders will be empowered to work independently with data using simple, free and easy-to-understand tools, assisted by LLM chats and agents.

Next steps: moving from closed formats to open data

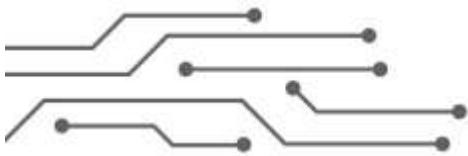
When working with the design data of the future, it is unlikely that anyone really needs to understand the geometric kernels of proprietary tools or learn hundreds of incompatible formats containing the same information. However, without understanding why the move to open structured data is important, it is difficult to argue for the use of new free tools, open data, and approaches that are unlikely to be promoted by software vendors.

In this chapter we have discussed the key features of CAD (BIM) data, their limitations and opportuni-

ties, and that despite the marketing promises of vendors, engineers and designers every day face difficulties in extracting, transferring and analyzing design information. Understanding the architecture of these systems and learning about alternative approaches - based on open formats and automation through LLM - can make life much easier for even a single professional, let alone companies. To summarize this part, it is worth highlighting the main practical steps that will help you apply these approaches to your daily tasks:

- Expand your toolkit for working with project data
 - Explore available plug-ins and utilities to extract data from the CAD - (BIM-) systems you use
 - Explore available SDKs and APIs that allow you to automate data extraction from closed formats without having to manually open specialized software
 - Master basic skills in working with open non-parametric geometry formats (OBJ, glTF, USD, DAE) and corresponding open source libraries
 - Try to think about a system for storing project metadata separate from geometry outside of CAD (BIM) solutions to simplify analysis and integration with other systems
 - Use LLM to automate data conversion issues between formats
- Create your own processes for handling project information
 - Begin to describe tasks and modeling requirements through parameters and their values in simple and structured formats
 - Create a personal library of scripts or code blocks for frequently performed operations
- Promote the use of open standards in your work
 - Invite colleagues and partners to share data in open formats that are not restricted by the software vendor ecosystem
 - Demonstrate the benefits of using structured data with specific examples
 - Initiate discussions about problems with closed formats and possible solutions

Even if you cannot change your company's policy regarding CAD - (BIM-) platforms, a personal understanding of the principles of working with project data in open formats will allow you to significantly increase the efficiency of your work. By creating your own tools and methods for extracting and transforming data from different formats, you not only optimize your workflows, but also gain the flexibility to bypass the limitations of standard software solutions.



VII PART

DATA-DRIVEN DECISION-MAKING, ANALYTICS, AUTOMATION AND MACHINE LEARNING

The seventh part is dedicated to data analytics and process automation in the construction industry. It discusses how data becomes the basis for decision-making and explains the principles of visualizing information for effective analysis. Key performance indicators (KPI), methods for evaluating return on investment (ROI) and creating dashboards for project monitoring are described in detail. Special attention is given to ETL processes (Extract, Transform, Load) and their automation using pipelines (Pipeline) to turn disparate data into structured information for analysis. Workflow orchestration tools such as Apache Airflow, Apache NiFi and n8n, which allow building automated data pipelines without deep programming knowledge, are discussed. Large Language Models (LLMs) and their use to simplify data analysis and automate routine tasks are playing a significant role

CHAPTER 7.1.

DATA ANALYTICS AND DATA-DRIVEN DECISION-MAKING

After the steps of collecting, structuring, cleaning and verifying the information, a coherent and analyzable data set has been formed. The previous parts of the book covered the systematization and structuring of heterogeneous sources - from PDF documents and text records of meetings to CAD models and geometric data. The process of checking and aligning information with the requirements of various systems and classifiers, eliminating duplicates and inconsistencies is described in detail.

All the calculations performed on this data (third, fourth parts of the book) - from simple transformations to calculations of time, cost, and ESG metrics (fifth part) - are aggregated analytics tasks. They form the basis for understanding the current state of a project, evaluating its parameters, and then making decisions. As a result, the data, as a result of calculations, turns from a set of disparate records into a manageable resource capable of answering key business questions.

The previous chapters have detailed the data collection and quality control processes for use in typical business cases and processes specific to the construction industry. Analytics in this context is in many ways similar to applications in other industries, but it also has a number of specific features.

In the following chapters, an enlarged data analysis process will be discussed in detail, including the stages of automation - from the initial acquisition of information and its transformation to its subsequent transfer to target systems and documents. First, a theoretical part will be presented, focusing on selected aspects of data-driven decision making. Then, in the following chapters, the practical part related to automation and building ETL -Pipeline will begin.

Data as a resource in decision making

Data-driven decision-making is often an iterative process and begins with the systematic collection of information from a variety of information sources. Like nature's cycle, individual data elements and entire information systems gradually fall into the soil - accumulating in companies' information repositories (Fig. 1.3-2). Over time, this data, like fallen leaves and branches, is transformed into valuable material. The mycelium of data engineers and analysts organizes and prepares information for future use and turns fallen data and systems into valuable compost, to grow new shoots and new systems (Fig. 1.2-5).

Trends in the widespread use of analytics in various industries, marks the beginning of a new era, where working with data becomes the basis of professional activity (Fig. 7.1-1). It is important for construction industry professionals to adapt to these changes and be prepared to move into a new era - the era of data and analytics

Manually moving data between tables and performing calculations manually are gradually becoming a thing of the past, giving way to automation, data flow analysis, analytics and machine learning. These tools are becoming key elements of modern decision support systems.

In McKinsey's book "Rebooting. McKinsey's Guide to Overcoming Competition in the Age of Digital Technology and Artificial Intelligence" [130] cites a study of 1,330 senior executives from various regions, industries, and functional areas in 2022 [130], cites a study conducted in 2022 with 1,330 senior executives from different regions, industries, and functional areas. According to its results, 70% of leaders use advanced analytics to generate their own ideas, and 50% implement artificial intelligence to improve and automate decision-making processes.



Fig. 7.1-1 Data analysis and analytics is the main tool to increase the speed of decision making in a company.

Data analytics, like the spreading of mycelium, penetrates the humus of past decisions, helping to connect individual systems and guiding managers to valuable insights. This knowledge, like nutrients from decayed data system trees, feeds new decisions in the company, leading to effective change and quality information growth, like new shoots and sprouts emerging from rich and healthy soil (Fig. 1.2-5).

Numbers have an important story to tell. They are counting on you to give them a clear and compelling voice [131].

— Stephen Few, Data Visualization Expert

In medium-sized and small companies, the work of extracting and preparing information for further analysis is today an extremely labor-intensive process (Fig. 7.1-2), comparable to eighteenth-century coal mining. Until recently, the work of data mining and preparation was rather reserved for adventurers working in a highly specialized niche with a small and limited set of tools for working with different

types of data from unstructured, loosely structured, mixed and closed sources.

Decision makers and managers are often inexperienced with heterogeneous data and systems, yet need to make data-driven decisions. As a result, data-driven decision making in the modern construction industry over the past decades has felt less like an automated process and more like the multi-day manual labor of a miner in the early coal mines.

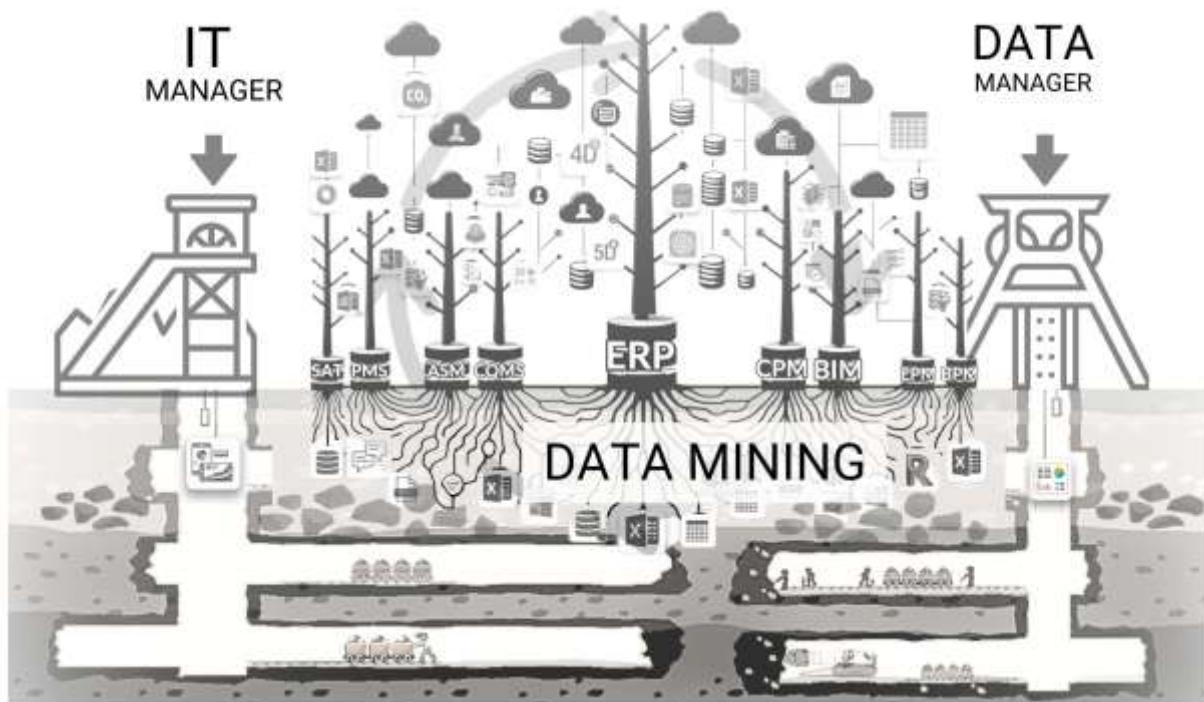


Fig. 7.1-2 In the data mining process, experts go through a complex path of data preparation - from cleaning to structuring for further analytics.

While modern methods of data extraction in the construction industry are certainly more advanced than the primitive techniques of 12th century miners, it is still a complex and high-risk task that requires significant resources and expertise that only large companies could afford. The processes of extracting and analyzing data from the accumulated legacy of past projects have until recently been predominantly handled by large, technologically advanced companies that have been collecting and storing data consistently for decades.

Previously, the leading role in analytics was played by technologically mature companies that had been accumulating data for decades. Today, the situation is changing: access to data and data processing tools is becoming democratic - previously complex solutions are now available to everyone for free.

Applying analytics allows companies to make more accurate and informed decisions in real time. The

following case study illustrates how historical data can help make financially sound decisions:

- ⌚ **Project Manager** - "Now the average price of concrete in the city is 82€ /m³, we have 95 €/m³ in the estimate."
- ⌚ **Estimator** - "On previous projects the cost overrun was about 15%, so I backed up."
- ⌚ **Data manager or customer-side control engineer** - "Let's look at the analytics for the last three tenders."

After analyzing DataFrames from past projects, we get:

- **Average actual purchase price:** 84.80 /m³€
- **Average overspend ratio:** +4.7%
- ⌚ **Recommended rate in the estimate:** ~ 85 /m³€

Such a decision will no longer be based on subjective feelings, but on specific historical statistics, which helps to reduce risks and increase the validity of the tender bid. Analysis of data from past projects becomes a kind of "organic fertilizer" from which new, more accurate solutions germinate.

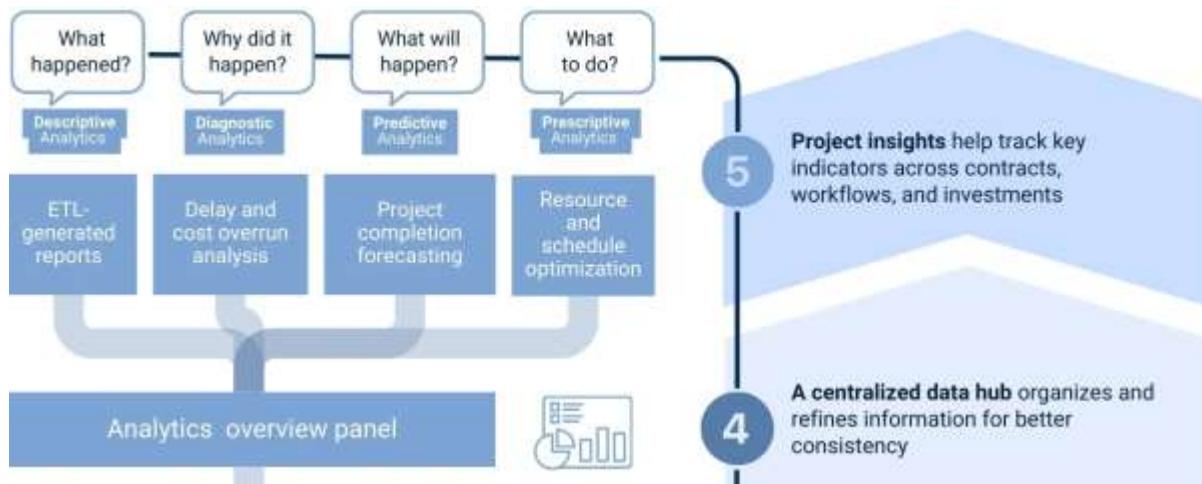


Fig. 7.1-3 Data analytics answers three key questions: what happened, why it happened, and what should be done next.

Decision makers and managers are often faced with the need to work with heterogeneous data and systems with little technical background. In such situations, visualization - one of the first and most important steps in the analytical process - becomes one of the key aids in understanding data. It allows to present information in a visual and understandable form.

Visualizing data: the key to understanding and decision making

In today's construction industry, where project data is characterized by complexity and multi-level structure, visualization plays a key role. Visualization of data allows project managers and engineers to visualize complex patterns and trends hidden in large, heterogeneous volumes of data.

Visualizing data makes it easier to understand the status of a project: resource allocation, cost trends or material usage. Graphs and charts make complex and dry information accessible and understandable, allowing you to quickly identify key areas that need attention and identify potential problems.

Visualization of data not only facilitates the interpretation of information, it is a critical step in the analytical process and informed management decision-making, helping *to answer the questions "what happened?" and "how did it happen?"* (Fig. 2.2-5).

Graphics are visual tools for solving logical problems [132].

— Jacques Bertin, "Graphics and Graphical Information Processing"

Before making key decisions, project managers are more likely to use visual representations of data rather than dry and difficult to interpret numbers from spreadsheets or text messages.

Data without visualization is like building materials strewn haphazardly on a construction site: their potential is unclear. It is only when they are visualized clearly, like bricks and concrete in a house, that their value becomes clear. Until the house is built, it is impossible to say whether the pile of materials will become a small hut, a luxury villa or a skyscraper.

Companies have data from various systems (Fig. 1.2-4 to Fig. 2.1-10), financial transactions, and extensive text data. However, utilizing this data for business benefit is often challenging. In such situations, visualization becomes an important tool for communicating the meaning of the data, which helps present the information in formats that any expert can understand, such as dashboards, graphs, and charts.

PwC's study "What Students Need to Succeed in a Fast-Changing Business World" (2015) emphasizes [9] that successful companies go beyond data analysis and actively use interactive visualization tools such as graphs, infographics, and analytical dashboards to support decision making. According to the report - data visualization helps customers understand the story that data tells through charts, graphs, diagrams, dashboards and interactive data models.

The process of converting information into visual graphic forms such as charts, graphs, and diagrams improves the human brain's understanding and interpretation of data (Fig. 7.1-4). This allows project managers and analysts to more quickly assess complex scenarios and make informed decisions based not on intuition, but on visually recognizable trends and patterns.

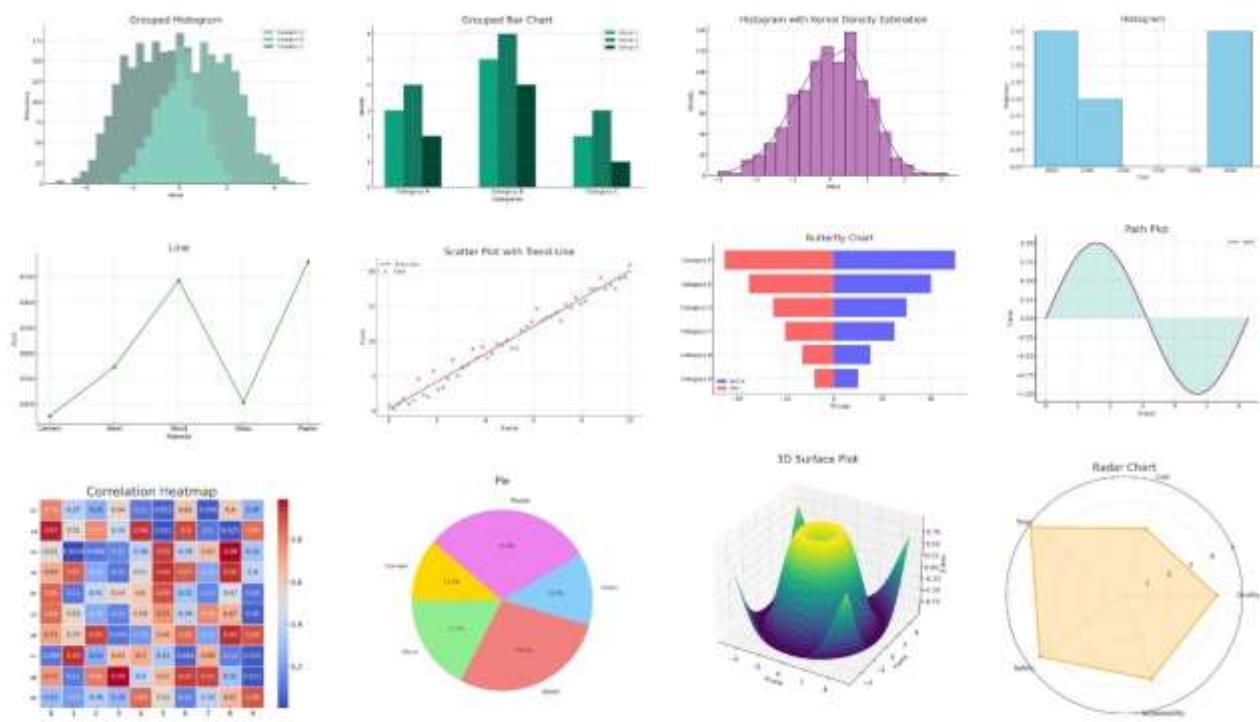


Fig. 7.1-4 Various types of visualizations are designed to help the human brain better understand and make sense of the dry information of numbers.

The issues of creating visualizations from data, and the use of various free visualization libraries, will be discussed in more detail in the next chapter on ETL-processes.

Visualization is becoming an integral element of working with data in the construction industry - it helps not only to "see" data, but also to understand its meaning in the context of management tasks. However, for visualization to be truly useful, it is necessary to determine in advance what exactly should be visualized and what metrics are really important for assessing project performance. This is where performance metrics such as KPIs and ROI come into play. Without them, even the most beautiful dashboards run the risk of being just "information noise".

KPIs and ROI

In today's construction industry, the management of performance indicators (KPI) and their visualization through reports and dashboards play a key role in improving productivity and project management efficiency.

As in any business, in construction, it is necessary to clearly define the metrics by which success, return on investment and performance are measured. When obtaining data on various processes, a data-driven organization must first learn to identify key **KPIs (Key Performance Indicators)** - quantitative measures that reflect the extent to which strategic and operational goals are being achieved.

To calculate a KPI, a formula is usually used (Fig. 7.1-5) that includes actual and planned indicators. For example, to calculate an individual KPI for a project, employee or process, divide the actual performance by the planned performance and multiply the result by 100%.

$$\text{index KPIs} = \frac{\text{actual performance}}{\text{target performance}} \times 100$$

Fig. 7.1-5 KPIs are used to measure the success of a project or process in achieving key objectives.

At the construction site level, more detailed KPI metrics can be used:

- **Timing of key milestones** (foundation, installation, finishing) - allows you to monitor compliance with work plans.
- **Material Overrun Percentage** - helps to manage purchases and minimize wastage.
- **Number of unscheduled machine downtimes** - affects productivity and costs.

Choosing the wrong metrics can lead to erroneous "what to do?" decisions (Fig. 2.2-5). For example, if a company focuses only on cost per square meter but does not consider the cost of remodeling, savings on materials can lead to poorer quality and higher costs in future projects.

When setting objectives, it is important to be clear about what is being measured. Vague wording leads to incorrect conclusions and complicates control. Let's look at examples of successful and unsuccessful KPIs in construction.

Good KPIs:

- ❑ "By the end of the year, reduce the percentage of reworked finishes by 10%."
- ❑ "Increase the speed of façade installation by 15% without compromising quality by next quarter"

- ❷ "Reduce equipment downtime by 20% by optimizing work schedules by year-end"

These metrics are clearly measurable, have specific values and timeframes.

Bad KPIs:

- ❷ "We will build faster" (How much faster? What does "faster" mean?)
- ❷ "We will improve the quality of concrete work" (How exactly is quality measured?)
- ❷ "We will improve contractor interaction at the site" (What criteria will show improvement?)

A good KPI is one that can be measured and objectively assessed. In construction, this is especially important, because without clear indicators it is impossible to monitor performance and achieve stable results.

In addition to KPI, there is an additional metric for assessing the effectiveness of investments: **ROI (Return on Investment)** - a return on investment indicator that reflects the ratio between profit and invested funds. ROI allows you to assess whether the implementation of new methods, technologies or tools is justified: from digital solutions and automation (e.g. Fig. 7.3-2) to the use of new building materials. This indicator helps to make informed decisions about further investments based on their real impact on business profitability

In the context of construction project management, ROI (return on investment) can be used as one of the key performance indicators (KPI) if a company's goal is to measure the return on investment of a project, technology or process improvement. For example, if a new construction management technique is being implemented, ROI can show how much it has improved profitability.

Regularly measuring KPIs and ROIs based on data collected from various sources, such as material consumption, labor hours and costs, allows project management to effectively manage resources and make quick decisions. Storing this data for the long term allows for analyzing future trends and optimizing processes.

Various charts and graphs are used to visualize KPIs, ROIs, and other metrics and are usually combined into dashboards.

Dashboards and dashboards: visualization of indicators for effective management

A variety of charts and graphs are used to visualize indicators and metrics, which are typically combined into data showcases and dashboards. These dashboards provide a centralized view of the status of a project or parts of a project, displaying key indicators (ideally in real time). Up-to-date and continuously updated dashboards allow the team to respond quickly to changes.

Dashboards are tools that visualize quantitative assessments, making them easily accessible and understandable to all project participants.

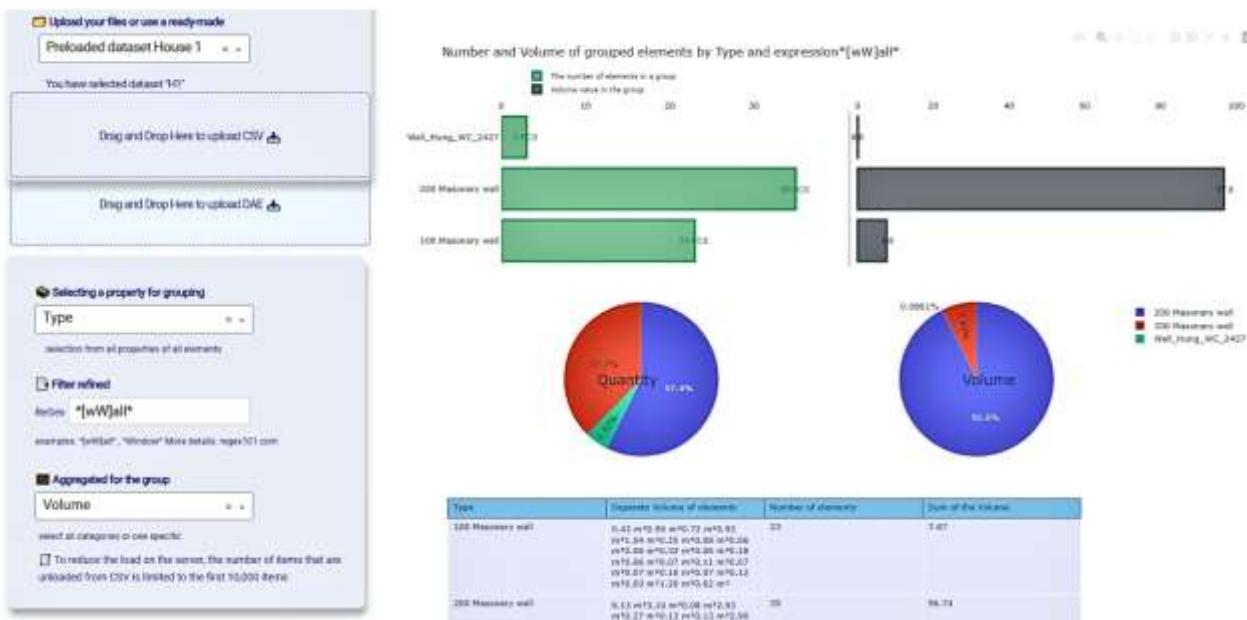


Fig. 7.1-6 Managing KPIs and visualizing them through dashboards is key to improving project productivity and efficiency.

Here are some examples of popular tools in which you can create dashboards:

- **Power BI** is a tool from Microsoft for creating interactive reports and dashboards.
- **Tableau and Google Data Studio** are powerful tools for visualizing data and creating dashboards without having to write code.
- **Plotly** (Fig. 7.1-6, Fig. 7.2-12) is a library for creating interactive graphs, and Dash is a framework for creating web applications for data analysis. They can be used in combination to create interactive dashboards.
- **Multiple Python libraries** (Fig. 7.2-9 - Fig. 7.2-11) - Python has many open source and free libraries for data visualization, such as Matplotlib, Seaborn, Plotly, Bokeh and others. These can be used to create graphs and integrate them into a web application using frameworks such as Flask or Django.
- **JavaScript Libraries**: allow you to create interactive dashboards using Open Source JavaScript libraries such as D3.js or Chart.js and integrate them into a web application.

To evaluate KPIs and create dashboards, you need up-to-date data and a clear timeline for collecting and analyzing information.

In general, KPI, ROI and dashboards in the construction industry form the basis for an analytical approach to project management. They not only help to monitor and evaluate the current status, but also provide valuable insights for future planning and process optimization - processes that depend directly on interpreting data and asking the right and timely questions.

Analyzing data and the art of asking questions

Data interpretation is the final stage of analysis, where information makes sense and begins to "speak". This is where the answers to the key questions are formulated: "*what to do?*" and "*how to do?*" (Fig. 2.2-5). This stage allows summarizing results, identifying patterns, establishing cause-and-effect relationships and drawing conclusions based on visualization and statistical analysis.

Perhaps the time is not far off when the realization will come that to fully become an effective citizen of one of the new great complex new world states that are now developing, it is as necessary to be able to calculate, to think in terms of averages, maxima and minima, as it is now necessary to be able to read and write [133].

— Samuel S. Wilkes, quoted in a 1951 presidential address to the American Statistical Association

According to the report "Data Analytics and Artificial Intelligence in the Implementation of Government Projects" (2024) published by the UK government [83], the implementation of analytics data and artificial intelligence (AI) can significantly improve project management processes, increasing the accuracy of time and cost forecasting, as well as reducing risk and uncertainty. The paper emphasizes that public organizations that use advanced analytical tools achieve higher performance in infrastructure initiatives.

The modern construction business operating in the highly competitive and low margin environment of the fourth industrial revolution can be compared to military operations. Here, the company's survival and success depend on the speed of obtaining resources and quality information - and thus on timely and informed decision-making (Fig. 7.1-7).

If data visualization is the "intelligence" that provides the overview, then data analytics is the "ammunition" needed for action. It answers the questions: *what to do?* and *how to do it?*, forming the basis for gaining a competitive advantage in the market.

Analytics turns disparate data into structured and meaningful information on which to base decisions.

The task of analysts and managers is not just to interpret information, but to offer informed decisions,

identify trends, determine relationships between different types of data and classify them in accordance with the goals and specifics of the project. Using visualization tools and statistical analysis methods, they turn data into a strategic asset for the company.



Fig. 7.1-7 It is the data analysis that ultimately turns the information collected into a source for decision making.

In order to make truly informed decisions in the analytics process, it is necessary to learn how to correctly formulate the questions that are asked of the data. The quality of these questions directly affects the depth of insights gained and, as a consequence, the quality of management decisions.

The past exists only insofar as it is present in the records of today. And what these records represent is determined by the questions we ask. There is no other history than this one [134].

— John Archibald Wheeler, physicist 1982.

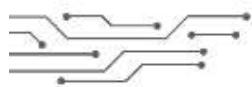
The art of asking deep questions and thinking critically is a critical skill in working with data. Most people tend to ask simple, superficial questions that require little effort to answer. However, true analysis begins with meaningful and thoughtful questions that can uncover hidden relationships and cause-and-effect relationships in information that may be hidden behind multiple layers of reasoning.

According to the study "Data-Driven Transformation: Accelerating at Scale Now" (BCG, 2017) [135], successful digital transformation requires investments in analytic capabilities, change management programs, and alignment of business goals with IT initiatives. Companies that create a data-driven culture should invest in data analytics capabilities and launch change management programs to instill new thinking, behaviors, and ways of working.

Without investment in developing an analytical culture, improving data tools and training professionals, companies will continue to risk making decisions based on outdated or incomplete information - or relying on the subjective opinions of HiPPO managers (Fig. 2.1-9).

Realizing the relevance and the need to constantly update analytics and dashboards inevitably leads management to understand the importance of automating analytical processes. Automation increases the speed of decision-making, reduces the influence of the human factor and ensures data relevance. With the exponential growth of information volumes, speed becomes not just a competitive advantage, but a key factor for sustainable success.

Automation of data analysis and processing processes in general is inextricably linked to the topic of ETL (Extract, Transform, Load). Just as in the automation process we need to transform data, in the ETL process data is extracted from various sources, transformed according to the necessary requirements and loaded into target systems for further use.



CHAPTER 7.2.

DATA FLOW WITHOUT MANUAL EFFORT: WHY ETL

ETL automation: lower costs and faster data handling

When key performance indicators (KPI) stop growing, despite the increase in data volumes and team size, company management inevitably comes to the realization of the need to automate processes. Sooner or later this realization becomes an incentive to launch complex automation, the main goal of which is to reduce the complexity of processes, speed up processing and reduce dependence on the human factor.

According to McKinsey's study "How to Build a Data Architecture to Drive Innovation - Today and Tomorrow" (2022) [136], companies using streaming data architectures gain a significant advantage because they can analyze information in real time. Streaming technologies allow direct analysis of real-time messages and the application of predictive maintenance in manufacturing through real-time analysis of sensor data.

Process simplification is automation, where traditional manual functions are replaced by algorithms and systems.

The issue of automation, or rather "minimizing the role of humans in data processing", is an irreversible and highly sensitive process for every company. Specialists in any professional field are often hesitant to fully disclose their methods and subtleties of work to fellow optimizers, realizing the risk of losing their jobs in a rapidly evolving technological environment.

If you want to make enemies, try to change things [137].

— Woodrow Wilson, speech to a convention of salesmen, Detroit, 1916

Despite the obvious benefits of automation, many companies still have a high proportion of manual labor in their daily practices, especially in the area of engineering data. To illustrate the current situation, let's look at a typical example of sequential data processing within such processes.

Manual data handling can be illustrated by the example of interaction with information obtained from CAD databases. Traditional data processing ("manual" ETL -process) in CAD (BIM) departments for creating attribute tables or creating documentation based on design data takes place in the following order (Fig. 7.2-1):

1. Manual **Extract**: the user manually opens the project - by launching the CAD application (BIM) (Fig. 7.2-1 step 1).

2. **Verification:** the next step usually involves manually running several plug-ins or auxiliary applications for data preparation and quality assessment (Fig. 7.2-1 step 2-3).
 3. **Manual transformation (Transform):** after preparation, data processing begins, which requires manual operation of various software tools in which the data are prepared for upload (Fig. 7.2-1 step 4).
 4. **Manual upload (Load):** manual upload of converted data to external systems, data formats and documents (Fig. 7.2-1 step 5).

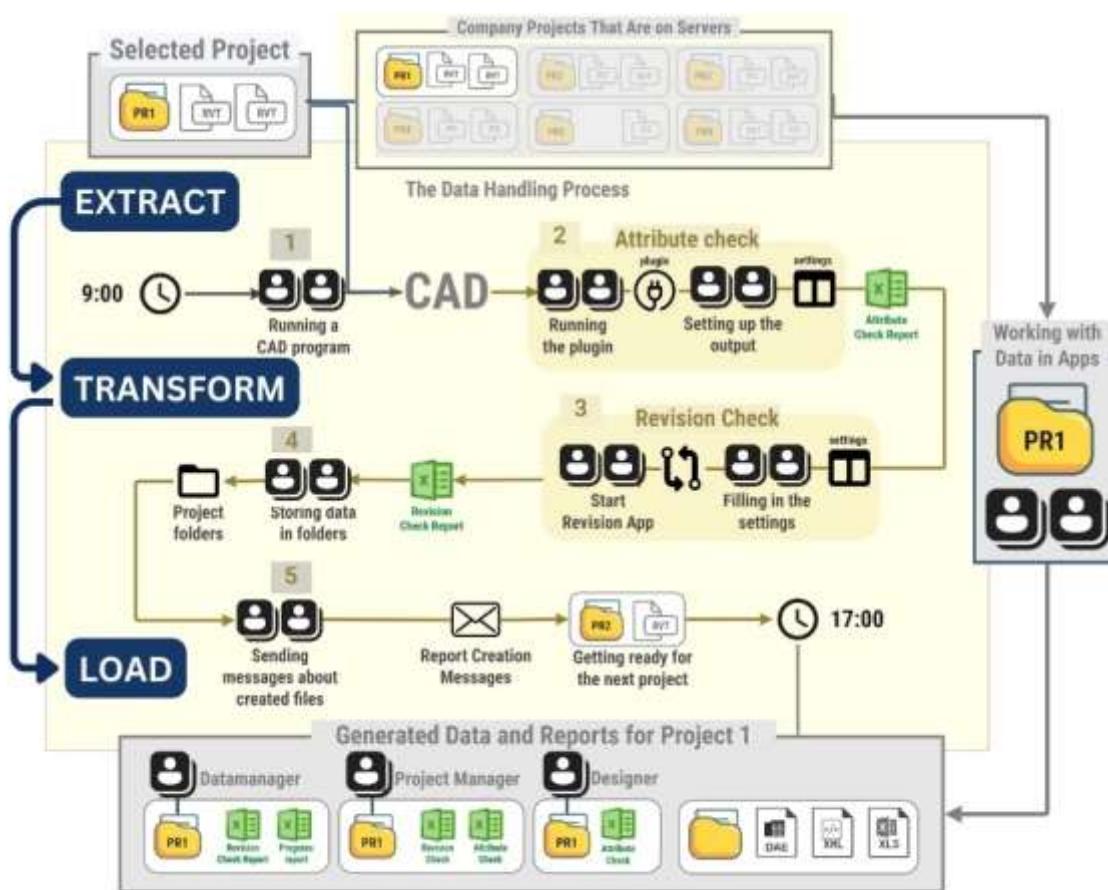


Fig. 7.2-1 Traditional manual ETL processing is limited by the desires and physical capabilities of the individual technician.

Such a workflow is an example of a classic ETL -process - extraction, transformation and loading (ETL). Unlike other industries where automatic ETL pipelines have long been the standard, the construction industry is still dominated by manual labor, which slows down processes and increases costs

ETL (Extract, Transform, Load) is the process of extracting data from various sources, transforming it into the desired format and loading it into the target system for further analysis and use.

ETL is a process that stands for three key components of data processing: Extract, Transform, and Load (Fig. 7.2-2):

- **Extract** - extract data from different sources (files, databases, API).
- **Transform** - data cleaning, aggregation, normalization and logical processing.
- **Load** - load structured information into a data warehouse, report, or BI system.

Earlier in the book, the concept of ETL was touched upon only occasionally: in the conversion of an unstructured scanned document into a structured tabular format (Fig. 4.1-1), in the context of formalizing requirements to systematize the perception of both life and business processes (Fig. 4.4-20), and in the automation of data validation and data processing from CAD solutions. Let us now look at ETL in more detail in the context of typical workflows.

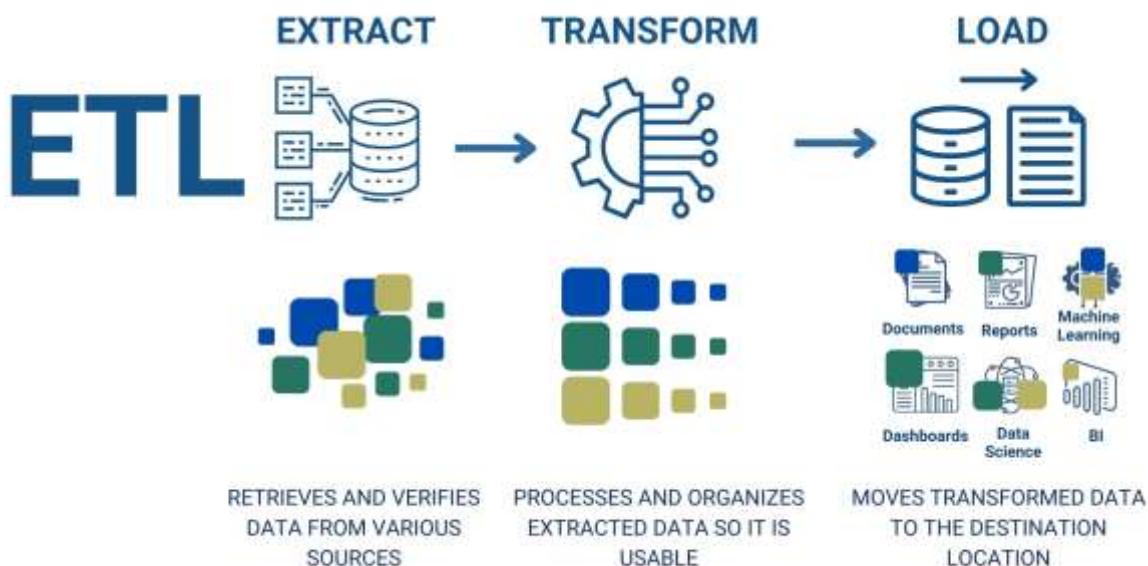


Fig. 7.2-2 ETL automates repetitive data processing tasks.

Manual or semi-automated ETL-process implies a manager or technician who manages all steps manually - from data collection to report generation. Such a process takes a significant amount of time, especially when working hours are limited (e.g. 9:00 to 17:00).

Companies often seek to solve the problem of low efficiency and slow speed by purchasing modular integrated solutions (ERP, PMIS, CPM, CAFM, etc.), which are then customized by external vendors and consultants. But these vendors and third-party developers often become a critical dependency point: their technical limitations directly affect the performance of the entire system and the business as a whole, as detailed in previous chapters on proprietary systems and formats. The problems created by fragmentation and dependency were discussed in detail in the chapter "How Construction Businesses Drown in Data Chaos".

If a company is not ready to implement a large modular platform from one of the vendors, it starts looking for alternative ways of automation. One of them is to develop their own modular open ETL -conveyors, where each stage (extraction, transformation, validation, loading) is implemented as scripts executed on a schedule.

In the automated version of the same ETL workflow (Fig. 7.2-1), the work process looks like a modular code that starts with processing data and translating it into an open structured form. Once the structured data is received, various scripts or modules are run automatically, on a schedule, to check changes, transform and send messages (Fig. 7.2-3).

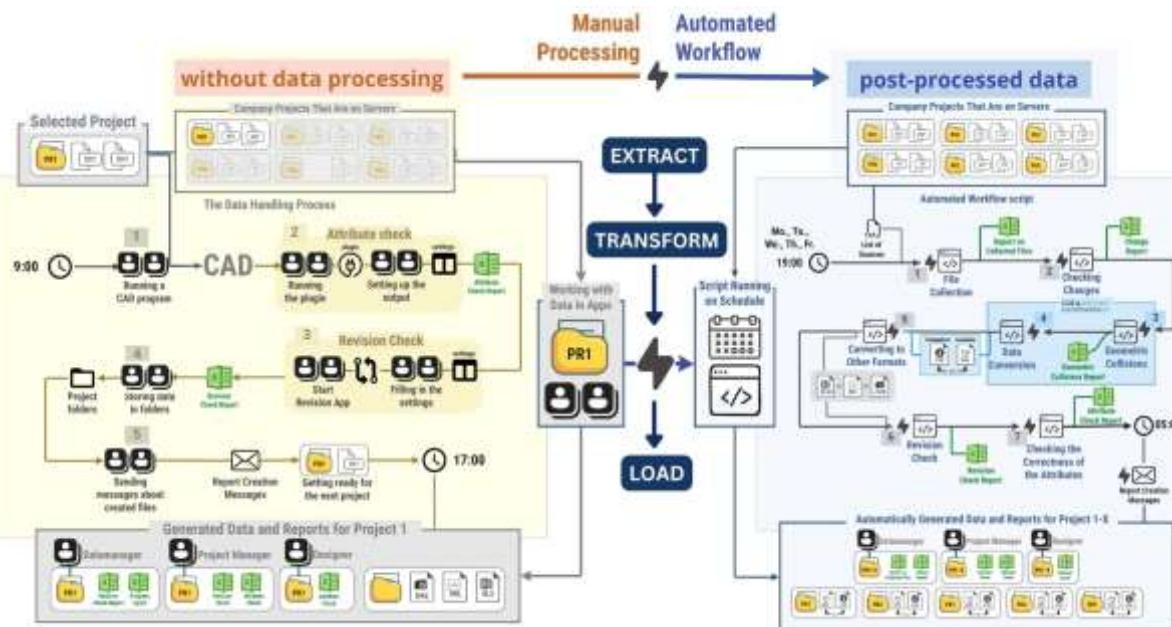


Fig. 7.2-3 On the left is manual machining, on the right is an automatic process that, unlike traditional manual machining, is not limited by user capabilities.

In an automated workflow, data processing is simplified by ET(L) data preprocessing: structuring and unification.

In traditional processing methods, specialists work with data "as is" - as it is retrieved from systems or software. In automated processes, by contrast, data often first pass through an ETL-Pipeline, where it is brought to a consistent structure and format suitable for further use and analysis.

Let's take a practical ETL example, demonstrating the data table validation process described in the chapter "Validating data and validation results" (Fig. 4.4-13). To do this, we use the Pandas library in conjunction with the LLM for automated data analysis and processing processes.

ETL Extract: data collection

The first stage of the ETL process - Extract) - starts with writing code to collect data sets to be further checked and processed. To do this, we scan all the folders of the production server, collect documents of a certain format and content, and then convert them into a structured form. This process is discussed in detail in the chapters "Converting unstructured and textual data into structured form" and "Converting CAD data (BIM) into structured form" (Fig. 4.1-1 - Fig. 4.1-12).

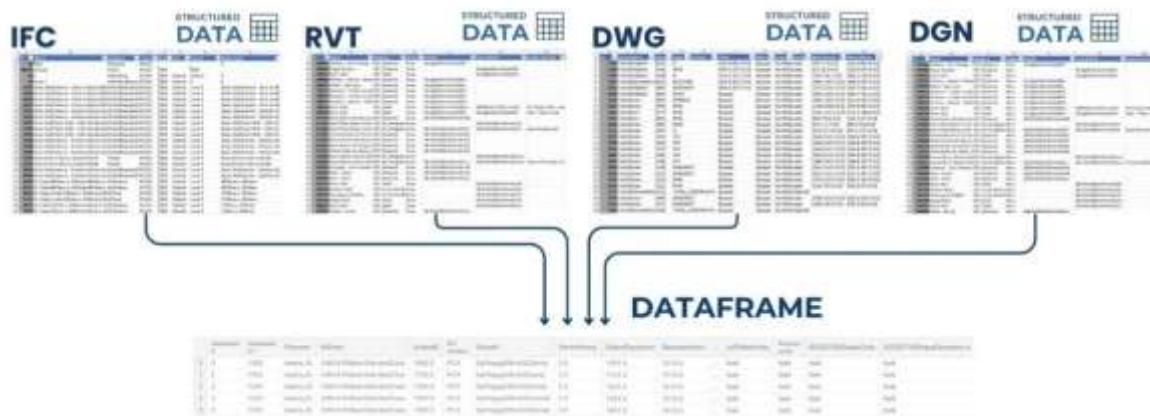


Fig. 7.2-4 Convert CAD data (BIM) into one large data frame that will contain all project sections.

As an illustrative example, we use the Extract data loading step and obtain a table of all CAD- (BIM-) projects (Fig. 7.2-4) uses reverse engineering-enabled converters [138] for RVT and IFC formats to obtain structured tables from all projects and combine them into one large DataFrame table.

```

1 import os
2 import subprocess
3 import time
4 import pandas as pd
5
6 path_conv = r'C:\DDC_2023\' # Where RvtExporter.exe|IfcExporter.exe is located
7 path = r'C:\IFCprojects\' # Where Revit|IFC files are stored
8
9 def convert_and_wait(path_conv, exporter_name, file_path, extension):
10     # Start the conversion process
11     subprocess.Popen([os.path.join(path_conv, exporter_name),
12                      file_path], cwd=path_conv)
13     output_file = os.path.join(path,
14                               f"{os.path.splitext(os.path.basename(file_path))[0]}_{extension}.xlsx")
15
16 # Conversion process for RVT and IFC files
17 for file in os.listdir(path):
18     full_path = os.path.join(path, file)
19     if file.endswith('.ifc'):
20         convert_and_wait(path_conv, 'IfcExporter.exe', full_path, 'ifc')
21     elif file.endswith('.rvt'):
22         convert_and_wait(path_conv, 'RvtExporter.exe', full_path, 'rvt')
23
24 # Combine converted Excel files into one dataframe
25 df = pd.concat([pd.read_excel(os.path.join(path, f)) for f in os.listdir(path)
26 if f.endswith('.xlsx')], ignore_index=True)

```

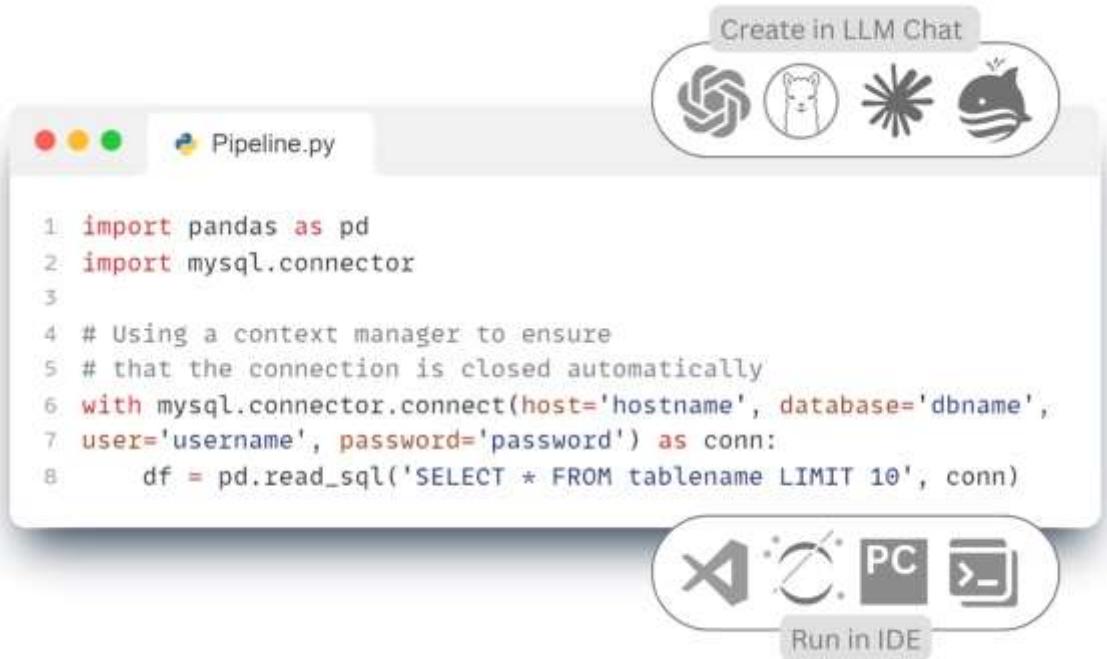
Fig. 7.2-5 Converting using Python code and SDK reverse engineering tool for RVT and IFC files into one large structured (df) DataFrame.

Pandas DataFrame can load data from a variety of sources, including CSV text files, Excel spreadsheets, JSON - and XML - files, big data storage formats such as Parquet and HDF5, and from MySQL, PostgreSQL, SQLite, Microsoft SQL Server, Oracle and other databases. In addition, Pandas supports loading data from APIs, web pages, cloud services and storage systems such as Google BigQuery, Amazon Redshift and Snowflake.

- ☞ To write code to connect and collect information from databases, send a similar text request to the LLM chat room (CHATGP, LLaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

Please write an example of connecting to MySQL and converting data to

💡 LLM's response:



The screenshot shows a user interface for generating code. At the top right is a button labeled "Create in LLM Chat" with four icons: a swirl, a llama, a star, and a whale. Below this is a code editor window titled "Pipeline.py". The code inside is:

```

1 import pandas as pd
2 import mysql.connector
3
4 # Using a context manager to ensure
5 # that the connection is closed automatically
6 with mysql.connector.connect(host='hostname', database='dbname',
7     user='username', password='password') as conn:
8     df = pd.read_sql('SELECT * FROM tablename LIMIT 10', conn)

```

At the bottom right is a "Run in IDE" button with icons for VS Code, PyCharm, and others.

Fig. 7.2-6 Example of connecting via Python to a MySQL database and importing data from the MySQL database into a DataFrame.

The resulting code (Fig. 7.2-5, Fig. 7.2-6) can be run in one of the popular IDEs (integrated development environments) we mentioned above in offline mode: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse with PyDev plugin, Thonny, Wing IDE, IntelliJ IDEA with Python plugin, JupyterLab or popular online tools: Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

By loading the multiformat data into the variable "df" (Fig. 7.2-5 - row 25; Fig. 7.2-6 - row 8), we converted the data to the Pandas DataFrame format, one of the most popular structures for data processing, which is a two-dimensional table with rows and columns. We will talk more about other storage formats used in ETL -Pipelines such as Parquet, Apache ORC, JSON, Feather, HDF5, and modern data warehouses in the chapter "Data Storage and Management in the Construction Industry" (Fig. 8.1-2).

After the stage of data extraction and structuring (Extract), a single array of information is formed (Fig. 7.2-5, Fig. 7.2-6), ready for further processing. However, before loading this data into target systems or using it for analysis, it is necessary to ensure its quality, integrity and compliance with the specified requirements. It is at this stage that data transformation (Transform) is performed, a key step that

ensures the reliability of subsequent conclusions and decisions.

ETL Transform: application of validation and transformation rules

The Transform step is where the data is processed and transformed. This process may include correctness checking, normalization, filling in missing values and validation using automated tools

According to the PwC study "Data-Driven. What Students Need to Succeed in a Rapidly Changing Business World" (2015) [9], modern auditing companies are moving away from random data verification and are moving to analyzing massive amounts of information using automated tools. This approach allows not only to identify inconsistencies in reporting, but also to offer recommendations for optimizing business processes.

In construction, similar methods can be used, for example, for automatic validation of design data, construction quality control and contractor performance evaluation. One tool to automate and speed up data processing is the use of regular expressions (RegEx) in the data transformation phase (Transform) of the ETL process. RegEx allows you to efficiently validate data strings, detect inconsistencies and ensure the integrity of information with minimal resource consumption. We talked more about RegEx (Fig. 4.4-7) in the chapter "Translating Requirements into Structured Form".

Let's consider a practical example: in the real estate object management system (RPM) the manager sets the requirements to the key attributes of objects (Fig. 7.2-7). At the transformation stage, the following parameters need to be validated:

- checking of object identifier formats (attribute "ID")
- control of replacement warranty period values (attribute "Warranty period")
- verification of the element replacement cycle (attribute "Maintenance Requirements")



Property Manager:
Long-term Management

ID	Element	Warranty Period	Replacement Cycle	Maintenance Requirements
W-NEW	Window	-	20 years	Annual Inspection
W-OLD1	Window	8 years	15 years	Biannual Inspection
W-OLD2	Window	8 years	15 years	Biannual Inspection
D-122	Door	15 years	25 years	Biennial Varnishing

Fig. 7.2-7 Verification of quality begins with setting requirements for attributes and their boundary values.

To set boundary values for parameter validation, for example, suppose we know from our experience that valid values for the "ID" attribute can only include string values "W-NEW", "W-OLD1" or "D-122" or similar values where the first character is a letter followed by a dash and then three alphabetic characters 'NEW', 'OLD' or any three-digit number (Fig. 7.2-7). The following regular expression (RegEx) can be used to validate these identifiers:

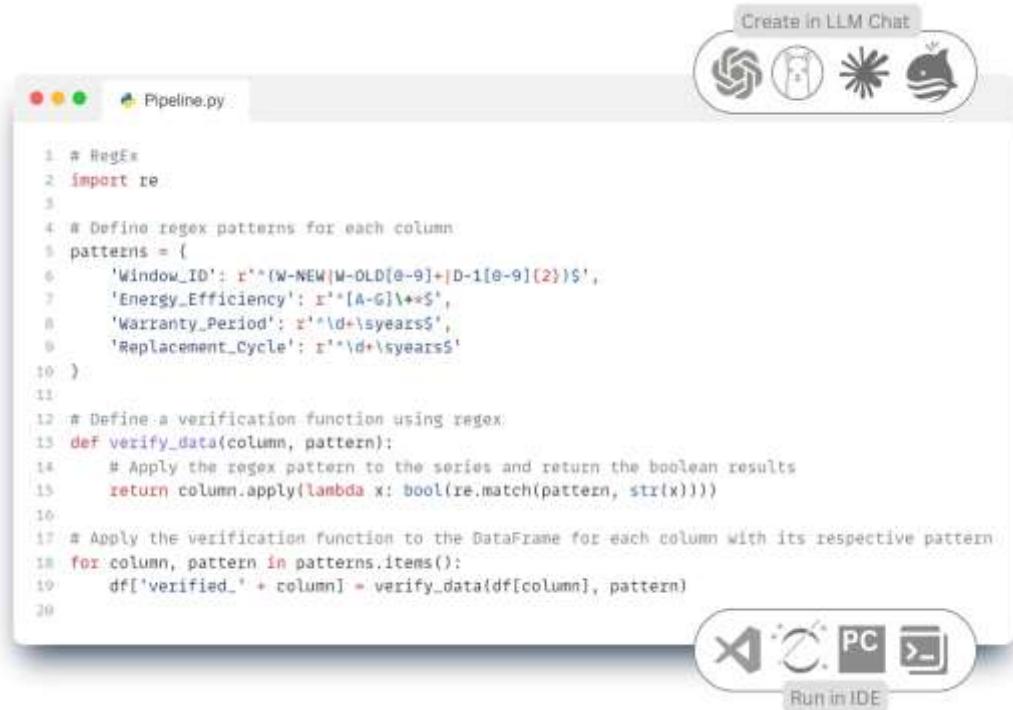
```
^W-NEW$| ^W-OLD[0-9]+$| ^D-1[0-9]{2}$
```

This template makes sure that all identifiers in the data match the specified criteria. If any value fails the check, the system logs an error. To create Python -code to transform the data and use the resulting data to create a results table, simply formulate a query in the LLM chat.

🔗 Text request to LLM:

```
Write code to validate the columns of DataFrame using regular expressions, which checks identifiers in the format 'W-NEW' or 'W-OLD' via RegEx, energy efficiency with letters 'A' through 'G', warranty period and replacement cycle with numeric values in years ↴
```

💡 LLM's response:



```

1 # RegEx
2 import re
3
4 # Define regex patterns for each column
5 patterns = {
6     'Window_ID': r'^W-NEW|M-OLD[0-9]+[D-1][0-9]{2})$',
7     'Energy_Efficiency': r'^[A-G]\w*$',
8     'Warranty_Period': r'^\d+\w{years}$',
9     'Replacement_Cycle': r'^\d+\w{years}$'
10 }
11
12 # Define a verification function using regex
13 def verify_data(column, pattern):
14     # Apply the regex pattern to the series and return the boolean results
15     return column.apply(lambda x: bool(re.match(pattern, str(x))))
16
17 # Apply the verification function to the DataFrame for each column with its respective pattern
18 for column, pattern in patterns.items():
19     df['verified_' + column] = verify_data(df[column], pattern)
20

```

Fig. 7.2-8 Code automates the validation process by applying RegEx templates to dataframe parameter columns.

The above, automatically generated Python -code (Fig. 7.2-8), uses the "re" library (RegEx regular expressions) to define a function that checks each attribute of a data item in a DataFrame. For each specified column (attribute), the function applies a RegEx pattern to check that each record conforms to the expected format and adds the results as new values (False/True) to the new attribute-column of the DataFrame.

Such automated validation ensures formal data compliance and can be used as part of a quality control system during the transformation phase.

After successful completion of the Transform step and quality check, the data is ready to be uploaded to target systems. Transformed and validated data can be uploaded to CSV, JSON, Excel, databases and other formats for further use. Depending on the task, the results can also be presented in reports, graphs or analytical dashboards.

ETL Load: Visualize results in charts and graphs

After completion of the Transform stage, when the data have been brought to a structured form and verified, the final stage - Load, where the data can be both loaded into target system and visualized for

analysis. Visual presentation of data allows to promptly identify deviations, analyze distributions and communicate key conclusions to all project participants, including those with no technical background.

Instead of presenting information as tables and numbers, we can use infographics, graphs, and dashboards (dashboards). One of the most common and flexible tools for visualizing structured data in Python is the Matplotlib library (Fig. 7.2-9, Fig. 7.2-10). It allows you to create static, animated, and interactive graphs, and supports a wide range of chart types.

- To visualize the results of attribute checking from the RPM system (Fig. 7.2-7), you can use the following query to the language model:

Write code to visualize the DataFrame data, above (Fig. 7.2-7), with a histogram for the results to show the error rate of the attribute



- LLM response in code form and ready visualization directly in the LLM chat room of the code execution results:

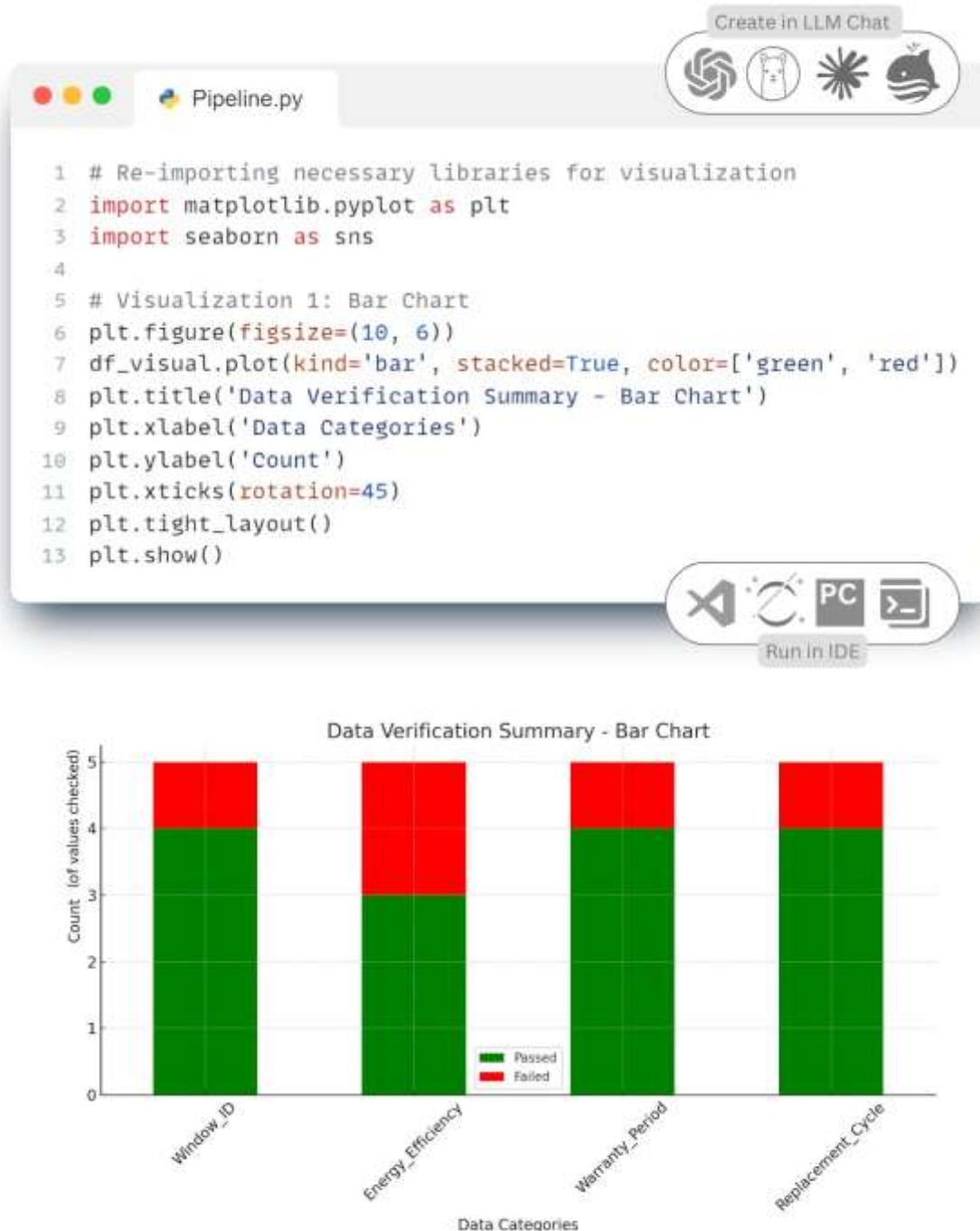


Fig. 7.2-9 Visualization of the results of the Transform step of checking attribute values from the RPM system (Fig. 7.2-7) as a histogram in the Load step.

- 💡 There are many open source and free visualization libraries that allow you to represent structured data in various formats. Let's continue visualizing the results with another type of graph with the following prompt in the chat:

Depict the same data as a graph of lines ↵

- 💡 LLM's response:

```

1 # Visualization 3: Line Chart
2 df_visual.plot(kind='line', marker='o', figsize=(10, 6))
3 plt.title('Data Verification Trend - Line Chart')
4 plt.xlabel('Data Categories')
5 plt.ylabel('Count')
6 plt.grid(True)
7 plt.show()

```

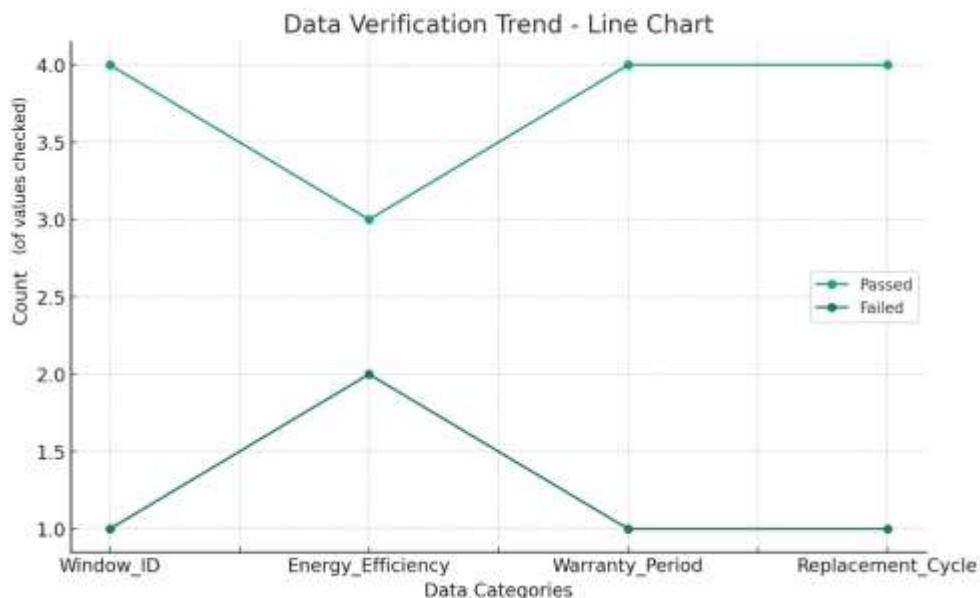


Fig. 7.2-10 Visualization of the validation data (Fig. 7.2-8) as a line diagram obtained using the Matplotlib library.

There are many open source and free visualization libraries such as:

- Seaborn - for statistical graphs (Fig. 7.2-11)
- Plotly - for interactive web visualizations (Fig. 7.2-12, Fig. 7.1-6)
- Altair - for declarative visualization
- Dash or Streamlit - to create full dashboards

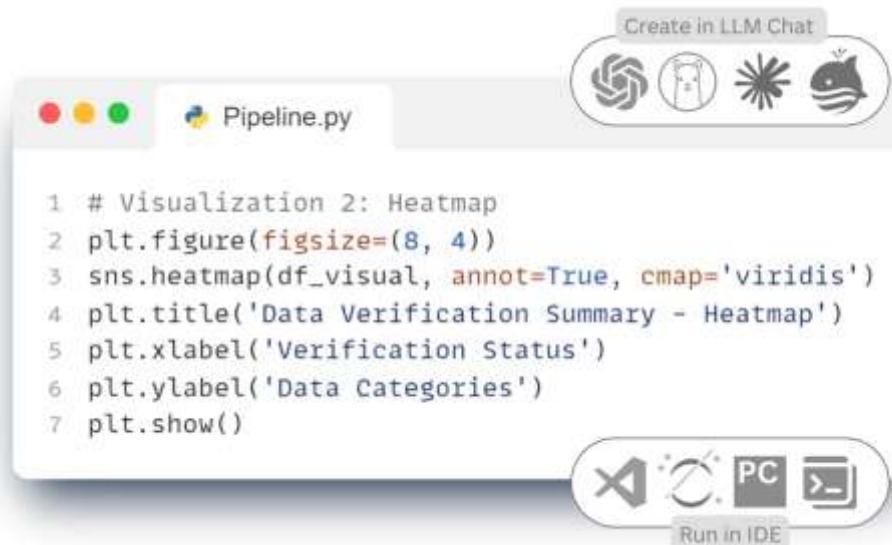
Knowledge of specific libraries for visualization is not necessary - modern tools, including LLM, allow you to automatically generate code for graphs and entire applications based on the task description.

The choice of tool depends on the objectives of the project: whether it is a report, a presentation or an online dashboard. For example, the open source library Seaborn is particularly good for working with categorical data, helping to identify patterns and trends.

- 2 To see the Seaborn library in operation, you can either ask the LLM to use the desired library directly or submit a similar text request in the LLM continuation at:

Show heatmap for ↴

- LLM's answer in the form of code and a finished graph, the plotting code of which can now be copied into the IDE, and the graph itself can be copied or saved for insertion into a document:



The screenshot shows a software interface with a light gray background. At the top, there are three colored dots (red, yellow, green) and a file icon labeled "Pipeline.py". To the right of the file name are four circular icons: a blue square with a white dot, a white square with a blue dot, a white square with a red dot, and a white square with a green dot. Above these icons is a button labeled "Create in LLM Chat". Below the file name, the Python code for generating a heatmap is displayed:

```

1 # Visualization 2: Heatmap
2 plt.figure(figsize=(8, 4))
3 sns.heatmap(df_visual, annot=True, cmap='viridis')
4 plt.title('Data Verification Summary - Heatmap')
5 plt.xlabel('Verification Status')
6 plt.ylabel('Data Categories')
7 plt.show()

```

At the bottom of the interface is a horizontal bar containing four icons: a red square with a white X, a blue square with a white circle, a white square with a blue square labeled "PC", and a white square with a black arrow pointing right. Below this bar is a button labeled "Run in IDE".

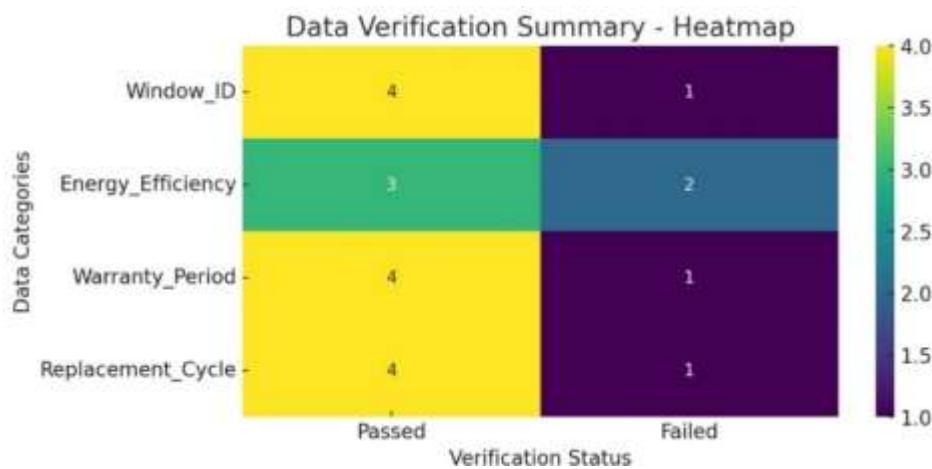


Fig. 7.2-11 Visualizing the results of validating (Fig. 7.2-8) the data using the Seaborn library.

For those who prefer an interactive approach, there are tools that allow you to create dynamic charts and panels with the ability to interact. The Plotly library (Fig. 7.1-6, Fig. 7.2-12) offers the ability to create highly interactive charts and panels that can be embedded in web pages and allow the user to interact with the data in real time.

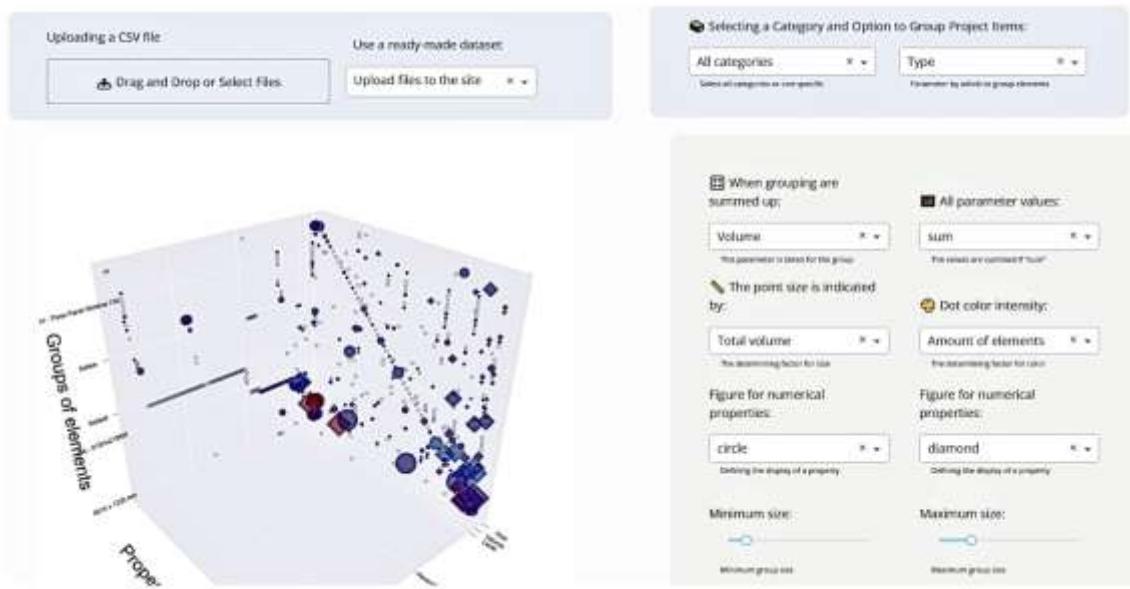


Fig. 7.2-12 Interactive 3D -visualization of element attributes from a CAD- (BIM-) project using the Plotly library.

Specialized open source libraries Bokeh, Dash and Streamlit provide a convenient way to present data without requiring deep knowledge of web development. Bokeh is suitable for complex interactive graphs, Dash is used to build full-fledged analytical dashboards, and Streamlit allows you to quickly create web applications for data analysis.

With visualization tools like these, developers and analysts can effectively disseminate results to colleagues and stakeholders, enabling intuitive interaction with data and simplifying decision-making.

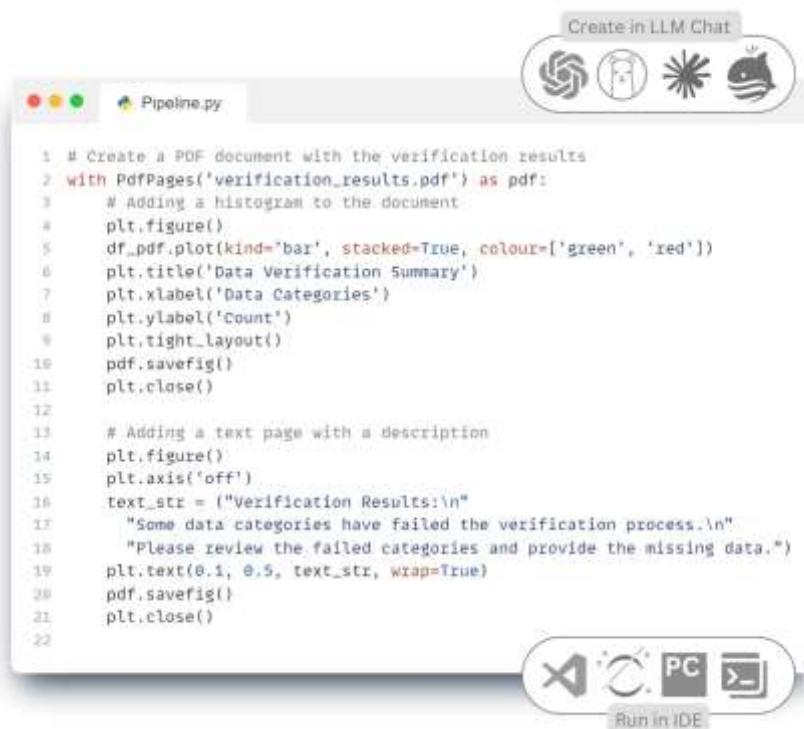
ETL Load: Automatic creation of PDF documents

At the data loading stage it is possible not only to visualize data, upload them to tables or databases, but also to automatically generate reports, including the necessary graphs, charts and key analytical indicators that are necessary for a manager or specialist waiting for inspection results. Automated reports can contain both comments and textual interpretation of data, as well as visualization elements - tables, graphs.

- To create a PDF -report with a histogram (Fig. 7.2-9) and a description of the analysis based on the validation we performed in the previous chapters, simply formulate a request in the continuation of the dialog with the LLM, e.g.:

Write code to create a PDF -file with a histogram and description of the results of the data validation above (in the chat), and write a text warning that some categories were not validated and that missing needs to be filled in ↴

- LLM's response as code and finished PDF with results:



The screenshot shows a software interface with a code editor and various toolbars. The code editor contains Python code for generating a PDF document with a histogram and a text page describing verification results. The toolbar includes icons for 'Create in LLM Chat' (with a AI icon), 'Run in IDE' (with a PC icon), and other development-related icons.

```

1 # Create a PDF document with the verification results
2 with PdfPages('verification_results.pdf') as pdf:
3     # Adding a histogram to the document
4     plt.figure()
5     df_pdf.plot(kind='bar', stacked=True, colour=['green', 'red'])
6     plt.title('Data Verification Summary')
7     plt.xlabel('Data Categories')
8     plt.ylabel('Count')
9     plt.tight_layout()
10    pdf.savefig()
11    plt.close()
12
13    # Adding a text page with a description
14    plt.figure()
15    plt.axis('off')
16    text_str = ("Verification Results:\n"
17                "Some data categories have failed the verification process.\n"
18                "Please review the failed categories and provide the missing data.")
19    plt.text(0.1, 0.5, text_str, wrap=True)
20    pdf.savefig()
21    plt.close()
22

```

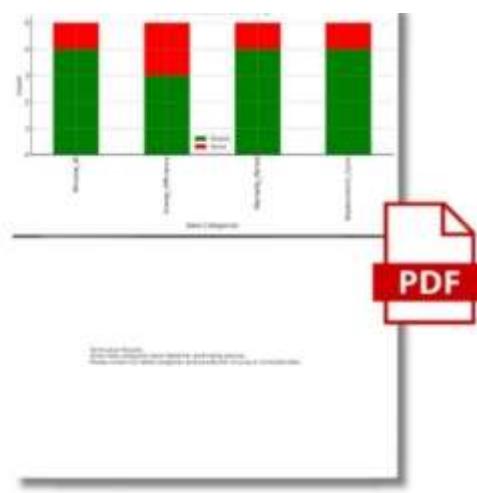


Fig. 7.2-13 The automated code creates a PDF -document containing a histogram with test data and text with test results.

An automatically written solution of only 20 lines of code using LLM instantly creates the desired PDF (or DOC) document with visualization in the form of an attribute histogram (Fig. 7.2-13) showing the number of data that passed and failed validation, and with the addition of a text block summarizing the results and recommendations further action.

Automated document generation is a key element of the Load stage, especially in a project environment where speed of reporting and accuracy are critical.

ETL Load: automatic document generation from FPDF

Automating reporting at the ETL stage Load is an important step in data processing, especially when the results of the analysis need to be presented in a format that is easy to communicate and understand. In the construction industry, this is often relevant for progress reports, project data statistics, quality assurance reports or financial documentation.

One of the most convenient tools for such tasks is the open source library, FPDF, available for both Python and PHP.

The **FPDF** open source library provides a flexible way to generate documents through code, allowing you to add headers, text, tables and images. Using code instead of manual editing reduces errors and speeds up the process of preparing reports in PDF format.

One of the key stages of creating a PDF -document is adding headings and main text in the form of comments or description. However, when creating a report, it is important not only to add text, but also to structure it properly. Headings, indents, line spacing - all this affects the readability of the document. Using FPDF, you can set formatting parameters, control the arrangement of elements and customize the style of the document.

FPDF is very similar in principle to HTML. Those who are already familiar with HTML can easily generate PDF documents of any complexity using FPDF, as the code structure is very similar to HTML markup: headers, text, images and tables are added in a similar way. Those who are not familiar with HTML need not worry - you can use LLM, which will instantly help you compose the code to generate the desired document layout.

- The following example demonstrates how to generate a report with a header and body text. Executing this code in any IDE with Python support creates a PDF -file containing the desired header and text:

```
from fpdf import FPDF      # Import the FPDF library
pdf = FPDF()    # Create PDF -document
pdf.add_page()   # Add a page

pdf.set_font("Arial", style='B', size=16)  # Set font: Arial, bold, size 16
pdf.cell(200, 10, "Project Report", ln=True, align='C')  # Create title and center it
pdf.set_font("Arial", size=12)  # Change the font to regular Arial, size 12
pdf.multi_cell(0, 10, "This document contains data on the results of project file verification . . .")  #
Add multi-line text
pdf.output(r "C:\reports\report.pdf") # Save PDF -file
```



Fig. 7.2-14 With a few lines of Python code, we can automatically generate the PDF text document we need.

When preparing reports, it is important to take into account that the data from which the document is formed is rarely static. Headers, text blocks (Fig. 7.2-14) are often formed dynamically, receiving values at the Transform stage in the ETL process.

Using the code allows you to create documents that contain up-to-date information: project name, date of report generation, as well as information about participants or current status. The use of variables in the code allows you to automatically insert this data in the required places in the report, completely eliminating the need for manual editing before sending.

In addition to simple text and headings, tables occupy a special place in project documentation. Al-

most every document contains structured data: from object descriptions to inspection results. Automatic generation of tables based on data from the Transform stage allows not only to speed up the process of document preparation, but also to minimize errors when transferring information. FPDF allows to insert tables into PDF -files (as text or pictures), setting cell borders, column sizes and fonts (Fig. 7.2-15). It is especially convenient when working with dynamic data, when the number of rows and columns can vary depending on the document tasks.

- The following example shows how to automate the creation of tables, e.g. with a list of materials, estimates or parameter test results:

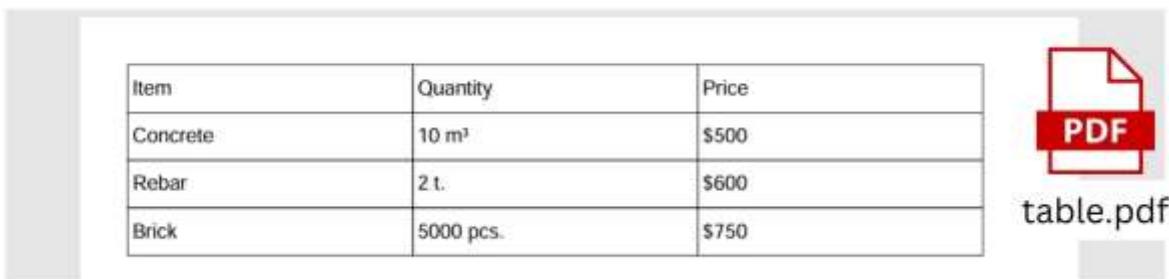
```

data = [
    ["Item", "Quantity", "Price"], # Column headings
    ["Concrete", "10 m³", "$ 500."], # First row data
    ["Rebar", "2 tons", "$ 600"], # Second row data.
    ["Brick", "5,000 pieces", "$ 750."], # Line 3 data.
]

pdf = FPDF () # Create PDF -document
pdf.add_page() # Add a page
pdf.set_font("Arial", size=12) # Set the font

for row in data: # Search table rows
    for item in row: # Go through the cells in the row
        pdf.cell(60, 10, item, border=1) # Create a cell with a border, width 60 and height 10.
    pdf.ln() # Move to the next line
pdf.output(r "C:\reports\table.pdf") # Save PDF -file

```



A screenshot showing a table generated by FPDF and a PDF file icon labeled 'table.pdf'. The table has three columns: Item, Quantity, and Price. The data rows are Concrete (10 m³, \$500), Rebar (2 t., \$600), and Brick (5000 pcs., \$750).

Item	Quantity	Price
Concrete	10 m³	\$500
Rebar	2 t.	\$600
Brick	5000 pcs.	\$750

Fig. 7.2-15 You can automatically generate not only text, but also any table information obtained in the Transform step.

In real reporting scenarios, tables are usually dynamically generated information obtained at the data transformation stage. In the example shown (Fig. 7.2-15), the table is inserted into the PDF -document in a static form: the data for the example was placed in the data dictionary (the first line of the code), in real conditions such data variable is filled in automatically after e.g. grouping of the dataframe.

In practice, such tables are often built on the basis of structured data coming from various dynamic sources: databases, Excel -files, API -interfaces or results of analytical calculations. Most often, at the Transform (ETL) stage, data is aggregated, grouped or filtered - and only then transformed into totals

in the form of graphs or two-dimensional tables displayed in reports. This means that the table content can change depending on the selected parameters, analysis period, project filters or user settings.

The use of dynamic dataframes and datasets in the Transform stage makes the reporting process in the Load stage as flexible, scalable and easily repeatable as possible without the need for manual intervention.

Besides tables and text FPDF also supports adding graphs of tabular data, which allows you to embed images generated with Matplotlib or other visualization libraries we have considered above into the report. You can supplement the document with any graphs, charts and diagrams using the code.

- 2 Using the Python library FPDF, let's add a graph pre-generated with Matplotlib. to the PDF document:

```
import matplotlib.pyplot as plt # Import Matplotlib to create plots

fig, ax = plt.subplots() # Create the Fig. and axes of the chart
categories = ["Concrete", "Rebar", "Brick"] # Category names
values = [50000, 60000, 75000] # Category values
ax.bar(categories, values) # Create a bar chart
plt.ylabel("Value,$.") # Sign the Y axis
plt.title("Cost Distribution") # Add a title
plt.savefig(r "C:\reports\chart\chart\chart.png") # Save the chart as an image.

pdf = FPDF () # Create PDF -document
pdf.add_page() # Add a page
pdf.set_font("Arial", size=12) # Set the font
pdf.cell(200, 10, "Cost Chart", ln=True, align='C') # Add a header

pdf.image(r "C:\reports\chart\chart\chart.png", x=10, y=30, w=100) # Insert image
into PDF (x, y - coordinates, w - width)
pdf.output(r "C:\reports\chart_report.pdf") # Save PDF file
```

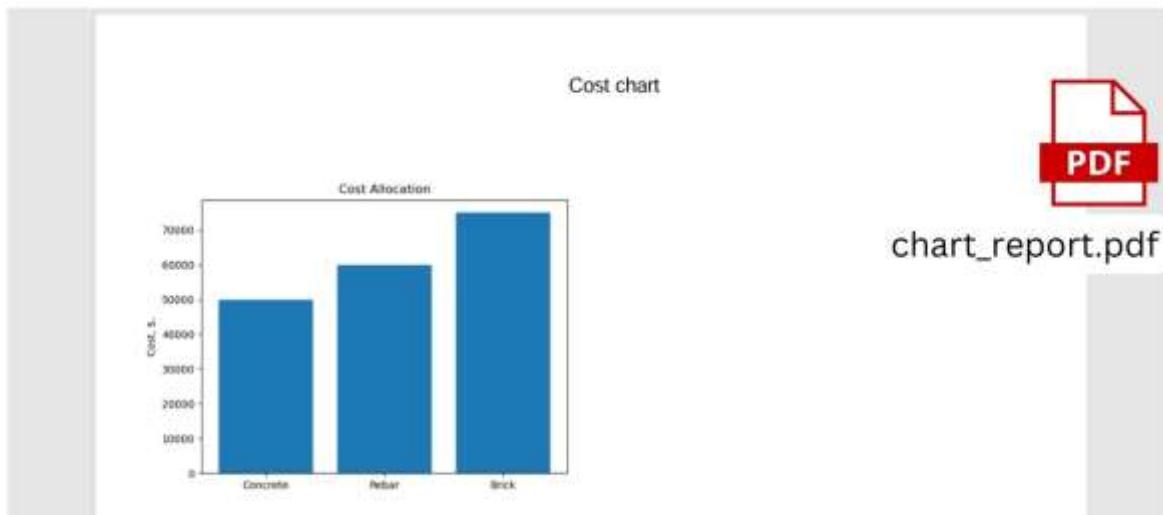


Fig. 7.2-16 With a dozen lines of code, you can generate a graph, save it, and then paste it into a PDF document.

Fpdf makes the process of document preparation and logic transparent, fast and convenient. Templates built into the code allow generating documents with up-to-date data, eliminating the need for manual filling.

Using ETL automation - instead of time-consuming manual reporting, professionals can focus on analyzing data and making decisions, rather than choosing the right tool to work with a particular data silo with a clear user interface.

Thus, the FPDF library provides a flexible tool for automated creation of documents of any complexity - from short technical reports to complex analytical summaries with tables and charts, which allows not only to speed up the document flow, but also significantly reduce the probability of errors associated with manual data entry and formatting.

ETL Load: Reporting and loading to other systems

At the Load stage, the results were generated in the form of tables, graphs and final PDF reports prepared in accordance with the established requirements. Further it is possible to export this data into machine-readable formats (e.g. CSV), which is necessary for integration with external systems such as ERP, CAFM, CPM, BI platforms and other corporate or industry solutions. In addition to CSV, uploads can be made to XLSX, JSON, XML or directly to databases that support automatic information exchange.

- ❸ To generate the appropriate code to automate the Load step, simply query the LLM -interface, for example: ChatGPT, LlaMa, Mistral DeepSeek, Grok, Claude or QWEN:

Write code to generate a report of data validation results in Data-Frame, where columns prefixed with 'verified_' are counted, renamed to 'Passed' and 'Failed', missing values are replaced with 0, and then only those rows that pass all validations are exported to a CSV -file.

2 LLM's response:



The screenshot shows a code editor window titled "Pipeline.py". The code is as follows:

```

1 # Create a report on the results of data validation
2 report = df.filter(like='verified_').apply(lambda x: x.value_counts());T
3 report.columns = ['Passed', 'Failed']
4 report.fillna(0, inplace=True)
5
6 # Export the cleaned and verified data to a CSV file
7 verified_df = df[df.filter(like='verified_').all(axis=1)]
8 verified_df.to_csv('verified_data.csv', index=False)

```

At the top of the window, there are icons for "Create in LLM Chat" and several AI-related icons. At the bottom, there are icons for "Run in IDE" and other system-related icons.

Fig. 7.2-17 Validated data obtained in the Transform step from the final dataframe is exported to a CSV -file for integration with other systems.

In the given code (Fig. 7.2-17) the final stage of ETL -process - Load - is realized, during which the checked data are saved in CSV format, compatible with most external systems and databases. Thus, we have completed the full cycle of the ETL -process, including extraction, transformation, visualization, documentation and export of data to the systems and formats we need, which ensures reproducibility, transparency and automation of work with information.

ETL - pipeline (pipeline) can be used both for processing single projects and for large-scale application - when analyzing hundreds and thousands of incoming data in the form of documents, images, scans, CAD -projects, point clouds, PDF -files or other sources coming from distributed systems. The ability to fully automate the process makes ETL not just a technical processing tool, but the foundation of a digital construction information infrastructure.

ETL with LLM: Visualize data from PDF -documents

It's time to move on to building a full-fledged ETL process that covers all key stages of data handling in a single scenario - extraction, transformation and loading. Let's build an automated ETL-Pipeline

that allows you to process PDF documents without manual work - extract data from documents, visualize, analyze and transfer to other systems.

The ETL process in our example will be described through prompts, which will have to explain to the language model (LLM) all the ETL processes with a description of the end result to be obtained. In this case, the task is to find all PDF files in the specified folder and its subfolders, extract relevant information from them - e.g. material names, quantities and costs - and present the result as a structured table (DataFrame) for further analysis

- ❷ First text request to LLM to automatically extract data from multiple PDF documents and create a dataframe of the data for the Extract step:

```
Write code to extract material information from PDF files in a given
folder and its subfolders. The data in the PDF includes material
name, quantity and cost. The result should be saved in a DataFrame ↵
```

2 LLM's response:

```

1 import os
2 import pdfplumber
3 import pandas as pd
4
5 def extract_data(base_folder_path):
6     """
7         Extract material data from PDF files in the specified base
8         folder and all its subfolders. Assumes each line in the PDF contains
9         'Material' followed by material name, quantity, and cost.
10    """
11    data = []
12    # Walk through each folder and subfolder in the base folder
13    for folder_path, _, files in os.walk(base_folder_path):
14        for file in filter(lambda f: f.endswith('.pdf'), files):
15            with pdfplumber.open(os.path.join(folder_path, file)) as pdf:
16                # Process each page in the PDF
17                for page in pdf.pages:
18                    # Splitting text into lines
19                    for line in page.extract_text().split('\n'):
20                        # Check if the line contains material data
21                        if 'Material' in line:
22                            parts = line.split()
23                            # Append data to list; convert quantity and cost to float
24                            data.append({'file': file, 'material': parts[1],
25                                         'quantity': float(parts[2]), 'cost': float(parts[3])})
26    # Convert list of dictionaries to pandas DataFrame
27    return pd.DataFrame(data)
28
29 # Specify the base folder path containing PDF files and subfolders with PDFs
30 df = extract_data(r'C:\')

```

Fig. 7.2-18 LLM creates Python -code to extract data from PDF -files in a specific folder and all its subfolders.

The LLM response (Fig. 7.2-18) is a ready-made Python -script that automatically traverses all folders, opens the found PDF -files, extracts textual information from them and converts it into a table. The resulting chat code can be run in one of the popular IDEs PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse with PyDev plugin, Thonny, Wing IDE, IntelliJ IDEA with Python plugin, JupyterLab or popular online tools Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

Alternatively, to simplify the collection process, instead of copying the code from the LLM and using the code in the IDE, we can also upload a dozen PDF -files directly to the LLM chat room (Fig. 7.2-19) and get a table as output, without having to see the code or run it. The result of executing this code will be a table with the attributes we have selected.

ChatGPT 4 ~

You
Write code to extract material information from PDF files in a given folder. The data in the PDF includes material name, quantity and cost. The result should be saved in a pandas Dataframe and displayed on the screen.

ChatGPT
Below is a sample Python script that demonstrates this process using 'PyPDF2'. Please note that you might need to adjust the text parsing part based on the specific format of your PDF files.

```
python
import os
import PyPDF2
import pandas as pd

# Create a DataFrame from the extracted data
df = pd.DataFrame(data)

# Display the DataFrame
print(df)
```

Here's an example of how the output would look in a pandas DataFrame after extracting and processing the material information from the PDF files:

Material Name	Quantity	Cost
Copper Wire	100	250.50
Aluminum Sheet	50	150.75
Plastic Pipes	200	100.00
Steel Rods	80	300.00

This table shows the material name, quantity, and cost as columns, with each row representing the data extracted from a PDF file. The actual content and structure of your DataFrame will depend on the specific format and data in your PDF files. [.]

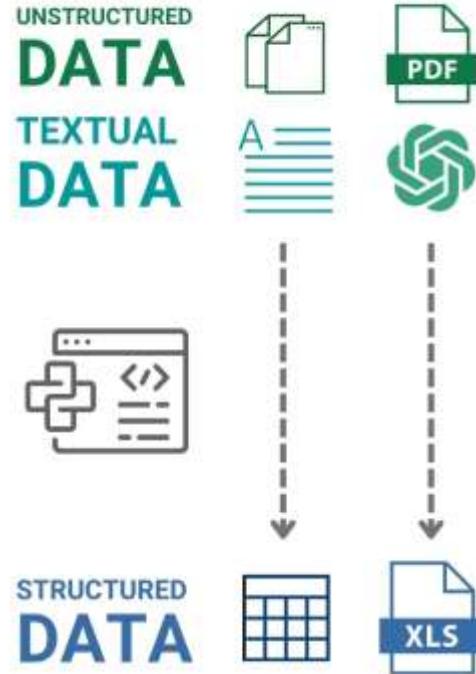
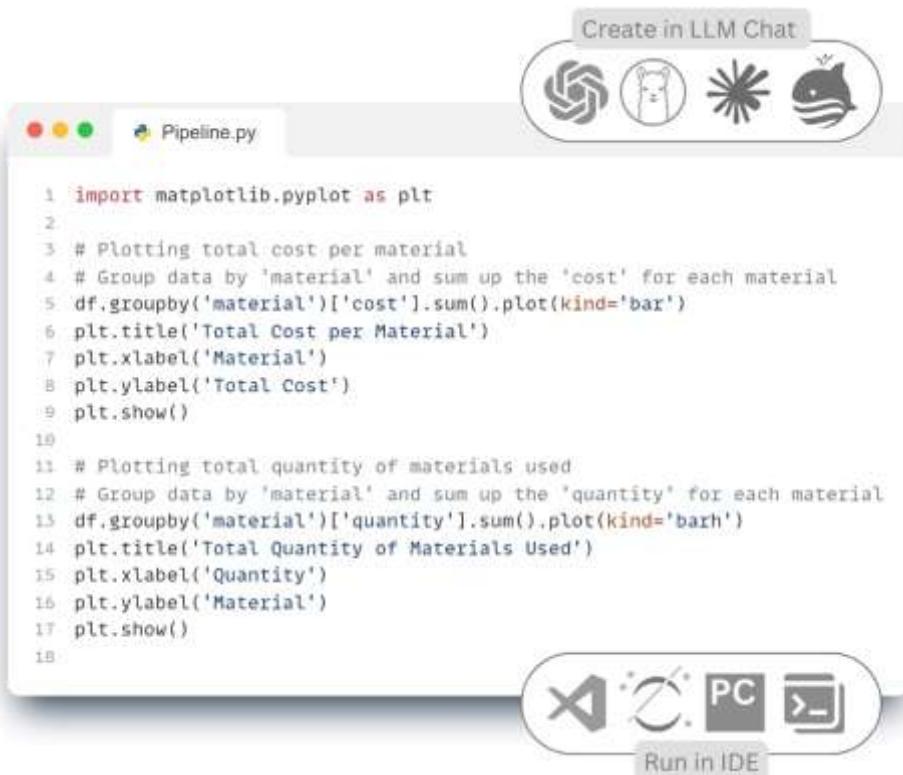


Fig. 7.2-19 The result of executing code in LLM, which extracts data from PDF -files in a structured dataframe view with selected attributes.

In the next step, we ask for a language model on the obtained data - for example, to compare the cost and volume of material usage and create some visualization examples that will serve as a basis for further analysis.

- Ask in a continuing chat with the LLM to plot some graphs from the tables that were produced in the Transform step (Fig. 7.2-18):

Visualize the total cost and quantity of each material from DataFrame (Fig. 7.2-18) ↴



The screenshot shows a chat interface with a "Create in LLM Chat" button at the top. Below it is a code editor window titled "Pipeline.py" containing the following Python code:

```

1 import matplotlib.pyplot as plt
2
3 # Plotting total cost per material
4 # Group data by 'material' and sum up the 'cost' for each material
5 df.groupby('material')['cost'].sum().plot(kind='bar')
6 plt.title('Total Cost per Material')
7 plt.xlabel('Material')
8 plt.ylabel('Total Cost')
9 plt.show()
10
11 # Plotting total quantity of materials used
12 # Group data by 'material' and sum up the 'quantity' for each material
13 df.groupby('material')['quantity'].sum().plot(kind='barh')
14 plt.title('Total Quantity of Materials Used')
15 plt.xlabel('Quantity')
16 plt.ylabel('Material')
17 plt.show()
18

```

At the bottom right of the code editor is a "Run in IDE" button.

Fig. 7.2-20 LLM Response -models as Python code to visualize data from a data frame using the matplotlib library.

LLM automatically generates and executes Python -code (Fig. 7.2-20) using the matplotlib library. After executing this code, we get graphs of costs and material utilization in construction projects directly in chat (Fig. 7.2-21), which greatly simplifies the analytical work.

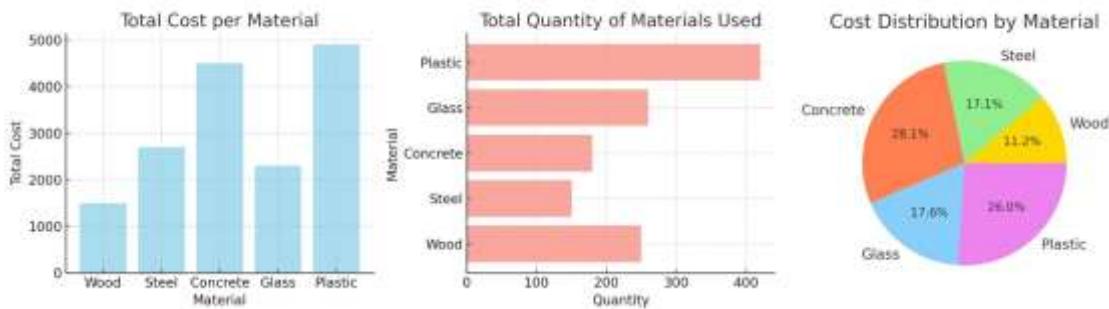


Fig. 7.2-21 Visualizing the LLM response as graphs based on data collected in the DataFrame.

Support in developing ideas for writing ETL code, analyzing and executing code, and visualizing results is available through simple text queries in LLM, without the need to learn the basics of programming. The emergence of AI tools such as LLM is definitely changing the approach to programming and automating data processing (Fig. 7.2-22).

According to the PwC report "What is the real value of artificial intelligence for your business and how can you capitalize on it?" (2017) [139], process automation and productivity improvements will be the main drivers of economic growth. And productivity improvements are expected to account for more than 55% of all AI-driven GDP growth between 2017 and 2030."

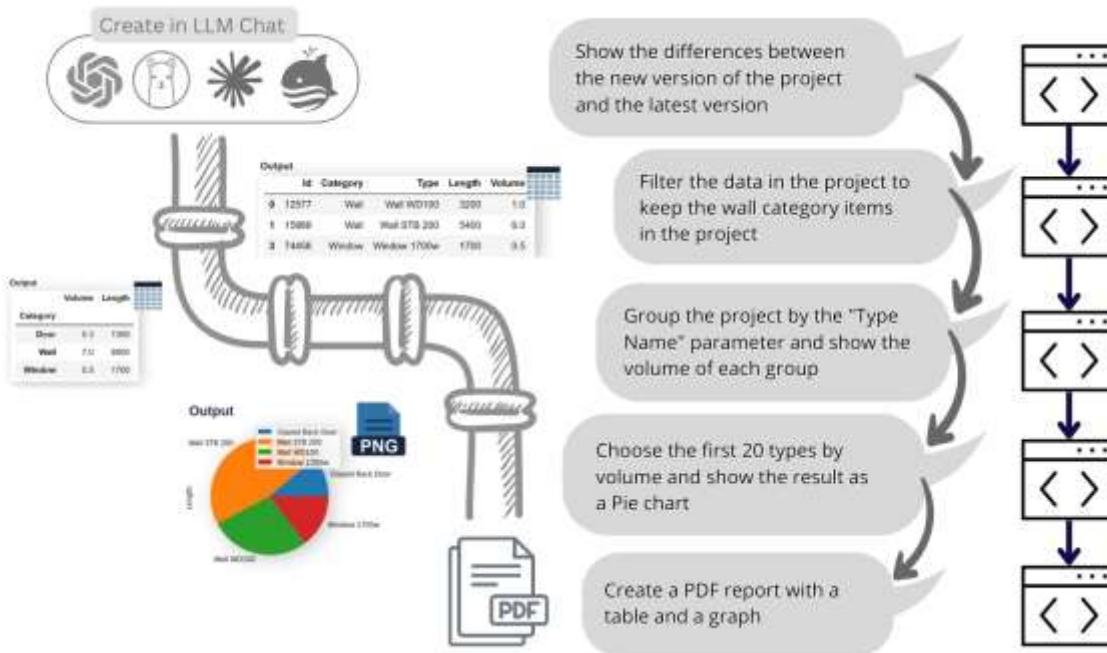


Fig. 7.2-22 AI LLM helps generate draft code that is applied to future projects without the need for LLM.

Using tools like ChatGPT, LLaMa, Mistral, Claude, DeepSeek, QWEN, Grok, as well as open data and open source software, we can automate processes that were previously only done with specialized, high-cost and difficult to maintain modular proprietary systems.

In the construction context, this means that companies that are the first to implement automated Pipeline -data processes will realize significant benefits, from increased project management efficiency to reduced financial losses to the elimination of fragmented applications and siloed data warehouses.

The described logic of executing business tasks in the ETL process is a crucial part of automating analytics and data processing processes, which is a specific variation of a broader concept - pipelines (Pipelines).

CHAPTER 7.3.

AUTOMATIC ETL CONVEYOR (PIPELINE)

Pipeline: Automatic ETL

The ETL process has traditionally been used to process data in analytical systems, covering both structured and unstructured sources. However, in today's digital environment, a broader term is increasingly used - Pipeline (conveyor), which describes any sequential processing chain where the output of one stage becomes the input for the next.

This approach applies not only to data, but also to other types of automation: task processing, building reporting, integration with software and digital workflow (Fig. 7.3-1).

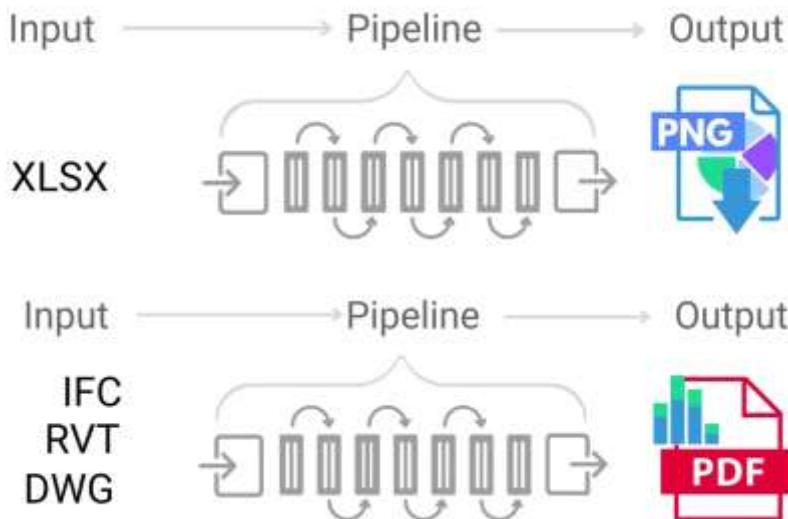


Fig. 7.3-1 Pipeline is a processing sequence in which the output of one stage becomes the input to the next stage.

The use of Pipeline is a major element of automation, especially when working with large amounts of heterogeneous data. Pipeline architecture allows complex processing steps to be organized in a modular, consistent, and manageable format, which increases readability, simplifies code maintenance, and enables incremental debugging and scalable testing.

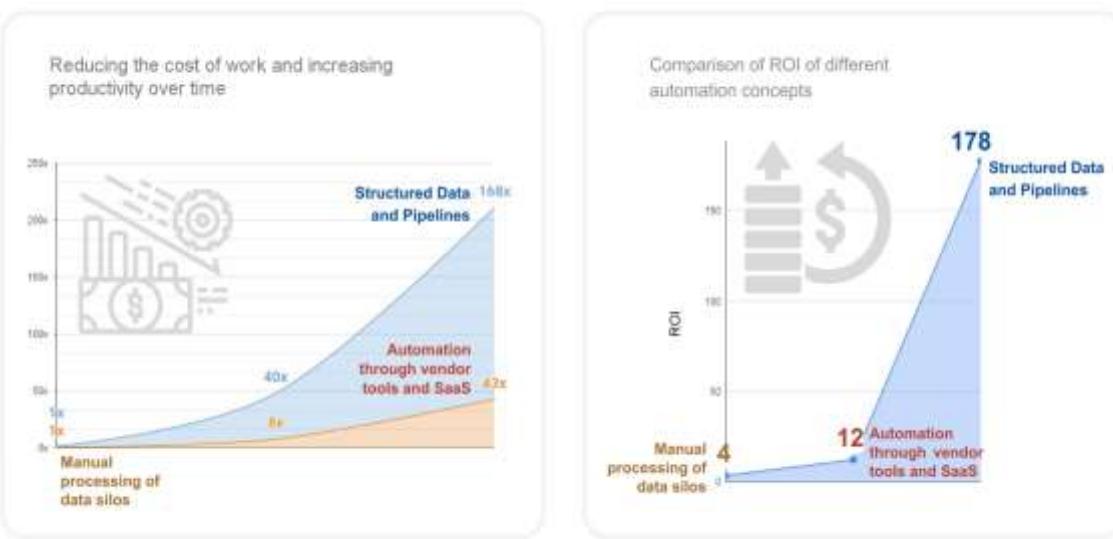


Fig. 7.3-2 ROI Pipeline data validation process reduces execution time by tens and hundreds of times compared to processing using classical tools [74].

Unlike manual work in proprietary systems (ERP, PMIS, CAD, etc.), pipelining allows you to significantly (Fig. 7.3-2) increase the speed of tasks, avoid repetitive work, and automate the start of processes at the right time (Fig. 7.3-3).

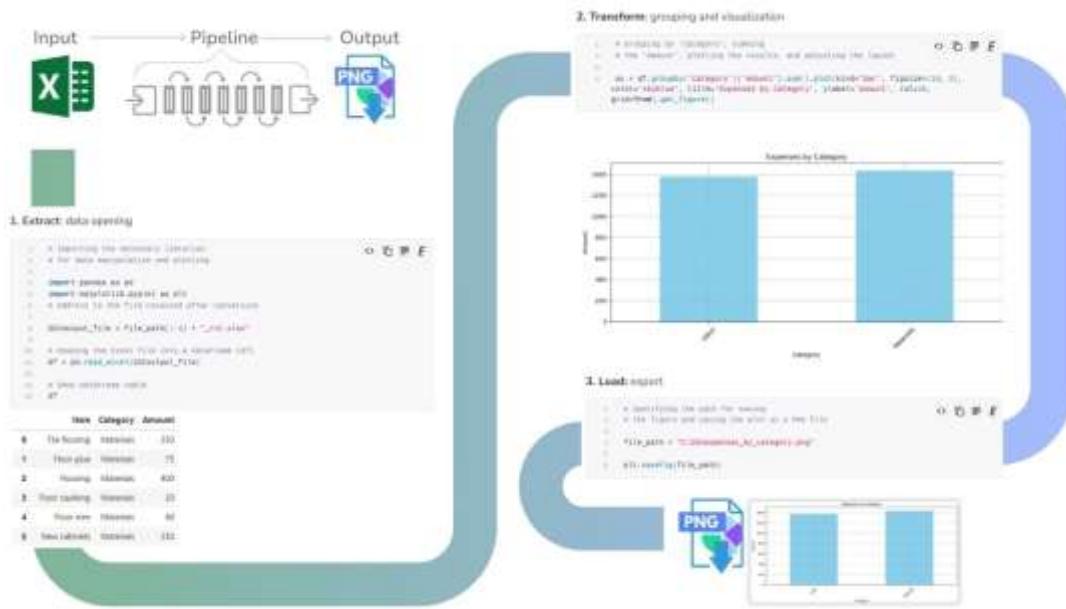


Fig. 7.3-3 ETL example Pipeline on automatically retrieving a chart from tabular data in an XLSX file without opening Excel.

To process streaming data and build an automated Pipeline, similar to the ETL process, you need to identify the sources of data in advance, as well as the timeframe for collecting it - either for a specific business process, or across the entire company.

In construction projects, data comes from many heterogeneous sources with different update intervals. In order to create a reliable data showcase, it is critical to record when information is retrieved and updated. This enables timely decision making and improves the efficiency of project management.

One option is to start the assembly process at a fixed time - for example, at 19:00, at the end of the working day. At this point, the first script responsible for aggregating data from various systems and storages is activated (Fig. 7.3-4 step 1). This is followed by automated processing and transformation of data into a structured format suitable for analytics (Fig. 7.3-4 step 2-4). At the final stage, using the prepared data, reports, dashboards and other products described in the previous chapters are automatically generated (Fig. 7.3-4 step 6-7). As a result, by 05:00 a.m. managers already have up-to-date project status reports in the required format (Fig. 7.3-5).

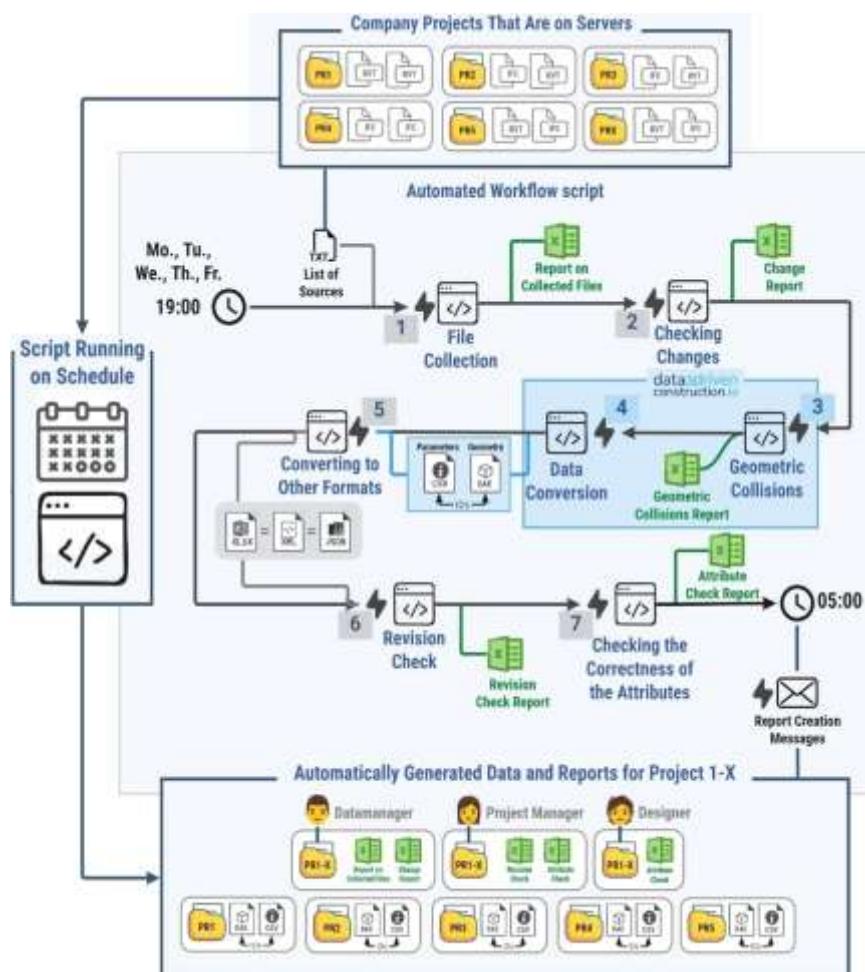


Fig. 7.3-4 Data in Pipeline, automatically collected in the evening, is processed overnight so that managers have up-to-date reports and fresh reports by morning.

Timely data collection, KPI definition, automation of transformation processes and visualization through dashboards are key elements of successful data-driven decision making.

Such automated processes (Fig. 7.3-4) can be executed with full autonomy: they run on a schedule, process data unattended and can be deployed either in the cloud or on the company's own server (Fig. 7.3-5). This allows such ETL pipelines to be integrated into the existing IT infrastructure, maintaining control over data and providing flexibility in scaling.

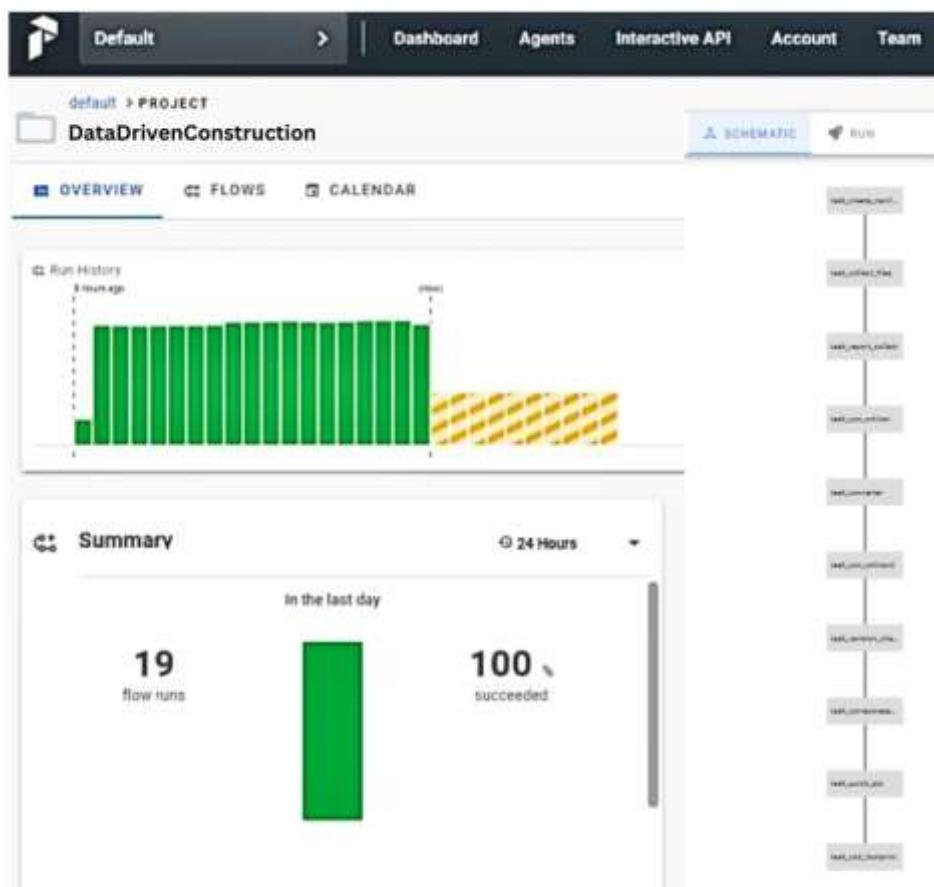


Fig. 7.3-5 Automatic ETL -conveyor processes (Fig. 7.3-4) on the Prefect platform, in which 10 python scripts are run alternately after 7:00 p.m. each business day.

Automating workflows not only increases team productivity by freeing up time for more meaningful and less routine tasks, but also serves as an important first step toward incorporating artificial intelligence (AI) technologies into business processes, which we will discuss in more detail in the chapter on Predictive Analytics and Machine Learning.

Pipeline-ETL data validation process with LLM

In the previous chapters on creating data requirements and automating ETL, we step-by-step broke down the process of data preparation, transformation, validation, and visualization. These activities were implemented as separate code blocks (Fig. 7.2-18 - Fig. 7.2-20) each performing a specific task.

Now we have the next goal - to combine these elements into a single, coherent and automated pipeline of data processing - a pipeline, ETL -Pipeline - in which all stages (loading, validation, visualization, export) are executed sequentially in a single auto-executable script.

In the following example, the full cycle of data processing will be realized: from loading the source CSV file → to checking the structure and values using regular expressions → calculating the results → generating a visual report in PDF format.

- 💡 The following text query to the LLM. can be used to retrieve the appropriate code:

Please write a code sample that loads data from CSV, validates Data-Frame with regular expressions, checks identifiers in 'W-NEW' or 'W-OLD' format, energy efficiency with letters 'A' to 'G', warranty period and replacement cycle with numerical values in years and at the end creates a report with a count of passed and failed values, generates a PDF with a histogram of results and adds a text description.



 LLM's response:



```

1 # Importing necessary libraries
2 import pandas as pd
3 import re
4 import matplotlib.pyplot as plt
5 from matplotlib.backends.backend_pdf import PdfPages
6
7
8 # Load data
9 df = pd.read_csv('path_to_csv_file.csv')
10
11 # Define patterns and verify data
12 patterns = {
13     'Window_ID': r'^W-NEW|W-OLD[0-9]+|D-1[0-9]{2})$',
14     'Energy_Efficiency': r'^[A-G]\+*$',
15     'Warranty_Period': r'^\d+\syears$',
16     'Replacement_Cycle': r'^\d+\syears$'
17 }
18
19 for column, pattern in patterns.items():
20     df[f'verified_{column}'] = df[column].apply(lambda x: bool(re.match(pattern, str(x))))
21
22 # Create report
23 report = df.filter(like='verified_').apply(lambda x: x.value_counts()).T
24 report.columns = ['Passed', 'Failed']
25
26 # Generate PDF
27 pdf_path = '/mnt/data/verification_results_full_pipeline.pdf'
28 with PdfPages(pdf_path) as pdf:
29     # Plot and save bar chart
30     report.plot(kind='bar', stacked=True, color=['green', 'red'])
31     plt.title('Data Verification Summary')
32     plt.xlabel('Data Categories')
33     plt.ylabel('Count')
34     plt.tight_layout()
35     pdf.savefig()
36     plt.close()
37
38     # Add text page
39     plt.figure()
40     plt.axis('off')
41     text_str = "Verification Results Summary"
42     plt.text(0.1, 0.5, text_str, wrap=True)
43     pdf.savefig()
44     plt.close()
45
46 pdf_path
47

```

Fig. 7.3-6 Pipeline (ETL) automates the complete data processing cycle, from downloading and validation to creating a structured PDF report.

The automated code (Fig. 7.3-6) inside the LLM chat room or in DIE, after copying the code, will validate the data from the CSV -file using the specified regular expressions, create a report on the number of passed and failed records, and then save the validation results as a PDF-file.

This structure of ETL-conveyor, where each step - from data loading to report generation - is implemented as a separate module, ensures transparency, scalability and reproducibility. Presenting the validation logic as easy-to-read Python code makes the process transparent and understandable not only for developers, but also for specialists in data management, quality, and analytics.

The Pipeline -approach to automating data processing allows you to standardize processes, increase their repeatability and simplify adaptation to new projects. This creates a unified methodology for analyzing data, regardless of the source or type of task - whether it's compliance testing, generating reports, or transferring data to external systems.

Such automation reduces the impact of human error, reduces reliance on proprietary solutions, and increases the accuracy and reliability of results, making them suitable for both operational analytics at the project level and strategic analytics at the company level.

Pipeline-ETL: verification of data and information of project elements in CAD (BIM)

Data from CAD systems and databases (BIM) are some of the most sophisticated and dynamically updated data sources in the business of construction companies. These applications not only describe the project using geometry, but also supplement it with multiple layers of textual information: volumes, material properties, room assignments, energy efficiency levels, tolerances, life expectancies and other attributes.

Attributes assigned to entities in CAD -models are formed at the design stage and become the basis for further business processes, including costing, scheduling, life cycle assessment and integration with ERP- and CAFM -systems, where the efficiency of processes largely depends on the quality of data coming from design departments.

The traditional approach to attribute validation in CAD- (BIM-) models involves manual validation (Fig. 7.2-1), which becomes a long and costly process when the volume of models is large. Considering the volume and number of modern construction projects and their regular updates, the process of data validation and transformation becomes unsustainable and unaffordable.

General contractors and project managers are faced with the need to process large amounts of project data, including multiple versions and fragments of the same models. The data comes from design organizations in RVT, DWG, DGN, IFC, NWD and other formats (Fig. 3.1-14) and requires regular review for compliance with industry and corporate standards

The dependence on manual actions and specialized software makes the data validation process a bottleneck in workflows related to data from company-wide models. Automation and the use of structured requirements can eliminate this dependency, multiplying the speed and reliability of data validation (Fig. 7.3-7).

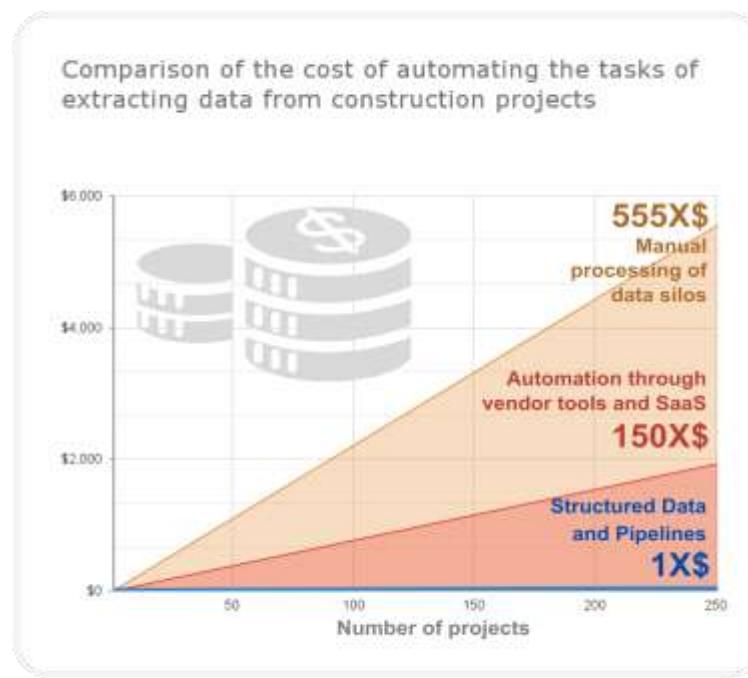


Fig. 7.3-7 Automation increases the speed of data verification and processing, which reduces the cost of work by dozens of times [140].

CAD data validation process includes data extraction (ETL stage Extract) from various closed (RVT, DWG, DGN, NWS, etc.), open semi-structured and parametric formats (IFC, CPXML, USD) or open semi-structured and parametric formats (IFC, CPXML, USD), in which rule tables can be applied to each attribute and its values (Transform stage) using regular expressions RegEx (Fig. 7.3-8), a process which we discussed in detail in the fourth part of the book.

The creation of a PDF error report and successfully validated records should be finalized with output (Load step) in structured formats that only consider validated entities that can be used for further processes.

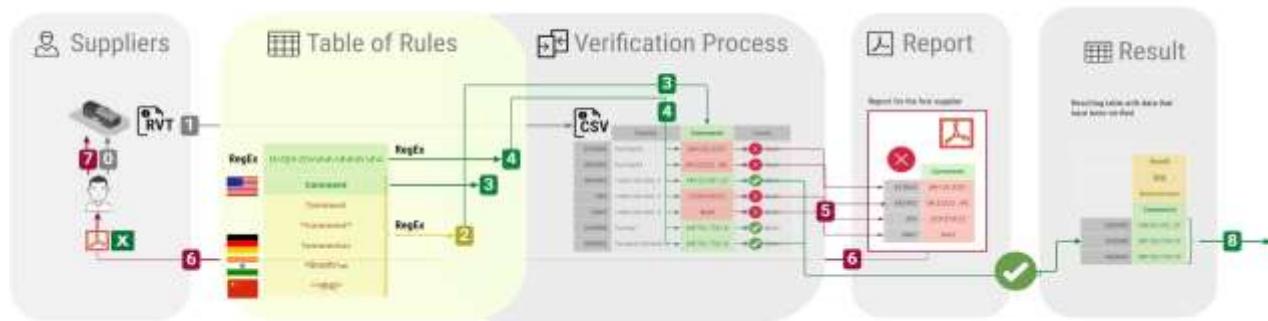


Fig. 7.3-8 Data validation process from project data providers to the final report validated using regular expressions.

Automating the validation of data from CAD (BIM) systems with structured requirements and streaming new data that are processed through ETL-Pipelines (Fig. 7.3-9) reduces the need for manual involvement in the validation process (each of the validation and data requirements processes have been discussed in previous chapters).

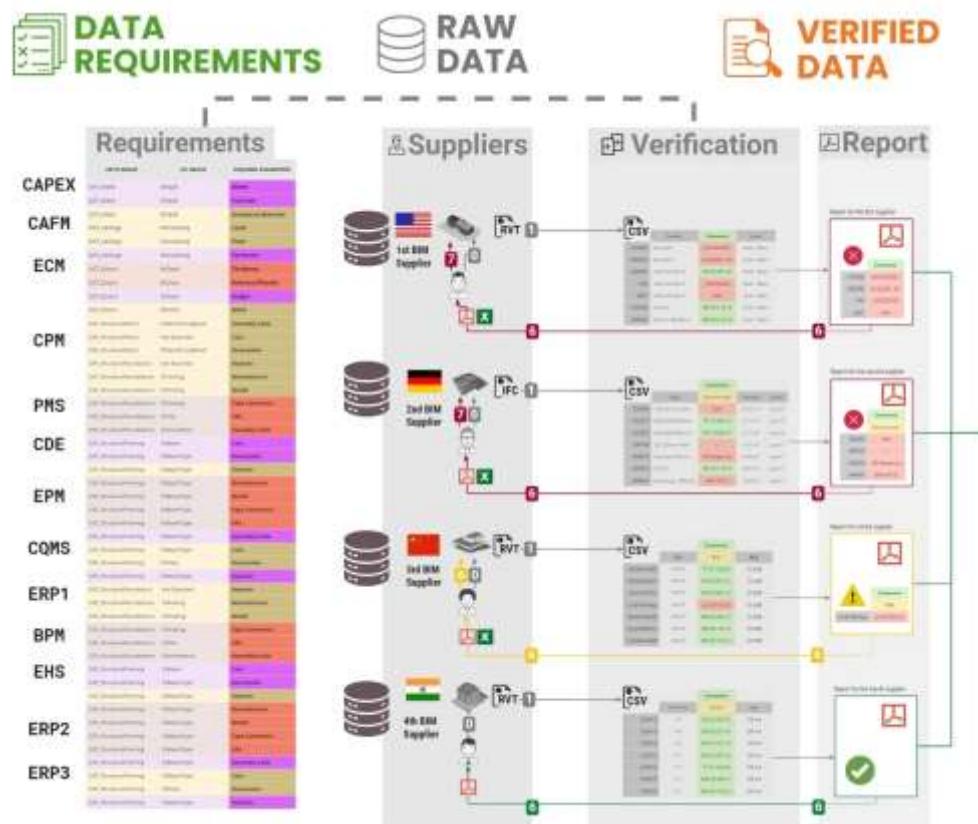


Fig. 7.3-9 Automating data validation through ETL simplifies construction project management by speeding up processes.

Traditionally, validation of models provided by contractors and CAD (BIM) specialists can take days to weeks. However, with the introduction of automated ETL processes, this can be reduced to a few minutes. In a typical situation, the contractor states: "*The model is validated and compliant.*" This statement starts the chain of verification of the contractor's data quality claim:

- ☛ Project Manager - "Contractor states, '*The model has been tested, everything is fine!*'".
- ☛ Data Manager - Load Validation:
 - A simple script in Pandas detects a violation in seconds. Automation eliminates disputes:
 - Category: OST_StructuralColumns, Parameter: FireRating IS NULL.
 - Generate list of violation IDs → export to Excel/PDF.

A simple script in Pandas detects the violation in seconds:

```
df = model_data[model_data["Category"] == "OST_StructuralColumns"] # Filtering
issues = df[df["FireRating"].isnull()] # Empty values
issues[["ElementID"]].to_excel("fire_rating_issues.xlsx") # Export IDs
```

- ☛ Data Manager to Project Manager - "A check shows that 18 columns do not have the FireRating parameter filled in."
- ☛ Project manager to contractor - "The model is returned for revision: the FireRating parameter is mandatory, without its acceptance is impossible"

As a result, the CAD model does not undergo validation, automation eliminates disputes, and the contractor almost instantly receives a structured report with a list of IDs of problematic elements. In this way, the validation process becomes transparent, repeatable and protected from human error (Fig. 7.3-10).

This approach turns the data validation process into an engineering function rather than a manual quality control process. This not only increases productivity, but also makes it possible to apply the same logic to all of the company's projects, enabling end-to-end digital transformation of processes, from design to operations.

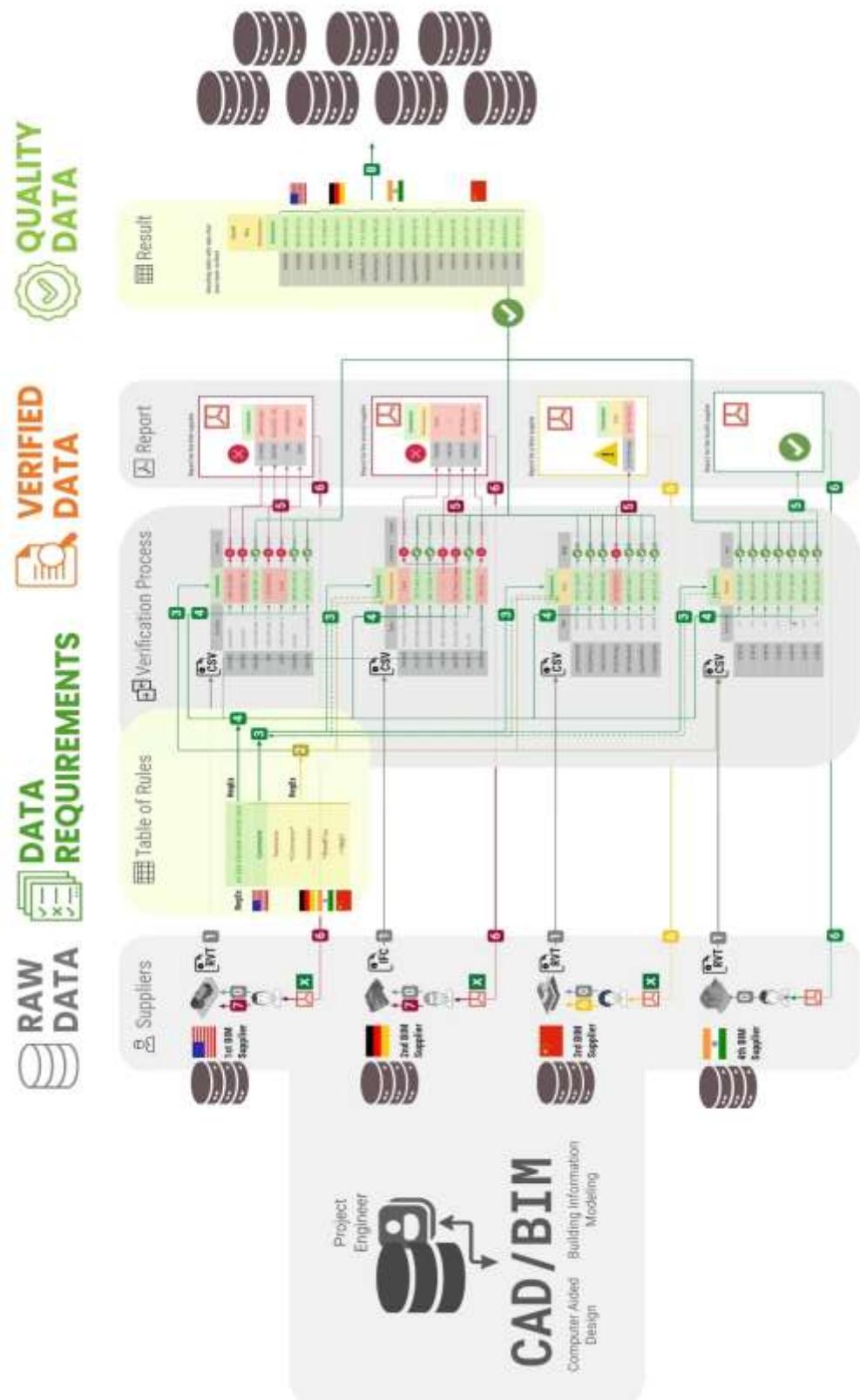


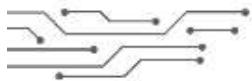
Fig. 7.3-10 Automating element attribute checking eliminates human error and reduces the likelihood of errors.

Through the use of automated pipelines (Fig. 7.3-10), system users expecting quality data from CAD-(BIM-) systems can instantly get the output data they need - tables, documents, images - and quickly integrate it into their work tasks.

The automation of control, processing and analysis is driving a change in the way construction project management is approached, especially the interoperability of different systems, without the use of complex and expensive modular proprietary systems or closed vendor solutions.

While concepts and marketing acronyms come and go, the data requirements validation processes themselves will forever remain an integral part of business processes. Rather than creating more and more specialized formats and standards, the construction industry should look to tools that have already been proven effective in other industries. Today, there are powerful platforms for automating data processing and process integration that allow companies to significantly reduce time for routine operations and minimize errors in Extract, Transform and Load.

One of the popular examples of solutions for automation and orchestration of ETL processes is Apache Airflow, which allows you to organize complex computational processes and manage ETL pipelines. Along with Airflow, other similar solutions such as Apache NiFi for data routing and streaming and n8n for business process automation are also actively used.



CHAPTER 7.4.

ORCHESTRATION OF ETL AND WORKFLOWS: PRACTICAL SOLUTIONS

DAG and Apache Airflow: workflow automation and orchestration

Apache Airflow is a free and open source platform, designed to automate, orchestrate and monitor workflows (ETL -conveyors).

Working with large amounts of data is required every day:

- Download files from different sources - Extract (for example, from suppliers or customers).
- Transform this data into the desired format - Transform (structure, cleanse and validate)
- Send results for verification and create reports - Load (upload to required systems, documents, databases or dashboards).

Manual execution of such ETL processes takes considerable time and leads to the risk of human error. A change in the data source or a failure at one of the steps can cause delays and incorrect results.

Automation tools, such as Apache Airflow, allow you to build a reliable ETL -conveyor, minimize errors, reduce processing time and ensure data correctness at every step. At the heart of Apache Airflow is the concept of DAG (Directed Acyclic Graph) - a directed acyclic graph in which each task (operator) is connected to other dependencies and executed strictly in a specified sequence. DAG eliminates cycles, which provides a logical and predictable structure of task execution.

Airflow takes care of orchestration - managing dependencies between tasks, controlling execution schedules, tracking status and automatically responding to failures. This approach minimizes manual intervention and ensures the reliability of the entire process.

Task Orchestrator is a tool or system designed to manage and control task execution in complex computing and information environments. It facilitates the process of deploying, automating, and managing task execution to improve performance and optimize resources.

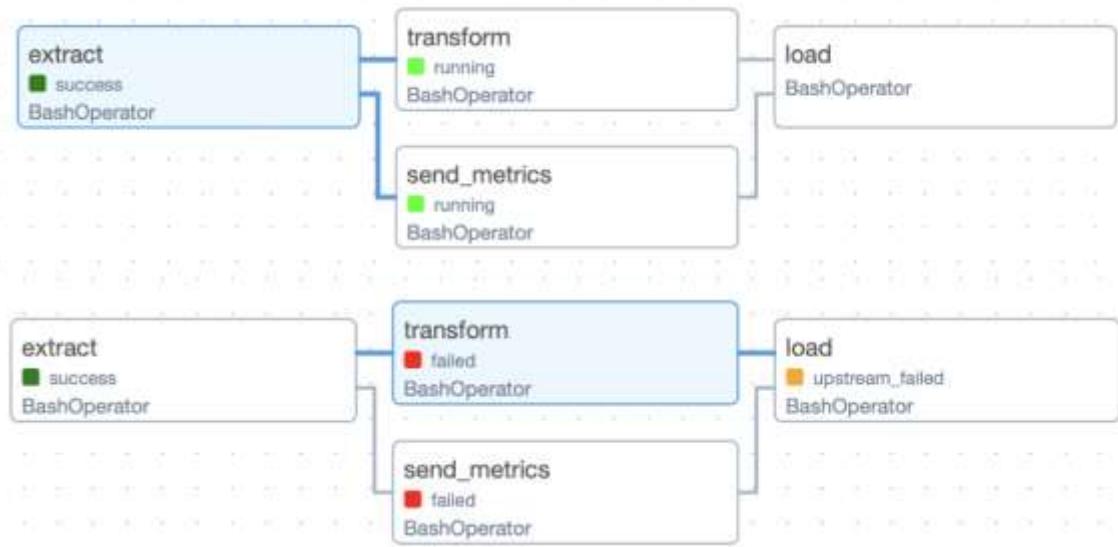


Fig. 7.4-1 Apache Airflow provides a user-friendly interface where you can visualize DAG -ETL, view execution logs, task startup status, and more.

Airflow is widely used for orchestration and automation of distributed computing, data processing, ETL (Extract, Transform, Load) process management, task scheduling, and other data scenarios. By default, Apache Airflow uses SQLite as the database.

An example of a simple DAG, similar to ETL, consists of tasks - Extract, Transform and Load. In the graph, which is controlled through the user interface (Fig. 7.4-1), the order of execution of tasks (code fragments) is defined: for example, extract is executed first, then transform (and sending_metrics), and load task completes the work. When all tasks are completed, the data loading process is considered successful.

Apache Airflow: practical application on ETL automation

Apache Airflow is widely used to organize complex data processing processes, allowing to build flexible ETL -conveyors. Apache Airflow can be run either through a web interface or programmatically through Python code (Fig. 7.4-2). In the web interface (Fig. 7.4-3), administrators and developers can visually track DAGs, run tasks, and analyze execution results.

Using DAG, you can set a clear sequence of tasks, manage dependencies between them and automatically react to changes in the source data. Let's consider an example of using Airflow to automate reporting processing (Fig. 7.4-2).

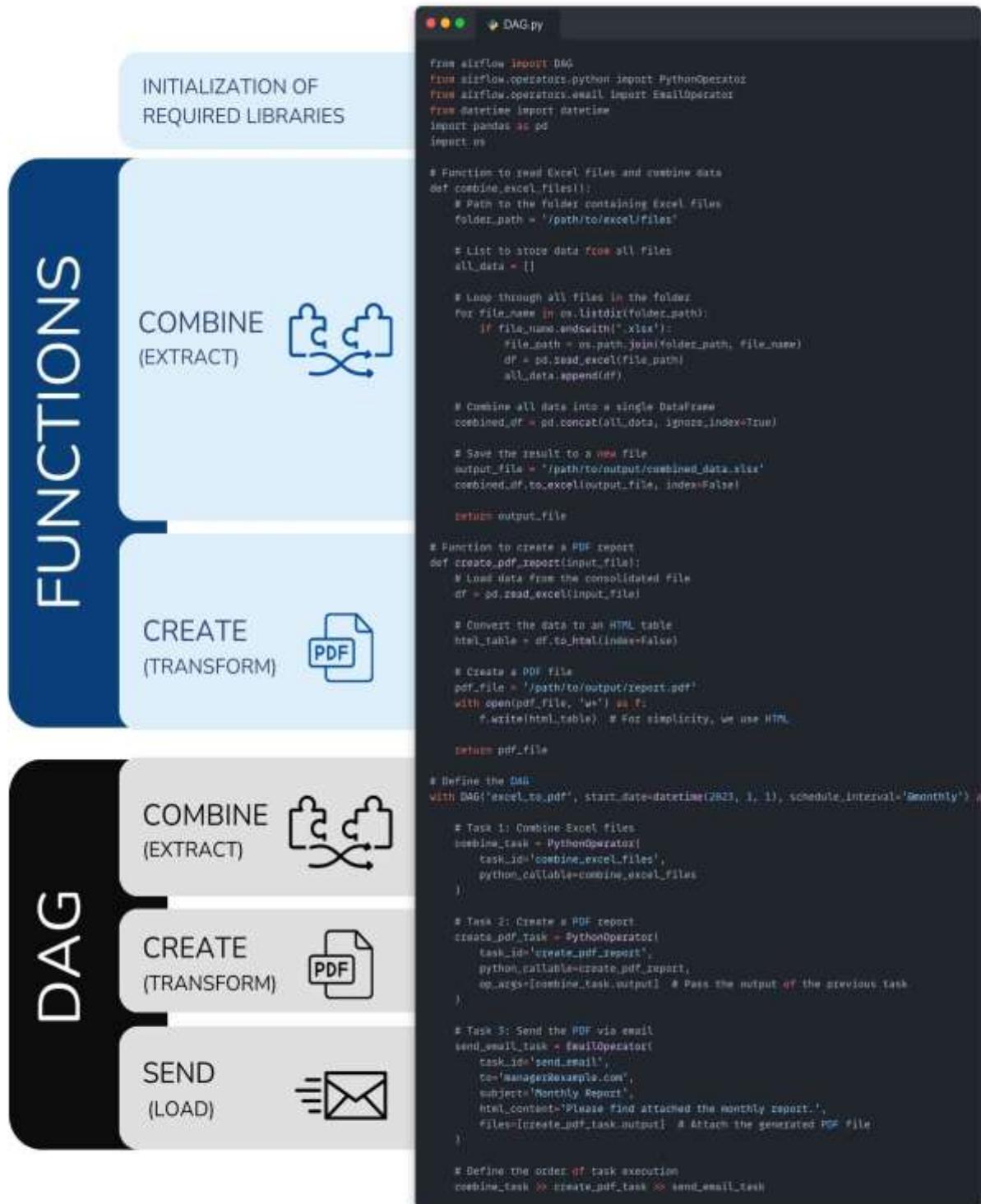


Fig. 7.4-2 ETL -conveyor concept for data processing using Apache Airflow.

This example (Fig. 7.4-2) considers the DAG, which performs key tasks within the ETL -conveyor:

ETL (EXTRACT, TRANSFORM, LOAD) PIPELINE

■ Read Excel -files (Extract):

- Sequential traversal of all files in a given directory.
- Read data from each file using the pandas library.
- Combining all data into a single DataFrame.

■ Create PDF -document (Transform):

- Transform the merged DataFrame into an HTML -table.
- Save the table as PDF (in the demo version - via HTML).

■ Sending a report by e-mail (Load):

- Apply EmailOperator to send PDF -document by email.

■ Customizing DAG:

- Defining the sequence of tasks: extracting data → generating report → sending.
- Assigning a launch schedule (@monthly - first day of each month).

The automated ETL -example (Fig. 7.4-2) shows how to collect data from Excel -files, create a PDF -document, and email it. This is just one of many possible use cases for Airflow. This example can be adapted to any specific task to simplify and automate data processing.

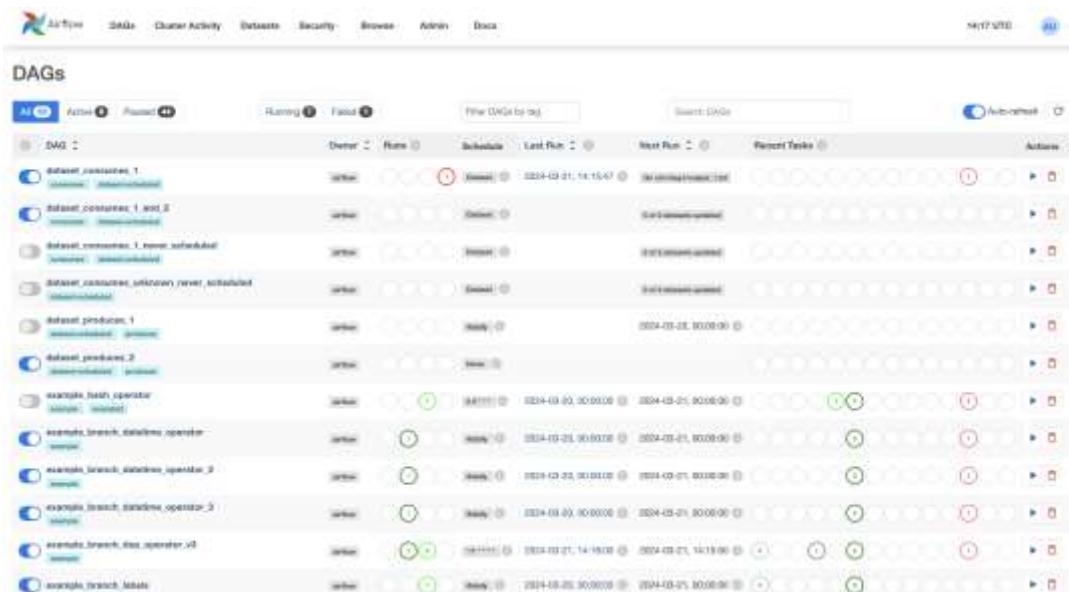


Fig. 7.4-3 Overview of all DAGs in the environment with information about recent runs.

The Apache Airflow web interface (Fig. 7.4-3) provides a comprehensive visual environment for managing data workflows. It displays DAGs as interactive graphs, with nodes representing tasks and edges representing dependencies between them, making it easy to keep track of complex data workflows. The interface includes a dashboard with information on task execution status, run history, detailed logs, and performance metrics. Administrators can manually start tasks, restart failed operations, suspend DAGs, and customize environment variables, all through an intuitive user interface.

Such architecture can be supplemented with data validation, notifications on execution status, integration with external APIs or databases. Airflow allows flexible customization of DAG: add new tasks, change their order, combine chains - which makes it an effective tool for automating complex data processing. When running DAG in the Airflow web interface (Fig. 7.4-3, Fig. 7.4-4), you can monitor the status of task execution. The system uses color indication:

- Green - the task has been successfully completed.
- Yellow - the process is in progress.
- Red - an error while performing the task.

In case of failures (e.g., missing file or broken data structure), the system automatically initiates sending a notification.

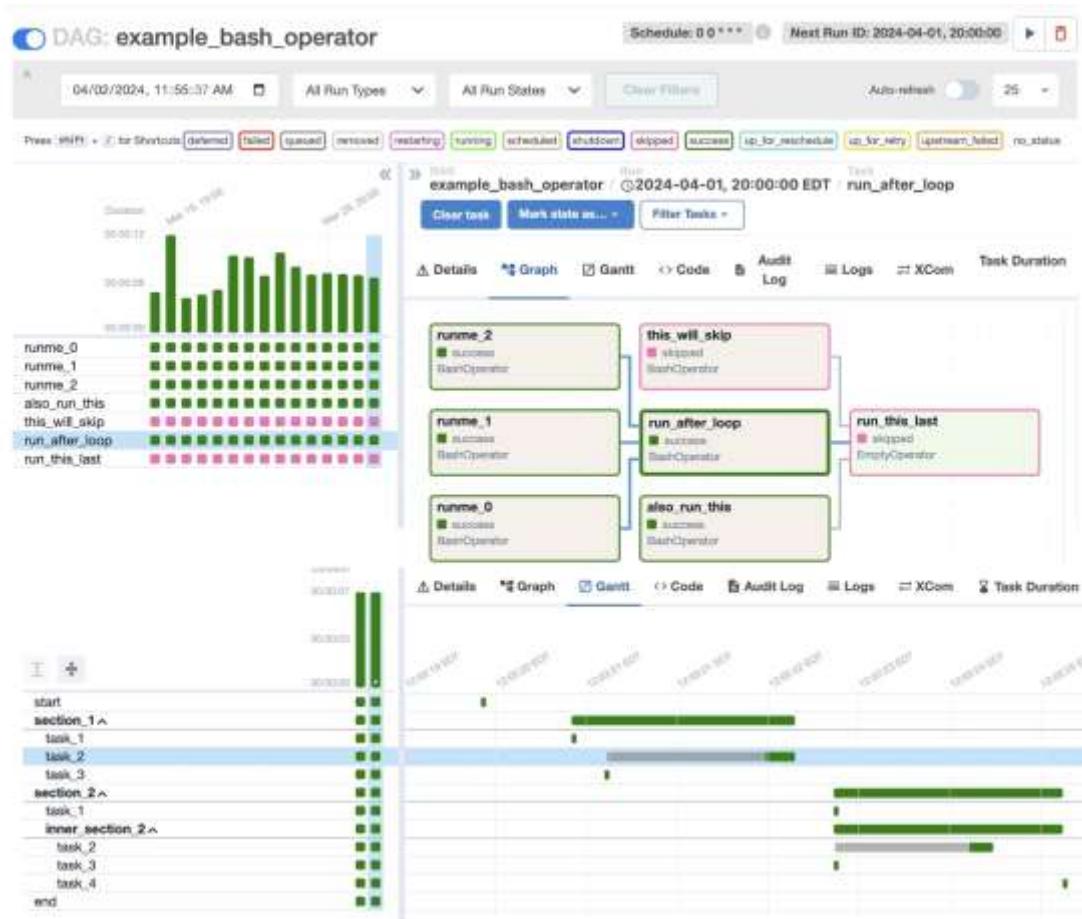


Fig. 7.4-4 Apache Airflow greatly simplifies problem diagnosis, process optimization, and team collaboration on complex data processing pipelines.

Apache Airflow is convenient because it automates routine tasks, eliminating the need to perform them manually. It provides reliability by monitoring process execution and instant error notification. The flexibility of the system makes it easy to add new tasks or modify existing ones, adapting workflows to meet changing requirements.

In addition to Apache Airflow, there are similar tools for orchestrating workflows. For example the open source and free Prefect (Fig. 7.3-5) offers a simpler syntax and integrates better with Python, Luigi, developed by Spotify, provides similar functionality and works well with big data. Also worth noting are Kronos and Dagster, which offer modern approaches to building Pipeline with a focus on modularity and scalability. The choice of task orchestration tool depends on the specific needs of the project, but they all help automate complex ETL data processes

Of particular note is Apache NiFi, an open source platform, designed for streaming and routing data. Unlike Airflow, which focuses on batch processing and dependency management, NiFi focuses on real-time, on-the-fly data transformation and flexible routing between systems.

Apache NiFi for routing and data conversion

Apache NiFi is a powerful open source platform, designed to automate data flows between different systems. Originally developed in 2006 by the US National Security Agency (NSA) under the name "Niagara Files" for internal use. In 2014, the project was open-sourced and transferred to the Apache Software Foundation, becoming part of their technology transfer initiatives [141].

Apache NiFi is designed to collect, process and transmit data in real time. Unlike Airflow, which works with batch tasks and requires well-defined schedules, NiFi operates in a stream processing mode, allowing for uninterrupted data transfer between different services.

Apache NiFi is ideal for integration with IoT devices, construction sensors, monitoring systems, and e.g. stream validation of CAD formats on a server where immediate response to changes in data may be required.

With built-in filtering, transformation and routing tools, NiFi allows you to standardize data (Transform) before transferring it (Load) to storage or analytics systems. One of its main advantages is its built-in security support and access control, making it a reliable solution for handling sensitive information.

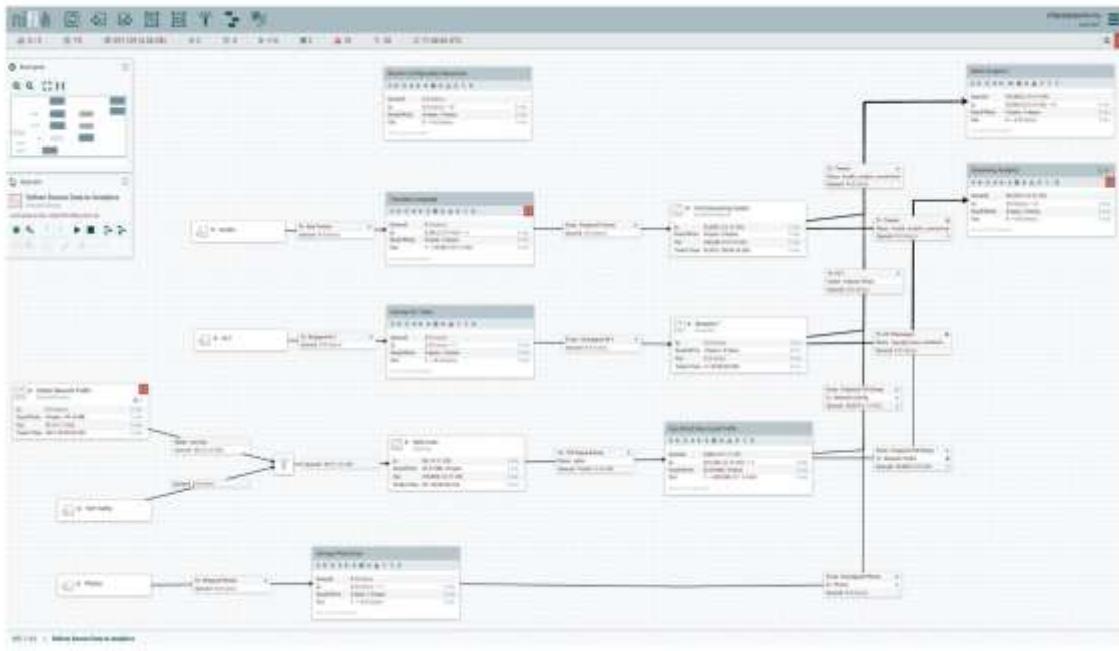


Fig. 7.4-5 Graphical representation of data flow in the Apache NiFi interface.

Apache NiFi efficiently handles real-time data streaming, filtering, and routing tasks. It is ideal for technically intensive scenarios where stable information transfer between systems and high throughput are important.

However, in cases where the main goal is to integrate various services, automate routine operations and quickly set up workflows without deep programming knowledge, solutions with a low entry threshold and maximum flexibility are in demand. One such tool is n8n - a Low-Code /No-Code class platform focused on business automation and visual process orchestration.

n8n Low-Code, No-Code process orchestration

n8n is an Open Source Low-Code / No-Code platform for building automated workflows, characterized by ease of use, flexibility and the ability to quickly integrate with a wide range of external services.

No-Code is a method of creating digital products without writing code. All elements of the process - from logic to interface - are realized exclusively with visual tools. No-Code platforms target users with no technical background and allow you to quickly create automations, forms, integrations and web applications. Example: a user configures automatic notification sending or Google Sheets integration through a drag-and-drop interface without programming knowledge.

With open source and local deployment capabilities, n8n in automation and ETL Pipelines creation

processes gives companies complete control over their data while ensuring security and independence from cloud providers.

Unlike Apache Airflow, which is oriented toward computational tasks with tight orchestration and requires knowledge of Python, n8n provides a visual editor that allows scripting without requiring knowledge of programming languages (Fig. 7.4-6). Although its interface allows the creation of automated processes without writing code (No-Code), in more complex scenarios users can add their own JavaScript and Python -functions to extend the capabilities (Low-Code).

Low-Code is an approach to software development in which the basic logic of an application or process is created using a graphical interface and visual elements, and program code is used only to customize or extend functionality. Low-Code platforms allow to significantly accelerate the development of solutions by involving not only programmers, but also business users with basic technical skills. Example: a user can build a business process from ready-made blocks and, if necessary, add their own script in JavaScript or Python.

Although n8n is positioned as a platform with a low entry threshold, basic programming knowledge, understanding of web technologies and skills in working with API. The flexibility of the system allows it to be adapted to a wide range of tasks - from automated data processing to integration with messengers, IoT -devices and cloud services.

Key features and benefits of using the n8n:

- **Open source** and local deployment capabilities ensure complete data control, security compliance, and independence from cloud providers.
- **Integration with over 330 services** including CRM, ERP, e-commerce, cloud platforms, messengers and databases.
- **Scenario flexibility:** from simple notifications to complex chains with API processing of -requests, decision logic and connection of AI -services.
- **Support for JavaScript and Python:** users can embed custom code as needed, extending automation capabilities.
- **Intuitive visual interface:** allows you to quickly configure and visualize all process steps.

Low-Code class platforms provide tools to create digital solutions with minimal code, making them ideal for teams that don't have deep technical expertise but need to automate processes.

In construction, n8n can be used to automate a variety of processes such as integrating with project management systems, streaming inspections, writing out-of-the-box reports and letters, automatically updating material inventory data, sending task status notifications to teams, and more. A customized Pipeline in n8n can reduce manual operations by multiples, reduce the likelihood of errors, and speed up decision making for project execution.

You can choose from nearly two thousand off-the-shelf, free and open source n8n Pipeline, available at: n8n.io/workflows, to automate both construction workflows and personal tasks, reducing routine operations.

Take one of the ready-made Pipeline templates, available for free at n8n.io [142], which automatically creates draft responses in Gmail (Fig. 7.4-6), helping users who receive a high volume of emails or have difficulty composing responses.

This n8n "Gmail AI Auto-Responder: Create draft responses to incoming emails" template (Fig. 7.4-6) analyzes incoming emails using LLM from ChatGPT, determines if a response is needed, generates a draft from ChatGPT and converts the text to HTML and adds it to the message chain in Gmail. This does not automatically send the email, allowing you to manually edit and approve the response. Setup takes about 10 minutes and includes OAuth configuration of the Gmail API and OpenAI API integration. The result is a convenient and free solution for automating routine email communication without losing control over the content of emails.

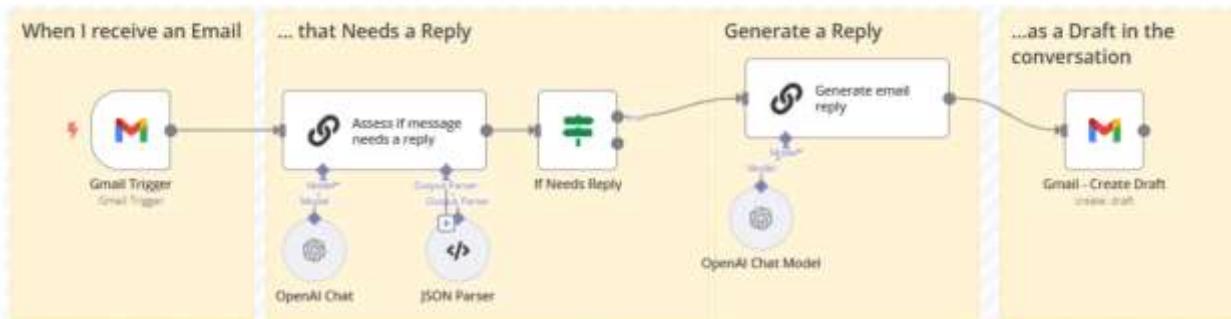


Fig. 7.4-6 Automated email response generation process using n8n.

Another example of automation with n8n is finding great deals in the real estate market [143]. N8n Pipeline "Automating Daily Real Estate Deals with Zillow API, Google Sheets and Gmail", collects daily relevant offers matching specified criteria using Zillow API. It automatically calculates key investment metrics (Cash on Cash ROI, Monthly Cash Flow, Down Payment), updates Google Sheets and sends a summary report to email (Fig. 7.4-7), allowing investors to save time and respond quickly to the best offers.

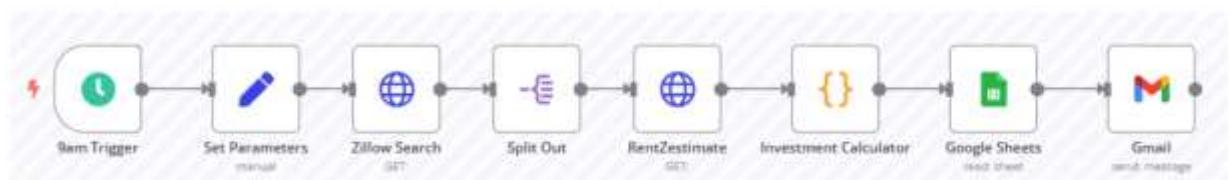


Fig. 7.4-7 Automated process for assessing the investment attractiveness of real estate.

With its flexibility and extensibility, n8n becomes a valuable tool for companies seeking to digitally transform and become more competitive in the marketplace with relatively simple and free open

source tools.

Tools such as Apache NiFi, Airflow, and n8n can be thought of as three layers of data processing (Fig. 7.4-8). NiFi manages the flow of data, ensuring its delivery and transformation, Airflow orchestrates task execution by aggregating data into processing pipelines, and n8n automates integration with external services and manages business logic.

	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

Fig. 7.4-8 Apache Airflow, Apache NiFi and n8n can be viewed as three complementary layers of modern data management architecture.

Together, these free and open source tools potentially form an example of an effective ecosystem for data and process management in the construction industry, allowing companies to leverage information for decision making and process automation.

Next steps: moving from manual operations to analytics-based solutions

Today's construction companies operate in an environment of high uncertainty: changing material prices, delayed deliveries, labor shortages and tight project deadlines. The use of analytical dashboards, ETL -conveyors and BI systems helps companies quickly find problem areas, assess resource efficiency and predict changes before they lead to financial losses.

To summarize this part, it is worth highlighting the main practical steps that will help you apply the discussed technologies in your daily tasks:

- Implement data visualization and analytics dashboards
 - Master the process of creating dashboards to monitor key performance indicators (KPIs)
 - Use visualization tools for your data (Power BI, Tableau, Matplotlib, Plotly)
- Automate data processing through ETL -processes
 - Set up automatic data collection from various sources (documentation, tables, CAD)

via ETL processes

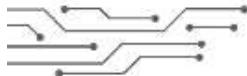
- Organize data transformation (e.g., regular expression validation or calculation) using Python scripts
- Try setting up automatic PDF (or DOC) reporting with the FPDF library, using data from Excel files or extracting information from other PDF documents

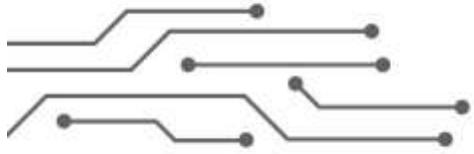
■ Use language models (LLM) to automate the

- Use large language models (LLMs), to generate code to help extract and analyze data from unstructured documents
- Familiarize yourself with n8n's automation tool and explore ready-made templates and case studies on their website. Determine which processes from your work can be fully automated using the No-Code/Low-Code approach

Analytical approach to data and automation of processes not only reduces time for routine operations, but also improves the quality of decisions. Companies that implement visual analytics tools and ETL - conveyors get an opportunity to react quickly to changes

Automating business processes using tools like n8n, Airflow and NiFi is only the first step to digital maturity. The next step is the quality storage and management of the very data that underpins the automation. In Part 8, we take an in-depth look at how construction companies can build a sustainable data storage architecture, moving from a chaos of documents and multi-format files to centralized storage and analytics platforms.





VIII PART

DATA STORAGE AND MANAGEMENT IN CONSTRUCTION

Part 8 explores modern data storage and management technologies for the construction industry. It analyzes efficient formats for handling large amounts of information - from simple CSV and XLSX to the more productive Apache Parquet and ORC with a detailed comparison of their capabilities and limitations. The concepts of data warehouses (DWH), data lakes (Data Lakes) and their hybrid solutions (Data Lakehouse), as well as the principles of data governance (Data Governance) and data minimalism (Data Minimalism) are discussed. The problems of Data Swamp and strategies to prevent chaos in information systems are covered in detail. New approaches to working with data are presented, including vector databases and their application in construction through the concept of Bounding Box. This part also touches upon the DataOps and VectorOps methodologies as new standards for organizing data workflows.

CHAPTER 8.1.

DATA INFRASTRUCTURE: FROM STORAGE FORMATS TO DIGITAL REPOSITORIES

Data atoms: the foundation of effective information management

Everything in the Universe consists of the smallest building blocks - atoms and molecules, and over time all living and non-living things inevitably return to this initial state. In nature, this process occurs with astonishing speed, which we are trying to transfer to the processes controlled by man.

In the forest, any living organisms are eventually transformed into a nutritious substance that serves as the basis for new plants. These plants, in turn, become food for new living things made up of the same atoms that created the universe millions of years ago.

In the business world, it is also important to break down complex, multi-layered structures into their most fundamental, minimally processed units - much like atoms and molecules in nature. This allows atoms of data to be efficiently stored and managed, turning them into a rich, fertile foundation that becomes a key resource for the growth of analytics and decision quality.



Fig. 8.1-1 Analysis and decision making is based on reused data that was once processed and stored.

Musical compositions are made up of notes that combine to create complex pieces of music, while

words are created from a primitive unit, the letter-sound. Whether it is nature, science, economics, art or technology, the world exhibits remarkable unity and harmony in its pursuit of destruction, structure, cyclical and creation. Similarly, processes in costing systems are broken down into tiny structured units - resource items - at the level of costing and schedules. These units, like notes, are then used to form more complex calculations and schedules. The same principle is used by computer-aided design systems, in which complex architectural and engineering projects are built from basic elements - individual elements and library components, from which a complete 3D -model of the project of a complex building or structure is created.

The concept of cyclical and structure inherent in nature and science is also reflected in the modern world of data. Just as in nature all living things revert to atoms and molecules, so in the world of modern data processing tools, information tends to revert to its most primitive form.

The smallest elements with their finite indivisibility are the basic building blocks of business processes. It is important to carefully consider from the very beginning how to collect, structure (break down into atoms) and store these tiny building blocks from various sources. At the same time, organizing and storing data is not just a matter of breaking it down into its constituent parts. It is equally important to ensure that they are integrated and stored in a structured way so that data can be easily retrieved, analyzed and used for decision-making whenever it is needed.

To process information efficiently, the format and methods of data storage must be carefully chosen - just as the soil must be prepared for the growth of trees. Data warehouses should be organized to ensure high quality and relevance of information, eliminating redundant or irrelevant data. The better this "information soil" is structured, the faster and more accurately users can find the data they need and solve analytical problems.

Information storage: files or data

Data warehouses allow companies to collect and combine information from different systems, creating a single center for subsequent analytics. Collected historical data enables not only deeper analysis of processes, but also the identification of patterns that can affect business performance.

Let's say a company is working on several projects at the same time. An engineer wants to understand how much concrete has been poured and how much is still to be purchased. In a traditional approach, he would have to manually search the server and open several estimate tables, compare them with the certificates of completed work and check the current stock balances. This takes hours, if not days. Even with ETL processes and automatic scripts, the task remains semi-manual: the engineer still has to manually specify the path to folders or specific files on the server. This reduces the overall effect of automation, as it continues to take up valuable work time.

When switching to data management, instead of working with the server file system, the engineer gets access to a unified storage structure where information is updated in real time. A single query - in the

form of a code, SQL -query or even a call to an LLM -agent - can instantly provide accurate data on current balances, volumes of work performed and upcoming deliveries, if the data has been prepared in advance and combined in a data warehouse where there is no need to wander through folders, open dozens of files and manually compare values.

For a long time construction companies used PDF -documents, DWG -drawings, RVT -models and hundreds and thousands of Excel -tables and other disparate formats, which are stored in certain folders on the company's servers, making it difficult to find information, check and analyze it. As a result, the files left after projects are completed are most often moved back to the server into archival storage folders, which are practically not used in the future. Such traditional file-based data storage loses relevance with increasing data flow due to its vulnerability to human error.

A file is just an isolated container in which data is stored. Files are created for people, not systems, so they require manual opening, reading and interpretation. An example would be an Excel -table, a PDF -document, or a CAD -drawing that needs to be specifically opened in a certain tool to access the desired information. Without structured retrieval and processing, the information in it remains unused.

Data, in turn, is machine-readable information that is linked, updated, and analyzed automatically. In a single data warehouse (e.g., database, DWH or Data Lake), information is represented as tables, records, and relationships. This enables uniform storage, automated queries, value analysis, and real-time reporting.

Using data instead of files (Fig. 8.1-1) makes it possible to eliminate the manual search process and unify processing. Companies that have already implemented such an approach gain a competitive advantage due to the speed of access to information and the ability to quickly integrate it into business processes.

The shift from file-based to data-driven is an inevitable change that will shape the future of the construction industry.

Every company in the construction industry will face a key choice: continue to store information in disparate files and silos that must be read by humans using special programs or transform it in the first stages of processing into structured data, creating a single integrated digital foundation for automated project management.

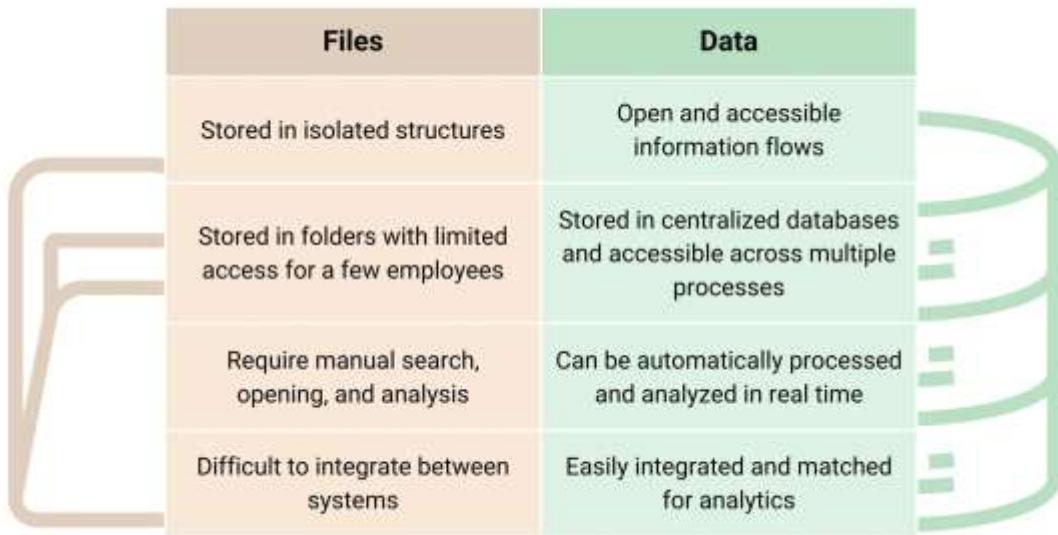


Fig. 8.1-1 Evolution of information flow: from isolated files to integrated data.

With the explosion of information, traditional methods of storing and processing files are becoming less and less efficient. In the construction industry, as in other sectors, it is no longer sufficient to rely on disparate file folders with different file formats or unconnected databases.

Companies seeking to remain competitive in the digital age will inevitably shift to integrated digital platforms, utilize big data technologies and automated analytics systems.

Moving from file storage to data will require a rethinking of information management approaches and a conscious choice of formats suitable for further integration into centralized repositories. This choice will determine how efficiently data can be processed, how quickly it can be accessed, and how easily it can be integrated into a company's digital processes.

Big Data Storage: Analyzing Popular Formats and Their Effectiveness

Storage formats play a key role in the scalability, reliability, and performance of analytics infrastructure. For data analysis and processing - such as filtering, grouping, and aggregation - our examples used Pandas DataFrame - a popular structure for working with data in RAM.

However, Pandas DataFrame does not have its own storage format, so after processing is complete, the data is exported to one of the external formats - most often CSV or XLSX. These tabular formats are easy to exchange and compatible with most external systems, but have a number of limitations: low storage efficiency, lack of compression and poor versioning support:

- **CSV (Comma-Separated Values):** a simple text format widely supported by various platforms and tools. It is easy to use, but does not support complex data types or compression.
- **XLSX (Excel Open XML Spreadsheet):** a Microsoft Excel file format that supports sophisticated features such as formulas, charts, and styling. While it is convenient for manual data analysis

and visualization, it is not optimized for large-scale data processing.

In addition to the popular tabular XLSX and CSV, there are several popular formats for efficiently storing structured data (Fig. 8.1-2), each with unique advantages depending on specific data storage and analysis requirements:

- **Apache Parquet:** a columnar data storage file format optimized for use in data analysis systems. It offers efficient data compression and encoding schemes, making it ideal for complex data structures and big data processing.
- **Apache ORC (Optimized Row Columnar):** Similar to Parquet, ORC provides high compression and efficient data storage. It is optimized for heavy read operations and is well suited for storing data lakes.
- **JSON (JavaScript Object Notation):** although JSON is not as efficient in terms of data storage compared to binary formats such as Parquet or ORC, it is very accessible and easy to work with, making it ideal for scripts where readability and web compatibility are important.
- **Feather:** a fast, lightweight, and easy-to-use analytics-oriented binary columnar data storage format. It is designed to efficiently transfer data between Python (Pandas) and R, making it an excellent choice for projects involving these programming environments.
- **HDF5 (Hierarchical Data Format version 5):** designed for storing and organizing large amounts of data. It supports a wide range of data types and is well suited for working with complex collections of data. HDF5 is particularly popular in scientific computing due to its ability to efficiently store and access large datasets.

	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)

Fig. 8.1-2 Comparison of data formats showing the main differences in storage and processing aspects.

To conduct a comparative analysis of the formats used at the Load stage of the ETL process, a table showing file sizes and reading times was created (Fig. 8.1-3). Files with identical data were used in the study: the table contained 10,000 rows and 10 columns filled with random values.

The following storage formats are included in the study: CSV, Parquet, XLSX and HDF5, as well as their compressed versions in ZIP archives. The raw data were generated using the NumPy library and represented as a Pandas DataFrame structure. The testing process consisted of the following stages:

- File saving: the dataframe is saved in four different formats: CSV, Parquet, XLSX, and HDF5. Each format has unique features in the way it stores data, affecting file size and read speed.
- ZIP file compression: to analyze the effectiveness of standard compression, each file was further compressed into a ZIP archive.
- File reading (ETL - Load): reading time was measured for each file after its unpacking from ZIP. This allows estimating the speed of data access after extraction from the archive.

It is important to note that Pandas DataFrame was not used directly in the size or read time analysis, as it does not represent an independent storage format. It served only as a intermediate structure for generation and subsequent saving of data into different formats.



Fig. 8.1-3 Comparison of storage formats by size and read speed.

CSV and HDF5 files exhibit (Fig. 8.1-3) high compression efficiency, significantly reducing their size when packed in ZIP, which can be particularly useful in scenarios requiring storage optimization. XLSX files, on the other hand, are virtually uncompressible and their size in ZIP remains comparable to the original, making them less favorable for use in large data volumes or in environments where speed of data access is important. In addition, the read time for XLSX is significantly higher compared to other formats, making it less preferable for fast data read operations. Apache Parquet has demonstrated high performance for analytical tasks and large data volumes due to its columnar structure.

Optimize storage with Apache Parquet

One of the popular formats for storing and processing big data is Apache Parquet. This format is designed specifically for columnar storage (similar to Pandas), which allows you to significantly reduce memory footprint and increase the speed of analytical queries. Unlike traditional formats such as CSV and XLSX, Parquet supports native compression and is optimized for big data systems including Spark, Hadoop, and cloud storage.

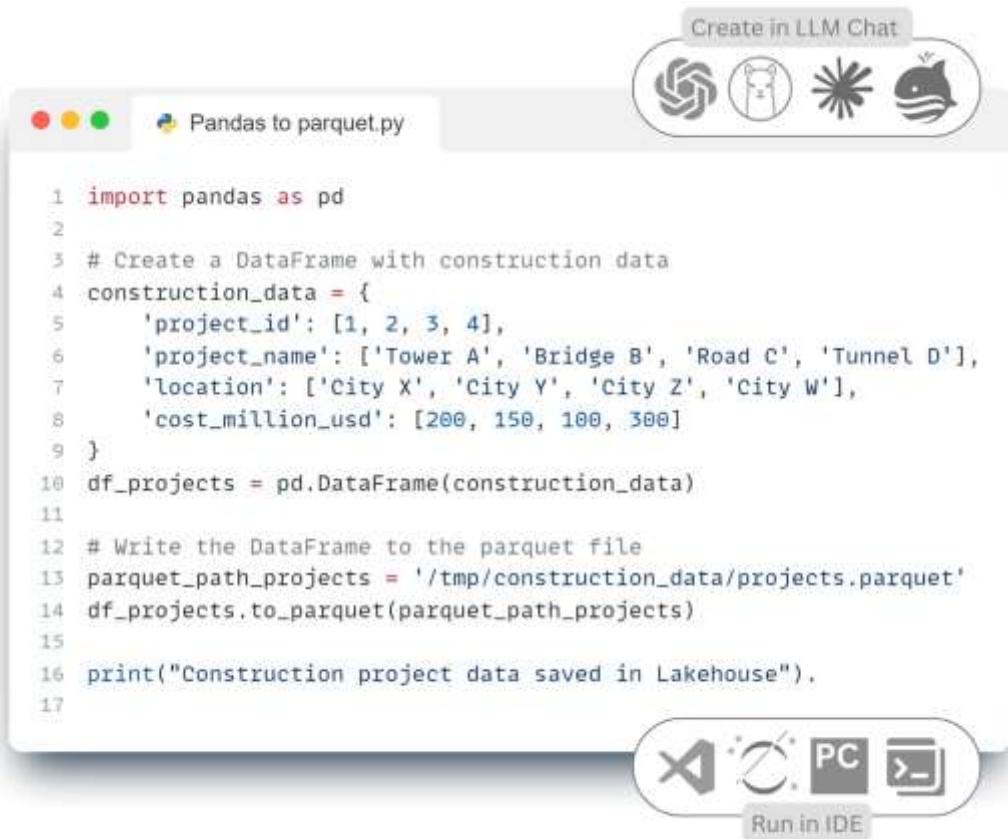
Key features of Parquet include support for data compression and encoding, which significantly reduces storage size and speeds up data read operations by working directly with the desired columns rather than all rows of data.

For an illustrative example of how easy it is to get the necessary code to convert data to Apache Parquet, let's use the LLM.

- Send a text request to LLM chat (CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN):

Write code to save data from Pandas DataFrame to Apache Parquet. ↵

- LLM's response:



The screenshot shows a LLM chat interface with a message input field containing the text "Write code to save data from Pandas DataFrame to Apache Parquet. ↵". Below the input field, a response is displayed in a code editor window titled "Pandas to parquet.py". The code is as follows:

```

1 import pandas as pd
2
3 # Create a DataFrame with construction data
4 construction_data = {
5     'project_id': [1, 2, 3, 4],
6     'project_name': ['Tower A', 'Bridge B', 'Road C', 'Tunnel D'],
7     'location': ['City X', 'City Y', 'City Z', 'City W'],
8     'cost_million_usd': [200, 150, 100, 300]
9 }
10 df_projects = pd.DataFrame(construction_data)
11
12 # Write the DataFrame to the parquet file
13 parquet_path_projects = '/tmp/construction_data/projects.parquet'
14 df_projects.to_parquet(parquet_path_projects)
15
16 print("Construction project data saved in Lakehouse").
17

```

At the bottom of the code editor, there are icons for "Run in IDE" and "PC".

Fig. 8.1-4 Transferring Dataframe data from RAM to a storage-efficient format Apache Parquet using a few lines of Python.

Next example: let's simulate the ETL process with data stored in Parquet format to filter projects by a certain value of one of the attributes "cost_million_usd" (Fig. 8.1-4).

- As the chat continues, send a text request to LLM:

Write code where we want to filter the data in a table and save only those projects (table rows) from Apache Parquet data, whose cost (parameter cost_million_usd) exceeds 150 million dollars. ↴

🔗 LLM's response:

```

1 # Loading construction project data
2 df_loaded_projects = pd.read_parquet(parquet_path_projects)
3
4 # Filtering out projects with a value greater than $150 million
5 df_filtered_projects = df_loaded_projects[df_loaded_projects['cost_million_usd'] > 150]
6
7 # Saving the filtered data to a new Parquet file
8 filtered_parquet_path_projects = '/tmp/construction_data/filtered_projects.parquet'
9 df_filtered_projects.to_parquet(filtered_parquet_path_projects)
10
11 print("Filtered construction projects data saved to Lakehouse.")

```

Fig. 8.1-5 The ETL process when working with data in the Apache Parquet format looks the same as with other structured formats.

Using the Parquet format (in relation to XLSX, CSV, etc.) significantly reduces the amount of information stored and speeds up search operations. This makes it excellent for both storing and analyzing data. Parquet integrates with various processing systems, providing efficient access in hybrid architectures.

However, an efficient storage format is only one element of a complete data experience. To create a sustainable and scalable environment, a well-designed data management architecture is required. This is exactly what DWH (Data Warehouse) class systems do. They provide aggregation of data from heterogeneous sources, transparency of business processes and the possibility of complex analysis using BI-tools and machine learning algorithms.

DWH: Data Warehouse data warehouses

Just as the Parquet format is optimized for efficient storage of large amounts of information, the Data Warehouse is optimized for integrating and structuring data to support analytics, forecasting and management decision making.

In today's companies, data comes from many disparate sources: ERP, CAFM, CPM, CRM systems, accounting and warehousing, digital CAD models of buildings, IoT sensors and other solutions. To get

a holistic picture, it's not enough to just collect data - it needs to be organized, standardized and centralized in a single repository. This is exactly what DWH does - a centralized storage system that allows you to aggregate information from various sources, structure it and make it available for analytics and strategic management.

DWH (Data Warehouse) is a centralized data warehouse system that aggregates information from multiple sources, structures it, and makes it available for analytics and reporting.

In many companies, data are scattered across different systems, which we discussed in the first parts of the book (Fig. 1.2-4). DWH integrates these sources, ensuring complete transparency and reliability of information. A DWH data warehouse is a specialized database (a large database) that collects, processes, and stores data from multiple sources. The main characteristics of a DWH are:

- **Using ETL -processes** (Extract, Transform, Load) - extracting data from sources, cleaning it, transforming it, loading it into the repository, and automating these processes discussed in Part 7 of the book.
- **Data granularity** - data in DWH can be stored in both aggregated form (summary reports) and in granular form (raw data). From 2024 onwards, it is CAD-vendors that have started talking about granular data [125], possibly indicating that the industry is preparing for the transition to specialized cloud storage for handling digital building model data.
- **Supporting analytics and predictive analytics** - data warehouses provide the foundation for BI tools, Big Data -analysis and machine learning.

DWH serves as the foundation for business intelligence, enabling analysis of key performance indicators, forecasting of sales, purchases, and costs, and automated reporting and data visualization (Fig. 8.1-6).

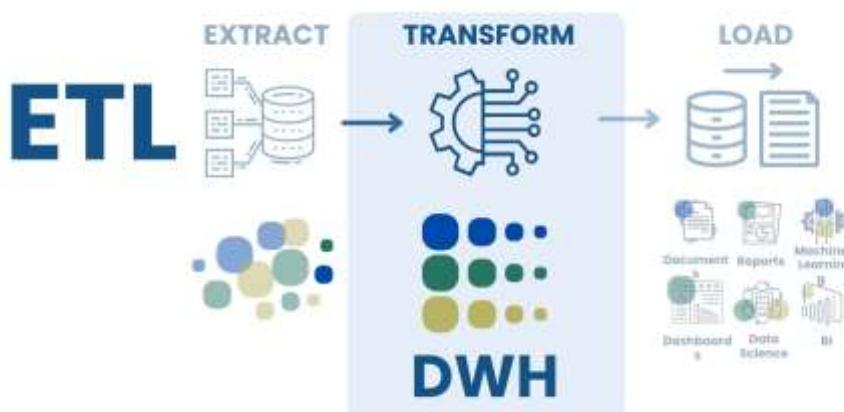


Fig. 8.1-6 In an ETL -process, DWH can act as a central repository where data extracted from various systems undergoes transformation and offloading steps.

DWH plays a key role in integrating, cleansing and structuring information, providing a solid foundation for business intelligence and decision-making processes. However, in today's environment where data volumes are growing rapidly and data sources are becoming increasingly diverse, the traditional DWH approach to information storage often requires extension in the form of ELT and Data Lake approaches.

Data Lake - evolution of ETL to ELT: from traditional cleaning to flexible processing

Classic DWH - data warehouses, designed to store structured data in a format optimized for analytical queries, have faced limitations in handling unstructured data and scalability. In response to these challenges, Data Lakes) have emerged, offering flexible storage for large amounts of heterogeneous data.

Data Lake offers an alternative DWH -approach that allows working with unstructured, semi-structured and raw data without a prior hard schema. This storage method is often relevant for real-time data processing, machine learning and advanced analytics. Unlike DWH, which structures and aggregates data before loading, Data Lake allows information to be stored in its raw form, thus providing flexibility and scalability

It was frustration with traditional data warehouses (RDBMS, DWH) and interest in "big data" that led to the emergence of data lakes, where instead of complex ETL, data is now simply loaded into a loosely structured repository, and its processing takes place already at the analysis stage:

- In traditional data warehouses, data is typically pre-processed, transformed and cleaned (ETL - Extract, Transform, Load) before being loaded into the warehouse (Fig. 8.1-6). This means that data is structured and optimized for specific future analytics and reporting tasks. The emphasis is on maintaining high query performance and data integrity. However, this approach can be costly and less flexible in terms of integrating new data types and rapidly changing data schemas.
- Data lakes, on the other hand, are designed to store large amounts of raw data in its original format (Fig. 8.1-7). The ETL (Extract, Transform, Load), process is being replaced by ELT (Extract, Load, Transform), where data is first loaded into storage "as is" and only then can be transformed and analyzed as needed. This provides greater flexibility and the ability to store heterogeneous data, including unstructured data such as text, images and logs.

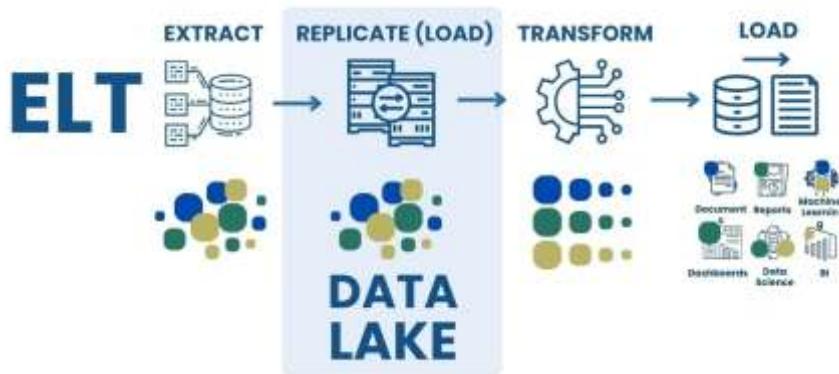


Fig. 8.1-7 Unlike ETL, Data Lake uses ELT, in which information is first loaded in "raw" form, and transformation is performed at the upload stage.

Traditional data warehouses focus on pre-processing data for high query performance, while data lakes prioritize flexibility: they store raw data and transform it as needed (Fig. 8.1-8).

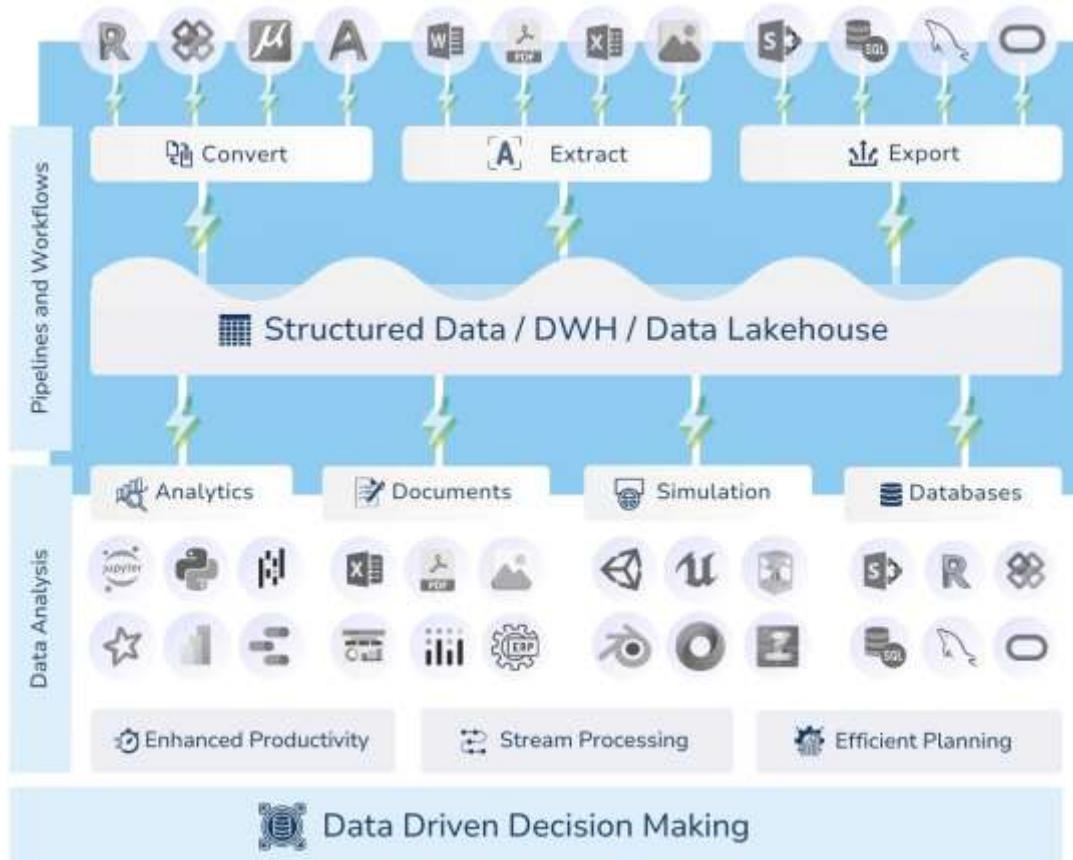


Fig. 8.1-8 Modern storage concepts aim to store and process all types of data for decision-making purposes.

However, despite all the advantages and data lakes are not without disadvantages. The lack of a strict structure and the complexity of information management can lead to chaos in which data is duplicated,

contradicts each other or loses its relevance. In addition, searching and analyzing data in such a repository requires considerable effort, especially when dealing with heterogeneous information. To overcome these limitations and combine the best features of traditional data warehouses and data lakes, the Data Lakehouse architecture was developed.

Data Lakehouse architecture: synergy of warehouses and data lakes

To combine the best features of DWH (structured, manageable, high performance analytics) and Data Lake (scalability, handling heterogeneous data), the Data Lakehouse approach was developed. This architecture combines the flexibility of data lakes with the powerful processing and management tools typical of traditional warehouses, striking a balance between storage, analytics and machine learning. Data Lakehouse is a synthesis of data lakes and data warehouses, combining the flexibility and scalability of the former with the manageability and query optimization of the latter.

Data Lakehouse is an architectural approach that seeks to combine the flexibility and scalability of data lakes with the manageability and query performance of data warehouses (Fig. 8.1-9).

Key features of the Data Lakehouse include:

- **Open storage format:** using open formats for data storage, such as Apache Parquet, provides efficiency and query optimization.
- **Read-only schema:** in contrast to the traditional approach of a write-only schema in DWH, Lakehouse supports a read-only schema, which allows more flexibility in managing the data structure.
- **Flexible and scalable:** supports storage and analysis of structured and unstructured data, providing high query performance through storage-level optimization.

Data Lakehouse offers a compromise solution that combines the benefits of both approaches, making it ideal for modern analytic workloads that require flexibility in data processing.

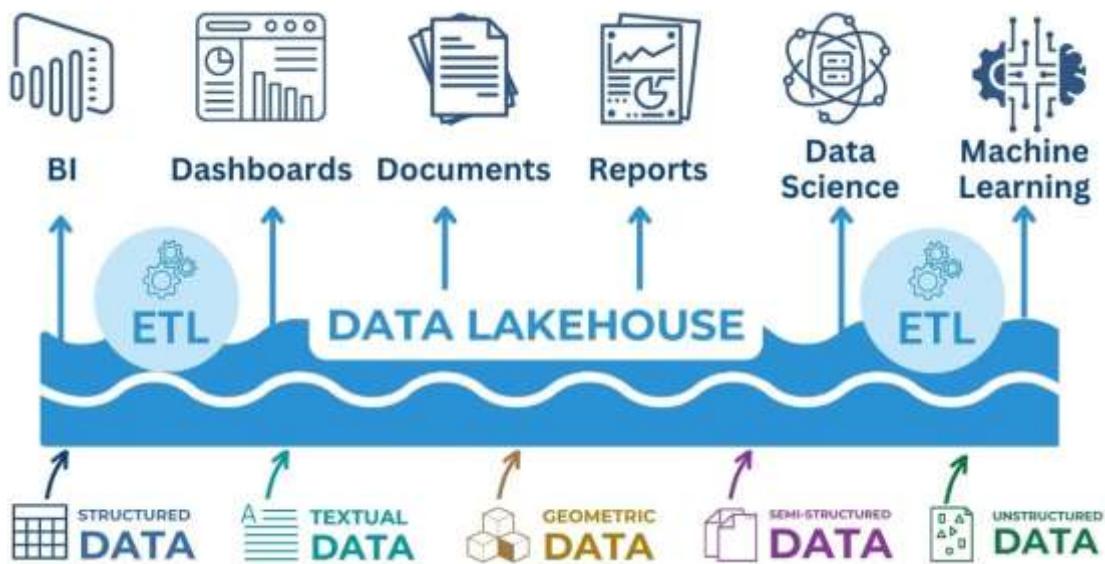


Fig. 8.1-9 Data Lakehouse is the next generation of storage systems designed to meet complex and ever-changing requirements.

The idea of modern data warehouses seems simple: if all the data is in one place, it is easier to analyze it. However, in practice everything is not so smooth. Imagine that a company decides to completely abandon the usual accounting and management systems (ERP, PMIS, CAFM or others), replacing them with one huge data lake to which everyone has access. What will happen? Most likely chaos will ensue: data will be duplicated, contradictory, and critical information will be lost or corrupted. Even if the data lake is used only for analytics, without proper management, it will be severely compromised:

- Data is difficult to understand: in conventional systems, data has a clear structure, but in a lake, it is just a huge accumulation of files and tables. To find something, specialist has to Fig. out what each row and column is responsible for.
- Data can be inaccurate: if many versions of the same information are stored in one place, it is difficult to know which version is up to date. As a result, decisions are made based on outdated or erroneous data.
- It is difficult to prepare data for work: the data must not only be stored, but also presented in a convenient form - in the form of reports, graphs, tables. In traditional systems this is done automatically, but in data lakes it requires additional processing.

The bottom line is that each data warehousing concept has its own characteristics, processing approaches and business applications. Traditional databases focus on transactional operations, data warehouses (DWH) provide a structure for analytics, data lakes (Data Lake) store information in raw form, and hybrid warehouses (Data Lakehouse) combine the advantages of DWH and Data Lake (Fig. 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

Fig. 8.1-10 DWH, Data Lake and Data Lakehouse: key differences in data types, usage scenarios, processing methods and storage approaches.

Choosing a storage architecture is a complex process, depending on business needs, information volume and analytics requirements. Each solution has its pros and cons: DWH provides structure, Data Lake provides flexibility, and Lakehouse provides a balance between the two. Organizations are rarely limited to a single data architecture.

Regardless of the chosen architecture, automated data management systems are significantly superior to manual methods. They minimize human errors, speed up information processing, and ensure transparency and traceability of data at all stages of business processes.

And while centralized data warehouses have already become an industry standard in many areas of the economy, the situation in construction remains fragmented. Data here is distributed across different platforms (CDE, PMIS, ERP, etc.), which makes it difficult to create a unified picture of what is happening and requires architectures capable of integrating these sources into a holistic, analytically usable digital environment.

CDE, PMIS, ERP or DWH and Data Lake

Some construction and engineering companies are already using the concept of Common Data Environment (CDE) according to ISO 19650. In essence, the CDE performs the same functions as a data warehouse (DWH) in other industries: centralizes information, provides version control, and provides access to validated information.

A Common Data Environment (CDE) is a centralized digital space used to manage, store, share and collaborate on project information throughout all phases of the facility lifecycle. CDE is often implemented using cloud-based technologies and integrated with CAD (BIM) systems.

The financial, retail, logistics and industrial sectors have been using centralized data management systems for decades, integrating information from different sources, controlling its relevance and providing analytics. CDE takes these principles further by adapting them to the challenges of building design and lifecycle management.

Like DWH, CDE structures data, captures changes and provides a single point of access to validated information. With the move to the cloud and integration with analytical tools, the differences between the two are becoming less and less apparent. Adding to CDE granular data, the concept of which has been discussed by CAD -vendors since 2023[93, 125], one can see even more parallels with classic DWH.

Earlier in the chapter "Construction ERP and PMIS systems" we have already discussed PMIS (Project Management Information System) and ERP (Enterprise Resource Planning). In construction projects, CDE and PMIS work together: CDE serves as a repository for data including drawings, models and project documentation, while PMIS manages processes such as controlling deadlines, tasks, resources and budgets.

ERP, responsible for business management as a whole (finance, procurement, personnel, production), can integrate with PMIS, providing control of costs and budgets at the company level. For analytics and reporting, DWH can be used to collect, structure and aggregate data from CDE, PMIS and ERP to evaluate financial KPIs (ROI) and identify patterns. In turn, Data Lake (DL) can complement DWH by storing raw and unstructured data (e.g., logs, sensor data, images). This data can be processed and loaded into DWH for further analysis.

Thus, CDE and PMIS focus on project management, ERP focuses on business processes, and DWH and Data Lake focuses on analytics and data operations.

In comparing CDE, PMIS and ERP systems with DWH and Data Lake, significant differences can be seen in terms of vendor independence, cost, integration flexibility, data independence, speed of adaptation to change, and analytical capabilities (Fig. 8.1-11). Traditional systems such as CDE, PMIS, and ERP are often tied to specific vendor solutions and standards, making them less flexible and increasing their cost due to licenses and support. In addition, data in these systems are often encapsulated in proprietary, closed formats, which limits their use and analysis.

		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)

Fig. 8.1-11 DWH and Data Lake offer greater flexibility and data independence than systems like CDE, PMIS and ERP.

In contrast, DWH and Data Lake provide greater flexibility in integrating with different data sources, and their use of open technologies and platforms helps reduce total cost of ownership. Moreover, DWH and Data Lake support a wide range of analytical tools, which enhances analytics and management capabilities.

With the development of reverse-engineering tools for CAD formats and the availability of access to CAD application databases, the question becomes more and more acute: how justified is it to continue using closed, isolated platforms if design data must be available to a wide range of specialists working in dozens of contractors and design organizations?

This vendor-specific technology dependency can significantly limit data management flexibility, slow responses to project changes, and inhibit effective collaboration between participants.

Traditional approaches to data management - including DWH, Data Lake, CDE and PMIS - have focused primarily on storing, structuring and processing information. However, with the development of artificial intelligence and machine learning, there is a growing need for new ways to organize data that not

only aggregate, but also identify complex relationships, find hidden patterns, and provide instant access to the most relevant information.

Vector databases - a new type of storage optimized for high-dimensional embeddings - are beginning to play a special role in this direction.



CHAPTER 8.2.

DATA WAREHOUSE MANAGEMENT AND CHAOS PREVENTION

Vector Databases and the Bounding Box

Vector databases are a new class of repositories that do not just store data, but allow searching by meaning, comparing objects by semantic proximity, and creating intelligent systems: from recommendations to automatic analysis and context generation. Unlike traditional databases that focus on exact matches, vector databases find similar objects based on attributes - even if there is no exact match

A vector database is a specialized type of database that stores data as multidimensional vectors, each representing certain characteristics or qualities. These vectors can have different numbers of dimensions, depending on the complexity of the data (in one case it may be a few dimensions, and in another — thousands).

The main advantage of vector databases is search by semantic relevance rather than by exact matching of values. Instead of SQL- and Pandas -queries with "equals" or "contains" filters, the search of nearest neighbors (k-NN) (we will talk more about k-NN in the next part of the book) in the feature space is used.

With the development of LLM (Large Language Models) and generative models, interaction with databases is beginning to change. It is now possible to query data in natural language, get semantic searches on documents, automatically extract key terms, and build contextual relationships between objects - all without the need for SQL proficiency or knowledge of table structure. This was discussed in more detail in the section "LLMs and their role in data processing and business processes".

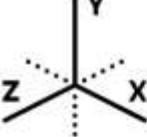
However, it is important to realize that LLMs do not automatically structure and organize information. The model just floats through the data and finds the most relevant piece of data based on the context of the query. If the data has not been pre-cleaned or transformed, deep search will be like trying to find an answer in digital "garbage" - it may work, but the quality of results will be lower. Ideally, if the data can be structured (e.g., translate documents into Markdown) and loaded into a vector database. This significantly increases the accuracy and relevance of the output.

Initially, vector databases were used in machine learning, but today they find more and more applications outside of it - in search engines, content personalization, and intelligent analytics.

One of the most obvious examples of the vector approach in construction is the Bounding Box (bounding parallelepiped). It is a geometric construction that describes the boundaries of an object in three-dimensional space. The Bounding Box is defined by the minimum and maximum X, Y and Z coordinates, forming a "box" around the object. This method allows you to estimate the size and placement

of an element without having to analyze the entire geometry.

Each Bounding Box can be represented as a vector in a multidimensional space: for example, [x, y, z, width, height, depth] - already 6 dimensions (Fig. 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

Fig. 8.2-1 Bounding Box -element coordinate information and their location in the project model is analogous to a vector database.

This data representation facilitates many tasks, including checking for intersections between objects, planning the spatial distribution of building elements, and performing automated calculations. Bounding Box can serve as a bridge between complex 3D models and traditional vector databases, allowing you to effectively use the advantages of both approaches in architectural and engineering modeling

Bounding Box is "vectorization of geometry", and embedding (a way of transforming something abstract) is "vectorization of meaning". Both approaches allow you to move from manual search to intelligent search, be it 3D -objects in a project model or concepts in a text.

Search of objects in the project (for example, "find all windows with width > 1.5 m") is similar to the search of nearest neighbors (k-NN) in a vector database, where the criteria define a "zone" in the feature space. (we will talk more about k-NN nearest neighbors search in the next part about machine learning) (Fig. 8.2-2). If we add additional parameters (material, weight, production time) to the bounding box attributes, the table turns into a high-dimensional vector, where each attribute is a new dimension. This is closer to modern vector bases, where dimensions are counted in hundreds or thousands (e.g., embedding from neural networks).

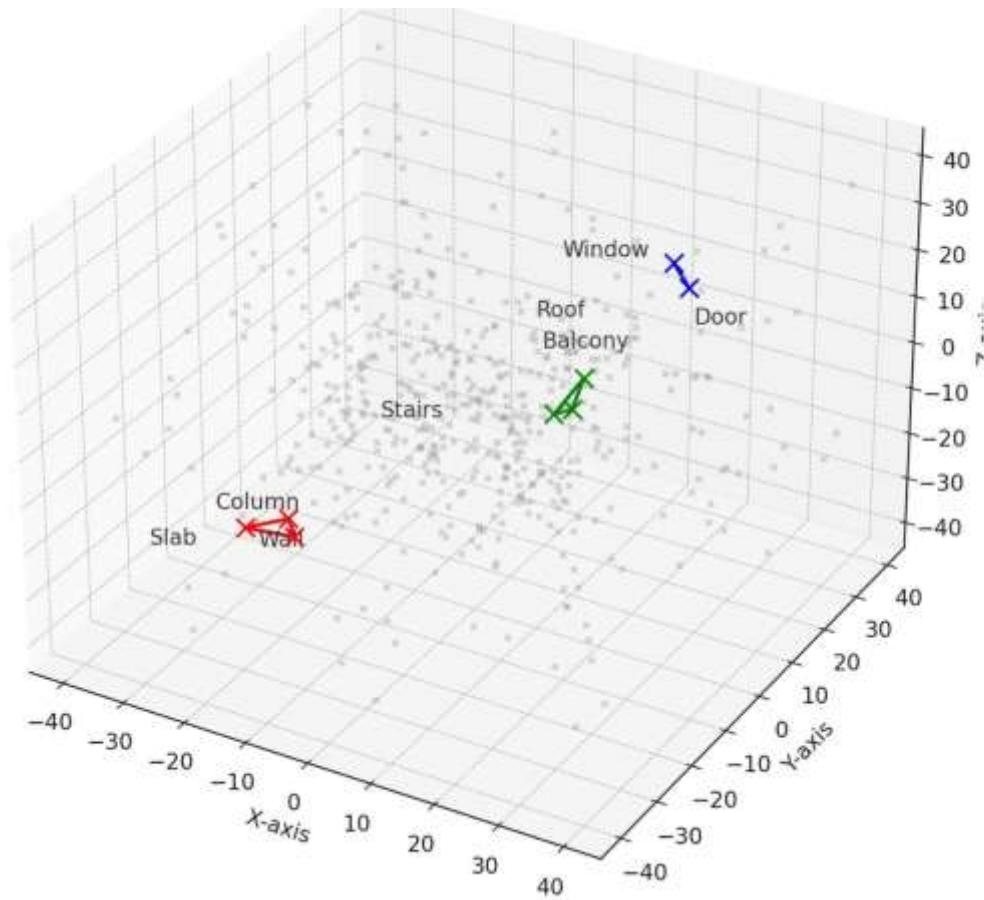


Fig. 8.2-2 Search for objects in the project using vector databases.

The approach used in Bounding Box, is applicable not only to geometric objects, but also to text and language analysis. Vector representations of data are already actively used in natural language processing (NLP). Just as objects in a construction project can be grouped by spatial proximity (Fig. 8.2-2), words in text can be analyzed by their semantic and contextual proximity.

For example, the words "architect", "construction", "design" will be next to each other in vector space because they have a similar meaning. In LLM this mechanism allows automatic, no manual categorization required:

- Identify the subject matter of a text
- Perform semantic searches on the content of documents
- Generate automatic annotations and summaries of text
- Find synonyms and related terms

Vector databases allow you to analyze text and find related terms in it in the same way that Bounding Box helps you analyze spatial objects in 3D -models. The example of Bounding Box of project elements helps to understand that vector representation is not a purely "artificial" concept from ML, but a natural

way of structuring data for solving applied problems, whether it is searching for columns in a CAD project or semantically close images in a database.

Specialists working with databases should pay attention to vector stores. Their proliferation indicates a new stage in database development, where classical relational systems and AI-oriented technologies begin to intertwine, forming hybrid solutions of the future.

Users developing complex and large-scale AI-applications will use specialized databases for vector search. At the same time, those who need only separate AI-functions for integration into existing applications are more likely to choose built-in vector search capabilities in the databases they already use (PostgreSQL, Redis).

Although systems such as DWH, Data Lake, CDE, PMIS, vector databases and others offer different approaches to storing and managing data, their effectiveness is determined not only by their architecture, but also by how well the data itself is organized and managed. Even when using modern solutions - be it vector databases, classical relational DBMSs or Data Lake-type warehouses - the lack of clear rules for managing, structuring and updating data can lead to the same difficulties faced by users working with disparate files and multi-format data.

Without Data Governance), even the most powerful solutions can become chaotic and unstructured data, turning data lakes into Data Swamps). To avoid this, companies must not only choose the right storage architecture, but also implement Data Minimalism), access management and quality control strategies to turn data into an effective decision-making tool.

Data Governance, Data Minimalism and Data Swamp

Understanding and implementing the concepts of Data Governance, Data Minimalism, and preventing Data Swamp are key to successfully managing data warehouses and delivering business value (Fig. 8.2-3).

According to a study by Gartner (2017), 85% of big data projects fail, and one of the key reasons is insufficient data quality and data governance [144].

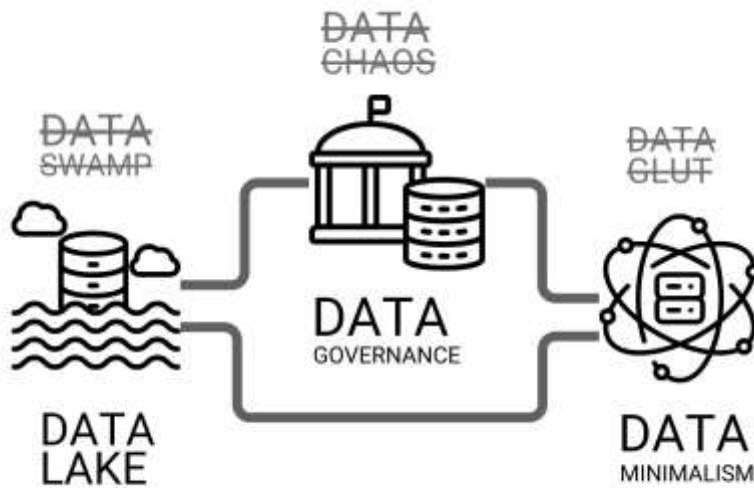


Fig. 8.2-3 Some of the key aspects of data governance are Data Governance and Data Minimalism.

Data Governance (Data Governance) is a fundamental component of data management, ensuring that data is used appropriately and effectively in all business processes. It is not only about establishing rules and procedures, but also about ensuring the availability, reliability and security of data:

- Defining and classifying data: clearly defining and classifying entities allows organizations to understand what entities are needed in the company and determine how they should be used.
- Access rights and management: developing policies and procedures for data access and management ensures that only authorized users can access certain data.
- Protecting data from external threats: Protecting data from external threats is a key aspect of data management. This includes not only technical measures, but also training of employees in the basics of information security.

Data Minimalism (Data Minimalism) is an approach to reduce data to the most valuable and meaningful attributes and entities in the formation (Fig. 8.2-4), thereby reducing costs and improving data utilization:

- Simplifying decision making: reducing the number of objects and their attributes to the most relevant simplifies the decision making process by reducing the time and resources required to analyze and process data.
- Focusing on what's important: selecting the most relevant entities and attributes allows you to focus on the information that really matters to the business, eliminating noise and unnecessary data.
- Efficient resource allocation: data minimization enables more efficient resource allocation, reducing data storage and processing costs, improving data quality and security.

The logic of working with data should not start with its creation as such (Fig. 8.2-4), but with under-

standing of future scenarios of using this data even before the generation process starts. This approach allows to determine in advance the minimum necessary requirements for attributes, their types and boundary values. These requirements form the basis for creating correct and stable entities in the information model. Preliminary understanding of the purposes and uses of the data contributes to the formation of a suitable structure for analysis. More details about approaches to data modeling at the conceptual, logical and physical levels were discussed in the chapter "Data modeling: conceptual, logical and physical model".

In the traditional business processes of construction companies, data processing more often resembles dumping data into a swamp, where data is first created and then specialists try to integrate it into other systems and tools.

Data Swamp (Data Swamp) is the result of uncontrolled collection and storage of data without proper organization, structuring and management, resulting in data that is unstructured, difficult to use and of little value.

How to prevent the flow of information from becoming a quagmire:

- **Data structure management:** ensuring that data is structured and categorized helps prevent data swamping by making it organized and easily accessible.
- **Understanding and interpreting** data: a clear description of data origins, modifications and meanings ensures that data are understood and interpreted correctly.
- **Maintaining data quality:** regular data maintenance and cleansing helps maintain data quality, relevance, and value for analytics and business processes.

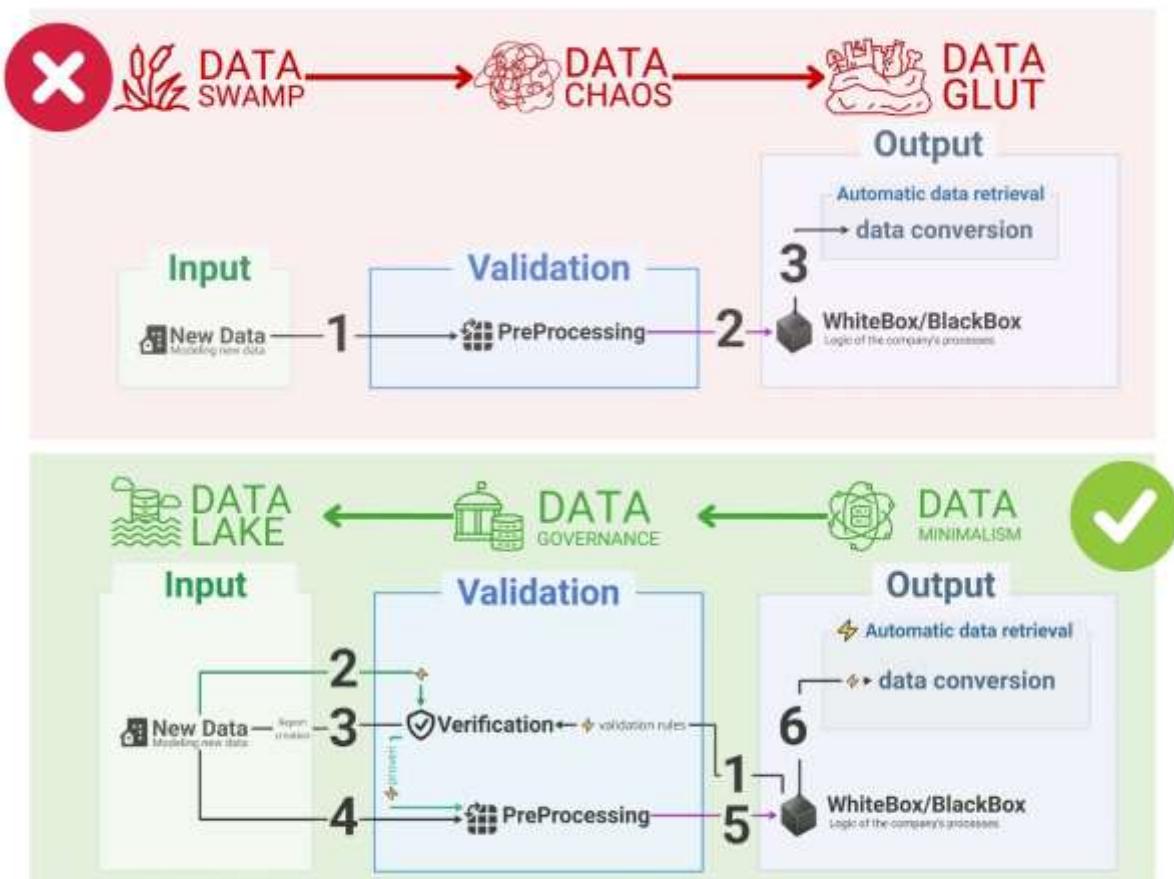


Fig. 8.2-4 To avoid clutter in the data warehouse, start the data creation process by gathering attribute requirements.

By integrating the principles of data governance and data minimalism into data management processes, and actively preventing data warehouses from becoming data swamps, organizations can maximize the potential of their data.

The next stage in the evolution of working with data, after solving the issues of management and minimalism, is the standardization of automatic processing, quality assurance and the implementation of methods that make data usable for analysis, transformation and decision-making. This is what the DataOps and VectorOps methodologies are doing, which are becoming important tools for companies working with big data and machine learning.

DataOps and VectorOps: new data standards

While Data Governance is responsible for controlling and organizing data, DataOps helps ensure its accuracy, consistency and smooth flow within the company. This is especially critical for a number of business cases in construction, where data is generated continuously and requires timely processing.

For example, in situations where building information models, project requirements and analytical reports need to be synchronized between different systems within a single business day, the role of DataOps can be key. It allows you to build stable and repeatable data processing processes, reducing the risk of delays and loss of relevance of information.

Data Governance alone is not enough - it is important that data is not just stored, but actively used in daily operations. This is where DataOps - a methodology focused on automation, integration and continuous data flow - comes to the fore.

DataOps focuses on improving collaboration, integration, and automation of data flows in organizations. Adopting DataOps practices promotes data accuracy, consistency, and availability, which is critical for data-centric applications.

Key tools in the DataOps ecosystem are Apache Airflow (Fig. 7.4-4) for workflow orchestration, and Apache NiFi (Fig. 7.4-5) for routing and transforming data flows. Together, these technologies enable flexible, reliable, and scalable data pipelines, enabling automatic processing, control, and integration of information between systems (more details in the chapter "Automatic ETL -conveyor"). When implementing the DataOps approach in construction processes, it is important to consider four fundamental aspects:

1. **People and tools are more important than data:** siloed data repositories may be seen as a major problem, but the reality is more complex. In addition to data fragmentation, the isolation of teams and the disparate tools they use play a significant role. In construction, specialists from different disciplines work with data: data engineers and analysts, BI and visualization teams, and project management and quality experts. Each of them has different ways of working, so it becomes important to create an ecosystem where data flows freely between participants, providing a single, consistent version of information.
2. **Automate testing and error detection:** Construction data always contains errors, be it inaccuracies in models, calculation errors or outdated specifications. Regularly testing data and eliminating recurring errors can significantly improve data quality. As part of DataOps, you need to implement automated controls and validation mechanisms that monitor data correctness, analyze errors and identify patterns, and capture and address system failures in every workflow. The higher the degree of automated validation, the higher the overall data quality and the lower the likelihood of errors in the final stages.
3. **Data should be tested in the same way as program code:** most building applications are based on data processing, but its control is often left to secondary roles. If machine learning models are trained on inaccurate data, it leads to incorrect predictions and financial losses. As part of DataOps, data should be subjected to the same scrutiny as software code: logic checks, stress tests, and evaluation of model behavior when input values change. Only validated and reliable data can be used as the basis for management decisions.
4. **Data observability without sacrificing performance:** data monitoring is not just a collection of metrics, but a strategic quality management tool. For DataOps to work effectively, observability must be built in at all stages of data handling, from design to operation. At the same time, it is

important that monitoring does not slow down the system. In construction projects, it is critical to not only collect data, but to do so in such a way that the work of the professionals (e.g. designers) creating the data is not disrupted in any way. This balance allows you to control data quality without sacrificing productivity.

DataOps is not an additional burden for data scientists, but the backbone of their work. By implementing DataOps, construction companies can move from chaotic data management to an efficient ecosystem where data works for the business.

In turn, VectorOps represents the next stage in the evolution of DataOps, focused on processing, storing, and analyzing multidimensional vector data (which were discussed in the previous chapter). This is particularly relevant in areas such as digital twins, neural network models and semantic search, which are starting to come to the construction industry. VectorOps relies on vector databases to efficiently store, index, and search multidimensional representations of objects.

VectorOps is the next step after DataOps, focused on processing, analyzing and using vector data in construction. Unlike DataOps, which focuses on data flow, consistency and quality, VectorOps focuses on managing the multidimensional object representations needed for machine learning.

Unlike traditional approaches, VectorOps allows you to achieve more accurate object descriptions, which is critical for digital twins, generative design systems, and automatic error detection in CAD data converted to vector format. The combined implementation of DataOps and VectorOps forms a solid foundation for scalable, automated work with large volumes of information - from classic tables to semantically rich spatial models

Next steps: from chaotic storage to structured storage

Traditional approaches to building data warehousing often result in the creation of disparate "silos of information" where important insights are inaccessible for analysis and decision making. Modern storage concepts, such as Data Warehouse, Data Lake and their hybrids, can unify disparate information and make it available in a centralized way for data streaming and business intelligence. It is not only important to choose the right storage architecture, but also to implement Data Governance) and Data Minimalism) to prevent storage from becoming uncontrollable Data Swamps).

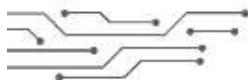
To summarize this part, it is worth highlighting the main practical steps that will help you apply the concepts discussed to your daily tasks:

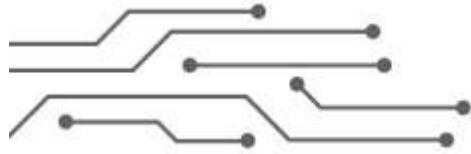
- Select efficient data storage formats
 - Move from CSV and XLSX to more efficient formats (Apache Parquet, ORC) for storing large amounts of data
 - Implement a data versioning system to track changes

- Use metadata to describe the structure and provenance of information
- Create a unified company data architecture
 - Compare different storage architectures: RDBMS, DWH, and Data Lake. Choose the one that best meets your needs for scalability, source integration, and analytical processing
 - Design a process map for extracting, loading, and transforming data (ETL) from various sources for your tasks. Use visualization tools such as Miro, Lucidchart or Draw.io to visually represent key steps and integration points
- Implement Data Governance practices and Data Minimalism
 - Follow the Data Minimalism approach - store and process only what is truly valuable
 - Implement Data Governance principles - define responsibility for data, ensure quality and transparency
 - Learn more about data management policies and DataOps concepts, VectorOps
 - Define data quality criteria and procedures for data validation within DataOps

Well-organized data storage creates the basis for centralizing the company's analytical processes. The transition from chaotic accumulation of files to structured storages allows turning information into a strategic asset that helps to make informed decisions and improve the efficiency of business processes.

Once the processes of data collection, transformation, analysis and structured storage are automated and standardized, the next stage of digital transformation is the full-fledged work with Big Data.





IX PART

BIG DATA, MACHINE LEARNING AND FORECASTS

Part nine is dedicated to big data, machine learning and predictive analytics in the construction industry. It explores the transition from intuitive decision making to objective analysis based on historical data. Practical examples are used to demonstrate big data analysis in construction, from parsing the San Francisco building permit dataset to processing CAD -projects with millions of elements. Special attention is given to machine learning methods for predicting the cost and schedule of construction projects, with a detailed discussion of linear regression and k-nearest neighbor algorithms. It is shown how structured data become the basis for predictive models to assess risks, optimize resources and improve project management efficiency. The part also provides recommendations on how to select representative data samples and explains why large data sets are not always required for effective analysis.

CHAPTER 9.1.

BIG DATA AND ITS ANALYSIS

Big data in construction: from intuition to predictability

The term "big data" does not have a strict definition. The concept originally appeared when the volume of information began to exceed the capabilities of traditional methods of its processing. Today, the volume and complexity of data in many industries, including construction, has increased so much that it does not fit into the local memory of computers and requires the use of new technologies to process it.

The essence of working with big data is not only storage and processing, but also predictive capabilities. In the construction industry, Big Data opens the way from intuitive decisions based on subjective interpretation of tables and visualizations (as discussed earlier) to reasonable forecasts supported by real observations and statistics.

Contrary to popular belief, the goal of working with big data is not to "make a machine think like a human", but to apply mathematical models and algorithms to analyze massive data sets in order to identify patterns, predict events and optimize processes.

Big Data is not a cold world of algorithms devoid of human influence. On the contrary, big data works in conjunction with our instincts, mistakes and creativity. It is the imperfection of human thinking that allows us to find non-standard solutions and make breakthroughs.

With the development of digital technologies, the construction industry has started to actively use data processing techniques that have come from the IT field. Thanks to tools such as Pandas and Apache Parquet, structured and unstructured data can be combined, simplifying access to information and reducing loss to analysis, while large datasets from documents or CAD projects (Fig. 9.2-10 - Fig. 9.2-12) allow data to be collected, analyzed and predicted at all stages of the project lifecycle.

Big Data is having a transformative impact on the construction industry, influencing it potentially in a variety of ways. The application of Big Data technologies is yielding results in a number of key areas, including, for example, the following:

- **Investment potential analysis** - forecasting of profitability and payback periods of projects based on data from previous facilities.
- **Predictive maintenance** - identifying likely equipment failures before they actually occur, which reduces downtime.
- **Supply chain optimization** - predicting disruptions and improving logistics efficiency.
- **Energy efficiency analysis** - assisting in the design of low energy buildings.

- **Safety monitoring** - the use of sensors and wearable devices to monitor site conditions.
- **Quality control** - real-time monitoring of compliance with process standards.
- **Labor management** - performance analysis and forecasting of staffing needs.

It is hard to find an area in construction where data analytics and predictions are not in demand. The main advantage of prediction algorithms is their ability to self-learn and continuously improve as data accumulates.

In the near future, artificial intelligence will not just assist builders, but will make key decisions - from design processes to building operation issues.

More about how predictions are generated and learning models are used will be discussed in the next part of the book, "Machine Learning and Predictions".

The transition to full-fledged work with big data requires a change in the approach to analytics itself. Whereas the classical systems we have considered so far focused on cause-and-effect relationships, in big data analytics the emphasis shifts to the search for statistical regularities and correlations that allow us to identify hidden relationships and predict the behavior of objects even without a full understanding of all factors.

Questioning the feasibility of big data: correlation, statistics and data sampling

Traditionally, construction was based on subjective hypotheses and personal experience. Engineers assumed - with a certain degree of probability - how the material would behave, what loads the structure would withstand and how long the project would last. These assumptions were tested in practice, often at the cost of time, resources and future risks.

With the advent of big data, the approach is changing dramatically: decisions are no longer made on the basis of intuitive guesses, but as a result of analyzing large-scale data sets. Construction is gradually ceasing to be an art of intuition and becoming a precise science of prediction.

The transition to the idea of using big data inevitably raises an important question: how critical is the amount of data and how much information is really necessary for reliable predictive analytics? The widespread belief that "the more data, the higher the accuracy" does not always prove to be statistically valid in practice.

Back in 1934, statistician Jerzy Neumann proved [145] that the key to the accuracy of statistical inference lies not so much in the amount of data as in its representativeness and randomness of sampling.

This is especially true in the construction industry, where large masses of data are collected using IoT

-sensors, scanners, surveillance cameras, drones and even multi-format CAD -models, increasing the risk of blind spots, outliers and data distortions.

Let's consider an example of road pavement condition monitoring. A complete data set of all road segments may take X GB and take about a day to process. At the same time, a random sample including only every 50th road section will take only $X/50$ GB and will be processed in half an hour, while providing similar accuracy of estimates for certain calculations (Fig. 9.1-1).

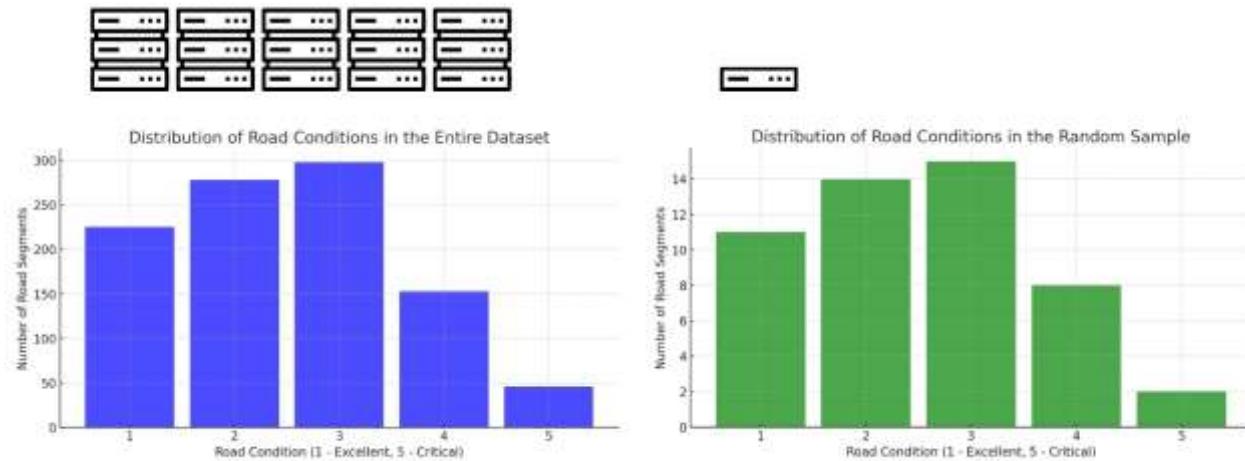


Fig. 9.1-1 Pavement condition histograms: full data set and random sampling show identical results.

Thus, the key to successful data analysis may often not be the amount of data, but the representativeness of the sample and the quality of the processing methods used. The move to random sampling and a more selective approach requires a shift in thinking in the construction industry. Historically, companies have followed the logic of "the more data the better," believing that covering all possible indicators would maximize accuracy.

This approach is reminiscent of a popular misconception from project management: "the more specialists I attract, the more effective the work will be". However, as with staffing, it is quality and tools that are more important than quantity. Without considering the interrelationships (correlations) between data or project participants, increasing volume can only lead to noise, distortion, duplication, and unnecessary waste.

In the end, it often turns out that it is much more productive to have a smaller, but qualitatively prepared data set capable of producing stable and reasonable forecasts than to rely on massive but chaotic information containing many contradictory signals.

Excessive data volume not only does not guarantee greater accuracy, but can also lead to distorted conclusions due to the presence of noise, redundant features, hidden correlations and irrelevant information. In such conditions, the risk of overfitting models increases and the reliability of analytical results decreases.

In construction, a major challenge in dealing with big data is determining the optimal quantity and quality of data. For example, when monitoring the condition of concrete structures, using thousands of sensors and collecting information every minute can overwhelm the storage and analysis system. However, if you perform a correlation analysis and select the 10% most informative sensors, you can get almost identical prediction accuracy, spending many times, sometimes tens or hundreds of times, fewer resources.

Using a smaller subset of data reduces both the amount of storage required and the processing time, which significantly reduces the cost of storing and analyzing data and often makes random sampling an ideal solution for predictive analytics, especially in large infrastructure projects or when working in real-time. Ultimately, the efficiency of construction processes is not determined by the amount of data collected, but by the quality of its analysis. Without a critical approach and careful analysis, data can lead to incorrect conclusions.

After a certain amount of data, each new unit of information yields less and less useful results. Instead of endlessly collecting information, it is important to focus on its representativeness and methods of analysis (Fig. 9.2-2).

This phenomenon is well described by Allen Wallis [146], who illustrates the use of statistical methods using the example of testing two alternative U.S. Navy projectile designs.

The Navy tested two alternative projectile designs (A and B) by conducting a series of paired rounds. In each round, A receives a 1 or 0 depending on whether its performance is better or worse than that of B, and vice versa. The standard statistical approach involves conducting a fixed number of trials (e.g., 1000) and determining the winner based on a percentage distribution (e.g., if A gets a 1 more than 53% of the time, it is considered the best). When Allen Wallis discussed such a problem with (Navy) Captain Garrett L. Schuyler, the captain objected that such a test, to quote Allen's story, might be useless. If a wise and experienced ordnance officer such as Schuyler had been on the spot, he would have seen after the first few hundred [shots] that the experiment need not be terminated either because the new method is clearly inferior or because it is clearly superior to what was hoped for [146].

— U.S. Government Statistical Research Group at Columbia University, World War II period

This principle is widely used in various industries. In medicine, for example, clinical trials of new drugs are conducted on random samples of patients, which makes it possible to obtain statistically significant results without the need to test the drug on the entire population of people living on the planet. In economics and sociology, representative surveys are conducted to reflect the opinion of society without the need to interview every person in the country.

Just as governments and research organizations conduct surveys of small populations to understand general social trends, companies in the construction industry can use random data samples to effectively monitor and create forecasts for project management (Fig. 9.1-1).

Big data may change the approach to social science, but it will not replace statistical common sense [147].

— Thomas Lansall-Welfair, "Predicting the nation's current mood," Significance v. 9 (4), 2012

From a resource-saving perspective, when collecting data for future predictions and decision-making, it is important to answer the question: does it make sense to spend significant resources to collect and process huge data sets when a much smaller and cheaper test data set that can be scaled up incrementally can be used? The effectiveness of random sampling shows that companies can reduce costs by tens or even thousands of times in collecting and training models by choosing data collection methods that do not require comprehensive coverage, but still provide sufficient accuracy and representativeness. This approach allows even small companies to achieve results on par with large corporations using significantly fewer resources and data volumes, which is important for companies looking to optimize costs and accelerate informed decision making using small resources. In the following chapters, explore examples of analytics and predictive analytics based on public datasets using big data tools.

Big data: analyzing data from San Francisco's million building permit dataset

Working with open datasets provides a unique opportunity to put into practice the principles discussed in previous chapters: judicious feature selection, representative sampling, visualization, and critical analysis. In this chapter, we will explore how complex phenomena such as construction activity in a large city can be investigated using open data—specifically, over one million building permit records in San Francisco.

Publicly available data on over one million building permits (Fig. 9.1-2) (records in two datasets in CSV format) from the "San Francisco Department of Buildings" [148] allow us to use the raw CSV -table to analyze not only construction activity in the city, but also to critically analyze recent trends and history of San Francisco's construction industry over the past 40 years, from 1980 to 2019.

The code examples used to create the dataset visualizations (Figures 9.1-3- Figures 9.1-8), as well as visual graphs with code, explanations, and comments, can be found on the Kaggle platform by searching for "San Francisco. Construction Sector 1980-2019." [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 880 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 (37.7903468760490,-122.4322541443574)
570817	200412131233	06/13/2005	9000.0	9000.0	apartments	94127.0 (37.729258516003388,-122.4644245057462)
198411	9202396	09/18/1992	9000.0	NaN	apartments	94111.0 (37.79506002552974,-122.39583224461805)

Fig. 9.1-2 The datasets contain information on issued building permits with various object attributes.

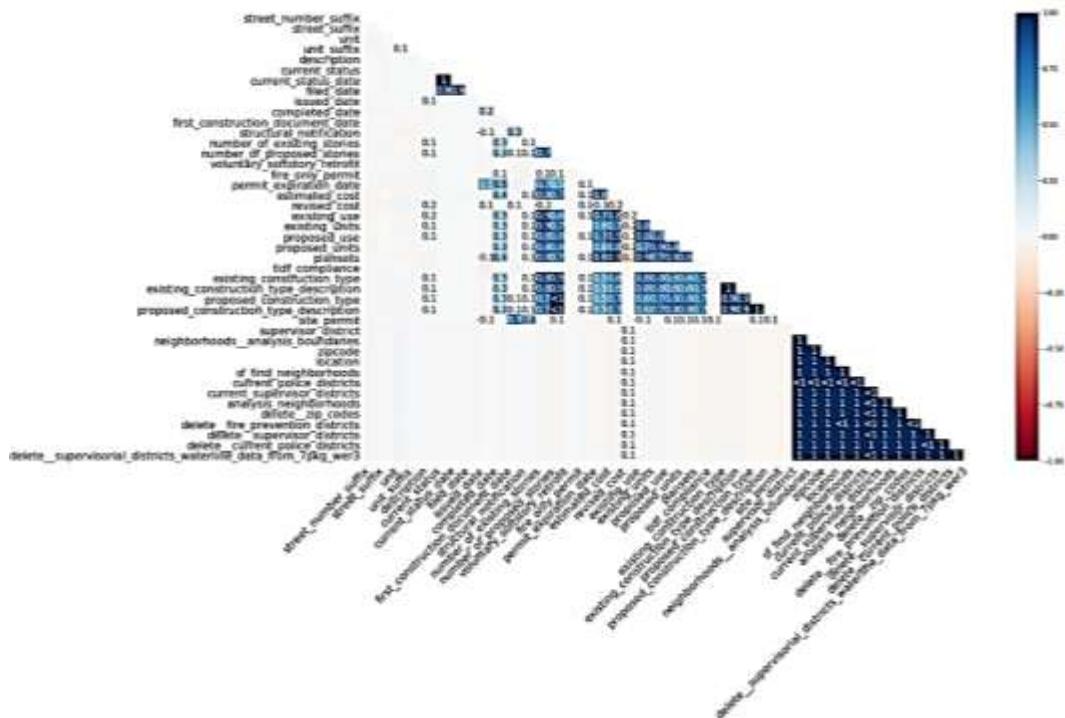


Fig. 9.1-3 A heat map (Pandas and Seaborn) that visualizes all attributes of a dataset and helps to identify relationships between attribute pairs.

No trends or conclusions are apparent from the table provided by the San Francisco Department of Buildings (Fig. 9.1-2). Dry numbers in tabular form are not a basis for decision making. To make the data visually understandable, as discussed in detail in the chapters on data visualization, it must be visualized using the various libraries discussed in Part Seven of the book on "ETL and visualizing results as graphs".

By analyzing data, using Pandas DataFrame and Python visualization libraries, on the value of 1,137,695 permits [148], we can conclude that construction activity in San Francisco is closely tied to economic cycles, especially in Silicon Valley's booming technology industry (Fig. 9.1-4).

Economic booms and busts have a significant impact on the number and value of construction projects. For example, the first peak in construction activity coincided with the electronics boom of the mid-1980s (used Pandas and Matplotlib), and subsequent peaks and declines were associated with the dot-com bubble and the technology boom of recent years.

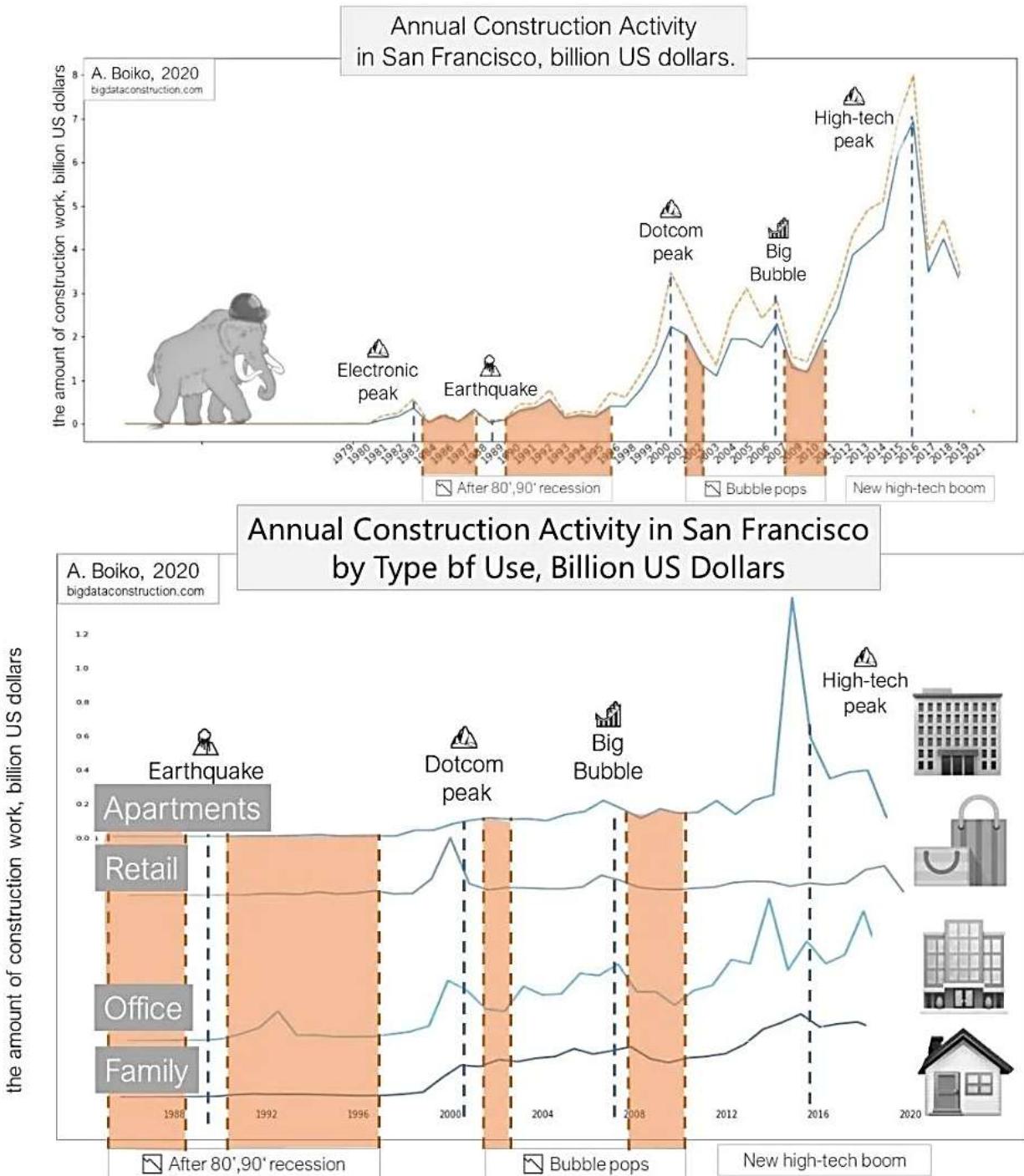


Fig. 9.1-4 In San Francisco real estate, investment correlates with Silicon Valley's technological development.

Data analytics estimates that in San Francisco, most of the \$91.5 billion invested in construction and redevelopment over the past decade - nearly 75% - is concentrated in downtown (Fig. 9.1-5 - used Pandas and Folium visualization library) and within a 2 km radius of downtown, reflecting the higher density of investment in these central zones.

The average cost of building permits varies greatly by neighborhood, with applications in the downtown area costing three times as much as those outside the city limits due to higher costs of land, labor, materials, and strict building codes requiring more expensive materials for energy efficiency.

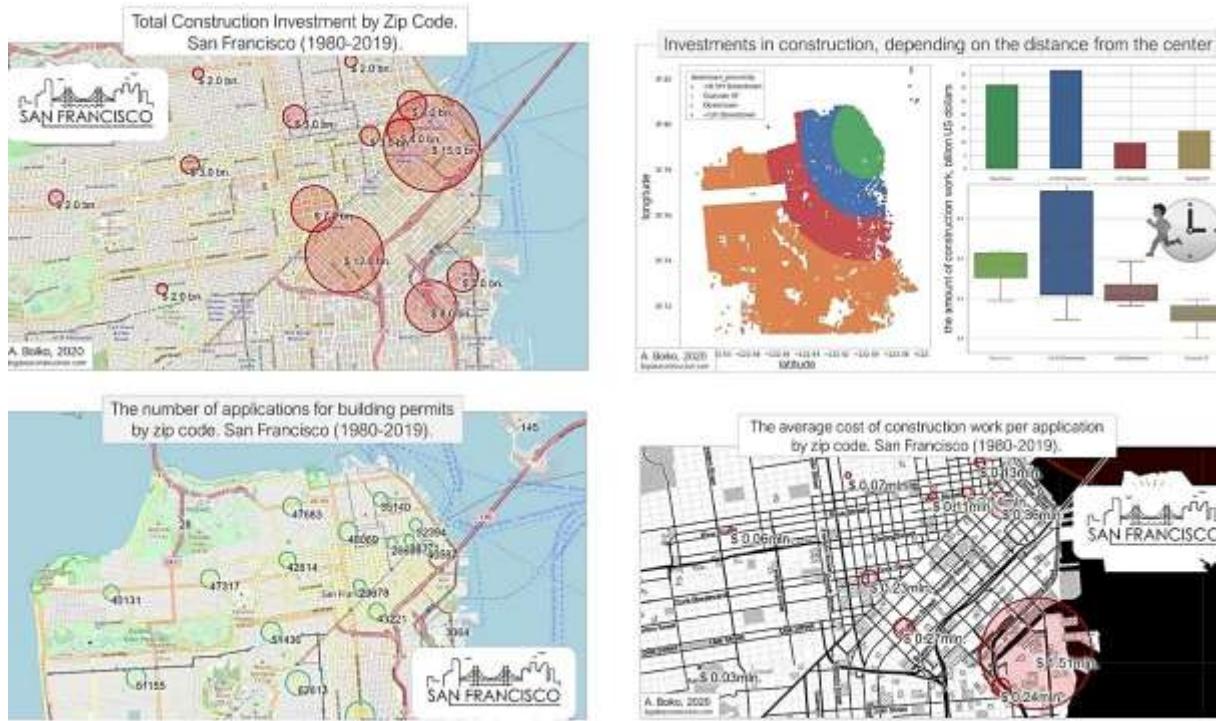


Fig. 9.1-5 In San Francisco, 75 percent of construction investment (\$91.5 billion) is concentrated in downtown.

The dataset also allows us to calculate average repair prices not only by house type, but also by city neighborhoods and individual addresses (zip codes). In San Francisco, the dynamics of home renovation costs show distinct trends for different types of renovations and housing (Fig. 9.1-6 - used Pandas and Matplotlib). Kitchen renovations are noticeably more expensive than bathroom renovations: the average kitchen renovation in a single-family home costs about \$28,000 compared to \$25,000 in a two-family home.

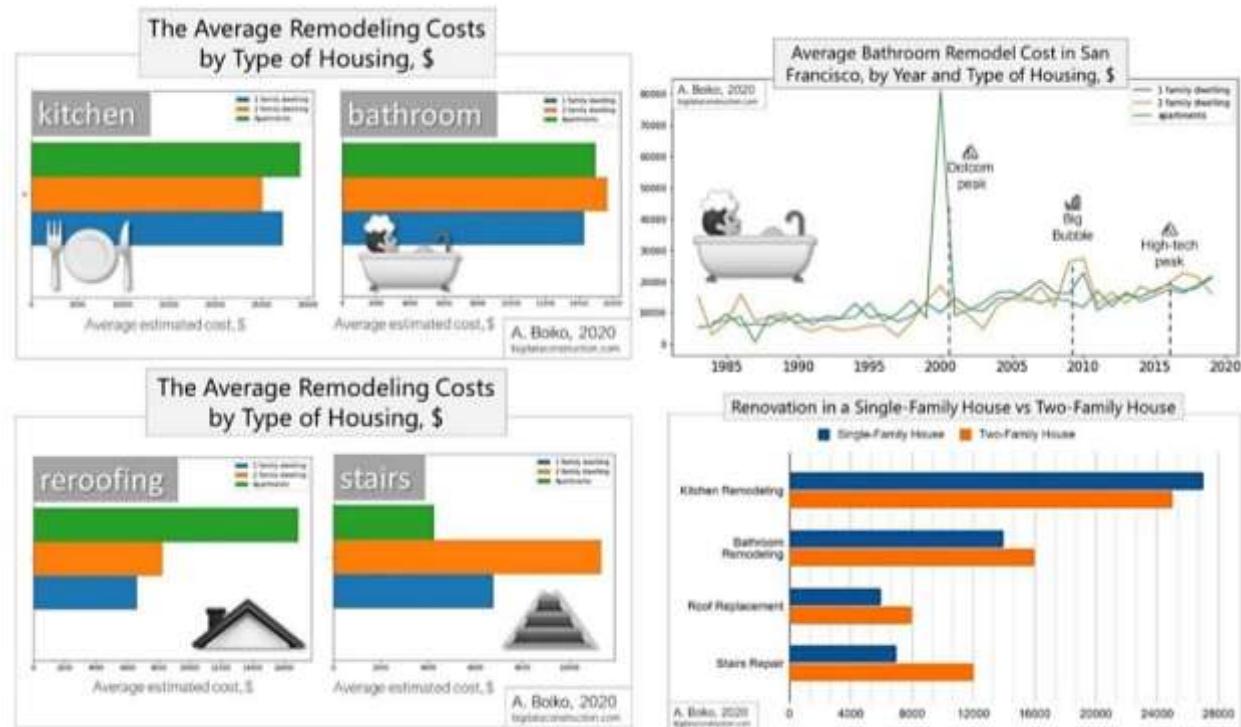


Fig. 9.1-6 In SF, kitchen renovations cost nearly twice as much as bathroom renovations and homeowners need to set aside \$ 350 each month for 15 years to cover the cost of major home repairs.

Construction cost inflation in San Francisco over the years can be seen by analyzing data grouped by housing type and year (Fig. 9.1-7 - used Pandas and Seaborn), which shows a steady increase in average repair costs since 1990 and reveals short-term three-year cycles in multifamily repair costs.

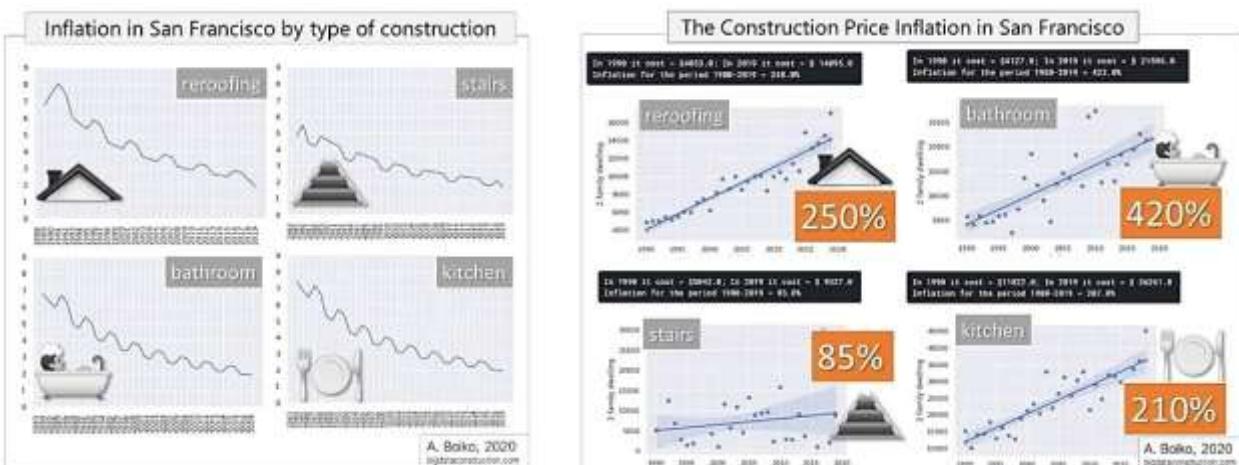


Fig. 9.1-7 From 1980 to 2019, the cost of bathroom renovations in SF has increased fivefold, while roof and kitchen renovations have tripled in price and stair renovations have only increased 85%.

A study of public data from the San Francisco Building Department (Exhibit 9.1-3) reveals that construction costs in the City are extremely variable and often unpredictable, being influenced by a variety of factors. These factors include economic growth, technological innovation, and the unique requirements of different housing types.

In the past, this kind of analysis required in-depth knowledge of programming and analytics. However, with the advent of LLM- tools, the process has become accessible and understandable to a wide range of professionals in the construction industry, from engineers in design departments to senior management.

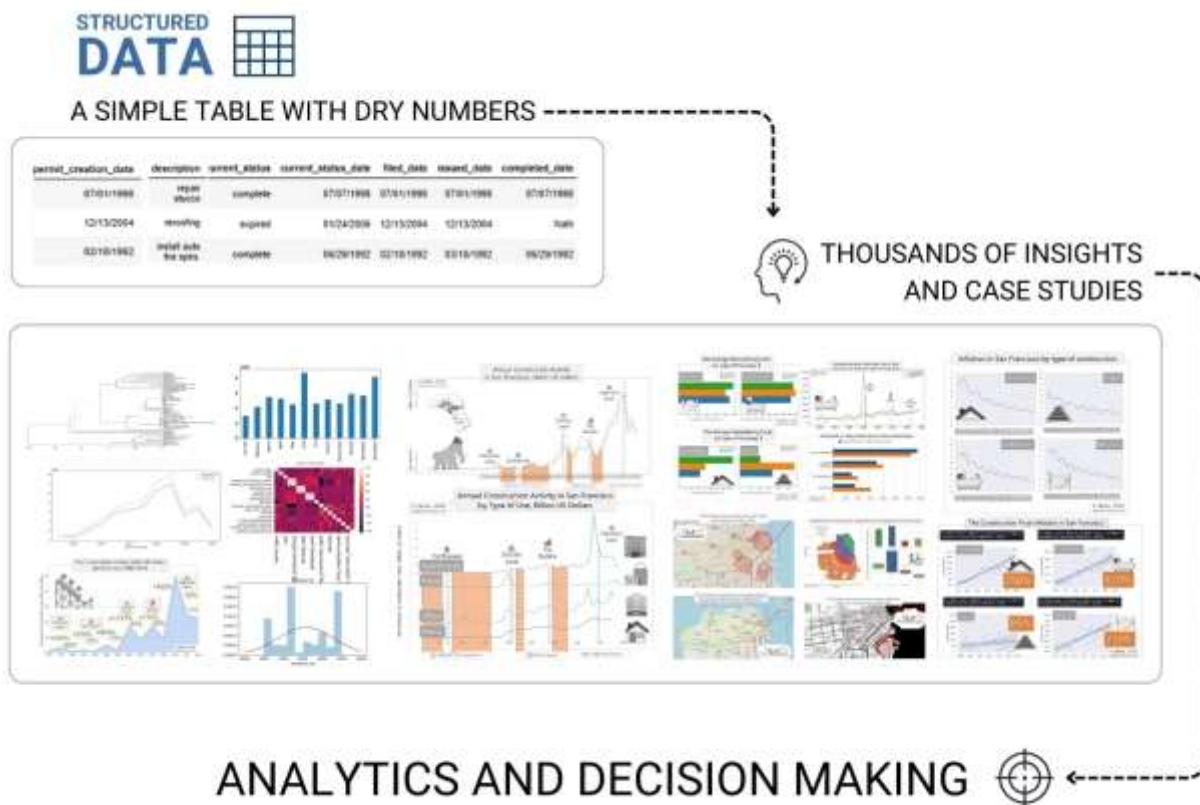


Fig. 9.1-8 The shift to visually comprehensible data enables automated decision making by recognizing hidden patterns.

Just as we analyzed data from the "San Francisco Building Authority" tabular dataset, we can visualize and analyze any dataset - from images and documents to IoT data, or data from derived CAD databases.

Example of big data based on CAD data (BIM)

In the following example we will analyze a large dataset using data from different CAD tools (BIM). To collect and create the large dataset, a specialized automated web crawler (script) was used, configured to automatically search and collect design files from websites offering free architectural models

in the formats RVT and IFC. In a few days, the crawler successfully found and downloaded 4,596 IFC files and 6,471 RVT files and 156,024 DWG files [149].

After collecting projects in RVT and IFC formats of different versions and converting them to a structured CSV format using the free reverse engineering SDKs, nearly 10 thousand RVT and IFC projects were collected into one large Apache Parquet table file and uploaded to Pandas DataFrame for analysis (Fig. 9.1-9).

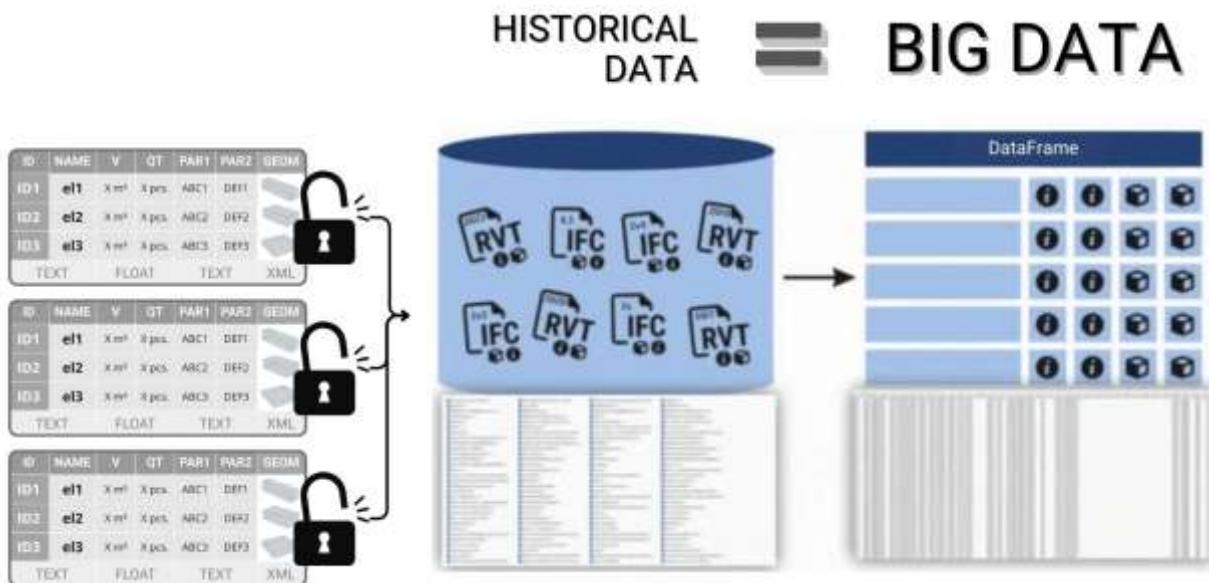


Fig. 9.1-9 Structured Data project data allows you to combine any number of projects into a single two-dimensional table.

The data from this large-scale collection contains the following information: the IFC file set contains about 4 million entities (rows) and 24,962 attributes (columns), and the RVT file set, consisting of about 6 million entities (rows), contains 27,025 different attributes (columns).

These information sets (Fig. 9.1-10) cover millions of elements, for each of which the coordinates of the Bounding Box geometry (the rectangle that defines the boundaries of the object in the project) were additionally obtained and added to a common table - the coordinates of the Bounding Box geometry (the rectangle that defines the boundaries of the object in the project) and images of each element in PNG format and the geometry in the open XML format - DAE (Collada) were created.

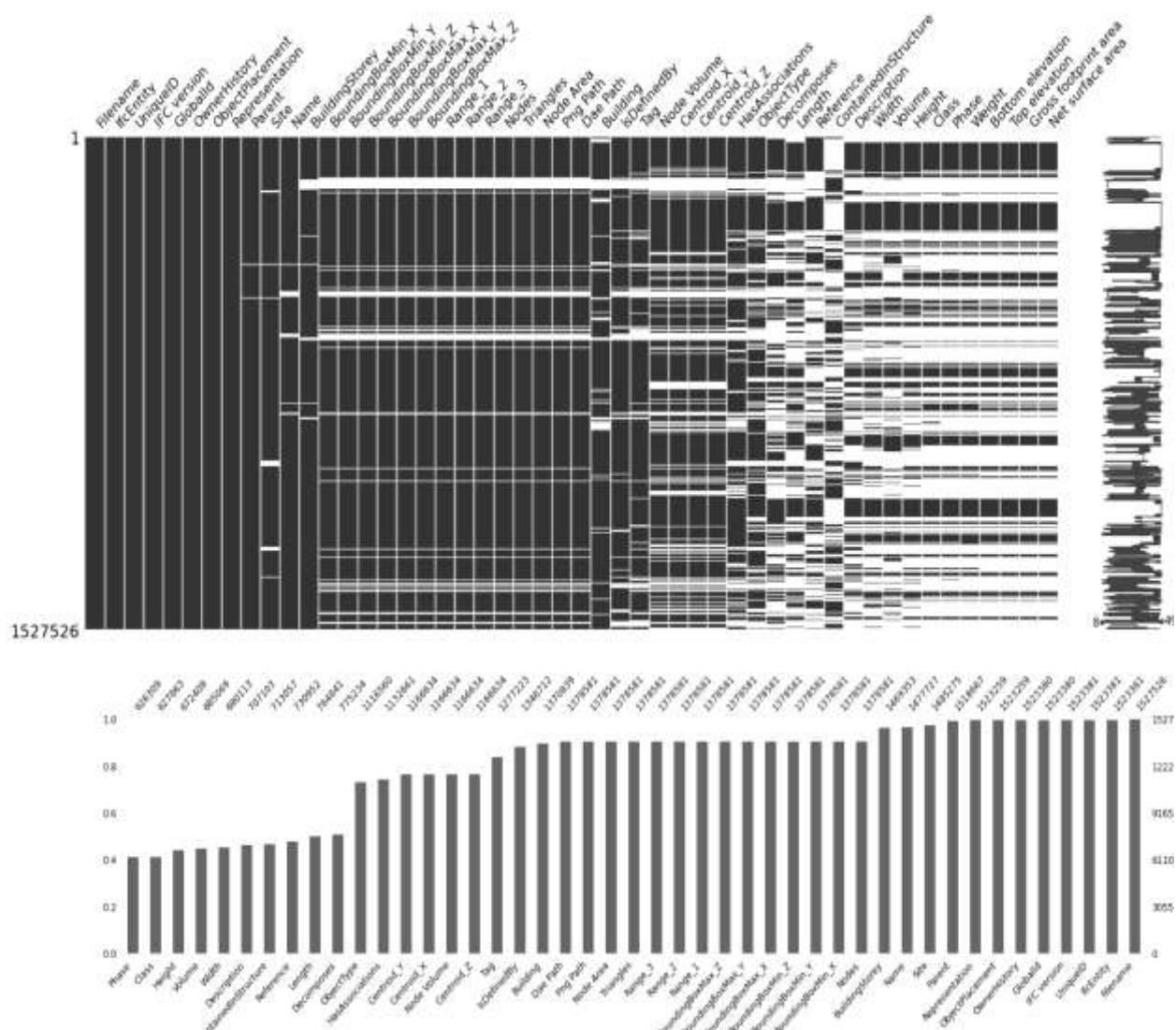


Fig. 9.1-10 Subset of 1.5 million elements and visualization (missingno library) of the occupancy of the first 100 attributes as a histogram.

Thus, we obtained all information about tens of millions of elements from 4,596 IFC projects and 6,471 RVT projects, where all attributes-properties of all entity elements and their geometry (Bounding Box) were translated into a structured form of a single table (DataFrame) (Fig. 9.1-10 - data about dataframe populations appear as histograms).

Histograms (Fig. 9.1-10, Fig. 9.2-6, Fig. 9.2-7) plotted during the analysis process allow a quick assessment of the data density and frequency of occurrence of values in columns. This gives a first insight into the distribution of features, the presence of outliers and the potential usefulness of individual attributes in analyzing and building machine learning models.

One example of a practical use of this dataset (Fig. 9.1-10) is the "5000 IFC and RVT " project. [149],

available on the Kaggle platform. It presents Jupyter Notebook with full Pipeline solution: from pre-processing and data analysis to visualization of results using Python libraries - pandas, matplotlib, seaborn, folium and others (Fig. 9.1-11).



Fig. 9.1-11 Examples of analyzing data from CAD formats (BIM) using Python visualization libraries and the pandas library.

Based on meta-information, it is possible to determine in which cities certain projects have been developed and display this on a map (e.g., using the folium library). In addition, time stamps in the data allow you to explore patterns in the time when files were saved or edited: by day of the week, time of day, and month.

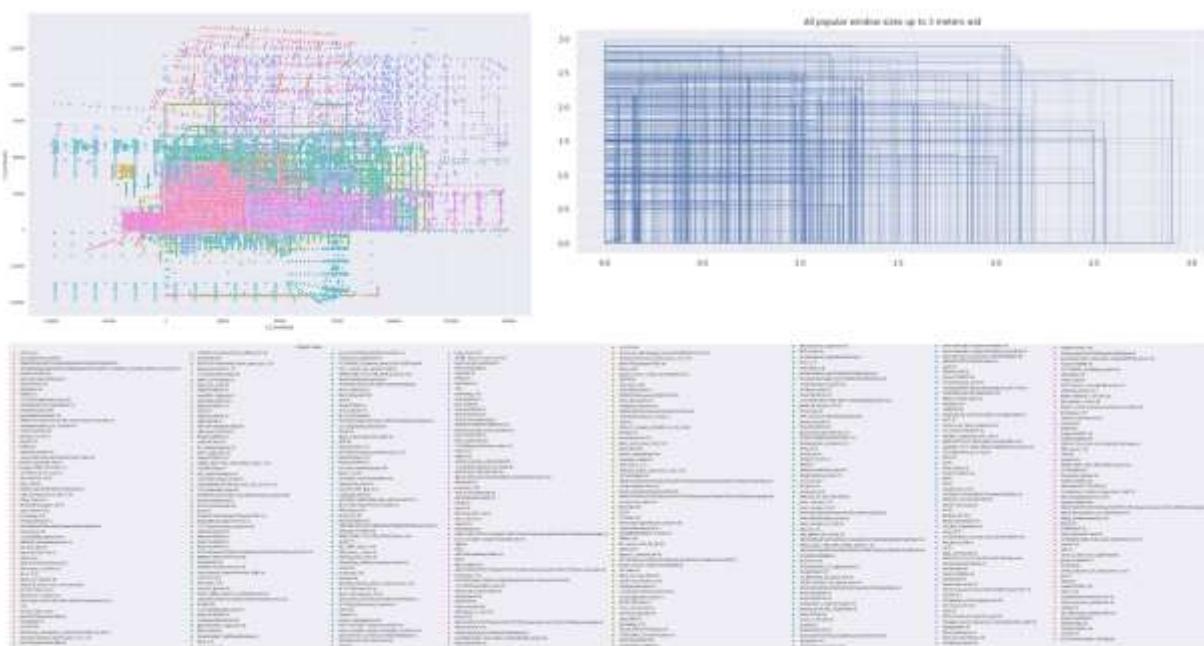


Fig. 9.1-12 Visualization of the geometric position of all columns and the dimensions of all windows up to 3 meters in projects from the list at the bottom of the chart.

Geometric parameters in the form of Bounding Box extracted from the models also lend themselves to aggregate analysis. For example, Fig. 9.1-12 shows two graphs: the left one shows the distribution of distances between columns for all projects relative to the zero point, and the right one shows the sizes of all windows up to 3 meters high in a sample of tens of thousands of window elements (after grouping the whole dataset by the parameter "Category" with the value "OST_Windows", "IfcWindows").

The analytical Pipeline code for this example and the dataset itself are available on the Kaggle website under the title "5000 IFC and RVT | DataDrivenCo-nstruction.io projects" [149]. This finished Pipeline along with the dataset can be copied and run free online for free on Kaggle or offline in one of the popular IDEs: PyCharm, Visual Studio Code (VS Code), Jupyter Notebook, Spyder, Atom, Sublime Text, Eclipse with PyDev plugin, Thonny, Wing IDE, IntelliJ IDEA with Python plugin, JupyterLab or popular online tools Kaggle.com, Google Collab, Microsoft Azure Notebooks, Amazon SageMaker.

Analytical data derived from the processing and examination of huge amounts of structured data will play a crucial role in decision-making processes in the construction industry.

With this kind of information analysis based on past project data, experts can effectively forecast, for example, material and labor requirements and optimize design solutions before construction begins

However, while design data or building permits are relatively static information that changes relatively slowly, the construction process itself is rapidly becoming saturated with a variety of sensors and IoT-devices: cameras, automated monitoring systems that transmit data in real time - all of this turns the construction site into a dynamic digital environment where data needs to be analyzed in real time.

IoT Internet of Things and Smart Contracts

IoT The Internet of Things represents a new wave of digital transformation in which every device gets its own IP address and becomes part of a global network. IoT is a concept that involves connecting physical objects to the Internet to collect, process and transmit data. In construction, this means the ability to control construction processes in real time, minimize material loss, predict equipment wear and tear and automate decision-making.

According to the CFMA article "Preparing for the Future with Connected Construction" [150], the construction industry will undergo a major digital transformation in the next decade, culminating in the concept of Connected Construction - a fully integrated and automated construction site.

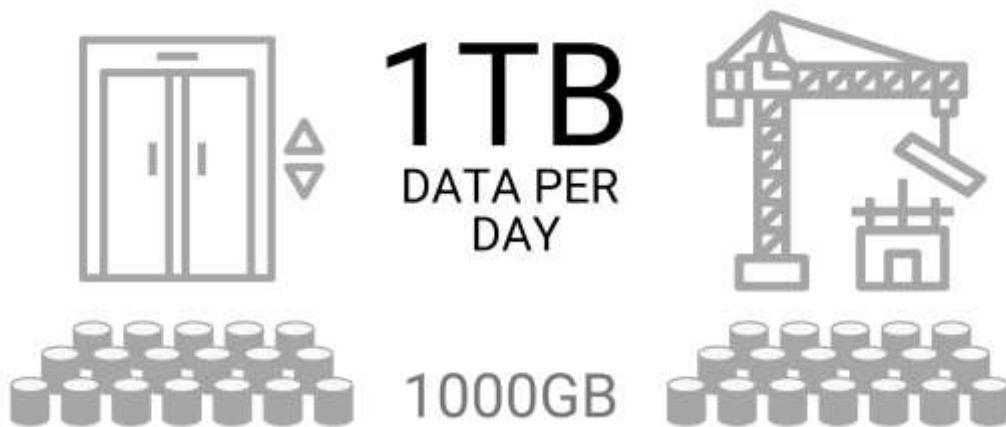


Fig. 9.1-13 IOT devices or construction site data devices can produce and transmit terabytes of data per day.

A digital construction site implies that all elements of construction - from planning and logistics to work execution and quality control at the construction site using fixed cameras and quadcopters - will be integrated into a single dynamic digital ecosystem. Earlier, in Part 7 of this book, we have already looked at the capabilities of Apache NiFi (Fig. 7.4-5), a free and open source tool that allows you to organize real-time data streaming - from collection from various sources to transfer to storage or analytics platforms.

Data on construction progress, material consumption, equipment status and safety will be transmitted in real time to analytical systems (Fig. 9.1-13). This allows predicting potential risks, promptly reacting to deviations and optimizing site processes. Key components of a digital construction site include:

- IoT -sensors - tracking environmental parameters, monitoring construction equipment and controlling labor conditions.
- Digital twins - virtual models of buildings and infrastructure to predict possible deviations and prevent errors.
- Automated logistics systems - real-time supply chain management to reduce downtime and costs.
- Robotic construction complexes - using autonomous machines to perform routine and hazardous tasks.

Robotization use of IoT and the Connected Site (Construction) digital construction site concept will not just increase efficiency and reduce costs, but also usher in a new era of safety, sustainable construction and predictive project management.

RFID (Radio Frequency Identification) tags are also an important component of IoT. They are used to identify and track materials, machinery and even personnel on a construction site, increasing transparency and control of project resources.

RFID -technology is used to automatically recognize objects using radio signals. It consists of three

key elements:

- RFID -tags (passive or active) - contain a unique identifier and are attached to materials, tools or machinery.
- Scanners are devices that read information from tags and transmit it to the system.
- Centralized database - stores information about the location, status and movement of objects.

Application of RFID in construction:

- Automatic material accounting - tags on ready-mixed concrete products, rebar or sandwich panel packages allow inventory control and prevent theft.
- Personnel work control - RFID - employee badges record shift start and end times, providing a record of working hours.
- Equipment Monitoring - RFID - system tracks the movement of equipment, preventing down-time and improving logistics efficiency.

Complementing this technology suite are blockchain-based smart contracts that automate payments, delivery control and contract compliance without the need for intermediaries, reducing the risk of fraud and delays.

Today, in the absence of a common data model, smart contracts are simply code that participants agree upon. However, with a Data-Centric approach, it is possible to create a common model of contract parameters, encode it in a blockchain, and automate the fulfillment of the terms.

For example, in a supply chain management system, a smart contract can track the delivery of a shipment from IoT -sensors and RFID -tags and automatically transfer payment when it arrives. Similarly, on a construction site, a smart contract can record the completion of a work phase - such as installing rebar or pouring a foundation - based on data from drones or construction sensors and automatically initiate the next payment to the contractor without the need for manual checks and paper certificates.

But despite new technologies and the efforts of international standardization organizations, a multitude of competing standards complicate the IoT landscape.

According to a Cisco study published in 2017 [151], almost 60% of Internet of Things initiatives (IoT) stop at the proof-of-concept stage, and only 26% of companies consider their IoT projects to be fully successful. Moreover, a third of completed projects do not achieve their stated goals and are not recognized as successful even after implementation.

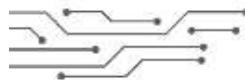
One of the key reasons is the lack of interoperability between platforms that process data from different sensors. As a result, data remains isolated within separate solutions. An alternative to this approach, as in other similar cases (which we have covered in this book), is an architecture built around the data itself as the primary asset.

IoT sensors play a key role not only in monitoring the technical condition of equipment, but also in predictive analytics to reduce risks on the construction site and improve overall process performance by predicting failures and deviations.

The data collected by IoT sensors and RFID tags can be processed in real time by machine learning algorithms that can detect anomalies and alert engineers to potential malfunctions in advance. This can range from micro-cracks in concrete structures to uncharacteristic pauses in tower crane operation, indicating technical failures or regulatory violations. Furthermore, advanced behavioral analysis algorithms can capture behavioral patterns that may indicate, for example, physical fatigue of personnel, enhancing proactive management of safety and employee well-being on site.

In the construction industry, accidents and failures - whether of machinery or people - rarely happen suddenly. They are usually preceded by minor deviations that go unnoticed. Predictive analytics and machine learning make it possible to detect these signals at an early stage, even before critical consequences occur.

While documents, project files, and data from IoT devices and RFID tags form the digital footprint of construction projects, machine learning can help extract useful insights from it. With the growth of data and democratization of data access, the construction industry is gaining new opportunities in analytics, predictive analytics and artificial intelligence applications.



CHAPTER 9.2.

MACHINE LEARNING AND PREDICTIONS

Machine learning and artificial intelligence will change the way we build

The databases of the various systems in the construction business - with their inevitably decaying and increasingly complex infrastructure - are becoming a breeding ground for future solutions. Company servers, like a forest, are rich with a biomass of important information, often hidden underground in the bowels of folders and servers. The masses of data from the various systems being created today - after use, after falling to the bottom of the server and after years of fossilization - will fuel machine learning and language models in the future. Internal company chat rooms (e.g., a separate instance of locally configured ChatGPT, LLaMa, Mistral, DeepSeek) will be built on these in-house models using centralized storage to quickly and conveniently retrieve information and generate the necessary graphs, dashboards, and documents.

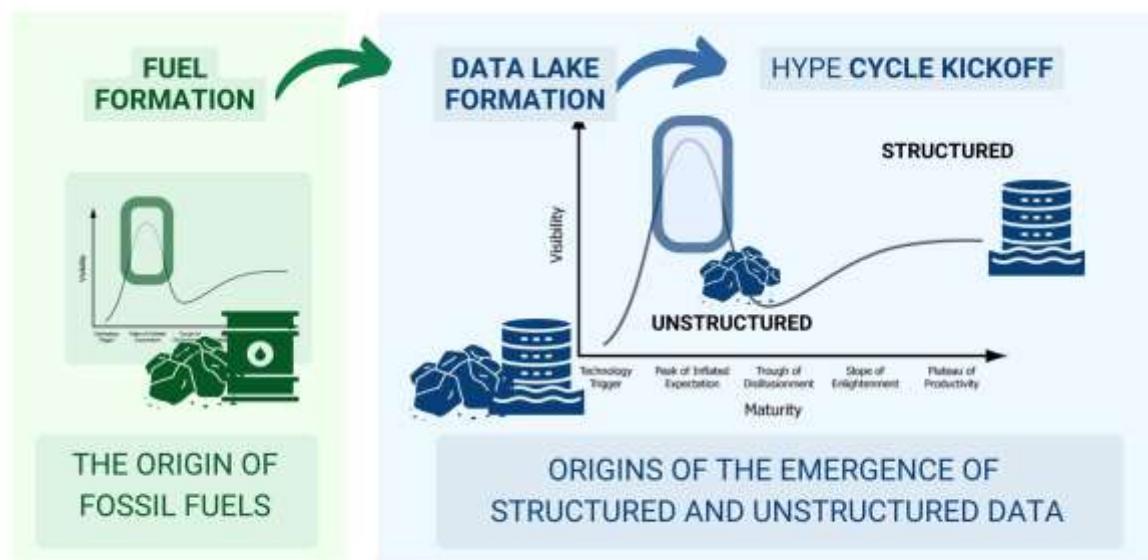


Fig. 9.2-1 Just as trees turn into coal, so too does information turn into valuable business energy over time under the pressure of time and analytics.

Fossilization of plant mass in combination with pressure and temperature creates a homogeneous and uniquely structured homogeneous mass of trees of different species that lived at different times - charcoal [152]. In the same way, information recorded on hard disks in different formats and at different times under the pressure of analytics departments and temperature of quality management eventually forms a homogeneous structured mass of valuable information (Fig. 9.2-1).

These layers (or more often isolated nuggets) of information are created through painstaking data organization by experienced analysts who begin to gradually extract valuable information from seemingly long irrelevant data.

The moment these mature layers of data are no longer just "burned" in reports, but begin to circulate in business processes, enriching decisions and improving processes, the company becomes ready for the next step - the transition to machine learning and artificial intelligence (Fig. 9.2-2).

Machine learning (ML - Machine learning) is a class of methods for solving artificial intelligence problems. Machine learning algorithms recognize patterns in large data sets and use them to learn themselves. Each new data set allows the mathematical algorithms to improve and adapt according to the information obtained, which allows to constantly improve the accuracy of recommendations and predictions.

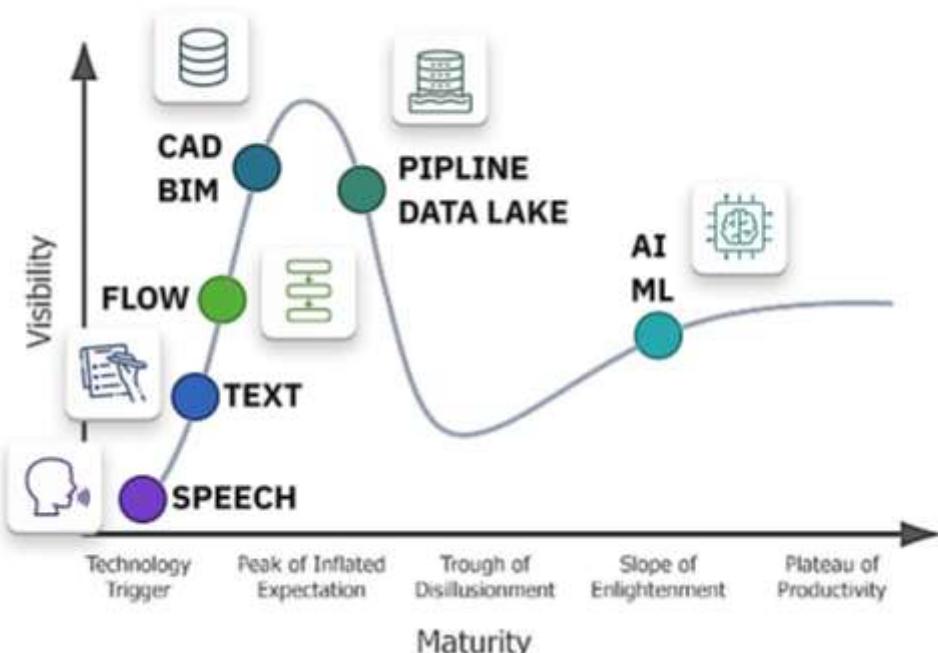


Fig. 9.2-2 The fading of data creation technologies and the application of analytics tools opens the door to the topic of machine learning.

As the influential CEO of the world's largest investment fund (which owns key stakes in almost all of the largest construction software companies, as well as the companies that own the most real estate in the world [55]) said in an interview with 2023) - machine learning will change the world of construction.

AI has enormous potential. It will change the way we work, the way we live. AI and robotics will change the way we work and the way we build, and we will be able to use AI and robotics as a means to create much greater productivity [153].

– CEO of the world's largest investment fund, interview, September 2023.

Machine Learning (ML) works by processing large amounts of data, using statistical techniques to mimic aspects of human thinking. However, most companies do not have such datasets, and if they do, they are often not sufficiently labeled. This is where semantic technologies and transfer learning, a technique that allows ML to be more effective when dealing with small amounts of data, the feasibility of which has been discussed in previous chapters of this part, can help.

The essence of transfer learning is that instead of processing each task from scratch, you can use knowledge gained in related fields. It is necessary to realize that patterns and discoveries from other industries can be adapted and applied in the construction industry. For example, methods of optimizing logistics processes developed in retail help to improve the efficiency of construction supply chain management. Big data analysis, which is actively used in finance, can be applied to cost forecasting and risk management in construction projects. And computer vision and robotics technologies being developed in industry are already finding application in automated quality control, safety monitoring and construction site facilities management.

Transfer learning allows not only to accelerate the introduction of innovations, but also to reduce the cost of their development, using the already accumulated experience of other industries.

**labor productivity in
construction = f(AI)**

Fig. 9.2-3 Artificial Intelligence technologies and robotics will be the main driving force of the future to increase productivity in the construction industry.

Human thinking is organized on a similar principle: we rely on previously acquired knowledge to solve new problems (Fig. 4.4-19, Fig. 4.4-20, Fig. 4.4-21). In machine learning, this approach works too - by simplifying the data model and making it more elegant, we can reduce the complexity of the problem for ML algorithms. This in turn reduces the need for large amounts of data and reduces computational cost.

From subjective assessment to statistical forecast

The era when strategic decisions depended on the intuition of individual managers (Fig. 9.2-4) is a thing of the past. In an increasingly competitive and challenging economic environment, a subjective approach is becoming too risky and inefficient. Companies that continue to rely on personal opinions instead of objectively analyzing data, lose the ability to respond quickly to change.

The competitive environment demands accuracy and repeatability based on data, statistical patterns and computable probability. Decisions can no longer be based on a feeling, they must rely on correlations, trends and predictive models derived from analytics and machine learning. This is not just a change in tools - it is a change in the logic of thinking: from assumptions to evidence, from subjective probabilities to statistically calculated deviations, from feeling to facts.



Fig. 9.2-4 The era of decisions made by HiPPO (the opinion of the highest paid employee) will become a thing of the past with the advent of big data and machine learning.

Managers who used to rely solely on their own feelings will inevitably face a new reality: authority no longer determines choices. At the center of management are now systems that analyze millions of parameters and vectors, identifying hidden patterns and suggesting optimal strategies.

The main reason why companies today still avoid implementing ML is its lack of transparency. Most models work as "black boxes" for managers, without explaining how exactly they come to their conclusions. This leads to problems: algorithms can reinforce stereotypes and even create ridiculous situations, as in the case of Microsoft's chatbot, which quickly turned into a toxic communication tool [154].

In "Deep Thinking", Garry Kasparov, former world chess champion, reflects on his defeat by the IBM Big Blue computer [155]. He argues that the true value of AI is not in copying human intelligence, but in complementing our abilities. AI should perform tasks in which humans are weak, while humans bring creativity. Computers have changed the traditional approach to analyzing chess. Instead of creating fascinating stories about games, computer chess programs evaluate each move impartially, based only on its actual strength or weakness. Kasparov notes that the human tendency to see events as coherent stories rather than individual actions often leads to wrong conclusions - not only in chess, but in life in general.

Therefore, if you plan to use machine learning for prediction and analysis, it is important to understand its basic principles - how algorithms work and how data is processed - before you start using machine learning tools and AI in your work. The best way to get started is through hands-on experience.

One of the most convenient tools for an initial introduction to the topic of machine learning and prediction is the Jupyter Notebook and the popular classic Titanic dataset, which will provide a visual introduction to key methods of data analysis and ML model building.

Titanic dataset: Hello World in the world of analytics data and big data

One of the most famous examples of using ML in data analytics is the analysis of the Titanic dataset, which is often used to study the probability of survival of passengers. Studying this table is analogous to the "Hello World" program when learning programming languages.

The sinking of the RMS Titanic in 1912 resulted in the deaths of 1502 out of 2224 people. The Titanic dataset contains not only information about whether a passenger survived, but also attributes such as: age, gender, ticket class and other parameters. This dataset is available for free and can be opened and analyzed on various offline and online platforms.

Link to Titanic dataset:

<https://raw.githubusercontent.com/datasets/master/titanic.csv>

Earlier in the chapter "LLM-enabled IDEs and future programming changes" we already discussed Jupyter Notebook - one of the most popular development environments for data analysis and machine learning. Free cloud-based counterparts to Jupyter Notebook are the Kaggle and Google Collab platforms, which allow you to run Python code without installing software and provide free access to computing resources.

Kaggle is the largest data analytics, machine learning competition platform with a built-in code execution environment. As of October 2023, Kaggle has over 15 million users [156] from 194 countries.

Download and use the Titanic dataset on the Kaggle platform (Fig. 9.2-5) to store the dataset (a copy

of it) and run Python code with pre-installed libraries directly in a browser, without having to install a dedicated IDE.



Fig. 9.2-5 Statistics Titanic table, the most popular training dataset for learning data analytics and machine learning.

The Titanic dataset includes data on the 2224 passengers on board the *RMS Titanic* at the time of its wreck in 1912. The dataset is presented as two separate tables, a training (*train.csv*) and a test (*test.csv*) sample, allowing it to be used both for training models and for evaluating their accuracy on new data.

The training dataset contains both attributes-attributes of passengers (age, gender, ticket class and others) and information about who survived (column with binary values "Survived"). The training dataset (Fig. 9.2-6 - file *train.csv*) is used to train the model. The test dataset (Fig. 9.2-7 - file *test.csv*) includes only passenger attributes without survivor information (without a single "Survivor" column). The test dataset is designed to test the model on new data and to evaluate its accuracy.

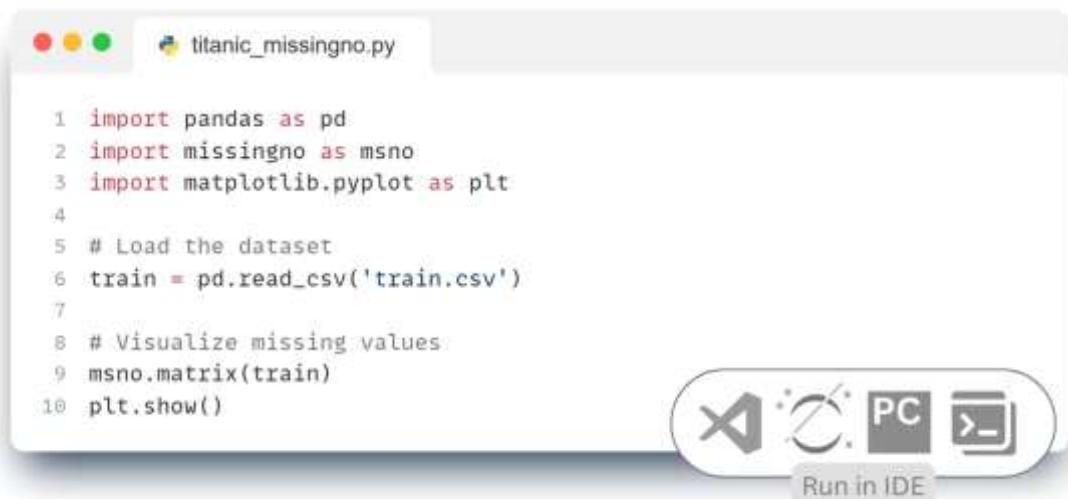
Thus, we have almost identical attributes of passengers in the training and test datasets. The only key difference is that in the test dataset we have a list of passengers who do not have the "Survivor" column - the target variable, which we want to learn to predict using various mathematical algorithms. And after building the model, we will be able to compare the output of our model with the real parameter "Survivor" from the test dataset, which we will take into account to evaluate the results.

The main columns of the table, passenger parameters in the training and test dataset:

- **PassengerId** - unique passenger identifier
- **Survived** - 1 if the passenger survived, 0 if dead (not available in the test set)
- **Pclass** - ticket class (1, 2 or 3)

- **Name** - passenger's name
- **Sex** - sex of the passenger (male/female)
- **Age**
- **SibSp** - number of brothers/sisters or spouses on board
- **Parch** - number of parents or children on board
- **Ticket** - ticket number
- **Fare** is the cost of a ticket
- **Cabin** - cabin number (many data are missing)
- **Embarked** - port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

To visualize missing data in both tables, you can use the missingno library (Fig. 9.2-6, Fig. 9.2-7), which displays missing values in the form of a histogram, where white fields show missing data. This visualization allows a quick assessment of data quality before processing.



```
1 import pandas as pd
2 import missingno as msno
3 import matplotlib.pyplot as plt
4
5 # Load the dataset
6 train = pd.read_csv('train.csv')
7
8 # Visualize missing values
9 msno.matrix(train)
10 plt.show()
```

The screenshot shows a Jupyter Notebook cell with the title "titanic_missingno.py". The cell contains Python code for visualizing missing values in the "train.csv" dataset using the "missingno" library. The code imports pandas, missingno, and matplotlib.pyplot, loads the "train.csv" dataset, and then uses msno.matrix() to create a matrix plot of missing values. A "Run in IDE" button is visible at the bottom of the cell.

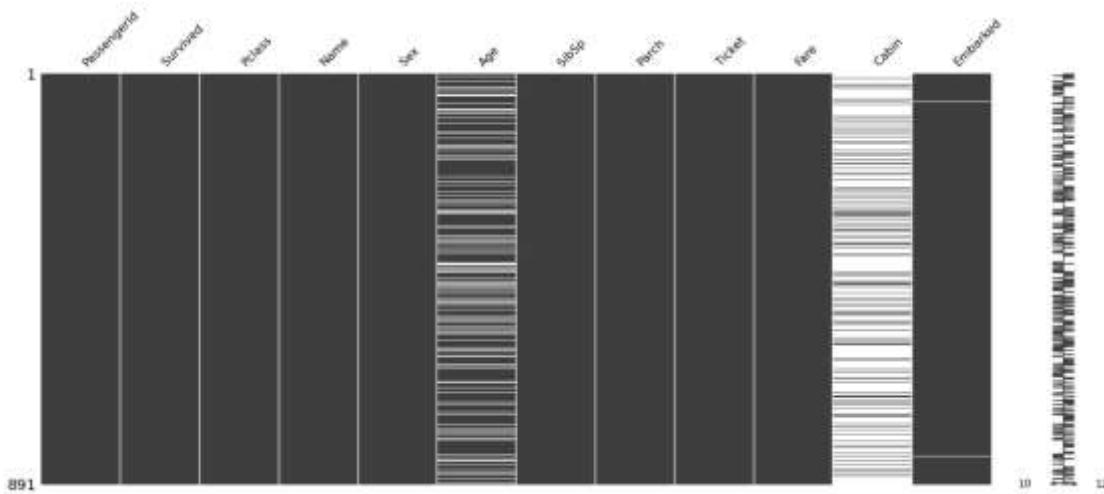


Fig. 9.2-6 A few lines of code are used to visualize missing data in the Titanic training dataset, where the key parameter for training is the "Survived" parameter.



Fig. 9.2-7 Visualization of missing data in the Titanic test dataset, which contains only passenger characteristics with no information.

Before formulating hypotheses and making predictions based on the dataset, visual analysis helps to identify key patterns in the data, assess its quality, and identify possible dependencies. There are many visualization techniques that help you better understand the Titanic dataset. You can use distribution plots to analyze passenger age groups, survival charts by gender and class, and missing data matrices to assess the quality of information and understand the data.

- 💡 Let's ask LLM to help us visualize the data from the Titanic dataset by sending the following text request to any LLM model (CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN or any other):

```
Please show some simple graphs for the Titanic dataset. Download the  
dataset yourself and show the ↵
```

- 2 LLM response in the form of ready-made code and graphs visualizing the dataset parameters:

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3
4 # Load the dataset
5 file_path = "/mnt/data/titanic.csv"
6 df = pd.read_csv(file_path)
7
8 # Convert column names to lowercase for consistency
9 df.columns = df.columns.str.lower()
10
11 # Basic Plots
12 plt.figure(figsize=(10, 5))
13
14 # Bar Chart: Survival Count
15 plt.subplot(1, 3, 1)
16 df['survived'].value_counts().plot(kind='bar', color=['red', 'green'])
17 plt.title("Survivors vs. Non-Survivors")
18 plt.xlabel("Survived (0 = No, 1 = Yes)")
19 plt.ylabel("Count")
20
21 # Histogram: Age Distribution
22 plt.subplot(1, 3, 2)
23 df['age'].dropna().plot(kind='hist', bins=20, color='blue', edgecolor='black')
24 plt.title("Age Distribution")
25 plt.xlabel("Age")
26 plt.ylabel("Count")
27
28 # Scatter Plot: Fare vs. Age
29 plt.subplot(1, 3, 3)
30 plt.scatter(df['age'], df['fare'], alpha=0.5, color='purple')
31 plt.title("Fare vs. Age")
32 plt.xlabel("Age")
33 plt.ylabel("Fare")
34
35 # Show the plots
36 plt.tight_layout()
37 plt.show()
```

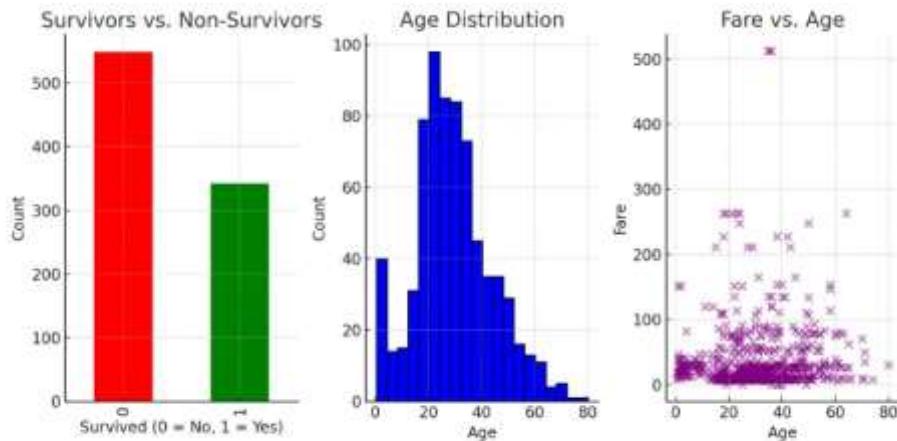


Fig. 9.2-8 LLM helps you get an instant visualization of your dataset data.

Data visualization is an important step to prepare the dataset for the subsequent construction of a machine learning model, which can only be accessed by understanding the data.

Machine learning in action: from Titanic passengers to project management

The main hypothesis used to explore the machine learning framework based on the Titanic dataset is that certain groups of passengers had a higher chance of survival.

The small table of Titanic passengers has become popular all over the world, and millions of people use it for training, experimentation and model testing to find out what algorithms and hypotheses will maximize the accurate survival prediction model based on the training dataset for Titanic passengers.

The appeal of the Titanic dataset is due to its compactness: with several hundred rows and twelve columns (Fig. 9.2-6), it provides ample opportunity for analysis. The dataset is, relatively simply, a classic example of a binary classification solution, where the goal of the problem - survival - is expressed in the convenient format 0 or 1.

John Wheeler in "It from Bit" [7] argues that the universe is based on binary choices. Similarly, a business run by people made up of molecules is actually built on a series of binary binary choices.

In addition, the data is based on a real historical event, which makes it valuable for research, unlike artificially created examples. On the Kaggle platform alone, one of the largest Data Pipeline and ETL, 1,355,998 people participated in the Titanic dataset-based challenges, developing 53,963 unique Data Pipeline solutions [157] (Fig. 9.2-9).

It seems unbelievable, but just 1000 lines of data on the Titanic passengers with 12 parameters have become a field for millions of hypotheses, logical chains and unique Data-Pipelines. From a small dataset are born endless insights, hypotheses and interpretations - from simple survival models to complex ensembles that take into account hidden patterns and complex labyrinths of reasoning.

The screenshot shows the Kaggle interface for the 'Machine Learning from Disaster' competition. At the top, there's a navigation bar with 'Data', 'Code', 'Models', 'Discussion', 'Leaderboard', and 'Rules'. A 'Submit Prediction' button is also visible. Below the navigation, there are five listed solutions:

- Titanic Tutorial**: Updated 3y ago, 29858 comments, Score: 0.80143, 16916 upvotes, Gold level.
- Titanic competition w/ TensorFlow Decision Forests**: Updated 2y ago, Score: 0.80143, 318 comments, 1098 upvotes, Gold level.
- Titanic Data Science Solutions**: Updated 6y ago, 2590 comments, 10723 upvotes, Gold level.
- Exploring Survival on the Titanic**: Updated 7y ago, Score: 0.80382, 1072 comments, 3968 upvotes, Gold level.

Fig. 9.2-9 The first five solutions out of a total of 53,963 ready and open source Pipeline solutions. Almost 1.5 million people have already tried to solve this problem on Kaggle alone [157].

If even such a small table can generate millions of unique solutions (Fig. 9.2-9), what can we say about real industrial construction datasets where parameters are measured in tens of thousands?

A standard CAD -design of a relatively small building contains tens of thousands of entities with thousands of parameters - from geometric characteristics to cost and time attributes. Imagine how many potential insights, relationships, predictions and management hypotheses are hidden in the data from all of your company's projects collected over the years. Historical project data is not just an archive - it is the living memory of an organization, its digital footprint that can be analyzed to build a large number of unique hypotheses.

Most importantly, you don't have to wait for the Kaggle community to take an interest in your company or your data. You can start working with what you have today: run analytics on your own data, train

models on your own data, identify repeats, anomalies, and patterns. Where it used to take years of experimentation and expensive consulting, today all you need is initiative, an LLM, an open approach to data, and a willingness to learn.

- 💡 To build a machine learning algorithm, which will predict passenger survival rates based on the train.csv passenger training dataset let's ask LLM to solve this problem for us:

Based on the Titanic passenger training dataset, build a machine learning model to predict survivability ↴

2 LLM's response:

```

1 import pandas as pd
2 import seaborn as sns
3 import matplotlib.pyplot as plt
4 from sklearn.model_selection import train_test_split
5 from sklearn.ensemble import RandomForestClassifier
6 from sklearn.metrics import accuracy_score

# Load the dataset
train = pd.read_csv('train.csv')

# Convert categorical data
train['Sex'] = train['Sex'].map({'male': 0, 'female': 1})
train = train.dropna(subset=['Age']) # Remove missing age values

# Select relevant features
features = ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']
X = train[features]
y = train['Survived']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Train the model
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Model Accuracy: {accuracy:.2f}')

```

Fig. 9.2-10 LLM built a prediction of Titanic survivors using the machine learning algorithm Random Forest.

The resulting code from LLM (Fig. 9.2-10) loads Titanic passenger data, cleans it, converts categorical variables (e.g., gender to numeric format), and trains the model through the RandomForestClassifier algorithm to predict whether a passenger survived or not (we will talk more about popular algorithms in the following chapters).

The code separates the training data into training and test sets (Kaggle's website has already created ready-made test.csv (Fig. 9.2-7) and train.csv (Fig. 9.2-6) for training, then the model is trained on the training data and tested on the test data to see how good a particular prediction model is. After training, the test data from test.csv (with real data about those who survived or did not survive) is fed into the

model and it predicts who survived and who did not. In our case, the accuracy of our machine learning model is about 80%, which shows that it captures the patterns quite well.

Machine learning can be compared to a child trying to fit a rectangular block into a round hole. In the initial stages, the algorithm tries many approaches, encountering errors and inconsistencies. This process may seem inefficient, but it provides important learning: by analyzing each error, the model improves its predictions and makes increasingly accurate decisions.

Now this model (Fig. 9.2-10) can be used to predict the survival rate of new passengers and for example, if you feed it with passenger information using the `model.predict` function the parameters: "male", "3rd class", "25 years old", "no relatives on board", the model will produce a prediction - that the passenger with 80% probability will not survive the catastrophe if he was on the Titanic ship in 1912 (Fig. 9.2-11).

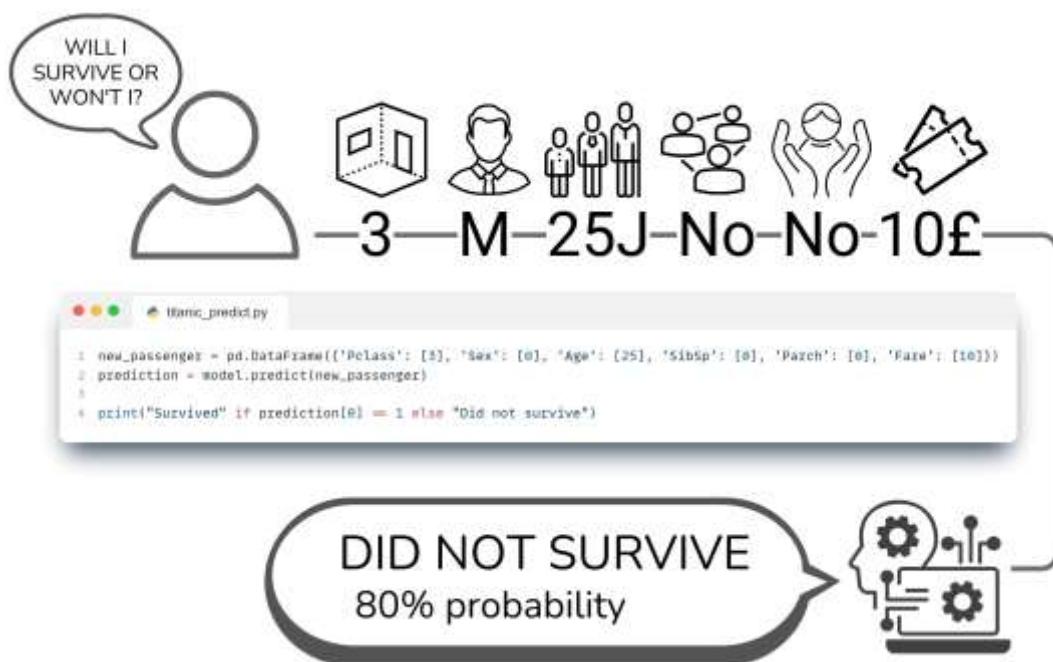


Fig. 9.2-11 The model we created above can now predict with 80% probability whether or not any new Titanic passenger will survive.

The Titanic passenger survival prediction model illustrates a much broader concept: every day, thousands of professionals in the construction industry make similar "dual" decisions - the life or death of a decision, a project, an estimate, a tool, profit or loss, safety or risk. As in the Titanic example, where the outcome depended on factors (gender, age, class), in construction each aspect of the decision is influenced by many of its own factors and variables (columns of tables): cost of materials, skill of workers, timing, weather, logistics, technical risks, comments and hundreds of thousands of other parameters.

In the construction industry, machine learning follows the same principles as in other fields: models are trained on historical data - from projects, contracts, estimates - to test various hypotheses and find the most effective solutions. This process is much like teaching a child through trial and error: with each cycle, the models adapt and become more accurate.

The use of accumulated data opens up new horizons for construction. Instead of time-consuming manual calculations, models can be trained that can predict key characteristics of future projects with a high degree of accuracy. In this way, predictive analytics transforms the construction industry into a space where you can not only plan, but also confidently predict developments.

Predictions and forecasts based on historical data

The data collected on the company's projects opens up the possibility of building models capable of predicting the cost and time characteristics of future, not yet realized objects - without time-consuming manual calculations and comparisons. This makes it possible to significantly speed up and simplify valuation processes, relying not on subjective assumptions, but on sound mathematical forecasts.

Earlier, in the fourth part of the book, we have considered in detail traditional methods of project cost estimation, including the resource-based method, and also mentioned parametric and expert approaches. These methods are still relevant, but in modern practice they are beginning to be enriched with tools of statistical analysis and machine learning, which allows to significantly improve the accuracy and reproducibility of estimates.

The processes of manual and semi-automatic calculation of prices and temporal attributes will in the future be complemented by the opinion and predictions of ML models capable of analyzing historical data, finding hidden patterns and offering informed decisions. New data and scenarios will be generated automatically from already available information - similar to how language models (LLM) create texts, images and code based on data collected over the years from open sources [158].

Just as humans today rely on experience, intuition and internal statistics to assess future events, in the coming years the future of construction projects will be increasingly determined by a combination of accumulated knowledge and mathematical machine learning models.

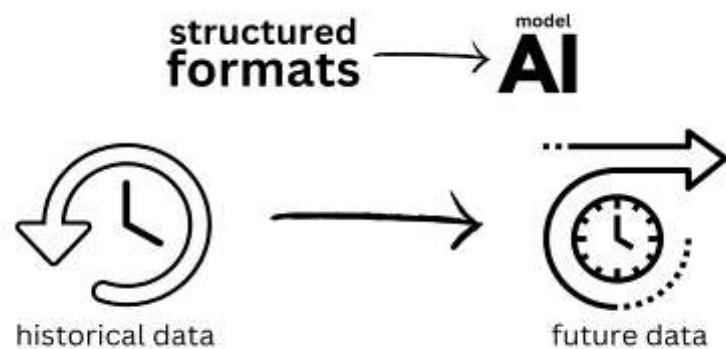


Fig. 9.2-12 Qualitative and structured historical company data is the material upon which machine learning models and predictions are built.

Consider a simple example: predicting the price of a house based on its area, plot size, number of rooms and geographical location. One approach is to build a classical model that analyzes these parameters and calculates the expected price (Fig. 9.2-13). This approach requires a precise and known formula in advance, which is practically impossible in real practice.

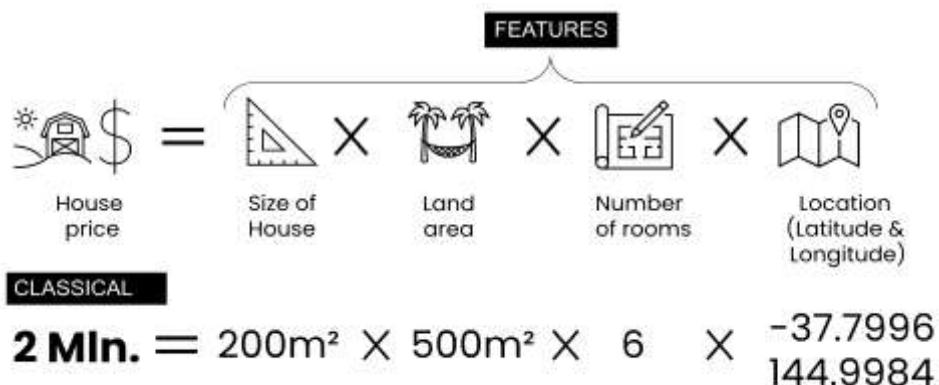


Fig. 9.2-13 A classical algorithm can be used to estimate the value of a house with a fixed formula to be found.

Machine learning eliminates the manual search for formulas and replaces them with trained algorithms that independently identify dependencies that are many times more accurate than any predetermined equations. Alternatively, let's create a machine learning algorithm, which will generate a model based on a prior understanding of the problem and historical data that may be incomplete (Fig. 9.2-14).

Using the pricing problem as an example, machine learning allows you to create different types of mathematical models that do not require knowledge of the exact mechanism of cost formation. The model "learns" from the data on previous projects, adjusting to real patterns between building parameters, their cost and deadlines.

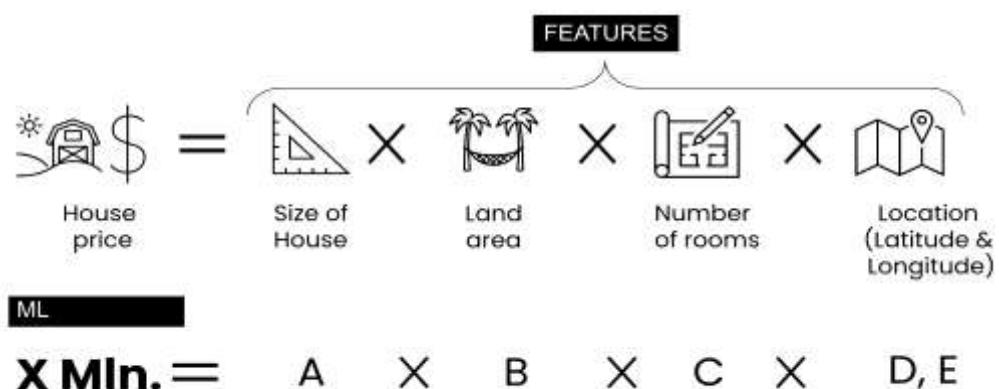


Fig. 9.2-14 Unlike classical formula-based estimation, the machine learning algorithm is trained on historical data.

In the context of supervised machine learning, each project in the training dataset contains both input attributes (e.g. cost and time data for similar buildings) and expected output values (e.g. cost or time). A similar dataset is used to create and customize a machine learning model (Fig. 9.2-15). The larger the dataset and the higher the quality of the data in it, the more accurate the model will be and the more accurate the prediction results will be.

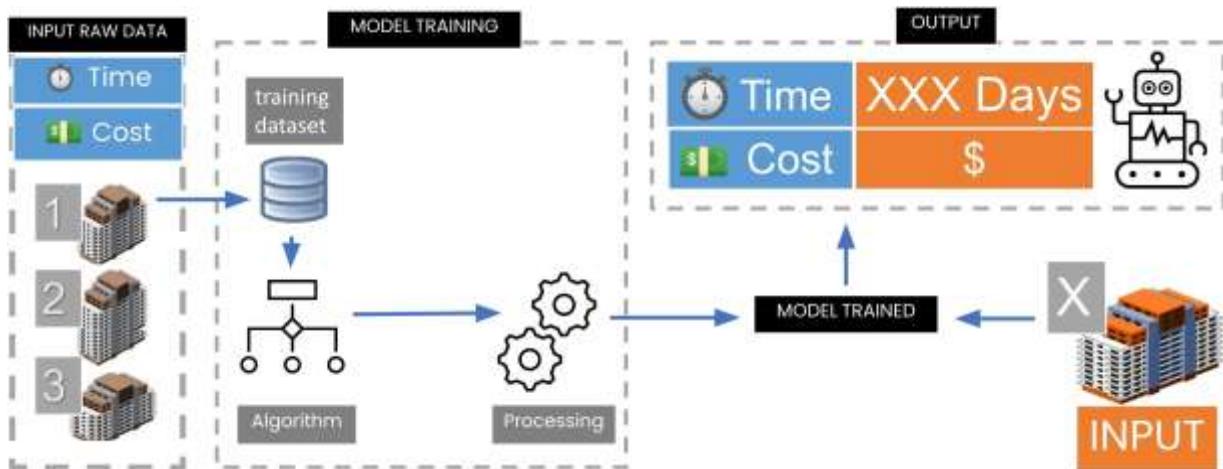


Fig. 9.2-15 An ML model trained on cost and schedule data from past projects will determine the cost and schedule of a new project with a certain probability.

Once the model is created and trained to estimate the construction of a new project, simply provide the model with new attributes for the new project, and the model will provide estimated results based on previously learned patterns with some probability.

Key concepts of machine learning

Machine learning is not magic, it's just math, data and finding patterns. It has no real intelligence, but

is a program trained on data to recognize patterns and make decisions without constant human involvement.

Machine learning uses a number of key concepts to describe its structure (Fig. 9.2-16):

- **Labels** are target variables or attributes (the "Survivor" parameter in the Titanic dataset) that the model should predict. Example: construction cost (e.g., in dollars), duration of construction work (e.g., in months).
- **Features** are independent variables or attributes that serve as inputs to the model. In a forecasting model, they are used to predict labels. Examples: plot area (in square meters), number of floors of a building, total floor area of a building (in square meters), geographic location (latitude and longitude), type of materials used in construction. The number of characteristics also determines the dimensionality of the data.
- **A model** is a set of different hypotheses, one of which approximates the target function to be predicted or approximated. Example: machine learning model, which uses regression analysis techniques to predict the cost and timing of construction.
- **Learning Algorithm** **Learning Algorithm** is the process of finding the best hypothesis in a model that exactly matches the target function using a set of training data. Example: A linear regression, KNN or random forest algorithm that analyzes cost and construction time data to identify relationships and patterns.
- **Training** - During the training process, the algorithm analyzes the training data, finding patterns that correspond to the relationship between input attributes and target labels. The result of this process is a trained machine learning model, ready for prediction. Example: a process in which an algorithm analyzes historical construction data (cost, time, facility characteristics) to create a predictive model.

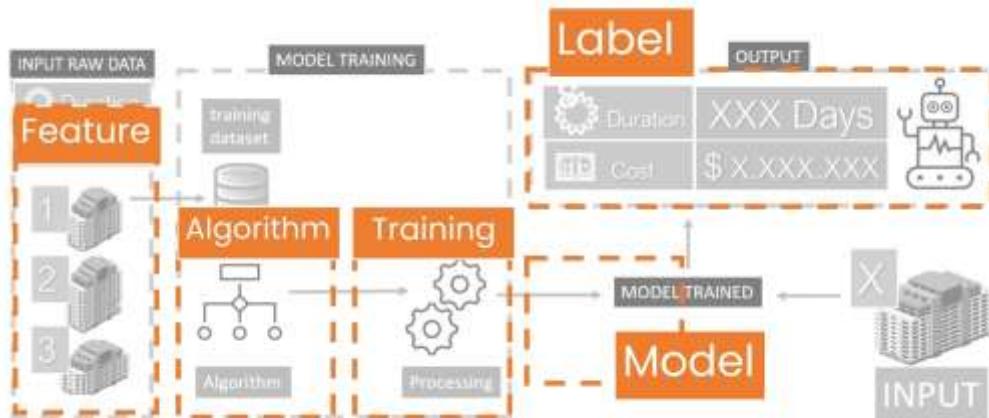


Fig. 9.2-16 ML uses labels and attributes to create models that are trained on data using algorithms to predict outcomes.

Machine learning does not exist in isolation, but is part of a broader ecosystem of analytical disciplines including statistics, databases, data mining, pattern recognition, big data analytics, and artificial intelligence. Fig. 9.2-17 demonstrates how these fields overlap and complement each other, providing a

comprehensive framework for modern decision-making and automation systems.

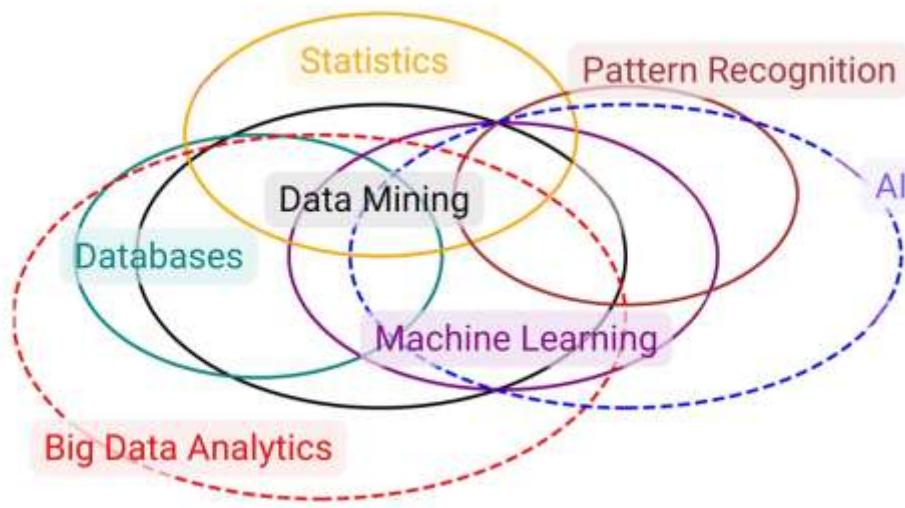
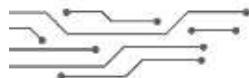


Fig. 9.2-17 The relationship between the different areas of data analysis: statistics, machine learning, artificial intelligence, big data, pattern recognition, and data mining.

The main goal of machine learning is to endow computers with the ability to automatically learn knowledge without human intervention or assistance and adjust their actions accordingly [159].

Thus, in the future, the human's role will only be to provide the machine with cognitive capabilities - they will set the conditions, weights and parameters, and the machine learning model will do the rest.

In the next chapter we will consider concrete examples of algorithm application. Real tables and simplified models will be used to show how the forecast is built step by step.



CHAPTER 9.3.

COST AND SCHEDULE FORECASTING USING MACHINE LEARNING

An example of using machine learning to find project cost and schedule

Estimation of construction time and cost is one of the key processes in the activities of a construction company. Traditionally, such estimates are made by experts based on experience, reference books and regulatory databases. However, with digital transformation and increasing data availability, it is now possible to use machine learning (ML) models to improve the accuracy and automation of such estimates.

The introduction of machine learning into the process of calculating the cost and timing of construction not only makes it possible to increase the efficiency of planning, but also becomes a starting point for integrating intelligent models into other business processes - from risk management to optimizing logistics and procurement.

It is important to be able to quickly determine how long it will take to build a project and what its total cost will be. These questions about project time and cost have traditionally been at the forefront of the minds of both clients and construction companies since the birth of the construction industry.

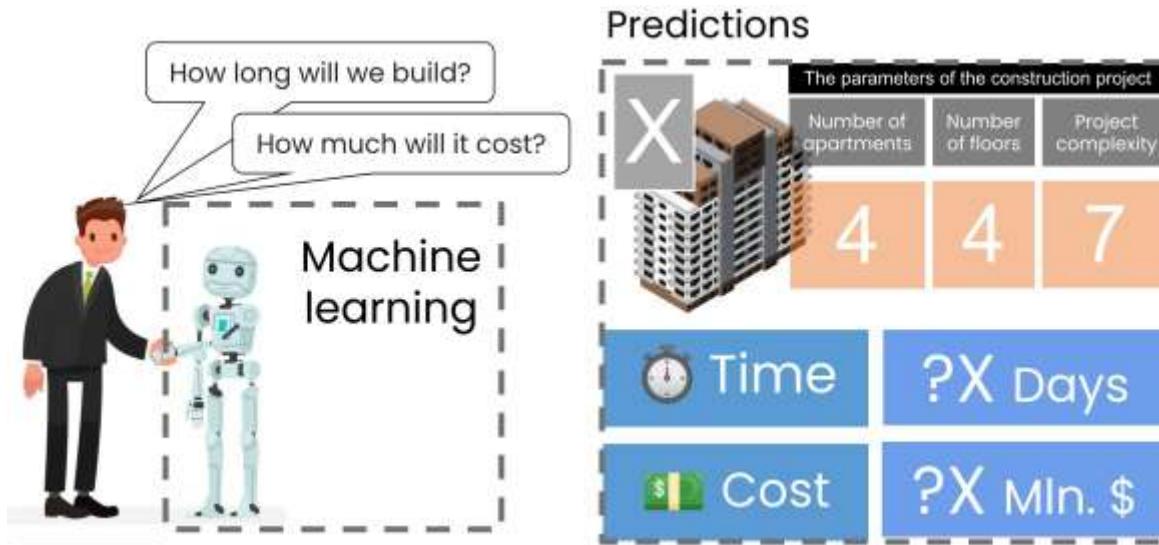


Fig. 9.3-1 In construction projects, the speed and quality of estimating construction time and cost are key success factors.

In the following example, key data from past projects will be extracted and used to develop a machine learning model, which will allow us to use the model to estimate the cost and timing of new construction projects with new parameters (Fig. 9.3-1).

Consider three projects with three key attributes: the number of apartments (where 100 apartments is equivalent to the number 10 for ease of visualization), the number of floors, and a conditional measure of construction complexity on a scale of 1 to 10, where 10 is the highest complexity score. In machine learning, the process of converting and simplifying values such as 100 to 10 or 50 to 5 is called "normalization".

Normalization in machine learning is the process of bringing different numerical data to a common scale to facilitate its processing and analysis. This process is especially important when the data has different scales and units.

Suppose that the first project (Fig. 9.3-2) had 50 apartments (after normalization, 5), 7 floors and a complexity score of 2, which meant a relatively simple construction. The second project already had 80 apartments, 9 floors and a relatively complex project. Under these conditions, construction of the first and second apartment building took 270 and 330 days, and the total project cost was \$4.5 million and \$5.8 million, respectively.

Construction project	The parameters of the construction project			The key parameters of the project	
	Number of apartment	Number of floors	Project complexity	Time	Cost
1	5	7	2	270	\$ 4.502.000
2	8	9	6	330	\$ 5.750.000
3	3	5	3	230	\$ 3.262.000
X	4	4	7	?X	\$?X. XXX.XXX

Fig. 9.3-2 An example of a set of past projects that will be used to estimate the time and cost of future project X.

When building a machine learning model for such data, the main task is to identify critical attributes (or labels) for prediction, in this case, construction time and cost. With a small dataset, we will use information about previous construction projects to plan new ones: using machine learning algorithms, we have to predict the construction cost and duration of a new project X based on given attributes of the new project, such as 40 apartments, 4 floors, and a relative high project complexity of 7 (Fig. 9.3-2). In a real-world setting, the number of input parameters can be much larger, ranging from several tens to hundreds of factors. These may include: type of construction materials, climatic zone, qualification level of contractors, availability of utilities, type of foundation, season of commencement of works, comments of foremen, etc.

To create a predictive machine learning model, we need to choose an algorithm to create it. An algorithm in machine learning is like a mathematical recipe that teaches the computer how to make predictions (mix in the right order of parameters) or make decisions based on data.

To analyze data on past construction projects and predict the timing and cost of future projects (Fig. 9.3-2), one popular machine learning algorithm can be used:

- **Linear regression (Linear regression):** this algorithm tries to find a direct relationship between attributes, for example between the number of floors and the construction cost. The goal of the algorithm is to find a linear equation that best describes this relationship, which allows making predictions.
- **Algorithm k-nearest neighbors (k-NN):** this algorithm compares a new project with past projects that were similar in size or complexity. The k-NN classifies the data based on which of the k (number) training examples are closest to it. In the context of regression, the result is the mean or median of the k nearest neighbors.
- **Decision Trees:** is a predictive modeling model that divides data into subsets based on different conditions using a tree structure. Each node of the tree represents a condition or question leading to further division of the data, and each leaf represents the final prediction or outcome. The algorithm divides the data into smaller groups based on different characteristics, such as first by number of stories, then by complexity and so on, to make a prediction.

Let's take a look at machine learning algorithms for estimating the cost of a new project using two popular algorithms as examples: linear regression and the K-nearest neighbors algorithm.

Project cost and time prediction using linear regression

Linear regression is a fundamental data analysis algorithm that predicts the value of a variable based on a linear relationship with one or more other variables. This model assumes that there is a direct linear relationship between the dependent variable and one or more independent variables, and the goal of the algorithm is to find this relationship.

The simplicity and clarity of linear regression has made it a popular tool in a variety of fields. When dealing with a single variable, linear regression is about finding the best fitting line through the data points.

Linear regression finds the best straight line (red line) that approximates the dependence between input variable X and output variable Y. This line allows predicting Y values for new X values based on the linear relationship identified (Fig. 9.3-3).

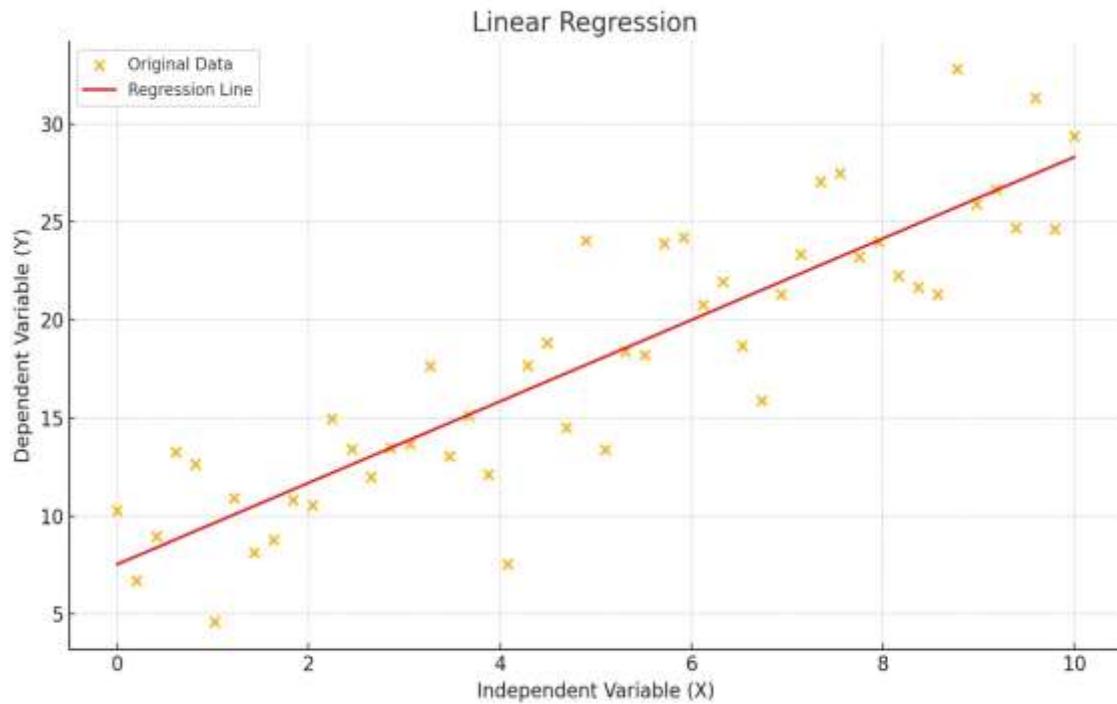


Fig. 9.3-3 The principle of linear regression is to find the best straight line through the training values.

This line is represented by an equation in which inputting a value of the independent variable (X) produces a predicted value of the dependent variable (Y). This process allows Y to be effectively predicted from known values of X using a linear relationship between them. An example of finding such a statistically averaged line can be seen in the San Francisco Building Permit Data Assessment (Fig. 9.1-7), where inflation using linear regression was calculated for different types of facilities.

Let's load the project data table (Fig. 9.3-2 from the previous chapter) directly into the LLM and have it build a simple machine learning model for us.

- 👉 Send a text request to LLM chat (CHATGP, LlaMa, Mistral DeepSeek, Grok, Claude, QWEN):

We need to show the construction of a simple machine learning model for predicting the cost and time of a new project X (Fig. 9.3-2 as attached image) ↴

2 LLM's response:

```

1  from sklearn.linear_model import LinearRegression
2  import numpy as np
3
4  # Data from the image
5  # Features: Number of apartments, number of floors, project complexity
6  X = np.array([
7      [5, 7, 2],
8      [8, 9, 6],
9      [3, 5, 3]
10 ])
11
12 # Targets: Time (days), Cost ($)
13 y_time = np.array([270, 330, 230])
14 y_cost = np.array([4502000, 5750000, 3262000])
15
16 # Creating linear regression models for time and cost
17 model_time = LinearRegression()
18 model_cost = LinearRegression()
19
20 # Training the models
21 model_time.fit(X, y_time)
22 model_cost.fit(X, y_cost)
23
24 # The new project's features
25 X_new = np.array([[4, 4, 7]])
26
27 # Predicting the time and cost for the new project
28 predicted_time = model_time.predict(X_new)
29 predicted_cost = model_cost.predict(X_new)
30
31 predicted_time, predicted_cost

```

Result
(array([238.44444444]), array([3042337.77777778]))

Fig. 9.3-4 LLM chose linear regression to create a machine learning model to predict project cost and time.

The LLM automatically recognized the table from the attached image and converted the data from a visual format to a table array (Fig. 9.3-4 - line 6). This array was used as the basis for creating features and labels from which a machine learning model was created (Fig. 9.3-4 - 17th-22nd line), which used linear regression.

Using a basic linear regression model that was trained on an "extremely small" data set, predictions were made for a new hypothetical construction project, labeled Project X. In our problem, this project

is characterized by having 40 apartments, 4 floors, and a complexity level of 7 (Fig. 9.3-2).

As predicted by a linear regression model based on a limited and small data set for the new Project X (Fig. 9.3-4 - line 24-29):

- The **construction duration** will be approximately 238 days (238.4444444)
- The **total cost** will be approximately \$3,042,338 (3042337.777)

To further explore the project cost hypothesis, it is useful to experiment with different machine learning algorithms and methods. Therefore, let's predict the same cost and time values for a new project X based on a small set of historical data using the K-Nearest Neighbours algorithm (k-NN).

Project cost and time predictions using the K-nearest neighbor algorithm (k-NN)

We use the k-Nearest Neighbors (k-NN) algorithm as an additional predictor to estimate the cost and duration of a new project. The K-Nearest Neighbors (k-NN) algorithm is a supervised machine learning (supervised machine learning) method for both classification and regression. We have also previously discussed the k-NN algorithm in the context of vector database search (Fig. 8.2-2), where it is used to find the closest vectors (e.g., texts, images, or technical descriptions). In this approach, each project is represented as a point in a multidimensional space, where each dimension corresponds to a specific attribute of the project.

In our case, given the three attributes of each project, we will represent them as points in a three-dimensional space (Fig. 9.3-5). Thus, our upcoming project X will be localized in this space with coordinates ($x=4$, $y=4$, $z=7$). It should be noted that in real conditions, the number of points and the dimensionality of the space may be orders of magnitude larger.

The K-NN (k-nearest neighbors) algorithm works by measuring the distance between the desired project X and the projects in the training database. By comparing these distances, the algorithm determines the projects that are closest to the point of the new project X.

For example, if the second project ($x=8$, $y=9$, $z=6$) from our original dataset is much farther away from X (Fig. 9.3-5) than the other projects, it can be excluded from further analysis. As a result, only the two ($k=2$) nearest projects can be used for calculations, based on which the average value will be determined.

Such a method, through a neighborhood search, allows us to assess the similarities between projects, which in turn helps us to draw conclusions about the possible cost and timing of a new project based on similar projects that have been implemented previously.

k-nearest neighbors algorithm

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

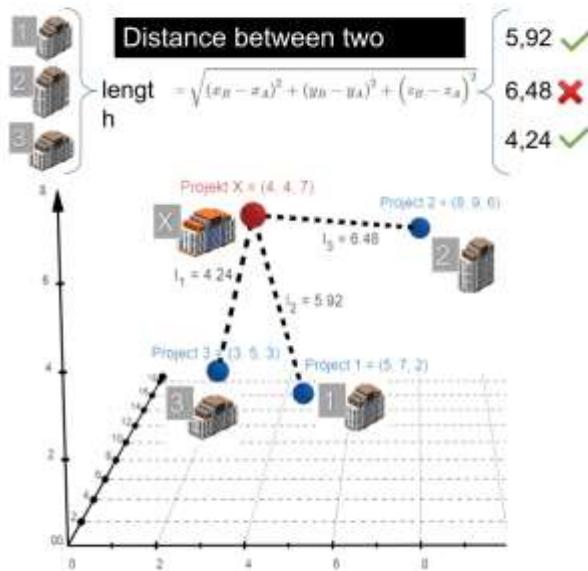
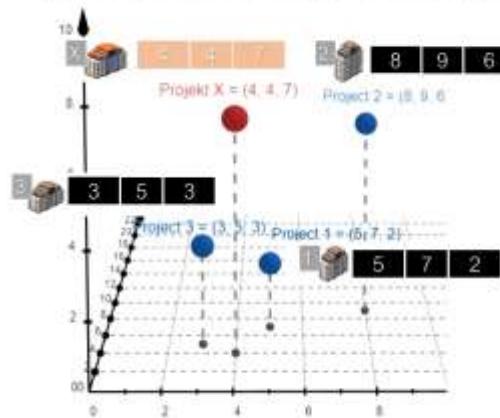


Fig. 9.3-5 In the K-NN algorithm, projects are represented as points in a multidimensional space, and nearby projects are selected based on distances to evaluate similarity and make predictions.

The work of k-NN involves several key steps:

- **Data preparation:** training and test data sets are loaded first. Training data is used to "train" the algorithm, and test data is used to check its efficiency.
- **Selecting the parameter K:** a number K is selected, which indicates how many nearest neighbors (data points) should be considered in the algorithm. The value of "K" is very important because it affects the result.
- Classification process and regression for test data:
 - **Calculating distances:** for each element from the test data, the distance to each element from the training data is calculated (Fig. 9.3-5). Different distance measurement methods can be used for this, such as Euclidean distance (the most common method), Manhattan distance or Hamming distance.
 - **Sorting and selecting K nearest neighbors:** after calculating the distances, they are sorted and K nearest points to the test point are selected.
 - **Determining the class or value of a test point:** if it is a classification task, the class of the test point is determined based on the most frequent class among K selected neighbors. If it is a regression task, the mean (or other measure of central tendency) of the K neighbors' values is calculated.
- **Completion of the process:** once all test data has been classified or predictions have been made for it, the process is complete.

The algorithm k-nearest neighbors (k-NN) is effective in many practical applications and is one of the main tools in the arsenal of machine learning specialists. This algorithm is popular due to its simplicity and efficiency, especially in tasks where relationships between data are easy to interpret.

In our example, after applying the K-nearest neighbor algorithm, the two projects (from our small sample) with the shortest distance to project X were identified (Fig. 9.3-5). Based on these projects, the algorithm determines the average of their price and construction duration. After analysis (Fig. 9.3-6), the algorithm, by averaging the nearest neighbors, concludes that project X will cost approximately \$3,882,000 and take about 250 days to complete.

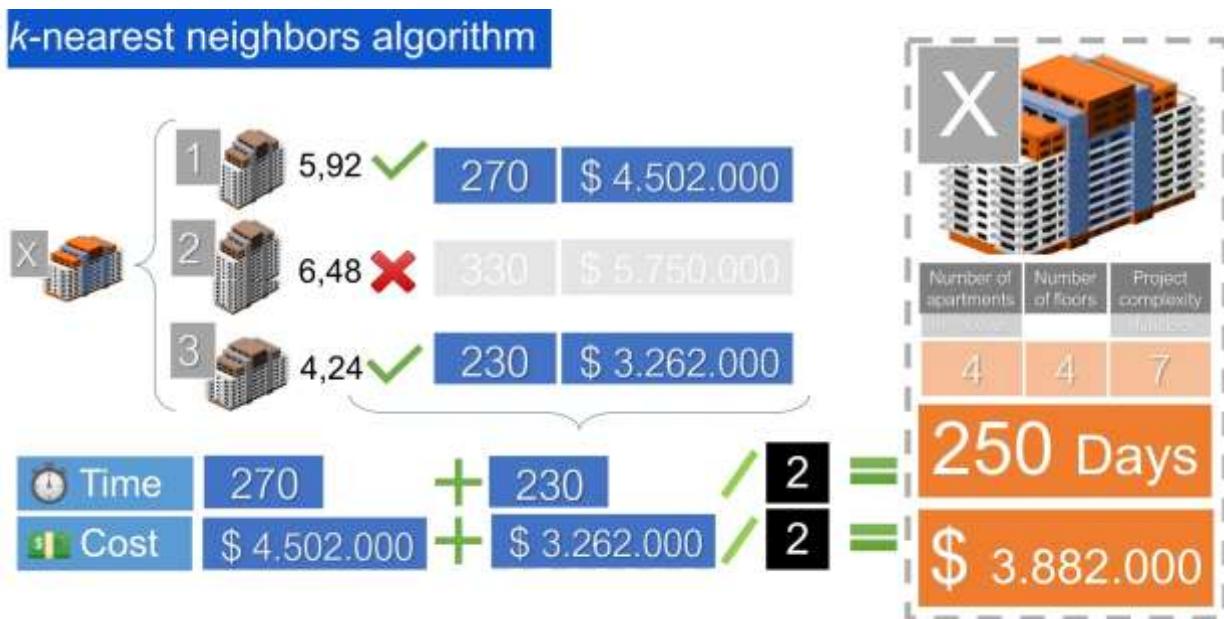


Fig. 9.3-6 The K-nearest neighbors algorithm determines the cost and schedule of project X by analyzing the two closest projects in the sample.

The k-Nearest Neighbors (k-NN) algorithm is particularly popular in classification and regression tasks, such as recommendation systems, where it is used to suggest products or content based on preferences similar to the interests of a particular user. In addition, k-NN is widely used in medical diagnostics to classify types of diseases based on patient symptoms, in pattern recognition, and in the financial sector to assess the creditworthiness of customers.

Even with limited data, machine learning models can provide useful predictions and significantly enhance the analytical component of construction project management. As historical data is expanded and cleaned up, it is possible to move to more complex models - for example, taking into account the type of construction, location, season of construction start and other factors.

Our simplified problem used three attributes for visualization in three-dimensional space, but real projects, on average, include hundreds or thousands of attributes ((see the dataset from the chapter "Example of CAD (BIM) based big data"), which greatly increases the dimensionality of the space and the complexity of representing projects as vectors (Fig. 9.3-7).

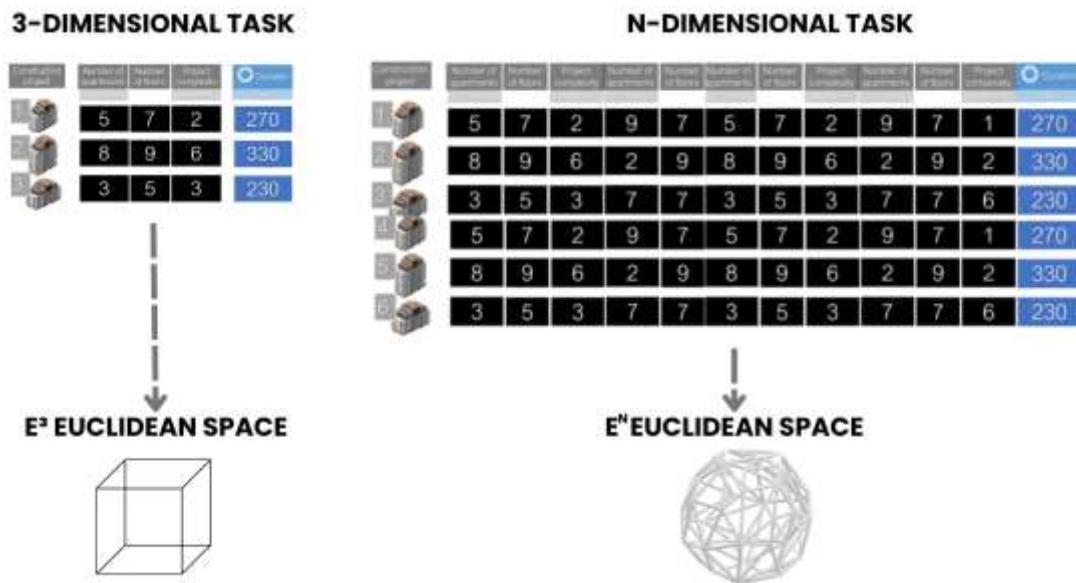


Fig. 9.3-7 In the simplified example, three attributes were used for 3D -visualization, while real projects have more.

Applying different algorithms to the same data set for project X, which has 40 apartments, 4 floors, and a complexity level of 7, yielded different predicted values. The linear regression algorithm predicted a completion time of 238 days and a cost of \$ 3,042,338 (Fig. 9.3-4), while the k-NN algorithm predicted 250 days and \$ 3,882,000 (Fig. 9.3-6).

The accuracy of predictions obtained using machine learning models, directly depends on the volume and quality of the input data. The more projects are involved in training, and the more completely and accurately their characteristics (attributes) and results (labels) are represented, the higher the probability of obtaining reliable predictions with minimal error values.

Data preprocessing techniques play an important role in this process, including:

- Normalization to bring features to a common scale;

- Outlier detection and elimination that eliminates model distortion;
- Coding of categorical attributes to allow manipulation of textual data;
- Filling in missing values, increasing the stability of the model.

In addition, cross-validation methods are used to assess the generalizability of the model and its robustness to new datasets to detect overfitting and improve the reliability of the prediction.

Chaos is an order to be deciphered [160].

— José Saramago, "The Double"

Even if it seems to you that the chaos of your tasks cannot be described formally, know that any event in the world and especially construction processes are subject to mathematical laws, which may need support for calculating values not through strict formulas but with the help of statistics and historical data.

Both traditional estimates performed by estimating departments and machine learning models inevitably face uncertainty and potential sources of error. However, when sufficient quality data is available, machine learning models can demonstrate comparable and sometimes even higher prediction accuracy than expert estimates.

Machine learning is likely to become a reliable complementary tool for analysis, allowing to: refine calculations, propose alternative scenarios and identify hidden dependencies between project parameters. Such models will not claim to be universal, but they will soon occupy an important place in calculations and project decision-making processes. Machine learning technologies will not exclude the participation of engineers, estimators and analysts, but, on the contrary, will expand their capabilities by offering an additional point of view based on historical data.

If properly integrated into the business processes of construction companies, machine learning has the potential to become an important element in the management decision support system - not as a replacement for humans, but as an extension of their professional intuition and engineering logic.

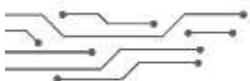
Next steps: from storage to analysis and forecasting

Modern approaches to working with data are beginning to change decision-making in the construction industry. Moving from intuitive assessments to objective data analysis not only improves accuracy, but also opens up new opportunities for process optimization. To summarize this part, it is worth highlighting the main practical steps that will help you apply the discussed methods in your daily tasks:

- Building a resilient storage infrastructure

- Try to combine disparate documents and project data into a single tabular model, aggregating key information into a single dataframe for further analysis
 - Use efficient data storage formats - for example, columnar formats like Apache Parquet instead of CSV or XLSX - especially for those sets that could potentially be used to train machine learning models in the future
 - Create a data versioning system to track changes throughout the project
- Implementation of analysis and automation tools
- Start analyzing historical project data - by documentation, models, estimates - to identify patterns, trends and anomalies
 - Master ETL processes (Extract, Transform, Load) to automatically load and prepare data
 - Learn how to visualize key metrics using various free Python visualization libraries
 - Begin to apply statistical methods and random sampling to produce representative and reproducible analytical conclusions
- Increasing maturity in working with data
- Learn a few basic machine learning algorithms with simple and straightforward examples like the Titanic dataset
 - Analyze current processes and identify where you can move from rigid cause-and-effect logic to statistical methods of forecasting and estimation
 - Start treating data as a strategic asset rather than a by-product: build decision-making processes around data models rather than around specific software solutions

Construction companies that have realized the value of data are entering a new phase of development where competitive advantage is determined not by the amount of resources, but by the speed of decision-making based on analytics.



MAXIMUM CONVENIENCE WITH THE PRINTED VERSION

You are holding a free digital version of **Data-Driven Construction**. For more convenient work and quick access to the materials, we recommend you to pay attention to the **printed edition**:



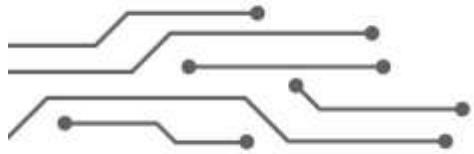
■ **Always at hand:** the book in printed format will become a reliable working tool, allowing you to quickly find and use the necessary visualizations and diagrams in any working situation

■ **High quality illustrations:** all images and graphics in the printed edition are presented in maximum quality

■ **Quick access to information:** easy navigation, the ability to make notes, bookmarks and work with the book anywhere.

By purchasing the full printed version of the book, you get a convenient tool for comfortable and efficient work with information: the ability to quickly use visual materials in everyday tasks, quickly find the necessary schemes and make notes. In addition, your purchase supports the dissemination of open knowledge.

Order the print version of the book at: datadrivenconstruction.io/books



X PART

THE CONSTRUCTION INDUSTRY IN THE DIGITAL AGE. OPPORTUNITIES AND CHALLENGES

The final tenth part is a comprehensive look at the future of the construction industry in the era of digital transformation. It analyzes the shift from causal analysis to working with big data correlations. Parallels are drawn between the evolution of fine art and the development of data work in construction, demonstrating how the industry is moving from detailed control to a holistic understanding of processes. The concept of the "uberization" of the construction industry is examined, where data transparency and calculation automation can radically change traditional business models, eliminating the need for intermediaries and reducing opportunities for speculation. Unresolved issues, such as the universal classification of elements, are discussed in detail, giving construction companies time to adapt to the new environment. The part concludes with specific recommendations for shaping a digital transformation strategy that includes analyzing vulnerabilities and expanding services to remain competitive in a changing industry.

CHAPTER 10.1.

SURVIVAL STRATEGIES: BUILDING COMPETITIVE ADVANTAGE

Correlations instead of calculations: the future of construction analytics

Due to the rapid digitalization of information (Fig. 1.1-5), modern construction is undergoing a fundamental transformation where data is becoming not just a tool but a strategic asset that can fundamentally change traditional approaches to project and business management.

For millennia, construction activities have relied on deterministic methods - precise calculations, detailing and strict control of parameters. In the first centuries AD, Roman engineers applied mathematical principles to the construction of aqueducts and bridges. In the Middle Ages, architects strived for ideal proportions of Gothic cathedrals, and in the industrialization of the 20th century, systems of standardized norms and regulations were formed, which became the basis for mass construction

Today, the vector of development is shifting from the search for strictly cause-and-effect relationships to probabilistic analysis, the search for correlations and hidden patterns. The industry is entering a new phase - data is becoming a key resource, and analytics based on it is replacing intuitive and locally optimized approaches.



Fig. 10.1-1 Hidden potential of construction data: existing calculations in the company are only the tip of the iceberg available for management to analyze.

A company's information system is like an iceberg (Fig. 10.1-1): only a small part of the data's potential is visible to the company's management, while the main value is hidden in the depths. It is important to evaluate data not only by its current use, but also by the opportunities it will unlock in the future. It is those companies that learn how to extract hidden patterns and create new insights from data that will create a sustainable competitive advantage

Finding hidden patterns and making sense of data is not just about working with numbers, but a creative process that requires abstract thinking and the ability to see the whole picture behind disparate elements. In this sense, the evolution of working with data can be compared to the evolution of the visual arts (Fig. 10.1-2).

The development of construction is remarkably similar to the progress of fine arts. In both cases, mankind has progressed from primitive methods to sophisticated visualization and analysis technologies. In prehistoric times, people used cave drawings and primitive tools to solve everyday tasks. During the Middle Ages and Renaissance, the level of sophistication in architecture and art increased dramatically. By the early Middle Ages, the tools of construction had evolved from a simple axe to extensive tool kits symbolizing the growth of technical knowledge.

The Age of Realism was the first revolution in the visual arts: artists learned to reproduce the smallest details, maximizing the verisimilitude of the image. In construction, the counterpart of this period was precise engineering techniques, detailed drawings and strictly regulated calculations, which became the basis of design practice for centuries.

Later, Impressionism changed the very perception of artistic reality: instead of literally rendering form, artists began to capture mood, light, and dynamics, aiming to reflect an overall impression rather than absolute accuracy. Similarly, machine learning in building analytics is moving away from rigid logical models to pattern recognition and probabilistic patterns that allow you to "see" hidden dependencies in the data that are inaccessible in classical analysis. This approach echoes Bauhaus' ideas of minimalism and functionality, where meaning (function) is more important than form. Bauhaus sought to remove the superfluous, to abandon ornamentation for the sake of clarity, utilitarianism and mass appeal. Things had to be understandable and useful, without excess - aesthetics was born from the logic of design and purpose.

With the advent of photography in the late 19th century, art gained a new tool to capture reality with unprecedented accuracy and turned attitudes towards fine art upside down. Similarly, in construction, the industrial revolution in the 21st century is leading to the use of robotic technologies, lasers, IoT, RFID and concepts like Connected Construction, where the collection of individual parameters has evolved to a scalable intelligent capture of the full reality of a construction site.

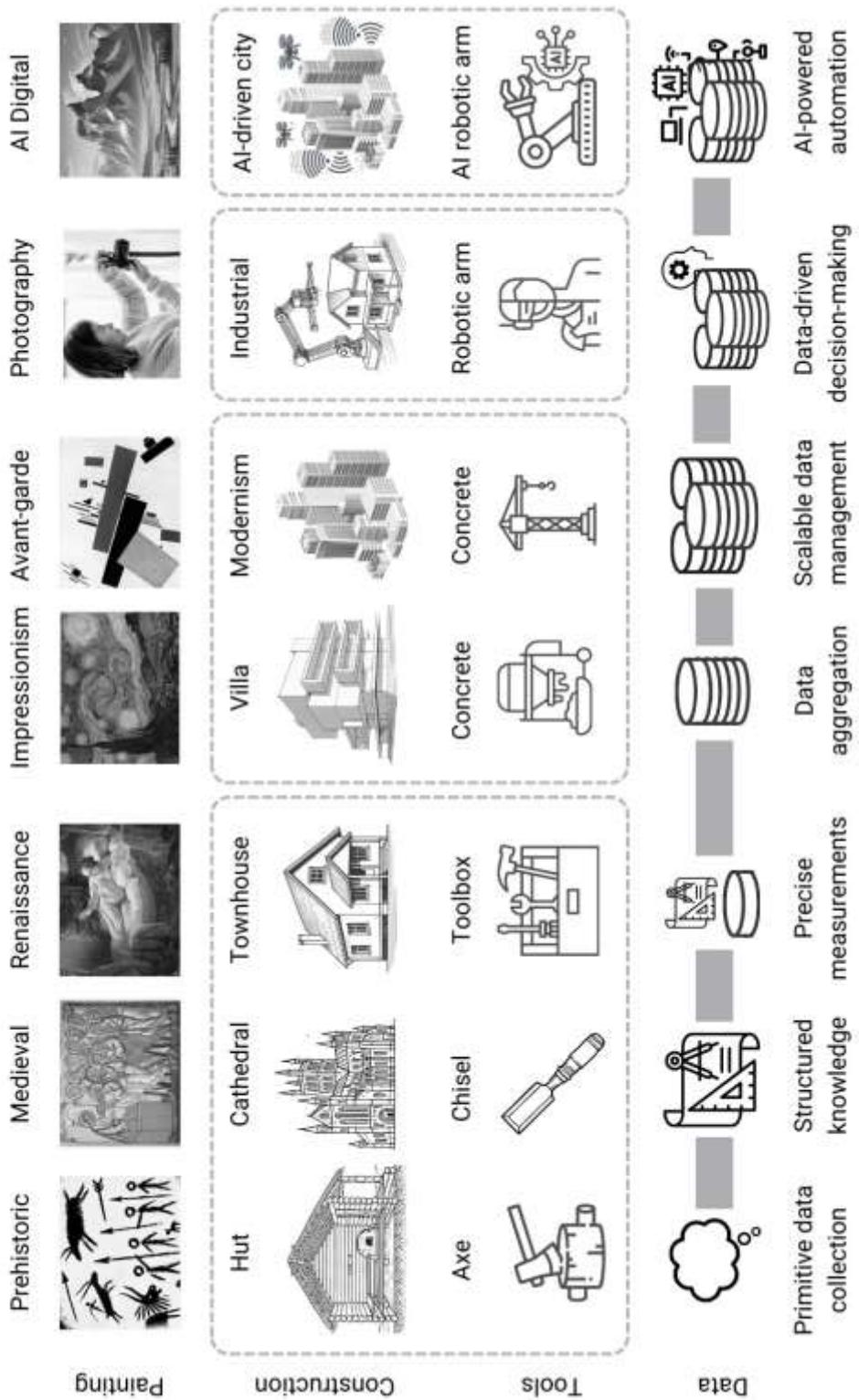


Fig. 10.1-2 Era Evolutions the visual arts are consistent with developments in approaches to working with data in the construction industry.

Today, just as the visual arts are experiencing a rethinking with the arrival of AI tools and LLM, the construction industry is experiencing another quantum leap: intelligent systems driven by artificial intelligence (AI), LLM chats allow predicting, optimizing and generating solutions with minimal human intervention.

The role of data in design and management has changed radically. Whereas knowledge used to be transmitted verbally and empirical in nature - just as reality was captured by hand painted pictures until the 19th century - today the focus is on the complete digital capture of the construction "picture". With the help of machine learning algorithms, this digital picture is transformed into an impressionistic representation of construction reality - not an exact replica, but a generalized, probabilistic understanding of the processes.

We are rapidly approaching an era in which the processes of designing, constructing and operating buildings will not just be augmented, but largely driven by artificial intelligence systems. Just as modern digital art is created without a paintbrush - using textual prompts and generative models - architectural and engineering solutions of the future will be shaped by key queries and parameters set by the user.

In the 21st century, access to data, its interpretation and the quality of analytics are becoming indispensable to the success of a project. And the value of data is determined not by its volume, but by the ability of specialists to analyze, verify and turn it into action.

Data-driven approach in construction: a new level of infrastructure

In the history of mankind, each such technological leap has brought fundamental changes to the economy and society. Today, we are witnessing a new wave of transformation comparable in scale to the industrial revolution of the 19th century. However, while a hundred years ago the main driver of change was mechanical forces and energy technologies, now it is data and artificial intelligence.

Machine learning, LLM and AI agents change the very essence of applications, making traditional software stacks (discussed in the second part of the book) unnecessary (Fig. 2.2-3). All data logic is centered in AI agents rather than in hard-coded business rules (Fig. 2.2-4).

In the data age, traditional views of applications are being fundamentally transformed. We are moving toward a model where bulky, modular enterprise systems will inevitably give way to open, lightweight, specialized solutions.

In the future, only the underlying data structure will remain, and all interaction with it will be through agents working directly with the database. I truly believe that the entire application stack will disappear because it is simply not necessary when artificial intelligence interacts directly with the underlying database. I've spent my entire career working in SaaS – building companies, working in them, and to be honest, I probably wouldn't launch a new SaaS business right now. And I probably wouldn't invest in SaaS companies right now either. The situation is too uncertain. That's not to say that there won't be software companies in the future, just that they will look very different. Future systems will be databases with business logic brought into [AI] agents. These agents will work with multiple data repositories at the same time, not limited to a single database. All logic will move into the AI layer [46].

– Matthew Berman, CEO Forward Future

The key difference of the new paradigm is the minimization of technological ballast. Instead of monumental complex and closed software systems, we will get flexible, open and quickly customizable modules that literally "live" inside the data flow (Fig. 7.4-1 - Apache Airflow, NiFi). The architecture of future process management envisions micro applications - compact, targeted tools, fundamentally different from massive and closed ERP, PMIS, CDE, CAFM systems. The new agents will be as adaptive, integrated and business task-oriented as possible (e.g. Low-Code/No-Code Fig. 7.4-6).

All the business logic will go to these [AI] agents, and these agents will perform CRUD [Create, Read, Update, and Delete] operations on multiple repositories, meaning they will not distinguish which backend is being used. They will update multiple databases, and all the logic will end up in the so-called AI -level. And once the AI layer is where all the logic is, people will start replacing backends. We're already seeing a pretty high percentage of market wins in Dynamics backends and agent utilization, and we're going to be moving aggressively in that direction, trying to bring it all together. Whether it's in customer service or in other areas, like not just CRM, but our finance and operations solutions. Because people want more AI-driven business applications where the logic layer can be driven by AI and AI agents. [...]. One of the most exciting things for me is Excel with Python, which is comparable to GitHub with Copilot. I mean, what we've done is: now that you have Excel, you should just open it up, run Copilot, and start playing with it. It's no longer just understanding the available numbers – it will make a plan on its own. Just like the GitHub Copilot Workspace creates a plan and then executes it, it's like a data analyst, using Excel as a tool to visualize rows and columns for analysis. So Copilot uses Excel as a tool with all its capabilities because it can generate data and has a Python interpreter.

– Satya Nadella, CEO, Microsoft, interview with BG2 channel December 2024 [28]

The transformation we are witnessing in the logic of office applications - moving from modular, closed systems to AI agents working directly with open data - is only part of a much larger process. It is not just a matter of changing interfaces or software architecture: the changes will affect the fundamental principles of work organization, decision-making, and business management. In construction, this will

lead to a data-driven logic in which data becomes the centerpiece of processes, from design to resource management to construction monitoring.

The next-generation digital office: how AI is changing the workspace

Almost a century ago, mankind was already experiencing a similar technological revolution. The transition from steam engines to electric motors took more than four decades, but ultimately catalyzed an unprecedented increase in productivity - primarily due to the decentralization of energy capacity and the flexibility of new solutions. This shift not only changed the course of history, moving the bulk of the population from rural to urban areas, but also laid the foundation for modern economies. The history of technology is a journey from physical labor to automation and intelligent systems. Just as the tractor replaced scores of tillers of the soil, today's digital technologies are replacing traditional office-based construction management methods (Fig. 10.1-3). As recently as the early 20th century, most of the world's population worked the land by hand until the mechanization of labor with machines and tractors began in the 1930s.

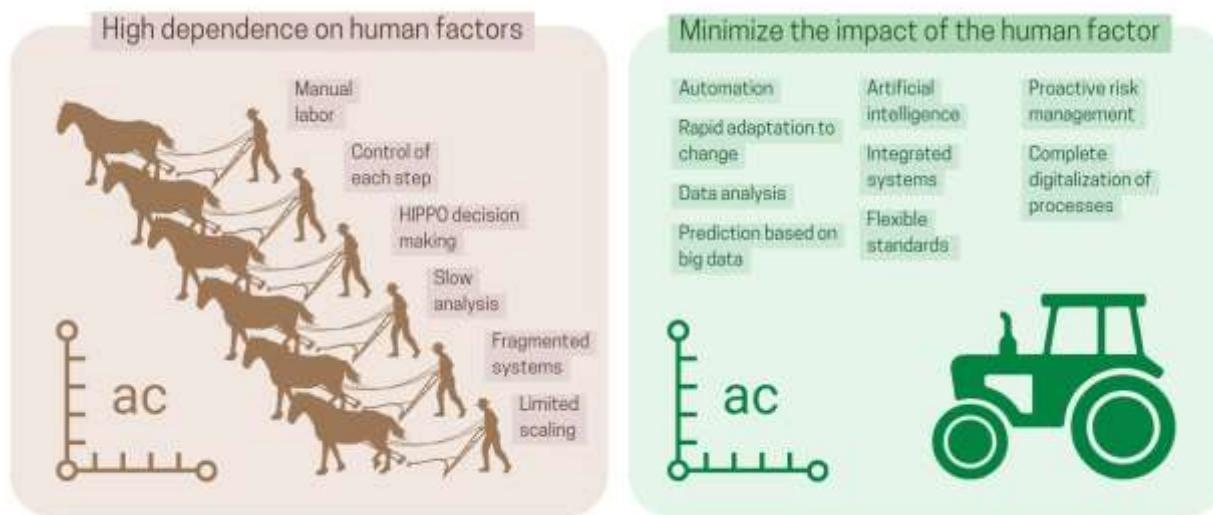


Fig. 10.1-3 Just as the tractor replaced dozens of people in the early 20th century, machine learning will replace traditional business and project management methods in the 21st century.

Just as humanity a hundred years ago moved from cultivating individual plots of land with primitive tools to large-scale agriculture using machinery, today we are making the transition from processing disparate "silos" of information to working with arrays of data using powerful "tractors" - ETL-pipeline and artificial intelligence algorithms.

We are on the cusp of a similar leap - but on the digital plane: from traditional, manual business management to data-driven models.

The path to a full-fledged data-driven architecture will require time, investments and organizational efforts. But this path opens the way to not just gradual improvement, but a qualitative leap towards greater efficiency, transparency and manageability of construction processes. All this is subject to

systematic implementation of digital tools and abandonment of outdated business practices.

Task parameterization, ETL, LLM, IoT components, RFID, tokenization, big **data** and machine learning will transform traditional construction into **data-driven construction**, where every detail of the project and construction business will be controlled and optimized by data.

It used to take thousands of man-hours to analyze information. Now these tasks are performed by algorithms and LLMs that turn disparate data sets with the help of prompts into strategic sources. In the tech world, the same thing that happened to agriculture is happening: we are moving from hoeing to automated agribusiness. So too, office work in construction - from Excel files and manual summarization - is moving to an intelligent system where data is collected, cleaned, structured and turned into insights.

Already today, companies should start "cultivating" information fields with the help of quality data collection and information structuring, and "fertilize" them with cleaning and normalization tools, and then "harvest" them in the form of predictive analytics and automated solutions. If a modern farmer with a machine is able to replace a hundred tillers of the soil, intelligent algorithms will be able to remove the routine from employees and transfer them to the role of strategic managers of information flows.

However, it is important to understand that creating a truly data-driven organization is not a quick process. It is a long-term strategic direction, similar to creating a new site for growing a new forest (Fig. 1.2-5) of systems, where each "tree" in this ecosystem is a separate process, competence or tool that takes time to grow and develop. As in the case of a real forest, success depends not only on the quality of planting material (technology), but also on soil (corporate culture), climate (business environment) and care (systems approach).

Companies will no longer be able to rely solely on closed, out-of-the-box solutions. Unlike previous stages of technological development, the current transition to open data access, artificial intelligence, and Open Source is unlikely to be supported by large vendors because it directly threatens their established business models and core revenue streams.

As shown by the Harvard Business School study [40], which has already been discussed in the chapter on the fourth and fifth technological revolutions, the cost of creating the most used Open Source solutions from scratch for all companies would be about 4.15 billion dollars. However, if we imagine that each company would develop its own alternatives without access to existing Open Source tools, which is what has been happening for the last decades, the total business costs could reach a colossal 8.8 trillion dollars - this is the price of irrational demand that the software market can be valued at.

Technological progress will inevitably lead to a rethinking of established business models. Whereas companies used to be able to make money from complex, opaque processes and closed data, with the development of AI and analytics this approach is becoming less and less viable.

As a result of democratization of access to data and tools, the traditional software sales market may shrink significantly. However, at the same time, a new market will grow - the market for digital expertise, customization, integration and solution design. Here, value will not come from license sales, but from the ability to build flexible, open and adaptable digital processes. Just as electrification and the advent of tractors have spawned new industries, so too will the application of big data, AI and LLM open up entirely new horizons for business in the construction industry, which will require not only technological investments but also a profound transformation of mindsets, processes and organizational structures. And those companies and professionals who realize this and start acting today will be the leaders of tomorrow.

In a world where open data is becoming a major asset, the availability of information will be a game changer. Investors, clients and regulators will increasingly demand transparency, and machine learning algorithms will be able to automatically identify discrepancies in estimates, timelines and costs. This creates the conditions for a new stage of digital transformation, which gradually leads us to the "uberization" of the construction industry.

Open data and Uberization is a threat to the existing construction business

Construction is becoming an information management process. The more accurate, quality and complete the data, the more efficient the design, calculations, cost estimates, erection and operation of buildings. In the future, the key resource will not be a crane, concrete and rebar, but the ability to collect, analyze and use information.

In the future, construction company clients - investors and clients financing construction - will inevitably utilize the value of open data and analytics of historical data. This will open up opportunities to automate the calculation of project timelines and costs, without involving construction companies in costing issues, which will help control costs and identify redundant costs more quickly.

Imagine a construction site where laser scanners, quadrocopters and photogrammetry systems collect accurate real-time data on concrete volumes used. This information is automatically converted into simple flat MESH -models with metadata, bypassing cumbersome CAD (BIM) systems, without dependencies on complex geometric kernels, ERP or PMIS. These data, collected from the construction site, are centrally transferred to single structured repositories, available to the client for independent analysis, where real prices from different construction stores are uploaded and for example various parameters - from the rate of credit financing to dynamically changing factors such as weather conditions, stock exchange quotations of construction materials, logistics tariffs and statistical seasonal fluctuations in labor prices. Under such conditions, any discrepancies between the design and actual volumes of materials become instantly obvious, making it impossible to manipulate estimates both at

the design stage and when the project is delivered. As a result, the transparency of the construction process is achieved not through an army of supervisors and managers, but through objective digital data, in which the human factor and the possibility of speculation will be minimized.

In the future, this kind of data control work will rather be done by data managers on the customer side (Fig. 1.2-4 CQMS manager). This is especially true for calculations and project estimates: where there used to be a whole department of estimators, tomorrow there will already be machine learning and forecasting tools that will set price limits for construction companies to fit into.

Given the fragmented nature of the [construction] industry, where most of the systems and subsystems are supplied by SMEs, the digital strategy must come from the customer. Clients must create the conditions and mechanisms to unlock the digital capabilities of the supply chain [20].

— Andrew Davis and Giuliano Denicol, Accenture "Creating more value through capital projects"

Such openness and transparency of data poses a threat to construction companies, which are used to making money from opaque processes and confusing reports, where speculation and hidden costs can be hidden behind complex and closed formats and modular proprietary data platforms. Therefore, construction companies, as in the case of vendors promoting Open Source solutions, are unlikely to be interested in fully implementing open data into their business processes. If the data is available and easy to process for the customer, it can be checked automatically, which will eliminate the possibility of overestimating volumes and manipulating estimates.

According to the World Economic Forum report "Shaping the Future of Construction" (2016) [5], one of the key problems of the industry remains the passive role of the client. Nevertheless, it is the customers who should take greater responsibility for the outcome of projects - from early planning, to selecting sustainable interaction models, to monitoring performance. Without the active participation of project owners, systemic transformation of the construction industry is impossible.

The loss of control over volume and cost calculations has already transformed other industries over the past 20 years, allowing customers to directly, without intermediaries, to stymie their goals. Digitalization and data transparency have transformed many traditional business models, such as cab drivers with the advent of Uber (Fig. 10.1-4), hoteliers with the arrival of Airbnb and retailers and stores with the rise of Amazon, and banks with the rise of neo-banks and decentralized fintech ecosystems, where direct access to information and the automation of time and cost calculations have significantly reduced the role of intermediaries.

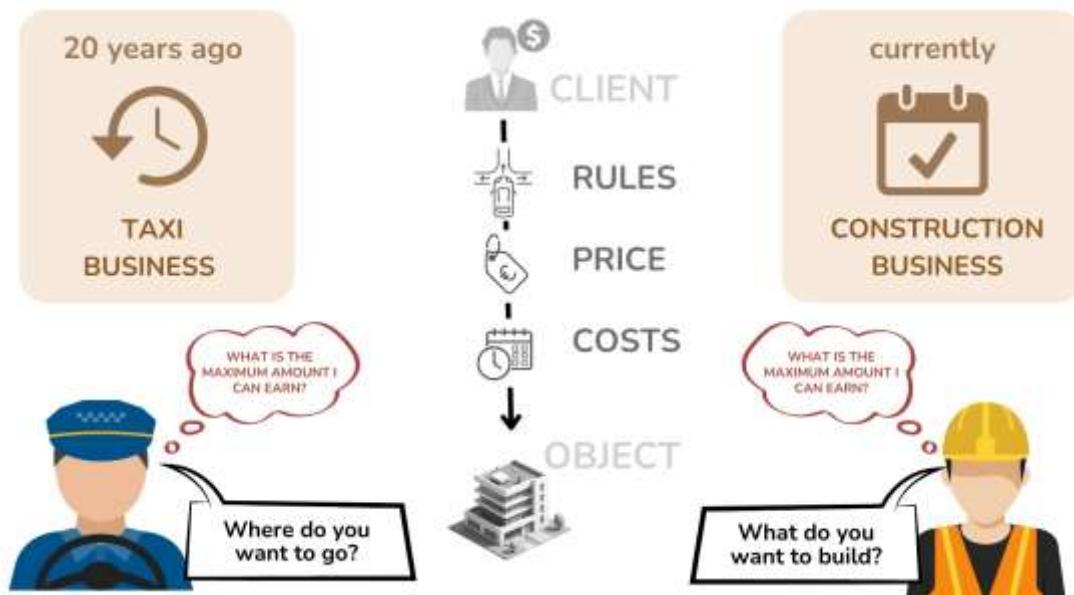


Fig. 10.1-4 The construction business will face the uberization that cab drivers, hoteliers and retailers had to face 10 years ago.

The process of democratization of access to data and tools for their processing is inevitable, and over time, open data on all project components will become a customer requirement and a new standard. Therefore, the introduction of open formats and transparent calculations will be promoted by investors, customers, banks and private equity funds (private equity) - those who are the end users of the built objects and then operate the object for decades.

Major investors, clients and banks are already demanding transparency in the construction industry. According to Accenture's study "Creating more value through capital projects" (2020) [20], transparent and reliable data is becoming a decisive factor for investment decisions in construction. As experts note, trustworthy and effective project management is impossible without transparency, especially in times of crises. In addition, asset owners and contractors are increasingly moving toward contracts that incentivize data sharing and collaborative analytics, reflecting growing demands from investors, banks and regulators for accountability and transparency.

The movement of the investor, the customer from idea to finished building, in the future will be akin to traveling on autopilot - without a driver in the form of a construction company, promises to become independent of speculation and uncertainty.

The era of open data and automation will inevitably change the construction business just as it has already done in banking, commerce, agriculture and logistics. In these industries, the role of intermediaries and traditional ways of doing business are giving way to automation and robotization, leaving no room for unjustified mark-ups and speculation.

The data and processes in all human economic activities are no different from what professionals in the construction industry have to deal with. In the long term, construction companies, which today dominate the market by setting price and service quality standards, may lose their role as a key intermediary between the customer and their construction project.

Unresolved problems of uberization as last chance to use time for transformation

But let's return to the realities of the construction industry. While self-driving cars, decentralized financial systems and artificial intelligence-based solutions are emerging in some sectors of the economy, a significant part of construction companies still remain paper-based organizations where key decisions are made rather on the basis of intuition and experience of individual specialists.

In this paradigm, a modern construction company can be compared to a 20-year-old taxi company that controls resources, routes and delivery times, and is responsible for the timing and cost of the "trip" - from the project idea (logistics and installation process) to project delivery. Just as GPS (in construction IoT, RFID) and machine learning algorithms in time/costing calculations once transformed transportation, data, algorithms and AI -agents are poised to transform construction management - from intuitive assessments to predictive, guided models. Over the past 20 years, many industries - finance, agriculture, retail, and logistics - have gradually eliminated the ability to speculate through the opacity of data. Prices, shipping costs or financial transactions are calculated automatically and statistically sound - in just a few seconds on digital platforms.

Looking to the future, construction companies must realize that democratizing access to data and the tools to analyze it will disrupt the traditional approach to estimating project costs and timelines and eliminate the ability to speculate on opaque volume and pricing data.

Like driving on a regulated road without driver intervention, the construction processes of the future will increasingly resemble an "Uberized" system - with automated time and cost estimation, transparent task routing and minimal dependence on human factors. This will change the very nature of the 'journey' from idea to realization - making it more predictable, manageable and data-driven.

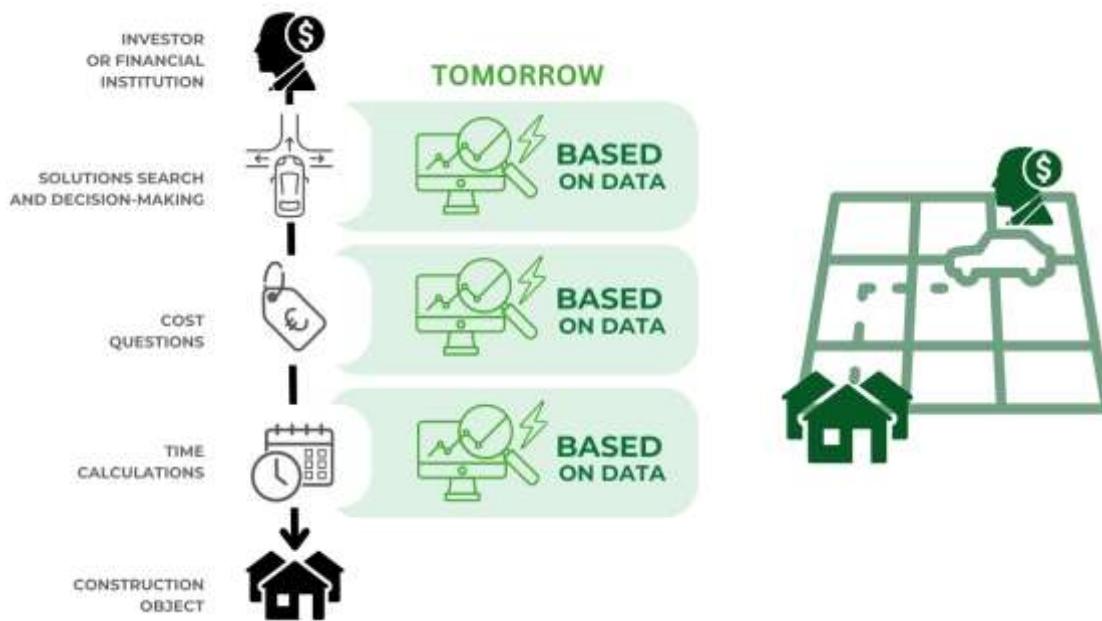


Fig. 10.1-5 Cost and "travel" time during construction will be determined using machine learning and statistical tools.

With the gradual introduction of new regulations and requirements in almost every country in the world, which oblige to transfer CAD- (BIM)-models to clients or banks financing construction projects, the client and customer have the opportunity to independently ensure transparency of cost and scope calculations. This is especially relevant for large customers and investors, who have sufficient competencies and tools for operational analysis of volumes and monitoring of market prices. For companies implementing large-scale standard projects - stores, office buildings, residential complexes - such practices are becoming standard.

As the information content of models becomes more complete and standardized, the possibility of manipulation and speculation virtually disappears. Digital transformation is gradually changing the rules of the game in the construction industry, and companies that do not adapt to these changes may face serious challenges.

Increased competition, technology disruption and shrinking margins have the potential to impact business sustainability. With liquidity constraints, more and more industry players are turning to automation, analytics and data technologies as a way to improve efficiency and process transparency. These tools are becoming an important resource to remain competitive in a changing economic environment.

It may not be worth waiting for external circumstances to force urgent action - it is much more effective to start preparing today by strengthening digital competencies, implementing modern solutions and building a data-centric culture.

One of the last key technological barriers to the large-scale digital transformation of the construction

industry, which will affect every company in the coming years, is the problem of automatic classification of elements of construction projects.

Without reliable, accurate and scalable classification, it is impossible to create a basis for full-fledged analytics, process automation and lifecycle management using AI and predictive models. As long as object classification still depends on manual interpretation by experienced specialists - foremen, designers, estimators - the construction industry still has a window of opportunity. This time can be used to prepare for the inevitable changes: increasing demands for transparency, democratization of tools and data, and the emergence of automatic classification systems that will radically change the rules of the game.

The task of automatic classification of elements of the construction world is comparable in its complexity to object recognition in unmanned driving systems, which is one of the main challenges. Let us imagine an unmanned car traveling from point A to point B (Fig. 10.1-5). Current automatic driving systems are bogged down by the problem of classifying objects that are recognized by lidars and cameras. It is not enough for a car to simply "see" an obstacle or landmark; it must know without error whether it is a pedestrian, a road sign, or a garbage can.

A similar fundamental challenge faces the entire construction industry. Project elements - such as windows, doors or columns - can be captured in documentation, represented in CAD models, photographed on the construction site or recognized in point clouds from laser scanning. However, just their visual or rough geometric recognition is not enough to build a truly automated project management system. It is necessary to ensure accurate and stable classification of each element by type, which will be unambiguously identifiable in all subsequent processes - from estimates and specifications to logistics, warehousing and most importantly - operation (Fig. 4.2-6).

It is at this stage - the transition from recognition to meaningful classification - that one of the key obstacles arises. Even if digital systems are technically capable of distinguishing and identifying objects in models and on the construction site, the main difficulty lies in the correct and contextually stable definition of the element type for different software environments. For example, a door may be designated by the designer in a CAD model as an element of the category "door", but when transferred to an ERP or PMIS system, it may be incorrectly typed - in case of an error on the part of the designer or due to inconsistencies between the systems. Moreover, the element often loses some important attributes or disappears from the system accounting during data exports and imports. This leads to a gap in the data flow and undermines the principle of end-to-end digitalization of construction processes. Thus, a critical gap between "visible" and "understandable" semantic meaning is formed, which undermines data integrity and significantly complicates the automation of processes throughout the entire life cycle of a construction project.

Solving the problem of universal classification of building elements using big data and machine learning technologies (Fig. 10.1-6) will be a catalyst for industry-wide transformation - and perhaps an unexpected discovery for many construction companies. A unified, trainable classification system will

be the foundation for scalable analytics, digital management and the adoption of AI into the daily practices of construction organizations.

NVIDIA and other technology leaders are already offering solutions in other industries that can automatically categorize and structure vast amounts of textual and visual information.

NVIDIA's NeMo Curator model [161], for example, specializes in automatically classifying and categorizing data into predefined categories, playing a key role in optimizing information processing pipelines for fine-tuning and pre-training generative AI models. The Cosmos platform is trained on real-world video and 3D -scenes [162], providing a foundation for autonomous systems and digital twins that are already being built in the NVIDIA ecosystem. NVIDIA Omniverse, which by 2025 has become the leading tool for working with the USD format, a universal scene description that could eventually replace the IFC format in design information delivery processes. Together with Isaac Sim, a robotic process simulator [163], solutions such as NeMo Curator, Cosmos, and Omniverse represent a new level of automation: from data cleaning and filtering to training set generation, object property modeling, and robot training on the construction site. And all of these tools are free and open source, significantly reducing barriers to adoption in engineering and construction practices.

Automatic classification of data at the level of structured tables is not as difficult a task as it may seem at first glance. As we have shown in the previous chapter (Fig. 9.1-10), it is possible to make up for missing or incorrect class values on the basis of similar parameters of other elements if there is accumulated historical data. If elements with similar characteristics have already been classified correctly in several completed projects, the system can suggest a suitable value for a new or incomplete element with a high probability (Fig. 10.1-6). Such logic, based on averaged values and context analysis, can be particularly effective when mass processing tabular data coming from estimates, specifications or CAD models.

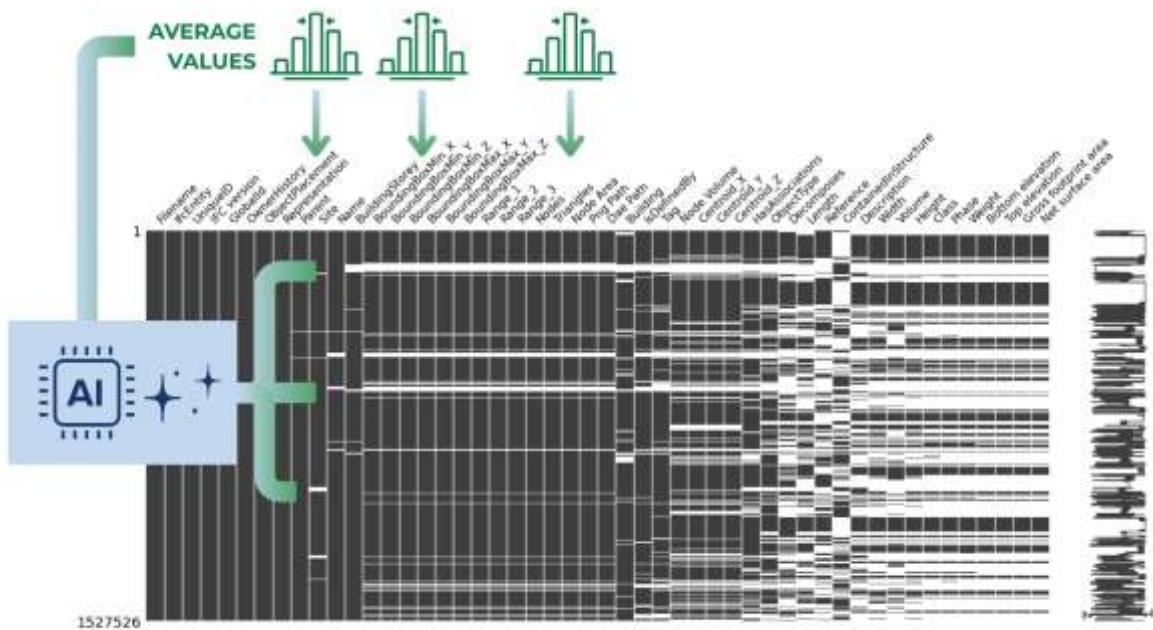
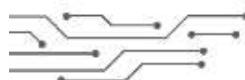


Fig. 10.1-6 Machine learning can help you automatically find average values for unfilled (white boxes) table parameters based on past projects.

Against the backdrop of such rapid progress in machine learning, it is clear: in 2025, it is naive to believe that the problem of automatic classification of building elements will remain unsolved for a long time. Yes, current algorithms have not yet reached full maturity, especially with incomplete or heterogeneous data, but the window of opportunity for adaptation is rapidly closing.

Companies that are already investing in collecting, cleaning and organizing their data, as well as adopting ETL automation tools, will be at an advantage. The rest risk falling behind - just as companies once failed to meet the challenges of digital transformation in the transportation and financial industries

Those who continue to rely on manual data management and traditional methods of cost and time estimation risk finding themselves in the position of the taxi fleets of the 2000s, unable to adapt to the era of mobile apps and automated route calculations by the early 2020s.



CHAPTER 10.2.

A PRACTICAL GUIDE TO IMPLEMENTING A DATA-DRIVEN APPROACH

From Theory to Practice: A Roadmap for Digital Transformation in Construction

The construction industry is gradually entering a new phase of development, where familiar processes are increasingly being supplemented - and sometimes even replaced - by digital platforms and transparent interaction models. This presents companies not only with challenges, but also with significant opportunities. Those organizations that are already building a long-term digital strategy today will not only be able to maintain their position in the market, but also expand it by offering customers modern approaches and reliable, technologically backed solutions.

It is important to realize that knowledge of concepts and technologies is only a starting point. Managers and specialists face a practical question: where to start implementation and how to turn theoretical ideas into real value. In addition, the question increasingly arises: what will the business be based on if the traditional methods of costing and timing can be revised by the customer at any time.

The answer probably lies not so much in technology, but in the formation of a new professional culture where working with data is perceived as an integral part of everyday practice. It is the lack of attention to digital technologies and innovations that has inoculated the construction industry into the serious backwardness that has been observed over the last decades [43].

According to McKinsey, R&D expenditures in the construction industry account for less than 1% of revenue, while in the automotive and aerospace industries this Fig. reaches 3.5-4.5%. Similarly, IT costs in construction remain at less than 1% of total revenue [107].

As a result, not only the level of automation, but also labor productivity, in construction is declining, and by 2020 a construction worker is producing less than half a century ago (Fig. 10.2-1)

Such productivity problems in the construction sector are common to most developed and developing countries (construction productivity has fallen in 16 out of 29 OECD countries (Fig. 2.2-1)), and point not only to a lack of technology, but also to the need for systemic changes in the very approaches to management, training and innovation.

The success of digital transformation depends not so much on the number and availability of tools, but rather on the ability of organizations to review their processes and develop a culture that is open to change. It is not the technology itself that is key, but the people and processes that ensure its effective use, support continuous learning and foster acceptance of new ideas.

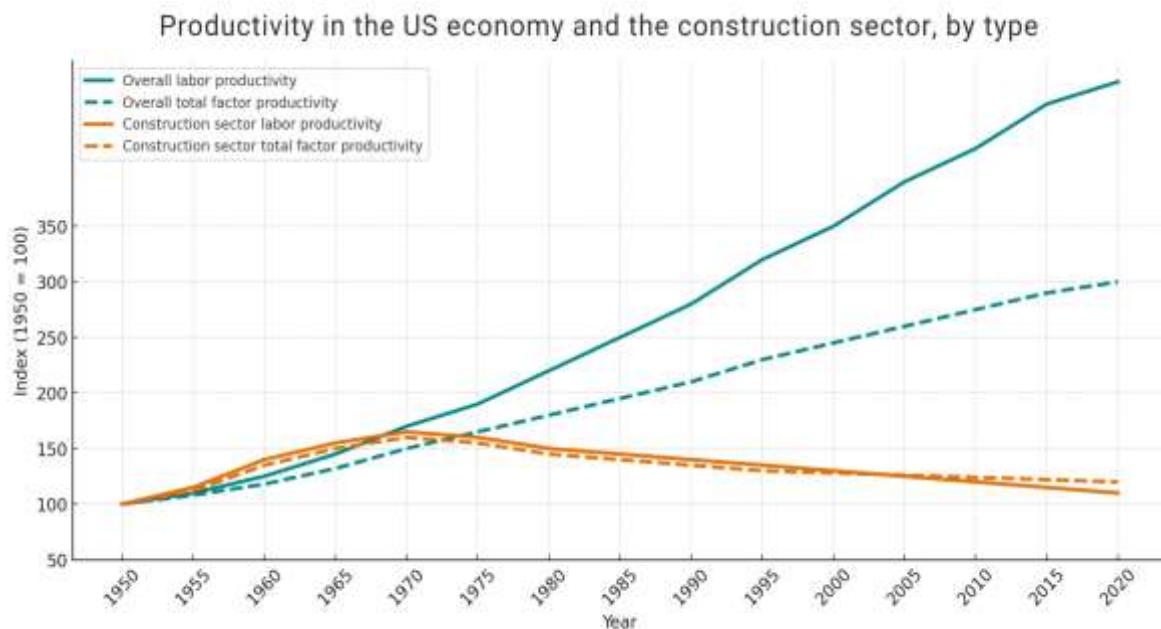


Fig. 10.2-1 The paradox of labor productivity and total resource productivity in the U.S. economy and construction sector (1950-2020) (based on [43]).

In the early parts of the book, the business environment model was compared to a forest ecosystem (Fig. 2.1-2, Fig. 1.2-4, Fig. 1.3-2). In a healthy forest, periodic fires, for all their destructive power, play a key role in long-term renewal. They clear the soil of old vegetation, return stored nutrients, and create space for new life. Some plant species have even evolved so that their seeds only open when exposed to high fire temperatures, a natural mechanism that provides the ideal time for germination.

Similarly in business, crises can play the role of "controlled burnout", encouraging the emergence of new approaches and companies that are not tied to outdated systems. Such periods force the abandonment of inefficient practices, freeing up resources for innovation. Just as a forest after a fire starts with pioneer plants, so business after a crisis forms new, flexible processes that become the basis for a mature information environment.

Companies that manage to correctly interpret these "signal fires" and transform their disruptive energy into constructive change will reach a new level of performance - with more transparent, adaptive data processes that enhance the organization's natural ability to renew and grow.

The growing influence of artificial intelligence and machine learning on the business environment is no longer in doubt. This is not just a temporary trend, but a strategic necessity. Companies that ignore AI, risk losing competitiveness in a market that increasingly encourages innovation and flexibility.

The future belongs to those who see AI not just as a tool, but as an opportunity to rethink every aspect of their business - from process optimization to management decision-making.

Laying the Digital Foundation: 1-5 Steps to Digital Maturity

In this chapter, we look at the digital transformation roadmap and identify the key steps required to implement a data-driven approach that can help transform both the corporate culture and the company's information ecosystem.

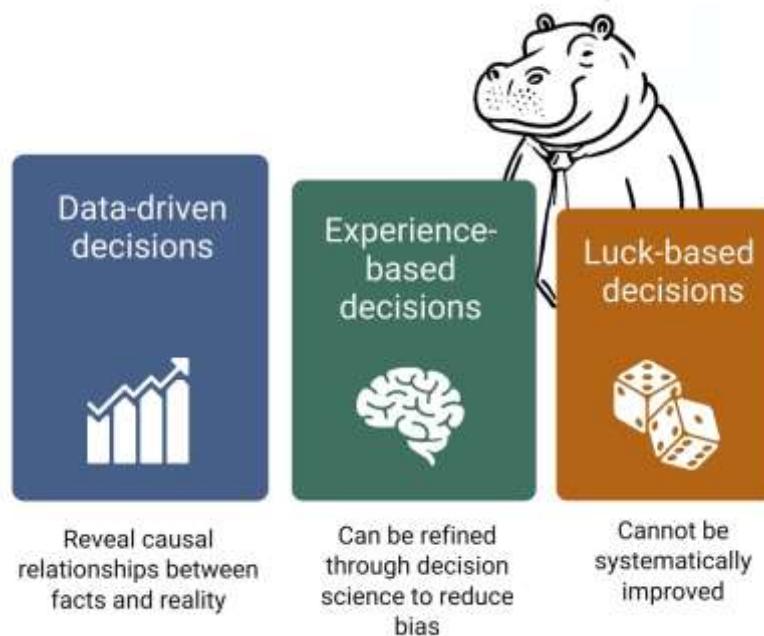


Fig. 10.2-2 Controlled updating and strategy selection: case, experience, or data.

According to McKinsey's study "Why Digital Strategies Fail" (2018), there are at least five reasons [164] why companies fail to achieve digital transformation goals

- **Blurred definitions:** executives and managers have different understandings of what "digital" means, leading to misunderstandings and inconsistencies.
- **Misunderstanding the digital economy:** many companies underestimate the magnitude of the changes that digitalization is bringing to business models and industry dynamics (Fig. 10.1-6).
- **Ignoring ecosystems:** companies focus on individual technology solutions (data silos), overlooking the need to integrate into broader digital ecosystems (Fig. 2.2-2, Fig. 4.1-12).
- **Underestimation of digitalization by competitors:** managers do not take into account that competitors are also actively adopting digital technologies, which may lead to a loss of competitive advantage.
- **Missing the duality of digitalization:** CEOs delegate responsibility for digital transformation to other executives, which bureaucratizes control and slows down the change process.

Addressing these challenges requires a clear understanding and alignment of digital strategies at all

levels of the organization. Before building a digital strategy, it is important to understand the starting point. Many organizations tend to adopt new tools and platforms without having a complete picture of the current state.

Step 1: Conduct an audit of your current systems and data.

Before changing processes, it's important to understand what's already in place. Conducting an audit allows you to identify weaknesses in data management and understand what resources can be used. An audit is a kind of "X-ray" of your business processes. It allows you to identify areas of risk and determine which data is critical to your project or business and which is secondary.

Key Actions:

- Map your IT environment (in Draw.io, Lucidchart, Miro, Visio or Canva). List the systems used (ERP, CAD, CAFM, CPM, SCM and others) involved in your processes and which we discussed in the chapter "Technology and Management Systems in Modern Construction" (Fig. 1.2-4)
- Assess data quality issues for each system for the frequency of duplicates, possible missing values, and format inconsistencies in each system.
- Identify "pain points" - places where processes can break down or often require manual intervention - imports, exports and additional validation processes.

If you want the team to trust the reports, you need to make sure the data is correct from the start.

A quality data audit will show you what data:

- Needs further development (automatic purification processes or additional transformation needs to be set up)
- They are "garbage" that only clogs up systems and can be gotten rid of by not using them in processes anymore.

It is possible to conduct such an audit on your own. But sometimes it is useful to engage an external consultant - especially from other industries: a fresh perspective and independence from construction "peculiarities" will help to soberly assess the status quo and avoid the typical pitfalls of bias towards certain solutions and technologies.

Step 2: Identify key standards for data harmonization.

After the audit, it is necessary to create common rules for working with data. As we discussed in the chapter "Standards: From Random Files to an Intelligent Data Model", this will help eliminate siloed information flows.

Without a single standard, each team will continue to work "their own way" and you will maintain a "zoo" of integrations where data is lost with every conversion.

Key Actions:

- Select the data standards to exchange information between systems:
 - For tabular data, this can be structured formats like CSV, XLSX or more efficient formats like Parquet
 - For exchange of loosely structured data and documents: JSON or XML
- Master working with data models:
 - Start by parameterizing the tasks at the level of the conceptual data model - as described in the chapter "Data modeling: conceptual, logical and physical model" (Fig. 4.3-2).
 - As you delve deeper into the business process logic, move to formalizing requirements using parameters in the logical and physical models (Fig. 4.3-6)
 - Identify key entities, their attributes and relationships within processes, and visualize these relationships - both between entities and between parameters (Fig. 4.3-7).
- Use regular expressions (RegEx) to validate and standardize data (Fig. 4.4-7), as we discussed in the chapter "Structured Requirements and RegEx Regular Expressions". RegEx is not a complex but an extremely important topic in the work of creating requirements at the physical data model level.

Without data-level standards and process visualization, it is impossible to provide a consistent and scalable digital environment. Remember, "bad data is expensive." And the cost of error increases as a project or organization becomes more complex. Unifying formats, defining naming, structure and validation rules is an investment in the stability and scalability of future solutions.

Step 3. Implement DataOps and automate processes.

Without a well-defined architecture, companies will inevitably be faced with disparate data contained in siloed information systems. Data will be unintegrated, duplicated in multiple locations, and costly to maintain.

Imagine that data is water, and the data architecture is the complex system of pipelines that transports that water from its storage source to its point of use. It is the data architecture that determines how information is collected, stored, transformed, analyzed, and delivered to end users or applications.

DataOps (Data Operations) is a methodology that integrates the collection, cleansing, validation, and

utilization of data into a single automated process flow, as we discussed in detail in Part 8 of the book.

Main Actions:

- Create and customize ETL -conveyors to automate processes:
 - Extract: organize automatic data collection from PDF documents (Fig. 4.1-2, Fig. 4.1-5, Fig. 4.1-7), Excel spreadsheets, CAD -models (Fig. 7.2-4), ERP -systems and other sources you work with
 - Transform: set up automatic processes to transform data to a single structured format and automate calculations that will take place outside of closed applications (Fig. 7.2-8)
 - Load: try creating automatic data uploads to summary tables, documents, or centralized repositories (Fig. 7.2-9, Fig. 7.2-13, Fig. 7.2-16)
- Automate the calculation and QTO (Quantity Take-Off) processes as we discussed in the chapter "QTO Quantity Take-Off: Grouping Project Data by Attributes":
 - ConFig. automatic extraction of volumes from CAD -models, using APIs, plug-ins or reverse engineering tools (Fig. 5.2-5).
 - Create rules for grouping elements for different classes by attributes in the form of tables (Fig. 5.2-12)
 - Try to automate frequently repeated volume and cost calculations outside of modular closed systems (Fig. 5.2-15)
- Start using Python and Pandas to process data, as we discussed in the chapter "Python Pandas: an indispensable tool for working with data":
 - Apply DataFrame to work with XLSX files and automate the processing of tabular data (Fig. 3.4-6)
 - Automate information aggregation and transformation through various Python libraries
 - Use the LLM to simplify writing ready-made code blocks and entire Pipelines (Fig. 7.2-18)
 - Try building a Pipeline in Python, which finds errors or sees anomalies and sends a notification to the responsible person (e.g., the project manager) (Fig. 7.4-2)

Automation based on DataOps principles allows you to move from manual and fragmented data handling to sustainable and repeatable processes. This not only reduces the burden on employees who deal with the same transformations every day, but also dramatically increases the reliability, scalability and transparency of the entire information system.

Step 4: Create an open data governance ecosystem.

Despite the development of closed modular systems and their integration with new tools, companies face a serious problem - the growing complexity of such systems outpaces their usefulness. The initial idea of creating a single proprietary platform covering all business processes has led to excessive centralization, where any changes require significant resources and time to adapt.

As we discussed in the chapter "Corporate Mycelium: How Data Connects Business Processes," effective data management requires an open and unified ecosystem that connects all sources of information.

Key elements of the ecosystem:

- Select an appropriate data store:
 - For tables and calculations use databases - for example, PostgreSQL or MySQL (Fig. 3.1-7).
 - For documents and reports, cloud storage (Google Drive, OneDrive) or systems that support JSON format may be suitable
 - Familiarize yourself with the capabilities of Data Warehouse, Data Lakes, and other tools for centralized storage and analysis of large amounts of information (Fig. 8.1-8)
- Implement solutions to access proprietary data:
 - If you use proprietary systems, set up API or SDK access to them to retrieve data for external processing (Fig. 4.1-2)
 - Familiarize yourself with the potential of reverse engineering tools for CAD formats (Fig. 4.1-13)
 - Set up ETL-Pipelines that periodically collect data from applications or servers, convert it into open structured formats, and save it to repositories (Fig. 7.2-3)
 - Discuss within the team how to provide access to data without the need for proprietary software
 - Remember: data is more important than interfaces. It is the structure and availability of information, not specific user interface tools, that provide long-term value
- Think about creating a Center of Excellence (CoE) for data, as we discussed in the "Center of Excellence (CoE) for Data Modeling" chapter, or how you can provide data expertise in other ways (Fig. 4.3-9)

The data management ecosystem creates a unified information space in which all project participants work with consistent, up-to-date and verified information. It is the basis for scalable, flexible and reliable digital processes

Unlocking the potential of data: 5-10 steps to digital maturity

In addition to technical integration, an important factor in the successful implementation of digital

solutions is their adoption by end users. Engaging customers or users in performance measurement is both a challenge of improving user experience and managing change in the company. If a solution doesn't fit into a familiar workflow or doesn't solve real user or customer problems, it won't be used, and no amount of additional measures and incentives will fix that.

Transformation is an iterative process based on analyzing user interaction data with new processes, with frequent testing cycles, constant feedback and refinements.

Step 5: Build a data culture, train staff and collect feedback

Even the most advanced system won't work without employee engagement. You need to create an environment where data is used on a daily basis and the team understands its value.

The published UK government report "Data Analytics and AI in Government Project Delivery" 2024 notes [83] that the training of professionals with the necessary competencies in data processing and interpretation is critical to the successful implementation of data analytics and AI.

Lack of data analytics expertise is one of the key issues limiting digital transformation. Leaders are used to established routines: quarterly cycles, prioritized initiatives, and traditional ways of moving projects forward. Change requires a distinctive leader - high enough in rank to have influence, but not so high that he or she has the time and motivation to lead a long-term transformation project.

Main Actions:

- Recognizing the need to move from subjective decisions based on the opinion of a highly paid employee (HiPPO) to a decision-making culture based on facts and data, as discussed in the chapter "HiPPO or the Danger of Opinion in Decision Making" (Fig. 2.1-9).
- Organize systematic training:
 - Hold trainings on how to use structured data, and invite experts from other industries who don't have a bias toward products and concepts popular in the construction industry today
 - Discuss data analysis approaches and tools with colleagues, and independently learn hands-on work with tools such as Python, pandas, and LLM (Fig. 4.1-3, Fig. 4.1-6)
 - Create a library of tutorials (preferably with short videos) on the topic of structuring data (Fig. 3.2-15) and creating data models (Fig. 4.3-6, Fig. 4.3-7)
- Utilize modern learning technologies:
 - Use language models (LLMs) to support code and data manipulation, including code generation, refactoring, and analysis, as well as processing and interpreting tabular information (Fig. 3.4-1)
 - Explore how LLM-generated code can be customized and integrated into a complete Pipeline solution when working in an offline development environment (IDE) (Fig. 4.4-

14, Fig. 5.2-13)

When a manager continues to make decisions the "old fashioned way," no amount of training will convince people to take analytics seriously.

Building a data culture is impossible without constant feedback. Feedback allows you to identify gaps in processes, tools and strategies that cannot be discovered through internal reports or formal KPIs metrics. Complimentary comments from users of your solutions will not provide practical value. It is critical feedback that is valuable, especially if it is based on specific observations and facts. But obtaining such information requires effort: you need to build processes where participants - both internal and external - can share comments (it may make sense to do so anonymously) without distortion and without fear that their opinions may affect their own work. It is important that they do so without distortion and without fear of negative consequences for themselves.

Any learning is ultimately self-learning [165].

— Milton Friedman, American economist and statistician

Implementation of analytical tools should be accompanied by regular verification of their effectiveness in practice (ROI, KPIs), which can only be achieved through structured feedback from employees, customers and partners. This allows companies not only to avoid repeating mistakes, but also to adapt faster to changes in the environment. Having a mechanism for collecting and analyzing feedback is one of the signs of maturity of an organization moving from episodic digital initiatives to a sustainable model of continuous improvement (Fig. 2.2-5).

Step 6: From pilot projects to scaling up

Pick battles big enough to matter and small enough to win.

— Jonathan Kozol

Launching digital transformation "all at once and everywhere" is extremely risky. A more effective approach is to start with pilot projects and gradually scale up successful practices.

Main Actions:

- Choose the right project for the pilot:

- Define a specific business objective or process with measurable results (KPI, ROI) (Fig. 7.1-5)
- Select an ETL automation process, such as automatic data validation or workload calculation (QTO) using Python and Pandas (Fig. 5.2-10)
- Establish clear metrics for success (e.g., reduce the time to write inspection specifications or data validation reports from a week to a day)
- Take iterative approaches:
 - Start with simple data conversion processes and create streaming conversions of multi-format data into the formats you need for your processes (Fig. 4.1-2, Fig. 4.1-5)
 - Gradually increase the complexity of tasks and expand the automation of processes by forming a complete Pipeline in the IDE based on documented code blocks (Fig. 4.1-7, Fig. 7.2-18).
 - Document and record (preferably with short videos) successful solutions and share them with colleagues or in professional communities
- Develop templates and accompanying documentation to replicate such solutions so that they can be used effectively by your colleagues (or members of the professional community, including social media users)

Phased "rolling up" allows you to maintain the high quality of changes and not fall into the chaos of parallel implementations. The "from small to big" strategy minimizes risks and allows you to learn from small mistakes without letting them grow into critical problems.

Moving from a project-based approach, where employees are only partially involved, to the formation of permanent teams (e.g., centers of expertise - CoEs) allows for sustainable product development even after the first version of the product is released. Such teams not only support existing solutions, but also continue to improve them.

This reduces dependence on lengthy approvals: team members are empowered to make decisions within their area of responsibility. As a result, managers are freed from the need to micromanage, and teams can focus on creating real value.

Developing new solutions is not a sprint, but a marathon. Those who succeed in it are those who are initially focused on long-term, consistent work.

It is important to realize that technology requires constant development. Investing in the long-term development of technological solutions is the basis for successful work.

Step 7: Use open data formats and solutions

As we discussed in the chapters on modular platforms (ERP, PMIS, CAFM, CDE, etc.), it is important to

focus on open and universal data formats that ensure independence from vendor solutions and increase the availability of information for all stakeholders.

Main Actions:

- Move from closed formats to open formats:
 - Use open formats instead of proprietary formats, or find a way to set up automatic upload or conversion of closed formats to open formats (Fig. 3.2-15).
 - Implement tools to work with Parquet, CSV, JSON, XLSX, which are the exchange standards between most modern systems (Fig. 8.1-2)
 - If working with 3D geometry plays an important role in your processes, consider using open formats such as USD, glTF, DAE, or OBJ (Fig. 3.1-14)
- Use vector databases databases for efficient analysis and information retrieval:
 - Use Bounding Box and other methods to simplify the 3D -geometry (Fig. 8.2-1)
 - Think about where you can implement data vectorization - converting texts, objects, or documents into numerical representations (Fig. 8.2-2)
- Apply big data analytics tools:
 - Organize the storage of accumulated historical data (e.g. PDF, XLSX, CAD) in formats suitable for analysis (Apache Parquet, CSV, ORC) (Fig. 8.1-2).
 - Begin to apply basic statistical methods and work with representative samples - or at a minimum, familiarize yourself with the fundamental principles of statistics (Fig. 9.2-5)
 - Implement and learn tools to visualize data and the relationships between data to visualize the results of analysis. Without good visualization, it is impossible to fully understand the data itself or the processes based on it (Fig. 7.1-4).

The move to open data formats and the introduction of tools for analyzing, storing and visualizing information lays the foundation for sustainable and independent digital governance. This not only reduces dependence on vendors, but also ensures equal access to data for all stakeholders.

Step 8. Start implementing machine learning for prediction

Many companies have accumulated vast amounts of data - a kind of "information geysers" that are still unused. This data has been collected in hundreds or thousands of projects, but has often been used only once or not at all in further processes. Documents and models stored in closed formats and systems are often perceived as obsolete and useless ballast. In reality, however, they are a valuable resource - the basis for analyzing mistakes made, automating routine operations, and developing innovative solutions for auto-classification and feature recognition in future projects.

The key challenge is to learn how to extract this data and transform it into actionable insights. As

discussed in the chapter on Machine Learning and Predictions, machine learning techniques have the potential to significantly improve the accuracy of estimates and predictions in a variety of construction-related processes. Fully utilizing the accumulated data opens the way to improving efficiency, reducing risk, and building sustainable digital processes.

Main Actions:

- Start with simple algorithms:
 - Try applying linear regression - using hints from the LLM - to predict recurring performance in data sets where dependencies on a large number of factors are absent or minimal (Fig. 9.3-4)
 - Consider at which stages of your processes the k-nearest neighbor (k-NN) algorithm could theoretically be applied - for example, for classification tasks, object similarity assessment, or forecasting based on historical analogs (Fig. 9.3-5).
- Collect and structure data to train models:
 - Collect historical project data in one place and in a single format (Fig. 9.1-10)
 - Work on the quality and representativeness of training samples, through automated ETL (Fig. 9.2-8)
 - Learn to separate the data into training and test sets, as we did in the Titanic dataset example (Fig. 9.2-6, Fig. 9.2-7)
- Consider expanding the application of machine learning techniques to solve a wide range of problems, from predicting project timelines to optimizing logistics, resource management, and early identification of potential problems

Machine learning is a tool to turn archived data into a valuable asset for prediction, optimization, and informed decision making. Start with small datasets (Fig. 9.2-5) and simple models, gradually increasing in complexity.

Step 9. Integrate IoT and modern data collection technologies

The construction world is rapidly becoming digital: every construction photo, every Teams post is already part of a larger process of parameterizing and tokenizing reality. Just as GPS once transformed logistics, IoT, RFID and automated data collection are changing the construction industry. As discussed in the chapter "IoT Internet of Things and Smart Contracts," a digital construction site with sensors and automated monitoring is the future of the industry.

Main Actions:

- Implement IoT -devices, RFID -tags and detail the processes associated with them:
 - Evaluate which areas or project phases of a project can benefit from sensor installation with the greatest return on investment (ROI) - for example, for monitoring temperature, vibration, humidity or movement

- Consider using RFID to track materials, tools and equipment throughout the supply chain
- Consider how the collected data can be integrated into a unified information system, such as Apache NiFi, for automated processing and real-time analysis (Fig. 7.4-5)
- Establish a real-time monitoring system:
 - Develop dashboards to track key process or project metrics using visualization tools such as Streamlit, Flask or Power BI)
 - Set up automatic notifications to signal critical deviations from the plan or norms (Fig. 7.4-2)
 - Assess the potential for predictive maintenance of equipment based on the data collected and patterns identified (Fig. 9.3-6)
- Combine data from different sources:
 - Start by visualizing the data model at the physical level - reflect the structure of information flows and key parameters coming from CAD systems, IoT devices and ERP platforms (Fig. 4.3-1)
 - Start by creating a rough outline of a unified platform for data analysis and management decision support. Capture key functions, data sources, users, and envisioned application scenarios (Fig. 4.3-7).

The sooner you start connecting real-world processes to the digital world, the faster you can manage them with data - efficiently, transparently and in real time.

Step 10. Prepare for the future of changes in the industry

Construction companies are constantly under pressure from the external environment: economic crises, technological leaps, regulatory changes. Just like a forest that has to withstand rain, snow, drought and scorching sun, companies live in an environment of continuous adaptation. And just as trees become resilient to frost and drought through deep root systems, only those organizations that have a solid foundation of automated processes, the ability to anticipate change and flexibly adapt strategies remain viable and competitive.

As mentioned in the chapter "Survival Strategies: Building Competitive Advantage", the construction industry is entering a phase of radical transformation. The interaction between client and contractor is moving towards a model of persuasion, where transparency, predictability and digital tools are replacing traditional approaches. In this new reality, it is not the largest, but the most flexible and technologically mature that win.

Main Actions:

- Analyze business vulnerabilities in the context of open data:

- Assess how democratizing access to data as part of Uberization could have a devastating impact on your competitive advantage and your business (Fig. 10.1-5)
- Think about a strategy to move from opaque and siloed processes to business models based on open solutions, system interoperability and data transparency (Fig. 2.2-5).
- Develop a long-term digital strategy:
 - Determine whether you aspire to be an innovation leader or prefer a "catch-up" scenario in which you will conserve your resources
 - Describe the stages: short-term (process automation, data structuring), medium-term (implementation of LLM and ETL), long-term (digital ecosystems, centralized repositories)
- Think about expanding your service portfolio:
 - Consider offering new services (focused on energy efficiency, ESG, data services). We will talk more about new business models in the next chapter
 - Strive to position yourself as a reliable technology partner that accompanies the entire life cycle of a facility, from design to operation. Your credibility should be based on your systematic approach, transparent processes and ability to provide sustainable technology solutions

In a transformational environment, it is not those who simply react to change who win, but those who are proactive. Flexibility, openness and digital maturity are the foundations of sustainability in building tomorrow.

Transformation roadmap: from chaos to data-driven company

The following plan can serve as an initial benchmark - a starting point for shaping your own data-driven digital transformation strategy:

- **Audit and standards:** analyze current state, unify data
- **Data structuring and classification:** automate the transformation of unstructured and loosely structured data
- **Automate groupings, calculations and calculations:** use open source tools and libraries for automation
- **Ecosystem and COE:** build an internal team to form a unified data ecosystem in the company
- **Culture and learning:** move away from HiPPO -solutions to data-driven solutions
- **Pilots, feedback and scaling:** act iteratively: test new methods on a limited scale, gather valid feedback and gradually scale solutions
- **Open formats:** use universal and open formats for independence from software vendors

- **Machine learning:** embed ML algorithms into processes for prediction and optimization
- **IoT and the digital construction site:** integrate modern data collection technologies into processes
- **Strategic adaptation:** prepare for future industry changes

Most importantly, remember that "data alone doesn't change a company: it's the people who know how to work with that data that change it". Emphasize culture, transparent processes, and a commitment to continuous improvement

A systems approach enables a shift from siloed digital initiatives to a full-fledged data-driven management model where decisions are not based on intuition or assumptions, but on data, facts and mathematically calculated probabilities. The digital transformation of the construction industry is not just about adopting technology, but shaping a business ecosystem where project information is transferred seamlessly and iteratively between different systems. In doing so, machine learning algorithms provide automatic, continuous analysis, forecasting and optimization of processes. In such an environment, speculation and hidden data become irrelevant - only proven models, transparent calculations and predictable results remain.

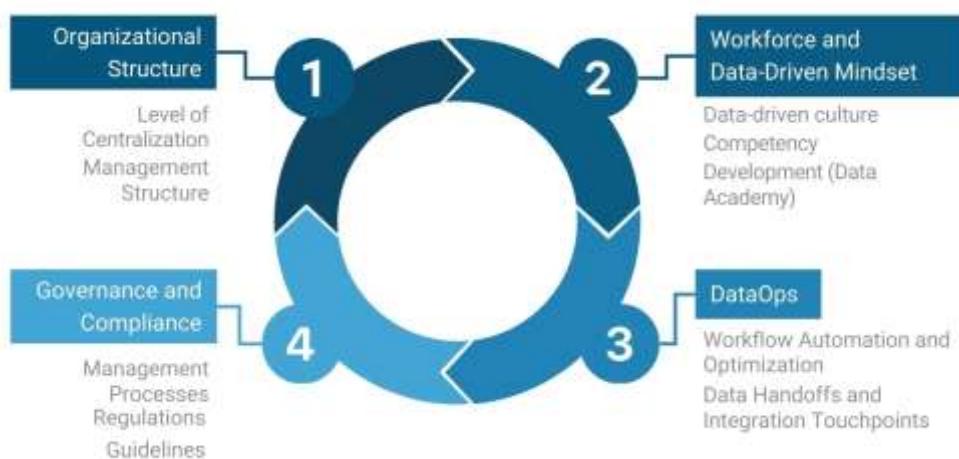


Fig. 10.2-3 Key elements of successful company-level data management.

Each part of the book corresponds to a specific stage of data processing and analysis in construction projects (Fig. 2.2-5). If you want to return to one of the topics discussed earlier and look at it from a holistic understanding of the data utilization flow, you can refer to the part titles in Fig. 10.2-4.

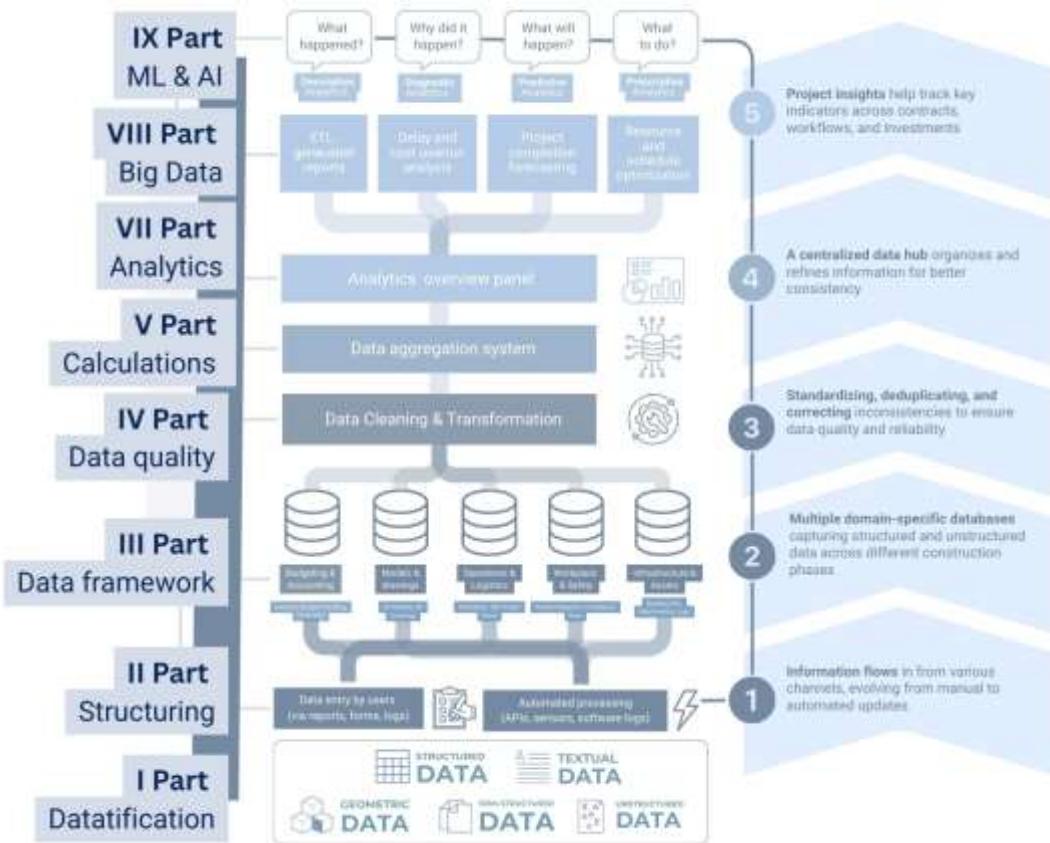


Fig. 10.2-4 Parts of the book in the context of the data processing pipeline (Fig. 2.2-5): from the digitalization of information to analytics and artificial intelligence.

Regardless of your organization's size, level of technology maturity, or budget, you can start moving toward a data-driven approach today. Even small steps in the right direction will produce results over time.

Data-driven transformation is not a one-time project, but a continuous, iterative improvement process that includes implementing new tools, redesigning processes, and developing a culture of data-driven decision-making.

Building in Industry 5.0: How to make money when you can't hide anymore

For a long time, construction companies have been making money on non-transparent processes. The main business model was speculation - overestimating the cost of materials, scope of work and percentage mark-ups in closed ERP - and PMIS - systems that are inaccessible to external audits. The limited access of customers and their authorized persons to the original project data created the ground for schemes in which it became almost impossible to verify the reliability of calculations.

However, this model is rapidly losing relevance. With the democratization of data access, the emergence of LLMs, the arrival of open data, and ETL automation tools, the industry is moving to a new standard of operation.

As a result, opacity is no longer a competitive advantage - soon it will become a burden that will be hard to part with. Transparency turns from an option into a prerequisite for staying in the market.

Who will clients - banks, investors, physical customers, private equity, government customers - work with in the new digital reality? The answer is obvious: with those who are able to provide not only the result, but also the justification of each step on the way to it. As the volume of open data grows, partners and customers will choose companies that guarantee transparency, accuracy and predictability of results.

Against this backdrop, new business models are emerging that are based on data management and trust rather than speculation:

■ **Selling processes instead of square meters:** the key asset becomes trust and efficiency instead of discounted concrete agreements. The main value will be predictability of the result based on reliable and verified data. Modern companies will not sell the construction object as such, but:

- accurate deadlines and transparent work schedules;
- reasonable estimates, supported by calculations;
- full digital traceability and control at all stages of the project.

■ **Engineering and analytics as a service:** the "Data-as-a-Service" model (a way to deliver ready-made data to users via the Internet, as a service), where each project becomes part of a digital data chain, and business value is in the ability to manage this chain. Companies are transforming into intelligent platforms offering solutions based on automation and analytics:

- Automated and transparent preparation of estimates and plans;
- Risk and timing assessment based on machine learning algorithms;
- calculation of environmental indicators (ESG, CO₂, energy efficiency);
- generation of reports from audited open sources.

■ **Productization of engineering experience:** the company's developments can be used repeatedly within the company and distributed as a separate product - forming an additional source of income through digital services. In the new environment, companies create not only projects, but also digital assets:

- libraries of components and estimate templates;
- automated verification modules;
- Open-source plugins and scripts (selling consulting) for working with data.

■ **A new type of company: the Data-Driven Integrator:** a market player that does not depend on specific software vendors or modular systems and is not "locked" into a single software interface. It operates freely with data - and builds its competitiveness on it. The construction company of the future is not just a contractor, but an information integrator capable of performing the following functions for the customer:

- Combine data from disparate sources and perform analytics;
- Ensure transparency and credibility of processes;
- advise on optimization of business processes;
- develop tools that work in the open data ecosystem, LLM, ETL and Pipelines.

Industry 5.0 (Fig. 2.1-12) marks the end of the "era of manual averaging ratios" and evening meetings between CEOs and the estimating and accounting department. Everything that was previously hidden - calculations, estimates, volumes - becomes open, verifiable and understandable even to non-experts. Those who are the first to reorient themselves will be the winners. All others will be left out of the new digital economy of the construction sector.

CONCLUSION

The construction industry is entering an era of fundamental change. From the first records on clay tablets to the massive amounts of digital data flowing from project servers and construction sites, the history of information management in construction has always reflected the maturity of the technology of its time. Today, with the advent of automation, open formats and intelligent analytics, the industry is facing not a gradual evolution, but a rapid digital transformation.

As in other sectors of the economy, construction will have to rethink not only the tools but also the principles of work. Companies that used to dictate the market and serve as the main intermediary between the client and the project are losing their unique position. Trust and the ability to work with data - from collecting and structuring it to analyzing, forecasting and automating decisions - are coming to the fore.

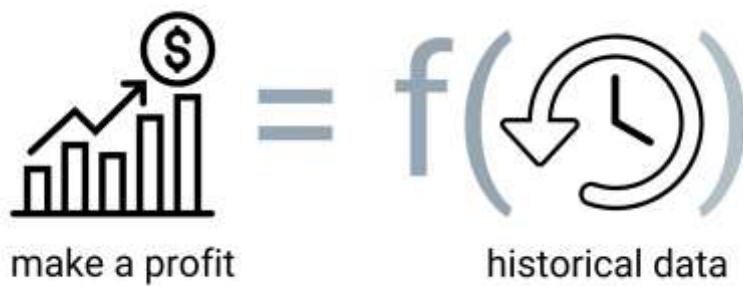


Fig. 10.2-1 Structured historical data is the fuel for an efficient and manageable business.

This book detailed the key principles of data management in the construction industry, from auditing and standardization to process automation, the use of visualization tools, and the implementation of intelligent algorithms. We looked at how, even with limited resources, you can build a working data architecture and start making decisions based on verifiable facts rather than intuition. Working with data is no longer just the task of the IT department - it becomes the foundation of the management culture, which determines the company's flexibility, adaptability and long-term sustainability.

The application of machine learning technologies, automatic processing systems, digital twins and open formats already today makes it possible to eliminate the human factor where it used to be critical. Construction is moving towards autonomy and controllability, where the movement from idea to project realization can be compared to navigation in autopilot mode: without dependence on subjective decisions, without the need for manual intervention at every stage, but with full digital traceability and control (Fig. 10.2-2).

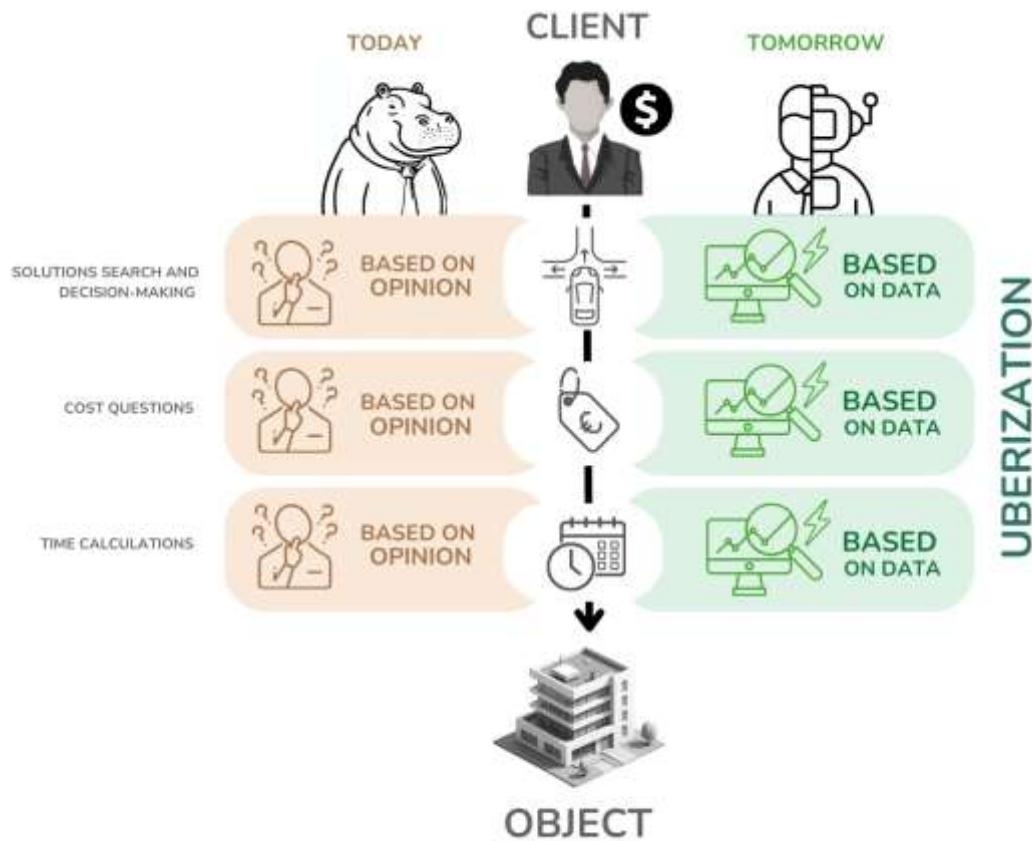


Fig. 10.2-2 The shift from making decisions based on the opinions of important experts (HIPPO) to analyzing data will be primarily promoted by the customer.

By learning the methods, principles, and tools presented in this book, you will be able to start making data-driven, rather than intuitive, decisions in your company. You will also be able to run module chains in LLMs, copy out-of-the-box ETL Pipelines into your development environment (IDE), and automatically process data to get the information you need in the form you want. Later, building on the book's chapters on big data and machine learning, you will be able to implement more complex scenarios - extracting new insights from historical data and applying machine learning algorithms to predict and optimize your processes.

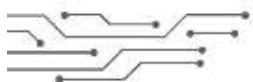
Open data and processes will provide the basis for more accurate estimates of project costs and timelines, preventing construction companies from speculating on opaque data. This is both a challenge and an opportunity for the industry to rethink its role and adapt to a new environment where transparency and efficiency will become key success factors.

The willingness to take knowledge and put it into practice is key to success in the age of digital transformation.

Companies that realize this first will have an advantage in the new digital competition. But it is important to realize that data alone will not change anything. Many people will need to change the way

they think, and that requires an incentive. Your company must rethink the way it shares data. The people who change the company are the people who know how to work with this data, interpret it, use it for optimization, and create a new process architecture based on it.

If you are reading these lines, you are ready for a change and you are already one step ahead. Thank you for choosing this path. Welcome to the era of digital transformation!



ABOUT THE AUTHOR

My name is Artem Boiko. My journey on the construction site started in 2007 - with a job as a miner at an oil shale mine, in my hometown, while studying at the Mining University, specializing in mine and underground construction. On the back cover of this book you can see an explosionist in the face where we mined and blasted hundreds of cubes of oil shale. My career has taken me in many different directions, from working as a mine and subway construction worker to industrial climber, roofing and elevator installer. I have had the honor of participating in projects of various scales: from construction of private houses to large industrial facilities in different regions of the world.



Over time, my work has shifted from physical construction to information management and digital processes. Since 2013, I have worked in various positions in small, medium and large construction companies in several regions of Germany, from planner to data management manager. As far as data management is concerned, my experience consists of working with data in various ERP systems, CAD (BIM), MEP, FEM, CMS. I have been involved in optimization, process automation as well as analysis, machine learning, data processing in the planning, calculation and execution phases of construction works in industrial, residential, infrastructure and utility construction companies.

I have been working with open source software and open data since 2003. During this time I have realized many web projects - from websites and online stores to full-fledged web applications using open source solutions and open CMS. These platforms, similar in many ways to modern building ERPs, have modular architecture, high adaptability and accessibility. This experience laid the foundation for my professional approach - a focus on open source technologies and a culture of collaborative development. Respect for open source and the free exchange of knowledge is something I strive to promote in the construction industry. My work to improve data accessibility in the construction industry has translated into the creation of several social media communities to discuss data openness and the use of Open Source in construction, as well as the launch of several startups developing solutions to provide access to data from various closed systems and platforms.

My contribution to the professional community is through participation as a speaker at conferences covering CAD interoperability (BIM), ERP, 4D-5D, LLM Machine Learning and Artificial Intelligence, as well as articles published in European construction industry publications. One of my notable achievements is the creation of the "History of BIM" [111], a comprehensive map of important software solutions for data management in the construction industry. My 7-part article series "BIM Development and Lobbying Games", translated into several languages, has been widely recognized as an attempt to illuminate the hidden dynamics of digital standards.

This is how I went from mining rock - to mining and systematizing construction data. I am always open to professional dialog, new ideas and joint projects. I will gratefully accept any feedback and look forward to your messages or seeing you among my social media followers. Thank you so much for reading this book to the end! I would be happy if this book helps you better understand the topic of data in the construction industry.

FEEDBACK

Readers' opinions play an important role in the further development of publications and the selection of priority topics. Comments on which ideas have proved useful and which have raised doubts and require further clarification or citation of sources are particularly valuable. The book includes a wide range of contributions and analytical assessments, some of which may seem controversial or subjective. If in the course of reading you find inaccuracies, incorrectly cited sources, logical inconsistencies or typos, I would appreciate your comments, thoughts or criticisms, which you can send to: boikoartem@gmail.com. Or through messages on LinkedIn: linkedin.com/in/boikoartem

I would be extremely grateful for any mentions of the book Data-Driven Construction on social media - sharing your reading experience helps spread the word about open data and tools and supports my work.

TRANSLATION COMMENTARY

This book was translated using artificial intelligence technology. This has made the translation process much faster. However, as with any technological operation, errors or inaccuracies may occur. If you notice anything that appears to be incorrect or incorrectly translated, please email me. Your comments will help improve the quality of the translation.

DATADRIVENCONSTRUCTION COMMUNITIES

It's a place where you can freely ask questions and share your problems and solutions:

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>

OTHER SKILLS AND CONCEPTS

In addition to the key principles of working with data in the construction industry, DataDrivenConstruction addresses a wide range of additional concepts, programs, and skills that are essential for the data-driven professional. Some of these are presented only in an overview but are critical to the practice.

The interested reader can visit the DataDrivenConstruction.io website for links to additional materials on key skills. These materials include working with Python and Pandas, building ETL -processes, examples of data processing in construction CAD projects, big data systems, and modern approaches to construction data visualization and analytics.

Many open source tools and software. were used in the preparation of the book "DataDrivenConstruction" and all the case studies. The author would like to thank the developers and co-authors of the following solutions:

- Python and Pandas - the foundation of data manipulation and automation
- Scipy, NumPy, Matplotlib and Scikit-Learn - libraries for data analysis and machine learning
- SQL and Apache Parquet - tools for storing and processing large amounts of construction data
- Open Source CAD (BIM) open data tools in open formats
- N8n, Apache Airflow, Apache NiFi - orchestration and workflow automation systems
- DeepSeek, LLaMa, Mistral - Open Source LLM

Special thanks to all the participants in discussions on the topic of open data and tools in professional communities and social networks, whose criticisms, comments, and ideas helped to improve the content and structure of this book.

Follow the development of the project on the DataDrivenConstruction.io website, where not only book updates and corrections are published, but also new chapters, tutorials, and practical examples of applying the described techniques.

MAXIMUM CONVENIENCE WITH THE PRINTED VERSION

You are holding a free digital version of **Data-Driven Construction**. For more convenient work and quick access to the materials, we recommend you to pay attention to the **printed edition**:



■ **Always at hand:** the book in printed format will become a reliable working tool, allowing you to quickly find and use the necessary visualizations and diagrams in any working situations

■ **High quality illustrations:** all images and graphics in the printed edition are presented in maximum quality

■ **Quick access to information:** easy navigation, the ability to make notes, bookmarks and work with the book anywhere.

By purchasing the full printed version of the book, you get a convenient tool for comfortable and efficient work with information: the ability to quickly use visual materials in everyday tasks, quickly find the necessary schemes and make notes. In addition, your purchase supports the dissemination of open knowledge.

Order the print version of the book at: datadrivenconstruction.io/books



UNIQUE OPPORTUNITY FOR STRATEGIC POSITIONING

We offer you to place advertising materials in the free version of DataDrivenConstruction. The paid version of the publication has attracted the attention of specialists from more than 50 countries of the world - from Latin America to Asia-Pacific region - for the first year after its publication. To discuss individual terms of cooperation and get detailed information about placement opportunities, please fill out the feedback form on the official portal datadrivenconstruction.io or write to the contacts listed at the end of the book.



THE BOOK CHAPTERS ARE AVAILABLE AT DATADRIVENCONSTRUCTION.IO

You can read chapters of Data-Driven Construction on the Data-Driven Construction website, where sections of the book are gradually published so that you can quickly find the information you need and use it in your work. You will also find many other publications on similar topics, as well as examples of applications and solutions to help you develop your skills and apply data to construction.



LATEST VERSIONS OF THE BOOK DOWNLOAD FROM THE OFFICIAL WEBSITE

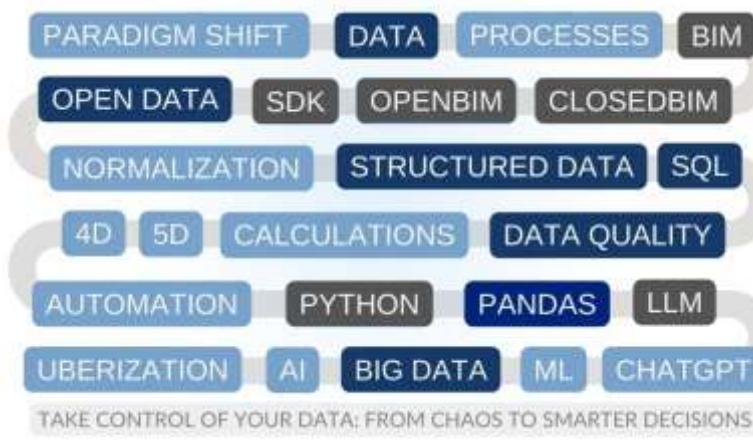
Current and latest versions of the DataDrivenConstruction book are available for download at datadrivenconstruction.io. If you would like to receive updates with new book chapters, practical tips or reviews of new applications, subscribe to the newsletter:

- You will be the first to be introduced to the new sections of the book
- Receive practical cases and tips on analytics and automation in the construction industry
- Keep up to date with trends, publications and application examples

Go to datadrivenconstruction.io to subscribe!

DATADRIVENCONSTRUCTION: CONSULTING, WORKSHOPS AND TRAINING

DataDrivenConstruction training programs and consulting have helped dozens of leading construction companies around the world to increase efficiency, reduce costs and improve the quality of solutions. DataDrivenConstruction's clients include some of the largest players in the billion-euro market, including construction, consulting and IT companies.



Why choose us?

- **Relevance:** talking about the main trends and insights of the industry
- **Hands-on:** helping professionals to efficiently solve everyday tasks through PoC.
- **Individual approach:** we take into account the specifics of your business, ensuring maximum benefit from training and consultations

The DataDrivenConstruction team's main areas of focus:

- **Data quality management:** help parameterize tasks, gather requirements, validate and prepare data for automated processing.
- **Data Mining - extracting and structuring data:** set up ETL processes and extract data from emails, PDF, Excel, images and other sources.
- **BIM and CAD analytics:** collect, structure and analyze information from RVT files, IFC, DWG and other CAD (BIM) formats.
- **Analytics and data transformation:** turning disparate information into structured data, analytics, insights and decisions.
- **Data integration and process automation:** from automated document creation to integration with internal systems and external databases.

Contact DataDrivenConstruction.io to learn how utilizing automation can help your company achieve tangible business results.

GLOSSARY

AI (Artificial Intelligence) - Artificial intelligence; the ability of computer systems to perform tasks that normally require human intelligence, such as pattern recognition, learning, and decision making.

Apache Airflow is an open source workflow orchestration platform for programmatically creating, scheduling, and tracking workflows and ETLs using DAGs (directed acyclic graphs).

Apache NiFi is a tool for automating data flows between systems, specializing in data routing and transformation.

Apache Parquet is an efficient file format for columnar data storage optimized for use in big data analytics systems. It provides significant compression and fast processing.

API (Application Programming Interface) - a formalized interface that allows one program to interact with another without access to the source code, exchanging data and functionality through standardized requests and responses.

Attribute - a characteristic or property of an object that describes its features (e.g., area, volume, cost, material).

Databases are organized structures for storing, managing and accessing information, used for efficient data retrieval and processing.

BEP (BIM Execution Plan) - A building information modeling implementation plan that defines the goals, methods, and processes for implementing BIM in a project.

Big Data - arrays of information of significant volume, variety and update rate, requiring special technologies for processing and analysis.

BI (Business Intelligence) - Business Intelligence; the processes, technologies and tools to transform data into meaningful information for decision making.

BIM (Building Information Modeling) - Building Information Modeling; the process of creating and managing digital representations of physical and functional characteristics of construction projects, including not only 3D models, but also information about characteristics, materials, time and cost.

BlackBox/WhiteBox - approaches to understanding the system: in the first case, the internal logic is hidden, only inputs and outputs are visible; in the second case, the processing is transparent and available for analysis.

Bounding Box is a geometric construct that describes the boundaries of an object in three-dimensional space through the minimum and maximum coordinates on the X, Y and Z axes, creating a "box" around the object.

BREP (Boundary Representation) is a geometric representation of objects that defines them through the boundaries of surfaces.

CAD (Computer-Aided Design) is a computer-aided design system used to create, edit and analyze accurate drawings and 3D models in architecture, construction, engineering and other industries.

CAFM (Computer-Aided Facility Management) is real estate and infrastructure management software that includes space planning, asset management, maintenance and cost monitoring.

CDE (Common Data Environment) - a centralized digital space for managing, storing, sharing and collaborating with project information at all stages of the facility life cycle.

Center of Excellence (CoE) - a specialized structure in an organization responsible for developing a specific area of knowledge, developing standards and best practices, training staff and supporting the introduction of innovations.

CoClass is a modern, third-generation building element classification system.

A conceptual data model is a high-level representation of basic entities and their relationships without attribute granularity, used in the initial stages of database design.

CRM (Customer Relationship Management) is a customer relationship management system used to automate sales and service processes.

DAG (Directed Acyclic Graph) is a directed acyclic graph used in data orchestration systems (Airflow, NiFi) to determine task sequences and dependencies.

Dash is a Python framework for creating interactive web-based data visualizations.

Dashboard - A **dashboard** that visually presents key performance indicators and metrics in real time.

The Data-Centric approach is a methodology that prioritizes data over applications or software code, making data the central asset of the organization.

Data Governance - a set of practices, processes and policies that ensure the appropriate and effective use of data in an organization, including access, quality and security controls.

Data Lake is a storage facility designed to store large amounts of raw data in its original format until it is used.

Data Lakehouse is an architectural approach that combines the flexibility and scalability of data lakes (**Data Lake**) with the manageability and performance of data warehouses (DWH).

Data-Driven Construction is a strategic approach in which every stage of the facility lifecycle - from design to operations - is supported by automated, interconnected systems. This approach provides continuous, fact-based learning, reduces uncertainty, and enables companies to achieve sustainable industry leadership.

Data-Driven integrator - a company specializing in combining data from disparate sources and analyzing it to make management decisions.

Data-Driven approach - A methodology where data is viewed as a strategic asset and decisions are made based on objective analysis of information rather than subjective opinions.

Data Minimalism - an approach to reducing data to the most valuable and meaningful, allowing for simplified processing and analysis of information.

Data Swamp - A scattered array of unstructured data that occurs when information is collected and stored in an uncontrolled manner without proper organization.

DataOps is a methodology that combines DevOps principles, data and analytics, focused on improving collaboration, integration and automation of data flows.

Information digitalization is the process of converting all aspects of construction activities into a digital form suitable for analysis, interpretation and automation.

DataFrame - A two-dimensional tabular data structure in the Pandas library, where rows represent individual records or objects and columns represent their characteristics or attributes.

Descriptive Analytics - Analyzing historical data to understand what happened in the past.

Diagnostic Analytics - Analyze data to determine why something happened.

A Gantt chart is a project planning tool that represents tasks as horizontal bars on a timeline, allowing you to visualize the sequence and duration of work.

DWH (Data Warehouse) is a centralized data warehouse system that aggregates information from multiple sources, structures it and makes it available for analytics and reporting.

ESG (Environmental, Social, Governance) - a set of criteria for assessing the environmental, social and governance impacts of a company or project.

ELT (Extract, Load, Transform) is a process where data is first extracted from sources and loaded into a repository and then transformed for analytical purposes.

ETL (Extract, Transform, Load) is the process of extracting data from various sources, transforming it into the desired format and loading it into the target storage for analysis.

ER-diagram (Entity-Relationship) - a visual diagram showing entities, their attributes and the relationships between them, used in data modeling.

ERP (Enterprise Resource Planning) is a comprehensive modular enterprise resource planning system used to manage and optimize various aspects of the construction process.

Features - In machine learning, independent variables or attributes used as inputs to a model.

Physical data model - a detailed representation of the database structure, including tables, columns, data types, keys and indexes, optimized for a particular DBMS.

FPDF is a Python library for creating PDF documents.

Geometry Core is a software component that provides basic algorithms for creating, editing and analyzing geometric objects in CAD, BIM and other engineering applications.

HiPPO (Highest Paid Person's Opinion) is an approach to decision making based on the opinion of the highest paid person in the organization rather than objective data.

IDE (Integrated Development Environment) - integrated development environment, a comprehensive tool for writing, testing and debugging code (e.g. PyCharm, VS Code, Jupyter Notebook).

IDS (Information Delivery Specification) is an information delivery specification that defines the data requirements at different stages of a project.

IFC (Industry Foundation Classes) is a BIM data exchange format that provides interoperability between different software solutions.

Industry 5.0 is a concept for industrial development that combines the capabilities of digitalization, automation and artificial intelligence with human potential and environmental sustainability.

Data integration is the process of combining data from different sources into a single, coherent system to provide a unified view of information.

Information silos are isolated data storage systems that do not share information with other systems, creating barriers to efficient data utilization.

IoT (Internet of Things) is the concept of connecting physical objects to the internet to collect, process and transmit data.

k-NN (k-Nearest Neighbors) is a machine learning algorithm that classifies objects based on similarity to the nearest neighbors in the training sample.

Kaggle is a platform for data analytics and machine learning competitions.

Calculation - calculation of the cost of construction works or processes for a certain unit of measurement (e.g. for 1 m² of plasterboard wall, 1 m³ of concrete).

KPIs (Key Performance Indicators) are key performance indicators, quantifiable metrics used to evaluate the success of a company or a specific project.

Labels - In machine learning, the target variables or attributes that the model should predict.

Learning Algorithm - The process of finding the best hypothesis in a model corresponding to a target function using a set of training data.

Linear Regression - A statistical method for modeling the relationship between a dependent variable and one or more independent variables.

LLM (Large Language Model) - Large Language Model, an artificial intelligence trained to understand and generate text from huge data sets, capable of analyzing context and writing program code.

LOD (Level of Detail/Development) - the level of detail of the model that determines the degree of geometric accuracy and information content.

A logical data model is a detailed description of entities, attributes, keys, and relationships that reflects business information and rules, an intermediate stage between the conceptual and physical models.

Machine Learning - A class of artificial intelligence techniques that allow computer systems to learn and make predictions from data without explicit programming.

Masterformat is a first generation classification system used to structure construction specifications by section and discipline.

MEP (Mechanical, Electrical, Plumbing) - Building engineering systems that include mechanical, electrical, and plumbing components.

Mesh is a mesh representation of 3D objects consisting of vertices, edges and faces.

Model - In machine learning, a set of different hypotheses, one of which approximates the target function to be predicted or approximated.

Data modeling is the process of creating a structured representation of data and their relationships for implementation in information systems, including conceptual, logical and physical levels.

n8n is an open source tool for automating workflows and integrating applications through a low-code approach.

Normalization - in machine learning, the process of bringing different numerical data to a common scale to facilitate their processing and analysis.

Reverse engineering - the process of studying the device, functioning and manufacturing technology of an object by analyzing its structure, functions and operation. In the context of data - extracting information from proprietary formats for use in open systems.

OCR (Optical Character Recognition) is an **optical character** recognition technology that converts text images (scanned documents, photos) into a machine-readable text format.

OmniClass is a second-generation international classification standard for construction information management.

Ontology - A system of interrelationships of concepts that formalizes a particular field of knowledge.

Open Source - a model for developing and distributing open source software that is available for free use, study and modification.

Open BIM is the concept of open BIM, which implies the use of open standards and formats for data exchange between different software solutions.

Open standards are publicly available specifications for accomplishing a specific task that allow different systems to interoperate and exchange data.

Pandas is an open source Python library for data processing and analysis, providing DataFrame and Series data structures for efficient handling of tabular information.

The open data paradigm is an approach to data processing in which information is made freely available for use, reuse and dissemination by anyone.

Parametric method is a construction project estimation method that uses statistical models to estimate cost based on project parameters.

PIMS (Project Information Model) is a digital system designed to organize, store and share all project information.

Pipeline - A sequence of data processing processes, from extraction and transformation to analysis and visualization.

PMIS (Project Information Management System) is a project management system designed for detailed control of tasks at the level of an individual construction project.

Predictive Analytics - A section of analytics that uses statistical methods and machine learning to predict future outcomes based on historical data.

Prescriptive Analytics - A section of analytics that not only predicts future outcomes, but also suggests optimal actions to achieve the desired results.

Proprietary formats are closed data formats controlled by a particular company that limit the ability to share information and increase dependence on specific software.

QTO (Quantity Take-Off) is the process of extracting quantities of elements from the design documents to calculate the quantities of materials required for the project.

Quality Management System - a quality management system that ensures that processes and results meet established requirements.

RAG (Retrieval-Augmented Generation) is a method that combines the generative capabilities of language models with the extraction of relevant information from corporate databases, improving the accuracy and relevance of answers.

RDBMS (Relational Database Management System) is a relational database management system that organizes information in the form of interrelated tables.

RegEx (Regular Expressions) is a formalized language for searching and processing strings, allowing you to specify templates for checking text data for compliance with certain criteria.

Regression is a statistical method of analyzing the relationship between variables.

CO₂ calculations are a method of estimating carbon dioxide emissions associated with the production and use of construction materials and processes.

Resource method - a method of making estimates based on a detailed analysis of all necessary resources (materials, labor, equipment) to perform construction work.

RFID (Radio Frequency Identification) is a technology for automatically identifying objects using radio signals, used for tracking materials, machinery and personnel.

ROI (Return on Investment) - an indicator reflecting the ratio between profit and invested funds, used to assess the effectiveness of investments.

SaaS (Software as a Service) is a model of software as a service where applications are hosted by a provider and made available to users over the Internet.

SCM (Supply Chain Management) - supply chain management, which includes the coordination and optimization of all processes from the purchase of materials to the delivery of finished products.

Data silos are isolated stores of information in an organization that are not integrated with other systems, making it difficult to share data and inefficient.

SQL (Structured Query Language) is a structured query language used to work with relational databases.

SQLite is a lightweight, embeddable, cross-platform DBMS that does not require a separate server and supports basic SQL functions, widely used in mobile applications and embedded systems.

Structured data - information organized in a specific format with a clear structure, such as in relational databases or tables.

Loosely structured data - information with partial organization and flexible structure, such as JSON or XML, where different elements may contain different sets of attributes.

An entity is a concrete or abstract object of the real world that can be uniquely identified, described and represented in the form of data.

Supervised Learning - A type of machine learning in which an algorithm is trained on labeled data where the desired outcome is known for each example.

Taxonomy - A hierarchical classification system used to systematically categorize elements based on common features.

Titanic Dataset is a popular dataset for training and testing machine learning models.

Training - The process in which a machine learning algorithm analyzes data to identify patterns and form a model.

Transfer learning is a machine learning technique in which a model trained for one task is used as a starting point for another task.

Data Transformation - The process of changing the format, structure, or content of data for later use.

Data requirements are formalized criteria that define the structure, format, completeness and quality of information needed to support business processes.

Uberization of the construction industry is the process of transformation of traditional business models in construction under the influence of digital platforms that provide direct interaction between customers and contractors without intermediaries.

Uniclass is a second and third generation building element classification system widely used in the UK.

USD (Universal Scene Description) is a data format developed for computer graphics, but has gained application in engineering systems due to its simple structure and independence from geometric cores.

Data validation is the process of checking information against established criteria and requirements to ensure accuracy, completeness and consistency of data.

Vector Database - A specialized type of database that stores data as multidimensional vectors for efficient semantic search and object comparison.

Vector representation (embedding) is a method of transforming data into multidimensional numerical vectors that allows machine algorithms to efficiently process and analyze information.

VectorOps is a methodology focused on processing, storing and analyzing multidimensional vector data, especially relevant in areas such as digital twins and semantic search.

Visualization - Graphical representation of data for better perception and analysis of information.

The alphabetical categorization of terms was done by their names in English.

LIST OF REFERENCES AND ONLINE MATERIALS

- [1] Gartner, "IT Key Metrics Data 2017: Index of Published Documents and Metrics," December 12, 2016. [Online]. Available: <https://www.gartner.com/en/documents/3530919>. [Date of access: 1 March 2025].
- [2] KPMG, "Familiar challenges - new approaches. 2023 Global Construction Survey," January 1, 2023. [Online]. Available: <https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2023/06/familiar-challenges-new-solutions-1.pdf>. [Date of access: 5 March 2025].
- [3] F. R. Barnard, "A picture is worth a thousand words," 10 Mapi 1927. [Online]. Available: https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words. [Date of reference: 15 March 2025].
- [4] M. Bastian, "Microsoft CEO Satya Nadella says self-claiming AGI is 'nonsensical benchmark hacking,'" 21 Feb. 2025. [Online]. Available: <https://the-decoder.com/microsoft-ceo-satya-nadella-says-self-claimingagi-is-nonsensical-benchmark-hacking/>. [Date of access: March 15, 2025].
- [5] W. E. Forum, "Forum Shaping the Future of Construction - A Landscape in Transformation;" January 1, 2016. [Online]. Available: https://www3.weforum.org/docs/WEF_Shaping_the_Future_of_Construction.pdf. [Date of address: 2 March 2025].
- [6] C. D. Gillespie, "Clay: Earth's entanglement in the age of clay," 2024. [Online]. Available: <https://ufl.pb.unizin.org/imos/chapter/clay/>.
- [7] "Papyrus, 3rd century B.C. Language is Greek," 2024. [Online]. Available: <https://www.facebook.com/429710190886668/posts/595698270954525>.
- [8] "Monitoring: making use of the tools which are available," 1980. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/10246720/>. [Date of address: 15 March 2025].
- [9] PWC, "Data driven What students need to succeed in a rapidly changing business world," February 15, 2015. [Online]. Available: <https://www.pwc.com/us/en/faculty-resource/assets/PwC-Data-driven-paper-Feb2015.pdf>. [Date of access: 15 March 2025].
- [10] Skanska USA, "Fall Construction Market Trends," 2 November 2023. [Online]. Available: <https://x.com/SkanskaUSA/status/1720167220817588714>.

- [11] "Oxford Essential Quotations (4th ed.)," Oxford University Press, 2016. [Online]. Available: <https://www.oxfordreference.com/display/10.1093/acref/9780191826719.001.0001/q-or-ed4-00006236>. [Date of reference: 1 March 2025].
- [12] "Quote: Sondergaard on Data Analytics," [Online]. Available: <https://www.causeweb.org/cause/resources/library/r2493>. [Date of access: March 15, 2025].
- [13] "How global AI interest is boosting the data management market," May 28, 2024. [Online]. Available: <https://iot-analytics.com/how-global-ai-interest-is-boosting-data-management-market/>. [Date of access: 15 March 2025].
- [14] И. McCue, "History of ERP," 2024. [Online]. Available: <https://www.netsuite.com/portal/resource/articles/erp/erp-history.shtml>.
- [15] erpscout, "ERP Price: How much does an ERP system cost?," [Online]. Available: <https://erpscout.de/en/erp-costs/>. [Date of access: March 15, 2025].
- [16] softwarepath, "What 1,384 ERP projects tell us about selecting ERP (2022 ERP report)," 18 Jan. 2022. [Online]. Available: <https://softwarepath.com/guides/erp-report>. [Accessed March 15, 2025].
- [17] Deloitte, "Data-Driven Management in Digital Capital Projects," December 16, 2016. [Online]. Available: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Real%20Estate/us-engineering-construction-data-driven-management-digital-capital-projects.pdf>. [Date of address: 1 March 2025].
- [18] McKinsey, "The data-driven enterprise of 2025," 28 Jan. 2022. [Online]. Available: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025>. [Date of reference: May 22, 2024].
- [19] Wikipedia, "Moore's law," [Online]. Available: https://en.wikipedia.org/wiki/Moore%27s_law. [Date of access: 15 March 2025].
- [20] Accenture, "Building More Value With Capital Projects," 1 January 2020. [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>. [Accessed March 3, 2024].
- [21] B. Marr, "How much data do we create every day? The Mind-Blowing Stats Everyone Should Read," 2018. [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] "How much data is produced every day?," 2024. [Online]. Available:

- [https://graduate.northeastern.edu/resources/how-much-data-produced-every-day/.](https://graduate.northeastern.edu/resources/how-much-data-produced-every-day/)
- [23] T. Sullivan, "AI and the global "datasphere": how much information will humanity have by 2025?" 2024. [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, "Total number of printed books produced in various regions of Western Europe in each half century between 1454 and 1800," [Internet]. [Online]. Available: <https://www.statista.com/statistics/1396121/europe-book-production-half-century-region-historical/>. [Date of reference: March 1, 2025].
- [25] "Pricing examples," 2024. [Online]. Available: <https://cloud.google.com/storage/pricing-examples>.
- [26] M. Ashare, "Enterprises outsource data storage as complexity rises," 10 May 2024. [Online]. Available: <https://www.ciodive.com/news/enterprises-outsource-data-storage-complexity-rises/715854/>. [Date of reference: 15 March 2025].
- [27] JETSOFTPRO, "SaaS is Dead? Microsoft CEO's Shocking Prediction Explained," January 13, 2025. [Online]. Available: <https://jetsoftpro.com/blog/saas-is-dead/>.
- [28] BG2 Pod, "Satya Nadella | BG2 w/ Bill Gurley & Brad Gerstner," December 12, 2024. [Online]. Available: https://www.youtube.com/watch?v=9NtsnzRFJ_o. [Date of access: March 15, 2025].
- [29] GoodReads, "Tim Berners-Lee," [Online]. Available: <https://www.goodreads.com/quotes/8644920-data-is-a-precious-thing-and-will-last-longer-than>. [Date of access: 15 March 2025].
- [30] KPMG, "Cue Construction 4.0: Make-or-Break Time," 1 Jan. 2023. [Online]. Available: <https://kpmg.com/ca/en/home/insights/2023/05/cue-construction-make-or-break-time.html>. [Date of access: 5 March 2025].
- [31] I. Deininger, B. Koch, R. Bauknecht, and M. Langhans, "Using digital models for decarbonizing a production site: An example of connecting a building model, a production model, and an energy model," 2024. [Online]. 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] McKinsey, "REINVENTING CONSTRUCTION: A ROUTE TO HIGHER PRODUCTIVITY," February 1, 2017. [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] Construction Task Force to the Deputy Prime Minister, "Rethinking Construction," 1 October 2014. [Online]. Available: https://constructingexcellence.org.uk/wp-content/uploads/2014/10/rethinking_construction_report.pdf.
 - [34] Forbes, "Without An Opinion, You're Just Another Person With Data," March 15, 2016. [Online]. Available: <https://www.forbes.com/sites/silberzahnjones/2016/03/15/without-an-opinion-youre-just-another-person-with-data/>. [Date of access: March 15, 2025].
 - [35] Wikiquote, "Charles Babbage," [Online]. Available: https://en.wikiquote.org/wiki/Charles_Babbage. [Date of access: March 15, 2025].
 - [36] SAP, "New Research Finds That Nearly Half of Executives Trust AI Over Themselves," 12 Mar. 2025. [Online]. Available: <https://news.sap.com/2025/03/new-research-executive-trust-ai/>. [Date of access: March 15, 2025].
 - [37] The Canadian Construction Association and KPMG in Canada, 2021, "Construction in a digital world," 1 May 2021. [Online]. Available: <https://assets.kpmg.com/content/dam/kpmg/ca/pdf/2021/05/construction-in-the-digital-age-report-en.pdf>. [Date of access: 5 March 2025].
 - [38] ZCS, "Decoding the Fifth Industrial Revolution," [Online]. Available: <https://www.pwc.in/decoding-the-fifth-industrial-revolution.html>. [Accessed March 15, 2025].
 - [39] M. K, Private Rights and Public Problems: The Global Economics of, Peterson Inst. for Intern. Economics,, 2012.
 - [40] F. N. a. Y. Z. Harvard Business School: Manuel Hoffmann, "The Value of Open Source Software," 24 Jan. 2024. [Online]. Available: <https://www.hbs.edu/faculty/Pages/item.aspx?num=65230>. [Date of access: 15 March 2025].
 - [41] Naval Center for Cost Analysis Air Force Cost Analysis Agency, "Software Development Cost Estimating Handbook," 1 September 2008. [Online]. Available: <https://www.dau.edu/sites/default/files/Migrated/CopDocuments/SW%20Cost%20Est%20Manual%20Vol%20I%20rev%2010.pdf>.
 - [42] McKinsey, "Improving construction productivity," [Online]. Available: <https://www.mckinsey.com/capabilities/operations/our-insights/improving-construction-productivity>. [Date of address: 15 March 2025].
 - [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>

strange-and-awful-path-of-productivity-in-the-us-construction-sector/. [Accessed March 1, 2025].

- [44] McKinsey, "Delivering on construction productivity is no longer optional," August 9, 2024. [Online]. Available: <https://www.mckinsey.com/capabilities/operations/our-insights/delivering-on-construction-productivity-is-no-longer-optional>. [Date of access: 5 March 2025].
- [45] ING Group, "Lagging productivity in construction is driving up construction costs," December 12, 2022. [Online]. Available: <https://think.ing.com/articles/lagging-productivity-drives-up-building-costs-in-many-eu-countries/>. [Date of address: 15 March 2025].
- [46] M. Berman, "Microsoft CEO's Shocking Prediction: 'Agents Will Replace ALL Software,'" December 19, 2024." [Online]. Available: <https://www.youtube.com/watch?v=uGOLYz2pgr8>. [Date of access: 15 March 2025].
- [47] Business Insider, "Anthropic's CEO says that in 3 to 6 months, AI will be writing 90% of the code software developers were in charge of," 15 Mar 2025. [Online]. Available: <https://www.businessinsider.com/anthropic-ceo-ai-90-percent-code-3-to-6-months-2025-3>. [Date of access: March 30, 2025].
- [48] Statista, "Popularity comparison of database management systems (DBMSs) worldwide as of June 2024, by category," June 2024. [Online]. Available: <https://www.statista.com/statistics/1131595/worldwide-popularity-database-management-systems-category/>. [Date of address: 15 March 2025].
- [49] DB-Engines, "DB-Engines Ranking," [Online]. Available: <https://db-engines.com/en/ranking>. [Date of reference: March 15, 2025].
- [50] "Stack Overflow Developer Survey 2023," 2024. [Online]. Available: <https://survey.stackoverflow.co/2023/>.
- [51] "SQL," 2024. [Online]. Available: <https://en.wikipedia.org/wiki/SQL>.
- [52] "Structured and unstructured data: What's the Difference?," 2024. [Online]. Available: <https://www.ibm.com/blog/structured-vs-unstructured-data/>.
- [53] DataDrivenConstruction, "PDF COMPARISON OF DATA FORMATS 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] "Building Information Modeling Whitepaper site," 2003. [Online]. Available: <https://web.archive.org/web/20030711125527/http://usa.autodesk.com/adsk/servlet/item?id=>

- 2255342&siteID=123112.
- [55] A. Boiko, "Lobbying wars and BIM development. Part 5: BlackRock is the master of all technologies. How corporations control open source code," 2024. [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 reference: 02 2025].
 - [57] C. Eastman and A. Cthers, "Eastman, Charles; And Cthers," September 1974. [Online]. Available: <https://files.eric.ed.gov/fulltext/ED113833.pdf>. [Date of access: March 15, 2025].
 - [58] D. Ushakov, "Direct Modeling - Who and Why Needs It? A Review of Competitive Technologies," November 11, 2011. [Online]. Available: https://isicad.net/articles.php?article_num=14805. [Date of address: 15 March 2025].
 - [59] D. Weisberg, "History of CAD," 12 Dec. 2022. [Online]. Available: https://www.shapr3d.com/blog/history-of-cad?utm_campaign=cadhISTORYnet. [Date of reference: 15 March 2025].
 - [60] ADSK, "White Paper Building Information Modeling," 2002. [On the Internet]. Available: https://web.archive.org/web/20060512180953/http:/images.adsk.com/apac_sapac_main/files/4525081_BIM_WP_Rev5.pdf#expand. [Date of address: 15 March 2025].
 - [61] ADSK, "White Paper Building Information Modeling in Practice," [Online]. Available: https://web.archive.org/web/20060512181000/http:/images.adsk.com/apac_sapac_main/files/4525077_BIM_in_Practice.pdf. [Date of access: 15 March 2025].
 - [62] A. Boiko, "Lobbying wars and BIM development. Part 2: open BIM VS closed BIM. Europe VS the rest of the world," 2024. [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. Boiko, "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, "Whitepaper BIM," 2002. [On the Internet]. Available: https://web.archive.org/web/20060512180953/http:/images.autodesk.com/apac_sapac_main/files/4525081_BIM_WP_Rev5.pdf#expand. [Date of address: 15 March 2025].
 - [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. [Date of access: March 15, 2025].
- [66] M. Shacklett, "Structured and unstructured data: Key differences," 2024. [Online]. Available: <https://www.datamation.com/big-data/structured-vs-unstructured-data/>.
- [67] K. Woolard, "Making sense of the growth of unstructured data," 2024. [Online]. 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, "Cost Analysis of Inadequate Interoperability in the," 2004. [Online]. Available: <https://nvlpubs.nist.gov/nistpubs/gcr/2004/nist.gcr.04-867.pdf>. [Date of access: 02 2025].
- [69] CrowdFlower, "Data Science Report 2016," 2016. [Online]. Available: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf. [Date of access: March 15, 2025].
- [70] Analyticsindiamag, "6 Most Time Consuming Tasks For Data Scientists," 15 May 2019. [Online]. Available: <https://analyticsindiamag.com/ai-trends/6-tasks-data-scientists-spend-the-most-time-doing/>.
- [71] BizReport, "Report: Data scientists spend bulk of time cleaning up," 06 July 2015. [Online]. Available: <https://web.archive.org/web/20200824174530/http://www.bizreport.com/2015/07/report-data-scientists-spend-bulk-of-time-cleaning-up.html>. [Date of access: 5 March 2025].
- [72] S. Hawking, "Science AMA Series: Stephen Hawking AMA Answers!," July 27, 2015. [Online]. Available: https://www.reddit.com/r/science/comments/3ny5i/science_ama_series_stephen_hawking_a_ma_answers/. [Accessed March 15, 2025].
- [73] B. Cyphers and K. Doctorow, "Privacy without Monopoly: Data Protection and Interoperability," 2024. [Online]. Available: <https://www.eff.org/wp/interoperability-and-privacy>.
- [74] McKinsey Global Institute, "Open data: Unlocking innovation and performance with liquid information," October 1, 2013. [Online]. Available: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information>. [Date of address: 15 March 2025].
- [75] A. Boiko, "The struggle for open data in the construction industry. The history of AUTOLISP, intelliCAD, openDWG, ODA and openCASCADE," 15 05 2024. [Online]. Available: <https://boikoartem.medium.com/the-struggle-for-open-data-in-the-construction-industry-2b97200e6393>. [Date of address: 16 02 2025].

- [76] Wikipedia, "Microsoft and open source," [Online]. Available: https://en.wikipedia.org/wiki/Microsoft_and_open_source. [Date of access: 15 March 2025].
- [77] TIME, "The Gap Between Open and Closed AI Models Might Be Shrinking. Here's Why That Matters," 5 November 2024. [Online]. Available: <https://time.com/7171962/open-closed-ai-models-epoch/>. [Date of access: 15 March 2025].
- [78] The Verge, "More than a quarter of new code at Google is generated by AI," October 29, 2024. [Online]. Available: <https://www.theverge.com/2024/10/29/24282757/google-new-code-generated-ai-q3-2024>. [Date of access: 15 March 2025].
- [79] McKinsey Digital, "The business case for using GPUs to accelerate analytics processing," December 15, 2020. [Online]. Available: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/the-business-case-for-using-gpus-to-accelerate-analytics-processing>. [Accessed: 15 March 2025].
- [80] PwC, "PwC Open Source Monitor 2019," 2019. [Online]. Available: <https://www.pwc.de/de/digitale-transformation/open-source-monitor-research-report-2019.pdf>. [Accessed 15 Mar. 2025].
- [81] Travers Smith, "The Open Secret: Open Source Software," 2024. [Online]. Available: <https://www.traverssmith.com/knowledge/knowledge-container/the-open-secret-open-source-software/>. [Date of access: March 15, 2025].
- [82] Deloitte, "The data transfer process in corporate transformations," 2021. [Online]. Available: <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/finance/us-the-data-transfer-process-in-corporate-transformations.pdf>. [Date of reference: March 15, 2025].
- [83] gov.uk, "Data Analytics and AI in Government Project Delivery," 20 Mar. 2024. [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] "Quote Origin: Everything Should Be Made as Simple as Possible, But Not Simpler," May 13, 2011. [Online]. Available: <https://quoteinvestigator.com/2011/05/13/einstein-simple/>. [Date of access: March 15, 2025].
- [85] "Transformer (deep learning architecture)," [Online]. Available: [https://en.wikipedia.org/wiki/Transformer_\(deep_learning_architecture\)](https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)). [Date of access: March 15, 2025].
- [86] "Python Packages Download Stats," 2024. [Online]. Available: <https://www.pepy.tech/projects/pandas>.

- [87] Interview Bit, "Top 10 Python Libraries," 2023. [Online]. Available: <https://www.interviewbit.com/blog/python-libraries/#:~:text=With%20more%20than%20137%2C000%20libraries,data%20manipulation%2C%20and%20many%20more>. [Date of access: March 30, 2025].
- [88] "NVIDIA and HP Supercharge Data Science and Generative AI on Workstations," 7 Mar. 2025. [Online]. Available: <https://nvidianews.nvidia.com/news/nvidia-hp-supercharge-data-science-generative-ai-workstations>. [Date of access: March 15, 2025].
- [89] P. Orac, "How to process a DataFrame with millions of rows in seconds," 2024. [Online]. Available: <https://towardsdatascience.com/how-to-process-a-dataframe-with-millions-of-rows-in-seconds>.
- [90] Ç. Uslu, "What is Kaggle?," 2024. [Online]. Available: <https://www.datacamp.com/blog/what-is-kaggle>.
- [91] "NVIDIA CEO Jensen Huang Keynote at COMPUTEX 2024," June 2, 2024. [Online]. Available: <https://www.youtube.com/live/pKXDVsWZmUU?si=Z3Rj1Las8wiPII2w>. [Date of access: March 15, 2025].
- [92] "Members: Founders and corporate members," 2024. [Online]. Available: <https://www.opendesign.com/member-showcase>.
- [93] A. Boiko, "The Age of Change: IFC is a thing of the past or why ADSK and other CAD vendors are willing to give up IFC for USD in 14 key facts," 24 November 2024. [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 access: February 23, 2025].
- [94] A. Boiko, "The post-BIM world. Transition to data and processes and whether the construction industry needs semantics, formats and interoperability," December 20, 2024. [Online]. Available: <https://boikoartem.medium.com/the-post-bim-world-7e35b7271119>. [Date of reference: February 23, 2025].
- [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: February 23, 2025].
- [96] Harvard Business Review, "Bad Data Costs the U.S. \$3 Trillion Per Year," September 22, 2016. [Online]. Available: <https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year>.
- [97] Delpha, "Impacts of Data Quality," 1 Jan. 2025. [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," in Design for Maintainability: The Importance of Operations and Maintenance Considerations During the Design Phase of Construction Projects. [Online]. Available: <https://www.wbdg.org/resources/design-for-maintainability>. [Date of access: 15 March 2025].
- [99] O. o. D. C. P. a. Oversight, "Corrosion Prevention and Control Planning Guidebook for Military Systems and Equipment," April 2014. [Online]. Available: <https://www.dau.edu/sites/default/files/Migrated/CopDocuments/CPC%20Planning%20Guidebook%204%20Feb%202014.pdf>. [Date of access: 15 March 2025].
- [100] Gartner, "Data Quality: Best Practices for Accurate Insights," 1 Jan. 2025. [Online]. Available: <https://www.gartner.com/en/data-analytics/topics/data-quality>.
- [101] "For Want of a Nail," [Online]. Available: https://en.wikipedia.org/wiki/For_Want_of_a_Nail. [Date of access: March 15, 2025].
- [102] McKinsey Global Institute, "Open data: Unlocking innovation and performance with liquid information," October 2013. [Online]. Available: https://www.mckinsey.com/~media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/open%20data%20unlocking%20innovation%20and%20performance%20with%20liquid%20information/mgi_open_data_fullreport_oct2013.pdf. [Date of reference: March 15, 2025].
- [103] EY, "The path to carbon neutrality," 10 Mar. 2023. [Online]. Available: https://www.ey.com/ru_kz/services/consulting/the-path-to-carbon-neutrality. [Date of access: 15 March 2025].
- [104] PWC, "ESG Awareness," 1 July 2024. [Online]. Available: <https://www.pwc.com/kz/en/assets/esg-awareness/kz-esg-awareness-rus.pdf>. [Date of access: March 15, 2025].
- [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_calculating the embodied carbon," 2024. [Online]. Available: https://github.com/datadrivenconstruction/CO2_calculating-the-embodied-carbon.
- [107] McKinsey, "Imagining Construction's Digital Future," June 24, 2016. [Online]. Available: <https://www.mckinsey.com/capabilities/operations/our-insights/imagining-constructions-digital-future>. [Date of reference: February 25, 2025].

- [108] Bund der Steuerzahler Deutschland e.V., "Das Schwarzbuch," 10 Oct. 2024. [Online]. Available: <https://steuerzahler.de/aktuelles/detail/das-schwarzbuch-202425/>. [Date of address: 15 March 2025].
- [109] SAS, "Data lake and data warehouse - know the difference," [Online]. Available: https://www.sas.com/en_is/insights/articles/data-management/data-lake-and-data-warehouse-know-the-difference.html. [Date of reference: 15 March 2025].
- [110] ADSK, "Building Information Modeling," 2002. [On the Internet]. Available: https://www.laiserin.com/features/bim/autodesk_bim.pdf. [Date of address: 15 March 2025].
- [111] A. Boiko, "BIM History Map," 2024. [Online]. Available: <https://bigdataconstruction.com/history-of-bim/>.
- [112] A. S. Borkowski, "Definitions of BIM by Organizations and Standards," December 27, 2023. [Online]. Available: <https://encyclopedia.pub/entry/53149>. [Date of address: 5 March 2025].
- [113] CAD Vendor, "OPEN BIM Program," 2012. [Online]. Available: https://web.archive.org/web/20140611075601/http://www.graphisoft.com/archicad/open_bim/. [Date of address: 30 March 2025].
- [114] Wikipedia, "Industry Foundation Classes," [Online]. Available: https://en.wikipedia.org/wiki/Industry_Foundation_Classes. [Date of access: 15 March 2025].
- [115] Wikipedia, "IGES," [Online]. Available: <https://en.wikipedia.org/wiki/IGES>. [Date of reference: 30 March 2025].
- [116] A. Boiko, "History of CAD (BIM)," 15 Dec. 2021. [Online]. Available: https://miro.com/app/board/o9J_laML2cs=/. [Date of address: February 24, 2025].
- [117] T. K. K. A. O. F. B. C. E. L. H. H. E. L. P. N. S. H. T. J. v. L. H. H. G. D. H. T. K. C. L. A. W. J. S. Francesca Noardo, "Reference study of IFC software support: the GeoBIM benchmark 2019 - Part I," 8 Jan. 2021. [Online]. Available: <https://arxiv.org/pdf/2007.10951.pdf>. [Date of access: 5 March 2025].
- [118] И. Rogachev, "Let's Talk BIM: Maxim Nechiporenko | Renga | IFC | Domestic BIM," April 13, 2021. [Online]. Available: <https://www.youtube.com/watch?t=3000&v=V03Y9uuzF9M&feature=youtu.be>. [Date of address: 5 March 2025].
- [119] D. Ares, "RETS in Real Estate: Why It's Crucial for Efficiency & Growth," December 17, 2024. [Online]. Available: <https://www.realpha.com/blog/rets-importance-in-real-estate-explained>.

- [Date of reference: 5 March 2025].
- [120] "Flex token cost," 2024. [Online]. Available: <https://www.adsk.com/buying/flex?term=1-YEAR&tab=flex>.
- [121] A. Boiko, "Forget BIM and democratize access to data (17. Kolloquium Investor - Hochschule - Bauindustrie)," 2024. [Online]. Available: <https://www.bim.bayern.de/wp-content/uploads/2023/06/Kolloquium-17-TUM-Bauprozessmanagement-und-Bauindustrie.pdf>.
- [122] D. 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," 1 Jan. 2018. [Online]. Available: <https://www.ise.fraunhofer.de/de/forschungsprojekte/scope.html>. [Accessed: 2 Mar 2025].
- [124] Apple.com, "Pixar, Adobe, Apple and NVIDIA form Alliance for OpenUSD to drive open standards for 3D content," August 1, 2023. [Online]. Available: <https://www.apple.com/newsroom/2023/08/pixar-adobe-apple-adsk-and-nvidia-form-alliance-for-openusd/>. [Accessed: 2 March 2025].
- [125] AECmag, "ADSK's granular data strategy," 25 Jul 2024. [Online]. Available: <https://aecmag.com/technology/autodesks-granular-data-strategy/>. [Date of access: 15 March 2025].
- [126] A. Boiko, "The Age of Change: IFC is a thing of the past or why ADSK and other CAD vendors are willing to give up IFC for USD in 14 key facts," 24 11 2024. [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-3f9a82cccd10a>. [Date of address: February 23, 2025].
- [127] A. Boiko, "ENG BIM Cluster 2024 | The Battle for Data and Application of LLM and ChatGPT in the Construction," August 7, 2024. [Online]. Available: ENG BIM Cluster 2024 | The Battle for Data and Application of LLM and ChatGPT in the Construction. [Accessed: March 15, 2025].
- [128] "Jeffrey Zeldman Presents," May 6, 2008. [Online]. Available: <https://zeldman.com/2008/05/06/content-precedes-design/>. [Date of access: 15 March 2025].
- [129] A. Boiko, "DWG Analyse with ChatGPT | DataDrivenConstruction," 5 Mar 2024. [Online]. Available: <https://www.kaggle.com/code/artemboiko/dwg-analyse-with-chatgpt-datadrivenconstruction>. [Date of reference: 15 March 2025].

- [130] McKinsey, "The McKinsey guide to outcompeting in the age of digital and AI," 2023. [Online]. Available: <https://www.mckinsey.com/featured-insights/mckinsey-on-books/rewired>. [Date of reference: March 30, 2025].
- [131] Forbes, "Data Storytelling: The Essential Data Science Skill Everyone Needs," March 31, 2016. [Online]. Available: <https://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs/>. [Accessed March 15, 2025].
- [132] J. Bertin, "Graphics and Graphic Information Processing," 8 September 2011. [Online]. Available: https://books.google.de/books/about/Graphics_and_Graphic_Information_Process.html?id=csqX_xnm4tcC&redir_esc=y. [Date of access: 15 March 2025].
- [133] CauseWeb, "Wells/Wilks on Statistical Thinking," [Online]. Available: <https://www.causeweb.org/cause/resources/library/r1266>. [Date of access: March 15, 2025].
- [134] Ministrymagazine, "How science discovered Creation," January 1986. [Online]. Available: <https://www.ministrymagazine.org/archive/1986/01/how-science-discovered-creation>. [Date of reference: March 15, 2025].
- [135] BCG, "Data-Driven Transformation: Accelerate at Scale Now," May 23, 2017. [Online]. Available: <https://www.bcg.com/publications/2017/digital-transformation-transformation-data-driven-transformation>. [Date of access: May 15, 2024].
- [136] "How to build a data architecture to drive innovation-today and tomorrow," June 3, 2020. [Online]. Available: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/how-to-build-a-data-architecture-to-drive-innovation-today-and-tomorrow>. [Date of address: 15 March 2025].
- [137] Oxford, "Woodrow Wilson 1856-1924," [Online]. Available: <https://www.oxfordreference.com/display/10.1093/acref/9780191866692.001.0001/q-oro-ed6-00011630>. [Date of address: 15 March 2025].
- [138] "Convertors," 2024. [Online]. Available: <https://datadrivenconstruction.io/index.php/convertors/>.
- [139] PWC, "Sizing the prize What's the real value of AI for your business and how can you capitalize?" 1 January 2017. [Online]. Available: <https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf>. [Date of access: February 18, 2025].
- [140] "Pipeline in Construction," 2024. [Online]. Available: <https://datadrivenconstruction.io/index.php/pipeline-in-construction/>.

- [141] Wikipedia, "Apache NiFi," 1 January 2025. [Online]. Available: https://de.wikipedia.org/wiki/Apache_NiFi. [Date of access: 5 March 2025].
- [142] n8n, "Gmail AI Auto-Responder: Create Draft Replies to incoming emails," May 1, 2024. [Online]. Available: <https://n8n.io/workflows/2271-gmail-ai-auto-responder-create-draft-replies-to-incoming-emails/>. [Date of access: March 15, 2025].
- [143] n8n, "Real Estate Daily Deals Automation with Zillow API, Google Sheets and Gmail," 1 Mar 2025. [Online]. Available: <https://n8n.io/workflows/3030-real-estate-daily-deals-automation-with-zillow-api-google-sheets-and-gmail/>. [Accessed March 15, 2025].
- [144] B. T. O'Neill, "Failure rates for analytics, AI, and big data projects = 85% - yikes!" 1 Jan. 2025. [Online]. Available: <https://designingforanalytics.com/resources/failure-rates-for-analytics-bi-iot-and-big-data-projects-85-yikes/>.
- [145] J. Neyman, On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, Oxford University Press, 1934.
- [146] T. J. S. a. J. S. Jesse Perla, "A Problem that Stumped Milton Friedman," Quantitative Economics with Julia, 1 Jan. 2025. [Online]. Available: https://julia.quantecon.org/dynamic_programming/wald_friedman.html. [Date of access: 1 May 2024].
- [147] T. Landsall-Welfair, Predicting the nation's current mood, Significance, 2012.
- [148] A. Boiko, "San Francisco. Construction Sector 1980-2019," 2024. [Online]. Available: <https://www.kaggle.com/search?q=San+Francisco.+Building+sector+1980-2019>.
- [149] A. Boiko, "Kaggle: RVT IFC Files 5000 Projects," 2024. [Online]. Available: <https://www.kaggle.com/datasets/artemboiko/rvtifc-projects>.
- [150] CFMA, "Preparing for the Future with Connected Construction," [Online]. Available: <https://cfma.org/articles/preparing-for-the-future-with-connected-construction>. [Accessed March 15, 2025].
- [151] Cisco, "Cisco Survey Reveals Close to Three-Fourths of IoT Projects Are Failing," May 22, 2017. [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] "Conditions Required for Plant Fossil Preservation," 2024. [Online]. Available: <https://ucmp.berkeley.edu/IB181/VPL/Pres/PresTitle.html>.

- [153] "BlackRock's Fink on bonds, mergers and acquisitions, the U.S. recession, and the election: Full Interview," 2023. [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 famous AI disasters," 02 October 2024. [Online]. Available: <https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html>. [Date of access: March 15, 2025].
- [155] G. Kasparov, Deep Thinking, PublicAffairs, 2017.
- [156] Wikipedia, "Kaggle," 1 January 2025. [Online]. Available: <https://en.wikipedia.org/wiki/Kaggle>. [Date of access: March 15, 2025].
- [157] Kaggle, "Titanic - Machine Learning from Disaster," 1 Jan. 2025. [Online]. Available: <https://www.kaggle.com/competitions/titanic/overview>. [Accessed March 10, 2025].
- [158] Ш. Johri, "Creating ChatGPT: From Data to Dialogue," 2024. [Online]. Available: <https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/>.
- [159] П. Domingos, "A few useful things to know about machine learning," 2024. [Online]. Available: <https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>.
- [160] J. Saramago, "Quotable Quote," [Online]. Available: <https://www.goodreads.com/quotes/215253-chaos-is-merely-order-waiting-to-be-deciphered>. [Date of access: March 17, 2025].
- [161] NVIDIA, "Enhance Your Training Data with New NVIDIA NeMo Curator Classifier Models," December 19, 2024. [Online]. Available: <https://developer.nvidia.com/blog/enhance-your-training-data-with-new-nvidia-nemo-curator-classifier-models/>. [Accessed March 25, 2025].
- [162] "NVIDIA Announces Major Release of Cosmos World Foundation Models and Physical AI Data Tools," March 18, 2025. [Online]. Available: <https://nvidianews.nvidia.com/news/nvidia-announces-major-release-of-cosmos-world-foundation-models-and-physical-ai-data-tools>. [Accessed March 25, 2025].
- [163] NVIDIA, "NVIDIA Isaac Sim," [On the Internet]. Available: <https://developer.nvidia.com/isaac/sim>. [Accessed March 25, 2025].
- [164] M. Quarterly, "Why digital strategies fail," 25 Jan. 2018. [Online]. Available: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/why-digital-strategies-fail>. [Accessed March 15, 2025].

- [165] M. J. Perry, "My Favorite Milton Friedman Quotes," 17 November 2006. [Online]. Available: <https://www.aei.org/carpe-diem/my-favorite-milton-friedman-quotes/>. [Date of access: March 1, 2025].
- [166] J. A. Wheeler, "Information, physics, quantum: the search for links," 1990.
- [169] A. Boiko, "Lobbying wars and BIM development. Part 5: BlackRock is the master of all technologies. How corporations control open source code," 2024. [Online]. Available: <https://boikoartem.medium.com/lobbyist-wars-and-the-development-of-bim-d72ad0111a7d>.
- [170] T. Krijnen and J. Beetz, "A SPARQL query engine for binary-formatted IFC building models," *Advanced Engineering Informatics*, 2024.
- [171] "Number of businesses in the construction sector in the UK in 2021, by size of business," 2024. [Online]. Available: <https://www.statista.com/statistics/677151/uk-construction-businesses-by-size/>.
- [172] "5,000 IFC&RVT projects," 2024. [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/>. [Date of reference: February 2025].
- [174] *Lobbying Wars Over Data in Construction / Techno-Feudalism and the History of BIM's Hidden Past*. [Film]. Germany: Artem Boiko, 2023.
- [175] A. Boiko, "CHATGPT WITH REVIT AND IFC | Automatic retrieval of documents and data from projects," 16 Nov. 2023. [Online]. Available: https://www.youtube.com/watch?v=ASXolti_YPs&t. [Date of access: March 2, 2025].
- [176] M. & Company, "Three new mandates for capturing a digital transformation's full value," 22 Jan. 2022. [Online]. Available: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/three-new-mandates-for-capturing-a-digital-transformations-full-value>. [Date of access: February 15, 2025].
- [177] KPMG, "Construction in a Digital World," 1 May 2021. [Online]. Available: <https://assets.kpmg.com/content/dam/kpmg/ca/pdf/2021/05/construction-in-the-digital-age-report-en.pdf>. [Date of access: April 5, 2024].
- [178] LLP, KPMG, "Cue Construction 4.0: Make-or-Break Time," 17 Mar. 2023. [Online]. Available: <https://kpmg.com/ca/en/home/insights/2023/05/cue-construction-make-or-break-time.html>.

[Date of access: February 15, 2025].

- [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>. [Date of access: 15 March 2025].
- [180] Forbes, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," March 23, 2016. [Online]. Available: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>. [Date of access: 15 Mar 2025].
- [181] UK Foreign, Commonwealth and Development Office, "Digital development strategy 2024 to 2030," 18 Mar. 2024. [Online]. Available: <https://www.gov.uk/government/publications/digital-development-strategy-2024-to-2030/digital-development-strategy-2024-to-2030>. [Date of access: 15 March 2025].
- [182] "Vision and Strategy in the Building Design Industry," 7 November 2003. [Online]. Available: <https://web.archive.org/web/20030711125527/http://usa.adsk.com/adsk/servlet/item?id=2255342&siteID=123112>. [Date of address: 5 March 2025].
- [183] M. Bocharov, "Information modeling," March 2025. [On the Internet]. Available: <https://www.litres.ru/book/mihail-evgenevich-bocharov/informacionnoe-modelirovanie-v-rossii-71780080/chitat-onlayn/?page=5>. [Date of address: 15 March 2025].
- [184] "Integrated Design-Through-Manufacturing: Benefits and Rationale," 2000. [Online]. Available: https://web.archive.org/web/20010615093351/http://www3.autodesk.com:80/adsk/files/734489_Benefits_of_MAI.pdf. [Date of reference: March 25, 2025].
- [185] CAD Vendor, "Open BIM Program is a marketing campaign," 12 Mar. 2012. [Online]. Available: <https://web.archive.org/web/20120827193840/http://www.graphisoft.com/openbim/>. [Date of address: 30 March 2025].

SUBJECT INDEX

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

AIA, 289

AIM, 3, 289

AMS, 14, 84, 151, 153

Apache Airflow, 311, 361, 362, 363, 364, 366, 367, 369, 371, 399, 494

Apache NiFi, 116, 311, 361, 367, 368, 371, 399, 482, 494

Apache ORC, 62, 330, 378

Apache Parquet, 62, 67, 380, 381, 494

API, 54, 94, 95, 97, 109, 137, 138, 154, 168, 218, 219, 256, 257, 260, 271, 273, 294, 295, 296, 297, 300, 308, 326, 329, 342, 366, 369, 370, 488

B

BDS, 258, 259

Big Data, 9, 67, 245, 383

BIM, 2, 3, 4, 6, 3, 14, 17, 24, 56, 58, 60, 63, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 84, 92, 137, 138, 139, 140, 141, 142, 144, 150, 154, 156, 166, 172, 183, 184, 186, 187, 190, 195, 196, 208, 210, 211, 213, 216, 217, 218, 219, 220, 221, 222, 227, 228, 237, 238, 239, 242, 243, 246, 250, 251, 252, 255, 256, 257, 258, 259, 260, 261, 262, 263, 266, 271, 272, 273, 275, 276, 277, 278, 279, 280, 282, 285, 287, 288, 289, 290, 291, 292, 293, 294,

295, 296, 297, 298, 299, 300, 301, 308, 309, 324, 328, 337, 351, 356, 358, 361, 389, 413, 416, 448, 462, 466, 475, 476, 492, 494, 497

BlackBox, 240, 242, 243.

BMS, 8

Bokeh, 320, 337

BOM, 76, 77, 79, 257, 263

Bounding Box, 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, 406, 413, 416, 431, 448, 462, 466, 475, 476, 480, 492, 494, 497

CAE, 16, 283

CAFM, 14, 24, 62, 84, 151, 153, 172, 233, 278, 326, 356, 387, 458, 473

CAM, 16, 78

CAPEX, 14, 82

CDE, 84, 175, 388, 389, 390

ChatGPT, 103, 104, 109, 110, 124, 129, 221, 303, 307, 343, 349, 421

Claude, 103, 104, 109, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 429, 443

CO₂, 72

CO₂, 229, 234, 235, 236, 237, 238, 239, 243

COBie, 156, 289, 292.

CoE, 56, 168, 169, 170, 477

Copilot, 114, 116, 459

CPIXML, 143, 272, 273, 276, 277, 279, 285, 296

CPM, 14, 17, 62, 166, 175, 233, 326, 473

CQMS, 14, 84, 177, 178, 462

CRM, 109, 369, 459.

CRUD, 51, 459

CSG, 263

CSV, 61, 62, 88, 89, 120, 123, 128, 129, 130, 131, 135, 168, 268, 272, 280, 329, 333, 344, 354, 354, 356, 373, 377, 378, 379, 380, 401, 407, 414, 474, 480

D

DAE, 276, 277, 278, 280, 281, 284, 285, 414

DAG, 362, 363, 365, 366

Dash, 320, 336, 337.

Data Governance, 373, 395, 396, 398, 400, 401

Data Lake, 214, 373, 376, 384, 385, 386, 387, 388, 389, 390, 400

Data Lakehouse, 67, 373, 386, 387, 388

Data Minimalism, 373, 395, 396, 400, 401

Data Swamp, 373, 395, 397, 401

Data Warehouse, 382, 383, 400

Data-as-a-Service, 487

data-driven, 50, 170, 460, 461, 484, 486

DataFrame, 67, 117, 121, 122, 123, 125, 129, 130, 131, 133, 134, 135, 137, 219, 220, 224, 237, 308, 328, 329, 330, 332, 333, 344, 345, 347, 348, 354, 365, 377, 381, 409, 414, 415, 475

DataOps, 170, 373, 398, 399, 400, 401, 475

DeepSeek, 103, 104, 107, 109, 110, 124, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 421, 429, 443, 494

DGN, 8, 140, 186, 227, 357

DWG, 8, 70, 71, 73, 97, 140, 186, 211, 227, 272, 287, 302, 303, 304, 307, 357, 376, 497

DWH, 67, 373, 376, 382, 383, 384, 386, 387, 388, 389, 390

DXF, 8, 73, 277

E

ECM, 58, 175

ECS, 142

EIR, 289

eLOD, 289

ELT, 384, 385

EPM, 14, 166, 198

ERP, 2, 11, 12, 14, 17, 18, 24, 25, 58, 62, 109, 153, 166, 172, 175, 196, 198, 210, 232, 239, 240, 241, 242, 243, 244, 245, 246, 247, 249, 272, 277, 278, 279, 282, 326, 351, 356, 361, 369, 387, 388, 389, 390, 458, 462, 473, 475, 487, 492

ESG, 196, 235, 236, 238

ETL, 6, 1.1-8, 32, 49, 81, 113, 116, 119, 128, 188, 193, 219, 291, 311, 312, 317, 323, 324, 325, 326, 327, 328, 330, 331, 333, 338, 339, 340, 343, 344, 348, 349, 350, 351, 353, 354, 356, 361, 362, 363, 364, 365, 367, 371, 372, 381, 382, 383, 384, 385, 399, 409, 430, 475, 481, 494

Excel, 57, 61, 62, 65, 66, 85, 88, 111, 120, 123, 125, 132, 154, 167, 187, 210, 223, 224, 226, 227, 228, 278, 291, 329, 333, 342, 351, 365, 376, 378, 459, 475, 497

Extract, 81, 128, 134, 193, 311, 323, 324, 326, 328, 330, 345, 361, 362, 363, 365, 383, 384, 450, 475

F

Feather, 62, 123, 330, 378

FPDF, 339, 340, 341, 342, 343

G

GDPR, 109

GIS, 58

GLTF, 143, 278

Google Sheets, 368, 370

Grok, 103, 104, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 429, 443

H

HDF5, 62, 67, 123, 329, 330, 378, 379, 380

HiPPO, 29, 37, 95, 424, 477, 484, 490

HTML, 123, 340, 365, 370

I

IDS, 289, 290, 291.

IFC, 8, 73, 138, 142, 186, 227, 261, 262, 263, 264, 265, 266, 267, 268, 272, 273, 273, 276, 277, 278, 279, 280, 284, 286, 292, 296, 302, 329, 357, 414, 415, 417, 497

IGES, 262, 263, 276.

iLOD, 289

IoT, 10, 18, 67, 271, 367, 369, 405, 413, 417, 417, 418, 419, 455, 460, 465, 482, 484

ISO 19650, 388

J

JavaScript, 320, 369, 378

JSON, 88, 89, 90, 92, 123, 128, 142, 269, 272, 280, 329, 330, 333, 378, 474, 480

Jupyter Notebook, 114, 115, 116, 130, 187, 224, 330, 346, 417, 425

K

Kaggle, 115, 121, 130, 187, 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

Load, 81, 128, 134, 193, 311, 323, 325, 326, 333, 334, 338, 339, 342, 343, 361, 362, 363, 365, 367, 383, 384, 450, 475

LOD, 287, 289

LOI, 287

LOMD, 287

Low-Code, 368, 369

M

Matplotlib, 123, 306, 320, 333, 335, 342, 372, 409, 411, 494

MCAD, 77, 78, 257, 284

MEP, 14, 175, 492

MESH, 234, 283, 284, 285, 296, 462

Microsoft SQL, 65

Mistral, 103, 104, 107, 110, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 421, 429, 443, 494

MRP, 11, 12

MS Project, 70

MySQL, 63, 64, 65, 291, 329, 330

N

n8n, 116, 311, 361, 368, 369, 370, 371, 372

NLP, 69, 394

No-Code, 368, 369

NURBS, 142, 282, 283, 284, 285

NWC, 8, 276

O

OBJ, 143, 273, 276, 277, 278, 280, 281, 284, 285

OCCT, 273

OCR, 69, 128, 132, 134

OmniClass, 154, 155, 156.

open BIM, 142, 216, 219, 256, 261, 278, 291

Open Source, 45, 55, 97, 98, 107, 108, 273, 275, 494

OWL, 267, 268, 269.

P

Pandas, 56, 67, 103, 117, 118, 119, 120, 121, 122, 123, 125, 125, 130, 134, 138, 186, 187, 220, 221, 225, 227, 269, 300, 303, 328, 329, 330, 377, 378, 380, 381, 403, 408, 409, 410, 411, 412, 414, 475, 479, 494

Parquet, 67, 123, 329, 330, 373, 378, 379, 380, 381, 382, 386, 401, 403, 414, 474, 480

PDF, 69, 70, 71, 85, 111, 126, 127, 128, 129, 130, 131, 132, 134, 146, 147, 177, 186, 190, 191, 211, 215, 278, 338, 339, 340, 341, 342, 342, 344, 345, 346, 347, 354, 355, 356, 357, 365, 376, 475, 480, 497

PDM, 16

PHP, 63, 339

Pipeline, 44, 53, 115, 128, 183, 303, 307, 308, 311, 312, 349, 350, 351, 352, 354, 355, 356, 367, 370, 417, 430, 476

PLM, 16, 246

PLN, 8, 272, 296

Plotly, 320, 336, 337, 372

PMIS, 3, 24, 32, 62, 152, 196, 198, 211, 239, 240, 245, 246, 247, 248, 249, 250, 272, 326, 351, 387, 388, 389, 390, 458, 462, 487

PMS, 84, 151, 233

PostgreSQL, 63, 64, 65, 329, 395

Power BI, 320, 372, 482

private equity, 464, 487

Python, 56, 63, 103, 105, 112, 114, 115, 117, 118, 119, 129, 130, 131, 133, 134, 135, 166, 167, 179, 187, 219, 224, 225, 227, 303, 304, 308, 320, 329, 330, 332, 333, 339, 340, 346, 348, 356, 367, 369, 372, 378, 409, 417, 426, 459, 475, 476, 479, 494

Q

QTO, 72, 196, 214, 215, 216, 217, 218, 219, 221, 223, 225, 226, 228, 237, 238, 242, 243, 251, 301, 302, 475, 479

QWEN, 103, 104, 124, 129, 133, 166, 219, 221, 300, 329, 343, 349, 381, 429, 443

R

RAG, 111, 116
 RDBMS, 63, 64, 65, 82, 89.
 RDF, 267, 268, 269
 RegEx, 126, 136, 177, 179, 331, 332, 333, 357, 474
 RFID, 8, 18, 58, 84, 418, 419, 455, 460, 465, 482
 ROI, 311, 317, 319, 321, 351, 370, 389, 479
 RPM, 14, 84, 151, 331, 333, 334.
 RVT, 8, 73, 77, 140, 186, 227, 272, 296, 300, 302, 322, 329, 357, 376, 414, 415, 417, 497

S

SaaS, 24, 50, 51, 52, 458
 SCOPE, 273, 277
 SDK, 139, 141, 257, 264, 273, 281, 286, 296, 329, 414
 Seaborn, 123, 320, 336, 337, 408, 412
 SPARQL, 269
 SQL, 63, 65, 66, 88, 89, 103, 105, 119, 123, 166, 168, 268, 269, 279, 276, 277, 300, 329, 376, 392, 494
 SQLite, 63, 64, 65, 166, 167, 296, 329, 363
 STEP, 261, 262, 263, 266, 272, 276, 277, 292
 Streamlit, 336, 337
 SVF, 142, 276, 285

T

Transform, 128, 193, 311, 323, 325, 326, 330, 331, 333, 334, 340, 341, 342, 344, 347, 361, 362, 363, 365, 367, 383, 384, 450, 475

U

Uniclass, 154, 155, 156.
 USD, 142, 143, 255, 276, 277, 278, 279, 280, 281, 284, 285, 286, 292, 296

V

VectorOps, 373, 398, 400, 401
 VR, 84, 271, 285

W

WhiteBox, 240, 242, 243.

X

XLSX, 8, 61, 62, 123, 128, 129, 231, 268, 276, 277, 280, 296, 299, 308, 351, 373, 377, 378, 379, 380, 401, 474, 480
 XML, 61, 88, 89, 92, 128, 143, 269, 272, 277, 280, 291, 299, 329, 378, 414, 474



EXPLORE HOW DATA IS RESHAPING CONSTRUCTION

What's Inside

- 100+ key data-related topics in AEC
- 300+ unique visualizations and diagrams
- 50+ real-world business cases
- Hands-on LLM & AI applications
- Code examples & ready-to-use workflows

Topics

- Data-Driven Construction
- Digital Transformation in AEC
- Data Analytics & Automation
- Data Quality & Governance
- CAD, BIM, and Data Interoperability
- LLMs, AI, and Machine Learning in Construction
- Project Cost and Time Forecasting

Audience

- Construction Project Managers
- Architects and Civil Engineers
- BIM Coordinators and Data Managers
- Digital Transformation Leads
- IT and AI Experts in AEC
- Urban Planners and Sustainability Consultants
- Students in Architecture and Engineering
- Educators and Researchers

Website

- www.datadrivenconstruction.io

Comments on the first edition:



"Boiko is the James Carville of IT – in the latter's much-quoted "It's the economy, stupid", only one word needs to be exchanged for this famous book. "It's the data, stupid." (not the software) And to find one's way in the data universe, a saying of the ancient Romans dating back to Greek is still valid today: "Navigare necesse est". The author navigates his readers through all the depths and shallows of the data ocean with a sure hand and an unwavering compass, not to mention a comprehensive historical approach and, last but not least, highly original graphics..."

— Dr. Burkhard Talebitari



"Artem Boiko's book is a milestone for the democratization of digitization in the construction industry – and a real game changer for small and medium-sized enterprises (SMEs). This book is a call to action! It is a valuable guide for anyone who not only wants to understand the digital transformation in the construction industry, but also wants to actively shape it – pragmatically, efficiently and in a forward-looking way. Now is the time to work together to share this knowledge and sustainably increase the productivity of the construction industry..."

— Dr. Michael Max Buehler

This practical guide empowers both professionals and newcomers to navigate the fast-evolving world of data-driven construction. From the foundations of data management to advanced digital workflows, AI tools, and real-world applications – this book is your roadmap to smarter, faster, and more efficient construction processes.



ISBN: 978-3-00078-229-9



9 783000 782299