Deep Science Publishing
https://doi.org/10.70593/978-93-49910-67-6_1

DeepScience
Open Access Books

Chapter 1: The evolution of data engineering and its impact on global digital transformation

1.1. Introduction

Data Engineering is a relatively new term, which has resulted from the necessary protection of the engineering profession and the natural evolution of Digital Technology, in this way Data Engineers have to work closely with Data Science. The use of Data Science has been available for a short period of time, and plenty of times, the Data Scientists are mistaken and considered Data Engineers, therefore these tasks are separated and come back to repeat its original meanings and function; therefore, it is necessary to take a look through history to better understand these two new professions.The data-driven organizations are at the top of their respective markets, considering the point of view of digital transformation that gives organizations innovate through Digital Technology. As Digital Transformation is a new way of thinking how organizations utilize Digital Technology, Data Engineering has to work to define the foundations of assured data, which is essential for the algorithms necessary for these organizational strategies and tactics. Digital Technologies evolve at a high rate, in the production, storage, and transport of data; we are talking about the Fourth Industrial Revolution, where the most important assets are information and the processes that generate knowing-how are more efficient and lean, but having their foundations based on accurate data. How do organizations use their organizational strategies focused on digital transformation? Having the backlogs of their data transformation processes developed by a certified information technology professional, the Data Engineer (Chen et al., 2012; Chen & Lien, 2014; Einsiedler et al., 2015).

Data Engineers are professionals who are responsible for building the management chains of Data in Business Intelligence solutions, from Data Extraction, Transformation, and Loading, to the Data Visualization tools, they design and optimize Data Pipelines

and Datamarts, Data Warehouses and Data Lakes, Data Robotics scripts, Cloud Computing, developing automation scripts, using Application Programming Interfaces and connecting data from different databases, in Data Platforms. But continuing our approach, Data Engineering is within Business Intelligence and never outside as Data Science (Panwar et al., 2016; Kumar et al., 2024).

1.1.1. Background and Significance

The digital era perceives a rapid economic, societal, and political transformation worldwide. These changes also accelerate the growth of digital data, resulting in an unification of the economy and society. Data generation, storage, and sharing processes have, correspondingly, increased exponentially in size and complexity. The volume and structure of data have made data management and processing a prominent practice for organizations striving for innovation. Digital transformation thus calls for a new art of using organizational data and revelatory effort to produce digital insights for all organizational practices. At the same time, annual statistics show that organizations rely heavily on marketing, sales, and customer service in data-driven strategies to drive digital transformation. Despite the exuberance of digital transformation, the nature of research themes – brand transformation and managerial-level digital decisions – diverges from the process of choosing how to move forward with digitalization.

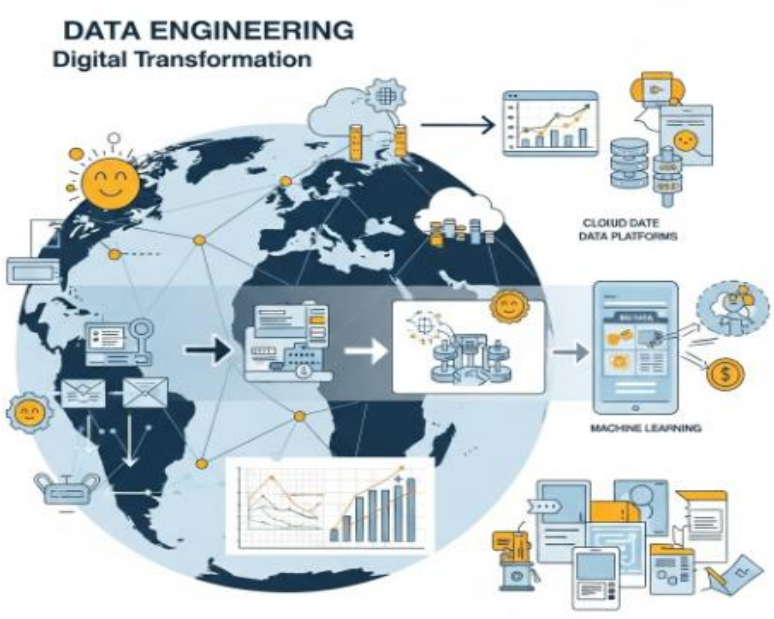

Fig 1.1: Data Engineering and Its Impact on Global Digital Transformation**.**

To overcome digital transformation challenges, organizations rely on data engineering and its related technologies and accompanying processes to design, create, manage, and optimize functional and technical data capabilities. Emerging tech that enable data engineering include analytics-driven data management, applied machine learning, advanced statistical techniques and technologies, artificial intelligence, natural language generation and processing, data visualization, robotics, cognitive process automation, and others. Relying on technology capabilities, data engineering provides lightweight and agile processes to deduce knowledge discoverable from data in a timely way. With organizational digital transformation firmly rooted from business strategy down towards business design and structure itself, the data-driven design of analog business creates the condition for speed and accelerates time-to-market.

1.2. Historical Overview of Data Engineering

Data has been a part of humankind's life for centuries. However, people and organizations did not have vast amounts of data available at their disposal. Only as of late, the exponential growth of collected data began. This growth was fueled by factors as: the globalization of businesses worldwide; the democratization of technology; the cheap price to store vast amounts of data; the impact of social networks; the Internet of Things; the widespread use of mobile phones; the technological advancements in sensors; and the aggregate time that people spend online.

The first accounts of data engineering appeared with Information Technology pioneers, who started storing data in computing individual, proprietary data stores. Early methods to manage decision-oriented data were based on computer programming, through the development of ad hoc solutions. With them, end-users obtained access to their data without institutional support. Methods to produce trustworthy decision-oriented reports were virtually nonexistent. In the early 1980s, tools available for reporting were few, mostly proprietary, and not user-friendly. Specific production query instructions were required for each report.

The first actual commercial solutions for decision support systems are data warehouses. These coding databases were originally theory-laden, with academic computing scientists developing peculiar, singular academic solutions using their data. The first commercial data warehouse was developed in the late 1980s, a remarkable and inflexible implementation of the original relational data warehouse theory, also known as star schema for decision support business transaction systems. In the latter half of the 2000s, the sheer volume, diversity, and data velocity began overwhelming traditional data warehouses. Data engineering solutions evolved past data warehouses. Emerging solutions were classified as Big Data businesses.

1.2.1. Early Data Management Techniques

During the early computing era, the management of data was primarily driven by the need for record-keeping and database support for application software. As early information technology systems were developed to support different business functions, such as accounting, manufacturing, sales and distribution, and HR, the need for maintaining higher volumes of data and integrating data between different applications was being felt. Initial management of data in the 1960s focused on supporting the business functions. Business functions were able to provide an initial set of requirements – what data needed to be collected from users, the methods for maintaining the data, and the use of data for applications. With the escalating demands for high-speed online access to critical corporate data, information systems in business organizations grew in complexity through the use of large-scale data processing, online transaction processing especially for consumer banking and financial services, and eventually, data warehouses.

The increasing demands for record-keeping in support of critical business functions and the emergence of different styles of data collection and usage until the early 1980s saw the first wave of early data management objectives and capabilities. First came the early diagrams into the operational aspect of organizational activities with a natural emphasis around the multiple applications that used different aspects of the same set of root data entities. A second set of diagrams were the data cycle conceptualizations or the data hierarchy diagrams. These diagrams tried to show that business data came from natural relationships amongst different data entities and that every data entity had intrinsic physical storage needs, semantic value, and life-time states. Based on the earlier work, different data models and management systems were implemented as specific commercial requirements.

1.2.2. Emergence of Data Warehousing

By the end of the 1970s, organizations had begun investing in more sophisticated data management technologies. Business intelligence systems were predicting the need for more complex data infrastructure solutions, helping corporate management make strategic decisions. Traditional hierarchical databases were becoming ever more incapable of handling the decision support requirements of multidimensional analysis for this new wave of business systems. In fact, data in operational systems on online transaction processing databases was becoming a bottleneck. Complex calculations and reporting processes were dragging operational systems down.

For this reason and others, many companies began to create additional databases specifically for the purpose of running queries. Companies soon began to realize that they could offload some of their processing to report generation systems. Once the

systems were in place, it became apparent that pulling data from multiple databases and integrating them into a central repository brought even greater benefits. These early warehouse implementations were designed as separate data repositories with query tools on top of them. By the late 1980s, enterprise data warehouse technology was forming around the integration of and query tools. All forms of decision support querying-age report generation, budgeting and planning, statistical analysis, and financial consolidation and forecasting-quickly coalesced around various versions of a decision support database.

Today, the data warehouse industry has evolved into one of the hottest areas of information technology. Enterprises across all industries are applying data warehousing technology to a wide variety of decision support applications. Large companies have invested millions of dollars building massive warehouse databases that serve the entire enterprise. These data warehouses are highly integrated central repositories that contain large volumes of historical transaction data, which organizations retrieve, analyze, and use to make critical strategic business decisions.

1.2.3. The Rise of Big Data

The late 2000s ushered in a seismic shift in the ways data was created, shared, stored, processed, and consumed. This shift was characterized by social networks, mobile platforms, sensor networks, and an expanding network of connected devices—often referred to today as the Internet of Things. The products of these networks generated massive volumes of diverse new data: web log files, social chatter, sensor signals, to name a few. Each day, millions of cell phone users: tweet, message, tag pictures, check in to venues, search for directions, take pictures, post blog entries, record videos. Such signals combined with vast stores of structured data create complex data lakes—pools of digital data that encapsulate the world's natural and social systems.

These new data assets take immensely diverse forms: raw and processed data; structured, semi-structured, and unstructured data; transactional, temporal, and spatial data; consumer, business, and public data; direct and indirect data. And every day, huge new volumes are generated. Modern devices are equipped with GPS chips and accelerometers that collect real-time streams of data about their owners' movements. Nested atop this system is a software and services ecosystem that fetches, consolidates, analyzes, and executes actions based on this data. How many new tweets are generated each moment? What do people talk about? What geographical and social areas are they talking about? How often? At what times? What actors are involved? Are they combined with pictures, videos? These data questions apply to not just Twitter, but to every aspect of our digital world—telematics signals, RFID movements, cellular events, phone traffic records,

human population data, credit card transactions, web search and clickstream logs, social network votes, newswire stories.

1.3. Key Technologies in Data Engineering

Research has shown that data is generated, manipulated, and incorporated into systems using several key technologies. These were initially focused on data warehousing solutions but have evolved to include newer concepts like cloud services that have changed the data engineering landscape. In this section, we will discuss some of the important technologies that support most commercial offerings related to data engineering.

The storage of data has been an integral part of any data-centric activity. Until the late '80s and early '90s, data was primarily stored on-premises. Companies housed large computers that stored all their data and had local control of it. The problem of growing data sizes was handled by increasing the size and capacity of these machines using hardware and technology advancements. Data that had structures and strict formats used Relational Database Management Systems to store the data. These storage options offered consistency and durability advantages while processing data through the ACID rules. Another feature that differentiated these systems was the implementation of the SQL programming language to manipulate data. The second category of solutions were purpose-built NoSQL systems that had eventually supplanted RDBMS-based solutions as the primary solution to store data. NoSQL databases were built around the CAP theorem and accepted designs with URS rule implementations that were less convoluted to handle large and semi-structured data. These systems had their programming interfaces built around key-value pairs, or for some document stores, implemented their own custom scripts.

The past decade has seen a huge increase in data and the processing of that data from disparate sources to provide intelligence to drive decisions. Businesses have used this data to test ideas and hypotheses quickly and determine Go or No-Go strategies in a fraction of the time and cost spent in traditional and conservative environments.

1.3.1. Data Storage Solutions

The growth of data volumes has necessitated increasingly scalable and flexible data storage solutions. Early data storage solutions – such as enterprise data warehouses, data marts, transactional databases and other silos housing discrete data sets – attempted to bridge, within the enterprise, the data domain chasms created by the proliferation of use case driven systems sustaining a variety of data types and formats. However, even with

growing adoption of extract, transform, and load processes to move business activity data to a centralized transaction processing or business intelligence system for analytics or reporting, the advent of Internet-scale distributed data generation and Digital Transformation heightened the demand for less rigid forms of data storage solutions.

Emerging first within the Online Data Processing realm, NoSQL and NewSQL databases – offering variably constrained storage for application driven, generally unstructured, data generation and operations – soon were witnessed to be adapted by Internet businesses for data analytics as well. Eventual Model Consistency mechanisms allowed organizations to operate on very large data sets of variable and low granularity data types more akin to that typically generated across systems of engagement within the Digital Economy. Wide Column, Document, Object and Graph databases designed specifically to address challenges associated with Digital Transformation related to both Data Engineering and Data Analytics are now ubiquitous and available within the Cloud Computing ecosystem via the Database-as-a-Service offering framework.

The Data Lake analogy ultimately delineates a broad concept of evaporating enterprise data warehouses in favor of heterogeneous mass storage repositories from which structured, semi-structured and raw data can be accessed and utilized for diverse Data Engineering and Data Science use cases, including Data Analytics. The Data Lake Oil Theory attempts to address by analogy the questions of governance, control and performance posed by the allowed lack of structure inherent in a Data Lake approach to Data Engineering while Data Fabric Technology efforts attempt to formalize processes and tools on top of Data Lakes and other such systems for Enterprise Data Engineering embedding DataOps workflow automation.

1.3.2. Data Processing Frameworks

Batching refers to the execution of a sequence of tasks on a unit of data at once. A batch processing framework typically has a scheduler that the user can use to indicate when functions should run and a storage solution for ensuring that the input data is available in a timely manner. Originally, the input data for a batch framework consisted of sets of data files, each identified by a name pattern; or a potentially infinite stream of updates, being deposited into one of a set of lists. Early frameworks were technology-specific, while a programming model popularized the MapReduce programming model. At its heart, MapReduce maps out the data, distributes it across many machines, runs a separate task on each piece, and then reduces the results.

Hadoop later gained an ecosystem of supporting technologies, such as Hive for data warehousing, Pig and Spark for alternatives to MapReduce, HBase for fast random access to data stored on HDFS, Oozie for task scheduling and dependency management,

ZooKeeper for distributed coordination, Sqoop for efficient SQL-import and export, and Flume for bulk loading log data from various sources. Most of these were written in Java. The last decade has, however, seen dozens of alternatives emerge, often based on Python; great progress has been made, for example, by Airflow and Dagster. And more recently, as memory became steadily cheaper and more abundant, Spark popularized a new model of "in-memory lazy evaluation" which accelerated batch operations.

1.3.3. ETL and ELT Processes

Extract, Transform, Load (ETL) and Extract, Load, Transform (ELT) processes are two paradigmatic routes for moving and processing data within systems. Both paradigms have been gaining relevance in the current data engineering landscape and are essential concepts in the field. ETL processes predominate in traditional data warehouses and the data pipelines used to import data from operational systems into them. ETL is an old technique used to move data extracted from source systems into repositories for operational and business intelligence purposes. ETL usually refers to a set of data pipelines that move large volumes of data on a daily basis and require several hours of processing before the resulting data is ready for reporting and analysis-oriented tasks. ETL moves data into a target database, which contains all the data transformed and conformed for analysis. ETL is an old concept in databases, which was implicitly used by most on-premise data warehouse solutions in the pre-digital-transformation era. ETL tools have moved on to become practical solutions to constantly design, deploy, and execute complex data integration solutions.

ETL processes grew out of ad hoc scripts developed on-demand inside a data warehouse by the same teams developing reports, dashboards, and analysis queries on top of it. Bulk inserts were usually used to load and transform any on-premises database management system, which was the de-facto technology used to develop those early data warehouses. With the advent of cloud computing, the demand for huge volumes of analytical, reporting, and dashboarding tasks became a consolidated market and attracted specialized companies in the analysis domain, which began offering distinct cloud-based solutions built on top of their own cloud data warehouses or data lakes. Over the last two decades, the availability of cloud computing services has changed the way organizations built their data management and data science ecosystems, encouraging business ventures to externalize and exploit data.

1.3.4. Cloud Computing and Data Engineering

While the previous topics discussed building blocks for data engineering systems like data storage and processing, this section will discuss a technology that eased the building

of data engineering systems and service worldwide: cloud computing. Cloud computing is a diverse set of tools and services that runs under a model in which secure and on-demand computing services are offered remotely, often charging consumers only for the computing time they consume, as open-source IT services tend to be more common in other technical areas.

Cloud computing is a model for enabling ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. The resources mentioned in the definition can be related to the core components of data engineering, such as data storage and processing. Solutions for data storage include cloud-based data lakes and data warehouse as a service solutions. Other cloud-based infrastructure services also are being used to build on-premises data warehouses. Solutions for data processing include cloud-based servers managed by cloud services and serverless computing services.

Cloud computing has made building more complex data engineering systems—the so-called data pipelines—easier, in part because all services necessary to build and run them are available in an integrated way, and in part because the way we acquire IT services is different—companies no longer need to hire and maintain a full technology team to build services to handle the storage and processing of the mass data generated by their business. With that, small and medium enterprises can leverage data engineering services that were unavailable or too complex to build and manage.

1.4. The Role of Data Engineers

Data engineers are responsible for creating the infrastructure necessary for optimal extraction, transformation, and loading of data for analytics by data scientists. They design the architecture and data pipelines that allow data scientists to spend the maximum amount of time conducting analyses. In addition, data engineers are often tasked with maintaining the infrastructure of data analytics, including the hardware, databases, pipelines, cloud infrastructure, orchestration of data movement, and monitoring of data. Data engineers should possess a software engineer's skill set, as much of their day-to-day work involves coding to solve problems in data. Data engineers cooperate closely with data scientists to determine how to structure and process the data needed for various analyses. Frequent communication is necessary to ensure that data pipelines deliver data that meets the needs of data scientists' analyses. Although pipelines are designed by data engineers, data scientists spend a great deal of time defining how the data should be structured, the intricacies of the required analyses, and the required deadlines for data delivery. Data scientists may also consult with data engineers when reviewing and selecting the most effective tools and technologies for

their analyses. Creating the pipeline that instances a data delivery and supporting framework is not a trivial matter. Decision makers expect insight gleaned from data analyses to be correct and should be able to trust in the provision of structured data and that the associated methods of delivery are accurate. If a data scientist analyzes the business impact of a new pricing strategy to enhance sales volume and increase gross profit that is based upon a dataset which contains samples that are incorrectly labeled, the proposed recommendations may be potentially disastrous for the company.

1.4.1. Skills and Competencies

Data engineers are responsible for preparing data so that it can be effectively used by data scientists. Usually working in tandem with data scientists, they relate closely with data management and development teams. Usually involved in large-scale data warehouse design and implementation, data engineers bridge the gap between data management and data analytics. Data engineers apply technologies such as data management, data integration, databases, data warehousing, machine learning, and business intelligence to move raw data to the point where data scientists can use them for model building. Building a data pipeline is their main contribution to a project.

The growth of big data technologies, tools, and techniques has led to a growing need for expertise in building and deploying data pipelines – which convert data into a format that the models created by data scientists can work with. They design and build systems that allow massive amounts of data to be accessed and analyzed. Data engineering involves leveraging the right storage, ETL, and data processing technologies to allow businesses to scale analytics and data science. The skills needed for building large-scale, high-performance systems include experience in data warehousing, data modeling, coding, distributed computing, and working on cloud platforms. Overseeing the processes involved in collecting, moving, cleaning, and organizing data, they build and maintain the architecture. Proficient at scalable data manipulation, data quality and validation checks, batch and streaming processing, and distributed computing, data engineers work with databases at scale, message queueing systems, and workflow orchestration platforms.

1.4.2. Collaboration with Data Scientists

A crucial aspect of a data engineer's job is collaborating with data scientists to optimize the usage of data and to solve challenging business problems. Data scientists need fast access to clean and well-formed data that expresses the business problem clearly. More often than not, they start from data already available, and pursue lower-quality techniques to navigate to the business goal. Intervening at this stage may avoid delays,

high costs, and achieve better business outcomes. To maximize the value of the available data, data engineers also need to be knowledgeable about machine learning and data science techniques, besides building the tools needed to store, transport, and clean the data. At the same time, it is important for data scientists to have some knowledge about the data pipeline to describe the characteristics of the data and to understand the data processing limitations. Various tools have emerged that make these alternatives work, such as data catalogs or other semantic web technologies. Tools from the data integration and big data domains help describe and schedule the data transformation process so that data scientists also understand these limitations. Data scientists can also use notebooks that use data visualization libraries, becoming examples of data exploration techniques to share for the whole organization. Using the same tools also facilitates interactions. Sharing notebooks through version control systems can clarify what the data scientist's expectations are about the data transformation. In the end, a collaborative environment that also features shared blame for the quality of the data adds security in the production process. Just as in software engineering, data engineering projects also need to follow specific methodologies to manage the functional and non-functional requirements that impact data quality. Various data engineering tools and platforms from the data integration and big data domains showed up that further increased the ease of providing reliable storage and transformation of data.

1.4.3. Impact on Business Decisions

The growth and popularity of Big Data have transformed the role of data engineers, and consequently, their work has evolved too. Companies have realized that there is inherent value in data that they generate through different processes whether it is transactional data, campaign data or other sources. Data engineers have effectively partnered with the business in leveraging this data to make better and smarter business decisions. Organizations have recognized this contribution and are also changing the way they assess data engineers. Traditional competency models do not apply anymore. The ongoing technology experiments and advancements in areas like cloud computing, in-memory databases, distributed frameworks, data wrangling technologies, data quality, machine learning and digital ecosystems, are providing the necessary encouragement for organizations and tech giants to flatter their career paths.

The new-age data engineering professional is being assessed not only for their skills and competencies but also for their contribution to business decisions especially with implementation of data products, data integration with business solutions, evolution of data ecosystems and enablement of self-service solutions. Career paths are being built that provide flexibility for the data engineer to learn new-age analytics capabilities like data storytelling, data strategy and be at the forefront of the complex intersection of data,

business and technology. They continue to be architects of scalable data pipelines, seeding data insights for fast business action ensuring organizations expedite their Digital Transformation journey while alleviating the risks associated with adverse business decisions. Reporting structures are evolving under the C-suite functions of the organization where data engineers are now looking up to Chief Data Officers or Chief Technology Officers for career progression, along with their collaboration with Chief Information Officers to operate the organizations' IT and Data infrastructure.

1.5. Data Engineering in Various Industries

As we discuss the evolution of data engineering in the present day, it is important to reflect on some examples across different industries to uncover how these advancements are applicable. The healthcare industry serves to implement some of the most modern use cases of data engineering such as the ability to sequence a complete human genome in less than a day. Data pipelines in the healthcare space have been heavily influenced by the need to act as a source of truth for different downstream ML and AI workloads. Working with healthcare data is tedious and requires immense privacy concerns and compliance with very strict regulations. These solutions allow data engineers to focus on solving important business problems, while automating mundane tasks. The work being accomplished here will serve to facilitate structures and pipelines needed to support the next revolutions in medtech such as dramatically reducing the costs of genome sequencing or even embedding genome sequencing as a part of wedding packages. The healthcare advancements seen there will serve to guide and create a ripple effect across various other industries.

The financial services industry is an early adopter of technology such as better access to high performance computers, GPUs, and more recently TPUs. This specific industry harnesses data engineering and a well built data infrastructure to serve as conduits for major trading firms to identify new trades and strategies on events happening across the world. As the ties between different economies across Europe tighten, major financial services firms are looking at events happening across Europe such as news announcements from their own governments to determine when to sell and when to buy Euros. Data pipelines here act as important sets of rules for determining whether these events will be attractive enough to act upon and when such events are taking place.

Fig 1.2: Data Engineering in Various Industries.

1.5.1. Healthcare

People invest a great amount of time gathering and assembling various forms of health-related data via metrics and methods from quantitative and qualitative human-centered design research. In addition to existing forms of digital data, new forms of health-related data are generated through social media, smart devices and objects, and other non-digitized, formal and informal facets of the health and wellness ecosystem. Data from these and other sources are new forms of big health data that can be applied to HCI research to gain insights into health-related issues and aid design that puts personal and social health into action. In fact, there are many important and influential stakeholders in the health and wellness ecosystem who need to be better involved in the design process.

Crowdsourcing is a popular and widely adopted research and design method in the HCI community. Its application to HCI exhibits many benefits for accomplishing many tasks within any phase of the design process. For example, crowdsourcing enables crowds or masses, or people from civilian and diverse backgrounds, to assist in activities that will advance research by providing patches of data since dealing with big data is not something one person or group can accomplish easily. As a result, the collaborative and collective feature of crowdsourcing is leveraged in HCI by employing a diverse group of users and/or experts to conduct activities that can assist in the successful completion of the design process in ways that transform research ideas into solutions and novel design. Online data collection such as crowdsourcing has become a cost-effective and timely way to gather data for various types of qualitative and quantitative HCI research.

However, data collection is demanding, complex, and takes time to complete. The ability to obtain an abundance of health-related data with relative ease is particularly alluring. Moreover, obtaining data crowdsourced from individuals located in various public venues or with various backgrounds can lead to having a diverse group of people who can assist researchers in completing a particular task. Recently, crowdsourcing platforms have offered researchers a new way to conduct studies. Specifically, such platforms utilize the internet in combination with traditional experiment methods to recruit a stipulated number of participants for an online task-design data collection and analysis inexpensively or free of charge.

1.5.2. Finance

Finance is another vital industry that has evolved and been significantly impacted by data engineering. The Financial Services industry has also undergone revolution through digital transformation and shift from physical to digital-first businesses fueled by adoption of data, analytics and optimization. From banks and credit unions, investment management and insurance companies, and centralized to decentralized finance, the Finance industry is heavily reliant on their ability to collect, consolidate, store, manage, transmit, integrate, and interact digitally with their customers, internal and external stakeholders, while leveraging data analytics and data science to maximize profitability and shareholder value.

The basis of all financial services offered by Banks, Insurance, Capital Markets, and other businesses is transactions. Financial transactions are conducted each day among individuals and businesses through accounts – ranging from current and savings accounts at Banks, through death, health, and property insurance policies, to mortgage payments and the buying and selling of Staples by Publicly Listed Corporates on Capital Markets. Each transaction generates data pertaining to several fields such as region, demographics, timescale, nature, amount, etc.

While individual transactions may only be a few decibels worth of data, there are billions and trillions of transactions taking place every day, and considering a decade-long timescale, petabytes worth of transaction data are generated. Hence, the Finance industry demands state-of-the-art data engineering capabilities to access, process, manage, and analyze all transaction data generated. Moreover, the Finance Industry comprises highly regulated businesses, mandated to comply with procedures advocating audit readiness, and the internal control framework for the prevention of fraud, corruption, and money laundering; hence, these businesses require the most advanced data engineering support for availability, consistency, and security of data.

1.5.3. Retail

In the retail sector, data engineers facilitate the collection, aggregation, and conversion of consumer behavior data into business intelligence. Retailers use the data to better understand what consumers buy and why, to create ideal product assortments by season and locale, and to maintain optimal stocks on their shelves. The data helps also inform selection of retail partners to cheer, or to boo.

Moreover, analytics provides solutions to the classic retailing dilemma of deciding on which of the many promotional offers available to use to make product sales more enticing at different times and across different target customers. In the domain of online merchandising, a growing number of businesses rely on product information such as data feeds from manufacturers and suppliers, websites and social media to layer additional product images and descriptions on to products pulled from online catalogs. In-store retailers create a digital front door to their business through mobile and online shopping apps, web chat, and social media. Data creation has proliferated at these businesses. Retailers can deploy specific data engineering infrastructures and tools to capture the highest volumes of online clickstream and media data, while managing the deluge of marketing campaigns and sales data.

Data engineering is having an impact on the world of retailing over and above creating business intelligence and building digital storefronts. Retailers and CPG brands are competing on whether they can create the most intelligent supply chains. Data engineering also informs decisions on when and how to connect the online and in-store shopping experiences. Over the past few years, many CPG brands have intertwined the management of their manufacturer and retailer e-commerce channels. They have used search-engine optimization techniques to bid down keyword prices on their product pages for those retailers that take in product feed data from them, while representing an overload of online customer reviews and images, curating influencer content across social media, and bombarding customers with promotions, product logos, newsletters, and emails. The goal? To persuade online shoppers to buy directly rather than via a discerning e-commerce retailer. Brands are competing on whether they can make a faster direct-to-consumer delivery.

1.5.4. Telecommunications

Telecommunications has emerged as one of the most dynamic sectors in the global economy through both technological advancements and deregulation. Since approximately 1990, when American AT&T ceased to monopolize long-distance voice service in the United States and began to face competition from overseas, telephone companies have faced aggressive competition from internal communications systems,

international satellite consortiums, and alternative telephone networks as well as from large suppliers of communications equipment and proprietary software. Telecommunications has long been recognized as strategically important in economic development, and a clear distinction between public, commercial, and private telecommunications systems has not been maintained. Many telecommunications services and facilities needed for supporting global production and distribution, either because of rapid growth in developed-country markets or because of new technological and institutional capabilities in developing countries, are among those that have a practical commercial prospect during the 1990s and beyond.

Many of the basic large, heavy, and expensive private wired telecommunications service were pioneered by the multinational corporation, and continue to be heavily utilized, but less costly and lighter satellites as well as exploding domestic markets for personal communication systems and telephone services have opened substantial new opportunities for telecommunications suppliers to individual companies and to groups of local enterprise customers in developed as well as developing countries. In performing this mission of supporting global economic development using telecommunications, foreign direct investment, control, and knowledge transfer, telecommunications supply can also add to the rapid growth of the global economy as well as to the internal economic development of the countries making up that economy. Telecommunications has added a new dimension to the process of globalization.

1.6. Challenges in Data Engineering

Data engineering is not without its own set of challenges. New technologies and vast quantities of information present new hurdles for data engineers. The creation of massive data repositories and the demand for instantaneous availability of information, while daunting, are also real challenges. In addition, technology advances and the need for more sophisticated algorithm development have changed what data engineers do. They are no longer just expected to create massively scalable systems that make use of information. As companies increasingly rely on both internal and external data for business intelligence, data engineers need to take responsibility for the quality and reliability of the information pipelines they build. A data engineer's role has become much more than a simple engineer.

One of the most important tasks performed by a data engineer is figuring out what data pipeline should be built to facilitate the use of the data. Most functions utilize algorithms, but very few companies have a mature algorithm development function. The challenge is providing engineers with the pipeline that functions can put to use to create models and take them into production. Data modelers need easy access to data stored in a variety of sources and formats. Access does not guarantee that the data is reliable or of sufficient

quality for model development purposes. Some companies and researchers have chosen to build their own tools that can handle ad-hoc query requirements. While this can solve the need for instant data queries, it puts additional pressure on a data engineer to support the requirements of a wider range of users than an internal client development team.

1.6.1. Data Quality and Governance

Effective and accurate decision-making processes need high-quality and reliable data, but these types of high-quality data do not always exist in a data-driven organization. These challenges are centered on poor data quality metrics, such as accuracy, timeliness, consistency, validity, reliability, and relevance. Additionally, there is no effective and efficient method or process to govern and quality-check all the available data resources. Issues with data quality are intrusive and, more likely than not, could be classified under the bigger umbrella of governance. How data are collected, transformed, consumed, and reviewed, including data accessibility, usability, and credibility, must be schooled and structured within a trusted framework. Data governance processes define policies and structures with corresponding roles and responsibilities that dictate how an organization manages its data operations, assets, and ethics. The rules set up under this data governance framework augment and enhance data transformation and storage processes, monitor data flow, and sustain data quality and integrity across the organization.

As such, organizations need to democratize access for all business-end users to leverage the data along the conversion and flow cycle, which is usually restricted to data or information technology work biases. Enabling a wide set of business users to access data and carry out design and development work expands the possible number of applications. After all, there could be thousands of processes available for utilization if data is available to everyone. Data engineering's nine functions can be entrusted to be performed by many data-savvy business users, including creating and managing data warehouses, data lakes, and data marts; building and managing data pipelines; creating and managing data quality rules; modifying transformations; and creating and managing data access methods.

1.6.2. Scalability Issues

Abstract: Big Data has emerged as one of the most significant factors influencing digital transformation. Data volumes are exploding, data is being generated faster than ever before, and data types are evolving constantly. All of these factors have contributed to the biggest challenges in the data lifecycle: scalability and performance. In this paper, we explore the evolution, creation, and growth of Data Engineering as a discipline in its own right, and Data Engineers as role players in the broader data ecosystem. We propose

that Data Engineering has become a critical enabler of global digital transformation initiatives.

Big data analytics has grown exponentially during the last two decades. With business data growing at an unprecedented rate, so too has the investment in the discipline of Data Engineering, and rightly so, for poorly managed data cannot provide information at scale. Current challenges are not limited to extracting, transforming, and loading data into data lakes, data marts, and data warehouses – data is often engineered using pipelines that process and perform analyses on external-on-the-fly data, or at-rest data that has just been ingested; and hundreds or even thousands of pipeline components that process both historical and real-time transactional data on their way to models that predict the future.

The evolution of Data Engineering, therefore, is directly driven by the continued growth and demand for better and faster processing, and deeper analyses of larger and larger volumes of data, and increasing pressure to uncover insights at speed with precision. Models are being used that have billions of odds parameters, and demand for the entire analytic lifecycle to be executed hundreds of times a day with ever-increasing accuracy, on ever-increasing data volumes combined together with diverse data types, using both at-rest and real-time transaction processing data, arguably adds up to data volumes that are far above what is commonly termed "big data".

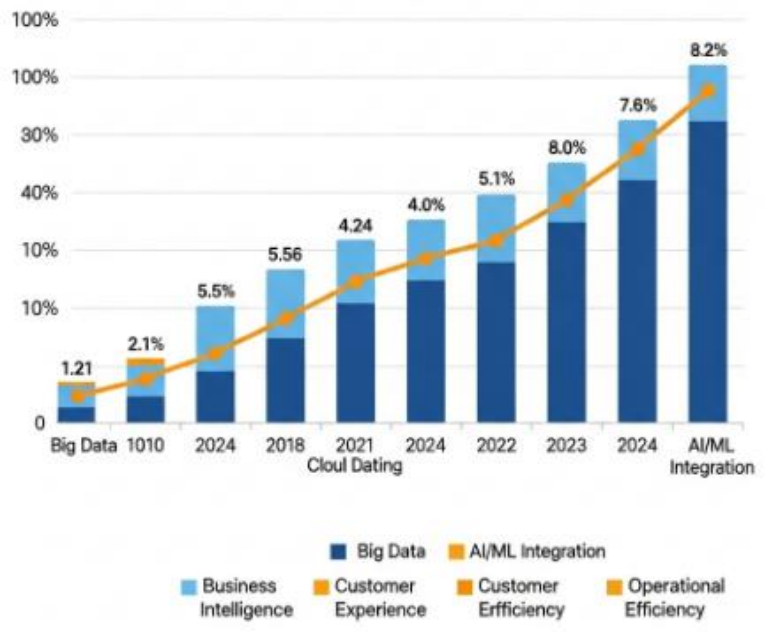

Fig : The Evolution of Data Engineering and Its Impact on Global Digital Transformation.

1.6.3. Integration of Legacy Systems

The demand for digital transformation and enhanced data management has accelerated the implementation of innovative cloud data solutions. However, these new systems are forced to coexist with existing legacy systems that still contain a large amount of business-critical data. The need for data engineering teams to connect valuable interacting datasets present in these legacy systems becomes crucial to provide actionable analytical insights. Connecting these previously inaccessible datasets within Data Lakes exposes new business opportunities, maximizes the utility of existing business processes, and often helps to avoid the cost of legacy system replacement.

But, as opposed to new data sources natively designed for distributed cloud systems, legacy data sources are often incorrectly designed for the problem domain and implemented on costly proprietary software solutions developed in silos and for security concerns tend to expose limited integration capabilities, presenting unique challenges for data engineering teams to solve. Legacy systems that need to be integrated may include mainframes, Data Warehouses, ERPs, flat files, and APIs from third-party applications, among several others. Integrating them to new cloud data platforms while moving the ETL processes usually present in the older systems to allow a more scalable ELT approach is one of the biggest challenges for data engineers.

Moreover, data engineers must keep in mind that data consumers may require data as close to real-time as possible, as more organizations adopt a real-time analytics strategy to support business operations that operate on real-time data. However, these legacy systems mainly use either batch data ingestion, where data moved from legacy solutions to cloud systems may be weeks or months out of date or streaming data ingestion, where the legacy system streams the data to cloud data platforms as the changes occur. These considerations need to be addressed and often require costly workarounds to make them comprehensive.

1.7. Conclusion

The sum of our review suggests that data engineering will continue to grow and evolve in step with the development of a range of data-driven products and solutions that enable users to explore their data, gain and interact with insights, and model their business using AI/ML technology. The rise of generative AI is transforming the cost economics of many industries, in addition to creating an explosion of data content, further accelerating infrastructure, data pipeline, and data solution innovation. Over the last decade, the core underpinnings of data engineering have transitioned into a mature set of domain solutions, capabilities, and products. Globally, a number of new companies are being created to commercialize data engineering capabilities in a wider range of areas.

Organizations in verticals such as retail, travel, ecommerce, consumer services, media and entertainment, and technology that are investing have quickly recognized that investing in their data, and data security, is critical to both customer retention and customer acquisition. Customer engagement is central to revenue growth. The evolving and growing requirements of customers today require organizations to enable their data engineers to further digitize key business functions and processes, moving to a real-time, closed loop decision support model, with high levels of efficiency enabled by the potential for automation. Our experience has shown that investments in support of data engineering capabilities, combined with investments to enable a new generation of citizen developers and knowledge workers across the organization, are capable of unlocking significant incremental productivity. These efforts in turn can further support the revenue growth objectives of a company, and accelerate any organization's digital transformation efforts. This in turn is likely to open up new revenue sources for a number of companies involved in the data engineering landscape over time.

1.7.1. Future Trends

The future of Data Engineering is expected to be shaped by numerous factors, predominantly the growth of Data Science, Artificial Intelligence, Cloud Computing, and the increase in Social Networks and the Mixture of Data Sources. Data Engineering is projected to grow and develop in key areas that will lay the foundation for Data Science and Analytics, facilitating the creation and manipulation of diverse Data Objects. Solutions for Data Engineering will cover additional fields beyond traditional Data Warehousing, allowing discovery, and deeper exploitation of diverse Data Engineering solutions for different data types. Data Engineering encompasses Databases, Data Lakes, Data Pipelines, and actively participates in the construction of AI Solutions and in the management of Cloud Processing Solutions.

Current Data Engineering solutions and the surrounding technology have their origin in the key solutions built from the 1970s until today. Technology is equipped to evolve but faces points of friction. The specialized and complex management of the different Data Object Managers, which are evolving as Embedded and Stand-Alone components and specialized Data Management Solutions, amplifies Data Technical Debt, as well as the quality of the Data Objects, limiting the economic value of Data Science and Analytics. Move-fast and Break-things Data Management approaches are being overcome by concepts derived from Physical Sciences Areas, partially explaining Data Friction. A need to alleviate Data Friction entails the formulation of a new Architecture, or better, several new Architectures of Distributed, Mixed, Hazard Native Data Object Managers, that will encapsulate or hide complexity. This new Architecture or Architectures should

include concepts like complete Data Operations tools, Automation, Assisted-AI, Model Based, and No Code Data Engineering Solutions.

References

Kumar, A., Chakravarty, S., K., Aravinda, & Sharma, M. K. (2024). 5G-Based Smart Hospitals and Healthcare Systems: Evaluation, Integration, and Deployment. CRC Press.routledge.com

Chen, M., Wan, J., & Li, F. (2012). Machine-to-machine communications: Architectures, standards, and applications. KSII Transactions on Internet and Information Systems, 6(2), 480–497.link.springer.com

Chen, K. C., & Lien, S. Y. (2014). Machine-to-machine communications: Technologies and challenges. Ad Hoc Networks, 18, 3–23.link.springer.com

Einsiedler, H. J., Gavras, A., Sellstedt, P., Aguiar, R., Trivisonno, R., & Lavaux, D. (2015). System design for 5G converged networks. In 2015 European Conference on Networks and Communications, EuCNC 2015 (pp. 391–396). Piscataway: IEEE.link.springer.com

Panwar, N., Sharma, S., & Singh, A. K. (2016). A survey on 5G: The next generation of mobile communication. Physics Communication, 18, 64–84.link.springer.com