Deep Science Publishing
https://doi.org/10.70593/978-93-7185-253-1_1

Chapter 1: Understanding the critical role of auditing and compliance in complex and regulated environments

1.1. Introduction

"In 2020, the world underwent an unprecedented transformation; consumer and company behaviors shifted dramatically as a response to the pandemic, inspiring a digitally enabled world." For many, a search or a meeting became part of everyday life. Furthermore, in 2020, estimates show about three billion people worldwide were under full or partial lockdown. Staying connected became critical to everyone during the lockdowns, and social media emerged as the primary communication channel between people, contributing to increased usage of various social media platforms, many times exploding in popularity, usage, and outreach. Daily time spent online skyrocketed during the pandemic. This sudden and extreme push for digitalization led to exponential growth within numerous companies and sectors. Stocks gained significantly in value as opposed to a decrease in value of the aviation sector over the same time frame, a broad indicator of the industry impacted by the pandemic (Brazel et al., 2019; Alles, 2020; Appelbaum et al., 2021).

Remote conclaves became the norm during the pandemic period, and perceptions of the threat of contagion led people to rethink their social interactions, their travel intentions, and many agents of economic transactions in favor of digitally enabled interactions and exchanges. As a result, the pandemic spurred new trends related to social isolation, health and sanitation, and other aspects of the consumer experience. Travel, entertainment, health, and work became more digital than ever. The shift to digital displayed how much data flows and communication networks mattered for life and everyday experiences. Yet, even with this phase being uncanny, trends related to data, cloud, privacy and security had already taken precision and direction. In fact, before the pandemic, a slow shift towards digital was taking root propelled by further exponential growth of social media and the decreasing costs of bandwidth. Data had begun to transform itself into more than just a byproduct of businesses, but more as a critical asset

contributor to revenue generation, value creation, and return on investment (Cao et al., 2018; Yoon et al., 2018).

1.1.1. Background and Significance

The instantaneous and explosive growth of digital technology occurred in step with the Industrial Revolution in the mid-nineteenth century. Ingenious inventors developed revolutionary inventions ranging from the telegraph to railroads to machines capable of unprecedented production of textiles. Similar ingenuity created, nurtured, and reinvented computers in the twentieth century, laying the foundation for the sudden explosion of software applications and services in the twenty-first century. Collectively, these digital technologies accelerated the pace of globalization, putting power in the hands of every individual on the planet. Social media, mobile applications, e-commerce platforms, and on-demand business services became fixtures of everyday life.

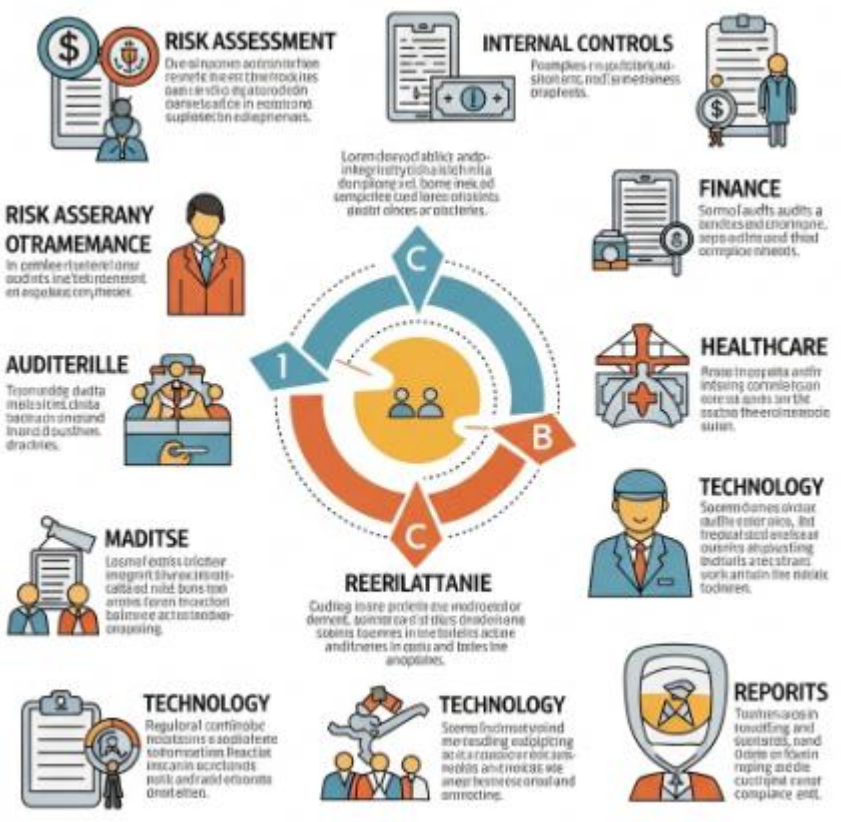

Fig 1.1: Auditing and Compliance in Complex and Regulated Environments.

The development and availability of these tools coincided with, relied upon, and laid the foundations of unimaginable growth in the generation of data. Humans created 59 zettabytes of digital data in 2020, and this will increase to 175 zettabytes by 2025. Data was categorized into what was referred to as a "data sphere," the world of data created each year from the Internet of Things, social media, business transactions, and database management systems. The collection and storage of data from these myriad sources eventually gave rise to a completely new discipline of data engineering. By design and necessity, database software evolved into what would become the discipline of data engineering. These data management platforms became massive, capable of storing trillions of bytes of information in a form useful to users. They were the technical backbone of Big Data storage and analytics, developed to capture, process, and analyze the large data sets of data-driven enterprises.

1.2. Historical Overview of Data Engineering

Data engineering is an engineering discipline that has been recognized informally for several decades at least. However, although data engineering as a formal field is relatively new, aspects of it have been practiced since the earliest years of digital data processing. Well-informed people, even among the ones not in the field, assert that data engineering is the key to our progress in our economy. Its work is also often described as key to the development and use of Artificial Intelligence and, by extension, enabling the Fourth Industrial Revolution.

The roots of data engineering reach back into the early days of computing, well before the emergence of electronic computers. For example, punched card data processing might have been the first data engineering effort. It not only captured data but also allowed data cleansing and the application of simple algorithms. After the invention of electronic computers, users quickly recognized the important tasks of programming address tables and building programs for allocation and cleansing. However, task specialization was rare, and programming for data access and data management were still typically performed by both users and computer center staff.

With the emergence of mainstream commercial demand for computing resources, the demand for rapid development of application programs with a variety of functions increased. Organizations grew their in-house computer skill sets, which resulted in the birth of development tools, librarian programs, and master data depositories of metadata. These facilitation tools simplified and accelerated the development and maintenance of application programs that utilized multiple data stores. However, many organizations hired outside services to develop and maintain their application programs.

1.2.1. Early Data Processing Techniques

The broad theme of information processing has existed since the dawn of civilization, with the emergence of cleaved stone for tallying absolute quantities, or cave wall paintings for tracking serial behavioral experiences such as hunting. The first complex techniques disseminated with the invention of writing up to 5000 years ago, in the form of clay tokens for tax keeping. This development allowed the transformation of knowledge creation, as books became a means for transference of thought across generations.

The earliest known ancestors of modern databases are inventors of punched card systems utilized for tabulation and analyzing information collected by the census. The development and widespread adoption of the tabulating machine for census data processing took place from the late-1890s through the following decade. These devices, primarily designed for the U.S. census, were discovered for use in business applications, which came to dominate the industry after 1909. Following the widespread adoption of keypunches for both government and business applications by the early 1920s, businesses began using tabulators to process transactions. They were still using ledgers for recordkeeping, however, often creating a backlog of work piling up in the ledgers for weeks or months.

Large oil companies were the first to develop and implement the use of office solutions that would first gain massive popularity during and after the 1930s with full scale implementation of electric tabulators and keypunch machines for all steps of the office data processing cycle. In tandem, they also developed and implemented the first mechanical devices for highly labor intensive processes, such as payroll processing, where the tasks of preparing work time cards, computing payment amounts, and preparing payrolls were all done manually, and were created monthly.

1.2.2. The Rise of Database Management Systems

The earliest data management efforts focused on data processing due to the high computational and storage costs. This changed with the introduction of magnetic core memory followed by magnetic disk, which made both data storage and online data access relatively inexpensive. As a result, developers shifted focus from data processing to efficient data storage and access, which became practical concerns for database system developers rather than application developers. In addition to supporting traditional transaction processing for commercial use, these early systems also had to support a wide variety of applications in diverse areas such as government and healthcare, including those requiring extensive batch oriented data analysis. These early data management systems allowed data items of various types residing in files on disk to be

shared and accessed easily by many users across different applications. After more than four decades of continued research, development, and commercial productization in both industry and academia, we now have a family of industrially standard commercialized DBMS products for transactional, analytical, and hybrid processing. Transactional DBMS products primarily use the relational model, while analytical DBMS products primarily use varying forms of a columnar model for data representation and processing.

The evolution of data engineering has been driven by continued advances in hardware and software technology. During initial data management efforts, database management systems developed for traditional data processing were essentially extensions of file management systems. Traditional DBMS enabled application developers to create systems that supported both online transaction processing and periodic batch-oriented data analysis. A couple of decades later, as the cost of data storage and access continued to decline, companies began storing detailed operational data in DBMSs and performing data analysis for business decision making on a frequent, regularly scheduled basis rather than on an ad hoc basis. This change in requirements marked the advent of data warehousing – a new area of data engineering applied to business decision support.

1.2.3. Advent of Big Data Technologies

The massive growth of the internet after the dot.com bubble saw an increase in data generation from a wide variety of sources. The advent of Social Media generated huge volumes of data from unstructured sources. Data from Social sensors along with data from Natural Sensors was growing exponentially. Data from Businesses and Industries, which mainly existed in structured form in Enterprise Databases, was being created, collected and kept both internally as well as externally in large volumes. This wide variety and massive growth in volume of data being generated and collected was termed "Big Data". However, the traditional mechanisms for handling structured data stored in relational databases began to crack under the weight of these new sources of Big Data. Conventional databases could not efficiently handle the new scale on the investments for updating and maintaining existing infrastructures in the face of increasingly competitive corporate environments. These potential costs related to not being able to leverage Big Data led many organizations to redefine their Data Management strategies. Traditional Data Processing Technologies such as ETL and Data Warehouses were not being rendered useless, but were beginning to be augmented with newer, diverse Big Data Technologies to implement Modern Data Processing Strategies.

MDPS brought together several Big Data Technologies to turn Data Lakes into Enterprise Goldmines by leveraging Modern ITOps, DevOps, and Site Reliability Engineering Practices. The concept of DataOps was born from these combined strands. Positive ROI on investments in these Big Data Technologies was faster and easier to

demonstrate than it had been for many traditional Data Management Initiatives. These trends have been gaining traction in organizations of all sizes around the world, as their decision-makers hurry to adopt, implement, and respond to the potential advantages brought by Big Data Technologies. The success of early Big Data implementation was primarily owing to large internet companies.

1.3. Key Concepts in Data Engineering

The domain of data engineering possesses its own concepts, whose clear understanding is crucial for the full comprehension of the work. In this section, we present these key concepts, which include: data warehouses and data lakes; ETL processes; and data governance. Importantly, it should be noted that these concepts are crystallized from the plethora of methods and technologies that are found in both the industry and academia.

A data warehouse is a collection of data centralized in terms of information about an organization and its main components, internal and external, mostly historical in nature. It is managed by a DBMS specialized at data warehousing and, in essence, data stored in a data warehouse can be utilized either for operational purposes or for decision support. A data warehouse is a centralized data repository that is accessible to various processes that belong to the data engineering and data science domains.

When compared with operational databases, data stored in a data warehouse is optimized for analysis, and, consequently, query processing is accelerated, because data is stored in a denormalized format; more specifically, a third normal form or a star schema/cube architectural style. Furthermore, data is periodically refreshed, new data is appended to existing data, and no updating or deletion of existing data is allowed. Finally, data owners are typically different, being operational databases owned by users who generate and interact with data, while data warehouses are created and maintained by a centralized data engineering function.

A data lake is a collection of data whose aim is the storage of original information from various sources, mostly current in nature, and is employed for operational purposes, or for decision support, or both. It is managed by a distributed system specialized at storing big data from heterogeneous sources and, in essence, data stored in a data lake can be utilized either for operational or analytical purposes, and it can be either consistent or inconsistent.

1.3.1. Data Warehousing

Data warehousing is a de facto infrastructure aiming to store all business-relevant data at a single and coherent location. The first thoroughly studied and documented warehousing architecture is called the Corporate Information Factory. The idea of enterprises storing most of their data in coherent data model structures thereby reducing or eliminating cross-system redundancy is generally acknowledged when implementing business applications. The goals of data redundancy reduction are firstly to reduce operational expenses for data management and, secondly, to improve data delivery reliability. A consequence of efficient data management is the removal of the business users' burden for facilitating easy data consumer access. Considering that in the early days of business systems, application data were exclusively stored in operational systems, it became evident that the combination of large enhancements in data volume and the technical limitations of those systems led to drastic performance degradation for mission-critical business functions. More and more enterprises started to implement additional large operational systems unable to deliver data for analysis in the required time frame. Last but not least, the diversity of query types, query volumes and performance, the diversity in terms of user expectations and the differing service levels between applications increased with the growing data management portfolio, contradicting typical efficiency measures. These issues led to increasingly complex not only operational architectures but also data model designs. A concept-driven approach integrating data from various business functions into a single coherent model structure on the basis of de-normalized data model structures alleviates the user access issues and leads to predictable OLAP performance. A consequence of positive data management experiences is that data users having access to data within and across the different applications become convinced of the value of accessing a central database for analysis instead of their operational databases, due to typically more tolerant data currency requirements and the ability to aggregate over longer time periods. It is the first step to transition to a central enterprise database architecture.

1.3.2. Data Lakes

Data lakes are a relatively recent evolution in the data storage and processing field that addresses some of the complexities of modern data pipelines and architecture. While data warehouses house curated, enriched, and structured data, often in a star schema framework, data lakes store both structured and unstructured data, often in the format and schema as originally received. Accessibility of data inside a data lake is typically better than a formalized data warehouse schema but requires significantly more work to ensure data meaning is understood.

The purpose of data lakes is to provide a centralized storage location for as much disparate data as possible, encouraging exploration and discovery of value present in the original raw data. Data lakes achieve this capstone of purpose primarily through the extension of cheap storage media and cloud technology. Modern data lakes are housed in cloud environments that enable pure storage providers, as well as cloud-native Data Lake solutions. The business goal of saving all relevant data and the execution and investment in cloud technology are the primary innovations of the data lake movement enabling a "store-all" philosophy on the front end of the data pipeline.

We can categorize data lakes into two main data types: file-oriented and data-oriented. File-oriented Data Lakes are typically less formalized environments that do not enforce schema on write, allowing users to ingest and store arbitrary files on storage services. Data-oriented Data Lakes are platforms that provide a more structured, UI-driven environment for storing data in different formats and structures.

Data lakes are powerful tools for ensuring a store-all philosophy at the front end of the data pipeline. With the growth of cloud technologies and increased storage capacity at reduced costs, the ability to scrape and store all relevant data is now entirely possible and, in many cases, a necessity.

1.3.3. ETL Processes

ETL processes have become one of the core principles of data engineering, along with warehouse technology as its first by-product. ETL is often confused with Data Warehousing. The first ETL tool was created just after the first Data Warehouse. However, the data warehouse was not vibrant, and other technologies were created for the analytical space. ETL began as a proprietary technology for specific vendors of Data Warehouses and relational databases. The first key concept that we mentioned was ETL. We introduced it for the following reasons. ETL was the principal way to populate a warehouse or a lake. It became complex, laborious, and costly. Organizations were unhappy with this situation. Why was the bridge too wide? Because the purpose of ETL was to move data from every possible legacy source to every possible warehouse. Connector technology gained momentum. Once again we had silos; but they were connected. ETL was re-designed from on-prem to cloud. Serverless ETL emerged – dynamic scalability. Instead of using thousands of servers, PETL is hosting the application on runbot. We also need to reconsider if we consider ETL with non-analytical storage. We need clean, clear data to do Business Intelligence. EAI does exactly what we always needed: integrate data, on-the-fly, into a common memory that has no coherence issue. Data Integrate. Composed Data. Buffer. Outbound Envelopes. Data Integration (DI) and ETL are currently converging. ETL vendors recognize the

revolution by offering EAI-style DI software as a separate product or an option within their suite.

1.3.4. Data Governance

In the world of rapidly evolving data environments, ensuring that an organization's data remains its most trusted asset is a growing challenge. Most organizations venture into a make-it-up-as-you-go approach without proper guidelines, leaving executives and information stewards guessing at which data should be trusted and which should not. In those cases, it's hardly surprising when people turn to damaged goods and use erroneous, inconsistent, outdated, and incomplete data when making important decisions. Damage to an organization's integrity can be great when the decisions inform financial investments, budget issues, clinical trials, product launches, and corporate acquisitions. The lack of trust can spread throughout the organization as a vicious cycle that causes people to spend immense amounts of time reconciling resulting in discrepancies between business units or departments. Handling data the traditional trial-and-error way encourages a culture of mistrust having detrimental results companywide. A company's many divisions cannot work independently without a common basis. Disagreements between business units inhibit communication, teamwork, and cooperation. This lack of coordination and communication leads to dissimilar and discordant reports that foster irritation and skepticism. To remove that cynical culture, organizations have embarked on formalizing the guidelines governing their data assets – data governance – and it's working.

A data governance program outlines policies and procedures for handling data and implements the practice of taking care of data as an asset. This allows everyone in the organization to share a common view of the company's data, reducing inconsistencies for a variety of external and internal data sources like corporate directories, addresses, products, partners, and customers, and maximizing the advantages of using data for making analytical, operational, tactical, and strategic decisions. Companies of all sizes are being dragged by a combination of technology enablers and aggressive vendors in the direction of instituting data governance best practices and tools.

1.4. Modern Data Engineering Practices

With the explosive growth of data created every second, data engineering, the technical implementation of data management strategies, is in the spotlight. Data engineers design and maintain infrastructure and architectures for data generation and use in data science, machine learning, and analytics. Data engineering is the foundation layer of the analytics and AI-driven economy. Without fast and reliable data sources, any data fueled strategy

would fail. In fact, the success of data-intensive applications directly depends on the ability of data engineers to create complex tools for data management and manipulation to enable data scientists, business analysts, and data consumers to analyze data and create models which guide the next steps in business transformation or product enhancement. Because of their mission-critical role in the current economy, data engineering is evolving rapidly and in parallel with the way we do business, and with developments in enabling technology through cloud solutions and the democratization and automation of data management steps.

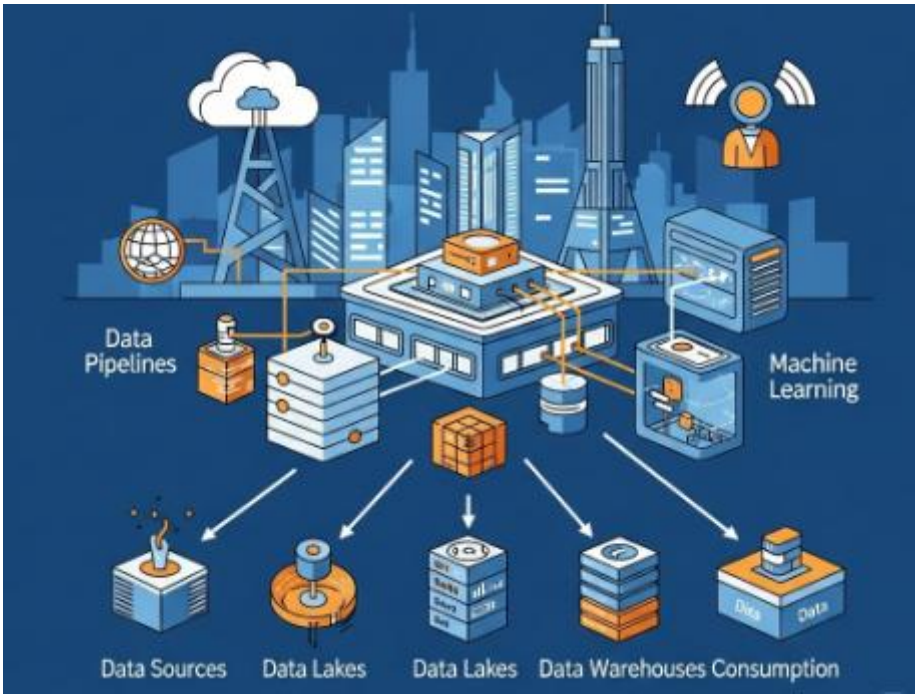

Fig 1.2: Modern Data Engineering Practices.

Modern data engineering is a twist on classical data engineering, incorporating cloud computing, open-source big data tools, near real-time data processing, automation solutions, and current security and governance practices. Unlike classical data management practices, modern data engineering embraces the philosophy of DataOps, Data Contracts, and Agile Data Engineering methodologies while adopting new tools and technologies for automating the preparation, storage, and delivery of data to enable faster and easier analytics. Today, you are hard-pressed to find a digital application, which does not rely on current data flows and processes for analyses, modeling, and automated or human-driven decision-making. To cope with the data deluge, modern data engineering incorporates best practices from other core engineering and IT disciplines: software development, DevOps, and systems engineering.

1.4.1. Cloud-Based Data Solutions

Today, data engineering teams deploy data lakes, data warehouses, and or hybrid solutions based on the current and short-term data needs of organizations but scale onto established commercial cloud-based solutions offered by the major cloud service providers. Some of the high-level benefits to be gained from any of these cloud-based data solutions include Lower Infrastructure Cost and Maintenance Overhead, Lower Data Accessibility Threshold, Data Deepening and Shared Data Sets. Lower Infrastructure Cost and Maintenance Overhead Cloud implementations incur less lifting and labor than standard on-premise bare-metal setups. Staff do not have to worry about disk drives filling up, power and connectivity issues with bare-metal servers, or constantly customizing on-premise systems as workloads change. Cloud storage costs reduce drastically on the basis of volume. This lowering of costs and labor allow data engineers to focus on higher-level tasks such as ingestion of new data source types, metadata management, and democratization of data access. Data archaeologists have lots of help from the cloud, as new data volumes and data sharing demands usually incur lower infrastructure costs than on-premise solutions.

Lower Data Accessibility Threshold – Especially third-party or partner-sourced data can literally incur variables and roadblocks that decrease sample rates to the point where costs to acquire, clean, ingest, and transform the data becomes pointless and the threshold for difficulty of obtaining outside data becomes very high. In the cloud, this process can usually be reduced to merely deciding how to align sensitive data files with the right permissions to the right cloud storage solution and with what level of encryption. Data Deepening and Shared Data Sets Companies want to share the data that they own with other companies or industry federations to provide insight to the growing pool of data scrapers, or even make money off renting access to databases. Sharing of external data sets with metadata documentation and management protocols sounds simple, but such seamless data flights are one of the massive pains of today's data engineers, and few companies do a good job at doing this.

1.4.2. Real-Time Data Processing

Real-time data processing means that data is processed almost immediately after it is engineered. The speed of data or data freshness is what distinguishes data engineering from established areas like data science or database management systems. Real-time or streaming data has special requirements: First, the service must offer low-latency processing, and thus minimizes the time between receiving the data and generating a result. Second, it must be scalable, supporting data streams with ever-increasing data loads. Third, it must guarantee reliability and the accuracy of the results in cases of

failure. Fourth, it must be cost-effective, to allow everyone to monetize the produced data.

The examples for real-time data are numerous: Data from sensors monitoring the condition of machines. Monitoring web access logs for cybersecurity issues. Watching traces from online banking. Monitoring trades from a stock exchange. Watching vehicle tracking messages or credit card transactions, or other money transfer transactions. Monitoring content uploaded to news agencies. Watching postings for sentiment analysis. Monitoring voice or text chat for issue resolution. Monitoring weather events for natural disasters. Monitoring the loading of a cluster for cybersecurity issues. Watching aggregators for visualization.

All of these scenarios need availability, monitoring, and analytics. Data is connected through a network and is brought together in a distributed data architecture. Events are stored in a distributed ledger providing immutability and security. Centralized analytics support data exploration, enrichment, optimization, and prediction. Offering thanks to a registered identity is model-free. Data can flow between distributed parts in an ecosystem through email, file exchange, or messaging.

1.4.3. Data Pipeline Automation

The provision and availability of a modern set of technologies for data engineering have contributed to accelerate development of new data infrastructures, but also to realize investment in and modernization of existing ones. However, acceleration is not a desideratum in the field of data pipelines design and creation. An important problem with modern data infrastructures is data pipeline sprawl: companies are diversely and too much investing in the development of ad hoc data pipelines that end up being, or are being foreseen to become bottlenecks for data consumption, maintenance and scalability. Companies want faster access to data but also need governance and cataloging of data. In this frame, an important activity is data pipeline automation.

The study indicates that over half of the data pipelines are still created from scratch, mainly because organizations use different destinations for their business data. Some of the top frustrations of data engineers center around maintaining and automating data pipelines, with data engineers up to spending a significant amount of their time on data quality issues and correcting broken data pipelines. However, there is interest in low code automation tools and pipeline orchestration tools. Future tools will evolve to automatically fix broken data pipelines and to monitor them for data quality issues, and to allow users to build pipelines with less or no code.

Data pipelines are responsible for collecting, cleaning, organizing, monitoring, and loading into storage the data that a company consumes and what the company's pipeline

consumers use. Without data pipelines, pipeline consumers would be unable to analyze data and extract insights. Data pipelines are the foundation of the data economy, supporting the most in-demand functions of organizations: machine learning and analytics.

1.5. The Role of Data Engineering in Digital Transformation

The qualitative and quantitative aspects related to masses of data produced on a daily basis, or Big Data, tend to derive in a Paradigm Shift that changes the business dynamics. Consequently, enhancing the decision-making process, as well as the knowledge about customers, products, competitors, and the market, relies on the capability to identify relevant data and deal with its specifics. In this context, Data Engineering, as responsible for making data ready for analytics, customers, or systems, plays a pivotal role in Digital Transformation, supporting three main fields that drive improvements in business results: Business Intelligence, Artificial Intelligence and Machine Learning, and Customer Experience.

Business Intelligence (BI) refers to the set of processes, technologies, and tools that transform unstructured data into structured information about the company's performance and generate insights that support the decision-making process. Data Engineering plays a key role in BI by helping Data Analysts and Data Scientists collect, clean, and manipulate data for business-oriented visualizations and reports. Additionally, they build infrastructure capable of handling the volume, variety, veracity, and velocity that characterize Big Data, so that Data Analysts and Business Intelligence tools access all relevant historical data whenever needed, and easily update queries and dashboards with more current data.

Artificial Intelligence (AI) and Machine Learning (ML) in turn, have been gaining traction from companies throughout the world. With that being said, the role of Data Engineers in AI and ML initiatives is to collaborate with Data Scientists by preparing the data needed for building ML models in the development and deployment stages. Besides that, Data Engineers also monitor the performance of those models and the underlying data, while ensuring that the right stakeholders have the right data at the right moment. Therefore, the contribution of Data Engineering in AI initiatives hinges on its capability to make good quality data consistently available and easily accessible for Data Scientists and ML Engineers.

1.5.1. Enhancing Business Intelligence

One place in which digital transformation has enabled significant accomplishment is in the B2B realm of Business Intelligence (BI). One of the primary roles of internal data is to accentuate a company's understanding of its own markets and operations. Robust BI development and dashboards are increasingly the by-product of platforms that consider data transformation a separate service. In addition, many businesses take advantage of the platform's infrastructure for managing data transfer and storage. Thus user firms "own" the BI product, while the BI platforms increasingly function as enabling simplified development and access to additional tools.

While the rapid development of pre-built solutions and BI platforms has eased the development of dashboards, for most firms the relationship is more nuanced. Without adequate plumbing and preparation of their data, inner-circle users cannot accurately produce BI visualizations. Thus, the transition from reporting to serious BI and analytics is predominantly a collaborative effort between the Data Engineers and internal data scientists or power users. These prototypes guide the Data Engineers as they put in place the foundational capabilities needed to empower decentralized, iterative, experimental, rapid prototyping if not permanent solutions. BI is like advertising for data and requires a close collaboration between Data Engineers and Data Scientists, as well as close connection with end-users, the internal clients.

1.5.2. Driving AI and Machine Learning Initiatives

Many companies implement AI and Machine Learning initiatives. Initiatives have simplified development and streamlined deployment. Business functions have adopted pre-built models from cloud and third-party vendors to optimize processes, such as document management, sales forecasting, and optimized marketing campaigns. Each of these ML initiatives has its data connection and lifecycle to add and optimize ML experiences. New model management platforms are being developed, aiming for a holistic view of the ML landscape and the ability to answer critical questions for organizations.

Yet, with all this activity, very few companies have truly achieved AI. For one, mainstream enterprise use of Machine Learning for critical business functions is still rare. Organizations tend to reserve the most critical decisions for humans while using Machine Learning as a heuristic check. Deep Learning research has made major advances in topics such as natural language processing and predictive state labeling. Illustrations show how Machine Learning tools such as recommendation engines, predictive purchase forecasting for e-tailers, and fraud scoring for credit card companies use Machine Learning in business practice to reduce significant costs and risk. But many

of the concrete production uses come from building engineering-based approaches that learn with a Machine Learning component. Those companies that have applied AI solutions have not implemented an enterprise-wide strategy. Instead, these implementations have focused on targeted business problems and used specific techniques and solution components. Rather than coming from an enterprise-wide need, the specific needs have driven ad hoc data access requests through unprivileged viewers and country disk consumption by target users. In large multi-site operations, multiple local efforts often overlap, exacerbating confusion. Data owners with no experience of the technical challenges involved have become overwhelmed, while local efforts often use poorly conceived and maintained methods. Duplication of effort within and across business units prevents the reuse of analytic steps and results.

1.5.3. Facilitating Customer Experience Improvements

Data and data management activities play a central role in creating a seamless customer experience. There are many facets to using data to advance on the customer experience journey, including knowing the customer, personalizing the communication and messaging to the customer, improving products and services, improving customer service, deploying chatbots and other self-service for customers, and customer experience management.

There are many attractive technology options available today for facilitating a better user experience. Organizations invest significantly in using these technologies to create better products and services, improve engagement through communication, and improve the quality of service to the customer. Research shows that for the customer experience to be truly transformational for a business, using data can help organizations drive a comprehensive, holistic approach. Depending on the digital focus for the organization, using data requires a depth of understanding to design a customer experience strategy that meets customer expectations consistently, even through low latency digital channels.

Having a data strategy cannot just be an advocacy for innovation and automation; those investment decisions must be made with a keen eye for tradeoffs that treat humans with respect. Customers want their data leveraged to create a better experience; businesses demonstrate that they respect their customers' data and its use when they base their data strategies in line with organizational values and ethics, both in terms of the digital footprint and the business decisions enabled by the insight derived from data against the digital footprint. Data engineering is the key discipline in realizing the promise of data in improving customer experience across the digital landscape.

1.6. Challenges in Data Engineering

Data engineering, while enabling the promise of Big Data environments, carries with it distinct challenges. The biggest are those of data quality: a fundamental premise of data engineering is that the data is store-ready for its purpose, yet many processes simply do not guarantee this. Data quality systems – data profiling, monitoring, auditing – are not included in most data engineering systems. Diverse aspects of data quality can be affected: if data is to be readily assimilated by machine learning models, it needs to be transformed to conform to the constraints of these algorithms: for example, numerical data should not have missing values, and a categorial variable should be converted into binary indicators. Data quality problems occur and continue undetected until they invisibly influence the result, clouding the overall data-driven strategy. The second challenge, scalability, is also fundamental. Many of the key processes for moving and transforming data – ETL systems, data cataloging, and data cleansing – have not clearly evolved, and struggle with the demands of current and foreseeable data volume, variety, velocity, and variability challenges.

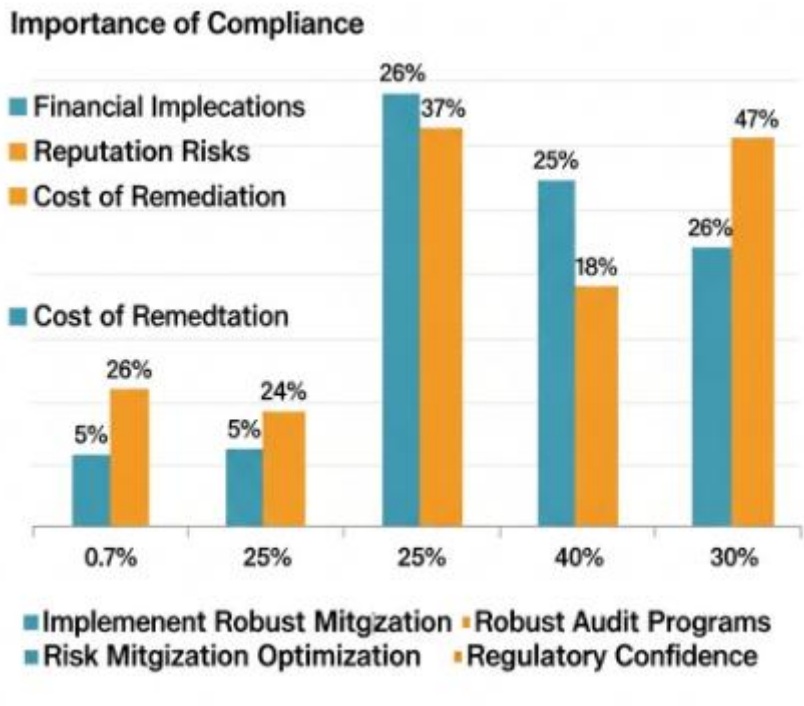

Fig : Critical Role of Auditing and Compliance in Complex and Regulated Environments.

Additionally, much data comes from increasingly diverse, heterogeneous sources, from different cloud providers, enterprise data centers, data lakehouses, IoT devices, API

endpoints, web scraping, or any of the other myriad tactics for acquiring data. This diversity, both in terms of logical and physical architecture, and in terms of technology, makes effectively discovering, integrating, managing, and orchestrating the diverse data sources and platforms difficult. These sources are compounded by both a lack of common semantics and metadata to understand the data, a task sometimes handled by external metadata standards, which seldom have enterprise adoption, and a lack of universal APIs for data platform interactions. Current ecosystem players help solve some of these challenges, for example, towards scalable transformation and integration, but do not provide clear guides for data engineers, who frequently have to rely on their expertise. For the field to truly emerge, it must provide a greater sense of methodology and strategy.

1.6.1. Data Quality Issues

The reduction of manual processes in favor of automated pipelines must be accompanied by careful data quality monitoring, as fully automated solutions carry the undeniable risk of poisoning the overall data assets of a company, which could involve making countless irrelevant images available, maybe even without advertising what their original purpose was. The company was then faced with the challenge of developing an internal solution that would clear out frames that were incorrect after the automated process. As noted, this does lead to a demand for hybrid, semi-automated solutions: Much of the process must be automated, but human involvement must remain essential if data accuracy and relevance are to be maintained.

Data quality monitoring services must then be set up, and their recommendations translated into actions taken in a timely manner. The amount of data requiring quality monitoring can, however, be immense – with an example provided by mentioning a company that monitors log files across data centers in several countries for about 200 billion messages daily. Such massive log collecting requires censoring out any unnecessary information, including false positives. Again, such a task provides an easy example of needing human assistance in order to avoid censoring out information that appears to be a false positive, as it may represent a major cyber security issue and consumers are fully expecting such companies to be able to protect their cyber security.

1.6.2. Scalability Concerns

Scalability is a term we all read or hear everywhere. It is a matter of scale. When we can provide a seemingly limitless amount of resources on demand, or a seemingly limitless request for resources, then we have scalability. In the context of data engineering, scalability means the ability to deal with ever-increasing amounts of data in a timely

manner. As companies move to the digital world, data is everywhere. Users leave data everywhere like a trace of their movements in the digital world. Devices leave data everywhere. Services leave data everywhere. A company that is delivering services to its users leaves data associated with that service everywhere. So what does that mean in regard to scalability?

When a company decides to invest in a data strategy, and sources data within its own organization, as well as external data, the amount of data residing within its various storage systems will grow. A company has various reasons to source or create data to integrate into a data strategy: being able to deliver better service to users, keeping those users excited, being able to connect those users to other users who can help them, and of course monetization of services are all reasons to invest in data. As those services become fruitful and conduct large numbers of transactions, the underlying data supporting those services increases and becomes ever more complex and interconnected. The processes and systems that support sourcing, integrating, validating and cleansing, and storage and management of that data also need to be complex and interconnected and allow for scale. Scalability is an important word to remember during every step and every decision made with a data strategy in mind.

1.6.3. Integration of Diverse Data Sources

The mere presence of data from several distinct sources makes the concept of data integration one of the most important issues in the data engineering field. In fact, the integration of data from diverse sources is crucial to adding valuable information to some application domain. On a big picture approach, the simple goal behind the different data sources integration is constructing a unique global instance of an application domain in a local context. Such a goal is achieved by a consolidated global schema supplying the needed information, and a collection of sources for each sub-domain within the local context, each one with its defined schema, possibly presenting semantically compatible elements with those of other source schemas, but physically distinct tuples representing the same facet of an application domain. By this means, the global schema makes transparent to the final user the existence of the different sources, as well as the data distribution.

However, this huge job of issuing the data sources associated with an application domain, together with the mapping to the local or source schemas and fulfilled by the global schema, is painstakingly labor-intensive. More troublesome is the main source of the data integration problems and what makes this effort enormous: is the data heterogeneity. The constituents of the information contained in the different sources may refer to the same conceptual domains but have different representations, which provokes the integration application to involve an error-prone process. Such a data integration

process consists mainly of resolving the following three basic steps: schema integration, data matching, and data fusion. Typically, new heterogeneous information is added to a traditional database, either by the mass importing of records during a large update operation or by integrating local updates into the database.

1.7. Conclusion

The digitalization of all aspects of daily life and business operations paves the way for the increasing adoption of cloud technologies, artificial intelligence, blockchain, machine learning, and the Internet of Things. Data science-assisted technologies will be increasingly associated with or replace software systems in support of optimization problems in finance, operations, and supply chain management. Technologies and innovation cycles are aligned to enable decision systems to improve continuously. Thus, bottom-line implications of digital transformation will include sustainable corporate performance outcomes. From a technology perspective, improved speed-to-value, decreased costs, and lowered risks can only accrue from more broadly deployed technology and resources for development. As such, along with the decreased cost of physical and technology resources, for the digitalization push to proceed, the barriers of configuration, development, and deployment must drop. However, this is not only a question of technological enablement but also of market structure, internal corporate environment, and ecosystem solutions. Ecosystem architecture and theory will suggest that decision processes will be key to both market structure and ecosystems in shaping the speed and investment models needed for this investment and digital transformation. Finally, the scientific legacy of data engineering needs to continue to build systems and supporting workforce capabilities for deep-domain expertise and applied understanding that support successful lower-cost, low-risk digital transformation decision-making and solutions. Future trends in data engineering suggest that digital transformation is likely to emphasize engineering-supported data systems that can enhance the data science effort in decision-making and applied action. The nature of data science-pooled engines working inside intelligent manufacturing, big data, intelligent financial services, and intelligent healthcare will reinforce data engineering as a shared corporate-critical technology that impacts each aspect of decision-making. Further, analytics systems will pervade all aspect of business management and operations, including data-supported preparations for climate change impacts as conversations enable the evolution of business strategy by suggesting environmental and other considerations, preparing for systems of intelligence, assessing their risks, and making the best decisions possible.

1.7.1. Future Trends

The recent explosion of interest in generative AI has laid the foundations for disruptive uses of structured data. The pre-trained transformer encoding methods on contextualized content feed into users' need to quickly convert the information available in LLMs into domain-specific action items. For example, a system was built to scrape product information into databases to feed into LLMs fine-tuned in the marketing language of companies selling similar products. Clients could then instantly ask questions about their competitors' pricing, description, and availability; and ask for action items to fix issues pointed out in short product review texts, or to improve the quality of their product descriptions. These action items would fit perfectly into the marketing plan that companies update and follow regularly. In this example, the engineers creating and maintaining the data and the associated data pipelines are taking on new functions where data usage, data interpretation, and data action are increasingly automated.

These trends are not limited to the marketing domain, nor LLMs as the sole answer engine to be used. As organizations create large amounts of structured data that come from enterprise functions such as HR, Sales, or Customer Service, we can connect data to users' daily activities via LLMs or other models that interpret domain-specific structured data. For example, companies engaging in complex employee attrition and recruiting processes can track major hiring trends, pipeline bottlenecks, and timelines of both incoming and outgoing candidates by departments. LLMs can help them write hiring pages that fit the segments they are trying to recruit. It can automate urgent actions dealing with upcoming interview dates and interviewers.

References

Alles, M. G. (2020). "The Evolution of Auditing: From the Traditional Approach to the Future." Journal of Information Systems, 34(2), 5–20.

Brazel, J. F., Agoglia, C. P., & Hatfield, R. C. (2019). "The Effects of Audit Review Format on Review Team Judgments." The Accounting Review, 94(5), 1–25.

Appelbaum, D., Kogan, A., & Vasarhelyi, M. A. (2021). "Analytics in External Auditing: An International Perspective." Managerial Auditing Journal, 36(2), 189–212.

Cao, M., Chychyla, R., & Stewart, T. (2018). "Big Data Analytics in Financial Statement Audits." Accounting Horizons, 32(3), 31–45.

Yoon, K., Hoogduin, L., & Zhang, L. (2018). "Big Data as Complementary Audit Evidence." Accounting Horizons, 32(3), 75–90.