UNIVERSITY OF TURKU
Turku School of Economics

Data Governance on Data Platforms

Designing Playbook for Data Platforms

Master's thesis
in Information Systems Science

Author:
Miikka Luosmaa

Supervisor:
Ph.D. Tiina Nokkala

20.2.2022
Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Master's thesis

Subject: Information Systems Science
Author: Miikka Luosmaa
Title: Data governance on data platforms – designing playbook for data platforms
Supervisor(s): Ph. D. Tiina Nokkala
Number of pages: 50 pages + appendices seven pages
Date: 20.2.2022

Abstract

The amount and value of data are increasing (Nokkala 2020). Data can be seen as one of the key enterprise assets (Weill and Ross 2005). Organizations have started to build data platforms to have the data from all sources available for use in cross-organization and cross-industry business ecosystems ((Feibus 2021; Panian 2010). The constraints with data are no longer the amount of data or the technology. It is the structures and processes. Data needs governing like any other enterprise asset. Data professionals interviewed for this research had rarely experienced successful data governance. The academic literature on data governance frameworks is fragmented and lacks holistic guidance on how data governance on data platforms should be designed, implemented, and monitored.

This master's thesis contributes to the lack of holistic guidance on designing data governance on data platforms through design science research. The research supports the theoretical framework (Nokkala 2020) and advances it with data analysis of 13 semi-structured interviews with data professionals. This research produces a canvas tool, Playbook for Data Platforms, to help people working with data design successful data governance in the data platform context. The canvas tool aims to increase data value and minimize the data-related cost and risk in the platform context, a definition of data governance presented by Abraham et al. (2019). The research also considers how cloud data affects the data governance framework. As a result, 12 areas and 31 guiding questions are proposed to be included in the data governance framework on data platforms.

Key words: data governance, data platforms, data governance on data platforms

TABLE OF CONTENTS

1 Introduction 7

1.1 Research area 7

1.2 Research gap 7

1.3 Research questions 8

1.4 Research scope and justification 9

2 Theoretical background 10

2.1 Data governance 10

2.2 Data governance on data platforms 12

2.3 Data governance frameworks 14

3 Research design and methodology 18

3.1 Research strategy 18

3.2 Data collection 20

3.3 Data analysis 21

3.4 Ethics 22

4 Data governance on data platforms 24

4.1 The purpose of data governance 24

4.2 Areas of data governance framework 25

4.2.1 Platform context/strategy 26

4.2.2 Value of data 28

4.2.3 Shared data ontology 29

4.2.4 Data provenance 31

4.2.5 Data ownership & stewardship 32

4.2.6 Data access and security 33

4.2.7 Data risks 34

4.2.8 Regulatory environment 35

4.2.9 Data structure 36

4.2.10 Data quality 37

4.2.11 Platform (data) business and governance model 38

4.3 Cloud data governance 38

5 Playbook for Data Platforms 40

5.1 Demonstration **40**

5.2 Evaluation **42**

5.2.1 Completeness 43

5.2.2 Accuracy 43

5.2.3 Usability 45

6 Discussion and conclusions **46**

6.1 Discussion and conclusions **46**

6.2 Future research **47**

7 References **48**

8 Appendix **51**

8.1 Appendix 1. Research questions 1 **51**

8.2 Appendix 2. Research questions (evaluation) **52**

8.3 Appendix 3. Guiding questions **53**

LIST OF FIGURES

Figure 1 DIKW hierarchy (adopter from Ackoff 1989) 10

Figure 2 The relationships between different governance concepts (adopted from Otto 2011; Nokkala 2020) 11

Figure 3 Generic architecture of a data platform (adopted from Microsoft 2022) 13

Figure 4 Design science research model (adopted from Hevner et al. 2004) 19

Figure 5 Governance of Data, Data Governance, Data management (adopted from Nokkala 2020) 24

Figure 6 Data governance of Platform Data framework (adopted from Nokkala 2020) 26

Figure 7 Data types (adopted from Nokkala 2020) 30

Figure 8 Playbook for Data Platforms 41

LIST OF TABLES

Table 1 Areas of data governance framework deducted from existing literature 16

Table 2 Interviewee profiles and interviewee durations rounded to the closest quarter of an hour 20

Table 3 Example of data structure and coding scheme for data analysis 21

Table 4 Changes to the guiding questions during the evaluation 45

1 Introduction

1.1 Research area

Calling data as the new oil would be a mistake (c.f Humby 2006). Unlike oil, the value of data grows together with the usage, and storing it in silos does not make sense. Data reaches its full potential when used in cross-organization and cross-industry business ecosystems. Organizations have enormous amounts of data from ERP, CRM, SCM, and other enterprise systems (Niemi 2015). Also, the data can be brought in from sources controlled by partner organizations or unknown entities (Janssen et al. 2020). Many organizations have prioritized having a data platform that combines the data from all data sources and serves data from the platform intra- and inter-organizationally (Feibus 2021; Panian 2010). The role of these platforms is growing (Nokkala 2020). All of the above can be concluded that what is possible to do with data is no longer dependent on the technology or the amount of data. (Luoma-Aho 2021)

Data can be seen as a part of Information and IT assets that, together with financial resources, human resources, intellectual property, physical structure, and organizational relationships, form the key enterprise assets (Weill and Ross 2005). Involving business management with data enables data to serve the strategic interests of an organization (Feibus 2021; Otto 2015). Just like any other key enterprise asset, data needs governing. Underestimating the complexities of data will lead to missing opportunities (Panetta 2021). A good data governance framework improves trust in an organization's data and helps organizations maintain a clear mission, scope, and focus aligned with the organization's strategy and values (Al-Ruithe et al. 2019). Still, almost half of the organizations do not implement a data governance program (Alhassan et al. 2019). Leading Finnish data professionals argue that the constraints for data usage can be found in the structures and governance of the organizations (Luoma-Aho 2021). Many organizations recognize the value of data and declare themselves as data-driven - without acting so (Nokkala 2020).

1.2 Research gap

Academic data governance literature is fragmented with two main streams: one focusing on data quality and setting up roles and responsibilities to deal with poor data quality. The second stream values the role of data governance in maximizing the value of data in

organizations. (Benfeldt et al. 2020) There is no standard definition for data governance. Abraham et al. (2019) define data governance as "-- 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."

When the number of data sources, end-users, and variety of use cases and tools grows, data platforms are seen as a solution to ingest, manipulate, combine and share data. The growing number of data platforms and the lack of data governance programs are the primary motivation for this research. A proven data governance framework for platform data seems not to be established. This research gap has been pointed out by Fu *et al.* (2011), Niemi (2015), Lee et al. (2017), and Nokkala (2020). The literature review conducted for this master's thesis revealed the research gap still existing. Seven data governance frameworks were analyzed, and only two deal with data governance on data platforms (Lee et al. 2018; Nokkala 2020).

1.3 Research questions

The research gap in data governance on data platforms and the need for holistic data governance frameworks have been highlighted. For this reason or another, implementing data governance in organizations is proven to be complicated. When interviewing data professionals about data governance, many of them said that they have never experienced successful data governance or there is no such thing. Data governance was seen as a dusty and challenging term with no right answers. The research goal is to understand how to design, implement, and monitor data governance on data platforms. To contribute to the academic literature on this subject and to contribute to solving a real-life problem, a design science research was conducted with the following research questions:

1. What questions should be considered when designing, developing, and monitoring data governance on data platforms?

To define the solution objective and to be able to design and develop a solution to the problem identified, three sub-questions support the main research question:

   a. What is the purpose of data governance?

b. What are the areas of data governance framework in the data platform context?

c. How does cloud data affect data governance?

1.4 Research scope and justification

Wilson (2002) presents three essential questions when considering potential research: Is it interesting? Is it new? Is it true? The first question is the most important one and should be answered first. My interest in data and its value rose when writing my bachelor's thesis about the possibilities of data in managing the costs and environmental effects of last-mile deliveries. Soon after, I started working as a Business Intelligence developer and project manager in data platform projects in a multinational consulting company. Seeing it close to the powerful asset the data can be, fueled my interest. But at the same time, I saw the difficulties in the daily work of data professionals due to unclear or missing policies, structures, and processes. When studying the subject through academic literature and interviews with data professionals, I learned that I am not alone with this notion.

Data governance as a topic is nothing new. However, the lack of a holistic framework in data governance literature and the lack of holistic guidance for the practitioners have been shown earlier. Especially the need for data governance research on data platforms has been highlighted. The interviews were conducted with Finnish data professionals who mostly are not working as data governance specialists but instead working with the data daily in different roles. The scope is to evaluate the existing framework (Nokkala 2020) for data governance on data platforms and design a canvas tool to help implement the framework. Later, this canvas tool will be referred to as "Playbook for Data Platforms" or "playbook". The playbook will be the artifact of this research.

2 Theoretical background

A brief literature review was conducted on data platforms, data governance, and data governance on data platforms. “Data governance”, “information governance”, “data platform”, “data governance on data platforms”, “data governance framework” were used to review the existing literature in Volter and Google Scholar databases. In addition, the chain-referral sampling method was used to track more relevant references. This literature review aims to provide background and define objectives for the desired solution to be produced in this master’s thesis.

2.1 Data governance

To define what data governance means, we must first define what is meant by data. Russel Ackoff (1989) has been credited for introducing the Data-Information-Knowledge-Wisdom hierarchy (DIKW), one of the best-known and fundamental models. The hierarchy is illustrated in figure 1.

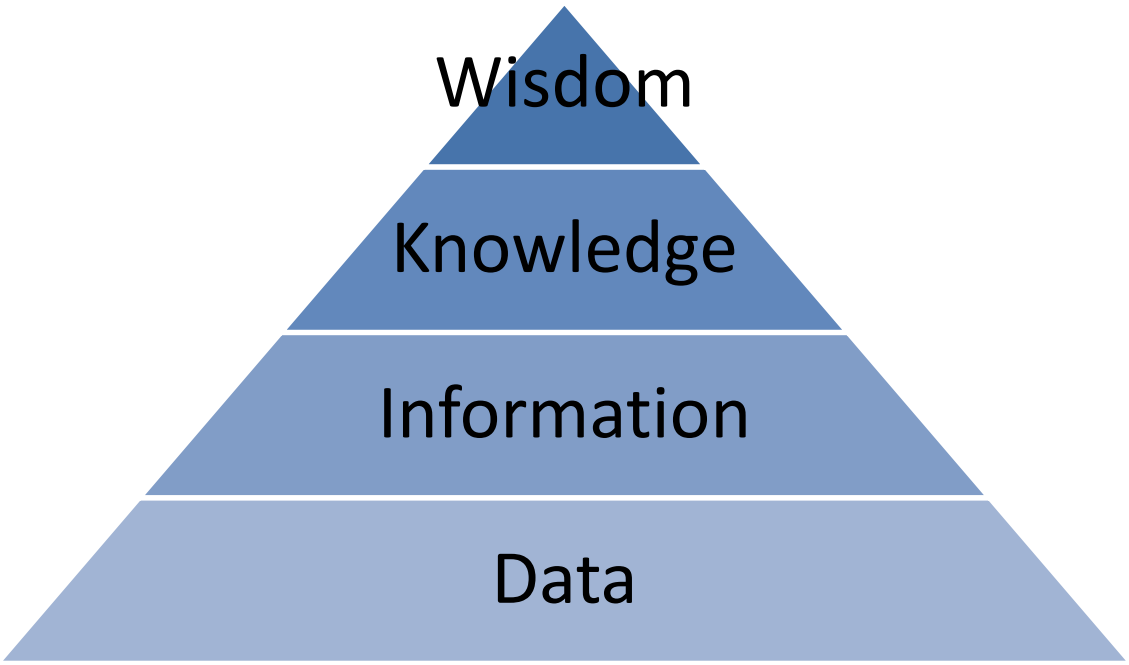

Figure 1 DIKW hierarchy (adopter from Ackoff 1989)

The hierarchy presents the process from data to wisdom. First, data represents facts as symbols with no meaning in itself. Second, information is data in a context containing descriptions and answering questions like who, what, when, and how many. Third, information transforms into knowledge when it is integrated with know-how. Last,

wisdom is the result of adding value to knowledge and increasing effectiveness. (Ackoff 1989; Rowley 2007)

After defining data, we must define the governance of data. The relationship between different terms is presented in figure 2. First, it is necessary to distinguish between governance and management. Governance refers to how the organization ensures that strategies are set, monitored, and achieved, while management refers to planning, doing, and managing actions accordingly. Therefore, data management is one area of data governance. (Al-Ruithe et al. 2019) Second, IT governance and data governance must be differentiated. While data governance relates to information assets that are documented and have value or potential value, IT governance relates to technologies like computers, communication, and databases. IT governance is quite a mature area compared to data governance (Al-Ruithe et al. 2019). Data governance literature is fragmented, and case studies have primarily selected IT and data management executives as interview partners. (Abraham et al. 2019)

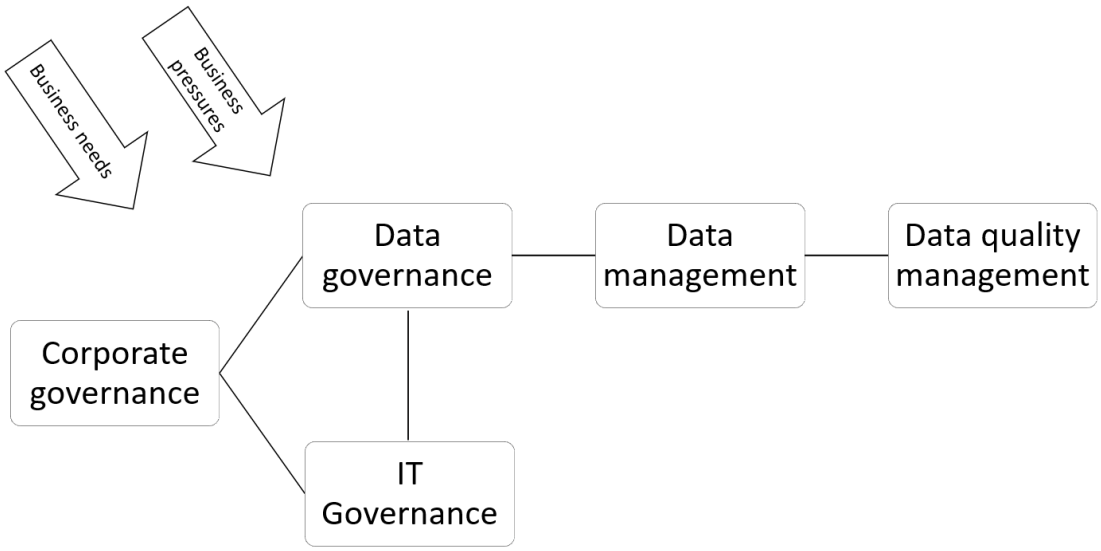

Figure 2 The relationships between different governance concepts (adopted from Otto 2011; Nokkala 2020)

A generally accepted definition of data governance is lacking. Most of the definitions acknowledge the role of data governance in managing data as a strategic asset by specifying decision rights and accountabilities and formalizing and monitoring data policies, standards, and procedures (Abraham et al. 2019). This research adopts the definition presented by (Abraham et al. 2019): “The purpose of data governance is to increase the value of data and minimize the data-related cost and risk.”

Data governance is a continuous improvement process due to changing internal and external needs (Janssen et al. 2020). Actions within data governance can be categorized into 1) define, 2) implement, 3) monitor (Alhassan et al. 2019). Other namings for the same actions are often called also 1) evaluate 2) direct 3) monitor (Nokkala 2020). Due to the growth of data volume and complexity, data management solutions have become expensive and heavy to maintain, emphasizing the role of data governance (Al-Ruithe & Benkhelifa 2020). While data governance used to be a nice to have feature in the past, today, the importance of data governance has grown significantly (Abraham et al. 2019).

The less time spent on data-related issues, the more time can be spent running the business and improving the processes. Lack of data governance might lead to inconsistent data, poor performance, little accountability, and unhappy IT customers. (Al-Ruithe & Benkhelifa 2020) Successful data governance positively affects data quality, data utilization, operations of an organization and helps find new ways to compete in the market. (Abraham et al. 2019) It also improves the confidence in the usage of the organization's data. However, while data governance strategy is integral to creating value from data, the phrase 'culture eats strategy for breakfast' might hold. Too much governance can constrain users so that the value of data will not be maximized. Overly complex and bureaucratic data governance can limit innovation based on innovation. In addition, it can motivate users to create shortcuts within the policies and take unnecessary risks with data. (Abraham et al. 2019; Otto 2011) Communication, education, and skill development are required to motivate people about the boring but essential data governance issues. (Chakravorty 2020)

2.2 Data governance on data platforms

The rise of platform companies like Facebook and Uber has drawn attention in the academic and practitioner's community (Otto & Jarke 2019). The role of digital platforms is growing in the world, where networks and ecosystems are playing an increasingly important role (Nokkala 2020). Platforms, in general, can be single-sided, two-sided, and multi-sided, depending on the number of actors inside the platform. At the same time, the need for combining, manipulating, storing, and presenting information has risen when organizations have gotten more sophisticated with data. (Abraham et al. 2019) Data platforms have emerged to answer this need by exchanging and sharing data (Otto & Jarke 2019). Today, public and private organizations work together with their clients,

vendors, consulting companies, and cloud service providers to build data ecosystems (Janssen et al. 2020). The amount of platforms ecosystems and the value of data in platforms has increased in recent years. (Lee et al. 2017)

What is a data platform? In a business sense, it is a tool that enables data exchange within one or many organizations serving a product, an organization, or the whole industry (Lee et al. 2017). Academic literature seems not to provide a technical definition for data platforms. However, defining it is crucial for this research to help narrow the scope of data governance and to help navigate between different data-related terms. In a technical sense, a data platform is an integrated solution that combines the tools and functionalities as one manageable product to meet the organization's data needs instead of multiple point-to-point sets of tools (MongoDB 2022). This definition is aligned with the definitions by data professionals interviewed. Data platforms are sometimes referred to with different names or adding terms, like Cloud Data Platform, Customer Data Platform, Big Data Platform, or Enterprise Data Platform (MongoDB 2022). These names refer to different types of data the data platform processes or technology they use, but they all are in the scope of the data platform in this research. The high-level architecture of a data platform is presented in figure 3.

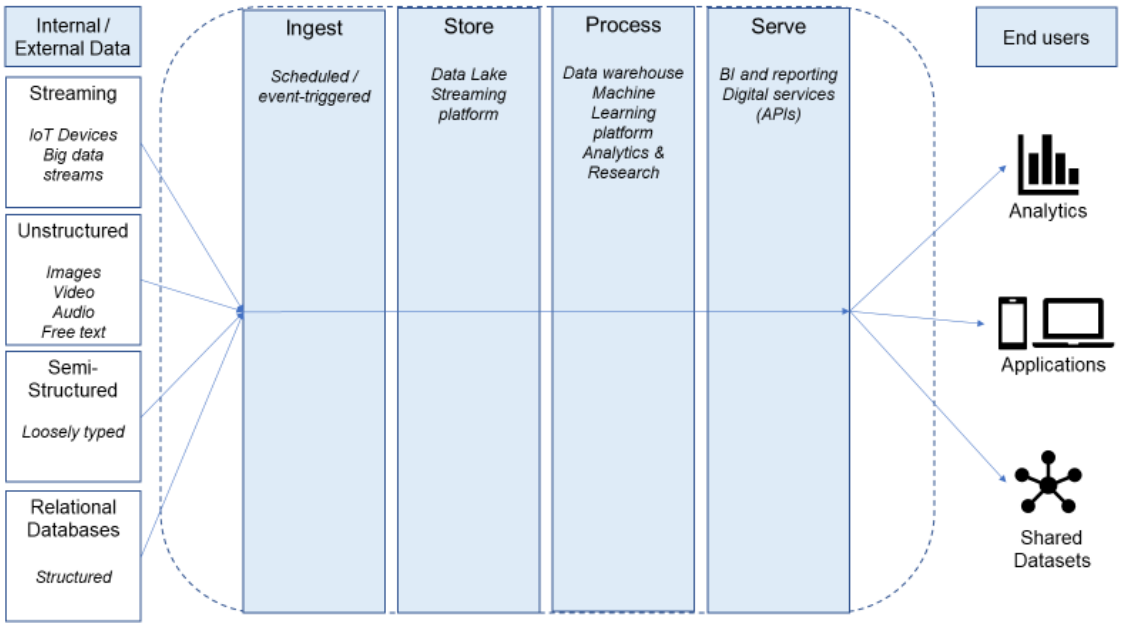

Figure 3 Generic architecture of a data platform (adopted from Microsoft 2022)

Platforms and especially cloud platforms enable data sharing intra-organizationally and inter-organizationally. A proven data governance framework for platform data has not

been established, and every platform has its specific requirements to adjust to (Lee et al. 2017). Nokkala (2020) argues that ownership and data access, usage and value, and data stewardship are domains that differentiate single-organizational data governance from data governance on platforms.

The amount of data used and stored has grown massively, and the growth seems never-ending (Nokkala 2020). Cloud computing is seen as an answer to coping with the growing data volume and complexity. A cloud data platform is an attractive option due to its instant scalability, cost efficiency, agility, data processing capability, security, and real-time data (Velayutham 2021). At the same time, data governance and its challenges are the main reasons not to adopt cloud computing. The risk of loss of data control, security and privacy, data quality and assurance, data stewardship and data governance are all risks associated with cloud computing. (Al-Ruithe et al. 2019).

Cloud data governance is closely related to cloud governance. Cloud governance is a new term that has not been studied much yet. Cloud governance refers to managing cloud services in a compliant way, and data governance is an essential part of it. (Al-Ruithe & Benkhelifa 2020) Several challenges can be faced when implementing cloud data governance. Securing and classifying data, especially sensitive data, in the cloud is a delicate topic for many. Privacy and control of the data are lost to some extent due to cloud providers when consumers are concerned about who can access their data. (Al-Ruithe & Benkhelifa 2020) They argue that loss of governance is one of the top risks in cloud computing and suggest that organizations develop a cloud data governance program before shifting to the cloud. The area of cloud data governance will require more research in the future as more and more organizations are adopting cloud computing (Al-Ruithe et al. 2019). This research contributes to this demand by considering the cloud aspect in each area of the data governance framework.

2.3 Data governance frameworks

While a data strategy sets a high-level plan based on strategic business objectives, a good data governance framework can help maintain a clear mission, scope, and focus (Al-Ruithe & Benkhelifa 2020). Focusing on single aspects of data governance leads to isolated solutions (Niemi 2015). A good data governance framework supports compliance and legal efforts and creates value from data Fu et al. (2011). Even though data governance has been researched since the 1980s, organizations still lack holistic guidance

to plan and develop their data governance (Niemi, 2015). This research gap has been pointed by Fu et al. (2011), Nokkala (2020), Niemi (2015)

For this literature review, seven data governance frameworks were analyzed. Most of these seven frameworks have data quality, data stewardship, roles and ownership, metadata, data access, and some kind of data standards, policies, or principles aspect included in their framework. Also, the value of data is included. These areas and references are summarized in table 1, and a short description of each framework is offered below.

Khatri & Brown (2010) introduce a data governance framework consisting of five interrelated decision domains: data principles, data quality, metadata, data access, data lifecycle. The authors propose roles and decision-makers for each domain and questions to assess the data governance. The model was tested in large insurance companies. While the framework (Khatri & Brown 2010) seems to have provided a base for many of the frameworks offered since its publishing, the framework presented by Panian (2010) has gotten less attention. Panian (2010) presents a data governance framework based on practical experiences. It consists of six key data attributes and four components of data governance. The key data attributes are accessibility, availability, quality, consistency, auditability, and security. The four components are standards, policies and processes, organization, and technology.

Janssen et al. (2020) focus on data governance from an Artificial Intelligence point of view and provide a data governance framework for trustworthy Big Data Algorithmic Systems. The framework has 13 design principles proposed for single organizations and multiple networked organizations. In this literature review, those 13 principles were clustered into six themes: data security, data principles, the value of data, data stewardship, data lifecycle, data access, data quality, and bias.

Abraham et al. (2019) develop the conceptual framework for data governance from six dimensions: governance mechanisms, organizational scope, data scope, domain scope, antecedents, consequences. The purpose of the conceptual framework is to assist practitioners by introducing an overview of antecedents scoping the parameters and governance mechanism in a structured format. The framework includes both intra- and inter-organizational scopes. The concepts of antecedents and consequences emphasize the internal and external factors on data governance and the measurability of the data

governance program. The concepts of antecedents and consequences align with Otto (2011), who divides data governance into two dimensions: goals and structure. Goals include measurable formal business and IS/IT-related goals and functional goals such as data quality, data stewardship, data standards, metadata, data lifecycle, and data architecture management. Data governance structure identifies the locus of control, organizational form, and roles and committees.

The only data governance frameworks that focus on platform context were Lee et al. (2018) and Nokkala (2020). Lee et al. (2018) propose four principles and seven decision domains for the platform data governance framework. The principles align with platform governance concepts, meet the needs of all participating groups, address all types of data, and consider the platform's context. The seven decision domains are data ownership and access, regulatory environment, contribution estimation, data use case, conformance, monitoring, and data provenance. Besides the proposed framework, Lee et al. (2018) have presented different data governance models, including contingency models and centralized/decentralized models. Nokkala (2020) proposes a new data governance framework to platform contexts that utilize the framework presented by Lee et al. (2018) and other existing data governance frameworks as a basis to build on. The framework holds seven data governance issues on platforms: data value, data provenance, shared data definitions, data ownership and access, data risks, data structure, and data quality. In addition, it has metadata, data types, platform context, and specific features related to data sharing.

Table 1 Areas of data governance framework deducted from existing literature

Topic	Definition	References
Contextual factors	Internal and external needs for data governance	(Abraham et al. 2019; Janssen et al. 2020; Lee et al. 2018; Nokkala 2020)
Data access	Specifying access requirements of data	(Abraham et al. 2019; Janssen et al. 2020; Khatri and Brown 2010; Lee et al. 2018; Nokkala 2020; Otto 2011; Panian 2010)
Data lifecycle	Decisions that define the collecting, creating, using, maintaining, archiving, and deleting of data	(Abraham et al. 2019; Janssen et al. 2020; Khatri and Brown 2010; Nokkala 2020; Otto 2011)
Data principles	Data principles set the direction for all other decisions and activities to	(Abraham et al, 2019; Janssen et al. 2020; Khatri and Brown 2010; Lee et al.

	govern data through data standards and policies	2018; Nokkala 2020; Otto 2011; Panian 2010)
Data quality	Ensuring data quality and its management Accuracy, availability, completeness, consistency, and timeliness of data Limiting errors due to data inconsistencies	(Abraham et al. 2019; Janssen et al. 2020; Khatri and Brown 2010; Nokkala 2020; Otto 2011; Panian 2010)
Data technologies	Data security, data architecture, data storage, and infrastructure	(Abraham et al. 2019; Khatri and Brown 2010; Nokkala 2020; Panian 2010)
Data types	Different data types have different requirements for management and governance	(Abraham et al. 2019); (Nokkala 2020; Panian 2010)
Metadata	Establishing the semantics or content of data to make it interpretable by the users	(Khatri and Brown 2010; Nokkala 2020; Otto 2011; Panian 2010)
Roles and responsibilities	Defining roles and responsibilities regarding data and functions around it	(Abraham et al. 2019; Janssen et al. 2020; Nokkala 2020; Otto 2011; Panian 2010)
Value of data	The value of data must be understood and assessed	(Abraham et al. 2019; Janssen et al. 2020; Nokkala 2020; Otto 2011)

This chapter has attempted to summarize the literature on data governance, data platforms, and data governance on data platforms. The lack of a proven holistic data governance framework for platform data has been shown earlier in this master's thesis. As the most holistic framework, Nokkala (2020) combines the intra-organizational data governance frameworks, some of the first platform data governance frameworks presented in the literature, and the findings of her case studies. It does not neglect the earlier frameworks but is complementary by bringing a new point of view to data governance. Nokkala (2020) takes an inter-organizational perspective to data governance in her framework. Many organizations collaborate with other companies, outsourcing vendors, and cloud service providers (Abraham et al. 2019). For the above reasons, the framework (Nokkala 2020) will be used as a theoretical framework in this master's thesis.

3 Research design and methodology

This design science research was concerned with two interrelated objectives: studying the data governance on data platforms and how it can be designed, developed, and monitored. As stated earlier, neither technology nor the amount of data are the constraints for data usage, but organizations' structures and governance are (Luoma-Aho 2021). The literature review showed the fragmentation of data governance literature and exposed the gap in the research regarding data governance in the data platform context. This chapter describes the research strategy to achieve the objectives of this research. First, the methodology and the research process are described and justified. Second, the methods for collecting data are presented. Finally, the limitations and ethics of the research are discussed.

3.1 Research strategy

Data governance in the platform context is complex and contextual, with relatively little academic research. For this reason, the questions and the problems on this subject are not yet structured or standardized. Qualitative research was chosen to achieve a better and more holistic understanding of this area. (Eriksson & Kovalainen 2008)

After interviewing data professionals, it became clear that data governance rarely, or never, is seen as successful in data platform projects. Data governance was seen as a complex and unapproachable phenomenon, where too much governance leads to restricting data innovations. However, too little governance leads to data risks and the value of data diminishing. The lack of holistic guidance for data governance on data platforms has been shown in the literature review. The areas of the data governance framework (Nokkala 2020) were seen as accurate in interviews for this master’s thesis, but implementing the areas in real life and getting started with the process was seen as the most challenging part. Design science research aims to develop a solution to a real-life practical problem. Thus, the research approach chosen for this study can be defined as design science research (DSR) (Hevner et al. 2004).

DSR has been widely used in Information Systems (IS) research. It combines practice and theory by constructing tools for modeling, decision support systems, IS governance strategies, methods for IS evaluation, and change interventions in IS. (Hevner et al. 2004) Successful design science relies on rigorous methods in constructing and evaluating

design artifacts. DSR contributes to the knowledge base while developing relevant solutions to business problems. (Hevner et al. 2004) The design science research model for this master's thesis is presented in figure 4. The descriptive knowledge base for constructing the artifact is the previous academic literature, the theoretical framework (Nokkala 2020), and other data governance frameworks. Most of the current academic literature is generally descriptive (Al-Ruithe et al. 2019). The data collected for this research supported the existing descriptive knowledge but suggested that the prescriptive knowledge in this topic is relatively modest. Prescriptive research is an approach where the contribution to the knowledge is not the description of how things are but how the things should be or how to do something. Prescriptive knowledge complements the descriptive aspect. Thus, more prescriptive knowledge was collected from interviews done for this research, and this research aims to contribute to the prescriptive knowledge.

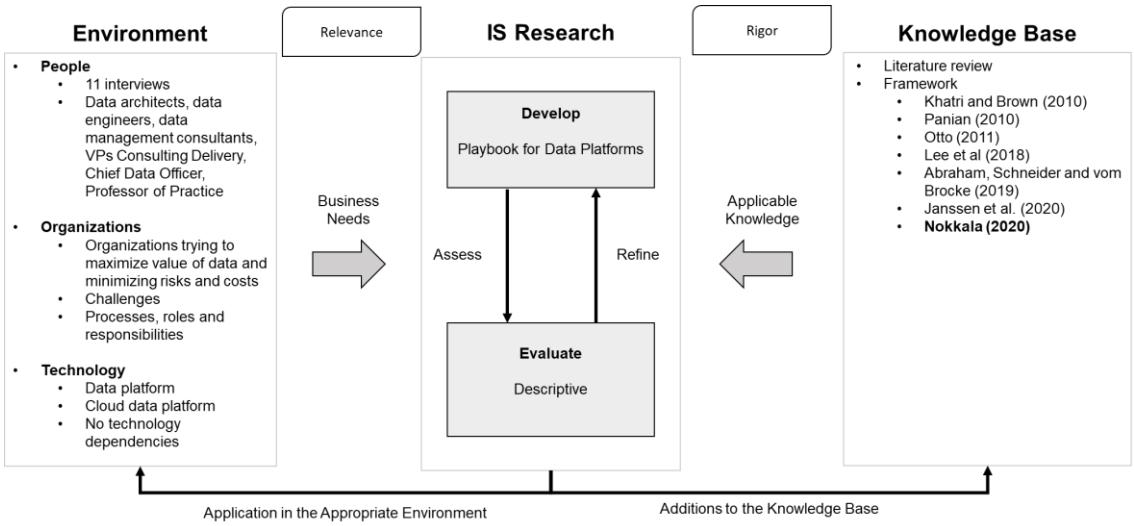

Figure 4 Design science research model (adopted from Hevner et al. 2004)

DSR allows a different level of abstraction, completeness, and maturity of knowledge for the artifact and the contribution to be produced. The contribution can be a new solution for a new problem, a new solution for known problems, extending known solutions to new problems, or applying known solutions to known problems. (Gregor & Hevner 2013) With chosen methodology, efficient use of relevant knowledge, and understanding the business need, this research aims to develop a canvas tool, "Playbook for Data Platforms", as the artifact of this research. The artifact can be considered as a new solution for a known problem. The six activities of DSR suggested by Peffers et al. (2007) will be used as a structure for this research:

1. Identification of the research problem and justification. (Chapter 1)

2. Definition of objectives of the desired solution. (Chapter 2)

3. Creation of an artifact. (Chapter 4)

4. Demonstration of the artifact by solving a problem. (Chapter 5.1)

5. Observation of applicability of the solution. (Chapter 5.2)

6. Communication of results to the proper audience. (Chapter 6)

3.2 Data collection

The study analyzes the experiences of 11 data professionals with diverse backgrounds, experiences, and competencies. The interviewees and their profiles are presented in table 2. The interviewees were from different projects and organizations. They all have worked as consultants, which gave them experience from various environments. All interviews were approached with a personalized email. The interview questions were not sent beforehand. Just the high-level interview themes “data governance”, “data platforms”, and “data governance on data platforms” were sent to them when scheduling the interview through email. The first round of semi-structured interviews was done remotely during a four-week timespan. Each interview lasted from 40 minutes to 65 minutes. All interviews were recorded. The second round of semi-structured interviews was done remotely during a two-week timespan to evaluate the created artifact. Four interviewees were selected, from which two had participated in the first round, and two had not. These interviews lasted 30 minutes.

Table 2 Interviewee profiles and interviewee durations rounded to the closest quarter of an hour

Interviewee	**Number of years working with data**	**Job role**	**Duration of interview 1 (creation of artifact)**	**Duration of interview 2 (evaluation)**
Interviewee 1	7	Vice President Consulting Delivery, Analytics & Automation	1 hour	
Interviewee 2	15	Cloud data architect	1 hour	
Interviewee 3	21	Data engineer	1 hour	
Interviewee 4	4	Data engineer	1 hour	
Interviewee 5	7	Chief Data Officer	1 hour	

Interviewee 6	15	Senior Data Management Consultant	1 hour	
Interviewee 7	17	Senior Data Management Consultant	1 hour	
Interviewee 8	21	Vice President Consulting Delivery, Analytics & Automation	1 hour	30 minutes
Interviewee 9	6	Vice President Consulting Delivery, Analytics & Automation	45 minutes	
Interviewee 10	8	Senior Consultant	1 hour	30 minutes
Interviewee 11	26	Professor of Practice, Entrepreneur	45 minutes	
Interviewee 12	7	Director of Consulting Services		30 minutes
Interviewee 13	3	Project Manager		30 minutes

In the interviews, the researcher made notes using the framework (Nokkala 2020) as a template to collect the relevant areas of the data governance framework to data platforms. At the end of the interview, the researcher showed the template and comments made giving interviewee possibility to reflect on what had been discussed and if there were any misunderstandings or if something was missed during the interview. There were no significant misunderstandings. In most cases, some areas were added after seeing the framework.

3.3 Data analysis

The coding scheme for the data analysis follows the theoretical framework areas presented in chapter 3. Five new areas, or themes, were needed in addition to the themes from the framework. After analyzing the themes, more specific patterns inside these themes were analyzed, which resulted in categories. Table 3 is presenting the data structure and coding scheme created for this analysis.

Table 3 Example of data structure and coding scheme for data analysis

Interviewee	Coding example	Category	Theme

1	The client did not know what to do with the data, but wanted to build a big data platform, have all data there, and figure out its value later. It was pretty complex. Nowadays, it usually starts from known use cases.	Use case	Cloud transformation
1	Currently, having personal information in a data platform requires specific mechanisms and procedures to be compliant with GDPR	GDPR	*Regulatory environment*
1	Data lineage and data catalog are some of the new themes that are currently relevant	Data lineage, data catalog	Data provenance
1	For regulatory reasons, some data can only be on-premise. There might be a requirement that data must be able to access even if all the connections from Finland to other countries are broken from Finland to other countries.	Cloud vs. on-prem	Regulatory environment
1	In a cloud environment, storage of data is cheaper, so managing the data lifecycle needs extra emphasis in order to avoid a situation where there is a lot of data, and no one knows what it is and where is it used	Data lifecycle	Data structure

The analysis started by reading each transcript twice thoroughly to get the first impression of the complete set. This already led to an understanding of what kinds of artifacts could be needed. Then, transcripts were analyzed more deeply by starting to find what interviewees had said about the different themes in the framework (Nokkala 2020) and what new themes arose from the interviews. All gripping narratives were captured into an Excel worksheet and coded. After the first set of the coding scheme was ready, they were reorganized into 23 themes and 78 categories.

3.4 Ethics

Research ethics give a framework for conducting and reporting research (Eriksson & Kovalainen 2008). The research has been conducted and reported according to the Guidelines of Finnish Advisory Board on Research Integrity for responsible conduct of research (TENK 2012). Ethical considerations are even more critical when studying human objects. Confidentiality and anonymity of the individual interviewees and organizations were ensured. Every referring to any organization or name of an individual was left out from the transcription. Audio recordings were deleted after the transcriptions were conducted. The transcriptions will be stored for two years before deletion. Informed

consent was granted by accepting the interview invitation. Each individual was given a chance to end the interview at any point by leaving the virtual meeting.

Most of the interviewees had direct or indirect contact with the researcher. Thus, the researcher can be considered as an insider. The challenge with insider position usually is getting confused with what the researcher knows intuitively and what the researcher knows based on the research evidence. Also, presumptions in a familiar environment and topic can be misleading. (Eriksson & Kovalainen 2008) In this research, due to short working experience with data platforms, the researcher does not have much knowledge or presumptions on data governance - which was the trigger for doing this research in the first place. The variety of roles, experiences, and backgrounds of the interviewee mitigate the risk of normative beliefs of the researcher. Instead, the insider role provided excellent access to the core of data professionals who work with these issues daily.

4 Data governance on data platforms

The artifact's creation will be structured by using the three sub-questions of this master's thesis. First, the purpose of data governance is analysed. Second, the areas that should be included in the playbook are analysed. Third, the cloud data aspect to the model will be discussed. Previous academic literature and data collected for this research are utilized for all of these analyses. The artifact resulting from these is demonstrated and evaluated in chapter 5.

4.1 The purpose of data governance

The first sub-question is *"What is the purpose of data governance?"* To design a tool for designing data governance, one must first understand why data governance is needed and its purpose. Definition of data, governance, and data governance were conducted in chapter 2.1 and summarized in figure 10, adopted from Nokkala (2020).

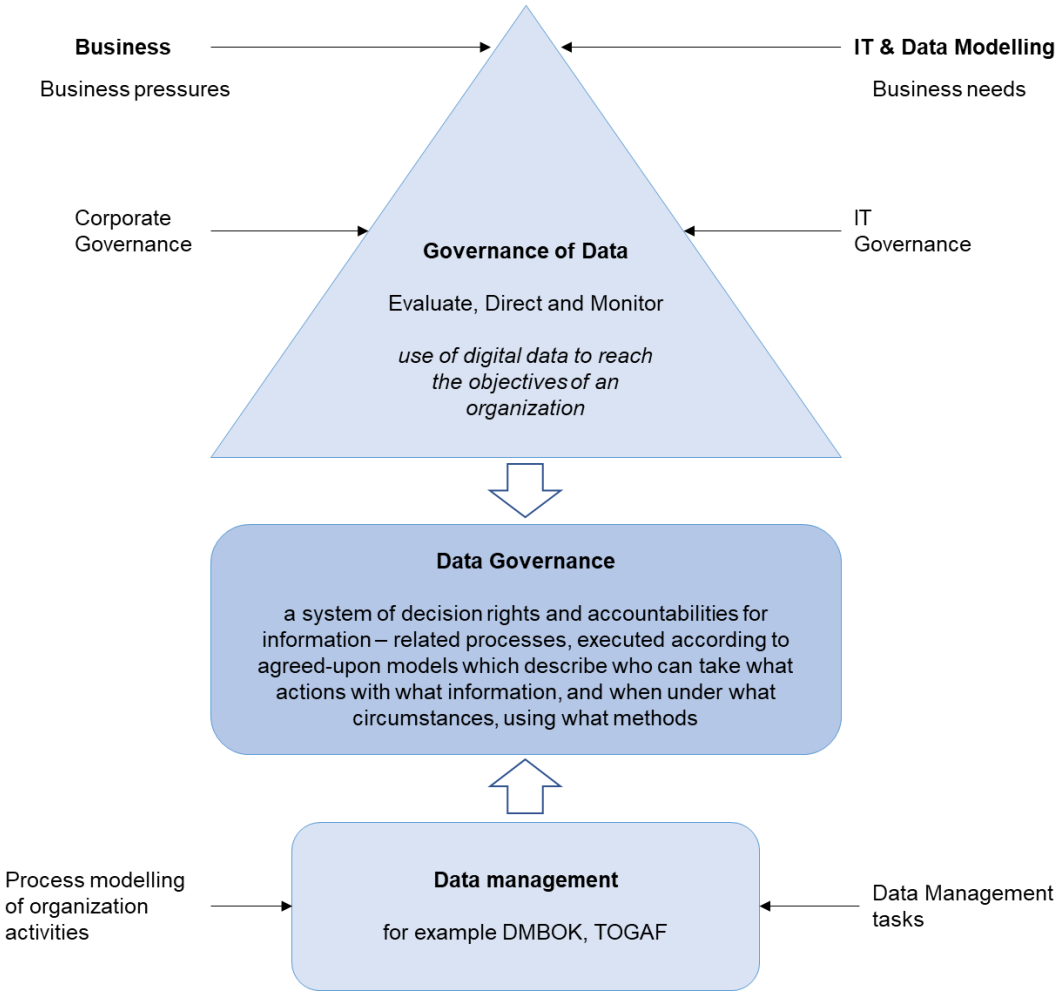

Figure 5 Governance of Data, Data Governance, Data management (adopted from Nokkala 2020)

Different academic literature definitions recognize the need for managing data as a strategic asset, which requires specifying accountabilities and decisions rights, and implementing and monitoring data policies, standards, and procedures (Abraham et al. 2019). This master's thesis adopts the definition presented by (Abraham et al. 2019): "The purpose of data governance is to increase the value of data and minimize the data-related cost and risk." In practical sense, the interviewees emphasized that data governance should not be a dusty book just for the sake of it. It should be scalable, maintainable and cross-functional. Consultants from outside of the platform can come and say how things should be done but the people involved the platform are the ones making the data governance happen. It should be in use and it should be challenged as well.

4.2 Areas of data governance framework

The second sub-question is "What are the areas of data governance framework in the data platform context?" The seven data governance frameworks analysed for this master's thesis are summarized in chapter 2.3. Most of these seven frameworks include value of data, data quality, data stewardship, roles and ownership, metadata, data access, and some type of data standards, policies, or principles aspect in their framework. For this master's thesis, the framework introduced by Nokkala (2020) and presented in figure 6 serves as the theoretical framework. It is one of the few frameworks proposed for this area and it complements the earlier literature on the topic. Each area of the framework (Nokkala 2020) was analysed through data collected and the previous academic literature. This analysis resulted 12 areas of data governance in data platform context. There did not rise any new areas to the framework (Nokkala 2020) but some of the areas were merged into one due to the inseparable nature of them.

However, the framework as such is presented from too high-level to help practitioners implement it in practice as such, according to the interviews. The challenge of getting started in the process of implementing data governance and proceed iteratively step by step demands a model in more detailed level. To response for this need, the Playbook for Data Platforms is presented in chapter 5.

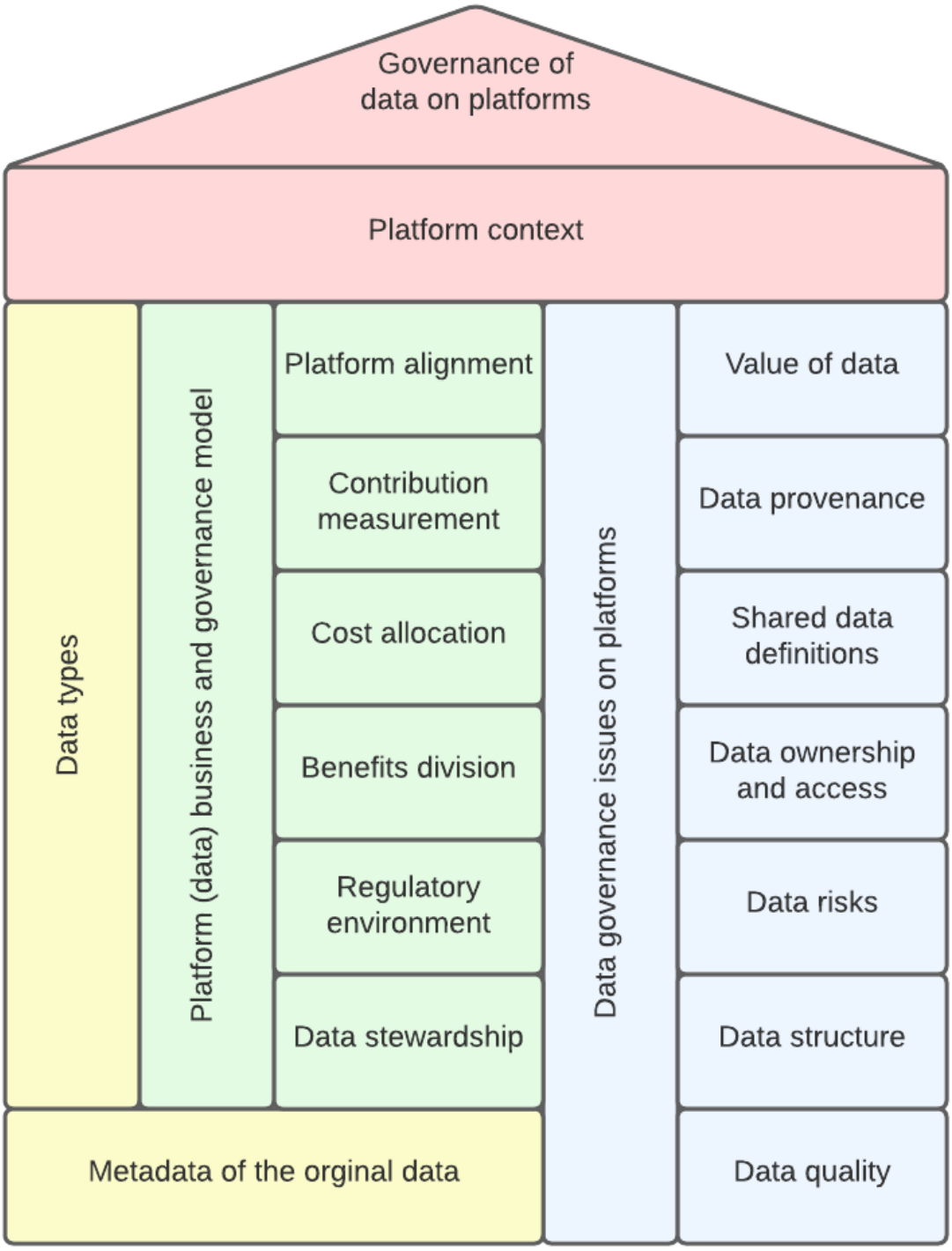

Figure 6 Data governance of Platform Data framework (adopted from Nokkala 2020)

4.2.1 Platform context/strategy

Data governance aims at maximizing the value of data (Otto 2015). Competitive advantage is described as something that is valuable, rare, inimitable, and non-substitutable. This idea of data as an asset is not a new one. It can be traced back to the 1980s when the focus of management shifted from tangible goods to intangible goods,

like information and data. (Otto 2015) Considering data as a valuable asset must be linked with business and lead to actions (Khatri & Brown 2010; Nokkala 2020). **Platform strategy** is an important part of the platform data governance program by providing the platform's objective and context. For what purpose is the platform existing should be one of the first questions to ask when designing a data platform. This purpose carries to data governance on data platforms as well. Interviewees argued that there should not be a separate data governance strategy. Instead, it should share the same goals with the organization. This idea is supported by Chakravorty's (2020) who argued that understanding what to govern and why is the starting point for data governance. The same question arose from the interviews with data professionals:

> "Start with why. It does not make sense to have a platform if there is no answer to why the platform exists. How is the data used and what is it used for? What is the value that is wanted from data?" (Interviewee 8)

The business/IT alignment model (Henderson & Venkatraman 1993) shows the importance of interaction between business and IT strategies. The same alignment model can be applied to data (Nokkala, 2020). The visualization of the framework (Nokkala 2020) was criticized in interviews for not showing this connection to business clear enough. Nokkala (2020) included platform alignment as one area of data governance to illustrate the different relations of power within the platform and the role in business. Centralized and decentralized design choices to balance ownership and power within the platform are essential choices (Lee et al. 2017). However, the platform strategy was seen as better wording for this. Data governance strategies need to be aligned with the business goals to define and develop the requirements and functions needed. These goals can vary and they should affect the data governance if it does. The goal can be for example to increase revenue, manage risks or improve performance quality, which affects the platform design (Chakravorty 2020).

> "The 'why', the priorities and chosen must-win battles should flow down to the data governance program to truly maximize the value of data for the organization" (Interviewee 11)

Besides the business and digital strategy, **the platform context** affects the data governance. The platform context includes the internal and external factors such as characteristics of the industry, market dynamics, the company's governance approach, which can all change over time making the data governance contingent Otto (2013). A growth-oriented organization might adopt a decentralized approach, while profit-oriented

organizations choose a centralized approach. The willingness to take risks rose in the interviews and can also be a significant factor. In addition, different industries have different requirements regarding data security, quality, retention, and archiving. Monitoring and updating these requirements is a continuous process (Al-Ruithe & Benkhelifa 2020). The regulatory environment was mentioned in all interviews and it will be discussed in its subsection.

4.2.2 Value of data

The purpose of data governance is to maximize the value of data while minimizing the cost and risks. During the interviews, this was experienced as a double-edged sword. On the one hand, controlling risks and being compliant with regulation is often seen as the purpose of data governance. This is the preventing role of it. On the other hand, governance is an agreement between the providers of funds and the users of the funds on how to achieve the objectives set. Ensuring the return on investments usually requires having the right data available fast, data to be high quality and understandable, having possibly guidelines on how to use it, where it can be used, and the tools around it to be effective. This is because of data's nature: it has only value when used (Otto 2011). According to the interviewees, this role of enabler has been under-valued in data governance projects.

Often data governance discussions jump straight to technical and socio-technical questions while forgetting the generally accepted goal of data governance: maximizing the value of data while minimizing the costs and risks. Nokkala (2020) acknowledges the importance of the value of data. For that reason, it is included in the framework as its own area. However, Nokkala (2020) argues that the meaning and the value of data are contextual, supported by the interviews. In some of the interviews, the platform strategy was seen as an input of platform data governance and the value of data as its output. Hence, the platform strategy and context set data and data governance goals. Failing in implementing effective data governance can lead to losing any competitive advantage of an organization (Abraham et al. 2019)

The value of data and its measurement is one issue of platform data governance (Nokkala 2020). Defining the natural data value and the measures for data value are still open questions (Abraham et al. 2019). Especially in multi-sided ecosystems, the contextual value of data makes it even more difficult, as Otto (2015) argues. Interviewees agreed

that the value of data governance and the value of data to an organization are equal. When the data is starting to be valuable, also the role of data governance gets more valuable. Thus, measuring the value of data and data governance complement each other. Otto (2011) proposes that organizations should have measurable formal goals for data governance, including business goals and IS/IT goals. These goals can be for example, related to compliance, customer satisfaction, operational efficiency, data quality, or integrations from a business and IS perspective. These goals can help to measure both the value of data and data governance. Interviewees proposed measures for the data governance, such as how well project lead times can be predicted.

4.2.3 Shared data ontology

The cloud transformation and how it eliminates the restrictions of size and format of data did arise in most interviews. Still, pouring all the data to "data lake" or such is not enough to make it interoperable according to Nokkala (2020) nor the interviewed data professionals, who often referred to the term "data swamp": "Managing the processes and controlling the formats is important to prevent data lake from transforming into a data swamp" (Interviewee 2)

In DMBOK, Earley & Henderson (2017) argue that having many different data types makes it difficult to manage and govern data. This domain gets even greater emphasis in the inter-organizational context, according to Nokkala (2020). One challenge is integrating traditional data and big data due to the siloed nature of traditional data (Abraham et al. 2019). The challenge of making integrated data interoperable was often mentioned in interviews, but data types were not seen as their own topic. Instead, data types, including metadata and shared data definitions, were seen as a whole when making integrated data interoperable within the data platform. Thus, they are all discussed under the area of shared data ontology.

Nokkala (2020) divides data into six data types: transactional data, reporting data, historical data, master data, reference data, metadata. These data types are illustrated in figure 6. When considering the different data types in the data governance context, Abraham et al. (2019) have a broader perspective on categorizing data. They divide data into two clusters: traditional data and big data. Traditional data mean all the master, transactional, and reference data. Big data can be described with five Vs: variety, velocity, volume, veracity, and value (Lee et al. 2014)

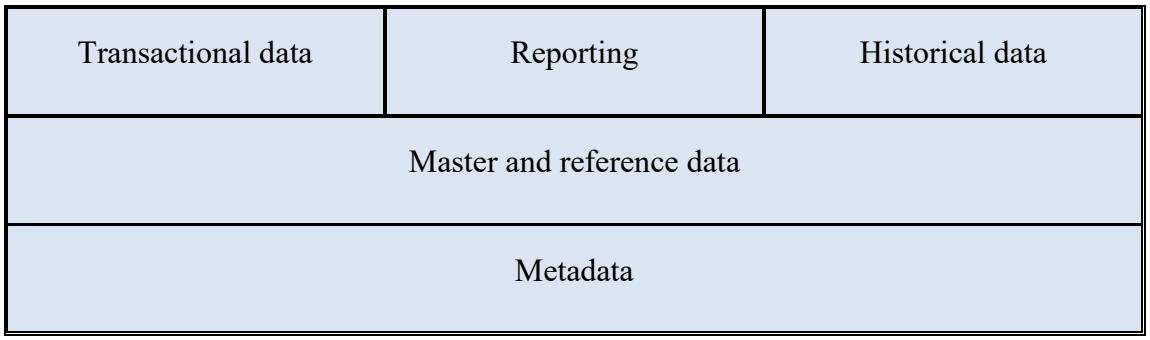

Figure 7 Data types (adopted from Nokkala 2020)

Master data is the core of an organization's data used as the basis in transaction data. Reference data is standards that are used to check data's validity. (Nokkala 2020) Data management and master data management are subordinate to data governance. Thus, they are not discussed too detail in this master's thesis. When trying to combine data from different data sources in a data platform, the problem of "many versions of the truth" is often faced according to the interviews, making data interoperability extremely difficult. The current data management thinking is to have one universal truth to data or a so-called golden record for each data (Nokkala 2020).

> "What is the master for each data? Where is the master data? Who is allowed to modify it? Who is responsible for keeping it up-to-date? These all are questions to answer, especially in the beginning of building data platforms." (Interviewee 3)

The problem with master data is its contextuality. Many organizations aim to have centralized master data records even though it is usually created in different contexts that vary from each other (e.g. accounting, procurement, sales). Nokkala (2020) argues that "single truth" is not always needed in integrations. Nokkala (2020) offers a federative approach, which emphasizes the role of shared data ontology to enable interoperability, and can be seen as a complementary, or in some cases alternative, approach to one single truth approach. The federative approach builds on the idea that different contexts remain when networks grow and changes in the way the business is done happen. The approach resonates with the need for scalability highlighted in the interviews since the needs might change in the future and the organizations merge, for example. In such cases, ensuring interoperability often requires governance related to **metadata**.

In general, effective use of data requires an understanding of data (Chakravorty 2020). Metadata explains what the data is about, and it holds the technical information, processing information, and socio-contextual information about a data attribute (Nokkala

2020). With metadata, one can understand data better, making it easier to decide whether the data is valuable or not for one and the data sensitivity level and data retention periods. (Abraham et al. 2019; Nokkala 2020) According to the interviews, the role of metadata has grown in the cloud environment:

> When there are not just one or two teams working with data and producing proof of concepts and aiming for quick wins, the need for systematically scalable data operation is needed (Interviewee 9).

Metadata documentation helps to make the data interoperable (Nokkala 2020). Many interviewees called “data catalogs” (i.e., metadata repositories) one of the most relevant topics in practical data governance today. However, finding the equilibrium in data documentation can be a challenging task. There should be enough documentation that, for example, losing a specific employee does not hurt the organization and its processes, but also not too much documentation that it still supports people's daily work (Chakravorty 2020). Metadata documentation calls for roles and responsibilities. First is the responsibility of deciding what metadata needs to be documented and who is responsible for defining it. Second is the accountability for producing metadata. Third, maintaining, using, and accessing metadata needs to be governed. (Chakravorty 2020)

4.2.4 Data provenance

The number of data sources and tools has grown due to cloud transformation. For this reason, data provenance was seen in interviews as an essential and relevant part of creating value from data, which is supported by academic literature (Nokkala 2020). Many interviewed did not recognize the term “data provenance” but instead used “data lineage”. However, data provenance is a broader concept than just data lineage. In addition, data provenance is widely used in academic literature. Thus, it is also used as a term in this master’s thesis.

Data provenance presents the history of data by showing the trail from the origin of the data record to the current place and explaining how and why it is in its current place (Gupta 2009; Nokkala 2020). Data provenance helps to understand the value and credibility of data as creation, usage, changes, and deletion for each data entry affect data the end-user uses (Nokkala 2020). In interviews, understanding the value of data, helping to evaluate the credibility, being compliant with legislation were seen as factors for data provenance being relevant today. Interviewee 7 referred to data provenance as the data

supply chain. The data is produced in one place, and then there are people using the data in another place - understanding what happens between matters.

4.2.5 Data ownership & stewardship

Fragmented attempts to govern data system-by-system have mostly failed due to working in silos with no organization-wide support and structure. (Al-Ruithe & Benkhelifa 2020) These silos lead to unclear responsibility, accountability, and data quality issues (Janssen et al. 2020). Thus, Abraham et al. (2019) see data governance as a cross-functional effort that overcomes functional boundaries and data subject areas. Chakravorty (2020) acknowledges the difficulty of implementing roles and accountabilities across the organization if a siloed organization is the starting point. It can be the most challenging task in deploying a data governance program.

Data ownership is a common area of data governance frameworks. “Owning” as a concept gives a contradictory impression related to the idea of data as an organizational asset. Still, the term has been established by practitioners and researchers. Deciding whether the owner of data is a dedicated data owner, data producer, or data stewardship, can be ambiguous. (Otto 2011) The difficulty of the term ‘ownership’ and what goes below it was also apparent in interviews. All interviewees mentioned ownership as one topic under data governance without specifying the stewardship and ownership. One of them divided ownership into two categories based on interviews: technical ownership and business ownership. He described technical ownership as something that helps to cope with issues regarding data quality, rules for cleaning the data, and deleting the data, for example. In contrast, business ownership is interested what data presents and how it can be used. Based on the academic literature, technical ownership can be referred to as data stewardship and business ownership as data ownership.

Data owners are accountable for certain data but the actual development of rules for handling the data should be the responsibility of data stewards (Otto 2011). The data owner can be determined based on application, location of data storage, or the process using the data. In addition, the scope of data ownership is a contextual issue. For an organization aiming at holistic data analytics, data ownership can have a much broader scope in data domains than an organization focusing on regulation. (Abraham et al. 2019) Data ownership is highlighted on multi-sided platforms where data can be shared outside of the organization. Abraham et al. (2019) suggest that retaining control over data shared

intra-organizationally requires more research. They also highlight the increasing complexity of data sharing in one-to-one, one-to-many, or many-to-many settings.

Data stewardship promotes data awareness (Nokkala 2020). Data stewardship refers to data management over the acquisition, storage, aggregation, and de-identification of data and the processes for releasing and using it (Rosenbaum 2010). Data stewardship should belong to the people defining, producing, and using the data. (Chakravorty 2020) Having data management as close to the action is logical since the benefits and disbenefits of data are also experienced there (Niemi 2015). Platforms collect data from various sources with very heterogeneous data. Nokkala (2020) argues that despite the general governance model of the platform, someone should be there to take care of the data's lifecycle and metadata documentation.

4.2.6 Data access and security

Thinking about how the data can be accessed and with which tools are part of the data governance in data platforms. Data access refers to who, when, and on what conditions can access the data objects. Data usage rarely follows the organizational structure, making it challenging to control data over its entire lifecycle (Janssen et al. 2020). Especially when sharing data outside of the organization, the fear of providing confidential data is a concern for organizations. Thus, precise controlling of data access is essential. (Nokkala 2020). Nokkala (2020) had paired data access with data ownership. However, during the data analysis for this master's thesis, access and ownership were separated as separate areas. Instead, data security was paired with data access because that is a driver for restricting data access.

The importance of designing the user groups and levels of data access were highlighted in the interviews. The user groups should be scalable because fixing them is slow and challenging. Failures in data access management can lead to wrong people accessing data and right people not being able to access data. One interviewee referred to the CIA Triad model when designing and governing data access management. CIA stands for confidentiality, integrity, and availability. Data should not be accessed or read without authorization to ensure confidentiality. Data should not be modified or compromised to remain in its intended state and edited only by authorized parties to ensure integrity. Nevertheless, at the same time, it should be accessible unimpededly for those who should have access. (Center for Internet Security 2022)

Too restrictive data access management was seen as a problem in many interviews. The nature of data is that it does not cost more after production if more people are using it. Instead, the more people use the data, the more valuable it gets, proving that it is the correct data, according to interviewee 8. Maximizing the value of data might require sharing it outside the organizations. There occurred a conflict between the opinions of the interviewees on sharing data outside of the organization. Most interviewees thought "what is shared and to where" is a more relevant question than whether the data is accessed internally or externally.

> "Sharing data externally does not affect the framework. It is just another end-user group, and just like for users within the organization, the mechanisms to access the data, monitoring those and what data is being shared is something that needs to be considered." (Interviewee 4)

However, one interviewee argued that the risk level increases exponentially when data is shared outside the organization, affecting data governance, tools, and many more things. This is an important topic when aiming to maximize the value of data through ecosystems. Future research is required to understand better these different aspects.

Another concern related to data accessing arose from the academic literature and was also discussed in interviews. For cloud services, data access has been a concern since organizations feel like they are not aware of who is accessing their data (Al-Ruithe et al. 2019). If some regulatory factors or data needs to be accessible 100% of the time, even in extreme cases, data should be kept on-premises according to all interviewees. In addition, the interviewees advocated for thinking about the data accesses before even bringing the data to the data platform. If there is data accessible on a data platform that no one is needed to access, it should not be brought to the data platform in the first place. This preventive thinking did not differ whether the data platform is in the cloud or on-premises.

4.2.7 Data risks

The consequences of data governance can be divided into intermediate performance effects (i.e., the value of data) and risk management (Abraham et al. 2019). This distinction was apparent in both the interviews and the academic literature. Too little governance can lead to unclear responsibilities, uncontrolled risks, and not taking the right actions (Janssen et al. 2020). However, over-bureaucratic data governance can limit

the innovations, motivate users to bypass the policies, and take risks when using data (Abraham et al. 2019).

Many data risks within data platforms are risks with any data, not just with data on the data platform. Often organizations still lack the understanding of all the data they own and handle every day, let alone the value and risks associated with the data (Nokkala 2020). Not knowing what there is in the platform was supported in the interviews as the most considerable risk. At the same time, perceived risks are preventing data sharing on platforms. (Nokkala 2020) Data risks are caused by poor data policies or the absence of data ownership regarding data quality. Having a clear data governance structure helps to reduce these risks. (Nokkala 2020) Standardized, documented, and repeatable processes reduce the risk together with constant monitoring of compliance (Abraham et al. 2019). Interviewees proposed some checklist before a new use case to ensure that the use case is valid, predefined things are verified, permissions are clarified and there is an owner on that data and use case. (Janssen et al. 2020) supports the idea of stewardship to manage risks. Identifying the most significant risks for the platform and having an annual clock for specific processes to ensure that risk management is up to date was suggested in the interviews.

Big data has brought new concerns regarding privacy leaks and data inconsistencies, which should be handled without preventing innovation (Abraham et al. 2019). Without control over data and its quality and compliance, decisions based on data might be risky. Missing, stolen, outdated, inaccurate, or biased data are examples of these risks. (Janssen et al. 2020). Toxic data is a relatively new concept referring to modifying the data before it is ingested for analyzing or AI applications, mentioned in one interview.

4.2.8 Regulatory environment

The platform strategy and context section 4.3.1 analyses the reasons for having data governance on data platforms. One of the drivers often is ensuring compliance with regulation (Panian 2010). The regulatory environment affects the level of governance needed. The regulatory environment is one of the platform context factors, but since its role today is significant and just growing, it is discussed on its own topic.

Organizations meet a growing number of external regulations (Panian 2010) General Data Protection Regulation European Union based law affects all the data targeted or collected

of people in the European Union (European Union 2022). For data platforms, it means that if one asks for information about all the data there is about one; there needs to be procedures and mechanisms to provide that information. Then if the person asks for deleting all data about one, there must be processes for that one. However, interviewees pointed out that regulation is much more than just GDPR. European Commission is proposing the AI Act to harmonize rules regarding AI applications. Staying compliant with these legal and regulatory provisions is about governing the data. Regulation affects the data throughout its lifecycle and touches many areas of the data governance framework. Many interviewees advocated for the idea of leaving all the unnecessary data that is under regulation outside of the platform because it is the safest way. The importance of knowing the data use cases is supported by (Lee et al. 2018).

The regulation also affects the decision between cloud and on-premise. The data is transferred to third parties for storage and processing in the public cloud. Checking the data handling practices of the cloud provider or other involved actors can be experienced difficult. The data professionals interviewed argued that the regulation is the only restrictive reason not to move data to the cloud. Usually, what to consider did not differ as much as how to do those considered things when designing a data governance framework in the cloud and on-premise settings.

4.2.9 Data structure

One of the interviewees argued that bad enterprise architecture usually leads to bad data governance. When people do not know what they should do or how they should do certain things, it quickly leads to bad data governance. One of the key benefits of the data governance framework is that it provides a structure for data management (Abraham et al. 2019). Data management usually focuses on how data element is defined, stored, structured, and moved (Al-Ruithe et al. 2019). Data policies provide guidelines throughout the whole data lifecycle, while standards ensure consistent data activities throughout the organization. Bringing new datasets or users, managing the changes in the data supply chain, setting new entities, or accessing the data with new tools are examples of processes where data policies and standards are needed.

The naming of data structure as one data governance issue in platforms was criticized during the interviews as an unclear title in the framework. Nokkala (2020) describes the data structure not only as defining the data but also as how new parties can join the

platform and start to contribute. It also refers to the storage and infrastructure of data assets, and the policies and guidelines for data creation. The data structure is closely connected to data types. (Nokkala 2020) Establishing a clear data structure through a data governance program makes it easier to join the platform and start contributing but also reduces risks and increases trust and cost efficiency (Nokkala 2020). Also, bringing in new workers gets easier when data policies tell what to do and what not to do (Chakravorty, 2020).

Data policies and standards are usually done by architects, data stewards, or externally by standardization organizations, but they should be monitored, evaluated, and revised continuously (Abraham et al. 2019). One interviewed architect said that the technologies and tools are the easy part of data governance, but the processes and getting the organization involved are complex.

4.2.10 Data quality

A third of an employee's workday might be spent searching for the correct data, and it is not uncommon that data cannot be found or has inaccuracies (Niemi 2015). Bad data quality destroys trust in data and the KPIs based on the data. Interviewees had experiences with insufficient quality data, which led to changing the tools and providers because of the trust issues on data quality. Data quality is usually one main domain of data governance frameworks, which aims to develop, monitor, and measure data quality. The impact of data quality becomes increasingly important when organizations decide based on data. The role of data governance is to ensure a good enough data quality that maximizes the value of data (Abraham et al. 2019). According to some, having a perfect data quality can be impossible to reach, while others blame data quality for false outcomes (Janssen et al. 2020).

All interviewees agreed that improving data quality is extremely difficult. One interviewee argued that aiming for close to 100% data quality does not make sense. Instead, adding resilience and ways to react on bad quality is cheaper and more efficient. Especially when looking at big data literature, there is a research gap on how data quality metrics and how accurate the big data even has to be (Abraham et al. 2019). While monitoring the data quality is important for the data platform, all interviewees agreed that the data should be fixed in the original data source supported by Nokkala (2020). Interviewees argued that the data platform's responsibility is to monitor if data has been

ingested normally, how much data were here, whether there was an average amount of data, and whether the data looked like it should be. Hence, the original data source owners are responsible for the data quality, data management, and data life cycle issues.

Interviewees saw data quality as an essential part of the data governance but felt that its role has been labeling data governance too much. Data governance is much more than data management, and its sub-function data quality management, which aims to maximize data quality, is part of data governance. Data governance specifies what decision and by whom should be taken regarding data management. (Otto 2011) Decision rights and accountabilities about an organization's data and its quality are pretty covered in the literature (Abraham et al. 2019).

4.2.11 Platform (data) business and governance model

The evolution of multi-sided platforms, i.e. data ecosystems, is a complex process that requires future research (Otto & Jarke 2019). Nokkala showed the need for a business model of the data platform in an inter-organizational context. The balance of power between the different actors within the platform affects the platform alignment. Defining the relationships and accountabilities between these different actors builds trust (Al-Ruithe et al. 2019). Cost allocation, distribution of benefits, and contribution measurement can be essential topics when data is being shared inside the platform inter-organisationally. However, to scale and grow activities around the platform, it must be designed, adopted, and used by the first user groups (Otto & Jarke 2019). The interviews could not provide insights into the business and governance model area when scaling the platform into a data ecosystem. Most interviewees said that this type of inter-organizational data sharing and business models around data platforms are still rare and in the launching phase in Finland. Sharing data outside the organization focuses mainly on technical questions such as what data can be shared, how the data is accessed, regulation aspects, and monitoring the data usage. Future research is needed to contribute to the area of the data ecosystem.

4.3 Cloud data governance

Cloud computing makes IT services available as a commodity, improving cost efficiency, removing storage constraints, enabling quick deployments, and delivering new services with dynamic scalability (Al-Ruithe et al. 2019). The theoretical framework (Nokkala

2020) did not differentiate between cloud and on-premise data from a governance point of view. However, Al-Ruithe et al. (2019) claim that data governance and security are some of the main reasons organizations do not adopt cloud computing. According to them, the risk of loss of data control, security and privacy of data, data quality and assurance, data stewardship, and data governance are all risks associated with cloud computing. For this reason, it is vital to understand how the created artifact differs in cloud and on-premise contexts, which resulted in the third sub-question: *"How does cloud data affect data governance?"* In other words, can the model be applied to both data platforms and cloud data platforms?

Al-Ruithe et al. (2019) claim that data governance designed for on-premises IT infrastructure cannot be implemented into a cloud environment. Academic literature regarding cloud governance and cloud data governance, let alone data governance on cloud data platforms is under-developed but will most likely be a topic of this decade. However, based on the data collected for this master's thesis, the difference between cloud and on-prem data is relatively small. The main differences are regulation and the lack of constraints with cloud data. Cloud transformation has enabled all formats and versions of data, both small data and big data, to be utilized cross-functionally. So, controlling the data, staying compliant with the regulation, and enabling scalability of data usage does get emphasized when data platform is utilizing cloud computing. Interviewees agreed that when looking at the data governance framework for data platforms at this abstraction level, the "what needs to be governed" does not differ, unlike "how it is governed". This playbook focuses on what areas and what questions need to be considered when implementing data governance without taking a stance on how they should be, governed. Thus, the playbook can be applied in both contexts.

The concept of cloud governance is introduced in chapter 2.2. Cloud governance defines policies to manage availability, security, privacy, location of cloud services, and monitoring these factors. One interviewee referred to it as the parent to data governance, which is aligned with (Al-Ruithe et al. 2019), who consider data governance as one crucial aspect of cloud governance. Its role and impacts on cloud data governance need to be understood, but they are outside of the model created in this master's thesis.

5 Playbook for Data Platforms

The artifact of this research is the "Playbook for Data Platforms". In this chapter, the artifact is demonstrated and evaluated.

5.1 Demonstration

The importance of data governance has been highlighted in both this master's thesis and in previous academic literature (Abraham et al. 2019; Alhassan et al. 2016). The practice-oriented literature and the interviewed data professionals call for holistic guidance to help define, implement, and monitor data governance. Organizations and top-level management lack a holistic theoretical model of guidance to govern data as an asset (Nokkala, 2020). The research goal was to design a solution that solves this problem and provides a holistic understanding of how to design data governance on data platforms. The main research question for this master's thesis is *"What questions should be considered when designing, developing, and monitoring data governance on data platforms?"*Figure 7 presents the artifact of this DSR, which is a "Playbook for Data Platforms". Playbook consists of 12 data governance areas relevant to data platforms based on the analysis in chapter 4.2. The guiding questions under each area are derived from the academic literature and the interviews for this research. Appendix 3 lists the guiding questions and the description and references for each of them.

Gregor and Hevner (2013) illustrate the contribution of DSR to knowledge by dividing the types of contribution into three levels. Level 1 is a situated implementation of artifact, which is very specific, limited, and less mature knowledge. Level 2 can be a construct, method, model, design principle, or technological rule that is more mature and complete theory than level 1. Finally, level 3 is a well-developed design theory called a grand theory. "Playbook for Data Platforms" can be categorized as level 1 or level 2 type of contribution. The knowledge is still less mature, but it is not specific or limited. The guiding questions can be utilized in many contexts and help practitioners form design principles for data governance.

Figure 8 Playbook for Data Platforms

Playbook for Data Platforms

Platform strategy

- For what purpose(s) is the platform existing?
- How is data governance aligned with the business, IT and data strategy?

Platform context

- What are the internal/external factors affecting data governance?

Value of data

- What are the measurable business goals for the data platform?
- What are the measurable IS/IT goals for the data platform?
- How is the value of data governance measured and communicated to the organization?

Shared data ontology

- How the data can be connected to other datasets in the data platform?
- How is data managed in the data platform?
- Are there common data definitions across the organization? Can these definitions change over time?
- What metadata is needed to be documented? Who are accountable and responsible for it?

Data quality

- What are the standards for data quality regarding accuracy, timeliness, and credibility?
- How is the data quality established, monitored and communicated in the data platform?

Data provenance

- How well is known where the data is coming from, where is it used, and what is been done to the data in between?

Data ownership

- How clear is the owner of each dataset?
- What is the role of the data owner in the organization in the data platform?
- How well is it known and communicated what is allowed to do with each dataset and how sensitive is it?

Data risks

- How well is it known what data there is in the data platform?
- What are the most relevant risks for the data and how those risks are managed?
- How are the members and stakeholders of the platform trained and educated on data governance issues?
- What is the process for bringing new datasets to cloud?

Data access & security

- How is the data access controlled in the data platform?
- How does the platform enable data sharing intra- and inter-organizationally?
- What are the security policies to ensure data security within the data platform?

Data structure

- What are the process for defining data, bringing new datasets to the platform, change management, release management, providing data to new use cases, and retention of data?
- What are the policies for cost control?
- How scalable is the platform?

Regulatory environment

- What regulation is the data platform or dataset within the platform affected by?
- How well is the data platform compliant with the regulation?
- How is the compliancy maintained and what roles and responsibilities are there for this purpose?

Data stewardship

- How is the responsiblity for managing the data, its quality and rules for handling the data ensured over its lifecycle?

The purpose of design science research is to solve a real-life problem. This playbook provides holistic guidance for data professionals on what to consider when developing data platforms. There is no specific role or context within the data professionals that this is targeted towards. Instead, it can be utilized for different purposes by different people. Consultants can use this to assess the data platform's current state of data governance. In contrast, C-level executives can use this by not answering these questions themselves but instead asking them from the data platform team. The project manager can utilize this to understand different aspects of data governance to take actions based on that. The contribution of this master's thesis will be discussed in more detail in chapter 6.

5.2 Evaluation

A design science research project aims to develop a solution to a practical problem in real-life (Hevner et al. 2004). To justify that this master's thesis has succeeded in that goal, evaluation of the artifact needs to be done. Evaluation is an essential part of rigorous design science research (Peffers et al. 2012). Due to the iterative nature of DSR, the artifact is developed further during the evaluation phase until it satisfies the requirements for the solution. (Hevner et al., 2004)

The artifact needs to be evaluated according to defined metrics: completeness, accuracy, and usability for this research. First, evaluating completeness helps to understand whether the created artifact has successfully summarized all the relevant data governance areas into the playbook. Second, accuracy is evaluated by how well the guiding questions in each area help to take the areas to a concrete level. Third, usability is evaluated by 1) how usable is the visualization of the artifact in canvas format 2) how well the playbook helps data professionals to implement data governance on data platforms. (Hevner et al. 2004) divide design evaluation methods into five categories: observational, analytical, experimental, testing, and descriptive. The evaluation method chosen for this research is descriptive and conducted as an expert assessment. The research questions used for the evaluation are documented in Appendix 2. The interviews were conducted as described in chapter 3.2.

5.2.1 Completeness

Evaluating completeness through expert assessments focused on the areas of the playbook. The theoretical framework (Nokkala 2020) provided the base for the areas of data governance on data platforms. After the data analysis, the number of areas reduced from the original 16 to 12. The biggest reason for this was a lack of experience in the inter-organizational platform ecosystem. The areas from the framework (Nokkala 2020) that were left out were contribution measurement, cost allocation, and benefits division. During the evaluations, the role of cost management was brought up as missing from the playbook. However, this was not related to cost allocation but rather managing the costs and monitoring cost efficiency. It was not seen as its area in the playbook but was added under the data structure as the guiding question: What are the policies for cost control?

Some naming changes and splitting of areas were also done when creating the artifact. Platform alignment was in the framework (Nokkala 2020) as the balance of power within the platform and the platform's role in business. This was replaced with platform strategy which rose as an essential topic in the interviews. Platform strategy was especially appreciated in the evaluation since it is often overlooked and would improve the base for building the data platform. Data ownership and access were as one area in the framework (Nokkala 2020), which were during the analysis divided as their own areas. This was seen as a successful change in evaluation. However, the lack of data security was mentioned by two interviewees. For this reason, it was added together with data access with an additional guiding question: “What are the security policies to ensure data security within the data platform?” In addition, data types, metadata, and shared data definitions were combined as one after the data analysis.

After the iterative process of evaluating and developing the artifact, the playbook was seen as complete from the perspective of areas in the canvas. It is acknowledged that most of the areas are connected, and the borders between them are pretty overlapping.

5.2.2 Accuracy

When collecting data to create the playbook, many interviewees mentioned the difficulty of taking concrete actions to improve data governance. The guiding questions in the playbook are designed for taking the areas to a concrete level and helping organizations in this iterative process. Some minor additions and changes were made to guiding

questions during the evaluation. The improvements were made with the red-colored font to the canvas to support the iterative nature of evaluation. Thus, the interviewees could see the changes made to the playbook and comment on those. The changes that were made are presented in table 4.

Area	*Before evaluation*	*After evaluation*
Data risks	How well is it known what data there is in the data platform? How is the risk management done and monitored in the data platform? How are the members and stakeholders of the platform trained and educated on data governance issues? What is the process for bringing new datasets to cloud?	How well is it known what data there is in the data platform? What are the biggest risks for the platform and how are they managed? How are the members and stakeholders of the platform trained and educated on data governance issues? What is the process for bringing new datasets to cloud?
Data access and security	How is the data access controlled in the data platform? How scalable is the access control for the data platform? How does the platform enable data sharing inter-organizationally?	How is the data access controlled in the data platform? How does the platform enable data sharing intra- and inter-organizationally? What are the security policies to ensure data security within the data platform?
Data structure	What are the process for defining data, bringing new datasets to the platform, change management, release management, providing data to new use cases, and retention of data? How easy is it to bring new users to the platform or new data sources to the platform?	What are the process for defining data, bringing new datasets to the platform, change management, release management, providing data to new use cases, and retention of data? What are the policies for cost control? How scalable is the platform?
Shared data ontology	How the data can be connected to other datasets in the data platform? Are there common data definitions across the organization? Can these definitions change over time? What metadata is needed to be documented? Who are accountable and responsible for it?	How the data can be connected to other datasets in the data platform? How is data managed in the data platform? Are there common data definitions across the organization? Can these definitions change over time? What metadata is needed to be documented? Who are accountable and responsible for it?
Regulatory environment	What regulation is the data platform or dataset within the platform affected by? How well is the data platform compliant with the regulation?	What regulation is the data platform or dataset within the platform affected by? How well is the data platform compliant with the regulation? How is the compliancy maintained and what roles and responsibilities are there for this purpose?

Table 4 Changes to the guiding questions during the evaluation

5.2.3 Usability

This playbook provides holistic guidance for data professionals on what to consider when developing data platforms. The target group is people working with data in general with no specific role or context. Its usability was evaluated through interviews with two main focus areas: visualization and effectiveness.

The visualization of the playbook in canvas format was seen as a good solution that serves as a checklist or "bingo game", as one interviewee referred to it. One interviewee saw the visualization of the playbook as an enabler for planning on what parts of the data governance are centralized and what is decentralized in data platforms. Adding meaning to how the areas are placed on the y and x-axis and the size of the "boxes" can be future improvements. These additions would make visible the hierarchy of each area and support the iterative process by visualizing the areas that the design should start from. However, these additions would require quantitative research and more data to be rigorously implemented on the artifact.

The effectiveness of the playbook for data professionals was appreciated in the evaluation. There were many use cases for the playbook for any data platform project. Project manager interviewed saw this as a good tool for project managers in data platforms to get a clear overview on data governance and what it means in practice. On the other hand, interviewed data professionals who are often involved with new data platform projects or taking over existing data platform projects saw the playbook valuable for their work.

> "This canvas would need to be in use every time something new is about to be built. It gives a checklist on what to take into account. If it is not done in the beginning, then sooner or later this tool will become useful to check that everything is robust and taken into account" (Interviewee 12)

The abstraction level of the tool was seen as appropriate. The tool was seen as too complex or detailed for people not working with data daily. Also, the tool was considered rather unnecessary for smaller development teams that are very familiar with these areas and work with them daily. These notions are pretty anticipated and do not neglect the importance of the tool for the targeted group.

6 Discussion and conclusions

In this chapter, the purpose of the research, methods used and results are summarized before the contribution in theory and practice are discussed. In addition, future research will be suggested.

6.1 Discussion and conclusions

The amount and value of data are increasing. Recognizing data as a strategic asset for the organization requires governing it like any other enterprise asset. This master's thesis has studied data governance, data platforms, and data governance on data platforms. The literature review shows the research gap in holistic data governance frameworks and data governance on data platforms. The data collected for this research highlighted the lack of holistic understanding and guidance in practice for this area. Data governance is rarely seen as flourishing, and getting started is challenging. Design science research was conducted to create a canvas tool, "Playbook for Data Platforms", to support data professionals in designing, implementing, and monitoring data governance on data platforms. The created tool needs to consider the purpose of data governance, all the relevant data governance areas on data platform context, and acknowledge differences between cloud and on-premise data.

The theoretical framework (Nokkala 2020) provided the base for this artifact. Data was collected with 11 interviews to create the artifact. Nokkala (2020) takes an inter-organizational approach to data platforms. However, the development of such data platforms is in such an early phase that data collected for this research could not provide insights into the platform business and governance model to the inter-organizational setting. However, sharing data between different organizations is considered at a technical level. Playbook for Data Platforms consists of 12 areas and 31 guiding questions evaluated by expert assessments in four interviews. During the evaluation, the playbook was developed to satisfy the requirements set for the artifact.

Design science research combines practice and theory. The practical contribution of this research is the artifact produced that provides holistic guidance for professionals working with data governance on data platforms. The playbook is a new solution to a known problem. The playbook suggests areas to consider and provides guiding questions to take these areas to a concrete level in daily work. The academic literature has called a proven

data governance framework. This research contributes to the theoretical knowledge base by supporting the framework (Nokkala 2020) and advancing it to a more concrete level with the guiding questions used in the playbook.

6.2 Future research

It is evident that the playbook created is only the first version. At least in academic literature, Playbook for Data Platforms is the first attempt to provide holistic guidance for designing data governance on data. If taken to use, the tool will be iteratively improved. The designed artifact could be improved with quantitative research. The size of the "boxes" within the canvas to follow their importance for data governance on data platforms and the placing of areas to match the prioritization order would bring value for the canvas. Also, it was acknowledged during the research that the value of data increases when shared inter-organizationally. Future research is needed to support organizations to take steps forward in this area.

7 References

Abraham, R., Schneider, J., vom Brocke, J., 2019. Data governance: A conceptual framework, structured review, and research agenda. Int. J. Inf. Manag. 49, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Ackoff, R.L., 1989. From Data to Wisdom. Journal of Applied Systems Analysis, pp. 3–9.

Alhassan, I., Sammon, D., Daly, M., 2019. Critical Success Factors for Data Governance: A Theory Building Approach. Inf. Syst. Manag. 36, 98–110. https://doi.org/10.1080/10580530.2019.1589670

Alhassan, I., Sammon, D., Daly, M., 2016. Data governance activities: an analysis of the literature. J. Decis. Syst. 25, 64–75. https://doi.org/10.1080/12460125.2016.1187397

Al-Ruithe, M., Benkhelifa, E., 2020. Determining the enabling factors for implementing cloud data governance in the Saudi public sector by structural equation modelling. Future Gener. Comput. Syst. 107, 1061–1076. https://doi.org/10.1016/j.future.2017.12.057

Al-Ruithe, M., Benkhelifa, E., Hameed, K., 2019. A systematic literature review of data governance and cloud data governance. Pers. Ubiquitous Comput. 23, 839–859. https://doi.org/10.1007/s00779-017-1104-3

Benfeldt, O., Persson, J.S., Madsen, S., 2020. Data Governance as a Collective Action Problem. Inf. Syst. Front. 22, 299–313. https://doi.org/10.1007/s10796-019-09923-z

Center for Internet Security, 2022. CIA Triad [WWW Document]. Cent. Internet Secur. URL https://www.cisecurity.org/spotlight/ei-isac-cybersecurity-spotlight-cia-triad/ (accessed 3.1.22).

Chakravorty, R., 2020. Common challenges of data governance 22.

Eriksson, P., Kovalainen, A., 2008. Qualitative Methods in Business Research. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom. https://doi.org/10.4135/9780857028044

European Union, 2022. What is GDPR? [WWW Document]. GDPR.EU. URL https://gdpr.eu/what-is-gdpr/ (accessed 4.1.22).

Feibus, M., 2021. The rise of the cloud data platform 4.

Fu, X., Wojak, A., Neagu, D., Ridley, M., Travis, K., 2011. Data governance in predictive toxicology: A review. J. Cheminformatics 3, 24. https://doi.org/10.1186/1758-2946-3-24

Gregor, S., Hevner, A.R., 2013. Positioning and Presenting Design Science Research for Maximum Impact. MIS Q. 37, 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Gupta, A., 2009. Data Provenance.

Henderson, J.C., Venkatraman, N., 1993. Strategic alignment: Leveraging information technology for transforming organizations. IBM Syst. J. 32.

Hevner, A.R., March, S.T., Park, J., Ram, S., 2004. Design Science in Information Systems Research. Manag. Inf. Syst. Res. Cent. Vol. 28, 32.

Humby, Clive, 2006. Data is the New Oil. URL https://ana.blogs.com/maestros/2006/11/data_is_the_new.html (accessed 9.9.21).

Janssen, M., Brous, P., Estevez, E., Barbosa, L.S., Janowski, T., 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. Gov. Inf. Q. 37, 101493. https://doi.org/10.1016/j.giq.2020.101493

Khatri, V., Brown, C.V., 2010. Designing data governance. Commun. ACM 53, 148–152. https://doi.org/10.1145/1629175.1629210

Lee, S., Zhu, L., Jeffery, R., 2017. Design Choices for Data Governance in Platform Ecosystems – A Contingency Model 10.

Lee, S.U., Zhu, L., Jeffery, R., 2018. Designing Data Governance in Platform Ecosystems 51.

Lee, Y., Nadbucj, S., Wang, R., Wang, F., Zhang, H., 2014. A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data. MIS Q. Exec. 13, 1–13.

Luoma-Aho, V., 2021. Hypen jälkeen. Hels. Sanomat.

MongoDB, 2022. What is a data platform? [WWW Document]. MongoDB. URL https://www.mongodb.com/what-is-a-data-platform (accessed 8.1.22).

Niemi, E., 2015. Working Paper: Designing a Data Governance Framework 15.

Nokkala, T., 2020. Governance of Platform Data: From Canonical Data Models to Federative Interoperability. Turku School of Economics, Turku.

Otto, B., 2015. Quality and Value of the Data Resource in Large Enterprises. Inf. Syst. Manag. 32, 234–251. https://doi.org/10.1080/10580530.2015.1044344

Otto, B. (Ed.), 2013. On the Evolution of Data Governance in Firms: The case of Johnson & Johnson, in: Handbook of Data Quality. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36257-6

Otto, B., 2011. Data Governance. Bus. Inf. Syst. Eng. 3, 241–244. https://doi.org/10.1007/s12599-011-0162-8

Otto, B., Jarke, M., 2019. Designing a multi-sided data platform: findings from the International Data Spaces case. Electron. Mark. 29, 561–580. https://doi.org/10.1007/s12525-019-00362-x

Panetta, K., 2021. Top 10 Data and Analytics Trends for 2021. Gartner.

Panian, Z., 2010. Some Practical Experiences in Data Governance 8.

Peffers, K., Rothenberger, M., Tuunanen, T., Vaezi, R., 2012. Design Science Research Evaluation, in: Peffers, K., Rothenberger, M., Kuechler, B. (Eds.), Design Science Research in Information Systems. Advances in Theory and Practice, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 398–410. https://doi.org/10.1007/978-3-642-29863-9_29

Rosenbaum, S., 2010. Data Governance and Stewardship: Designing Data Stewardship Entities and Advancing Data Access: Data Governance and Stewardship. Health Serv. Res. 45, 1442–1455. https://doi.org/10.1111/j.1475-6773.2010.01140.x

Rowley, J., 2007. The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. 33, 163–180. https://doi.org/10.1177/0165551506070706

TENK, 2012. Responsible conduct of research and procedures for handling allegations of misconduct in Finland.

Velayutham, S. (Ed.), 2021. Challenges and Opportunities for the Convergence of IoT, Big Data, and Cloud Computing:, Advances in Web Technologies and Engineering. IGI Global. https://doi.org/10.4018/978-1-7998-3111-2

Weill, P., Ross, J., 2005. A Matrixed Approach to Designing IT Governance. MIT Sloan Manag. Rev. 46, 12.

Wilson, J.R., 2002. Responsible authorship and peer review. Sci. Eng. Ethics 8, 155–174. https://doi.org/10.1007/s11948-002-0016-3

8 Appendix

8.1 Appendix 1. Research questions 1

Introduction

1 What is your current role?

2 Tell shortly about your background of working with data

3 Do you have experience on both cloud and on-premises data?

Definitions

4 Define what is meant with the term "data platform"?

   a. What kind of data can there be on data platform?

5 Define what is meant with the term "cloud data platform"?

   a. What kind of data can there be on cloud data platform?

6 Define what is meant with the term "data governance"?

Data governance on data platforms

7 What are the areas of data governance?

   a. Do these areas differ between cloud data platform and data platform?

8 How would you describe a successful data governance? You can provide examples.

   a. What are the benefits of successful data governance?

   b. How can the success of data governance be measured?

9 How would you describe a unsuccessful data governance? You can provide examples.

   a. What are the biggest pain points when designing, implementing or monitoring data governance on data platform?

10 Are there areas in data governance that are misunderstood or overemphasized?

11 What is the goal of data governance?

12 What are the risks of data in data platforms?

   a. Do these risks differ between cloud data platform and data platform?

   b. How can these risks be minimized?

13 How do the internal and external factors affect the data governance?

14 Do you find data governance as relevant topic?

   a. How has the relevancy changed during the last 5-10 years?

Platform ecosystems

15 Does sharing data outside the organization affect the areas of data governance? If yes, please describe.

16 Do you have experience on measuring the benefits of data and how the benefits are divided between different actors? If yes, please describe.

17 Do you have experience on measuring costs of data platforms and the allocation of these costs between different actors? If yes, please describe.

Framework analysis

18 Is there an area that requires explanation?

19 Do you agree on the areas I have marked as relevant based on this interview? Do those areas bring any thoughts?

20 What do you think about areas that are not marked during this interview? Should them be included in to the data governance framework?

21 Would this framework work as a model to be used in designing, implementing and monitoring data governance on data platforms? Please reason your answer.

8.2 Appendix 2. Research questions (evaluation)

Introduction

1 What is your current role?

2 Tell shortly about your background of working with data

Accuracy

3 How accurate do you see the areas of the canvas for designing, implementing and monitoring of data governance on data platforms?

    a. How accurate do you see the guiding questions?

Completeness

4 How complete do you see the canvas as a tool?

5 How could the canvas be improved?

Usability

6 How useful do you see the canvas as a tool?

7 Who do you think this tool is useful for?

8 How suitable do you see the abstraction level of the canvas for this use?

8.3 Appendix 3. Guiding questions

Area	Guiding Question	Description	Reference
Data access & security	How is the data access controlled in the data platform?	Who, when, and on what conditions can access the data objects is needed to control data accesses through its lifecycle. Too bureaucratic processes can prevent the value of data from being maximized if people cannot access the data, but too loose access control can lead to many risks with data.	Interviews, Janssen et al (2020)
Data access & security	How does the platform enable data sharing intra- and inter-organizationally?	In what ways can the data be accessed, and by which tools? Sharing data outside of one's organizations requires that it is clear what data is being shared, to where it is shared, permissions needed for that, regulation affecting and possibly monitoring how the data is being used.	Interviews
Data access & security	What are the security policies to ensure data security within the data platform?	Data security is unique to each organization and context. Security requirements need to be established and policies applied across networks, data, and configurations.	Interviews
Data ownership	How clear is the owner of each dataset?	Siloed data leads to bad quality data with unclear responsibility and accountability. The data owner can be determined based on application, location of data storage, or the process of using data.	Otto (2011)
Data ownership	What is the role of the data owner in the organization in the data platform?	The scope of the data owner's responsibilities differs between different contexts and should be determined within the platform.	Abraham, Schneider and vom Brocke (2019)
Data ownership	How well is it known and communicated what is allowed to do with each dataset, and how sensitive is it?	Data owners should know and communicate further about the data, its sensitivity, its fitness to different use cases.	Interviews
Data provenance	How well is it known where the data is coming from, where it is used, and what has been done to the data in between?	Data provenance (i.e., data lineage) is needed to understand the value of data and evaluate its credibility. The history of data from the origin to the current place can be needed for regulatory purposes.	Interviews, Nokkala (2020)

Data quality	What are the standards for data quality regarding accuracy, timeliness, and credibility?	Searching for the correct data, not finding the right data, or having inaccuracies in data are examples of bad data quality, which quickly destroy the trust in data and its value.	Khatri and Brown (2010), Niemi (2015), (Abraham, Schneider, and vom Brocke, 2019)
Data quality	How is the data quality established, monitored, and communicated in the data platform?	Validating if data has been ingested normally, how much data were here, was it the average amount, and did it look like it should be are possible areas for the data quality monitoring in the data platform context. Fixing the data in the data platform is less efficient and more expensive. Hence, improving the data quality is the responsibility of the owner of the original data source. There need to be processes and responsibilities for doing that.	(Khatri and Brown, 2010), Interviews
Data risks	How well is it known what data there is in the data platform?	Knowing what data there is in the data platform also helps assess the risks related to it. One of the most significant data-related risks is lacking the understanding of what data there is and how it is used daily. This is strongly related to data ownership, shared data ontology, and data quality. Having unnecessary data in the platform can lead to “data swamps” and data risks and cause unnecessary work if the data is under regulation.	Interviews, Lee, Zhu and Jeffery (2018), Nokkala (2020)
Data risks	What are the most relevant risks for the data, and how are those risks managed?	Procedures and stewardship to ensure that risk management is updated need to be covered. Having an annual clock for certain processes was suggested in the interviews.	Interviews, Janssen et al (2020)
Data risks	How are the members and stakeholders of the platform trained and educated on data governance issues?	Data governance can easily be seen as a dusty and uninteresting topic, so getting people to buy in is the hard part. Training, education, and awareness event for data, data governance, and data risks are part of the effective implementation of data governance.	Interviews, (Al-Ruithe et al., 2019)
Data risks	What is the process for bringing new datasets to the cloud?	Cloud is often an attractive option for storing, processing, and sharing data. However, specific regulation or platform context factors can affect the decision whether to bring data to the cloud or not.	Interviews

Data stewardship	How is the responsibility for managing the data, its quality, and rules for handling it ensured over its lifecycle?	Data owners are accountable for data, but the actual development of rules for handling the data should be the responsibility of data stewards. The rules might refer to the quality of data, cleaning the data, responsibilities for cleaning it, and deletion rules.	Interviews, Otto (2011)
Data structure	What is the process for defining data, bringing new datasets to the platform, change management, release management, providing data to new use cases, and data retention?	Bringing new datasets or users, managing the changes in the data supply chain, setting new entities, or accessing the data with new tools are examples of processes where data policies and standards are needed. Standardized, documented, and repeatable processes and constant compliance monitoring reduce the risk and increase trust and cost-efficiency.	Interviews, Abraham, Schneider and vom Brocke (2019), Nokkala (2020)
Data structure	What are the policies for cost control?	Establishing an explicit data structure through a data governance program makes it easier to join the platform and start to contribute but also reduces risks, increases trust, and cost-efficiency	Nokkala (2020)
Data structure	How scalable is the platform?	Overly complex or poorly scalable user groups and levels of data accesses can be difficult and slow to fix. Failures in data access management lead to wrong people accessing the data or right people not being able to access the data.	Interviews
Platform context	What are the internal factors affecting data governance?	Internal factors, such as the strategic, organizational, system-related, and cultural factors, affect the data governance. A growth-oriented organization might adopt a decentralized approach, while profit-oriented organizations choose a centralized approach. The willingness to take risks rose in the interviews and can also be a significant factor.	Interviews, Otto (2013), Abraham (2019), Nokkala (2020)
Platform context	What are the external factors affecting data governance?	Industry characteristics, market dynamics, legal and regulatory requirements affect data governance. Different industries have different requirements regarding data security, quality, retention, and archiving.	Interviews, Abraham (2019), Nokkala (2020)

Platform strategy	For what purpose(s) is the platform existing?	How is the data used, and what is it used for? What is the value that is wanted from data?	Interviews
Platform strategy	How is data governance aligned with the business, IT, and data strategy?	The understanding of the value of data as an asset and the risks related to it should derive from the organization's business and data strategy	Henderson and Venkatraman, (1993), Nokkala (2020), interviews
Regulatory environment	What regulation is the data platform or datasets within the platform affected by?	General Data Protection Regulation is the best known but just one law affecting data processing. Platforms might be affected by the regulation of multiple countries, for example.	Panian (2010), interviews
Regulatory environment	How well is the data platform compliant with the regulation?	Regulation might demand, for example, having processes for deleting a person's data from all places or showing the audit trail. These are capabilities that the platform must be able to do.	Interviews
Regulatory environment	How is the compliance maintained, and what roles and responsibilities are there for this purpose?	The maintainability is important since regulation, and use cases change. For example, GDPR requires that data is used only for the purposes the consumer has given consent. Even though the data is already available on the platform, it cannot be used for new use cases without consumer's consent.	Interviews
Shared data ontology	How can the data be connected to other datasets in the data platform?	Documenting data assets and metadata becomes necessary when data is used as an asset. Effective use of data requires an understanding of data. Knowing what data the organization has helps mitigate data-related risks and costs.	Interviews, Chakravorty (2020)
Shared data ontology	Are there common data definitions across the organization? Can these definitions change over time?	Having many versions of truth makes the interoperability of data difficult. In some contexts, having a single version of truth can be impossible. Metadata can be used to make such data interoperable.	Interviews, Nokkala (2020)

Shared data ontology	What metadata is needed to be documented in the data platform? Who are accountable and responsible for it?	Metadata documentation calls for visible data stewardship.	Chakravorty (2020)
Shared data ontology	How is the data managed in the data platform?	Processes for data management are a prerequisite for successful data sharing across the platform. It affects the data quality, shared data ontology, data provenance, and data value. Data management is one of the areas under data governance.	Interviews, Nokkala (2020)
Value of data	What are the measurable business goals for the data platform?	The priorities and requirements from platform strategy and context should flow into the goals. For example, increasing operational efficiency and ensuring compliance	Interviews, Otto (2011)
Value of data	What are the measurable IS/IT goals for the data platform?	The IS/IT goals should be aligned with the business goals and support them. For example, improving data quality or reducing data project lead times.	Interviews, Otto (2011)
Value of data	How is the value of data governance measured and communicated to the organization?	The value of data governance can be measured by measuring the value of data. Setting the goals, measuring the success of meeting those goals, and communicating those to the organization supports the implementation of data governance.	Interviews