Migration Letters

Volume: 19, No: S8 (2022), pp. 2046-2068

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)

www.migrationletters.com

AI-Driven Data Engineering: Automating Data Quality, Lineage, And Transformation In Cloud-Scale Platforms

Phanish Lakkarasu

Abstract

In this book, we examine the need for automated data engineering. As organizations accelerate their adoption of cloud-scale analytics platforms, they increasingly rely on centralized core teams to deliver trusted data domains quickly and publish them for consumption by data analysts and business users. However, as the complex network of data domains increases, the cost and time to create and maintain them become unmanageable. Moreover, tensions grow between the core teams and the business teams who are most dependent on timely and relevant data, due to the long turnaround times for updates and the low tolerance for data quality issues.

Increasingly, organizations are recognizing the need for self-service data democratization to empower users close to business operations to curate their data domains. These business users best understand the specifics of their domain and know best how to enrich the quality of the domain. Accordingly, organizations need to ensure that the right tooling is provided to the business user so that they can publish relevant and trusted business data while ensuring central governance and security compliance. Technology must be used to ensure high data quality at scale and mitigate data usage and compliance risks. This need has recently become the focus of renewed efforts by cloud provider companies and third parties. These companies are investing in solutions that automate many parts[1] of the data engineering process. Other new companies are accelerating the introduction of AI-led autonomous systems. Increasingly, solutions are moving closer to the business users, lowering the barrier for them to build trusted data domains. These market moves represent directions that data engineers should look to thoughtfully embrace as partners, not competitors.

Keywords: *Automated Data Engineering, Cloud-Scale Analytics, Trusted Data Domains, Self-Service Data Democratization, Business-Led Data Curation, Centralized Data Governance, Data Quality at Scale, AI-Led Data Automation, Autonomous Data Systems, Domain-Specific Data Ownership, Data Compliance Tooling, Low-Code Data Publishing, Data Engineering Acceleration, Scalable Data Management, Data Platform Enablement, User-Centric Data Tools, Cloud-Native Data Architecture, Business-Driven Analytics, Intelligent Data Pipelines, Data Democratization Frameworks.*

1. Introduction

Recent years have seen a proliferation of new products and services in the area of data engineering. New-age organizations are scaling their data needs with the adoption of multiple

Staff Engineer, phanishlakarasu@gmail.com, ORCID ID: 0009-0003-6095-7840

products, such as customer data tools, sales automation platforms, marketing stacks for analytics and marketing automation, next-generation cloud-native databases, cloud platforms for analytics, and numerous foundational technologies that enable the deployment of scalable data engineering and processing pipelines within hours. The diverse set of use cases is forcing for transformation of the previous way of doing data engineering that relied on defined hand-rolled processes, carried out by skilled data engineers and manually monitored for reruns in cases of failure or substandard completion. Cloud-native data engineering is primarily about the automation of these traditional processes by reducing data silos and breaking down the barriers between traditional tools and data science, automating the building of data flows on infrastructure provided by the service providers, and using the industry standards for discovery, accessibility, and security.

Since the advent of transformative technologies, data is mostly dumped in a centralized or semi-centralized architecture layer where people have the option to enrich plain legacy data with additional non-legacy and external data and keep it posted for the uses in need.

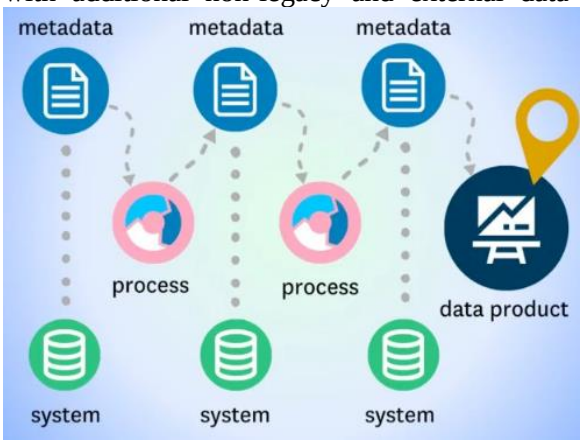

Fig 1 : Data Lineage

While this democratization of access, driven by the paradigm of self-service analytics, has augmented the strategy of making insightful analytics available to a wider audience of business decision-makers, opening up raw data is a process that has security, trust, and liability implications. Data quality and suitability for identified use cases become even more paramount to avoid spurious business decisions founded on duplicate, inconsistent, or just plain erroneous data. Thus, amidst this explosion, there emerges the need for high-quality foundation data pipelines that address key facets of enterprise data management like data quality, data integration, data lineage, security, privacy, and ethical data monetization. The governance of such pipelines is what we refer to as DataOps.

1.1. Background and Significance of AI in Data Engineering

Data engineering, an essential yet often overlooked component of data-driven business work, focuses on preparing data for analysis and modeling and is often prohibitively expensive, fraught with challenges and errors, and becoming increasingly important for modern businesses. Companies with vast amounts of data usage face growing pressure to ensure the utmost quality, safety, and reliability of their data pipelines while minimizing overall operation costs. As analytical models increasingly become the centerpiece of modern business, misalignment of data in terms of quality, standards, testing, and best practices, can lead to wasted resources and increase time to market. The data engineering lifecycle can be decomposed into Data Quality Management, Data Lineage Management, and Data Transformation Management. Data Quality Management is responsible for ensuring that the

data being moved, integrated, and analyzed is accurate in terms of correctness, validity, completeness, and coherence.

Traditional Data Quality Management processes rely on costly manual coding efforts to define the data quality constraints and leverage them for monitoring data quality issues frequently seen across disparate and dynamic cloud infrastructure stacks. Data Lineage Management focuses on automatically capturing the data flow, its dependencies, and the transformations applied to the data. Data Transformation Management focuses on the multitude of data integration patterns and efficiently observing, detecting, and proposing solutions for user requirements for the transformation of data across disparate databases, schemas, tables, and structured and unstructured data.

There are four motivations for deploying AI and ML in solving these domain problems. Data-driven solutions have several advantages over conventional solutions, such as accuracy, robustness to the size and diversity of the data, the use of semi-supervised and few-shot learning tools, and the potential for continual learning, where models are deployed in production systems and can be incrementally adaptive to changes in the systems over time. However, the contention on processing time and modeling resource constraints requires domain-specific modeling techniques combined with domain knowledge to improve efficiency and accuracy.

2. The Role of AI in Data Engineering

Today we are witnessing a rapid advancement of artificial intelligence (AI) tools that use enormous amounts of data and compute. Such AI tools are revolutionizing many industries by affecting both the market and the costs of research and development. AI-generated content can allow companies to cut out a substantial portion of their workforce while doing repeated high-level cognitive functions, e.g., generating textual answers, transcribing a video, or finding and highlighting useful content for a specific objective among a sea of papers. Conversely, many AI companies became the new arms of software companies by embedding state-of-the-art AI into their software offering. Examples include data observability, large language model-based application development, customer service, communication, and so forth.

In the above context of major advancements in AI, which were enabled by large-scale R&D and infrastructural efforts, a vital question arises: can AI assist (or disrupt) data engineering? If so, how? Is AI an additional (or alternative) tool for data engineering or does it become an embedded component in the essence of data engineering? The dramatic increase in data volume, variety, and velocity implies that manual data engineering at scale is neither possible nor desirable. To neither overburden data engineers nor compromise on agility, innovation, quality, and availability, data engineering can leverage advanced AI techniques that have recently become available and become embedded as components in the data engineering lifecycle.

Equation 1 : Automated Data Quality Scoring (DQS):

$$Q_d = \alpha \cdot C + \beta \cdot A + \gamma \cdot V$$

where:

- Q_d = Data quality score
- C = Completeness ratio
- A = Accuracy estimation
- V = Validity score (schema/type conformity)
- α, β, γ = AI-learned weights (sum to 1)

2.1. Overview of AI Techniques

From its earliest roots, AI was deeply entrenched in computer science and pursued primarily through algorithmic methods. Some of these early techniques, such as symbolic systems and

expert systems, focused on human reasoning but limited the scope of automation. During the era of big data, activities gained more momentum, with search and information retrieval engaging and attracting a more diverse audience, and machine learning in various guises focused on the practical task of replicating smart human perception-driven research into new data-driven methods, such as deep learning, that took advantage of the accelerating acceleration in computation and available data. Machine learning, especially deep learning, remains the most exciting area of AI, propelling the field into application areas, such as computer vision, natural language understanding, speech recognition, robotics, and game playing. AI's most mainstream functions are recommendations and personalization, mapping, and search. Since the emergence of deep learning, AI has also encompassed foundational models for vision, language, and multi-modal capabilities, such as neural networks that are trained on extensive amounts of unannotated data, with techniques, such as self-supervision, and are then finetuned for specific tasks with smaller amounts of labeled data. As deep neural networks become more ubiquitous, new insights regarding convergence times have been revealed and connections, such as provable generalization in stochastic models based on convex losses, have emerged. But many open problems remain, such as approximating low-dimensional structures and representations in high-dimensional data spaces, scaling to large amounts of unlabeled data, and algorithmic bias, questions about these properties notably date back to the initial explorations of statistical learning theory more than thirty years ago. Other areas of foundational AI, such as knowledge-based reasoning and symbolic systems, have received renewed attention and focus.

2.2. Machine Learning Applications in Data Engineering

Analytics and machine learning (ML) have increasingly been adopted within Data Engineering, such as for guiding data gathering, enrichment, and transformation, to support analytics systems and data-heavy applications. While the goals for using ML may be to optimize BI dashboards, streamline ML training data preparation, or formally check and manage data quality to avoid costly analytics mistakes, we argue that these ML applications are a clear sub-domain of Data Engineering rather than a new domain in its own right. The focus of Analytics and ML within Data Engineering, in this view, is to enhance data readiness for consumption and use rather than to directly apply Intelligence Engineering techniques to drive computing processes, even if those processes also lay some of the foundations for Intelligence Engineering life-cycle pipelines.

The goal of Intelligence Engineering, in contrast to Intelligence Engineering, is to make analytics-ready data artifacts that serve as ML input targets more readily available and understandable to Data Engineers as well as to analytics and application practitioners, and their specific technologies and methods. This requires knowing not only to which applications and tools one’s ML projects are exposed but also the ML workflow steps followed. Guided by these ML readiness and readiness enhancement objectives and extending the Toolkit use example described earlier, in this chapter, we summarize various recent methods and projects that have applied specialization, NLP, machine learning, modeling, and inference-based explanations to four key Data Engineering tasks: analytics process data forensic technology artifact discovery, qualification, and scaling; data augmentation; data storage format optimization; and ML readiness errors uncovering. We conclude with a discussion of open issues and directions and challenges ahead.

3. Data Quality Automation

Applying AI for automation of the Data Quality (DQ) pipeline is one of the key areas to achieve the objective of DQ at scale. DQ automation can enable organizations to narrow the gap between business expectations of DQ and what system architectures, processes, and practices deliver in reality, at an unprecedented scale. In this chapter, we describe research and practice

in the automation of DQ tasks. This DQ automation work includes DQ metric definition and discovery, DQ assessment, DQ data curation, DQ augmentation, and DQ quality data ecosystems. The outcome aims to provide organizations with enterprise-quality DQ solutions built upon open-source frameworks and services for DQ-related workflow streams.

DQ metrics are critical to DQ automation technologies and practice. The work starts with our observation on the increasing importance of DQ metrics, apparent in almost all major software DQ vendors. This increase promotes demand for automation, yet is still a labor-intensive, fallback process based upon heuristics and experts. It is still very limited because it is non-automated, difficult, and expensive to do. Creating a solid base, or foundation, of DQ metrics would help all organizations to make DQ automation possible and have better and more accurate data analytics and decision-making outcomes. At scale, it would be impossible to rely on expert knowledge in an organization to build the entire DQ metric catalog for their DQ operations. Methods to solve this task for organizations with disparate DQ practices. By good fortune, AI methods have recently proven to succeed in tackling related challenges. Therefore, research must include AI-based solutions tackling DQ metrics as needed by organizations. Solving this will enable DQ data practices, enterprises, and ecosystems to flourish at scale.

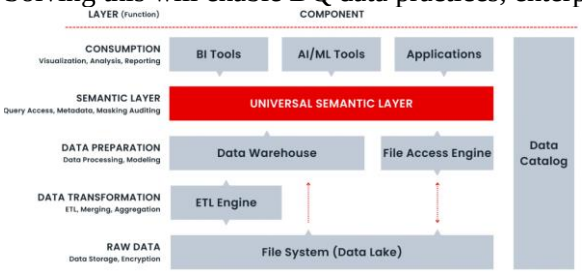

Fig 2 : Data Quality Automation

3.1. Defining Data Quality Metrics

Data quality work has a relatively long history, with such works generally predating the wider data engineering and engineering observability area of concern. Many definitions of data quality are possible, and many dimensions of quality metrics are defined; however, a commonly cited definition is the "data quality dimensions" definition—where four intrinsic data quality dimensions and six contextual dimensions are defined. The four intrinsic dimensions are accuracy, completeness, consistency, and timeliness. The contextual dimensions are: (1) congruence with intended uses, (2) quality of fit, or is the data appropriate for the task at hand, (3) accessibility, where is data available when needed (it, conceptually, could be of high quality yet unreachable), (4) representational consistency, where is the data formatted, encoded, or contained such that it allows for ingestion and processing (for instance, an empty CSV file could be of high dimensional accuracy yet effectively unusable in a processing context), (5) interpretability, or availability of information on how the data is structured (a file with no headers or metadata could be difficult to process), and (6) systems orientation, where the quality of the data may depend heavily on the systems about that collect them.

Several empirical evidence and projects highlight challenges with specific data quality dimensions and models derived from the framework---for example, the importance of representational consistency, with realistic examples from finance on transactions that were mis-encoded using differing formats and randomness for certain fields. Due to this, the work in this book offers an exploratory design lens, which aims to allow for efficient and deep exploration of what data qualities are important for specific data in a specific scientific or business context during production while understanding decisions made in prior systems-oriented approaches.

3.2. AI Approaches to Data Quality Assessment

Automating the entire spectrum of data quality assessment is a complex, resource-intensive endeavor that may not have a perfect answer. In the wide variety of proposed solutions, we highlight a couple of representative approaches we found in the literature. We also extend the discussion accounts of descriptive statistics approaches. We point out that the current trend in DataOps tools is to rely on relatively naive heuristics and automatic metadata collection and inference techniques for Data Quality Improvement. The definition of more advanced but accurate heuristics is a critical step for building more mature tools. Any qualitative risk assessment can exploit automatically generated tags and, when present, metadata attributes such as schema description, creativity level, content type, key profile, taxonomical and semantic descriptions, authorship and publishing dates, and so on. The heuristics cannot be other than heuristic: they will yield the best results for a niche or filtering but are much needed for creating a score – for ranking.

We divide the existing solutions into three classes: descriptive statistics, anomaly detection, and natural language processing. The first class of techniques for supporting data assessment is classical descriptive statistics/meta-data applications for data. Attempts to summarize the largest data properties and identify those that are naturally unlike but we track those that require an exhaustive description of all modes of creation across the user-model tuples. The first attempts to describe user-created object properties. They describe them as news artifacts and analyze their observable statistical characteristics, providing novel machine-assisted heuristics to assist News-Event detection, selection, and scoring using observed User-Content interaction statistics. Other considerations guide the choice of statistical probability required for invalidating the null hypothesis of a pair of event distributions being similar.

3.3. Case Studies in Data Quality Improvement

The below case studies illustrate how data quality automation creates business value. It's important to note that the data quality problems discussed in these case studies are those that proactively compromise the quality of a data product over its lifetime. These types of issues transfer between domains and technology stacks, especially in the spirit of data mesh architecture where we promote domains managing their data rather than a central IT group. Less frequent are insights into issues with data value assigned to any data product during its construction. The issues in this category create challenges for a data product consumer but are often an afterthought by the data product provider. There can be some additional domain knowledge required to identify why a data product may not meet expectations depending on consumer intent, and thus domain mesh principles would allow domain-level adjustments to this infrequency. Better algorithms would also lead to an increase in data value from a provider perspective and reduce conflicts in reasoning top-down and bottom-up.

There is a limit to the degree of automation and AI involvement in the monitoring and reporting phase which is not a semantic translation. The two case studies would focus on two problem areas in data: data value and business model intents and the trustworthiness of that data. These lie at the crossroads between technology investments and effective governance. The project is attempting to assign data bug counts at scale against the data sets and map those against the guidelines defined. In the model, a data bug is any issue that would not meet a data product consumer's needs consistent with the business model intent for that level of data abstraction.

4. Data Lineage Tracking

Data collection and resources within companies are today subjected to numerous daily changes, the result of the business's dynamic nature and the need to satisfy the company's legacy requirements. Volumes of data recorded for several years create two types of pressure on internal development teams; the first is the need to ensure that access mechanisms and data ingestion/processing pipelines work correctly and, consequently, are not affected by any

programmatic issues that have occurred over time (abandoned data structures, column name changes, data formats, etc.). Data architectures must be designed to guarantee redundancy, data availability, and cost optimization. The control of the automatic maintenance of these systems over time (data quality, transformation, etc.) is what characterizes an organization's data integrity at a particular time. The second is to have a business decision model that can be easily analyzed to create new use cases and new architectures, meeting modern advanced analytics requirements and respecting the historical requisites.

The availability of graphical representations of the data pipelines reveals the inherent connections of data warehouses and lakes and allows data engineers to be alerted about foreseeable errors and regularities along the pipelines and automatically perform optimized maintenance interventions, taking into account not only processes but also predictive modeling and data visualization processes. For these reasons, the tracking of data lineage, which allows us to represent these processes graphically, has become greatly relevant in its design phase. For services such as data ingestion and transformation, numerous techniques can be adopted to automate the activity of creating simplified graphical models that can be used to represent the analytical architecture of the system.

4.1. Importance of Data Lineage

Data engineering is at the heart of any AI-driven business, making data accessible and creating the systems for answering the questions that data science will explore. Data lineage is key to effective data engineering because questions about data itself — how it is ingested, transformed, and moved about — are the hardest, get the least attention, and critically depend on asking the right questions. Though the analysis of data provenance has been heavily studied in both security and database literature, little of that research has made it into the mainline data engineering pipeline, in part due to cultural reasons of service-oriented metrics that are much more frequently kept in mind. It seems that more often than not, ETL tools are turn-directed rather than data-directed. The introduction of data mesh makes data more focused and invests extra resources to get data right, which makes the gap between the service and data-oriented views apparent. One of the original ideas of data mesh was to have more data product thinking in the organization, and product thinking always demands awareness of the entire product lifecycle, especially data quality.

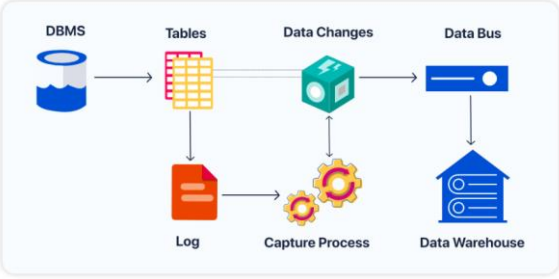

Fig 3 : Data Lineage tools and its Best Practice

A data product always implicitly exists upstream. All users of your output need to know how the product was built to install the implicit contract hedge against a data monstrosity. Data preparation and processing is a collaborative effort between multiple teams, which raises issues of trust and accountability if the resulting product is low quality. Data lineages tell a story of how data were created and modified within the stages of a pipeline, providing context that data customers need to evaluate whether the data can be trusted. A lineage is a network showing the relationship between the various data products, the lines being defined by their source and

destination information. The nodes are defined by fields in one or more datasets that share the same semantic meaning, and the content of the datasets is their data product value.

4.2. AI Techniques for Lineage Analysis

Lineage analysis involves gathering information about data states and changes at the appropriate granularity and defining properties to track their evolution. It has been studied for several decades in several communities, including databases, machine learning, and data visualization. Nonetheless, the task is still largely performed manually, even for critical uses like generating regulatory reports, and off-the-shelf solutions only scratch the surface. Herein, we cover ML and DL techniques for automating lineage analysis.

Security, privacy, and compliance dictate that the data available for inference can likely be very sparse and lacking information. An effective ML solution to the lineage inference task is using generative modeling approaches. Generative-based approaches have been demonstrated empirically to yield results comparable to or better than supervised and other unsupervised alternatives. In the scenario with well-annotated inputs, additional pre-training, sampling, and reinforcement fine-tuning cycles can yield improved performance metrics, over end-to-end solutions.

Moreover, auxiliary pre-trained sequence models trained overabundant source-label aligned text pairs can be applied to lineage prediction with remarkable results in a few examples. Despite previous evidence in other vision and text tasks, transfer learning from a single, vertical lineage-based prediction across similar data types, achieved with classifier-decoder architectures, still trade-offs computational resources and training cycles against classification performance. Also endowed with extensive knowledge about the data collected, poorly performing ML-based line disturbance inference on the vertical trigger could benefit substantially from domain transfer-based pipelines trained over previous procedure-label predictions. The results would be indirectly transferable to poorly performing unlabeled queries through joint classification across data types.

4.3. Tools for Data Lineage Visualization

Data analysis frequently requires that we find relationships among many datasets and tables. Documenting how a dataset was obtained or generated is a form of data lineage; it is an important property of the dataset and is needed to help other analysts understand the semantics of a dataset for correct use. Documenting data transformations is very tedious and difficult work. Several packages can track data lineage provided that we are working with supported packages.

Data lineage refers to the end-to-end process that describes the flow of data through various stages, tools, tables, datasets, queries, etc., by which data flows from its source to its destination. Lineage acts as an audit log for data processes operating in production, describing what happened to the data, when it happened, and who carried out the process. Keeping track of data flow is critical for organizations with strict compliance regulations. The tracking process is usually highly manual. In the case where sensitive information flows through various tools, it is important to track that flow and ensure that proper actions are taken.

A third-party integration can be integrated into several popular data tools, to make the tracking process automatic, making the information about the pipeline and lineage available in one single source of truth. Today’s tools allow for visual exploration to help overcome the difficulty of navigating very large data ecosystems. Visualization of the data ecosystem with a tool can answer questions like which are the most commonly used tables, how are they related to each other, are they related to any application workflows, what schemas do they have, what are the upstream and downstream dependencies, and so on. Tools can connect to your popular data tools to provide automatic lineage tracking and enable visualization.

5. Data Transformation Automation

Automation across the data engineering lifecycle is still nascent. While manifesting across diverse facets of a cloud data platform, one of the biggest opportunities for transformation is in its core component—data transformation. This is ironic considering that data transformation has been a fundamental principle of data engineering since its inception. Data transformation automation can have a significant impact on time and effort savings, increasing the percentage of time data engineers spend on high-level design versus tactical implementation. Even more so than data integration, data transformation acts as a core technology within a cloud data platform. All data pipelines leverage these transformation capabilities to create downstream data assets in different zones of the data architecture and with different personas leveraging such assets in mind. For business users or analysts looking at insights, this means cleaning, conforming, aggregating, summarizing, and modeling. For data scientists, this means performing point transformations to get data in a location and state that is true to their analysis. Often, though, the biggest challenge faced by these personas is not the pipeline design, but ensuring the right transformations exist; transformations that enforce appropriate data quality or lineage checks. Despite the ubiquity, enterprise transformation efforts are still stuck in a vicious cycle of development and validation, requiring inefficient manual processes across the lifecycle. Not only did such transformations require code to be written from scratch for every conceivable variation of a pipeline, but the data engineering team had to spend considerable effort validating and redeploying this code for every pipeline run. Predicting commonly executed transformations faster and more autonomously can speed up data operations significantly.

Equation 2 : Probabilistic Lineage Confidence (PLC):

$$L_{ij} = P(D_i \rightarrow D_j) = \frac{f(D_i, D_j)}{\sum_k f(D_k, D_j)}$$

L_{ij} = Confidence that dataset D_j originated from D_i

$f(D_i, D_j)$ = AI-inferred transformation relevance

$\sum_k f(D_k, D_j)$ = Normalization across candidate sources

5.1. Understanding Data Transformation Processes

The growing emphasis on data and its conversion into insights for the development of innovative products and services has triggered an unprecedented increase in the volume, variety, and velocity of data generated, resulting in a data deluge. Organizations across verticals are looking to prepare massive amounts of structured and unstructured data, generated both internally and externally, for machine learning and analytics. Ubiquitous access to cloud-scale data storage, compute power, and machine learning and analytics frameworks have enabled organizations to experiment, develop, and deploy a range of innovative data products. While some organization-specific data has continued to reside on-prem partly due to compliance and security reasons, the hybrid-cloud and multi-cloud architectures have become popular over the years.

The multi-tenant nature of cloud-scale data platforms has prompted organizations to adopt cloud data platforms that specialize in managing cost, scalability, compliance, and security. Most traditional solutions for data lifecycle management have evolved to operate on cloud data platforms involving multiple actor stakeholders. Such data lifecycle management is performed programmatically involving data ingest, optimization, preparation, storage, monitoring, and retirement. Data transformation is a critical step in the data analytics lifecycle, allowing the transformation and conversion of raw data into meaningful information. Properly designed data transformations are essential and often take a longer time with unacceptable costs, involving

manual coding, testing, and scheduling. The advent of cloud-scale data platforms has resulted in a plethora of transformation engines.

5.2. AI-Driven Transformation Frameworks

Architectures for AI-driven transformation fall into two categories: Compiler-like architectures and Executable-template architectures. Compiler-like architectures generate executables that implement data transformation pipelines. These architectures are closer to the role of a code generator generating a compilation unit. The generated transformation is a sequence of simple transformations that are each one-to-one and order-preserving for the input and output. The templates become source code for a programming language. These architectures are language-based because the template is similar to a program that implements the data transformation.

1. Transfomer and CodeGen use pre-trained LLMs and simple sets of transformations created manually by humans from user inputs. The transformation for converting a user query into SQL code is an example. Model transformation is created and then modified to transform user CTAs specifying specific input-output data format domains into functional code. It involves a language generation step but the output is a prototype that we use to generate an equivalent database function from user queries.
2. LLM2 Function takes the output of a conversational chatbot on top of a pre-trained large language model and converts it to a functional database function. This database function can be used to interact with data in a database.

Executable-template architectures generate templates that contain a small AI augmentation code that in turn expands the template to an actual executable. The advantage of template generation is that the template can efficiently generate multiple different executables depending on the size or characteristics of the data input during its execution. There are a few frameworks that take this approach:

1. APT generates executable templates for mapping data formats and data distributions from one type to another for one- or two-dimensional grid-like data such as images or tables. It uses functions creating building blocks used for template generation and uses example pairs of input data and corresponding outputs as training data.
2. Fused-LLM generates data-processing pipelines as templates that process data from relational databases as SQL tables and return the same SQL tables after modifications. The input is SQL queries describing the transformation of the data. These templates use the power of LLMs to generate SQL queries but require a human to design the pipeline to execute the template.

5.3. Case Studies in Automated Data Transformation

Numerous private and public sector organizations have had success implementing several of these data transformation automation methods. For instance, a fully automated Machine Learning Operations (MLOps) pipeline was introduced to deploy, validate, and monitor predictive ML models. Our case study examined all phases of data transformation and found AI to be central to the deployment of the MLOps pipeline. The core contribution of the case study is in how such MLOps pipelines enable organizations to deploy predictive ML models for an increasing number of applications in the financial services sector, such as risk score prediction, loan origination risk assessment, credit fraud detection, price prediction for support and settlement services, loan serviceability estimation, revenue prediction, and wholesale loan aggregation management. The key part of this case study is the cloud deployment of the MLOps pipeline, as a hosted product offering, which allows organizations to deploy predictive ML models for their clients in an increasingly efficient manner.

Moreover, the literature on data transformation automation also features research on thousands of AI/ML tools, infrastructure, and technology stacks. While many of those works recommend solutions to industry problems involving data quality and other data transformation steps, we

complement that body of work with a case study focusing on the deployment of transformational AI tools by an example business. In particular, our case is that of a digital consultancy firm who have been deploying numerous transformational AI tools for enterprise clients in the last couple of years as part of data science initiatives across industries in several continents. What is particularly interesting, in this context, is that those transformational AI/ML tools and technology stacks are deployed not for internal purposes, but as part of enterprise clients' data science initiatives, to allow disparate and decentralized departments to undertake enterprise AI/ML strategies autonomously.

6. Cloud-Scale Platforms for Data Engineering

1. Overview of Cloud Data Engineering Platforms

The previously discussed foundational AI solutions are deployed on top of data processing stacks and programs that turn the mass of incoherent, noisy data descriptive of most human processes into formats suitable for posterior AI analysis and promulgation. These programs begin by scouting the data for new phenomena, discoverable features, and mappings to user needs based on inceptive, unsupervised analysis and processing of the data. Discoveries are then managed from a data pipeline supply chain perspective, automating the update of their profiles in data lakes, while the detected changes are transformed into higher-level data products. Finally, these products become formalized, endorsed, and verified through successive approximation pipelines, increasingly automated by the feedback from validation using automatable QA and lineaging.

2. Comparative Analysis of Major Cloud Providers

The building blocks offered by the leaders in data cloud architecture differ rather significantly in their scope, granularity, native variety of definitions and operations, and software-defined pipeline control and monetization capacity. One provider started its services with the most basic building blocks, scalable persistent storage, low tables querying, and tools for AI. It rapidly expanded it into a considerable offering, where the joints are linked with standardized APIs and services for orchestrating extra functionality such as message queuing and event handling, and service deployment and monitoring. This provider does support functions to various extents using external commercial solutions and custom offerings.

3. Scalability Challenges and Solutions

While the cloud removes the burden of the physical setup of production systems and their GUI deployment, pipeline-scale industrial production at scale imposes its burdens. Designing and deploying the workflows, monitoring, diagnosing, and repairing production issues, and essentially running extremely complex data automation software designed to automate the automation are all challenging, error-prone, and often require extensive engineering toolkits, as well as human cognition inputs for tool feedback and errors.

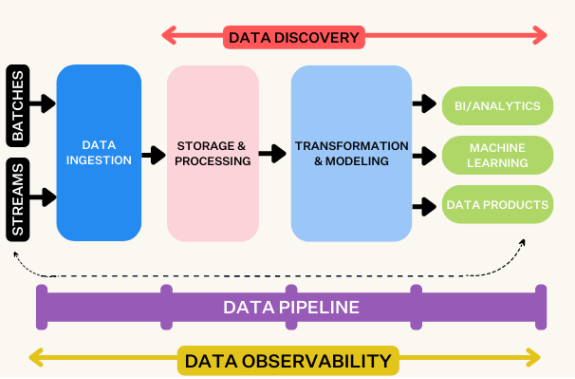

Fig 4 : A Data Platform

6.1. Overview of Cloud Data Engineering Platforms

Data is essential to every aspect of business today. Organizations have built systems that store large amounts of data and allow employees to access and operate on that data using software applications. Over the last two decades, as data volume has increased from gigabytes to terabytes and petabytes, the operational costs of on-premises data systems have increased exponentially. In contrast, the cloud has created new data infrastructures that distribute and share resources across many users, dropping the cost of infrastructure to the point where companies spend only a fraction of their employee time developing new software to extract value from data.

Cloud companies offer these cloud data systems such as cloud storage and cloud big data platforms that create an explosion in the amount of data that is captured and stored by organizations. Instead of costing millions and requiring months to deploy, these platforms allow companies to manage all of their data for a fraction of what it used to cost, allowing any company to store all of their data at less cost than it takes to store only some of it. Putting all of the data into one place allows companies to overcome the data silo problem. Cloud services then allow users to build big data pipelines that move the data from its origin systems into a centralized data lake or warehouse and then build data products or internal reports that other employees can easily access.

Data pipelines connect many different services that specialize in processing data and running machine learning models. Used in conjunction with machine learning APIs or services, data pipelines enable organizations to operate on large data volumes continually. Because each of these steps is built by connecting separate data services, adding new services or modifying existing services becomes easier, allowing companies to innovate and react to business changes faster than with traditional software development methods.

6.2. Comparative Analysis of Major Cloud Providers

For data engineers, the cloud providers offer comparable tooling and services. The largest cloud providers have similar core data engineering services. These services include managed computing, managed data lakes, managed distributed computing, data warehouse, managed ETL and ELT, managed data orchestration, notebook services for data engineering and data science, and private cloud options. In terms of general computing infrastructure, all of the leading cloud platforms provide similar capabilities such as virtual machines with suspend/resume, auto-scaling, implementation of firewalls, and integration with virtualization systems.

While the major cloud providers and their offerings are similar, they do have unique features, strengths, weaknesses, and proprietary services. This is especially true for the emerging components of cloud data engineering including functionality for ML-integrated data operations. Understanding these differences and planning their projects accordingly may lessen rework and support costs. The growing popularity of cloud providers has begun to intrude on the previously held dominance of on-premise big data capabilities in organizations. There are many reasons, both technical and non-technical, for adopting one cloud provider over the others. These reasons stem from unique features that address customer needs and customer ecosystem synergies. Major reasons to adopt a leading provider are organizational inertia, cloud vendor ecosystem synergies, and unique service features and capabilities that address customer needs and technical problems.

6.3. Scalability Challenges and Solutions

To scale effectively, data engineering workloads require workloads with high data-level parallelism, update workloads that work well with key-partitioned indexes, and programmable batch workloads that can accommodate per-row data or service level requirements. Other than the common requirement for extremely low latency, data engineering workloads do not usually have the frequently cited requirements for thousands of very high-compute nodes that are

required for training large deep learning or natural language models or for training various other sophisticated models using similar amounts of compute. Although HPC started gaining prominence with the use of models as feature sources for other models, the core capabilities of data engineering workloads are vast data processing, feature transformation and update, data quality, data augmentation, data hosting, and lineage.

Motivated by their volume-intensiveness, the cloud majors and their other hyperscaler competitors are currently focused on batch mode workloads equivalent to the ETL pipelines of old. However, some enterprises are running into scaling issues with frequent queries on very wide tables using older designs that rely on scanning hundreds of raw data files or parquet-glue file pairs on an event-triggered basis, followed by diverse row line-based transformations on very sparse like-event-distributed days. These usually involve several data vendor sources at once for special lunch or sale days or similar periods in the private user data sense. In response, some cloud majors are adding on-demand hourly, daily, or weekly transforms based on the external source federated query, which are then executed to detect less sparse updates to the target tables via the table store's changelog and on their terms.

7. Integrating AI into Existing Data Pipelines

Data pipelines are the lifeblood of the data engineering world, moving data from its source to its final destination for business intelligence or machine learning applications. These pipelines are already highly-extensible and easily automated tools, written in Python or similar languages using a library or framework. They are composed of units referred to as operators (or tasks), which each accomplish a task in the chain of data transformation. Integrating AI systems into data pipelines closely mirrors the process of building data pipelines. The goal is to take advantage of existing product-ionized resources while minimizing the expense of building these additional data flows.

Integrating AI into existing data pipelines has several advantages. Each operator runs often, commanding many resources in the aggregate to run the brainless data transformation task, but the steps are small and the output looms large in impact. AI systems have many features in common with the existing processes that populate pipelines with cleaned data. Often, the inputs and outputs are the same or similar: popular generative AI systems are rarely used in isolation, and often run as layered stacks happy to take requests or serve as the output when the code is triggered. Pretty much all AI models take as input a batch of bytes and output a batch of bytes, making it easy to plug them into existing workflows that rely on such processing. Like the data pipelines themselves, AI systems are designed to run and output chunks of transformation at different cadences.

7.1. Strategies for Integration

There are multiple ways to integrate AI, such as creating a new pipeline, layering on top, and creating a new part of an existing pipeline.

New AI Pipeline development: This is the simplest use-case where companies only need to create a new pipeline model to AI-enable one of their data quality, quality, transformation functions, etc. In this case, the company's existing pipeline is not working well or the part with the AI-enabled function would be too complex to wrap into a new part of an old pipeline.

This use-case is generally the most straightforward for implementation. Many companies first pilot-simple ML models to basically complete a key component of their existing pipeline on one of their critical accuracy or performance metrics. If the pilot phase is successful, the model gets several improvements to productionize it, and other companies then create external models.

Layering AI on Top: This is the second use-case where some part of the existing pipeline uses AI for the data-driven step. This use case is very popular with programmatic companies running their data pipelines as a service. Currently, the companies support hundreds of AI pipelines.

Their user's data requirements are usually completely undifferentiated. In this use case, we train a model to ego a model built in-house by the service provider on the common escape route metrics. The incentivized drivers for this use case are infrastructure issues and the flexibility of combining multiple different AI techniques.

7.2. Best Practices for AI Implementation

When it comes to implementing AI into existing data pipelines, there are a few important guidelines and principles that should be followed. The first, which moves us toward our goal of a correct integration without negatively impacting the quality of the existing solution, is to ensure that the new AI component does not degrade the integrity and quality of the overall pipeline. Using AI to speed up the data processing flow is one of the main objectives of the integration but it shouldn't come at the cost of lower data quality. Therefore, quality checks of the data throughout the ETL pipeline are extremely important as the AI system might not always deliver the best-tuned output data.

This concept also helps tune the AI models to ensure proper use of cost-benefit analysis of cloud environments. The cost of performing operations on the cloud is always proportional to the activity so it is best to enrich existing pipelines with quality-control points if they do not currently exist. This will help try to avoid performing very expensive processing tasks that are still not guaranteed to yield good quality results. Once the data is quality-checked, DataOps principles could also be used to replace the old manual processes of quality-checking with automated AI-based checks that quickly notify DataOps engineers of bottlenecks in the data pipelines. This could be done through a combination of better dashboards that promptly signal quality problems in the data or even by sending proactive notifications to an operational office of data problems. The idea behind such automated monitoring systems is to take standardized metrics inspection as well as problem detection and resolution out of the manual process that requires a lot of resources and time. Utilizing monitoring dashboards in the cloud will enable a dynamic and holistic monitoring of not only the internal KPIs of each single pipeline but also of the overall architecture KPIs.

7.3. Real-World Examples of Integration

Integrating AI into existing data platforms is not simply a vague research agenda; it is already mission-critical, production-volume work at many companies. Here, we give a small sample of examples to illuminate the vast scope of what AI integration means in the real world. As is the case for many advanced capabilities in the past, the first industry to adopt AI and machine learning capabilities are usually the innovators in advertising, especially search and social media. In the case of Generative AI, such companies had to respond not only to feedback from within their user base but also to the impact of free Generative AI tools offered by challengers. These solutions may involve not only adding new capabilities for text, image, and video generation but also providing added capabilities for quality control, such as for detecting AI-generated video deepfakes. Other rich opportunities are in text moderation or augmentation and spam detection, especially for internal tools, using large model embeddings for text similarity. Other rich opportunities also abound for companies that build product recommendation systems.

The financial industry is deploying advanced capabilities in real-time fraud detection and credit default prediction, and the healthcare industry is deploying novel NLP solutions for chatbots, customer relationship management, prior authorization, and medical billing. The manufacturing industry can reduce machinery downtime via AI-enhanced predictive maintenance models; the automotive industry predicts battery life more accurately with new models; logistics firms can optimize operations; and utilities are deploying AI solutions for predictive load modeling.

8. Ethical Considerations in AI-Driven Data Engineering

Artificial intelligence (AI) has emerged as a game changer in almost every aspect of the business and technological landscape, with AI-driven cloud-scale data platforms redefining data engineering. The convergence of cloud computing, big data, and AI is automating and democratizing data engineering, enabling businesses and users to scale their efforts and expedite their analytics aspirations. However, as with any technology breakthrough in a governed and regulated enterprise operating environment, the unintended consequences of this generation of new technology cannot be ignored. Indeed, matters concerning data privacy and security, model interpretability, and AI algorithmic bias threaten the ideal of tech valiantism and drive brave new conversations about dark tech. This chapter delves into ethical and governance questions surrounding AI-driven data engineering and raises questions concerning model accountability, privacy and security protections for personally identifiable information (PII) data, regulatory compliance, and cybersecurity concerns. We cover important considerations such as bias and fairness, privacy and data governance, algorithmic model accountability, cybersecurity, and regulatory compliance.

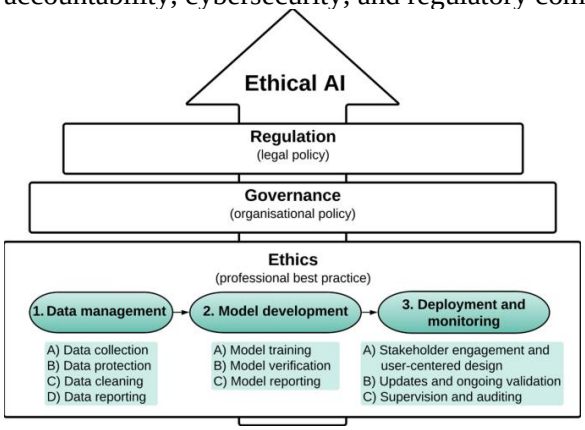

Fig 5 : Ethical Considerations in AI and Data Science

Data and digital privacy have been recent focal points of governments and industry regulators, as the digital revolution sparks angst about surveillance capitalism, especially the blurring of boundaries for data monetization and responsible data usage. This chapter concretizes the discussion on responsible data engineering by focusing on ethical considerations in the design of AIs for data engineering, digital platform governance and regulatory compliance, and cybersecurity. We explore the growing concerns regarding data privacy, security, and governance, and call for responsible data engineering as industry business firms collect, use, and curate data for their user base with ever-increasing laxness. We also raise questions about the increasing scrutiny by regulatory compliance bodies on issues within data engineering concerning PII usage, model interpretability, regulatory compliance, and algorithm accountability.

8.1. Bias in AI Algorithms

Algorithms take in various kinds of data and weight for varying degrees of importance in determining decisions. Algorithms are ultimately determined by the human beings who create them. Thus these advanced algorithms often simply reflect the biases of the creator or homogenous group of creators. Machine learning algorithms are trained on example-labeled datasets to predict outcomes from new data. As such, the quality of the training data, and how reflective of a diverse real-world population it is, determines the quality and validity of the algorithm's learned behaviors. AI can be biased if an unrepresentative group of labeled data has been assembled to train the model. For instance, if a dataset on facial recognition is created

using only white faces, a resulting trained AI finding will be biased against non-Caucasian faces. Thus there is a real problem with bias in labeled datasets.

Also, there is an even larger concern due to the possibility of feedback loops leading to self-reinforcing outcomes. For instance, if an AI is used to determine if an arrested suspect is likely to offend again, and if it bases that decision in part on whether the person has previous arrests for other offenses, and the person is arrested not because of living in a neighborhood where some individuals are gathering and selling illegal substances but rather because a police AI has promulgated that this is a high-crime area. Thus, the AI for arrestees being used for prediction helps determine the makeup of the inputs to subsequent predictions leading to an amplifying of outputs over time. In many applications such as crime or academic admissions, these factors can have major real-world implications creating moral hazards by unfairly benefitting or punishing individuals.

8.2. Privacy Concerns and Data Governance

Advancements in data engineering workloads and processes allow for the storage and processing of more sensitive information, such as unique User IDs or Personally Identifiable Information, at an unprecedented scale. Ethical dilemmas begin posing significant risks for individuals, data engineers, organizations, and society at large. Increased automation in data engineering may enable the algorithmic amplification of existing concerns and privacy issues. Examples include but are not limited to, tracking and storage of user behavior data and business use cases dependent on sensitive information. Organizations also seek to obtain specialized datasets for narrow domains. The lack of a clear framework for governance and principles around such datasets and development processes, however, raises privacy and security considerations. For example, recruitment data containing sensitive future information specific to individuals is stored and shared with third parties for engineering successful yet unethical predictive models, amplifying societal issues within certain domains. Although the ethics of building, deploying, and using such models is the responsibility of organizations, teams within organizations may also be compelled to build desirable products at the cost of both individuals and society.

Additional automation also entails more overall complexity within data pipelines. The mapping and understanding of dependencies across data tables within Data Lakes, Data Warehouses, Data Meshes, and Multimodal Platforms is not always straightforward. Sensitive information may be ingested and transformed through intermediate tables without the knowledge of users, teams, and organizations. Unintentional due diligence violations may be a possibility. Concerns stem from a lack of centralized governance around data assets that equips stakeholders with the capability to enforce rules, restrictions, and features, concerning privacy, over large-scale datasets, irrespective of the pipeline pathway.

8.3. Regulatory Compliance

Modern societies are experiencing historical shifts in how information is shared and democratized. Indeed, in the context of Western democracies, liberal values often oppose stronger government involvement in both private and common life, while in authoritarian regimes, this is generally not a problem. Normally, in democracies, economic freedom is paired with anti-monopoly legislation to prevent unregulated concentration of wealth, while companies, whose power and weight in society have been constantly increasing in the last 100 years, overpower governments in this realm. However, with the advent of AI and the attendant danger of abuse of power and sophistication of abuses, the global regulatory environment has been slowly but steadily changing.

Many countries have promulgated rules protecting both citizens and enterprises. The European Union, in this context, has taken the lead with the General Data Protection Regulation, which has become the reference framework for data protection. This regulation mandates that

organizations must protect the personal data and privacy of citizens for transactions that occur within member states. Other countries have followed suit, with legislation influenced or inspired by this regulation, like the Brazilian General Data Protection Law, the California Consumer Privacy Act, and several other current or proposed state laws. A distinguishing feature of this regulation is the empowerment of the user, who can grant companies consent for the collection and treatment of their data for a defined purpose, and have the right to revoke it afterward. It also foresees that if there is no consent granted by the user, data processing can happen in cases of legitimate interest and that anonymized data or that is stripped of the user's identifying traits, may be used. This is the reason why one of the greatest trends in the data business is the increased interest and investment in data anonymization, thanks to a growing market for data protection-compliant data harvesting and service vendors.

9. Future Trends in AI and Data Engineering

Developing an AI strategy for data engineering is a 20–30-year endeavor. As a striking but low-cost development, text generation will be much less expensive and pervasive. The possible consequences of very cheap text generation at internet-scale are difficult to assess: loss of unique human voices? Perpetuation of bias? Availability of phenomenal prompt engines capable of generating high-quality responses to complex prompts, making large language models equal to the task? I stated that while the future of cheap AI was a fantastic generator of basic thoughts, the next few years would see the extensive development of AI augmenting LLMs.

Based on predictions for advances in semantic multiplexing, with LLM augmentations, involved paid curators will build stacks of highly polished, premium-quality, conversational-language talent in every imaginable domain. These entities will be dynamic and driven, able to converge on requests for complex high-level thoughts in every imaginable domain, with the relatively low cost of prompting and polishing complex specialized text compared to giving equivalent instructions to present-future LLMs.

Consequently, I believe that the growth of the guided visual programming languages — GraphQL and DBT, at the near-end of the complexity-user frustration spectrum, Gen B, the outer-end stack at the far end, will drive a base-level formalization of the career data engineer, who plays a leading-edge role within and between data-product development teams. AI augmenters will not reduce the need for a data team of human-certified data professionals, but they may make the majority of redundant tasks performed by data engineers in product-focused teams less labor-intensive and more effective and enriched by data documentation.

9.1. Emerging Technologies

The world is on the verge of the next industrial revolution powered by the confluence of many different technologies: Infrastructure as code, Cloud reusable components, Cloud Artificial Intelligence services, Cloud Native Architecture, Digital twins, 5G Communications, Edge Computing, Blockchain, Distributed Cloud Data Platforms, Hybrid Cloud Architecture, Data Mesh, and of course, Artificial Intelligence.

In this chapter, we will elaborate more on how these different emerging technologies converge to create the AI-Driven Intelligent Data Engineering model. The advent of the Cloud has opened up the door to reusable components for most of the traditional enterprise IT functions and cloud IT vendors have made Cloud reusable components the centerpiece of their Cloud platforms, creating an explosion of solutions, services, and applications, most of the functionality built by combining the reusable components in different ways. Data Engineering and Data Management are no exception and this has opened the door to many cloud investment firms entering the data engineering market with industry and vertical-specific Data Analytics SaaS Applications. It is predicted that there will be thousands of industry and function-specific

Data Analytics SaaS Applications built by marrying cloud reusable components with the industry and function-specific use case templates.

Equation 3 : Dynamic Transformation Latency (DTL):

$$T_l = \frac{\sum_{i=1}^{n} \theta_i \cdot \lambda_i}{R}$$

where:

- T_l = Average latency of data transformation
- θ_i = Complexity weight of transformation i
- λ_i = Execution time of transformation i
- R = Total number of records processed

9.2. Predictions for the Next Decade

Data quality enforcement will become a core investment area for most cloud data platforms. The market will become increasingly competitive, with large incumbents and innovative new players alike needing to invest heavily to distinguish their offerings. AI-baked automation features will therefore become common across all market segments. Budget-conscious buyers will expect to see pilot projects in near-real-time, or they will look at leisure-scanning new devices from an array of vendors. Business tracking customers will demand higher-fidelity auditing interfaces, and demonstrating value will become paramount, as lower-fidelity "easy" piloting and demonstrations become table stakes.

The emergence of new usage and delivery paradigms such as Data Mesh, Data Fabric, and Data Hub will change how data knowledge workers think about data processing. Whether these frameworks are used to champion the movement of data knowledge work from centralized data teams into lines of business, or instead the intelligent brokering of multiple logical sources into centralized data views will frame decisions in tools and processes for coming years. Evolution in themes like data democratization will help facilitate this thought leadership change paradigm around how data becomes a product. Since many factors driven around centralized governance parceling out data to functional category-primed data customers are social and cultural, the degree to which emerging technology companies can influence these decisions through improved tooling will be variable.

Data engineering automation, oriented around the combination of a smaller set of reusable data assets into higher-order data products, representing autorouting the path of least resistance through a semantically enriched data funnel, will be a rapidly maturing technology area. The rapid accumulation of all data at lower cost, questioning the need for semantic use and design models to optimize efforts filtering into the best business decisions is coming to an end. Diverse forms of understanding- and workflow-bridging models will be needed, stimulating demand across a larger array of capabilities.

9.3. The Role of AI in Data Democratization

Data democratization democratizes enterprise data across a wider potential audience. More corporates and start-ups than ever want to harness their data in pursuit of a competitive advantage. This represents a powerful drive to radically transform data processes, turning them from something only specialists can do to what more people do more easily every day. While the term itself is more than a decade old, data democratization is now becoming real and tangible, especially through the lens of self-service analytics and data discovery tools. AI also plays a key role in enabling these goals, from automatically testing and cleansing data in the cloud pipeline to automatically recommending the key results to particular classes of business questions and users.

As businesses gather data and seek evidence to support their strategic decisions, a self-service reporting environment that enables users to perform analytics tasks without waiting for help

from specialized analysts is viewed as a necessity. Economic realities reinforce this perspective: surveys consistently indicate that business analysts work more than half of the time trying to compile or prepare data; corporations do not have budgets or need for armies of BI experts; rather they need to empower thousands of frontline workers. To be empowered, users need to feel risk-free to experiment and ask unexpected questions, while at the same time being confident that the guidance and policies provided by the data teams in terms of high-level dashboarding, reference reports, data catalogs, and cloud data pipelines are reinforced and enhanced by each individual's exploration.

10. Conclusion

The cloud and Big Data have opened a new frontier for artificial intelligence. As organizations look to make the most of the firehose of information within their data lakes, AI-driven capabilities are automating mundane, manual tasks, freeing up people to do more interesting, strategic work. For data consumers, this means improved data ecosystems, filled with quality data they can use with confidence, and less waiting for data engineers and analysts to perform simple data transformations. For data engineers and analysts, this means smart tools that accelerate the delivery of business insights by augmenting human efforts and automating the final mile of building and deploying data pipelines for industrial users of AI.

The tension between automation and oversights remains. Data quality, lineage, and transformation are areas with very high levels of human oversight for good reason. The stakes are high, and automated systems have a long way to go before they can be trusted entirely to take over from humans in these areas. However, the move to large-scale cloud architecture has also driven the development of new AI techniques, such as few-shot learning and natural language understanding, to help address long-standing healthcare analytics challenges. As a result, data engineers and analysts can use AI tools to scale their efforts and drive more value, avoiding the grind of enterprise data engineering and analytics work for the sake of it. With this new era of data ecosystem tooling built atop cloud infrastructure, the role of the data analyst and engineer can slowly shift from back-office construct to mission-critical brokers of organizational success. AI has the potential to be the glue that enables this shift while remaining conscious of human-specialized areas of expertise to bring the best of both worlds together in a unified DataOps cycle of success.

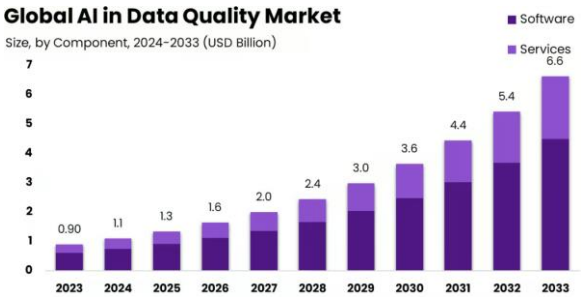

Fig 6 : AI in Data Quality Market Size

10.1. Final Thoughts and Key Takeaways

We hope you take away the following key lessons from this book. First, ML is key for successful data platform automation at scale. It must be the starting point to make data platforms practical for both data consumers and data systems engineers. As the amount and scale of data grows, it simply becomes infeasible to have humans in the loop to ensure data is always consumable and ready for analytical use. Second, data quality has to be defined in a consumable context. Many organizations approach data quality checking as a purely technical

engineering problem, applied to unconsumed data, leading to both irrelevant failure alerts for engineers and focused avoidance of alerts or metrics by consumers. Usually, the person requesting quality checking of data is the consumer of the data, at least indirectly, not the engineer ingesting the data. Third, data quality checking must be aware of the analytics tools or systems being used to analyze the data. Quality checks for data using batch analysis are not the same as data using streaming or batch analysis with SQL systems or NoSQL systems. Fourth, as with all ML tools and processes, data quality checking systems and automation will not be perfect. However, with flexible tolerances and quality enrichments, they can minimize costs and maximize usefulness over the long term.

Fifth, tools to manage and aid in lineage creation are required for a variety of reasons, but on top of compliance and assisting in analytical data review, they assist in defining data quality checks quickly and triggering them on relevant consumers. Sixth, data operations tooling for data transformation pipelines should aim to be as 'normal' a coding interface as possible, building on models from IDEs and software engineering processes while augmenting and monitoring with ML tools.

11. References

[1] Vamsee Pamisetty, Lahari Pandiri, Sneha Singireddy, Venkata Narasareddy Annapareddy, Harish Kumar Sriram. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Migration Letters, 19(S5), 1770–1784. Retrieved from https://migrationletters.com/index.php/ml/article/view/11808

[2] Jeevani Singireddy,. (2022). Leveraging Artificial Intelligence and Machine Learning for Enhancing Automated Financial Advisory Systems: A Study on AIDriven Personalized Financial Planning and Credit Monitoring. Mathematical Statistician and Engineering Applications, 71(4), 16711–16728. Retrieved from https://philstat.org/index.php/MSEA/article/view/2964

[3] Dodda, A. (2022). Strategic Financial Intelligence: Using Machine Learning to Inform Partnership Driven Growth in Global Payment Networks. International Journal of Scientific Research and Modern Technology, 1(12), 10–25. https://doi.org/10.38124/ijsrmt.v1i12.436

[4] Koppolu, H. K. R. (2022). Advancing Customer Experience Personalization with AI-Driven Data Engineering: Leveraging Deep Learning for Real-Time Customer Interaction. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3736.

[5] Pallav Kumar Kaulwar. (2022). Data-Engineered Intelligence: An AI-Driven Framework for Scalable and Compliant Tax Consulting Ecosystems. Kurdish Studies, 10(2), 774–788. https://doi.org/10.53555/ks.v10i2.3796

[6] Srinivasarao Paleti. (2022). Adaptive AI In Banking Compliance: Leveraging Agentic AI For Real-Time KYC Verification, Anti-Money Laundering (AML) Detection, And Regulatory Intelligence. Migration Letters, 19(6), 1253–1267.

[7] Balaji Adusupalli. (2022). Secure Data Engineering Pipelines For Federated Insurance AI: Balancing Privacy, Speed, And Intelligence. Migration Letters, 19(S8), 1969–1986. Retrieved from https://migrationletters.com/index.php/ml/article/view/11850

[8] Nandan, B. P., & Chitta, S. (2022). Advanced Optical Proximity Correction (OPC) Techniques in Computational Lithography: Addressing the Challenges of Pattern Fidelity and Edge Placement Error. Global Journal of Medical Case Reports, 2(1), 58–75. Retrieved from https://www.scipublications.com/journal/index.php/gjmcr/article/view/1292

[9] Recharla, M., & Chitta, S. (2022). Cloud-Based Data Integration and Machine Learning Applications in Biopharmaceutical Supply Chain Optimization.

[10] Pandiri, L., & Chitta, S. (2022). Leveraging AI and Big Data for Real-Time Risk Profiling and Claims Processing: A Case Study on Usage-Based Auto Insurance. In Kurdish Studies. Green Publication. https://doi.org/10.53555/ks.v10i2.3760

[11] Someshwar Mashetty. (2022). Enhancing Financial Data Security And Business Resiliency In Housing Finance: Implementing AI-Powered Data Analytics, Deep Learning, And Cloud-Based Neural

Networks For Cybersecurity And Risk Management. Migration Letters, 19(6), 1302–1818. Retrieved from https://migrationletters.com/index.php/ml/article/view/11741
[12] Anil Lokesh Gadi. (2022). Transforming Automotive Sales And Marketing: The Impact Of Data Engineering And Machine Learning On Consumer Behavior. Migration Letters, 19(S8), 2009–2024. Retrieved from https://migrationletters.com/index.php/ml/article/view/11852
[13] Pamisetty, A. (2022). Enhancing Cloud native Applications WITH Ai AND Ml: A Multicloud Strategy FOR Secure AND Scalable Business Operations. Migration Letters, 19(6), 1268-1284.
[14] Burugulla, J. K. R. (2022). The Role of Cloud Computing in Revolutionizing Business Banking Services: A Case Study on American Express's Digital Financial Ecosystem. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3720.
[15] Nuka, S. T. (2022). The Role of AI Driven Clinical Research in Medical Device Development: A Data Driven Approach to Regulatory Compliance and Quality Assurance. Global Journal of Medical Case Reports, 2(1), 1275.
[16] Challa, K. (2022). Generative AI-Powered Solutions for Sustainable Financial Ecosystems: A Neural Network Approach to Driving Social and Environmental Impact. Mathematical Statistician and Engineering.
[17] Malempati, M. (2022). Machine Learning and Generative Neural Networks in Adaptive Risk Management: Pioneering Secure Financial Frameworks. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3718.
[18] Chakilam, C. (2022). Generative AI-Driven Frameworks for Streamlining Patient Education and Treatment Logistics in Complex Healthcare Ecosystems. Kurdish Studies. Green Publication. Kurdish Studies. Green Publication. https://doi. org/10.53555/ks. v10i2, 3719.
[19] Komaragiri, V. B. (2022). AI-Driven Maintenance Algorithms For Intelligent Network Systems: Leveraging Neural Networks To Predict And Optimize Performance In Dynamic Environments. Migration Letters, 19, 1949-1964.
[20] Chava, K. (2022). Redefining Pharmaceutical Distribution With AI-Infused Neural Networks: Generative AI Applications In Predictive Compliance And Operational Efficiency. Migration Letters, 19(S8), 1905-1917.
[21] Harish Kumar Sriram. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Mathematical Statistician and Engineering Applications, 71(4), 16729–16748. Retrieved from https://philstat.org/index.php/MSEA/article/view/2966
[22] Kannan, S. (2022). The Role Of AI And Machine Learning In Financial Services: A Neural Networkbased Framework For Predictive Analytics And Customercentric Innovations. Migration Letters, 19(6), 985-1000.
[23] Annapareddy, V. N. (2022). Innovative Aidriven Strategies For Seamless Integration Of Electric Vehicle Charging With Residential Solar Systems. Migration Letters, 19(6), 1221-1236.
[24] Siramgari, D. (2022). Enhancing Telecom Customer Experience Through AI Driven Personalization - A Comprehensive Framework. Zenodo. https://doi.org/10.5281/ZENODO.14533387
[25] Challa, S. R. (2022). Optimizing Retirement Planning Strategies: A Comparative Analysis of Traditional, Roth, and Rollover IRAs in LongTerm Wealth Management. Universal Journal of Finance and Economics, 2(1), 1276.
[26] Daruvuri, R. (2022). An improved AI framework for automating data analysis. World Journal of Advanced Research and Reviews, 13(1), 863-866.
[27] Ganesan, P. (2021). Advancing Application Development through Containerization: Enhancing Automation, Scalability, and Consistency. North American Journal of Engineering Research, 2(3).
[28] Vamsee Pamisetty, Lahari Pandiri, Sneha Singireddy, Venkata Narasareddy Annapareddy, Harish Kumar Sriram. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Migration Letters, 19(S5), 1770–1784. Retrieved from https://migrationletters.com/index.php/ml/article/view/11808
[29] Sriram, H. K. (2022). AI Neural Networks In Credit Risk Assessment: Redefining Consumer Credit Monitoring And Fraud Protection Through Generative AI Techniques. Migration Letters, 19(6), 1017-1032.
[30] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved

Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29–41. Retrieved from https://www.scipublications.com/journal/index.php/gjmcr/article/view/1294
[31] Komaragiri, V. B., & Edward, A. (2022). AI-Driven Vulnerability Management and Automated Threat Mitigation. International Journal of Scientific Research and Management (IJSRM), 10(10), 981-998.
[32] Chakilam, C. (2022). Integrating Generative AI Models And Machine Learning Algorithms For Optimizing Clinical Trial Matching And Accessibility In Precision Medicine. Migration Letters, 19, 1918-1933.
[33] Malempati, M. (2022). AI Neural Network Architectures For Personalized Payment Systems: Exploring Machine Learning's Role In Real-Time Consumer Insights. Migration Letters, 19(S8), 1934-1948.
[34] Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55–72. Retrieved from https://www.scipublications.com/journal/index.php/ojms/article/view/1295
[35] Kishore Challa, Jai Kiran Reddy Burugulla, Lahari Pandiri, Vamsee Pamisetty, Srinivasarao Paleti. (2022). Optimizing Digital Payment Ecosystems: Ai-Enabled Risk Management, Regulatory Compliance, And Innovation In Financial Services. Migration Letters, 19(S5), 1748–1769. Retrieved from https://migrationletters.com/index.php/ml/article/view/11807
[36] Anil Lokesh Gadi. (2022). Connected Financial Services in the Automotive Industry: AI-Powered Risk Assessment and Fraud Prevention. Journal of International Crisis and Risk Communication Research , 11–28. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2965
[37] Botlagunta Preethish Nadan. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Mathematical Statistician and Engineering Applications, 71(4), 16749–16773. Retrieved from https://philstat.org/index.php/MSEA/article/view/2967
[38] Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101–122. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1297
[39] Srinivasarao Paleti. (2022). Fusion Bank: Integrating AI-Driven Financial Innovations with Risk-Aware Data Engineering in Modern Banking. Mathematical Statistician and Engineering Applications, 71(4), 16785–16800.
[40] Pallav Kumar Kaulwar. (2022). Securing The Neural Ledger: Deep Learning Approaches For Fraud Detection And Data Integrity In Tax Advisory Systems. Migration Letters, 19(S8), 1987–2008. Retrieved from https://migrationletters.com/index.php/ml/article/view/11851
[41] Singireddy, J., Dodda, A., Burugulla, J. K. R., Paleti, S., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Universal Journal of Finance and Economics, 1(1), 123–143. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1298
[42] Kurdish Studies. (n.d.). Green Publication. https://doi.org/10.53555/ks.v10i2.3785
[43] Satyaveda Somepalli. (2022). Beyond the Pill: How Customizable SaaS is Transforming Pharma. The Pharmaceutical and Chemical Journal. https://doi.org/10.5281/ZENODO.14785060
[44] Daruvuri, R. (2022). Harnessing vector databases: A comprehensive analysis of their role across industries. International Journal of Science and Research Archive, 7(2), 703-705.
[45] Sikha, V. K., Siramgari, D., Ganesan, P., & Somepalli, S. (2021). December 30. Enhancing Energy Efficiency in Cloud Computing Operations Through Artificial Intelligence. Zenodo.
[46] Somepalli, S. (2021). Dynamic Pricing and its Impact on the Utility Industry: Adoption and Benefits. Zenodo. https://doi.org/10.5281/ZENODO.14933981
[47] Ganesan, P. (2021). Advanced Cloud Computing for Healthcare: Security Challenges and Solutions in Digital Transformation. International Journal of Science and Research (IJSR), 10(6), 1865-1872.
[48] Satyaveda Somepalli. (2020). Modernizing Utility Metering Infrastructure: Exploring Cost-Effective Solutions for Enhanced Efficiency. European Journal of Advances in Engineering and Technology. https://doi.org/10.5281/ZENODO.13837482

[49] Ganesan, P. (2021). Leveraging NLP and AI for Advanced Chatbot Automation in Mobile and Web Applications. European Journal of Advances in Engineering and Technology, 8(3), 80-83.
[50] Kaulwar, P. K. (2022). The Role of Digital Transformation in Financial Audit and Assurance: Leveraging AI and Blockchain for Enhanced Transparency and Accuracy. Mathematical Statistician and Engineering Applications, 71 (4), 16679–16695.
[52] Anil Lokesh Gadi. (2021). The Future of Automotive Mobility: Integrating Cloud-Based Connected Services for Sustainable and Autonomous Transportation. International Journal on Recent and Innovation Trends in Computing and Communication, 9(12), 179–187. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11557
[53] Sondinti, L. R. K., & Yasmeen, Z. (2022). Analyzing Behavioral Trends in Credit Card Fraud Patterns: Leveraging Federated Learning and Privacy-Preserving Artificial Intelligence Frameworks.
[54] Ganti, V. K. A. T., & Valiki, S. (2022). Leveraging Neural Networks for Real-Time Blood Analysis in Critical Care Units. KURDISH. Green Publication. https://doi. org/10.53555/ks. v10i2, 3642.
[55] Kothapalli Sondinti, L. R., & Syed, S. (2022). The Impact of Instant Credit Card Issuance and Personalized Financial Solutions on Enhancing Customer Experience in the Digital Banking Era. Universal Journal of Finance and Economics, 1(1), 1223. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1223
[56] Vankayalapati, R. K., & Pandugula, C. (2022). AI-Powered Self-Healing Cloud Infrastructures: A Paradigm For Autonomous Fault Recovery. Migration Letters, 19(6), 1173-1187.
[57] Kalisetty, S., Vankayalapati, R. K., Reddy, L., Sondinti, K., & Valiki, S. (2022). AI-Native Cloud Platforms: Redefining Scalability and Flexibility in Artificial Intelligence Workflows. Linguistic and Philosophical Investigations, 21(1), 1-15.
[58] Gadi, A. L., Kannan, S., Nanan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87–100. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1296