MASTER THESIS
INFORMATION SCIENCES

RADBOUD UNIVERSITY

A comprehensive meta model for the current data landscape

Author:
Willem Boumans

Supervisor/assessor:
Dr. S. J. B. A. Hoppenbrouwers
stijnh@cs.ru.nl

Second assessor:
Dr. Ir. E. Herder
eelcoherder@cs.ru.nl

November 18, 2019

“Als je de beperkingen kent, kun je daarbinnen onbeperkt te werk gaan.”

-J.A. Deelder

Abstract

Many organizations struggle to turn their data into value. They do not know which data they can utilize, how they can utilize it and what they can do with the results. In this thesis we have identified the challenges faced by professionals in data management within organizations and have designed and created a meta model for the data landscape. Additionally, we have created a conceptual model based on the meta model to help modelers capture data landscapes in a model representation. This can help those professionals create the overview they were lacking and can help organizations create value from data.

We have discovered that the most important distinction to make in current data processes is whether or not the problem addressed is a big data-scale problem. These bring with them different data sizes and structures and require different applications to gather knowledge out of it. The combination of traditional database systems and big data-scale applications is represented in the meta model of the data landscape that we have developed. We show with example stories and their corresponding data landscapes how our model can help professionals capture their data landscapes in a model representation. In this validation step we again see the distinction between Big Data-scale and smalleer scale data application scenarios.

Contents

1 Introduction **3**
- 1.1 Scope and background . . . 3
- 1.2 Solution approach . . . 4
- 1.3 Further structure of this thesis . . . 5

2 Methodology **6**
- 2.1 Literature Review . . . 9

3 The Data Landscape **11**
- 3.1 Database systems . . . 11
  - 3.1.1 Retrieval and inference . . . 12
- 3.2 Data warehouses . . . 13
  - 3.2.1 Workings of data warehouses . . . 13
  - 3.2.2 Uses of data warehouses . . . 14
  - 3.2.3 Metadata . . . 14
  - 3.2.4 Data-drivenness of organizations . . . 15
- 3.3 Big Data . . . 16
  - 3.3.1 Data analytics . . . 17
  - 3.3.2 Pre-processing of data . . . 19

4 Identifying Challenges in Data Management **21**
- 4.1 Data is an intangible asset . . . 21
- 4.2 Value of data . . . 22
- 4.3 Quality of data . . . 22
- 4.4 Planning and data . . . 24
- 4.5 Metadata . . . 24
- 4.6 The people who need to manage data . . . 25
- 4.7 Perspectives on data . . . 25
- 4.8 Life cycle of data . . . 26
- 4.9 Data types and risks . . . 26
- 4.10 Technological advancements . . . 26
- 4.11 Overview . . . 27

5 Creating A Meta Model **29**
5.1 Model entities 29
5.1.1 Data 29
5.1.2 Data structure 30
5.1.3 Origin of data 31
5.1.4 Data storage 31
5.1.5 Data ownership and accountability 32
5.1.6 Applications 32
5.1.7 Business processes 32
5.1.8 Actionable knowledge 33
5.2 The reference model 34
5.2.1 Connections throughout the model 35
5.2.2 Instantiating the model 37
5.3 Towards a conceptual model 38
5.4 Relationship with other modeling approaches 41
5.5 Entities left out of model 42

6 Validation of the Model **44**
6.1 Reflection on validation 50

7 Discussion and Limitations **51**

8 Conclusions and Future Work **53**
8.1 Future Work 54

Bibliography **55**

Chapter 1

Introduction

1.1 Scope and background

During the last decade, the landscape of digital information has become much more diverse. The traditional database system in which information is stored that later could be recalled is still very much in use and of use, but new ways of storing and using digital data have come around. Currently, it is very feasible to link multiple data sources with each other to create a new, more expressive data system. This approach can for instance be used for administrative systems, which by extracting data from another database removes the need for redundantly stored information and makes it easier to switch from one organization to another because data does not need to be transferred [37].

Not only can data sources be linked to each other, there is simply much more data available to recall information from since storage has become much cheaper and because almost everything creates data nowadays. Many databases in organizations have even become so large that performing SQL-(like) queries on them will no longer be feasible because this will take too much time. These development have led to the implementation of data warehousing systems, which help organizations cope with the large amounts of data they store. A data warehousing system can often also process the stored data to give its users a useful representation of the data instead of a raw dump.

During recent years, the processing of large datasets has evolved quite rapidly with the development of data mining techniques that can extract information and create predictions out of a dataset that was previously unknown and impossible. This transition also identifies a transition in the relationship between stored data and queries on that data. No longer were queries formulated first and then answered with the data. Instead, the data can give answers to questions were not asked upfront by showing correlations in the information. We see these kinds of systems implemented more and more and sometimes decision making within organizations sis totally based on processed data and predictions nowadays.

Data collection is ultimately done by humans and humans are selective in this process, especially when regarding qualitative data [30]. This is why according to LeCompte, the first step in the data collection and analysis process should be the identification of sources of bias. As she puts it, data analysis is re-assembling a puzzle, which cannot be done if some of the puzzle pieces are missing, broken or warped. Biases in data collection and analysis can propagate to the research results and this is something that researchers and other data collectors should be careful of. Therefore it is important to have a clear data model in which all facets of the used data, their origin and the application using the data is modeled. This helps reduce bias and increase inclusivity in the gathered and processed data.

1.2 Solution approach

In this thesis we will work towards a meta model that comprises the current data landscape. This includes database systems, data warehousing, analytics and computer applications utilizing data. The goal of this model is to offer a model that can help information architects and organizations in their process of mapping and designing their data processes by requiring a structured approach to mapping and designing the flow of data through their organization and in what ways the gathered knowledge can be put to use. Our model provides modelers with the tools and concepts to create a solid understanding of the applications using data, how this is used, in what context and with what goals. Using this model in the design process of organizations and their IT applications, it forces a system or enterprise architect to think about the issues and implications that data processes bring with them. Our model can for instance be used in organizations to determine the extent to which (collected) data drives their primary processes or important decision mechanisms, or to reason about the applications within an organization, the data they use/create, how this is stored/retrieved and the further use of the outcomes of such applications.

Our model tackles amongst others the questions listed below in order to provide a solution which tackles the most common challenges faced within data management in organizations. The origin of these challenges is the Body of Knowledge from DAMA. This will be further elaborated upon in Chapter 4. Our classification model for the data landscape would need to address, amongst others, the following questions:

- What is the data that is used in the organizational process?
- Where does this data originate from?
- Where is this data stored?
- What properties does the data have?
    - Data volume

- Data structure
- Data velocity
- Etc.

- Is the data considered big data?
- Which applications use the data and with what goal?
- Who is responsible for the data management process?

We will come back to these questions and more challenges that are prevalent within data management that are to be addressed within our model in Chapter 4.

1.3 Further structure of this thesis

Chapter 2 will elaborate on the methods used for our research. For the creation of our data landscape meta model, we need to have a solid basis of concepts and terminology from the data landscape to build upon and to understand the challenges within the field. Chapter 3 of this thesis will therefore consist of a background chapter, in which we go through the aspects of data storage, retrieval and processing that are important to us. Chapter 4 will cover the literature review extracting the issues currently faced regarding data management. From these issues, we will derive the concepts needed in our meta model and create the model itself. Additionally, we also create a conceptual model based on the meta model. This is done in Chapter 5. Chapter 6 contains validations of our model by means of example stories and their corresponding data landscapes. Chapter 7 consists of the discussion of the research results and discussing the limitations of our research. Finally, Chapter 8 will present the conclusions of our research and possible future directions.

Chapter 2

Methodology

Our research, the creation of our meta model for the data landscape, follows the Design Science approach as put forward by Peffers *et al.* [38]. Peffers *et al.* state that in Information Systems research, there is often a focus on the creation of artifacts. They present a common framework for the creation of such artifacts; a road map for the Design Science process.

The approach from Peffers *et al.* is that the Design Science process is divided into six steps. These steps are as follows:

1. Problem identification and motivation
2. Definition of the objectives for a solution
3. Design and development of the artifact
4. Demonstration of the artifact
5. Evaluation of the artifact
6. Communication of the created artifact

When these six steps have been performed correctly, an artifact has successfully been created by means of Design Science. Peffers *et al.* explicitly state that these six activities do not have to be performed in order from 1 through 6. They find that the Design Science process can begin at nearly any of the six steps and then move outwards to the other steps. In their methodology, Peffers *et al.* mention different approaches; for instance the problem-centered approach starting at activity 1 but also an object-centered approach (starting at activity 2) and a design and development-centered approach (which starts with activity 3) as possible development approaches within Design Science. In our case, we did not start with activity 1 but rather with activity 3. What started as an idea to model the combination of "old school" data management and "new school" data science evolved into the creation of our meta model of the data landscape. This then needed iterations of a altering and improvement because the problem domain it would be applied in was not clear

at first because there initially were no clear defined boundaries and scope to the model. During the exploratory phase more and more items to potentially integrate arose and this needed structuring and justification from literature.

After our literature review (see section 2.1), the items to integrate and their justification became clear by deciding to focus on the most prevalent data management challenges as listed by the DAMA DMBOK and as such the complete trajectory from problem to solution took shape. In the following sections, we will fill in the six activities as put forward by Peffers *et al.* in the context of the creation of our data landscape model.

- *Activity 1: Identification and motivation of the problem*
  This activity comprises of defining the research problem we will be tackling and a justification of the value of our created solution. The problem definition will then be used to create our artifact which is the meta-model.

  We see that many organizations lack insight and understanding about the way data is and can be used within their business processes. This is further elaborated upon in Chapter 4. Organizations gather lots of data but struggle to create business value from it. Our research goal is to offer (data engineers / architects within) organizations a way to create a clear view of which data exists in their organization and by which means they can create value from it with regards to their business goals. In short: mapping out their data landscape and identifying points of improvement.

  Offering enterprises a structured view of their data not only enables them to spot points where data processing can offer them benefits but also forces them to think about their current situation with regards to data processing and how fitting it is to the organizations processes and goals. It can be the case that data is processed in ways which do not offer direct or optimal value to the organization or their clients. These processes can then be restructured and optimized by thinking about which ways of data processing *do* offer added value.

- *Activity 2: Objectives for a solution*
  What should our solution offer to be a good solution within the boundaries of what is feasible based on current knowledge? We define how our solution helps with the problems found in the literature review.

  In order for our model to be a usable and good solution, it should offer a complete yet concise overview of the data management process: it should include the different kinds of data storage, processing and applications from the creation of the data to the use of the outcomes of data processing applications within business processes. The level of abstraction of the model needs to be such that the entire data landscape can be covered within the model

but also such that more specific organizational issues and items can still be expressed clearly. It should consist of all the relevant entities and relations to let modelers capture the organizations data landscape and value creation through data processing.

- *Activity 3: Design and development*
  The third activity consists of creating the designed artifact. This includes determining the desired functionality and architecture of the artifact and then creating said artifact, which in our case is our model.

  In Chapter 4 we will present an overview of challenges faced by professionals active in data management, based on the Data Management Body of Knowledge from DAMA. These issues we will translate into model entities in Chapter 5. Together with basis structural entities and the relations between the entities, this will make up our meta-model for the data landscape. Design choices such as which entities to incorporate and which not to incorporate into the model are also elaborated upon in Chapter 5.

- *Activity 4: Demonstration*
  Activity 4 entails the demonstration of the created artifact to solve instances of the earlier stated research problem. To perform this activity, knowledge on how to use the model is required.

  We will demonstrate the use of our model by means of example stories; presenting different scenarios in which our created model will be evaluated based on real life cases[1]. This is done in Chapter 6. It will give several examples of how our created model can be applied in real life cases and how it can help solve organizational issues regarding data management. The goal of the demonstration is to show that our model is a helpful tool in solving the problems faced by professionals in data management, as put forward in Chapter 4.

- *Activity 5: Evaluation*
  This activity consists of observing and measuring how well the created artifact supports a solution to the research problem. It is to show that the designed and constructed model is a sufficient and correct one with regards to the stated research problem.

  We will perform the evaluation of our created model in Chapters 6 and 7. Based on the example stories made to resemble real world organizational scenarios as good as possible, we gain a view of how good our model helps

[1] Though the cases are not obtained in real-life situations, they are chosen carefully to match the real world as optimally as possible. Validating our model with actual organizational situations is outside of the scope of this thesis.

in dealing with problems regarding data management in organizations as described in Chapter 4. By seeing how good our model fits with the scenarios that professionals encounter, we can evaluate our meta model and identify possible points of improvement.

- *Activity 6: Communication*
  The final activity of the Design Science methodology consists of communicating all parts to the relevant public. This includes the results from all five of the previous activities from this methodology. The problem statement, justification of the problem and the solution, created artifact and evaluation thereof all need to be communicated to our intended audience. In our case, communication of our developed model is done by means of this thesis. In this thesis we go through the design steps of the model; what we did and did not include in the model; an evaluation of the model and example stories. We also cover limitations to and possible future directions for our model.

2.1 Literature Review

The first part of our research consists of a structured review of academic literature to identify the challenges in the data management field, both in (recent) history and currently. The method we used for collecting academic literature about data systems, data management, modeling and data mining/science is as follows:

- The starting point of our literature collection is the Data Management Association Data Management Body of Knowledge (DAMA-DMBOK) [22].
  DAMA is a global professional association of technical and business professionals. They are focused on improving the concepts behind and practices of information and data management. DAMA aims to be a resource for those professionals involved with information and data management. The Body of Knowledge provides foundational knowledge about all aspects of data management.

- From the list of data management challenges the DAMA-DMBOK lists, we have created a first set of keywords to enter into the Google Scholar[2] system for a broad search consisting of: “data assets”, “data value”, “data quality”, “metadata (management)”, “perspectives on data”, “data life cycle” and “risks in data (management)”. This set was expanded with general search terms such as “data warehousing”, “big data challenges”, “big data issues”, “data processing” “data management history”, etc.

- The results we obtained from these searches were used as first pool of literature to investigate.

[2] `https://scholar.google.com`

- Starting from this set of articles, we snowballed the references from the articles [47] to gather more information about the field. Eventually, our set of literature contained 18 articles [22, 11, 14, 5, 7, 8, 43, 17, 18, 24, 25, 27, 29, 31, 41, 42, 46, 48]. These articles formed the basis of scientific literature from which we derived the challenges which needed to be tackled by our created meta model.

- Because the data landscape, especially regarding big data and data science practices, is rapidly evolving, time of publication is an important metric for our literature. Some identified challenges might no longer be an issue even months later and many new challenges arise quickly. This does not mean that earlier publications have no use for us, on the contrary, they help us place the developments in the field in perspective. Many organizations still use large amounts of structured data, data warehouses, etc. New developments and technologies are often *added* to the existing data structure of organizations and do not replace them. This makes that modeling the combination of older and newer techniques is very much relevant to many organizations. Therefore our meta model can be a good support for modelers and architects in organizations that make use of multiple "eras" of data management and applications.

Chapter 3

The Data Landscape

In this chapter we will go through the important facets and developments over time of the data landscape. We will cover database systems, data warehouses, data mining and machine learning. Additionally, preprocessing of data and with it the anticipated use of systems utilizing data will also be discussed.

3.1 Database systems

As Donald Jardine wrote in his 1990 article, the real world can be mapped into a Universe of Discourse (UoD) which contains entities that are both real and abstract objects such as persons, departments or dates [23]. These entities and their relationships, abstractions, classifications and generalizations can be stored in a data system. Such a data system has a conceptual schema which the structure of the system is declared, in terms of statements about the UoD. Most data systems work as a table, containing rows and columns with information from the UoD. An example is given below in Table 3.1.

On top of the structure defined in a conceptual schema, a data system can operate with operations such as Create, Read, Update and Delete to insert, extract and alter information in the data system. Note that a change to the information in the data system does not change the UoD. The UoD can change separately from the data system and it is the responsibility of the owner of the data system to keep the data system up to data with the UoD (if this is necessary for the system to perform well).

First name	**Last name**	**Age**
John	Doe	25
Jane	Doe	76

Table 3.1: A simple example of a data system

Structuredness of data

Data can be categorized based on the level of structure in it. We distinguish between three categories:

- Structured data
  Structured data has a formal structure, defined via an explicit data model.
- Semi-structured data
  Semi-structured data has no formal structure, but does contain markers with which semantic elements can be separated and with which an hierarchy can be established. The structure is also seen as self-describing.
- Unstructured data
  Unstructured data has no data model and/or predefined organization of the data. It offers no clear way to structure or organize the contents.

Database structure

Jeffrey D. Ullman makes the interesting note in his 1988 book "Principles of Database and Knowledge-base Systems" that there is a difference between what he calls the plan of a database and the actual data in it [43]. He argues that when designing a database, the plan of the database is more focused on than the contents of the database. This has to do with the fact that the contents of a database are continuously changing, whilst the plan of a database is more or less rigid . This means that when designing a database system, one has to be able to more or less predict what kind of information is going to be stored in and retrieved from it.

3.1.1 Retrieval and inference

Donald Jardine writes about two types of retrieval being possible on data systems: retrieving information explicitly entered into the system via Read operations and deducing information with explicitly inserted inference rules.
For example the rule:

All fathers are men.

and the data system:

Person	**Father**	**Mother**	**Sex**
John	Tim	Martha	Male
Mary	Bill	Michelle	Female

Allows us to infer that Tim and Bill are men (meaning their sex is male). In 1990, it was still very much an open question where inference rules for a data system come from. Then, they had to be invented and entered into the system manually. In later years, efforts have been made to create mechanisms to automatically deduce inference rules from bodies of text, which then could also be used in data systems [32]. Lin and Pantels research for instance is based on deriving the similarity of words based on their context in a document to create rules such as:

$$X \text{ wrote } Y \approx X \text{ is the author of } Y$$

Which means that based on this rule, a user can query the system for the author of for instance a certain book and the system can retrieve this information if a document contains the expression *"Author A wrote book B"*. Such rules can help data systems find relationships between statements such as the ones above and use this in ranking documents in a retrieval system. If the inference rule mentioned above would not be known, then a document containing the sentence *"X is the author of Y"* would not be ranked highly with regards to a query *"Who wrote ... ?"* Inference rules help overcome this and are a first step in the direction of more intelligent information retrieval systems and are still very much in use.

One thing that should always be kept in mind when inferring information from other information is inclusivity and bias. As already mentioned in the introduction, data collection and processing is ultimately a human activity and humans have all sorts of (implicit) biases, which might influence results obtained through data [30]. Starting with data collection, it is important for professionals to try and check if the data they collect represents the complete situation as in the real world or if only parts of it are making their way into the collected data. If this is the case, then one needs to be very careful with the inferences made from this data, as they might be incomplete or incorrect. Determining whether data sets are collected in such a way that they are inclusive and unbiased is a very hard challenge.

3.2 Data warehouses

Not long after the advent of database systems, so-called data warehouses arose. Data warehouses bring together data from multiple sources and make it possible to query the combined sources all at once. Often, data warehouses also incorporate solutions to filter data that is put into the system and to infer information from the stored data.

3.2.1 Workings of data warehouses

For a data warehouse to work efficient and effective, both the data in it needs to be of high quality and the systems needs to be able to translate user queries into a correct set of procedures to run through the warehouse system. Regarding data

in the warehouse, there are two approaches to accessing the data sources bundled in the warehouse. One approach, according to Jennifer Widom is to only extract information from external sources once the system is asked a query [46]. She calls this an *on-demand* approach. The other approach is an *in-advance* approach, in which a centralized repository is created from all relevant data sources that can get queried instead of having to query all separate sources. This of course requires (pre-)processing of the data when putting it into the local repository. We will come back to this in Section 3.3.2. Zhou *et al.* also mention a *hybrid* approach to data warehouses. This approach will create a repository based on parts of some of the data sources, but will also query other sources on-demand as needed [48]. Such an approach can be useful if data from third parties is used or when many different data sources are used in the warehouse but not very often. It might then be more efficient to request the data not stored internally on an on-demand basis as opposed to importing it into the data warehouse. The *in-advance* approach as Widow calls it, is what is normally meant by a data warehouse. In an *on-demand* scenario, there would not really be a warehouse in which data is kept; there would only be a query translation mechanism which pulls data from outside storage when it is needed.

3.2.2 Uses of data warehouses

Data warehouses are primarily used in organizations that want to make use of the large collection of data they have. These can include for instance sales data, customer data, product information, etc.

The main use of the data stored in data warehouses is to support the process of decision making in the organization [46, 16]. As Foote puts it: "Somehow, the data needed to be integrated to provide the critical "Business Information" needed for decision-making in a competitive, constantly-changing global economy." Due to the large growth in the use of computer systems in organizations during the 1990's, much more data became available to them and the importance of using data for a competitive advantage became bigger. Because of this, many organizations required their data which previously was in fragmented databases to be in a single, easy to access system with a high data quality. Data warehouses were the solution that could help organizations gain intelligence out of their collected and stored data and make decisions based on this intelligence. Being able to measure and decide within your organization based on data sparked the Business Intelligence movement that nowadays is a very common practice within larger organizations. Business Intelligence makes it that organizations gain insight in their processes through collected data and helps them in the decision making to support and increase value creation.

3.2.3 Metadata

An important facet of data systems is metadata. Metadata, simply put, is data about other data. Examples of metadata include timestamps, file type, file size, order of

data, author of data etc. Metadata can help greatly in structuring, ordering and filtering data sets. For instance, when knowing at what time a file was added to a data system, one can decide to extract only those files younger than a certain time period. Without sufficient metadata, you would only have the raw data and this operation would be impossible because there is no opportunity to filter on the time at which files entered your system.

Chaudhuri and Dayal discern between different types of metadata in the context of data warehouses: administrative, business and operational metadata [7].

- Administrative metadata is the needed information to set up and use a warehouse. For instance the warehouse schema, access control and descriptions of the source databases.
- Business metadata contains business terms.
- Operational metadata is metadata that is collected during the usage of a warehouse. For instance error reports.

Within the context of data warehouses, there is often a separate repository to store and manage all the metadata [7].

3.2.4 Data-drivenness of organizations

Business Intelligence has led to organizations changing their business process to be more driven by metrics such as Key Performance Indicators (KPIs) and score cards, which are extracted from all the Business Intelligence data that the organization gathers [18]. This is done to gain more and a more direct insight into the performance of the organization. When an indicator shows that parts of the process are performed sub-optimally, management of the organization can choose to intervene.

In more recent years, data mining has proven to also be potentially useful for use in business intelligence and knowledge management in organizations because it can show additional patterns in the gathered data by an organization [45, 46]. With the advent of big data scale data collection and processing, Business Intelligence systems need to be able to turn data into information much quicker than before in order to be able to directly monitor and steer the organizational processes. To accomplish this, machine learning and data mining are used to create faster predictive and prescriptive analysis [29] which tries to predict the effects that changes to the business process will have. Larson and Cheng argue that analytics approaches using big data are different to the organization as more traditional Business Intelligence. This is because in faster analytics, the data in the Business Intelligence system *drives* the business instead of only supporting it. Therefore it is very important to make sure the data in the system is meaningful and not hindering the analysis process.

The way in which data drives the business process in an organization is by offering better precision in measuring and by offering more efficiency in the measuring of everything within the organization. As McAfee and Brynjolfsson put it:

"From intuition to rigor" [36]. The added insight trough big data, because everything within the business is measurable, makes organizations use it as their main method of decision making. McAfee and Brynjolfsson note that there is often an advantage for organizations that are "born" digital, meaning that they were established with (big) data measuring already in place, over older organizations that have to adopt these techniques within their structures to stay competitive. Enterprises adopting these new digital techniques often have to change their business processes and have to educate their employees on how to use these new systems. This can cost an organization large amounts of time and resources, reducing the competitive advantage of this company.

3.3 Big Data

With rapidly growing data generation from for instance social networks and digital sensors and with cheap storage solutions being available, more and more data is created and stored to (hopefully) be of use to organizations, governments or individuals. Chen *et al.* state in 2013 that due to the rapid generation of data, technology at the time could no longer process these amounts of data in a reasonable time [8]. This is what is commonly refered to as Big Data. Big data is identifiable by the so-called 3 *V*'s:

- *Volume*: How much data is stored and processed.
- *Velocity*: The speed with which data is created and needs to be stored and processed.
- *Variety*: The level of heterogeneity of the data sources, types and structures.

The ever increasing capabilities of technology make it such that the point at which a data set becomes big data is never the same. Therefore, each situation needs to be investigated in the context of the technological capabilities at the time of that situation. A big data problem from ten years ago might be processable by current technologies without a specific big data approach, because the data processing and storage techniques have gotten better over time.

Big data applications often have to deal with data coming from multiple sources. According to Chen *et al.* an important challenge is to integrate different data types of data (unstructured, semi-structured, structured) and still keep their meaning. An interesting note Chen *et al.* make is that with the rise of big data, data has become more uncertain. Therefore a good data cleansing and integration system is needed for applications using that uncertain data as inputs. Machine learning and parallel processing are seen as promising techniques to deal with the ever growing volumes of data, although Chen *et al.* in 2013 already stated that information technology cannot keep up with the growth of data. García, Luengo and Herrera agree with Chen *et al.* and mention that data mining requires correct *and* meaningful data. A

trade off has to be made between the overall data quality and the efficiency of the data mining program [17]. The largest problem that large data sets introduce is that they often also have a large amount of dimensions which makes data mining algorithms computational rather complex and thus slow. Techniques to reduce the dimensionality of data are amongst others feature selection, instance selection and discretization from data. Of course, which reduced representation works best depends on the specific data mining technique you are using and the connections you want to investigate in your data set. On the basis of this intention, you can decide which features of your data set to ignore. If many or all dimensions are relevant to your application, you can also apply techniques reducing the data size and/or variability such as sampling [17].

3.3.1 Data analytics

In 2017, David Donoho published his article "50 Years of Data Science" [11]. In this article, he presents an overview of the origins and rise of data analytics and what today is known as data science. This is an interesting viewpoint for us because it allows us to see where the new techniques fit into the existing landscape.

Donoho explains that during the last 50 years, statistics has made a change from more theoretical statistics towards a field that is more focused on prediction and presentation. Part of this development is also the emerging of machine learning and techniques to deal with ever growing data sets. Taken together with the traditional statistics, this is what Donoho refers to as Data Science.

David Donoho identifies the beginning of the data science field in 1962, when John Tukey released his article called "The Future of Data Analysis" in which he argued that statistics at the time were too mathematically focused [41]. Tukey was more interested in the application of statistics than in the theory behind which he found to be too focused on specific mathematical subjects. Tukey, together with Wilk listed four driving forces behind the development of data analytics in the 1960's [42]:

- Formal theories of statistics
- Developments in computers and display devices
- More and larger bodies of data
- Emphasis on quantification across disciplines

It is interesting to see that these four forces are today still very much applicable within data management and data science, especially the larger and larger data volumes being created.

In later years, the field of Data Science has gone trough a large development. The term "data mining" emerged during the 1990's when techniques such as support vector machines arose; not much later the discipline called "data science" came

into life. It was William Cleveland who first proposed a discipline called data science in 2001 [9]. This new discipline was focused on supplementing the field of statistics with data analytics. In Clevelands words: "Technical areas of data science should be judged by the extend to which they enable the data scientist to learn from data". In the years after, the field accelerated very quickly. According to Donoho, a part of this is thanks to the Common Task Framework (CTF) which arose around 2010. A CTF is a challenge put out to the general public to create the best prediction algorithm. Such challenges are open and the results can therefore help the entire field by providing new and more efficient solutions to big data scale problems. A few years earlier in 2006, Davenport, Cohen and Jacobson found that many organizations competing with each other for market share and profits did not longer rely on traditional ways to distinguish themselves but rather use data and predictive modeling to optimize their key business processes and thus reduce their costs and improve their earnings [10]. Many organizations try to apply some sort of data science into their practices to measure and improve their organizations. The quick development of data science as a discipline led to discussion about what is and what is not data science and if it is a real science or not[1][2] because there is no clear problem statement given upfront. Pete Warden argues that although data science often does not start with a clear problem in mind but instead has data and tries to find patterns in it, it offers a structured way to reason about data and challenges from it and thus *is* a science. The structured way to work with data (collection, processing, visualizing etc.) is the overarching set of activities that defines data science as a practice according to Warden. In Donohos eyes, data science is not much more than applied statistics. He objects against the arguments that data science is something new because it deals with large data sets and that it requires new technical skills. Traditional statistics already worked with very voluminous data sets such as census data and used to develop all kinds of specialist mathematical models, instead of applying techniques such as Hadoop. According to Donoho, there is no real new science in data science. It is the science of learning from data, which already was done by traditional statisticians. Donoho predicts that the most useful development in data science can be the usage of scientific publications as bodies of data to analyse and study, but this requires a less commercial focus of data scientists.

Currently, deep learning is a hot topic within data science. By using multiple layers to extract information from data, it offers many new insights, especially with image and speech data. Because processing of very large data volumes requires a large amount of computing power, we see a shift towards cloud computing to offer the needed amount of storage and processing capacity.

[1] `http://radar.oreilly.com/2011/05/data-science-terminology.html`

[2] `https://blog.revolutionanalytics.com/2011/05/data-science-whats-in-a-name.html`

3.3.2 Pre-processing of data

Jennifer Widom mentions that in data warehouses it might be needed to transform base data before integrating it into a data warehouse. Such transformations are called the "scrubbing" of data and can for instance be the summarizing, discarding or correcting of data [46]. The same goes when data mining, say Garcia *et al.* . Data needs to be made compatible with the algorithms and storage that you wish to apply to them [17]. An important note they make is that in order to know how to pre-process your data, you need to have an understanding and specification of the problem you are trying to solve via data mining on top of general cleansing of the data. Moreover, larger data sets have a large amount of noise in them and they can be high-dimensional meaning that pre-processing is necessary to ensure that applications are reasonably efficient and produce reliable results because current applications cannot efficiently process data sets with many dimensions [17].

Lee *et al.* discern in data scrubbing between correcting data errors and de-duplication of data [31]. The former is more about the correctness of data, whilst the latter enables data warehouses and other data processing applications to make more effective and efficient use of large bodies of data. Making sure that the data which enters your application is correct is very important according to Lee *et al.* since "garbage in" will most likely also result in "garbage out". Of course, knowing which parts of your data are incorrect and what their correct values are is a very hard challenge. In their paper, Lee *et al.* limit correction to removing typing errors and abbreviations. In the field, we see a demand for good pre-processing of data, which includes the scrubbing of data and also knowing what is relevant to store and what can be discarded [14, 7, 25, 5]. Especially with the advent of big data scale data sets, it has become very important to reduce data volumes before processing them because otherwise it would not be computationally feasible [17].

Anticipation of use

Data(base) systems do not have unlimited storage location and cannot be searched through very effectively if they are very large. Therefore it is useful to pre-process the data that enters the database system, to make sure you only store what is needed later. However, to be able to leave data out of the system, you need to know which data is needed in the future. For instance when considering retrieval systems, you need to know what kinds of queries are going to be put into the system such that you can store only relevant data and store it in an efficient structure. This is quite hard and requires a good prediction of what the system is going to be used for.

Often it is possible to optimize the data system after some time of use, because it provides you with the information of how the system is being used. For instance, observing the search strategy employed by users [34]. The data system could then be optimized by taking usage information into account. Another example is the Splunk system, which applies a time-sensitive data model because more recent data is accessed more than older data. This data model stores more recent data on faster

storage, making the system faster in general [4]. Bitincka *et al.* state however that a better anticipation is not possible, because often queries are entered in the system that have never been entered before and because it is impractical to extract all distinct queries for a system upfront. Splunk tackles this problem with a flexible retrieval-time scheme which adds parsing rules when new queries are entered. A system that is well configured to its intended use can offer large performance benefits. The more data a data system needs to process, the more important it gets to only do and use what is relevant.

Chapter 4

Identifying Challenges in Data Management

In this chapter we will present an overview of the most important challenges found within the data landscape, with a focus on both the more traditional data management and the more modern data science / big data approaches. Our goal is to identify points of concern for professionals and organizations with regards to the way they handle data. For this we have performed a literature review, as explained in Chapter 2.

The DAMA DMBOK[1] presents a solid list of issues that their professionals encounter within data management. The most important problem is that, according to Evans and Price many organizations recognize the potential of their data yet do not manage it as such [12]. Meaning that a lot of valuable data is not used to create value for organizations and their customers. Solid data management enables the value of data to be extracted and used. The DAMA mentions in their DMBOK a list of 13 challenges that data management presents to professionals active with data management [22]. We will go through these challenges in this section and will elaborate and expand upon them using more academic literature.

4.1 Data is an intangible asset

The DAMA states that data is not a physical asset of an organization. You cannot grasp it, but can still easily move it, copy it etc. Data can even be at multiple places at one time and used for multiple purposes, but when it is lost it is very hard or even impossible to recover again. This makes data unique for organizations, which additionally requires its own management strategy. There is no one-size-fits-all approach that is applicable to every enterprise because data, processes, applications and goals differ too much per organization.

Kumar and Palvia extend on this problem statement, stating that data is an orga-

[1] Data Management Community Data Management Body of Knowledge

nizational resource that needs to be managed as such. In their study they evaluated how a data system including data from both within and outside the organization for multinational organizations can help executives make decisions based on gathered data from all parts of the organization [27]. These decisions can have a high impact and therefore it is important to treat data as an organizational asset and not as a side issue. Additionally, this requires a global data architecture in order to gain the optimal result.

4.2 Value of data

It is not possible to put a precise and concrete (monetary) value on data, because the costs and benefits of it are unclear. Most of the added value that data can provide to an organization is highly time-dependent and depends on the process in which the data is being used. A list of sales data (if large enough) can for instance be used to predict trends and depend advertising campaigns on, which might provide a company with a lot of profit. On the other hand it can also be used to keep better stocks of products, cutting costs. Yet these applications being *possible* does not mean that they are being used and this has an effect on the value of the data. Is the data valuable to an organization if it is only stored and never put to use? Is it valuable data for competitors of the organization to own? Is data still valuable if it is outdated? The value of data strongly correlates with the potential gains in profit to be made by using the data in an application. Since this depends on many factors, it is very hard to determine the potential value of data upfront. Kumar and Palvia agree with this. Their approach to maximizing value gain from data uses internal organizational data to gain a more precise insight into the organization and uses external data to get a solid view of the business environment. To an organization, the internal data of another is often too complicated for the first organization and thus has less value to them. Understanding the data requires a full understanding of the organizational processes of the specific organization. Zhou *et al.* found in 1995 that combining data from heterogeneous and external sources into a data warehouse was an important challenge for organizations [48]. This issue is still very much applicable nowadays [26, 35].

4.3 Quality of data

Data quality is of high importance to data management. As an organization or individual you need to have trust in your data. This is especially the case if decisions within the organization are made based on data. If the data put into the decision system (which can be automated but this can also be a human acting based on information gathered from data) is of low quality, then the outcomes of the process have the risk to also be of low quality (this is also known as the "garbage in, garbage out" principle [31]). In turn, this can have all sorts of negative consequences: from recommending the wrong product to a consumer to putting the wrong person in jail.

The costs of having low quality data or too much trust in the data you use can be extremely high. It is therefore that there needs to be a good quality metric for data. Some aspects of the quality are more easily checkable than others, for instance the correctness of facts in data. According to Lee *et al.* most of the mistakes in data originate from missing data entries and mistakes in the data due to typing errors or different formats being used [31].

Data features deemed important to a specific organization process have to be graded by the person(s) overseeing that specific process. In the case of multinational organizations, assuring data quality can be a hard process since there can be cultural and/or political differences leading to differences in the data [27]. It is hard to use the same definitions throughout an international organization with departments in different countries. People with different (cultural) backgrounds might capture the same information differently, which makes it hard to directly compare the data. This is something that a data manager needs to keep in mind when combining data sets created in different cultural settings, combined with the earlier mentioned biases that occur in data collection.

Additionally, the concepts and definitions used by one part of the organization or an external source might be different than what another part uses. Therefore it is very important to have strong definitions for them and to process imported data before using it to base decisions on [7]. Another factor impacting data quality is the freshness of data, especially data imported from external sources [48]. If data is out of date, there is the risk that organizational decisions are made in a lagging fashion rather than a leading or current one. This can lead to a reduced competitive advantage for said organization. A data warehouse should therefore be active according to Zhou *et al.* which means that updates to the data in their sources are propagated to the warehouse. Often external data sources do not send updates to their data to systems using the data, meaning that a data warehouse should check for updates itself [7]. How often this is needed is dependent on many factors, including the capacity of the data source and the level of importance to have the most up to date data. Zhou *et al.* mention that an important question for organizations to ask themselves is how important it is for organizations to have local stored copies of external data they use. The benefits from storing data within your organization and keeping sure it is up to date have to outweigh the added retrieval cost in terms of both time and resources from utilizing external sources. Chaudhuri & Dayal agree with this, stating that organizations should determine how much of the data they want to have "materialized", i.e. stored locally on their own storage and then act according to this preference [7].

When mining from data / using data in your decision process, it is important that the data is as correct as possible, meaning that it is a good practice to pre-process the data. This way, most of the "simple" mistakes such as typing errors can be eliminated, resulting in better outcomes from the mining algorithm [31, 46, 7]. Determining which data entries are "wrong" when not considering empty fields or spelling mistakes is still quite a challenge. Attempts have been made to create error detecting mechanisms [40], but this still needs the correct information to determine

what is right and what is wrong. Therefore this is quite an unfeasible approach, because knowing the correct data already would remove the need to correct the faulty data.

4.4 Planning and data

Getting value out of data as an organization is not a single, straightforward task. It requires planning across the entire organization about how data is obtained/created, stored, used within processes and how it can add value to the organization. Enterprises might need to alter their structure and strategy to accommodate for the best rewards that can be gathered from data because of the financial potential of data within their company. For instance, in multinational organizations it needs to be decided if the gathered data will be stored at separate locations of all subsidiaries or if the data will all be transferred to the headquarters of the organization [27]. Such structural decisions can have quite a large impact on the structure of the organization and its processes, therefore this decision needs to be an educated one. This is where our model can be a good support mechanism, as it maps the way data is obtained, stored and used within business processes. Developing Business Intelligence applications requires a thorough understanding not only of the application but also of the organization [18]. Through the development of Business Intelligence, organizations have become more process driven because it allows to effectively measure and steer the processes within the organization where needed. This requires them and their employees to have a focus on the global business goal instead of only single tasks. Measuring through Key Performance Indicators (KPI) provides management with insight in how to structure their processes such that it will create more value and is a data source in itself. An important challenge with this is that there is a latency between events occurring and it showing through indicators. Additionally, defining such indicators can be a hard tasks since it needs to be very clear *what* you are actually measuring with a KPI as to not perform a wrong intervention within your business process, resulting in a less optimal result.

4.5 Metadata

Metadata, data about data, provides insight in the data that an organization owns and helps make abstract data more understandable. Metadata is everywhere, from databases to organizational processes and architectures. To manage data in a good way, you also need to manage metadata as such. Metadata is especially useful and necessary to add the right structure to the data and to add to the meaning of the data. Good metadata can help explain what is represented by the data.

This can be a challenge to organizations who do not yet have a good data management practice in place as there needs to be a start point. According to the DAMA, metadata management can be a good beginning to improve the entire data

management of an organization. This is because good metadata offers insight in the data you have, making more clear the potential and uses of the data.

4.6 The people who need to manage data

As data travels through the organization, there is often not a single data manager which oversees the process. Therefore various employees (often managers of a sub-process of the organization) within the enterprise will partake in the activity of data management. They will be responsible for the data going through and originating from their parts of the primary process. These employees do not only need to have the relevant competences to perform their tasks, but also need to be able to cooperate and communicate with each other. Especially in larger organizations problems might arise in which you do not know who is responsible for which part of the data management process, resulting in communication and responsibility issues.

Whilst there are many different persons overseeing the data flow through the organization, it is generally not a bad idea to appoint a central responsible person to be responsible for the data process / strategy as a whole. This person is then also responsible for keeping oversight and control. This person can oversee that the sub-processes do not get too locally optimized, meaning that data structure and applications can be made very fitting to one part of the process, but this might be not fitting to another part. A central responsible person can oversee this and structure the processes such that they all benefit.

4.7 Perspectives on data

As data flows trough all parts of the organization, there are many different processes and persons within the organization that use the data. Each has their own perspective on the data and the contents of it, some with more knowledge of data systems than others. This makes that it is important to represent data in such a way that it is understandable for those who have to use it [27]. When data mining, this is also very important, as you otherwise risk to "dredge" data instead of mining from it [14], meaning that found patterns have the risk to be meaningless because the data and used algorithms are not interpreted sufficiently. Inference is an important step in data mining and to infer patterns from data you need to have confidence in the inference rules that you apply [14] and to be unbiased as possible [30]. To determine the optimal structure of your data it is necessary to know the intended use of it from all business processes. This might also include external organizations that use your data as an input to their processes or vice versa data that you import from external sources.

When structuring a knowledge process such as information retrieval or data mining, it can be very useful to take up the customer perspective to structure the process [14]. It makes it such that the end goal of the process, what you want delivered out of it, is clear. The process can then be structured in such a way that

it fulfills this wish. Choosing the right techniques and algorithms for the task is of the essence in order to produce a reliable and useful result.

4.8 Life cycle of data

Data has a life cycle, similar to other organizational assets. This means that data at one point is created/gathered, moved, edited, stored, etc. then is used and eventually will be removed. The life cycle of data is closely related to the life cycle of the product that the organization puts out. Throughout the entire life cycle it is important to ensure (meta)data quality to get value out of it. However, there might be too much data to effectively manage with the organizations capacity; meaning that data needs to be prioritized in terms of importance and that you need regular cleanup of your data.

4.9 Data types and risks

The data within an organization consists of all sorts of different data types. For instance: metadata versus transactional data or sales data versus network activity data. All sorts of different data types play a unique role in the organization and need to be managed accordingly.

The role that certain data has also introduces risks. Data quality and security are important risks for example. Another very important risk with regards to data is the misuse of (especially personal) data for means that they were not intended to be used for and this could lead to privacy problems and concerns. For organizations it is important to assess the risks of their data and data processes in a good way such that they stick to the legislation and that they maximize the added value for their customers.

4.10 Technological advancements

The main driving force behind developments in data management are technological advancements. The invention of database systems sparked the data management movement, which later was strongly influenced by data warehouses, data mining and big data and there will no doubt be new data techniques in the future. Data and information technology nowadays is strongly entwined with the business processes of organizations. This makes it that data management requires expertise not only of management practices but also of technological advancements.

Often, technology is also a limit for many techniques. Limited storage capacity and computational complexity restrict what is possible with data sets in a given time [14]. This has always been the case, but with the rapid growth of data volumes in the current big data era, this is emphasized. This makes the interesting situation that technology is both an enabler and also a limiter for data processing.

4.11 Overview

From the issues addressed by DAMA in the previous sections, we can extract the important aspects of data management which need covering in our model. Those are the following:

- There is no single way to gain value from data for organizations.

  Even within a single organization there are many different sub processes; each with their own persons with their own perspectives on the data process and gaining value through data.

  All those perspectives need to be brought together or at least somebody needs to have the overview and responsibility over the processes. Some processes might also need redesigning in order to benefit more from data gathering and processing. Additionally, data introduces risks that need managing.

- The (possible) value of data highly depends on:

  Technical possibilities, i.e. the ability to store and process large data volumes which need their own specific approach.

  Data quality. In short: when garbage goes in; garbage comes out. It is also hard to determine the quality of a data set, especially with large data sets because they cannot be checked by hand. Additionally, you need metadata of high quality.

  Relevance of the data is also an important aspect to take into account. Data can have sufficient quality, but when the data is aged such that it is no longer relevant, it has no use or value anymore.

- Big Data brings with it quite a lot of specific issues. Therefore it is good to distinguish data scenarios based on whether or not you are dealing with Big Data issues. This can be done based on the 3 V's (Volume, Variety, Velocity). There are however no hard set criteria for when something is big data or not. The modeler/organization/field has to decide in this. They will have to ask themselves if the organization is dealing with big data level problems in its processes. This is the largest and primary distinction needed to make when modeling data processes. Big data scale data sets and applications ask for their own specific use (and modeling) approaches. Surrounding information and entities are more or less similar, yet there also remain some key differences. Data needs to be cleansed and processed in real time and there is need for a different data architecture to deal with the vast amounts of semi-structured and unstructured data that will be imported into the data system according to Katal, Wazid and Gouar [26].

- Concluding, the most important task for a modeler is to identify "old school" and "new school" kinds of data management within the complete process

based on indicators such as the 3 V's and data structure. They give the modeler insight into making this distinction, as there are no precise criteria to define big data scale problems.

Chapter 5

Creating A Meta Model

The challenge that remains after the previous chapter is to translate the different kinds of data management in an overarching model that captures all relevant entities in an overview in the form of a meta model. This we will do in the following sections. We will work towards a meta model that comprises the data landscape as it currently stands and addresses the challenges identified in the previous chapter; cover the entities that are going to be incorporated in our model and what they express; the model itself, what is left out of the model and how the model can be used. We will also construct a conceptual model based on the meta model to show how the model can be put to use in practice. Additionally, we will also shortly compare our modeling approach with a few other modeling methods / languages often used to model (parts of) data within organizations to put our modeling approach into perspective.

5.1 Model entities

In the literature and media we see many concepts and terms that might mean the same item or understanding, but which are not defined explicitly. Most persons have some idea of what is meant with the concept data for instance, but still those ideas may vary greatly between persons. Therefore it is important for us to explicitly define the concepts that we will be using in our model as to not have misunderstandings about what is and what is not meant with them. Another important aspect of creating the model is to define the boundaries of the model; what is included and which concepts are going to be excluded from it. This is done in the following sections.

5.1.1 Data

Data naturally is the core concept of the model we are creating. Defining data can be rather difficult because it is such a broad term and is easily confused with other concepts such as information and knowledge. The definition of data as used in the

FRISCO report by Falkenberg *et al.* is a good starting point for our definition of data [13]. The definition used by Falkenberg *et al.* is:
"Data denotes any set of representations of knowledge, expressed in a language."
This definition does not entirely fit our purpose, since Falkenberg *et al.* explicitly note that nonsense and corrupted sequences of symbols are not data. In current data(base) systems there is often no cleaning or correctness checking on what is put into the system. Therefore, using the definition of data from the FRISCO report, some entries in a database or other generated data might not be data. Because of this limitation, we choose to define data as stored records that are not (yet) interpreted or processed in any way. Data is only the form, without the meaning. As Ackhoff puts it: data is raw, information has been given meaning by relational connections and knowledge is the appropriate collection of information such that its intent is to be useful [1]. We will continue to use this definition of data in our meta model.

Subjective aspects

Data in a computer system also has some "subjective" aspects to it such as the size of a dataset. In this case subjective is meant as that the value of the aspect could be objectively measured, but the decision to put it into the big data category is a subjective one. We define the following aspects as such:

- Data volume, the size of datasets
- Data variety, the differences within a dataset
- Data velocity, the speed at which data is generated and enters the system

These 3 V's are the most common criteria to decide whether or not a case is big data scale [8]. Some authors choose to expand on the 3 V's with additional criteria. Jin *et al.* for instance have included two additional "V's" in the shape of *Value* and *Veracity* [24]. Value meaning the possible monetary value to be gained from the data and Veracity meaning the high amount of uncertainty within the data.

We choose to forgo additional criteria to categorize big data scale problems on top of the 3 V's. Many authors introduce extra concepts to the 3 V's, but many of them are quite different from one another. The 3 V's can be seen as the baseline that is agreed upon, therefore we will use this as the criterium for deciding what is a big data scale issue and what is not.

5.1.2 Data structure

As mentioned in Section 3.1, data can be structured, semi-structured or unstructured. This is also how we will express data structure in our model. So: structured data has a formal structure, defined via a data model; semi-structured data has no formal structure, but does contain markers and unstructured data has no data model or predefined structure. Data sets naturally also can be of a mixed nature, meaning

parts of the data are (semi-)structured whilst other parts are unstructured. Efforts have been made to perform information retrieval on data systems containing mixed structure data [33]. In our solution, we choose to omit mixed structure data because applications using such data will still have to work with the "lowest common denominator", which most of the time is unstructured data. The additional structure in some parts of the data can then be seen as an advantage, but not as a structural part of the data set.

5.1.3 Origin of data

Digital data that is put to use in organizational processes and actually all data has to originate from somewhere. This can be done in numerous ways, for instance, digital data can be automatically generated by sensors from the readings they perform or it could be manually entered into a data system by a person.

When looking at data origins from an organizational perspective, the creation of data can be an internal or external process. Internal meaning that the data is created within the organization itself, whereas external data creation means that the data originates from outside the organization and is imported into their data systems. Importing data can be done either in batches or as a continuous stream of data. Both have their advantages and disadvantages. Which type of data importing suits best to the organizational process depends on several factors. Most importantly, the process creating the data. If this process puts out a continuous flow of data, it makes more sense to also import a stream than if data is created in batches. From an organizational standpoint, it is important that the entry of data into the organizations systems is in tune with the applications that need to use the data. This reduces time spent waiting, processing or storing data. When importing data from external sources, it might prove to be useful if you clean the data before storing it on your own storage. This makes for more reliable data and also reduces the amount of storage space needed.

5.1.4 Data storage

Data that is generated and which is going to be used by applications needs to be stored somewhere before it can be used. Such a storage location can be implemented in many different ways, depending on the intended use. Storage can range from a simple spreadsheet or unstructured flat files to complex database structures. For us it is important to discern between local and remote storage. Especially with the advent of big data volumes, we see specialized cloud storage solutions emerge that offer storage solutions to organization on which they can store their data. These organizations specialize in storage of data and can offer larger and more reliable storage than if organizations had to create their own data center. In this solution of storing your data elsewhere, there is the tradeoff between reliable and large storage volumes and the cost of having to send data which is required over the network.

This latency can become an issue for applications if very large data volumes have to be transferred.

5.1.5 Data ownership and accountability

What does ownership of data mean? Is the person or organization that stored or created the data the owner? Does this apply to both personal and non-personal data (personal data being data (in)directly relatable to individuals)? Can data ownership be transferred? Is the owner of a data storage also the owner of the data in it? Such questions are all important to consider when discussing and modeling the ownership of data.

Data can be personal and non-personal data. For personal data, it is easier to define who the owner is than for non-personal data. We can say that the data subject of personal data is the owner of that data. The challenge is in defining which and when data is personal data, since this depends on context. Some data might in itself not be linkable to an individual person, but it could be when it is combined with different data. Therefore we choose not to define data ownership as legal ownership but rather as the accountability over the data as also done by Al-Khouri [2]. He states that the data owner is the agent that is responsible for utilizing the data. This is a good definition for us to use at this moment since we can abstract from the distinction between personal and non-personal data (see section 8.1 for more on this). Defining data ownership as the agent responsible for using the data is also a better definition than just calling the storage owner the owner of the data, because this is not always the case. For example in the case of cloud storage, the owner of the cloud servers is not the data owner per se. Our definition covers these issues and we will therefore choose to use it.

5.1.6 Applications

When data is only stored, it has no use for an organization or person. Only when extracting information from the data, it can add value. This can be done in all kinds of ways: from a simple query lookup to machine learning based prediction, all of which are digital applications. We define a data application as a digital application having data as input which extracts information from that data. We argue that pre-processing of datasets also is a data application since it naturally uses data as input and makes the data more valuable by removing waste from it. Therefore our model will have no separate entity to represent the pre-processing of data. This combination of both pre-processing and processing itself is the definition of a data application that we will use in our model.

5.1.7 Business processes

Randomly applying data applications to the data within an organization has no direct benefit to the organization; there needs to be a issue/question that can be solved

with the gathered knowledge from the data application. Within organizations, these issues arise within their organizational processes which can range from production or providing a service to management and directing the organization. Data and the knowledge that can be gathered from it gives you the possibility to see trends, to optimize processes, to save resources and to get insight in the business processes. When wanting to get extra insight into a certain situation, you can consult the data (if present, otherwise you can gather data) and apply an application to it that corresponds to the problem you are trying to solve.

5.1.8 Actionable knowledge

Actionable knowledge is in our eyes the goal of any data processing application and it is what the user of an application is using the application for. Actionable knowledge is the knowledge based on which decisions and actions can be taken in (organizational) processes. Such knowledge can vary from knowing how much of a certain item is still in stock and thus knowing how much to order to machine learning based predictions of who should be given a certain morgage and who should not. Note that some processes within organizations nowadays are fully automated, making the actionable knowledge more or less invisible because the action is taken by a machine instead of a person. For these kinds of systems, it needs to be very clear how a decision is made and on which information this was based. One can imagine that an automated morgage system that was trained with the "wrong" data can lead to all kinds of problems for its users and owners.

5.2 The reference model

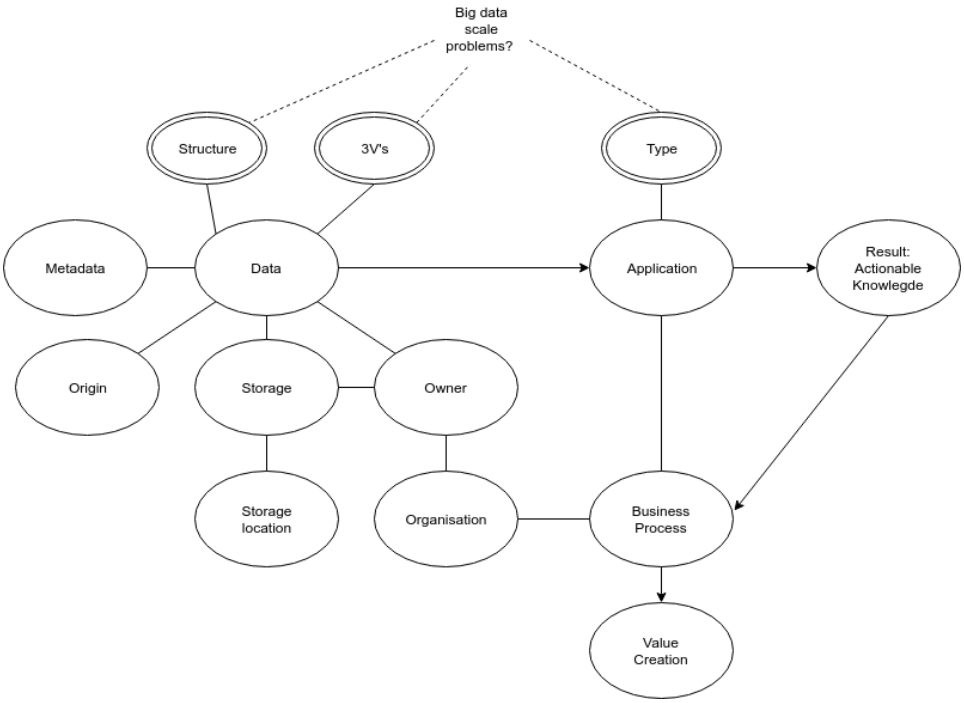

Figure 5.1: Our Meta Model

In this model, we see the entities from the previous section represented visually and connected to each other. We see data with its associated entities such as metadata, the 3 V's and data storage. Some entities are more concrete than others, which is because of the level of possible instantiations of the entities. We also see lines, dashed lines and arrows. The arrows are used to represent a certain degree of order from the knowledge gathering process in the model. The dashed lines do not represent a relationship between entities, but rather a choice the modeler has to make.

Another cluster is the application using data, which is of a certain type, is connected to a business process and which of course produces a result. The knowledge which is this result is then applied within the organization to create value.

When tackling big data challenges, some entities are instantiated differently from scenarios in which more classical data management practices are in place. These entities are: the 3 V's; Data Structure; Application type; Application; Storage and Storage location. For instance, an application can change from a query lookup to prediction using machine learning. We will go into the instantiation of the model entities in the next sections.

5.2.1 Connections throughout the model

Ethical considerations

Being ethical means doing the morally "right" thing with your actions. Ethical considerations are also quite applicable to data analytics. Gathering and analyzing data can bring many new insights, yet some of these actions might have implications which are not overseeable or have implications that might not be considered positive to the general public. This is why ethics are an important part of the data analytics process. In 2016 Luciano Floridi and Mariarosaria Taddeo have written about data ethics and what it entails [15]. In their view, data processes provide large opportunities but also pose ethical challenges. Examples they give are the processing of personal data and the increasing reliance of organizations on fully automated decisions made by algorithms combined with reduced human involvement. Floridi and Taddeo state that challenges are there to be solved which also applies to ethical challenges posed by data analytics. Data processing potentially offers such benefits to society that the issues it has need to be tackled. An ethical approach to data processing can help grow the information society and can help with the general acceptance of data-driven processes and decision making.

Floridi and Taddeo divide data ethics into three axis: the ethics of data, ethics of algorithms and the ethics of practices.

- **Ethics of data**
  Ethics of data applies to the collection and analysis of large datasets. Challenges on this axis are for instance profiling, personal advertising, open datasets and re-identification of persons. Additional issues are data trust and transparency of data.

- **Ethics of algorithms**
  The ethics of algorithms have to do with the ever growing complexity of data analysis algorithms and the increased level of autonomy that is given to these techniques. A large challenge in this context is accountability.

- **Ethics of practices**
  Ethics of practices address the issues regarding responsibility over data processes. People in organizations who oversee the data process, strategy or policy have to keep in mind the possible implications of their actions. Therefore, Floridi and Taddeo see it as a good idea to develop a professional code for data analysis in order to keep data collection and analysis responsible.

Of course, in practice these three axis are intertwined in data collection and analysis processes. Floridi and Taddeo do not prescribe concrete rules or guidelines to adhere to in data ethics, but rather points of consideration. It is up to the persons involved in the process to determine what is a morally right action and what is not.

In our model there is no specific entity representing ethics, yet we find it important to consider data ethics when applying data in an organization. Every step in the

process of gathering actionable knowledge from raw data has an ethical component and we would like to stress this by means of this section.

Data security

The act of data security entails protecting data from unwanted and unauthorized access and potential abuse. As DAMA puts it: achieving proper authentication, authorization, access and auditing of data and information assets [22]. Achieving data security is usually done through encryption and access control systems. Criminals might want to gain access to personal, financial or organizational data and use it for their own gain. This is why data security is necessary in organizations. Getting unlawful access can be done in all kinds of ways, such as hacking, fishing or physical theft. There is not a single solution for assuring data security, as there are many factors in the data process that can be a security risk and for the entire process to be secure (enough), these all need to be covered.

Depending on the considered value of the data, the level of security wanted and correspondingly the security methods change. This makes it hard to model data security into a separate model entity. Therefore, we choose to not include data security in our meta model as a concrete entity but rather as a factor that is present throughout the entirety of the model and data process.

Personal Data and Data Privacy

Much of the data collected nowadays is personal identifiable data, meaning the data leads to and can identify natural persons. Additionally, data can also *become* personal data when datasets are connected to each other. For instance, a list of sales transactions consisting of the sold item, price and customer number can become personal data if it is linked to a record of customer information which connects customer numbers to the name and address of the customers.

Many organizations process personal data in order to offer their services to their clients. How this data is gathered and for what purpose it it used is in many countries limited by law because of the privacy implications that the processing of personal data can bring. It can for instance show patterns in the daily life of people, can be sensitive medical or financial data or it can show a persons political or sexual orientation. In the European Union, the processing of personal data is limited by the General Data Protection Regulation[1].

What and when data is personal data is an ongoing discussion and is dependent on the country the organization gathering/processing the data and its clients are located. Because of this, we choose to not explicitly model privacy and personal data but stress the importance of it.

[1] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679`

5.2.2 Instantiating the model

In the following section, we will present a list of examples of instantiated model entities. These examples can be seen as typical items that a modeler might want to express in a model of our sort. These lists are only an illustration of the possibilities of our model and are by no means exhaustive.

- Data Origin
  - Sensors measuring the environment and putting out a data stream
  - Importing an open dataset
  - Checkout systems storing transactions
  - Manually entered spreadsheet with overview of library books
  - Etc.
- Data Storage
  - Relational databases
  - Data warehouses
  - Spreadsheet files
  - Flat files
  - Hadoop system
  - Etc.
- Metadata
  - Time of creation
  - Creator of the data
  - Access level
  - File type
  - Etc.
- Storage Location
  - Local PC
  - Locally hosted database
  - Externally hosted database
  - Cloud storage
  - Etc.
- Owner
  - Natural person (Not in the scope of this thesis)
  - Organization

- Organization
  - Local football club
  - University
  - Government
  - Etc.
- Application
  - Database management system such as Microsoft Access
  - Machine learning application
  - Data visualization tool
  - Error correction mechanisms
  - Data reduction tools
  - Etc.
- Business Process
  - Sales
  - Marketing
  - Stock keeping
  - Financial Administration
  - Production of goods
  - Etc.
- Value Creation
  - Increased Profits
  - Decreased Costs
  - Less environmental impact
  - More satisfied customers
  - Etc.

5.3 Towards a conceptual model

Complementing the meta model presented in Section 5.2 and the list of possible entity intstantiations, we have also constructed a conceptual model in the ORM language based on the meta model and its entity instantiations. This model is presented in Figure 5.3. The ORM model still presents the data from a high level, but with a finer granularity and with more established relations between the objects in the model. It allows modelers to represent their data landscape in for instance a database structure in which the landscape is stored. If questions arise about the landscape (for instance: which data is used by which application in which process?

Where does the organization get its data from? Where does the organization store this data?) the system can be queried for information.

The fact types in the ORM model represent the relations between the objects. Below is Table 5.1 which explains the use and meaning of each fact type. Note that some object types are annotated with a *. This means that there can exist more such objects, but we have chosen to leave them out of the model for the sake of readability. The same goes for the ORM constraints. Facts and objects of the model might be constrained by certain business rules and processes of the organization, but because of the differing nature in these rules and constraints, we have decided to not incorporate them into our conceptual model.

Fact type	Meaning
1,2	Data is of size Data Volume
3,4	Data has structure Data Structure
5,6	Data originates from Data Source
7,8,9	Data from Source is imported with Data Importing Method into Storage Solution
10,11	Storage Medium is located on Storage Location
12,13	Storage Solution is provided by Organization
14	Data is Big Data Y/N
15,16,17	Access Level of Data is stored in Metadata
18,19	Data Importing Method makes use of Pre-Processing Application
20,21	Organization incorporates Business Process
22,23	Application has a Throughput
24,25	Data is used in Data Processing Application
26,27	Data Processing Application is applied within Business Process
28,29	Results are applied in Business Process
30,31	Result Application is used to attain Value Creation
32,33	Data Processing Application puts out Results

Table 5.1: The fact types of our conceptual model and their meaning

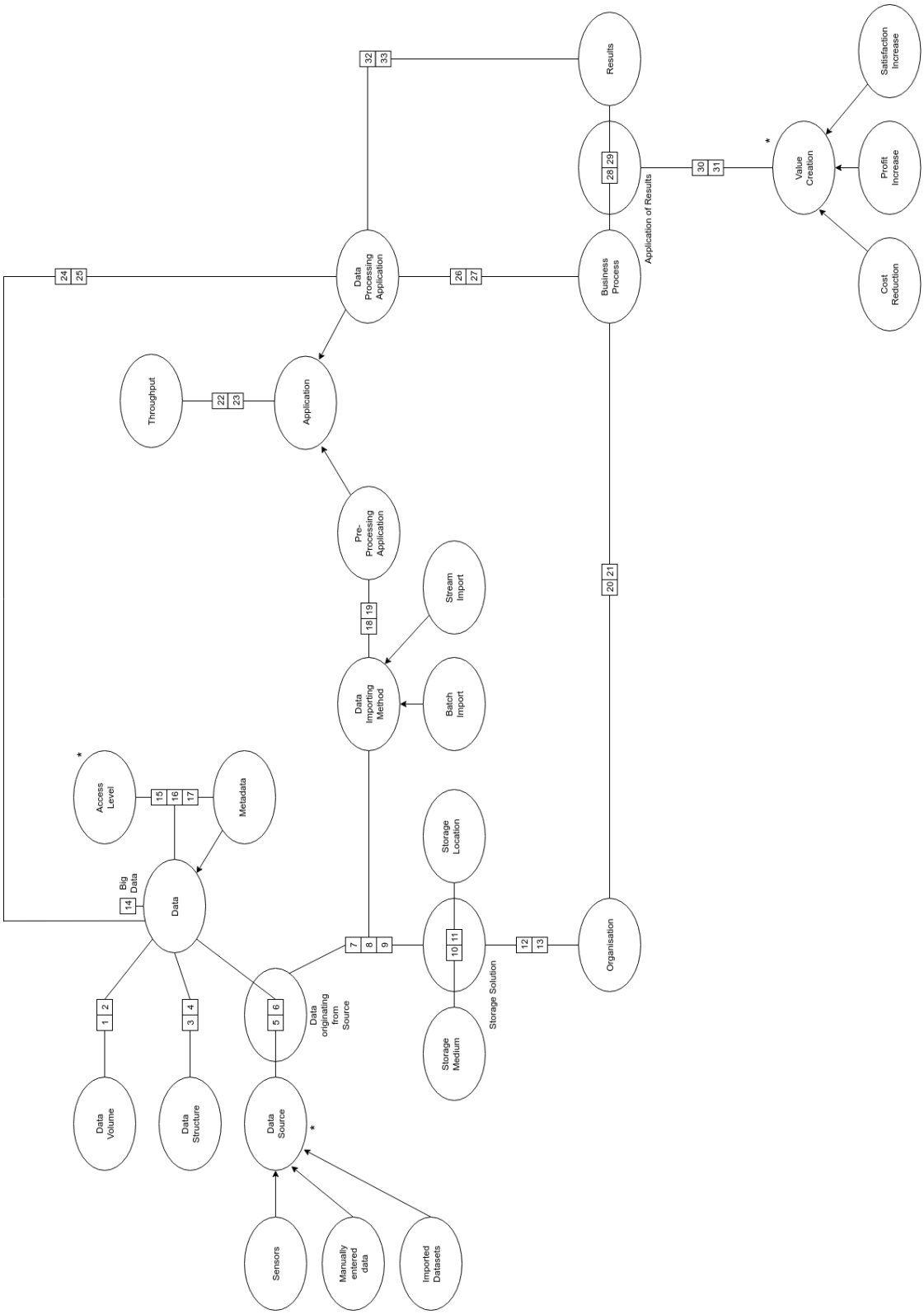

Figure 5.2: Our Conceptual Model

5.4 Relationship with other modeling approaches

A lot of other modeling techniques also incorporate modeling data into their models, albeit from a different perspective than our approach. It is interesting to compare our model to these established ones. We will look into the focus of these models, their abstraction levels and for whom the models are intended.

Enterprise architecture

Enterprise architecture can best be described by the following quote: "Enterprise Architecture tries to describe and control an organization's structure, processes, applications, systems, and technology in such an integrated way" [28]. One of the most used techniques to create such descriptions is the ArchiMate modeling language.[2] ArchiMate is a modeling language that allows an enterprise architect to model an organization from business, application and technology perspectives. It offers an extensive overview of the organization with multiple views that are more focused on some parts of the model and more abstract in others. For instance, a view for the business processes does not need to be as elaborate on the technical level and vice versa.

ArchiMate models have a clear separation into three layers: Business, Application and Technology. Additionally, there is also a separation between passive and active components in the organization (such as a data server or business process). This results in a clearly classified model, in which each entity is part of one of the layers. Our model does not have such a clear split into different layers, mostly because it does not focus on all aspects of an organization, but only on the data that traverses through it. This results in much less need for different abstracted view levels. Enterprise Architecture is intended to create a total overview of an organization, either in its current situation or in a goal situation. Our model has a much narrower focus, specializing only on the data processes.

Data modeling

Data modeling is a modeling method that is focused on correctly representing the Universe of Discourse and the structure in that data. As Calvanese, Lenzerini and Nardi put it, data modeling is the following activity: "Specifying the structure of the data to be managed within an application."[6] The population of a data model consists of the data records from the UoD, represented in a structured overview. These records are linked to each other in the form of relations between the entities.

Data modeling focuses on translating real world data from the Universe of Discourse into a structured representation that can later be used by applications. It comes in three types: conceptual, logical and physical data modeling. The conceptual data model is a high level overview of the connections and specifications of the data. The logical data model concerns the structures within that data and how this

[2] `https://www.opengroup.org/archimate-forum`

can be implemented in data structures. The physical data model looks at the data itself, how it is stored in tables and how these are stored on a storage medium, who can access this etc.

Compared to our model, there is a much larger focus on the contents of the data itself when looking at data modeling [44, 19]. Our model more or less takes the data that comes into the system for granted. It may contain errors and may be incomplete, but the generation of this data is outside of our scope.

Concern modeling

Another modeling approach that is related to our solution is concern modeling. Concern modeling is of modeling of so-called concerns of software applications. Concerns are abstract or concrete requirements put forward by stakeholders in the development process. These requirements must be fulfilled when finishing the development project [3][20]. An example of an abstract concern would be a feature that has to be implemented in the software application. A concrete concern would be a specific work product of the software development phase.

Concern modeling is the act of modeling all these stakeholder concerns into an overview with which software developers can go to work. The software product only is considered finished when all concerns are addressed. Concern modeling clearly is more focused on software applications and the development thereof, yet there is some interesting overlap with our modeling approach. We also address the stakeholder wishes to be quite important in the data management process and also concerns are not implementation-specific just as our gathering of knowledge is not application-specific.

5.5 Entities left out of model

Not all concepts from Chapter 4 made it into the model presented in Section 5.2. In this section we will go through these entities and explain why these are not included in our meta model.

Causal relationships in data

When inferring relationships between entities in large volumes of data, it is not often clear whether this relationship is just a correlation or if there is a provable causal relationship. It would be a very interesting feature of our model if we could show in our model which inferred data is proven to be in a causal relationship and which is only a correlation. However, causality analysis is rather hard because the methods that can prove causality still rely on several assumptions and specific knowledge in the field of the analyzed data which need to be taken into account [21, 39]. Therefore, we have chosen to omit this from our model. If causality of inferred relationships were to be easily proven, then incorporating this into our model would give a good measure of data reliability.

Persons managing data

We choose to omit employees overseeing the data process from our model. Including them would bring too much detail to modeling organizational processes, which has no additional benefit to the modeling of the data processes.

Evaluation of the application results

After data is used in an application, the results of that application are put to use in organizational processes. This is often not done directly, because the results of the application are evaluated first. If a data application has produced wrong or unusable results, it is of no use to try and apply these results in organizational processes.

This step of evaluation we choose to omit from our reference model. The application-entity is abstract to such a level that evaluation can be seen as a component of the application. Therefore there is no separate entity to express the step of evaluating application results.

Combining of data sources

In our created model, data has an origin and is stored on a certain storage on a certain storage location. This data is then later used within applications. What is not explicitly included in our model is the ability to join data sets from different origins into one larger data set. This does not mean that this action is impossible, rather we see data origin not as one single origin but also as mixed origins.

Level of trust in the data

Trusting data and outcomes from data applications is strongly connected to data quality. Good data with high quality can be considered trustworthy data and vice versa. Although trust is hard to quantify and prove, it can play an essential role in data management and data applications as those can be seen as a chain of trust. If the chain is broken at any point, the outcomes of the data application are also not trustworthy anymore. This can for instance happen through using low quality data or an unproven algorithm in your application.

It is very much a challenge to determine which data and/or outcomes can be trusted in large data processes. It requires the responsible person for the process to have a thorough understanding of both all the data used and the techniques with which it is processed. Because we have seen that creating good data quality controlling techniques is a rather hard process in Section 4.3, we have considered the trustworthiness of data and results to be out of scope for this thesis.

Chapter 6

Validation of the Model

In this chapter, we will validate the meta model we created in Chapter 5 in order to show the practical usability and relevance of our model. This will be done by means of example stories which portray populations of our model. We will complement these stories with a sketch of the corresponding data landscape in terms of a table containing the instantiated entities of our model. These tables can be seen as a database structure in which the data landscape is represented.

We have chosen to represent the data landscapes from the example stories in tables for several reasons:

1. It allows for a condensed text-based overview of the cases.
2. Objects being instantiated multiple times, for instance when considering multiple datasets, origins or applications; can more easily be represented in a table because a table is a two-dimensional structure. A new item such as another dataset can be put in a new column of the table. It shows which data is connected to which source, application, storage etc.
3. The ORM model we have constructed in Chapter 5 acts as a base for a database implementation of the landscape model. The tables in which we store our examples are basic databases containing all the relevant records in a single table. Depending on the database implementation and its intended use, the database can be split up into multiple tables representing the data landscape; each with their own focus.

The example stories in this chapter are not exhaustive but will attempt to cover the most important divisions we discussed in the previous chapter.

- Instantiations of our model can be structured as follows: *Organization O uses data D with application A in process P. Data D has the properties:* $3V's$*, storage S on location L, Origin Or.* We will loosely use this structure in the following examples which place our model in a practice-context.

- *Example 1:* A department store has their sales records and inventory stored in a local relational database which is ran on a server that is located in the back office of the store. The database records of the sales consist of which product is sold in which amount to which customer and for what price. The inventory database is made of records representing how many items of a certain product are still in stock in the store.

  Metadata for the sales data would be: the time of the sale, which register the transaction was made at and which employee was working at that register at the time of the transaction. The data stored in these databases is by no means big data scale. It is a small scale, locally stored database consisting of structured data.

  Using the information stored in the databases, the department store wants to improve their sales by finding out which items are frequently purchased together and locating them close to each other within the store. The data analysis therefore is rather straightforward: the database system is queried for those sets of items that are purchased in a single transaction by a single customer. The organization can then check if and how many times customers buy items together. If this is of a significant amount, the management of the department store can decide to change the layout of their store such that those items often sold together are closely located to each other.

  The data landscape from this example can be represented by the following overview:

Model Entity	Instantiations	
Organization	Local Department Store	
Business Process	Floor Planning	
Data	**1**: Sales Records	**2**: Inventory
Data Volume	Low-Volume	Low-Volume
Data Structure	Structured	Structured
Big Data Scale	No	No
Data Origin	Registers	Warehouse + Registers
Metadata	Time of sale, Register No., Employee No.	-
Data Storage	Local database at the store	Local database at the store
Application	Query system on the data storage: Finding coupled sales	-
Application Results	Overview of Coupled sales	-
Application of Results	Optimizing the floor plan w.r.t coupled items	-
Value Creation	Increasing sales	-

Table 6.1: Example 1

From the data landscape model we can gain information about the specific landscape and its possibilities. When we take a look at the landscape from Example 1 above, we see that the data used by the department store is stored locally in a database located at the store. For such small-scale, local applications this is more than sufficient because the application requires few resources and because it presents very low overhead costs. An implication of using separately ran servers and applications for each store is that the results of the application will be specific to the store itself. In case of the floor plan, this is not a large problem since the floor plan is very specific to each store. This is in contrast with Example 2, in which local sales information is translated to a more global overview of trends. Due to the low size of the datasets, the locally ran applications might have a lower reliability when it comes to presenting the items often sold together. It can be that items are only sold once or twice together, but still are presented as frequently sold together. This depends on the configuration of the application. Still, this should pose no real problem as the implications of minor changes in the floor plan are small and easily revertible. The output of the application is quite easily applicable by people in the process, as it is a list of coupled items which potentially can be placed close to each other within the store.

Whilst the results might not be one on one transferable to other stores, the process is. The step of data collection through the register system can be implemented in each kind of register system in a department store. Because the sales data is rather straightforward and has a low amount of features, there is little to no need for pre-processing of the data. Also, transmission of the data will not be a limiting factor to the system, something which is the case in Example 3. The collected and stored metadata does not offer much extra benefit with regards to the application in Example 1, but it can offer additional insights into the sales of the store. For instance: some items may be sold more or less on a certain time of the day. Such information could perhaps be used by the organization in the future. Additionally, with each transaction the corresponding client is stored. Is this really necessary? Most likely not, but the department store could put this data to use in future applications such as personalized advertising and discounts. Note that gathering personal data without stating a clear purpose for it upfront is not allowed by many forms of data protection regulation.

The goal of the application in Example 1 is to create value through extra sales by nudging clients towards buying items that are frequently bought together. By placing the beer next to the potato chips, clients are more inclined to buy both for example. The potential for increases in sales is dependent on the level of "optimization" of the sales already in place. Many customers already know what they want to purchase from a store and where it is located within that store. Therefore the gains to be made will most likely be from impulse buying. These sales can still amount to a large amount of extra sales.

- *Example 2: This example continues on the previous example.* Instead of a single department store storing their transactions, a national chain of department stores decides to organize the sales records of all their stores, their client databases, their inventories and financial data in a central data warehouse. In this warehouse, the separate transaction records of stores are transformed into a representation of (local) trends. The main organization and also local branches of the organization can use the warehouse to query for these trends and can use this information to time promotions of certain articles better.

  Additionally, by merging all the inventory databases, the store chain can cut costs because they will need less stock of products. With separate inventories for each store, each location will need to keep all its items in stock. With the knowledge of what other locations have in stock, this becomes less necessary as an item that is out of stock on one location but in stock on another can be transferred instead of having to add new items to the inventory.

Model Entity	Instantiations			
Organization	Department Store Chain			
Business Process	Marketing			
Data	**1**: Sales Records	**2**: Inventory	**3**: Client data	**4**: Financial data
Data Volume	Medium Volume	Medium Volume	Medium Volume	Medium Volume
Data Structure	Structured	Structured	Structured	Structured
Big Data Scale	No	No	No	No
Data Origin	Registers	Warehouse + Registers	Entered by clerk	Entered by bookkeeper
Metadata	Time of sale etc	-	Time of entry, Employee no.	-
Data Storage	Data warehouse at headquarters			
Application	Fitting **1-4** in warehouse	Identify trends from **1**	Creating general inventory from **2**	-
Results	List of trends in sales	Merged inventory	-	-
Result Application	Offering promotions through trends	Reduce needed stock	-	-
Value Creation	More sales through promotions	Needing less inventory	-	-

Table 6.2: Example 2

What we see here is that multiple data sources (the register system from each store) send their data to the centralized data warehouse. Naturally, this makes for an increase in data volume as we are no longer considering a single store. The largest difference compared to Example 1 is that the used data is not stored locally on a server, but instead is located in a centralized data warehouse. This means that the data imported from each store needs to be made fitting to the warehouse. In this example that will be a straightforward task since the structure present in the data can be kept in place. The data volume is not of a level that requires reduction, so importing the data into the warehouse will not be a large challenge.

A benefit for the store chain is that it can centrally order, store and structure its data instead of having a server in each store needing maintenance and supervision. Another positive aspect of a data warehouse for the department

store chain is the reduction of redundancy that it offers. Clients for instance do not need to be registered at each store; they can register once and then access that account in every store. Financial information that is centrally stored can be used by the store chain to have direct insight in the financial situation of the local stores, instead of having to wait on periodically sent financial reports. This makes the administration easier and allows for direct interventions if needed.

Contrary to Example 1, the aggregated sales data from Example 2 can be used to gain global insights about product sales. It is not big data, as the data volume is not large enough and as the data is structured data. Analysis of the sales records consists of identifying trends in sales with which stores can offer fitting promotions and fill their stocks in line with these trends. By trying to identify trends as early as possible, the organization can prepare itself and can offer better fitting deals to their customers than competitors, with which they hope to achieve more product sales.

- *Example 3:* The department of traffic control in a large city, e.g. New York wants to improve the flow of traffic in their city. To achieve this, they use the data sent by so-called "connected" cars to continuously keep track of the traffic and to adjust the traffic lights to potentially improve the flow of traffic.

  Connected cars in the city constantly send their location, direction and speed to the control center which has to process and analyze this in near real-time to identify traffic jams and to come up with flow improving configurations of their traffic lights. The vast stream of vehicle data is transmitted to a data center located elsewhere in the country. The capacity of this data center is rented by the traffic department of the city. In the data center, the vehicular data is processed and possible improvements in traffic light configurations are sent back to the traffic control department. This department can then decide to change the configuration of the traffic lights to one which is proposed by the application processing the vehicular data.

  Although the data that is sent by the cars is quite structured, analyzing data of hundreds of thousands of vehicles at near real-time clearly is a big data scale problem. This makes it that the data center which receives and processes the data needs to offer enough bandwidth and processing power to make this possible. Additionally, the application analyzing the data coming in needs to be efficient enough to produce results quickly. A better configuration of traffic lights for the traffic situation of an hour ago has no use for the traffic department. Therefore, the application might not produce the best result, but a better one which is computed quickly offers more benefits for the traffic department and ultimately the drivers in the city.

Model Entity	Instantiations
Organization	Traffic dept. New York
Business Process	Traffic control
Data	Vehicular data
Data Volume	Very High
Data Structure	Semi-Structured
Big Data Scale	Yes
Data Origin	Cars sent their data
Metadata	Car no., Time of transmission
Data Storage	Data center located remotely
Application	Finding possible better traffic light configurations
Application Results	Possible improved light configurations
Application of Results	Deciding to change traffic light configurations
Value Creation	Better flow of traffic in the city

Table 6.3: Example 3

When looking at this example, we can immediately identify it as a big data-scale applications because of the very large data volume. The data has some structure in it, but because of the large volume this scenario is clearly about big data. This brings with it several challenges that we do not see in Example 1 or Example 2.

One of those challenges is the performance of the process. Because so much data needs to be transmitted in near real-time, efficiency is of the essence. Not only does the data application need to process large amounts of data in very little time, the data also needs to be transmitted from the vehicle to the data center and from the data center to the traffic department very rapidly. So not only an efficient algorithm is needed, also a fast connection is necessary. If the process takes too much time, the results are no longer useful as the traffic configuration will have changed.

Compared to Examples 1 and 2, the infrastructure and complexity components are very important because this application can only be of added value if it can operate in real-time. If this is not possible for any reason, the traffic department could decide to use the gathered data to run scenarios of traffic light configurations to gain more optimal configurations that way. By doing this, they naturally lose the ability to deal with real time peaks and congestion in traffic.

6.1 Reflection on validation

The example stories and landscapes presented in the previous section present how different data landscapes can be represented in the (meta-)model we have created. Of course, because there are only three examples, they are constructed in such a way that they all highlight different parts of the data landscape. Looking back at chapters 1 and 4, we have stated several questions that our model should address in order to be fitting to help professionals with the issues they face. These include the origin of data used in organizations, where and how it is stored, in which processes it is used and for which goal. We have attempted to incorporate the earlier stated problems and questions in our model in a way that both helps professionals answer these questions but also *forces* them to answer them. We have achieved this by translating the issues faced by professionals into model entities. One of the most important issues we have extracted from the literature is the distinction between facing Big Data-scale problems in organizations or not. This scale of data volume and variety requires its own approach and algorithms, something we see in Example 3. This example presents problems regarding efficiency and data transmission speed, something we do not see in Example 1 and 2 as those applications are of a much smaller scale.

If modelers model the organization and its data landscape, they are required to for instance describe the origin of their data or the application of the data application results. Such a structure will by itself help ask the right questions about the data landscape. In Example 2 we see that the data is stored in a data warehouse at the department store chain headquarters, whereas in Example 1 the data is stored locally on servers at the stores. The data warehouse offers many benefits to the organization, so it might be useful for managers of the store in Example 1 to consider using a data warehouse solution where possible to streamline the data process.

When it comes to answering questions about data landscapes, there are two sides to it. The first is that a meta model in itself cannot answer any concrete questions about the contents of a populated model or the real-world situation it resembles. The model merely offers a structure in which the relevant information can be captured. On the other hand, an already constructed and populated landscape model can help greatly in answering the aforementioned questions because it presents a structured and condensed overview of the relevant information. Especially when communicating this information or when persons other than the modeler want to answer questions such as those stated in the earlier chapters, our model can be of great use. We see this in Example 1, where a floor planner is not involved in the data process but he can put the data to use to create a optimized floor plan. Concluding, we see that the example stories highlight the possible differences in data landscapes, how they are to be modeled within the framework of our landscape model and what can be learned from these model representations.

Chapter 7

Discussion and Limitations

In this thesis research we have created and presented a possible model in which organizations can express their challenges and solutions with regards to gathering value out of data. First, we have identified the challenges faced by professionals active in data management. This is done by means of the Data Management Body Of Knowledge (DMBOK) by DAMA, where possible and/or necessary supported by other scientific literature. The overview of issues faced by professionals is our first product. Based on this list, we have constructed our model. Although the DMBOK attempts to be an exhaustive knowledge base by and for professionals, this is of course not 100 percent possible. That is why we have supported and extended the results from DAMA with additional scientific literature. However, it might be that there exist some challenges within data management which have not yet made it into the scientific literature. We have tried to address this as much as possible by also consulting regular media and media aimed at data professionals. In our review of these media, we have not found additional challenges which needed addressing within our model.

From the overview of challenges, we have abstracted to the entities which our model requires. Because our model attempts to be a complete overview of the data landscape on a high level, there have been some difficult trade offs between abstracting too far from the issue and becoming too specific in certain areas of the model. An example would be the inclusion of data security into the model as an entity or splitting the entity Application into different types of applications. We have chosen to always remain at the meta-level in our meta model in order to stay consequent. In the conceptual model we see how concepts from the meta level and lower levels are linked and can be modeled. This offers the modeler who instantiates models of our kind to fill in the entities in such a way that matches the real-world scenario.

Validation of the meta model we have constructed is done by means of example stories. We would like to have put our model into the field and let it be tested by professionals, but due to the limited time and resources available for this thesis,

this was not possible. With the example stories we have given several examples from practice that each in their own way offer an instantiation of our model that fills in the entities in a different, yet relevant way. The example stories are by no means exhaustive and there might be real-world situations that are at this time not capturable within our model. Because of time constraints, it remains an opportunity for the future (see Chapter 8.1 for more on this) to see if such scenarios exist and to improve the model such that these situations can be modeled.

General limitations

Our model has some general limitations that we will go through in this section.

- The model might be too abstract to be of most effect for professionals within organization. Many organizations have their own specific configuration and ways of working which can be quite complex. In our model, this is all put under the entity "Business Process" for example, which might be too abstract to really be of use for a modeler that wishes to use our model to gain better insight in how his or her organization can gain more value through data.
- Outside influences in the field, for instance changing legislature can rapidly change the data landscape. These changes might be of such a nature that our model does not longer cover the entire data landscape. We have developed our model in the abstract nature it has to be as robust and future proof as possible, but the developments might be of such a nature that our model no longer is applicable to the entire data landscape.
- The model we have created offers an added overview of the data landscape, but does not propose solutions and/or best practices to professionals to actually help solve (big) data problems they face within their organization. Our model offers insight into the data landscape to professionals and we hope that gained information helps professionals in tackling the (implementation) challenges within their organization.

Chapter 8

Conclusions and Future Work

In the following chapter we will go through answering the research problem, presenting the conclusions drawn from our Design Science approach to developing a meta model for the data landscape. We will also cover the possibilities for future extensions and additions to this model.

While performing the design of the data landscape meta model, we have attempted to find out how we can create a model such that it helps organizations express their challenges and solutions with regards to gathering value out of data. We have found that the first step in this process is to identify these challenges and solutions before translating them into model entities and finding/mapping the connections between them. The final results of this process are the meta and conceptual models presented in Chapter 5. We have performed a validation in the form of example stories which show that real-world data application-scenarios can be mapped into our model.

We have created this model because many organizations and professionals struggle with keeping oversight over how data is put to use within their organization and subsequently do not unlock the full potential of the data they posses. No earlier concrete attempts have been made to bring all aspects of data processes into a single, concise overview. By means of this thesis we present our version of such a model created through a Design Science process. In our opinion this was a well-fitting methodology to our goal. Not only does the creation of a model fit well to the notion of designing and creating an artifact, the Design Science approach also gave us the opportunity to start with a prototype model and later find a justification for it in literature. This was exactly how the research process of this thesis was formed, as we started with the idea to make a model of the entire data landscape and only later started consulting the literature for possible problems encountered by professionals which needed to be included in our model.

Concluding, the chosen methodology fit well to the challenge we address in this research. We have identified challenges professionals face in data management and

have created a meta model for the data landscape which can help tackle these issues by offering oversight over the data process, from the creation of the data to the application of it in business processes. Due to the Design Science methodology used in this research, there is no theoretical conclusion; the models we have created are our end products.

8.1 Future Work

In this section we will briefly go through several possible directions to continue with or expand upon this research, which presented themselves during the research process.

- The largest future opportunity for our model is a more thorough validation process. It would be interesting to see if a field test of our model in practice would produce unforeseen results or would present needed changes or additions to our model.
- In Section 5.5 we have presented several entities which we explicitly not have included in the meta model. It can be an interesting future direction to see if those entities or perhaps other entities are relevant in other ways and if they can be added to the model. This could be as concrete entities or as concerns to keep in mind throughout the entire process such as data security and ethics.
- Some entities in the model are deliberately left quite abstract in order to let them encompass all different possible instantiations that present themselves in real world scenarios. The Data Ownership-entity however has been defined such that the agent responsible for utilizing the data is the owner. This is quite a rough abstraction which is not entirely true for personal data anymore (for instance with the GDPR in place in the European Union). This presents the opportunity to find a definition or entity which can express the differences in ownership between personal data and non-personal data.
- An aspect of the data landscape model that we were unable to complete within the scope of this thesis was the visualization of specific data landscapes. In our solution we have chosen to represent the data landscapes in Chapter 6 in terms of tables containing all relevant information. This is a rather functional solution which is not very visually pleasing. It would be a good follow-up to find out how data landscapes such as those from Chapter 6 can be visually represented.

Bibliography

[1] Russell L Ackoff. From data to wisdom. *Journal of applied systems analysis*, 16(1):3–9, 1989.

[2] Ali M Al-Khouri et al. Data ownership: who owns "my data". *Int. J. Manag. Inf. Technol*, 2(1):1–8, 2012.

[3] Omar Alam, Jörg Kienzle, and Gunter Mussbacher. Concern-oriented software design. In *International Conference on Model Driven Engineering Languages and Systems*, pages 604–621. Springer, 2013.

[4] Ledion Bitincka, Archana Ganapathi, Stephen Sorkin, Steve Zhang, et al. Optimizing data analysis with a semi-structured time series database. *SLAML*, 10:7–7, 2010.

[5] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 2015.

[6] Diego Calvanese, Maurizio Lenzerini, and Daniele Nardi. Description logics for conceptual data modeling. In *Logics for databases and information systems*, pages 229–263. Springer, 1998.

[7] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and olap technology. *ACM Sigmod record*, 26(1):65–74, 1997.

[8] Jinchuan Chen, Yueguo Chen, Xiaoyong Du, Cuiping Li, Jiaheng Lu, Suyun Zhao, and Xuan Zhou. Big data challenge: a data management perspective. *Frontiers of Computer Science*, 7(2):157–164, 2013.

[9] William S Cleveland. Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1):21–26, 2001.

[10] Thomas H Davenport et al. Competing on analytics. *harvard business review*, 84(1):98, 2006.

[11] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[12] Nina Evans, Louis Fourie, and James Price. Barriers to the effective deployment of information assets: The role of the executive manager. In *Proceedings of the European Conference on Management, Leadership & Governance*, volume 7, pages 162–169, 2012.

[13] ED Falkenberg, W Hesse, P Lindgreen, BE Nilsson, JLH Oei, C Rolland, RK Stamper, FJM Van Assche, AA Verrijn-Stuart, and K Voss. A framework of information systems concepts (the frisco report). *International Federation for Information Processing, Geneva*, 1998.

[14] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[15] Luciano Floridi and Mariarosaria Taddeo. What is data ethics?, 2016.

[16] Keith D. Foote. A brief history of the data warehouse, 2018. `https://www.dataversity.net/brief-history-data-warehouse`.

[17] Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. Springer, 2015.

[18] Matteo Golfarelli, Stefano Rizzi, and Iuris Cella. Beyond data warehousing: what's next in business intelligence? In *Proceedings of the 7th ACM international workshop on Data warehousing and OLAP*, pages 1–6. ACM, 2004.

[19] Terry Halpin and Anthony Bloesch. Data modeling in uml and orm: a comparison. *Journal of Database Management (JDM)*, 10(4):4–13, 1999.

[20] William Harrison, Harold Ossher, Stanley Sutton, and Peri Tarr. Concern modeling in the concern manipulation environment. In *ACM SIGSOFT Software Engineering Notes*, number 4, pages 1–5. ACM, 2005.

[21] Hossein Hassani, Xu Huang, and Mansi Ghodsi. Big data and causality. *Annals of Data Science*, 5(2):133–156, 2018.

[22] DAMA International. *DAMA Data Management Body of Knowledge*, volume 2. Technics Publications, 2017.

[23] DA Jardine. Concepts and terminology for the conceptual schema and the information base. *Computers and Standards*, 3(1):3–17, 1984.

[24] Xiaolong Jin, Benjamin W Wah, Xueqi Cheng, and Yuanzhuo Wang. Significance and challenges of big data research. *Big Data Research*, 2(2):59–64, 2015.

[25] Stephen Kaisler, Frank Armour, J Alberto Espinosa, and William Money. Big data: Issues and challenges moving forward. In *2013 46th Hawaii International Conference on System Sciences*, pages 995–1004. IEEE, 2013.

[26] Avita Katal, Mohammad Wazid, and RH Goudar. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, pages 404–409. IEEE, 2013.

[27] Anil Kumar and Prashant Palvia. Key data management issues in a global executive information system. *Industrial Management & Data Systems*, 101(4):153–164, 2001.

[28] Marc Lankhorst et al. *Enterprise architecture at work*, volume 352. Springer, 2009.

[29] Deanne Larson and Victor Chang. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5):700–710, 2016.

[30] Margaret D LeCompte. Analyzing qualitative data. *Theory into practice*, 39(3):146–154, 2000.

[31] Mong Li Lee, Hongjun Lu, Tok Wang Ling, and Yee Teng Ko. Cleansing data for mining and warehousing. In *International Conference on Database and Expert Systems Applications*, pages 751–760. Springer, 1999.

[32] Dekang Lin and Patrick Pantel. Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328. ACM, 2001.

[33] Robert M Losee. Browsing mixed structured and unstructured data. *Information processing & management*, 42(2):440–452, 2006.

[34] James Macgregor, Eric Lee, and Newman Lam. Optimizing the structure of database menu indexes: A decision model of menu search. *Human Factors*, 28(4):387–399, 1986.

[35] Sam Madden. From databases to big data. *IEEE Internet Computing*, 16(3):4–6, 2012.

[36] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data: the management revolution. *Harvard business review*, 90(10):60–68, 2012.

[37] Anne-Marie Oostveen and Peter Van den Besselaar. Linking databases and linking cultures. In *Towards the E-Society*, pages 765–774. Springer, 2001.

[38] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[39] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.

[40] Caroline Sporleder, Marieke Van Erp, Tijn Porcelijn, and Antal Van Den Bosch. Spotting the 'odd-one-out': Data-driven error detection and correction in textual databases. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, 2006.

[41] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[42] John W Tukey and Martin B Wilk. Data analysis and statistics: an expository overview. In *Proceedings of the November 7-10, 1966, fall joint computer conference*, pages 695–709. ACM, 1966.

[43] Jeffrey D Ullman. *Principles of database and knowledge-base systems*, volume 1. Computer Science Press, Incorporated, 1988.

[44] Patrick van Bommel, Arthur HM ter Hofstede, and Th P van der Weide. Semantics and verification of object-role models. *Information Systems*, 16(5):471–495, 1991.

[45] Hai Wang and Shouhong Wang. A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems*, 108(5):622–634, 2008.

[46] Jennifer Widom. Research problems in data warehousing. 1995.

[47] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, page 38. Citeseer, 2014.

[48] Gang Zhou, Richard Hull, Roger King, and Jean-Claude Franchitti. Data integration and warehousing using h2o. *IEEE Data Eng. Bull.*, 18(2):29–40, 1995.