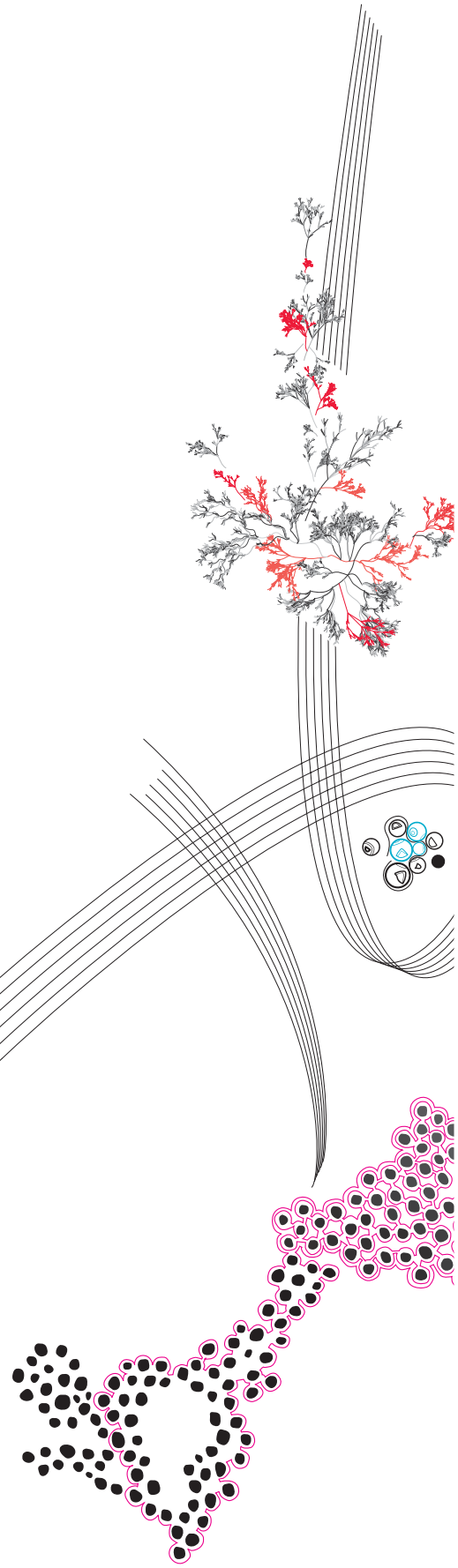


MSc Business Information Technology
Master Thesis

A Methodology for Developing and Maintaining Data Products within a Data Mesh Architecture

Mark Langedijk

Supervisors:
dr. L. Ferreira Pires, University of Twente, Chairman
dr. G. Sedrakyan, University of Twente, Second supervisor
M. van de Goor MSc, Info Support, Company supervisor

November, 2024

Department of Business Information Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Acknowledgments

In front of you lies my master's thesis and with it the end of my days as a student. I jumped into the deep end during my time as a student, especially when I decided to move to a new city across the country for my master's degree. I am grateful and proud of the things I learned during my time as a student, but I did not get here through myself, but through the support of the wonderful people around me.

First and foremost, I would like to thank my parents and my brother immensely for inspiring me and giving me the opportunity to discover myself. Throughout my studies and master's thesis, they have been a beacon of support.

I would like to thank Info Support for their guidance and support during my graduation period. In particular, I would like to thank Martin van de Goor for all the feedback, help, and tips that helped me get the most out of my thesis. I would also like to thank my colleagues for the sparring sessions and lunches.

To my thesis committee, I would like to extend my gratitude for sharing your expertise and offering constructive feedback. Dr. Luís Ferreira Pires thank you for all the help and guidance. Dr. Gayane Sedrakyan thank you for your enthusiasm and patience, which helped me through the thesis.

To all the individuals who graciously participated in my interviews and survey, thank you for giving your time and insights. This has enriched my work

A special thanks to Personal Trainer, Jungle, Royal Blood, Parcels, Franz Ferdinand, IDLES, The Hives, and many more. Your music helped me stay focused during long study sessions and helped me escape whenever I needed it most. Your music has given me energy and inspiration.

Finally, I would like to express my gratitude to my friends. You inspired me and without you I would not be who I am today. Thank you for all the study sessions together, for all the fun nights after those study sessions, and for coloring my days as a student.

Without the support and contributions of all these wonderful people, this thesis would not have been possible.

Abstract

Large organizations make use of data to improve decision-making. However, traditional data architectures can not keep up with the scale and complexity of data. Data mesh addresses these problems by decentralizing data management. In this data architecture, data is part of a data product, which contains data, metadata, code, interfaces, and infrastructure.

In this research, we employ Design Science Research to design a methodology for developing and maintaining data products within a data mesh architecture. In this study, we conducted a gray literature review, which enabled us to gain insight into existing data and software methodologies, as well as the development of data products in practice. This knowledge was used to design our methodology. Furthermore, we utilized the FEDS evaluation framework to conduct a formative and summative evaluation of the methodology.

The primary contribution of this study is the Development Methodology for Data Products within a Data Mesh Architecture (DPDM-DMA). This methodology was perceived as useful and easy to use by experts in the field. To our knowledge, the DPDM-DMA is the first structured methodology that focuses on developing data products within a data mesh architecture. This work lays the foundation for creating high-quality, sharable data products. Further research can give insights into how this can be applied in various sectors.

Contents

Chapter 1

Introduction

This chapter introduces the design problem that motivates our study. The chapter outlines our research objectives and formulates the research questions that guide our study.

1.1 Motivation

A data architecture is a blueprint for data management within an organization. Traditionally, most organizations have relied on centralized data architectures, where a single team is responsible for managing data within an organization. However, as the quantity and complexity of data continue to expand, this approach has proven inadequate for large data-driven organizations, resulting in bottlenecks within the central data teams.

A novel data architecture called data mesh addresses these challenges by decentralizing data management. First proposed in a blog article by Dehghani [13], data mesh can play an important role in addressing the issue of scale and complexity in data management at large organizations. Data mesh is a socio-technical solution that fosters scalability and interoperability. The topic of data mesh has been of interest to researchers and organizations alike, as these characteristics are seen as important drivers for keeping up with the increased demand for data-driven decision-making. Importantly, the issue of data sharing is no longer a technical challenge but an organizational one. Data mesh addresses this by introducing a completely new approach to handling data in organizations.

The data mesh architecture utilizes the concept of "data product thinking", a growing trend in data management. This philosophy applies product thinking to data, making data producers responsible for delivering value to their users. This reasoning can be compared to traditional products, where organizations are responsible for making an appealing product. Data product thinking encourages data producers to consider how they format and deliver data to maximize value for data consumers. Data products are prepared datasets and services that are easily usable by parties who don't directly interact with teams responsible for the data. In the context of data mesh architecture, data products are supported by an organizational structure and technology.

Data mesh and data products are novel topics, and these concepts are still relatively new in the field of data management [18]. Consequently, there is still limited guidance for organizations on how to implement a data mesh architecture or how to develop high-quality data products. This study set out to find a methodology for developing and maintaining data products within a data mesh data architecture, providing practical insights for

organizations seeking to adopt data mesh architectures.

1.2 Problem statement

In today's data-driven landscape, a comprehensive understanding of data products and data mesh architectures enables us to further understand and facilitate effective data sharing among various stakeholders, both within and outside an organization [7]. Data can be incredibly valuable for organizations, as described by Svensson & Taghavianfar [4] who identified potential benefits and challenges of data-driven organizations. Foremost, data can enhance decision-making by possibly making decisions more informed, accurate, specific, faster, and reliable. Data can help organizations understand the customer. Finally, it can improve creativity, productivity, and the market position of organizations. However, to realize these benefits, an organization has to have a data-driven culture and access to a significant volume of data. Maturity assessment models can help to assess organizations' readiness for migrating to a data mesh architecture [8].

Recent advancements in data analytics, artificial intelligence and big data offer interesting opportunities for companies. These technologies lead to enhanced decision-making, reduced costs, and increased efficiency [43]. Organizations are seeking to gather as much data as possible in order to capitalize on these emerging opportunities. Concurrently, there is a growing trend of organizations sharing data between different departments and with competitors. Technologies such as Multi-Party Computation, Federated Learning, and Data Spaces are facilitating this collaborative data exchange.

To further understand how we can exchange data between data providers and data consumers, research into data mesh and data products is really valuable [16, 48].

Data mesh offers a promising solution to the limitations of centralized data architectures. However, data mesh is a novel paradigm, and as a consequence, principles and guidelines for working with a data mesh architecture are still lacking [7, 18]. The effective utilization of analytical data for critical decision-making is dependent upon the availability of data that meets certain quality standards. A data architecture follows principles and standards for managing data, which facilitate quality data management. However, these principles and standards must be followed correctly to ensure quality data.

Within a data mesh, data is part of a data product. These data products are nodes within the data mesh, which includes data, metadata, code, and infrastructure. Data products represent the central elements of the data mesh architecture. However, to the best of our knowledge there is currently no guidance or research that focuses on developing and maintaining data products. To ensure data quality, organizations must establish guidelines to effectively build and manage data products. These guidelines would result in enhanced efficiency, quality, and consistency in data product management.

1.3 Research goals

Data products and data mesh are becoming an important factor in data-driven decision making. As large organizations struggle with the increasing volume and complexity, these concepts can offer promising solutions for data management. This research aims to address a gap in the current knowledge by designing a methodology to help data producers in

building and maintaining high-quality data products within a data mesh architecture.

To define our research goal precisely, we defined our design problem using Wieringa's design-problem-template [49]:

Improve *the development and maintenance of data products within a data mesh architecture* ***by*** *designing a methodology for data producers* ***that*** *enhances the efficiency, consistency and quality of data products* ***in order to*** *improve data-driven decision making in large data-driven organizations.*

The primary objective of this research has been to offer insights into the development and maintenance process for data products. This understanding can help both practitioners and researchers better understand data products. Additionally, our research aimed to enhance the efficiency and quality of data products in organizations. To achieve this, we developed a methodology to establish guidelines, expand knowledge, and ultimately improve the quality of data products.

Ultimately, this research gives new perceptions on managing shareable data. While our focus was on data products within a data mesh architecture, our findings are expected to be applicable beyond the data mesh context. Therefore, our research can contribute to the advancement of data management practices in general.

To address the design problem, we have formulated the following research goals:

- To determine the activities necessary for developing and maintaining data products within a data mesh.
- To design a methodology that describes activities for building and maintaining data products in detail.
- To validate the ease of use and usefulness of the proposed methodology.

1.4 Research questions

We translated the design problem and research goals into our main research question:

- ***RQ:*** *How can we design a methodology to assist organizations in developing and maintaining high-quality data products within a data mesh architecture?*

We defined sub-questions to help us answer our main research question. Firstly, it is essential to have a proper understanding of data products and how they are built and maintained before we started to design our methodology. The objective of the methodology is to assist organizations in building and maintaining high-quality data products. However, to achieve this, it was necessary to have a clear understanding of the characteristics that characterize a high-quality data product. This information facilitates the formulation of objectives and requirements for the methodology.

- **SQ1:** What are the essential steps in the design and maintenance of data products in a data mesh architecture?
- **SQ2:** What are the characteristics of high-quality data products?

Once the characteristics and development of a high-quality data product have been fully understood, the next step is to design a methodology that supports organizations in applying this knowledge. This entailed identifying the appropriate method for designing a methodology and drawing inspiration from existing methods and guidelines for working with data.

- **SQ3:** How can we design an effective methodology for developing and maintaining data products based on insights from existing literature?

In order to demonstrate the applicability and efficiency of the methodology, we demonstrated the methodology in a fictitious case. Subsequently, we utilized expert interviews to evaluate the methodology and ascertain its compliance with the established requirements.

- **SQ4:** To what extent can the designed methodology be applied to a use case?
- **SQ5:** To what extent is the designed methodology applicable and useful?

1.5 Thesis structure

This report is further structured as follows:

- Chapter 2 offers an in-depth explanation of our research strategy and methods. This chapter describes how we addressed our research questions.
- Chapter 3 provides an overview of the data mesh data architecture. Additionally, the chapter examines the concepts of data products and reviews relevant data and software methodologies.
- Chapter 4 explicates the design problem, ensuring the problem is precisely defined. Furthermore, the chapter defines the requirements for the methodology used for the development and evaluation of our design.
- Chapter 5 describes the designed methodology on a conceptual and detailed level. This chapter justifies the choices made during the development of the methodology.
- Chapter 6 demonstrates the designed methodology by describing the use of the methodology on a fictitious case.
- Chapter 7 presents the findings of the evaluations. The evaluation consists of two components: a formative evaluation and a summative evaluation.
- Chapter 8 gives our final remarks, summarizing our key findings and contributions. Further, this chapter outlines limitations to the study and identifies potential future research.

Chapter 2

Research methodology

This chapter presents the research methodology, providing a comprehensive overview of the methods used for our study. First, we explore Design Science Research (DSR), which is used to address our research questions. Subsequently, we examine the research strategies and methods used to carry out our Design Science research project and build the DPDM-DMA (Data Product Development Methodology within a Data Mesh Architecture). We discuss the research methods used, which include a review of gray literature and evaluation methods such as interviews with industry experts and surveys.

2.1 Design Science Research

In this research, we employed Design Science Research (DSR), as described by Johannesson and Perjons [23], to address our research questions. DSR is a framework that facilitates the development of artifacts, with Johannesson and Perjons focusing on its applications in information systems and IT. These artifacts are objects that address a practical problem. Our research aims to improve the design and maintenance of data products within a data mesh architecture, which will be done by designing a methodology (the artifact).

Implementing a data mesh architecture presents significant challenges, one of which is the lack of guidelines for implementing a data mesh and developing data products. Practitioners are still exploring ways to implement this data architecture. In this research project, we employ DSR as a structured approach to design a methodology for developing and maintaining data products. The proposed methodology is called the Data Product Development Methodology within a Data Mesh Architecture (DPDM-DMA). DDSR is particularly applicable to our research because we are trying to solve a real-world problem by developing a methodology. The structured approach of DSR ensures that we thoroughly investigate the problem and context and evaluate the designed methodology.

Johannesson and Perjons developed a method framework for Design Science Research [23] that consists of five activities:

1. Explicate the problem.
2. Define requirements.
3. Design and develop the artifact.
4. Demonstrate the artifact.

5. Evaluate the artifact.

Next to these activities, Johannesson and Perjons provide guidelines for carrying out the activities, selecting research methods, and relating the research to an existing knowledge base.

Figure 2.1 illustrates the activities in the Design Science method framework [23]. In their study, Johannesson & Perjons [23] employed the function modeling methodology, IDEF0, to describe the research methodology visually. This model utilizes a visual notation to represent information, with boxes denoting activities and arrows representing data and objects. Figure 2.1 shows the types of data and objects used in the model. The model consists of:

- Input: Knowledge at the beginning of the activity.
- Output: Knowledge at the end of the activity.
- Controls: Information used for governing the activity.
- Resources: Knowledge used as the basis for an activity.

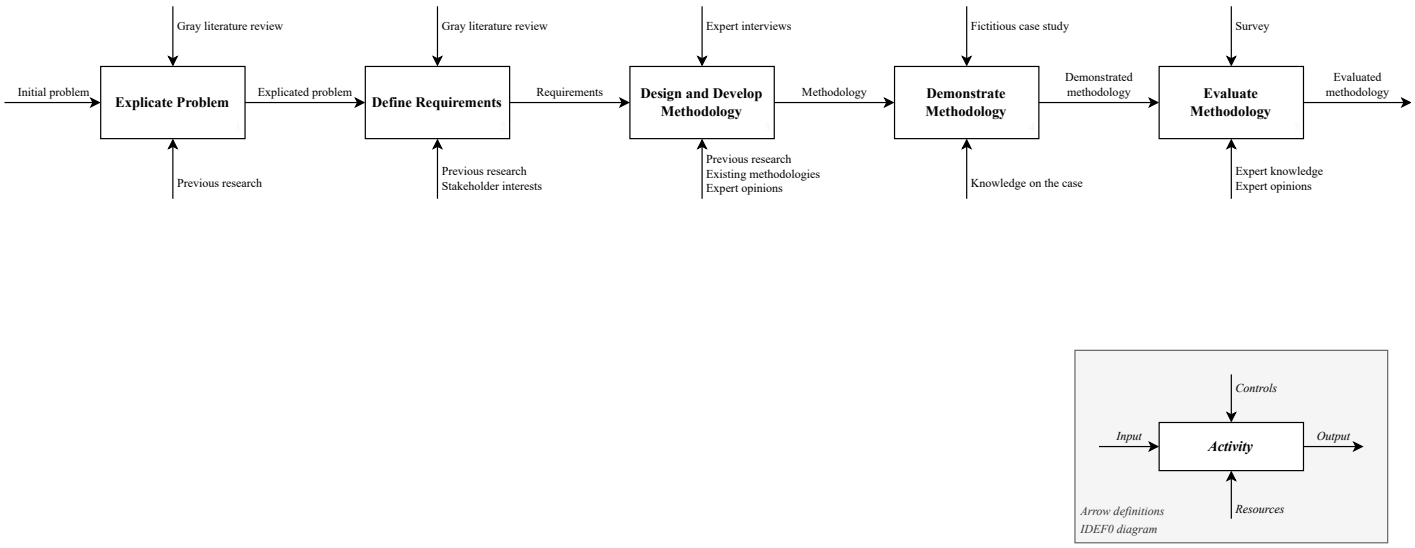


FIGURE 2.1: Method framework for Design Science applied to this research [23]

The framework proposed by Johannesson & Perjons [23] provides clear guidelines for each Design Science research process activity using the IDEF0 modeling methodology to enhance its clarity and usability. Additionally, Johannesson & Perjons [23] recognize the challenges of socio-technical systems, which aligns with the concept of data mesh as a socio-technical system. In comparison to other DSR methodologies, the framework offers a more detailed and structured set of guidelines, making it particularly suitable for our research on developing a methodology.

The sections below explain the activities performed in each phase.

2.1.1 Explicate problem

In the first activity, *explicate problem*, we investigated the practical problem [23]. This activity answers the first subquestion by examining the essential steps in designing and maintaining data products in a data mesh architecture. This activity builds on the information gathered during our previous scoping review [27], presented in Chapter 3, which provided an initial overview of data mesh concepts and associated challenges.

In order to gain a deeper understanding of challenges faced by organizations and data producers, we performed a Gray Literature Review (GLR). This approach enabled us to capture industry practices and trends not yet reflected in literature, which is particularly suitable for a novel topic like data mesh.

2.1.2 Define requirements

Before designing our methodology, we *defined requirements*. Requirements serve as guidelines during the design and as criteria for the evaluations [49].

During this activity, we used insights from our GLR and the results of the explicate problem activity. The GLR findings were synthesized to identify common themes and issues. Based on these themes, the requirements were formulated to guide the design phase. Furthermore, the requirements were used to define an initial outline of the DPDM-DMA.

2.1.3 Design and develop artifact

During the third activity, we *designed and developed* a methodology for developing data products by reusing and adapting parts of existing data management methods and by adopting new ideas.

We started by sketching and creating a first concept, which was discussed with experts to obtain valuable feedback and refine the design.

Johannnesson & Perjons [23] describe the following components of an artifact:

- Structure, the components of the artifact.
- Behavior, actions that the artifact can carry out.
- Function, what the artifact does for the user.
- Environment, the surroundings and conditions in which the artifact will operate.

- Effects, how the artifact will change its environments; this can both be intended and unintended effects.

The methodology is described in accordance with these components in Chapter 5.

2.1.4 Demonstrate artifact

In the fourth activity, the applicability and usefulness of the designed methodology are *demonstrated* through its application in a fictitious case study. This case study illustrates how the methodology addresses the problem at hand. The demonstration is documented to illustrate the methodology's practical application. This documentation serves to explain the methodology's goal and usage to users.

Ideally, an artifact should be demonstrated in a real-life case to show that the artifact can solve an instance of the problem [23]. However, due to the novelty of the data mesh architecture, we faced challenges in finding organizations where we could perform a use-case study within the limited time frame.

We acknowledge this as a limitation of our research. The methodology should be evaluated in a real-life scenario to determine risks and evaluate efficacy. We address this limitation and propose future work in Chapter 8.

2.1.5 Evaluate artifact

During the *evaluate artifact* activity, we assessed how well our designed methodology solves the problem and fulfills the set requirements. We employed the Framework for Evaluation in Design Science [45] to guide our evaluation process. Our evaluation included both formative and summative evaluations.

During the formative evaluations, we interviewed industry experts to gain insights from practice. The experts were selected from different industries and backgrounds to provide a wide range of feedback. These evaluations were used to improve our design iteratively.

For the summative evaluation, we utilized the Method Evaluation Model (MEM) [33]. The MEM is designed to validate information systems design methods, making it particularly suitable for our research. The MEM is based on the popular Technology Acceptance Model (TAM) and Methodological Pragmatism [33].

2.2 Literature Review

This section discusses the literature review, which addresses the first, second, and third sub-research questions of our study. We utilized a Gray Literature Review (GLR) for the *explicate problem* and *define requirements* activities.

2.2.1 Gray literature review

During the *explicate problem* and *define requirements* activities, we utilized a Gray Literature Review (GLR) to further investigate the problem and set requirements for the proposed data product development methodology. A GLR is valuable as it bridges the gap between academic and professional practice. A GLR can help to get industry insights

on the topic, especially in our case with a novel paradigm such as data mesh. By using a GLR, we can incorporate not yet peer-reviewed publications in our research, enabling us to capture all current knowledge on data product development within a data mesh architecture.

We used a structured approach for our GLR, following the guidelines presented by Garousi et al. [17]. This approach, based on Kitchenham & Charters [26], enables us to close the gap between academic and professional practice. The research protocol used for our literature review can be found in Appendix A.

Search strategy

To establish a focus scope, we first defined research questions using PICOC (Population, Intervention, Comparison, Outcome, Context) criteria as described by Kitchenham & Charters [26]. This structured approach allowed us to define the scope of our search.

Our search contained a wide range of gray literature resources, such as expert panels, (video) blogs, and industry reports. This broad search allowed us to capture the most recent developments.

Selection criteria

The selection of resources was based on criteria and a quality assessment. A quality assessment ensured that our resources are valid and free of bias [26].

We utilized a structured document to maintain consistency in data extraction across all sources. This approach ensured a consistent output of information for each resource.

Data extraction and synthesis

The data extraction involved reviewing the key concepts, methodologies, steps, and best practices. The GLR findings were synthesized to identify common themes and issues. These themes were utilized to formulate requirements and create an outline for our design. These requirements are later used to guide our design and evaluation.

The GLR established a solid theoretical foundation for our research.

2.3 Evaluation methods

This section discusses the research methods utilized to evaluate the Data Product Development Methodology within Data Mesh Architectures (DPDM-DMA). We first discuss the framework used for the evaluation. Then, we will dive deeper into the specific methods.

We utilized expert interviews during the *design and development* of the DPDM-DMA to improve our design. Finally, a survey was used during our final *evaluation* to assess the perceived usefulness and perceived ease of use of the DPDM-DMA.

2.3.1 Evaluation framework

Evaluation is a crucial part of design science. An evaluation can determine how well an artifact performs, identify possible side effects, and determine the weaknesses and potential

improvements for an artifact [44]. We applied the Framework for Evaluation in Design Science (FEDS) [45] to guide our evaluation process. This framework guided us through our evaluation using four steps [45]:

1. Explicate the goals of the evaluation
2. Choose the evaluation strategy or strategies
3. Determine the properties to evaluate
4. Design the individual evaluation episodes

By applying a FEDS strategy to our research, we use a structured approach and are guided through the evaluation processes.

Evaluation goal

During the evaluation of the DPDM-DMA, we try to answer the following question: "How well does the methodology solve the explicated problem and fulfill the defined requirements?" [23]. Utilizing the explicated problem and defined requirements from Chapter 4.

Evaluation strategy

Evaluations can be categorized as ex-ante evaluation or ex-post evaluation. Ex-ante evaluations are assessments of an artifact before it is designed and employed. This can be done to understand the need for an artifact better and determine possible components of an artifact. Ex-post evaluations are assessments after an artifact is designed and assess the value of an implemented artifact [23, 45].

Another way we can distinguish evaluations is by discerning the evaluation as an artificial- or naturalistic evaluation, which separates evaluation based on its setting. A naturalistic evaluation takes place in a natural environment, while an artificial evaluation includes laboratory experiments or simulations [44]. Due to limited time and no access to a real-world example, we have chosen only to perform artificial evaluations. This is, however, a limitation to our study, as due to the organizational challenges of a data mesh architecture, a naturalistic evaluation would be needed to evaluate social risks.

We perform both a formative evaluation during the design and a summative evaluation after the finished design.

Evaluation properties

The requirements defined in Chapter 4 are utilized when applying the evaluation method. The designed methodology should align with the data mesh principles. Furthermore, it should be usable, comprehensible, accessible, and relevant.

Additionally, we utilized the Method Evaluation Model (MEM) [33] for the summative evaluation. The MEM is designed to validate information systems design methods, making it particularly suitable for our research. The model is based on the popular Technology Acceptance Model (TAM) and Methodological Pragmatism [33]. The method is used to assess the perceived usefulness and ease of use. We have mapped our explicated problem and defined requirements to the constructs of MEM.

Evaluation episodes

As this is a master's thesis research, we have limited resources regarding time, money, and people. Furthermore, the topic of data mesh is relatively untried, limiting the number of organizations and people with experience in implementing and working with data mesh. These factors helped us narrow down the possible evaluation methods.

We performed expert interviews during the methodology design to get feedback on the process. Next, we utilized a survey to get feedback on the final design of the methodology.

2.3.2 Expert-interviews

We conducted semi-structured interviews with industry experts to gain in-depth insights during our *design and development*. Interviews enabled direct feedback from industry experts, allowing us to refine our methodology iteratively. This type of evaluation is called a formative evaluation, where we utilize feedback for the improvement of our methodology.

Interviews are a valuable tool for discussing complex information in a conversation. By conducting interviews, we were able to explain our methodology and ensure a shared understanding of key concepts. This is especially crucial when exploring novel topics such as data mesh architecture.

We identified experts with experience in data architectures, data mesh, and data products. We tried to obtain a broad selection of experts to gain as much different feedback as possible. To get a wide selection, we utilized our network, and a community focused on data mesh, LinkedIn, Google, and authors of relevant literature.

An interview can be classified by level of structure [38]. There are three levels of standardization:

1. Structured interviews, which follow a predefined protocol with identical questions.
2. Semi-structured interviews, which utilize a list of questions but allow flexibility in order and permit to go off-topic.
3. Unstructured interviews, which are only guided by topic.

For our research, we used semi-structured interviews. The predefined questions enabled us to compare results from different reviews. Additionally, this choice allowed us to dive into the expert's specific knowledge and points of feedback.

While semi-structured interviews offer numerous advantages, it also has some limitations. Potential biases can arise from the interviewer or interviewee. Furthermore, taking interviews is a time-consuming practice. Additionally, the semi-structured nature of the interviews can lead to inconsistencies between interviews, which can affect the results. Finally, the interview results depend on the interviewer's interview and communication skills.

The interviews aimed to better understand how data products are created in large organizations, understand the steps involved and verify whether the steps are similar to our methodology. Furthermore, we gathered feedback on the structure and quality of our methodology.

Each interview lasted approximately one hour and took place online and offline. The interviews were automatically transcribed to ensure efficiency. We analyzed the feedback collected to identify themes and improvements, which were used to improve and refine the methodology.

2.3.3 Survey

This study utilizes a survey to evaluate the usefulness, ease of use, and defined requirements of the final version of the DPDM-DMA. A major advantage of surveys is that they can be easily sent to a large audience. Further, a survey has the same set of questions for each respondent, allowing us to easily compare the answers of each respondent. The responses can be statistically analyzed.

For the survey, we targeted experts with experience in data mesh architectures by approaching the interviewees of our formative evaluation and sending the survey to a community channel with data mesh enthusiasts and experts. The survey was conducted using a web questionnaire [32].

The questionnaire is divided into five sections:

1. Information on the methodology.
2. Demographic information and data mesh experience (5 questions).
3. Perceived usefulness (6 questions, using a Likert scale [33]).
4. Perceived ease of use (5 questions, using a Likert scale [33]).
5. Intention to use (2 questions, using a Likert scale [33]).
6. Open-ended feedback (5 questions).

The survey questions were developed based on the requirements set for the methodology and on the Method Evaluation Model (MEM) [33]. By using this proven model, we ensure the validity of the evaluation and survey. The questionnaire starts with questions on the experience and demographics of the respondents. Further, the questionnaire contained mandatory closed-ended and optional open-ended questions on the DPDM-DMA, allowing us to collect quantitative and qualitative data.

Chapter 3

Background

The data mesh architecture can be described by explaining the four underlying principles of this architecture. This chapter explains these principles and both the social and technical properties of data mesh. Furthermore, the concept of data product is examined. This chapter ends with a review of relevant methodologies that were utilized as inspiration for the developed methodology.

3.1 Data mesh architecture

In this section, we summarize the most important findings of recent studies to better understand the current state of research on data mesh. This synthesis is done through a gray literature review. This work builds on an earlier review by Goedegebuure et al. [18] by considering the most recent research and focusing on approaches for designing and implementing data meshes.

Large organizations commonly use data to gain insights and achieve a competitive advantage. Data architectures are used as guidelines for working with data. A paradigm shift in data architectures is being driven by the challenges large organizations face in dealing with poor data governance and bottlenecks in current data architectures due to the increasing volume and complexity of data. Currently, data are often managed in a central place within an organization. However, some organizations have shifted to data mesh as a socio-technical solution that works with analytical data in a decentralized way. Data mesh is not solely a technical solution; instead, it also focuses on how an organization should implement the roles within the organization and is thus considered a socio-technical solution.

Data mesh follows from four core principles for successfully implementing a decentralized architecture:

1. Domain-Oriented Data Ownership: introduces decentralized data management by splitting an organization into business domains and making these domains responsible for data management.
2. Data as a Product: introduces data products by applying "product thinking" to data, making the business domains responsible for providing high-quality data to other business domains.
3. Self-Serve Data Platform: a domain-agnostic platform built by a centralized platform team that provides autonomous business domains with the tools they need for the entire data product lifecycle.

4. Federated Computational Governance: manages decentralized data, ensures compliance with rules and maximizes data quality through federated decision-making.

In the following sections, we give a detailed explanation of these four principles of data mesh. The main insights for this chapter are derived from Dehghani's book [12], who first coined the term data mesh. Additionally, a significant amount of inspiration came from Goedegebuure et al. [18], who structurally analyzed and synthesized information from gray literature in a literature review.

3.1.1 Principle 1 — Domain-oriented data ownership

The first principle of data mesh is domain-oriented data ownership, which introduces decentralized data management to analytical data by using Domain-Driven Design (DDD) to structure an organization into business domains. Data architectures before data mesh use a monolithic or silo approach to data, where the data are stored and managed by a central data team. In a data mesh, the responsibility of gathering, maintaining, and offering data and (a part of the) data governance is decentralized to business domains, bringer tasks closer to people with expertise in the data [18].

By decentralizing, data can be managed by people who are closest to them [12]. These users have domain knowledge of their specific business domain and can use their expertise in the business domain when working on the data. As a result, the quality of the data should increase, and the domains will have a quicker way to access and modify the data. Each business domain now has many more responsibilities, such as providing data to other domains, which presents numerous challenges that will be addressed in a later section.

3.1.2 Principle 2 — Data as a product

The second principle applies product thinking to data. As business domains become responsible for offering data to other business domains, it is essential to adopt a new mindset regarding providing data. Data are offered to consumers by creating data products. Business domains should provide high-quality data to other domains, and data consumers must be happy and pleased [12]. This is done with the help of new roles that are created within the business domains: the domain data product owner and the data product developer. With data products, as with normal products, the producer is responsible for warranty, quality, information, and sales.

We must ensure the usefulness of data products, as the business domain is responsible for making a useful data product. Dehghani identified eight attributes of high-quality data products, shown in Figure 3.1. Business domains can use these attributes to test if their data products are valuable for their consumers, inside or outside the domain. We further dive into data product quality and the DAUTNIVS attributes in Chapter 4.

3.1.3 Principle 3 — Self-serve data platform

The third principle of data mesh is a self-serve data platform and is about building a domain-agnostic infrastructure [18]. All business domains must make use of infrastructure to store, process, and visualize data in order to create data products [13]. A central data team is responsible for creating this domain-agnostic infrastructure, relieving business domains from the task of building their own infrastructure.

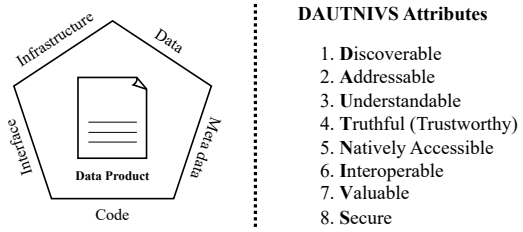


FIGURE 3.1: Components and attributes of high-quality data products

Goedegebuure et al. [18] and Machado et al. [29] described the components of the data platform, which are shown in Table 3.1. Besides relieving business domains from the work of creating and maintaining infrastructure, the platform should empower its users. In particular, the platform should be created so that users can work with data without knowing the jargon, thereby making it more accessible for people with a limited technical background. This enables software developers and business teams to work on data products and prevents organizations from hiring scarce and costly data engineers [13]. Finally, the platform helps with the cooperation and interoperability between different business domains.

A data mesh platform should facilitate working with data and the product-thinking philosophy. It should allow business domains to exchange data products and allow consumers to effortlessly discover data products. The platform should automate and assist in governance tasks and help its users comply with governance policies.

TABLE 3.1: Components of a self-serve data platform

Goedegebuure et al.	**Machado et al.**
Polyglot storage	Storage
Distributed query engine	Data visualization
Service for data product componentes	Integration
Security and privacy	Machine learning
Metadata repository	Processing
Data catalog	Software development
BI tools	
Monitoring	
Compute	
Networking	
Product lifecycle management	
Policy enforcement	

Dehghani divides the data platform into three different planes based on capabilities [12]:

- Data infrastructure plane, for managing the underlying infrastructure.
- Data product experience plane, for developing and consuming data products. For example, setting local policies, creating and deploying a data product.
- Data mesh plane, for services that are on a mesh level, such as discovering data

products, or setting global policies.

3.1.4 Principle 4 — Federated computational governance

"Data governance specifies a cross-functional framework for managing data as a strategic enterprise asset. In doing so, data governance specifies decision rights and accountabilities for an organization's decision-making about its data. Furthermore, data governance formalizes data policies, standards, and procedures and monitors compliance." – Abraham et al. [1].

The fourth principle of data mesh is federated computational governance. Effective data governance results in improved performance effects while mitigating risks [1]. However, governing a distributed data architecture is a complex task, and Bode et al. [7] identified this as one of the main challenges for professionals in the adoption of a data mesh.

TABLE 3.2: Global and local governance activities, defined by Goedegebuure et al. [18]

Global governance	**Local governance**
Define organization-wide standards and guidelines	Managing the data models of the product
Define and enforce global governance policies	Managing data access control
Define data quality assessment methodology	Managing compliance and conformance
Business glossary modeling	Managing data quality
Monitoring data mesh	Monitoring data product health
Creating incentive models	

Organizations should look at a data mesh as a connected ecosystem using systems thinking. A balance between local and global decision-making is required to successfully govern a data mesh [12]. Goedegebuure et al. [18] defined three types of governance in a data mesh: global, local, and automated governance. Responsibilities for data design are decentralized in a data mesh; however, sometimes data from multiple business domains need to be interoperable, and global governance can facilitate this. Furthermore, global governance is needed to set global quality standards and organization-wide policies. Business domain teams get the control to set local policies. These local policies ensure data quality on a local level utilizing domain expert knowledge. A data platform can be used to automate governance policies. Regulation, security, and errors can be detected (and resolved) using automated governance. Business domain teams are responsible for local domain models and quality. Global and local governance tasks are summarized in Table 3.2.

3.1.5 Benefits of data mesh

Moving to a data mesh architecture has multiple potential benefits. Firstly, a data mesh helps reduce bottlenecks [7, 18, 28] by shifting responsibility from a central team to business domain teams. This shift allows users to quickly request and access data without the need for a middle person, significantly reducing the lead times. Secondly, business domains can produce data products more rapidly, which can help shorten the time required to bring new products and services to the market [7]. As a consequence of reducing bottlenecks, business domain teams can react quickly to trends and create new analyses, thereby facilitating agility. These characteristics make data mesh an optimal solution for addressing the expanding scale and complexity of data, which are challenging to maintain in current data architectures.

Thirdly, data mesh increases the value of the data by increasing the data quality, interoperability, and accessibility. Centralized data warehouses often lead to complexity. Moving away from a central solution helps create comprehensible data sources.

Fourthly, applying product thinking to data encourages business domain teams to consider the users of the data, thereby increasing the quality of the data. Business domain teams are encouraged to consider users and data quality by using service-level objectives (SLOs). Dehghani introduced the DAUTNIVS principles (see Figure 3.1), which should be used to build high-quality data products [12]. Fifthly, the quality of data increases by moving the responsibilities of the creation and management of data products closer to the source of the data and, thus, closer to the knowledge of the data.

Finally, business domain teams' knowledge can also help create better governance policies for data [7]. The data mesh approach aligns with businesses moving toward an agile approach, particularly in combination with microservice architecture and domain thinking. Data-driven decision-making helps organizations gain a competitive advantage, which can be further increased using a data mesh as a result of increased agility and data quality.

3.1.6 Challenges of data mesh

Migrating to a data mesh architecture presents both organizational and technical challenges. These challenges include cultural shifts, employee acceptance and readiness, skill gaps, and financial investments.

The shift towards a decentralized architecture requires a holistic approach, as it can involve changes to the organization's culture [7]. Implementing data mesh in an organization calls for changes in the skills and mode of operation of certain employees, making the implementation of change management strategies a challenging task.

Business domain teams need to be skilled enough to perform the new tasks assigned to them. Moreover, they should be motivated to create high-quality data products, even though these efforts may not directly benefit their business domains. This situation can potentially lead to friction and resistance [7, 2].

The shift from a single centralized team to multiple decentralized teams requires good governance to avoid errors in data quality, duplication, and policy compliance. Decentralizing responsibilities creates security challenges that can be mitigated by putting good governance in place [18].

Organizations must carefully evaluate whether the transition to a data mesh architecture would be beneficial, as this shift can be quite challenging, costly, and highly influential on the organization's structure. While data mesh is a current trend, it is essential to understand that it is not a one-size-fits-all solution [7]. It should only be considered when organizations are already familiar with a data-driven way of working and feel that their current architecture lacks agility or scalability.

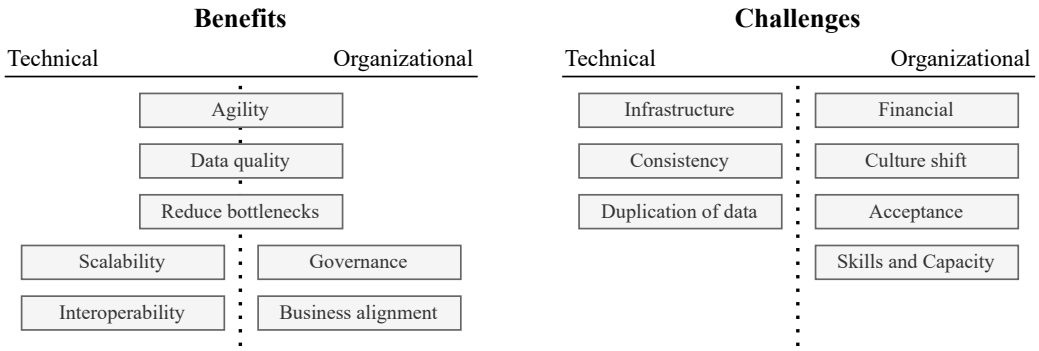


FIGURE 3.2: Benefits and challenges mapped to technical or organizational

3.1.7 Implementation of a data mesh architecture

Data mesh is not a silver bullet. Organizations should carefully consider whether the benefits of a data mesh architecture outweigh those of their current data architecture. Furthermore, organizations must possess a certain level of understanding and readiness before implementing a data mesh within their organization. In their thesis on data mesh, De Boer [8] and Jonkman [25] developed a readiness model and a maturity model, respectively. These models can be used by companies to assess their current situation and their need for data mesh.

A readiness assessment "measures an organization's ability to undertake a transformational change by means of a systematic analysis, while identifying potential challenges that might arise when implementing new procedures and structures within the current organizational context" [8]. De Boer [8] developed a data mesh readiness model, which enables users to assess their readiness to migrate to a data mesh architecture. The assessments examine the necessity, capacity, and preparedness for change and if the principles of data mesh are implemented correctly within the organization.

Jonkman [25] created a maturity model to assess the maturity of a data mesh implementation. In defining maturity models, Jonkman employs the definition of DAMA International. Maturity assessments are utilized to enhance a process based on a model that describes the evolution of characteristics across levels, which indicate an organization's current capabilities and the desired states [25]. Jonkman's maturity model provides insights into the overall maturity of the data mesh implementation and offers detailed insights into five dimensions related to data mesh migration. Organizations may utilize a self-assessment tool by responding to questions corresponding to the five dimensions of data mesh. Additionally, the model offers an overview of maturity across the People, Process, and Technology perspectives.

3.2 Data products

The term 'data products' encompasses various meanings, so we need to precisely define data products and the context of our methodology. According to a definition provided by Dehghani [12], data product "is the node on the mesh that encapsulates three structural components [code, (meta-)data, infrastructure] required for its function, providing access

to the domain's analytical data as a product." However, the term 'data product' is also used outside the data mesh context.

We distinguish three types of data products:

1. **Data-driven services:** Products or services that are built on data, such as AI models, dashboards, and APIs.
2. **Consumption-ready datasets:** A prepared dataset that is easily usable by other parties but is not part of a data mesh architecture.
3. **Data mesh node:** A node within a data mesh architecture consisting of data, metadata, code, and infrastructure.

In our research, we focus on data products within a data mesh architecture. However, consumption-ready datasets can be similar to a data mesh node. The main difference between consumption-ready datasets and data mesh nodes is the context in which they are built. Data mesh nodes are built within a data mesh architecture, which facilitates the socio-technical structure for sharing data between parties.

Data mesh architectures are built on four principles. In addition to the principle of "Data as a Product," three other principles add significant value and ensure the quality of data products. Firstly, the domain-oriented ownership principle helps prevent data siloing and establishes clear responsibilities by dividing an organization into business domains. Secondly, the self-serve data platform provides data users with the necessary tools to develop and access data products efficiently. Finally, federated computational governance addresses governance challenges and ensures the interoperability of data products by, for example, building organization-wide data models and creating policies. The combination of these three principles makes data products within a data mesh architecture effective.

Although the designed methodology can provide insights for practitioners building consumption-ready datasets, this research focuses on building data products for a data mesh architecture, where data management is done by business domains using a domain-agnostic platform.

Data contracts can help build discoverable, trustworthy, and interoperable data products. Data contracts are agreements between data producers and consumers containing promises and rules. Data providers are responsible for providing the set promises. By creating a data contract, a data product can become more discoverable and interoperable. Ultimately, by utilizing data contracts within your data mesh architecture, you standardize the way of storing information about the data product.

3.3 Existing methodologies and approaches

In this section, we explore existing methodologies that could be adopted for data mesh and have inspired our data product development methodology.

3.3.1 CRISP-DM

CRISP-DM is a well-known methodology for data mining projects. In the 1990s, data mining began to become more mainstream. Wirth & Hipp [51] recognized that the success of data mining depended on the expertise of the involved analysts. To make data mining

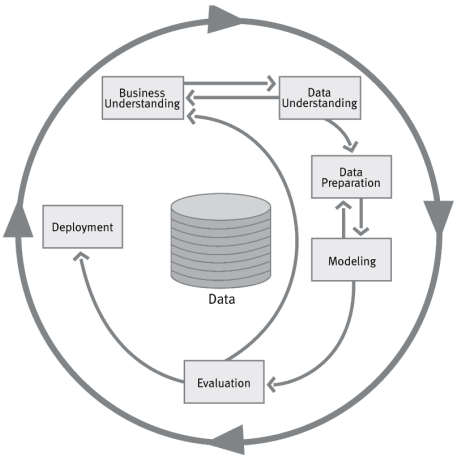


FIGURE 3.3: CRISP-DM Process Model for Data Mining [51]

more comprehensible, they designed the Cross Industry Standard Process for Data Mining (CRISP-DM) process model. The process model was still the most widely-used analytic methodology in 2019, according to various opinion polls [30].

The CRISP-DM methodology consists of four levels of abstraction: phases, generic tasks, specialized tasks, and process instances [51]. A generic reference model presents a quick overview of the methodology, while a user guide dives deeper into step-by-step instructions. CRISP-DM consists of six phases, which can be executed in any order but are visualized using the most frequent dependencies. The visualization of the methodology is shown in Figure 3.3.

In the first phase, *business understanding*, the business objectives are determined, and a project plan is created. The second phase, *data understanding*, is used to get a better understanding of the available data, which can help to better formulate the project plan. Next, in the *data preparation* phase, the initial data is transformed into a final data set, which can be used in the *modeling* phase. During the modeling phase, data mining models are selected, built, and assessed. The results from the models are evaluated, and a decision is made about the deployment of the model. This is done during the *Evaluation* phase. Finally, in the *deployment* phase, the model is deployed, and the project is reviewed. Table 3.3 gives an overview of the phases and their tasks.

TABLE 3.3: Overview of the CRISP-DM reference model

Business understanding	**Data understanding**	**Data preparation**	**Modeling**	**Evaluation**	**Deployment**
Determine business objectives	Collect initial data	Select data	Select modeling technique	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test designs	Review process	Plan monitoring and maintenance
Determine data mining goals	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model	Review project	
Format data					

The CRISP-DM process model is "considered the most complete data mining methodology in terms of meeting the needs of industrial projects" [30]. This process model has several strengths that make it a valuable starting point for developing a data product methodology. First, CRISP-DM is well-documented and complete [30] and has been successfully adapted for various other data mining tasks and machine learning applications [42]. Second, the *business understanding* phase of CRISP-DM aligns well with the principles of a data mesh

architecture, as it focuses on understanding the business objectives and creating a plan before starting on the project. This focus aligns with the core value of data products, which is understanding and delivering customer value. Finally, the iterative nature of CRISP-DM, with its cyclic approach, is well-suited to the continuous evolution and improvement of data products [37].

Despite these strengths, three major shortcomings are of CRISP-DM are identified that make it impossible to directly apply the model to data products: First, the CRISP-DM process model focuses specifically on data mining projects, making it not directly applicable to building data products. Second, CRISP-DM focuses mainly on data, whereas data products commonly consists of pipelines, APIs, governance and documentation. Finally, CRISP-DM is somewhat outdated and may not fully address the challenges of modern data architectures.

While CRISP-DM offers valuable insights and a complete and structured approach, its limitations do not make the process model directly applicable within data mesh architectures.

3.3.2 Design thinking

Design thinking has emerged as a concept with diverse interpretations across various disciplines. [24]. In our research, we utilize design thinking as defined by Brown [9] and Stanford University's D.School [40]. Design thinking applies practices designers use to other practices, including services, products, and strategy development. The steps in design thinking are shown in Table 3.4.

Brown identifies five characteristics of design thinkers: empathy, integrative thinking, optimism, experimentalism, and collaboration [9]. These characteristics are also relevant for data producers in a data mesh architecture. Data producers must deliver value to data consumers, which requires empathy and integrative thinking. Moreover, developing data products is an iterative process that demands collaboration, optimism, and experiments.

Table 3.4: Steps in design thinking

Brown [9]	**Stanford's D.School** [40]
Inspiration	Emphatize
Ideation	Define
Implementation	Ideate
Protoypte	
Test	

Table 3.4 shows the steps involved with design thinking.

Design thinking's customer-centric approach, iterative nature, and focus on prototyping make it a compelling framework for data product development. These characteristics align with the data product thinking principle of data mesh architectures. However, design thinking may be too generic and lacks structure making it not useful as a standalone solution.

3.3.3 Agile, DevOps, DataOps

Agile, DevOps, and DataOps are popular methodologies for software and data engineering. These methodologies build upon each other:

Agile is a software development methodology formalized in the Manifesto for Agile Software Development in 2001. The manifesto focuses on iterative development, customer collaboration, and responding to change. These principles address the limitations of traditional Waterfall development models.

DevOps extends on Agile principles by enhancing communication and collaboration between developers and operators [22]. Jabbari et al. [22] define DevOps as: "a development methodology aimed at bridging the gap between Development (Dev) and Operations, emphasizing communication and collaboration, continuous integration, quality assurance and delivery with automated deployment utilizing a set of development practices."

As organizations adopt data-driven decision-making, DataOps emerged as a new methodology focusing on data engineering. DataOps is often seen as the data alternative to DevOps. Munappy et al. worked towards a definition in his study into DataOps, he defines DataOps as: "as an approach that accelerates the delivery of high-quality results by automation and orchestration of data life cycle stages. DataOps adopts the best practices, processes, tools, and technologies from Agile software engineering and DevOps for governing analytics development, optimizing code verification, building and delivering new analytics, thereby promoting the culture of collaboration and continuous improvement" [34].

All these methodologies focus on collaboration, iterative development and automation. These are all factors that can be seen in the Data as a Product principle of data architectures. However, these methodologies are not directly applicable to build data products, as these methodologies don't focus on data product specific topics like governance. We use these popular data methodologies as inspiration for a new methodology for data product development.

3.3.4 Project management

The development of a data product can be seen as a project. Thus, we investigated project management practices, specifically the Project Management Book of Knowledge (PMBOK) [20]. PMBOK is a structured and comprehensive project framework applicable to any industry and type of project. The PMBOK project management principles guide behavior throughout the project lifecycle. The key principles of PMBOK focus on delivering value and engaging with stakeholders, which are relevant to data products.

The PMBOK is divided into eight project performance domains, which are groups of related activities. These domains provide an approach to project management, including stakeholder performance, project work, and measurement performance. Furthermore, the PMBOK provides models and methods that can be used in these performance domains. Next, deliverables and artifacts are given, that can be used as project output.

Project management is a mature field, and PMBOK is an extensive resource. However, the PMBOK primarily focuses on traditional project management, which may not fully align with the iterative nature of the data as a product princple. We can combine the structured

processes with agile principles to create a more tailored fit for data product development.

3.4 Conclusion

This chapter delves into the data mesh data architecture and its four underlying principles: domain-oriented data ownership, data as a product, self-serve data platform, and federated computational governance. Combining these four principles enables the creation of a decentralized data architecture, where domain experts focus on delivering value from their data by focusing on product thinking. This is achieved by implementing a self-serve data platform, which is created and maintained by a central team. In addition, local and global policies are balanced in order to ensure quality and compliance with regulations. Data mesh is not a silver bullet; therefore, this chapter describes the benefits and challenges of data mesh and provides insights into models that can be used to assess your readiness for data mesh. One of the major challenges is the cultural shift and the need for employee acceptance.

Exploring different types of data products enables us to create a clear definition of data products. Finally, we looked at existing methodologies that can be used as inspiration and adapted for a data product development methodology. The explored methodologies do not address the challenges of data product development. Most importantly, all methodologies lack a specific focus on data products. There is a clear need for a specialized methodology that combines the technical and organizational aspects, incorporates product thinking, and provides an understandable explanation of data product development for domains. We used the existing methodologies as the foundation for the development of a data product development methodology.

Chapter 4

Problem and Requirements

This chapter examines the challenges of developing and maintaining high-quality data products within a data mesh architecture. The first section of the chapter extends our initial problem statement. The following section defines the requirements for the DPDM-DMA methodology. Finally, we dive into the concepts of high-quality data and high-quality data products.

4.1 Problem

The data mesh architecture is a novel paradigm that offers solutions to challenges encountered in traditional data architectures. The shift to decentralization of data management is key to solving organizational challenges faced with data management. Furthermore, decentralization and product thinking align with trends toward sharing data within and between organizations. However, the way of sharing data changes drastically in a data mesh architecture compared to centralized data architectures.

In this section, we dive into technical and organizational challenges associated with developing data products.

4.1.1 System thinking

We try to understand the problem holistically by taking a system perspective of the problem. We employed the CATWOE analysis derived from Soft System Methodology (SSM) [3, 10]. SSM is designed to address problems at a strategic level, which can help us understand our complex problem [36]. SSM is concerned with the study of Human Activity Systems (HAS), which can be defined as a set of activities that interact in order to achieve a specific goal. It focuses on the socio-technical aspects of systems, making it applicable to information systems and, in our case, to data mesh, as data mesh is a socio-technical solution.

'CATWOE' is a mnemonic for Customers, Actor, Transformation process, Weltanschauung (world view), Ownership, and Environment. We utilized CATWOE to define the problem and the relationship between the problem and the actors involved with the development of data products. The analysis is shown in Table 4.1.

Thus, to utilize the advantages of a decentralized data architecture, we require social and technical infrastructure, skills, and governance. The current lack of guidelines and standards for implementing a data mesh architecture leads to challenges when creating

Table 4.1: CATWOE analysis applied to our problem

Customers	Data providers, data consumers
Actor	Data product developer, data product owner, business domain teams, platform team, data governance team
Transformation process	From domain data and customer needs (input) to a value-driven data product (output)
Weltanschauung	A decentralized data architecture offers a number of advantages, including high-quality, scalable, and agile data management, which in turn positively affects decision-making
Ownership	Domain teams
Environment	Available social and technical infrastructure (data maturity), company culture, skills, and governance

high-quality data products. The development of data products is dependent on the implementation of the other principles of data mesh. In the following sections, we explore the problem further.

4.1.2 Relation data mesh principles

The four principles of data mesh are interconnected, so building high-quality data products relies on the effective implementation of each principle. This interdependence underscores the complexity inherent to migrating to a data mesh architecture. Consequently, organizations should have a clear data strategy and migration plan to be capable of building high-quality data products. We dive into the different principles of data mesh and how they are interrelated.

- Domain-oriented ownership establishes the correct organizational structure and organizational strategy. This principle is concerned with creating boundaries within an organization, thereby clarifying responsibilities and data ownership within an organization. By defining business domains, we create clear responsibilities for groups to both produce and consume data.
- Federated computational governance ensures interoperability and adherence to standards across data products. We want to create consistent data products across domains. One approach to interoperability and consistency is the implementation of a template for building data products, providing consistency in the structure of data products. Standards and policies are needed to build interoperable data products.
- Self-serve data platform provides the essential tools and platforms required for developing data products. Data producers rely on appropriate tooling to build and deploy data products. Without the correct infrastructure, it is impossible to create, store, and share data products efficiently.

The interrelation between these principles highlight the holistic nature of the data mesh approach. Organizations must first develop a data strategy that incorporates all four principles: domain-oriented ownership, data as a product, self-serve data infrastructure, and federated computational governance. After the implementation of these principles, we are able to start creating high-quality data products.

4.1.3 Problem definition

Data products in a data mesh architecture are significantly more complex than traditional datasets. They do not only contain data but also code, metadata, infrastructure, and interfaces. The shift from centralized to decentralized data management is a mostly a organizational change, where new actors work on developing data products. Existing methodologies and practices are not applicable to the data mesh architecture.

This process of developing data products is complex, involving numerous steps and skills. High quality data products are of great importance for the entire organization, as decisions are made based on the information they provide. Furthermore, the transition to data mesh is often driven by the desire for enhanced agility. It is essential to ensure the implementation of processes will facilitate this increase in agility.

The primary challenges in the design and maintenance of data products are balancing design and governance for domain and organization-wide needs. While there is research on migration and implementation to data mesh, this is not yet the case for working with data mesh in practice. Ensuring the quality of data products, the data literacy of domain experts, and the interoperability and ownership of data lead to high-quality data.

Many organizations need to work with data and may consider data mesh as their data architecture. This makes the problem relevant to a big audience. Working with data is universal. Helping organizations build high-quality data products can benefit a broad range of industries.

Acceptance by business domains is one of the challenges the novel data architecture faces [2, 7]. According to the Technology Acceptance Model (TAM) [11], perceived usefulness and perceived ease of use are crucial factors for the adoption of technology. A development methodology could help business domain teams to better understand the usefulness of data products. Furthermore, when data products are easy to build, the perceived ease of use should increase. Ultimately, the development of a methodology could help the acceptance of the data mesh architecture.

4.2 Requirements

We defined requirements that were used as guidance during the design and evaluation of the methodology. These requirements can be used to assess the effectiveness of the proposed solution in addressing the problem.

4.2.1 Functional requirements

In order to ensure the successful implementation and operation of the DPDM-DMA, it is important to outline the functional requirements. These requirements provide clear guidance regarding the capabilities and features that the DPDM-DMA must support to achieve its objectives.

- The DPDM-DMA should define the different phases of the data product lifecycle.
- The DPDM-DMA should clarify the goals and tasks of different stakeholders for the data product lifecycle.

- The DPDM-DMA should present a suite of tools for each phase of the data product lifecycle.
- The DPDM-DMA should present expected input and prerequisites for each phase of the data product lifecycle.
- The DPDM-DMA should present expected output and outcomes for each phase of the data product lifecycle.

4.2.2 Structural Qualities

The DPDM-DMA should focus on helping data producers create value for their customers. This is done by listening to the customer and utilizing feedback. We make use of design thinking principles to set the structural qualities of the methodology.

- **User-Centric Focus:** The DPDM-DMA should prioritize understanding and addressing the needs and challenges of end-users to ensure it meets their requirements.
- **Iterative Development:** The DPDM-DMA should support iterative development through prototyping and frequent feedback loops to facilitate continuous improvement.
- **Scalability:** The DPDM-DMA should support large organizations.
- **Modularity:** The DPDM-DMA should support modularity, allowing components to be independently developed, maintained, and reused.

4.2.3 Environmental Qualities

The methodology should be usable, comprehensible, and accessible to different stakeholders:

- **Data mesh alignment:** The DPDM-DMA should align with the core data mesh principles.
- **Usability:** The DPDM-DMA should be usable by data product developers, data product owners, and management.
- **Comprehensibility:** The DPDM-DMA should be easy to understand for data product developers, data product owners, and management.
- **Accessibility:** The DPDM-DMA should be accessible to data product developers, data product owners, and management.
- **Relevance:** The DPDM-DMA should be relevant across diverse types of organizations and sectors.

4.3 Data quality

In this section, we answer our second research question, "What are the characteristics of high-quality data products?" We do this by first delving into the fundamental concept of data quality and then into data products and their quality characteristics.

4.3.1 Data quality

Data quality impacts organizational efficiency and decision-making effectiveness [47]. There are a lot of definitions and attributes for data quality. Wang & Strong [47] take the customer viewpoint of quality and define data quality as *"data that are fit for use by data consumers."*

In their paper, Wang & Strong [47] establish a framework capturing four data quality domain categories, which each represent a construct of data quality (Table 4.2). However, over the years, many different quality domains have emerged. One reason for this could be the use of data in many different contexts, and the need for a quality attribute can vary depending on the context. More recently, Sidi et al. [41] identified 40 different quality domains and their definitions from the literature.

TABLE 4.2: Data quality domains, categorized by Wang & Strong [47].

Category	**Data quality domain**
Intrinsic data quality	Believability, accuracy, objectivity, reputation
Contextual data quality	Value-added, relevancy, timeliness, completeness, appropriate amount of data
Representational data quality	Interpretability, ease of understanding, representational consistency, concise representation
Accessibility data quality	Accessibility, access security

4.3.2 Data product quality

Extending Wang & Strong's definition, high-quality data products can be seen as *data products that are fit for use by data consumers.*

As introduced in Chapter 3, Dehghani [12] defined data product attributes, which serve as the baseline for useful data products within a data mesh architecture. Driessen et al. [15] ensured the relevance of these attributes through the use of semi-structured interviews. Additionally, they identified a new attribute of data products, namely "feedback-driven," making it DAUTNIVS+.

The DAUTNIVS+ attributes focus mainly on data accessibility, which is a key issue that data mesh tries to address. The attributes are widely adopted, and most books, blogs, and other gray literature use the DAUTNIVS+ attributes. By using these attributes in our study, experts can quickly recognize their meaning.

While alternatives exist, such as the widely-cited FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [50], we selected DAUTNIVS+ for several reasons.

Firstly, FAIR focuses on improving the Findability, Accessibility, Interoperability, and Reusability of scientific data. These principles are well-known in the scientific community. However, the specific focus on scientific data can make this principle less generally applicable. Secondly, DAUTNIVS+ attributes are specifically designed for data products within a data mesh architecture, making it more suitable for our study. Thirdly, the DAUTNIVS+ attributes already cover most FAIR principles, making them more complete. This is shown in Table 4.3.

Table 4.3: FAIR principles related to DAUTNIVS+ attributes

Findable		
F1	(Meta)data are assigned a globally unique and persistent identifier	Addressable
F2	Data are described with rich metadata (defined by R1 below)	Discoverable, understandable
F3	Metadata clearly and explicitly include the identifier of the data they describe	
F4	(Meta)data are registered or indexed in a searchable resource	Discoverable
Accessible		
A1	(Meta)data are retrievable by their identifier using a standardised communications protocol	(Natively) accessible, Secure
A2	Metadata are accessible, even when the data are no longer available	
Interoperable		
I1	(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Understandable, Interoperable
I2	(Meta)data use vocabularies that follow FAIR principles	Discoverable, Interoperable
I3	(Meta)data include qualified references to other (meta)data	Interoperable
Reusable		
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes	Understandable, Trustworthy, Secure

4.3.3 DAUTNIVS+

For our research, we utilized DAUTNIVS+ as the quality attributes for data products. We describe these quality attributes.

Discoverable

Data consumers should easily find the data they need. Data producers can facilitate this by sharing information about source, quality, sample datasets, and use cases. In addition, metadata helps make data products more discoverable. A data catalog is often implemented within a data mesh architecture to make data products easily discoverable.

Addressable

In a data mesh architecture, there should be a clear structure of how to address data products, and there should be a consistent pattern throughout the architecture. Data products should have a permanent and unique address. Next to retrieving data, it should be easy to adjust and remove a data product.

Understandable

Data products are made understandable, by explaining use cases and providing sample data, notebooks, and possible use cases. This is needed as data consumers should be able to understand the data product on their own. This means that they should be able to get a good understanding of the attributes and the use of the data product. The data user should be able to understand the attributes, and any ambiguity should be removed by providing documentation.

Trustworthy and trustful

The data provided by the data product should be truthful and credible. The producer of the data product should ensure that the product conforms to data quality domains that

can be established using a data quality framework. In addition, data consumers should be able to quickly trust the data, which can be accomplished by using service level objectives (SLOs) that include information about quality attributes of the data, such as timeliness, completeness, and lineage.

Natively accessible

A data product should be easily accessible to data consumers. This can be done by providing a way to access the data product that is intuitive or familiar with the data consumer's skills and tools.

Interoperable

Data products should be interoperable with other data products, as this can provide powerful insights. Organization-wide standards facilitate interoperability and consistency, which should be applied to various aspects of the data product. There should be standards for describing data products, metadata, and data quality. One way to facilitate this is to use templates to build data products. Rules and policies that facilitate interoperability should be created globally as part of federated governance within a data mesh architecture.

Valuable

A data product should be relevant and add value for its consumers. Furthermore, it should be valuable on its own without using other data products. It is important to track the value of the data product during its lifecycle. This can be done by monitoring indicators such as usage.

Secure

Data products should be secure so that unauthorized users do not have access to data products.

Feedback-driven

A data product should use feedback from consumers to improve the data product. This should be done as early as possible by building prototypes and gathering early feedback. By getting feedback early and employing an iterative approach to data product development, the data product brings more value to the consumer. Data producers are responsible for delivering value, meaning they should continually adjust their data products and process feedback.

Chapter 5

Design and Development

In this chapter, we delve into the structure and function of the Data Product Development Methodology in Data Mesh Architecture (DPDM-DMA). In the first section, we provide an overview of the designed methodology and how it addresses the design problem. Next, we dive into the details of the design.

5.1 Design overview

In this study, we aimed to provide guidelines for building data products within a data mesh architecture to facilitate its adoption, improve data quality, and more effectively utilize data assets.

Johanneson and Perjons [23] describe two essential ways of thinking for generating ideas that can be utilized for the Design Science process: divergent and convergent thinking. Divergent thinking entails imaginative and innovative approaches to creating new ideas. This approach should be combined with convergent thinking, which is a more analytical approach where the designer uses existing ideas and applies this to their design [23].

In this study, we integrate new and innovative ideas with the established methods, principles, and guidelines described in Chapter 3. We employ concepts that have proven effective in other contexts and apply them to the data mesh context.

This section explores the key components of the Data Product Development Methodology within Data Mesh Architecture (DPDM-DMA). The methodology offers a framework that outlines the steps for developing a data product. The methodology provides inputs and outcomes for each step of the development process and offers the necessary skills and tools to execute each step successfully.

5.1.1 Prerequisites

For our methodology, we expect organizations to have fully implemented the data mesh architecture. They should have a self-serve data platform and a data strategy. Furthermore, we expect organizations to have implemented an organization-wide data contract or data model to facilitate interoperability between data products and to provide guidance on the components of a data product.

5.1.2 Structure

We have taken inspiration from the CRISP-DM model [51], which utilizes different levels of abstraction from general to specific. We also describe our model from generic to more specific, beginning with stages, phases, and tasks.

The DPDM-DMA methodology consists of three main stages: a decision-making starting stage, an iterative development stage, and an end stage. In the starting stage, we identify the need for a data product and determine the specific needs of the customer. During the iterative stage, we develop the data product iteratively by gathering feedback from the data consumer. Finally, during the end stage, we retire a data product in case it is not used anymore.

We identified these stages through analysis of our gray literature review, which emphasized the initial identification of problems and business objectives. The initial and iterative stage aligns with steps of Agile methodologies and DevOps and DataOps [22, 34].

5.1.3 Phases

We systematically derived the phases of the methodology by analyzing 20 articles from our gray literature review on building data products (see Appendix B). We employed a thematic analysis for identifying and grouping similar development activities. We grouped terminologies with a similar meaning.

In our analysis, we discovered four primary phases (Table 5.1). The ideation and design phase was observed in 60% of the articles. Next, a develop phase was seen in 30% of the sources. Finally, a deployment and maintenance phase was observed in 45% of the sources. Only 20% of the sources discussed a separate activity for data exploration.

TABLE 5.1: Analysis of data product development phases in gray literature

Phase	Key terms	Source IDs (table B.1)	Frequency
Ideation and Design	"identify", "understand", "design"	G2, G3, G4, G5, G6, G7, G8, G12, G13, G17, G19, G20	12/20
Explore	"collect"	G3, G6	2/20
Develop	"develop", "create", "build"	G1, G6, G10, G13, G16, G19	6/20
Deploy and Maintain	"deploy", "release", "maintenance", "monitor"	G3, G5, G7, G14, G15, G16, G18, G19, G20	9/20

We validated our initial phases through interviews and identified the following considerations for the phases of our methodology. Firstly, we chose to distinguish between data exploration and development by further splitting the development phase into an exploration and a build phase. Secondly, gray literature (G18) and the book of Deghani [13] emphasize the importance of a retirement strategy, so we also added a retirement phase.

Ultimately, our methodology contains the following five phases:

1. Ideation phase assesses the need and feasibility of the data product.
2. Exploration phase explores data and its quality.
3. Build phase constructs different components of the data product.
4. Deploy phase combines these components and deploys and monitors the data product.

5. Retire phase encompasses activities related to the end of the data product lifecycle.

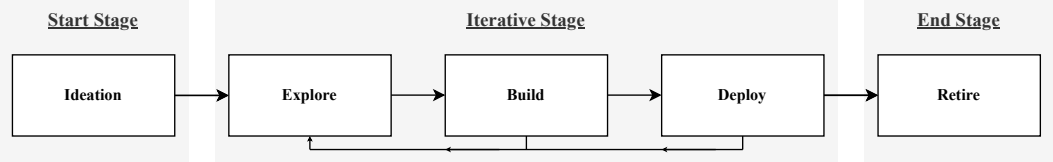


FIGURE 5.1: Phases of the DPDM-DMA

5.1.4 Tasks

Each phase consists of one or more tasks that detail the required actions. These tasks provide structure to the phases and make them accessible. These tasks are based on frameworks and user guides, including DAMA DMBOK, CRISP-DM, and the Design Science Research Framework [21, 23, 51].

To guide the data producer, each task is described using five key factors:

- Task description.
- Skills needed.
- Expected information for the task (input).
- Goal and outcomes of the task (output).
- Models and methods that support the execution of the task.

5.2 Detailed design

This section provides a detailed explanation of the designed methodology. We explain the goals of each phase and describe their tasks. Figure 5.2 gives an overview of the stages, phases, and tasks of DPDM-DMA.

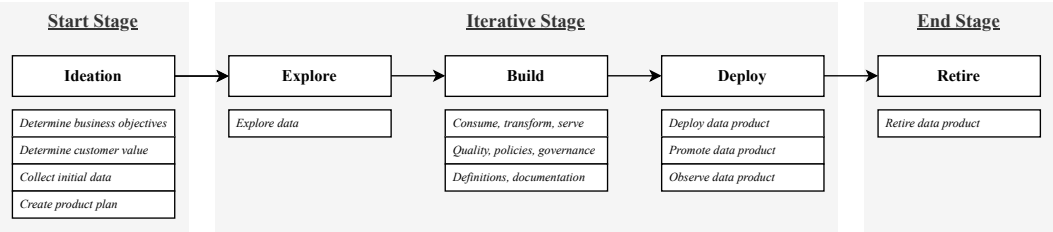


FIGURE 5.2: Overview of stages, phases, and tasks of DPDM-DMA

5.2.1 Ideation phase

In the *ideation* phase, the data producer is tasked with identifying the business objectives, customer needs, and available data. These factors are crucial to determining how to proceed with the build of the data product. Having a well-defined plan and business understanding is crucial before starting a project. In this phase, the data producer should investigate if there is a need for a data product and evaluate whether developing a data product is an

appropriate solution for the problem. Furthermore, the data producer should assess the availability and ownership of the required data. Without this first phase, data producers will be unable to create a successful data product. By first investigating feasibility, we mitigate the risk of spending time and resources in building a data product that may not be feasible or valuable.

Ideation phase - Tasks:

Determine business objectives	DAUTNI**V**S+
It is the responsibility of the data producer to clearly define the problem or opportunity that a data product is designed to address. Each data product needs to add value (on its own), so the data producer should analyze the current landscape and the organization's needs. A well-structured business case should describe how a data product can add value, taking into account both costs and benefits.	
Skills	• Benefits management • Business process improvement • Business situation analysis • Demand management • Innovation • Strategic planning
Input	• Organization-wide vision and goals • Data strategy
Output	• Business objectives – Describe the problem or opportunity the organization wants to solve • Background information – Explain background information • Ubiquitous language – Define common language

Determine customer value	DAUTNI**V**S+
The data producer is tasked with creating a data product that adds value for its customers (data consumers). The data producer is responsible for creating value with their data product. They need to empathize with the customer and understand their needs.	
Skills	• Innovation • User research • Strategic planning • Requirements definition • Risk management
Input	• Background information
Output	• Requirements – Have a clear understanding of the requirements for the data product • Assumptions & constraints • Risks – Identified risks that should be avoided and a plan on how to tackle these risks • User story – Customer point of view and understanding of customer's needs

Collect initial data	DAU**T**N**IVS**+
The data producer is responsible for identifying data that could meet the customer's needs. This is done by collecting data from source systems or from other data products. The data producer should be the owner of the data that is needed for the data product.	
Skills	• Data engineering • Data management • Data science • Knowledge management
Input	• Business objectives • Requirements • User story
Output	• Initial data collection report of customer's needs

Create product plan	DAUTNI**V**S+
The data producer needs to create a plan for developing the data product. The data producer combines the business goals and requirements into a plan, in which they also assign responsibilities. The plan summarizes the goal, stakeholders, schedule, value & benefits, and cost of the data product. The data producer should clearly outline the data product and decide if building a data product is profitable and a suitable solution for our case. When in doubt or inexperienced, they should ask the organization's expert on data mesh and data products for guidance.	
Skills	• Data management • Feasibility assessment • Investment appraisal • Project management • Strategic planning
Input	• Business objectives • Requirements • User story • Risks • Initial data collection report
Output	• Inflection point – Choose to develop the data product or not • Product plan – Document outlines DPDM-DMA, including details on responsibilities, schedule, and budget. • Data product canvas – A one-page summary for data product development

5.2.2 Explore phase

In the *explore* phase, the data producer identifies potential sources for the data product. Data producers gain a better understanding of the available data by performing an in-depth exploration of the data. Data is explored to understand how it should be transformed, how it should be stored, its quality, and what level of governance should be applied.

Explore phase - Tasks:

Explore data	DAU**T**N**IVS+**
The data producer examines the data to better understand what data transformation, data integration, data cleansing, and governance policies to be applied.	
Skills	• Data engineering • Data science • Governance • Quality management
Input	• Initial data collection report • Global governance policies
Output	• Data reports – Data description report – Data exploration report – Data quality report • Governance plan

5.2.3 Build phase

In the *build* phase, data producers utilize the collected information of the explored data and the data consumer to collect, transform, store, and serve the data. Data producers should ensure quality and compliance by establishing governance rules. Ultimately, data producers are responsible for creating understandable data products. This is achieved by adding documentation, a data sample, and computational notebooks to help data consumers understand and gain insight into the data product. After this phase, data producers combine all the information to deploy the data product.

Build phase - Tasks:

Consume, transform, serve	DAUTNIVS+
The data producer needs to consume the data, transform it into an appropriate form, store it, and create interfaces for data consumers. The self-serve data platform should facilitate the infrastructure for these actions. The data producer should consider the input of the data, the best way to store the data, and the form and interface that are best for consumers to access the data. They should carefully consider the data model of the data product.	
Skills	• Data engineering • Data modeling and design • Database design
Input	• Product plan • Data reports • User story
Output	• Definition of input interfaces – The interfaces of the data product need to be defined; this is usually done in the form of code • Definition of output interfaces – The interfaces of the data product need to be defined; this is usually done in the form of code • Data transformations – Code for transforming data in the preferred format • Data storage – Storage for data

Qualities, policies, governance	DAU**T**N**I**V**S+**
Data products should conform to both local and global policies to ensure interoperability, quality, and compliance. Data producers should consider the data we consume and determine whether governance should be applied. Additionally, they should consider the users of the data and whether they should be able to consume the data.	
Skills	• Availability management • Governance • Information assurance • Personal data protection • Service level management • Quality management
Input	• Data quality report • Governance plan
Output	• Policy as code • SLO (Service Level Objectives) – Targeted levels of service • SLI (Service Level Indicators) – Metrics used to measure quality

Definitions, descriptions, documentation	**DAUT**NIVS+
Data providers and consumers need to understand the value and use of the data product. Data producers must present the data in the best possible way. Data consumers are customers, so we should make it easy for them to understand the data. This can be achieved by describing the data (fields), providing data samples, and providing computational notebooks with usage examples.	
Skills	• Data science • Data visualisation • Information content authoring • Knowledge management
Input	• Data description
Output	• Metadata • Data product description • Examples – Such as computational notebooks, sample data, visualizations

5.2.4 Deploy phase

In the *deploy* phase, the information built and collected in the previous phases is leveraged on the self-service data platform to create a version of the data product and deploy a data contract.

After deploying a data product, the data producer should ensure that potential data consumers are aware of the existence and the purpose of the data product. They should actively promote the data product. Additionally, the data producer is responsible for delivering valuable data products, which conveys that they should monitor the data product and address feedback.

Deploy phase - Tasks:

Deploy data product	**DAUTNIVS+**
Use the information from the build phase to create a data contract and deploy the data product by utilizing the self-serve data platform.	
Skills	• Acceptance testing • Product management
Input	• Code • Data • Metadata • Interfaces • Infrastructure
Output	• Data product • Data contract

Determine business objectives	**DAUT**NIVS+
It is the responsibility of data producers to create valuable data products. After creating a data product, data producers should ensure that it is easy to use. A data product is appealing when promises are kept and the product is maintained. Data producers should present and explain their data products on platforms and in places that are available to the entire organization.	
Skills	• Marketing • Stakeholder relationship management
Input	• Organization-wide vision and goals • Data strategy
Output	• Feedback • New users

Observe data product	DAUTNI**V**S+
Data producers are responsible for continually improving and adapting their data products based on the needs of data consumers. Data producers should be open to feedback and continually improve their data products. Data products become of a higher quality when data producers communicate with the data consumers and incorporate their feedback.	
Skills	• Product management • Service level management • Stakeholder relationship management
Input	• Feedback • SLI
Output	• Improvement plan

5.2.5 Retire phase

In the *retire* phase, data producers create a plan for retiring the data product. There can be multiple reasons to retire a data product. Primarily, the data product may no longer be of value to customers, and making adjustments does not add value to the data product. Secondly, a data product may no longer be utilized by data consumers.

Retire phase - Tasks:

Retire data product	DAUTNI**V**S+
It is the responsibility of the data product owner to decide when a data product should be retired. This may be due to the migration to one or multiple new data products, or because the data product is no longer required. The self-serve data platform should be able to facilitate the retirement of data products.	
Skills	• Product management
Input	• SLO • SLI • Product plan
Output	• Plan for retiring data product • Communication to stakeholders

Chapter 6

Demonstration

This chapter demonstrates how the Data Product Development Methodology can help data producers develop high-quality data products with a fictitious case study.

6.1 Case study background

To demonstrate the DPDM-DMA, the case and sample databases AdventureWorks from Microsoft are used [31]. The AdventureWorks database contains data from Adventure Works, a fictional multinational company that manufactures and sells bicycles (parts) in multiple countries on three continents. The company sells both directly to consumers and to other businesses. The database consists of multiple schemas, which lend themselves to being structured into different business domains.

In this section, we describe the use case and the application of the data mesh architecture.

6.1.1 Business domains

A data mesh architecture is decentralized, meaning data management efforts are shared among the business domains of an organization. In this case study, we defined five business domains:

- Purchasing
- Production
- Sales
- Finance
- Human Relations (HR)

Each of these business domains is responsible for producing its data products. The business domains and potential data products are shown in Figure 6.1.

6.1.2 Self-Serve Data Platform

Adventure Works employed a central platform team to implement a self-serve data platform. The platform enables data providers to create, adjust, monitor, and remove data products. It also allows data consumers to search and access data products. The self-serve platform facilitates (autonomous) policy checking.

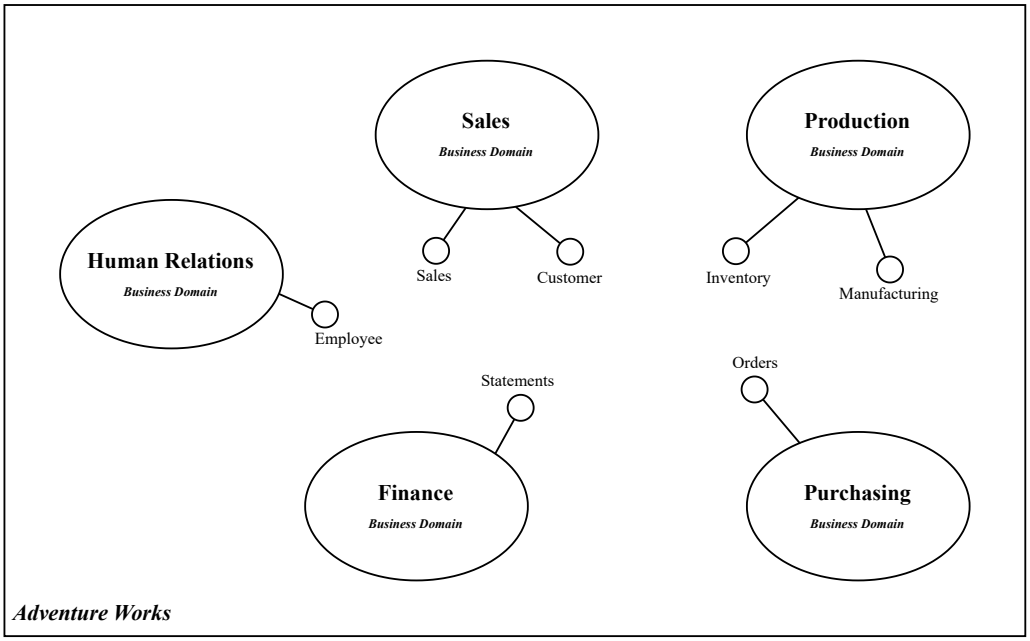


FIGURE 6.1: Business domains and their data products

In a previous study on data mesh, Machado et al. [29] proposed a technical architecture for a self-serve data platform as shown in Figure 6.2. This architecture provides insight into which possible technologies could be adopted by organizations such as Adventure Works.

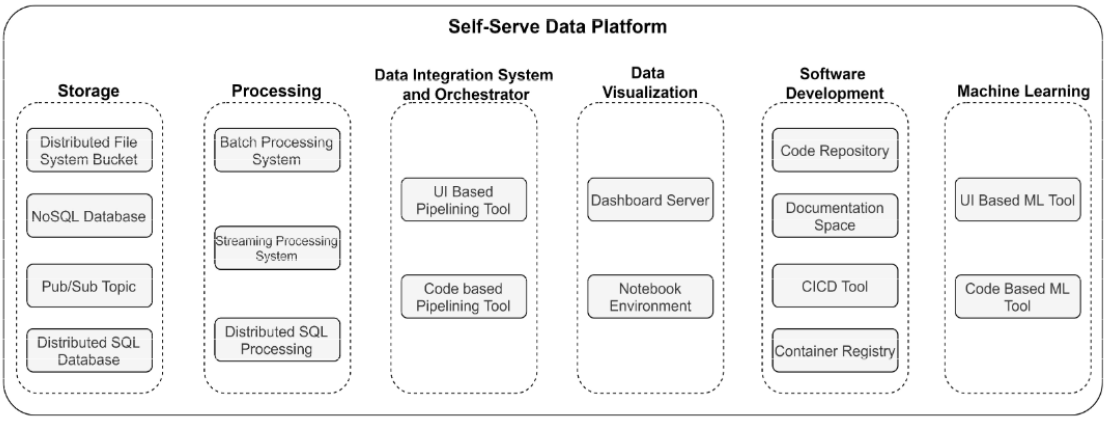


FIGURE 6.2: Self-serve data platform technological architecture (copied from [29])

6.1.3 Data Product Template

Adventure Works utilizes the ProMoTe meta-data model for data products [15] to align data products across the organization. The model can be used to instantiate the different components of data products [15]. Furthermore, the model helps keep data products consistent and thus interoperable across the organization.

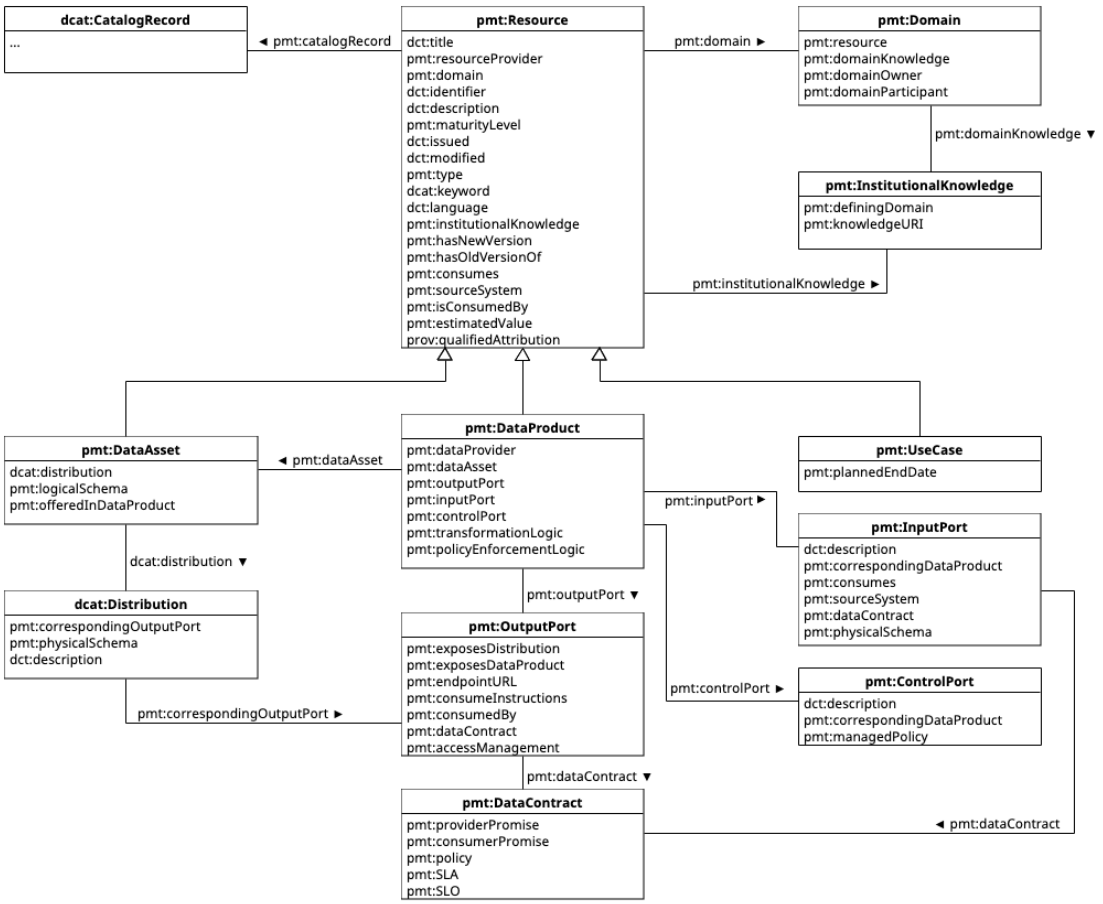


FIGURE 6.3: A UML-representation of ProMoTe (copied from [15])

Adventure Works chose this model because it is designed for a data mesh architecture context and is built with the DAUTNIVS+ attributes in mind. The model can be found on GitHub [14]. The model is a good starting point for the structure of the data product.

We have chosen to use the ProMoTe model, as it is academically validated. However, there are also other models for data products created from the collaboration of multiple practitioners, like the Open Data Contract Standard (ODCS) [6] by the Bitol project [5] and the Open Data Product Specification (ODPS) [35].

6.2 DPDM-DMA application

The sales business domain of Adventure Works goes through all five phases of the DPDM-DMA to build a Consumer Data Product. Figure 6.4 gives an overview of the phases and tasks of the case study. This section discusses the use of the DPDM-DMA, starting with the ideation phase.

6.2.1 Ideation

In the starting stage, the organization establishes the need and added value of developing a data product. In this case, the sales business domain received numerous requests from

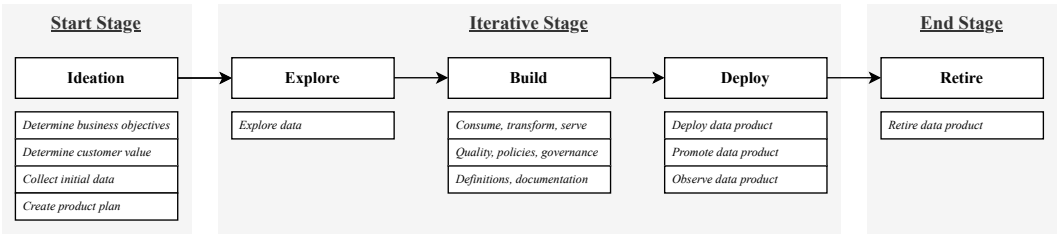


FIGURE 6.4: Overview phases and tasks of demonstration.

other business domains to share customer data. This led the sales domain to decide to develop a data product.

The sales business domain, which is the data producer, employs the DPDM-DMA to develop its data product. This process starts with determining the need and feasibility of developing a data product during the ideation phase.

Determine business objectives

Firstly, the sales business domain studies the company strategy (input) and discovers that Adventure Works aims to be customer-oriented, develop new products that align with customer needs, and attract new customers. Furthermore, the company is committed to becoming even more data-driven.

The data producer decides to write a business case (tool) based on the data requests and company strategy. In this business case, they describe a Consumer Data Product that can help the organization with its marketing analysis and customer relationships (output).

Determine customer value

The data producer discusses the need for a data product with various potential data consumers (tool) to validate this need. From these conversations, they create the following user stories (output):

- As a marketing analyst, I want access to up-to-date demographic data within the Consumer Data Product so that I can create targeted marketing campaigns that align with our customers.
- As a board assistant, I want the ability to analyze customer data within the Consumer Data Product so that I can provide insights for next year's strategy.
- As a financial analyst, I want to gain a clearer understanding of customer purchasing behavior from the Consumer Data Product so that I can develop a price strategy.

These user stories will help out later when developing a product plan and exploring and building the data product. The data producer defines the following requirements (output):

- The data product should contain:
  - Customer demographics
  - Purchase history
- The data product should be accessible by the marketing team, board assistants, and the finance team.

Collect initial data

The data producer analyzes the available data in the different systems. They find customer data in the CRM and ERP systems. The data producer utilizes the user stories (input) to check if the available data could answer their business questions. Luckily, this is the case.

They create a report (output) of the available data, which is used to confirm the availability and ownership of data needed to develop the data product.

Create product plan

The data producer decides to build the data product after they analyze the business objectives, user stories, and the data report (input). Since there is a need for a data product and the data is available, the sales business domain decides (output) to build the Consumer Data Product.

The sales domain utilizes the data product canvas [19] (tool/output) to create an overview of the data product project. This overview can be easily consulted and communicated. The canvas is shown in Figure 6.5.

Data Product Name: Customer data product

Data Objects	Key Tasks	Resource Requirements	Value Propositions	Consumers

Objects:
- Customer demographics
- Purchasing behavior
 	
Sources:
- CRM
- ERP
	- Explore data
	- Build data product
	  - Interfaces
	  - Governance policies
	  - Documentation
	- Deploy
	- Promote
		People:
		- Domain owner
		- Product developer
		- Business analyst
		
		System:
		- Self-serve data platform
			This data product provides:
			Access to accurate consumer demographics & behavior
			
			which will realize:
			Improved company strategy and aligned marketing and customer help.
				- Finance
				- Marketing
				- Board
				- Customer service

Cost	Benefits
Fixed: FTE *Variable:* Infrastructure, Maintenance	Easier data access, improved marketing, better customer help, improved decision-making

FIGURE 6.5: Data product canvas, based on Hasan & Legner [19]

6.2.2 Explore

After deciding to build a data product, the sales business domain completed the starting stage. Next, they move on to the iterative stage, during which the domain iteratively explored, built, and deployed the data product.

Explore data

The sales domain gathers data from multiple sources and analyzes the data types, relations, and quality of the data. To better understand the data, they utilize data exploration and visualization techniques (tools).

The domain utilized the gathered information to create data reports, which summarize

the findings of data relationships, data quality, and needed governance in multiple reports (output). These reports include the data model for the data included in the data products.

6.2.3 Build

The data producers, the sales business domain, now have a clear picture of the business objectives, available data, and the data characteristics. In the following phase, the data producer is responsible for building the different components of the data product. This phase consists of three tasks based on three focus areas:

1. Data producers create the infrastructure for the data in the *consume, transform, serve* task.
2. Data producers focus on policies and governance in the *quality, policies, governance* task.
3. Data producers focus on documentation and communication during the *definitions, description, documentation* task.

Consume, transform, serve

The data producer uses the product plan, data model, and data reports (input) to determine the data needed. This data is then gathered from the source systems, which is done using the self-serve data platform (tool). The platform provides a user-friendly way of consuming, transforming, and storing data.

Data products should be natively accessible. Therefore, the data producer decides to make the data available through an API interface (tool), which complies with the needs of the marketing analyst, board assistant, and financial analyst.

Quality, policies, governance

The data producer uses the data strategy, data reports, and organization-wide guidelines (input) to determine which policies should be applied to the data product. In this case, the data contains Personal Identifiable Information (PII), which can be linked back to a single customer.

The data producer decides to anonymize the dataset by tokenizing sensitive data. However, this means that the data producer has to make some changes to the data. This is done by changing the transformation process, which they do in the *consume, transform, serve* task.

The data producer agrees with potential data consumers that the data is updated monthly and there are no null key fields. These agreements are written down in the SLO (output).

Definitions, descriptions, documentation

The data producer writes documentation (output) on the data product. This helps the consumers understand the data from the data product and clarifies the data quality.

6.2.4 Deploy

In the next phase, the data producer uses the information collected in the build phase to create a data contract and deploy the data product. A deployed data product needs to be promoted and observed.

Deploy data product

Adventure Works uses the ProMoTe meta-data model for data products. The self-serve data platform (tool) helps the producer deploy the data product by providing user-friendly templates to fill in the meta-data model.

To deploy the data product, the data producer utilizes the gathered information from the *build* phase (input). The data producer combines the input- and output interfaces, the data transformations, data storage, policies by code, SLOs, metadata, and descriptions into one single data product.

Promote data product

The data product is launched in the data catalog on the self-serve data platform. However, the data producer is responsible for creating a valuable data product. They decide to contact all interviewed stakeholders (the marketing analyst, board assistant, and financial analyst) about the data product deployment.

The marketing analyst is really enthusiastic about the new data product. However, they ask for data on postal codes (tool). The data producer initially anonymized this data.

Observe data product

The data producer receives feedback from the marketing analyst (input) and decides to change the data product. To do this, the data producer goes through the *build* phase again. They reconsider the governance rules and create a new access point specifically for the marketing analyst to access sensitive data.

6.2.5 Retire

After some years, the data product is no longer used. The marketing analyst uses a new strategy to determine targeted campaigns, the board assistant utilizes other data to determine his strategy, and the financial analyst uses other sources to do their job.

The business domain decides to retire the data product.

Retire data product

The Service Level Indicators (tool) show reduced usage of the data product. The data producer decides to retire it. They inform all stakeholders that the data product will retire after 1 year (output).

Chapter 7

Evaluation

This chapter discusses the evaluation of the Data Product Development Methodology within a Data Mesh Architecture (DPDM-DMA). The DPDM-DMA has been evaluated during the iterative design and development phase (formative evaluation) and after the final design (summative evaluation). Furthermore, this chapter considers the ease of use and assesses whether the methodology meets the defined requirements.

7.1 Formative evaluation: expert-interviews

During the design phase, we utilized semi-structured interviews to evaluate the methodology and gather practitioner feedback. The feedback is used to refine the DPDM-DMA. This section describes the collected feedback and recurring themes from the expert interviews.

We have interviewed a total of seven experts on two different versions of our methodology. Table 7.1 shows an overview of the seven interviewed experts, including roles and years of data-related experience. With the interviews, we assessed the functional requirements and the structural and environmental qualities of the DPDM-DMA. The interview script can be found in Appendix C.

Table 7.1: Interview participant overview

Expert	Role	Industry sector	Years of experience	Familarity Data Mesh
I	PhD student	-	1 - 3 years	Expert
II	Data engineer	Retail	1 - 3 years	Very familiar
III	Consultant	Technology	10+ years	Expert
IV	Executive	Technology	10+ years	Very familiar
V	Consultant	Technology	10+ years	Very familiar
VI	Executive	Technology	10+ years	Expert
VII	Architect	Government	10+ years	Very familiar

7.1.1 Interview round 1

During the first interviews, the overwhelming majority of interviewees agreed with the strength of the structure and found the methodology well-documented and clear. While several interviewees mentioned that their steps in developing a data product differ in name

or number, they agreed that their steps could be mapped to the phases of the DPDM-DMA.

Several interviewees emphasized that there needs to be a more concrete distinction between determining the need for a data product and building one. As one interviewee stated, "You need a go/no-go moment." We incorporated this feedback by changing the relationship between the phases of the methodology and by introducing a higher level of abstraction to the methodology. Further, we specifically mentioned that the ideation and retire phase are the start and ending phase, respectively we changed the visualization to clarify that these phases are not part of the iterative cycle.

Most interviewees indicated that certain areas needed more detail or better explanations, particularly regarding specific tools or steps that were not explained in full detail. We gathered all points for which interviewees asked for clarification and improved the description or naming accordingly.

A few interviewees were curious about our definition of data product and data mesh architecture. Due to the novelty of the topic and the different interpretations of practitioners, there is a lack of a commonly agreed definition for data products. The interviewees suggested adding our definition to the methodology.

Additionally, one interviewee asked to improve the stakeholders section in the methodology and suggested adding skills to minimize ambiguity.

7.1.2 Interview round 2

One interviewee highlighted the importance of using a template or model to guide data producers when building data products. A model can illustrate the components of a data product. Furthermore, data products are better interoperable when built using the same model. An interviewee in round 1 also mentioned the usage of models.

Another interviewee commented that he would benefit from a priority on the tools used in each task, so he suggested distinguishing the importance of each tool. We applied this by adding MoSCoW prioritization method to the tools.

One interviewee with a background in the public sector provided interesting insights. He stressed the importance of an ethical assessment when building data products. Next, he pointed out the importance of monitoring and checking (ethical) policies. This is especially important for organizations that need to justify their operations, such as organizations in the public sector.

7.2 Summative evaluation: Survey

We evaluated the final design by utilizing a survey, which is a way to collect responses to the same set of questions. This section first discusses the response rate and analyzes the findings.

The overall response to the survey was poor, with only four respondents. We targeted interviewees from the formative evaluation, colleagues, and experts from a data mesh community, but despite pre-inviting interviewees from the formative evaluation to participate in the survey, the number of responses was lower than the number of interviews performed.

Table 7.2: Survey questions

Perceived usefulness	
U1	Using the data product methodology would reduce the effort required to develop data products.
U2	Using the data product methodology would improve the quality of produced data products.
U3	Using the data product methodology would increase my productivity.
U4	Using the data product methodology would make it easier to do my job.
U5	Using the data product methodology would make the migration to data mesh easier.
U6	Overall, I found this data product methodology useful.
Perceived ease of use	
E1	Using the data product methodology would be easy for me.
E2	I found the data product methodology clear and understandable.
E3	I found the phases of the data product methodology easy to understand.
E4	I found the tasks of the data product methodology easy to understand.
E5	Overall, I found this data product methodology easy to use.
Intention to use	
I1	I intend to use this data product methodology if I have to develop data products in the future.
I2	I plan to incorporate this methodology into my data product development practices.

We undertook multiple attempts to gain more responses by reminding the target group, however, without success. One of the reasons for the low response rate may be the length of the survey [39]. Further causes could be the specialized nature of the target group or potential survey fatigue among participants.

The design of the questionnaire was based on the Method Evaluation Model (MEM). The survey includes questions about respondents' demographics and data mesh experience, feedback on the DPDM-DMA using questions on a Likert scale questions, and optional open-ended questions for extra feedback. Table 7.2 shows the survey questions on perceived usefulness, perceived ease of use, and intention to use

All respondents are based in the Netherlands and have a role in data. The sample comprises three data engineers and one data architect. The participants had between one and three years of experience (n = 2), seven and ten years of experience (n = 1), and over ten years of experience (n = 1). They indicate that they are very familiar with data mesh architectures.

The responses were too low to draw any statistically justified conclusions. However, the respondents provided positive feedback across all dimensions of the survey. On the 7-point Likert scale, all items received scores at the positive end, as shown in Table 7.3.

Table 7.3: Average score for each survey question, rated on a 7-point Likert scale

Perceived usefulness	5.38	Perceived ease of use	5.95	Intention to use	5.25
U1	5.25	E1	5.75	I1	5.75
U2	5.50	E2	6.25	I2	4.75
U3	5.25	E3	6.50		
U4	5.25	E4	5.75		
U5	5.25	E5	5.50		
U6	5.75				

Respondents indicated that the DPDM-DMA would reduce the effort required to develop data products (U1). Further, the respondents agreed that the methodology would improve the quality of data products (U2) and increase productivity (U3). Finally, the respondents indicated that the DPDM-DMA would make the migration to data mesh easier (U5), which aligns with our requirement to develop a methodology that focuses on data mesh architectures. Overall the proposed methodology is found useful (U6).

The DPDM-DMA is perceived as easy to use (E5). Respondents found the methodology easy to use, clear, and understandable (E1, E2). The phases of the DPDM-DMA were partially well received by the respondents (E3), which shows the successful implementation of the different phases of the data product lifecycle. These strong scores indicate that the methodology complies with the usability, comprehensibility, and accessibility requirements.

The respondents suggest that they intend to use the methodology (I1) and incorporate it into their development practices (I2), which shows the relevance of the methodology.

The participants reported that they found the phases quite clear. Further, one respondent mentioned this would also benefit management as it would give a great overview of the steps that should be taken by an entire company. Additionally, one participant indicated that it would be challenging to implement this methodology as these choices are made by high-level management. One participant observed that the methodology is quite detailed, which can present challenges in terms of implementation.

Although the response rate was low, we note that data mesh is a novel and specialized field, consequently, the insights of these four experts are valuable. We have chosen only to target experts in data mesh architectures to ensure the high quality of the responses. However, this survey could have been forward to general data experts.

Despite the low response rate, the survey yielded positive responses, suggesting a positive reception amongst the respondents. Consequently, the findings underscore the potential value of conducting further studies on DPDM-DMA.

Chapter 8

Final Remarks

This chapter presents our final remarks on developing a development methodology for data products within a data mesh architecture (DPDM-DMA). First, we discuss our study's primary findings and their implications, explaining the implications to theory and practice. Then, we critically reflect on the study and examine its limitations. Finally, we draw the final conclusions and propose directions for future research.

8.1 Discussion

8.1.1 Interpretation of findings and implications

In this study, we designed and evaluated the Data Development Methodology within a Data Mesh Architecture (DPDM-DMA) using a gray literature review, interviews, and a survey. Existing data and software methodologies do not properly align with the socio-technical structure of data mesh. The DPDM-DMA solves this by focusing on data products within a data mesh architecture.

We first needed to determine the definition and attributes of high-quality data products. We have chosen to utilize the DAUTNIVS+ attributes as defined by Dehghani [12] and validated by Driessen et al. [15]. We considered different attributes, such as the FAIR principles [50]. Nevertheless, the DAUTNIVS+ attributes are the standard for describing high-quality data products, making them most suitable for a methodology for beginners. Additionally, we found that there is a lot of overlap between the DAUTNIVS attributes and the FAIR principles, which is an interesting finding.

We looked at different abstract levels for data product development. From general to specific, we defined stages, phases, and tasks for developing data products. We identified three key stages using existing methodologies and expert interviews: (1) a decision-making starting stage, (2) an iterative development stage, and (3) an end stage. This structure aligns with other methodologies such as agile development and design thinking [9, 22]. We emphasize the importance of first determining the need and feasibility of the data product, which is the first stage of the DPDM-DMA.

Through our research, we identified five distinct phases for developing data products. These phases were designed based on the analysis of gray literature on developing data products, existing methodologies [20, 22, 34, 51], and the formative evaluation. Both interviews and a questionnaire validated these phases' utility. The phases provide a useful

overview for management and data producers.

DPDM-DMA provides structured and practical guidelines by introducing tasks and their inputs, outputs, needed skills, and useful tools. These are practical and actionable steps as shown in the demonstration. The demonstration shows the applicability of the DPDM DMA. The demonstration also shows the importance of an iterative process. Finally, both the DPDM-DMA and the demonstration demonstrate the importance of and reliance on an appropriate self-serve data platform.

The data product development methodology addresses two critical gaps: the lack of structured guidelines for data product development and the organizational challenges of implementing data mesh.

The evaluation of the methodology indicates that the DPDM-DMA would be useful and easy to use. The methodology was well received by respondents, who agreed on its usefulness and perceived it as a great guide and overview of the phases in data product development. These results may have implications for the organizational and cultural friction as described in the use case study from Vestues et al. or the study from Bode et al. [7, 46].

8.1.2 Limitations

The study has the following limitations:

- The novelty of the topic forced us to use gray literature in the literature review. Although the gray literature was aligned with the most recent practices and provided insights into the latest state of practice, it focused on more abstract levels of the data mesh implementation, which influenced the knowledge required to explicate the problem.
- The requirements and first methodology outline were developed through a synthesis of the gray literature. The involvement of experts in this research phase would have improved the problem definition and requirements.
- During the design phase, we had difficulty finding suitable experts to interview. Due to the limited number of experts interviewed, we may have missed insights.
- Due to limited time and resources, it was not possible to conduct an evaluation based on a real use case. However, due to the socio-technical nature of the research, a study in a naturalistic setting is needed to evaluate the impact of the designed methodology in a social setting. The fictional case study provided insight into the use of our methodology. However, a real-world case study would have provided stronger empirical evidence.
- Despite the large number of experts we reached out to, the questionnaire had a small number of respondents, which limited the evaluation and may influence the validity and confidence with which a conclusion could be drawn. We considered reaching out to a wider audience by consulting data experts who are not familiar with data mesh architectures. However, this could impact the quality of the responses because the methodology is closely tied to data mesh architecture, which is complex and not easy to grasp.

Although the current study is based on a small sample of participants, the findings suggest the usefulness of the DPDM-DMA. Future research should aim to address these limitations through real-world case studies and broader expert engagement.

8.2 Conclusion

This study advances the field of data management by developing a comprehensive methodology for data product development within data mesh architectures, addressing challenges posed by decentralizing data management. We applied the method framework for Design Science Research by Johannesson and Perjons [23] to structurally design and evaluate the DPDM-DMA, which is our methodology to guide data producers to build high-quality data products.

This research showed that traditional software and data management methodologies are unsuitable for building data products within a data mesh architecture. Due to their fundamentally different organizational and technical structures. Data mesh is a decentralized data architecture that is characterized by domain-driven ownership and federated governance. Further, it relies on the availability of a self-serve data platform. Data products encompass not only data but also code, metadata, and infrastructure. Ultimately, these factors limit the applicability of traditional methodologies, and DPDM-DMA serves as a starting point for research on building data products within a data mesh architecture.

This research has identified the relations, importance, and dependency of the data mesh principles in the context of data product development. Additionally, we explored and integrated the DAUTNIVS+ attributes (Discoverable, Addressable, Understandable, Trustworthy, Natively accessible, Interoperable, Valuable, Secure, Feedback-driven) into DPDM-DMA. This study shows that by utilizing the DAUTNIVS+ attributes and data mesh principles, we can create data products that roughly comply with the FAIR data principles.

The designed methodology provides a first guideline for developing data products. It was designed with the help of experts to ensure completeness. We have shown its applicability by demonstrating the methodology in a fictitious case study. Furthermore, we utilized a survey to assess the perceived usefulness and ease of use. Our evaluations indicate the usefulness of the proposed methodology.

The methodology provides a structured approach and much-needed guidelines [18] for data product development. DPDM-DMA makes data product development more comprehensible and can therefore reduce the organizational challenges of data mesh migration.

Further studies are needed to validate and refine DPDM-DMA. These studies should focus on:

1. Conducting real-world case studies to assess the methodology's impact over time and investigate its utility and complications.
2. Exploring applications of the methodology in different sectors and industries, as interviews already showed different focus points between public and private organizations.
3. Investigating the main differences between the development of datasets in a centralized architecture and a data product in a decentralized architecture, to further clarify the differences and challenges presented by decentralized data architectures.

This future research will help create a deeper understanding of decentralized architectures and sharing data as a product in other environments, such as international data spaces and data markets.

In conclusion, this research contributes a methodology for data product development within data mesh architecture. The DPDM-DMA offers a structured approach to building high-quality data products adhering to the DAUTNIVS+ attributes.

Bibliography

[1] Rene Abraham, Johannes Schneider, and Jan Vom Brocke. "Data governance: A conceptual framework, structured review, and research agenda". In: *International journal of information management* 49 (2019), pp. 424–438.

[2] Astri Moksnes Barbala, Geir Kjetil Hanssen, and Tor Sporsem. "Towards a Common Data-Driven Culture: A Longitudinal Study of the Tensions and Emerging Solutions Involved in Becoming Data-Driven in a Large Public Sector Organization". In: *Available at SSRN 4658343* ().

[3] Richard Baskerville, Jan Pries-Heje, and John Venable. "Soft design science methodology". In: *Proceedings of the 4th international conference on design science research in information systems and technology*. 2009, pp. 1–11.

[4] Richard Berntsson Svensson and Maryam Taghavianfar. "Toward becoming a data-driven organization: challenges and benefits". In: *Research Challenges in Information Science: 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23–25, 2020, Proceedings 14*. Springer. 2020, pp. 3–19.

[5] Bitol. *Bitol*. URL: https://bitol.io/.

[6] Bitol. *GitHub - bitol-io/open-data-contract-standard: Home of the Open Data Contract Standard (ODCS)*. URL: https://github.com/bitol-io/open-data-contract-standard.

[7] Jan Bode et al. "Data mesh: motivational factors, challenges, and best practices". In: *arXiv preprint arXiv:2302.01713* 10 (2023).

[8] Willemijn de Boer. "A Meshed Up Data Architecture Design". In: (2022).

[9] Tim Brown et al. "Design thinking". In: *Harvard business review* 86.6 (2008), p. 84.

[10] Peter Checkland and Jim Scholes. *Soft systems methodology in action*. John Wiley & Sons, 1999.

[11] Fred D Davis, RP Bagozzi, and PR Warshaw. "Technology acceptance model". In: *J Manag Sci* 35.8 (1989), pp. 982–1003.

[12] Z. Dehghani. *Data Mesh*. O'Reilly Media, 2022. ISBN: 9781492092360.

[13] Zhamak Dehghani. *Data Mesh Principles and Logical Architecture*. URL: https://martinfowler.com/articles/data-mesh-principles.html (visited on 02/20/2024).

[14] Stefan Driessen. *GitHub - Stefan-Driessen/ProMoTe: Repository for maintaining the Data Product Model template*. GitHub. URL: https://github.com/Stefan-Driessen/ProMoTe (visited on 03/15/2024).

[15] Stefan Driessen, Willem-Jan van den Heuvel, and Geert Monsieur. "ProMoTe: A Data Product Model Template for Data Meshes". In: *International Conference on Conceptual Modeling*. Springer. 2023, pp. 125–142.

[16] Stefan W Driessen, Geert Monsieur, and Willem-Jan Van Den Heuvel. "Data market design: A systematic literature review". In: *Ieee access* 10 (2022), pp. 33123–33153.

[17] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering". In: *Information and software technology* 106 (2019), pp. 101–121.

[18] Abel Goedegebuure et al. "Data mesh: a systematic gray literature review". In: *arXiv preprint arXiv:2304.01062* (2023).

[19] M Redwan Hasan and Christine Legner. "Data Product Canvas: A visual inquiry tool supporting data product design". In: *International Conference on Design Science Research in Information Systems and Technology*. Springer. 2023, pp. 191–205.

[20] P.M.I.P.M. Institute. *A Guide to the Project Management Body of Knowledge (PMBOK® Guide) – Seventh Edition and The Standard for Project Management (KOREAN)*. PMBOK® Guide. Project Management Institute, 2021. ISBN: 9781628257120. URL: https://books.google.nl/books?id=aZBFEAAAQBAJ.

[21] D. International, D. Henderson, and S. Earley. *DAMA-DMBOK: Data Management Body of Knowledge*. DAMA-DMBOK: Data Management Body of Knowledge. Technics Publications, 2017. ISBN: 9781634622349. URL: https://books.google.nl/books?id=YjacswEACAAJ.

[22] Ramtin Jabbari et al. "What is DevOps? A systematic mapping study on definitions and practices". In: *Proceedings of the scientific workshop proceedings of XP2016*. 2016, pp. 1–11.

[23] Paul Johannesson and Erik Perjons. *An introduction to design science*. Vol. 2. Springer, 2021.

[24] Ulla Johansson-Sköldberg, Jill Woodilla, and Mehves Çetinkaya. "Design thinking: Past, present and possible futures". In: *Creativity and innovation management* 22.2 (2013), pp. 121–146.

[25] Christian Jonkman. "Organisational Maturity Assessment during the Paradigm Shift from Monoliths to Data Mesh: Design Science Research in Developing a Data Mesh Maturity Assessment Model". In: (2023).

[26] Barbara Kitchenham and Stuart Charters. *Guidelines for performing systematic literature reviews in software engineering*. 2007.

[27] M.L. Langedijk. *Mesh up your Data Architecture - Data Mesh: A Scoping Review*. 2024.

[28] Antti Loukiala et al. "Migrating from a centralized data warehouse to a decentralized data platform architecture". In: *International Conference on Product-Focused Software Process Improvement*. Springer. 2021, pp. 36–48.

[29] Inês Machado, Carlos Costa, and Maribel Yasmina Santos. "Data-driven information systems: the data mesh paradigm shift". In: (2021).

[30] Fernando Martínez-Plumed et al. "CRISP-DM twenty years later: From data mining processes to data science trajectories". In: *IEEE transactions on knowledge and data engineering* 33.8 (2019), pp. 3048–3061.

[31] Microsoft. *AdventureWorks Sample Databases*. URL: https://github.com/Microsoft/sql-server-samples/tree/master/samples/databases/adventure-works.

[32] Microsoft. *Microsoft Forms*. URL: https://forms.office.com/.

[33] Daniel L Moody. "The method evaluation model: a theoretical model for validating information systems design methods". In: (2003).

[34] Aiswarya Raj Munappy et al. "From ad-hoc data analytics to dataops". In: *Proceedings of the International Conference on Software and System Processes*. 2020, pp. 165–174.

[35] *Open Data Product specification - Linux Foundation*. URL: https://opendataproducts.org/.

[36] A Platt and S Warwick. "Review of soft systems methodology". In: *Industrial Management & Data Systems* 95.4 (1995), pp. 19–21.

[37] Jeffrey S Saltz. "CRISP-DM for data science: strengths, weaknesses and potential next steps". In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE. 2021, pp. 2337–2344.

[38] M.N.K. Saunders, P. Lewis, and A. Thornhill. *Research Methods for Business Students*. Pearson Education, 2019. ISBN: 9781292208794. URL: https://books.google.nl/books?id=TMGYDwAAQBAJ.

[39] Mark Saunders. "Research methods for business students". In: *Person Education Limited* (2019).

[40] Scott et al. *Design Thinking Bootleg*. Tech. rep. d.school at Stanford University and Doorley, 2018.

[41] Fatimah Sidi et al. "Data quality: A survey of data quality dimensions". In: *2012 International Conference on Information Retrieval & Knowledge Management*. IEEE. 2012, pp. 300–304.

[42] Stefan Studer et al. "Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology". In: *Machine learning and knowledge extraction* 3.2 (2021), pp. 392–413.

[43] Davide Tosi, Redon Kokaj, and Marco Roccetti. "15 years of Big Data: a systematic literature review". In: *Journal of Big Data* 11.1 (2024), p. 73.

[44] John Venable, Jan Pries-Heje, and Richard Baskerville. "A comprehensive framework for evaluation in design science research". In: *Design Science Research in Information Systems. Advances in Theory and Practice: 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings 7*. Springer. 2012, pp. 423–438.

[45] John Venable, Jan Pries-Heje, and Richard Baskerville. "FEDS: a framework for evaluation in design science research". In: *European journal of information systems* 25.1 (2016), pp. 77–89.

[46] Kathrine Vestues et al. "Agile data management in NAV: a case study". In: *International Conference on Agile Software Development*. Springer International Publishing Cham. 2022, pp. 220–235.

[47] Richard Y Wang and Diane M Strong. "Beyond accuracy: What data quality means to data consumers". In: *Journal of management information systems* 12.4 (1996), pp. 5–33.

[48] Arif Wider, Sumedha Verma, and Atif Akhtar. "Decentralized data governance as part of a data mesh platform: Concepts and approaches". In: *2023 IEEE International Conference on Web Services (ICWS)*. IEEE. 2023, pp. 746–754.

[49] Roel J Wieringa. *Design science methodology for information systems and software engineering*. Springer, 2014.

[50] Mark D Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3.1 (2016), pp. 1–9.

[51] Rüdiger Wirth and Jochen Hipp. “CRISP-DM: Towards a standard process model for data mining”. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1. Manchester. 2000, pp. 29–39.

Appendix A

Protocol for Gray Literature Study on Data Products

Background

We undertake a gray literature review to identify the challenges for managing data products. Furthermore, we want to identify the activities, characteristics, and tools used for working with data products. The activities will be characterized.

A gray literature review can help get close the gap between academic and professional practice. In our case data mesh is a novel paradigm, and there is a lack of academic literature on the topic. A gray literature study can help in this case to get industry insights on the topic. Thus we chose to use both white and gray literature in our literature review, thus choosing for a gray literature study.

We will use the extracted data for creating objectives and requirements for a artefact that helps organizations with managing data products.

Research questions

To design our research questions we use PICOC criteria as described by Kitchenham & Charters [26].

- **Population:** Domain teams or organizations looking to adopt data mesh.
- **Intervention:** Applying product-thinking to data in a data mesh architecture
- **Comparison**: /
- **Outcome:** Improved data quality and data maintaince. Improved efficiency of domain teams.
- **Context:** Data mesh architectures

These criteria are used to define the following research questions:

- What are the essential activities in the design and maintenance of data products in a data mesh architecture?

Inclusion Criteria	Exclusion Criteria
Published after 2019	Article from low-reputation websites
Language is English	Article doesn't focus on data products
	Article doesn't focus on data mesh
	Article is of low quality (quality assessment)

TABLE A.1: Inclusion and exclusion criteria

- What steps are characterized for the design and maintenance of data products in a data mesh architecture?
- What tools are used in the design and maintenance of data products in a data mesh architecture?
- What roles are involved with the design and maintenance of data products in a data mesh architecture?
- What problems do organizations experience when designing and maintaining data products in a data mesh architecture?

Search strategy

We will use both gray literature. For our gray literature review we will be using a general web search engine: Google, specialized websites: www.datameshlearning.com, and snowballing.

We use the following search terms: 1:`(build OR design OR maintenance OR maintain) AND "data product"`

In the case of the gray literature, a different approach is necessary in comparison with white literature, given that the search on Google yielded more than 4000 results. Consequently, a search will be conducted until no further results are yielded.

Selection

The inclusion and exclusion criteria are stated in Table A.1. We will determine the quality of the papers during the selection using the quality assessment described in the following section.

The selection will be performed using the following steps:

1. Reading the title.
2. If there are doubts, we will read the introduction or abstract of the article.
3. If there are doubts, we will further decide during the data collection phase.

Quality assessment

We have developed the following quality checklist based on Garousi et al. [17]. In order to align the checklist with the search strategy, adjustments were made. The following criteria and questions are used to access the quality of an article:

- Authority of the producer
- Methodology
- Objectivity
- Date
- Position
- Novelty
- Impact
- Outlet type

Data collection

The following data will be extracted:

- The source
- Publication date
- Date of extraction
- Author and/or Organization
- Quality Assessment
- Context of data architecture
- Activities
  - Description
  - Objectives
  - Tools
  - Stakeholders
  - Steps
- DP Characteristics
- Challenges in data product design

Data analysis

We will answer the research questions by using the collected data, Table A.2 shows the relation between the collected data and the research questions.

Research Question	**Data Collection**
What are the essential activities in the design and maintenance of data products in a data mesh architecture?	Design or maintenance activitiy
What steps are characterized for the design and maintenance of data products in a data mesh architecture?	Design or maintenance activitiy, objectives
What tools are used in the design and maintenance of data products in a data mesh architecture?	Tools
What roles are involved with the design and maintenance of data products in a data mesh architecture?	Stakeholders
What problems do organizations experience when designing and maintaining data products in a data mesh architecture?	Challenges in data product design

Table A.2: Correlation research questions and data collection

Appendix B

Gray Literature Review - Sources

Table B.1: Sources of the gray literature review

ID	Title	Author	Year	Type	URL
G1	Building an Analytics Ecosystem of Global Data Products at Adevinta; Data Mesh Learning Meetup #12	Real S. & Gumara Rigol X.	2021	Webinar	https://www.youtube.com/watch?v=av6cT_r4orQ
G2	5 essential steps to building great data products	Seow I.	2023	Blog post	https://www.thoughtspot.com/data-trends/product-management/data-product
G3	How to Build Data Products: Strategies for 2024 and Beyond		2023	Blog post	https://atlan.com/how-to-build-data-products/
G4	Building Effective Data Products: A Step-by-Step Guide	Dahlager T.	2024	Blog post	https://www.analytics8.com/blog/building-effective-data-products-a-step-by-step-guide/
G5	How to Build Data Products	Kumar A., Shubhanshu J., Ghosh S.	2023	Blog post	https://moderndata101.substack.com/p/how-to-build-data-products-design
G6	The Ultimate Guide to Data Products	Horner M	2023	Blog post	https://www.timextender.com/blog/product-technology/the-ultimate-guide-to-data-products
G7	A Not-to-Miss Opportunity: How to Build Data Products	West C.	2023	Blog post	https://www.instinctools.com/blog/building-data-products/
G8	Designing Data Products			Blog post	https://www.datamesh-architecture.com/data-product-canvas
G9	Build data products in a data mesh			Technical Documentation	https://cloud.google.com/architecture/build-data-products-data-mesh
G10	Data Product Thinking: Treating Data as a Product in a Data Mesh Environment	Owczarek D.	2023	Blog post	https://nexocode.com/blog/posts/data-as-a-product-in-data-mesh/
G11	Using DataOps To Build Data Products and Data Mesh	Segner M.	2023	Blog post	https://www.montecarlodata.com/blog-how-roche-uses-dataops-to-build-data-products-and-data-mesh/
G12	Building An "Amazon.com" For Your Data Products	Moses B., Jain M., Porto P.	2023	Blog post	https://www.thoughtworks.com/insights/blog/data-strategy/building-an-amazon-com-for-your-data-products
G13	Data Product Examples To Get Inspired By		2023	Blog post	https://www.keboola.com/blog/data-product-examples
G14	Data product toolkit			Professional guide	https://www.dataproductbusiness.com/toolkit
G15	What is a Data Product?		2023	Blog post	https://www.k2view.com/what-is-a-data-product/
G16	The data product lifecycle: Getting the most out of your data investments	Mohan S, Khan S.	2024	Blog post	https://www.starburst.io/blog/data-product-lifecycle/
G17	Data mesh in practice: Product thinking and development (Part III)	Gafoor A., Murdoch I., Prakash K.	2022	Blog post	https://www.thoughtworks.com/insights/articles/data-mesh-in-practice-product-thinking-and-development
G18	A streamlined developer experience in Data Mesh (Pt. two)	Jain M.	2023	Blog post	https://www.thoughtworks.com/insights/blog/data-strategy/dev-experience-data-mesh-product
G19	Unleashing the data mesh revolution: Empowering business with cutting-edge data products	O'Riordan D.	2023	Blog post	https://www.capgemini.com/insights/expert-perspectives/unleashing-the-data-mesh-revolution-empowering-business-with-cutting-edge-data-products/
G20	Data Fabric vs Data Mesh: Demystifying the Differences	Perlov Y.	2023	Blog post	https://www.k2view.com/blog/data-fabric-vs-data-mesh/

Table B.2: Activities for developing data products

ID	Activities
G1	Data product ownership, Build data products, Discover data
G2	Identify problem, define objective, Decide architecture and framework, Design data product, Launch data product, Iterate data product
G3	Identify business objectives, Data collection, Data cleaning and transformation, Data analysis and modeling, Prototyping, Production deployment, Continuous monitoring and improvement
G4	Identify current pain points and needs, Define the target audience, Build a cross-functional team, Develop a plan, Test and iterate
G5	Design, Develop, Deploy, Evolve
G6	Understand internal needs, Data collection and preparation, Develop and model DP, Create User-Friendly Interfaces, Implement governance and compliance, Beta testing and feedback, Launch and train
G7	Ideation, Design, Engineering, Release, Maintenance
G8	Design
G9	Consumer requirements, Curation, Provide DP through interfaces
G10	Creation of data products, Developing data products
G11	Configure and publish, Access data product, Governance
G12	Identifying data products, Creating data product SLOs, Implementing SLOs, Monitor as code, Monitoring and visualizing DP health
G13	Understand value that needs to be delivered, Build product, Test value delivery with end users
G14	Need, Draft, Validate, Implementation, Test & Deploy
G15	Define, Engineer, Test, Deploy
G16	Build, Maintain, Operate, Retire
G17	Identify value-oriented use case, Identify DPs that satisfy use-case, Define SLIs SLOs, Analyze bigger picture
G18	Bootstrap, Deploy, Retire
G19	Identify problem and reason, Build DP, Maintain DP
G20	Definition and design, Engineering, Quality assurance, Support and maintenance, Management

Appendix C

Interview Protocol

C.1 Interviewees selection and invite

I made use of multiple sources to get experts from different organizations with different backgrounds and experiences. I have sent the following invites to participants.

Direct email

Dear [Name],

My name is Mark Langedijk, and I am a master's student in Business Information Technology at the University of Twente (NL). I am contacting you because of your expertise in the area of data/data mesh.

As part of my thesis research, I am developing a methodology for building and managing data products within a data mesh architecture. I was wondering if you would help me by participating in an interview to gather feedback on this methodology.

The interview details are as follows:

- *Duration: 30-60 minutes*
- *Format: Video call, scheduled in the upcoming two weeks*
- *Topics: Your experience with data mesh and data products, and your feedback on the proposed methodology*

Your participation will help shape this research. In return for your time, I would be happy to share the final version of the methodology with you, which may offer new insights for data product development in your organization.

All responses will be anonymized in the research to ensure confidentiality. If you are willing to participate, please reply to this email, and I will follow up with potential dates and times for the interview.

Thank you for considering this request.

Kind regards,

Mark Langedijk
Master Student, Business Information Technology
University of Twente

Group message

Hey Data Mesh Learning Community!

I'm Mark Langedijk, a master's student at the University of Twente (NL), and I'm working on my master thesis focusing on data products in a data mesh architecture.

I am developing a methodology for building and managing data products within a data mesh architecture. I was wondering if you would help me by participating in an interview to gather feedback on this methodology.

The interview details are as follows:

- *Duration: 30-60 minutes*
- *Format: Video call, scheduled in the upcoming two weeks*
- *Topics: Your experience with data mesh and data products, and your feedback on the proposed methodology*

What's in it for you?

- *Be part of cutting-edge research on data mesh*
- *Get access to the results*
- *Contribute to advancing data product management practices*

Interested? Drop me a DM or reply here, and we'll set up a time to chat! (All responses will be anonymized in the research to ensure confidentiality.)

C.2 Script of Interview

The following script is used for the interview:

- Introduction (3 min)
  - Thanks for the time
  - Explain purpose of the interview
  - Assure confidentiality
- Background questions
  - Can you briefly describe your role and experience with data products and/or data mesh architecture?
- Overview of Data Product Lifecycle (methodology) (7 min)
  - Where you able to read the Data Product Lifecycle before hand?
  - Brief explanation of the Data Product Lifecycle
- Functional requirements evaluation (7 min)
  - How well does the methodology define the different phases of the DPDM-DMA?
  - To what extent does the methodology clarify goals and tasks for different stakeholders?
  - How useful are the presented tools for each phase of the lifecycle?
  - How useful are the inputs/prerequisites of each phase?
  - How useful are the outputs/outcomes of each phase?
- Structural qualities evaluation (7 min)
  - How well does the frame work prioritize user-centric focus and product thinking?
  - To what extent does the methodology support iterative development?
  - How effectively does the methodology address scalability for large-scale data?
  - How well does the methodology support modularity in data product development?
- Environmental qualities evaluation (7 min)
  - How usable do you find the methodology for different stakeholders (data product developers, owners, and management)?
  - How well does the methodology ensure the following qualities of data products: a) Discoverability b) Addressability c) Understandability d) Trustworthiness e) Accessibility f) Interoperability g) Value h) Security
- Overall evaluation and suggestions? (15 min)
  - What do you consider to be the main strengths of this methodology?
  - What areas of the methodology do you think need improvement?
  - Are there any important aspects of data product lifecycle management that you feel are missing from this methodology?

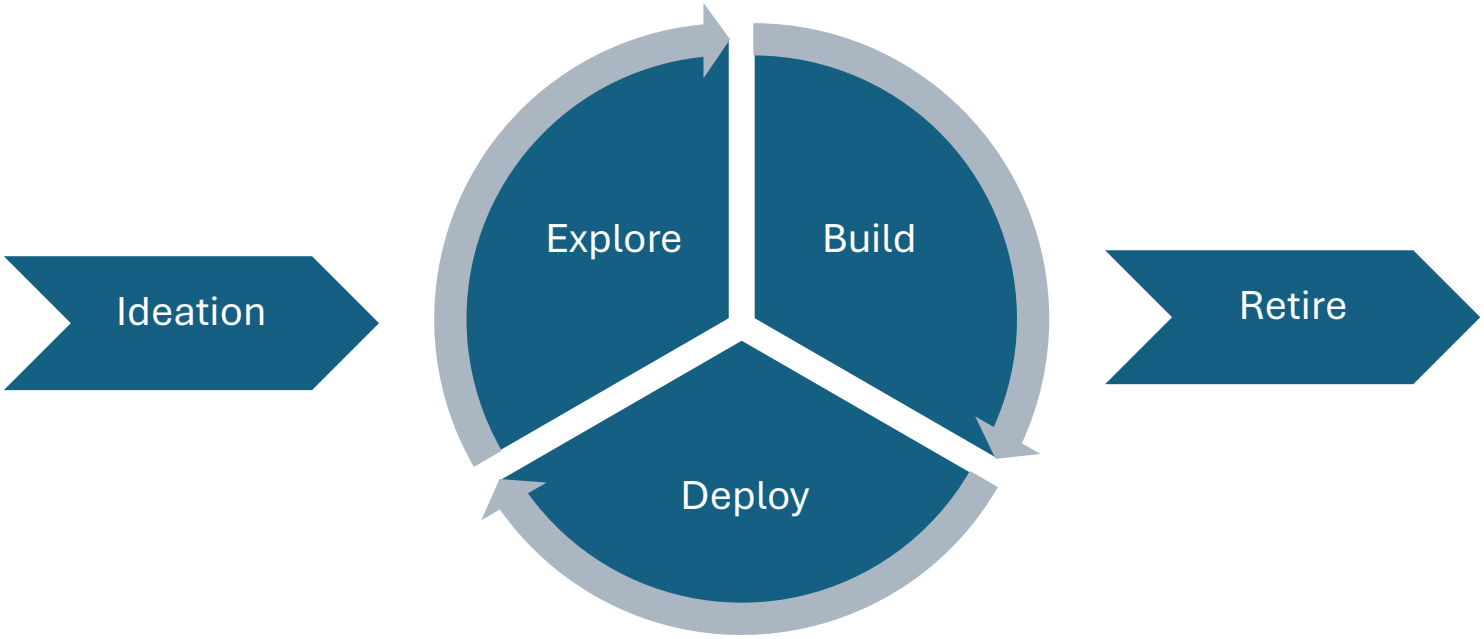
  - How likely would you be to recommend or use this methodology in your organization?

- Conclusion (5 min)
  - Any final thoughts or comments?
  - Would you be available for a second interview of a second version?
  - Explain next steps

DPDM-DMA

Data Product Development Methodology

within Data Mesh Architecture



Introduction

The Data Product Development Methodology within Data Mesh Architecture (DPDM-DMA) is a methodology that provides guidelines for the development and maintenance of data products within a data mesh architecture. It is comprised of three stages: a starting stage, an iterative cycle, and an ending stage.

We first start this document by explaining our definitions of data mesh and data products. Then we will give a bird's eye view of the DPDM-DMA. Thereafter, we will take a deep dive in each phase of the DPDM-DMA.

Data Mesh

Data mesh is a decentralized data architecture. The architecture consists of organizational and technological concepts to manage data in an organization. Data mesh emerged from the challenges of centralized data architectures, where centralized data teams are a bottleneck and data management become increasingly complex due to the complexity of the organization and the growing volume of data in an organization.

Data mesh follows four core principles:

1) Domain-Driven Ownership: introduces decentralized data management by splitting an organization into business domains and making these domains responsible for data management.
2) Data as a Product: introduces data products by applying "product thinking" to data, making the business domains responsible for providing high-quality data to other business domains.
3) Self-Serve Data Platform: A domain-agnostic platform built by a centralized platform team that provides autonomous business domains with the tools they need for the entire data product development.
4) Federated Computational Governance: manages decentralized data, ensures compliance with rules, and maximizes data quality through federated decision-making.

Want to learn more about Data Mesh and the motivation for this decentralized architecture?

- Dehghani, Z. (2020). Data Mesh Principles and Logical Architecture. https://martinfowler.com/articles/data-mesh-principles.html
- Dehghani, Z. (2022). Data mesh. "O'Reilly Media, Inc."

Data Products

Data mesh makes uses of the term 'data product' and defines data product as:

The node of the data mesh that encapsulates structural components (code, data, meta-data, and infrastructure) required for providing access to business domain's analytical data products.

However, data products do not solely exist in data mesh, data product is a term used for products that utilize data, such as dashboards, AI tools, and APIs. In addition, data product thinking is applied as strategy in different organizations. In this document, "data product" is used in the context of data mesh architecture.

A data product consists of code, data, meta-data, and infrastructure. High-quality data products are dependent on a good data strategy and data platform.

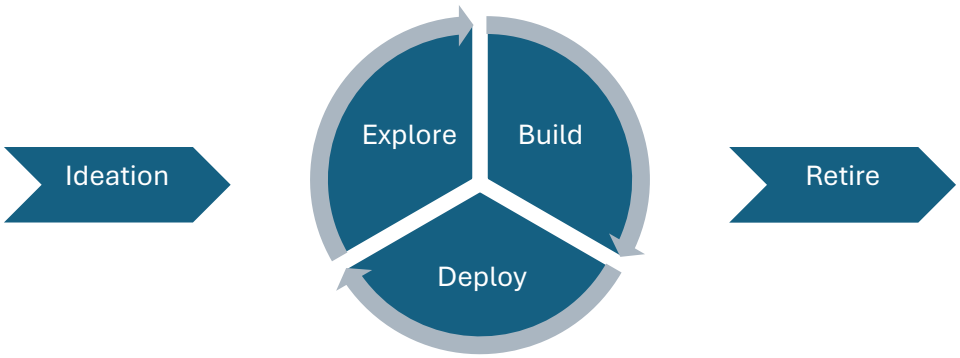
This DPDM-DMA focuses on building high-quality data products. At a minimum, a data product should meet the following attributes:

Discoverable	Data products should be easily discoverable by data consumers
Addressable	Data products should have a permanent and unique address
Understandable	Data products should be easily understood by data
Trustful	Data products should provide trustful and credible data
Natively accessible	Data products should be easily accessible by data consumers
Interoperable	Data products should be interoperable with other data products
Valuable	Data products should be relevant and add value for its consumer
Secure	Data products should be secure
+ Feedback driven	Data products should use feedback from data consumers to improve

Our methodology indicates how tasks relate to these quality attributes.

DPDM-DMA from a bird's eye view

The DPDM-DMA (Data Product Development Methodology for Data Mesh Architecture) consists of three stages, further divided into five phases:



Starting stage: Ideation

In the ideation phase we will explore the business case for our data product, investigate the need and how it can add value to the data consumer, which aligns with the product-thinking philosophy of data mesh. Before building a data product, we should consider the benefits and necessity for building a data product and if data products are the best solution.

After the ideation phase, you will have understood the need, benefits, and viability of the data product.

Iterative stage: Explore, Build, Deploy

After determining the need for a data product, we create the data product by exploring, building, and deploying. This process is iterative, allowing use to move between phases while we create prototypes and gather feedback from data consumers.

Data products are never finished, as a data provider you will be responsible for continuously delivering value with your data product.

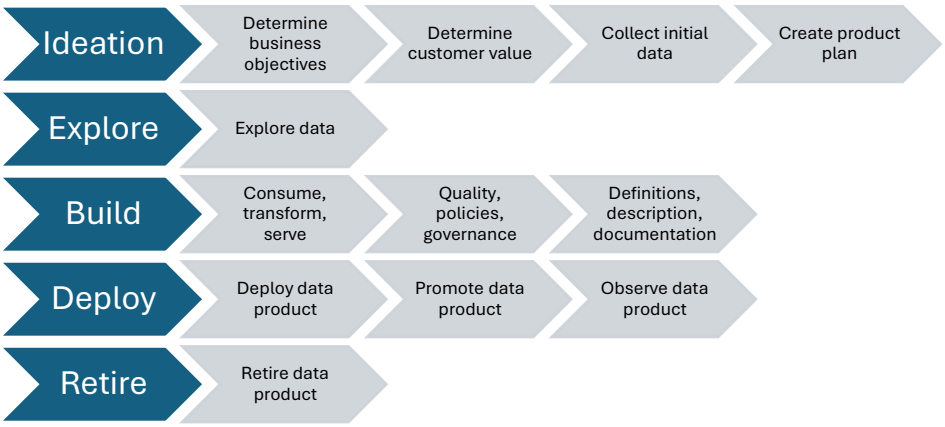
Ending stage: Retire

There may come a time when, after taking feedback from data consumers, you decide to end or split the data product. This is the final phase of the DPDM-DMA.

You should communicate this decision with the data consumers of the data product and develop a plan for the retirement of the data product.

Structure

The DPDM-DMA contains the following phases and tasks:



For each task we dive deeper into the following topics:

Each task is described in detail, and the stakeholders are identified to facilitate a clear understanding of roles involved in each task. Some tasks need information from other tasks to be executed most efficiently, this is described in the input of each task. The artifacts that result from the completion of a task are described in the output. Each task description contains tools or methods that can facilitate the execution of a task.

Roles and Skills

The methodology makes use of the following roles.

- Data product owner, person accountable for the data product, including its governance and quality.
- Data product developer, persons responsible for building and deploying data products.
- Data consumer, person that utilizes the data product.

We make use of Skills Framework for the Information Age 8 (SFIA 8) for indicating skills needed for each task. This can provide insight into the skills needed to successfully perform certain tasks. For organizations, this provides insight into possible future training or talent recruitment.

Prerequisites

We expect that you have the following things in place when using the DPDM-DMA:

- Your organisation has a data strategy.
- Your organisation is migrating or has migrated to a data mesh architecture.
- Your organisation has a self-serve data platform in place.
- Your organisation has a template/model for building data products. A template or model can help data producers and ensure interoperability between data products.

1. Ideation

Introduction:
In the phase, the data producer is tasked with identifying the business objectives, customer needs, and available data. These factors are crucial to determining how to proceed with the build of the data product. Having a well-defined plan and business understanding is crucial before starting a project. In this phase, the data producer should investigate if there is a need for a data product and evaluate whether developing a data product is an appropriate solution for the problem. Furthermore, the data producer should assess the availability and ownership of the required data.

In this phase we decide if it is profitable to build a data product.

Desired outcomes:
> Business objectives
> Customer value proposition
> Initial data collection
> Inflection point
> Data product plan

1.1 Determine business objectives	DAUTNIVS+
Task	Determine business objectives It is the responsibility of the data producer to clearly define the problem or opportunity that a data product is designed to address. Each data product needs to add value (on its own), so the data producer should analyze the current landscape and the organization's needs. A well-structured business case should describe how a data product can add value, taking into account both costs and benefits.
Stakeholders	Domain experts, Data Product Manager Skills: • Benefits management • Business process improvement • Business situation analysis • Demand management • Innovation • Strategic planning
Input	• Organization wide vision and goals • Data strategy

Output	• Business objectives
	  ○ Describe the problem or opportunity the organizations want to solve
	• Background information (Ubiquitous language)
	  ○ Explain the information known about the situation, including ubiquitous language (Domain-Driven Design) to make sure stakeholders have a common language and understanding.

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Business case	A value proposition for data products	- Ward, J., & Daniel, E. (2012). *Benefits management: how to increase the business value of your IT projects.* John Wiley & Sons.	M
Cost-benefit analysis	An analysis to determine financial and non-financial benefits and costs	- Ward, J., & Daniel, E. (2012). *Benefits management: how to*	S

			increase the business value of your IT projects. John Wiley & Sons.	
Business model canvas	One-page template for creating a value proposition using nine building blocks that we need to think about			C
Stakeholder interviews	Interviews with (domain) experts to better understand background of opportunity or problem			C
Wardley mapping	A map for a business strategy to better assess a situation and improve decision making	-	Wardleymaps. Medium. https://medium.com/wardleymaps	C
BCG matrix	Framework to rank products based on market share (data product usage) and growth potential, to better understand and prioritize data products.	-	Stern, C. W., & Deimler, M. S. (Eds.). (2012). *The Boston consulting group on strategy: Classic concepts and new perspectives*. John Wiley & Sons.	C

1.2 Determine customer value

DAUTNIVS+

Task	<u>Determine customer value</u> The data producer is tasked with creating a data product that adds value for its customers (data consumers). The data producer is responsible for creating value with their data product. They need to empathize with the customer and understand their needs
Stakeholders	Data product owner, data product developer, data consumer Skills: • Innovation • User research • Strategic planning • Requirements definition • Risk management
Input	• Background information

Output	• Requirements
	  ○ Have a clear understanding of the requirements for the data product.
	• Assumptions & Constraints
	• Risks
	  ○ Identified risks that should be avoided and a plan on how to tackle these risks.
	• User story
	  ○ Customer point of view and understanding of customer's needs.

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Customer interviews	Get a better understanding of the needs of the data consumers by discussing needs.		M
Requirements documentation	Document the requirements for the data products. These requirements can be related to schedule, quality, governance, etc.	- Robertson, S., & Robertson, J. (2013). *Mastering the requirements process: Getting requirements right*. Pearson Education.	S
Risk review	Analyse and identify possible risks for the data product and during the creation of the data product.	- Chapman, C., & Ward, S. (2003). *Project risk management processes,*	S

			techniques and insights. John Wiley & Sons Ltd,.	
User story	Description of the objectives from a data consumers point of view	-	Patton, J., & Economy, P. (2014). *User story mapping: discover the whole story, build the right product.* " O'Reilly Media, Inc.".	S
Jobs to Be Done	Framework for capturing and defining customer's needs	-	Ulwick, A. W. (2016). *Jobs to be done: Theory to Practice.*	C

1.3 Collect initial data

DAUTNIVS+

Task	Collect initial data The data producer is responsible for identifying data that could meet the customer's needs. This is done by collecting data from source systems or from other data products. The data producer should be the owner of the data that is needed for the data product.
Stakeholders	Data product developer Skills: • Data engineering • Data management • Data science • Knowledge management
Input	• Business objectives • Requirements • User story
Output	• Initial data collection report of customer's needs.

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Data exploration	Identify and explore the available data		M

1.4 Create product plan

DAUTNIVS+

Task	Create product plan The data producer needs to create a plan for developing the data product. The data producer combines the business goals and requirements into a plan, in which they also assign responsibilities. The plan summarizes the goal, stakeholders, schedule, value and benefits, and cost of the data product. The data producer should clearly outline the data product and decide if building a data product is profitable and a suitable solution for our case. When in doubt or inexperienced, they should ask the organization's expert on data mesh and data products for guidance.
Stakeholders	Data product owner, data product developer, data product expert Skills: • Data management • Feasibility assessment • Investment appraisal • Project management

	• Strategic planning
Input	• Business objectives • Requirements • User story • Risks • Initial data collection report

Output	• Inflection point
	• Product plan
	  ○ Document outlines the DPDM-DMA, including details on responsibilities, schedule, and budget.
	• Data product canvas
	  ○ A one-page summary for data product development

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Schedule	A schedule is a calendar of activities, milestones, dependencies, resources, and responsibilities.	- Gantt chart	M
Budget	Plan for number of resources that can be used for building and maintaining data product.		S
Responsibility Assignment Matrix (RAM)	A matrix that maps stakeholder responsibilities.	- RACI chart	S
DPDM-DMA	Guidelines and principles for building data products		C

2. Explore

Introduction:
In the explore phase, the data producer identifies potential sources for the data product. Data producers gain a better understanding of the available data by performing an in-depth exploration of the data. Data is explored to understand how it should be transformed, how it should be stored, its quality, and what level of governance should be applied.

Exploring data helps us with the decision of which data to share and under which conditions.

Desired outcomes:
> Data reports (collection, description, exploration, quality)

2.1 Explore data				DAUTNIVS+
Task	Explore data The data producer examines the data to better understand what data transformation, data integration, data cleansing, and governance policies to be applied.			
Stakeholders	Data product developer Skills: • Data engineering • Data science • Governance • Quality management			
Input	• Initial data collection report • Global governance policies			

Output	• Data reports
	  ○ Data description report
	  ○ Data exploration report
	  ○ Data quality report
	• Governance plan			

Tools:				
What	*Explanation*	*Sources / Tools*		*MoSCoW*
Data exploration	Analyse and evaluate data			M
Data visualisation	Creating a graphical representation of data to get a better understanding of the data	-	Wilke, C. O. (2019). *Fundamentals of data visualization: a primer on making informative and compelling figures.* O'Reilly Media.	M
Data quality analysis	Make use of a data quality framework to define product's quality on multiple dimensions	-	Choosing a framework: *Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. Ieee Access, 7, 24634-24648.*	M

3. Build

Introduction:
In the build phase, data producers utilize the collected information of the explored data and the data consumer to collect, transform, store, and serve the data. Data producers should ensure quality and compliance by establishing governance rules. Ultimately, data producers are responsible for creating understandable data products. This is achieved by adding documentation, a data sample, and computational notebooks to help data consumers understand and gain insight into the data product.

After this phase, data producers combine all the information to deploy the data product

Desired outcomes:
> Code (extraction, transformation, policies)
> Governance plan
> Access plan
> Documentation and examples

3.1 Consume, transform, serve	DAUTNIVS+
Task	Consume, transform, serve The data producer needs to consume the data, transform it into an appropriate form, store it, and create interfaces for data consumers. The self-serve data platform should facilitate the infrastructure for these actions. The data producer should consider the input of the data, the best way to store the data, and the form and interface that are best for consumers to access the data. The data producer should carefully consider the data model of the data product.
Stakeholders	Data product developer Skills: • Data engineering • Data modelling and design • Database design
Input	• Data reports • User story • Product plan

Output	• Definition of input interfaces
	  ○ The interfaces of the data product need to be defined; this is usually done in the form of code.
	• Definition of output interfaces
	  ○ The interfaces of the data product need to be defined; this is usually done in the form of code.
	• Data transformations
	  ○ Code for transforming data in the preferred format.
	• Data storage
	  ○ Storage of the data

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Self-serve data platform	A self-serve data platform should provide tools to data producers for developing data products	- *Dehghani, Z. (2022). Data mesh. "O'Reilly Media, Inc."*	M

	Data integration system	The data platform supports integration of incoming and outcoming data using UI tools.	M
	Data processing	The data platform supports the transformation from the data.	M
	Polyglot storage	The data platform helps with storing polyglot data.	M

3.2 Quality, policies, governance

DAUTNIVS+

Task	Qualities, policies, governance Data products should conform to both local and global policies to ensure interoperability, quality, and compliance. Data producers should consider the data we consume and determine whether governance should be applied. Additionally, they should consider the users of the data and whether they should be able to consume the data.
Stakeholders	Data product owner, data product developer Skills: • Availability management • Governance • Information assurance • Personal data protection • Service level management • Quality management
Input	• Data quality report • Governance plan

Output	• Policy as code
	• SLO
	  ○ Targeted levels of service
	• SLIs
	  ○ Metrics used to measure quality

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Self-serve data platform	A self-serve data platform should provide tools to data producers for developing data products	- *Dehghani, Z. (2022). Data mesh. "O'Reilly Media, Inc."*	M
Security and privacy	The data platform helps set security and privacy requirements for the data, these requirements should be based on the consumed data and to whom we provide data		M
Monitoring	The data platform helps enforce security and privacy rules, the data producers should define how they monitor security, privacy, and quality		M

3.3 Definitions, description, documentation

DAUTNIVS+

Task	Definitions, description, documentation Data providers and consumers need to understand the value and use of the data product. Data producers must present the data in the best possible way. Data consumers are customers, so we should make it easy for them to understand the data. This can be achieved by describing the data (fields), providing data samples, and providing computational notebooks with usage examples.
Stakeholders	Data product developer Skills: • Data science • Data visualisation • Information content authoring • Knowledge management
Input	• Data description
Output	• Metadata • Data product description • Examples (computational notebook, data)

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Self-serve data platform	A self-serve data platform should provide tools to data producers for developing data products	- *Dehghani, Z. (2022). Data mesh. "O'Reilly Media, Inc."*	M
Data catalog	The data platform provides a catalog service to publish data, metadata, and information		M

4. Deploy

Introduction:
In the deploy phase, the information built and collected in the previous phases is leveraged on the self-service data platform to create a version of the data product and deploy a data contract.

After deploying a data product, the data producer should ensure that potential data consumers are aware of the existence and the purpose of the data product. They should actively promote the data product. Additionally, the data producer is responsible for delivering valuable data products, which conveys that they should monitor the data product and address feedback.

Desired outcomes:
> Detailed contract
> Data product

4.1 Deploy data product			***DAUTNIVS+***
Task	Deploy data product Use the information from the build phase to create a data contract and deploy the data product by utilizing the self-serve data platform.		
Stakeholders	Data product developer Skills: • Acceptance testing • Product management		
Input	• Code • Data • Metadata • Interfaces • Infrastructure		
Output	• Data contract • Data product		
Tools:			
What	*Explanation*	*Sources / Tools*	*MoSCoW*
Self-serve data platform	A self-serve data platform should provide tools to data producers for developing data products. The data platform should guide the developer on the deployment of the data product	- *Dehghani, Z. (2022). Data mesh. "O'Reilly Media, Inc."*	M

4.2 Promote data product

DAUTNIVS+

Task	Promote data product It is the responsibility of data producers to create valuable data products. After creating a data product, data producers should ensure that it is easy to use. A data product is appealing when promises are kept and the product is maintained. Data producers should present and explain their data products on platforms and in places that are available to the entire organization.
Stakeholders	Data product owner, data consumer Skills: • Marketing • Stakeholder relationship management
Input	• Data product documentation • Data product examples
Output	• Feedback • New users

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Communication channels	You should make use of the available communication channels in the organization for promoting the data product.		M

4.3 Observe data product

DAUTNIVS+

Task	Observe data product Data producers are responsible for continually improving and adapting their data products based on the needs of data consumers. Data producers should be open to feedback and continually improve their data products. Data products become of a higher quality when data producers communicate with the data consumers and incorporate their feedback.
Stakeholders	Data product owner Skills: • Product management • Service level management • Stakeholder relationship management
Input	• Feedback • SLI
Output	• Improvement plan

Tools:

What	*Explanation*	*Sources / Tools*	*MoSCoW*
Communication channels	Data consumers should be able to easily ask question to the data provider		M
Service Level Indicators (SLI)	SLIs can be used to understand the quality of your data product, which can be a reason to make adjustments to your data product		S

5. Retire

Introduction:
In the retire phase, data producers create a plan for retiring the data product. There can be multiple reasons to retire a data product. Primarily, the data product may no longer be of value to customers, and making adjustments does not add value to the data product. Secondly, a data product may no longer be utilized by data consumers.

This is the final phase of the DPDM-DMA. You should communicate this decision with the data consumers of the data product and develop a plan for the retirement of the data product. This plan contains the specific details for retiring the current data product, however the organization should have a general plan for retiring data products and an implementation for this in the self-serve data platform.

Desired outcomes:
> Retirement plan

5.1 Retire data product			DAUTNIVS+
Task	Retire data product It is the responsibility of the data product owner to decide when a data product should be retired. This may be due to the migration to one or multiple new data products, or because the data product is no longer required. The self-serve data platform should be able to facilitate the retirement of data products.		
Stakeholders	Data product owner Skills: • Product management		
Input	• SLO • SLI • Product plan		
Output	• Plan for retiring data product • Communication to stakeholders		
Tools:			
What	*Explanation*	*Sources / Tools*	*MoSCoW*