Data Mesh for Managing Complex Big Data Landscapes and Enhancing Decision Making in Organizations

Otmane Azeroual[1][a] and Radka Nacheva[2][b]
[1]German Centre for Higher Education Research and Science Studies (DZHW), 10117 Berlin, Germany
[2]Department of Informatics, University of Economics, Varna, 9002 Varna, Bulgaria

Keywords: Big Data, Data Management, Data Warehouse, Data Lake, Data Swamp, Data Lakehouse, Data Mesh, Data Fabric, Data Discovery, Decision Making.

Abstract: In the age of digitization, data is of the utmost importance. Organizations can gain competitive advantage by being ahead of the curve in organizing data, deriving insights from it, and turning those insights into action. In practice, however, many organizations fail to meet this challenge. Far too many decisions are made without data, decision makers don't trust their own data. The data warehouse, later the data lake and more recently the data lakehouse have been propagated as solutions to these problems in recent decades. In some cases, this actually succeeds, in other cases challenges remain. The recently prominent data mesh approach changes the perspective on data and in this respect provides valuable impulses for data architectures in general. Data mesh is a new architectural concept for data management in organizations. Therefore, in this paper, we introduce this new data concept and provide a clear overview of the design of a data mesh architecture. We will then show how it can be technically implemented and what potential there is for using data mesh in organizations. Our methodology is a type of investigation that provides a helpful and practical guide to understanding the principles and patterns of data mesh and their implementation in organizations. Our research result has shown that the data mesh approach is therefore a very good tool for organizations where data sharing and reuse is crucial. In addition to facilitating scalability, data mesh can enable better data integration and data management, improving data quality while fostering a culture of data-driven decision-making.

1 INTRODUCTION

With increasing digitization in numerous areas, e.g. research, industry or health care etc., data-driven process optimization and cost reductions are made possible (Buer, Fragepane & Strandhagen, 2018). For this purpose, a large amount of data is collected, which is very extensive and often heterogeneous, i.e. structured differently (structured, semi-structured and unstructured). That's called big data. Big data is characterized by several characteristics, including high volume, large variety of data, high speed of collection and the potential value that the data contains (Fan, Han & Liu, 2014), (Silva, Diyan & Han, 2019). In this case, we speak of potential value because at the time of collection it is not always clear how and whether these can be used to create value in later use cases. In order to even In order to even potentially utilize the value it contains, the data must be stored, managed and processed (Malik, 2013).

In recent years, organizations have realized that data is at the core. Data enables new efficient solutions, promotes innovations, opens up new business models and increases customer satisfaction (Antikainen, Uusitalo & Kivikytö-Reponen, 2018). Becoming a data-driven organization (leveraging data at scale) remains a top priority for most organizations.

Traditional concepts (as shown in Figure 1) combine that they connect decentralized, operational source systems, load data into a centrally managed system, have it processed by a central team and then return results in the form of reports or results of analytical models.

Data warehouses and data lakehouses are not always suitable for managing this data, especially since the data warehouse only contains a cleaned and

[a] https://orcid.org/0000-0002-5225-389X
[b] https://orcid.org/0000-0003-3946-2416

Azeroual, O. and Nacheva, R.
Data Mesh for Managing Complex Big Data Landscapes and Enhancing Decision Making in Organizations.
DOI: 10.5220/0012195700003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 3: KMIS*, pages 202-212
ISBN: 978-989-758-671-2; ISSN: 2184-3228

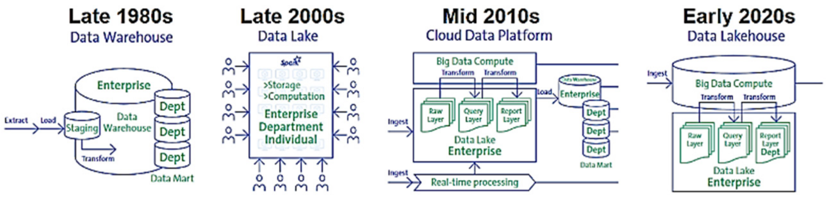

Figure 1: Evolution of data architectures (According to Microsoft via Ralph Kemperdick).

prepared section of the data whose usefulness is known (Borjigin & Zhang, 2021). Other data and information is lost during this processing. Due to the complexity of data warehousing and information systems, the distribution of data across different locations poses a challenge for companies (Shin, 2003). Integrating new data is time consuming and encourages data duplication. Additionally, the point-to-point connectivity makes it difficult for organizations to monitor the full data landscape. The company underestimated the sheer need for intensive data usage. New use cases are introduced quickly and successively. Data governance (e.g. data ownership and quality) and costs are difficult to control (Ladley, 2019). Maintaining ongoing compliance with applicable regulations is difficult because organizations don't know exactly where their data resides. Organizations using a data mesh approach can address the challenges of optimizing data as a strategic asset. The data mesh was therefore developed by Zhamak Dehghani as a new concept alongside data lake for the management and use of big data (Machado, Costa & Santos, 2022). Data mesh is a new architectural concept for data storage in larger companies (Strengholt, 2020). In contrast to the centralization of company data that is common today, the data mesh approach strives for an increasing decentralization of data sovereignty. With the help of data mesh, large amounts of data can be easily structured. Data can be found more quickly, is generally accessible and secure. This architectural approach also helps organizations in decision-making and ensures a faster value chain. Data mesh is not only a technical concept, but also an organizational one (Araújo Machado, Costa & Santos, 2022). Data mesh is currently one of the most discussed hype terms in IT and especially in the data and analytics context.

Based on the current challenges, the following research question arises for organizations that want to implement a data mesh: How must a data mesh be structured in order to support the management and processing of big data in practice? In order to answer this research question, with our contribution we try to consolidate the background knowledge of the term data mesh and the currently known approaches and functions and to show the implementation possibilities of data mesh. So far, this new topic has only been considered in practice and companies have benefited from it when managing data. In the scientific literature, most authors in research have treated and discussed the topic in a very abstract way with data warehouse and data lake. Therefore, we would like to demonstrate the relevance of this topic, because the data mesh concept is currently being discussed so interestingly in the data community that it could actually become the next widespread design pattern for data. The big innovation here is not that a new technology is introduced, but that the problems of centralization are to be solved by changed organizational, data governance and data culture measures.

The paper is divided into eight sections. After the introduction in Section 1, Section 2 introduces the theoretical foundations and the potential for using them on new topics such as data mesh and data fabric, and discusses the related work in the literature. Section 3 defines the term data mesh and explains its architecture in detail using four principles. Section 4 gives an overview of the possible potential use of the topic of data mesh. Section 5 explains and describes the possible implementation of data mesh and presents a practical example of the Snowflake Data Cloud Platform. This is followed by a guide to data mesh strategy and execution. Section 6 presents a practical application of our data mesh proposal through a case study. Section 7 discusses the best practices of data mesh deployment through theoretical and practical implications. Section 8 summarizes the main findings of the paper and outlines future work on data mesh.

2 LITERATURE REVIEW

Organizations must continually rethink and adapt their data strategies, architectures, and management systems to create value from an ever-growing volume of data and remain competitive in the field of data science. Various terminologies have emerged in connection with the related concepts in the past, including terms such as data warehouse, data lake and more recently data lakehouse, data mesh and data fabric (Shrivastava, et al. 2022). Among modern data architectures, data mesh and data fabric stand out (Strengholt, 2020). These approaches are frameworks that can help to master these new challenges in different organizations. Because the concepts are abstract, they cannot be used only for a specific product, technology, or industry (Pithadia, et al. 2023). Depending on the use case, data mesh and data fabric can instead take different forms. The terms and their differences are clearly defined in the literature (Strengholt, 2020), (Butte & Butte, 2022), (Bode et al. 2023). Data fabric is an architectural framework that enables simplified access to enterprise data and delivers it in the right way, at the right time, and to the right user (Bode et al. 2023), (Macías, et al. 2023). In this way, data fabric ensures a clear and uniform view of different services and technologies. Technologically, the data structure consists of a service package that sits between the data source and the user. The integration of the individual services takes place via different processes that influence the life cycle of the data and can be divided into different layers. This approach can provide several benefits (Hechler, Weihrauch & Wu, 2023):

- At the enterprise level, users can make data-driven decisions and take action, making the experience faster and more personalized
- Data management can benefit from automated and less expensive data lifecycle activities
- From an organizational perspective, the gap between data professionals and the enterprise level is narrowing

Data mesh is referred to in the literature as an architectural framework based on the concept of the domain (Machado, Costa & Santos, 2022). The data is treated as a product and maintained by the team that has the functional understanding of that data. A domain can be viewed as a high-level category associated with a specific business function and not systems or applications. Each domain is defined by its own internal process and pipelines. These run on a common infrastructure. In addition, each domain is unique in terms of the data it provides and the operations that can be performed on it. This approach can benefit various areas (Bode, et al. 2023), (Hechler, Weihrauch & Wu, 2023):

- At the enterprise level, it enables the democratization of data using a self-service approach
- It helps with data management by simplifying the way data can be retrieved
- Within the organization, it enables faster data exchange between producers and consumers

Data mesh and data fabric are approaches in data architecture that aim to improve the effectiveness and efficiency of data management within an organization. The main difference between the two approaches lies in the way data is processed and used (Strengholt, 2020).

3 DATA MESH ARCHITECTURE

3.1 Data Mesh vs. Data Lake

Data mesh is an organizational concept for data and for the organization that manages the data (Dončević, et al. 2022). Data mesh was first developed by Zhamak Dehghani, who worked at Thoughtworks at the time of initial publication (Dehghani, 2022). In principle, it is similar to the domain-driven design approach used in software development for some time and uses the insights gained from building robust, Internet-based solutions to unlock the true potential of enterprise data (Dehghani, 2022). The basic idea is to achieve decentralization of data, maximum technological support from one platform and minimum centralized governance to ensure interoperability and scale-out for data.

A data mesh is a distributed data architecture in which data is organized by domain to provide better access for users in an organization (Machado, Costa & Santos, 2022). A data lake is a low-cost storage environment that typically stores petabytes of structured, semi-structured, and unstructured data for business analytics, machine learning, and other large-scale applications (Liu, Isah & Zulkernine, 2020). A data mesh is an architectural approach to data in which a data lake can be embedded (Castro, et al. 2020). However, a central data lake is typically used more as a dumping ground for data, as it is often used to house data that does not yet have a defined purpose. This can result in it becoming a data swamp, i.e. a data lake that lacks the appropriate data quality and data governance practices to generate meaningful insights.

3.2 Data Mesh Architecture with Four Principles

The data mesh concept includes data, technology, processes and organization. At the conceptual level, this is a democratized approach to data governance, with different domains operationalizing their own

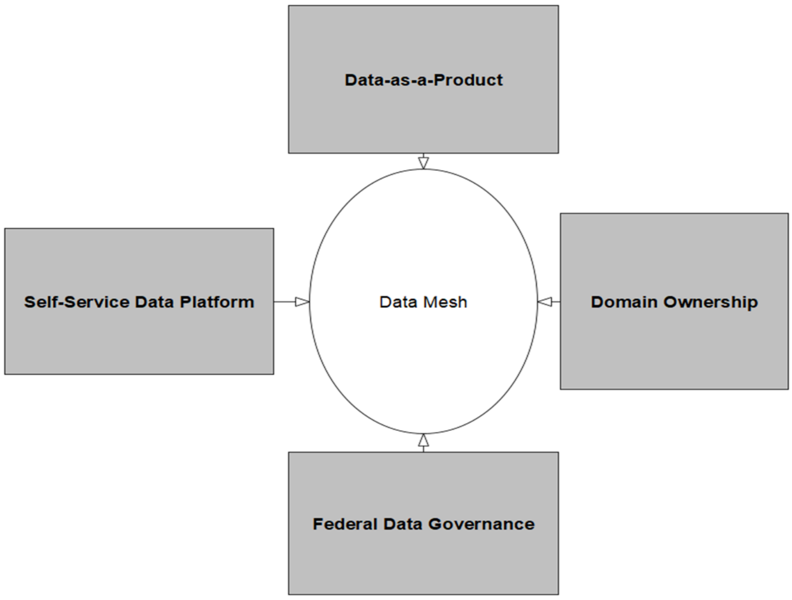

Figure 2: Data Mesh Principles.

data (Strengholt, 2020). Data mesh challenges the idea of traditional centralization of data: instead of looking at data as one big repository, data mesh looks at the decomposition of independent data products (Podlesny, Kayem & Meinel, 2022). This shift from centralized to federated ownership relies on a modern, self-service data platform typically built with cloud-native technologies. Regardless of the technology, a data mesh concept is based on four principles (see Figure 2).

The four principles of the data mesh concept will be explained here in individual modules:

1. Principle 1 - Data-as-a-Product: Data is seen as a product owned by the team that publishes it. Data mesh obliges the specialist teams to be responsible for their data. The team owns this data and must ensure the quality, consistency and presentation of their data. Only the use of the data products shows whether the development process was successful. Data products should not suffice the developers, but justify themselves through the application. This principle projects a philosophy of product thinking onto analytical data.
2. Principle 2 - Domain Ownership: Data is segmented by line of business, down to the line of business that is closest to the data - either the source of the data or its primary consumers. Following this principle, organizations should ideally define and model each data domain node within the network using domain-oriented design. It must decompose the analytical data logically and based on the business domain it represents, and independently manage the life cycle of the domain-oriented data.
3. Principle 3 - Federal Data Governance: The primary goal of this principle is to create a data ecosystem that adheres to organizational rules and industry regulations while ensuring the interoperability of all data products. The interaction of the various principles makes it clear that a major challenge lies in an efficient framework that largely automatically ensures the implementation of the high requirements. Topics such as data protection, data lineage, uniform interfaces must be considered and tested before implementation. Due to the decentralized responsibility and development of the various data products, there is a risk of data silos that can no longer be resolved or dependencies between the individual products. To ensure that each data owner can trust the others and share their data products, a data governance department must be established in the organization to implement data quality, centralized data ownership visibility, data access management, and privacy policies.
4. Principle 4 - Self-Service Data Platform: Data is available in a data mesh virtually anywhere in the organization. For example, it can create a sales forecast for a specific product in a German market. In this case, all the data required for a meaningful report should ideally be available within a few minutes. There is no need to wait

until the requirement is prioritized, planned and implemented. Data mesh starts with a self-service data platform that allows users to abstract away the technical complexity so they can focus on their unique data use cases.

The four principles enable data to be held accountable by those with domain-specific knowledge and empowered to process and disseminate that data through a self-service data platform, skills are used more effectively, and data quality is improved. Data users can independently retrieve and reuse the data they need. In this way, added value can be independently generated from the data. Creating this possibility of being able to participate in the data company-wide is one of the central building blocks for the successful use of data science in the company. The data mesh concept contributes to this and is more effective and scalable than a central collection in a data warehouse or data lake. The end result is a network of data products that are made available to others by the domain teams so that the data products can be shared across teams. Figure 3 shows a paradigm by adding business domains and their domain data products with their interfaces.

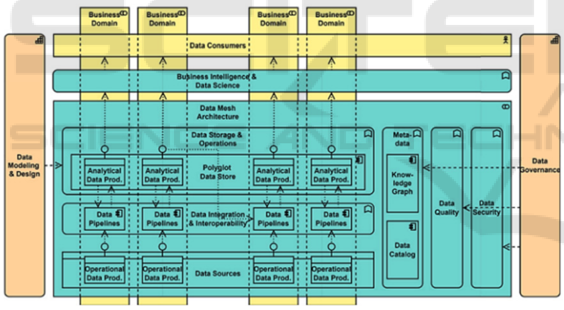

Figure 3: Data Mesh Architecture (Priebe, Neumaier & Markus, 2022).

The authors (Priebe, Neumaier & Markus, 2022) explain that the data pipelines also belong to the business domains, which means that each domain is responsible for its own data transformations. A domain can consume data products from another domain. As with data fabric, the focus is on metadata with a data catalog that is a cross-domain inventory of available data products. As with data fabric, reporting and analytics tools are not part of the focus (hence "business intelligence and data science" is outside the data mesh architecture box). However, unlike the other data architecture paradigms presented, the data mesh concept takes the data sources into account. Operating data is processed via operating data products (or their interfaces) and analysis data products.

4 DATA MESH POTENTIAL USE

Data science and data engineering teams are relieved enormously in the development of models, the analysis of the data and the maintenance of the platform, since the responsibility for data processing and data quality is transferred to the domain teams (Marr, 2016). The data quality is increased because the data is evaluated by the data producers themselves, who have the domain-specific expertise (Koltay, 2016). From a product perspective, too, the domain teams have an incentive to ensure high data quality (Callegaro, et al. 2014).

The creation of new data-driven solutions is made easier because the teams are empowered to evaluate their data independently. Additionally, domain teams can leverage each other's data products to drive their own work. Responsibility for the data is clearly divided between the respective teams and data analysis and the development of data-driven solutions are accelerated. In addition, the data mesh concept enables more employees to participate in the process of data evaluation and use, which is becoming increasingly important given the growing importance of data in companies. Overall, this decentralization results in a more scalable solution.

Organizations can adopt a data mesh architecture by recognizing the fact that the way data is organized best meets modern business needs and overcomes many of the challenges. Other uses of data mesh are summarized below:

- The decentralized data ownership model accelerates time to insight and time to value by enabling business units and operations teams to quickly and easily access and analyze non-core data. This means that companies are becoming more flexible and agile.
- The data mesh architecture helps organizations make real-time decisions by minimizing the temporal and spatial gap between an event and its analytics processing. The business model becomes significantly more efficient and reacts more quickly to changing trends.
- Data mesh also overcomes the shortcomings of data warehouses and data lakes by allowing data owners more autonomy and flexibility and more data experimentation. It also reduces the burden on data teams who must meet the needs of all data consumers through a single pipeline.

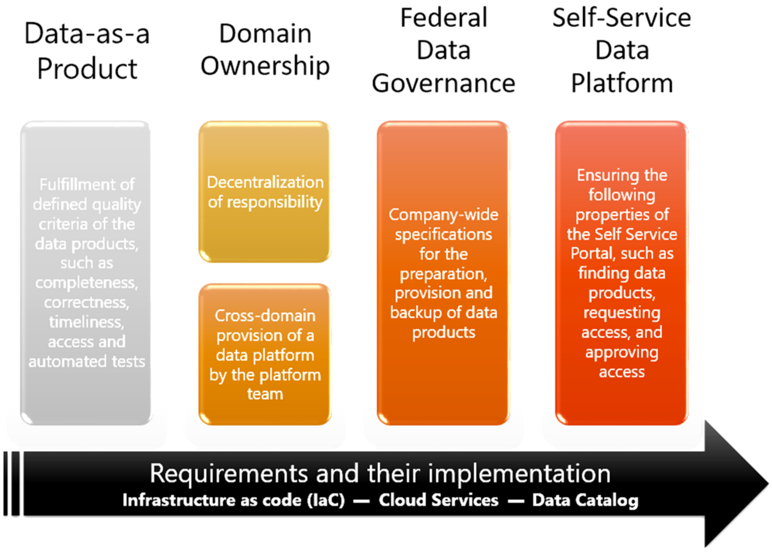

Figure 4: Data Mesh, Requirements and their Implementation.

5 DATA MESH IMPLEMENTATION

More data is generated in the data landscapes of organizations today (Ballard, et al. 2014), (Black, et al. 2023). At the same time, there is a growing desire to harness the value of data for basic and advanced analytics applications. However, data organizations and data architectures do not yet correspond to the new requirements in the field of data analytics and data science. The complexity and size of organizations has created a situation where the agility and immunity with which organizations can create value from data is reduced - unless the data management approach is changed.

From Figure 4, we have highlighted the requirements resulting from the four data mesh principles. The question now arises as to how these requirements can be implemented in a data platform. In this section, we present three technologies that enable the efficient construction of a data mesh.

- Infrastructure as Code (laC): In a data mesh, the platform team must provide each domain with an instance of the data platform. Each of these instances must meet functional requirements for data storage and analytics, as well as core requirements for security, audit, and governance. To ensure every platform is compliant for every domain, platform deployment needs to be automated. The services that make up a data platform are described and configured in a formal language via IaC. The deployment tool processes this formal description and thus ensures that the defined specifications are met in every domain and in every staging environment (development, test, production). The formal description of the platform by IaC guarantees that there are no manual configuration errors in any environment. IaC is also the foundation for delivering a self-service data platform to a domain team.
- Cloud Services: A data platform consists of a number of components that enable data storage and processing, access protection, monitoring and auditing. All of these components must be managed using consistent user administration. All major hyperscalers offer cloud service solutions for this that optimally meet the required requirements, can be integrated and regulate comprehensive access protection. The cloud services can also be provided via IaC and thus represent the ideal basis for the implementation of a data mesh.
- Data Catalog: In a data mesh, responsibility for the data lies with the individual domain teams. The area of responsibility includes the data transmission to the data platform, the processing of the data, the analyzes and the provision of data products. In order to be able to make data-based decisions company-wide, data processing must not end at domain boundaries. Other domains need to be aware of the existence of high-quality data products, enrich that data with their own data and thereby create higher quality

analysis and data products. A data catalog takes on exactly this task in a data mesh. Each domain advertises and manages its data products in the company-wide data catalogue. This describes each data product based on the following information, such as content of the data, quality, and frequency of change, interfaces, and data owner. If another team has found a data product that is of interest to them in this way, access to the data can be requested via the data catalogue. If the data owner agrees to this use, the corresponding authorizations are released by the data catalogue. The data catalog thus represents the central interface to ensure the cross-domain reuse of data products. Without a data catalogue, useful data is often hidden from end users (Buranarach, et al. 2017). As organizations collect more and more data, it tends to be scattered across different data stores. When business and analytics users can't find relevant data, business operations and analytics initiatives are less effective. This is a big problem as companies increasingly want and need to make data-driven business decisions. Data catalogs help eliminate this problem by providing a unified view of data assets with built-in search and data discovery capabilities.

In summary, data mesh addresses the pain points and underlying characteristics that caused failures in older generations of data warehouses and data lakes. Therefore, we propose to switch to these three technologies in total. Infrastructure as code is proven to deliver the services and configure them according to governance requirements, and cloud services are ideally suited to implement a data mesh, and the data catalog ensures that the data products can be reused across domains to become enterprise-wide data-aware and will make decisions. The implementation in any existing data landscape requires a deep understanding and a well-stocked toolbox. These technologies and their right combination for each individual organization require knowledge and know-how at the highest level. They lay the foundation for converting data into values.

Data mesh opens countless possibilities for organizations in various usage scenarios, including behavioral modeling, data analysis and business intelligence. From development to production, all teams can benefit from this decentralized architecture model. Snowflake Data Cloud provides organizations and their lines of business with an excellent foundation for establishing and managing a decentralized data mesh architecture. In this way, local teams can not only share their data with each other as products, but also process data with the same logic and treat it like products. Organizations should have access to tools that help them create, deliver, and consume data products at every stage of the lifecycle, from accessing the right data, through processing and preparation, to analyzing, modeling, and delivering data products to users throughout Company. A powerful self-service infrastructure platform should provide elastic performance to allow departments to access different applications at the same time. This includes rich data pipelines, ad hoc exploration, BI reports, feature engineering, and interactive applications. With such a powerful platform, enterprise architecture can be simplified without sacrificing speed or flexibility. Whether the teams work with SQL, code (e.g. Java, Scala or Python) or a mixture of these, the self-service platform should support them all equally. As data variety and size explodes, a platform must be able to accommodate large amounts of data in different formats. The data must be able to come from different sources and be accessible as products for different users. The platform should also be flexible enough that certain data can be used and made available at the same time. This flexibility or openness that allows a platform to interact with the rest of the organization's ecosystem does not necessarily have to be open source. Snowflake Data Cloud thus ensures that all organizations and their departments as well as central data teams have access to all relevant data at all times without being trapped in silos or complex structures. This is what the Snowflake Data Cloud Platform is based on, which thanks to its cloud capacity stands for scalable performance, user-friendliness, regulated data exchange and collaboration. The platform is ideally suited to support both centralized standards and decentralized data ownership, both essential to a successful deployment of the data mesh. Implementing a data mesh in Snowflake Data Cloud can be based on a variety of topologies: departments or domains can be account-based and leverage secure data sharing capabilities to break down silos across regions and clouds with a single copy of data work. Alternatively, departments or domains can be based on databases or schemas and use catalogs like Collibra's (https://www.collibra.com) to make products discoverable and accessible. In any case, Snowflake Data Cloud can provide independent resources to the lines of business in an organization to load, process, and list their data products using third-party virtual warehouses. These products can then be shared and used via data sharing within the account or database.

6 PRACTICAL APPLICATION BASED ON A CASE STUDY

Data analysis enables organizations to make evidence-based decisions, for example to identify high-risk customers and take countermeasures.

The challenge is that informed decisions require a holistic view of the data. For example, not only does a customer switch suppliers because of the occasional defective part that needs to be replaced, but this risk could increase in combination with delays in delivery due to predictable maintenance intervals of production machines. As a rule, however, the required information is spread across many different applications and thus data sources and is owned by different departments. It is also often not transparent which data from other areas of the organization is available at all. The following example outlines such a use case (see Figure 5). The aim of the customer service department is to use data analysis to identify dissatisfied customers and to proactively initiate countermeasures to ensure customer loyalty. To get a complete picture of the situation, information from different areas of the company is helpful. In the example, data from production and quality assurance are to be used.

A lot of different data is generated in the production area. Information about the production volume and sensor data about the machine condition should be used. Because the department knows their data very well, they know that this raw data is difficult for other departments to understand and use. However, the information about necessary maintenance measures can provide valuable information about production interruptions. Therefore, a data product should be made available for planned maintenance intervals that can be used by data consumers for higher value analysis. For this purpose, the raw data is extracted from the source systems and, in a transformation step, a data set about planned maintenance measures is created. This transformation can be carried out with conventional processing methods, but the use of modern AI methods (such as predictive maintenance) would also be conceivable. The finished data product is made available in the organization via standardized interfaces. Similar to the situation in the production area, quality control also has different types of information. In the example, registered product defects are stored in a relational database and logs of parts replaced due to quality defects are stored in Excel reports. In the transformation step, these two data sources are correlated customer-related and the results are provided as a new product quality data set. This data product is also made accessible to other departments via an interface.

The availability of high-quality, curated data products is in itself an added value for organizations. However, the full potential only unfolds when several data products are linked. In the practical application example, customer service wants to identify customers who are at risk of leaving. The department's analysts can use a data catalog to find the two data products described and use the meta information to get an idea of how they can be used for their own application. The data sets can be easily used via the interfaces offered. It does not matter whether this is done using BI tools, using source code or in

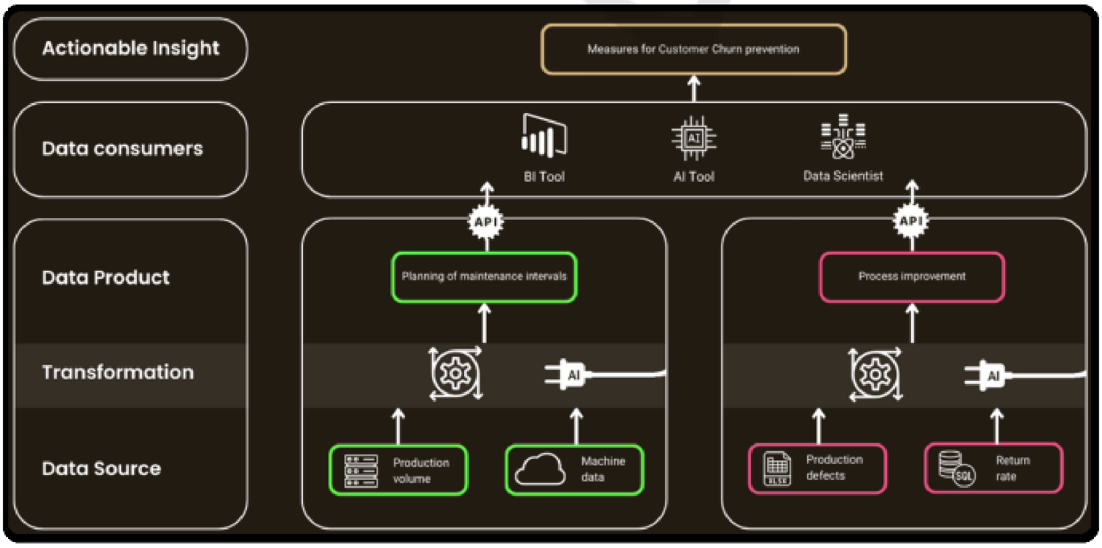

Figure 5: Practical application for data mesh (taken from https://de.steadforce.com/).

some other way. In addition to using the results for their own application, customer service can also make the new data set available to other departments in the organization as their own data product. Using the data mesh approach in our practical application brings four benefits, such as:

1. The creation and administration of the data products is carried out by the departments. It can accurately assess what information is valuable and address compliance issues head-on. A high quality of the data is guaranteed for the users of the customer service.
2. Questions about the data or change requests can be clarified directly with the responsible department without having to go through the central IT department.
3. The raw data is processed at the point of origin. A time-consuming relocation to a central location is no longer necessary. Current data is available more quickly.
4. By searching the central data catalogue, the datasets could be easily found and identified as useful. The data can be used directly without any IT requirements having to be made first. The time until insight is reduced.

Data mesh can be used in a wide variety of industries. For example, an e-commerce company could use data mesh to create different domains for customer data, product data, order data, and marketing data. Each area would independently manage its individual information and make it available to other areas for a deeper understanding of customer needs, product performance and marketing effectiveness. Another example: Data mesh can be used in healthcare organizations. By implementing it, a healthcare organization could create multiple domains for patient data, clinical information, and financial data. This would enable effective organization and management of this data to provide better patient care and more efficient business operations. With data mesh, a healthcare organization could adopt a data-driven approach to their processes, thereby improving their performance and competitiveness. Each domain manages its individual data and makes it accessible to other domains to promote a better understanding of patient care, clinical outcomes and financial performance. The implementation also allows the organization to ensure that the data is updated in real-time and is therefore always up-to-date. This is especially important in the financial industry, where quick and accurate decisions need to be made. Overall, the use of data mesh offers an innovative solution to the challenges facing financial services companies today. Each department is responsible for managing its own data and making it available to other departments to gain a more complete understanding of clients' needs, their transaction history and risk profile. Such an approach could help make more informed lending, fraud prevention and investment decisions.

7 DISCUSSION

This part discusses some theoretical and practical implications of running data mesh as a concept. To decide whether organizations should invest in a data mesh architecture, organizations must consider the number of data sources, the size of the data teams, the number of data domains, and data governance. In general, the larger and more complex these factors are, the more demanding the organization's data infrastructure requirements are, and the more likely the organization is to benefit from a data mesh approach.

Typically, moving to a data mesh architecture is a sensible consideration for teams that need to manage large amounts of data sources and process them into clean data. However, unless the organization's data needs are complex and demanding, data mesh should not be considered just yet. For organizations looking to rapidly evolve and adapt to data modernization, it makes more sense to first adopt some data connectivity best practices and concepts to facilitate migration at a later date.

In the data area, in addition to data mesh, there is also the term data fabric and both are interesting trend topics. Data fabric describes the combined use of several existing technologies to enable metadata-based implementation and advanced design of orchestrations. While the data fabric is based on a flexible ecosystem of software solutions for data use, the data mesh is a special way of data organization. With a data mesh, the data is stored decentrally in its respective area within an organization. Each node has local storage and processing power, and no central point of control is required for operation. In contrast, with a data structure, data access is centralized, with clusters of high-speed servers for networking and sharing powerful resources. There are also differences at the data architecture level. In this way, the data mesh introduces an organizational perspective that is independent of specific technologies. Its architecture follows a domain-oriented design and product-related thinking. Although data mesh and data fabric follow different logics, they serve the same goal: the optimal use of their data stocks and improve access to data. Therefore, despite their differences, one should not weigh them against each other, but see them as complementary.

As reported by (Mikalef, et al. 2020), the biggest challenge in getting value from their investments is not so much the technical issues as embedding these technologies into the organizational structure and using them to drive strategic outcomes. This requires investing in resources that are not purely technical in nature, such as human skills and establishing a data-driven culture and continuous learning. Data mesh as a new technical solution concept promotes the democratization of data, i.e. data should be made accessible to all employees. This can be accomplished by providing tools and resources that enable employees to access and use data across the organization. By enabling employees to access and use data more easily, data mesh can help improve data literacy and data-driven decision-making within the organization. With data mesh, business units gain more control over the data they use and the quality of that data. This can help ensure that the data is aligned with the needs of the business and is more easily accessible and usable by the people who need it.

8 CONCLUSIONS

When does it make sense to integrate data mesh into an organization's data landscape? Data mesh's approach is to keep pace with organizations' ever-increasing data volume and complexity. Enterprises should operate in such a way that decentralization has the potential to improve data architecture and does not introduce unnecessary complexity. A multi-heterogeneously scattered system landscape, increasingly complex data structures, large amounts of data and a diverse group of data consumers can be the consequences of the change that makes data accessible on a large scale. Additionally, there are significant operational costs associated with creating cross-process insights.

By bringing process and data governance together, data mesh can help the organization reduce complexity by breaking down the architecture into smaller pieces. Prioritizing data democratization and striving for data governance as a core activity will help break down data silos in organizations that may also consider moving to data mesh when the information cycle is measured in months or weeks rather than days or hours.

Not choosing the right tools and infrastructure can limit the benefits of a data mesh. Added complexity slows value creation and increases costs. SaaS platforms like Snowflake Data Cloud remove this complexity and reliance on expertise. Provisioning and management of Snowflake Data Cloud resources can be fully automated, using infrastructure as code with the highest level of security and governance, interoperable with any public cloud. The next level is abstracting the complexity of data workflows. Snowflake Data Cloud can also help here by automating data workflows, so departments can provide their data more easily as products and integrate it directly with the tools available. Other important tools that should be part of a data mesh architecture are ingestion, as well as large-scale automation, machine learning, and related technologies. In summary, it can be said that a suitable platform for the data mesh architecture must have the following properties: it should provide scalable computing power, be usable from any location, be able to make all data in the organization accessible and also help to approach by setting up product pipelines and providing all the tools needed to use, process and control data, as well as to ensure centralized governance and data security.

As limitations, implementing data mesh in any organization requires knowledge of data architecture and data processing. If these skills are not available in organizations, they must be made available at an early stage and with foresight so that the development towards the data mesh concept does not fail due to a lack of know-how.

In future work, we want to show how data mesh can achieve higher data quality and data availability using data catalog technology (the basis of a data fabric structure) as a solution. Data mesh encourages the establishment of clear data governance frameworks that help ensure data is used responsibly and ethically. This includes defining roles and responsibilities for data management, setting standards for data quality and accuracy, and defining processes for data access and data usage. Only with data governance can data controllers help plan and understand problems in storing large amounts of data (Tallon, 2013).

REFERENCES

Antikainen, M., Uusitalo, T., & Kivikytö-Reponen, P. (2018). Digitalisation as an enabler of circular economy. Procedia Cirp, 73, 45-49. https://doi.org/10.1016/j.procir.2018.04.027

Araújo Machado, I., Costa, C., & Santos, M. Y. (2022, May). Advancing Data Architectures with Data Mesh Implementations. *In Intelligent Information Systems:*

CAiSE Forum 2022, Leuven, Belgium, June 6–10, 2022, Proceedings (pp. 10-18). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-07481-3_2

Ballard, C., Compert, C., Jesionowski, T., Milman, I., Plants, B., Rosen, B., & Smith, H. (2014). Information governance principles and practices for a big data landscape. IBM Redbooks.

Black, S., Davern, M., Maynard, S. B., & Nasser, H. (2023). Data governance and the secondary use of data: The board influence. *Information and Organization*, 33(2), 100447. https://doi.org/10.1016/j.infoandorg.2023.100447

Bode, J., Kühl, N., Kreuzberger, D., & Hirschl, S. (2023). Data Mesh: Motivational Factors, Challenges, and Best Practices. arXiv preprint arXiv:2302.01713.

Borjigin, C., & Zhang, C. (2021). Data Science: Trends, Perspectives, and Prospects. https://doi.org/10.21203/rs.3.rs-1014621/v2

Buer, S. V., Fragapane, G. I., & Strandhagen, J. O. (2018). The data-driven process improvement cycle: Using digitalization for continuous improvement. *IFAC-PapersOnLine*, 51(11), 1035-1040. https://doi.org/10.1016/j.ifacol.2018.08.471

Buranarach, M., Krataithong, P., Hinsheranan, S., Ruengittinun, S., & Supnithi, T. (2017, November). A scalable framework for creating open government data services from open government data catalog. *In Proceedings of the 9th International Conference on Management of Digital EcoSystems* (pp. 1-5). https://doi.org/10.1145/3167020.3167021

Butte, V. K., & Butte, S. (2022, October). Enterprise Data Strategy: A Decentralized Data Mesh Approach. *In 2022 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 62-66). IEEE.

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.). (2014). Online panel research: A data quality perspective. John Wiley & Sons.

Castro, A., Machado, J., Roggendorf, M., & Soller, H. (2020). How to build a data architecture to drive innovation—today and tomorrow. *McKinsey Technology*. Retrieved November, 21, 2021.

Dehghani, Z. (2022). Data Mesh - Delivering Data-Driven Value at Scale. O'Reilly Media, Inc.

Dončević, J., Fertalj, K., Brčić, M., & Kovač, M. (2022). Mask-Mediator-Wrapper architecture as a Data Mesh driver. arXiv preprint arXiv:2209.04661.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2), 293-314. https://doi.org/10.1093/nsr/nwt032

Hechler, E., Weihrauch, M., & Wu, Y. (2023). Data Fabric and Data Mesh Business Benefits. In Data Fabric and Data Mesh Approaches with AI: A Guide to AI-based Data Cataloging, *Governance, Integration, Orchestration, and Consumption* (pp. 71-85). Berkeley, CA: Apress.

Koltay, T. (2016). Data governance, data literacy and the management of data quality. *IFLA journal*, 42(4), 303-312. https://doi.org/10.1177/0340035216672238

Ladley, J. (2019). Data governance: How to design, deploy, and sustain an effective data governance program. Academic Press.

Liu, R., Isah, H., & Zulkernine, F. (2020). A big data lake for multilevel streaming analytics. In *2020 1st International Conference on Big Data Analytics and Practices (IBDAP)* (pp. 1-6). IEEE. https://doi.org/10.48550/arXiv.2009.12415

Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: concepts and principles of a paradigm shift in data architectures. *Procedia Computer Science,* 196, 263-271. https://doi.org/10.1016/j.procs.2021.12.013

Macías, A., Muñoz, D., Navarro, E., & González, P. (2022, November). Digital Twins-Based Data Fabric Architecture to Enhance Data Management in Intelligent Healthcare Ecosystems. *In Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2022)* (pp. 38-49). Cham: Springer International Publishing.

Malik, P. (2013). Governing big data: principles and practices. *IBM Journal of Research and Development,* 57(3/4), 1-1. https://doi.org/10.1147/JRD.2013.2241359

Marr, B. (2016). Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results. John Wiley & Sons.

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2020). The role of information governance in big data analytics driven innovation. *Information & Management*, 57(7), 103361. https://doi.org/10.1016/j.im.2020.103361

Pithadia, H., Fenoglio, E., Batrinca, B., Treleaven, P., Echim, R., Bubutanu, A., & Kerrigan, C. (2023). Data Assets: Tokenization and Valuation. Available at SSRN 4419590.

Podlesny, N. J., Kayem, A. V., & Meinel, C. (2022, July). Cok: A survey of privacy challenges in relation to data meshes. In Database and Expert Systems Applications: *33rd International Conference, DEXA 2022*, Vienna, Austria, August 22–24, 2022, Proceedings, Part I (pp. 85-102). Cham: Springer International Publishing.

Priebe, T., Neumaier, S., & Markus, S. (2022). Von Data Warehouse bis Data Mesh. *BI-SPEKTRUM*. https://doi.org/10.48550/arXiv.2212.03612

Shin, B. (2003). An exploratory investigation of system success factors in data warehousing. *Journal of the association for information systems*, 4(1), 6. https://doi.org/10.17705/1jais.00033

Shrivastava, S., Srivastav, N., Sheth, R., Karmarkar, R., & Arora, K. (2022). Solutions Architect's Handbook: Kick-start your career as a solutions architect by learning architecture design principles and strategies. Packt Publishing Ltd.

Silva, B. N., Diyan, M., & Han, K. (2019). Big data analytics. *In Deep learning: convergence to big data analytics* (pp. 13-30). Singapore: Springer.

Strengholt, P. (2020). Data Management at Scale. O'Reilly Media, Inc.

Tallon, P.P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Computer*, 46(6): 32-38. https://doi.org/10.1109/MC.2013.155.