OULUN AMMATTIKORKEAKOULU

Timo-Joel Piippola

Data Strategy Handbook as Guide Towards Data-Driven Organization

Data Strategy Handbook as Guide Towards Data-Driven Organization

Timo-Joel Piippola
Master Thesis
Spring 2024
The Degree Programme in Data Analytics and Project Management
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
The Degree Programme in Data Analytics and Project Management

Author: Timo-Joel Piippola
Title of the thesis: Data Strategy Handbook as Guide Towards Data-Driven Organization

Thesis examiner: Ilpo Virtanen
Term and year of thesis completion: Spring 2024 Pages: 90

The need for an organizational data culture is evident in the digital era. More organizations are making data-driven decisions, viewing data as a crucial business asset. This thesis aimed to help a case company enhance its data maturity by defining a data strategy, examining data architecture, and exploring tools. The company's leadership showed interest in becoming data-driven, propelling this thesis. The theoretical basis was built on literature and webinars on data strategy, focusing on Data Governance, Data Management, and Data Analysis.

Given that the case company is a rapidly expanding start-up with a lean business team managing operations, the organization's maturity in data and data use capabilities were assessed. This involved an assessment done together with business and support function team leaders, as well as examining the current data architecture and data analytics tools in use. Necessary actions were initiated based on these findings.

The significance of data roles was acknowledged, leading focus to enhancements in data quality and governance in various business platforms. A new strategy of treating data as a product was introduced. A data catalog was selected as the documentation tool for this task.

The development of a data strategy faced challenges like managing multiple projects in a small organization. The focus was primarily on projects upgrading the business platforms. However, these upgrades also improved data analysis capabilities as new features became available in the original systems. For example, the Odoo ERP platform was upgraded, introducing new data analytics and forecasting features.

The data quality significantly improved with the help of the established data and integration team. The next step would be to adopt a Data Mesh approach, involving business teams as accountable data owners and providers, and fostering a mindset of treating data as a product. In the future, appointing a Chief Data Officer or Data Analytics Officer could be necessary to encourage a data-driven culture.

About the use of Artificial Intelligence in this thesis. A Generative AI tool was employed to enhance spelling and grammar. It was also used to find references and assist with ideation. Perplexity.ai served as a companion to generate the Data Glossary introduced in Chapter 7.

Keywords: data, data strategy, data maturity, data-driven organization, data-driven decision making, data culture, data management, data governance, data analytics, data architecture , data literacy, data roles, data as product, data ownership, data fabric, data mesh, Artificial Intelligence, AI strategy, Generative AI

CONTENTS

1 INTRODUCTION

This information era produces a huge amount of data every day, either automatically in processes or manually, by knowledge workers who input data into business systems for everyday operations. Without a systematic approach to data management, you may end up in a chaotic situation where you are unaware of the data you have, who is using it, and where it's located. To leverage your data for business advantage, you need to gather it for analysis and visualize business insights for decision-making. Ultimately, you'll want to ensure that your data is accurate and reliable. (Google 2023).

Accurate and timely data is crucial for making strategic business decisions. Trustworthy data is needed by marketing and sales professionals to understand customer needs. Accurate data is also essential for maintaining inventories stocked and minimizing costs in manufacturing companies. Additionally, compliance officers need to ensure that data is handled correctly and complies with laws. (Google 2023).

This thesis aims to help the case company in data governance and management by creating a data strategy. Goals were set also to enhance the organization's ability to utilize data in decision-making and establish a data culture. The leadership team of the case company leaders has been interested in building a data-driven company and is supporting improvements in this area. This motivation led to the subject of this thesis. This thesis will help any company in a data-driven path by introducing data strategy and its three dimensions Data Governance, Data Management and Data Analytics.

The thesis introduces a handbook-first approach to defining what data-drivenness means, defining the role of data strategy and its three dimensions data governance, data management and data analysis. The thesis aims to provide a foundational knowledge basis using a literature review as a tool and introducing practices in strategic data management for the organization. As a result, data literacy within the organization will improve, enabling individuals to read, understand, create, and communicate data as information. This growth in data literacy occurs, for example, when people actively participate in the data strategy implementation process and observe how data is used to impact the bottom line through real-world examples.

1.1 Research objectives

The primary goal of this thesis is to investigate the elements that define a data-driven company and to develop a company data strategy for enchaining data literacy, data maturity, and overall data management practices withing an organization. Research questions were set to define what data-driven means, what is data strategy, what is company's level of data maturity and what data company have, who owns it and how to utilize it.

1.2 Methods

The theoretical basis was built upon a literature review of both becoming a data-driven organization and defining a data strategy. These topics included the theory section of the data strategy handbook where the audience is introduced to strategic data management practices, technologies, and tools. The data strategy design work was started with small projects and aligned with the company's business strategy and values.

The current state of the organization's data maturity journey was investigated by conducting analysis and workshops with business stakeholders. These stakeholders were asked about the data, and reports they use, as well as about any information they felt was missing. Interviews helped to define the necessary data roles within the organization and highlighted areas for improvement in data literacy and accountability. An additional analysis was conducted to examine the current infrastructure, data access, and tools. This was done in collaboration with the IT team. The team also explored the solution architecture for a new data platform using learning and offering material from three major cloud providers. Then four alternative paths to continue and the project plan were presented to stakeholders.

The organization was introduced to data-driven decision-making using data from the competence management system. The case company operates in the IT consultancy business, with most employees working as consultants on customer projects. These customers demand the availability of skilled consultants at competitive rates. The data was extracted and aggregated, then presented with a hypothesis concerning the need for certain competencies to expand the business. An example is available in Chapter 4.3.

The centralized team established utilized Agile methods and used the Objective Key Results (OKR) method to set goals for each three months in one one-year time frame.

1.3 Introduction of the case company

The case company is a technology firm that develops secure solutions and products for real-time processes and offers related consulting services. While its consultancy business primarily serves the automotive and telecom sectors, it also provides IT consulting for other fields. The company operates globally and has experienced rapid growth over the past few years. In its early years, the organization maintained a relatively flat hierarchy, with a Chief Executive Officer, Chief Technical Officer, Chief Operative Officer, Chief Finance Officer, and a few other business directors. The largest business domains are sales and recruitment, with small Human Resources (HR) and IT departments supporting operations. A couple of industry-focused domains also exists. In seven years, the company's headcount grew from a few dozen to over 400 people.

2 BECOMING DATA-DRIVEN ORGANIZATION

Data-driven decision-making has become a hot topic in the business world in recent years. In 2020, it was reported that 50% of organizations worldwide had implemented data-driven decision-making, a 12% increase compared to the previous two years (Needham, 2022; Taylor, 2022). This development can be attributed to various factors. There is more data available to analyze and gain revenue. There are big data platforms available in the cloud at affordable prices. (Serra 2024 Chapter 1).

The advantage cloud platforms offer contrasts with the poorly executed and expensive master data management projects of the early 2000s and beyond. A survey, to support that those projects didn´t achieve goals, from 2019 revealed that only 31% of organizations considered themselves data-driven, and just 28% believed they had a data culture. These numbers fell even further in a 2022 survey, with only 26.5% of respondents identifying as data-driven and 19.3% stating they had established a data culture. The same survey revealed that 91.7% said they will be increasing investments in data and AI. These results carry weight as they involve 94 well-known, publicly listed companies and leading organizations across various industries. (NewVantage Partners, 2022). The survey shows that data may been collected but not used to gain business benefits.

2.1 Data-driven organization

A data-driven organization treats data as an asset. Instead of keeping data in silos or as the property of individual departments, it is made accessible and securely available to anyone who needs it. Data is actively utilized in analytics and Machine Learning applications to facilitate better decision-making, improve efficiency, and drive innovation. (Amazon AWS training, 2022).

Data-driven organization utilizes data across the company. The capability of understanding data and providing insights based on data is spread to all organization levels. A data-driven organization can make well-informed and data-driven decisions to continuously improve the company's capability to grow and compete in business. For example, without data-driven decisions, Netflix would be gone by staying in the DVD-sharing business without growing to Internet streaming and without becoming one of the known and successful companies. It is the same situation with Amazon which

started as an online bookstore and applied the model to other products and finally utilized investments in their platform's infrastructure by selling cloud services to everyone. (Joubert 2019).

Data-drivenness is about building tools, abilities and culture that use and acts on data. A data-driven organization involves three dimensions: data, technology, and people & culture. These are illustrated in the Venn Diagram below (Figure 1). Each dimension is crucial for becoming a data-driven organization. The growth of data-drivenness at the center of these dimensions signifies that all data initiatives should consider each dimension. Without technology timely decision-making is impossible. Without data, informed decisions cannot be made. Without a people and culture that utilize data and tools, you are left with unused systems and tools. Each dimension should be strategically tailored to the organization's needs. (Anderson 2015, Chapter 1. Allouin 2022)

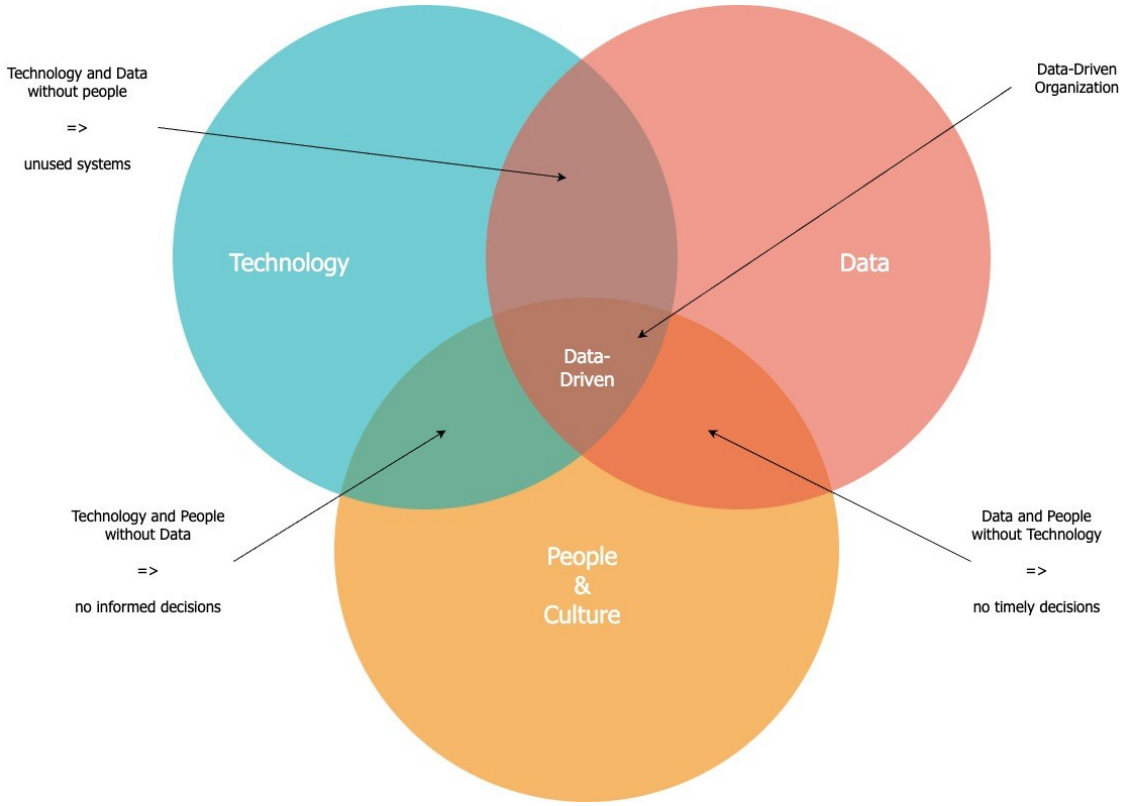

Figure 1 A Data-driven organization has three dimensions (Adapted from Allouin 2022)

Chapter 4 introduces the approach to becoming a data-driven company using Data Strategy and its three dimensions: Data Governance, Data Management, and Data Analytics strategies and how data, technology and people & culture are part of the strategic approach to Data.

The next chapters discuss the concepts of data and data-driven decision-making. Furthermore, the need for Data Literacy and Data Culture in the people dimension is explained to progress in the Data Maturity journey.

2.2 What is data

Data is often referred to as the new gold or oil of the 21st century. More data is being collected and stored than ever before, often for future use or regulatory compliance. Each person generates 1.7 MB of data every second. Data surrounds us, and in this era of cloud computing, affordable solutions are finally available to collect, prepare, analyze, and utilize data in decision-making processes. This also includes Generative AI solutions built on data. Data consists of observations and facts presented in forms such as numbers, words, pictures, audio files, videos, and maps. It can be collected from transactional systems as structured data, or from external systems, wearables, sensors, and devices as unstructured data. (Klidas, Hanegan 2022). Essentially, structured data is stored in databases while unstructured data is kept in file-based storage. Data-Information-Knowledge-Wisdom (DIKW) model (Figure 2) is commonly used to describe raw data transformation into valuable insights (Cotton 2023, Rowley 2007).

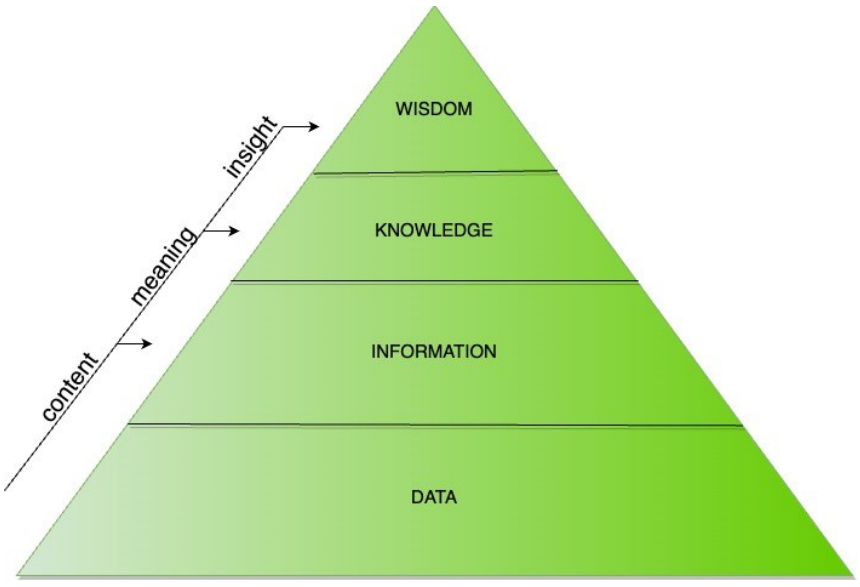

Figure 2 DIKW pyramid. (Adapted from Cotton 2023, Rowley 2007)

In the DIKW model, data serves as the starting point. It begins as unstructured raw facts and figures without context and is refined into information by providing context, structure, and purpose. Analyt-

ical methods are used to give this information meaning and usefulness, enabling informed decisions based on data. The information stage asks questions like 'who', 'what', 'where', and 'when'. During the knowledge phase, the information is analyzed, and learnings are applied to experiences and relationships with other information. The questions at this stage involve 'how' and 'why'. Wisdom is viewed as the ability to make strategic actions based on insights and learning from data. Decision-making no longer looks to history, instead, it defines your future with well-informed, data-driven decisions. (Cotton 2023, Rowley 2007).

2.3 Data-driven decision-making

All decision-making is data-oriented, even those based on gut feelings. We draw on our experiences and beliefs. For instance, choosing what to wear based on current and forecasted weather is a data-driven decision. By gathering and analyzing data, we can gain new insights to make better decisions.

In a company, data-driven decision-making is a process of making organizational decisions that are backed by actual data, rather than relying solely on intuition or observation. It involves using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives. While many data-driven approaches may seem profit-driven or greedy, focusing on maximizing revenue rather than improving the customer experience, they should align with the organization's values and strategy. Concentrating on short-term profitability rather than long-term goals might neglect the environment and humanity. Thus, a strategic approach to data collection and analytics can ensure the right data is collected, and the results align with predetermined objectives. (Rouse 2018, Joubert 2019).

Data quality is fundamental to data-driven decision-making, which can enhance outcomes in both business and real-life situations. Decisions based on flawed data can lead to significant financial losses and missed opportunities. Poor data quality can also result in service delivery delays and financial reporting inaccuracies. Furthermore, machine-made decisions based on low-quality data can create hazardous situations in autonomous traffic or automated factory environments. (Anderson 2015 Chapter 2).

2.4 Data literacy

> *"Data literacy is the ability to read, write and communicate data in context, including an understanding of data sources and constructs, analytical methods and techniques applied, and the ability to describe the use-case application and resulting value." (Panetta & Gartner 2021).*

Everyone has skills in data literacy. In our daily lives, we see and hear signals of data and use them in decision-making. By improving our skills in data literacy, we can act better, reduce noise, and focus on what is important. Poor data literacy is said to be the second-biggest internal roadblock to success according to the Gartner annual Chief Data Office Survey. Gartner continues that by 2023, data literacy will become essential in driving business value (Klidas & Hanegan 2022 Chapter 1; Panetta & Gartner 2021).

In an organization data literacy journey should be in line with the data maturity journey and business objectives. A training program is needed as well as an assessment to identify current maturity and skills among business analysts, data stewards and architects. Also, skilled translators are needed to act as mediators for business domains. Data literacy training could be gamified through a company learning platform to make it as fun as possible. (Klidas & Hanegan 2022 Chapter 3; Panetta & Gartner 2021).

Matt Crabtree (Crabtree 2023) presents the Datacamp data competency framework (Figure 3) it divides data skills into foundational data literacy skills: reading data, working (writing and analyzing) with data, communicating with data, and reasoning with data.

Figure 3 The Data Competency framework by Datacamp (Crabtree 2023)

Data literacy journey and learning needs sponsorship and support from data leaders like the Chief Data Officer, Chief Information Officer or whoever is responsible for data in the organization. Becoming a Data-Literate organization is more than adopting tools or technologies. It requires a culture that sees data as an asset and encourages people to use and analyze data and utilize it in data-driven decision-making. (Crabtree 2023).

2.5 People in data-driven organization

Data-driven organizations employ individuals skilled in extracting and interpreting data to produce useful metrics. These individuals create data products such as dashboards for reporting and analysis, identify key performance indicators, and set alerts for when these metrics are triggered. Coupled with business domain experts who pose the right questions to data, these teams can drive significant organizational insight. However, to truly harness data-driven strategies, a shift in organizational mindset is needed. According to the NewVantage survey conducted in 2022, only 26.5% of respondents were identified as data-driven, while 19.3% reported having established a data culture (NewVantage Partners, 2022). (Anderson 2015 Chapters 4 and 10).

2.6 Data-driven culture

The importance of establishing a data culture in an organization together with a strategical approach to data management and governance in this modern, digital world is clear. In 2020, it was

reported that 50% of organizations worldwide were implementing data-driven decision-making, a 12% increase from two years prior (Needham, 2022; Taylor, 2022). Data needs to be governed to comply with data regulations and ensure the ethical and right use of data collected but also have an open culture for data utilization and data-literate organization.

> *A data culture is not just about deploying technology alone, it is about changing culture so that every organization, every team, and every individual is empowered to do great things because of the data at their fingertips. (Satya Nadella 2014)*

Cultivating a data culture begins at the top. Senior managers must lead by example, utilizing data to support their communication and decision-making. When leaders habitually use data, it encourages the wider team to follow suit, fostering a data-driven environment. Leaders can also influence behavior by carefully selecting metrics that enable the organization to make predictive business decisions related to future needs and direction. Also, data-driven initiatives need sponsorship and support from top management. (Bean & Davenport 2021 Chapter 5, Waller 2020).

The next step involves promoting data literacy activities. This means educating on what data is, where it can be found, and how it can be utilized. Common ground in data language should be established by using the company's own data glossary. Learning happens when people are given access to data and tools to analyze data, and by promoting a Proof-of-Concept attitude towards data-driven initiatives. This not only fosters innovation but also helps develop fundamental skills such as coding and analytical tool proficiency. Encouraging employees to use data independently allows them to find more efficient work solutions without relying solely on IT department support. They can utilize strategically chosen platforms and tools delivered by the company to drive their projects in data and analytics. Learnings and findings are shared to promote further data literacy across the company. (Bean & Davenport 2021 Chapter 5, Waller 2020).

Lastly, organizations should promote the habit of storytelling with data. Data-backed decisions create a narrative explaining what happened, why it happened, and what is predicted to occur. With an analytical mindset, alternative solutions are explored, and decisions are made data-driven and well-informed. (Bean & Davenport 2021 Chapter 5, Waller 2020).

2.7 Data maturity

Data maturity is a critical measure of an organization's ability to use data for decision-making. It reflects an organization's progress towards becoming a data-driven company. Various data maturity models can evaluate a company's data maturity. A data maturity model can serve as a framework for initiating improvement projects within a company. Understanding the starting point is crucial for successful strategy implementation (Wallis, 2021, Chapter 3.2).

Based on the literature review, commonly used data models are the Dell Data Maturity model, the IBM Data Governance Maturity Model, and the DAMA Data Management Book of Knowledge 2.0. Most of these models follow the Capability Maturity Model (CMM), resulting in similarities between them. The choice of the right model depends on industry and cost (Firican 2020 1.).

Since the DAMA model is commercial, it will not be covered here. Instead, the Dell Data Maturity model and a similar model from heap.io will be introduced as complementary models. The IBM model will also be briefly mentioned, with a source provided for further information.

2.7.1 Dell data maturity model

Dell Data Maturity mode consists of four levels: Data Aware, Data Proficient, Data Savvy and Data Driven (Figure 4).

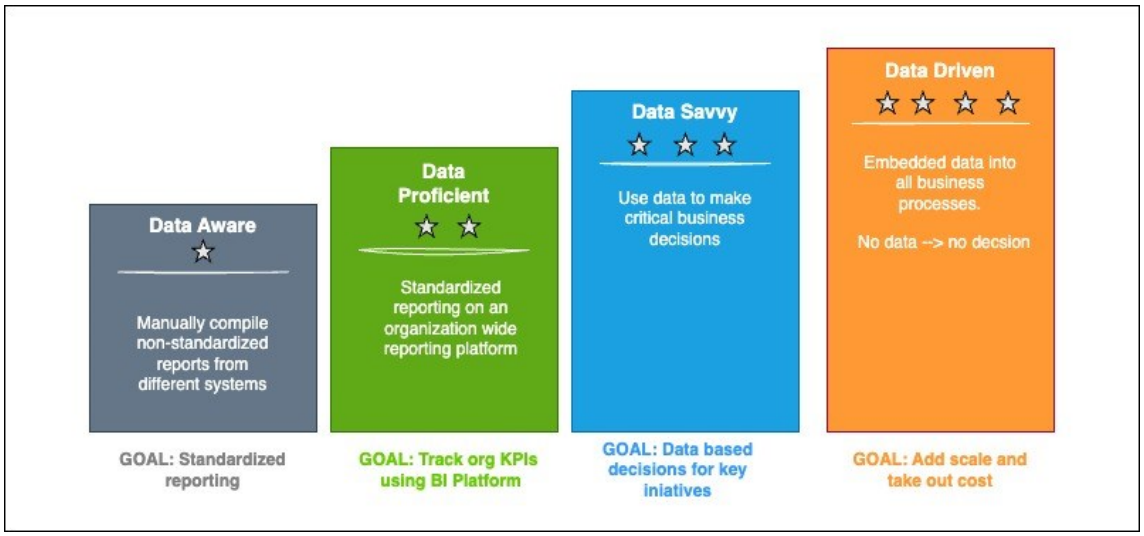

Figure 4 Dell Data Maturity Model (Adapted from Onis 2016.)

The Dell Data Maturity Model outlines steps to ascend the maturity curve. In a Data Aware organization, non-standardized reports are manually compiled from various systems. There is a lack of data aggregation and application integration. The goal should be to standardize reports as customized ones may lead to data being used merely to justify decisions post hoc. To advance to the next level, an organization should

- develop a data model reflecting the company's value chain.
- consolidate data into a single datastore.
- standardize the reporting system and create dashboards.

At the Data Proficient level, data quality improves, and dashboards are available. Challenges still exist with incomplete data collection in a single data warehouse and a lack of application integration. Organizational KPIs are tracked, and the organization is ready to pilot a data initiative. However, there may be a lack of executive sponsorship and skills to handle or use unstructured data. Business demands may exceed IT department capabilities or capacity. (Onis 2016).

Focus areas at the Data Proficient level should be data quality, application integration, and speed of data provision to users. This requires action on ETL data pipelines and data warehouse optimization. The goal should be to serve standardized reports on a centralized platform accessible company wide. Establishing a Master Data Management strategy is key to advancing to the next level. (Onis 2016).

Data Savvy organizations use data for critical business decisions. The business and IT departments collaborate under leadership to minimize data fragmentation. Data is viewed as an asset, and new technologies are integrated across all data sources and applications for data production, storage, and analytics. (Onis 2016).

The Data-Driven level implies a "no data – no decision" approach. The goal is to extend data-driven practices to all organizational levels using a data strategy while continuing to reduce costs. In such an organization, all users leverage data analytics as a self-service on strategically chosen platforms. (Onis 2016).

2.7.2 Simple data maturity model by heap.io

Rachel Obstler (at heap.io) introduces a simpler data maturity model. The four levels of the data maturity model introduce four levels where each level is defined by the combination of strategic, operational, and cultural data-driven practices.

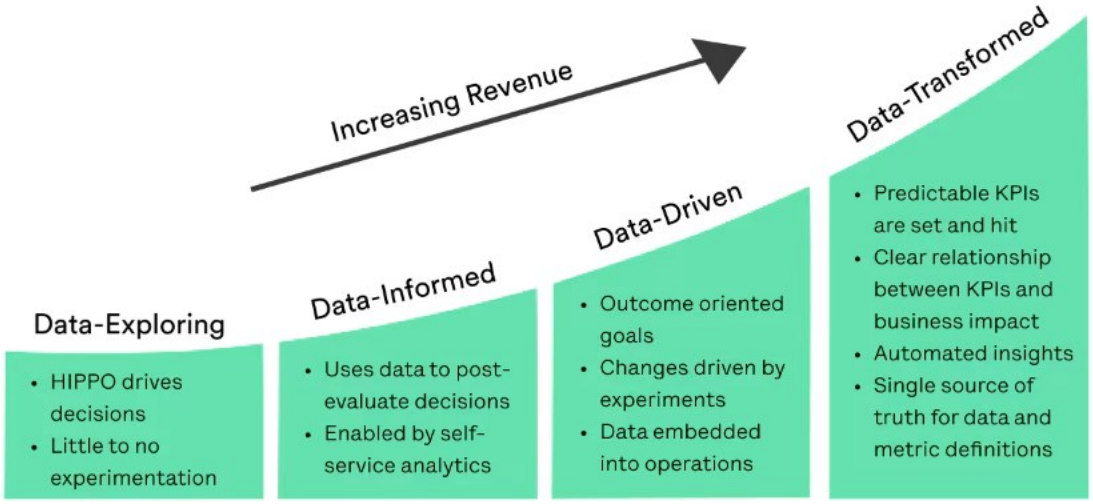

Figure 5 Four Levels of Data Maturity (Obstler 2023.)

To determine where the organization is in its data maturity journey Obstler has a few questions to ask:

- *How do you measure the success of digital projects?*
- *Are data and analytics easily accessible to the team? How quickly can they find and analyze data to answer their questions about product performance?*
- *How does your organization connect product changes to business performance?*
- *How does your organization experiment with new ideas?*
- *Does your team effectively use both qualitative and quantitative data for analysis?*

Obstler suggests that to get started, make data accessible in real-time through self-service and train people in Data Literacy (Obstler 2023).

2.7.3 IBM data governance maturity model

IBM Data Governance Maturity Model published in 2007 introduces five levels of maturity (Figure 7) and works as a framework for data governance. (Figure 8). Firican introduces the model well in his blog. (Firican 2020. 2)

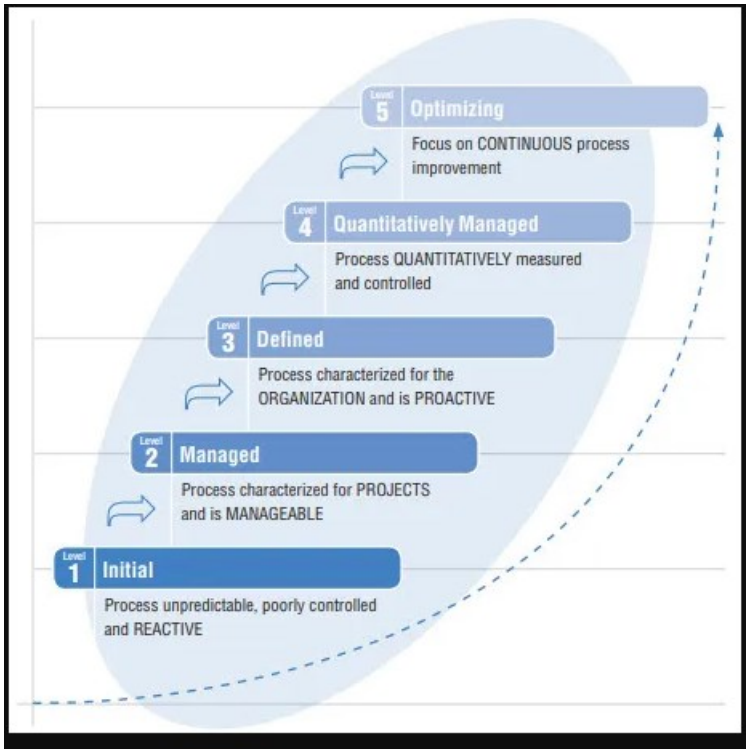

Figure 6 IBM Five Levels of Data Governance Maturity (Firican 2020 2.)

The model introduces the 11 data governance domains (Figure 7) and helps assess and measure progress in each.

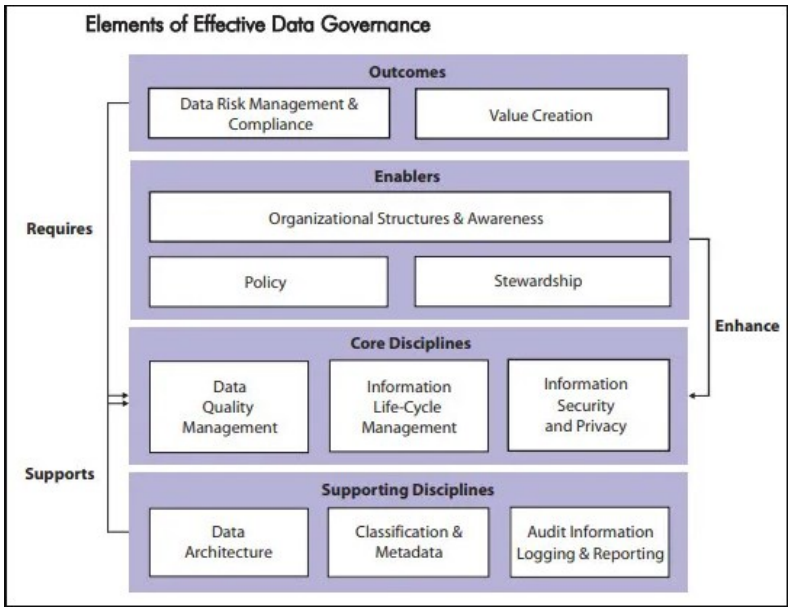

Figure 7 The Elements of Effective Data Governance (Firican 2020 2.)

2.7.4 Benefits of using a data maturity model

Using a data maturity model as a framework to evaluate an organization's data maturity has two main advantages. Firstly, it provides a roadmap to plan for progress to higher levels. Secondly, by examining an organization's capabilities and data usage, it identifies potential cost savings and areas to increase revenue and improve decision-making. Dell, for instance, reported numerous benefits from this approach. They reduced their expenditure on BI by 50%, boosted predictive analytics capabilities by 20%, eliminated non-standardized reports and KPIs, and increased revenue by millions of dollars. They also sped up data-driven decision-making and reduced IT development costs by 25% (Onis, 2016). These results align with other sources advocating the strategic approach to data management and data-driven decision-making.

The next chapter introduces a common story in the data maturity journey, as elaborated by the thesis writer.

2.7.5 Common story: awakening to data awareness

Most people have seen Excel spreadsheets filled with numbers and some graph visualization. This is typically how organizations aim for data-driven decision-making. Start-up companies often seek Software as a Service (SaaS) solutions for business operations, such as payroll, sales, HR, and

other processes. As the business grows, a Customer Relationship Management (CRM) tool is acquired, which may complement or overlap the existing SaaS services.
If data can be exported, it is usually done so into Excel, where complexity is managed using pivot tables. Eventually, the need to centralize all data in a data warehouse for cleaning and analysis is recognized, leading to the setup of a cloud platform. The process of building this single source of truth can take months, with work and maintenance often outsourced to another company.
In the meantime, someone might discover a business insight tool, typically a desktop-only solution, to replace Excel. Sometimes, this tool is even connected directly to the source databases.

The steps in this data-driven path are usually driven by management's need for monthly reports. These efforts often depend on a single individual in finance or IT to gather and present the data. While Key Performance Indicators (KPIs) are set and monitored, the insights are usually retrospective and may not directly link to the company's strategy and goals.

This ad-hoc approach can lead to increased costs and potential chaos with data accuracy and reports. Therefore, a strategic approach to improve data maturity is necessary.

2.8 Common challenges in becoming a data-driven organization

NewVantage Partners' annual survey tracks the progress of corporate data initiatives. The survey reveals that many companies have been striving to become data-driven for years, but success has been mixed. Cultural changes pose significant challenges, and this is no exception when organizations are trying to adapt to being data-driven. (Joubert 2019).

The amount of data collected is growing exponentially. The survey indicates that only 26.5% of organizations have established data-driven operations. 91.9% of executives state that cultural roadblocks are the main barriers to becoming data-driven. Only 40.2% of companies report a well-established and successful Chief Data and Analytics Officer role, which is considered foundational for becoming data-driven. (NewVantage Partners 2022).

The growth of collected data introduces governance and compliance challenges. Without strategic actions, the ownership, quality, responsibility, and ethical use of data remain unclear (Bean 2022; DalleMule, Davenport 2017).

Common challenges also include data quality and access to data in legacy systems. These systems often lack well-defined APIs, and data may be entered without a validation process to ensure its correctness. Data quality is not just a technical issue; it requires a commitment to a cultural shift in processes where data originates. This cultural change also ensures data privacy and security so that only authorized individuals have access to the appropriate data. (Precisely 2024).

To become a data-driven organization, a systematic and strategic approach is needed to enhance processes, data literacy, cultural change, and organizational commitment to a strategical approach to data, technology, and people & culture building blocks of a data-driven organization. In the next chapters, a Data Strategy is introduced as the next step to build data-driven organization capabilities in Data Governance, Data Management and Data Analytics.

3 DATA STRATEGY

While the term 'strategy' has its roots in the military, it has been adapted for business since the 1960s. In 1979/80, Michael Porter defined strategy as the broad formula for how a business is going to compete, what its goals should be, and what policies will be needed to carry out those goals. It is the combination of the ends (goals) for which the company is striving and the means (policies) by which it seeks to get there. (Wallis 2021, chapter 1).

Data is essential for formulating a business strategy. As outlined by Tienari and Meriläinen (2021), the process of building a business strategy comprises five steps: 1) collecting and analyzing data, 2) defining a strategy, 3) planning projects, 4) scaling the strategy into actions, and 5) regularly measuring, monitoring, evaluating, and updating the strategy. Notably, data collection and analysis form the foundation. This is crucial for data-driven decision-making also, as introduced in Chapter 2.

Data strategy is a highly dynamic process that supports data collection, organization, analysis, and delivery to align with business objectives (Gartner 2023). Data strategy should help businesses deliver value, achieve primary goals, and comply with laws and regulations. Organizations must find a balance in their data strategy, considering the offensive or defensive nature of data use and the trade-offs between control and flexibility. This balance will influence how data is viewed as a strategic asset in the future (DalleMule, Davenport 2017).

A strategic data plan, or data strategy, is essential for using data effectively to achieve organizational goals. From a defensive perspective, data usage must comply with laws, regulations, and the organization's policies, such as information security policy. This alignment ensures ethical data usage and protects sensitive business information, personal data, and data assets from unauthorized access. This is how organizations in industries with high regulations, such as healthcare, insurance and finance tend to focus their efforts on data. From an offensive perspective, the data strategy should support the organization's growth and competitive goals. It must align with the company's vision, goals, and objectives. Organizations in the commercial industry usually have a more offensive approach to data since they look more often for ways to monetize the data they collect and try to predict customer behavior. (DalleMule, Davenport 2017; Wallis 2021).

Balance between defensive and offensive as well as control and flexibility usage is accomplished through a strategic approach to data strategy (DalleMule, Davenport 2017). In this thesis, key elements of a data strategy are presented through these three dimensions of a data strategy - data governance, data management, and data analytics. Their relationship is presented in a Venn diagram in Figure 8. Data strategy seeks value from data but also sets objectives for how data is governed, managed, and used.

Data Governance include processes, roles, policies, standards, and metrics that ensure the effective, efficient, and secure use of data. Policies define the terms and standards for data usage. In a data-driven organization, data governance gives guidance for people and processes on how data is managed and used securely with quality in mind. (Qlik 2024, Google 2024, Data Centric Inc 2021).

Data Management focuses on implementing collecting, organizing, protecting, and storing an organization's data so it can be analyzed for business decisions in a data-driven organization data is harvested with strategically chosen technologies. (Strengholt 2023 Chapter 1, Data Centric Inc 2021).

Data Analytics focuses on using data and technology to support business objectives. This can involve descriptive analytics ("What happened?"), diagnostic analytics (Why it happened?) and predictive analytics (What will happen?). Results are visualized in dashboards and reports together with Key Performance Indicators (KPI). (Edx 2023).

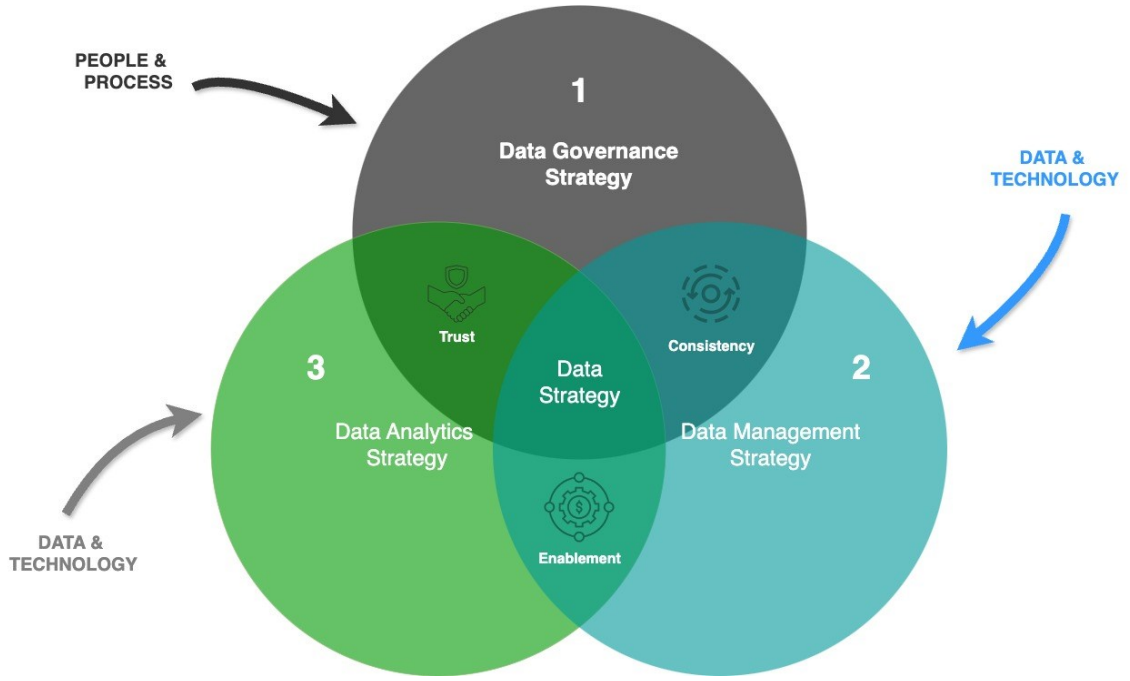

Figure 8 Three Dimensions of Data Strategy (Adapted from Curry 2018, Data Centric Inc 2021.)

Data governance ensures *consistency* in treating data as an asset, maintaining quality and adherence to policies. Data management and technologies *enable* data exploration in analytics, building *trust* in the data through alignment with data governance. These dimensions are equally important. Data governance without data management is simply paperwork. Data management without data governance leads to insecure and non-compliant data chaos. Without data analytics, well-informed and data-driven decisions cannot be made. (Data Centric Inc 2021, Allouin 2022).

The three dimensions of data strategy are represented in the enriched data-driven organization Venn diagram, as seen in Figure 23. This was introduced in Chapter 3.5 as a communication tool during the data strategy design phase.

The next chapters will introduce these three dimensions of data strategy.

3.1 Data governance strategy

Data governance includes processes, roles, policies, standards, and metrics that ensure the effective and efficient use of data. It helps organizations meet strategic objectives that depend on data, by defining who can act on which data, under what circumstances, and using which methods. (Data Governance Institute 2024).

The benefits of data governance include a shared understanding of data, improved data quality, and compliance with regulatory requirements. For example, a well-defined data governance framework can help satisfy the demands of government regulations, such as the EU General Data Protection Regulation (GDPR). Since violations of GDPR can result in fines, establishing processes and policies around data can save a significant amount of money. (Data Governance Institute 2024).

Creating a Data Governance strategy involves several steps such as securing an executive sponsor, forming a data governance committee, establishing an implementation team, and creating a data management unit for enforcement. Also, it includes aligning data governance activities with strategic objectives and implementing a data governance framework. (Data Governance Institute 2024).

The Data Governance Institute presents the Data Governance Framework (Figure 9), which introduces data rules, role assignments, and processes to align everyone within the organization. The development of a data governance strategy starts with identifying the 'why', usually including one or more value statements forming the program's basis. The goal is to deliver value. Next, it is crucial to identify 'what & how', the outputs, processes, and data products delivered by data governance. Tools such as data catalogs, definitions, and metadata management facilitate data discovery within the company. The 'how' component is tackled through policies, rules, guidelines, and guardrails, while decision rights dictate who makes decisions, when, and based on what criteria. (Data Governance Institute 2024).

The framework also introduces a data governance office with key roles such as Data Stewards, who are embedded within business and compliance functions, and Custodians, who are part of technology and data management functions. The 'whom' component describes the beneficiaries. These can include the organization's products, services, processes, capabilities, and assets. For instance, data-driven decision-making benefits outputs like well-defined data products like dashboards and data catalogs. (Data Governance Institute 2024).

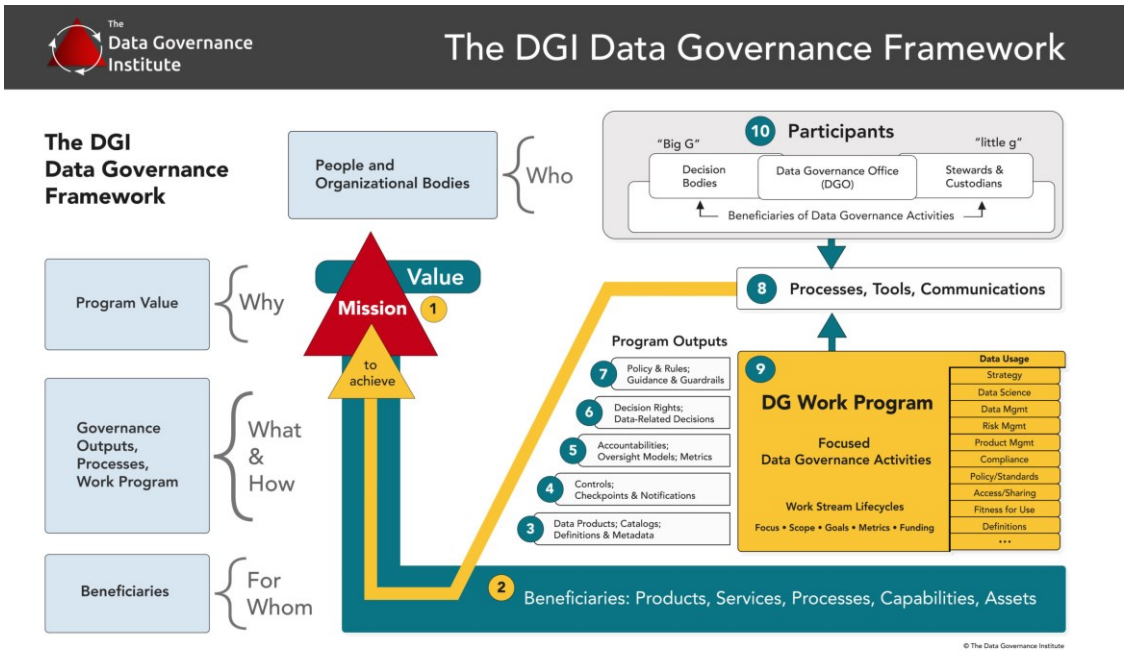

Figure 9 Data Governance Framework by Data Governance Institute (Data Governance Institute 2024.)

3.1.1 Roles and responsibilities

Data governance is unique to each organization. Strategic approach crafts data governance activities with organization-specific demands. Also, each organization has its structure and titles for roles acting in data governance responsibilities. Defined data governance roles are essential, and it is essential to assign ownership on the corresponding levels. The goal is that everyone's responsibilities are clear, to achieve the best results (Treder 2020 Chapter 9). Below are the most common roles and responsibilities in data governance introduced.

3.1.1.1 Chief data officer

A chief data officer (CDO) is responsible for overall data strategy and data management practices. This person is responsible for that data strategy vision, goals and standards are aligned with business strategy and organization values as well as with the goals and objectives. The role requires both technical and business-oriented skills with a focus on diplomatic and communicative skills. The challenge for CDO is to get people, both executives and employees, engaged. (Treder 2020 Chapter 9).

3.1.1.2 Data owner

A data owner or a process owner is a person in the organization who is accountable for the data processed and stored in a business domain context. This person oversees data quality, definitions, classifications, and where data is used. In an organization, there are several Data Owners. (Strengholt *202*3 Chapter 8).

3.1.1.3 Data steward

A company's data steward typically has a leadership role within the data governance team and is responsible for making final decisions in data governance policies. Data stewardship focuses on tactical coordination and implementation. The role requires both technical and business-oriented skills, like programming, data modelling and technology as well as strong communication and collaboration skills (Pratt, Luna 2024). A data steward is often a subject matter expert for a specific type of data where the Chief Data Officer has overall responsibility (Strengholt *202*3 Chapter 8).

3.1.1.4 Data custodian

An application owner, also known as a data custodian, is in charge of maintaining the application's core and its interfaces. They are responsible for business delivery, operation, and services. Additionally, they oversee the maintenance of application information and access control. Having a separate application owner may be useful when several business domains use the same application or parts of it. For example, the company ERP platform has several modules for many use cases like Sales, Recruitment, Finance operations, Human resources, Inventory etc. Application owner is responsible for coordinating the whole platform lifecycle and data owners are responsible for specific data in the application. (Strengholt 2023 chapter 8.)

3.1.1.5 Other roles and organization processes

Strengholt (2023, Chapter 8.) introduces in his book Data Management at Scale a high-level data governance framework breaking the organization into domain teams using different roles (Figure 11). No doubt it has been influenced by the Data Mesh concept of federated governance and data as a product which will be introduced in its chapter later.

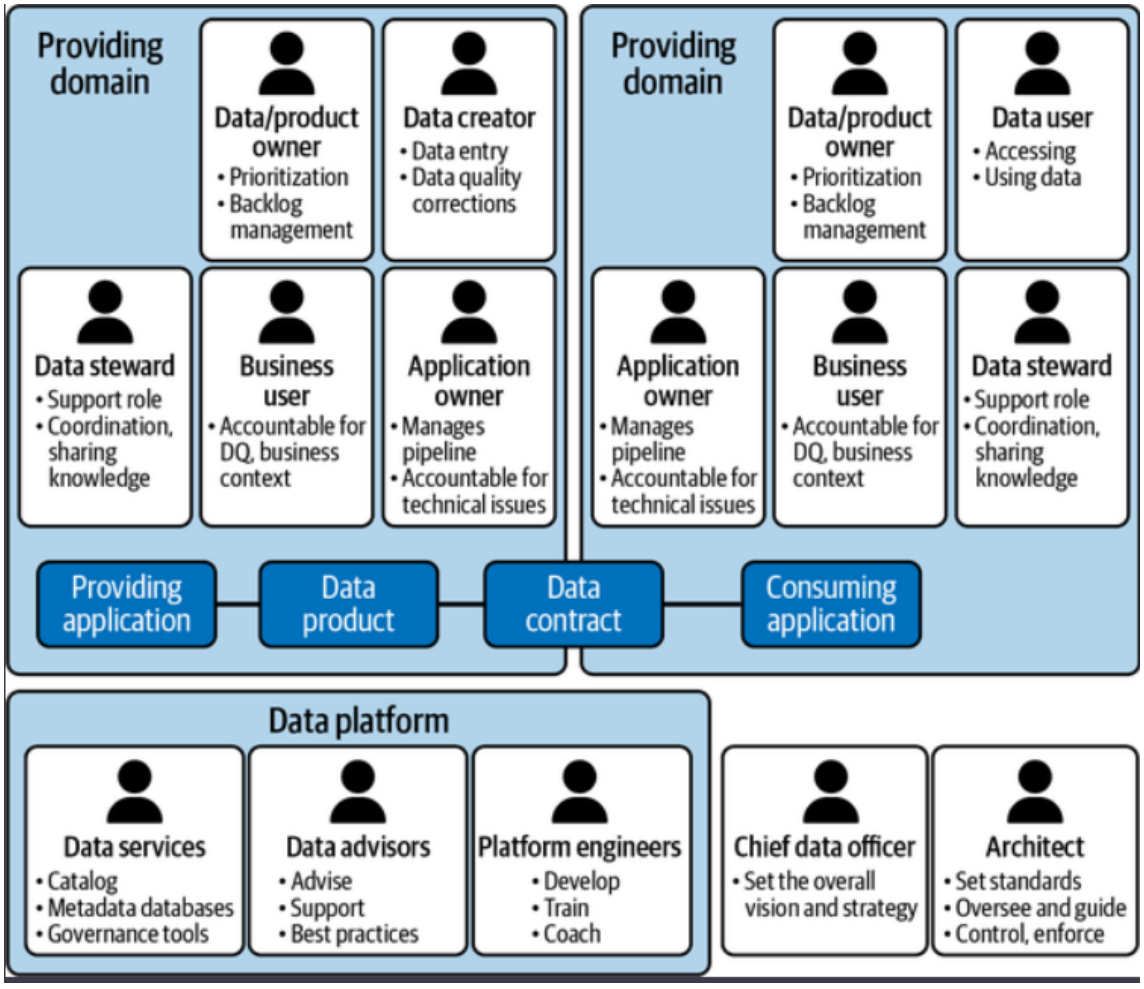

Figure 10 A High-level data governance and organization with different roles in business domains. (Strengholt 2023 Chapter 8.)

Instead of Data Mesh decentralized approach, Strengholt has a hybrid model in his Data Governance framework where Data Platform and roles like Data engineers and advisors are centralized functions as well as Chief Data Officer and Architect act at the company level. Business domains have their own roles responsible for their data and business processes. Close collaboration is required to get data from other business domains and data contracts between domains created to commonly understand what purpose that data will be used. (Strengholt 2023 Chapter 8).

Treder (2020, Chapter 9.) in his book, The Chief Data Officer Management Handbook: Set Up and Run an Organizations Data Supply Chain, divides an organization into two groups. Data Owners, who are members of a business team in a particular area and responsible for the entire data domain across all business functions. For example, Data owner of "Customer". The other role is Data

Champions member of one single business function in any data discussion, like business users in the Marketing unit.

Then Treder (2020, Chapter 9.) introduces two stakeholder groups Data Creators and Data Consumers. Data Creators maintain data on a daily basis. They are responsible for the content and quality of data. Data creators include data stewards as a subgroup. Data Consumers are users and customers of a data product. This group includes Data Analytics people and business processes like HR, Sales, Finance, Customer service as well as performance and regulatory reporting. Treder also introduces the Chief Data Officer, Business application and process owners as part of the data governance organization.

3.1.2 Data governance tools

Data Governance tools are essential for organizations aiming to monetize data or comply with laws and regulations. They help identify the data an organization owns and its location. These tools facilitate data discovery, data classification, policy creation and management, and metadata maintenance. (Strengholt 2023 Chapter 8.)

Implementing tools like a data catalog and data standard is necessary. With them, establish policies for handling sensitive information like Personally Identifiable Information (PII), names, phone numbers, and credit card numbers. While these details may be required in operational business systems, they should be redacted or censored when the data is transferred to central storage for storage and analysis. Data models like conceptual, logical, and physical data models, are tools to set common ground of concepts and communicate across an organization's domains. Other tools to concern are data scanners and profiling tools to examine the level of data quality, schema information, and transformation and lineage information. These functionalities are part of the Data Catalog or dedicated offerings. Metadata repositories are tools to help users find and manage information about the data. (Strengholt 2023 Chapter 8.)

The subsequent chapters will provide a brief introduction to data catalog, data standardization and data modelling.

3.1.2.1 Data catalog

A data catalog provides a comprehensive view of an organization's data assets. It serves as an inventory located in one easily accessible place and uses metadata to describe the location, security levels, and content of each dataset. It is designed to assist data professionals in quickly locating data for any analytical or business purpose. (IBM 2024 1).

The data assets included in the data catalog can be for example the following:

- Structured (tabular) data
- Unstructured data, including documents, web pages, email, social media content, mobile data, images, audio, and video
- Reports and query results
- Data visualizations and dashboards
- Machine Learning models
- Connections between data sources

Tools data catalog should provide are search datasets and data products, automation to discover potentially relevant data side-by-side search results and information about policies, rules, and compliance to follow with data. (IBM 2024 1).

The benefits of using a data catalog are in its capability to offer self-service inventory to organization data. It increases operational efficiency when data professionals can access and analyze data faster without interfering with centralized IT team or business domains. Also, all data sets have policies on who can access the data and rules on how long data can be stored to maintain compliance with data regulations. (IBM 2024 1).

There are several data catalog tools available. There are open-source tools like Apache Atlas and Amundsen. Commercial data catalog tools may be included in business insights tool like Tableau, or in offerings in Azure, Google Cloud and Amazon AWS cloud platforms.

3.1.2.2 Data standardization and data models

An underrated tool in data governance is data standardization. This tool describes data in a more detailed manner. The depth of data standardization relates to the attributes of database tables and the data types they use. This information is crucial when performing transformations in data preparation for analytics or ensuring interoperability.

Data standards also describe the conceptual relationships between data sets, such as the connection between an address entity and a customer entity. Conceptual and logical data models facilitate communication within an organization, establishing a shared understanding of data, such as the attributes of a person entity across different business systems. Where the Physical data model describes how data is structured in a database. (Perälä 2018).

Data standardization establishes common attributes among data entities in various locations, such as databases in distinct business systems. It details the extensions an entity has in different applications, going beyond common attributes at the business domain level. Essentially, data standardization promotes uniformity of data across an organization. (Perälä 2018).

A data standard document is a system-level guide used within an organization or business for enhancing operation efficiency and enabling interoperability. Industry-wide data standards also exist to facilitate interoperability among different organizations within the same sector. For instance, the Finnish public administration shares data models using the suomi.fi Data Vocabularies Tool, which can be found at http://tietomallit.suomi.fi (Digital and Population Data Services Agency 2024). The healthcare industry uses Health Level 7 (HL7) global standards to transfer clinical and administrative health data between applications (hl7.org,2024). (Perälä 2018).

Data standardization and cataloguing not only improve self-service and operational efficiency but also ensure compliance with regulations and laws, such as the European Union's data acts and the General Data Protection Regulation (GDPR). Data standards and data modelling are vital for enhancing interoperability and communication within an organization. A prime example is the handling of Personally Identifiable Information (PII). Laws and regulations dictate how to store, share, and retain PII. In an organization, a 'person' entity can be named differently based on the business context. For instance, a 'person' could be an employee, customer, job applicant, or subcontractor in a partner organization (see Figure 11). These entities share common attributes such as first

name, last name, email, and phone number but the data type and format of how attribute is presented may vary by business system. Therefore, it is essential to identify the master data source in all cases and determine if there is a need for standardizing the data types of these entity attributes and apply data integration between systems.

Integration between business systems for operational purposes can be automated to achieve operational efficiency by updating data across all systems simultaneously. For example, in a project sales scenario, you may have a list of candidates who are both employees and applicants without a contract. Keeping the contact information and availability status of these candidates up to date is crucial to match talents into sales cases.

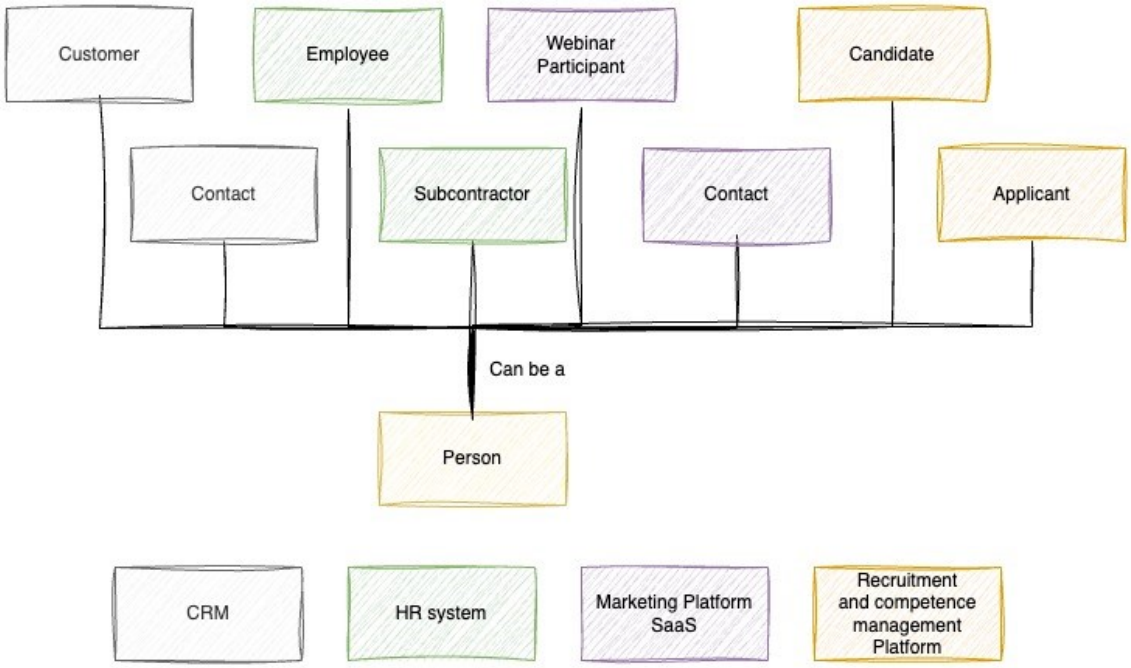

Figure 11 A Conceptual Model of a Person

3.2 Data management strategy

Data management refers to the practice of collecting, organizing, protecting, and storing an organization's data so it can be analyzed for business decisions. Effective data management is crucial for driving value from data and achieving strategic objectives. (Tableau 2024 1.).

Data management is an important part of data strategy. Organizations that collect, store, and analyze customer behavioural insights outperform peers by 85% in sales growth and more than 25% in gross margin (McKinsey 2017).

Effective data management:

1. **Improves decision-making**: Well-managed data supports organizations to make informed decisions based on accurate, trustworthy, and up-to-date information.

2. **Enables innovation**: Efficient data management fuels innovation by supporting advanced analytics with Machine Learning and Artificial Intelligence

3. **Reduces risks**: Robust data management mitigates risks associated with data loss, theft, and misinterpretation.

4. **Boosts competitiveness**: Effective data management enhances organizational agility and resilience, helping companies stay ahead of competitors.

5. **Supports compliance**: Comprehensive data management promotes compliance with legal requirements and industry standards. (Tableau 2024 1.).

The Data Management Survey 24 lists the main benefits of data management are increasing value from data (40% achieved this to a high degree) and improved decision support (39%).

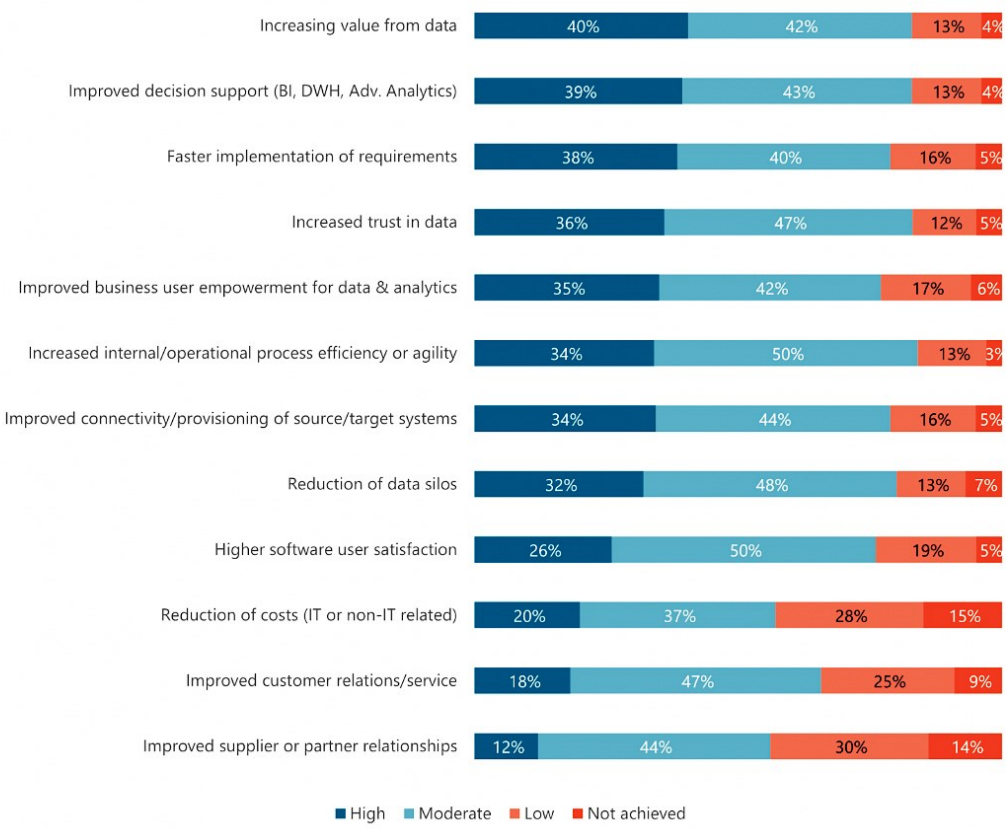

Figure 12 Business benefits achieved using data management software (n=664) BARC GmbH 2023.)

A strategic approach to data management pays off for example in improving the data supply chain which means how data is captured, acquired, managed stored, transformed, shared, and then consumed. For example, it is said that correcting data is 10x expensive after data is collected and stored to be prepared for data analysis. It is even more expensive if data is not correct, or it is applied in the wrong place as elaborated earlier. Therefore, organizations should keep focus on data quality in origin business systems and seek tools to ensure that. (Treder, 2020; Mäenpää, M & Vihervaara T 2021).

Next common technologies and methods in data management are introduced through both traditional centralized and decentralized approaches to data management practices.

3.2.1 Data architecture

Data Architecture provides a framework for organizing and managing data to support an organization's business objectives and decision-making processes. It establishes the blueprint for how data

is stored, processed, accessed, and securely governed securely. Data architecture needs to specify how data is stored, and the data structures used to organize the data. This means tools like data catalog and data modelling. Data access needs a mechanism from data architecture to provide access to data through user interfaces and well-defined APIs to enable data queries and analysis. A mechanism for access control, encryption, and data redaction is needed to enable data security and privacy. Data architecture provides tooling for managing data quality, lineage tracking, and retention policies. (IBM 2024 2.; Serra 2024 Chapter 2).

3.2.2 Centralized vs. decentralized data architecture

Data Centralization is a traditional method in data management where monolithic infrastructure is used to collect, store, clean, transform, and consume data. While it minimizes data silos, fosters collaboration, and provides a single source of truth, it has its challenges. The processes to access and transform data can be slow. It has been observed that when data management and data team is centralized in an organization, requests take longer to get done. In today's real-time business world, modern data-driven companies cannot afford to wait for data. (Serra 2024 Chapter 2).

Data decentralization has emerged as a solution to the challenges of centralized data management. The real-time demands of data analytics have led organizations to re-think new strategies for data management and data governance. Data decentralization aims to give responsibility and ownership back to the business domains, enabling them to impact their operations with the data they own. (Serra 2024 Chapter 2).

Data decentralization refers to an approach where the collection, storage, cleaning, and transformation of data are distributed, eliminating the need for a central repository. Business domains take responsibility for their data, offering it as data products for other business domains to use. The advantages of decentralization include reduced complexity and improved data quality, as the data is managed by the owner and the business team that produces it. (Serra 2024 Chapter 2).

Next, tools such as relational databases, relational data warehouses, Data Lakes, modern data warehouses, and Data Lake houses are introduced. These tools are utilized to construct either centralized or decentralized data architectures using frameworks like Data Fabric and Data Mesh. Refer to Table 1 for the timeline of when these data architectures were introduced.

Table 1 High-level characteristics of data architectures (Adapted from Serra 2024 Chapter 2.)

	Relational Data Warehouse	Data Lake	Modern data warehouse	Data Fabric	Data Lakehouse	Data Mesh
Year Introduced	1984	2010	2011	2016	2020	2019
Centralized/decentralized	Centralized	Centralized	Centralized	Centralized	Centralized	Decentralized
Storage type	Relational	Object	Relational and Object	Relational and Object	Object	Domain specific
Total cost of solution	High	Low	Medium	Medium to high	Low to high	High

3.2.2.1 Relational database

Relational databases are used for storing structured data in online transaction processing (OLTP) systems. Common operations in these systems include creating, reading, updating, and deleting (CRUD) data in a database. These systems typically require fast response times, so generating reports with queries to an actively used database is not ideal. As the amount of data began to grow exponentially, processing and analyzing data in relational databases became a challenge, leading to the development of relational data warehouses. (Serra 2024 Chapter 2).

3.2.2.2 Relational data warehouse

A relational data warehouse is a centralized storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The data warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs.

The five components of a data warehouse are:

- Production data sources
- Data extraction and conversion

- Data warehouse database management system
- Data warehouse administration
- Business intelligence (BI) tools

A data warehouse contains data arranged into abstracted subject areas with time-variant versions of the same records, with an appropriate level of data grain or detail to make it useful across two or more different types of analyzes most often deployed with tendencies to third normal form. A data mart contains similarly time-variant and subject-oriented data but with relationships implying dimensional use of data wherein facts are distinctly separate from dimension data, thus making them more appropriate for single categories of analysis. The data warehouse is centralized storage serving a single version of truth. (Gartner 2024, Serra 2024 Chapter 2).

3.2.2.3 Data lake

A Data Lake is a concept that involves storing vast amounts of raw data without transforming it into a different format. The data assets in a Data Lake are schema-on-read, which means that a schema must be defined by creating or pulling from a separate file when reading. (Gartner 2024).

Data Lakes are a cost-effective choice for storing logs and unstructured data, with cloud services further reducing costs and providing suitable storage technology. Structured data from databases can be stored in the Data Lake in semi-structured formats like CSV, XML, or JSON files. Data Lake offers performance and easy no maintenance need data storage. (Serra 2024, Chapter 2)

However, the challenge with Data Lakes lies in querying. This process requires advanced skills in a data warehouse system and analytics tool for Data Lakes like Apache Hive. The common issue of all data stored in a Data Lake approach turning into a data swamp. This has led to the development of Delta Lake and Data Lakehouse architectures. (Serra 2024, Chapter 2).

To avoid a data swamp, some organization or layers in the file system for raw, cleansed, transformed and presentation data is needed. Serra (2024) introduces in his book Deciphering Data Architectures a data zone approach as folder structure seen in Figure 13.

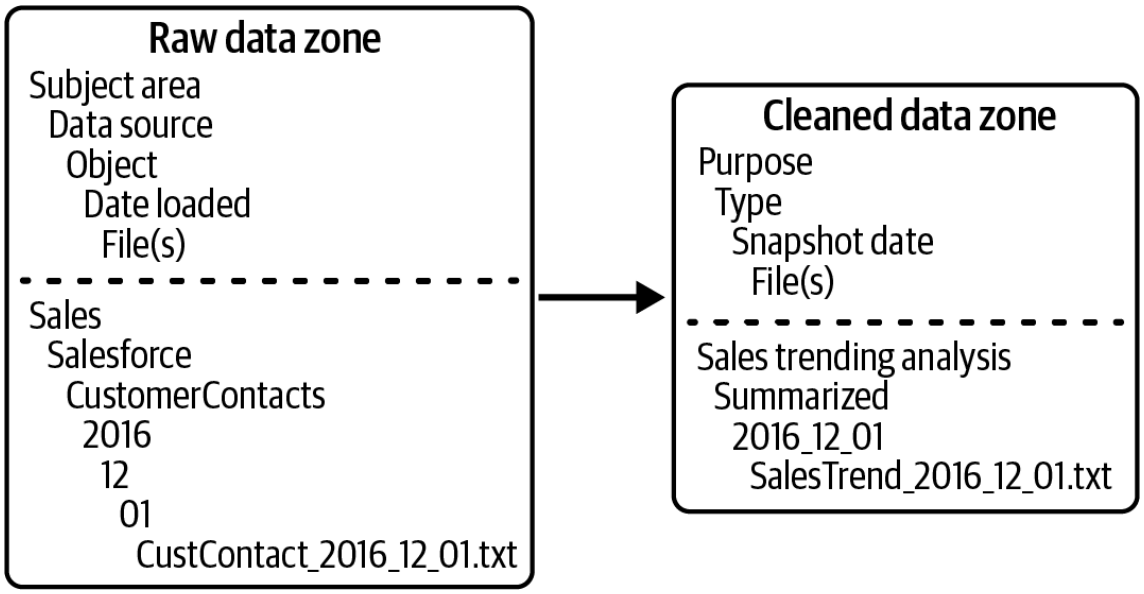

Figure 13 Folder Structure in two Data Lake zones. (Serra 2024 Chapter 5.)

Databases, Data Warehouses, and Data Lakes have evolved into components of modern data architectures. These include the Modern Data Warehouse, the Data Lakehouse, the Data Fabric, and the Data Mesh, which will be introduced in the following chapters.

3.2.2.4 Modern data warehouse

A modern data warehouse incorporates new technologies and strategies for data warehousing. This process involves ingesting data from various sources, storing it in a Data Lake, transforming it, and then storing it in a relational data warehouse to be modelled and made available for querying. Figure 14 illustrates a modern data warehouse's architecture, showing how data in all formats is ingested into a Data Lake, cleaned, transformed, and modelled into a relational data warehouse. The visualized data is then used for data-driven decision-making. This workflow, known as a data pipeline, should always be automated. (Serra 2024 Chapter 10).

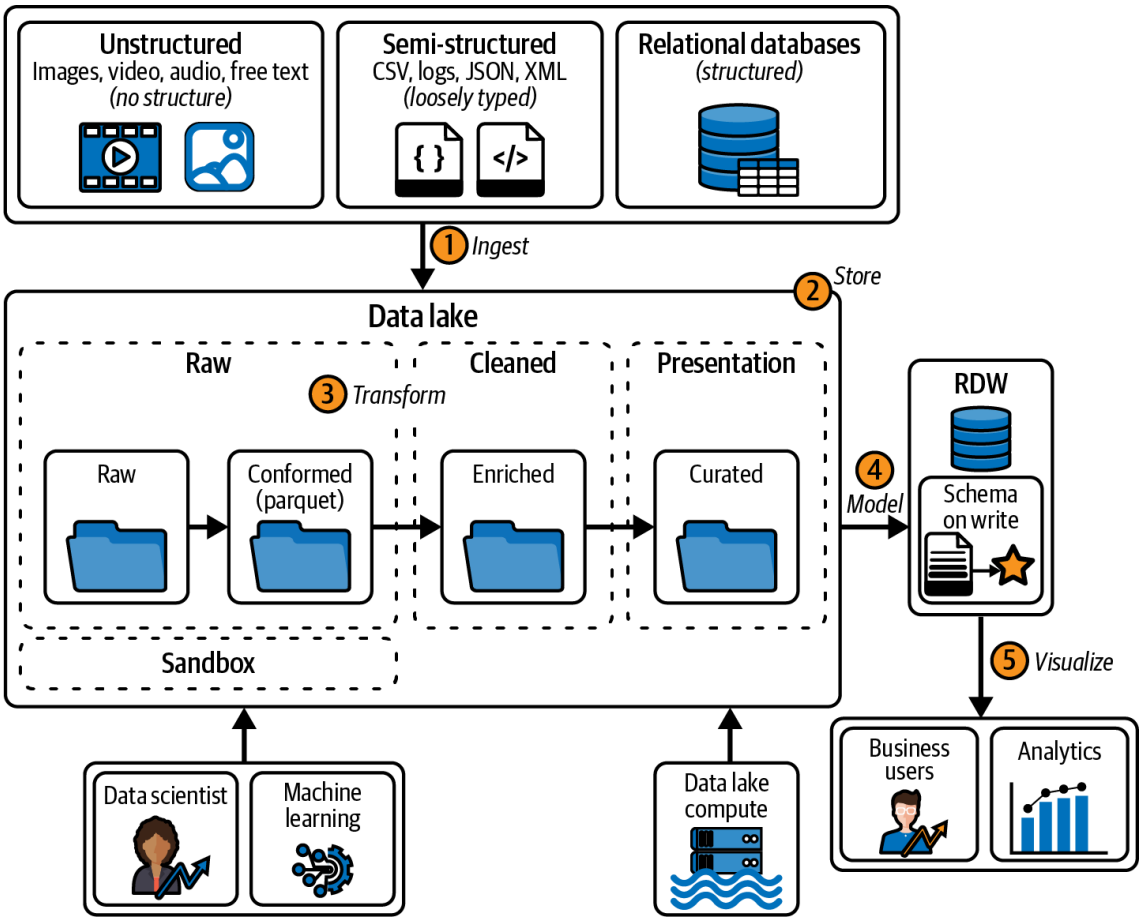

Figure 14 Modern Data Warehouse Architecture (Serra 2024 Chapter 10.)

3.2.2.5 Data lakehouse

Data Lakehouse combines the concepts of Data Lakes and data warehouses. The idea is to simplify architecture by using only a Data Lake to store all data. This is accomplished with Delta Lake, a transactional storage software layer developed and then open-sourced by Databricks. Delta Lake improves the reliability, security, and performance of the underlying Data Lake. It is important to note that Delta Lake is a software layer, not a storage itself. It includes a Delta Lake format where Parquet files are stored in folders and a transaction log that tracks all changes made to the data. Alternatives to Delta Lake include Apache Iceberg and Apache Hudi, which offer very similar features. (Serra 2024 Chapter 12).

The Data Lakehouse architecture can include an optional relational serving layer to help users understand the data using a relational data model. Metadata is included in this layer, but it can cause the relational serving layer to misrepresent the data if the same dataset is used with two different metadata. Many companies create SQL views on top of files in Delta Lake. Then reporting

tool to use those views making it easy for end users to create reports and dashboards. (Serra 2024 Chapter 12).

3.2.3 Data fabric

A Data Fabric is a framework that can be seen as an evolution of modern data warehouse architecture. It incorporates more technology to cover all aspects of data management and governance. By providing a modular platform for all data activities, it supports more use cases and data sources. While a Data Fabric is technology-focused, a data mesh focuses on organizational change. For example, Gartner's definition of a Data Fabric is as follows:

> *"A Data Fabric is an emerging data management design for attaining flexible, reusable, and augmented data integration pipelines, services, and semantics. A Data Fabric supports both operational and analytics use cases delivered across multiple deployment and orchestration platforms and processes. Data Fabrics support a combination of different data integration styles and leverage active metadata, knowledge graphs, semantics, and ML to augment data integration design and delivery." (Gartner 2024).*

Data Fabric is not a single product but is often constructed using various technologies from several vendors. This presents a challenge as it requires acquiring multiple licenses and running products separately. Recently, there have been initiatives to offer Data Fabric under a single license model. Pioneers in this field include IBM with their IBM Cloud Pak for Data, Tibco Software, Hewlett Packard Enterprise with HPE Ezmeral Data Fabric and more recently, Microsoft, which announced its Microsoft Data Fabric and generated significant hype in 2024. (Serra, 2024, Chapter 11; Gartner 2024).

Figure 15 provides an overview of the Data Fabric architecture. Data Fabric supports ingestion from various data sources, storing data in a Data Lake, transforming raw data into a standardized parquet format, and modelling transformed data before presenting it. Data Fabric has support for a broad range of use cases, from analytics to Machine Learning and AI solutions. The next chapter will introduce the eight key components of Data Fabric beyond modern data warehouse to govern and manage data. (Serra, 2024, Chapter 11).

.

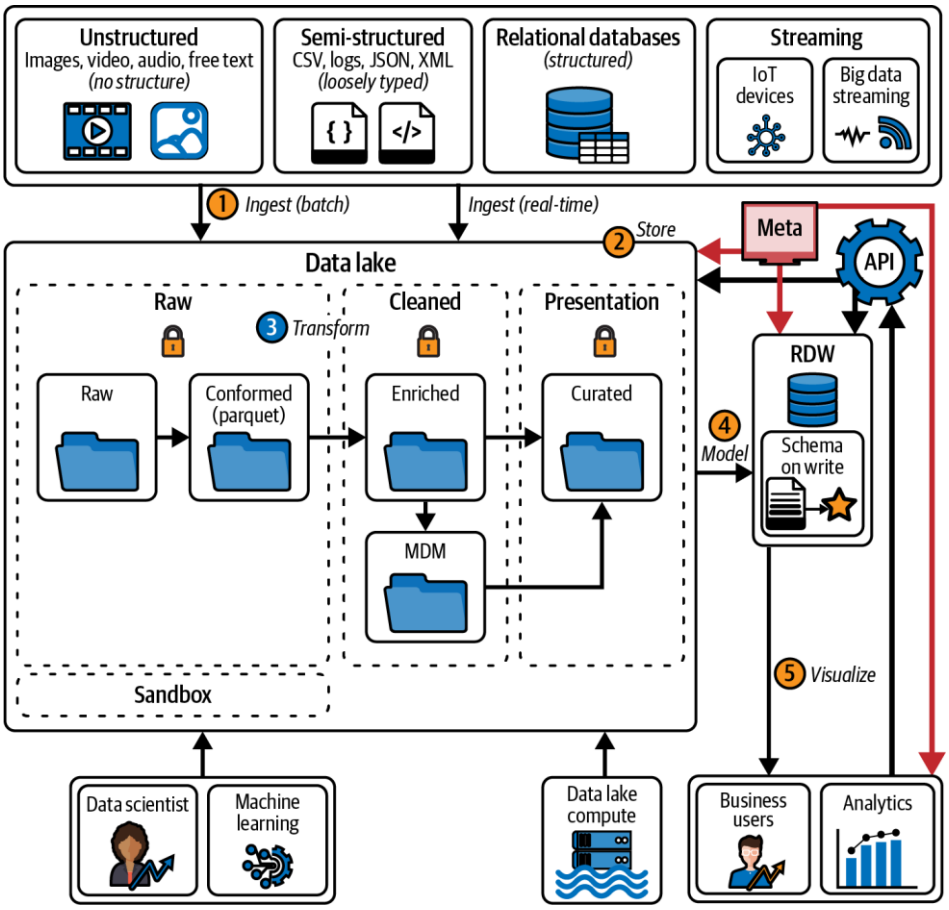

Figure 15 An overview of the journey data takes through a Data Fabric architecture (Serra 2024 Chapter 11.)

3.2.3.1 Data fabric eight key components

A Data Fabric has eight key components: data access policies, a metadata catalog, master data management, data virtualization, real-time processing, APIs, services, and products.

1. *Data access policies* ensure the security, privacy, and integrity of sensitive data. They are key to data governance with a set of guidelines, rules, and control procedures. Data access policies help organizations comply with regulations and laws to cover data classification, access, encryption, retention, backup and recovery, and data disposal. (Serra, 2024, Chapter 11)

2. A *metadata catalog* is a repository for information about data, including structure, relationships, and characteristics. It helps organizations manage and discover data. It also offers

a data lineage tool to record the data journey, its transformations and where it is stored. (Serra, 2024, Chapter 11)

3. Data Fabric has the capability for *master data management*. It is a technology-enabled process to ensure the uniformity, accuracy, consistency, and accountability of company-shared master data sets. (Gartner, 2024)

4. One of the key components in Data Fabric is *data virtualization* which allows access to data from multiple locations without the need to store or move it in a single location. This virtual layer acts as a single point of access that simplifies the complexity of data sources and allows users to access and combine data for multiple data sources in real-time. (Serra, 2024, Chapter 11).

5. Data Fabric enables *real-time processing* of data and producing immediate results as soon as data becomes available (Serra, 2024, Chapter 11).

6. *APIs* provide flexibility to share data securely without exposing underlying solutions or information where data is stored (Serra, 2024, Chapter 11).

7. *Services* are building blocks to share and reuse code for data collecting and cleansing (Serra, 2024, Chapter 11).

8. An entire Data Fabric can be bundled as a *product* and sold (Serra, 2024, Chapter 11).

3.2.3.2 Mitigating challenges in data fabric

Transitioning from a modern data warehouse to a Data Fabric can be a complex and overwhelming experience due to integration, training, and cost factors. Data Fabric is still an emerging framework and a new concept for many. Despite being introduced in 2016, Google trends show that interest in Data Fabric has been growing (Figure 16).

Data Fabric requires expertise in many areas. For example, IBM Cloud Pak For Data may require various skills to install, run, and maintain on a container platform. On the other hand, Microsoft's new Data Fabric appears promising and allows to start small and scale later using its SaaS model.

Small businesses may not need all the capabilities bundled in Data Fabric offerings from vendors. Licensed Data Fabric can be costly, but it has its advantages, and vendors provide training and industry-proven components to start with. Each business must evaluate its needs and the potential return on investment against its long-term strategic goals.

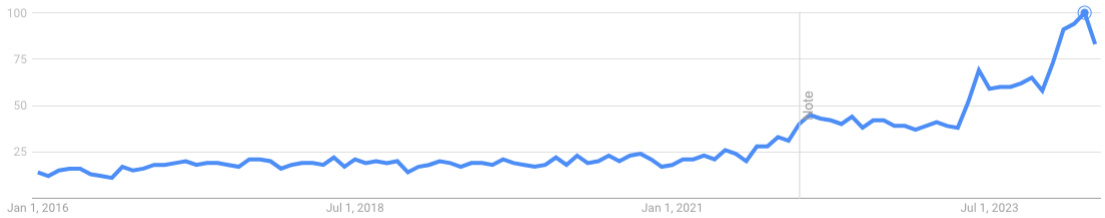

Figure 16 Google trends for the word "Data Fabric"

3.2.4 Data mesh

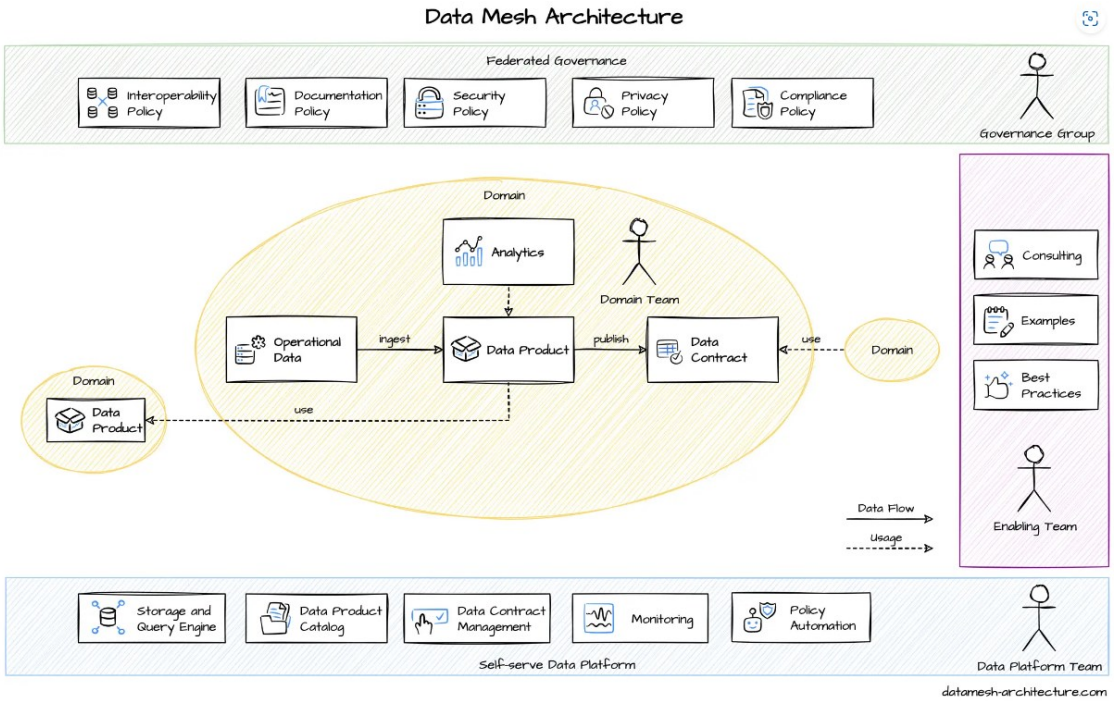

Figure 17 Data Mesh Architecture (Christ, Visengeriyeva, Harrer 2023)

Data Mesh is an architectural approach that decentralizes data management see Figure 17. This means distributing data ownership and responsibility to those closest to the data within an organization, typically to individual teams or domains grouped by business domain, as opposed to a central authority or a single IT team. Data Mesh enables domain teams to perform cross-domain data analysis on their own. It is important to understand that Data Mesh is a concept, not a technology. The technology architecture used to build a Data Mesh can vary, from the modern data warehouse, Data Fabric, and lakehouse, to data stream platforms like Confluent Kafka. (Gartner 2024; Serra 2024 Chapter 13).

Data Mesh is to data analytics what microservice is to software architecture. In a microservice architecture, applications are interconnected using well-defined APIs, sharing, and managing data within their respective domains. This principle of shared responsibility is applied to data analytics through Data Mesh, where data is offered by data products. (Dehghani 2022).

Key principles of Data Mesh include domain ownership, data as a product, self-service data platform, and federated governance (Dehghani 2022; Bellmare 2022). In this approach, subject matter experts who best understand the data in their domains are responsible for managing the data architecture, aligning with Data Mesh principles.

Data Mesh fundamentally changes how data is created, stored, and shared within an organization, representing a cultural and organizational shift. The next chapters introduce Data Mesh's four principles.

3.2.4.1 Principle 1: Domain ownership

Domain Ownership idea has similarities with Domain-driven Development. The business domain has control and responsibility for its data without the need to seek approval to change it from others. But at the same time domain data must be shareable across the organization. Through data contracts, domain data owners become aware of the data users, enabling federated data governance. (Bellamore 2023 Chapter 2; Dehghani 2022 Chapter 2).

3.2.4.2 Principle 2: Data as a product

Data as a product applies product thinking to domain-oriented data. Data provided by the domains is treated as a product, and the consumers of data are acting as customers. Data products are built collaboratively with other domains using well-defined data interfaces (APIs). This concept introduces new roles within domains, such as the domain data product owner and data product developer. These individuals are responsible for creating, serving, and maintaining data products. (Dehghani 2022 Chapter 3).

An example of a data product could be organizational competencies. Human Resources and competence development leads maintain data in the competence management system. They ensure that the data is trustworthy and up to date, encouraging employees to skill and experience growth. This data is offered as a product to the recruitment unit, which assists salespeople in talent matching for new sales cases. An analytic approach examines the characteristics of successful and unsuccessful cases and their relation to the competencies that the company owns or has in the applicant pipeline.

3.2.4.3 Principle 3: Self-service platform

The third principle of Data Mesh includes a self-service platform. This platform provides self-service to the organization through a self-serve data infrastructure. In the Data Mesh model, this platform is offered as domain-agnostic by the Data Platform team. This does not mean a single solution from a single provider, but rather a set of independent technologies that work well together. The platform should offer capabilities such as analytical data storage in the form of a Data Lake, Data Warehouse or Data Lakehouse. Other capabilities are a data processing framework, data querying, a data catalog, and data pipeline workflow management. The main user roles of the data platform are data consumers, data creators, and data product owners. (Dehghani 2022 Chapter 3; Bellamore Chapter 4).

3.2.4.4 Principle 4: Federated governance

The goal of federated governance is to create a data ecosystem where domains comply with universal policies. This ensures data interoperability, security, privacy, and adherence to guiding principles and values. The data mesh concept introduces a governance group consisting of representatives from all domains, including subject matter experts and data platform specialists. Federated Governance strikes a balance between the needs of data consumers, the autonomy of data product owners, and compliance and security requirements. (Dehghani 2022 Chapter 5; Bellamore 2023 Chapter 3).

3.2.4.5 Data stream platform enabling Data mesh architecture

A data stream platform such as Apache Kafka can address common challenges in business applications. Some of these applications may lack APIs, or developing APIs to legacy applications might be too costly. Often, business domain experts rely on multiple domain-specific applications to manage their operations. A company might use dozens of SaaS applications to run its business. Enterprises may also have hundreds of older, monolithic platforms without modern APIs. These legacy systems can lack timestamped data, which is crucial for time-based data analysis. These challenges can be efficiently handled through an event-driven streaming platform. (Confluent 2024).

Figure 18 presents an architecture to implement Data Mesh in an event streaming platform. Data is continuously ingested from various sources to streams, with each change timestamped. An event starting a stream could be a change in a legacy application table in the database or a modification in file storage. The data is then streamed to the event streaming platform where it is cleaned, and sensitive information is reduced. Data from several sources is joined and introduced as a data product. These data products are offered to consumers who register themselves to the streams they are interested in. Data Lineage is built in since producers and consumers of data are known. (Confluent 2024).

Apache Kafka includes pre-built connectors for common business platforms, aiding in data collection and analysis. The Confluent Kafka platform enhances Apache Kafka with data governance tools to ensure data quality and to understand data relationships. For instance, using a stream catalog can boost collaboration and productivity through data discovery, while a shared schema

registry ensures quality and interoperability. Commercial platforms built on Apache Kafka offer data governance tools like catalog, schema registry, data protection, audit logs and more. (Confluent 2024).

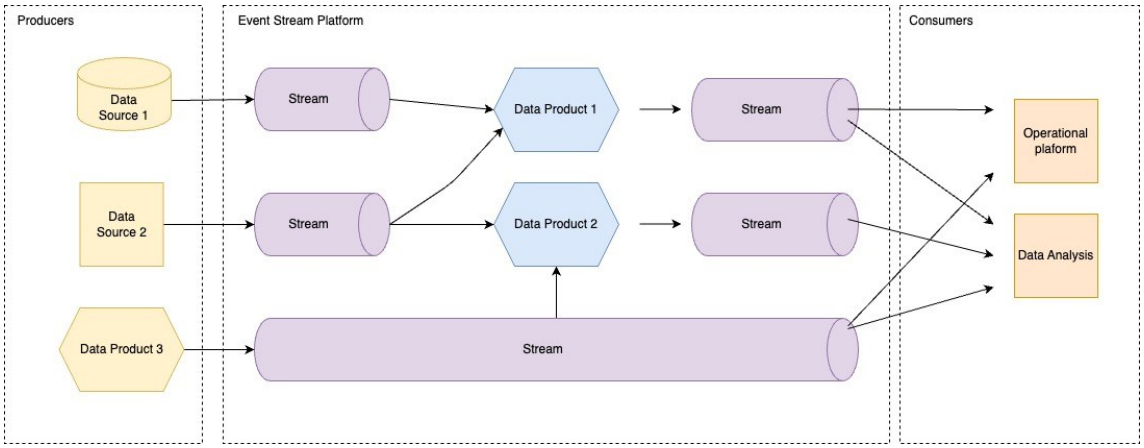

Figure 18 Data stream platform sourcing data sources, joining, aggerating data to new data products

3.2.4.6 Advantages and disadvantages of data mesh

The advantage of the Data Mesh approach is that it can be applied to operational data exchange as well as to analytical data. Data is managed by people who know the regulations and sensitivity of their data. Data quality is the responsibility of the origin of the data in its source. By applying Data Mesh organizations can start building data products from day one by documenting and describing data and developing APIs and scale as needed. Data Mesh allows scalability, agility, and flexibility in managing data. With federated data governance with subject matter experts from domains, organizations can make informed decisions about data usage itself. This improves quality, privacy as well and ethics of data usage. Data Mesh helps discover trustworthy and reliable data and eliminates data inconsistencies between analytics and operational data by using data streams as the single source of truth. (Bellamore 2023 Chapter 2).

There are also some disadvantages of a Data Mesh. The decentralized model of a Data Mesh can lead to coordination challenges. It relies on domain expertise within teams, which, depending on the size of the organization, may result in a lack of skilled personnel. In many cases, data engineers are centralized while data analysts are spread across various business domains. A key challenge

is that people in a business unit tend to focus on their responsibilities. The complexity of the business structure, domains with their own goals, markets, and varying regulations under which the businesses operate can also present challenges. (Dehghani 2022, chapter 7).

3.2.4.7 Mitigating challenges in data mesh

Data Mesh is a relatively new approach to enterprise data architecture as Google Trend shows (Figure 19). It signifies a shift away from a monolithic, centralized method of governing and storing data. The term "Data Mesh" was first introduced in 2019 by Zhamak Dehghani, a principal technology consultant at Thoughtworks at the time, in a blog post on martinflower.com. (Dehghani 2019; Wikipedia, Data Mesh). I discovered it in 2022 during a ThoughtWorks webinar while researching for my thesis. The Data Mesh architecture and methodology were chosen as the foundation for implementing the data strategy and governance at a case study company in this thesis.

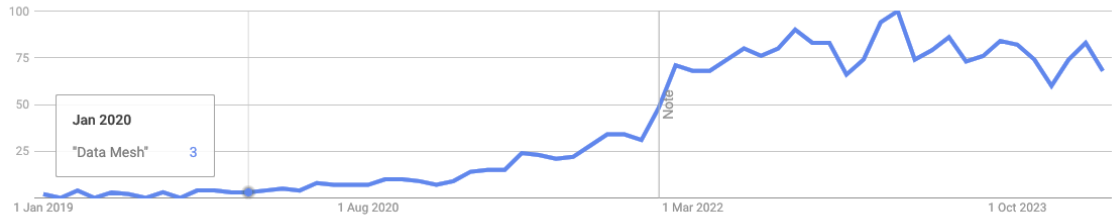

Figure 19 Google trends for the word "Data Mesh"

Despite its recent introduction, the Data Mesh has been adopted by corporations like Zalando, Netflix, Vistaprint, and PayPal (Wikipedia 2024). The challenge remains in explaining this new concept to people who may not be tech-savvy.

Data Mesh has the advantage of starting small and scaling later, making it a good approach for start-ups and agile companies. Challenges can be mitigated by promoting data literacy. Organizations must educate their workforce about data. Likewise, a strategic approach to data usage is necessary. Proper data governance practices ensure a common understanding of data, enhanced data quality, and compliance with regulatory requirements. Implementing principles of domain ownership and federated data governance from the Data Mesh model is part of the solution. Data management tools such as domain-specific data catalogs and data standards help in documenting and facilitating data discovery within the organization. Instead of relying on point-to-point integrations

between systems, the exchange of data can be managed through well-defined APIs via a streaming platform.

Many companies are interested in solving problems such as lack of data ownership, quality, and scalability with centralized teams and technologies through the Data Mesh approach. Often, they start by implementing domain data ownership and adopting a decentralized function mindset while understanding data as a product. They maintain a centralized self-serve platform and governance. This approach addresses challenges related to skill shortage and standardizes the technologies used at the organizational level. (Serra, 2024, Chapter 13).

3.3 Data analytics strategy

All companies collect data in some form. At a minimum, financial reports are gathered to comply with annual reporting regulations. To gain a business advantage, organizations need a strategic approach to data analytics. Data analytics is a key element in an offensive data strategy, where data is collected to analyze customer behavior, the market situation, competitor information, and conduct A/B testing. This helps determine which option performs better in the market and allows for data-driven decision-making. (DalleMule, Davenport 2017).

To be truly data-driven organization needs forward-looking analysis. Reports organize data into informational summaries to monitor how business is performing while Analysis is transforming data assets into business insights to fuel data-driven decision-making. Reporting says what happened. Analysis why it happened. Figure 20 summarizes the difference between the two. (Anderson 2015, chapter 1).

	Past	Present	Future
Information	What happened? (Reporting)	What is happening now? (Alerts)	What will happen? (Extrapolation)
Insight	How and why did it happen? (Modeling, experimental design)	What's the next best action? (Recommendation)	What's the best/worst that can happen? (Prediction, optimization, simulation)

Figure 20 Key questions addressed by analytics (Davenport, Harris, Morison 2010.)

In his book Carl Anderson (2015), while elaborating on characteristics of data-driven organization, suggests shifting the use of data from mere reporting to comprehensive analysis. He outlines two prerequisites:

#1: The organization must be collecting data.
#2: Data must be accessible and queryable.

But it is not just any data, Anderson emphasizes. The data set must be relevant to the question at hand. It should be timely, accurate, clean, unbiased, and trustworthy. For data to be accessible and queryable, it must also be joinable with other organizational data and shareable within a data-sharing culture. Moreover, there should be tools to query, slice, and dice the data.

For effective reporting and analysis, the raw data must be filtered, grouped, and aggregated into meaningful information. This needs a lot of human work. It said that data scientists spend 80% of their time collecting, cleaning, and preparing data and only 20% of their time is spent building models, analyzing, visualizing, and drawing conclusions from that data (Forbes 2016). See Figure 21. To reverse this 80-20 rule and challenge a strategic approach to data and data analytics is needed.

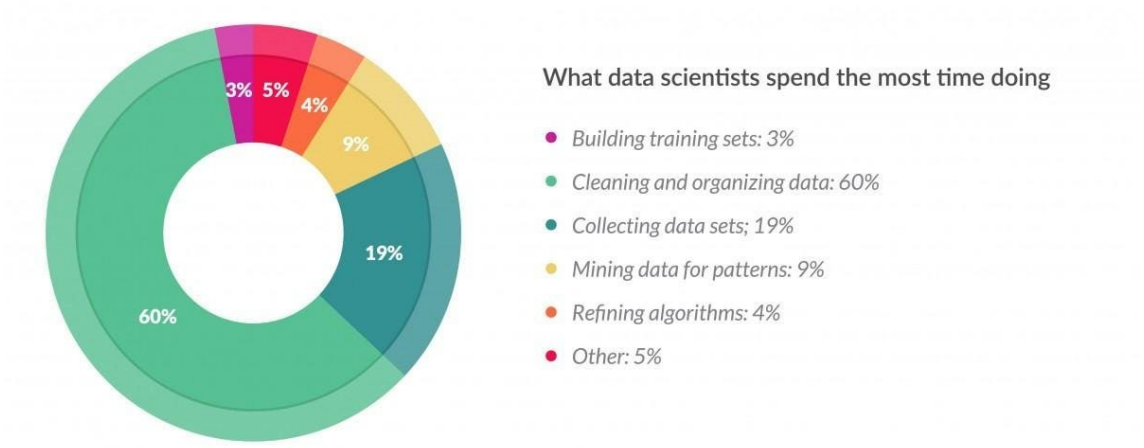

Figure 21 Data preparation accounts for about 80% of the work of data scientists (Forbes 2016.)

Data analytics strategy makes no difference in other strategies. It is a well-defined plan with goals and objectives to leverage data effectively and gain valuable insights. It needs to be aligned with business goals and objectives. It needs to be aligned with Data governance by following policies and rules set about data usage. It will use tools and platforms defined in Data management strategy. (Data Centric Inc 2021, Tableau 2024, Anderson 2015).

Organizations can strategically decentralize, centralize, or use a hybrid model as an operating model for people in data analytics. A decentralized model enables quick and tailored decision-making by divided authority and operational responsibilities across business domains. In a centralized model, authority is centralized and needs a high level of standardization and coordination within the organization. In the Hybrid model, a suitable approach is chosen for specific situations balancing flexibility and coordination. (Serra, 2024, Data Centric Inc 2021, Tableau 2024, Anderson 2015).

The outcome of Data analytics is well-explained data products like Data visualization dashboards and metrics align with organization goals, key performance indicators, values, and objectives. These data products are available for everyone who needs them for business insights and data-driven decision-making. (Data Centric Inc 2021, Tableau 2024, Anderson 2015).

3.4 Data & AI strategy

As said data is often referred to as the new oil. It serves as fuel for data-driven decision-making and, with the current focus on Generative AI, for Artificial Intelligence applications. Organizations need an AI strategy aligned with business and data strategy. Despite the risk issues like data privacy, security, bias risk, job displacement, and intellectual property protection just 21 percent of companies reporting AI usage say that they have policies governing people´s use of AI technologies (McKinsey 2024).

AI is not just a switch-off tool. It is already incorporated into many tools in the form of Machine Learning (ML) algorithms support, and AI assistants are becoming ubiquitous. The latest advancements in Natural Language Processing and cloud services have made Generative AI accessible to all. It is challenging to prevent its usage in an organization, making strategy and governance crucial for managing these tools. (McKinsey 2024).

The earlier described data strategy covers the entire value and supply chain of data, from collecting and managing datasets to generating business insights through analytics and business intelligence, ultimately delivering value through business use cases and data-driven decision-making. AI assists in automation and even in creating forecasts based on historical data using Machine Learning (ML) algorithms. As with a data strategy, there is need to balance between defensive and offensive approaches to AI. But first, let's establish a common understanding of what AI is.

3.4.1 What is artificial intelligence?

The term AI has been in use since the 1950s. In movies, it is often portrayed as a dominant power threatening human life. But the definition of AI has evolved, transitioning from early rule-based computer systems to modern AI algorithms that mimic human behavior. However, the current stage is still referred to as narrow AI, with the goal being human-level AI, also known as Artificial General Intelligence (AGI) (Petersson, 2023).

Types of AI are seen in Figure 22.

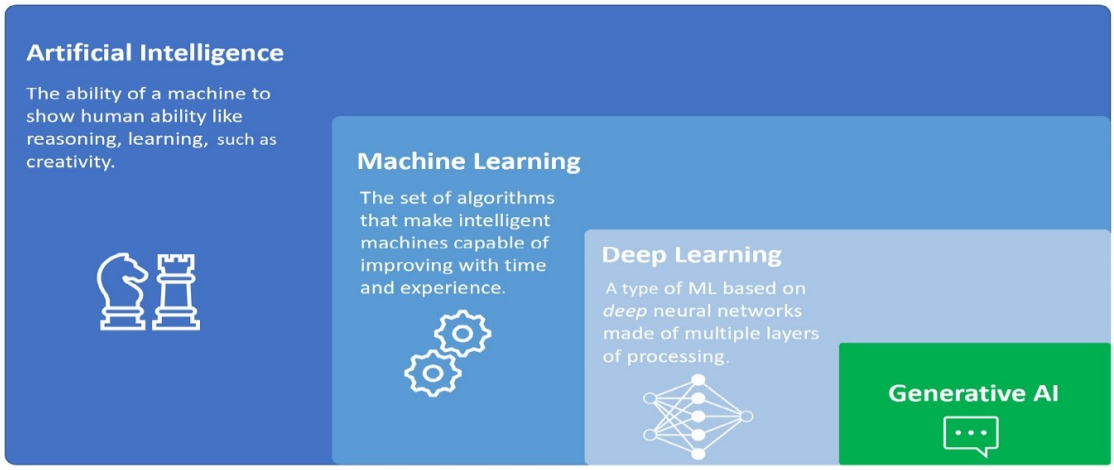

Figure 22 Relationship between AI, ML, DL, and Generative AI (Alto 2023.)

AI encompasses a broad field of creating systems that are capable perform tasks that require human abilities such as reasoning, learning, creativity, and ecosystem interaction. Machine Learning (ML), a branch of AI, focuses on creating algorithms and models that enable systems to learn and improve over time with training. These models are trained with existing data and update their parameters automatically as they evolve. (Alto, 2023; Plain Concepts, 2024).

Deep Learning is a subset of Machine Learning that includes in-depth models known as neural networks. These are used in areas like computer vision and Natural Language Processing (NLP). Typically, Machine Learning and Deep Learning models are discriminative models; their goal is to make predictions or infer patterns from the data. These models are built on three pillars: supervised, unsupervised, and reinforcement learning, with information provided based on clustering outcomes. (Alto, 2023; Plain Concepts, 2024).

Generative AI, a segment of Deep Learning, refers to systems that can generate new content such as text, images, music, and videos using neural networks. Generative AI creates new examples from scratch based on patterns in the training data. This is why the results can sometimes be misleading, as Generative AI is predicting the next best outcome. (Alto, 2023; Plain Concepts, 2024).

3.4.2 Common generative AI use cases

Companies seeking a competitive edge are turning to Generative AI due to its relative ease of use. Most initiatives focus on Copilot-style exercises where AI assists as a companion. Use cases span from software development, document reviews, and to the generation of text, code, images, and videos. Generative AI tools hold the advantage of working well with unstructured data. Additionally, customer services are increasingly utilizing bots with AI capability for online interactions. Although a leap in productivity is observed in these cases, it does not always translate into a competitive advantage. The challenge lies in generating revenue from increased productivity. Nevertheless, AI's near-term value is clearly in its ability to enhance work efficiency. (McKinsey 2024).

3.4.3 Monitor AI opportunities in data management

AI assistants, such as those available in data analytics tools like Tableau, can aid in storytelling to explain data within analytics dashboards. Machine Learning algorithms are commonly used in predictive data analysis. They help companies anticipate factory machine maintenance needs, monitor systems, and identify deficiencies. They also assist in sales and financial forecasting. (Tableau 2024).

In the next 2-5 years, we can expect Machine Learning and Generative AI to enhance data management with features such as augmented data quality, and augmented data catalog and metadata management. Over the next 5-10 years, active metadata management and Generative AI for data management are expected to become mainstream. Generative AI, with capable of learning from large source content datasets, will facilitate content discovery, ensure authenticity, and maintain regulatory compliance. It will also introduce natural language interfaces to data management, making managing activities more accessible. (Gartner 2023, Hype Cycle for Data Management).

Adoption of AI tools in data management and how vendors are developing offerings in these should be monitored. Even as these tools evolve, there are always opportunities to enhance traditional data management and governance practices. This will ensure readiness for the next phases of Generative AI.

3.5 Design your data strategy

To design a data strategy, you first need to understand your business objectives. This understanding will allow you to align your data strategy with other strategies like IT and Marketing. This is the first key step. Engage with C-suite and business stakeholders to discuss business objectives and treat data as an asset. Ultimately, these leaders should set an example to enable the organization to become truly data-driven. The executive support you need is established in this phase. (IBM 2024 3.)

The information-gathering phase should be collaborative and involve asking the right questions. IBM provides a template for what to ask stakeholders.

1. *What are your top business goals and indicatives that require data and AI use?*
2. *What are the biggest challenges preventing you from achieving those priorities?*
3. *What are data privacy and security challenges do you have related to self-service data access?*
4. *How much time do you spend integrating tools to build solutions?*
5. *What do you wish you could use data for that you cannot quite hack right now?*
6. *How do you measure success yourself and your teams?*

Understanding the organization's position in its data maturity journey and what data it has is another key factor. When figuring out the first data initiatives with stakeholders, start by assessing the applications, business systems, and SaaS services used in business domains. This involves building a data catalog and inventory. List existing infrastructure and technologies, as well as current strategies, together with the IT team. This information is crucial to understand what can be done with existing technologies and how to strategize for adopting new ones. (Wallis, 2021).

To assess the state of data maturity, consider using the questions introduced in Chapter 2.7.2.

- *How do you measure the success of digital projects?*
- *Are data and analytics easily accessible to the team? How quickly can they find and analyze data to answer their questions about product performance?*
- *How does your organization connect product changes to business performance?*
- *How does your organization experiment with new ideas?*
- *Does your team effectively use both qualitative and quantitative data for analysis? (Obstler 2023)*

Developing a data strategy is a journey that begins with small projects with defined business goals. Attempting to investigate all infrastructure and data sources at once can be time-consuming and often results in a loss of interest due to distant goals and unseen results. Therefore, accumulating small wins with compact projects targeted at the near future has proven to be a cost-effective way to test hypotheses and justify continued investments. The key tools are defining a Minimum Viable Product (MVP) and initiating proof of concept work. Data literacy and practices are built step by step in three dimensions introduced earlier: data governance, data management, and data analytics. This approach continues as projects scale to other business domains. (Wallis, 2021 Chapter 6).

Consider your data strategy as a living document where overall goals and objectives align with the organization's business objectives and values. Developing a data strategy is a team effort. With a data strategy, create a roadmap to continuously improve data governance policies, data asset inventory, data literacy, and lessons learned. The data strategy document should be kept between 12-20 pages, considering the audience, the necessary level of detail to ensure everyone reads it, and the importance of keeping ambitions realistic. (Wallis, 2021 Chapter 6).

Wallis (2021 Chapter 6.2) in his book Data Strategy, lists the following key elements expect to find in a data strategy:

- *a strong data management vision*
- *a coherent business case to support investment (even if that is just resources within the organization, these still come at a cost)*
- *some guiding principles, values, and alignment to the corporate vision*
- *clearly articulated goals related to data*
- *evaluation criteria and metrics to track success*
- *clarity on the data strategy program vision to be delivered*
- *clarity of roles and responsibilities.*

The Data Strategy should include an executive summary that outlines the purpose, context, and scope of the strategy, along with the necessary resources for delivery, timelines, and methodologies. It is crucial to highlight the expected value as well. In the data strategy document, you should clarify the 'why', 'what', 'when', and 'how' of three dimensions of data strategy: Data Governance, Data Management, and Data Analytics. Furthermore, you can discuss three aspects of becoming

a data-driven organization: Data, People & Culture & Processes, and Technology. The Venn Diagram below (Figure 23) summarizes these dimensions and provides a visual guide to maintaining balance among them. (Wallis, 2021 Chapter 6.2).

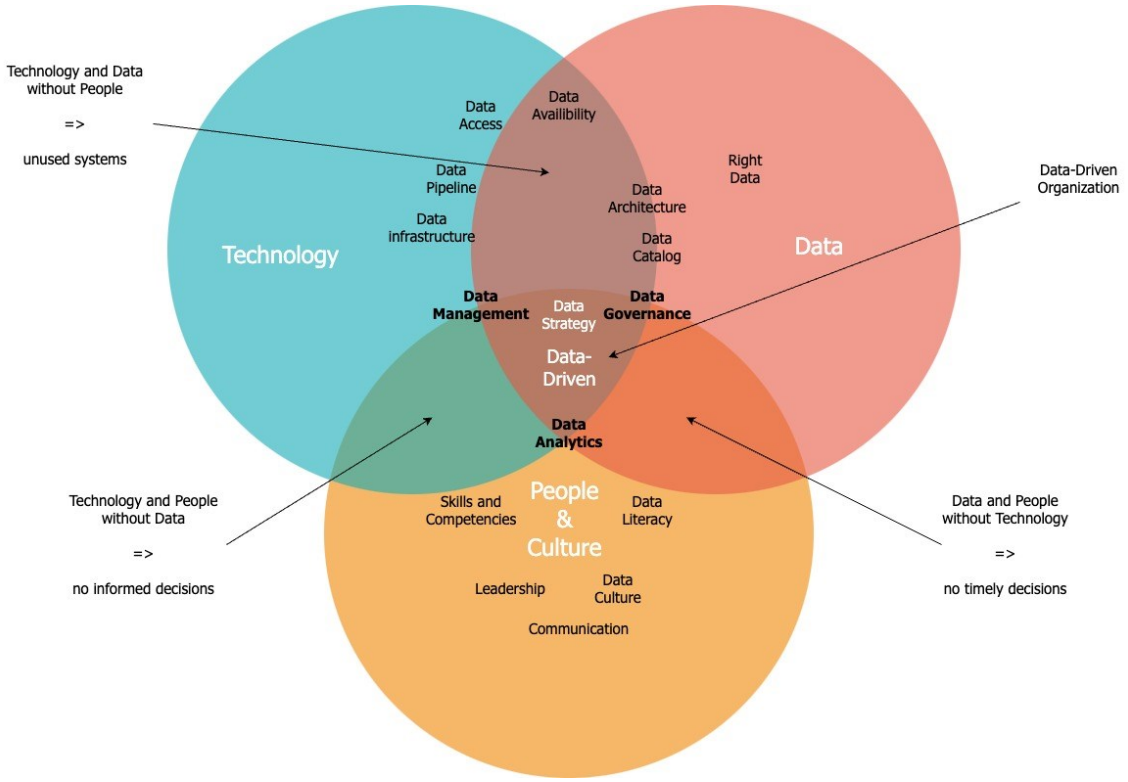

Figure 23 A Data-Driven Organization with Data Strategy (Adapted from Allouin 2022.)

The length of a data strategy document requires a balance in the details included. It should not be too high-level or overly detailed. Include details in the plan used for the implementation phase, as well as in documentation related to the data products and platforms in use. Data governance tools, such as data catalogs and data standards, will provide these details. A comprehensive coverage of governance needs should include defining data access policies, data redaction, and data retention. It should also specify who owns the data, and record the data journey (data lineage), including who uses the data, how, and where. (Wallis, 2021 Chapter 6).

3.6 Challenges with data strategy

Starting with a data strategy can be an overwhelming experience due to the multitude of methods, tools, laws, and regulations. It is crucial to align with the business strategy, which often is not communicated clearly within the organization. There may also be a lack of skilled resources. Common

challenges include also managing unstructured data, controlling data access, and the time-consuming task of data cleansing and preparation for analysis. It has been observed that less than one percent of unstructured data is effectively catalogued and utilized in organizations. Data access is often improperly restricted, with over 70% of employees having access to data they should not. Additionally, analysts often spend 80% of their time discovering and cleaning data (DalleMule, Davenport 2017). Along with data governance and management, addressing data lineage challenges is also necessary. This involves understanding, tracking, and visualizing data as it moves from its sources to its consumption points (IBM 2024 3.).

To overcome these challenges, it is important to set a business-related goal, secure a sponsor, and start small, scaling as necessary. Appointing a chief data officer and data evangelist can help improve data literacy within the organization. And if you have laws and regulations to follow you can always start with a defensive approach to data strategy, data governance and data management. (DalleMule, Davenport 2017).

4 CREATING DATA STRATEGY IN A CASE COMPANY

4.1 Data maturity of a case company

When evaluating the case company's current state, maturity in data-related practices and data literacy focus is organization business domains running the company. Company employees working on customer projects are excluded. Work related to establishing the data strategy in a company was done from March 2022 to the end of February 2023.

An assessment of the current data maturity level was conducted in collaboration with business domains and support functions. As the company is relatively small, all stakeholders joined the workshops. These workshops provided information about the company's current state, including data collection, usage, governance, management, and literacy. Findings were saved in the data catalog, and the technologies, tools, and business platforms in use were documented in the handbook. The IT team helped investigate the data within these systems.

A key observation was that the company has followed a typical path, transitioning from Excel sheets to non-standardized reports, and using multiple tools to visualize numbers for decision-making. A centralized data architecture project was ongoing to collect data in one storage to create a standardized single source of truth but failed to lack of support and time from stakeholders. The company became aware that action needs to be taken to reduce costs and gain benefits from the collected data.

In Dell's data maturity model level discussed in chapter 2.7.1, the data maturity state was observed to be between data proficient and early data savvy. Some business domains were in the data-aware stage, while others were successfully using data, aggregating datasets, and even performing predictive forecasting. Unsurprisingly, these were the finance and sales departments. The leadership set an example by presenting data during companywide monthly meetings and using data to justify the financial and sales situations and needs.

4.1.1 First steps in the data maturity journey

The outcome of the assessment with business domains was the need to educate people in data literacy and establish a culture where business domains collaborate more with data. The first milestones were set in Data Glossary, Data Discovery, Data catalog and Data modelling. Also, data quality was promoted to be solved in origin business systems.

In collaboration with business domains established task force team was given access to all data. Data quality issues in recruitment data as well as in company competence management systems were fixed first and development work was done to improve data quality by adding data validation in business systems.

The next step was to take over the data warehouse implemented by an external consultant. In the end findings with that led to freeze data warehouse building since that project was missing leadership support and the results were confusing. The decision was to make a strategic analysis of alternative platforms and ways to implement data collecting, storing, transforming and visualization. A surprising finding was that it is quite cost-effective to collect and store data but tools for visualizing and sharing dashboards were licensed expensively by cloud providers.

The next goal was set to standardize reports in data analytics and business insights tools. Dashboards were built to Tableau Business insights tool connected directly to the company Odoo ERP database. Business domain using another BI tool was encouraged to move to this tool since Tableau was licensed and could share dashboards through a web portal. The company was using heavily Odoo ERP and several modules there. There were several table-based reports available, and investigation justified that an attractive path was to upgrade Odoo ERP to the newer version which is offering lots of improvements in data analysis, forecasting and visualizing data.

Upgrading Odoo ERP was chosen for several reasons. At the same competence management platform met a problem that it was planned to put down by the company providing it. At the same time, the finance department started its project to change the company's HR and finance platforms. These events set the direction of data strategy work and slowed it down. But at the same time offered the possibility to create data governance policies since migration projects needed heavy focus on data: What data where and how to transfer it to new platforms.

In this state, it was easily proven that a company needs a data management process and educate people on data management roles. Handbook first approach was chosen since the company had that model in other areas already. Business domain leads were designated as Data Owners, and a Data Mesh type federated governance was introduced along with the concept of data as a product.

4.2 Data strategy development

4.2.1 Handbook first approach

Often, documentation is treated as an afterthought or something that happens after an event, or after a product or service has been delivered. This traditional approach sees documentation as a secondary process, usually occurring in the aftermath of a significant event or completion of a task. However, the "Handbook first" approach advocates for a shift in this perspective, suggesting that documentation should be the first step in any process. It argues for the importance of creating documentation in a structured and organized manner before any information is disseminated through other communication channels such as email or chat. This approach emphasizes the necessity of having a well-documented and readily accessible reference that can guide actions and decisions, thereby streamlining processes and improving efficiency. (Gitlab, 2022).

In the case company, the handbook servers as the single source of truth in information about company policies and practices. Data strategy will be communicated trough that handbook. The data strategy will give guidelines, policies and tools for data discovery and management and help the organization grow in data literacy.

The next eight steps were taken in case the company data strategy implementation journey and the handbook table of contents were introduced.

4.2.2 Start the data literacy journey!

~~Like~~ a handbook-first approach starting with data literacy development activities from the beginning is a good way to grow an organization's capability to read, write, analyze, reason, and communicate with data. While data literacy is developed it allows the organization and individuals to gain innovation and advantages from data and make better data-driven decisions. Data strategy work plans and documentation were available for everyone. Data Glossary was written first. Status and resources related data strategy development were communicated in companywide monthly meetings.

4.2.3 Set business goals

Business goals were set around data products enabling ongoing data-driven decision-making. Odoo ERP as an operative data platform needed more focus and a development plan. People in business units needed to push to take an active role in governing their data. First steps were taken to make sure data quality in its origin business systems. Data Mesh advised approach to treat data as a product was chosen as the method and data owners were identified in Sales, HR, Finance, Recruitment and in the Project management office. Another goal was set to enable robust data integration between business systems. Apache Kafka was chosen to be tested with the Proof-of-concept method.

4.2.4 Discover data sources

Planning for data products was based on data collected from business systems and business insights platforms. Both the Data Lake and warehouse were excluded due to challenges with data quality and accuracy as the data pipelines were halted. The focus was redirected towards operational data in origin systems such as Odoo ERP and its available tools. Data in business domains was identified at an operational level along with owners and potential consumers. The goal was to ensure data quality and accuracy through data integration between business systems.

4.2.5 Start data governance

A data catalog was compiled, and data owners were identified and documented in a handbook. Data modelling was used to illustrate the company's data and its location. Pertinent laws and regulations were identified, and actions were taken to comply with the ISO 27001 data security standard and ISO 9001:2015 certification, while also ensuring adherence to GDPR. Data access was investigated, with improvements made to ensure people have appropriate access for their roles.

4.2.6 Build a data architecture for data management and infrastructure

An analysis of the current infrastructure and alternatives for the data platform was conducted and reported to managers. As all efforts were directed towards business system upgrade projects, the Data Platform project was put on hold. Meanwhile, an alternative approach using a data mesh architecture with Apache Kafka Data Streaming was explored, and a proof of concept was completed.

4.2.7 Introduce data ownership and data products

Data ownership was established, leading to the identification and innovation of potential data products at both operational and analytical levels. The integration of sales opportunities, competence management data, and recruitment data and processes brought about substantial business benefits. Several data products were subsequently developed around this integration.

4.2.8 Improve data analytics

The company's dashboard and reporting were continuously improved. As awareness of tool availability spread throughout business domains, data literacy and the skill to create analytical analyzes also increased. Analytical data products were innovated to enhance data quality and availability for example by enabling gamification in the competency management system to enhance the readiness of employee CV data. An example of how to use data and data analytics in decision-making was developed as part of this thesis. You will find the example in chapter 4.3.

4.2.9 Discover improvement opportunities

The data platform was reviewed, and the decision on how to proceed was postponed. Alongside the review, alternative paths were analytically investigated using organizational competency management, required platform skills, and costs. The pros and cons were listed and presented to the managers.

Additionally, the wider use of Odoo ERP capabilities was introduced and was well-received within the organization. This will improve data awareness and analytical usage.

The Object Key Results (OKR) method was used to gauge the success of planned, projected, and implemented actions.

4.2.10 Data strategy handbook table of contents

1. Introduction
2. Data Glossary
3. Learning Material
    a. Becoming a Data-driven company
    b. Data strategy and its three dimensions
    c. Data ownership and data products
    d. Data culture
4. Company Data strategy
    a. Purpose and alignment to business strategies and values
    b. Goals, objectives, and metrics
    c. Data Governance Guidelines
    d. Data Management architecture and tools
    e. Data Analytics tools
    f. AI strategy and guidelines
5. Data Catalog

4.3 Example of how data can be used in decision-making

4.3.1 Introduction

This is an example of how data can be used in decision-making. For sensitive and business privacy reasons, the example is limited to publicly available information. The data is from the year 2022 and does not represent the current state of the case company.

The company has a strong presence in embedded technology and has been successful in servicing this area. As the amount and use cases of data expand, local processing has been accompanied by cloud platforms and data is seen further processed and integrated into business systems to accelerate and enhance data-driven decision-making. It is more than just intuition that there is a growing demand for new skills in customer projects and our product development.

4.3.2 Research questions

A survey conducted in March 2020 reported that by 2022, 90% of global enterprises will rely on a hybrid cloud (Deloitte 2020, IDC 2020). The term "hybrid cloud" implies that companies manage workloads in both private and public clouds. For instance, manufacturing companies with factories located overseas often deal with slow internet connections. To address this, they process data on a local private cloud platform and then transfer it to the public cloud when connections improve, usually on a nightly basis. This use of edge locations has been growing as data is collected everywhere through cameras, LIDAR radars, and other IoT sensors. Whether it is in a mill, parking lot, highway, or supermall, data is collected and processed locally to support real-time automated decision-making. It is also sent to the public cloud for storage and further analysis.

The research questions are: How can we enhance our role as a trusted partner within our organization? Do we have the required skills and competencies to broaden the services we provide, or should our focus remain on strengthening our core competencies?

4.3.3 Analysis of competence data

Data is crucial for data-driven decision-making. The first step is to identify the needed skills. Following this, conduct a competency analysis within the company. Consider whether there are external market analysis reports that can aid our decision-making. Also, evaluate whether we have our data to conduct market research.

The data sources for this research are listed in the company's established data catalog. The CRM provides sales data including a list of skills requested in both won and lost sales cases. This information gives us insight into customer demands and the reasons for lost cases. We can then determine whether the losses were due to a shortage of consultants or a deficiency in skills.

The company data catalog shows that there is a competency management system where employees and subcontractors have listed their skills. To gain insights and obtain global and country-specific market reports, we will use external market research data from companies like Statista and Gartner. For the sake of simplicity and privacy, we are only presenting employees' skills and experiences here. This information can also be found on LinkedIn, a fact we discovered at the end of this study.

The data was collected through a Python application, using APIs in the competency management system, and analyzed using the Pandas analysis library in Python and Matplotlib for visualization. Results are shown in Figures 24 and 25.

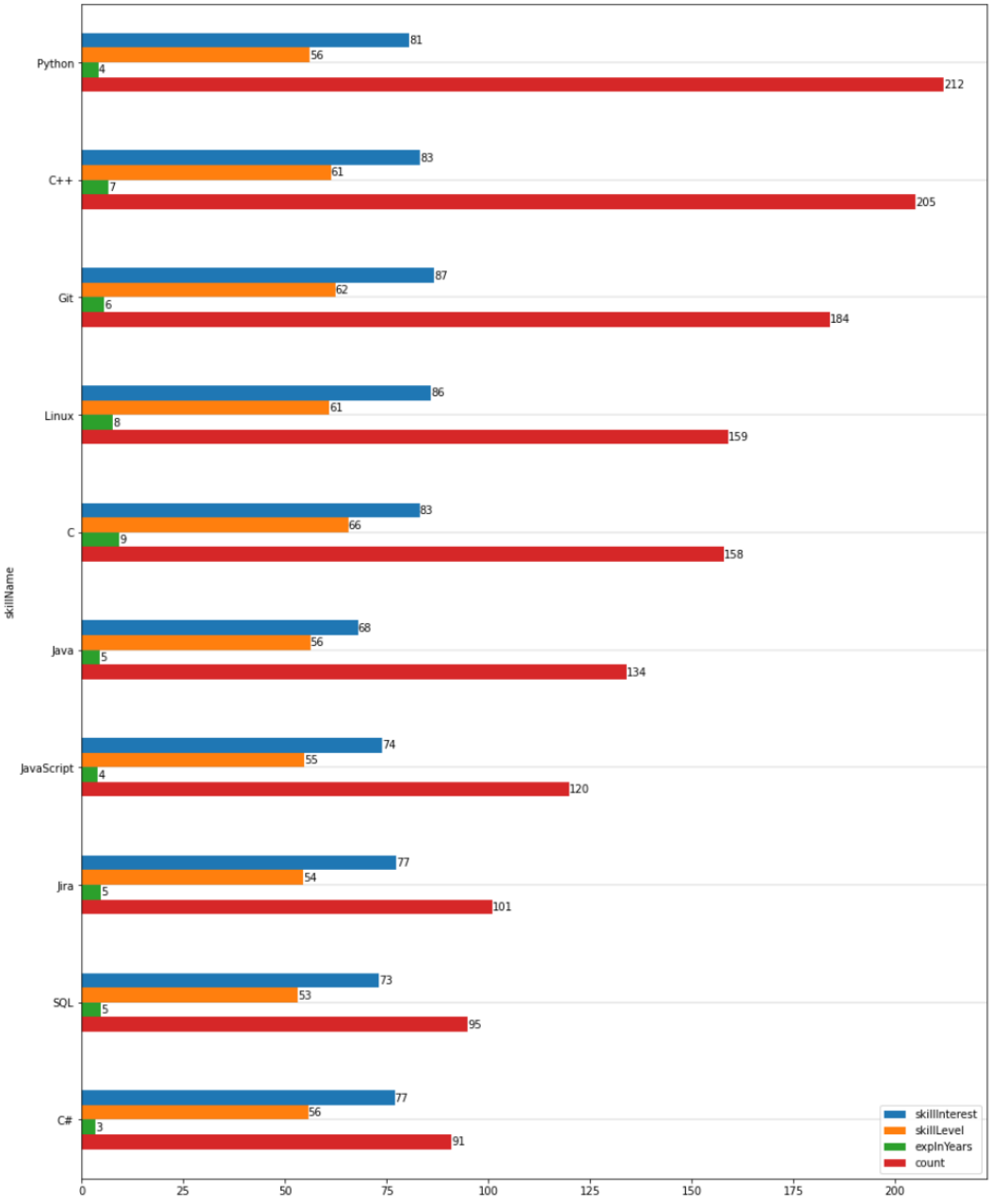

Figure 24 The most mentioned competencies in the company's competence management system that are ranked 1-10.

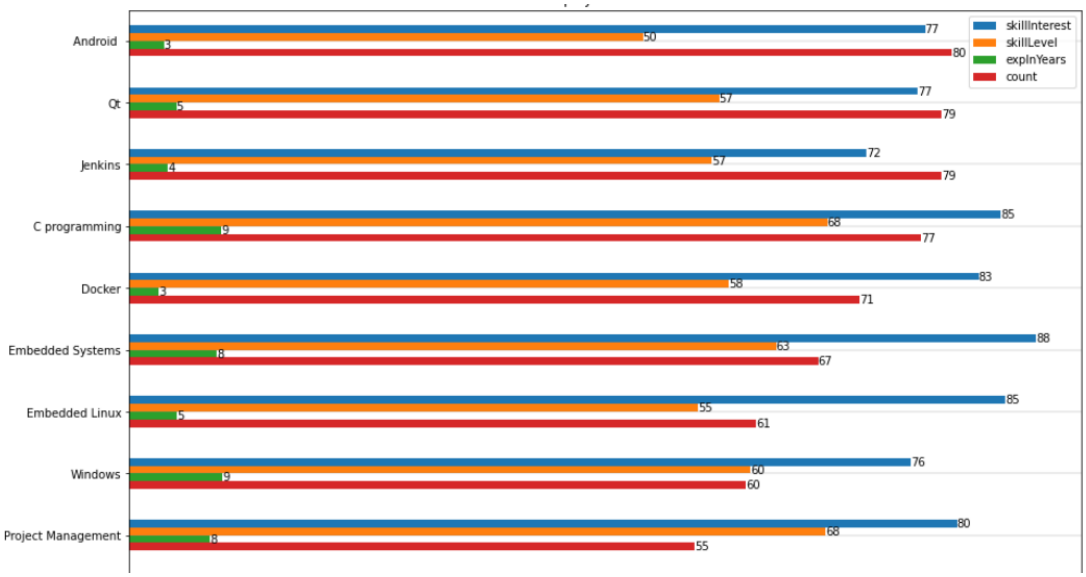

Figure 25 The most mentioned competencies in the company's competence management system are ranked 11-19.

The results reveal significant experience in embedded systems and their related programming languages. Employees who have listed a cloud provider are split into smaller groups between AWS, Azure, and Google, suggesting that cloud services are not commonly mastered.

Why is this the case? Likely, there are two reasons. First, customers typically either maintain an in-house platform engineering team or outsource platform upkeep to an external service provider. Moreover, the services used can change from project to project.

Second, the company mainly hires software developers, many of whom work in the embedded technology sector. However, about 80 individuals are recognized as working with business systems. This group could be leveraged to share experiences and knowledge in the platform area.
Is there still a skill gap that requires attention? Further study is necessary to compile a list of employees' certifications and to gather more data on cloud services usage and skills.

Does this analysis support decision-making related to exploring opportunities for growth as a trusted service provider in new areas? Results could be utilized to strategize and estimate the necessary efforts. Suppose the company sets a business goal to, within a year, assemble a team of experts within to build a center of excellence in one of the major cloud platforms (Google Cloud Platform, Amazon AWS, Microsoft Azure) in the next 12 months. Let's specify that the target market area is Finland, serving both the public and private sectors. The first question is, how do we choose the right platform? Often, the first customer sets the path, but let's consider an alternative, data-

driven solution. Investigation of what is happening in the Finnish market may give answers. Microsoft is known to be building a data center in Finland, while Google, which already has a local data center, is gaining popularity among industries that value local data storage. AWS, on the other hand, continues to be a leading cloud provider. Microsoft is also actively promoting OpenAI Generative AI solutions to enterprises.

The next step could be to determine the effort needed to achieve partner status with each cloud provider. Additionally, a survey should be conducted within the organization to understand employee perspectives and interests. Azure may appear to be the obvious choice in the Finnish market. It is also important to conduct some competitor analysis.

This was an example of how data could be utilized in data-driven decision-making.

4.4 Results

The "Handbook First" approach challenges the traditional view of documentation as an afterthought. It was recognized as an effective strategy for developing practices and processes in strategic data management. This approach also establishes a conceptual basis for understanding data concepts and theory.

A key outcome of this strategic approach to data governance and management was the increased awareness of the company's data and its location. Emphasizing data ownership and viewing data as a product engaged people to consider data as an asset.

Another lesson learned was that data quality is crucial. Good practice involves ensuring quality in original business systems. This can be accomplished by collecting and validating the right data immediately upon entry into the business system. Improvements were made to the forms used in business systems to achieve this.

The data mesh approach was applied for operational data, which led to a degree of decentralization in data architecture. This was achieved by using data visualization in dashboards and analytics features in the new version of the ODOO ERP. However, more detailed data analysis was still

performed in Tableau and Power BI with minimal integration. Meanwhile, the task force team influenced the selection of a new competence management system, Agileday.io and proposed new features to be included in the product. They introduced the concept of integration with the organization's data sources and aggregation data from sales, projects and competence management to meet specific use cases innovated during data strategy development.

The company's data-drivenness has entered a new phase with the development of a data strategy. Improvements in data quality have been achieved through various actions, including upgrading the ODOO ERP platform and implementing a new competence management system. Business leaders have embraced the concept of data ownership and shown interest in adopting a mindset that treats data as an asset.

To enhance the organization's way of working, the Chief Process Owner role was established. This role recognizes data as an asset and emphasizes the importance of data quality.

New ideas were collected on how to keep employee competence data relevant and up to date, as well as how to measure it. Gamify method with key metrics was introduced. Additionally, improvements were made in integrating competence, sales, and project data. Data literacy course material was planned to be implemented in the company learning portal. A successful future vision was set with a roadmap set for future enhancements.

5 IMPROVEMENT PROPOSALS

This chapter introduces improvement ideas in five areas for the case company: Data Strategy, People and Culture, Data Architecture and Methods, Tools, and Data Skills and Literacy.

First, continually align the data strategy with the business strategy. The foundation of data strategy is understanding the value of company data, discovering, modelling, cataloguing, ensuring data quality, and governing it. Once the foundation is set, start discovering useful external datasets to enhance market analysis and find business opportunities. Treat the data strategy as a living entity and continuously update the handbook for new use cases and improving content.

Second, promote open communication and a decision-making culture based on data. Leadership should set an example in this.

Third, monitor the maturity of data technology and methods. Use the Data Fabric framework as a guide in the world of data architecture. Maintain data ownership and data products by consistently updating responsibilities as the organization changes.

Fourth, use the handbook as a communication tool and provide open data access for everyone in the company. Ensure ethical and correct usage is understood.

Fifth, educate people in data literacy using the company learning portal and handbook. Use gamification as a motivator. Provide IT, Data Analytics, Data Science, and Data Leaders with opportunities to grow their skills through learning and applying their new skills. Promote data roles like Chief Data officer and data evangelist as needed.

Lastly, choose key metrics and monitor progress in the data maturity journey.

6 CONCLUSION

Like any other asset in an enterprise, data requires careful maintenance, governance, and lifecycle management. Despite these needs, there is significant potential to increase operational efficiency and gain a competitive advantage. The benefits are not always achieved through monetization. A defensive data strategy can reduce operating expenses and assist with risk management and fraud detection. Also, alignment with laws and regulations, such as GDPR, can help minimize the risk of sanctions. On the other hand, an offensive data strategy aims to better understand customer behavior and the market situation, leading to improved strategic decisions and innovation in product offerings.

Implementing data strategy, data governance, and data management can seem like an overwhelming long-term process. Therefore, it is advisable to split the implementation into smaller pieces. This minimum viable product (MVP) approach can help achieve results faster and verify if the expected business outcomes are accurate. The old method of doing everything in advance will only incur costs and eat away business outcomes over time.

On the other hand, you may want to look at the Data Fabric offerings from vendors like IBM, Microsoft, HPE and others to make things easier and implement a hybrid model what comes to centralization and decentralization of data architecture. The data mesh model of federated governance and data products can be implemented together with Data Fabric data architecture.

AI needs its own strategy and governance but can offer business outcomes by automating routine tasks and in the near future will help in data management also. The current state is to govern Generative AI usage in organization and how company data is used with Generative AI tools in public services and local use cases like AI assistants in office tools and operating systems in laptops. If seen as a business opportunity should AI have its own strategy.

7 DATA GLOSSARY

A collection of data definitions was generated with the help of Perplexity.ai a Generative AI tool by giving the list of definitions.

Data-driven decision:

A data-driven decision is a decision that is based on data and analysis rather than intuition or guesswork.

Example:

A retail company might use data-driven decisions to optimize pricing, inventory, and marketing strategies.

Data-driven leadership:

Data-driven leadership is the practice of using data and analysis to inform leadership decisions and strategies.

Example: A healthcare organization might use data-driven leadership to optimize patient care, research, and operations.

Data-driven organization:

A data-driven organization is an organization that uses data and analysis to inform its decisions and strategies.

Example: A financial services company might be a data-driven organization that uses data and analysis to optimize risk management, fraud detection, and customer experience.

Data Strategy:

A data strategy is a plan that outlines how an organization will manage, use, and protect its data to achieve business goals. It includes data governance, data management, data analytics, mining, and AI use.

Example:

A company might develop a data strategy to improve customer insights by integrating data from multiple sources, implementing data quality measures, and using data analytics tools.

Data Governance:

Data governance is the process of managing the availability, integrity, and security of data throughout the data lifecycle and across domains of an organization. It involves policies, roles, standards, metrics, and accountabilities.

Example:

A data governance team might establish policies for data access, data quality, and data sharing to ensure that data is consistent, trustworthy, timely, and not misused.

Data Governance Framework:

A data governance framework is a set of policies, procedures, and standards that guide the management and use of data within an organization. It includes roles, responsibilities, and accountabilities for data governance.

Example:

A data governance framework might include a data governance council, data stewards, data owners, and data users, with clear roles and responsibilities for data management and use.

Data Literacy:

Data Literacy describes competencies when working with data, i.e. the ability to work with, analyze, communicate, and argue with data with an understanding of the data sources and constructs, analytical methods and techniques applied, as well as of the use case application and resulting business value or outcome.

Example:

A company might invest in data literacy training for its employees to enable them to make data-driven decisions.

Business Insight

Definition: A business insight is a piece of information or knowledge that is derived from data analysis and can be used to inform decision-making, improve processes, or achieve business objectives.

Example:

A business insight for a manufacturing company might be the identification of a bottleneck in the production process that can be optimized to increase efficiency.

Business Insight tool

Definition:

A business insight tool is a software application or platform that is used to analyze data and derive business insights.

Example:

A business insight tool for a retail company might be a dashboard that shows customer, product, and sales data and provides insights into customer behavior and preferences.

Data Owner:

A data owner is a person or group responsible for managing and governing a specific data asset, including data quality, data security, and data access.

Example:

A data owner might be a data steward, a data manager, a business analyst, or a subject matter expert, with clear roles and responsibilities for data management and governance.

Data Asset:

A data asset is any data or information that has value to an organization. It includes structured and unstructured data, metadata, reports, query results, data visualizations, dashboards, Machine Learning models, and connections between databases.

Example:

A data asset might include customer data, sales data, financial data, or supply chain data, with clear definitions, attributes, and metadata.

Data Management:

Data management is the collection, storage, protection, organization, correction, management, and distribution of the enterprise's data. It includes data governance, data architecture, data warehousing, BI management, and document and content management.

Example:

A data management team might implement a Data Lake, data warehouse, or Data Lakehouse to store and manage data and use data analytics tools to extract insights.

Data Product:

A data product is a data asset that is designed, developed, and delivered to meet specific business needs, for analytics, reporting, or decision-making.

Example:

A data product might include data visualization, data reporting, data analysis, or data storytelling features, to help users understand and use data effectively.

Data Catalog:

A data catalog is a data and metadata management tool that companies use to inventory and organize the data within their systems. It uses metadata to create an informative and searchable inventory of all data assets.

Example:

A data catalog might include data from various sources, such as databases, Data Lakes, data warehouses, APIs, and cloud storage, and provide search, discovery, and lineage features.

Data Lineage:

Data lineage is the process of recording, visualizing, and understanding data as it flows from data sources to consumers, including all the transformations of data along the way.

Example:

A data lineage tool might provide a map of the data journey, from origination to consumption, with details on data origin, transformations, and destinations.

Data Quality:

Data quality is the overall accuracy, completeness, and consistency of data. It also refers to how well the data complies with regulatory requirements and security standards.

Example:

A data quality team might implement data quality measures, such as data profiling, data cleansing, data validation, and data monitoring, to ensure that data is accurate, complete, and consistent.

Data Cleansing:

Data cleansing is the process of identifying and correcting errors, inconsistencies, and inaccuracies in data.

Example:

A retail company might use data cleansing tools to ensure the accuracy and completeness of customer data.

Data Consumption:

Data consumption is the process of using data to support decision-making, business planning, and compliance. It includes data visualization, data reporting, data analysis, and data storytelling.

Example:

A data consumption tool might provide data visualization, data reporting, data analysis, and data storytelling features, to help users understand and use data effectively.

Data Consumer:

An individual or system that accesses and uses data for analysis, reporting, or decision-making.

Example:

A person in a business organization uses the Business Insight tool dashboard in decision-making. Or system uses data from another system to enrich its process data with weather data.

Data Producer:

A data producer is an individual or group that creates or generates data. A Producer can be a person, sensor, bot, or other source of data.

Example:
A manufacturing company might have data producers for production data, supply chain data, and customer data. People produce data as part of daily work. Sensors produce data from manufacturing processes.

Data Interoperability:

Data interoperability is the ability of different data systems and applications to exchange and use data effectively.

Example:

A data interoperability tool might provide features for data integration, data transformation, data mapping, and data synchronization, to ensure that data can flow seamlessly between different systems and applications.

Data Sharing:

Data sharing is the process of making data available to different users, groups, or organizations, for collaboration, analysis, or decision-making.

Example:

A data-sharing tool might provide features for data access, data sharing, data collaboration, and data security, to ensure that data can be shared effectively and securely.

Data Architecture:

Data architecture is the design, implementation, and management of data systems and applications, to support business goals and objectives.

Example:

A data architecture tool might provide features for data modelling, data integration, data transformation, data mapping, and data synchronization, to ensure that data can be managed and used effectively and efficiently.

Data Warehouse:

A data warehouse is a central repository of data that is integrated from various sources, for reporting, analysis, and decision-making.

Example:

A data warehouse tool might provide features for data integration, data transformation, data mapping, data synchronization, and data reporting, to ensure that data can be managed and used effectively and efficiently.

Data Mart:

A data mart is a smaller, more focused version of a data warehouse that is used for specific business functions or departments.

Example:

A healthcare organization might have a data mart for patient data, clinical data, and research data for specific departments, such as oncology or cardiology.

Data Lakehouse:

A Data Lakehouse is a new data management architecture that combines the best features of Data Lakes and data warehouses, for data storage, processing, and analytics.

Example:

A Data Lakehouse tool might provide features for data storage, data processing, data analytics, data integration, data transformation, data mapping, and data synchronization, to ensure that data can be managed and used effectively and efficiently.

Data Lake:

A Data Lake is a large, centralized repository of raw data, in its native format, for analytics and Machine Learning.

Example:

A Data Lake tool might provide features for data storage, data processing, data analytics, data integration, data transformation, data mapping, and data synchronization, to ensure that data can be managed and used effectively and efficiently.

Data Pipeline:

A data pipeline is a workflow for moving and transforming data from one system or application to another, for analytics, reporting, or decision-making.

Example:

A data pipeline tool might provide features for data integration, data transformation, data mapping, and data synchronization, to ensure that data can be moved and transformed effectively and efficiently.

Data Supply Chain:

A data supply chain is the process of collecting, processing, and delivering data from various sources, for analytics, reporting, or decision-making.

Example:

A data supply chain tool might provide features for data integration, data transformation, data mapping, and data synchronization, to ensure that data can be collected, processed, and delivered effectively and efficiently.

Data Fabric:

Data Fabric is an emerging data management design for attaining flexible, reusable, and augmented data management through metadata.

Example: A Data Fabric tool might provide features for data integration, data transformation, data mapping, data synchronization, and data governance, to ensure that data can be managed and used effectively and efficiently.

Data Mesh:

Data mesh is an emerging data management architecture that emphasizes decentralized data ownership, data as a product, and data as a platform.

Example:

A data mesh tool might provide features for data management, data governance, data integration, data transformation, data mapping, and data synchronization, to ensure that data can be managed and used effectively and efficiently.

API:

An API (Application Programming Interface) is a set of protocols, routines, and tools for building software and applications, by connecting different systems and applications.

Example:

An API might provide features for data integration, data transformation, data mapping, and data synchronization, to ensure that data can be moved and transformed effectively and efficiently.

Data Analytics:

Data analytics is the process of examining, cleaning, transforming, and modelling data to discover useful information, draw conclusions, and support decision-making.

Example:

A data analytics team might use Machine Learning algorithms to analyze customer data and predict buying behavior or use data visualization tools to present insights to stakeholders.

Data Visualization:

Data visualization is the process of representing data in a graphical or visual format.

Example:

A marketing company might use data visualization tools to represent customer data in charts, graphs, and dashboards.

GDPR:

GDPR (General Data Protection Regulation) is a regulation that protects the privacy and personal data of individuals in the European Union.

Example:

A GDPR tool might provide features for data privacy, data security, data access, data deletion, and data portability, to ensure that data can be managed and used effectively and ethically.

European Union's Data Act:

The European Union's Data Act is a regulation that aims to create a single market for data in the European Union, by promoting data sharing, data access, and data portability.

FAIR Principles:

The FAIR principles are a set of guidelines for making data findable, accessible, interoperable, and reusable. It includes data discovery, data access, and data reuse.

Example:

The FAIR principles might be used to ensure that data is accessible and reusable by researchers and other stakeholders.

Data Glossary references:

European Commission 2024. Data Act. Search date 13.04.2024. https://digital-strategy.ec.europa.eu/en/policies/data-act

Gdrp.eu; Proton AG; Wolford, B. What is GDPR, the EU's new data protection law? Search date 13.04.2024. https://gdpr.eu/what-is-gdpr/

Genestack 2024, Glossary of data management terms. Search date 13.04.2024. https://genestack.com/resources/library/glossary-of-data-management/

National Library of Medicine 2024. Data Glossary. Search date 13.04.2024. https://www.nnlm.gov/guides/data-glossary

Novisto 2024. Data Management Glossary. Search date 13.04.2024. https://novisto.com/data-management-glossary/

SAP 2024. Data management glossary. Search date 13.04.2024. https://www.sap.com/uk/insights/data-management-glossary.html

Tietoevry, Mäkeläinen, J 2020. Learn the key concepts of data. Search date 13.04.2024. https://www.tietoevry.com/en/blog/2020/05/use-our-data-glossary-to-master-the-terms-of-the-data-world

REFERENCES

Allouin, Alexandre 2022. Data-driven organization with managers on board. Search date 20.12.2023. https://towardsdatascience.com/data-driven-initiatives-unpacked-bringing-managers-to-the-table-2c3e8f8f0b0

Alto, Valentina 2023. Modern Generative AI with ChatGPT and OpenAI models. Packt Publishing. https://learning.oreilly.com/library/view/modern-generative-ai/9781805123330/

Amazon AWS training 2022. Webinar: Fundamentals of Modern Data Strategy. Search date 16.02.2022. https://training.resources.awscloud.com/on-demand-events/vod-ve-fundamentals-of-modern-data-strategy

Anderson, Carl. 2015. Creating a Data-Driven Organization Search date 20.12.2023. O'Reilly Media, Incorporation. https://learning.oreilly.com/library/view/creating-a-data-driven/9781491916902/

BARC GmbH 2023. Generating Value with Data: World's Largest Survey Reveals Data Management Trends. Search date 12.03.2024. https://barc.com/news/generating-value-with-data-worlds-largest-survey-reveals-data-management-trends/

BARC GmbH 2023. Press release: Generating Value with Data: World's Largest Survey Reveals Data Management Trends. Search date 12.03.2024. https://barc-research.com/press-release-data-management-survey-24/

Bean, Randy & Davenport, Thomas 2021. Fail Fast, Learn Faster. Wiley. https://learning.oreilly.com/library/view/fail-fast-learn/9781119806226/

Bean, Randy 2022. Why Becoming a Data-Driven Organization Is So Hard. Harvard Business Review 24.02.2022. Search Date 19.02.2023. https://hbr.org/2022/02/why-becoming-a-data-driven-organization-is-so-hard

Bellmare, Adan. 2022. Practical Data Mesh. Building a Decentralized Data Architectures with Event Streams. Confluent Inc. https://www.confluent.io/resources/ebook/data-mesh-architectures-with-event-streams/

Christ, Jochen & Visengeriyeva, Larysa & Harrer, Simon 2023. Data Mesh from an Engineer Perspective. Search date 16.03.2024. https://www.datamesh-architecture.com/

Confluent 2024. Stream Governance on Confluent Cloud. Search date 16.03.2024. https://docs.confluent.io/cloud/current/stream-governance/index.html

Cotton, Richie 2023. The Data-Information-Knowledge-Wisdom pyramid. Search date 02.04.2024. https://www.datacamp.com/cheat-sheet/the-data-information-knowledge-wisdom-pyramid

Crabtree, Matt 2023. What is Data Literacy? A Guide for Data & Analytics Leaders. Datacamp. Search date 20.04.2024. https://www.datacamp.com/blog/what-is-data-literacy-a-comprehensive-guide-for-organizations

Curry, Maritza 2018. Enterprise Data Strategy. Search date 06.01.2024. https://slideplayer.com/slide/14814824/

DalleMule, Leandro & Davenport, Thomas 2017. What is your Data Strategy? Harvard Business Review May-June issue 2017. Search Date 19.02.2023. https://hbr.org/2017/05/whats-your-data-strategy

Data Centric Inc 2021. How to Create a Data Strategy and Vision. Search date 06.01.2024. https://youtu.be/kQmVHAQnBX0?si=IfF6vId2F9CM3vxG

Davenport, Thomas & Harris, Jeanne G & Morsion Robert 2010. Analytics at Work: Smarter Decisions, Better Results. Harvard Business Review Press

Data Governance Institute 2024. DGI Data Governance Framework Search date 24.02.2024. https://datagovernance.com/the-dgi-data-governance-framework/

Dehghani, Zhamak 2022. Data Mesh. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/data-mesh/9781492092384/foreword01.html

Dehghani, Zhamak 2019. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. Search date 16.03.2024. https://martinfowler.com/articles/data-monolith-to-mesh.html

Deloitte 2020. The cloud migration forecast: Cloudy with a chance of clouds. Search date 06.01.2024. https://www2.deloitte.com/xe/en/insights/industry/technology/technology-media-and-telecom-predictions/2021/cloud-migration-trends-and-forecast.html

Digital and Population Data Services Agency 2024. Tietomallit. Search date 24.3.2024. http://tietomallit.suomi.fi

Edx 2023. What is Data Analytics? Search date 14.05.2024. https://www.mastersindatascience.org/learning/what-is-data-analytics/

Firican, George 2020. 1. How to select a data governance maturity model. Search date 14.04.2024. https://www.lightsondata.com/how-to-select-a-data-governance-maturity-model/

Firican, George 2020. 2. IBM data governance maturity model. Search date 14.04.2024. https://www.lightsondata.com/data-governance-maturity-models-ibm/

Gartner 2024. Future of Data Architecture. Search date 16.03.2024. https://www.gartner.com/en/data-analytics/topics/data-architecture

Gartner 2024. Information Technology Glossary. Search date 24.02.2024. https://www.gartner.com/en/information-technology/glossary

Gartner 2023. What is Data Strategy? Search date 19.12.2023. https://www.gartner.com/en/information-technology/glossary/data-strategy

Gartner 2024. Master Data Management (MDM). Search date 05.04.2024. https://www.gartner.com/en/information-technology/glossary/master-data-management-mdm

Gartner 2023. Hype Cycle for Data Management. https://www.gartner.com/en/documents/4573399 Document available only for Gartner clients.

Google 2023. What is Data Governance? Search date 25.12 2023 https://cloud.google.com/learn/what-is-data-governance

HL7 International 2024. HL7 Standards. Search date 24.3.2024. https://www.hl7.org/implement/standards/index.cfm?ref=nav

IBM 2024. 1. What is a data catalog? Search date 24.02.2024. https://www.ibm.com/topics/data-catalog

IBM 2024. 2. What is Data Architecture? Search date 24.02.2024. https://www.ibm.com/topics/data-architecture

IBM 2024. 3. Design your strategy in six steps. Search date 01.05.2024. https://www.ibm.com/resources/the-data-differentiator/data-strategy

Joubert, Shayna 2019. Data-driven Decision making: A Primer for Beginners. Search date 27.12.2023. https://graduate.northeastern.edu/resources/data-driven-decision-making/

Klidas Angelika & Hanegan Kevin 2022. Data Literacy in Practice. Packt Publishing. https://learning.oreilly.com/library/view/data-literacy-in/9781803246758/

McKinsey 2017. Capturing value from your customer data. Search date 11.03.2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/capturing-value-from-your-customer-data

McKinsey 2024. A Generative AI reset: Rewiring to turn potential into value in 2024. Search date 15.05.2024. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/a-generative-ai-reset-rewiring-to-turn-potential-into-value-in-2024

Mäenpää, Markku & Vihervaara Tommi 2021. Datapääoma podcast: Tätä Datatoimitusketju on. Search date 21.03.2023. https://open.spotify.com/episode/4qrQpu9o3Vj0GGxFBGLMqi?si=10d9d8d78b98478e

Nadella, Satya 2014. A data culture for everyone. Search date 20.04.2024. https://blogs.microsoft.com/blog/2014/04/15/a-data-culture-for-everyone/

Needham, Graham 2022. What Is the State of Data-Driven Culture for Better Business Decisions? Search date 23.12.2023. https://www.accurity.ai/blog/what-is-the-state-of-data-driven-culture-for-better-business-decisions

NewVantage Partners a Wavestone Company. Data And Leadership Executive Survey 2022. Search date 10.03.2024. https://www.wavestone.com/app/uploads/2022/01/Wavestone-2022-Data-and-AI-Leadership-Executive-Survey-Report-1.pdf.

Obstler, Rachel 2023. The four stages of data maturity – how to ace them. Search date 14.04.2024. https://www.heap.io/blog/the-four-stages-of-data-maturity

Onis, Teresa de 2016. The Four Stages of the Data Maturity Model. CIO article. Search date 15.04.2024. https://hservers.org/kobo/Papers/The%20Four%20Stages%20of%20the%20Data%20Maturity%20Model.pdf

Panetta, Kasey & Gartner 2021. A Data and Analytics Leader's Guide to Data Literacy. Search date 20.04.2024. https://www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy

Petersson, David 2023. AI vs Machine Learning vs. Deep Learning: Key differences. Search date 09.05.2024. https://www.techtarget.com/searchenterpriseai/tip/AI-vs-machine-learning-vs-deep-learning-Key-differences

Perälä, Arto 2018. Datan standardoinnin viisi hyötyä liiketoiminnalle. Search date 15.05,2024. https://www.loihdeadvance.com/blogi/datastandardi-hyodyt-liiketoiminnalle

Plain Concepts, 2024. Discriminative AI vs Generative AI: Keys to understanding them. Search date 09.05.2024. https://www.plainconcepts.com/discriminative-ai-vs-generative-ai/

Pratt, Mary K & Luna, Melanie. 2024. What is Data Stewardship? Search date 07.04.2024. https://www.techtarget.com/searchdatamanagement/definition/data-stewardship

Precisely 2024. The Role of Technology in Making Data-Driven Strategic Decisions. Search date 07.04.2024. https://www.precisely.com/blog/datagovernance/data-driven-strategic-decisions

Qlik 2024. Data Governance. Search date 14.05.2024. https://www.qlik.com/us/data-governance

Rouse, Margaret 2018. Data-Driven Decision Making. Search date 27.12.2023. https://www.techopedia.com/definition/32877/data-driven-decision-making-dddm

Rowley, Jennifer 2007. The wisdom hierarchy: representations of the DIKW hierarchy. Search date 02.04.2024 https://api.semanticscholar.org/CorpusID:17000089

Serra, James 2024. Deciphering Data Architectures. Search date 23.03.2024. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/deciphering-data-architectures/9781098150754/ch13.html

Strengholt, Piethein 2023. Data management at Scale 2nd edition. O'Reilly Media Inc. Search date 08.04.2024. https://learning.oreilly.com/library/view/data-managementt-at/9781098138851/ch08.html

Tableau 2024. Data management: What it is, Importance, and Challenges. Search date 11.03.2024. https://www.tableau.com/learn/articles/what-is-data-management

Tableau 2024. 5 Key steps to creating a Data Management strategy. Search date 11.03.2024. https://www.tableau.com/learn/articles/data-management-strategy

Tableau 2024 Tableau AI: Propel your data culture with Generative AI. https://www.tableau.com/products/tableau-ai

Taylor, Petroc 2022. Data-driven decision-making in organizations worldwide in 2018 and 2020 https://www.statista.com/statistics/1235409/worldwide-data-driven-decision-making-companies/

Tienari, Janne. & Meriläinen, Susan. 2021. Johtaminen ja organisointi globaalissa taloudessa. WSOYpro, Alma Talent Oy.

Treder, Martin 2020. The Chief Data Officer Management Handbook. APress. Search date 07.04.2024. https://learning.oreilly.com/library/view/the-chief-data/9781484261156/html/499948_1_En_BookFrontmatter_OnlinePDF.xhtml

Waller, David 2020. 10 Steps to Creating a Data-Driven Culture. Harvard Business Review 06.02.2020. Search date 19.02.2023. https://hbr.org/2020/02/10-steps-to-creating-a-data-driven-culture

Wallis Ian 2021. Data Strategy. BCS, The Charted Institute for IT. https://learning.oreilly.com/library/view/data-strategy/9781780175430

Wikipedia 2024. Search date 16.03.2024. https://en.wikipedia.org/wiki/Data_mesh