

MSc Business Information Technology

Master's Thesis:

Towards a Data Mesh Reference Architecture

Daniel van der Werf

Supervisors:

João Luiz Rebelo Moreira (University of Twente)

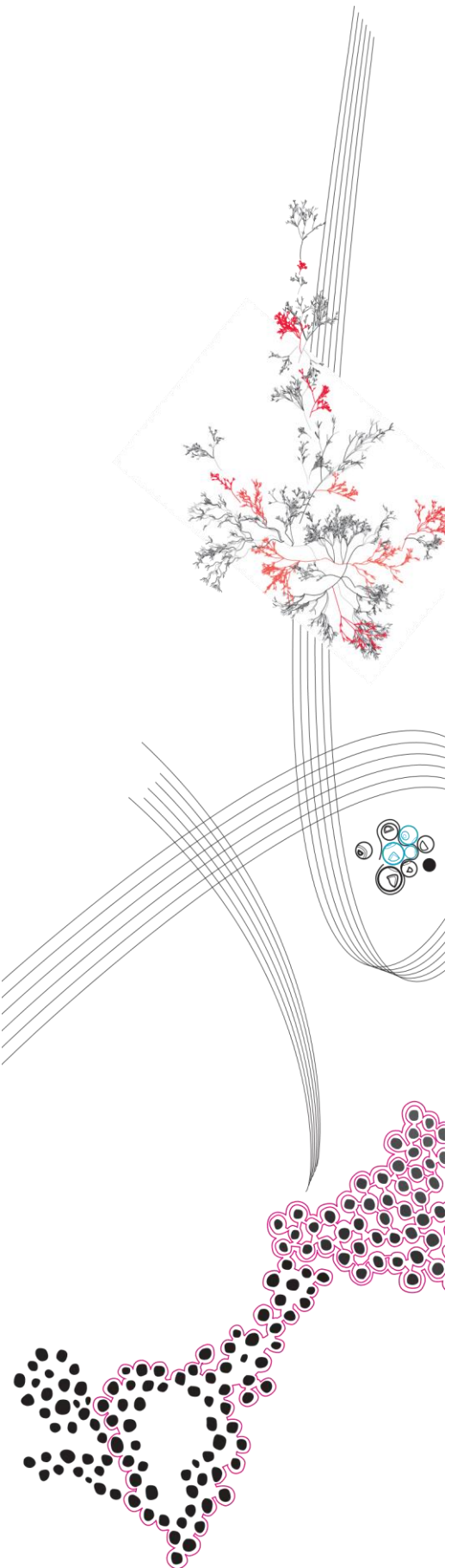
Sebastian Piest (University of Twente)

Jesse Struick (KPMG)

May 21, 2024

Faculty of Electrical Engineering
Mathematics and Computer Science

**UNIVERSITY
OF TWENTE.**



Executive Summary

The increasing complexity and volume of data within organization has created the need for a paradigm shift in data architectures. The monolithic architectures with central data teams are becoming a bottleneck and thus the data mesh paradigm emerged. This thesis explores the Data Mesh concept, a new way of structuring the enterprise data architecture by decentralizing the data capabilities and positioning those at a domain level, with an overarching governance framework, supported by a self-serve data platform. The data mesh paradigm emphasizes treating data as a product, in a domain-oriented decentralized data architecture, supported by a self-serve data platform, and with federated computational governance. These principles aim to solve the problems of traditional monolithic data architecture, such as scalability issues, data siloes, data quality issues and inefficient data processing.

This thesis aims to develop a comprehensive data mesh reference architecture, based on the ArchiMate enterprise architecture modelling language, to guide organizations in designing data mesh solution architectures. The data mesh reference architecture consists of 3 parts, which represent the main components of a data mesh, the domain architecture, the self-serve data platform architecture, and the federated governance architecture. The data mesh RA was developed using a 6 step method to develop empirically grounded reference architectures.

First, a systematic literature review was performed to identify the main data mesh structures and components. This resulted in the establishment of 4 data mesh archetypes with varying level of maturity, decentralization and domain independence: Pure Data Mesh, Semi-Pure Data Mesh, Hybrid Data Mesh and Distribution data mesh. Next, challenges and limitations of data meshes were analysed and based on those possible solutions and mitigation techniques were proposed. Following this, motivational factors driving organizations to adopt a data mesh and prerequisites for data meshes were identified. The impact of data meshes onto the organization was assessed and additionally, other data methodologies were compared with the data mesh to provide organization alternative data architecture approaches because data mesh is not a solution that fits every organization. Lastly, existing data reference architectures, reference architecture development methodologies and validation methods were analysed to guide the design and validation of the data mesh reference architecture.

The resulting data mesh reference architecture was validated through questionnaire distributed among experts. The results of the questionnaire validated that the developed data mesh reference architecture is a useful tool guiding the design of solution architectures, is of sufficient quality and has good variability.

This research contributes to practice by providing a data mesh reference architecture, providing a comprehensive blueprint for solution architects to design data mesh solution architecture. It details elements and their relationship, also serving as a checklist to examine concrete designs. This the first data mesh reference architecture using the ArchiMate language providing a foundation for future improvements, extensions and derivatives.

This research contributes to literature by conducting a literature review on data mesh structures, components, challenges, limitations, motivational factors and prerequisites, organizational and technical impact, and alternative data architectures. This study also demonstrates how an empirically grounded RA can be created using an enterprise architecture modelling language and how expert opinion can be used as a validation method through a questionnaire.

A limitations of this research is that the data mesh reference architecture was not validated in a practical case study thus leaving practical applicability untested. Additionally, relevant literature may have been missed due to the formulation of search queries and boundaries set by the inclusion and exclusion criteria. Next, the use of a single respondent group with a small number of respondents for some of the work roles may allow for personal bias and omit the

possibility more advanced statistical analyses. The questionnaire's closed questions limit response depth and the inflexible nature may oversimplify complex issues.

Future work should validate the data mesh reference architecture in practice through case studies in various industries. Additionally, a study comparing the efficiency of designing a data mesh solution architecture with the RA, compared to a group not using the RA, can be performed. Research could be performed to assess if the usefulness, quality and variability of the model improve after improvements have been made to the model. Lastly, future research is needed to update this research with new findings from theory and practice, for example, regarding best practices or by identifying different archetypes.

Keywords: Data Mesh, Reference Architecture, Data Architecture, Data Mesh Archetype, ArchiMate

Acknowledgements

This master's thesis "Towards developing a Data Mesh Reference Architecture" written as part of the Business Information Technology master program, concludes my academic journey at the University of Twente.

First of all, I want to express sincere gratitude towards my supervisors from the University of Twente, João Luiz Rebelo Moreira and Sebastian Piest, for their valuable guidance, continuous support, and encouragement, throughout the course of this research. Their insightful feedback kept me motivated and has been crucial to the completion of this thesis.

Additionally, I want to thank KPMG for providing me with an internship position and providing resources and support during my research. A special thank you goes out to Jesse Struick whose experience and insightful ideas have contributed substantially to the completion of this thesis.

Next, I want to thank the experts who participated in the questionnaire. Without their participation, I would not have obtained the necessary insights for this research.

Lastly, I would like to thank you as a reader of my thesis I hope it is informative and enjoyable.

Daniel van der Werf
Amersfoort, May 2024

Table of Contents

Executive Summary	1
List of Figures.....	7
List of Tables	8
List of Abbreviations	9
1 Introduction	10
1.1 Rising Interest in Data.....	10
1.2 Different Data Platforms.....	10
1.3 Data Mesh Paradigm	12
1.4 Scientific Relevance.....	13
1.5 Research Goal and Questions	14
1.6 Research Methodology	15
1.7 Thesis Structure.....	16
2 Theoretical Background.....	17
2.1 Data Mesh	17
2.1.1 Data as a Product.....	18
2.1.2 Domain in a Data Mesh	18
2.1.3 Self-Serve Platform	19
2.1.4 Federated Computational Governance	19
2.1.5 Data Mesh Benefits and Challenges.....	19
2.2 Enterprise Architecture.....	21
2.2.1 Benefits and Challenges of Enterprise Architecture	21
2.2.2 ArchiMate	22
2.3 Reference Architectures.....	22
3 Systematic Literature Review	23
3.1 SLR Planning.....	23
3.1.1 Search Queries	23
3.1.2 Inclusion and Exclusion Criteria.....	24
3.2 Literature Search and Selection	24
3.3 Data Mesh Structures, Components and Limitations.....	26
3.3.1 Data Mesh Archetypes	26
3.3.2 Towards Four Archetypes	29
3.3.3 Data Mesh Components.....	32
3.3.4 Challenges, Limitations and Mitigations.....	34
3.3.5 Data Mesh Structures, Components and Considerations.....	37
3.4 The Shift to Data Mesh	37
3.4.1 Data Mesh Prerequisites	37
3.4.2 Impact of the Data Mesh Transition	39
3.4.3 Other Data Methodologies.....	41

3.4.4	When to and When not to Data Mesh	43
3.5	Data Reference Architectures	43
3.5.1	Data Reference Architecture Characteristics	43
3.5.2	Reference Architecture Parts	46
3.6	Developing a Reference Architecture	46
3.6.1	Goals and Requirements of a Reference Architecture	46
3.6.2	Reference Architecture Design Methodologies	47
3.7	Validating a Reference Architecture	50
4	Artifact Design	51
4.1	Type of Reference Architecture	51
4.2	Reference Architecture Design Strategy	51
4.3	Empirical Acquisition of Data.....	51
4.4	Construction of the Reference Architecture	52
4.4.1	Domain Reference Architecture	52
4.4.2	Self-Serve Data Platform	53
4.4.3	Federated Governance Reference Architecture	53
5	Artifact Validation	54
5.1	Questionnaire	54
5.1.1	Questionnaire Introduction.....	54
5.1.2	Usefulness Assessment.....	54
5.1.3	Quality Assessment	55
5.1.4	Variability Assessment.....	55
5.1.5	Additional Feedback	55
5.2	Questionnaire Distribution.....	55
5.3	Participant Profiles	56
6	Results	59
6.1	Usefulness Assessment Results	59
6.2	Quality Assessment Results.....	61
6.3	Variability Assessment Results.....	63
6.4	Additional Feedback from the Questionnaire.....	64
6.4.1	Comments Usefulness Section.....	64
6.4.2	Comments Quality Section	65
6.4.3	Comments Variability Section	65
6.4.4	Comments Domain Architecture	66
6.4.5	Comments Self-Serve Data Platform Architecture	66
6.4.6	Comments Federated Governance Architecture	66
6.5	Main Questionnaire Takeaways	67
6.5.1	Takeaways Usefulness	67
6.5.2	Takeaways Quality.....	67

6.5.3 Takeaways Variability.....	67
6.5.4 Suggested Improvements to the Model	68
6.5.5 Results Summary	68
7 Conclusion	69
7.1 Data Mesh Structures, Components and Considerations	69
7.2 The Transition to a Data Mesh	70
7.3 Data Reference Architectures	71
7.4 Developing a Reference Architecture	71
7.5 Reference Architecture Validation	72
7.6 Main Research Question.....	72
7.7 Contributions of the Research.....	72
7.7 Limitations of the Research.....	73
7.8 Future Research	73
References.....	74
Appendix	78
A Domain Architecture and Component Explanation	78
B Data Mesh Reference Architecture Evaluation Questionnaire	81
B.1 Questionnaire Introduction.....	81
B.2 Questionnaire Section Usefulness.....	84
B.3 Questionnaire Section Quality	87
B.4 Questionnaire Section Variability	91
B.5 Questionnaire Additional Feedback	94
C Questionnaire Likert Scale Answers Per Respondent	97
C.1 Likert Scale Answers Usefulness Section.....	97
C.2 Likert Scale Answers Quality Section	98
C.3 Likert Scale Answers Variability Section	99

List of Figures

Figure 1 Different Generation Data Platforms adapted from (Zaharia et al., 2021) 11

Figure 2 Separation of Data Planes..... 11

Figure 3 Data Mesh Paradigm.....12

Figure 4 Data Product Attributes (Driessen et al., 2023).....18

Figure 5 Data Mesh Topologies (Strengholt, 2022).....26

Figure 6 Data Mesh Logical Architecture (Dehghani, 2020).....27

Figure 7 Pure Data Mesh, adapted from (Strengholt, 2022).....30

Figure 8 Semi Pure Data Mesh adapted from (Strengholt, 2022)30

Figure 9 Hybrid Data Mesh adapted from (Strengholt, 2022).....31

Figure 10 Distribution Data Mesh adapted from (Strengholt, 2022)31

Figure 11 Factors Blocking Data Mesh Adoption (Hokkanen, 2021)38

Figure 12 Different Generation Data Platforms adapted from (Zaharia et al., 2021).....41

Figure 13 Domain Architecture52

Figure 14 Self-Serve Data Platform Architecture53

Figure 15 Federated Governance Architecture53

Figure 16 Respondent Work Role Distribution56

Figure 17 Respondent Data Mesh Experience57

Figure 18 Respondent Enterprise Architecture Experience.....57

Figure 19 Respondent ArchiMate Experience.....58

Figure 20 Usefulness Responses Bar Charts59

Figure 21 Quality Responses Bar Charts.....61

Figure 22 Variability Responses Bar Charts63

List of Tables

Table 1 Issues wit Current Data Platforms	13
Table 2 Data Mesh Core Principles.....	17
Table 3 Data Mesh Benefits.....	20
Table 4 Enterprise Architecture Layers	21
Table 5 Knowledge Question Search Queries	24
Table 6 Inclusion and Exclusion Criteria	24
Table 7 Initial Search Result	25
Table 8 Reference Summary Table.....	25
Table 9 Data Mesh Archetypes Advantages and Disadvantages	32
Table 10 Data Mesh Main Components and Elements	34
Table 11 Data Mesh Challenges and Mitigations	36
Table 12 Data Platform Comparison	42
Table 13 Reference Architecture Characteristics	45
Table 14 Framework for Reference Architecture Design (Angelov et al., 2012).....	48
Table 15 Respondent Work Role	56
Table 16 Participant Company.....	56
Table 17 Median and Mode Usefulness Section	59
Table 18 Respondent Sentiment Usefulness Section	60
Table 19 Median Per Role Usefulness Section	60
Table 20 Median and Mode Quality Section.....	61
Table 21 Respondent Sentiment Quality Section	62
Table 22 Median Per Role Quality Section.....	62
Table 23 Median and Mode Variability Section.....	63
Table 24 Respondent Sentiment Variability Section	64
Table 25 Median Per Role Variability Section.....	64

List of Abbreviations

Abbreviation	Meaning
BPMN	Business Process Model and Notation
DaaS	Data as a Service
DATSIS	Discoverable, Addressable, Trustworthy, Self-Describing, Interoperable and Secure
DDD	Domain-Driven-Design
EA	Enterprise Architecture
ETL	Extract, Transform, Load
IT	Information Technology
KQ	Knowledge Question
RA	Reference Architecture
SLR	Systematic Literature Review
TOGAF	The Open Group Architecture Framework
UML	Unified Modelling Language

1 Introduction

The first chapter of this thesis will be dedicated to highlighting how the growing data-driven culture imposes new challenges for organizations and current data platforms. A new data paradigm, the data mesh, is proposed as a solution. Next, the scientific relevance of this study is explained and the research goal and questions are introduced. The introduction concludes with an explanation of the research methods used and a description of the thesis structure.

1.1 Rising Interest in Data

Over the last years, organizations have become more and more data-driven. This is resembled by the increasing interest in data analysis capabilities (I. A. Machado, 2022). According to market research by Fortune Business Insights, the global big data analytics market size was valued at 271.83 billion US dollars in 2022 (Fortune Business Insights, 2024) and the market size is expected to keep growing in the coming years. Data-driven decision-making has many benefits, like improved efficiency in decision-making, the ability to make more informed decisions, increased understanding of customer needs, more innovations, better market decisions, and improved business development (Berntsson-Svensson & Taghavianfar, 2020). Therefore, to stay competitive (Hooshmand et al., 2022), organizations have to follow this trend. To achieve this organizations have made considerable investments in data warehouses, data lakes, and analytical platforms. The rising interest in, and collection of data, gives rise to new challenges. Kim and Park (2022), for example, found that organizations are overwhelmed by the sheer volume of data and therefore, using it efficiently becomes challenging. Additionally, Graetsch et al. (2023) found that data intensive solutions provide challenges in resolving quality issues, understanding data, managing access to data, and aligning data with business needs.

1.2 Different Data Platforms

The need to stay competitive (Bode et al., 2023) in today's fast-moving business environment has thus created the need for companies to invest in data storage and processing capabilities to keep up with business requirements. Throughout the years data platform architectures have evolved from data warehousing solutions with some reporting capabilities, to data fabrics (Priebe et al., 2021) incorporating different data sources and making machine learning and AI capabilities possible.

The first generation of data platform architectures, the data warehouse platform (Azeroual. & Nacheva., 2023) (Zaharia et al., 2021), is visualized in Figure 1-a. It consists of structured data from various sources being, by virtue of the ETL (Extract, Transform, Load) processes, stored in data warehouses (Zaharia et al., 2021). The reporting and Business Intelligence dashboards are created based on the data residing in the data warehouses. Problems with first generation data platforms are their stale nature, difficulties with processing semi- and unstructured data, troubling users with incorrect data, and high costs as the volume of data grows. To tackle these problems a new architecture was proposed.

The second generation of architectures are two-tier architectures, as shown in Figure 1-b, combining data lakes (first tier) and data warehouses (second tier). The combination of data warehouse storages and data lake storages makes it possible to store semi- and unstructured data. It also includes the incorporation of data science and machine learning capabilities. Eventually, the two-tier architectures, also started to fail meeting increasing requirements. Challenges of the two-tier architecture are the complexity of implementing data pipelines, the separate ETL process not being able to meet the demand for timely data, and rising costs. Additionally, it requires separate management of the data warehouse and data lake storages.

To be able to combine the benefits of data warehouses and data lakes without needing separate storage capabilities new architectural solutions were proposed. For example, the data lakehouse, which is shown in Figure 1-c.

The data lakehouse is defined by Zaharia et al. (2021) as a data management approach that allows for the low-cost storage of raw data while simultaneously allowing for data warehouse capabilities by providing structure and schemas by virtue of a semantic and indexing layer.

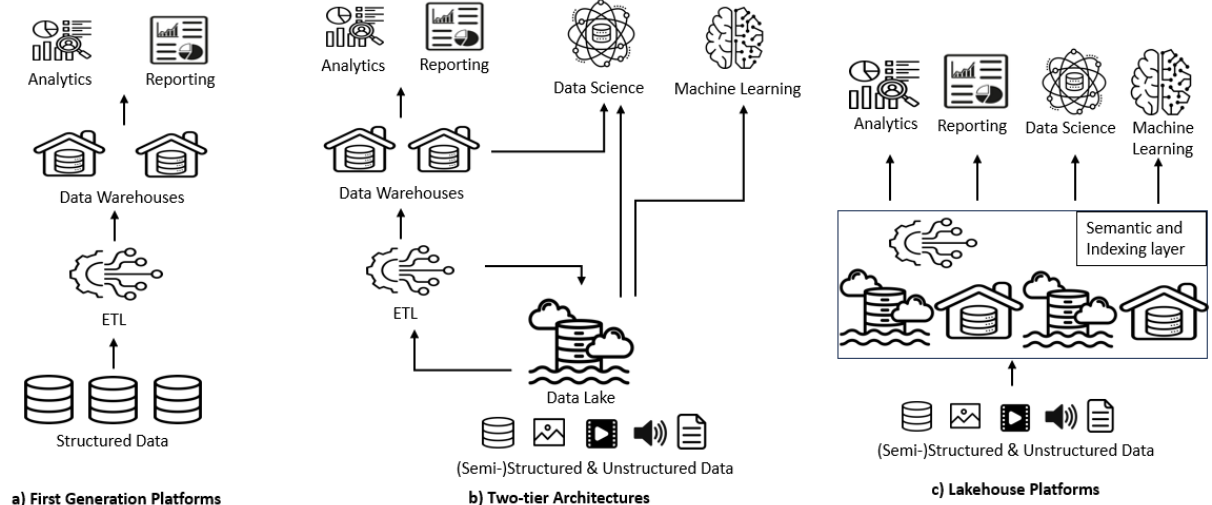


Figure 1 Different Generation Data Platforms adapted from (Zaharia et al., 2021)

With the IT landscape of organizations growing, and the increasing volumes of data to be processed, problems arise. Data lakes are overflowing with data in all kinds of different formats and are becoming data swamps (Li et al., 2022). The different data platforms presented above are all monolithic data structures, centralized and managed by central data teams. These central data teams are becoming the bottleneck for analytics (Hendriks, 2023) in today’s fast-changing business environments and technological landscapes. The central data teams are unable to keep up with the increasing volumes of data to be processed, and the increasing demands for analytical purposes (Vestues et al., 2022) (I. A. Machado, 2022). More effort is spent on data cleaning and discovery than on creating value from data. Additionally, the domain-agnostic data team lacks the in-depth domain knowledge to serve the specific needs of different business units or to quickly understand the data. The monolithic structures give rise to multiple other challenges for bigger organizations. The centralized management and storage create a lack of ownership by the domain teams resulting in lower quality data (Hendriks, 2023). This also creates a problem for data consumers who do not know who to turn to when problems arise. Other issues are that the monolithic structures lack scalability (Bode et al., 2023), and that increasing capacity or enhancing performance often requires upgrading the entire architecture making it costly solutions (Vlasiuk & Onyshchenko, 2023). The centralized architecture also increases time to market and makes it hard to access the data of other business domains.

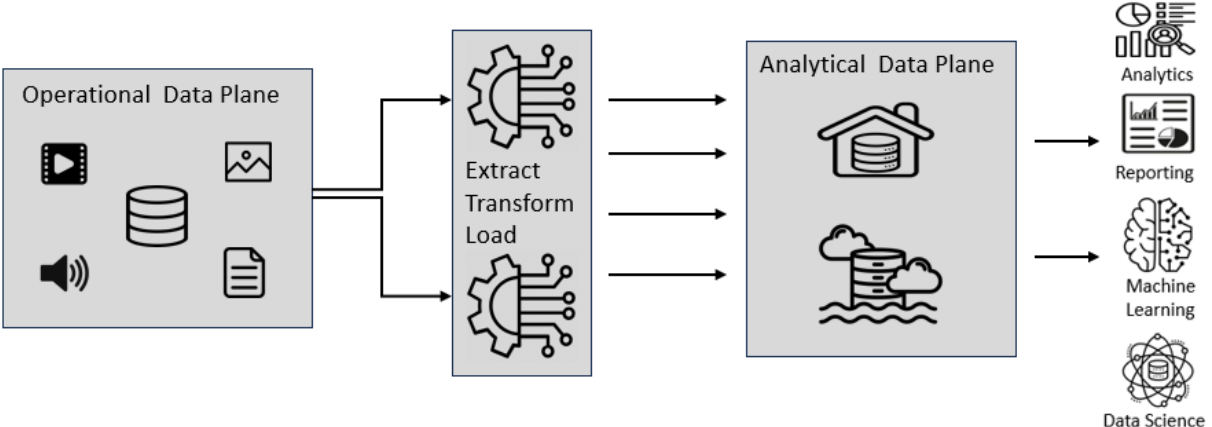


Figure 2 Separation of Data Planes

Figure 2 visualizes the problem of monolithic architectures. The ETL pipelines have to be maintained by a central data team which is isolated from the operational data plane and the analytical data plane. This separation of duties also shows the reason for the lack of ownership. The operational team and analytical team are not in direct contact with each other and the operational teams are not focused on the ETL process of their data. The ETL process is mainly performed by data engineers. The employees operating in the analytical plane on the other hand, are concerned with creating dashboards and visualizations or implementing data science capabilities. This three way split of responsibilities ultimately leads to the creation of siloed data architectures.

To briefly summarize the following problems arise in monolithic data architectures: 1) Central data teams have too many responsibilities and lack domain knowledge of the data they work with; 2) there is a low level of ownership regarding data; 3) a lot of time is wasted on collecting, cleaning and preparing data; 4) there is a lack of scalability; 5) architectures become siloed; 6) separation of data planes; 7) it is hard to access data from other domains and 8) data lead times increase.

1.3 Data Mesh Paradigm

A paradigm shift in data was needed and thus, in 2019, Dehghani (2019) proposed a new way of structuring data architectures, the 'Data Mesh' concept. The data mesh paradigm shift 'is in the convergence of Distributed Domain Driven Architecture, Self-serve Platform Design, and Product Thinking with Data' (Dehghani, 2019). Data mesh is a new way of structuring the enterprise data architecture by decentralizing the data capabilities and positioning those at a domain level, with an overarching governance framework, supported by a self-serve data platform. It follows from 4 core principles set out by Dehghani (2020): 1) data as a product, 2) domain-oriented decentralized data ownership and architecture, 3) a self-serve data infrastructure as a platform, and 4) federated computational governance. The goal is to shift from a rigid and siloed architecture, to a distributed architecture focused on domain ownership and scalability. By decentralizing, data mesh attempts to get rid of the bottleneck the central data platform and the central data team can become. Data is not residing on a central platform anymore but within the domains itself. Resulting data products, self-contained data sets with a business purpose, are exchanged between domains directly and can be acted upon without any help from the domain that created the data product. This improves collaboration and reduces data lead times.

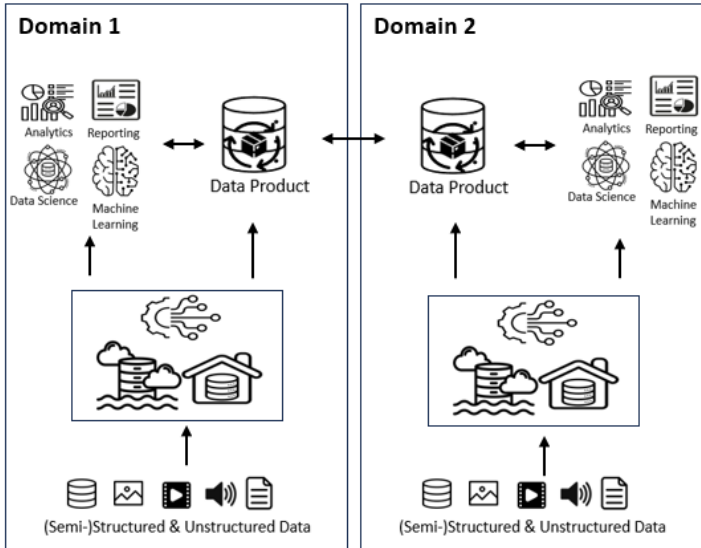


Figure 3 Data Mesh Paradigm

Another benefit of adopting data mesh is that it allows for better scalability than a centralized data architecture. By thinking about data as a product and making domains responsible for the data produced in their domain, the quality of data can be assured and stronger ownership is established. Data as a product in a data mesh means that data sets are readily available to provide business value. The self-serve platform on which a data catalog and other capabilities are provided facilitates collaboration and efficiency. Altogether, the idea of data mesh is to improve data sharing across the organization to improve data-driven decision-making.

Data mesh is not a one-size-fits-all approach and there are some limitations and challenges related to the concept. Implementing data mesh is a complex task. It requires organizational changes, management changes, and restructuring your data landscape. Data mesh can be seen as a cultural shift that significantly affects the way of working. Good data governance is an important part of making sure a data mesh functions properly. Standards and policies have to be defined to make sure data quality and consistency can be maintained. A data mesh may require investments in tooling and infrastructure. If data mesh is not implemented properly it can lead to more siloing and duplication of effort which is counterintuitive to what data mesh tries to achieve.

In short, data mesh attempts to tackle the shortcomings of the monolithic data platforms. Table 1 outlines the shortcomings of current data platforms, for organizations processing high volumes of data and requiring timely data-driven insights, and how data mesh tackles these problems.

Platform	Shortcomings	Data Mesh
<i>Data Warehouse Platform</i>	Inability to work with semi- and unstructured data. Limited advanced analysis capabilities.	Data mesh transfers responsibility to the domains alleviating the burden on the central data team.
<i>Two-Tier Architecture</i>	Difficulty of managing pipelines and separate ETL processes. Scaling costs.	Data mesh for domain level resource and cost allocation. Scaling can be realized at domain level instead of centrally.
<i>Data Lakehouse Platform</i>	Data lakes becoming data swamps. Central data team becoming a bottleneck.	Management of pipelines and ETL is performed on domain level.

Table 1 Issues wit Current Data Platforms

1.4 Scientific Relevance

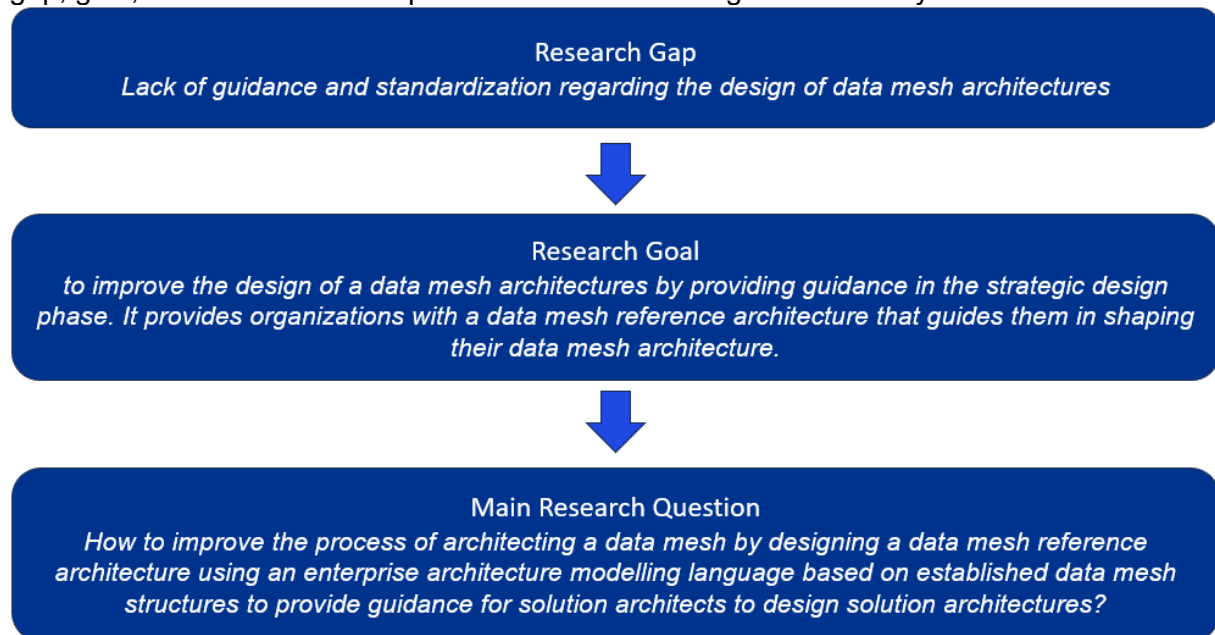
Being a relatively new concept data mesh benefits from additional research. While quite some research has been performed on the data mesh concept and on practical examples of data mesh implementations, Machado (2022) does stress the need for general models or methods related to the implementation of data mesh. There is a lack of guidance regarding the architectural design of a data mesh. This allows organizations to structure a data mesh freely according to their needs while simultaneously making it hard for other organizations and data architects to determine where and how to start designing their data mesh. Therefore, organizations often start with partial implementations not adopting all principles entirely (Lombardo, 2023). According to Bode et al. (2023), data mesh research would also benefit from research into data mesh archetypes as research lacks guidelines on how to determine the right strategy and architecture suitable for an organization. Without a good strategy fitting the organization, it can be difficult to make the implementation successful and realize the benefits. Machado et al. (2022) mentions that data mesh research would benefit from more concrete steps guiding the design and implementation of a data mesh. Goedegebuure et al. (2023) identified the challenge of standardizing data mesh in a multitude of examined studies and propose creating a data mesh reference architecture as a solution. Dončević et al. (2022) also mentioned the lack of standards regarding data mesh as future research direction which was confirmed by Wider et al. (2023).

The identified research gap is a lack of standardization and guidance regarding the design of data mesh architectures. Therefore, this research tries to fill a gap in research by *proposing 4 data mesh archetypes with different levels of maturity and by developing a data mesh reference architecture to make it easier for organizations to translate a data mesh design into an actual solution architecture.*

The reference architecture (RA), a template that serves as a blueprint to facilitate the creation of concrete or solution architectures (Sang et al., 2016) (Sang et al., 2017), will be modelled using an Enterprise Architecture (EA) modelling language. Enterprise architecture has proven its value in creating strategic alignment (Niemi, 2008) between the business objectives and the supporting IT infrastructure. Since a data mesh requires a cultural shift in organizations and does not only affect the data architecture of an organization, using an EA approach is a good fit. Enterprise Architecture (EA) provides a holistic view of the organization (Niemi, 2008) and includes different architectural layers of the organization including the data layer. Therefore it can capture how a data mesh architecture influences the whole internal business environment.

1.5 Research Goal and Questions

Based on the need for the data mesh paradigm and the lack of standard models regarding the design of mesh architectures, as presented in the previous sections, the following research gap, goal, and main research question were defined to guide this study:



The following knowledge questions (KQ) are defined to guide the literature review:

- KQ 1: What are the key components constituting a data mesh and what are the limitations?
 - KQ 1-a: What different kinds of data mesh archetypes exists?
 - KQ 1-b: What are common components of a data mesh?
 - KQ 1-c: What are the limitations of data mesh?
- KQ 2: Which factors determine if data mesh is a valid approach for an organization?
 - KQ 2-a: What are the main indicators to consider the switch to a data mesh?
 - KQ 2-b: What is the impact of data mesh on the existing architecture?
 - KQ 2-c: Which other data methodologies are there?
- KQ 3: Are there existing data mesh reference architectures?
 - KQ 3-a: What are characteristics of data reference architectures?
 - KQ 3-b: What parts of other data reference architectures can be re-used?

The following questions have been defined to understand more about the nature of the artifact to be designed:

- KQ 4: How to develop a reference architecture?
 - KQ 4-a: What are the goals and requirements of a reference architecture?
 - KQ 4-b: Which method can be used to design and develop the reference architecture?
- KQ 5: How can a reference architecture be validated?

1.6 Research Methodology

This study will make use of multiple research methods. The main research methodology used in this study is the Design Science Research Methodology by Wieringa (2014), and in particular, this study will follow the design cycle as proposed by Wieringa (2014). According to Wieringa Design science is: *“the design and investigation of artifacts in context”* (Wieringa, 2014). This means that in this type of research, an attempt is made to design an artifact to improve a problem context and accomplish the goals of stakeholders. Wieringa (2014) created a template to structure a design problem:

- Improve a problem context
 - By treating it with an artifact
 - That satisfies predefined requirements
 - In order to achieve stakeholder goals

For the purpose of this study and based on the main research objective the following design problem is created:

- Improve the process of architecting a data mesh
- By designing a data mesh reference architecture
- Based on established data mesh architectures and modelled in the ArchiMate Enterprise Architecture Modelling language
- In order guide architects to build and evaluate solution architectures

To support the research process additional research methods were used during different steps of the study. To build a theoretical foundation of the main concepts in this study, ‘Data Mesh’ and ‘Enterprise Architecture’ a quick literature scan was performed. During this initial exploratory review of (grey) literature, a starting point for further research was established.

To answer the knowledge questions as proposed in Section 1.5 the systematic literature review methodology by Carrera-Rivera et al. (2022) was used. Before designing and validating a treatment to attempt to solve a problem it is important to understand the problem context. To understand the problem context and the possible solutions this study starts with a Systematic Literature Review (SLR) (Carrera-Rivera et al., 2022) before going into the design phase. A SLR is divided into four main phases. The first phase is the planning phase in which the databases to search in are identified, keywords for search queries are selected and inclusion and exclusion criteria are defined. In the selection phase, the queries are executed and the inclusion and exclusion criteria are applied. The next phase, the extraction phase, is focussed on extracting data from the selected sources. The last phase, the interpretation phase, is where the extracted information gets analysed to identify common themes and establish an understanding of the problem context.

To design the treatment, the method proposed by Galster and Avgeriou (2011) was used because it provides a comprehensive method used in multiple studies to develop data reference architectures. The study proposes a 6 step approach to create empirically sound reference architectures. The method is explained in more detail in section 3.6.2.

Following the design of the treatment the artifact had to be validated. According to Wieringa (2014), the goal of treatment validation is to *“justify that the treatment will contribute to stakeholder goals when implemented in the problem context”* (Wieringa, 2014). The treatment was validated by virtue of a questionnaire. Questionnaires are a method used to gather expert opinions on the treatment by letting experts in the field think about how the treatment would interact with the problem context.

This study finishes with the treatment validation and omits the treatment implementation as this is outside of the scope of this study.

1.7 Thesis Structure

The thesis will be structured as follows. In section 2, the theoretical background, and the main concepts of this study, 'Data Mesh' and 'Enterprise Architecture' will be explained. The following section, section 3, will be a Systematic Literature Review that serves the purpose of answering the earlier introduced knowledge questions. Section 4, following the literature review, will be dedicated to the design of the treatment; the Data Mesh Reference Architecture. Section 5 will be the treatment validation, in which the designed artifact is validated by virtue of a questionnaire. The report will be finished with a discussion of the results, a conclusion, and recommendations for future research.

2 Theoretical Background

To create a common understanding of the main concepts involved in this study, the research goal, and the research questions, this chapter is aimed at providing general theoretical knowledge on the main concepts. First, the leading concept, Data Mesh, will be explained. Following this, the other main concepts, Enterprise Architecture, ArchiMate, and reference architectures, will be introduced.

2.1 Data Mesh

As briefly mentioned in the introduction data mesh is a new way of thinking about the enterprise data architecture. It is deemed to be a paradigm shift necessary to deal with the challenges encountered by large data-driven organizations. The concept can be classified as a domain-oriented decentralized architecture to manage data organization-wide (Jonkman, 2023).

The data mesh idea is loosely based on the microservice architecture used in software engineering (Ashraf et al., 2023). Data platforms are still mainly centralized and the data mesh paradigm can be seen as the microservice paradigm in the data architecture sphere. Each domain in a data mesh can be viewed as a service in the microservice architecture which collectively make up the application, or in case of a data mesh, the data architecture. Just like data mesh is a reaction to problems of monolithic data structures the microservice paradigm shift was a reaction to problems caused by monolithic application architectures. The decentralization and strong ownership at the domain level make the data mesh and microservices architectures more scalable and adaptable to meet fast changing business requirements.

Data mesh is based on 4 core principles originally set out by Dehghani (2020). The 4 core principles are: 1) to treat data as a product, 2) domain-oriented decentralized data ownership and architecture, 3) a self-serve data infrastructure as a platform, and 4) federated computational governance. The core principles are explained in more detail in Table 2 below.

Principle	Explanation
Data as a product	The data of a domain is treated as a product produced to serve a consumer. The product must be of high quality to be valuable for data consumers of other domains.
Domain-oriented decentralized data ownership	In a data mesh the data products are owned by the domain that produces it or the people closest to it. Each domain bears the responsibility for its own data and the quality of the data.
Self-serve data infrastructure as a platform	A key concept to differentiate data mesh from other approaches is the stimulation to build self-serve data platforms. On these platforms the data infrastructure and tools are made available as a service for the domain owners and users.
Federated computational governance	Data mesh makes use of a distributed governance model in which the data products are owned by the domains and interoperability is enforced by standardizations and policies which are organization-wide.

Table 2 Data Mesh Core Principles

To satisfy the core principles of a data mesh, a company has to go through a cultural as well as an organizational change (Driessen et al., 2023) (Vestues et al., 2022). Because of the architectural reorganization needed to decentralize the data architecture, responsibilities will shift. Teams need to start thinking about data products as products they own and provide to the other domains in the business. A culture of accountability and collaboration has to be established to make a data mesh effective. The way of thinking about data and managing data has to change accordingly.

2.1.1 Data as a Product

To think about data as a product closely resembles the concept of Data as a Service (DaaS). The idea of data as a product is to make data objects available that are ready to be used for different purposes by others. According to the original article by Dehghani (2019), data products have to adhere to 6 basic attributes. Data products have to be Discoverable, Addressable, Trustworthy, Self-describing, Interoperable and Secure (DATSIS). Driessen et al. (2023) went a step further by expanding these attributes and proposing 9 attributes for usability. Driessen et al. (2023) proposed to add Understandable, Valuable and Feedback-driven as additional attributes of a data product as shown in Figure 4.

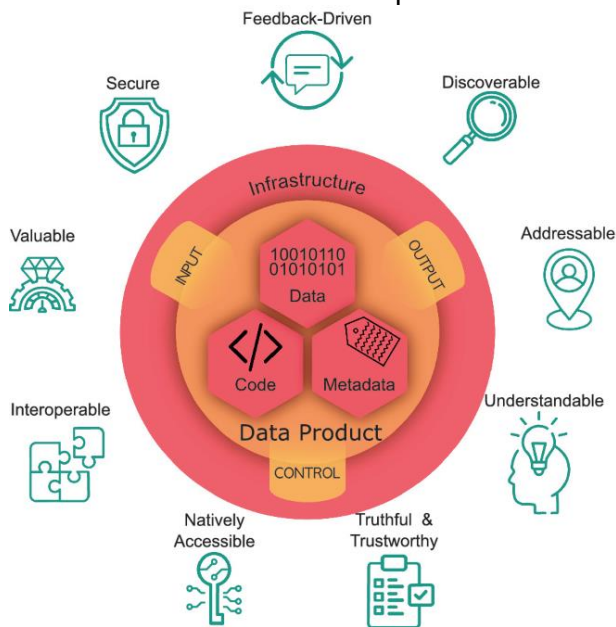


Figure 4 Data Product Attributes (Driessen et al., 2023)

The attributes of data products in a data mesh make sure all necessary information to understand and use the data, is present in the data product itself. The data products are accessible and self-explanatory while simultaneously quality is enforced.

Data products are thus self-explanatory entities which often consist of cleaned and transformed operational data including metadata, the origin history, and the necessary semantics to serve the needs of data consumers. Additionally, data products may include analytical data or be combined with other data products. Lastly, data products must be of value for the business.

2.1.2 Domain in a Data Mesh

The second, of the four main principles of a data mesh, is domain-oriented decentralized data ownership. The boundaries of a domain, are dependent on how an organization designs its data mesh (Vinnikainen, 2023). A domain can be a single team or business unit within the organization or it can be a department within the organization. When a company consists of multiple offices it could also be that one office is a domain, however usually a domain is aligned with a business capability or unit. The domain is the entity that is responsible for the ownership and publication of data products. The exact boundaries of a domain therefore vary based on how a data mesh has been constructed. Machado et al. (2021) propose a domain model in a situation in which the data mesh domains are aligned with business domains. According to Hooshmand et al. (2022) data mesh applies Domain-Driven-Design (DDD) to the data space. In DDD development is performed according to business processes and rules related to domain capabilities (Hooshmand et al., 2022). Domains can be divided into sub-domains allowing each domain to independently publish functionalities and services (Hooshmand et al., 2022), and in terms of a data mesh, data products. The main goal of applying DDD in data mesh is to align data products being developed in alignment with business requirements and objectives (Vinnikainen, 2023).

2.1.3 Self-Serve Platform

The self-serve platform within a data mesh corresponds to a platform providing capabilities that allows the autonomous domains to develop and maintain data products. These capabilities are for example computational resources, storage capabilities, connectivity services, and security related resources. It is important to note that the self-service platform is not a common distribution layer. The self-service platform's main purpose is to provide technologies to improve the efficiency of domain teams (Panigrahy et al., 2023). Typically the self-serve platform is managed by a central team whose responsibility is supporting the domains with infrastructure, tools, and other technologies to decrease their workload.

2.1.4 Federated Computational Governance

The decentralization of the creation and ownership of data products requires the need for federative components and agreements among the participating domains. Enforceable protocols are needed about the semantics, standards, policies, formatting of data, and means of communication to make sure the data mesh functions as intended. The purpose of federated governance is to institute governance at a centralized level while the domains have the responsibility and autonomy to apply these principles in a way that suits their needs (Panigrahy et al., 2023). The federated governance includes, for example, policies on the documentation of data products, policies to allow secure access to data products, and communication standards to ensure interoperability. Typically in a data mesh the federated governance is organized by a team consisting of members of all participating domains.

2.1.5 Data Mesh Benefits and Challenges

The table below summarizes some of the benefits of a Data Mesh compared to centralized data platforms found in literature.

Benefit	Elaboration
Good scalability <i>(Vinnikainen, 2023)</i> <i>(Kancharla & Madhu Kumar, 2023)</i> <i>(Pongpech, 2023)</i> <i>(Li et al., 2022)</i> <i>(Dahdal et al., 2023)</i> <i>(Dibouliya & Jotwani, 2023)</i>	Data mesh improves scalability because new domains can easily be added by following the policies and standards set by the federated governance layer and utilizing the tools and infrastructure from the self-serve platform.
Stronger data ownership <i>(Butte & Butte, 2022)</i> <i>(Bode et al., 2023)</i> <i>(Dibouliya & Jotwani, 2023)</i> <i>(Pakrashi et al., 2023)</i>	By designating the employees close to the origin of the data with the responsibility to provide the data as a product strong accountability is established.
Increased data quality <i>(Kancharla & Madhu Kumar, 2023)</i> <i>(Vestues et al., 2022)</i> <i>(Dibouliya & Jotwani, 2023)</i> <i>(Goedegebuure et al., 2023)</i>	This follows from the former benefit because the strong accountability and thinking about the data as a product improves the data quality. Domains have to adhere to certain quality standards.
Reduced data lead time <i>(Hendriks, 2023)</i> <i>(Kancharla & Madhu Kumar, 2023)</i> <i>(Panigrahy et al., 2023)</i>	Because data producers and data consumers can independently provision and use data through the self-service model the lead time of data throughout the organization reduces.
Better data governance <i>(Dahdal et al., 2023)</i> <i>(Pakrashi et al., 2023)</i> <i>(Hokkanen, 2021)</i>	Data governance is improved by the decentralized approach as day-to-day governance and responsibilities are transferred to the domains producing the data. The company wide governance is reduced to setting standards and policies.

<p>Domain expertise (Dibouliya & Jotwani, 2023) (Pakrashi et al., 2023) (Bode et al., 2023)</p>	<p>Domain experts make data-related decisions within their domains which leads to more relevant and informed data solutions.</p>
<p>Tackle data silos (Bode et al., 2023) (Kancharla & Madhu Kumar, 2023) (Dibouliya & Jotwani, 2023)</p>	<p>Data mesh promotes the interoperability between different business units and tries to break down data siloes.</p>
<p>Improved collaboration (Ashraf et al., 2023) (Falconi & Plebani, 2023) (Dahdal et al., 2023) (Dibouliya & Jotwani, 2023) (Sedlak et al., 2023)</p>	<p>Data mesh improves data sharing within the organization by promoting collaboration between business domains or between organizations by promoting data exchange.</p>
<p>Flexibility in technology stack (Araújo Machado et al., 2022) (Jonkman, 2023)</p>	<p>Domain teams have the flexibility to choose the technology that suits their needs best, and use tools and infrastructure form the self-service platform.</p>
<p>Cost efficiency (Ashraf et al., 2023) (Dibouliya & Jotwani, 2023) (Jonkman, 2023)</p>	<p>Data processing and storage capabilities can more easily be tailored to the requirements of different domains improving resource utilization and allowing for better cost management. Data that is not useful can be discarded or archived earlier in the data flow reducing storage and processing needs.</p>

Table 3 Data Mesh Benefits

As with any concept, the implementation of data mesh also comes with certain challenges:

- Transitioning to a data mesh is challenging because it requires restructuring of the organization's technical landscape and requires a shift in the way of working with, and thinking about data, within the organization (Bode et al., 2023) (Goedegebuure et al., 2023).
- The domain-oriented nature of data mesh requires teams to operate independently. This requires teams to have the necessary skills and capabilities to make this work (Hendriks, 2023) (I. A. Machado et al., 2022)
- Good governance is needed to make a data mesh effective. Establishing standards, protocols, and policies is essential to ensure interoperability (Krystek et al., 2023) (Sedlak et al., 2023).
- The decentralized data ownership creates challenges for the discoverability and accessibility of data. Implementing data catalogs, effective metadata management, common vocabulary, and coding conventions are vital to making a data mesh function properly (Vestues et al., 2022).
- As with any data related solution data security and compliance must be accounted for when sharing data between different departments or different companies (Podlesny et al., 2022).
- Where cost efficiency is a benefit of implementing data mesh, costs can also become an issue when effort is replicated and tools and technologies are duplicated (Falconi & Plebani, 2023).

Lastly a big challenge is that 'Data Mesh' is an abstract concept that lacks standard methods and models. It is not a tool you can add to your IT landscape but rather a methodology requiring a shift in how an organization treats data and how data is managed (Bode et al., 2023) (Hokkanen, 2021). It can therefore be challenging for organizations to determine how and where to start with their transition to a data mesh architecture.

Data mesh is a concept that can be used to transform the internal landscape of an organization and is often looked at from the perspective of a single organization (Falconi & Plebani, 2023). Another good application however, is to use data mesh structures for industry-wide application like the CowMesh (Pakrashi et al., 2023) and clinical trials example (Falconi & Plebani, 2023) show. By setting up these cross organizational data meshes, the participating organizations keep ownership of their data while simultaneously collaboration and sharing of necessary information is improved by having structures and policies in place to easily and securely share data among participants.

2.2 Enterprise Architecture

Enterprise Architecture has been a widely researched domain. Enterprise architecture provides a holistic view of the organization’s information technology (IT) infrastructure, application landscape, and data architecture (Niemi, 2008). The main goal of EA is to aid organizations in achieving their business targets by aligning their IT strategy with their business strategy. The enterprise architecture can be made visible by using a modelling language and can be built up out of different layers which gradually provide more details about the structure of the organization. EA can provide a framework of the current state of the organization and can be used to develop a blueprint for, and plan transitions to desired future states. Because a data mesh requires a shift from a monolithic to a decentralized architecture, an EA approach could help in guiding this transition.

According to the TOGAF standard, there are four architecture domains (TheOpenGroup, n.d.) which are commonly distinguished as the different layers that together make up the Enterprise Architecture as a whole. The four layers are the business, data, application, and technology architecture layers. Table 4 below explains the different architectural layers (Hermawan & Sumitra, 2019).

EA Layer	Explanation
Business Architecture	This layer is focused on the strategy, policies, business processes and business capabilities of the organization. It describes how the organization will operate to accomplish their business objectives.
Data Architecture	The data architecture is concerned with the establishing how data is collected, stored, processed and shared within the organisation. Additionally, standards for the quality, security and accessibility are determined.
Application Architecture	The goal of this layer is to determine which applications are needed to support the business processes and it specifies the interactions between applications.
Technology Architecture	This layer is focussed on managing the actual technology infrastructure, including hardware, software and networks among others. A well structured technology architecture enables scalability and adaptability to business needs.

Table 4 Enterprise Architecture Layers

2.2.1 Benefits and Challenges of Enterprise Architecture

Niemi (2008), investigated the benefits of Enterprise Architecture most widely mentioned in literature at the time in a paper on Enterprise Architecture benefits. Providing a holistic view of the organization, EA ensures improved alignment between business objectives and IT capabilities. This in turn improves decision-making processes and facilitates more effective change management. Furthermore, EA enhances risk management by making it easier to identify potential vulnerabilities. Enterprise Architecture also helps generate insight into business processes and optimizing them. Ultimately, EA fosters improved strategic agility allowing organizations to better navigate the complexity of the modern business landscape.

Designing and maintaining an Enterprise Architecture comes with challenges. LeanIX (Aldea, 2023) identified current challenges in the field of EA. The growing complexity of organizations and systems makes it difficult to model and maintain the EA. Organizations furthermore, lack proper EA practices and rely on fragmented information. This reduces the value EA can bring to an organization. Additionally, a lack of alignment between EA practice and business results in business units operating in silos, limited transparency, and decision-making based on narrow views. Despite these challenges EA is of value for organizations, aligning business processes and IT infrastructure with organizational goals, supporting decision-making, and facilitating integration of new technologies (Dela Cruz et al., 2011).

2.2.2 ArchiMate

The ArchiMate Enterprise Architecture modelling language, a standard by The Open Group (TheOpenGroup, n.d.) provides EA architects with components to describe and visualize the structure of different business layers and their relationships. ArchiMate has a wide range of elements divided into multiple layers. It provides elements to model business, application, and technology behaviour and processes. Furthermore, it provides an implementation and migration layer to map architecture transformations and a motivation aspect to model the motivation behind architectural changes.

The benefits of using the ArchiMate modelling language are that it provides a robust framework to visualize the architecture of the whole enterprise (Sanyoto & Saputra, 2023). It aids in gaining alignment between the business and the supporting IT infrastructure because it visualizes relationships between different domains. It is also a flexible tool that is widely used making it easier to communicate to others (Sanyoto & Saputra, 2023). It is mainly used to model higher-level processes which makes it less suitable for solution architectures. Its value is also in working alongside other modelling languages like BPMN and UML (Sanyoto & Saputra, 2023).

2.3 Reference Architectures

A reference architecture (RA) is a template that serves as a blueprint to facilitate the creation of concrete or solution architectures (Sang et al., 2016) (Sang et al., 2017). A reference architecture typically includes a common vocabulary, industry standards and best practices, standard architecture components, patterns, and architecture principles.

A reference architecture serves multiple purposes. It can be used as a guideline for designing solution architectures and it can help to standardize approaches within an industry or organization. It can also help architects and developers by encapsulating best practices and accelerating the design process. Next, it aids in maintaining consistency in architecture design. Lastly, it facilitates technology evaluation and provides organizations with a way to make informed decisions about which technologies suit their strategic needs.

The concepts discussed in this chapter provide knowledge of the main concepts involved in the research goal and research questions, and form a theoretical foundation for the following parts of this research to build upon.

3 Systematic Literature Review

This chapter is dedicated to answering the knowledge questions introduced in section 1.5, by performing a systematic literature review (SLR) (Carrera-Rivera et al., 2022). First, the databases to be explored will be determined and the search queries guiding the SLR will be defined. Inclusion and exclusion criteria will be established and the search results will be evaluated. After the relevant literature has been selected, the knowledge questions, as defined in section 1.5, will be answered.

3.1 SLR Planning

The first phase of a SLR is the planning phase. This phase is of high importance because it determines the quality and validity of the SLR. In this phase the databases to be used are identified and the search queries will be created. For this study the choice was made to use Scopus, IEEE, and Google Scholar. The choice was made to use two scientific databases, Scopus and IEEE, with Google Scholar as complementary database also including some non peer reviewed and non scientifically published resources to supplement the results found in the scientific databases. IEEE produces 83 results when searching for the keyword phrase “Data Mesh” and Scopus 111. Searching for “Data Mesh” in google scholar yields 3800 results thus Google Scholar results will be used to supplement the review. A cut of point will be determined for the number of articles to be evaluated for incorporation into this study. Lastly, some grey literature on ‘Data Mesh’, found during an exploratory literature review will be included because some widely referenced works, and the first notion of data mesh (Dehghani, 2019), are grey literature articles. This was also the reason for (Goedegebuure et al., 2023) to perform a systematic gray literature to investigate non-scientific literature.

3.1.1 Search Queries

All search queries including data mesh will search for the keyword phrase ‘Data Mesh’ because searching for ‘Data’ And ‘Mesh’ may yield irrelevant results to this study. For the search queries involving reference architecture the keyword phrase ‘Reference Architecture’ will be used fully written out because abbreviations or searching for ‘Reference’ And ‘Architecture’ may generate irrelevant results.

Table 6 shows the search queries identified to attempt to answer each of the knowledge questions:

Knowledge Questions	Search Queries
KQ 1: What are the key components constituting a data mesh and what are the limitations?	‘Data mesh’ AND (‘Archetype(s)’ OR ‘Topology’ OR Topologies’)
KQ 1-a: What different kinds of data mesh archetypes exists?	‘Data mesh’ AND (‘Structure(s)’ OR ‘Architecture(s)’ OR ‘Components’)
KQ 1-b: What are common components of a data mesh?	‘Data mesh’ AND (‘Limitations’ or ‘Challenges’)
KQ 1-c: What are the limitations of data mesh?	
KQ 2: Which factors determine if data mesh is a valid approach for an organization?	‘Data Mesh’ AND (‘Maturity’ OR ‘Adoption’ OR ‘Implementation’)
KQ 2-a: What are the main indicators to consider the switch to a data mesh?	‘Data Mesh’ AND ‘Impact’
KQ 2-b: What is the impact of data mesh on the existing architecture?	‘Data Mesh’ AND (‘Data Warehouse’ OR ‘Data Lake’ OR ‘Data Lakehouse’ OR ‘Data Fabric’)
KQ 2-c: Which other data methodologies are there?	

KQ 3: Are there existing data mesh reference architectures?	'Data Mesh' AND 'Reference Architecture'
KQ 3-a: What are characteristics of data reference architectures? KQ 3-b: What parts of other data reference architectures can be re-used?	'Data' AND 'Reference Architecture'
KQ 4: How to develop a reference architecture? KQ 4-a: What are the goals and requirements of a reference architecture? KQ 4-b: Which method can be used to design and develop the reference architecture?	Literature for this question will be gathered from the sources examined for KQ 3
KQ 5: How can a reference architecture be validated?	Literature for this question will be gathered from the studies examined for KQ 3 and KQ 4

Table 5 Knowledge Question Search Queries

The results of the executed queries were sorted on relevance to get the best matches first.

3.1.2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria serve the purpose of setting boundaries for the literature study and determining which articles will be included in the review.

Inclusion Criteria	Exclusion Criteria
<ul style="list-style-type: none"> • Peer-reviewed articles • English articles • Articles with detailed publication metadata • Open Access 	<ul style="list-style-type: none"> • Incomplete or poorly written articles • Duplicate articles • Articles with no relation to the research question • Search results Nr 21+

Table 6 Inclusion and Exclusion Criteria

An additional criterium was applied to the search results as some queries returned high amounts of records and thus reduction of the number of articles to include in the review was necessary to keep it manageable. Therefore, only the first 20 search results returned, sorted on relevance, were evaluated for the literature review. Incomplete articles, for example articles referencing figures which are not present in the article, or articles that mention to be still under review were not included in this study. To judge if an article is written poorly, grammatical errors, tone of voice, formatting and punctuation were taken into consideration among others.

3.2 Literature Search and Selection

The result of the selection process is a set of relevant articles based on the search queries and criteria defined in the planning phase. The first step of the selection process is to execute the search queries in the selected databases. Following the query execution, the inclusion and exclusion criteria will be applied. The duplicate search results and studies with no relevance to the research question will be excluded. The remaining studies will be assessed to determine their quality.

Table 8 summarises the initial outcome of performing each of the queries in the designated databases.

KQ	Query	Database – Nr of Results
1	'Data mesh' AND ('Archetype(s)' OR 'Topology' OR 'Topologies')	<ul style="list-style-type: none"> • Scopus – 1 • IEEE – 32 • Google Scholar – 276
	'Data mesh' And ('Structure(s)' OR 'Architecture(s)' OR 'Components')	<ul style="list-style-type: none"> • Scopus – 22 • IEEE – 36 • Google Scholar – 1080
	'Data mesh' AND ('Limitations' or 'Challenges')	<ul style="list-style-type: none"> • Scopus – 19 • IEEE – 8 • Google Scholar – 615
2	'Data mesh' AND ('maturity' OR 'adoption' OR 'implementation')	<ul style="list-style-type: none"> • Scopus – 13 • IEEE – 9 • Google Scholar – 769
	'Data mesh' AND 'impact'	<ul style="list-style-type: none"> • Scopus – 5 • IEEE – 2 • Google Scholar – 859
	'Data mesh' AND ('data warehouse' OR 'data lake' OR 'data lakehouse' OR 'data fabric')	<ul style="list-style-type: none"> • Scopus – 13 • IEEE – 7 • Google Scholar – 323
3	'Data Mesh' AND 'Reference Architecture'	<ul style="list-style-type: none"> • Scopus – 0 • IEEE – 1 • Google Scholar – 1
	'Data' AND 'Reference Architecture'	<ul style="list-style-type: none"> • Scopus – 1455 • IEEE – 1289 • Google Scholar – 79000 +/-

Table 7 Initial Search Result

After the execution of the queries on the database the inclusion and exclusion criteria were applied to narrow down the number of articles to consider for further analysis. After reviewing the abstract and conclusion of the remaining articles on relevance to the study, 61 articles remained. Table 9 shows the number of references remaining after the selection per knowledge question.

Knowledge Question	References
KQ 1	33
KQ 2	8 without duplicates (24 with duplicates from KQ 1)
KQ 3	20
Total	61

Table 8 Reference Summary Table

The following phases of the SLR are the extraction of information from the selected articles and the evaluation of the literature. The following sections will be dedicated to answering the knowledge questions by virtue of the selected articles in this section.

3.3 Data Mesh Structures, Components and Limitations

Even though the data mesh concept allows organisations freedom in how to structure their data mesh there are some common and key components which are required to create an effective data mesh. Additionally, a data mesh is not a one size fits all approach solving all of an organizations problems, there are limitations. This section will answer the first knowledge question ‘*What are the key components constituting a data mesh and what are the limitations?*’ by answering the sub-questions as defined in the introduction of this study.

3.3.1 Data Mesh Archetypes

The first sub-question to be answered is ‘*What different kinds of data mesh archetypes exist?*’ An ‘archetype’ according to the Cambridge dictionary is ‘*a typical example of something or model of something from which others are copied*’ (Cambridge University Press & Assessment, 2024) and a ‘topology’ according to the Cambridge dictionary is ‘*the way parts of something are organized or connected*’ (Cambridge University Press & Assessment, 2024). In data mesh literature the word ‘archetype’ and ‘topology’ are used to describe common architectural forms of data meshes. For example used in the studies performed by Strengtholt (2022), Pongpech (2023) and Bode et al. (2023). In this study, archetype is used, except when referring to articles that use a different terminology like (Strengtholt, 2022) for example.

When a data mesh is designed exactly like its theoretical description, it has one specific architectural form. Practice however, shows that organizations use the concept a bit more loosely and are willing to stretch the boundaries of data mesh to suit the specific needs of their organization or to make the implementation less difficult. Strengtholt (2022) distinguishes 6 different data mesh topologies, with different kinds of domain granularity. These topologies range from a fully decentralized and fined grained decoupling as originally intended by Deghani (2020) to a more managed approach. These 6 proposed topologies and their architectural design patterns are shown in Figure 5.

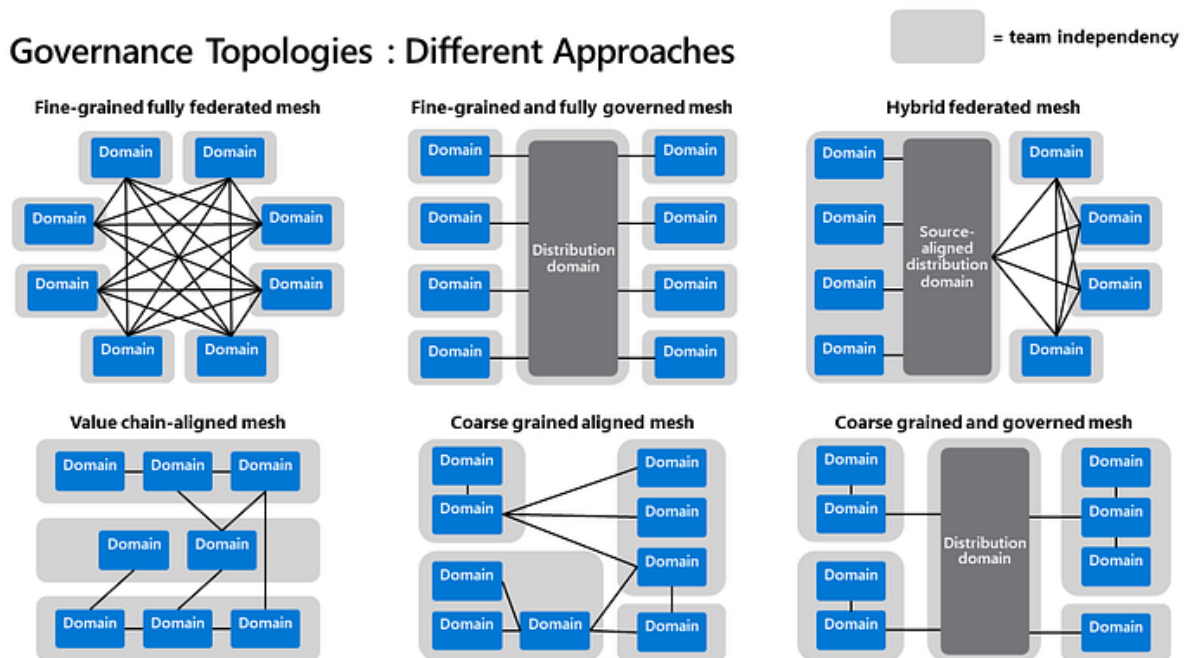


Figure 5 Data Mesh Topologies (Strengtholt, 2022)

3.3.1.1 Scenario 1 – fine grained fully federated mesh

Scenario 1 is the fine-grained fully federated mesh. This design resembles a data mesh in its most theoretical form. It is composed of small and independently deployable components. Every domain carries its own responsibility and the communication happens between the domains directly.

In such a data mesh architecture many small data product architectures are set up for sharing data between the different domains. This archetype is also mentioned by Pongpech (2023) as a ‘Fully Federated’ mesh.

The benefits of the first scenario are:

- It allows for exemplary domain specialization.
- High flexibility and limited dependencies
- Each data product becomes an architectural quantum

The fine-grained fully federated data mesh is the most theoretical data mesh structure. This comes with challenges and requires high maturity. First, because all responsibilities are decentralized this archetype requires agreement between all domains on standards for the storage of data objects and communication between domains to make the archetype interoperable. Second, this archetype creates the risk for capability duplication and inflicts high pressure on the network. Third, because there are many small data product architectures in this archetype it can lead to high costs. The fine grained fully federated mesh is the envisioned choice but hard to achieve. It is mainly found by organisations which have skilled in-house software engineers.

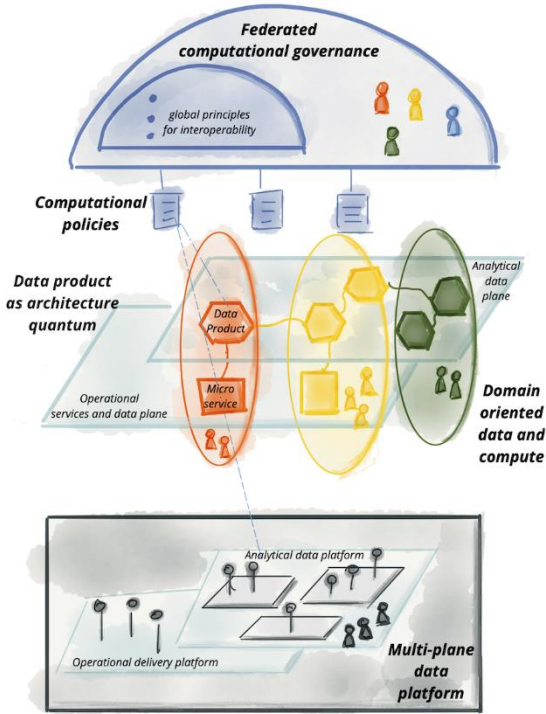


Figure 6 Data Mesh Logical Architecture (Dehghani, 2020)

Figure 6 shows how a data mesh architecture was envisioned from an architectural viewpoint when Dehghani (2020) came up with the concept. Hooshmand et al. (2022) used this archetype in their paper on transforming a monolithic PLM landscape into a landscape based on the Data Mesh principles. Another example of how this archetype is used in practice is found in the paper by Falconi and Plebani (2023) which used this archetype for peer-to-peer information exchange between different organizations. Additionally, this shows how a data mesh can be utilized as a cross-organizational solution with each organization acting as a domain. Adidas (Alcala, 2022) also implemented a data mesh based on this archetype. In the architecture of Adidas consumers have to request data products using a ticketing system. Lastly, Dahdal et al. (2023) show an example of a scenario 1 architecture to ensure real-time data for military operations. Each node is a standalone domain with all necessary components available to process data internally and communicate directly with other nodes in the data mesh.

3.3.1.2 Scenario 2 – Fine-grained and fully governed mesh

To deal with the challenges created by the fine-grained and fully federated mesh, the fully governed mesh incorporates a central data distribution layer into the architecture. Even though the inclusion of a central distribution domain does not adhere fully to the theoretical intentions of a data mesh the domains still have clear boundaries and autonomous ownership of their data products.

This archetype is also mentioned in the study by Pongpech (2023) as 'Fully Governed Mesh'. The main difference is that the data products are now provisioned by using a central distribution layer and not shared directly between the domains.

It addresses the challenges related to data distribution and data gravity of the fully federated mesh. In this archetype the domains create data products in their own domain spaces and share these using a central storage layer. In some cases companies using this archetype also provide computing and processing services which are managed centrally.

In this archetype standards are enforced more easily because non-compliant data products will not be allowed to be published on the central distribution domain. However, this piece of centralization leads to a higher time to market compared to the scenario 1 data mesh architecture. Compared to the first scenario, organizations trade some agility for more compliance and better enforced quality. This archetype is shown in the study by Kancharla and Madhu Kumar (2023) which put a unified Data Mesh layer in between consumers and producers in which the data products are combined and enriched. Zalando (Databricks, 2020) decided to implement a data mesh like this in which domains can opt-in with their data buckets that will be stored and processed on a central processing platform. This reduced the need to archive unvaluable data because only data deemed valuable by domains is plugged into the central data infrastructure as a data product.

3.3.1.3 Scenario 3 – hybrid federated mesh

The hybrid federated mesh can be seen as a combination between a centralized architecture and a federated architecture. In this archetype a single team is responsible for multiple less mature domains and the centralized distribution domain (Strengholt, 2022). On the consumer side of the architecture the domains have high autonomy and, next to communicating with the central domain, also communicate directly with each other. On the consumer side analytical domains take ownership of data and additionally share the newly created data peer-to-peer or distribute it using the central platform. This archetype is also mentioned by Pongpech (2023) as the 'Hybrid Federated Mesh'. However, in that paper each of the domains involved in the hybrid mesh are standalone domains which makes it slightly different from the topology presented by Strengholt (2022).

Both articles agree that the hybrid federated mesh includes more management overhead than the earlier mentioned approaches because the central platform has to be managed and governed, likely by a centralized team. The hybrid federated mesh can be viewed as a step towards the fine-grained and fully governed mesh in which the goal is to gradually increase the number of domains who function autonomously relating to data product creation and ownership. This archetype is mostly used by organizations that do not have a broad availability of high skilled software engineers or rely on legacy systems which are challenging to maintain and pull data from.

3.3.1.4 Scenario 4 – value chain-aligned mesh

The value chain-aligned mesh is intended for organizations which are part of the whole value chain of a product. The domains which operate in the same level of the chain are managed by one team. In this archetype centralization is organized within layers in the value chain, while the different layers are decentralized. Data is distributed between the domains directly, backwards and forwards. The different tiers in the value chains only have to adhere to central standards when crossing the domain boundaries (Strengholt, 2022). An example of organizations which would benefit from this particular data mesh architecture are fashion and retail companies who carry out the design, manufacturing, distribution and retailing of own their clothing.

3.1.1.5 Scenario 5 – Coarse grained aligned mesh

This is an archetype for organisations which have grown naturally in scale by mergers and acquisitions (Strengtholt, 2022). These type of organizations often have complicated IT landscapes. Within their architecture different levels of dependencies, governance and alignment exist. Domains consist of large groups of applications.

A challenge related to this archetype is that boundaries between domains can become unclear. Data is not particularly aligned according to the boundaries of business functions. Usually boundaries are based on organizational or geographical perspective making domains rather large. This creates challenges relating to data ownership.

Another challenge is capability duplication because each coarse grained domain uses its own data platform. Strong governance and guidance are needed to ensure that all domains consistently implement services which are required enterprise wide.

The coarse grained aligned mesh is characterized by requiring higher levels of autonomy, strong policies and strong self-service data platform capabilities. Additionally, it contradicts the principles of a theoretical data mesh. The larger nature of the domains introduces the risk of producing larger silos in which data is combined before publishing the products and data ownership is obfuscated because intermediary platforms are used to distribute data products.

This archetype additionally, is suitable for big investment management companies which possess a broad range of organizations in different industries which do not require much data traffic between each other but could benefit from having sector specific domains for example.

3.1.1.6 Scenario 6 – Coarse grained and governed mesh

The Coarse grained and governed mesh is another way for organisations with larger domains to implement a data mesh. In this mesh architecture a central platform is implemented to function as distribution platform for data product producers and marketplace for consumers (Strengtholt, 2022). An example of organizations that would benefit from this is when a parent company is divided into several banks and insurance companies. These organizations do not naturally have to share much data with each other but when a customer has a bank account at one of the subsidiaries and an insurance at another subsidiary it would be in the interest of both parties to have a central platform on which this data can be exchanged securely and timely.

3.3.2 Towards Four Archetypes

This study aims to consolidate the 6 topologies as presented in the article by Strengtholt (2022) and specifies the topologies with more detail than Pongpech (2023). Based on the articles in combination with examined literature this study proposes 4 different data mesh archetypes. The archetypes range from the most theoretical form of data mesh, to less theoretical data mesh architectures.

The first archetype, data mesh in its most theoretical form, looks like the fine-grained fully federated mesh as proposed by Strengholt (2022) and is shown in Figure 7. This study deems this to be the most mature data mesh archetype.

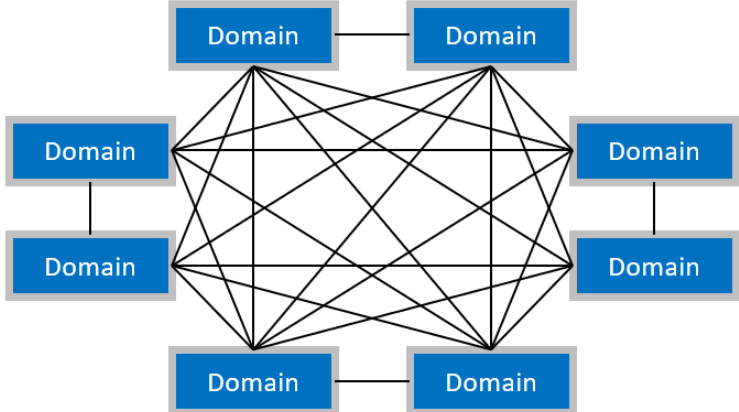


Figure 7 Pure Data Mesh, adapted from (Strengholt, 2022)

The second data mesh archetype is presented in Figure 8. This archetype is adapted from the value chain-aligned mesh and the coarse grained aligned mesh. This is the first change proposed by this study, related to the 6 topologies as presented by Strengholt (2022). This study deems the value chain-aligned mesh and coarse grained aligned mesh to be the same archetype. This is because, when viewed from an architectural point of view they are the same, they differ in how policies are set up, how capabilities are organized and values streams move internally. This does not influence the architectural principles of the data mesh, but only internal workings and data flows therefore these two topologies, as presented by Strengholt (2022), are combined into one archetype. The only difference with the first archetype, the 'Pure Data Mesh' is that one team in some cases is responsible for multiple domains which is chafing the data mesh principles in which each domain is supposed to be independent. This can however be a dedicated design decision for organisations who want each domain in a specific layer of their value chain to be managed by a single team.

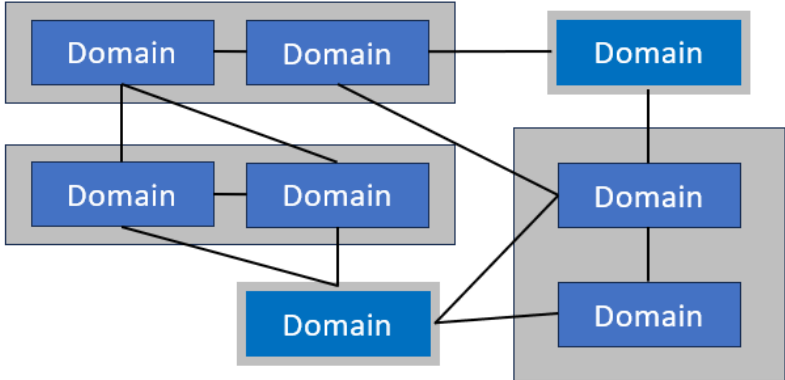


Figure 8 Semi Pure Data Mesh adapted from (Strengholt, 2022)

The third data mesh archetype presented is the hybrid federated mesh as presented by Strengholt (2022) and is shown in Figure 9. This archetype can be seen as a stepping stone towards the first and second archetype in which there is no need any more for a distribution domain.

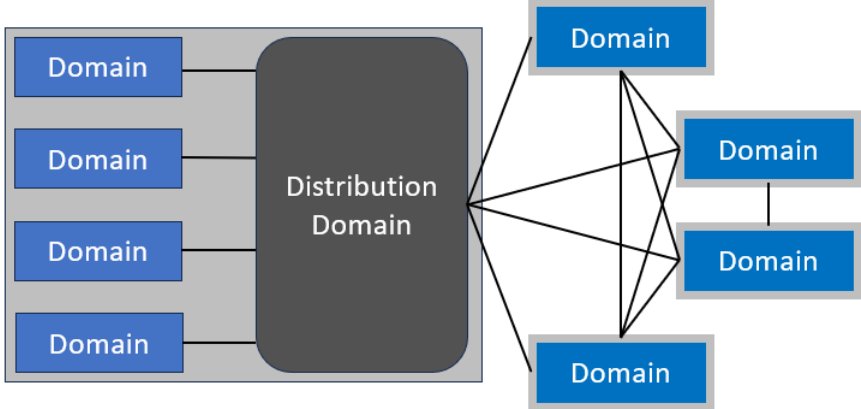


Figure 9 Hybrid Data Mesh adapted from (Strengholt, 2022)

The fourth data mesh is the fine-grained and fully governed mesh as presented in the article by Strengholt (2022). The choice was made to rename this data mesh archetype to Distribution Data Mesh as it is characterized by the inclusion of a distribution domain and is shown in the Figure 10 below.

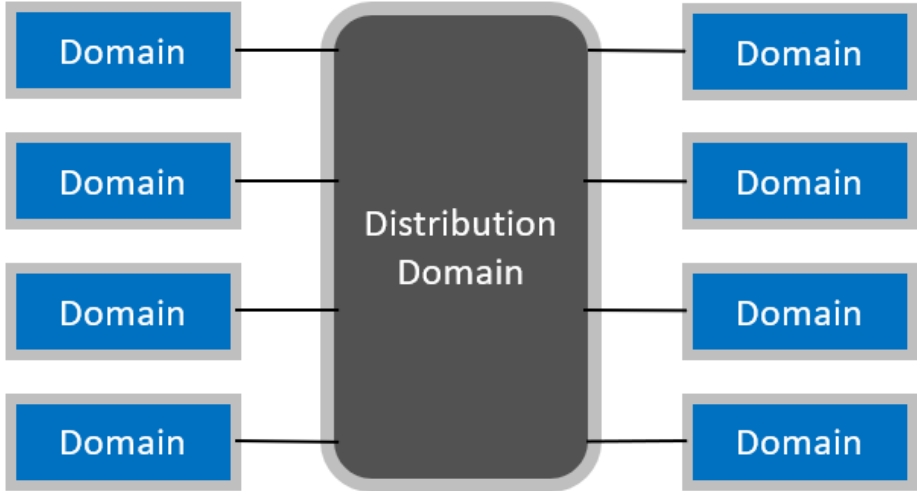


Figure 10 Distribution Data Mesh adapted from (Strengholt, 2022)

The final note is that the 'coarse grained and governed mesh' is not included in this list of archetypes as it stretches the boundaries of a data mesh to much. Because it has too much deviation from the intended ideas of a data mesh this study decided not to include it in a list containing Data Mesh archetypes.

This study therefore argues that the 4 different data mesh archetypes presented can be viewed as different levels of data mesh architecture maturity. The first level being the 'Pure Data Mesh' which is structured in full coherence with the theoretical data mesh principles. The second level is the 'Semi Pure Data Mesh'. This data mesh archetype differs from the Pure Data Mesh in a single way, which is the fact that some domains are not fully autonomous but are managed by one single team. The third level is the 'Hybrid Data Mesh'. This level is a hybrid between data mesh and more traditional data architectures because, for one part it adheres to data mesh principles however it still makes use of a centralized distribution domain for the other part. The fourth and lowest level of data mesh maturity is the 'Distribution Data Mesh'. In this archetype data exchange between domains goes through a centralized distribution domain.

This distribution domain is distinct from a central governance layer and a self-service platform. With a distribution domain a piece of infrastructure is meant which serves the purpose of storage and processing of data products which acts as additional node in the exchange of data between domains.

Each of the archetypes comes with its own advantages and challenges. The advantages and challenges related to each of the archetypes are summarized in Table 10.

<i>Archetypes</i>	Advantages	Challenges
<i>Pure Data Mesh</i>	Exemplary domain specialization High flexibility and limited dependencies Each data product becomes an architectural quantum	Risk for capability duplication Requires strong governance and agreement between all domains Can lead to high costs
<i>Semi Pure Data Mesh</i>	Less need for skilled personnel High flexibility and limited dependencies	Risk for capability duplication Requires strong governance
<i>Hybrid Data Mesh</i>	More control because of distribution domain Easier to transition to	Does not realize the full potential of data mesh Distribution domain could be a bottleneck
<i>Distribution Data Mesh</i>	More control because of the distribution domain Leverages domain knowledge Less need for technical skills Less investment needed	Does not realize the full potential of a data mesh Distribution domain could be a bottleneck Difficulty to add domains Less flexible and agile

Table 9 Data Mesh Archetypes Advantages and Disadvantages

Additional to the generic advantages and disadvantages, there are also other considerations which may influence the type of data mesh that is most suitable for an organization. For example, financial institutions may prefer a distribution data mesh or a hybrid data mesh because the distribution layer can act as an extra layer of security and governance. This could be beneficial for them because the organizations work with strictly confidential data and have strict compliance and regulatory requirements to adhere to. Another consideration which could make it unfeasible for an organization to strive for a pure data mesh is that they have legacy applications which are easier to integrate into their architecture by using a distribution domain. For an organization like this it could be too costly to change their current architecture because of the complexity of their landscape. Therefore, one archetype is not by definition better than another archetype because the archetype still has to fit the situation and requirements of an organization.

3.3.3 Data Mesh Components

This section is focussed on answering the sub-question: “*What are the common components of a data mesh?*” This will help in establishing what the key components are that an organization needs to think about when designing and implementing a data mesh.

3.3.3.1 Data Mesh Main Components

The main architectural components constituting a data mesh were set out by Dehghani (2020). The components that make a data architecture a data mesh architecture are the existence of domains in combination with a self-serve data platform and a federated governance layer. These main components themselves consist of multiple elements which together form the main component. However, Dehghani (2020) mainly spoke about data mesh in a theoretical sense so more clarity on data mesh components is needed.

Dibouliya and Jotwani (2023) reviewed how a data mesh architecture would look like. According to their study the 4 main components are: domains, a federated governance layer, a self-serve data platform and an enabling team. Their view extends the components originally proposed with an enabling team component. This study suggests that the enabling team is the team responsible for the self-serve data platform, as this is the enabling platform in the data mesh, and does not consider it an additional architectural component but as enablers of the self-serve platform. Vinnikainen (2023) and Lombardo (2023) depict a logical data mesh architecture with the 3 main architectural components domains, a self-serve data platform and a federated governance plane aligning with the original view. Butte and Butte (2022) model the federated governance plane and the domains but omit the self-serve data platform.

The consensus for this study, in line with the original data mesh idea, is that the main architectural components of a data mesh are: domains, a self-serve data platform and a federated governance layer. How these components look, and how they are organized does however vary depending on the organization that has implemented the data mesh. Literature also shows data mesh instances which include a central storage and/or processing layer (Kancharla & Madhu Kumar, 2023) or middleware even though this stretches the original intentions of a data mesh. This is also reflected in the archetypes discussed in the previous section.

3.3.3.2 Elements Constituting the Main Components

The main components and potentially a distribution platform are made up of different elements. Data products for example, the creation and exchange of which, is one of the main goals facilitated by a data mesh are an element of a domain. The data from which the data products are created is based on some operational process which generates data streams.

There are multiple actors (Pongpech, 2023) involved in a working data mesh, the data producer and the data consumer (Restel, 2023), the self-serve platform team and a federated governance group. How these roles function and are composed partly depends on how the data mesh is structured. For example, if a central distribution domain is present, the self-serve platform team becomes a centralized data team responsible for the central layer. The enabling team as discussed by Dibouliya and Jotwani (2023) can also be seen as a self-serve platform team. The actors can therefore be seen as elements in the main components.

Another vital element of a data mesh is the data catalog (Ashraf et al., 2023) (Butte & Butte, 2022) (Araújo Machado et al., 2022) (Vinnikainen, 2023) which is part of the self-serve data platform. In the data catalog information about available data products, and they way to access them is published.

Security cannot be overlooked in a data mesh, thus proper security mechanisms are required (Araújo Machado et al., 2022). Proper security policies are needed (Ashraf et al., 2023) (Vinnikainen, 2023) to ensure safe exchange of data. Security, next to safe exchange of information, is also concerned with access control (Dibouliya & Jotwani, 2023) to data products and securely storing data products. The security component is often realised through standards and policies in the federated governance layer of the data mesh.

Also related to the exchange of data are communication and interoperability policies (Butte & Butte, 2022) (Vinnikainen, 2023) setting the standards for publishing and accessing data products. This can be set up to facilitate peer-to-peer communication between domains or, in case there is a central distribution domain, to facilitate publication on, and access to the distribution domain. The policies relating to distribution of, and access to, data products are often defined in the federated governance layer. Meanwhile, infrastructure and tools needed to store and process the data products are made available through the self-serve data platform (Falconi & Plebani, 2023). Data visualization tools and data analytics capabilities are also provided to the domains by the self-serve data platform.

Lastly, an important component is to put monitoring capabilities in place (Butte & Butte, 2022). The monitoring component allows for visibility into metrics like which data products are accessed by whom, can be used to detect breaches of compliance policies, validate quality of data products, and other metrics to allow supervision over what is happening in the data mesh.

3.3.3.3 Main Data Mesh Components and Elements

This section summarizes the main components constituting a data mesh and groups important elements belonging to the main components together to provide an overview of what building blocks are typically part of a data mesh architecture.

Main Component	Elements
<i>Domain</i>	<ul style="list-style-type: none"> • Data product(s) • Domain team (Data producers) • Analytics (Data consumers) • Operational process
<i>Self-serve data platform</i>	<ul style="list-style-type: none"> • Self-serve platform team • Data catalog • Data storage infrastructure and tools • Data processing infrastructure and tools • Data analytics infrastructure and tools • Monitoring capabilities
<i>Federated governance layer</i>	<ul style="list-style-type: none"> • Federated governance group • Security policies • Communication policies • Interoperability policies • Documentation policies
<i>(not always included) Distribution domain</i>	<ul style="list-style-type: none"> • Data storage solutions • Data processing engines

Table 10 Data Mesh Main Components and Elements

3.3.4 Challenges, Limitations and Mitigations

This section will analyse the challenges and limitations of data meshes and determine possible solutions or mitigation strategies to tackle those. The sub-question to be answered in this section is: *'What are the challenges and limitations of a data mesh?'*

Designing and implementing a data mesh is accompanied by some challenges. These challenges exist in the organizational and technical layers of the company, as data mesh is a socio-technical approach that influences both layers. Therefore, the challenges related to data mesh can be viewed from 2 different perspectives. The first perspective is the organizational one, and the second perspective is the technical one. Next to challenges, there are also some limitations accompanying a data mesh which may be perceived as barriers for organizations considering transitioning their data ecosystem into a data mesh.

3.3.4.1 Organizational Challenges

First, the organizational challenges are discussed. The organizational culture of, and the way an organization works with data have to change. This is not only required from the perspective of the IT team(s), but organization-wide in all layers of the company. Companies need to align their business and technology needs (Divya et al., 2021) to establish a thriving data mesh ecosystem (Vestues et al., 2022).

Transitioning to a data mesh requires proper change management. Change management, including dealing with the resistance to change, is a frequently mentioned challenge in literature (Araújo Machado et al., 2022) (Divya et al., 2021) (Bode et al., 2023) (Hokkanen, 2021) (Goedegebuure et al., 2023). Moreover, a data mesh requires a shift in the way of working with, and thinking about data. A challenge following from this is that a common understanding about data mesh and its principles has to be established (Krystek et al., 2023) (Bode et al., 2023). Next, data mesh has an impact on its users (Araújo Machado et al., 2022) (Bode et al., 2023). Employees need to be trained to be able to adapt to shifting responsibilities. This gives rise to a limitation of a data mesh. A data mesh requires a certain skillset to be present in each of the domains and creates a need for a certain level of data literacy (Hendriks, 2023) (Hokkanen, 2021).

This can be a bottleneck for organizations lacking the required technical knowledge in house (Kraska et al., 2023) (Krystek et al., 2023) (Hendriks, 2023) (Hokkanen, 2021) (Panigrahy et al., 2023) (Goedegebuure et al., 2023). Additional to extra training on data mesh principles, companies also have to broaden the technical expertise within their staff.

Furthermore, a limitation of a data mesh is that it increases the data management complexity (Hendriks, 2023) as responsibilities over data and ownership of data change (Vestues et al., 2022). Security and privacy (Vestues et al., 2022) (Podlesny et al., 2022) (Bode et al., 2023) related challenges arise because data will be spread out over more nodes in the organization and data flows are harder to follow. Managing access to data, and keeping track of usage of data (Vestues et al., 2022) in a secure manner is difficult.

Clear data governance rules have to be defined to streamline the collaboration of all actors involved in the data mesh (Krystek et al., 2023) (Sedlak et al., 2023). It is a challenging task to reach agreement on governance principles and ensure compliance to agreed upon rules (Divya et al., 2021). A data mesh, for example, requires robust and clear Service Level Agreements (Dahdal et al., 2023) to enforce standards and clarify expected availability requirements. To supplement this quality requirements for metadata have to be agreed upon and update policies need to be defined (Sedlak et al., 2023) (Sedlak et al., 2023). If an organization is able to manage all these organizational challenges there are still some limitations which can be reasons to reconsider transitioning to a data mesh architecture. A company can be limited by its available resources (Bode et al., 2023) because there are investments needed to build a data mesh. Costs for infrastructure (Dibouliya & Jotwani, 2023) and investing in training (Hendriks, 2023) (Falconi & Plebani, 2023) can be a limiting factor. Lastly, on an organizational level, regulation and security considerations (Vestues et al., 2022) (Bode et al., 2023) may prevent an organization from realizing the full potential of a data mesh.

3.3.4.2 Technical Challenges

A data mesh does not only create challenges on an organizational level, but also on a technical level. First of all, implementing a data mesh has an impact on the existing data architecture (Araújo Machado et al., 2022). Choosing the right combination of systems and engines (Kraska et al., 2023) to support the data mesh is a challenging task. It involves questions regarding interoperability with the existing infrastructure (Araújo Machado et al., 2022) and uncertainties about tooling integration (Divya et al., 2021). A data mesh also puts a strain on the network (Dahdal et al., 2023) of an organization. The aforementioned challenges lead to data mesh being complex to implement (Dibouliya & Jotwani, 2023).

Another concern regarding data meshes is effort duplication (Araújo Machado et al., 2022) (I. A. Machado, 2022) (Goedegebuure et al., 2023) (Falconi & Plebani, 2023). This means doing repeating work in domains which has already been performed by other domains. Therefore it is important to maximize the value of the self-service platform by providing standardized tools and infrastructure limiting the replication of effort. Next, the risk of data duplication (Hendriks, 2023) (Goedegebuure et al., 2023) (Falconi & Plebani, 2023) is a known challenge in data meshes. This challenge is also related to the difficulty of establishing data products (Vestues et al., 2022) (Krystek et al., 2023) and the according standards. Standards and policies are needed to create data products of quality and minimize data duplication.

The final challenge on a technological level is how to deal with changes in data (Sedlak et al., 2023) so they are reflected in each of the data products in which this is necessary, keeping metadata in sync with federated data products, (Sedlak et al., 2023) and how to deal with deletion of data products and potential derivatives (Sedlak et al., 2023). All challenging cases that need careful consideration.

3.3.4.3 Possible Mitigations and Solutions

Table 12 provides an overview of the main mitigation techniques that can be used to tackle the most potent challenges and limitations of a data mesh approach.

Challenge / Limitation	Possible Mitigations / Solutions
Change management / resistance to change	<ul style="list-style-type: none"> • Carefully assess the need and readiness for a data mesh and create a project plan • Put emphasis on people management during the transition
Need for data literacy / need for technical knowledge	<ul style="list-style-type: none"> • Training existing employees • Hire external help or expand your IT staff
Data management complexity	<ul style="list-style-type: none"> • Put monitoring capabilities in place • Strong governance
Security concerns	<ul style="list-style-type: none"> • Have standardized security mechanisms in place • Actively enforce security policies • Encrypt data • Automated security scans
Privacy concerns	<ul style="list-style-type: none"> • Mask and anonymize data • Automated compliance checks
Governance and compliance	<ul style="list-style-type: none"> • Establish clear roles and responsibilities • Set clear governance rules • Actively monitor on compliance • Standardize communication
Data product quality	<ul style="list-style-type: none"> • Define clear quality requirements • Enforce inclusion of metadata • Quality monitoring
Cost concerns	<ul style="list-style-type: none"> • Start small and gradually expand the data mesh • Invest only in what is actually necessary
Regulatory restrictions	<ul style="list-style-type: none"> • Assure compliance with regulations and legislations before starting the transition
Impact on existing IT infrastructure	<ul style="list-style-type: none"> • Examine interoperability of required tooling and infrastructure with the current architecture • Investigate network requirements • Considerately choose the right combination of systems and engines
Effort replication	<ul style="list-style-type: none"> • Effectively design and use the self-service platform to share standardized services and infrastructure
Data duplication	<ul style="list-style-type: none"> • Leverage metadata • Put standards and policies in place
Data product consistency and maintainability	<ul style="list-style-type: none"> • Keep metadata up to date • Monitor and alert on data changes

Table 11 Data Mesh Challenges and Mitigations

3.3.5 Data Mesh Structures, Components and Considerations

The goal of the previous sections was to answer knowledge question 1: *'What are the key components constituting a data mesh and what are the limitations?'* by answering 3 sub-questions:

- What different kinds of data mesh archetypes exist?
- What are common components of a data mesh?
- What are the limitations of data mesh?

To start with the first sub-question, 4 different data mesh archetypes were identified. The 'Pure Data Mesh', 'Semi Pure Data Mesh', 'Hybrid Data Mesh', and the 'Distribution Data Mesh' in order of maturity. The Pure Data Mesh is a data mesh constructed in its most theoretical form but each of the archetypes comes with its own considerations and therefore an organization has to carefully assess which archetype fits its situation and environment best.

Next, the 3 main components of data meshes and a collection of elements making up these main components were identified. The main components constituting a data mesh are 'Domains', the 'Self-Serve Data Platform' and a 'Federated Governance' layer. Some of the most important elements composing the main components are Data Products, the domain team, self-serve platform team and federated governance group. Additionally, a data product catalog is a must have. Other important elements are monitoring capabilities, storage and processing tools and proper policies related to security, documentation and interoperability.

Lastly, the main challenges and limitations related to data mesh were identified and possible solutions and mitigation techniques were proposed. Some of the most potent challenges are the need for data literacy and technical expertise, establishing clear governance and standards, effort replication and data product maintainability.

3.4 The Shift to Data Mesh

A data mesh is not a one-size-fits-all approach. Organizations need to make a considerate choice whether to transition to a data mesh after assessing if it suits their needs and is effective for their strategy. This section will therefore be dedicated to answering the second knowledge question: *'Which factors determine if data mesh is a valid approach for my organization?'* by answering the sub-questions as defined in the introduction of this study.

3.4.1 Data Mesh Prerequisites

A data mesh is not a suitable approach for every organisation. Every organisation is different, and therefore the transition towards a data mesh should not be made without carefully assessing the motivational factors. This section will investigate what the main prerequisites are to make the shift towards a data mesh a valid approach and answer the sub-question *'what are the main indicators to consider the switch to a data mesh?'*

Bode et al. (2023) identified 6 motivational factors that drive companies to build a data mesh. The identified factors are:

- To reduce bottlenecks: this is related to the central data team which lacks the capacity to timely handle data requests from the business. Additionally, solving this bottleneck will improve the time to market and scalability of data use cases.
- Leverage domain knowledge: by bringing the responsibility of data back to the domains the quality of data will improve. Because employees with domain knowledge become responsible for providing domain data.
- Break down silos: because of the self-serve platform, business units can request the data they need by using the data catalog without the need to communicate with members of another domain to get access to the required data. This breaks down the barriers of silos.
- Establish data ownership: the shift of ownership back to the domains creates a strong sense of responsibility over the created data. This responsibility will improve data quality because poor quality data will reflect poorly on a domain.

- Adopt modern architecture: some organizations are persuaded to look into data mesh because other organizations in the industry transition to a data mesh. However, a data mesh is not a one-size-fits-all approach. Organizations must be careful not to adopt a data mesh for the wrong reasons.
- Reduce redundancies: a lack of transparency and communication in combination with siloed business units can lead to replicated effort. A data mesh can solve this by making good use of the self-serve platform and employing proper standards through federated governance so domains can reuse the work of other domains.

Bode et al. (2023) also mention that organisations feel pressure to adopt a more modern architecture because industry competitors are doing so. While the sole reason for adoption should not be based on the actions a competitor takes, it can be a reason for organizations to start investigating the possibility. If an organisation deems it necessary, a data mesh is a way to improve its technical maturity and build an architecture resilient for the future. Additionally, it allows for faster adaptation to changes in the market. If the technical knowledge is present within a company, and the transition to a data mesh has been carefully considered these are valid reasons to make the transition.

Hokkanen (2021) looked at it from the opposite point of view and distinguished factors which block the adoption of a data mesh. The study identified barriers on organizational, technological and industry level. These factors are shown in Figure 11.

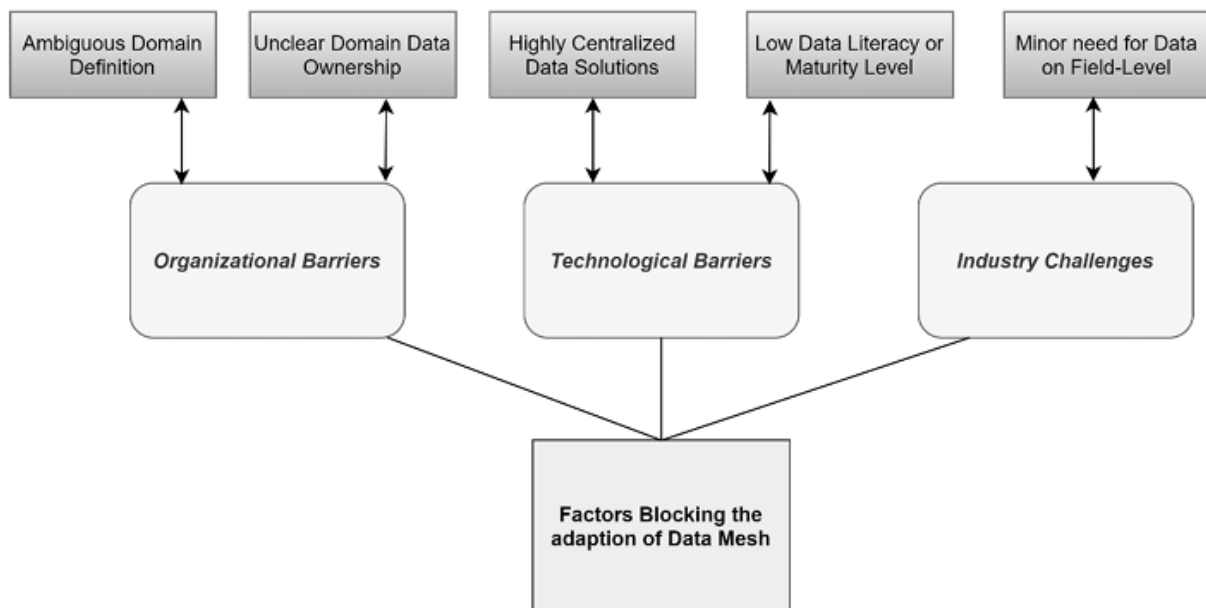


Figure 11 Factors Blocking Data Mesh Adoption (Hokkanen, 2021)

By reversing the factors identified by Hokkanen (2021) we can identify prerequisites for data mesh adoption. For example, from the ambiguous domain definition barrier, follows that clearly defined domains are a prerequisite if a company wants to adopt a data mesh. The clearly defined domains make it possible to leverage domain knowledge. The other organizational factor, unclear domain data ownership, can be turned around to function as a motivational factor, as also provided by Bode et al. (2023), to improve data ownership.

Highly centralized data solutions is a technological barrier and the breaking down of these silos as put forward by Bode et al. (2023) is thus a motivational factor. The barrier of low data literacy, and/or technological maturity level, creates the requirement to have a certain maturity level in-house and to have a certain level of data literacy before considering the switch to a data mesh.

If there is a low need for data on a business level, a data mesh is not the right solution for an organization. A motivational factor found in literature is the need to process high volumes of data in a variety of formats. Data lead times are generally shorter in a properly functioning data mesh which also has advantages when processing data in real time. This is in line with the argument that a data mesh reduces bottlenecks, and thus, decreases data lead time.

Next, transitioning to a data mesh improves the scalability and agility of the architecture. When policies and standards are well defined, and a well structured self-service platform is operational it is easy for domains to join the data mesh. This is not only true within a single organization but also for data meshes set up between multiple organizations within an industry like the CowMesh (Pakrashi et al., 2023) case. In this case data mesh was used to improve data sharing within the dairy industry to earlier detect possible diseases. This change was driven by a need for better collaboration between parties in the dairy industry.

McEachen and Lewis (2023) mention improving interoperability and collaboration between different business units as a reason to adopt a data mesh. Additionally, McEachen and Lewis (2023) point out the simplicity of joining the data mesh, and thus its scalable nature as a motivational factor. McEachen and Lewis (2023) conclude by stressing the enhanced data management following from strong domain ownership and briefly touch on cost reductions as motivational factors in favour of adopting a data mesh.

Lastly, Dončević et al. (2022) confirm the point of McEachen and Lewis (2023) that a data mesh improves manageability of data in the domains. Dončević et al. (2022) also point out the reduced lead time and improved access to domain knowledge. Finally, their study confirms the prerequisite that a data mesh is best suitable for companies that require more scalability in their data architecture.

A summary of the identified motivational factors and prerequisites is listed below:

1. Motivational factors
 - a. Need or want for a more scalable and agile architecture
 - b. Improve technical maturity
 - c. Governance and compliance needs which are easier to enforce by using a data mesh approach
 - d. The company has to change because of existing challenges like data siloes, low interoperability and low value of data
 - e. Strategic business objectives drive the organization to adopt a more data driven approach
 - f. Requirement to be able to adapt fast to the market
 - g. Want or need to improve internal communication and collaboration
 - h. Want or need to improve the quality of data and data operations
 - i. Improve collaboration with other parties in the ecosystem or industry
2. Prerequisites
 - a. Need to process high volumes of data in a variety of formats
 - b. A certain level of technical knowledge must be present in the company
 - c. Data literacy and culture are at a high level in the organization
 - d. Clearly defined domains
 - e. It needs to make sense to break up the architecture into different domains
 - f. Budget is available to make investments

3.4.2 Impact of the Data Mesh Transition

This section is dedicated to answering the sub-question '*what is the impact of data mesh on the existing architecture?*' As mentioned in earlier sections data mesh requires a transformation on multiple organizational levels and thus it is important for organizations to have a view on the impact the transition to a data mesh has on the organization and the existing architecture.

Implementing a new data strategy and architecture influences the organizational culture as well as the existing enterprise architecture. First of all, the decentralization of data ownership by moving away from a monolithic architecture requires both an organizational, and architectural reorganization. Domain boundaries need to be defined (Jonkman, 2023) (Dibouliya & Jotwani, 2023) and infrastructure has to be made available in each domain. A new governance model has to be set up, and new and changed roles and responsibilities have to be defined (Vestues et al., 2022) (Li et al., 2022). The new governance model must also entail the policies and standards which have to be established in the federated governance layer.

A self-serve data platform has to be designed and established. This requires decisions on which tools and services are to be made centrally available and investments have to be made. Additionally, the current data and IT architecture have to be examined and interoperability with new tools and services has to be studied (Pakrashi et al., 2023) (Krystek et al., 2023). Because a data mesh requires changes to the technological landscape it allows for infrastructure modernization. It provides an opportunity to build an architecture resilient for the future. The architecture will become more scalable and allows data to be leveraged more as a strategic asset (Jonkman, 2023) (Dahdal et al., 2023).

Consequently, the complexity of managing and coordinating the technological landscape will increase. Instead of managing a monolithic architecture, the architecture is broken down into domains that are autonomous. It is harder to gain a single overview of what is happening in the whole organisation. There will be need for continuous monitoring of data products and on compliance with company-wide policies (Vestues et al., 2022) (Kraska et al., 2023). Data mesh does however improve resource allocation, as it is easier to estimate the required resources on domain level than on company-wide level. If set up properly a data mesh will aid in leveraging data as a strategic asset.

Lastly, the transition to a data mesh has a long term strategic impact. Transitioning into a data mesh architecture is not a decision made for short term benefits. The decision has to be made after careful consideration and with a long term strategic plan to support it, as it impacts the way of working in the organization and the changes the technological landscape.

The following list summarizes how a data mesh impacts the existing architecture and culture of an organization:

1. Shifting from a monolithic architecture to a distributed architecture
2. Infrastructure reorganization and modernization
3. New governance models required
4. Improved scalability
5. Better resource allocation
6. Increased complexity in management and coordination
7. Demand for new skills and roles
8. Shifting responsibilities and tasks
9. Need for continuous monitoring
10. Long term strategic impact
11. Helps to leverage data as a strategic asset

3.4.3 Other Data Methodologies

As a data mesh is not a suitable approach for every organization alternative approaches have to be examined as well. Therefore this section is dedicated to answering the following sub-question: ‘which other data methodologies are there?’

In literature a distinction is made between 5 different data methodologies: data warehouse, data lake, data lakehouse, data mesh and data fabric. Each approach serves unique purposes and offers distinct advantages. The data warehouse, is the first generation of data platforms. Data warehouses are centralized data storages designed to integrate and store data from multiple sources (Bode et al., 2023). Data warehouses store structured data and allow for easy querying. Data warehouses enable quick data analytics and reporting capabilities (I. A. Machado et al., 2022). The limitations of a data warehouse platforms are the stale nature, difficulties with processing semi- and unstructured data, and high costs as the volume of data grows (Azeroual. & Nacheva., 2023) .

Therefore, to tackle problems arising with data warehouses two-tier architectures were designed combining data warehouses with data lakes. The inclusion of data lakes made it possible to store semi- and unstructured data. Additionally, the addition of data lakes to the architecture enabled incorporation of data science and machine learning capabilities (Vinnikainen, 2023). Eventually, two-tier architectures also started to fall short in meeting increasing requirements. Challenges of the two-tier architectures are the complexity of implementing data pipelines, the separate ETL process not being able to meet the demand for timely data, and rising cost (Voß, 2022). Additionally, it requires separate management of the data warehouse and data lake storages.

Following this, data lakehouse platforms came into existence trying to maintain the benefits of using both warehouses and lakes while reducing the management overhead of managing both storage solutions separately. The lakehouse approach allows for the low-cost storage of raw data while simultaneously allowing for data warehouse capabilities (Jonkman, 2023) (Priebe et al., 2021). It supports real-time data streaming and allows for comprehensive analytics. Figure 12 shows a visualization of the 3 different data platforms.

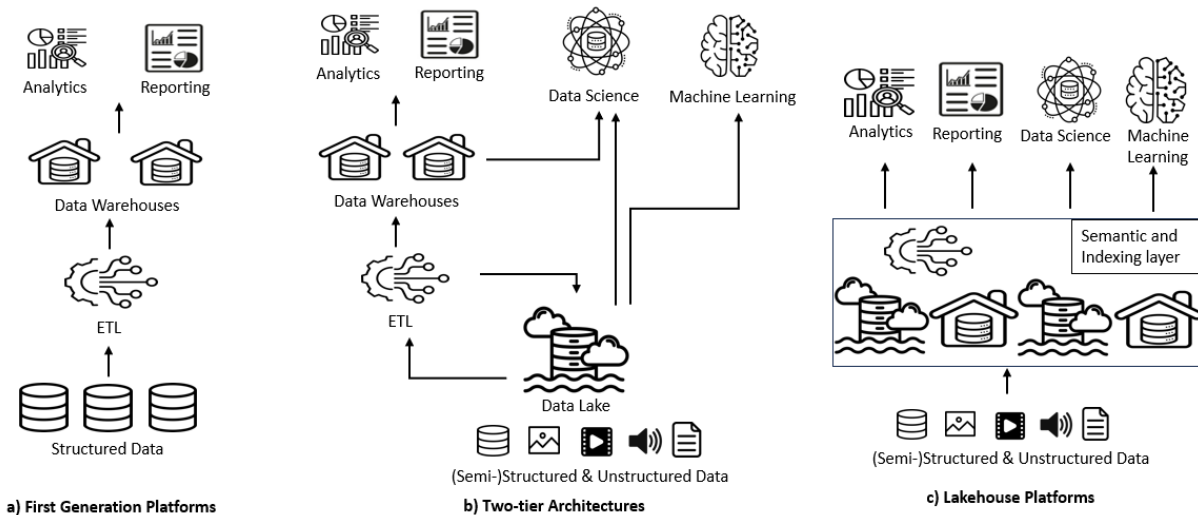


Figure 12 Different Generation Data Platforms adapted from (Zaharia et al., 2021)

The data warehouse, data lake and data lakehouse platforms are centralized solutions which are managed by central data teams. In these architectures the central teams are becoming a bottleneck for organizations dealing with large volumes of data, increasing demands for analytics and in need for more scalability. Therefore, other approaches have to be examined like the data mesh or data fabric.

The data mesh and data fabric are both approaches to deal with problems arising from the monolithic data architectures discussed above and in the introduction of this study. The data fabric is an approach that aims to create a unified data management framework by integrating data flows, and storage and processing technologies (Priebe et al., 2021). Data fabrics provide a holistic view of data improving data governance, accessibility and security. Data fabrics enhance the ability to efficiently leverage data from multiple sources (Dibouliya & Jotwani, 2023). Challenges of data fabrics are that it is complex to implement and manage (Jonkman, 2023). Additionally, maintaining data consistency and quality across the data sources is challenging. Lastly, scaling while maintaining performance can become costly and ensuring interoperability between different systems is a challenging task. Each of the different data platforms and architectures has its own strengths and weaknesses and they are also not mutually exclusive.

Table 13 summarizes the different data methodologies and what the main advantages and disadvantages of each of the approaches are.

	Data warehouse	Data lake	Data lakehouse	Data mesh	Data Fabric
<i>What is it</i>	A solution to centralise and consolidate data from multiple sources for analytical purposes	A centralized storage for large amounts of data in their original format for advanced analytical purposes	A centralized solution combining data warehouse and data lake principles	Decentralized data architecture based on 4 core principles: data as a product, domain ownership, federated governance and self-serve infrastructure	A unified network-based architecture providing real-time access to a distributed data layer.
<i>Advantages</i>	Optimized for query performance Highly structured data Mature technology Ideal for BI and reporting	Can store all types of data Scalable and cost effective	Combines the benefits of data lakes and warehouses Supports both structured and unstructured data Real time analytics and ML	Good scalability. Strong data ownership leading to higher data quality. Reduced data lead time and allows for better collaboration.	Unified data management and integration layer Supports real time processing and analytics Facilitates access to data across spread out sources
<i>Disadvantages</i>	Scaling can become costly and complex Not suitable for unstructured data Significant ETL effort needed	Less optimized for querying Can become a data swamp	Best practices and tooling still evolving Balancing between warehouse and lake features can be difficult	Requires a certain level of data literacy and technical knowledge. Requires organizational and technical changes.	Complex to implement and manage Requires advanced data integration tools and technology

Table 12 Data Platform Comparison

Each of the data methodologies has its own strengths and weaknesses and therefore the choice for a data platform has to be made based on the needs and requirements of an organisation, in line with its capabilities.

3.4.4 When to and When not to Data Mesh

The goal of the previous sections was to answer knowledge question 2: *'Which factors determine if a data mesh is a valid approach for an organization?'* by answering 3 sub-questions:

- What are the main indicators to consider the switch to a data mesh?
- What is the impact of data mesh on the existing architecture?
- Which other data methodologies are there?

To start with the motivating factors and prerequisites for data mesh adoption. The most prevalent motivational factors are having a more scalable and agile data architecture, improve collaboration between business units, and improve data quality and operations. Additionally, a data mesh is most effective for organizations that need to process high volumes of data in a wide variety of formats. To be able to make the transition to a data mesh the most important prerequisites are to have a certain level of data literacy and technical knowledge in-house and have enough budget to realize the transition.

The transition of a data mesh impacts the existing architecture and culture of an organization. The most prevalent impact is the shift from a monolithic data architecture to a distributed data architecture. This is paired with shifting responsibilities and tasks for employees, and requires new governance models. Additionally, a demand for new skills and roles is created, and continuous monitoring capabilities are needed. On the other hand, a data mesh will improve scalability of the architecture, it helps to leverage data as a strategic impact and allows for modernization of the data infrastructure.

Lastly, because a data mesh is not a fitting approach for every organization alternative data platform approaches were identified and compared to data mesh. Alternatives to a data mesh are traditional approaches like data warehouse, data lake, and data lakehouse platforms. Additionally, organizations can also consider looking into data fabric solutions.

3.5 Data Reference Architectures

This section is dedicated to answering the knowledge question: *'are there existing data mesh reference architectures?'*. During an earlier phase of this study, one reference architecture specifically tailored to a runtime structure of data mesh by (Goedegebuure et al., 2023) was identified. However, no other data mesh reference architectures were found in the databases included in this study within the set of inclusion criteria. Therefore additional ideas and inspiration have to be gathered from other data related reference architectures.

3.5.1 Data Reference Architecture Characteristics

To gain a better understanding of data reference architectures the first sub-question to be answered is: *'what are the characteristics of data reference architectures?'*

During the selection phase of the SLR 19 data reference architectures were identified and explored in more detail. The first noticeable thing about the examined reference architectures from the literature study is that there is a lot of difference in the modelling style chosen by the different authors. Many authors chose to stay away from an existing modelling language and create the RA in a free format. Another interesting finding is that most authors chose to validate their reference architecture by mapping it to an existing solution architecture.

To get a clear view of the examined Reference Architectures (RA) for this study Table 14 summarizes the key aspects of the investigated RAs. The aspects examined are:

1. The focus of the RA: the domain or industry the RA was created for.
2. The method of construction: the methodology, if any, used for the design and development of the RA.
3. The use of a modelling language: the modelling language, if any, used for visualizing the RA.
4. The validation method: the validation method, if any, used to examine the validity of the created RA.

Source	Focus	Method	Modelling Language	Validation
<i>(Goedegebuure et al., 2023)</i>	Data Mesh	Consolidating components from data mesh architectures found in literature	Language free	No validation
<i>(Sang et al., 2016)</i>	Big Data	No method used Divided key elements of big data use into 5 components and created mapping notations	Language free	Case study, mapping RA to solution architectures
<i>(Giray & Catal, 2021)</i>	Data Management (for agriculture)	DSR by (Hevner et al., 2004)	Language free	Mapping RA to a set of requirements established in literature
<i>(Klein et al., 2016)</i>	Big Data (national security domain)	No method used Established domain specific requirements The RA is a collection of modules decomposable into elements that realise functions or capabilities	Language free	Case study, demonstrate how RA is used to design an OSINT systems
<i>(Sang et al., 2017)</i>	Big Data Analytics	No method used Divided key elements of big data use into 5 components and created mapping notations	Language free	Case study, mapping to solution architecture
<i>(Wehrmeister et al., 2022)</i>	Big Data (energy sector)	No method used Combining Existing RAs to extend an RA to satisfy additional requirements	Language free, based on other RAs	No validation
<i>(Geerdink, 2013)</i>	Big Data	(Angelov et al., 2012)	TOGAF ArchiMate	Questionnaire

<i>(Pääkkönen & Pakkala, 2015)</i>	Big Data	(Angelov et al., 2012) (Galster & Avgeriou, 2011)	Language free	Case study, mapping to solution architectures
<i>(Arianyan et al., 2023)</i>	Big Data	Analysis of Big Data Reference Architecture standards	Not applicable	Mapping standards to RAs
<i>(Xiaofeng & Jing, 2020)</i>	Big Data	Use case modelling method	Language free	None, comparison with other models
<i>(Gollapudi, 2015)</i>	Data Aggregation (Financial Services)	Stating key design concerns	Language free	None
<i>(Roman & Stefano, 2016)</i>	Data marketplaces	Based on issues and concerns	Language free	None
<i>(Viana & Sato, 2014)</i>	Long term archiving, preservation and retrieval of Big Data	(Angelov et al., 2012)	Language free	In progress
<i>(Garises & Quenum, 2018)</i>	Big Data (healthcare)	(Galster & Avgeriou, 2011)	Language free	None
<i>(Maier, 2013)</i>	Big Data / Data Management	(Galster & Avgeriou, 2011)	Language free	Case Study, mapping to solution architectures
<i>(Iglesias et al., 2020)</i>	Big Data (emergency management)	None based on NIST Big Data RA	Language free	Case Study, mapping to solution architecture
<i>(El Arass et al., 2020)</i>	Big Data Application Provider	None extension on NIST Big Data RA	Language free	Case Study, mapping to solution architecture
<i>(De Almeida Neto & Castro, 2017)</i>	ETL stages of educational data mining and learning analytics	No method	Language free	None
<i>(Otto & Hüner, 2009)</i>	Master Data Management	4 phase approach as described in the report	Language free	Mapping to existing products, Case Studies

Table 13 Reference Architecture Characteristics

3.5.2 Reference Architecture Parts

This section is dedicated to answering the sub-question: *'what parts of other data reference architectures can be re-used?'*

A lot of the components found in the reference architecture by (Goedegebuure et al., 2023) were also identified in section 3.3.3 and therefore can be used as inspiration for the reference architecture to be designed later in this study. However, the RA by (Goedegebuure et al., 2023) is mainly focussed on data product exchange in a runtime environment and lacks a clear model of the domain even though this is a big part of data mesh architectures. Thus creating a new version of this model would not satisfy the goal set for this research.

The other data related reference architectures examined have no relation to data mesh and therefore do not provide the opportunity to copy components or to extend upon, however they provide some valuable contributions for this study. Based on the examined literature, 2 reference architecture development methodologies were identified which were further explored in the following section, section 3.6, in which the method used to develop the RA later in this study was determined. Next, the literature provided insight into 2 valuable validation methods. The validation methods were further examined in section 3.7.

3.6 Developing a Reference Architecture

When developing a reference architecture it is important to have a methodology or plan to follow to create an empirically sound reference architecture in a structured way that satisfies the goals of this study and of envisioned stakeholders. Therefore this section is dedicated to answering the knowledge question: *'how to develop a reference architecture?'*

3.6.1 Goals and Requirements of a Reference Architecture

This section is dedicated to answer the sub-question: *'what are the goals and requirement of a reference architecture?'*

A reference architecture's main purpose is to provide a template which outlines the structure of systems within a specific domain or for a specific type of platform (Angelov et al., 2012). It is often generalized and serves as a blueprint that guides the design and implementation of concrete architectures (Galster & Avgeriou, 2011). Reference Architectures are important to ensure consistency and efficiency across projects and managing quality by providing a set of standardized best practises and solutions (Cloutier et al., 2010).

The specific goals and requirements of a reference architecture are dependent on the envisioned stakeholders of the RA, the domain the RA is created for, and requirements from practice. However, there are some common objectives and criteria to follow when constructing a RA. The primary goal behind the construction of reference architecture is to standardize approaches for the design and development of solution architectures (Cloutier et al., 2010) (Nakagawa et al., 2012). The standardization is meant to ensure compatibility, interoperability and consistency across different projects. The incorporation of common frameworks and components enables system development according to industry or organizational standards.

Another objective of RAs is to encapsulate and advertise the best practices within a domain or industry (Cloutier et al., 2010) (Angelov et al., 2012).

Next, RAs facilitate communication and interoperability (Cloutier et al., 2010) (Galster & Avgeriou, 2011) as RAs provide a common language and model which improves communication between stakeholders. It also facilitates in creating a shared understanding about requirements, functionalities and components.

RAs accelerate the design and development of solution architectures (Nakagawa et al., 2012) because a RA offers a predefined structure and set of components.

When a RA is created by a governmental or legal body a RA can help achieve regulatory compliance (Heuser et al., 2018).

Additionally, a RA should be flexible to be useful in different use cases and suitable for systems of different sizes and varying complexity. Therefore it should remain neutral towards technology (De Almeida Neto & Castro, 2017). RAs are not effective without extensive documentation covering its components, patterns and guidelines for usage. The documentation must be understandable and accessible to all stakeholders (Cloutier et al., 2010) involved in the process.

Lastly, reference architectures need to be updated to stay relevant and keep up with changing requirements. Therefore RAs need to be maintainable (Cloutier et al., 2010) and allow for extensions or changes.

Based on the RAs examined in section 3.5, common steps involved in the design of a RA are: to determine the need in a domain for a standardized approach and identify the stakeholders. Next, the requirements for the RA have to be defined and the scope of the RA has to be determined. Then a design method must be followed to create the RA, and the final step is to validate the RA.

3.6.2 Reference Architecture Design Methodologies

This section is dedicated to answer the sub-question: ‘*which method can be used to design and develop a reference architecture?*’ based on the methodologies examined in this section a methodology will be chosen to design the RA in the next section of this study.

During the literature review for knowledge question 3, in section 3.5, 2 reference architecture development methods were identified. The identified methods are: a framework for analysis and design of software reference architectures by Angelov et al. (2012), and a method by Galster and Avgeriou (2011) to create empirically grounded RAs.

Angelov et al. (2012) propose a framework which has 3 different applications. To analyse an existing reference architecture, to design a reference architecture or to re-design a reference architecture due to changes in the environment.

The methodology by Angelov et al. (2012) contains the following steps:

Step 1: Analyse the relationship between the context, goals, and design of the reference architecture		
<i>Dimension / Type</i>	<i>Sub-Dimension</i>	<i>Description</i>
Context Dimension (C)	C1: where will it be used?	Will the RA be used in a single organization or multiple organization
	C2: Who defines it?	Refers to the stakeholders involved in creating the RA. This could be different types of organizations or in the case of a single organization organizational groups or on a lower level the types of people involved.
	C3: When is it defined?	1. Preliminary RA: created when the required tools and technologies for its concretization aren't available when it is designed. These reference architectures are characterized as research experiments. 2. Classical RA: takes input from concrete technological solutions. The tools and technology required to implement such RA's is readily available.
Goal Dimension (G)	G1: Why is it Defined?	There are two values for this dimension: standardization of concrete architectures or facilitation of the design of solution architectures.

Design Dimension (D)	D1: What is described?	Every RA should feature component and connector elements. Other elements like policies, interface and guidelines are texture.
	D2: How detailed is it described?	3 levels of detail are distinguished; detailed, semi-detailed and aggregated specification of elements. 2 methods to measure the level of detail can be used: 1) Simply counting the number of elements constituting the reference architecture, and 2) Counting the quantity of distinct aggregation layers establishing the specification of the RA.
	D3: How concrete is it described?	3 levels of abstraction values are defined; abstract, semi-concrete and concrete. These values can be assigned to each of the elements of D1. 1. Abstract: an abstract RA does not specify how elements should be implemented. 2. Semi-concrete: in a semi-abstract architecture there is a class of choices for each specific element. 3. Concrete: In a concrete architecture a specific choice is made for each element.
	D4: How is it represented?	3 levels of formalization have been defined; informal, semi-formal and formal. 1. Informal: an informal RA uses natural language or a free from graphical notation. 2. Semi-formal: a semi-formal RA is based on a modelling notation like UML which lacks a mathematical background. 3. Formal: a formal RA is based on an architecture specification language like ArchiMate.
Step 2: Determine the type of reference architecture to be created		
<i>Dimension / Type</i>	<i>Sub-Dimension</i>	<i>Description</i>
Standardization RA types	Type 1	Classical standardization architectures to be implemented in multiple organizations. Typically there are multiple organizations responsible for the creation of such a RA.
	Type 2	Classical standardization architectures to be implemented in a single organization. This type of architecture is used to standardize approaches within a single organization.
Facilitation RA types	Type 3	Classical facilitation architectures designed for multiple or organizations designed by an independent or software organization.
	Type 4	Classical facilitation architectures to be implemented in a single organization.
	Type 5	preliminary facilitation architectures to be implemented in multiple organizations. These type of RA's are usually developed by researchers.

Table 14 Framework for Reference Architecture Design (Angelov et al., 2012)

Galster and Avgeriou (2011) propose a method to create empirically grounded reference architectures. The methodology follows a 6 step approach:

Step 1: Decide on the type of RA

Step 2: Selection of the design strategy

Step 3: empirical collection of data

Step 4: construction of the RA

Step 5: enable the RA with variability

Step 6: evaluation the RA to check its validity.

The first step is to decide the type of RA to be created. Galster and Avgeriou (2011) group the types of RA based on 2 dimensions. The type of reference architecture is based on a combination of the usage context and the characterization framework dimensions proposed by Angelov et al. (2012). In terms of the usage context a distinction is made between 3 items.

1. Platform specific RA's: for example a reference architecture specifically focused on AWS cloud.
2. Industry specific RA's: a reference architecture focussed on a specific industry like healthcare, as seen in Garises & Quenum (2018) for example.
3. Industry cross cutting RA's: a reference architecture covering multiple industries like some of the big data reference architectures discussed in section 3.6.

The characterization framework by Angelov et al. (2012) proposed 5 types of reference architectures. The different types are discussed in earlier in this section, they are only listed here. The 5 types are:

1. Classical standardization architectures to be implemented in multiple organizations
2. Classical standardization architectures to be implemented in a single organization
3. Classical facilitation RA's for multiple organizations
4. Classical facilitation architectures to be implemented in a single organization
5. Preliminary facilitation architectures to be implemented in multiple organizations.

When the type of RA has been defined based on these 2 dimensions the next step of the process can be initiated.

The second step is to decide the design strategy for the reference architecture. Galster and Avgeriou (2011) distinguish 2 different design strategies. Designing a RA from scratch or design the RA based on existing artifacts. When designing an RA from scratch it is a research-driven reference architecture and when the design is based on existing artifacts it is practice-driven.

The third step is to acquire data empirically. The literature review of this study provided us with a literary basis for the construction of a Data Mesh Reference Architecture which is covered in more detail in section 4.

The fourth step is to construct the RA. The reference architecture is build based of the information found in literature or by extending or changing existing artifacts.

The fifth step is to enable the RA with variability. There are 3 ways to achieve variability. Annotation to elements can add variability to the RA, and the other two options are to create variability models or views.

The last step is to evaluate the RA empirically to assess its value and validity. Based on findings in literature on reference architectures in section 3.5 there are two methods commonly used to validate reference architectures. By conducting a survey and/or by mapping the reference architecture to a solution architecture in one or multiple case studies. The different validation methods will be described in more detail in the following section, section 3.7.

In line with the research methodology followed in this study the method by Galster and Avgeriou (2011) was chosen to design and create an empirically sound data mesh reference architecture in this study because it includes validation as part of the process which aligns with the treatment validation step of the engineering cycle (Wieringa, 2014).

3.7 Validating a Reference Architecture

This section is dedicated to answering the final knowledge question: *'how can the reference architecture be validated?'*

Two common validation methods for reference architectures were identified in section 3.5: validation by survey and validation by Case Study in which the Reference Architecture is mapped onto an existing solution architecture. We found that in 8 out of the 19 studies examined in section 3.5.1 one or multiple case studies were performed to validate the reference architecture. The goal in these case studies was to map the components of the reference architecture onto solution architectures to demonstrate, usefulness, compatibility and completeness. Geerdink (2013) used a questionnaire to validate the designed reference architecture. The goal of the questionnaire was to get expert opinion on different criteria of the reference architecture, which in the case of Geerdink (2013) were; maintainability, modularity, reusability, performance and scalability.

In line with the Design Science Research methodology by Wieringa (2014), which is the method followed in this study, the choice was made to validate the RA by using expert opinion. Therefore, the study by (Geerdink, 2013) will be used as inspiration for a questionnaire which will be used to validate the RA. Experts provide an understanding of how stakeholders of the RA perceive the model and can propose changes and points of improvement.

4 Artifact Design

In this chapter the development process of the data mesh reference architecture is discussed.

In the section 3.6 two methods to design RAs were identified. The RA in this study will be designed according to the method proposed by Galster and Avgeriou (2011). The choice was made to use this method because it includes validation which aligns with the treatment validation step of the engineering cycle (Wieringa, 2014) and additionally incorporates some steps of the framework created by Angelov et al. (2012). The method consists of 6 steps:

- Step 1: Decide on the type of RA
- Step 2: Selection of the design strategy
- Step 3: empirical collection of data
- Step 4: construction of the RA
- Step 5: enable the RA with variability
- Step 6: evaluation the RA to check its validity.

4.1 Type of Reference Architecture

The first step, is to decide the type of the envisioned reference architecture. For this, a decision had to be made on two dimensions. First, the usage context had to be determined. The usage context for the reference architecture envisioned in this study was 'industry-cross-cutting'.

Second, a characterization for the envisioned reference architecture had to be determined based on the characterization framework by Angelov et al. (2012). Galster and Avgeriou (2011) incorporate 3 questions from the framework by Angelov et al. (2012) into their method. A why question, a where question, and a when question. The following questions had to be answered establishing why the RA was created, where the RA will be used and when the RA was created. The following aspects determine the characterization of the envisioned reference architecture:

- Why? The envisioned reference architecture in this study is created as a 'facilitation' reference architecture to help the design of solution architectures.
- Where? The envisioned RA will be used in multiple organizations.
- When? The envisioned RA is a 'classical reference architecture' as the technologies necessary to create solution architectures are readily available.

Based on these 3 answers the type of reference architecture that was envisioned was a type 3 RA, 'classical facilitation reference architecture for multiple organizations'.

4.2 Reference Architecture Design Strategy

The second step of the process, was to select a design strategy for the envisioned reference architecture. Two design strategies are proposed for developing a reference architecture. Developing the RA from scratch or developing the RA based on existing architectures.

Because only a single data mesh RA was identified within the scope of this study, which did not cover all aspects, this research wants to cover, the design strategy for the envisioned Data Mesh Reference Architecture is to build the RA from scratch. The design of the RA will be practice driven, based on the findings from the literature review performed earlier in this study.

4.3 Empirical Acquisition of Data

The third step in the process was to collect empirically grounded data to support the design of the reference architecture. For the collection of data, for example about the main components of the RA, section 3, the SLR, is referred to.

4.4 Construction of the Reference Architecture

The fourth step was to create the reference architecture based on the collected data from the previous step. The envisioned architecture was mainly based on the section 3.3.3 about the main components of data mesh and based on the design decisions made in step 1 and 2 of this method.

For this study the choice was made to build the RA using the ArchiMate modelling language. ArchiMate was chosen as it is an extensive language and the components are clearly described in the ArchiMate specification (TheOpenGroup, n.d.). ArchiMate also has clearly described relationships making the connection between different elements clear. By virtue of using ArchiMate the reference architecture consist of clear building blocks and can easily be extended by adding other ArchiMate components.

Another deliberate design choice is to include a component only once in the architecture even if multiple instances of this component can be present in a solution architecture. For example, a domain team can be responsible for multiple business processes or applications but only one business process was modelled to keep the models clear and readable.

Next, the choice was made to refrain from specific technology and tools and only model components in a general sense. This leaves room for users of the RA to decide on the preferred technology and tools in a solution architecture. Additionally, this is in line with the 'classical facilitation RA for multiple organizations' determined in step 1 of the process.

4.4.1 Domain Reference Architecture

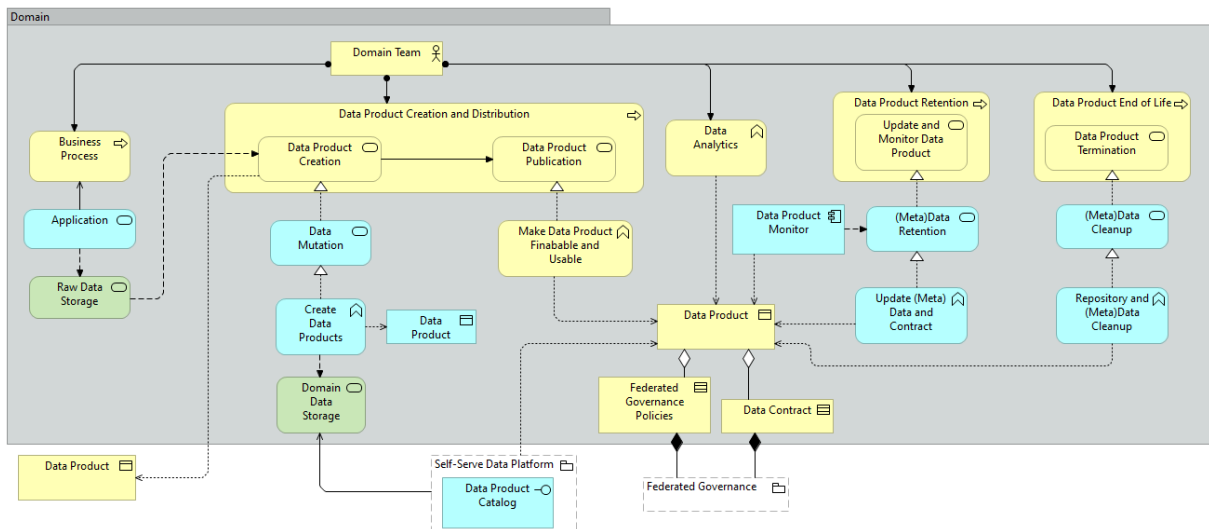


Figure 13 Domain Architecture

The Domain Reference Architecture describes the main processes performed within a domain in a data mesh and the main components supporting these processes. At the top of the architecture is the Domain Team which is responsible for carrying out the processes. Firstly, they are responsible for a 'Business Process' which generates operational data which is collected in some data storage. Secondly, the domain team is responsible for creation, distribution, retention and discontinuation of data products. This process is divided into multiple steps. First, the data product has to be created based on the operational data. Second, the data product has to be accompanied by a data contract and made compliant with the federated governance policies so it can be published in the data product catalog, to be used by other domains. After the data product has been distributed the Domain Team has to concern itself with the proper retention of the data product by monitoring and potentially updating it and eventually discontinuing the data product when it has reached its end of life. Finally, the Domain Team is responsible for performing data analytics.

The Domain team can make use of the capabilities provided on the Self-Serve Data Platform to realize the infrastructure needed and to make data products available to the other domains participating in the data mesh by virtue of the Data Product Catalog.

The Federated Governance layer is an overarching governance structure which specifies standards for communication, documentation of data products and other policies to establish a secure and interoperable data mesh. A more detailed explanation of each of the components can be found in Appendix A.

4.4.2 Self-Serve Data Platform

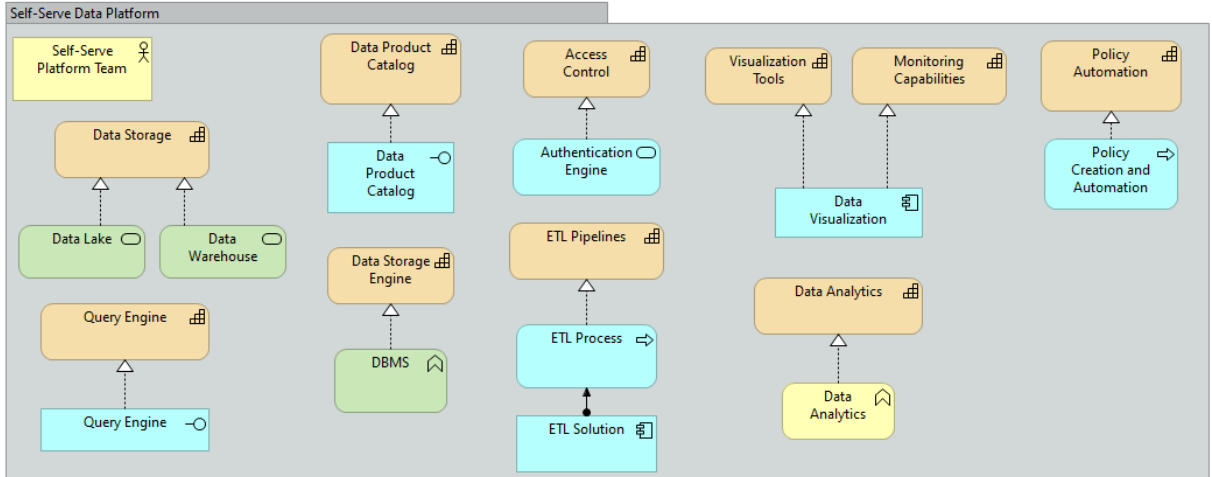


Figure 14 Self-Serve Data Platform Architecture

The Self-Serve Data Platform Architecture entails a collection of capabilities which are provided to the domains participating in the data mesh and is managed by the Self-Serve Platform Team. The capabilities provided by the Self-Serve Platform are not hosted on the platform, Domains still have to implement these technologies in their own environments. The capabilities are represented by the brown blocks, the other blocks are supporting, applications, technologies or processes.

4.4.3 Federated Governance Reference Architecture

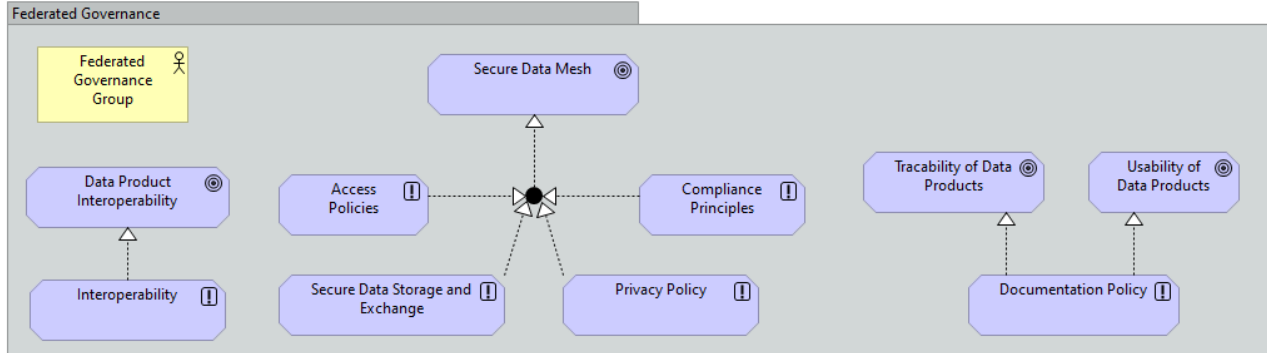


Figure 15 Federated Governance Architecture

In the Federated Governance Architecture of the Data Mesh Reference Architecture key principles to the functioning of the data mesh are entailed. These principles realize certain goals which are needed to make a data mesh function.

One goal is to create a secure data mesh because security can not be overlooked. This goal is supported by having access, compliance, privacy and exchange policies in place which define standards and requirements to achieve a secure data mesh. Additionally, interoperability is a key concern needed to realize interoperability between domains and the data products they exchange. Lastly, a documentation policy is needed to ensure traceability and usability of data products.

5 Artifact Validation

In this chapter the validation method used to validate the treatment, designed in the previous chapter, is discussed. In combination with chapter 6, the 6 and last step of the method by Galster and Avgeriou (2011), the designed data mesh RA will be validated.

To validate the proposed Data Mesh Reference Architecture a choice had to be made between performing a mapping case study or sending out a questionnaire to gather expert opinions. The choice was made to send out a questionnaire to experts with different kinds of roles and varying levels of knowledge regarding Data Mesh and Enterprise Architecture. The choice for a questionnaire was made because a mapping study would not provide many new insights. Additionally, using expert opinion as treatment validation method is in line with the research methodology followed in this study by Wieringa (2014). The reference architecture is developed from scratch based on existing architectures and literature on data mesh components. Therefore, if step 3 of the methodology by Galster and Avgeriou (2011) was performed well, a case study would only prove that all components of the Data Mesh Reference Architecture are also present in solution architectures. However, since this study is aimed at developing a reference architecture from scratch insights into other aspects like perceived usefulness, quality and variability would be more valuable and could help improve the model.

This study draws inspiration from Geerdink (2013) and tries to identify aspects related to the research goal on which to evaluate the Data Mesh Reference Architecture. The research goal for this study as presented in section 1.5 is *to improve the design of data mesh architectures by providing guidance in the strategic design phase. It provides companies with a data mesh reference architecture which guides them in shaping their data mesh architecture.*

Based on this research goal, and the design decisions made in chapter 4, the 3 aspects for evaluation were determined: 'usefulness', 'quality' and 'variability'. These aspects were chosen because to be a valid treatment, the usefulness to guide data mesh solution architecture design has to be assessed. Next, a sufficient quality level has to be achieved, and remarks on the quality can be used for improvement of the model. Lastly, variability is needed because the designed RA has to be useful for multiple organizations and use cases.

5.1 Questionnaire

The whole questionnaire can be found in Appendix B. The front page of the questionnaire contained an introductory text explaining that the questionnaire was conducted as part of a master's thesis, providing contact information, and explaining the purpose of the questionnaire. In the following sections the different parts of the questionnaire are briefly explained.

5.1.1 Questionnaire Introduction

The first section of the questionnaire, the introduction, contained 5 questions. This section of the questionnaire served the purpose of determining the role of the participant and establishing the knowledge of the participant with the concepts related to the Data Mesh Reference Architecture; 'Data Mesh', 'Enterprise Architecture' and 'ArchiMate'. 4 of the 5 introductory questions were mandatory and an option was given for the respondent to also provide the name of the company they work for. The question related to the role had some closed options and an open option for participants with a different role. The questions related to the experience of the participant with the concepts were entirely closed questions.

5.1.2 Usefulness Assessment

The second section of the questionnaire was dedicated to evaluating the perceived usefulness of the Data Mesh Reference Architecture. First an image of the Domain, Self-Serve Data Platform and Federated Governance Architecture were presented to the participants.

Following this the participants were asked to answer 4 mandatory Likert scale questions related to the perceived usefulness of the reference architecture. Each of the Likert scale questions had 5 response items, 1 always being the worst or lowest score and 5 the best or highest score. At the end of the usefulness section a text box was presented in which participants could leave additional remarks relating to the perceived usefulness of the reference architecture.

5.1.3 Quality Assessment

The third section of the questionnaire was dedicated to evaluating the perceived quality of the Data Mesh Reference Architecture. Just like in the usefulness section, the participants were first shown images of the Domain, Self-Serve Data Platform and Federated Governance Architecture, before answering the questions. The participants were asked to answer 5 mandatory Likert scale questions related to the quality of the model. Each of the Likert scale questions had 5 response items, 1 always being the worst or lowest score and 5 the best or highest score. At the end of the quality section a text box was presented in which participants could leave additional remarks relating to the perceived usefulness of the reference architecture.

5.1.4 Variability Assessment

The fourth section of the questionnaire was dedicated to evaluating the variability of the Data Mesh Reference Architecture. Again the participants were first shown images of the Domain, Self-Serve Data Platform and Federated Governance Architecture, before answering the questions. The participants were asked to answer 4 mandatory Likert scale questions related to the quality of the model. Each of the Likert scale questions had 5 response items, 1 always being the worst or lowest score and 5 the best or highest score. At the end of the variability section a text box was presented in which participants could leave additional remarks relating to the variability of the reference architecture.

5.1.5 Additional Feedback

The questionnaire concluded with 3 open and voluntary questions in which the participants were asked to leave any additional remarks or provide points of improvement for specifically the Domain, Self-Serve Data Platform, and Federated Governance Architectures.

5.2 Questionnaire Distribution

The questionnaire was distributed to employees within KPMG, the company at which this study was performed, with known experience of the concepts involved. Additionally, participants were engaged by using the 'Data Mesh Learning' (Data Mesh Learning, 2024) community slack platform.

5.3 Participant Profiles

This section will cover the profiles of the respondents and their knowledge and/or expertise with 'Data Mesh', 'Enterprise Architecture' and 'ArchiMate' based on the answers given in the introduction section of the questionnaire.

Table 15 shows the profiles of the respondents.

Response	Nr of Respondents
Enterprise Architect	4
Data Architect	5
Tech Consultant	7
Data Consultant	7
Data Engineer	4
Manager / Team Lead	3
Other: Professional trainer	1
Other: Chief Innovation Officer	1
Total	32

Table 15 Respondent Work Role

As can be seen in Table 15 for each of the predefined roles at least 3 respondents were found and additionally a Professional Trainer and Chief Innovation Officer filled out the questionnaire.

A total of 19 respondents also answered the optional question and provided the name of the company they work for. Table 16 shows the companies and the amount of respondents.

Company	Nr of Respondents
KPMG	12
CoWork	1
Decideo	1
b.Home	1
Alliander	1
AbeaData	1
Everest Engineering	1
Total	19

Table 16 Participant Company

Figure 16 shows how much of the total percentage of responses was given by respondents with a specific role. The most prevalent roles among the respondents were 'Data Consultant' and 'Tech Consultant'. This study did manage to collect a number of different profiles and at least 3 respondents in each of the main roles as shown in Table 15.

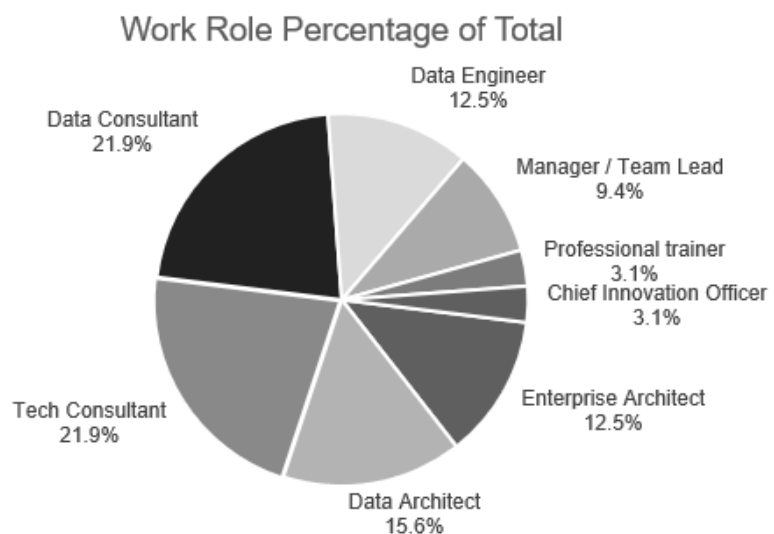


Figure 16 Respondent Work Role Distribution

Figure 17 shows the experience of the participants with the Data Mesh concept. Only 1 respondent was completely unfamiliar with the concept of data mesh. The other 31 respondents had at least heard of the concept. Over 75% of the respondents have theoretical knowledge or even hands on experience with the data mesh concept.

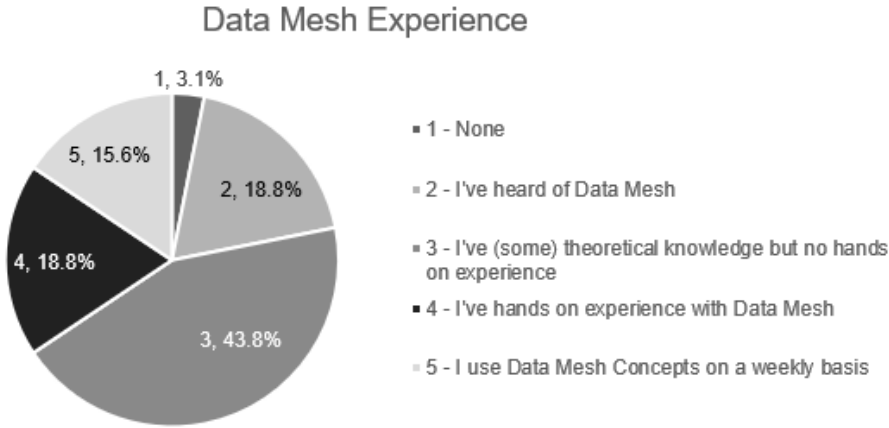


Figure 17 Respondent Data Mesh Experience

Figure 18 shows the experience of the respondents with EA. Again one respondent has no experience with the concept of EA. Almost one third of the respondent uses EA concepts on a weekly basis and over 90% of the respondents have at least theoretical knowledge with EA.

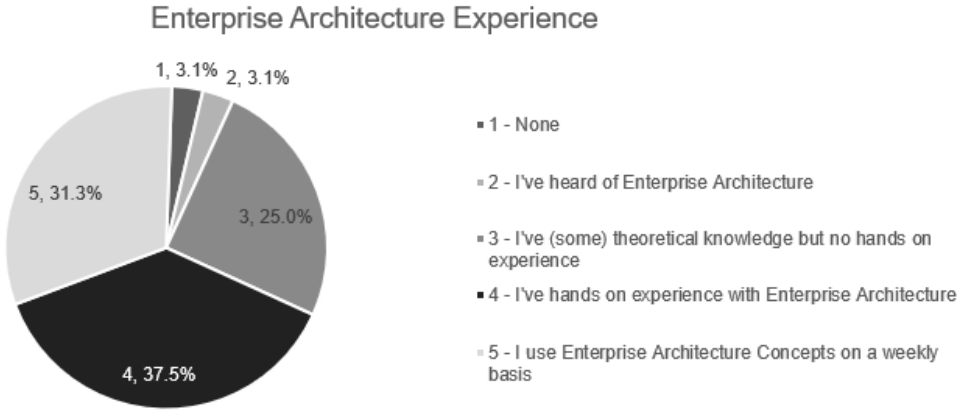


Figure 18 Respondent Enterprise Architecture Experience

Lastly, Figure 19 shows the experience of respondents with the ArchiMate EA modelling language. 6 of the respondents, are unfamiliar with ArchiMate. However, still almost 70% of the respondents have theoretical knowledge or have working experience with ArchiMate.

ArchiMate Experience



Figure 19 Respondent ArchiMate Experience

6 Results

This chapter will cover the results of the questionnaire. The Likert scale answers per respondent can be found in Appendix C so all the results can be verified.

6.1 Usefulness Assessment Results

Figure 20 below shows how the Likert scale responses to each of the questions in the usefulness section of the questionnaire were distributed.

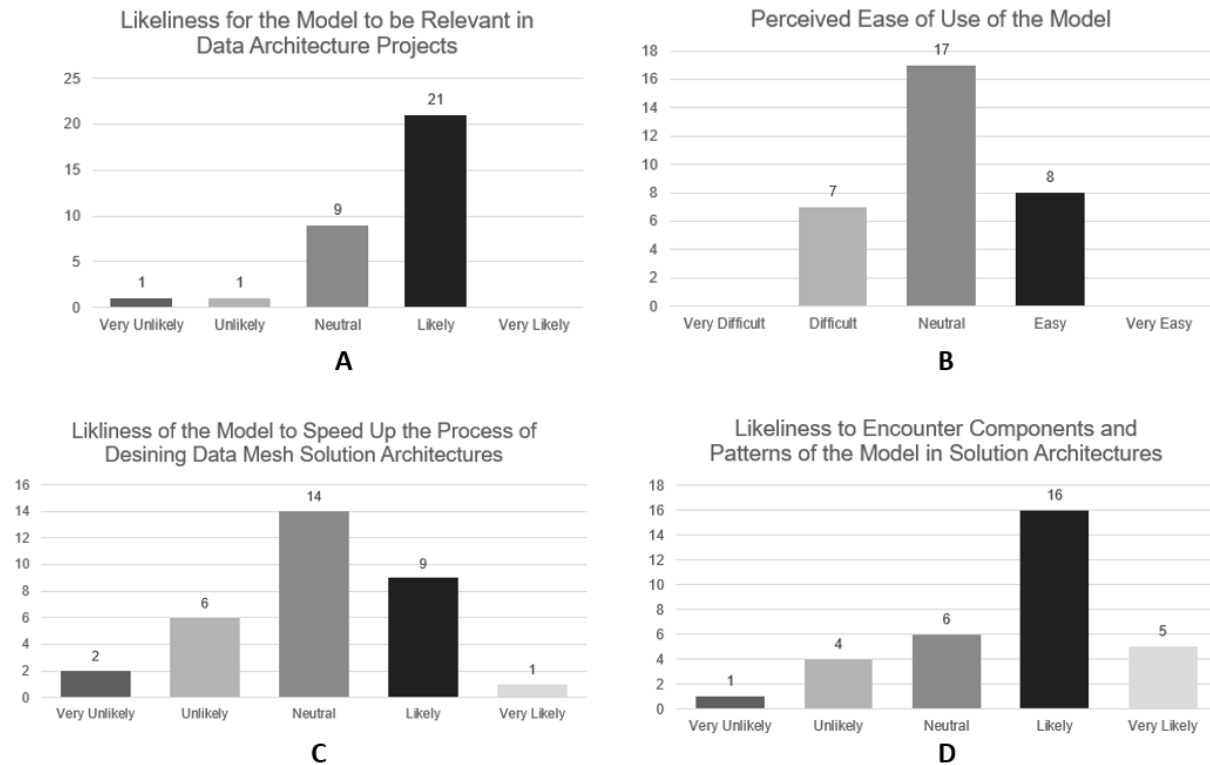


Figure 20 Usefulness Responses Bar Charts

The first noticeable thing is that a positive tendency can be seen in Figure 20-A and Figure 20-D. The majority of respondents thought it would be likely that the Data Mesh Reference Architecture as created in this study would be relevant in Data Architecture projects and accordingly, it is thus perceived likely to encounter components and patterns of the Data Mesh RA in solution architectures. Regarding the Ease of Use, Figure 20-B, the respondents were mostly neutral with almost equal spread to the difficult and easy side. Lastly, the respondents were quite spread out regarding the likeliness of the Data Mesh RA to speed up the process of the designing data mesh solution architectures.

The median and the mode were determined for each of the questions and are shown in Table 17. A slight preference towards the more positive side can be detected in the answers.

Question	Median	Mode
Likely Relevance (Figure 20-A)	4	4
Perceived Ease of Use (Figure 20-B)	3	3
Likeness to Speed Up Design (Figure 20-C)	3	3
Likelihood to Encounter Components and Patterns (Figure 20-D)	4	4
Average	3.5	3.5

Table 17 Median and Mode Usefulness Section

The mode being equal to the median indicates that the distribution of the given answers was relatively symmetrical. That the scores on usefulness of the model are mostly neutral but slightly skewed towards the positive side can also be seen in Table 18 laying out the percentages when dividing the answers into negative, neutral, and positive in which the lowest two scores are combined into a negative sentiment and the highest two scores into a positive sentiment.

<i>Question</i>	<i>Negative</i>	<i>Neutral</i>	<i>Positive</i>
<i>Likely Relevance (Figure 20-A)</i>	6,3%	28,1%	65,6%
<i>Perceived Ease of Use (Figure 20-B)</i>	21,9%	53,1%	25%
<i>Likeness to Speed Up Design (Figure 20-C)</i>	25%	43,7%	31,3%
<i>Likelihood to Encounter Components and Patterns (Figure 20-D)</i>	15,6%	18,8%	65,6%

Table 18 Respondent Sentiment Usefulness Section

What can also be observed in Figure 20-A and 20-D is that regarding those 2 questions the sentiments of the respondents is positive for the majority of respondents. However, for the second and third question, Figure 20-B and Figure 20-C the sentiment is a bit more neutral with a slightly higher tendency towards a positive sentiment than a negative sentiment.

Additionally an interesting take is, to analyse how the respondents from different work roles answered the questions. The median, per question, per role is shown in Table 19.

Role	Median				Average
	Q1	Q2	Q3	Q4	
Data Architect	3	3	2	4	3
Data Consultant	4	3	3	4	3.5
Data Engineer	4	3.5	3.5	4	3.8
Enterprise Architect	4	3.5	3.5	3.5	3.6
Tech Consultant	4	3	3	4	3.5
Manager / Team Lead	3	3	3	4	3.3
Chief Innovation Officer	3	3	2	2	2.5
Professional Trainer	4	2	3	2	2.8

Table 19 Median Per Role Usefulness Section

What can be derived from looking at the median of Q2 is that the data architects think it is unlikely, according to a median of 2, that the Data Mesh RA will speed up the design of data mesh solution architectures. This should not be taken lightly, as data architects can be perceived as the most knowledgeable in this area.

Additionally, we see that the Chief Innovation Officer (CIO) and the Professional Trainer had a more negative perception as both have average medians below 3. On the other hand, the other roles, except the data architects, had a more positive perception with average medians above 3. In terms of the CIO and the Trainer, it must be noted that this is based on the opinion of one person and not multiple (at least 3) like the other roles.

Additionally, it is interesting to examine the lowest scores. In Figure 20-A, 20-C and 20-D it can be observed that 4 times a score of 1 was given by one or multiple of the respondents. 3 of those were given by Data Architect 3. The other 1 score on the question, 'Likelihood for the model to speed up the design of data mesh solution architectures' was given by Data Architect 5. This corresponds well with the view of the data architects being more sceptical about the model and how helpful it will be in designing data mesh architectures. Data Architect 5 gave a score of 5 regarding the likelihood of the components and patterns of the model being encountered in solution architectures.

Data Architect 5 also opted to leave a comment regarding the usefulness: *'I am biased away from Reference Architectures in principal. In over 40 years of data architecture work I have built so many of these that I think they are completely useless artifacts.'*

And

'The diagrams you have drawn may serve as a checklist (as in have we covered this item)'

Thus, even though, in the perception of this Data Architect the model may not be useful during the design of solution architectures, the respondent still believes it can be used as a checklist to validate if a data mesh architectures contains the necessary components.

Lastly, the two most positive respondent groups were the Data Engineers and the Enterprise Architects.

6.2 Quality Assessment Results

Figure 21 show the distribution of the Likert scale responses in the quality section of the questionnaire.

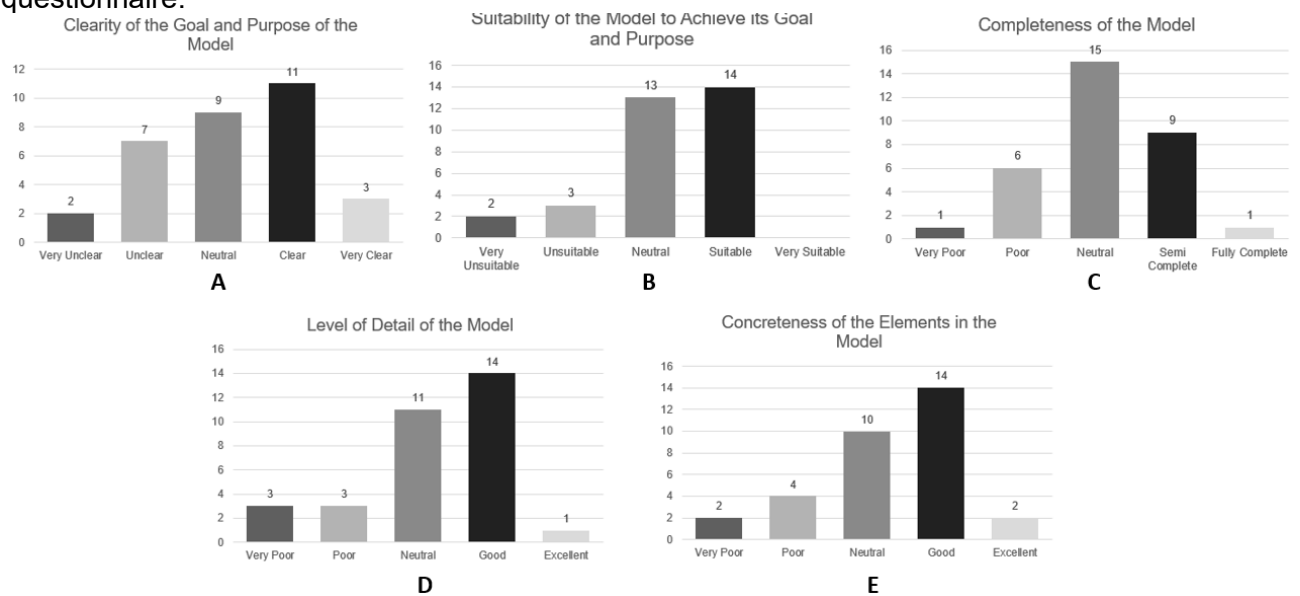


Figure 21 Quality Responses Bar Charts

The same tendency towards neutral and more positive responses can be observed in the quality section as was observed in the usefulness section. Although there were also more 1 and 5 scores given in this section of the questionnaire.

The median and the mode were determined for each of the questions and are shown in Table 20. A slight preference towards the more positive side can be detected in the answers as both medians are above 3. The median related to the quality is the lowest out of the 3 aspects covered in the questionnaire.

Question	Median	Mode
Clarity of Goal and Purpose (Figure 21-A)	3	4
Suitability to Achieve Goals and Purpose (Figure 21-B)	3	4
Completeness (Figure 21-C)	3	3
Level of Detail (Figure 21-D)	3	4
Concreteness of Elements (Figure 21-E)	3.5	4
Average	3.1	3.8

Table 20 Median and Mode Quality Section

The mode is quite a bit higher than the median which indicates that the distribution of the given answers may be negatively skewed.

That the scores on quality of the model are in some cases neutral and in some cases skewed towards the positive side can also be seen in Table 21 showing the percentages when dividing the answers into negative, neutral, and positive in which the lowest two scores are combined into a negative sentiment and the highest two scores into a positive sentiment.

Question	Negative	Neutral	Positive
Clarity of Goal and Purpose (Figure 21-A)	28,1%	28,1%	43,8%
Suitability to Achieve Goals and Purpose (Figure 21-B)	15,6%	40,6%	43,8%
Completeness (Figure 21-C)	21,9%	46,9%	31,2%
Level of Detail (Figure 21-D)	18,7%	34,3%	46,9%
Concreteness of Elements (Figure 21-E)	18,7%	31,3%	50%

Table 21 Respondent Sentiment Quality Section

A good note is that for each question related to the quality of the Data Mesh RA there is a more positive sentiment than negative sentiment. Additionally, in all cases there is also a more positive than neutral sentiment, except regarding the completeness of the model.

Again the median, per role, per question is analysed and show in table 22.

Role	Median					Average
	Q1	Q2	Q3	Q4	Q5	
Data Architect	3	3	3	2	3	2.8
Data Consultant	4	4	3	3	3	3.4
Data Engineer	3.5	3.5	3	4	3.5	3.5
Enterprise Architect	4	4	3	3.5	4	3.7
Tech Consultant	2	3	3	3	4	3
Manager / Team Lead	3	3	4	4	3	3.4
Chief Innovation Officer	3	2	2	2	3	2.4
Professional Trainer	4	4	4	3	2	3.4

Table 22 Median Per Role Quality Section

What can be derived from Table 22 above is that the CIO and the Data Architects have a slightly more negative view on the quality of the model. Again, it has to be noted that the CIO is just one respondent while the Data Architect group consisted of 5 respondents. Mainly the level of detail was poor according to the Data Architects. The Data Architects and CIO are just like in the usefulness section among the more negative. The Professional Trainer however, who had a more positive perception of the quality of the model than the usefulness. Just like in the usefulness section some 1 scores were given.

Data architect 5 gave a score of 1 for every question in the quality and referred back to his comment on the usefulness, *'I am biased away from Reference Architectures in principal. In over 40 years of data architecture work I have built so many of these that I think they are completely useless artifacts.'*

If the scores of Data Architect 5 were not included, because the respondent confirmed bias with his comment, the average of the data architects would be 3.2 instead of 2.8 being much more in line with the overall response to the questionnaire. This would also shift the overall response of the data architects towards the positive side. This can partly explain the negatively skewed distribution indicated by the average mode, 3.8 being higher than the average median, 3.1.

The other 1 score for Q1 was given by Data Architect 4. This respondent also answered Q4 with a score of 1 and provided some additional feedback to clarify these scores, *'This is not a data mesh reference architecture, this is an ontology of data mesh concepts and categories.'*

On Q2 the other 1 score, next to data architect 5, was given by data architect 3. This does show that the data architects had a more critical look on the quality of the model in general. The other 1 scores, were given by Tech Consultant 2 on Q4, and by Manager / Team Lead 1 on Q5.

The most positive respondent groups, just like in the usefulness section, were again the Data Engineers and the Enterprise Architects.

6.3 Variability Assessment Results

Figure 22 show the distribution of the Likert scale responses in the variability section of the questionnaire.

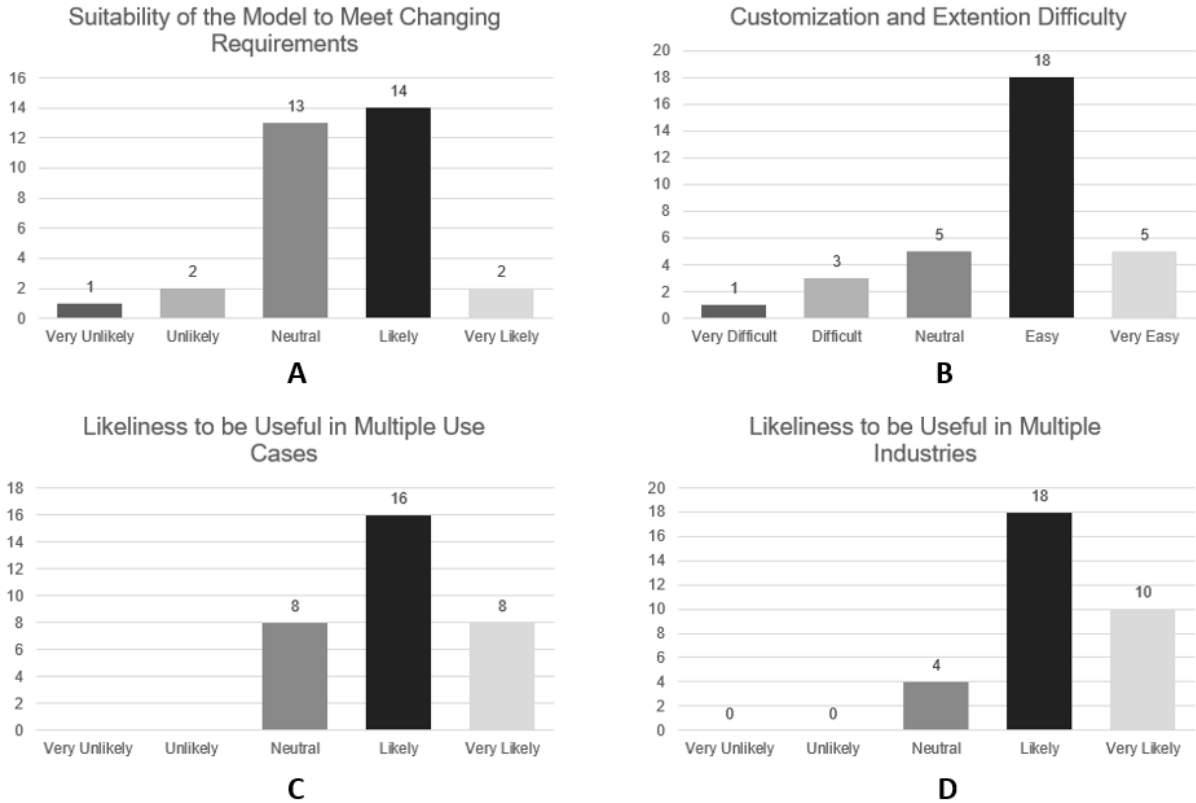


Figure 22 Variability Responses Bar Charts

In the variability section a prominent tendency towards the positive side can be observed in Figure 22-B, 22-C and 22-D. Figure 22-A about the suitability of the model to meet changing requirements is almost balanced between a neutral and a more positive perception. Regarding the likelihood to be useful in multiple use cases and industries, none of the respondents believed this to (very) unlikely.

The median and the mode were determined for each of the questions and are shown in Table 23. A preference towards the more positive side can be detected in the answers.

Question	Median	Mode
Suitability to Meet Changing Requirements (Figure 22-A)	3.5	4
Customization and Extension Difficulty (Figure 22-B)	4	4
Likely to be Useful in Multiple Use Cases (Figure 22-C)	4	4
Likelihood to be Useful in Multiple Industries (Figure 22-D)	4	4
Average	3.9	4

Table 23 Median and Mode Variability Section

The scores on the quality section of the questionnaire are mainly skewed to the positive side which can be observed in Table 24 showing the percentages when dividing the answers into negative, neutral, and positive in which the lowest two scores are combined into a negative sentiment and the highest two scores into a positive sentiment.

Question	Negative	Neutral	Positive
Suitability to Meet Changing Requirements (Figure 22-A)	9,4%	40,6%	50%
Customization and Extension Difficulty (Figure 22-B)	12,5%	15,6%	71,9%
Likely to be Useful in Multiple Use Cases (Figure 22-C)	0%	25%	75%
Likelihood to be Useful in Multiple Industries (Figure 22-D)	0%	12,5%	87,5%

Table 24 Respondent Sentiment Variability Section

What also can be observed in Figure 22-C, Figure 22-D and Table 24 is that none of the respondents had a negative perception regarding Q3 and Q4 in the variability section.

Additionally, it is interesting to analyse how the different roles responded to the questions in the variability section. The median, per role, per question is therefore shown in Table 25.

Role	Median				Average
	Q1	Q2	Q3	Q4	
Data Architect	4	3	4	4	3.8
Data Consultant	3	4	4	4	3.8
Data Engineer	3.5	4	3.5	4	3.8
Enterprise Architect	3.5	4	4	4	3.9
Tech Consultant	4	4	4	4	4
Manager / Team Lead	3	4	4	4	3.8
Chief Innovation Officer	3	4	3	4	3.5
Professional Trainer	5	5	5	5	5

Table 25 Median Per Role Variability Section

In the variability section the answers were quite positive by all of the roles. There is not a clear distinction between the roles other than the Professional Trainer being very positive with regards to variability scoring each of the questions at level 5. Additionally, the Data Architects and the CIO showed a more positive perception than in the usefulness and quality section of the questionnaire.

Both of the 1 scores, as seen in Figure 22-A and Figure 22-B were again given by Data Architect 5. The respondent which confirmed bias against reference architectures in his comment in the quality section.

6.4 Additional Feedback from the Questionnaire

In addition to the closed Likert scale questions each section of the questionnaire was concluded with a text box in which respondents could leave additional remarks related to the models and the theme of that section. At the end of the questionnaire each respondent could also leave additional feedback, or remarks on the ‘Domain Architecture’, the ‘Self-Serve Data Platform Architecture’ and the ‘Federated Governance Architecture’. This section will discuss some of the feedback and remarks by the respondents.

6.4.1 Comments Usefulness Section

Data Architect 1 left an interesting comment related to the usefulness section, “I think models are useful, but we must never fall into the pitfall of thinking that a model represents the whole truth. Even the models must evolve and adapt to learnings as we go.” ... “they are great for communication and facilitate discussions, and you might explore alternatives in early stages quite fast on various model abstraction layers”.

The lesson to be learned from this feedback is that in the perception of this respondent models are a useful tool but models need to be constantly improved and changed as the environment and use cases change. A model has to evolve to stay useful. Data Architect 1 also, agrees with literature that reference architectures are good tools for communication and can facilitate discussion.

Enterprise Architect 1 mentioned that the model would be relevant in situations in which no solution is foreseen or selected yet. Which aligns nicely with one of the goals of the RA which is to support architects in designing data mesh solution architectures, thus in situations in which no solution is foreseen yet.

A point of improvement mentioned by multiple different respondents, the Professional Trainer, Manager / Team Lead 1, Manager / Team Lead 2, Data Consultant 5 and Enterprise Architect 4 is that the model could benefit from more guidance. It was perceived by these respondents as hard to understand and follow. A point of improvement would be adding for example a step-by-step process or a legend explaining different components in the diagrams. A legend on the components in the Domain Architecture is present in this study in appendix A, however, the respondents had no access to this.

Data Engineer 4 noted that data visualization is part of data analytics, especially regarding monitoring of performance and models. The Chief Innovation Officer thought the model should illustrate more different actors. Enterprise Architect 4 agreed with the CIO, and additionally mentioned that the model lacks focus on stakeholders. Lastly, Manager / Team Lead 3 mentioned that it would be helpful to create a diagram showing how the 3 core components come together.

6.4.2 Comments Quality Section

Firstly, Tech Consultant 2 and Data Consultant 7 both mentioned that the model is very generic but it is a good start. However, as also mentioned by Data Architect 1, a RA must adapt and will never be complete and 100% correct. Therefore experimentation, for example, is needed to evolve the model. Thus circling back to the comment made by this respondent in the previous section.

Just like in the usefulness section, the Professional Trainer, Manager / Team Lead 1, Data Consultant 5 and Enterprise Architect 4 again stressed the lack of guidance accompanying the model.

Some points of improvement were also suggested. Enterprise Architect 4 pointed out that an aspect related to data quality is missing and that it would be helpful to map out how domains interact with each other. Enterprise Architect 1 noted that the implication on the business architecture could be more clear and Tech Consultant 3 mentioned that it is not specific enough what audience the model is useful for. The last point would be more clear if the respondents had the opportunity to read the steps taken during the construction of the model however this was not the case.

Tech Consultant 2 mentioned that the level of abstraction is sufficient to allow freedom in choosing technologies when implementing the data mesh while the need for certain technology, tooling and software is dependent on the company and context. Data Architect 5 agreed to this by mentioning that the model is only useful and valuable in the context of the problem that it solves, this is however true for every model.

Manager / Team Lead 3 thinks *“the model gives good guidance as where to work and what should be in place, but the how, is to be figured out by the people with experience.”*

Lastly, Data Architect 4 is of opinion that the created model is not a data mesh reference architecture but an *“ontology of data mesh concepts and categories”*.

6.4.3 Comments Variability Section

Tech Consultant 2 and Enterprise Architect 4 both hinted that for some use cases it would be required to make the model industry specific, or to create industry specific models based on the presented model.

According to Data Architect 1 *“it is not necessary the model itself that decides how easy it is to change the model and let it evolve. It is how the model is used and what you have derived from the model”*. The possibility exists that a shift in the environment requires changes while the model is “tied” to limitations outside it’s own control.

Manager / Team Lead 1 looks at the model differently, and points out that the general nature of the model allows it to be applied in multiple use-cases and industries. However, while also stressing that *“some opinionated elements”* may limit the extension possibilities.

Data Architect 4, again stresses the view that the model is not a RA but more an ontology and therefore the adaptability is quite high.

Data Engineer 4 noted that variability in general is hard. The results of the questionnaire show that even though variability is hard in general, it was managed to achieve variability in the model according to the responses on the questionnaire.

Lastly, Enterprise Architect 4 mentioned that it could be interesting to model the process of evolving data products.

6.4.4 Comments Domain Architecture

The main remark made in this section was on data quality. The Professional Trainer pointed out that data quality management was missing, just like Enterprise Architect 4. The CIO mentioned that the Domain Architecture *“should feature data QoS (Data Quality + Service Level Agreement)”*. Data Architect 4 mentioned maintaining the structure of data when the process or application changes. The main point of improvement regarding the domain architecture therefore is to include data quality management in an improved version of the Data Mesh Reference Architecture.

6.4.5 Comments Self-Serve Data Platform Architecture

Data Architect 2 recommended changing ETL to *“data processing or similar”* as *“ETL is too narrow of a term for all the potential solutions and processes in this area”*.

Data Architect 4, Enterprise Architect 4 and Data Engineer 3 pointed out missing data lineage, and in case of the data architect 4 also master data management components. Additionally, Data Architect 4 mentioned that data visualization tools and monitoring capabilities use different application components for visualization. From these first remarks the main points of improvement would be to rethink the term ETL and change this to a more broad term encompassing a broader range of solutions, adding data lineage components and possible splitting up the data visualization and data monitoring capabilities.

Manager / Team Lead 3 mentioned being confused about the choice for capabilities versus processes or functions. This has to do with trying to communicate the main function of the Self-Serve Data Platform which is to provide capabilities to the domains. However, while still showing the type of process, function or application component that can be used to realize certain capabilities to stay true to the nature of reference architectures.

6.4.6 Comments Federated Governance Architecture

According to Data Consultant 1 the federated governance group *“is very complete and covers all the important aspects”* however, Enterprise Architect 1 and Enterprise Architect 4 disagree. Both respondents pointed out to expect a principle regarding the ownership of data in the Federated Governance Architecture and the later respondent suggested a principle regarding the quality of data could be added.

The CIO mentioned that different actors should be identified to make the model more complete. Manager / Team Lead 3 expected a clear list of principles related to data mesh and found the Federated Governance Architecture vague.

Lastly, Data Consultant 5 left a more general remark about how the concepts of usefulness, quality and variability / adaptability are much related. *“For example, a model that has low quality and is not adaptable is not useful in practice. It is therefore hard to individually score these quality criteria of the model.”*

6.5 Main Questionnaire Takeaways

To conclude the results chapter, in this section the main takeaways and points of improvement will be discussed.

6.5.1 Takeaways Usefulness

The results on the usefulness section show that the model is more likely than not to be useful for developing data mesh solution architectures with an average mean of 3.5 out of 5 based on the 4 questions in the usefulness section of the questionnaire. Around 65% of the respondents thought the model would likely be relevant in data architecture projects. The respondents were more neutral towards the perceived ease of use, and likeliness for the model to speed up the design of data mesh solution architectures. The CIO and Professional Trainer had a negative sentiment regarding the usefulness of the model, a median of 2.5 and 2.8 respectively, however these roles contained only one respondent. The data architects were neutral with an average median of 3 while the data engineers and enterprise architects were the most positive with a median of 3.8 and 3.6 respectively.

The average mean of 3.5 shows there is room for improvement. The model provides a good start but has to evolve. This was also one of the remarks made on the usefulness of the model, models need to constantly improve and be changed to stay useful. Two other points of improvement are to add more guidance to the model to improve the ease of use and add more focus on the stakeholders involved. Lastly, a respondent confirmed the goal that RAs are good tools for communication and can help to facilitate discussions.

6.5.2 Takeaways Quality

The average median of the quality section, 3.1, was the lowest out of the 3 aspect covered in the questionnaire. The responses in the quality section had a slightly more positive sentiment but again there is room for improvement. The CIO again was the respondent with the most negative sentiment with an average median of 2.4. The Professional Trainer had a more positive view regarding the quality than on the usefulness. The data architects also had a more negative sentiment with a median of 2.8. However, it must be noted that one of the data architects confirmed bias away from RAs and only gave 1 scores. When the responses of this data architect are omitted, the average median of the data architects would be 3.2. The tech consultants were neutral with an average median of 3 and the data engineers and enterprise architects again had the most positive sentiments with an average median of 3.5 and 3.7 respectively. The average mode in this section was higher, 3.8, than the median which indicate a negative skew in the distribution of the answers. This can be explained by the answers of the CIO, Professional Trainer and Data Architect 5.

In the additional remarks on the quality of the model, respondents again stressed the need for a RA to evolve and adapt. The lack of guidance accompanying the model was also pointed out again. Additionally, one respondent recommend to map out how domains would interact with each other. Positive notes on the quality of the model are that a respondent mentioned that model gives good guidance about what should be put in place, but the how is to be figured out by experts. Another respondent pointed out the neutrality towards technologies. Both comments acknowledge that one of the goals of the developed model, to stay neutral towards technology and serve as a reference and not solution architecture are achieved according to some of the experts.

6.5.3 Takeaways Variability

The average median of the variability section, 3.9, was the highest out of the 3 aspects covered in the questionnaire. Not a single 1 or 2 score was given regarding the likeliness of the model to be useful in multiple use cases and industries, which proves that the goal to create a RA for general applicability was achieved. The most positive respondent was the Professional Trainer with an average median of 5 which is the highest score possible.

The CIO again had the lowest average median, 3.5, however, this time even the CIO had an overall positive sentiment. The other roles almost scored the same ranging from 3.8 to 4. The only 1 scores in this section were given by the data architect who in earlier sections of the questionnaire confirmed a bias away from RAs.

The same data architect that pointed out the model is not a RA but an ontology, left the same remark in the variability section. Other respondents mentioned that for some uses cases and industry specific model would be required or an industry specific version has to be derived for the general model. Lastly, a respondent mentioned that the general nature allows for application in multiple use cases and industry, but the model can be tied to limitations outside of its control.

6.5.4 Suggested Improvements to the Model

The main improvement point suggested by the respondents regarding the domain architecture, was to add data quality management to the model. Otherwise most respondents perceived the domain architecture to be complete.

Multiple interesting improvement points were suggested regarding the Self-Serve Data Platform. ETL was considered to be too narrow of a term for all the potential solutions and processes in this area thus, renaming this to a more broad term like 'data processing' would improve the model. Additionally, ETL could be modelled as being part of 'data processing'. Next, respondents mentioned missing data lineage and master data management components in the self-serve data platform architecture. Lastly, the remark was made that data visualization tools and monitoring capabilities use different application components for visualization thus a distinction between those two has to be made.

The self-serve data platform can cover as many components as the organization designing it wants, the challenge is in finding the right balance as to what capabilities are provided and what capabilities are not.

Two points of improvement were mentioned regarding the Federated Governance Architecture, which are in line with the improvements suggested on the other two parts of the RA. Respondents pointed out missing principles regarding data ownership and regarding the data quality.

Lastly, two more general improvement points that were suggested were, 1) identifying more actors, and 2) adding more guidance to the model to improve its usability and the quality.

6.5.5 Results Summary

To conclude, for each of the 3 aspects covered in the questionnaire the expert opinion was slightly skewed towards the positive side. The variability aspect scored the highest with an average median of 3.9 out of 5, followed by the usefulness with an average median of 3.5 and finished by the quality with an average median of 3.1. A good start has been made towards developing a data mesh reference architecture, however, there is room for improvement. One of the most important lessons learned is that a RA must keep evolving to stay relevant.

7 Conclusion

In this chapter the research is concluded by answering the main research question formulated in the introduction of this study. This chapter will reflect on the results of the literature review and the design of the Data Mesh Reference Architecture, and the results of the validation will be discussed. Lastly, the implications of this research, its limitations and the directions for future research will be discussed.

The main research question to be answered in this research is: *“How to improve the process of architecting a data mesh by designing a data mesh reference architecture using an enterprise architecture modelling language based on established data mesh structures to provide guidance for solution architects to design solution architectures?”*

7.1 Data Mesh Structures, Components and Considerations

The first knowledge question covered in this research was:

KQ 1: What are the key components constituting a data mesh and what are the limitations?

KQ 1-a: “What different kinds of data mesh archetypes exists?”

This study put forward 4 different data mesh archetypes with different levels of data mesh maturity. The ‘Pure Data Mesh’ archetype, the ‘Semi-Pure Data Mesh’, the ‘Hybrid Data Mesh’ and ‘Distribution Data Mesh’ in order of maturity. These four archetypes were established after analysing data mesh archetypes put forward by other authors, and based on those insights a consolidated list of data mesh archetypes was defined.

KQ 1-b: “What are common components of a data mesh?”

The main architectural components of a data mesh that were identified are domains, a self-serve data platform, and a federated governance layer. These components are made up of different elements.

The domain is the organizational structure in which data ownership and responsibility are decentralized, and in which the domain team is responsible for managing and distributing its data products. The domain team is also responsible for one or multiple operational processes.

The self-serve data platform is managed by the self-serve platform team. This component provides capabilities to the domains in a data mesh. The most important capability provided by the self-serve platform is the ‘data catalog’ on which information about data products and the way to access these is published. Next to this, the self-serve platform provides infrastructure and tools, like data storage and processing capabilities, and monitoring capabilities to the data mesh participants. The goal of the self-serve platform is to enable scalability, efficiency, and allowing domains to independently manage their data.

The federated governance layer is managed by the federated governance group. The federated governance layer entails policies and standards that ensure quality, security and compliance across the data mesh. Additionally, it serves the purpose of setting communication standards and maintaining consistency and interoperability through policies, for example on documentation.

Lastly, literature showed, in line with the findings on the data mesh archetypes, that some data meshes involve a distribution domain. Even though this is not in compliance with the theoretical approach to data mesh it can be a valid option for organizations to make a data mesh work. With a distribution domain a central platform is meant on which data products are published instead of hosting the data product within the domain itself.

KQ 1-c: “What are the limitations of data mesh?”

One of the biggest challenges related to data mesh is that it requires technical knowledge and a certain level of data literacy to be available within the organization. Proper training employees and an assessment of required skills is needed.

Additionally, security of privacy concerns are a challenging factor. Both concerns can be mitigated by establishing proper standards and policies in the federated governance layer, and by putting mechanisms in place to automate and standardize, security and privacy approaches. Management complexity increases because of the distributed nature of a data mesh, and compliance needs to be enforced. Monitoring capabilities and visualization aid in keeping an overview of what happens in the data mesh.

Enforcing data product quality, and data product maintainability are challenging tasks. Proper metadata management, documentation policies and quality standards need to be put in place. Another challenge, is replication of effort, which can be mitigated by establishing a complete self-serve data platform. Lastly, a data mesh will impact the organizations IT landscape and culture, which is costly. Good planning, change management, and a thorough readiness assessment can mitigate unforeseen consequences.

7.2 The Transition to a Data Mesh

The second knowledge questions covered in this research was

KQ 2: Which factors determine if data mesh is a valid approach for an organization?

KQ 2-a: "What are the main indicators to consider the switch to a data mesh?"

A total of 9 motivational factors that drive an organization to switch to a data mesh architecture were identified. 1) The organization needs or wants a more scalable and agile architecture, 2) the organization wants to improve its technical maturity, 3) the company has governance and compliance needs which are easier to enforce by using a federated governance and data mesh approach, 4) the company has to change to tackle existing challenges like data siloes, poor data quality and low interoperability, 5) strategic business objectives drive the organization to adopt a more data-driven approach, 6) an organization needs to be able to adapt faster to the market, 7) an organization wants to improve internal collaboration, 8) an organization wants or needs to improve the quality of data and data operations, and 9) an organization wants to improve collaboration with other parties in the ecosystem or industry.

Additionally, 6 prerequisites were identified that determine the transition to a data mesh is a feasible option. 1) The organization has a need to process high volumes of data in a variety of formats, 2) a certain level of technical knowledge must be present in the organization, 3) data literacy and culture are at a high level within the organization, 4) domains have to be clearly defined, 5) there must be value in breaking up the architecture into different domains, 6) there are sufficient financial resources to make required investments.

KQ 2-b: "What is the impact of data mesh on the existing architecture?"

A data mesh impacts the existing architecture and culture of an organization. The shift to a data mesh architecture means a shift from a monolithic data architecture to a distributed data architecture centred around domains. This creates a demand for new skills and new roles because knowledge needs to be present in each domain, instead of being present in a centralized data team. Responsibilities and tasks will transfer from a central team to decentralized domain teams. New governance models are required to manage the distributed teams and continuous monitoring is needed to ensure compliance and track activity involving data products. Management and coordination complexity will increase. A data mesh does aid in leveraging data as a strategic asset. Next, a data mesh has a long term strategic impact and enforces a culture revolving around data. Lastly, data mesh is a way to reorganize and modernize the data architecture improving scalability and resource allocation.

KQ 2-c: "Which other data methodologies are there?"

The first generation of data platforms were data warehouse platforms. Data warehouse solutions centralize and consolidate structured data from multiple sources for analytical purposes.

Its limitations are the stale nature, difficulties with processing semi and unstructured data and high scaling costs. To tackle these problems two-tier architectures were designed combining data warehouses with data lakes. This approach also allowed for storage of semi and unstructured data and enabled data science and machine learning capabilities. Challenges of two-tier architectures are the complexity of implementing data pipelines, not being able to meet the demand for timely data and separate management of the data warehouse and data lake storages. As a result data lakehouse platforms came into existence trying to maintain the benefits of using both warehouses and lakes while reducing the management overhead of managing both storage solutions separately. The lakehouse approach allows for the low-cost storage of raw data while simultaneously allowing for data warehouse capabilities. Lastly, a data fabric approach was coined with the aim to create a unified data management framework by integrating data flows, and storage and processing technologies. The data fabric approach aims to efficiently leverage data from multiple sources.

7.3 Data Reference Architectures

The third knowledge questions covered in this research was:

KQ 3: Are there existing data mesh reference architectures?

KQ 3-a: “What are characteristics of data reference architectures?”

19 data reference architectures were analysed on 4 different aspects. The focus of the RA, the methodology used to construct the RA, the use of a modelling language and the validation method used. The RAs differed in focus, some were broad and focussed on, for example, the whole big data suite, while others were industry specific, for example, to the energy sector. Not all of the examined RAs were developed according to a scientific methodology. The models that were, used either the framework by Angelov et al. (2012) or the 6 step methodology by Galster and Avgeriou (2011). The most interesting finding was that most authors chose to not use an existing modelling language but create the RA in a free format. Lastly, almost all authors chose to validate the proposed RA by performing a mapping case study, while expert opinion by virtue of a questionnaire was also used.

KQ 3-b: “What parts of other data reference architectures can be re-used?”

One RA tailored to data mesh solutions was identified which provided a lot of useful components when developing a new data mesh reference architecture. These components were also identified during an earlier part of the literature study. However, the analysed data mesh RA was focused on data product exchange in a runtime environment and lacked a clear domain model. The other RAs analysed, mainly focussed on big data solutions, did not provide components or elements to re-use but, did provide insights into RA development and validation of RAs.

7.4 Developing a Reference Architecture

The first design question was:

KQ 4: How to develop a reference architecture?

KQ 4-a: “What are the goals and requirements of a reference architecture?”

A reference architecture’s main purpose is to provide a template which outlines the structure of systems within a specific domain or for a specific type of platform. It is often generalized and serves as a blueprint that guides the design and implementation of concrete architectures. The goals and requirements of a RA depend on its stakeholders, the domain, and practical needs. There are some common goals of RAs. The primary objective, is to standardize architecture approaches, to ensure compatibility, interoperability, and consistency, while encapsulating best practices, facilitating communication among stakeholders, and accelerating the design and development of solution architectures. A RA should be flexible, technology-neutral, well-documented, and maintainable to adapt to different use-cases, systems, and complexities. Lastly, RAs must be updated and validated regularly to remain relevant.

KQ 4-b: "Which method can be used to design and develop the reference architecture?"

Two methods were identified which can be used to develop a reference architecture. The first method is a framework for the analysis and design of software reference architectures that serves 3 purposes, to analyse an existing reference architecture, to design a reference architecture, or to re-design a reference architecture due to changes in the environment. The second method, is a 6 step methodology to create empirically-grounded reference architectures. The later method incorporates some of the steps taken in the first method and has the following 6 steps: 1) decide on the type of RA, 2) selection of the design strategy, 3) empirical collection of data, 4) construction of the RA, 5) enable the RA with variability and, to conclude, 6) evaluate the RA to check it validity.

7.5 Reference Architecture Validation

The last knowledge question was:

KQ 5: How can a reference architecture be validated?

Two validation methods were identified, performing a case study and by using expert opinion. Validation by case study entailed mapping the RA onto solution architectures to validate completeness and compatibility. Expert opinion was gathered by virtue of a questionnaire and used to assess different aspects of the RA, like, maintainability, modularity, reusability, performance and scalability. Expert opinion can be applied to gain a better understanding of how stakeholders of the RA perceive the RA and what could be changed or improved.

7.6 Main Research Question

"How to improve the process of architecting a data mesh by designing a data mesh reference architecture using an enterprise architecture modelling language based on established data mesh structures to provide guidance for solution architects to design solution architectures?"

This research shows that the research methodology followed during this study can be used to design a reference architecture using an enterprise architecture modelling language. The results of this study provide evidence that based on expert opinion the data mesh RA fulfils the requirements on usability, quality and variability. The created data mesh RA will more likely than not be useful to design data mesh solution architectures with an average median of 3.5 out of 5. Next, the quality of the model is slightly above average with a median of 3.1 out of 5. The results additionally show, that the model has good variability with an average median of 3.9 out of 5, the model is thus applicable in multiple industries and for multiple use cases. To conclude, the proposed Data Mesh Reference Architecture improves the process of architecting data mesh solution architectures, while allowing an architect to decide on the type of tools and technologies.

7.7 Contributions of the Research

The first practical contribution of this research, is the data mesh reference architecture which provides a comprehensive blueprint for solution architects guiding them in designing data mesh solution architectures. It details the elements and their relationships and can serve as a checklist of covered elements to a concrete design. The data mesh reference architecture is, within the boundaries of this research, the first data mesh reference architecture developed using the ArchiMate language. Therefore, it can serve as a foundation for new or improved data mesh reference architectures or for the development of industry specific data mesh RAs. Second, this study proposed 4 data mesh archetypes that can help an organization to decide on a data mesh structure fitting to its capabilities and environment. The archetypes help in planning domain boundaries and determining the level of independence.

This research also made important contributions to data mesh literature. First, it extend the research on data meshy by performing a systematic literature review on data mesh structures and components. Next, it analysed challenges and limitations of data meshes and proposed solutions, and mitigation techniques. Following this, motivational factors were identified and perquisites for an organization that wants to transition to a data mesh were defined. Next the impact of a transition to data mesh on the culture and IT infrastructure of an organization was assessed. Additionally, it compared the data mesh approach to other data methodologies providing organizations with insights into alternative data architectures. Lastly, this study showed how an empirically sound reference architecture can be designed, using an EA modelling language, based on an established methodology. Finishing of with demonstrating how a questionnaire can be used to gather expert opinion to validate a RA.

7.7 Limitations of the Research

The first limitation of this research is that it could not fully complete all steps of the engineering cycle from the Design Science Research Methodology as the treatment implementation step was outside the scope. Therefore the practical applicability of the model has not been tested and conclusions regarding this are only based on expert opinions.

Second, even tough a systematic literature review has been performed potentially relevant literature may have been overlooked due to the formulation of the search queries, choice of databases, and the inclusion and exclusion criteria.

Next, using a single group of respondents with different roles and variability in the levels of knowledge and expertise regarding the subjects can lead to inconsistencies in responses. Additionally, for roles with a lower number of respondents, personal biases and subjectivity play a bigger part in the final results. Preferably, each of the roles has the same number of respondents and multiple groups of respondents would be used to enable more advanced statistical analyses.

Using a questionnaire to gather expert opinions also has some limitations. Closed questions limited the depth and richness of responses. Additionally, questionnaires have an inflexible nature, complex issues may be oversimplified and answers and questions may lack contextual understanding.

Lastly, while the developed data mesh reference architecture showed promising results, effectiveness in practice may be constrained by requirements and regulations outside of its control. Modification or adaptations may be required to fit certain use cases and it has to be stated that it is not a representation of the whole truth.

7.8 Future Research

As shortly mentioned in the previous section, future work is needed to validate the data mesh reference architecture in practice. To validate the general applicability of the model, case studies can be performed in multiple industries. Additionally, case studies could be performed in which one test group uses the model to design a data mesh solution architecture, and the other test group does not, to validate how it affects the process of designing a solution architecture in terms of efficiency.

Next, the improvement points discussed in the results of this research can be implemented to examine how these improvements influence the perceived usefulness, quality and variability of the model. Other validation methods like interviews can also be considered to gain a more in depth understanding related to different aspects of the model.

Further research could also explore potential extensions to the model or derive industry specific data mesh reference architectures.

Finally, because data mesh is a rapidly evolving concept, future research is needed to update this research with new findings from theory and practice, for example, regarding best practices or different archetypes.

References

- Alcala, J. L. (2022, May 13). *Adidas Data Mesh Journey: Sharing Data Efficiently at Scale*. <https://medium.com/adidoescode/adidas-data-mesh-journey-sharing-data-efficiently-at-scale-c50ee671fbd7>
- Aldea, A. (2023). Current Challenges and Opportunities in Enterprise Architecture: Insights from 950+LeanIX Customers. In S. and I. M. E. Griffo Cristine and Guerreiro (Ed.), *Advances in Enterprise Engineering XVI* (pp. 17–30). Springer Nature Switzerland.
- Angelov, S., Grefen, P., & Greefhorst, D. (2012). A framework for analysis and design of software reference architectures. *Information and Software Technology*, 54(4), 417–431. <https://doi.org/https://doi.org/10.1016/j.infsoc.2011.11.009>
- Araújo Machado, I., Costa, C., & Santos, M. Y. (2022). Advancing Data Architectures with Data Mesh Implementations. In A. De Weerd Jochen and Polyvyanyy (Ed.), *Intelligent Information Systems* (pp. 10–18). Springer International Publishing.
- Ashraf, A., Hassan, A., & Mahdi, H. (2023). Key Lessons from Microservices for Data Mesh Adoption. *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 1–8. <https://doi.org/10.1109/MIUCC58832.2023.10278300>
- Azeroual, O., & Nacheva, R. (2023). Data Mesh for Managing Complex Big Data Landscapes and Enhancing Decision Making in Organizations. *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KMIS*, 202–212. <https://doi.org/10.5220/0012195700003598>
- Berntsson-Svensson, R., & Taghavianfar, M. (2020). Toward Becoming a Data-Driven Organization: Challenges and Benefits. *Research Challenges in Information Science*. <https://api.semanticscholar.org/CorpusID:220254713>
- Bode, J., Kühn, N., Kreuzberger, D., Hirschl, S., & Holtmann, C. (2023). *Towards Avoiding the Data Mess: Industry Insights from Data Mesh Implementations*.
- Butte, V. K., & Butte, S. (2022). Enterprise Data Strategy: A Decentralized Data Mesh Approach. *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 62–66. <https://doi.org/10.1109/ICDABI56818.2022.10041672>
- Cambridge University Press & Assessment. (2024). *Cambridge Dictionary*.
- Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9, 101895. <https://doi.org/https://doi.org/10.1016/j.mex.2022.101895>
- Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E., & Bone, M. (2010). The Concept of Reference Architectures. *Systems Engineering*, 13(1), 14–27. <https://doi.org/https://doi.org/10.1002/sys.20129>
- Dahdal, S., Poltronieri, F., Tortonesi, M., Stefanelli, C., & Suri, N. (2023). A Data Mesh Approach for Enabling Data-Centric Applications at the Tactical Edge. *2023 International Conference on Military Communications and Information Systems (ICMCIS)*, 1–9. <https://doi.org/10.1109/ICMCIS59922.2023.10253568>
- Data Mesh Learning. (2024). *Data Mesh Learning*. <https://datameshlearning.com/>
- Databricks. (2020, July 28). *Data Mesh in Practice: How Europe's Leading Online Platform for Fashion Goes Beyond Data Lake*. Youtube.Com. <https://www.youtube.com/watch?v=eiUhV56uVUc>
- De Almeida Neto, F. A., & Castro, A. (2017). A reference architecture for educational data mining. *2017 IEEE Frontiers in Education Conference (FIE)*, 1–8. <https://doi.org/10.1109/FIE.2017.8190728>
- Dehghani, Z. (2019, May 20). *How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh*. MartinFowler.Com. <https://martinfowler.com/articles/data-monolith-to-mesh.html>
- Dehghani, Z. (2020, December 3). *Data Mesh Principles and Logical Architecture*. MartinFowler.Com. <https://martinfowler.com/articles/data-mesh-principles.html>
- Dela Cruz, N., Tobin, M., Schenz, G., & Barden, D. (2011). Enterprise Data Architecture: Development Scenarios Using ORM. In T. and H. P. Meersman Robert and Dillon (Ed.),

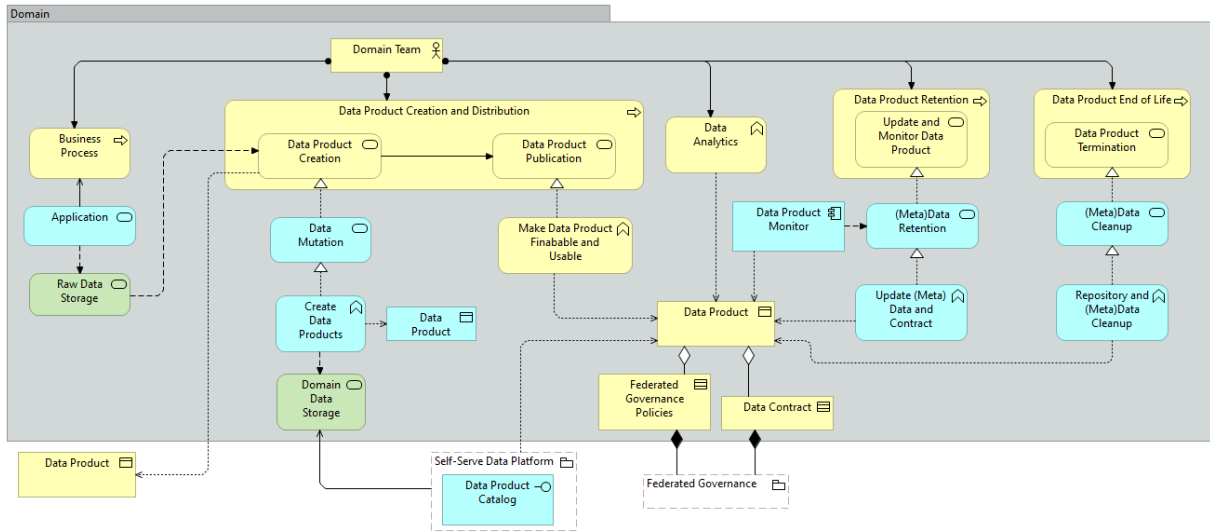
- On the Move to Meaningful Internet Systems: OTM 2011 Workshops* (pp. 278–287). Springer Berlin Heidelberg.
- Dibouliya, A., & Jotwani, Dr. V. (2023). Review on Data Mesh Architecture and its Impact. *Journal of Harbin Engineering University*, 44(7), 2353–2363. <https://harbinengineeringjournal.com/index.php/journal/article/view/809>
- Divya, J., Sheetal, P., & Madhu, P. R. (2021). Data Governance in Data Mesh Infrastructures: The Saxo Bank Case Study. *The 21st International Conference on Electronic Business*, 599–604. <https://aisel.aisnet.org/iceb2021/52>
- Dončević, J., Fertalj, K., Brčić, M., & Kovač, M. (2022). *Mask-Mediator-Wrapper architecture as a Data Mesh driver*. <https://doi.org/10.1109/TSE.2024.3367126>
- Driessen, S., van den Heuvel, W.-J., & Monsieur, G. (2023). ProMoTe: A Data Product Model Template for Data Meshes. In J. and G. G. and L. S. and Z. J. Almeida João Paulo A. and Borbinha (Ed.), *Conceptual Modeling* (pp. 125–142). Springer Nature Switzerland.
- Falconi, M., & Plebani, P. (2023). Adopting Data Mesh principles to Boost Data Sharing for Clinical Trials. *2023 IEEE International Conference on Digital Health (ICDH)*, 298–306. <https://doi.org/10.1109/ICDH60066.2023.00051>
- Fortune Business Insights. (2024). *Big Data Analytics Market Size, Share & COVID-19 Impact Analysis 2023-2030*. <https://www.fortunebusinessinsights.com/big-data-analytics-market-106179>
- Galster, M., & Avgeriou, P. (2011). Empirically-grounded reference architectures: a proposal. *Proceedings of the Joint ACM SIGSOFT Conference – QoSA and ACM SIGSOFT Symposium – ISARCS on Quality of Software Architectures – QoSA and Architecting Critical Systems – ISARCS*, 153–158. <https://doi.org/10.1145/2000259.2000285>
- Garises, V., & Quenum, J. (2018). The road towards big data infrastructure in the health care sector: The case of Namibia. *2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON)*, 98–103. <https://doi.org/10.1109/MELCON.2018.8379075>
- Geerdink, B. (2013). A reference architecture for big data solutions introducing a model to perform predictive analytics using big data technology. *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*, 71–76. <https://doi.org/10.1109/ICITST.2013.6750165>
- Goedegebuure, A., Kumara, I., Driessen, S., Di Nucci, D., Monsieur, G., Heuvel, W. van den, & Tamburri, D. A. (2023). *Data Mesh: a Systematic Gray Literature Review*.
- Graetsch, U. M., Khalajzadeh, H., Shahin, M., Hoda, R., & Grundy, J. (2023). Dealing With Data Challenges When Delivering Data-Intensive Software Solutions. *IEEE Transactions on Software Engineering*, 49(9), 4349–4370. <https://doi.org/10.1109/TSE.2023.3291003>
- Hendriks, K. W. (2023). *Data Governance Structures in Data Mesh Architectures*. <http://essay.utwente.nl/94999/>
- Hermawan, R. A., & Sumitra, I. D. (2019). Designing Enterprise Architecture Using TOGAF Architecture Development Method. *IOP Conference Series: Materials Science and Engineering*, 662(4), 42021. <https://doi.org/10.1088/1757-899X/662/4/042021>
- Heuser, L., Scheer, J., den Hamer, P., de Lathouwer, B., Cox, A., Parslow, P., Kempen, B., Klien, E., & Lonien, J. (2018). *REFERENCE ARCHITECTURE & DESIGN PRINCIPLES EIP SCC WORK STREAM 2-MAIN DELIVERABLE*.
- Hokkanen, S. (2021). *Utilization of Data Mesh Framework as a Part of Organization's Data Management* [Master's Thesis, University of Eastern Finland]. <http://urn.fi/urn:nbn:fi:uef-20211359>
- Hooshmand, Y., Resch, J., Wischnewski, P., & Patil, P. (2022). From a Monolithic PLM Landscape to a Federated Domain and Data Mesh. *Proceedings of the Design Society*, 2, 713–722. <https://doi.org/10.1017/pds.2022.73>
- Jonkman, C. (2023). *Organisational Maturity Assessment During the Paradigm Shift from Monoliths to Data Mesh* [Master Thesis, Delft University of Technology]. <http://resolver.tudelft.nl/uuid:294d7df5-511c-4149-9507-21be6379375d>
- Kancharla, J. R., & Madhu Kumar, S. D. (2023). Breaking Down Data Silos: Data Mesh to Achieve Effective Aggregation in Data Localization. *2023 International Conference on*

- Computer, Electronics & Electrical Engineering & Their Applications (IC2E3)*, 1–5.
<https://doi.org/10.1109/IC2E357697.2023.10262765>
- Kim, N., & Park, J. (2022, April). *Can Data Save Small Businesses? Benefits and Challenges of Big Data Analytics Adoption among Small-sized Clothing Retailers*.
<https://doi.org/10.31274/itaa.15739>
- Kraska, T., Li, T., Madden, S., Markakis, M., Ngom, A., Wu, Z., & Yu, G. X. (2023). Check Out the Big Brain on BRAD: Simplifying Cloud Data Processing with Learned Automated Data Meshes. *Proc. VLDB Endow.*, 16(11), 3293–3301.
<https://doi.org/10.14778/3611479.3611526>
- Krystek, M., Mazurek, C., Morzy, M., & Pukacki, J. (2023). Introducing Data Mesh Paradigm for Smart City Platforms Design. *The 56th Hawaii International Conference on System Sciences*, 6885–6892. <https://hdl.handle.net/10125/103468>
- Li, J., Cai, S., Wang, L., Li, M., Li, J., & Tu, H. (2022). A novel design for Data Processing Framework of Park-level Power System with Data Mesh concept. *2022 IEEE International Conference on Energy Internet (ICEI)*, 153–158.
<https://doi.org/10.1109/ICEI57064.2022.00032>
- Lombardo, G. (2023). *Data Friction in Data Sharing: a Physics Inspired Model* [Master Thesis, Politecnico Milano]. <https://hdl.handle.net/10589/215410>
- Machado, I. A. (2022). *Proposal of an approach for the design and implementation of a data mesh* [Master's Thesis, Universidade do Minho]. <https://hdl.handle.net/1822/82290>
- Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures. *Procedia Computer Science*, 196, 263–271.
<https://doi.org/https://doi.org/10.1016/j.procs.2021.12.013>
- Machado, I., Costa, C., & Santos, M. Y. (2021). Data-Driven Information Systems: The Data Mesh Paradigm Shift. In E. Insfran, F. Gonzalez, S. Abrahao, M. Fernandez, C. Barry, H. Linger, M. Lang, & C. Schneider (Eds.), *Information Systems Development: Crossing Boundaries between Development and Operations (DevOps)*. AISEL.
<https://aisel.aisnet.org/isd2014/proceedings2021/currenttopics/9/>
- McEachen, N., & Lewis, J. (2023). Enabling Knowledge Sharing By Managing Dependencies and Interoperability Between Interlinked Spatial Knowledge Graphs. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-4/W7-2023*, 117–124. <https://doi.org/10.5194/isprs-archives-XLVIII-4-W7-2023-117-2023>
- Nakagawa, E. Y., Oquendo, F., & Becker, M. (2012). RAModel: A Reference Model for Reference Architectures. *2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture*, 297–301.
<https://doi.org/10.1109/WICSA-ECSA.212.49>
- Niemi, E. I. (2008). Enterprise Architecture Benefits: Perceptions from Literature and Practice. *The 7th IBIMA Conference Internet & Information Systems in the Digital Age*.
<https://api.semanticscholar.org/CorpusID:14219660>
- Pakrashi, A., Wallace, D., Mac Namee, B., Greene, D., & Guéret, C. (2023). CowMesh: a data-mesh architecture to unify dairy industry data for prediction and monitoring. *Frontiers in Artificial Intelligence*, 6. <https://doi.org/10.3389/frai.2023.1209507>
- Panigrahy, S., Dash, B., & Thatikonda, R. (2023). From Data Mess to Data Mesh: Solution for Futuristic Self-Serve Platforms. *IJARCCCE*, 12(4), 677–683.
<https://doi.org/10.17148/IJARCCCE.2023.124121>
- Podlesny, N. J., Kayem, A. V. D. M., & Meinel, C. (2022). CoK: A Survey of Privacy Challenges in Relation to Data Meshes. In A. and K. G. and T. A. M. and K. I. Strauss Christine and Cuzzocrea (Ed.), *Database and Expert Systems Applications* (pp. 85–102). Springer International Publishing.
- Pongpech, W. A. (2023). A Distributed Data Mesh Paradigm for an Event-based Smart Communities Monitoring Product. *Procedia Computer Science*, 220, 584–591.
<https://doi.org/https://doi.org/10.1016/j.procs.2023.03.074>

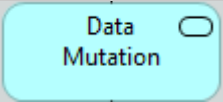
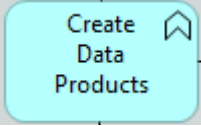
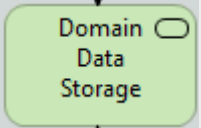
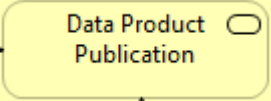
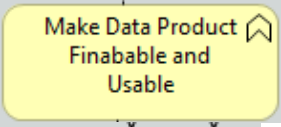
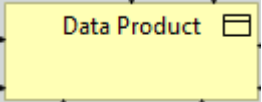


- Priebe, T., Neumaier, S., & Markus, S. (2021). Finding Your Way Through the Jungle of Big Data Architectures. *2021 IEEE International Conference on Big Data (Big Data)*, 5994–5996. <https://doi.org/10.1109/BigData52589.2021.9671862>
- Restel, H. (2023). *SimulationOps - Towards a simulation as-a-service platform for resilient societies using a cross-domain data mesh*. <https://publica.fraunhofer.de/handle/publica/443164>
- Sang, G. M., Xu, L., & de Vrieze, P. (2016). A reference architecture for big data systems. *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, 370–375. <https://doi.org/10.1109/SKIMA.2016.7916249>
- Sang, G. M., Xu, L., & de Vrieze, P. (2017). Simplifying Big Data Analytics Systems with a Reference Architecture. In H. and F. R. Camarinha-Matos Luis M. and Afsarmanesh (Ed.), *Collaboration in a Data-Rich World* (pp. 242–249). Springer International Publishing.
- Sanyoto, A., & Saputra, M. (2023). ArchiMate's Strengths and Weaknesses as EA Modeling Language: A Systematic Mapping Study. *International Conference on Informatics and Computing (ICIC)*, 1–6. <https://doi.org/10.1109/ICIC60109.2023.10381985>
- Sedlak, B., Casamayor Pujol, V., Donta, P. K., Werner, S., Wolf, K., Falconi, M., Pallas, F., Dustdar, S., Tai, S., & Plebani, P. (2023). Towards Serverless Data Exchange Within Federations. In J. and D. S. and L. F. Aiello Marco and Barzen (Ed.), *Service-Oriented Computing* (pp. 144–153). Springer Nature Switzerland.
- Strengtholt, P. (2022, May 24). *Data Mesh: Topologies and domain granularity*. Towards Data Science. <https://towardsdatascience.com/data-mesh-topologies-and-domain-granularity-65290a4ebb90>
- TheOpenGroup. (n.d.-a). *ArchiMate 3.2 Specification Language Structure*. TheOpenGroup. Retrieved March 20, 2024, from <https://pubs.opengroup.org/architecture/archimate3-doc/ch-Language-Structure.html>
- TheOpenGroup. (n.d.-b). *The ArchiMate Enterprise Architecture Modeling Language*. TheOpenGroup. Retrieved March 20, 2024, from <https://www.opengroup.org/archimate-forum/archimate-overview>
- TheOpenGroup. (n.d.-c). *The TOGAF Standard, Version 9.2*. TheOpenGroup. Retrieved March 20, 2024, from <https://pubs.opengroup.org/architecture/togaf9-doc/arch/>
- Vestues, K., Hanssen, G. K., Mikalsen, M., Buan, T. A., & Conboy, K. (2022). Agile Data Management in NAV: A Case Study. In K.-J. and P. M. and K. P. Stray Viktoria and Stol (Ed.), *Agile Processes in Software Engineering and Extreme Programming* (pp. 220–235). Springer International Publishing.
- Vinnikainen, O. (2023). *Data Mesh: a Holistic Examination of its Principles, Practices and Potential* [Master's Thesis, Lahti University of Technology]. <https://urn.fi/URN:NBN:fi-fe20230920133907>
- Vlasiuk, Y., & Onyshchenko, V. (2023). Data Mesh as Distributed Data Platform for Large Enterprise Companies. In I. and H. M. Hu Zhengbing and Dychka (Ed.), *Advances in Computer Science for Engineering and Education VI* (pp. 183–192). Springer Nature Switzerland.
- Voß, C. (2022). Identifying Alternatives and Deciding Factors for a Data Mesh Architecture. *SKILL 2022 Gesellschaft Für Informatik, Bonn*, 93–99. <https://dl.gi.de/items/324d48dd-373f-48e9-a602-cb82ffdf469>
- Wider, A., Verma, S., & Akhtar, A. (2023). Decentralized Data Governance as Part of a Data Mesh Platform: Concepts and Approaches. *2023 IEEE International Conference on Web Services (ICWS)*, 746–754. <https://doi.org/10.1109/ICWS60048.2023.00101>
- Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering* (1st ed.). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-43839-8>
- Zaharia, M. A., Ghodsi, A., Xin, R., & Armbrust, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. *Conference on Innovative Data Systems Research*. <https://api.semanticscholar.org/CorpusID:229576171>

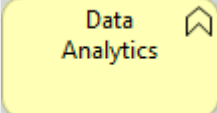
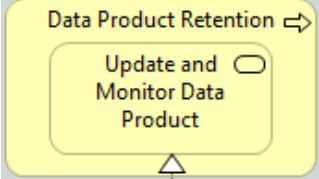
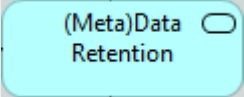

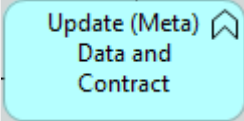
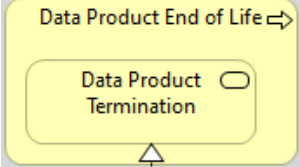
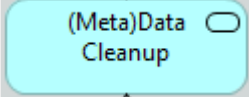
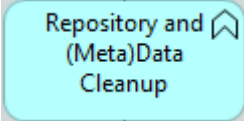
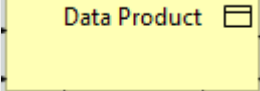
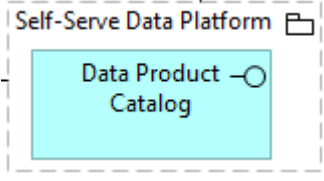
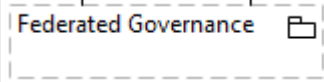
Appendix

A Domain Architecture and Component Explanation



Component	Explanation	Relationships
	Represents the actors within a domain. The domain team is responsible for 4 business processes and 1 business function	'Business Process' 'Data Product Creation and Distribution' 'Data Analytics' 'Data Product Retention' 'Data Product End of Life'
	A Business Process performed by the domain and served by an application.	'Domain Team' 'Application'
	The Application serves the business process and data flows to the Raw Data Storage	'Business Process' 'Raw Data Storage'
	The data created by the business process and the supporting application is collected in a Raw Data Storage. This data is the input for the Data Product Creation service	'Application' 'Data Product Creation'
	The Data Product Creation and Distribution process is realized by two business services, data product creation and publication	'Domain Team' 'Data Product Creation' 'Data Product Publication'
	Business service responsible for the creation of data products realized by the data mutation application. Part of the Data Product	'Data Product Creation and Distribution' 'Data Mutation'

	Creation and Distribution process.	
	Application service realizing the Data Product Creation and realized by the Create Data Products application function	'Data Product Creation' 'Create Data Products'
	Application function realizing the Data Mutation, creating the Data Product and its data flows to the Domain Data Storage	'Data Mutation' 'Data Product' 'Domain Data Storage'
	In the Domain Data Storage Data Products are stored. Data Products are accessible via the Data Product Catalog which is part of the Self-Serve Data Platform	'Create Data Products' 'Data Product Catalog'
	Business services responsible for making the data product accessible and usable for other domains. Part of the Data Product Creation and Distribution process.	'Data Product Creation and Distribution' 'Make Data Product Findable and Usable'
	The business function realizing the Data Product Publication	'Data Product Publication' 'Data Product'
	The Data Product is now a business object because it has value for the business. It is published in the data product catalog, compliant with policies and accompanied by a data contract. Additionally, it is continuously monitored and potentially updated or discontinued.	'Make Data Product Findable and Usable' 'Data Product Catalog' 'Federated Governance Policies' 'Data Contract' 'Data Analytics' 'Data Product Monitor' 'Update (Meta)Data and Contract' 'Repository and (Meta)Data Cleanup'
	The Federated Governance Policies from the Federated Governance layer are applied to the Data Product to make it compliant with agreed upon mesh standards	'Data Product' 'Federated Governance'
	Specifies information about the structure, semantics, usage policies and lineage.	'Data Product' 'Federated Governance'

	<p>Data Analytics function performed by the Domain Team on the Data Product</p>	<p>'Domain Team' 'Data Product'</p>
	<p>Business Process responsible for monitoring the Data Product and updating it and its metadata. Realized by the Update and Monitor Data Product Service.</p>	<p>'Domain Team' 'Update and Monitor Data Product'</p>
	<p>Application realizing the Data Product Retention process. Has a monitoring component and an update function.</p>	<p>'Update and Monitor Data Product' 'Data Product Monitor' 'Update (Meta)Data and Contract'</p>
	<p>Component of the (Meta)Data Retention application continuous monitoring access to and change of the Data Product</p>	<p>'(Meta)Data Retention' 'Data Product'</p>
	<p>Application Function realizing the (Meta)Data Retention</p>	<p>'(Meta)Data Retention' 'Data Product'</p>
	<p>Business Process responsible for discontinuing a Data Product. Realized by the Data Product Termination Service.</p>	<p>'Domain Team' 'Data Product' 'Termination'</p>
	<p>Application realizing the Data Product End of Life process. Has a cleanup function.</p>	<p>'Data Product Termination' 'Repository and (Meta)Data Cleanup'</p>
	<p>Application function cleaning up a data product and its metadata</p>	<p>'(Meta)Data Cleanup' 'Data Product'</p>
<p>(External Domain Data Product)</p> 	<p>Data Product produced by another domain which may be input for another data product</p>	<p>'Data Product Creation'</p>
	<p>The Self-Serve Data Platforms provisions the Data Product Catalog and provides capabilities, like storage and monitoring for example, to the domains.</p>	<p>'Domain Data Storage' 'Data Product'</p>
	<p>In the Federated Governance policies and standards are established</p>	<p>'Federated Governance Policies' 'Data Contract'</p>

B Data Mesh Reference Architecture Evaluation Questionnaire

B.1 Questionnaire Introduction

Data Mesh Reference Architecture Evaluation Questionnaire

All responses to this questionnaire are completely anonymous and no personal information (other than your work role) will be collected.

If you have any questions or remarks regarding this questionnaire feel free to contact me at: d.r.vanderwerf@student.utwente.nl

This questionnaire is conducted as part of a master's thesis research performed by, Daniel van der Werf, student at the University of Twente.

The research goal is to design a 'Data Mesh Reference Architecture to provide organizations guidance in desining their data mesh architecture'.

The purpose of this questionnaire is to present you with the 3 main architectural components which constitute a Data Mesh and evaluate these based on 3 aspects. The 3 aspects to evaluate are: 'usefulness', 'quality' and 'variability' of the proposed Data Mesh Reference Architecture.

The survey consists of a short introduction section in which you will be asked for your role and experience with Data Mesh and Enterprise Architecture.

The questions related to the 3 different aspects are '1 - 5' scales with '1' always being the lowest or worst score and '5' the highest or best score.

Each section will also have one (not required) open question in which you can leave additional comments related to the aspect covered in that section.

The survey will take around 10 minutes to complete.

Thankyou in advance!

* Indicates required question

B.1 Continued

Work Role *

- Enterprise Architect
- Data Architect
- Tech Consultant
- Data Consultant
- Data Engineer
- Data Scientist
- Manager / Team Lead
- Other: _____

Company (optional)

Your answer _____

Knowledge and/or expertise with 'Data Mesh' *

- None
- I've heard of Data Mesh
- I've (some) theoretical knowledge but no hands on experience
- I've hands on experience with Data Mesh
- I use Data Mesh concepts on a weekly basis

B.1 Continued

Knowledge and/or expertise with 'Enterprise Architecture' *

- None
- I've heard of Enterprise Architecture
- I've (some) theoretical knowledge but no hands on experience
- I've hands on experience with Enterprise Architecture
- I use Enterprise Architecture concepts on a weekly basis

Knowledge and/or expertise with 'ArchiMate' (EA Modeling Language) *

- None
- I've heard of ArchiMate
- I've (some) theoretical knowledge but no hands on experience
- I've hands on experience with ArchiMate
- I use ArchiMate on a weekly basis

Next



Page 1 of 5

Clear form

B.2 Questionnaire Section Usefulness

Usefulness

This section is dedicated to answering questions related to the perceived usefulness of the Data Mesh Reference Architecture

Data Mesh Reference Architecture viewpoints

The 3 pictures below show the 3 architectural components making up a Data Mesh

- Domain
- Self-Serve Data Platform
- Federated Governance Layer

For each of the questions take all 3 of the viewpoints in consideration.

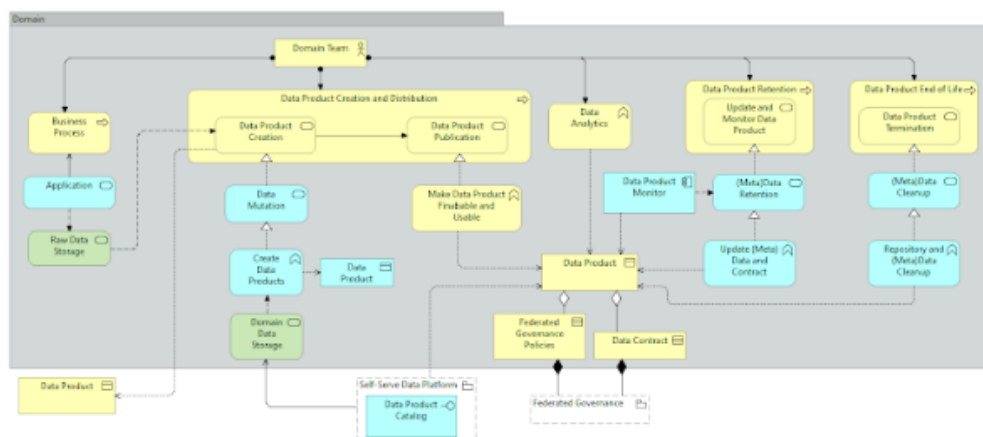
The Data Mesh Reference Architecture will hereafter be referred to as **'the model'**

Domain Architecture

A few important points regarding the Domain Architecture:

- A Domain Team has responsibility over a business process, the data product lifecycle (creation to end of life) and data analytics.
- for simplicity we assumed that a Domain is responsible for 1 business process but in practice this can be more
- A data product has two stages: as a data object and as a business object; this is because we assume that a data product becomes a business object after it has business value (it is accessible and useable by other domains)

Domain Architecture ([link for full size picture](#))

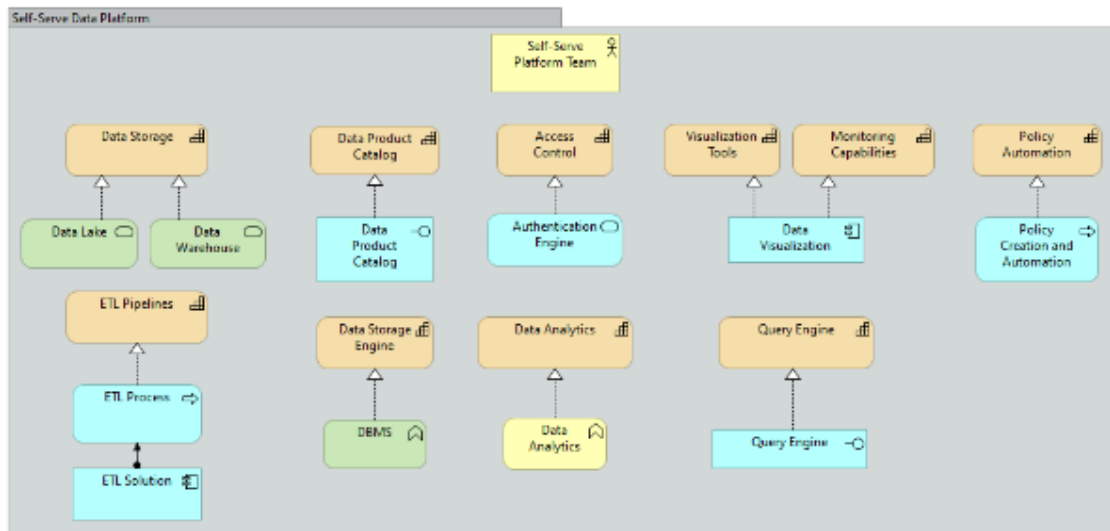


B.2 Continued

Self-Serve Data Platform Architecture

The Self-Serve Data Platform provides capabilities to the data mesh participants.

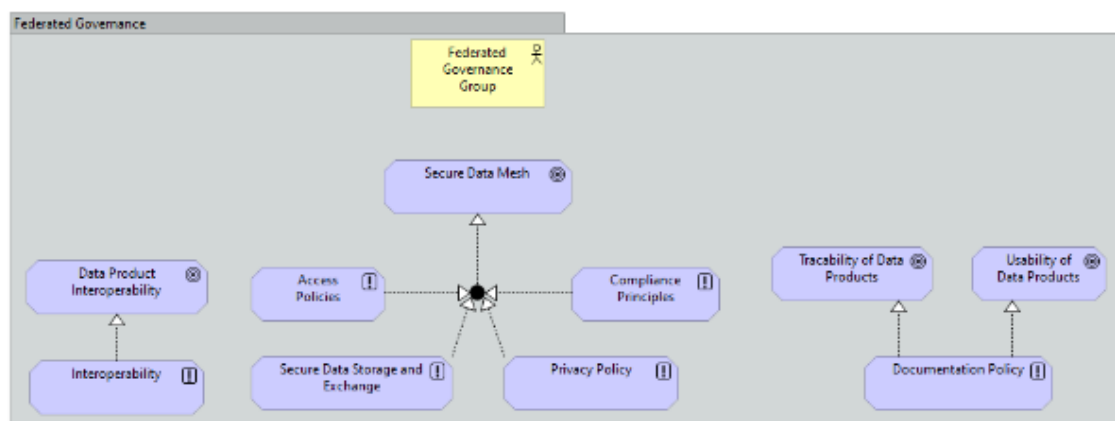
Self-Serve Data Platform Architecture ([link for full size picture](#))



The Federated Governance Architecture

The Federated Governance Architecture contains the main principles which apply to all participating domains in a data mesh.

Federated Governance Architecture ([link for full size picture](#))



B.2 Continued

Likelihood for the model to be relevant in data architecture projects *

	1	2	3	4	5	
Very Unlikely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Likely

Perceived ease of use of the model *

	1	2	3	4	5	
Very Difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Easy

Do you think the model will speed up the process of desining data mesh solution architectures? *

	1	2	3	4	5	
Very Unlikely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Likely

The possibility of encountering components and patterns of the model in solution architectures *

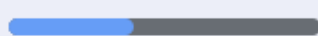
	1	2	3	4	5	
Very Low	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very High

In this section you can voice additional remarks with regards to the **usefulness** of the model

Your answer

Back

Next



Page 2 of 5

Clear form

B.3 Questionnaire Section Quality

Quality

This section is dedicated to answering questions related to the perceived quality of the Data Mesh Reference Architecture

Data Mesh Reference Architecture viewpoints

The 3 pictures below show the 3 architectural components making up a Data Mesh

- Domain
- Self-Serve Data Platform
- Federated Governance Layer

For each of the questions take all 3 of the viewpoints in consideration

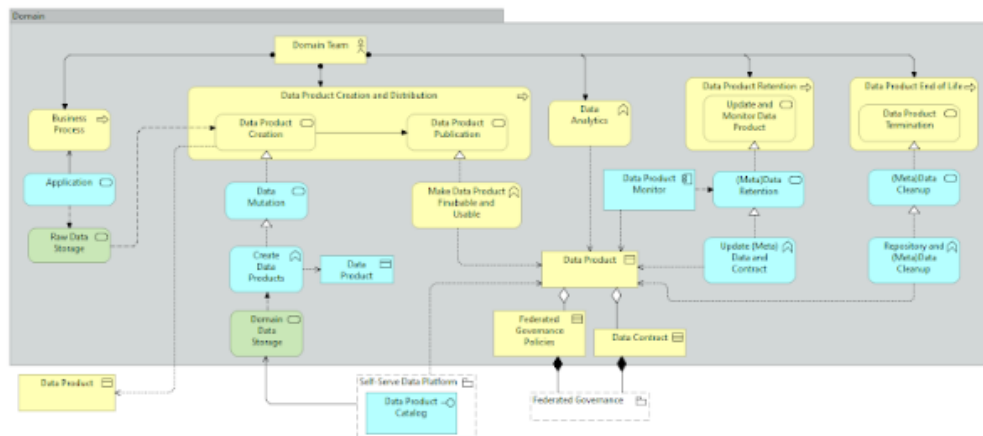
The Data Mesh Reference Architecture will hereafter be referred to as **'the model'**

Domain Architecture

A few important points regarding the Domain Architecture:

- A Domain Team has responsibility over a business process, the data product lifecycle (creation to end of life) and data analytics.
- for simplicity we assumed that a Domain is responsible for 1 business process but in practice this can be more
- A data product has two stages: as a data object and as a business object; this is because we assume that a data product becomes a business object after it has business value (it is accessible and useable by other domains)

Domain Architecture ([link for full size picture](#))

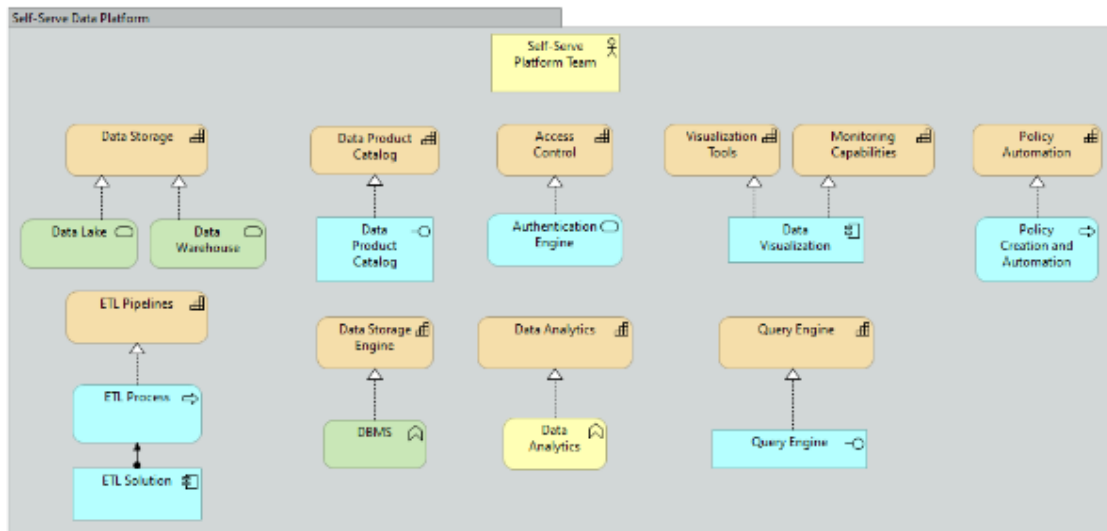


B.3 Continued

Self-Serve Data Platform Architecture

The Self-Serve Data Platform provides capabilities to the data mesh participants.

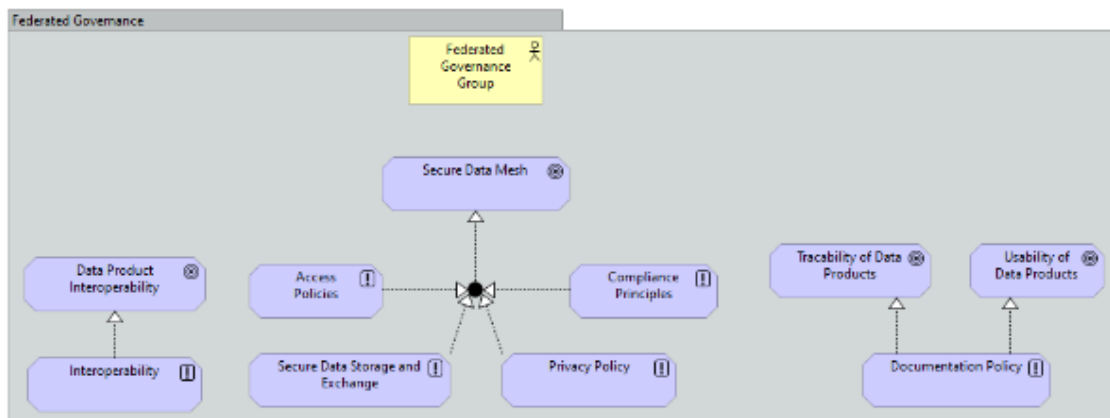
Self-Serve Data Platform Architecture ([link for full size picture](#))



The Federated Governance Architecture

The Federated Governance Architecture contains the main principles which apply to all participating domains in a data mesh.

Federated Governance Architecture ([link for full size picture](#))



B.3 Continued

Do you think the goal and purpose of the model are clear? *

	1	2	3	4	5	
Very Unclear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Clear

Do you think the model is suitable to achieve the goal and purpose? *

	1	2	3	4	5	
Very Unsuitable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Suitable

How complete do you think the model is? *

Would the inclusion of all components and principles be enough to create a working data mesh?

	1	2	3	4	5	
Very Poor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Fully Complete

What is your perception of the level of detail of the model? *

	1	2	3	4	5	
Very Poor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Excellent

B.3 Continued

What is your perception of the concreteness of the elements in the model? *
Is the level of abstraction sufficient to allow freedom in choosing technologies, tools and software while providing sufficient guidance in the specific need for certain technology, tooling and software?

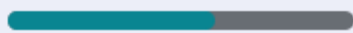
	1	2	3	4	5	
Very Poor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Excellent

In this section you can voice additional remarks with regards to the **quality** of the model

Your answer

Back

Next



Page 3 of 5

Clear form

B.4 Questionnaire Section Variability

Variability

This section is dedicated to answering questions related to the perceived variability of the Data Mesh Reference Architecture

Data Mesh Reference Architecture viewpoints

The 3 pictures below show the 3 architectural components making up a Data Mesh

- Domain
- Self-Serve Data Platform
- Federated Governance Layer

For each of the questions take all 3 of the viewpoints in consideration

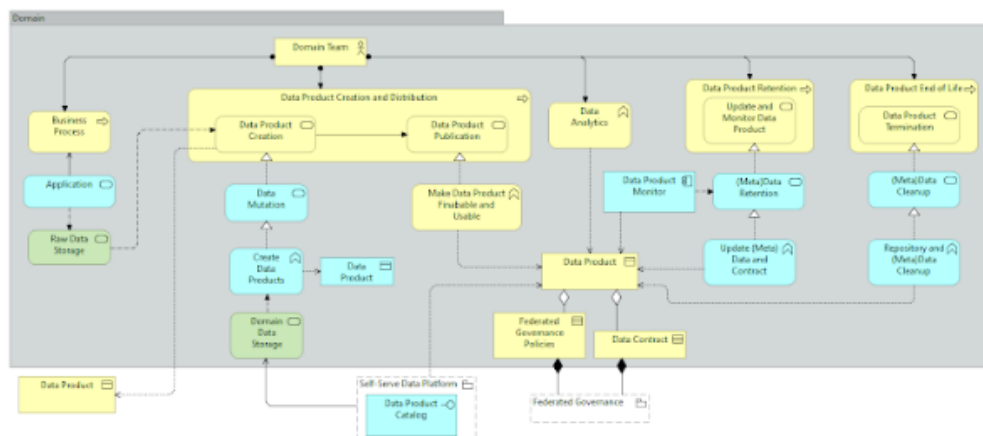
The Data Mesh Reference Architecture will hereafter be referred to as **'the model'**

Domain Architecture

A few important points regarding the Domain Architecture:

- A Domain Team has responsibility over a business process, the data product lifecycle (creation to end of life) and data analytics.
- for simplicity we assumed that a Domain is responsible for 1 business process but in practice this can be more
- A data product has two stages: as a data object and as a business object; this is because we assume that a data product becomes a business object after it has business value (it is accessible and useable by other domains)

Domain Architecture ([link for full size picture](#))

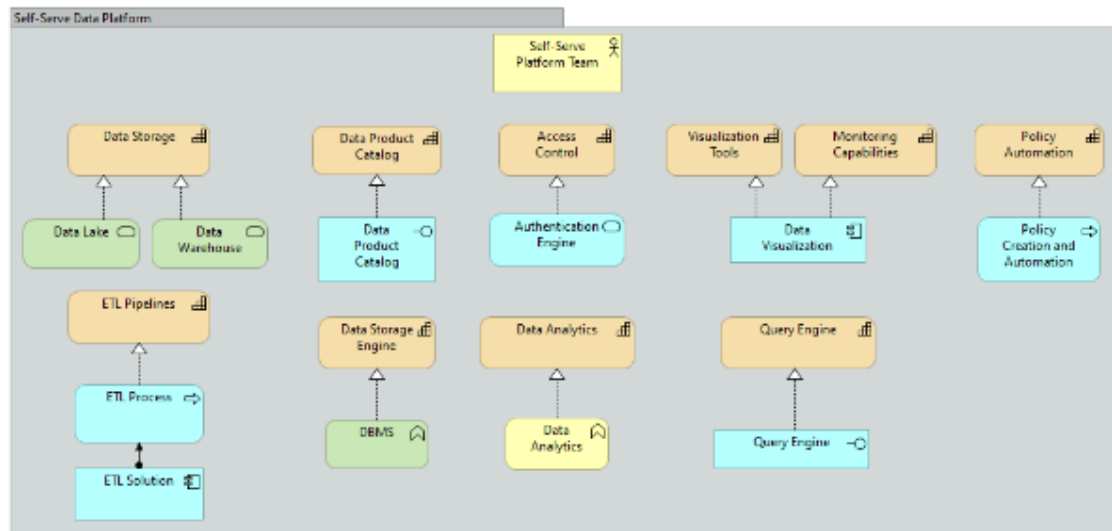


B.4 Continued

Self-Serve Data Platform Architecture

The Self-Serve Data Platform provides capabilities to the data mesh participants.

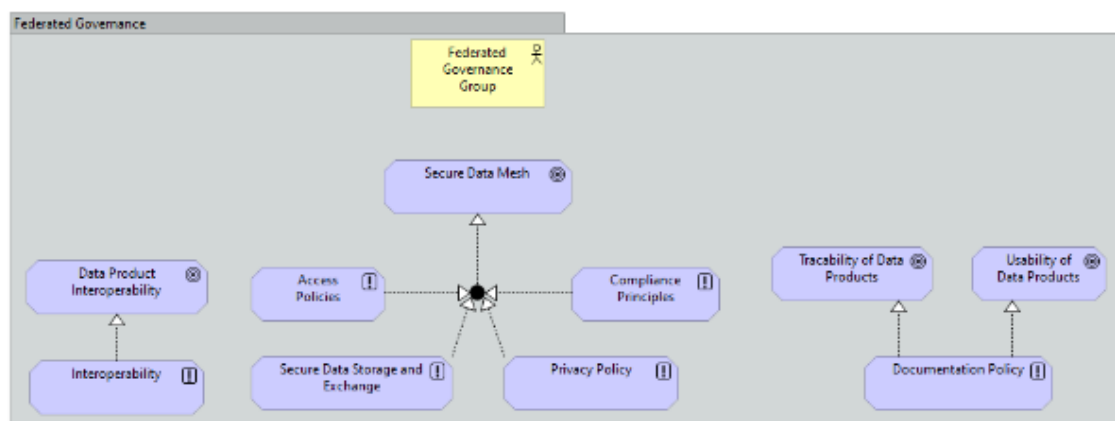
Self-Serve Data Platform Architecture ([link for full size picture](#))



The Federated Governance Architecture

The Federated Governance Architecture contains the main principles which apply to all participating domains in a data mesh.

Federated Governance Architecture ([link for full size picture](#))



Is the model suitable to meet changing requirements?

	1	2	3	4	5	
Very Unlikely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Likely

How easy is it to customize or extend the model? *

	1	2	3	4	5	
Very Difficult	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Easy

How likely is it that the model can be used in multiple use cases? *

	1	2	3	4	5	
Very Unlikely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Likely

How likely is it that the model can be used in multiple industries? *

	1	2	3	4	5	
Very Unlikely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Very Likely

In this section you can voice additional remarks with regards to the **variability** of the model

Your answer

Back

Next

Page 4 of 5

Clear form

B.5 Questionnaire Additional Feedback

Additional remarks and points of improvement

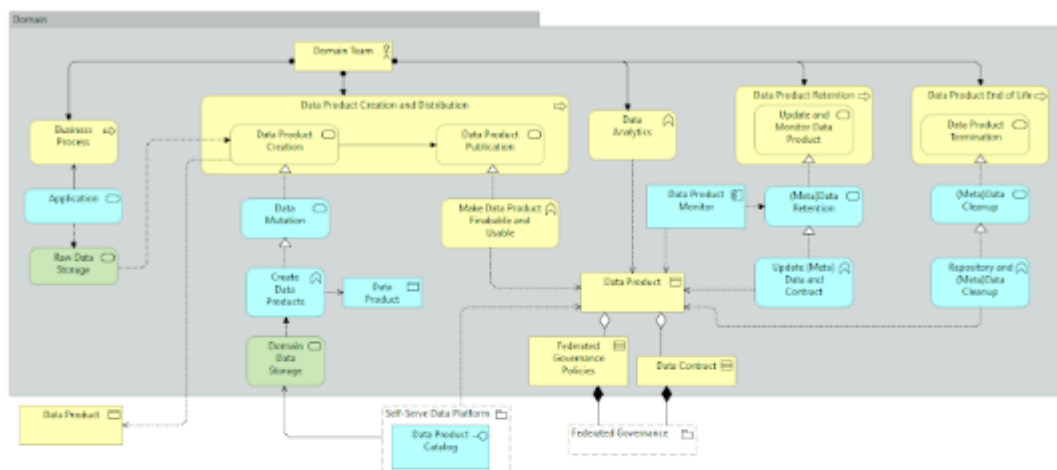
In this section you can leave additional remarks and points of improvement related to each of the viewpoints (not mandatory)

Domain Architecture

A few important points regarding the Domain Architecture:

- A Domain Team has responsibility over a business process, the data product lifecycle (creation to end of life) and data analytics.
- for simplicity we assumed that a Domain is responsible for 1 business process but in practice this can be more
- A data product has two stages: as a data object and as a business object; this is because we assume that a data product becomes a business object after it has business value (it is accessible and useable by other domains)

Domain Architecture ([link to full size picture](#))



If you have any remarks or points of improvement related to the Domain architecture leave them here

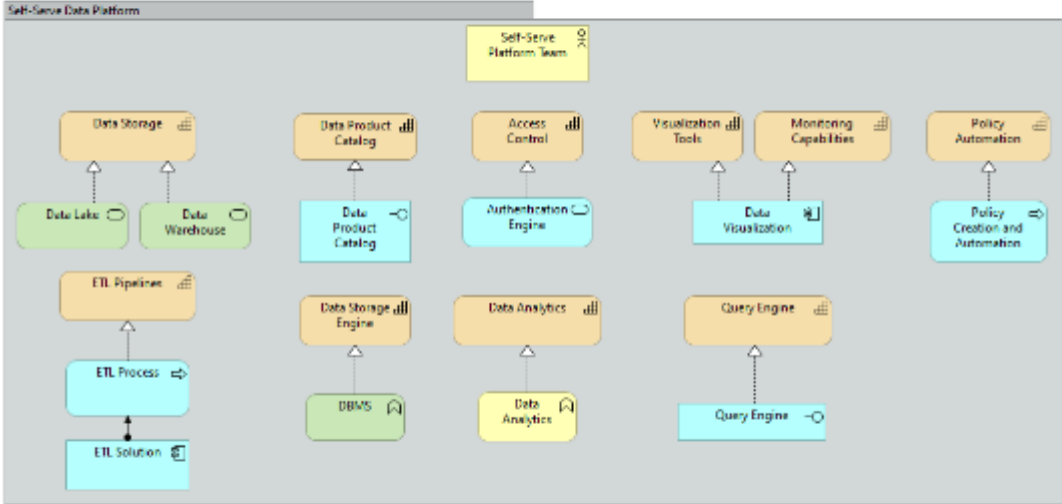
Your answer

B.5 Continued

Self-Serve Data Platform Architecture

The Self-Serve Data Platform provides capabilities to the data mesh participants.

Self-Serve Data Platform Architecture ([link to full size picture](#))



If you have any remarks or points of improvement related to the Self-Serve Data Platform architecture leave them here

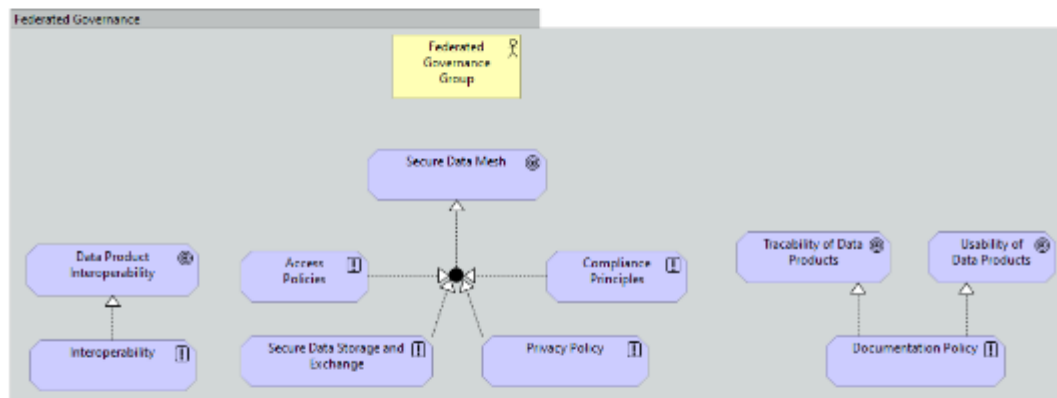
Your answer

B.5 Continued

Federated Governance Architecture

The Federated Governance Architecture contains the main principles which apply to all participating domains in a data mesh.

Federated Governance Architecture ([link to full size picture](#))



If you have any remarks or points of improvement related to the Federated Governance architecture leave them here

Your answer

Back

Submit

Page 5 of 5

Clear form

C Questionnaire Likert Scale Answers Per Respondent

C.1 Likert Scale Answers Usefulness Section

The questions:

- Q1: Likelihood for the model to be relevant in data architecture projects
- Q2: Perceived ease of use of the model
- Q3: Do you think the model will speed up the process of designing data mesh solution architectures?
- Q4: The possibility of encountering components and patterns of the model in solution architectures

<i>Role</i>	Q1	Q2	Q3	Q4
<i>Data Consultant - 1</i>	4	2	2	3
<i>Tech Consultant - 1</i>	4	2	3	2
<i>Tech Consultant - 2</i>	3	2	3	2
<i>Tech Consultant - 3</i>	4	3	4	3
<i>Data Architect - 1</i>	4	3	3	4
<i>Other: Professional Trainer</i>	4	2	3	2
<i>Data Architect - 2</i>	4	4	3	4
<i>Data Consultant - 2</i>	4	3	3	4
<i>Tech Consultant - 4</i>	4	3	2	4
<i>Enterprise Architect - 1</i>	4	4	2	3
<i>Data Consultant - 3</i>	4	4	4	4
<i>Data Engineer - 1</i>	4	3	3	4
<i>Data Architect - 3</i>	1	3	1	1
<i>Manager / Team Lead - 1</i>	4	3	2	5
<i>Data Consultant - 4</i>	4	4	4	4
<i>Manager / Team Lead - 2</i>	3	2	4	3
<i>Tech Consultant - 5</i>	4	3	4	4
<i>Enterprise Architect - 2</i>	4	3	3	4
<i>Tech Consultant - 6</i>	3	2	3	4
<i>Data Consultant - 5</i>	3	3	3	3
<i>Data Architect - 4</i>	2	2	2	5
<i>Data Consultant - 6</i>	4	4	3	5
<i>Tech Consultant - 7</i>	4	3	4	5
<i>Data Engineer - 2</i>	4	3	4	4
<i>Data Engineer - 3</i>	4	4	5	4
<i>Data Engineer - 4</i>	3	4	3	4
<i>Other: Chief Innovation Officer</i>	3	3	2	2
<i>Data Architect - 5</i>	3	3	1	5
<i>Enterprise Architect - 3</i>	4	3	4	3
<i>Enterprise Architect - 4</i>	4	4	4	4
<i>Manager / Team Lead - 3</i>	3	3	3	4
<i>Data Consultant - 7</i>	3	3	3	4

C.2 Likert Scale Answers Quality Section

The questions:

- Q1: Do you think the goal and purpose of the model are clear?
- Q2: Do you think the model is suitable to achieve the goal and purpose?
- Q3: How complete do you think the model is? (*Would the inclusion of all components and principles be enough to create a working data mesh?*)
- Q4: What is your perception of the level of detail of the model?
- Q5: What is your perception of the concreteness of the elements in the model? (*Is the level of abstraction sufficient to allow freedom in choosing technologies, tools and software while providing sufficient guidance in the specific need for certain technology, tooling and software?*)

Role	Q1	Q2	Q3	Q4	Q5
Data Consultant - 1	4	3	5	3	4
Tech Consultant - 1	2	3	3	4	3
Tech Consultant - 2	3	3	4	1	2
Tech Consultant - 3	4	3	2	3	5
Data Architect - 1	3	4	3	4	3
Other: Professional Trainer	4	4	4	3	2
Data Architect - 2	4	4	3	3	4
Data Consultant - 2	4	4	4	4	4
Tech Consultant - 4	2	2	2	3	4
Enterprise Architect - 1	2	3	2	2	4
Data Consultant - 3	3	4	3	4	3
Data Engineer - 1	3	4	3	4	3
Data Architect - 3	4	1	3	2	4
Manager / Team Lead - 1	5	3	4	3	1
Data Consultant - 4	4	4	4	4	3
Manager / Team Lead - 2	3	3	4	4	4
Tech Consultant - 5	3	3	4	4	4
Enterprise Architect - 2	4	4	3	4	4
Tech Consultant - 6	2	2	3	3	3
Data Consultant - 5	3	3	3	3	3
Data Architect - 4	1	3	2	1	2
Data Consultant - 6	5	4	2	3	2
Tech Consultant - 7	2	4	3	5	5
Data Engineer - 2	4	3	3	4	4
Data Engineer - 3	4	3	4	4	4
Data Engineer - 4	3	4	3	4	3
Other: Chief Innovation Officer	3	2	2	2	3
Data Architect - 5	1	1	1	1	1
Enterprise Architect - 3	5	4	3	3	4
Enterprise Architect - 4	4	4	4	4	4
Manager / Team Lead - 3	2	3	3	4	3
Data Consultant - 7	2	4	3	3	4

C.3 Likert Scale Answers Variability Section

The questions:

- Q1: Is the model suitable to meet changing requirements?
- Q2: How easy is it to customize or extend the model?
- Q3: How likely is it that the model can be used in multiple use cases?
- Q4: How likely is it that the model can be used in multiple industries?

<i>Role</i>	Q1	Q2	Q3	Q4
<i>Data Consultant - 1</i>	2	3	4	4
<i>Tech Consultant - 1</i>	3	5	4	5
<i>Tech Consultant - 2</i>	4	5	5	5
<i>Tech Consultant - 3</i>	5	4	4	4
<i>Data Architect - 1</i>	4	4	4	4
<i>Other: Professional Trainer</i>	5	5	5	5
<i>Data Architect - 2</i>	4	4	4	4
<i>Data Consultant - 2</i>	3	2	5	5
<i>Tech Consultant - 4</i>	4	4	4	4
<i>Enterprise Architect - 1</i>	3	4	4	4
<i>Data Consultant - 3</i>	4	4	4	4
<i>Data Engineer - 1</i>	4	4	3	4
<i>Data Architect - 3</i>	3	2	3	3
<i>Manager / Team Lead - 1</i>	3	3	5	5
<i>Data Consultant - 4</i>	4	4	4	4
<i>Manager / Team Lead - 2</i>	3	5	4	4
<i>Tech Consultant - 5</i>	3	4	4	4
<i>Enterprise Architect - 2</i>	4	5	5	5
<i>Tech Consultant - 6</i>	4	3	3	4
<i>Data Consultant - 5</i>	3	3	3	3
<i>Data Architect - 4</i>	4	3	5	5
<i>Data Consultant - 6</i>	3	4	5	5
<i>Tech Consultant - 7</i>	4	4	4	5
<i>Data Engineer - 2</i>	4	4	4	3
<i>Data Engineer - 3</i>	3	4	4	4
<i>Data Engineer - 4</i>	3	2	3	4
<i>Other: Chief Innovation Officer</i>	3	4	3	4
<i>Data Architect - 5</i>	1	1	5	5
<i>Enterprise Architect - 3</i>	2	4	3	3
<i>Enterprise Architect - 4</i>	4	4	4	4
<i>Manager / Team Lead - 3</i>	3	4	3	4
<i>Data Consultant - 7</i>	4	4	4	4