Utilization of Data Mesh Framework as a Part of Organization's Data Management

Simo Hokkanen

Master's thesis

UNIVERSITY OF
EASTERN FINLAND

School of Computing

Computer Science

September 2021

ITÄ-SUOMEN YLIOPISTO, Luonnontieteiden ja metsätieteiden tiedekunta, Joensuu
Tietojenkäsittelytieteen laitos
Tietojenkäsittelytiede

Hokkanen Simo Santeri: Data mesh viitekehyksen hyödyntäminen osana yrityksen datan hallintaa
Pro gradu –tutkielma, 63 s., 1 liite (3 s.)
Pro gradu –tutkielman ohjaajat: FT Virpi Hotti ja Antti Loukiala
Syyskuu 2021

Digitaalinen teknologia, jatkuvasti kehittyvät tietojärjestelmät, sekä erilaiset ohjelmistokokonaisuudet tuottavat valtavia määriä dataa. Tuotamme, kulutamme ja hyödynnämme dataa yhä useammalla tavalla. Jokainen moderni organisaatio, joka tahtoo pysyä nopean teknologisen kehityksen mukana, hyödyntää ja kehittää liiketoimintaansa datan avulla. Data-alan monipuolistuvat palvelut ja kehityksen nopeus vaativat uusia toimintamalleja. Data-alan uusimmat innovaatiot ja trendit kiinnostavat myös yritysten lisäksi tutkijoita. Tämän tutkielman tavoitteena on selvittää data meshin viitekehyksen pääpiirteitä ja ominaisuuksia. Lisäksi tutkielmassa selvitetään, miten kohdealue (domain) määritellään ja sen määrittelemiseen liittyvät haasteet. Tutkimusmenetelmällisesti tehdään kirjallisuuskatsaus -ja kyselytutkimus. Kirjallisuuden avulla selvitettiin data meshin ominaispiirteitä ja sen soveltamisen haasteita. Kirjallisuuskatsauksen pohjalta vastattiin seuraaviin kysymyksiin seuraavasti: kuinka domain määritellään ja toimivatko CDM ja data mesh yhdessä; domainin määritteleminen on haastavaa, mutta sen tulisi olla yhdenmukaista. Yleinen tietomalli (CDM) ei tue data meshin periaatteita. Empiirissä tutkimuksessa testataan tutkielmaa varten luodun data mesh -työkalun toimivuutta teemahaastatteluiden avulla. Kyselytutkimuksessa selvitettiin data meshin viitekehyksen sopivuutta organisaatioihin ja pyrittiin löytämään erilaisia tiedonhallinnan toimintamalleja. Kyselytutkimuksen perusteella voidaan todeta, että organisaatiot omaavat jo nyt erilaisia hajautetun arkkitehtuurin piirteitä ja kaikki kohdeorganisaatiot kykenevät hyödyntämään data meshin ominaispiirteitä haluamallaan tavalla. Tutkimus osoittaa, että organisaatioissa, joissa data meshiä jo sovellettiin, datan hyödyntäminen oli suoraviivaistunut. Tutkielma osoittaa myös erilaisia haasteita yritysten tiedonhallinnan tilanteista ja tuo esille data meshiä estäviä tekijöitä, kuten monitulkinnainen domainin määritelmä, epäselvä datan omistajuus, vahvasti keskittyneet dataratkaisut, sekä datanlukutaidon matala taso.

Avainsanat: data mesh, tiedonhallinta, data-analytiikka, hajautus, teemahaastattelu, datatuote

ACM-luokat (ACM Computing Classification System, 1998 version): C.2.4, D.2.11, E.0, H.2.0 & K.6.3.

UNIVERSITY OF EASTERN FINLAND, Faculty of Science and Forestry, Joensuu
School of Computing
Computer Science

Hokkanen Simo Santeri: Utilization of Data Mesh Framework as a Part of Organization's Data Management
Master's Thesis, 63 p., 1 appendix (3 p.)
Supervisors of the master's Thesis: PhD Virpi Hotti and Antti Loukiala
September 2021

Digital technology, constantly evolving information systems, and various software packages produce vast amounts of data. We produce, consume, and utilize data in more and more ways. Every modern organization that wants to keep up with the rapid technological development is utilizing and developing its business with the help of data. The diversifying services of the data industry and the speed of development require new operating models. The latest innovations and trends in the data industry are of interest not only to companies but also to researchers. The aim of this thesis is to elucidate the main features and principles of the data mesh framework. In addition, the thesis explains how to define a domain and the challenges associated with defining it. The research method is a literature review and a survey. The characteristics of the data mesh and the challenges of its application were investigated with the help of literature. Based on the literature review, the following questions were answered as follows: how the domain is defined and whether common data model (CDM) and data mesh work together; Defining a domain is challenging, but it should be consistent. The common data model (CDM) does not support data mesh principles. In the empirical study, the functionality of the data mesh tool created for this thesis is tested through theme interviews. The study examined the suitability of the data mesh framework for organizations and sought to find different information management operating models. Based on the study, it can be stated that organizations already have different features of a distributed architecture, and all case organizations are able to utilize the principles of the data mesh in the way they want. The study shows that in organizations where data mesh was already applied, data utilization was more streamlined. The thesis also points out various challenges in enterprise information management situations, and highlights factors that prevent data mesh, such as an ambiguous domain definition, unclear data ownership, highly centralized data solutions, and low data literacy.

Keywords: data mesh, data management, data analytics, distribution, theme interview, data product

CR Categories (ACM Computing Classification System, 1998 version): C.2.4, D.2.11, E.0, H.2.0 & K.6.3.

Acknowledgment

This thesis has been carried out in a collaboration with Solita Ltd in summer 2021. I would like to express my greatest appreciation to my thesis supervisors: PhD Virpi Hotti and Antti Loukiala for helping and pushing me forward during my thesis journey.

List of abbreviations

ETL	Extract, Transform & Load
CDM	Microsoft Common Data Model
IT	Information Technology
API	Application Programming Interface
AI	Artificial Intelligence
ML	Machine Learning
ERP	Enterprise Resource Planning
CRM	Customer Relationship Management
SOA	Service Orientated Architecture
DDD	Domain Driven Design
RQ	Research Question
DM	Data Mesh

Table of Contents

1 Introduction

Digital technology is present in almost every consumer's or enterprise's daily actions. For example, people have more and more technological innovations carried with them: smartphones, smart watches, and laptops are a good instance of this. The rapid development of software and systems engineering requires more data to reach the ambitions technology aims to solve. Therefore, massive amounts of data gathered through different systems have led the data business to grow rapidly for the past few years.

We stand on the brink of a technological revolution that will fundamentally alter the way we work, live, and relate to others around us. After three industrial revolutions introducing us to steam power, electricity, and automated production, we are transforming to the next industrial revolution: The fourth industrial revolution. This fourth revolution is the digital revolution that has been occurring since the middle of the last century. Its main characteristics are the internet of things (IoT), autonomous robots, cloud computing, and overall, the digital transformation towards the world of information systems (Schwab, 2016).

For the past few years, big data has been one of the most exciting and refreshing stimulating trends in the data engineering world. Big data has made enterprises develop their data strategies and projects to more complex levels. Almost every business has data to benefit from, which is why data engineering and information management are becoming a larger part of a successful business. However, for some companies' data are still an unclear object they are trying to tackle. Therefore, efficient data architecture is required for enterprises to achieve the full potential from the data they own.

Artificial intelligence (AI) is all around us, and we are already experiencing drones and self-driving cars with digital assistants and software that helps us in our daily tasks. Impressive progress has been made in AI research in recent years. The rapid growth of computing power and availability of vast amounts of data contribute together to developing the digital environment we live in (Schwab, 2016).

The introduction chapter gives a brief background towards data mesh and, overall, the world of IT (Information Technology). The introduction section also includes research

questions and the theoretical work behind this thesis. Used literature and information retrieval are decoded and explained. The main research keywords are reviewed. The first section likewise introduces the basics of data management, data models and reflects why data is so important.

The thesis proceeds in the following order. The second section aims to give a strong theoretical background in the data mesh paradigm. The main principles in data mesh, distributed systems, and service orientated architecture are covered. The previous studies are also examined, and the downs and upsides of the *data mesh* (DM) architecture are explained. The foundation of data mesh lies deep in *domain-driven design* (DDD). Domain-driven design will be tackled in numerous parts of this thesis. One research question is also formed around the question, what is domain and how to define domains in your organization. The data industry is full of different and versatile terms. This thesis does not aim to explain all possible data terms to regard. The third section will include the core framework from data mesh in questions formed for theme interviews. Question battery is formed around the most quantum questions of data mesh and distributed architecture. Theme interview results will give a comprehensive insight, how professionals in this field comprehend data mesh. The fourth chapter will define when and how data mesh could be applied in the customer organizations. Showcasing these organizations will be included in the fourth chapter. The case study will be carried out through interviews with the selected organizations. The question battery and *Data Mesh Suitability Reporting* tool will be put to the test. Results and the most interesting answers will be highlighted in the fourth chapter. The last chapter of this thesis will discuss the results, dive into the most important findings and draw conclusions on the research questions. The last chapter will also lay eyes on the future of data mesh and go through possible follow-up research.

The theoretical background for this study is assembled with the latest and the most relevant studies and reports in the data mesh paradigm. Data mesh as a concept is moderately new, making the theoretical point of view fractionally narrow and challenging to execute effectively. The concept of data mesh also keeps changing and restructuring at present, and this causes the viewpoints to reform behind the concept. It also creates an illusion that something we write, or state just now could be abrogated in just moments or few written articles. The literature part includes official reports,

statistics, whitepapers, and current news. Solita Ltd also provided great amounts of literature and statistics from data management and data engineering projects. Different hype trends and topics in software or data engineering fields usually gather people to the same place for open discussion and learning. One vendor-independent community in Slack was formed at the very early stages of the data mesh hype. This "Data Mesh Learning" Slack group performed a crucial role in gathering new information, standpoints, and opinions about the new paradigm.

Attention has been paid to the quality of references by examining the reference amount of the publications used. Also, *Julkaisufoorumi.fi* -website has been used, which offers a level classification for academic publications. The thesis background material has been searched throughout Google Scholar search engines. ACM, Scopus, and IEEE digital libraries have also been used to find previous literature. The most frequent keywords were data mesh, data management, information architecture, data product, domain, and data ownership.

It proved to be exceedingly difficult to include the subjects entirely via published academic literature throughout the thesis. After all, there are not too much official academic literature published yet. Overall, the field of data engineering is young, and it is getting closer to the side of software engineering. Because the field of data engineering is so nascent, much of the conversation on current challenges and the state-of-the-art is had throughout what is commonly known as "grey literature".

Although, grey literature is quite common in software engineering-related fields including data engineering and computer science. Grey literature usually includes different sources (e.g., blog posts, videos, podcasts, and whitepapers). Multivocal literature review (MLR) recognizes the need for several different sources of opinions or voices to be heard. Instead of constructing the evidence from only the knowledge accurately published in formal and official academic settings, multivocal literature also uses all accessible writings or other publications around a popular, often a current topic (Garousi et al., 2016).

Figure 1 shows us the key dependencies in data science and gives a great look at the overall structure of data lifespan. Data mesh core pursues to make all this data

engineering and processing more structured and efficacious (Dehghani, 2020a). Making data mesh a good research topic to deepen.

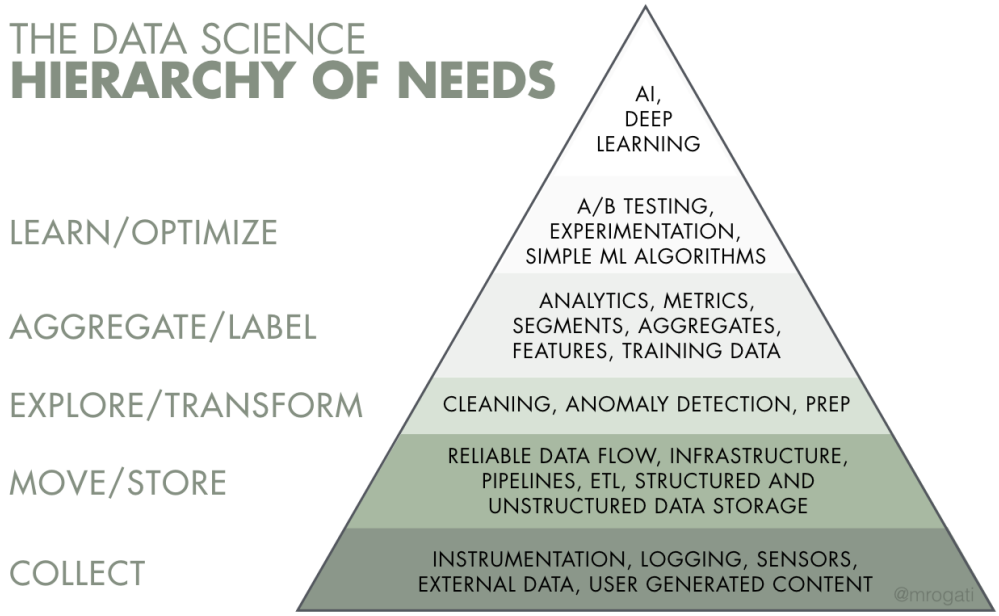


Figure 1: The data science hierarchy of needs, in the form of a pyramid, describes different data correlations. Credit Monica Rogati, Hackernoon.

Understanding the data we use is the key principle and aspect that must change in the data industry. Far too often ETL (Extract, Transform, Load) processes keep failing due to the constantly growing complexity of the labyrinth of data pipelines (Dehghani, 2020a). Data mesh is a high-level solution with decentralized and distributed responsibility of people nearest to the data to back up continuous transform and scalability (Dehghani, 2020a).

1.1 Research Questions

This thesis aims to answer the most ponderable questions about the data mesh paradigm. We are interested in studying how data mesh applies to different companies with different data management situations. Leaning towards this previous sentiment, the following research questions have been formed:

Research Question 1: What situations or organization data mesh can be applied into, and how to proceed to data mesh?

Research Question 2: How to define a domain in your organization? (What is domain)

Sub-Questions:

- Is data engineering more streamlined when using the data mesh procedure?
- What are the key challenges and benefits of data mesh?
- Does common data model (CDM) support data mesh framework?

The research questions consist of two main questions and three sub-questions. These questions provide an exploratory look into what requirements and challenges are found in the context of data mesh and distributed domain-driven design (DDD).

1.2 Data Management

Galetto (2016) defines data management as an administrative process that contains validating, acquiring, storing, securing, and processing the data to ensure reliability and accessibility of the data to its users. Bourque & Fairley (2014) state that one key concept in data management and database systems is data schema. Bourque & Fairley (2014) define schema as: "The relationships between the various entities that compose a database". Therefore, we can see schema as it is a description of the entire database structure and blueprints attached to the data.

Business questions acquire the data required to answer that question. Eventually, data needs to answer these business questions to generate the insights needed for data-driven decision making (Galetto, 2016). With the help of organisations' data management platforms, it is possible to gather, sort, and house their information and then re-package it in various ways to achieve the demanded analytics or insights. This way of information management will eventually create value for the company to grow their data business in the right direction.

A data pipeline is the structure and mapping of how the data is processed towards to use cases, such as building machine learning models. Data pipelines are known to impact machine learning performances and applications as much as algorithms. However, in practice, raw and unstructured data is infrequently prepared to be processed

and consumed and must be altered by a line of operations called a data pipeline (Quemy, 2019).

Modern technologies typically do not create bottlenecks for data management to succeed, as these new technologies usually can scale horizontally and vertically. One of the key aspects of good data management is to optimize functions and data processing in booming large software projects. Continuously scaling large software development is not a new problem to solve. However, the basic foundation of clear and effective data management is necessary to scale modern technologies and customers' needs.

1.3 Data Models

A data model can be seen as a high-end abstract premise that organizes data features and elements. These features include data entities and attributes. Model defines the data elements and the relationships between the elements and attributes. The goal of data models is to show how data is stored, connected, updated, and accessed. *Figure 2* highlights a simple example of a data model between customer and address.

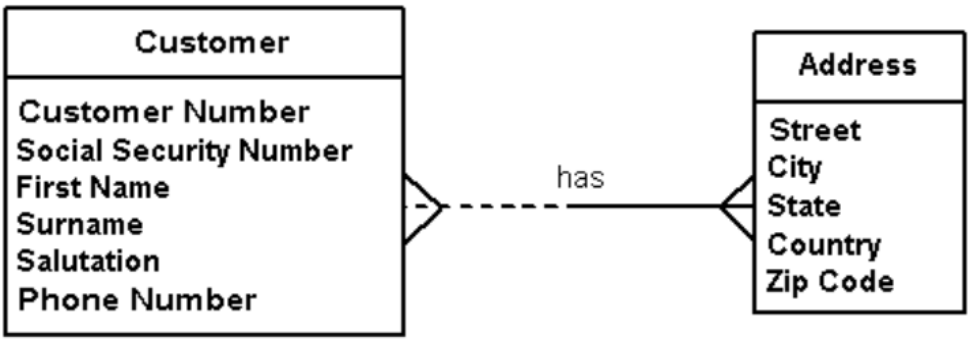


Figure 2: A simple logical example of the data model. Credit Scott W. Ambler, 2006.

The data model is built from the viewpoint of the raw information used in the specific concept. Data tables and relationships between the data define data models, such as entities and attributes (Bourque & Fairley, 2014).

Programs and applications work on data. Data must be organized and defined within computers, and after these, systems and applications can process applications or programs. Modern software development and data development practices can automate

multiple data modelling steps, producing the data to be available faster for consumption, causing a need for efficient data management. Development of data practices are a continuous task for organizations.

An algorithm is a set of precisive instructions for computers on how to complete a certain task. Algorithms are used to execute complicated programs and applications more cogently (Bourque & Fairley, 2014). Most AI and ML (Machine Learning) models require high level and swift algorithms to execute the tasks in a preferred way.

1.3.1 Microsoft Common Data Model

Adobe, Microsoft, and SAP published an "Open Data" -suggestion in Microsoft 2018 Ignite Event. The result of this suggestion, the Common Data Model (CDM), aims to model the common concepts in business into one homogeneous data model. Applications and systems could use this model as such or with small dilations (Hansen, 2020).

CDM defines a group of commonly used business objects (entities), attributes, and relations between objects. Typical entities for this data model are, for example account, Contact, Activity, Owner, Task, Product, and Order. The complete model includes roughly 700 entities, with about 100 fields per entity (Microsoft - CDM, 2020).

The florid published from these three companies reveals the main practical ambitions as follows, *a)* Getting rid of data silos, and *b)* Creating one unite data model, which illustrates the basic business concepts and relations with each other (Hansen, 2020).

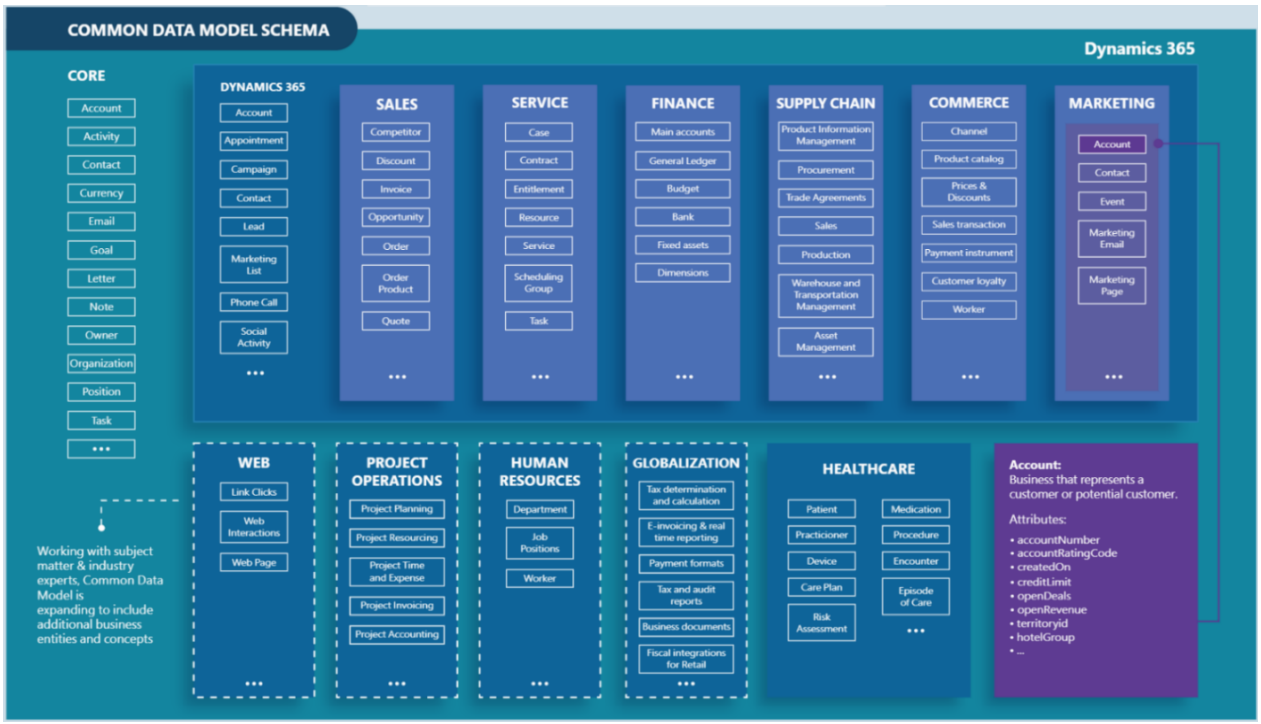


Figure 3: Microsoft CDM featured as a complete-scale example. Credit https://docs.microsoft.com/en-us/common-data-model/

CDM does not editorialise the actual storing method of the data. Information can be saved in any format or structure needed. The model just defines the structure of the data (schema), where data must be saved to guarantee compatibility. For example, if the data is stored in a traditional SQL (Structured Query Language) relational database, CDM would define the structure of the database, like the names of the tables, columns, and the foreign-key references between the tables (Hansen, 2020).

CDM strives to ease the issues caused by data centralization. A practical example of this silo-effect is a company with three massive operative systems: marketing, sales, and customer service applications. Every application has a data structure for the *Customer,* which is almost equal to the one in sister application. These systems have been built by different administrators in different periods. If these applications used the

same Common Data Model, they would all be having the same understanding of the *Customer,* what the *Customer* is and what data specifies with it. If these applications are correctly built, one storage structure, one Customer-view, one interface, and one technological tool would be enough to use all this data (Hansen, 2020).

CDM should offer a solution for having up-to-date data always in use. The second benefit is towards the integration and conversions between multiple systems. These conversions would be more efficient, and the general view is easier to understand with one unite Common Data Model (Microsoft - CDM, 2020). Overall, CDM and similar concepts of data management are a little old fashioned and are the traditional understanding. CDM shows us the classical and orthodox view of how data models, attributes and entities should be treated. These classic reckonings and the new pivotal data mesh differ in many ways. With this literature about CDM we can say that data mesh applies different patterns and logics for data than CDM. Data mesh heavily relies on domain-driven design: therefore, a united customer definition is not the solution. Domain-driven design and data mesh use domain specified customer definition that can be mapped together in the future, if even needed.

1.3.2 Conformed Dimensions

A conformed dimension is a dimension that has the same meaning to every fact that it relates to data warehousing. Therefore, using a conformed dimension can make the whole ETL process more efficient as it does not have to do various tasks to process the same dimension-related information more than once. (Serra, 2011).

McHugh (2017) defines that conformed dimensions are those dimensions that have been blueprinted so that the dimension can be used across many tables in distinct subject areas of the data warehouse or lake. Conformed dimensions can provide the customer with insights into their data that exceeds the initial needs and expectations. Eventually, this is exactly the main point behind strong and effective information management.

1.4 The Importance of Data

Data is increasingly seen as great wealth, such as oil or gold. However, data is more than just bits or information gathered in a specific form or structure. Data is information units, usually expressed in numeric form, and it is collected through observation. In modern business data can be gathered throughout applications and systems customers consume. Data itself always does not have high value in it, and the key thing is how to process, inspect and analyse the data.

Companies built specifically on data have been around for a long time. For example, gathering customer information and using it to make better decisions, products and services is an age-old strategy, but the complete process used to be slow and difficult to scale up (Hagiu & Wright, 2020). This low-speed process changed dramatically with the advent of cloud technologies and new IT innovations that allow companies to rapidly process and make sense of vast amounts of data on their hands (Hagiu & Wright, 2020).

Data is becoming the main driving force for digitalization, and it has been lifted in many companies as a pivotal key asset. Data is being used on an ever-increasing scale in applications, system development and decision making. Therefore, a constant demand for more data is insisted at better quality from a wider time horizon.

Then, where the data business gets interesting. Even when the data is proprietary or unique and it produces valuable insights, it is difficult to build a durable scaling advantage if the competitors can follow the resulting upgrade even without similar data. Another interesting factor is how fast the insights from customer data change. More repeatedly, they do so harder for others to imitate (Hagiu & Wright, 2020). These small but very notable factors make the data business complex and competitive. Changes in data business complexity and competitiveness results in enterprises investing in data consultants and technological experts. Data is on its way to being the main driving force in any business.

Though data business and industry is growing rapidly, it also has different interesting variations it is going through. In recent years, the vast amount of raw data produced has dramatically increased. In contrast, the use of such raw data for creating new value

and insights by organizations has been limited (Rodriguez et al., 2020). Organizations seeking for new value from data is a keen topic we will dive in to during this thesis.

Data is the most important tool for companies' administration which seeks to execute sustainable and cost-efficient business. Experienced data teams, state-of-the-art analytics, and technology solutions are an essential part of enterprises data pipelines. Previously mentioned factors are not always enough when the available data is desired to be used in the most productive way (Etlia, 2021). The key parts of the process must be sharpened together, and data must be seen in a new light.

Data projects need to change to be more cost-efficient and effective. Although companies use major parts of resources in data tools and projects, technology usually is not the issue because it bends on how users need it. Data mesh can be a solution for more effective data architecture and management to score completed data projects.

2 DATA MESH FRAMEWORK

Transforming to a successful data-driven enterprise remains as one of the key strategic goals for modern companies. These companies are valuing evermore static and efficient data architecture from organizations offering data consultation for them. Data mesh is a new and decision orientated paradigm with architectural data features. For the first time, data mesh was introduced by Thoughtworks technology consultant, Zhamak Dehghani on a highly appreciated blogsite by Martin Fowler (Dehghani, 2019). After this first impression of the data mesh paradigm, the concept has had a lot of enthusiasm around it. It has become the most relevant new data engineering topic in late 2020 and early 2021. Dehghani (2019) describes data mesh to be a paradigm shift in managing analytical data. Complexity and the size of software can scale rapidly out of hands. More complicated data requirements need clear design for enterprises to scale with the needs of the data-hungry customers. Domain-driven design-based data mesh is a new solution for this evolution.

More and more information systems and software require precise and comprehensive design for data utilization. Designing and planning out just the software itself is not enough anymore. The goal for data architecture is to design organizations information on different levels. Also, the overall picture from the data on centralized silos and on an explicit level is an objective to reach. The objective of data architecture is to show organizations crucial data content beyond organization and system borders.

Organizations must tackle multi-faceted complexity and challenges in the transformation to become more data-driven. Competing business priorities, migrating different legacy systems, and the culture relying on data are factors behind the data-driven movement. Data-driven means the transformation and the keen creating more value with data, placing importance of data in a core business position. Dehghani (2019) now suggests a new way to build a distributed data architecture at scale and focus on the importance of data domains. She introduces us to a new enterprise data architecture that aims to solve the current issues with centralized data architecture. Data mesh offers a new viewpoint to tackle the challenges in monolithic architecture (Dehghani, 2019). Commonly, it is seen that operative business software, such as ERP- systems,

are the main product. Data mesh wants to focus on the data itself and remove the habit of seeing data just as a "side product" (Hovi, 2021). Seeing data as a side product is still a common vision in any business. Data is just an obligatory case within acquired IT systems, a secondary business development requirement. This shift is the key fundament in data mesh.

Data mesh is presented as a framework (Dehghani, 2019). Regarding to ISO (2006) a framework is a specific structure expressed in text, diagrams, or formal rules. These relate to the components of an abstract entity to each other. Framework is important to define here, because data mesh is aimed to create new impact as the future framework for data management.

Data platforms are environments or applications that import data together and serve it across different business units. A data warehouse is data storage used, for example, around reporting and analytics. It is the key central repository of data integrated from different information sources. It remodels data into a common schema that enables easy data usage. Using this methodology can lead to more effective analytics and other valuable data products.

A data lake is also a central repository of already structured but unstructured data required in any format. The key benefits of a data lake are that it can speed up data availability because data is usually stored in raw format. Furthermore, the data schema is also defined at the usage time, allowing flexibility towards the data.

Data warehouses and lakes being the structured and meticulously conceived data management, data swamp is the opposite. According to Knight (2018), a data swamp has no clear organization form or system built around the enterprises' data. Data swamps have narrow curation, including little to no active management throughout the life cycle of data. Also, the metadata and data governance are usually poorly executed. We need to remember that someone's data lake can be another's data swamp, and this is because of the variety of data and businesses we face. Data swamps usually have the issue of being unusable and frustrating for data consultants and engineers (Knight, 2018). These examples of different data platforms show us that data requires effective management and an ability to create extensive insight directly to the organization on hands.

In data mesh, the data infrastructure is technically centralized, the same for everyone in the organization, but data pipelines are built in a distributed domain-driven fashion. Following this principle, every single of one data pipeline can be optimized for the specific needs in that business domain, for example, marketing or customer service (Hovi, 2021). Having unique and optimized data pipelines for domains does not mean that every business domain has or needs to build its own data lakes or warehouses, the side of which stays monolithic. Instead, these domains will be in charge and have full ownership of the data they consume.

Following the previous example operation model, the same specialist and employees are accountable for the complete data pipeline in its full form, all the way from the production of the data until its final usage. Using this model, business domain employees witness and understand the data they use at a completely new level (Hovi, 2021). When data pipelines are unique and understood by the same people that consume that data at the business domain side, massive value can be created continuously.

Zhamak Dehghani (2020a) ponders what we mean by data; she explains that we can divide the data landscape into analytical and operational data. Operational data lays in databases that support business capabilities through microservices and APIs. Operational data, for example, serves the needs of applications running the day-to-day business and has a transactional nature. Operational data typically comes from transactions between organizations and their customers (Dehghani, 2020a).

Analytical data is usually temporal and supports views of the business situations over time. Analytical data is traditionally modelled somehow, and future-perspective insights can be built with various available technological tools. For example, engineers can train machine learning models and create plots to support functions all around the business (Dehghani, 2020a). *Figure 4* shows the operational and analytical data cooperation, with ETL processes as the contactor for an effective data pipeline.

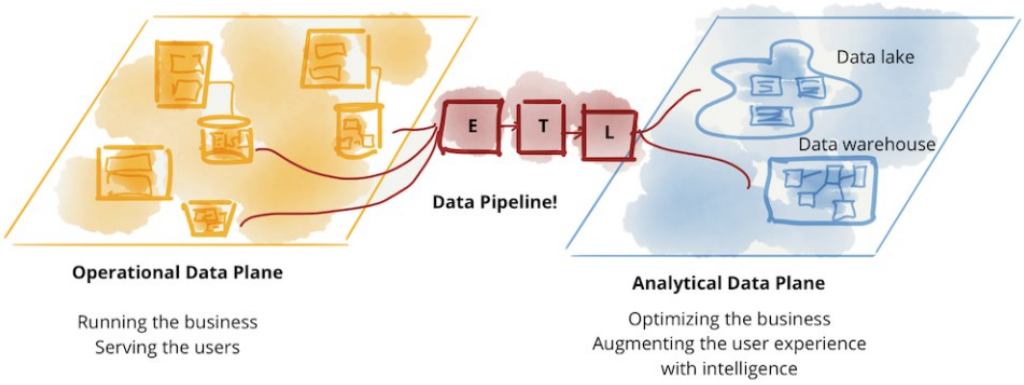


Figure 4. ETL Data pipeline. Credit Zhamak Dehghani, 2020. https://martin-fowler.com/articles/data-mesh-principles.html

In her presentation, Dehghani (2020b) explains that operational data creates API-based access to data, captures the current state of applications running, and serves later parts of ETL pipelines with graph and relational databases. The operational data plane creates platform for the analytical side of data utilization to scale towards different insights and future visions.

To summary the core message from Data Mesh together, we can highlight four main principles (Dehghani, 2020a):

1. *Domain-oriented decentralized data ownership and architecture*

2. *Data as a Product*

3. *Self-serve data infrastructure as a platform*

4. *Federated computational governance & Data Governance*

These four main principles are explained and tackled in various sections during this thesis. The domain-orientated decentralized data architecture principle and the third one, self-serving data infrastructure as a platform will be examined along with each other with distribution aspects. Data as a Product is an extremely interesting mindset, and it is undergone in section *2.4 Data as a Product.* The fourth principle, the data

governance part is processed in *2.3 Non-Invasive Data Governance.* This section is an apropos place to determine how data mesh observes data governance and how non-invasive actions help company data management. The meaning of domain and idea behind Domain-Driven Design will be observed first. Dehghani (2020a) states that Data Mesh has taken a lot of influence from DDD, so it makes the theories from Eric Evans natural to inspect first.

2.1 Domain-Driven Design

Before facing the current situation, pitfalls, and challenges in centralized monolithic architecture, we must focus on defining the meaning of the domain and core principles around domain-driven design. What does the data domain mean and do organizations vary with the views on the domain? This question will also be tackled on the later parts of this thesis, during the interviews performed on the selected customer organizations. Who could potentially apply data mesh -thinking into their data architecture?

In the context of data mesh, the domain does not mean a certain group of computers that can be accessed and administered with a common set of processes. Here, the domain does not touch the concept of domain names, network domains or Internet Protocol (IP) resources. However, in deeper levels of data management and distributed domain architectures, the domain concept has rather different and multidimensional definitions. These different definitions and viewpoints will be covered throughout the following section.

Domain-driven design is an approach for the software development industry that focuses on programming a clear domain model. This model has a rich understanding of the processes and rules of a domain. Domain-driven design name roots itself from an extensively honored book: Domain-Driven Design – Tackling Complexity in the Heart of Software by Eric Evans. Evans (2004) describes the approach through a catalogue of patterns. This approach is particularly suited for complex domains, where a lot of often-messy logic needs to be organized properly.

The concept of software systems based on a carefully developed domain model has been around since the software industry emerged (Fowler, 2020). Moreover,

throughout the 1980s and 1990s, representing the underlying domain was a fundamental part of much object-orientated and database development (Fowler, 2020). Overall, Evans tied up a complete and overwhelming contribution in developing a common vocabulary and identifying main conceptual elements beyond the diverse modeling notations that dominated the domain discussion.

Large-scaled agile software development neglects the proper architecting support in such development projects (Uludağ et al., 2018). Domain-driven design addresses solutions for an increasing number of large organizations developing evermore complex software systems while adopting agile and lean methods during the software development processes. Moreover, we can inspect the domain-driven design bottom theory: the business domain should match the language and structure used in software code (Evans, 2004). Definitive examples of software code structure repeatedly used are class names, class variables and methods. Domain-driven design is overall a very broad and heavy concept, signifying that it includes terms and abstracts. *Bounded Context,* for example, is a central pattern to make strategic decisions in DDD, where large domains and teams are on a linchpin.

2.1.1 Domain

Evans (2004) explains that every software program follows up to some activity or interest of its user to apply the product. That area of subject the user applies in the program is the domain of the software. For example, airline-booking program involves a domain of real people getting on a real aircraft. On the other hand, some domains are immaterial: An accounting programs domain is money and finance, IT system domains often have little to do with computers. Of course, there are few exceptions. For example, a source-code control system, the domain is software development itself (Evans, 2004).

Ability to solve domain-related tasks for its users is the core of the software. Software and systems have multiple functions, usually even vital when looking at the bigger picture of the software. In the end, these side features support the basic purpose of the application (Evans, 2004). When the domain is complex to identify, developing software is a difficult commission, and here the most sharpened effort of talented people is required during the software development. Programmers need to dive into the

domain itself to understand the business the software is implemented into. Developers must sharpen their modelling skills and master domain-driven design (Evans, 2004).

Although, that, for example, is not one of the key priorities on most IT projects. Commonly, developers do not find interest in understanding the certain domain in which they operate, much less making a significant effort to understand domain-modelling. Most highly technical software developers enjoy solving quantifiable problems that train technical skills and understanding (Evans, 2004). Computer scientist's capabilities and common interests do not seem to find messy domain work interesting. These preferences could come from the education or teaching of software development and programming. Developers see that their task is not on the domain side, but rather on the pure programming side. Talented developers can also have multiple projects running simultaneously, decreasing the time and interest to find a specific domain defining issue. Thus, there is a clear gap between the development and business units.

Also, Evans (2004) condenses that domain is: "a sphere of knowledge, influence, or activity". It clearly shows that a domain can be difficult to define, and the lack of definition for domains could cause obstacles in the software development process. Nowadays, Evans's domain definition is seen in a variety of ways. In this study, interviewed organizations can express how the domain is seen in their point of view. We will also aim to determine if some organizations do not have a clear meaning or definition for the domain. We might quickly find if the data mesh framework could or could not be applied to their data architecture at all.

Vaughn (2013) supports the domain definition from Eric Evans (2004) and discovers that domain to be one of the most important aspects of efficient software development and data processing. To design high-quality software products that meet core business objectives, tactical and strategic modelling tools are required to clear vision of the domains.

We need to dig deeper into the core difference between a business domain and a data domain. The different definitions between business or data domains are the most crucial parts to figure out in any modern data-utilizing enterprise.

2.1.2 Context Mapping

As previously explained, Bounded Context has different tools built around it. For example, Vaughn (2013) describes different ways annex several bounded contexts, context maps being the most explicit.

Context mapping is simply a tool that enables to recognize the relationship between bounded contexts and the relationship between the business units or teams being obligated for them. Vaughn (2013) particularizes that context maps are not a technique to be limited by drawing a diagram of a specific system architecture in use. Moreover, it is about understanding the relationships between the different Bounded Contexts in a business and then the patterns used to draw objects purely from model to another.

Overall, context mapping takes bounded context further in a notion of strategic design, and how to organize large domains. The context mapping principle on organizing large domains is one of the first instances to point out that domain-driven design. At the same time data mesh might just be intended to be used in large organizations.

Alternative ways for context modelling and mapping have also arrived in the IT industry. Event storming is a good example of this, and it is a workshop-based method for highlighting what is going on in the heart of the software program, the domain. If you compare it to other various methods out there, event storming is supremely lightweight and, on purpose, does not require support from computers. Results or an example process are attached to a wall with sticky notes. Event storming roots itself to show the focus on domain events, and the methods have similar aspects as brainstorming (Brandolini, 2013).

2.1.3 Ubiquitous language

Fowler (2020) imparts the topic of ubiquitous language as a key part of domain-driven design. This language is a major part of effective software development specialising in programming and application development around domain models. Ubiquitous language is originally a term expressed by Eric Evans (2004). It aims to create a pattern for the common and understandable language used in all business units across the organization. Software developers, business specialists and administrative users need a

unified language to understand the developed products in the same way. Evans (2004) states that domain terminology must be embedded straight into the software systems. The importance of domain terminology is one of the core roots in Domain-Driven Design. *Figure 5* below shows how Ubiquitous language could tie together various parties of an organization. Although Evans originally formed DDD and Ubiquitous language for software development, it is seen that it fits well into the concept of data development and the data industry.

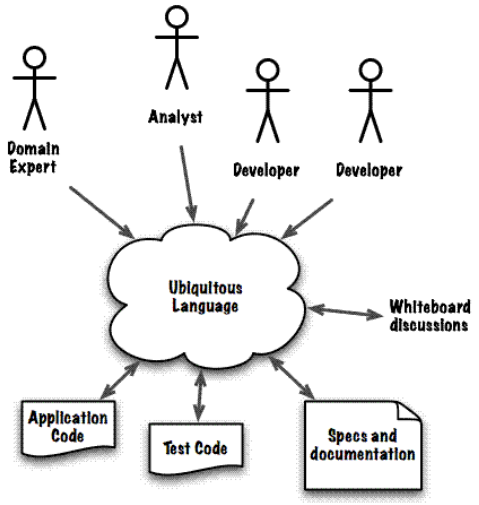


Figure 5: *The Ubiquitous Language*. Credit InfoQ, 2009.

Ubiquitous language offers multiple different participants and together they can form a united and efficient way of working. In addition, ubiquitous language touches the organizational culture, everyday conversations, and technical factors like code and documentation.

2.2 Service-Oriented Architecture

Architecture is one of the most intriguing and, at the same time, common terms around software development and data management. This part seeks to find answers on what architecture means and how the service-oriented architecture touches the data mesh concept. But, first, we need to find a proper answer for the question: What does architecture mean in the context of IT and data?

IT architecture can be seen as a set of structures needed to reason the system, which also comprises software elements, relations among them, and data properties. However, software architecture is not the same as data architecture, and they should be seen as separate architecture domains (Zhu, 2013). These architectures want to tackle slightly different concerns from their aspect.

Service-oriented architecture (SOA) instead means a logical way of blueprinting a software system to provide services either to applications that end-users consume or other distributed services in a network (Papazoglou et al., 2007). The distribution side articulately plays a vital role in a service-oriented architecture. New software applications and data systems are often seen as a service to end-users, which needs a strong architectural concept to lay on.

2.2.1 Microservices

The microservice architectural style is a vision to develop a singular application as a suite of many small services. Each service running its process and communicating with lite mechanisms, often seen as an API (Lewis & Fowler, 2014). Microservices has been commonly seen as the go-to method in modern software development.

Lewis & Fowler (2014) explain that the strength of microservices can be seen through a simple example, comparing it to a monolithic style single unit. Of course, software development varies clearly from data applications. However, the same mindset can still be set in both development environments.

Data mesh essentially refers to the concept of breaking down data siloes into smaller, more decentralized portions. Much like the shift from monolithic applications toward microservices architectures in the world of software development, data mesh can be described as a data-centric version of microservices (Furia, 2021). Data applications and solutions usually come an inch behind software business, which leads the direction of technology development and industry.

Data has undoubtedly massive potential in any modern business. Vast amounts of data push this market naturally towards smaller portions, easier to manage the complete picture of many microservices.

2.2.2 DataOps Culture

DataOps culture follows from a common term, Software Development (Dev) and IT Operations (Ops), known as DevOps. DevOps aims to clear the development cycle, boost continuous integrations, and ensure high software quality. DevOps includes aspects from Agile methodology. DataOps is a group of technical practices, cultural norms, workflows, and architectural patterns (DataKitchen, 2021). Shortly said, DataOps seeks to pursue more effective tools for data analytics and communication.

Overall goal of the data mesh is not to vaporize the benefits and utilization of data lakes and warehouses. Instead, the goal is to enhance productivity and to develop the teams consuming data. A clear object on the horizon is that technical experts, data production, and business units work together more efficiently. These same principles touch the overall work culture that DataOps wants to propel.

Rodriguez et al. (2020) also point out that DataOps is just one of the many tools or frameworks that emerge to attend the demanding requirements of a data-driven process that covers all the points from data collection to analysis and decision making.

2.2.3 Distributed Systems

As a simple definition, a distributed system is a group of computers working together meanwhile appears as one computer unit to the end-user. The core aspects of the data mesh framework rely on distribution and decentralization. Monolithic systems need to be replaced with microservices to be able to apply the Data Mesh thinking. A distributed system is a complete set of computers, networks, and processes, connected by a network, to work united to collectively execute a specific group of services (Neuman, 1994). This definition of a distributed system fits the distribution aspect of data architecture that data mesh strives to fulfil. Data mesh creates distribution in data ownership and data processing. Also, the mindset of distributed data architecture is a vital part of the mesh. Distribution sets a new and refreshing phase in the data world.

Distribution is a strong tool to enhance computing ability in the business globalisation around us. Distributed systems are used among various cloud databases and data systems. Distributed data systems provide distribution for data storage, infrastructure, and

cloud computing. These distributed data systems help companies with the continually growing need to model and analyse massive amounts of data.

2.3 Non-Invasive Data Governance

Data mesh follows distributed system architecture patterns with independent data products, self-serving data platform infrastructure and various deploying teams working with data. This data can include vital information from key processes, transactions, or customer engagements. These principles create a demanding requirement to implement a strong governance model for data (Dehghani, 2020a).

Dehghani (2020a) states that data mesh has different priorities regarding data governance models than traditional governance of analytical information management systems. In contrast, federated computational governance in data mesh contains the understanding of change management and multiple interpretative connections (Dehghani, 2020a).

Different governance models, laws, and standards challenge the data industry to have more transparent data usage and ownership. Data Governance has multiple different definitions, and it is widely seen in various ways. Commonly, data governance is seen as the process of managing the availability, integrity, security, and usability of data in enterprise systems used (Stedman & Vaughan, 2020). Data governance is based on internal data policies and standards that also control company data usage. Data governance is seen as increasingly critical for organizations that face new data privacy regulations. These companies usually rely evermore on data analytics and knowledge management to improve operative systems and decision-making (Stedman & Vaughan, 2020).

Non-invasive data governance means a set of practices of applying formal behaviour and accountability to secure effective use, security, compliance, and quality of data (Seiner, 2016). Non-invasive mindset helps companies to get a better grip on their data. Data mesh aims to support coequal data regulations and computational governance with clear instructions for business domain professionals, data experts and stewards.

A broader understanding of data politics and governance help organizations to scale with their data product development and analysis.

Different metadata management and standardizations fix the common problem in organizations, which is that data are often too difficult to find – almost like locked away in a system somewhere businesses cannot access it (Shahrin, 2021).

Data governance capabilities ensure the state of data. However, we should always remember that even after various high-quality checks, there will usually always be still a node that is contaminated. The danger of contaminated data is not a special case to remember, more like general knowledge to keep in mind while working in the world of data (Shahrin, 2021).

2.3.1 Data Ownership

Domain data ownership finds its place in data mesh core, where overall decentralization and responsibility distribution are key aspects of people nearest to the data they consume. Moreover, responsibility distribution is included to support scalability and continuous change of data business (Dehghani, 2020a).

Ownership of the specific business domain in the DM model means the ownership of data as well. However, data ownership creates different responsibilities and introductions to follow. Domain data owners must understand who is consuming that data, how it is being used, and what the common native methods users see comfortable carrying out (Dehghani, 2020a). Understanding these aspects creates a foundation for ethical working methods but also strives data as a product -thinking onwards.

2.3.2 Reshaping Data Teams

Data mesh strongly strives towards modern and agile data teams. These teams would be decentralized across the organization's domains, and they would serve the needs of business domain professionals. Dehghani (2020a) describes that when reshaping data teams and focusing on domain data, we need to accelerate the movement towards new data roles.

Data mesh implementation should support for domain data to be considered as a product. These changes on data teams also create a need for new data roles that organizations should introduce, such as data product developers and domain data product owners. Data product developers and domain data product owners are responsible for operative organizations which want to ensure that data is delivered as a product (Dehghani, 2020a). These new data roles divide future organizations between data as a product (DaaP) and data as a service (DaaS) operating model. For example, data product developers can be similar to data engineers but desire better data products. *Figure 6 (Adapted, Dehghani, 2020a)* explains new domain data nodes as cubes and different notations and authors affecting the bigger picture.

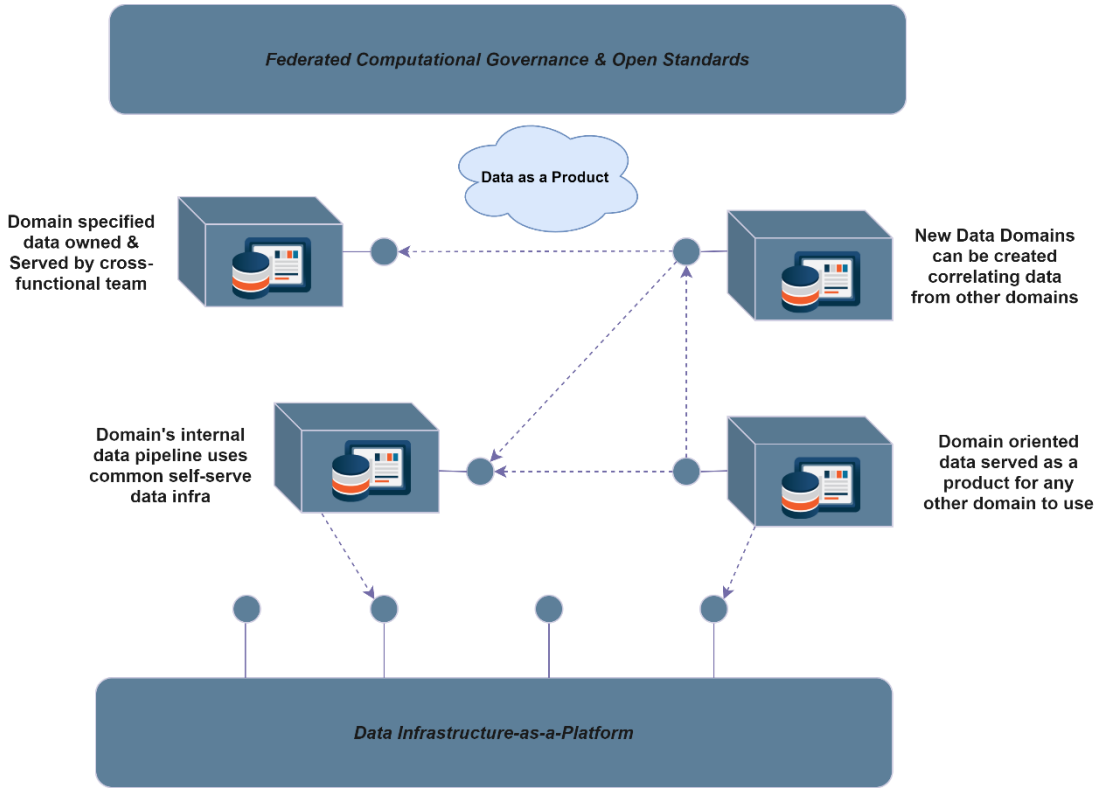


Figure 6. *Data Mesh as a Software Architecture*. Adapted, Zhamak Dehghani, 2020. https://martinfowler.com/articles/data-mesh-principles.html

Figure 6 shows us the architectural point of view behind data mesh. Software and information systems architecture is typically represented in such illustration. This method also suits data engineering and data management well. *Figure 4* presents all main principles in data mesh: Decentralization, A self-serving data platform, Data as a Product -thinking, and Federated Computational Governance parts. Cubes present

the domains that can self-serve data from the monolithic infrastructure (Data warehouse, data lake, etc.). New domains can also be created and work as cross-functional teams, owning their specific data. Data should be seen as a product and the complete organization can work together towards having more efficient and operant data products.

2.4 Data as a Product

One of the core principles of the data mesh framework is the data as a product -mindset. The view of how data should be treated is one of the most criticized parts of data mesh. Data as a product thinking leads to seeing data as an asset, even a possible product. Concept of data as a product is criticized because product thinking most definitely doesn't fit all businesses and data use cases. Data is a highly versatile commodity, and it varies largely between organizations.

Dehghani (2020b) points out that data needs to be easily discoverable, and a common implementation is to have a certain registry, data catalogue, for example. This registry shows all available data products with their meta-information, such as a source of origin, lineage, owners, and sample datasets.

Data mesh highly focuses on the efficient use of analytical data. Analytical data provided by the domains must be treated as a product. Consumers of that data should be treated as customers at the same time (Dehghani, 2020a). We also have to think about the fundamentals of a product and features used to create the product. Hovi (2021) state that good product has a clear concept, is produced through a certain production, has a standard value, and has an end-user or a customer.

Hovi (2021) defines data products as information formed from company data that value the customers with a standard product, building a data product has to start with the customer's requirements. Data products should always help a person, streamline operations, or do something that has value (Hovi, 2021).

Data as a product allows organizations to defeat the existing challenges of analytical data architectures. These challenges touch the high cost and friction of discovering, trusting, understanding, and ultimately using quality data (Dehghani, 2020a).

Data as a product mindset could possibly only fit technological organizations because most classic businesses have a physical product or operation to bring forth. This perception also came to prominence during our theme interviews, and it supports the statement from Hovi (2021) that agile technological organizations will be the first to implement this concept.

3 DATA MESH QUESTION BATTERY FOR HYPE LANDING

In this chapter, we focus on examining the purpose and objectives of the study on a deeper level. In addition, we will also lay our eyes on qualitative research, which is utilized as a vital part of the data collection of this study. The chapter also describes in detail the execution of theme interviews and assesses the reliability of the study. This chapter aims to provide the reader with a clear idea of the steps after which the implementation of the research and results have emerged.

Interviews are a typical way in qualitative research to collect data. This study uses theme interviews as a tool to find answers for the designed question battery. Theme interviews also set an advisable observation to the situation of case study organizations.

The goal of this case study is to find answers to our research questions from the Introduction chapter. Research question 1: "What situations or organization data mesh can be applied into, and how to proceed to data mesh?*"* & 2: "How to define a domain in your organization?". These research questions help us understand where data mesh could fit and the main principles to consider when moving towards distributed architecture.

The basic concept of new research is to create something new for science and people consuming the study. This study aims to create value through the new framework of data mesh. In addition, the study seeks to advance the know-how around data management, information architecture, and distributed data mesh paradigm. Case study supports and gives practical proof.

Because data mesh is a very young framework, we need to focus on creating the first steps towards understanding the concept from an academic perspective. This study fills the void of data mesh studies and creates a path for other studies to continue afterwards. We aim to find few starting points to review and lift them to the podium for further research. Our interview questions are justified based on data mesh principles

by Zhamak Dehghani (2020a) and the general assumptions on how distributed architecture would fit organizations.

Before interviews, we designed a question battery with a smaller set of questions or statements. However, soon we noticed that we need to dig deeper into the core of organizations and data management issues to find answers if the data mesh framework would fit. So, we structured the questions into six dimensions and this way, we found a way to prove why these exact questions should be asked. Similar questions and debates also came up on different social media platforms, such as Twitter, LinkedIn, and the Data Mesh Learning Slack group.

3.1 Data Mesh Suitability Report

Data mesh is a very new term, and massive hype around it also challenges the suggested architectural framework. *Data Mesh Suitability Reporting* tool was built to organize and justify the questions asked during the theme interviews.

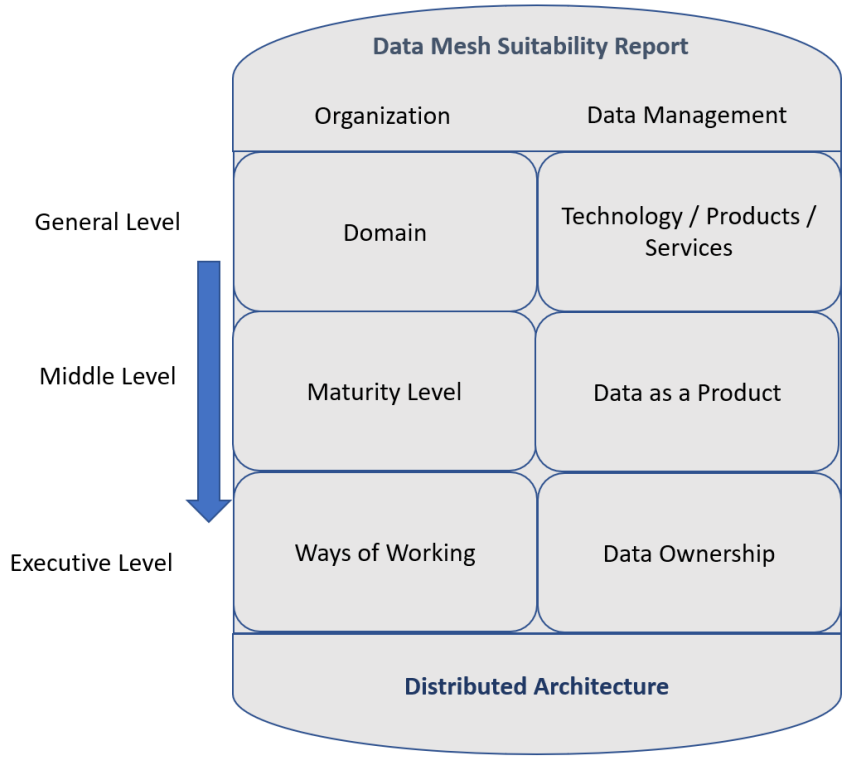


Figure 7: *Data Mesh Suitability Reporting Tool.*

Figure 7 shows us a suitable tool to find how well data mesh and distributed architecture fit into an organization's situation. This reporting tool includes six dimensions (Domain, Maturity Level, Ways of Working, Technology/Products/Services, Data as a Product, & Data Ownership). These dimensions all included 2-5 questions each to form the total 29 questions of the set. *Figure 7* shows three levels (General, Middle and Executive) on the left side. These levels include two dimensions each. Every interview started with the general questions, ending up with more specific executive questions about ways of working and data ownership.

After going through all the preprepared questions we could find some references if this organization could adapt data mesh. Also, the possible result of data mesh being difficult to adapt is an extremely important finding if obtained. Questions can be found in *Appendix 1*.

3.2 Question Layouts

Questions for this study are formed from various insights and standpoints of different professionals in the academic and business world. Questions are set in a neutral form with a little challenge at the same time. Questions aim to be eye-opening for interviewees to learn something and maybe find new viewpoints for their organizations data situation.

Questions are built to be answered even though the interviewed person does not profoundly understand data mesh architecture. The interviewed persons are IT and data professionals from different companies with variable industry backgrounds. The variety of different industries is a strong factor, and it lets us have a broad outlook towards the world of data on a practical and executive level. Theme interviews included previously mentioned and categorized 29 questions. The layout around these questions was rather neutral, and interviewees could answer the questions during an open-minded conversation that supports common qualitative research methods.

Most premier questions are highlighted in *Table 1*. These questions are opened more closely in chapter 4, where we inspect answers from interviews and make visions on how these affect the bigger picture of data mesh.

Domain (DQ) - Organization	
DQ1	*How is domain defined in your organization? How many domains are there?*
DQ2	*Do all business areas (domains) get the data utilized at the level of require? Do certain domains make more use of data than others?*
General Data (GD) – Data Management	
GD1	*Does your company utilize data? From how many different sources does your company collect data?*
GD2	*Who in your company utilizes and consumes this data?*
Technology / Products / Services (TQ) – Data Management	
TQ1	*Are your business products and services primarily physical or digital?*
TQ2	*What is the general situation of the company's digitalization? Is there a designed data strategy?*
Maturity Level (ML) - Organization	
ML1	*Do you feel that the company's data literacy/maturity level is high enough for a distributed model?*
ML2	*Are data team professionals (e.g., data engineer) overladen? Is the competence focused on a very small area or even into individuals at the moment?*
Data as a Product (DP) – Data Management	
DP1	*Does the company provide data for external use, or does it only utilize its own data?*
DP2	*Does company processes/operations generate data that could be utilized, but is not yet in use?*
Ways of Working (WW) – Organization	
WW1	*Is the development team responsible for the product being created, is the business involved in this?*
WW2	*What are the approximate sizes of the data teams? (How many data engineers, project managers, etc.)?*
Data Ownership (DO) – Data Management	
DO1	*Who owns the data in the company? Is data ownership in centralized solutions or decentralized in business areas?*

Table 1: The core questions of the thematic interview according to the framework.

3.2.1 Organization Questions

As *Figure 8* shows, the data mesh suitability tool has organizational questions formed into three dimensions. These three dimensions are seen as important parts of the organizational side of data mesh. Organizational questions tie up the organization defining of the domain, level of maturity and common ways of working. Key questions for these dimensions are highlighted in *Table 1*.

3.2.2 Data Management Questions

Technological questions are set up on the right side of *Figure 7.* Data management and its three dimensions form a strong base for questions focusing on the data technical side. We need to point out clearly that data mesh is not a technological framework, or it is not a new technical solution to supplant data lakes or warehouses. The goal is to strengthen the way data is handled. Data mesh has data management and technological viewpoints, so this needs to be a solid part of our question battery. These data management related questions also help us find essential information from the interviewed organizations and their data architecture situation. Data mesh creator Zhamak Dehghani (2019) states that data mesh is a new paradigm shift for organizations to adapt to the changing data world.

3.3 Study Reliability

Research studies commonly include different error and distraction factors that can affect the study results, and this way, the whole study reliability can also be in danger. Therefore, reliability assessment is a key part of scientific research, as it has certain standards and values that it should strive for (Saaranen-Kauppinen & Puusniekka, 2006).

In qualitative research, it is essential to access the credibility and reliability of research. For example, the results of a qualitative study must not be random, and the methods used in the study must be able to examine what the study is intended to investigate. In addition, the concepts used must fit the content of the research problem. One aspect of the reliability of qualitative research is generalizability or transferability: whether the research results can be generalized or transferred to other objects or situations (Tutkimuksen toteuttaminen - Jyväskylän yliopisto, 2010). As this study is qualitative research, attention has been paid towards reliability.

Theme interviews naturally have similar issues and error factors. Theme interviews fit the research when the issue to be explored is not very well known and the research design is not completely locked. It may also be clarified as the project progresses. Whereas, in the light of the answers received during the interviews, additional

questions will be asked. Data mesh is not extensively known or defined yet. Although we had questions prepared beforehand for the interviews, leaving the interview situations open-minded and interactive, we could reach just the right amount of conversation with each interviewee (Routio, 2020). Every interview situation was a little different, and the question battery prepared in advance could be modified to follow the direction of the conversation.

The following graph shows the progress process of this thesis. Four highlighted stages are explained, and the thesis progress bar is supported with the followed schedule of the thesis work.

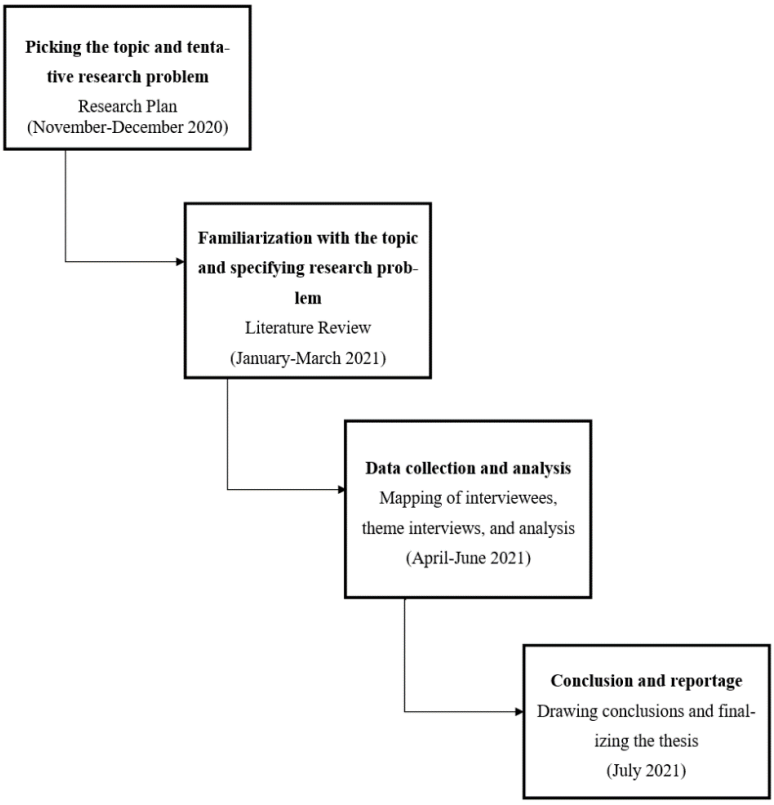


Figure 8: *Thesis progress process*

A well-documented and clarifying piece of writing is a key factor in achieving the trust of the reader. The study is nicely written and explicit to follow while reading, overall study reliability rises. These factors towards study reliability have been taken into account while writing the thesis methods and results.

4 CASE STUDY OF POSSIBLE DATA MESH ORGANIZATIONS

A total of seven interviews were conducted for the implementation of the thesis. The results from these interviews serve as the material for the empirical part of the research. As mentioned previously, theme interviews were chosen to be the form of data acquisition. This section will focus on the journey of the interviews, with the findings and results part. The chapter will include highlighting of the companies interviewed and critique of the method used during the interviews. The most important questions and answers are inspected. Results from the interviews and attachments towards the data mesh framework are drawn in verbal form.

As this thesis was done in collaboration with a business, thanks to Solita Ltd, we achieved a significant and respectful sampling of organizations for research use. Organizations were contacted via email with an invitation for an interview. The seven organizations were found rapidly, and we moved towards booking the interviews for each organization's representative. Interviews were carried out with a business communication platform, Microsoft Teams. Each interview's videoconference had a 1–2 -hour booking. The average duration of the interviews was 1 hour and 10 minutes. Interviews were recorded for later review and the interviewer took notes during the conversation.

The interviewees were professionals who have great responsibility and mission to develop their organizations data efficiency further. These people demonstrably know the significance and importance of data utilization and see pain points/challenges from the parade ground for their organization. Interviewees carry out different responsibilities and job titles, such as Data Management Manager, Data Lead and Head of Data & AI. All these positions aim to enhance the use of operational and analytical data. Interviewees being mostly senior and higher executives is a great advantage to get comprehensive insight from case organizations.

4.1 Case Companies

Seven major Finnish companies with a great view into their industry and data-driven thinking were chosen for the theme interview process. These companies have their unique situation with utilizing data. They all are challenging their industry to develop further and their ways of working with data. These seven case companies represent a wide scale of different industries that have a significant impact in Finland.

Case companies will be named anonymously (e.g., Case Organization X) during this thesis. Case organizations represent the following industries: *Wood/Forest industry, Telecommunications, Oil refining & Renewable products, Energy generation, Waste recycling and Construction industry.* All these companies have a specific way of demanding, producing, and consuming the data available. The differences between the organizations and industries are one reason why this theme interview builds a good variety of standpoints. At the same time, it shows us what similarities and common pitfalls organizations struggle with.

Organization	*Revenue (2020, million euros)*	*Number of Employees (2020)*	*Operating range by countries*
Case Organization 1	200 - 250	250 - 500	1
Case Organization 2	800 - 999	1 000 - 2 000	3
Case Organization 3	10 000 - 12 000	4 000 - 5 000	14
Case Organization 4	8 000 - 10 000	15 000 - 20 000	12
Case Organization 5	100 - 199	250 - 500	1
Case Organization 6	100 - 199	250 - 500	4
Case Organization 7	5 000 - 7 999	9 000 -10 000	30

Table 2: *Key indicators from interviewed organizations in 2020.*

Table 2 shows us some general information about case organizations. Some information is scaled to a certain range. For example, the number of employees and revenue information is set to a specific area for each case org. The table shows revenues, number of employees and operating range from all organizations. This information was selected to give a little foreknowledge towards case study sampling. All companies are major operators in their field in Finland. As *Table 2* opens the situation, Case organization 4 is the largest when watching the number of employees. Case organizations 1 and 5 represent small and medium-sized companies. Case organizations 3, 4 and 7

have the widest operating range, crossing over ten countries each. Revenue for case organizations differs a lot, as we can see completely different scales of revenues on the table.

These numbers strengthen the vision that our organizations are strong, modern, and willing to change over time. Every case organization has their unique way of producing data, and this way, we will find interesting standpoints, if and how data mesh could be implemented.

4.2 Theme Interview Study

The first booked interview situation was used as a test interview, and it helped us evaluate the question battery and its capability to support correct information regarding our research problem. The test interview was a success, and we could continue the interviews with the working question body. This first interview was recorded and documented in the same way as others. The questions did not change after this first test interview. All interviews had the same pattern that was followed, although different clarifying questions were expressed. Having free, and flexible structure is a big strength of open theme interviews; every situation is little different.

4.2.1 1st and 2nd Dimension Questions

This section includes the first two general level dimensions: Domain and Technology/Products/Services. The interviews began with going through practicalities and formatting the question battery concept for an interviewee. Generic questions and numbers are highlighted in *Table 2,* but we also asked if organizations products and services are primarily physical or digital. All seven organizations saw that their services and products were originally physical. However, without exception everyone had the desire to see data and digitalization as important as the basic physical business. This specific insight from organizations tells us that every industry is affected by digitalization. Organizations have a transparent vision to develop their business towards the complex world of information systems. For example, one of the interviewees stated the following:

"We have both digital and physical products/services, although it is very difficult anymore to see our activity as fully physical. Many services from the customer's point of view are digital, although it could still require physical processes from the company perspective. Digitalization revolutionizes the industry on a rapid phase." (Case Organization 5).

After getting answers to generic questions about organization size and industry, we headed straight to asking general questions about data usage. The first general data questions touched how many data sources the organization has, and who uses that data. The answers were very homogenous, and every case organization saw itself as having many data streams or sources. All companies were also able to explain what kind of data they collect and utilize.

"We have lots of data sources, for example more than 400 business applications. SAP creates a strong backbone for various applications and systems. We have transactional information, IoT data, image, video, and binary data, all data is mainly structured information." (Case Organization 3).

Other similar size operators stated the following:

"We utilize hundreds of data sources; external data is utilized in marketing and pricing. Our own systems bring us hundreds of sources of information. There are sensors, photography, video, external data, and relational data. Automation supports the vast amount of data." (Case Organization 4).

Previous comments proof us that large organizations deal with numerous types of data. Overall, data was seen to be used in most areas of an organization. We also found that larger organizations could have hundreds of data sources and various streams to give valuable information. Multiple data sources and streams create a flagrant need for effective data management, and architecture must not be allowed to become a bottleneck. To open the challenges with vast amounts of data, Case Organization 4 continued that:

"A large number of data sources creates a huge information blockage in factories operating systems. The large amount of IoTs and sensors is a challenge for us." (Case Organization 4).

The previous answer shows us that organizations also struggle with vast amounts of data. More isn't always the better, and effective discovery of important data is necessary in this case. Smaller organizations, such as case organizations 1 & 5, had fewer data sources, approximately between 5-15 each. Having fewer data sources does not mean they would not need suitable data architecture. Every modern organization needs explicit and convenient blueprints for data.

Next, we moved towards finding different definitions for the domain. Defining domains proved to be one of the most interesting dimensions for interviewees, and they found new viewpoints during the conversation. The definition of a domain is volatile and differs between companies. The number of domains ranged from a few to several dozen. Case organization 4 explained the following: "We see 12 domains; some domain volatility appears with shared data assets. Having multiple domains using and processing same the data sets creates an urgent need for understandable data architecture. Both business and IT need to see domains in a similar fashion."

Case Organization 6 continued the domain conversation as follows: "There is a lot of volatility in our domains, at least from a master data domain perspective. About 5-10 domains, which is also a big scale to give. So-called heavy users make better use of data than others." Following the previous answer, it tell us that domain definition isn't as clear as you could think. Every single interviewee also pointed out that domains can be seen in different ways. "Data domains, business domains, there is a big mountain to climb with these different definitions, at least for us." (Case Organization 1).

4.2.2 3rd and 4th Dimension Questions

This section opens Maturity and Data as a Product dimension answers. The third dimension included topics around maturity and skillset level, these questions brought up interesting unity between organizations. Question 20 opened the conversation about the organization's maturity level by asking if the interviewee saw maturity level or data literacy as high enough for distributed data teams and architecture. Most organizations told the same story – maturity level is not at the level it needs to be. Data literacy and lack of maturity around the organization seems to be one of the main challenges:

"Variable data literacy, the general level is under development. The overall level is advancing, our top data engineers and data stewards are overladen, and this is identified as a challenge." (Case Organization 7).

"The maturity level is not high enough. In general, it should be developed throughout the organization. Understanding what data is available, how to get it in your hands, and what to do with it. These are the vital things to achieve." (Case Organization 6).

These answers tell us about the harsh situation in various organizations. Organizations have so much data on their hands that getting a proper grip is difficult. Organizations also brought up knowledge spread as a perceived challenge. Domains don't seem to be on the same level across the organizations, some domains are independent when it comes to data utilization, and some are still starting their data journey.

Some organizations were already having some characteristics of distributed architecture and ownership among their data management. For example, the following organization explained their situation with maturity level differences between domains:

"Maturity level is now at a good level. In the past, the IT side has had clear challenges. Now we have some clear decentralization. Previously we had few strong units, but today the business side has stronger data expertise." (Case Organization 2).

During the interviews, my perception of successful change in a few organizations strengthened my belief that data literacy could increase in distributed architecture. Data mesh principles seem to fix some maturity level issues in organizations, but this most definitely needs great attention, and it does not happen automatically. On the other hand, pushing the data ownership and responsibility towards domains automatically increases data literacy.

Also, few interviewees pointed out the industry factor. Some industries are more agile and ready for the latest technological innovations, and some are very committed to traditional ways of working. Case organization 1 gave a really good example from their situation, where field-level is difficult for innovations to flourish. The same

organization also highlighted the importance of supportive and understanding management.

> *"The conditions and operating environment must be understood throughout. In our industry's field-level, data needs are the last needs overall. Support from management is very important. Attention towards data should be paid from the executive level." (Case Organization 1).*

Next, we went through data as a product sub-area. Data is a product questions seemed the hardest for the interviewees to answer. Data had different use cases, and of course, the needs between organizations varied highly. Overall, data was seen as a service. Still, some organizations handle data as a product when it was produced for internal use. These internal use cases include analyzes, statistics, and analytical charts. The definition of product emerged during a few interviews, and the definition for data product was challenging to generate. When asking if data had an assigned value for it, we received unanimous answers. Data did not have a clear assigned value and it does not receive attention from a business perspective. We asked if data was considered as a factor in financial statements or data account statements. Again, answers revealed that organizations had not defined a clear value for data.

> *"No specified value for data. Metadata should be given a specified value. Along with staff, data is an equally valuable intangible asset. It should most definitely get more attention" (Case Organization 3).*

> *There is no specified value for our data, but it would be necessary. This would bring more understanding and visibility towards data. (Case Organization 6).*

These thoughts sum up the great divide of data we are witnessing. Case Organization 3 impressively stated data to be an equally important intangible asset as staff. Case organization 3 commenting the importance of data shows how much attention data requires. Organizations have so many use cases and goals with data that it seems almost too difficult to see all data as a product within one organization. Though, some domains could most definitely implement this framework and mindset to their daily work. Data products could serve new demands and needs of the customer that data

services can't fulfil. Data, as is a product questions, left one big aggregating thought: How to create a data product if your data does not have a specified value?

4.2.3 5th and 6th Dimension Questions

This final section gathers together answers from two executive-level dimensions, Ways of Working & Data Ownership. According to literature and data mesh principles, data ownership should articulately be issued to a specific place or even an employee. A clear vision if data ownership fills the void of responsibility towards pivotal data assets. When domains consume and use the same data sets, clear ownership and accountability benefits, everyone. Data ownership should be defined directly within domains, and our interviews gave similar answers.

> *"Data ownership is difficult to identify if ownership of the data is taken too far from the entry-level or operating system. Business must be an active factor." (Case Organization 2).*

> *"Data ownership is decentralized. Ownership can be clearly found through core processes" (Case Organization 5).*

> *"Through the master data, the owners can be found. Ownership is commonly found, even if it is not always clearly displayed. Our ownership is mostly centralized. We also have business areas where ownership isn't clear. The different levels of domains can be seen here as well." (Case Organization 6).*

Most of the case organizations had decentralized data ownership. This decentralized situation tells us that ownership has been distributed among the domains and business functions. Few exceptions did observe. When data ownership was centralized, more confusion was in the air about where or to whom the data belongs. When the gap between business and IT was narrowed down, clearer ownership for data and processes was striking.

Large organizations with multiple domains consuming the same data also create challenges for data ownership. Our case organizations had a couple of solutions for this. Data governance units and different data catalogue factors clarify the ownership

muddle. These solutions and tools are great hands-on examples of how to improve your organizations data mess.

The last questions for our interviews sought answers for common ways of working procedures. Agile methods and different DataOps methodologies are clearly in use with all of our organizations. Data teams typically had something between 3-10 members, and smaller organizations had smaller teams. Larger organizations had more variability between their teams. However, the so-called standard data engineer role had a lot of diversity.

> *"The job description of a data engineer today is very broad, and the level of requirements has grown massively." (Case Organization 3)*

The job transformation for a specific employee is a typical transition in a newly developing industry, such as the data industry. The industry develops in such a rapid phase. Employees must receive continuous training. One organization also pointed out that their unique set of operative systems and data bases is a major challenge for recruiting new employees for data positions. "Data teams are a clear bottleneck; another clear bottleneck is the demanding nature of our operating factory systems. Finding the right architecture is a challenge. The flexibility of technology helps us in our development work." (Case Organization 7). This organization sees their development points and has a transparent vision to fix them. A transparent vision of change is extremely important while the industry and world around you change rapidly. A profitable organization is willing to change its ways of working and maybe even set new norms with their innovations.

Lastly, we finished the interview by asking if organizations see themselves having centralized data teams or not? Out of seven organizations, two had distributed architecture, one had centralized, and the remaining four had something in between. These results were not too surprising, understanding that we processed through a comprehensive question battery of 29 questions. After finding out what kind of architecture these organizations are running with, we specified our questions according to interviewees' answers. Finally, if the organization was adapting distribution and decentralized data teams, we asked if data utilization and management improved in a decentralized model? Both organizations with distributed data ownership and teams replied that their

data management has improved after moving towards decentralization. Of course, this doesn't always mean that the organization is fully using the data mesh framework, but it is most definitely adapting its common principles.

> *"We are currently fully decentralized. A common so-called "data handbook" is required for multiple data teams across domains. Business areas have benefitted directly from having decentralized teams. There must be an opportunity to make creative solutions". (Case Organization 3).*

> *"Decentralization has brought data closer to business. As a result, responsibility is given to business data experts. Our operations are more streamlined, and you don't have to ask every single thing from a centralized data unit." (Case Organization 5).*

These organizations are delighted with their decision to adopt a more distributed architecture. Although, we need to remember that distribution does not suit every organization. One notable point we need to consider is that these two organizations are highly different in size. Case organization 3 is large, meanwhile, organization 5 represents small and medium-sized enterprises. Although, they are both adapting decentralized architecture and ownership methodologies.

Our four organizations who were something in between centralization and decentralization, had few conjunctive factors. They all had a clear centralized data team or management, but data stewards and ownership were distributed across the organization in certain places. Some domains could usually demand more specific data than others, where few experts might be distributed in these cases. For all five organizations that did not have complete decentralization, we asked if they would move to distributed model with a specific business unit before moving the entire organization. For example, the transition would be done step by step, using a business unit with the required capabilities. All five organizations answered that they would prefer this step-by-step method, to get a good success rate and stories from it. This way, organizations can focus on a specific unit and find the most important challenges to tackle. Decentralization is a big change in work culture, and it shouldn't be done without good schemes.

4.3 Results & Observations

Now after we have gone through our theme interviews, we can set ourselves for making observations. First, we can draw upscale conclusions from our data acquisition. At the beginning of the interview, organizations answered a question about what changes their industry the most. 5 out of 7 answered that digitalization and rapid scaling with data applications are the main reasons their industry and day-to-day work changes. The other two organizations also pointed out digitalization and data as a big factor, but sectoral changes in their unique industry took the biggest role. For example, physical resources, such as oil, will vanish from the world at some point, and new materials will change how products or services are manufactured. These are things that data cannot change but having a good grip of your information is a great platform for future business applications.

Most of the large companies are feeling the agility/scalability pain that data mesh is designed to solve. Dehghani (2019) explained that data mesh in its entirety should usually only be considered if you are hitting the wall with getting a better grip on your analytical insights. Organizations have multiple different data sources and operational streams. An efficient analytical data plane is the biggest challenge. The popularity of knowledge management and data-based decision-making thrives many companies to innovations, and data mesh can be one of these.

Organizations struggle with different challenges. Bottlenecks are a typical way to describe a part of the process which slows down the production line. Different bottlenecks pointed out during interviews: Slow development, amount of data sources (or lack of them), data teams, the complexity of substance systems, huge amount of raw data, data quality, and workload (backlog challenges). These bottlenecks are a good example of organizations having a variety of different roadblocks.

Overall, multidimensional organization models and higher complexity of data domains seem to create a better breathing ground for data mesh principles and implementation. Before this research, we had some insight that larger organizations with complex domains would better fit data mesh. Our results most definitely support this finding, and

we can safely say that data mesh has certain organizational standards it requires to be completely efficient.

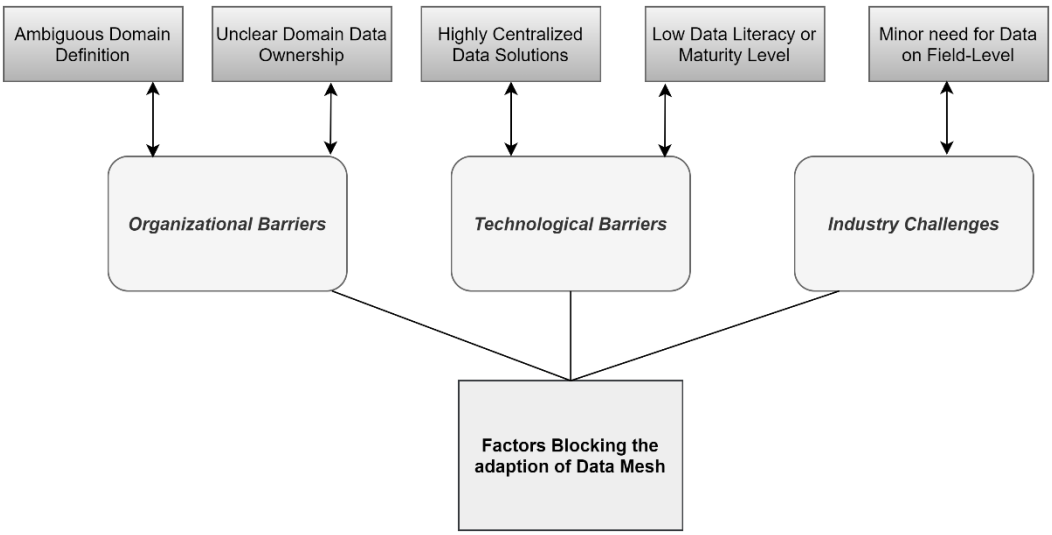
Case organization 3 interviewee described data mesh as a refreshing change to the data industry. Distribution isn't a new thing for enterprises to optimize their functions. Human resources and IT departments are classic instances of commonly distributed units. Data teams are overall very agile and adaptive towards new trends and features to advance ways of working.

> *"Data mesh is not a new model, and it is now rebranded" (Case Organization 3).*

While reading articles, blogs or whitepapers about data mesh and distributed architecture, you can always see someone saying that their organization has done things this way and adapted certain methods years ago. Organizations have done this previously might be true, but data mesh includes precisely designed patterns that have to match. Data mesh could be the next "Big Data" megatrend that everyone wants to understand and experiment with.

Domain questions and discussions also brought up different standpoints across the organizations. Definition of domain and differences between data domains and business domains mixed the ideas thoroughly. The core meaning of domain in every organization seems to be one of the most important things to define. Having an indefinite understanding of your domains can create big chaos when benefiting every bit of your data.

The graph below shows the most important barriers and challenges for data mesh adaption journey. These pieces have been fabricated from theme interviews.



Graph 1: *Factors blocking data mesh adaptation*

The graph shows us key indicators and factors blocking the use of distributed architecture. Decentralization requires heavy changes in any organization, and these organizational and technological barriers point out a few important ones. Industry challenges can also include various different factors, but the minor need for data on practical-level lift its head during our interviews. Organization 1 stated that their industry itself has challenges when trying to implement new data trends. This may result from different variations, but for example, employees on this industry could be typically against new technologies or find new trends difficult to use.

According to our research, we can point out the organizations that already implement some data mesh principles. Our results also show that the data mesh framework could be implemented in all case organizations. All 7 case organizations are willing to adapt their ways of working and data vision. Every organization also have found their specific challenges and barriers to tackle. Of course, it is much easier for organizations with decentralized data teams and distribution work is already done to implement more data mesh principles for their situation. Having a clear definition of your domains and the differences, responsibilities, data in use, and business goals between those domains seems to be just a few of the most important things. Data mesh requires strong domain definition, and a domain-driven design mindset helps towards this.

5 CONCLUSION

This thesis has addressed the features of the data mesh framework, examined the importance of domains, and determined what kind of enterprise a distributed architecture is suitable for. The goal of this thesis has been to find out what makes organizations suitable for data mesh implementation, and how to proceed towards distributed architecture. Our organizational spread and industry versatility allow us to see how different organizations adapt new data management frameworks.

In this chapter, we seek answers to our previously mentioned research questions and state if our hypotheses are correct. Organizations eager to adapt new data trends find data mesh principles more suitable for their use. Data mesh doesn't have to be fully implemented in all variations or forms, and every enterprise can pick up the most paramount formulas for their use. Data mesh is designed to enlighten the path for organizations towards more structural and efficient data management.

As a result of this study and answer to our first research question we can note that various organizations can adapt data mesh principles. Still, few key indicators need to be considered. The research question also sought an answer on how an organization can proceed towards distributed architecture. We can safely state that organization has to start by having a clear definition of domains, clarifying data ownership, and giving data more attention to bringing up the overall data literacy and maturity level.

Our second research question was about finding definition for domains. A clear domain definition is a starting point for any data-driven organization. Being data-driven does not rule out a domain-driven mindset. These together can form a supremely strong base for any organization to scale with data business. We can strongly indicate that data mesh suits organizations with multiple and complex domains. A structured specification for domains and the data possessed by those create a powerful basis for a data mesh framework. A well-designed data domain and entities for your most important data leads to a well-rounded data warehouse or master data management system, for example. These are examples on a practical level and every organization struggling with its core data can learn from this.

Following Sub-Questions helped us to form a comprehensive insight towards new data management methodologies. We tried to find answers to whether data pipelines were really more streamlined with data mesh principles. We found out that two of our case organizations stated their data processing and pipelines to be more streamlined with distributed architecture and data teams. Case organizations stating the improvement with distributed architecture tells us that data mesh can improve organizations data management on an executive level. We also tried to find the most important challenges and benefits in the data mesh framework. Our strong literature review supported the similar findings our theme interviews gave. Our literature section gave potent proof that data mesh also has issues but by focusing on the strengths of a framework, organization can achieve its data goals.

In data mesh, we can have multiple different data definitions for different attributes, and the variety of definitions creates the value within the domains that require a specific definition for their data in use. Therefore, CDM does not fit the same page with data mesh; the raw definition and goal of attributes, entities and data domains are way too different. Therefore, we can safely state that CDM does not support data mesh or domain-driven design principles.

We need to remember that when talking about data: more is not always better. It requires skill and profession to understand what data your organization need and what is secondary for the business. Data mesh is a tool for understanding what data organization truly needs. Data mesh challenges the traditional perspective that big data must be centralized to leverage the analytical potential from operational data. Data mesh gives options to deal with the issues big data hype created. Larger amounts of data are not always good for your organization or data management efficiency. Understanding what data is essential for your business and how to develop, process and manage that is the solution for many enterprises struggling with data-related issues.

The data industry and business are facing a rough skill cap challenge. On the competence level, data experts become too specialized in their area of expertise. They may create platform-level bottlenecks due to difficulty of finding specific data engineering talent. Organizations need to distribute the ownership and know-how across domains to prevent this.

This thesis and the results set up a strong foundation for future data mesh studies. Upcoming studies can continue with expressed conclusions and findings in further research. Our results were aimed to be a starting point for academic data mesh research, and this thesis achieved this goal. Future research could include topics such as, data fabric, data catalogue and data mesh on the practical level. It would be extremely interesting to learn how field-level data professionals find data mesh principles to work in their organization. Do data mesh, distributed ownership, and decentralized data teams ease the challenging tasks of data experts?

Usually, every research has some issues or critique-worth sections to point out, and we also want to find clear improvements for this study. Our sampling of seven organizations wasn't the highest, but it clearly gave us a first taste of decentralized aspects and feelings from successful organizations. With the scope and objectives of a master's thesis, the empirical sample can be considered excellent. However, we could have asked more specific and a little different questions during our theme interviews. This insight came after going through our answers and doing transcriptions. We found out that few different questions should have been asked to point out new aspects about decentralization on the executive level. Another aspect our research could have included is any practical demo or functional application of how data mesh is done on the data level. Functional demos and practical examples are difficult to find but, there will most definitely be plenty to adapt from in the future.

Scientific research is a long-term and systematic activity with no quick gains. Scientific knowledge is the basis for sustainable growth, well-being, and civilization, but its outcomes cannot and should not be precisely determined. Otherwise, the birth of new knowledge and understanding is killed. The previous statement touches the heart of data engineering, new data trends, as well as data mesh very sharply.

The starting point of every research is curiosity, and the end result is always uncertain. Scientific work is the creative joy of discovering new things and, if successful, it leads to new questions, perspectives, and learning - such actions are not possible with short-term guidance (Mönkkönen, 2021). Data mesh needs a patient approach for organizations to achieve the full potential to change the world of data engineering.

Data mesh is a new trend at the beginning of its lifecycle. We cannot predict the future or know how well data mesh will be implemented in the data world in the long term. However, data mesh can definitely change how organizations utilize their data.

The creator behind evolution theory, famous scientist Charles Darwin stated that: “In the long history of humankind (and animal kind, too) those who learned to collaborate and improvise most efficiently have prevailed”. This statement also applies to the data mesh framework changing the traditional monolithic solutions. Therefore, those organizations that are willing to adopt the new operating models will prevail and scale higher in the future.

Reference

Brandolini, Alberto. (2013). Introducing Event Storming - An independent blog on Software Development. Retrieved 11 May 2021, from http://ziobrando.blogspot.com/2013/11/introducing-event-storming.html

Bourque, P. & Fairley R.E., (Eds.). (2014). Guide to the Software Engineering Body of Knowledge, Version 3.0, IEEE Computer Society, www.swebok.org.

DataKitchen. (2021). What is DataOps? Retrieved 13 July 2021, from https://datakitchen.io/what-is-dataops/

Dehghani, Z. (2019) How to move beyond a monolithic data lake to a distributed data mesh. Martin Fowler's Blog, 2019. https://martinfowler.com/articles/ data-monolith-to-mesh.html as of May 20th, 2021.

Dehghani, Z. (2020a) Data Mesh principles and Logical Architecture. Martin Fowler's Blog, 2020 https://martinfowler.com/articles/data-mesh-principles.html as of February 15th, 2021.

Dehghani, Z. (2020b). Keynote - Data Mesh by Zhamak Dehghani. Presentation.

Etlia. (2021). Data Mesh - vaihtoehtoinen tapa hallita dataa. Retrieved 31 March 2021, from https://etlia.fi/data-mesh-vaihtoehtoinen-tapa-hallita-dataa/

Evans, E. (2004). Domain-driven design: tackling complexity in the heart of software. Addison-Wesley Professional.

Fowler, M. (2020). Domain-Driven Design. Martin Fowler's Blog. https://martinfowler.com/bliki/DomainDrivenDesign.html as of April 15th, 2021.

Furia, B. (2021). Data Mesh – Rethinking Enterprise Data Architecture. Cuelogic Technologies Ltd. Available at: https://www.cuelogic.com/blog/data-mesh Retrieved 15 September 2021.

Galetto, M. (2016). NGDATA | What is Data Management?. Retrieved 24 March 2021, from https://www.ngdata.com/what-is-data-management/

Garousi, V., Felderer, Michael. and Mäntylä, Mika V. (2016). The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature. In Proceedings of the 20th international conference on evaluation and assessment in software engineering, pages 1–6.

Hagiu, A., & Wright, J. (2020). When Data Creates Competitive Advantage. Retrieved 24 March 2021, from https://hbr.org/2020/01/when-data-creates-competitive-advantage

Hansen, J. (2020). Common Data Model - Unelma yhtenäisestä tietomallista. Retrieved 24 March 2021, from https://jannehansen.com/fi/cdm/

Hovi, J. (2021). Mikä on datatuote? - Ari Hovi. Retrieved 26 July 2021, from https://www.arihovi.com/mika-on-datatuote/

ISO (2006). ISO 19439:2006(E) Enterprise integration – Framework for enterprise modelling.

Jyväskylän yliopisto. (2010). Tutkimuksen toteuttaminen. Retrieved 14 July 2021, from https://koppa.jyu.fi/avoimet/hum/menetelmapolkuja/tutkimusprosessi/tutkimuksen-toteuttaminen

Knight, M. (2018). Data Lake vs. Data Swamp: Leveraging Enterprise Data - DATAVERSITY. Retrieved 29 April 2021, from https://www.dataversity.net/data-lake-vs-data-swamp-leveraging-enterprise-data/#

Lewis, J., & Fowler, M. (2014). Microservices. Retrieved 23 July 2021, from https://martinfowler.com/articles/microservices.html

McHugh, J. (2017). Data Warehouse Design Techniques - Conformed Dimensions –. Retrieved 24 March 2021, from https://www.nuwavesolutions.com/conformed-dimensions/

Microsoft - Common Data Model. (2020). Retrieved 24 March 2021, from https://docs.microsoft.com/en-us/common-data-model/

Mönkkönen, J. (2021). Mistä tieto tulee ja minne se menee?. Retrieved 3 August 2021, from https://www.savonsanomat.fi/paakirjoitus-mielipide/4225359?fbclid=IwAR2D0I6srwG5_J14rRmknAB9ipWZGQO--sWasRme2UWbo-HUBLuBssSGYU

Neuman, B. C. (1994). Scale in distributed systems. ISI/USC, 68.

Papazoglou, M., Traverso, P., Dustdar, S., & Leymann, F. (2007). Service-Oriented Computing: State of the Art and Research Challenges. Computer, 40(11), 38-45. doi: 10.1109/mc.2007.400

Quemy, A. (2019). Data Pipeline Selection and Optimization. In DOLAP.

Rodriguez, M., de Araújo, L. J. P., & Mazzara, M. (2020, December). Good practices for the adoption of DataOps in the software industry. In Journal of Physics: Conference Series (Vol. 1694, No. 1, p. 012032). IOP Publishing.

Routio, Pentti. (2005) Teemahaastattelu. Tuotetiede. Taideteollisen korkeakoulun virtuaaliyliopisto.

Saaranen-Kauppinen, Anita & Puusniekka, Anna. (2006). KvaliMOTV - Menetelmäopetuksen tietovaranto. Tampere: Yhteiskuntatieteellinen tietoarkisto https://www.fsd.tuni.fi/menetelmaopetus.

Schwab, Klaus. (2016). The Fourth Industrial Revolution: what it means and how to respond. Retrieved 18 March 2021, from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Seiner, R. (2016). What is Non-Invasive Data Governance?. Retrieved 13 July 2021, from https://tdan.com/what-is-non-invasive-data-governance/7354

Shahrin, N. (2021). Building a successful Data Mesh – More than just a technology initiative. Retrieved 26 July 2021, from https://www.linkedin.com/pulse/building-successful-data-mesh-more-than-just-nazia-shahrin/

Stedman, C., & Vaughan, J. (2020). What Is Data Governance and Why Does It Matter?. Retrieved 13 July 2021, from https://searchdatamanagement.techtarget.com/definition/data-governance

Serra, J. (2011). Conformed dimensions | James Serra's Blog. Retrieved 24 March 2021, from https://www.jamesserra.com/archive/2011/11/conformed-dimensions/

Uludağ, Ö., Hauder, M., Kleehaus, M., Schimpfle, C., & Matthes, F. (2018, May). Supporting large-scale agile development with domain-driven design. In International Conference on Agile Software Development (pp. 232-247). Springer, Cham.

Zhu, P. (2013). Software Architecture vs. Data Architecture. Retrieved 22 July 2021, from http://futureofcio.blogspot.com/2013/10/software-architecture-vs-data.html

Appendix 1: Theme Interview Questions Form

Questions for Data Mesh Suitability:

Question 1: What is your company's industry? In how many countries do you operate?

Question 2: Are your business products and services primarily physical or digital?

Question 3: How are the company's future determined, is there a clear change in the company's industry? For example, are the company's services and products developing into new ones, or maybe they stay the same?

Question 4: What is the general situation of the company's digitalization? Is there a designed data strategy?

Question 5: Does your company utilize data? From how many different sources does your company collect data?

Question 6: What is this data like?

Question 7: Who in your company utilizes and consumes this data?

Question 8: Is it possible to make analytical decisions with data?

Question 9: How domain is defined in your organization? Is the division of domains clear or, for example flickering? How many business domains does the organization have?

Question 10: Do all business areas (domains) get the data utilized at the level of require? Do certain domains make more use of data than others?

Question 11: How do data consumers know what data they need? Do those who need data from business areas know whether a potential centralized team can distribute the necessary data, so-called "automatically"?

Question 12: Has the company decentralized other functions in the past, such as human recourses or IT support?

Question 13: Do the company's business units have the ability to leverage data independently?

Question 14: Do you see the data specifically as a service or as a product itself? Is data just a "by-product", support function or the "main thing" itself?

Question 15: Does the company provide data for external use, or does it only utilize its own data?

Question 16: Does company processes/operations generate data that could be utilized, but is not yet in use?

Question 17: Is there a value assigned to the data? Is the data considered as a factor in the financial statements or data account statement?

Question 18: Do you have clear bottlenecks in data production?

Question 19: Do decision making units (higher management) that are important to the organization receive data and analysis to support decision making?

Question 20: Do you feel that the company's data literacy/maturity level is high enough for a distributed model?

- → If yes, has it been easy to find data experts to fill the needs of the company?
- → If not, should something specific be developed in information management?

Question 21: Are data team professionals (e.g. data engineer) overladen? Is the competence focused on a very small area or even into individuals at the moment?

Question 22: Do business and application development/data experts communicate clearly enough?

Question 23: Would it be possible to hold business units to take greater responsibility for data utilization?

Question 24: Who owns the data in the company? Is data ownership in centralized solutions or decentralized in business areas?

Question 25: Is it clear in the company who owns this specific data?

Continued: No, where would you feel these data owners to fit the best? In a centralized or decentralized part of the organization?

Continued: Yes, where are the data owners exactly? Is it even addressable by a domain?

Question 26: Does your organization use agile methods or agile development? Is the company also agile with data business?

Question 27: Is the development team responsible for the product being created, is the business involved in this?

Question 28: What are the approximate sizes of the data teams? (How many data engineers, project managers, etc.)?

Question 29: Does the company currently have centralized data teams? Is data management centralized in a specific business area or are the data experts decentralized throughout the organization?

→ If centralized (*1*), if distributed (*2*):

1. What about the possibility of decentralization in the organization? Thus, data teams would work closer to data consumers or business areas.

1. Would you prefer to try a decentralized model in specific business area first, before possibly moving the entire organization? So, would you rather make the transition step by step, using a business unit that has required capabilities?

2. Has data utilization and data management improved in a decentralized model?