Developing a Maturity Model for AI- Augmented Data Management

Author:
D. R. Defize

Organization:
University of Twente, Faculty of EEMCS, Master Business Information Technology

Committee:
Prof. Dr. Jos van Hillegersberg (Faculty of BMS, Universiteit Twente)
Dr. M. Daneva (Faculty of EEMCS, Universiteit Twente)
N. Vermeer (Manager, Deloitte)
C. Jacobs (Consultant, Deloitte)

UNIVERSITY OF TWENTE.

Deloitte.

Executive Summary

Data management is becoming more complicated due to the increase in data volume, variety, and velocity. In turn, the increase in time-consuming data management work is exponential, which means that it is now impossible to do it all manual. Augmented data management has the potential to overcome organizations' data management challenges by leveraging artificial intelligence to automate and enhance data management tasks and decisions. Meanwhile, organizations struggle to manage their data successfully. Maturity models are a proven approach to systematically assess and improve organizational capabilities towards achieving an organizational goal. In the context of managing data in organizations, a maturity model may help them navigate through the improvement options available and assess their relevance for the organization's goals. The objective of this master project is to develop a maturity model for augmented data management. To this end, the research in the present thesis adopted a Design Science grounded research process and, in turn, underwent three phases:

The first phase provides the scientific background through a systematic literature review on artificial intelligence, data management, and maturity model development. The results of this phase include: (i) an overview of the subfields of artificial intelligence and their applications, (ii) an overview of all data management and artificial intelligence maturity models available in current literature, and (iii) an overview of methods, methodologies, and guidelines on developing maturity models.

The second phase includes the initial design and development of the maturity model. To this end, the design choice is made to leverage the foundation of existing maturity models and build upon those with empirical research to develop a novel model. The development strategy used complementarily research techniques of three types: metamodel analysis, expert interviews, and market research. The metamodel analysis is used to systematically compare and synthesize existing maturity models. Through interviewing experts on artificial intelligence and data management, it is identified which data management processes can be augmented. The market research complemented this view by analyzing tools that provide these functionalities.

In the third phase, the initial model is evaluated and refined through a mixed-method validation approach. It includes experts' perception-based evaluation and case studies. The maturity model is operationalized by creating an Excel assessment tool that can be used to structure the assessment and assess the (sub) capabilities and processes. The model and assessment tool are evaluated with data management consultants, the expected users of the model. The case studies were conducted with the primary functional beneficiary of the model: organizations that want to improve their (augmented) data management practices. Based on the findings of the mixed-method validation, it is concluded that (1) the resulting Augmented Data Management Maturity Model (ADM^3) consists of sufficient and accurate maturity levels, (2) the processes and capabilities are relevant, comprehensive, mutually exclusive and accurate and (3) the model itself is understandable, easy to use, useful and practical. It can also be concluded that the recommendations on improving capabilities are understandable, easy to use, and useful. The recommendations on constructing a roadmap are understandable and easy to use.

The model consists of five capabilities: data quality, metadata management, data integration, master data management, and database management. Based on literature and expert interviews, these capabilities are essential to data management and are expected to have the largest impact by augmentation based on the amount of data and manual work involved. Each capability consists of

multiple sub capabilities and processes. The proposed ADM^3 consists of two maturity scales: one for data management and one for augmented data management. Because of the data management scale, the maturity model seamlessly complements existing data management maturity assessments while introducing a novel maturity scale for augmented data management.

The main strength of this research is the introduction of the novel ADM Maturity Model. The model fulfills all functional and non-functional requirements and can be operationalized using the assessment tool to assess and improve data management capabilities by leveraging AI. The main strength of the used research process is the combination of established methods for designing and validating the model, which combine current literature and empirical research from practice.

To conclude, the contribution of this research is fourfold:

1. Scientific – by presenting and demonstrating a maturity model development approach that combines existing frameworks and methodologies
2. Scientific – by designing and introducing the first maturity model for augmented data management.
3. Business – by introducing a maturity model and assessment tool that can be used to assess current (augmented) data management capabilities and improvement opportunities.
4. Business – by providing an evaluation with practitioners, data management consultants and organizations, which indicated that the proposed maturity model and assessment tool are promising.

Future work can improve the current limitations of the model. The model could be made more objective by using qualitative maturity measures. The protocol for selecting assessment participants should be improved to make it more multidisciplinary, so it covers all capabilities. Capabilities can be added, such as data governance, to make the model more comprehensive. The maturity model should be further validated, preferably during action research at an organization that wants to implement or improve its augmented data management. Future work should lead to revising the model every couple of years, as artificial intelligence and data management are fast-changing fields. The ADM^3 equips organizations with a framework that enables them to coordinate and synchronize their short-term and long-term improvement efforts concerning augmented data management.

Preface

Before you lies the master thesis 'Developing a Maturity Model for AI-Augmented Data Management'. It has been written to fulfill the graduation requirements for the master Business Information Technology at the University of Twente. The present research has been carried out in collaboration with Deloitte Netherlands from March to October 2020.

The topic of the thesis is augmented data management, which is the application of artificial intelligence (AI) to enhance data management capabilities. While data management is essential to create valuable insights from data, it is increasingly difficult due to the vast amount and complexity of collected data. Augmented data management is named one of the top trends to overcome these challenges. Despite the potential, little is known about the implementation and improvement of augmented data management. This research set out to create a maturity model that enables an organization to assess and improve current augmented data management capabilities.

I always had an interest in the impact that technology can make. After exploring the technology domain, I discovered that I was more interested in the human, social, and business aspects of IT. Technology is only as good as its user, so how do we leverage it to its full potential? This thought immediately came to mind when I read about augmented data management. I saw an opportunity to contribute to the adoption of a technology that still has to reach its full potential. The result is the Augmented Data Management Maturity Model (ADM^3).

The research process took place during a disruptive time. After barely two weeks, the whole country went in lockdown due to the rise of COVID-19. This was both a challenging and interesting time. Challenging, because of all the social restrictions and inability to go to the office. Interesting, because it imposed a new way of working where digital solutions and data are even more important. I especially enjoyed the empirical part of the research process. I was able to interview some of the best practitioners within the industry, and during the case studies I got a glimpse of what it is like to be one myself. While condensing the research into a concise and clear thesis was a challenge, I especially enjoyed how everything came together at the end.

From the University of Twente, I would like to thank Jos van Hillegersberg and Maya Daneva for their guidance as supervisors in this research. Through their critical feedback and interesting discussions, I was able to create a thesis that reflects my high ambitions.

I would like to thank Deloitte for the opportunity and a special thanks to Cas Jacobs and Niko Vermeer for their valuable support while working there. Through their experience, I learned a great deal about the data management practice and made my research relevant for the industry. I would also like to thank everyone within Enterprise Architecture and the EDM team for the challenging and pleasant (digital) working atmosphere. Finally, I wish to thank all of the interview and case study participants; without their cooperation, I would not have been able to perform this research.

I invite you to read the thesis, and I hope you enjoy reading it.

Dico Defize

Contents

Executive Summary ... *ii*
Preface ... *iv*
Contents ... *v*
List of Figures ... *viii*
List of Tables ... *ix*
List of Abbreviations ... *x*
1. Introduction ... *1*
1.1 Problem Statement ... **1**
1.1.1 Data Management ... 1
1.1.2 Limitations of Current Practices ... 1
1.1.3 Augmented Data Management ... 2
1.1.4 Maturity Models ... 3
1.1.5 Deloitte Enterprise Data Management ... 3
1.2 Research Goals and Requirements ... **4**
1.2.1 Design Science Research ... 4
1.2.2 Stakeholders and Goals ... 4
1.2.3 Relevance and Demand ... 4
1.2.4 Requirements ... 5
1.3 Research Questions ... **6**
1.4 Thesis Outline ... **7**
2. Theoretical Background ... *8*
2.1 Artificial Intelligence ... **8**
2.1.1 Machine Learning ... 8
2.1.2 Natural Language Processing ... 10
2.1.3 Expert Systems ... 10
2.1.4 Vision Recognition ... 10
2.1.5 Speech Recognition ... 10
2.1.6 Planning ... 11
2.1.7 Robotics ... 11
2.2 Systematic Literature Review ... **11**
2.2.1 Research Questions ... 12
2.2.2 Data Sources and Search Strategy ... 12
2.2.3 Data Extraction and Synthesis ... 13
2.3 Data Management Maturity Models ... **14**
2.4 Artificial Intelligence Maturity Models ... **18**
2.5 Maturity Model Development ... **20**
2.5.1 Maturity Model Types ... 20
2.5.2 Methodologies ... 21
2.5.3 Research Methods ... 22
2.5.4 Guidelines for Developing Maturity Models ... 22
3. Design and Development ... *25*
3.1 Mixed-Method Development Strategy ... **25**
3.1.1 Combining Existing Models ... 26
3.1.2 Metamodel Approach ... 26
3.1.3 Systematic Metamodel Comparison ... 29
3.1.4 Expert Interviews and Market Research ... 30

3.1.5 Qualitative Data Analysis 30
3.1.6 Constructing the ADM Maturity Model 31
3.2 ADM Maturity Model Version 1.0 32
3.2.1 Synthesizing Maturity Levels 32
3.2.2 ADM Maturity Model Levels 33
3.2.3 Selecting Capabilities 33
3.2.4 Synthesizing Capabilities 34
3.2.5 Selecting Sub Capabilities and Processes 39
3.2.6 Expert Interviews 39
3.2.7 Market Research 39
3.3 ADM Maturity Model Capabilities 41
3.4 Result of the Development Phase 43
4. Evaluation and Refinement 45
4.1 Mixed-Method Validation Strategy 45
4.2 Expert Interviews 46
4.2.1 Interview Participants 46
4.2.2 Interview Protocol 47
4.2.3 Qualitative Data Analysis 48
4.2.4 Evaluation Criteria 48
4.3 ADM Maturity Model Version 1.1 49
4.3.1 Introduction 49
4.3.2 Maturity Levels 50
4.3.3 Data Quality 51
4.3.4 Metadata Management 51
4.3.5 Data Integration 52
4.3.6 Master Data Management 53
4.3.7 Database Management 53
4.3.8 Universal Capabilities 54
4.3.9 Results 54
4.3.10 Improving Capabilities 55
4.3.11 Improvement Roadmap 56
4.3.12 Other Expert Feedback 57
4.4 Case studies 58
4.4.1 Case Study Participants 58
4.4.2 Case Study Protocol 58
4.4.3 Case 1: Health Insurer 59
4.4.4 Case 2: Bank 60
4.4.5 Case 3: Insurer 60
4.4.6 Evaluation of Maturity Model and Assessment Tool 61
4.4.7 Evaluation of Recommendations 62
5. Conclusion 63
5.1 Augmented Data Management 63
5.2 Existing Maturity Models 63
5.3 Maturity Model Development 64
5.4 ADM Maturity Model 65
5.5 Main Research Question 65
5.6 Contribution to Practice 66
5.7 Contribution to Research 66
6. Discussion 68
6.1 Reflection on the Chosen Research Methodology 68

6.2 ADM Maturity Model Reflection 69
6.3 Implications for Practice 71
6.4 Implications for Research 71
6.5 Research Limitations and Future Work 72
7. Bibliography 74
8. Appendix 80
A. Description of Data Management Literature 80
B. Systematic Comparison 81
C. Transcripts of Expert Interviews 86
D. Market Research Extended 99
E. Transcript of Expert Evaluation 104
F. Changes After Expert Evaluation 135
G. Transcript of Case Study Evaluation 138
H. Result of Case Studies Maturity Assessment 150

List of Figures

Figure 1: Visualization of the Human and Machine Intelligence Field 2
Figure 2: Research Framework 7
Figure 3: Overview of the AI field, Adapted From [27] 8
Figure 4: Research Methodology for the Systematic Literature Review 12
Figure 5: CMMI Maturity Levels Definition, Based on [51] 15
Figure 6: Model Capability Mapping, Adapted from [65] 17
Figure 7: Three Types of Maturity Models: Staged Fixed-Level (a), Continuous Fixed-Level (b), Focus Area (c), source [71] 20
Figure 8: Procedure model for guidelines based on Becker et al. [15] 25
Figure 9: Metamodels for AI Maturity Models 28
Figure 10: Metamodels of Data Management Maturity Models 28
Figure 11: Visualization of the Comparison Table and Systematic Metamodel Comparison 29
Figure 12: Construction of ADM Capabilities Simplified 31
Figure 13: ADM Maturity Axis 33
Figure 14: Maturity Assessment Tool v1.0 Tab for Data Quality 44
Figure 15: Maturity Assessment Tool v1.0 Results Tab 44
Figure 16: Evaluation Episodes, Based on [96] 45
Figure 17: Introduction Tab of ADM Maturity Assessment Tool v1.1 49
Figure 18: Maturity Levels and ADM Definition 50
Figure 19: ADM Maturity Assessment Tool 1.0 Results Tab 55
Figure 20: Example Timeline for Implementing Master Data Management, source [64] 57

List of Tables

Table 1: Stakeholders and Goals 5
Table 2: Evaluation Criteria as Design Goals [24] 6
Table 3: DM Models and Corresponding References 14
Table 4: Synthesis of the Analyzed Maturity Models Regarding Model Structure 14
Table 5: Synthesis of the analyzed maturity models regarding model assessment 15
Table 6: Synthesis of the analyzed maturity models regarding model support 16
Table 7: AI models and corresponding references 18
Table 8: Synthesis of the AI maturity models regarding model structure 18
Table 9: Synthesis of the AI maturity models regarding model assessment 18
Table 10: Synthesis of the AI maturity models regarding model support 19
Table 11: Model attribute mapping 19
Table 12: Model maturity level mapping 20
Table 13: Development guidelines overview 24
Table 14: Procedure Model Steps and Corresponding Sections 26
Table 15: Maturity levels of AI models 32
Table 16: Maturity levels of DM model 32
Table 17: Reference Table for Constructing ADM Capabilities 39
Table 18: ADM^3 v1.0 Data Quality Overview 41
Table 19: ADM^3 v1.0 Metadata Management Overview 42
Table 20: ADM^3 v1.0 Data Integration Overview 42
Table 21: ADM^3 v1.0 Master Data Management Overview 42
Table 22: ADM^3 v1.0 Database Management Overview 43
Table 23: Interview Participants 46
Table 24: Evaluation Criteria Scores from Interviews (N=11) 48
Table 25: ADM^3 v1.1 Data Quality Overview and Changes 51
Table 26 ADM^3 v1.1 Metadata Management Overview and Changes 52
Table 27: ADM^3 v1.1 Data Integration Overview and Changes 52
Table 28: ADM^3 v1.1 Master Data Management Overview and Changes 53
Table 29: ADM^{3}v1.1 Database Management Overview and Changes 53
Table 30: ADM^3 v1.1 Universal Capabilities Overview 54
Table 31: Overview of All Case Study Meetings 58
Table 32: Assessment Results DM Maturity of Case 1 59
Table 33 Assessment Result ADM Maturity of Case 1 60
Table 34: Assessment Result of Case 2 60
Table 35: Assessment Result of Case 3 61
Table 36: Evaluation Criteria Scores from the Case Studies (N=4) 62
Table 37: Recommendation Evaluation Criteria Scores from the Case Studies (N=5) 62
Table 38: Overview of Data Management Maturity Model Literature 80

List of Abbreviations

Abbreviation	Meaning
ADM	Augmented Data Management
ADM^3	Augmented Data Management Maturity Model
AI	Artificial Intelligence
AIMM	Artificial Intelligence Maturity Model
AMM	Algorithmic Maturity Model
CDO	Chief Data Officer
CIO	Chief Information Officer
CMMI	Capability Maturity Model Integration
DAMA-DMBOK	Data Management Association - Data Management Body Of Knowledge
DBMS	Database Management
DCAM	Data Management Capability Assessment Model
DGMM	Data Governance Maturity Model
DI	Data Integration
DL	Deep Learning
DM	Data Management
DMBOK	Data Management Body Of Knowledge
DMM	Data Management Maturity
DQ	Data Quality
DSR	Design Science Research
EDM	Enterprise Data Management
GAIM	Gartner Artificial Intelligence Model
IT	Information Technology
KPI	Key Performance Indicator
MD	Metadata
MD3M	Master Data Maturity Model
MDM	Master Data Management
ML	Machine Learning
NLP	Natural Language Processing
OAIM	Ovum Artificial Intelligence Model
SLR	Systematic Literature Review

1. Introduction

This chapter introduces the research topic and research goals. Section 1.1 introduces the problem statement and the main concepts of augmented data management and maturity models. Section 1.2 presents the research goals and requirements. Section 1.3 presents the research questions, and Section 1.4 outlines the thesis.

1.1 Problem Statement

1.1.1 Data Management

Everyone is talking about data. Organizations want to collect as much as possible to become data-driven, while consumers are becoming more vocal about their data rights and privacy. Data is essential because when data is processed and used within the right context, it becomes information that can lead to valuable insights. Organizations can use this information within their business processes to improve services, reduce costs, gain additional profits, and manage risks [1]. To realize these benefits, many organizations strive to collect as much data as possible, as timely as possible, and as precise as possible.

Only collecting and analyzing data is not enough. Organizations can spend tremendous effort analyzing their data to discover that the data itself is flawed or unusable. Undefined and fragmented data leads to increased complexity, costs, errors, and inefficiency. Projects, especially involving data integration, data quality, and reporting, depend on the strength of the underlying data models [2]. Data quality becomes a precondition to realize value from data effectively and, therefore, increasingly gains importance within organizations [3]. Working with incomplete or incorrect data can lead to incorrect and unsubstantiated insights. Data management is needed to realize the full potential that data has. Data management is defined as the business function of planning for, controlling, and delivering data and information assets [4]. The business functions related to data management vary per organization and various data management models, such as the Data Management Body of Knowledge (DMBOK), aim to define these functions in multiple disciples or capabilities [4].

Systematically integrating data management strategies proves to be effective and decreases the costs associated with decision making [5]. Market analyst Garter estimated that by 2020, 10% of organizations have a highly profitable business unit specifically for productizing and commercializing their information assets [6]. Next to financial and operational benefits, data management supports regulatory compliance in facilitating data security and regulatory reporting [7]. These applications illustrate the different drivers for organizations to adopt data management practices: to comply with regulations, to mitigate risk, or to strive for operational efficiency.

1.1.2 Limitations of Current Practices

While the benefits and necessities of data management are driving its adoption, the current approaches for data management are at their limits. Architectures and tools are breaking down due to the size, complexity, and distributed nature of data [8]. The Cisco Annual Internet Report forecasts that by 2023 the world will have 29.3 billion internet-connected devices, up from 18.4 billion in 2018, generating dozens of zettabytes of data. Businesses account for 24% of these devices, and consumers will own the other 76% with an average of 3.6 connected devices per capita [9]. This trend is growing exponentially. In turn, the effort and complexity associated with processing this

amount of data is growing too. Merely adding more data engineers and data scientists is not enough to keep up.

Methods used by data scientists and engineers are time-intensive and hard to scale. Manual activities cannot keep up with the volume, velocity, and variety of data, especially within streaming data architectures. With an increase in data sources and data pipelines, the complexity of the data landscape is increasing. Understanding such complex structures is difficult, which makes system integration and impact analysis time-consuming. These challenges result in projects that are prone to error and prolong the time to market. In addition, there is a substantial lack of knowledge and high demand for data experts. The professionals who do have this knowledge spend much time on preparatory and manual work rather than directly producing valuable insights [8]. To conclude, data management is becoming increasingly complicated by an increase in scope, volume, and architectural variety, having an exponential increase in time-consuming data management work as a consequence.

These limitations are driving the adoption of artificial intelligence to complement human capabilities. AI-augmented data management has the potential to overcome the current limitations and is one of the top data-trends that will change business in the coming years [8].

1.1.3 Augmented Data Management

Augmented data management is the application of augmented intelligence to enhance data management capabilities. Augmented intelligence or intelligence augmentation is the human-centered conceptualization of artificial intelligence, emphasizing human intelligence enhancement with cognitive technology. The goal is to leverage AI capabilities to complement human intelligence in learning and decision making, rather than replacing it [10]. In short, augmented data management is defined as the human-centered application of artificial intelligence to enhance data management capabilities.

Artificial intelligence is defined as intelligent behavior in artifacts that we associate with human thinking [11]. One of the subfields of artificial intelligence is machine learning, which refers to computer systems that use algorithms and statistical models to perform a task without explicit instructions [12]. Machine learning aims to mimic human-like learning to perform tasks and make decisions. Figure 1 represents how these fields relate to each other.

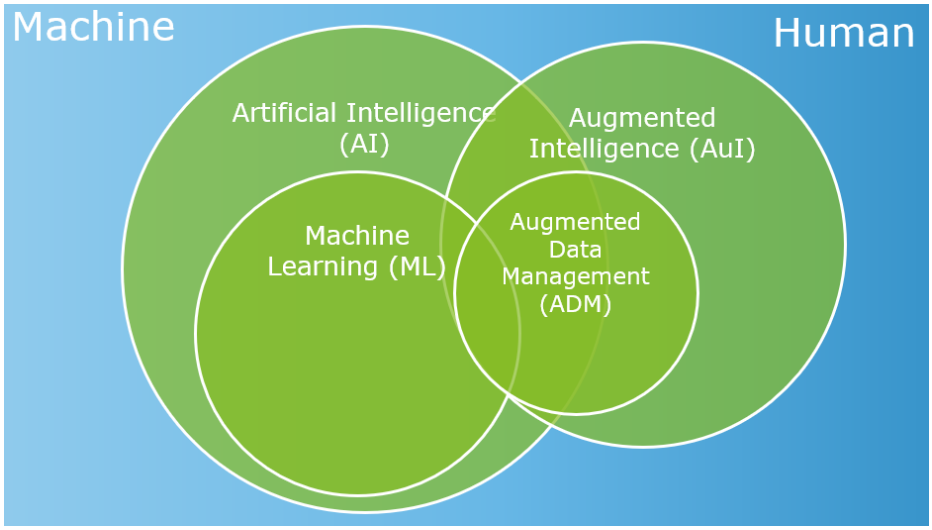


Figure 1: Visualization of the Human and Machine Intelligence Field

Augmentation is predicted to have an enormous impact on data management. Gartner predicts that by 2022 manual tasks will be reduced by 45% through the addition of artificial intelligence and automated service-level management [13]. In total, AI augmentation will create $2.9 trillion of business value and 6.2 billion hours of worker productivity in 2021 [8]. By automating and enhancing manual tasks, the acute talent shortage is eased, and experts can focus on more valuable tasks.

1.1.4 Maturity Models

To manage data effectively, organizations must recognize data as a tangible asset and manage it through data management [4]. Based on a survey by NewVantage among 70 worldwide leading organizations on data and AI, only 28% of CDOs are considered successful [14]. Gartner estimated this percentage to be around 50% [6]. These numbers reveal the need for practical and suitable data management approaches. Organizations are required to assess their current capabilities to continuously improve their data management. [15]. Maturity models are helpful tools to assess current capabilities in order to derive improvement measures[16] [17]. Maturity models provide a framework for assessing an organization's capabilities, strengths, and weaknesses, comparing processes between organizations, and identifying relations between maturity and business performance [18]. Generally, these models consist of multiple maturity levels, which correspond to the key maturity stages in the underlying capability within a functional domain. Based on the definition and description of these maturity levels, organizations can use the maturity model as a tool to assess current capabilities and identify incremental process improvements in relevant capabilities [19].

Augmented data management presents itself as a technical solution to the data challenges that organizations face today. Meanwhile, there appears to be a lack of systematic methods to implement and improve (augmented) data management successfully. What is takes organizations to make the success of employing augmented data management more predictable is hardly known. Leveraging a maturity model is a promising and proven approach to address both technical and managerial challenges faced in data management today by focusing on capabilities within the organization. Equipping organizations with a framework that allows them to assess where they stand and where they want to go concerning augmented data management will help them coordinate and synchronize their short-term and long-term improvement efforts. Therefore, the present research is set out to develop a maturity model for augmented data management.

1.1.5 Deloitte Enterprise Data Management

Deloitte is one of the largest technology consulting firms in the Netherlands and worldwide. The Enterprise Architecture service line consults client organizations to align business processes with information, applications, and integration technology. Within this service line, the Enterprise Data Management (EDM) team develops the strategy and essential capabilities needed to successfully manage and get value from data assets. One approach to achieving this is by performing a maturity assessment and using the results to construct a roadmap to improve data management capabilities. To realize this, the Data Management Body of Knowledge (DAMA-DMBOK) functional framework is, for example, used as a reference. Deloitte is continuously looking to leverage new technologies to help clients to make an impact. In collaboration with Deloitte, this research is expected to generate insight into the current and future of augmented data management. The maturity model developed in the present research is expected to provide Deloitte with a tool to implement augmented data management in maturity assessments.

1.2 Research Goals and Requirements

1.2.1 Relevance and Demand

In order to confirm relevance and demand in practice, a series of interviews were conducted. These interviews are additionally used to develop the model and are detailed in Section 3.1.4. The participants were asked whether a maturity model for augmented data management would be relevant and helpful. All seven respondents indicated high relevance and demand for the development of such a maturity model. Six out of seven indicated that they utilize data management maturity models as a useful and essential framework for improving data management capabilities. Furthermore, they confirmed that AI has an enormous potential in augmenting those capabilities, and organizations are currently not leveraging AI while there are options available. Current data management maturity models do not incorporate AI-augmented capabilities. The model is expected to serve as an instrument for assessing current capabilities and as a guideline to create a roadmap advance augmented data management.

1.2.2 Design Science Research

As indicated earlier, this research aims to create a maturity model for augmented data management. To realize this goal, the Design Science Research Methodology is used. Design science is the design and investigation of artifacts in their context of use [20]. The artifact interacts with the context in order to solve a design problem in that context. Within this research, the artifact is the maturity model, and the stakeholders are actors affected by the model. The problem context is already introduced in Section 1.1. An artifact that addresses this problem context can have many different designs, yet the usability is evaluated by the stakeholder goals. Therefore, this section introduces the social context with the stakeholders and their goals and corresponding requirements. The whole thesis is outlined in Section 1.4 and the multi-method development strategy is presented in Section 3.1.

1.2.3 Stakeholders and Goals

The domain of the model is enterprise data management, more specifically the application of data management maturity models. The main stakeholders are directly involved with the maturity model: data management consultants and organizations that want to improve their data management capabilities. The maturity model for augmented data management is intended to be used by the data management consultant to assess the capabilities of the organization. Within that context, data management consultants are the *intended users* or *normal operators* (according to the classification of Alexander [21]), as they directly interact with the maturity model. The participating organizations are *functional beneficiaries*; they interact with the data management consultant to conduct the assessment and benefit from the result. These two stakeholder groups directly interact with the maturity model or are in the immediate environment, and therefore the usefulness must be evaluated with respect to their goals [20].

Next to the main stakeholders, there are various stakeholders involved in the development of the maturity model. Deloitte EDM is the sponsor of the research. The University of Twente is the supplier of knowledge. Domain experts that participated in interviews served as consultants in the development. The author is the developer of the maturity model. Table 1 summarizes the stakeholders, their types, and their goals.

Maturity models in information systems are being applied as an informed approach for continuous improvement and benchmarking [22]. The model aims to assess and improve data management

capabilities by leveraging AI-augmentation without being an AI expert. AI-augmentation can be leveraged in one-off projects to reduce the workload of data management consultants or can be incorporated in continuous data management processes at a client organization.

Stakeholder	**Type** (Classification of Alexander [21])	**Goal**
Data Management Consultants	Normal Operators	Leverage maturity model to perform maturity assessment
Organizations Seeking Improvement in Data Management Capabilities	Functional Beneficiaries	Improve data management capabilities through maturity assessment
Deloitte EDM	Sponsor	Develop tools and capabilities to help clients successfully manage and get value from data assets
University of Twente	Supplier of Knowledge	Contribute to research and practice
Domain Experts	Consultant	Share knowledge within the domain or organization
Author	Developer	Develop a maturity model that fulfills the goals and requirements of the main stakeholders

Table 1: Stakeholders and Goals

1.2.4 Requirements

As presented in Section 1.2.3 the goals describe the desires of each stakeholder regarding the maturity model. The properties of the maturity model are detailed in the requirements. Therefore, the requirements must be fulfilled in order to realize the goals of the stakeholders. Consequently, the resulting model maturity model is evaluated with regard to these requirements. Functional requirements are a prerequisite for the desired function of the maturity model. Non-functional requirements, or quality properties, are global properties of the interaction between the maturity model and the [20]. The following two functional requirements for maturity models are derived from literature:

The requirements for the maturity model for augmented data management are derived from literature and expert interviews. The guidelines by Becker et al. [15] incorporate requirements for the development of maturity models. These requirements are supplemented with functional and design goals for the model itself. There are two functional requirements identified as relevant to enable continuous improvement:

1. **The maturity model must enable the assessment of the current state of capabilities:** what needs to be measured, how, what to compare it with, in order to assign the as-is situation to a specific degree of maturity. Furthermore, the assessment can be used for benchmarking within and between organizations if they utilize the same maturity model [15].
2. **The maturity model must enable the identification of improvement measures**: identify improvement potentials, deduce action measures, and their priority [15] [23].

The evaluation template for maturity models by Salah et al. [24] is used to identify non-functional requirements. This template combines requirements from various popular papers on maturity model development within design science research, such as Becker et al. [15], Mettler [25], De Bruin et al. [16] and Poppelbuss [26]. These requirements serve as design goals during the maturity model

development and are used as criteria during the evaluation. The design goals and evaluation criteria are presented in Table 2.

Criteria as Design Goals	Description
Sufficiency	The maturity levels are sufficient to represent all maturation stages of the domain
Accuracy	There is no overlap detected between descriptions of maturity levels, and processes can be assigned to every maturity level
Relevance	The processes are relevant to the domain
Comprehensiveness	Processes cover all aspects impacting/involved in the domain
Mutual Exclusion	Processes are clearly distinct
Understandability	The maturity levels, assessment guidelines, and documentation are understandable
Ease of Use	The scoring schema, assessment guidelines, and documentation are easy to use
Usefulness	The maturity model is useful for conducting maturity assessments
Practicality	The maturity model is practical for use in industry

Table 2: Evaluation Criteria as Design Goals [24]

1.3 Research Questions

The main research question to support the research goal is formulated as follows:

What constitutes a maturity model for Augmented Data Management that allows organizations to assess and improve their Data Management operations by leveraging AI?

To guide the research, the main research question is divided into the following sub-questions:

1. **How can Artificial Intelligence be leveraged to Augment Data Management capabilities?**
    a. What is Augmented Data Management?
    b. What is Artificial Intelligence?

2. **Which Data Management and artificial intelligence maturity models are available in current literature?**
    a. What does a Data Management model consist of, according to published literature?
    b. What does an Artificial Intelligence maturity model consist of, according to published literature?
    c. What are Data Management capabilities included in the reported models in the literature?

3. **How to design a maturity model for Augmented Data Management?**
    a. What are the maturity model's goals and requirements?
    b. Which method can be used to design and validate a maturity model?

4. **What constitutes the ADM maturity model?**
    a. Which maturity levels and definitions can be distinguished?
    b. Which capabilities can be distinguished?
    c. How to perform a maturity assessment?

1.4 Thesis Outline

To address the research questions, the research framework, as presented in Figure 2 was devised. The research approach consists of three phases: (1) theoretical background, (2) maturity model design and development, and (3) evaluation and refinement.

Chapter 2 covers the *theoretical background* in three building blocks. A literature review is performed on artificial intelligence, the enabling technology of augmented data management. A systematic literature review (SLR) is conducted to identify all data management and artificial intelligence models prevalent in literature. Another literature review is performed to identify maturity model development methodologies and guidelines.

Chapter 3 covers the *design and development phase* and starts with presenting the development strategy. The development strategy is based on the design science methodology and uses a mixed method of metamodel analysis and expert interviews. The metamodel for each DM and AI maturity model is constructed and used to compare and synthesize the models. Expert interviews and market research are conducted to identify relevant capabilities and processes for augmented data management. The result of this phase is the first version of the ADM Maturity Model, consisting of the selection of synthesized capabilities.

Chapter 4 covers the *evaluation and refinement stage*; the draft model is evaluated, validated, and improved using a mixed method of expert interviews and multiple case studies.

Chapter 5 covers the conclusion of the research, where the research questions are answered, and implications for practice and research are presented.

Chapter 6 covers the discussion of the research. The research methodology, the resulting ADM Maturity Model, its contributions, limitations, and future work are discussed.

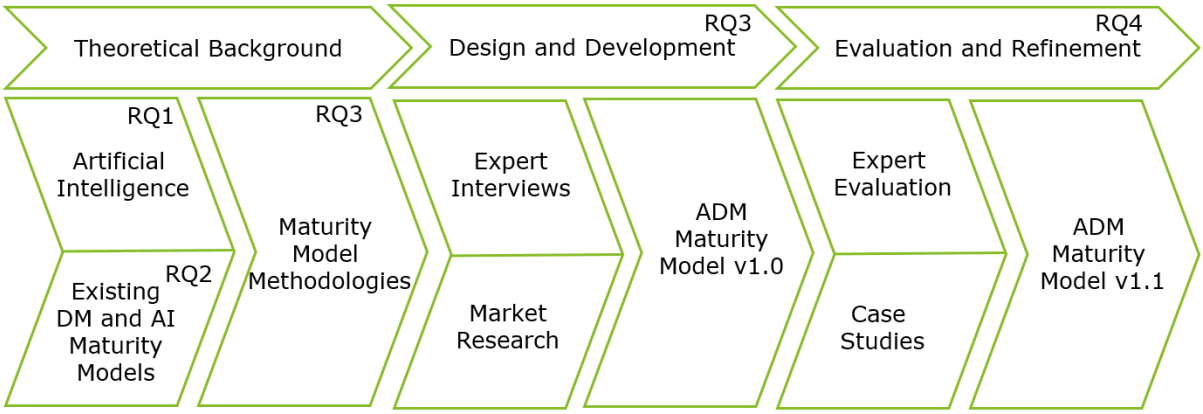


Figure 2: Research Framework

2. Theoretical Background

This chapter covers the scientific background of the research. Section 2.1 introduces the research discipline of artificial intelligence and its subfields. Section 2.2 describes the method used to perform a systematic literature review into maturity models for data management and artificial intelligence. Section 2.3 describes these data management maturity models, and Section 2.4 describes the artificial intelligence maturity models. Section 2.5 presents the background on maturity model types and methodologies.

2.1 Artificial Intelligence

The enabling technology of augmented data management is artificial intelligence. This section presents a brief analysis of the subfields within AI. We note that the goal is to provide background on how it can be leveraged and what tasks it can perform and not present a comprehensive explanatory or predictive theory. The overview is compiled by performing a literature review.

Artificial Intelligence is a broad term used by academics and practitioners worldwide to define intelligence displayed by machines. Beyond this general definition, current literature indicates only limited consensus on AI subfields, while much research is being done on AI techniques and applications. A common depiction of these subfields is displayed in Figure 3. These subfields combine both techniques and application domains as they are based on technical considerations, such as their goals, tools, or philosophical underpinnings and differences [27]. The remainder of this section outlines each sub-field.

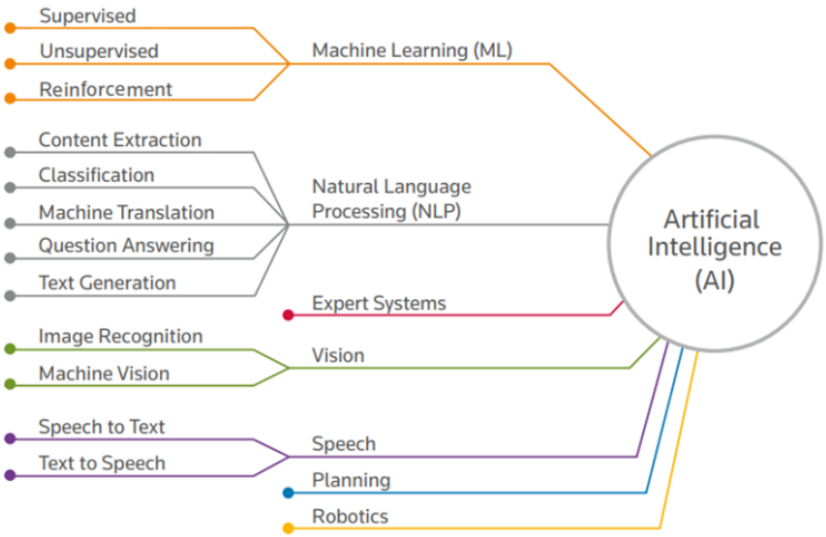


Figure 3: Overview of the AI field, Adapted From [27]

2.1.1 Machine Learning

Machine learning algorithms learn to perform tasks without explicit instructions. These algorithms differ in the learning technique and the underlying statistical model that is used. Within machine learning, many different subfields exist, such as deep learning and neural networks. Deep learning is a class of machine learning algorithms that uses multiple layers to extract higher-level features from the input and can be used to create complex models [28]. Neural networks are an example of a subfield within deep learning. Overarching for all these subfields are the learning types. The three classic types of machine learning techniques are supervised, unsupervised, and reinforcement

learning. A fourth, hybrid class can be considered, which combines multiple techniques. The most popular hybrid type is semi-supervised learning.

Supervised learning is the technique of learning a function that maps an input to an output based on example input-output pairs. The algorithm uses labeled or training data as an example to extract features or attributes for its function that correspond to the labeled class or category. The function can then be applied to new data to predict the output values based on the previous data sets [29]. The main types of supervised learning algorithms are classification and regression algorithms. Classification algorithms extract features that correspond to the labeled class to predict the class label of new data. One primary application of this type of algorithm is image recognition. Examples of classification algorithms are decision trees, random forest, and support vector machines. Regression algorithms extract features that correspond to a particular output in order to predict a value. Examples of regression algorithms are linear regression, multinarrative regression, and regression threes [29], [30].

Unsupervised learning is the technique of extracting inferences from data without labels to capture relationships between examples and uncover patterns. In contrast to supervised learning, there is no label or target given for the examples. The main types of algorithms are clustering and association rule learning algorithms. Clustering algorithms aim to group input data points into different classes using features derived from the input data. This algorithm can, for example, be used to group customer segments based on purchasing behavior. Examples of algorithms are k-means, k-medoids, and hierarchical clustering. Association rule learning algorithms are used for discovering relations between variables in large databases. For example, this algorithm can be used to identify products that are often bought together from an extensive sales dataset. Examples of algorithms are Apriori and GP growth [29], [30].

Semi-supervised learning is a combination of supervised and unsupervised learning, where unlabeled data is used to assist supervised learning. This technique is commonly applied, as it is the best technique in situations where there is limited training data available or the cost of manually labeling data is high. For this type of learning, both classification and clustering algorithms can be applied [30].

Reinforcement learning is a technique where an agent interacts with its environment and learns to take actions to maximize the reward and minimize the risk. The agent continuously learns from its experience of the environment in an iterative manner until it explores all possible ranges. These iterations follow several steps. First, the input state is observed by the agent. Then, the decision-making function is used to perform an action. After the action, the agent receives a reward from the environment, which leads to the new agent's state. The state-action pair of reward information is stored. For reinforcement learning, classification and control algorithms can be used. Examples of common algorithms are Q-Learning and Temporal Difference. Control algorithms are, for example, used in computer played board games or self-driving cars [29], [30].

Machine learning mimics human learning from training data to make predictions or decisions. This technique can subsequently take over routine tasks or tasks that are too complex for humans. For example, an experienced employee might recognize missing, incorrect, or duplicate customer data. Checking the data for every customer is time-intensive, and this employee has a limited ability to recognize duplicate data, as it is impossible to memorize all data. Machine learning can recognize and predict which files have missing or likely incorrect data and can scan the entire dataset to identify duplicates.

2.1.2 Natural Language Processing

Natural language processing (NLP) enables a computer system to understand and process human language. NLP can be classified into natural language understanding and natural language generation, where the input or output can be in human language. Modern NLP systems rely on machine learning to derive meaning from human language. NLP can be applied in various areas, like machine translation, information extraction, summarization, and dialogue systems, to understand and interpret human language in a similar way that humans do [31]. This technique can be used to process unstructured data such as plain text, for example, by recognizing topics or names.

2.1.3 Expert Systems

Expert systems embody expertise about particular domains and can use knowledge-based reasoning techniques to solve problems in those domains (i.e., problems that would usually need the assistance of a human expert in the real world) [32]. An expert system emulates the decision-making ability of human experts in order to solve complex systems through bodies of knowledge rather than conventional procedural code. Expert systems consist of a user interaction system, an inference engine, and a knowledge base. The inference engine is the control structure that allows the system to use search strategies to test different hypotheses and arrive at expert system conclusions. The knowledge base is the set of facts and heuristics about the expert system domain. Expert systems are most prevalent in medical diagnostics, engineering, and manufacturing applications [33]. The importance of these systems is paramount in areas and situations in which experiences employees might be scarce, and multiple experts' input might be urgent. In such cases, expert systems can be leveraged to complement or even replace experts' knowledge by identifying solutions to problems and explaining these solutions by presenting best practices and references.

2.1.4 Vision Recognition

The two key areas of vision recognition are machine vision and image recognition. Machine vision is the ability of computer systems to record and explore visual acuity. It captures and analyzes visual information using video cameras, analog-to-digital conversions, and digital signal processing. Image recognition applies machine learning techniques to identify and categorize computer vision input to recognize objects. Well-known examples are face recognition and medical image analysis [34]. Vision recognition enables the system to analyze images and video, which otherwise requires manual input. This technique can, for example, be used to recognize objects and people from text and video without human input.

2.1.5 Speech Recognition

Speech recognition is a technology that enables machines to process and produce spoken language. Speech recognition solutions implement either speech-to-text, or text-to-speech functionalities, or both. Speech to text functions enables computers to transform human language into commands that it can execute. There are various applications of speech to text technology, such as personal assistants like Amazon's Alexa, Google Virtual Assistant, and Apple's Siri. These applications further incorporate text to speech technology, which translates computer queries into human speech. Other examples of text-to-speech based systems are navigation systems and automated voice identification [11], [34]. Speech recognition enables vocal communication between humans and machines. Speech recognition can be combined with NLP to extract information from audio files without human input. For example, for transcribing recorded interviews or conveying information in a text to humans via speech.

2.1.6 Planning

Planning technology enables systems to find procedural action sequences in order to reach a goal while optimizing performance. The system's planning algorithm has an input of possible courses of actions, a predictive model for the system dynamics, and a measure for performance to evaluate the actions [35]. For unfamiliar environments where the performance of actions is unknown, reinforcement learning can be applied to find optimal solutions. An example of this combination are computer played board games, such as chess. Other AI planning applications can be found in supply chain planning, where advanced planning facilitates efficient production coordination. Advanced algorithms can incorporate external data like microeconomic cycles, geographic events, and weather to predict customer demand and automatically place purchase orders [36]. Planning uses certain factors to schedule procedural steps, similar to human planning. AI-assisted planning can incorporate more factors and complex calculations to find an optimal procedure, enhancing human efficiency.

2.1.7 Robotics

Robots are programmed physical machines that can perform a series of actions (semi) automatically. AI can be applied to robots to make them intelligent and let them perform more complex tasks, for which technologies such as computer vision and NLP can be leveraged. AI technology can drive new capabilities in robots for manufacturing as well as social robots, which interact with humans. Application areas include logistics, where robots are used to pick and transport orders [34]. Robotics automates manual tasks. With AI-assisted robotics, more complex tasks can be automated by leveraging computer vision to recognize objects, labels, or numbers to adapt actions accordingly.

2.2 Systematic Literature Review

In order to identify the leading data management and artificial intelligence maturity models, a systematic literature review is performed. For this research, the systematic review technique proposed by Kitchenham [37] is used to identify, evaluate, and interpret all available research related to data management and artificial intelligence maturity models. The methodology consists of three phases: planning, conducting, and reporting. First, Section 2.2.1 describes the underlying research questions for the review. Section 2.2.2 describes the planning phase, where the data sources and search strategy are defined. Section 2.2.3 covers the conducting phase, with data extraction and synthesis. Figure 4 visualizes the research methodology.

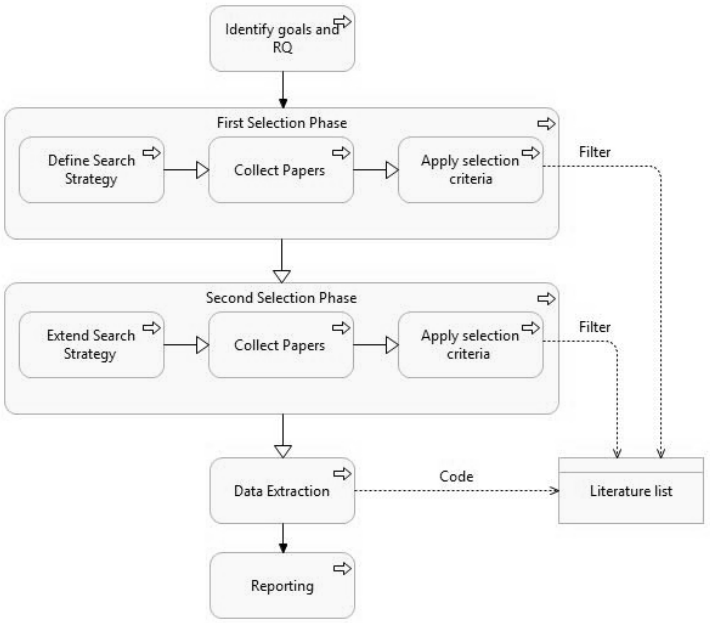


Figure 4: Research Methodology for the Systematic Literature Review

2.2.1 Research Questions

The goal of the systematic literature review is to answer sub-question 2 of the research.

2. **Which Data Management and artificial intelligence maturity models are available in current literature?**
    a. What does a Data Management model consist of, according to published literature?
    b. What does an Artificial Intelligence maturity model consist of, according to published literature?
    c. What are Data Management capabilities included in the reported models in the literature?

2.2.2 Data Sources and Search Strategy

The following data sources were selected to cover journals and books in the relevant subject fields of Information Systems and Computer Science: Scopus, Web of Science, ACM Digital Library, IEEE Xplore, and SpringerLink. Due to the high number of duplicate papers, we note that adding more digital repositories is not likely to result in additional relevant papers. To test this assumption, other digital libraries such as AISel were explored and confirmed this assumption.

Papers were searched and selected in two phases. The first phase aims to identify which models are prevalent in literature and what their names are. The following search terms are used to search for data management maturity models: “data management model”, “data management framework”, “data management maturity model”, or “data management capability model”. The following search terms are used to search for artificial intelligence maturity models: “artificial intelligence maturity model”, “Artificial intelligence capability model”, “AI maturity model”.

The second phase used the names of the identified models to find additional papers. If the original publications of data management models identified in the first phase were not among the second phase results, additional sources, such as the organization’s website, were consulted. The search query was applied to the title, abstract, and keywords of the articles. The data range was limited to the past ten years, from 2009 until 2020, to ensure relevance.

During both phases, papers were selected based on the following inclusion/exclusion criteria:

- To ensure academic quality, the document needs to be peer-reviewed; published in a journal, conference, workshop, technical report, thesis, or book (chapter). Due to the novelty of the subject and the limited amount of publications, the choice was made to include all document types, not only journals.
- To ensure relevance, the document needs to either propose a novel maturity model or report on the implementation of one.
- Software and database frameworks, such as Apache Hadoop for distributed data storage and processing, are excluded.
- Articles solely mentioning business or management process models are excluded.

2.2.3 Data Extraction and Synthesis

In order to compare the existing models found in the systematic literature review, each article is reviewed. Each article is classified using the classification proposed by [38] to extract relevant data. The maturity model analysis method by [39] is adopted as a systematic comparison approach. This thorough methodology considers three aspects for each model: the model structure, assessment, and support. Each aspect uses a set of variables which are detailed in [4] and [5] to define the model. The following variables and their definitions are used:

Model structure

1. Name: maturity model name and primary reference(s);
2. Number of levels: quantification of the maturity levels;
3. Name of attributes: definition of the attributes and sub-attributes that compose the maturity model. For data management, the attributes are the capabilities;
4. Number of attributes: number of attributes and sub-attributes used;
5. Maturity definition: indicates whether a detailed definition for capability maturity is given;
6. Practicality: provides practical or problem-specific recommendations.

Model assessment

1. Name: name of the maturity model and the primary references;
2. Assessment method described: whether the maturity model has an inherent method;
3. Assessment cost: the degree of expenditure of an assessment;
4. Strong/weak point identification: details about strong and weak points of the organization;
5. Continuous assessment: the pursuance of continuous improvement;
6. Improvement opportunities prioritization: the distinction between the order of improvement opportunities for the organization.

Model support

1. Name: name of the maturity model and the primary references;
2. Training available: the existence of training opportunities to become an expert;
3. Validation support availability: the degree of validation for the model based on the literature review. Only author support is ranked as low, validation with the organization as a medium, and validation outside the author's organization is ranked as high.
4. Tool support: whether the model includes data management tools or platforms;
5. Continuity from different versions shows the adaptability into newer versions of the model;
6. The origin of the model: academic or practical origin;
7. Accessibility: whether the documentation is freely available.

2.3 Data Management Maturity Models

The systematic literature review was performed in March and April 2020. The first search resulted in 981 articles, of which 11 were selected. The second search resulted in 841 articles, of which 14 were selected. A total of 10 data management models were identified. Three of those were only referred to once and did not show up in other searches. These models were disregarded, as it is hypothesized that these models were not adopted by the academic community or were only available in foreign languages. The resulting models and corresponding references can be found in Table 3. A more detailed description of every reference, according to the classification of Arnott and Pervan [38] can be found in Appendix A. The following two sections present the model characteristics and model capabilities.

Model	References
MD3M	[41], [42], [43], [44] ,[45], [46], [47]
DCAM	[48], [49]
CMMI DMM	[48], [1], [18], [50], [51]
IBM	[48], [52], [53], [54], [55]
Stanford	[18], [56], [57]
Gartner	[18], [6]
DAMA DMBOK	[58], [1], [59], [60], [61], [53], [62], [4]

Table 3: DM Models and Corresponding References

2.3.1 Model Structure

The majority of the maturity models have five levels. These levels correspond to the process level improvement model by the CMMI institute [51], as displayed in Figure 5. The lowest level covers undefined and unpredictable processes. The second level describes repeatable and reactive processes. The third level covers defined and proactive processes. The fourth level describes managed processes that are measured and controlled. The highest level strives for continuous improvement. DCAM is the only model with a sixth level, which is added below level 1 and is defined as 'not initiated'. These levels spread a selection of capabilities, which in turn can be split up into more variables. While the capability names vastly differ, there is overlap in the capabilities that they cover, as described in Section 4.2.1.4. About half the models provide a detailed description of each maturity level per capability, which can significantly improve the homogeneity across organizations and assessors. The other half provides a general description of the maturity levels based on CMMI and lets assessors define their definition per capability. All but one model provides specific recommendations for the defined (sub) capabilities. Almost every maturity model provides specific recommendations, while only Gartner provides general recommendations. An overview of all variables regarding model structure can be found in Table 4.

Maturity model	Nr. levels	Name of attributes	Nr. of (sub) attributes	Maturity definition	Practicality
Gartner[6]	5	Building blocks	7	Yes	General recom.
DCAM[49][63]	6	Components /Capabilities	7 / 31	Yes	Specific improv.
Stanford[39]	5	Dimensions	3	No	Specific improv.
IBM [64]	5	Categories	11	No	Specific improv.
CMMI DMM[51]	5	Categories/ Process areas	6 / 25	Yes	Specific recom.
MD3M[43]	5	Focus areas/Capabilities	13 / 65	No	Specific recom.
DMBOK[4]	5	Knowledge area	11	No	Specific recom.

Table 4: Synthesis of the Analyzed Maturity Models Regarding Model Structure

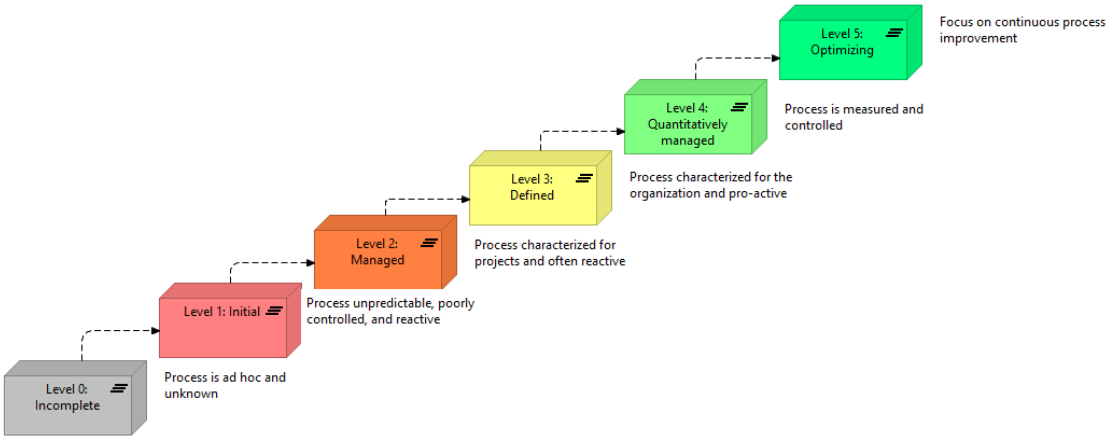


Figure 5: CMMI Maturity Levels Definition, Based on [51]

2.3.2 Model Assessment

All models suggest an assessment method, but the level of prescriptiveness differs. Like CMMI, DMBOK, and IBM, most models suggest doing a workshop with a representative participant to assess the maturity level collectively. However, the method does not provide assessment criteria other than the maturity definition. Other models like MD3M provide a questionnaire, which can be used as an assessment tool itself. Assessment costs are estimated by the level of detail in the assessment and the number of participants involved. Models with a high amount of capabilities and without guidance, such as a questionnaire, are estimated to have high assessment costs. About half the models mention strong and weak points per maturity level. These strong and weak points give the organization an indication of potential risks involved with low maturity and clearly states the benefits of advancing to higher maturity levels. Some models clearly state the iterative nature of the assessment, which can be used for continuous process improvements. Other models present themselves as one-time assessments or do not explicitly mention it at all. Inherent to all models is the pursuance to a higher maturity level. Some models present a hierarchy of capabilities within each maturity level, which presents a priority for improvement opportunities for organizations. An overview of all variables regarding model assessment can be found in Table 5.

Maturity model	Assess. method	Assess. cost	Strong/ weak points	Continuous assess.	Opportunity prior.
Gartner[6]	Yes	Medium	No	Yes	No
DCAM[49][63]	Yes	High	Yes	?	?
Stanford[39]	Yes	?	Yes	?	No
IBM [64]	Yes	High	No	Yes	Yes
CMMI DMM[51]	Yes	Medium	Yes	Yes	Yes
MD3M[43]	Yes	Medium	No	No	No
DMBOK[4]	Yes	High	Yes	Yes	No

Table 5: Synthesis of the analyzed maturity models regarding model assessment

2.3.3 Model Support

Models like DCAM and Garter were not supported by publications and only validated through claims made by the author(s). The IBM model was validated through multiple publications, but all authors were employed by or connected to IBM. The other models were peer-reviewed and applied in multiple cases by external authors. Most commercial organizations offer training opportunities to

become experts on their models. Training opportunities are often offered for a limited period or might be incidentally. Furthermore, third parties might offer training, which is not considered in the overview. The MD3M, CMMI DMM, and Stanford model have an academic origin. CMMI later gained a practical focus when the non-profit CMMI Institute was founded in 2012. DAMA (DMBOK) and EDM Council (DCAM) are both non-profit organizations with an international member base that shares best practices. Gartner and IBM are commercial parties that offer products, research, and advisory services. CMMI and DAMA offer their model through a (digital) book and charge a one-time fee. EDM Council charges substantial yearly corporate licenses for access to online material. Most models explicitly mention the use of tools. IBM includes many of their own products. Gartner refers to its own market reports on data management tools. The other models report on tool management and processes using tools but mention no specific capabilities or vendors. Most models have seen revisions and adjustments over the years, except for Stanford, IBM, and MD3M. An overview of all variables regarding model support can be found in Table 6.

Maturity model	Validation support	Training	Origin	Accessibility	Tool	Continuity
Gartner[6]	Low	No	Practitioner	Free	Yes	Yes
DCAM[49][63]	Low	Yes	Practitioner	Charged	Yes	Yes
Stanford[39]	High	No	Academic	Free	?	No
IBM [64]	Medium	No	Practitioner	Free	Yes	No
CMMI DMM[51]	High	Yes	Academic	Charged	Yes	Yes
MD3M[43]	High	No	Academic	Free	No	No
DMBOK[4]	High	Yes	Practitioner	Charged	Yes	Yes

Table 6: Synthesis of the analyzed maturity models regarding model support

2.3.4 Data Management Capabilities

A total of 32 capabilities are mentioned in the seven data management models, of which 13 overlap in two or more models. Despite the overlap, the capability definition can differ per model. Some models differentiate each capability into more detailed sub capabilities. Figure 6 presents an overview of all capabilities and the coverage per model. Figure 6 is adjusted to cover all the maturity models identified in the literature review. The data management capabilities can be grouped into governance, technology, data, data and system design, and related capabilities. The remainder of this section describes the 13 overlapping capabilities to introduce a general understanding of the most common capabilities.

The most covered capabilities are data quality, data architecture, data governance, stewardship, metadata, and master data management. Data quality is the planning, implementation, and control of activities that apply quality management techniques to data to ensure that it is fit for consumption and meets the need of data consumers [4]. Data architecture refers to the models, policies, and standards to guide data integration, control data assets, and align data investments with business strategy [4]. Data governance is defined as the exercise of authority and control (planning, monitoring, and enforcement) over the data management assets [4]. Stewardship is a quality-control discipline designed to ensure the custodial care of data for asset enhancement, risk mitigation, and organizational control [64]. Metadata describes the information about technical and business processes, data rules and constraints, and logical and physical structures. It describes the data itself and the relationships between the data and concepts [4]. Master data management is

defined as managing shared data to meet organizational goals, reducing risks associated with data redundancy, ensuring higher quality, and reducing data integration [4].

The following capabilities are mentioned in three of the models: Data management strategy, data storage and operation, and information life cycle management. Data management strategy defines the vision, goals, and objectives for the data management program and ensures that all relevant stakeholders are aligned on priorities and the program's implementation and management [51]. Data storage and operations is defined as the design, implementation, and support of stored data to maximize its value [4]. Information lifecycle management is a systematic, policy-based approach to information collection, use, retention, and deletion [64].

The following capabilities are mentioned in two models: organizational structures, awareness, security, technology infrastructure. Organizational structure and awareness refer to the level of mutual responsibility between business and IT and the recognition of fiduciary responsibility to govern data at different levels of management [64]. (Data) security refers to the definition, planning, development, and execution of security policies and procedures to provide proper authentication, authorization, access, and auditing of data and information assets [4]. Technology infrastructure focuses on the relationship of data with the physical IT infrastructure needed for operational deployment [49].

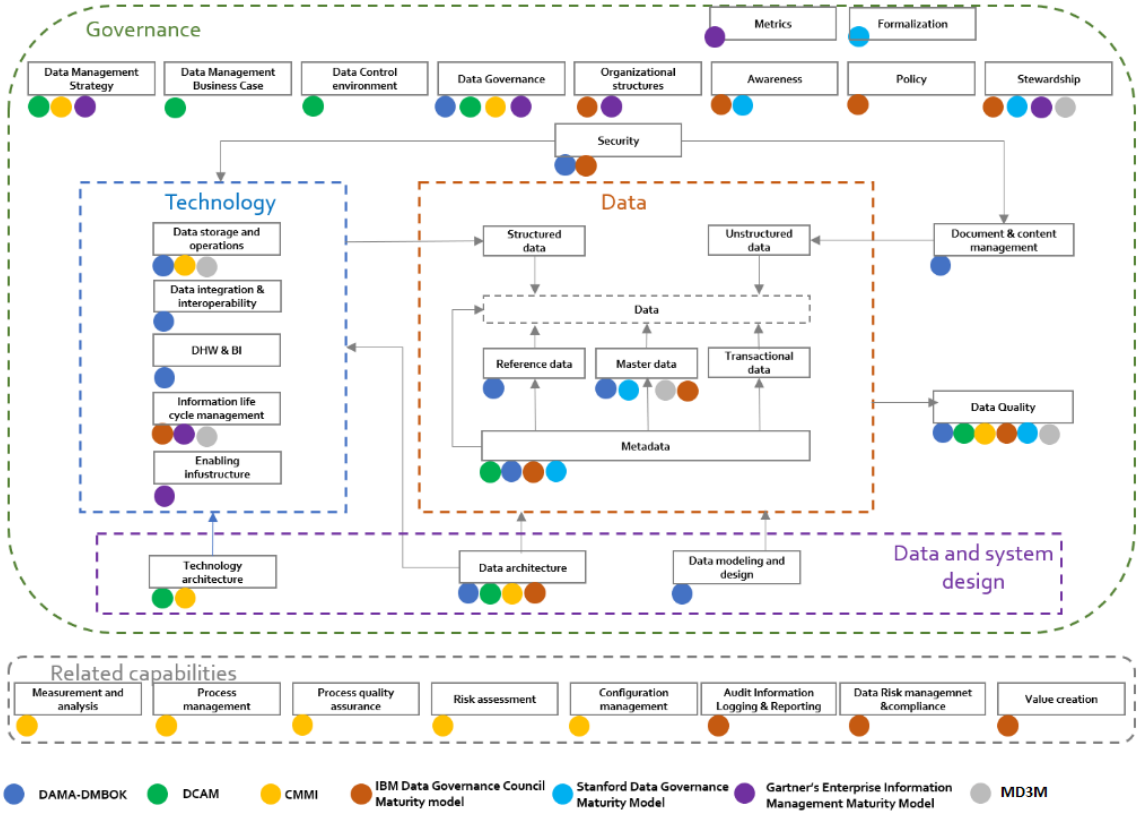


Figure 6: Model Capability Mapping, Adapted from [65]

2.4 Artificial Intelligence Maturity Models

The systematic literature review was performed in May and July 2020. The first search resulted in 1346 articles, of which two were selected. The second search resulted in 2 additional selected articles from non-academic sources. Based on the search, it can be concluded that maturity models for artificial intelligence are a novel field (for academics). Table 7 presents the AI maturity models and corresponding reference.

Model	References
AI Maturity Model (AIMM)	[66]
Algorithmic Maturity Model (AMM)	[67]
Gartner AIM	[68]
Ovum AIM	[69]

Table 7: AI models and corresponding references

2.4.1 Model Structure

AI maturity models either have four or five levels. All models describe the lowest level in a situation where no AI is applied, and a certain AI-readiness needs to be reached. Gartner AIM and AMM do not provide universal attribute definition, they only provide a definition per attribute maturity level. While these attributes differ per model, they can be categorized into people, technology, and processes. A more detailed comparison will follow in Section 3.4.4. All models except for AMM provide specific improvements per attribute. All variables regarding the model structure can be found in Table 8.

Maturity model	Nr. levels	Name of attributes	Nr. of (sub) attributes	Maturity definition	Practicality
AIMM [63]	5	Dimensions	4	Yes	General recom.
AMM [66]	4	N/A	5	Yes	Specific improv.
Gartner AIM [64]	5	Indicators	5	Yes	Specific improv.
Ovum AIM [65]	4	Core assess. pillars	5	Yes	Specific improv.

Table 8: Synthesis of the AI maturity models regarding model structure

2.4.2 Model Assessment

All models have little information on the maturity assessment itself. Only the Gartner AIM hints at an assessment method as ‘place yourself on the curve’. The other models do not mention any inherent assessment method. Gartner AIM hints at performing the assessment continuously and that potential improvements should be prioritized. These points are briefly mentioned but not central to the model. All variables regarding model assessment can be found in Table 9.

Maturity model	Assess. method	Assess. cost	Strong/ weak points	Continuous assess.	Opportunity prior.
AIMM [63]	No	?	No	No	No
AMM [66]	No	?	Yes	No	No
Gartner AIM [64]	Yes	Low	No	Yes	Yes
Ovum AIM [65]	No	?	Yes	No	No

Table 9: Synthesis of the AI maturity models regarding model assessment

2.4.3 Model Support

The two academic models have low validation support. The AIMM is research in progress. The original publication has no validation, and the research is still in progress at the time of writing. The book in which AMM is published has numerous citations, but these do not mention the model itself. Gartner and Ovum claim some validation as being industry leaders and working with experts. No training is available for any of the models as the book mentioning AMM has a moderate price tag, while the other models are freely available. For extensive documentation for the AIMM, a Gartner subscription is required. All models mention AI tools and platforms explicitly. Half the models have known revisions or announced them. All variables regarding model support can be found in Table 10.

Maturity model	Validation support	Training	Origin	Accessibility	Tool	Continuity
AIMM [63]	Low	No	Academic	Free	Yes	Yes
AMM [66]	Low	No	Academic	Charged	Yes	No
Gartner AIM [64]	Medium	No	Practitioner	Free	Yes	Yes
Ovum AIM [65]	Medium	No	Practitioner	Free	Yes	No

Table 10: Synthesis of the AI maturity models regarding model support

2.4.4 AI Dimensions

The four models combined have 19 attributes, which can be combined into seven overlapping attributes. The attributes and overlap are presented in Table 11. Only Ovum AIM and AIMM present a general definition of the attribute, which can be directly mapped onto each other. The other models provide a definition of each attribute per maturity level, for which the general definition must be reverse-engineered in order to map it to the others. Strategy refers to the plan of action and roadmap to support AI [69]. Data covers the availability of data assets and analytics capabilities [69]. The organization covers business characteristics such as culture, managerial structure, and decision making [69],[66]. People refer to individuals within an organization involved in AI [66]. Technology refers to the technologies and capabilities that are leveraged to implement AI [69]. Operations cover the where and how AI is supporting processes [69]. Budget & measures refer to the financial and structural involvement of AI [68].

AIMM	AMM	GAIM	OAIM
	Strategy	Vision & strategy	Strategy
Data structure	Data Analytics		Data
Organization	Organization/ People	Organization & Governance	Organization
People			
AI functions		Technologies employed	Technology
	Decisions	AI usage	Operations
		Budget & Measures	

Table 11: Model attribute mapping

2.4.5 AI Maturity Levels

The maturity levels differ across the different models, Table 12 presents an overview. Only the AIMM uses the five CMMI levels. Gartner AIM levels correspond roughly to these levels while using more business-oriented terms in the description. Gartner AIM deviates from the other models, which have an initial stage where no AI is present in any dimension. Instead, the lowest stage starts with planning for AI adoption by identifying first use cases and success criteria. Level 2 evolves around the discovery, experimentation, and assessment of AI technology. In the third stage, AI has a

defined position within the organization, and the first use cases are in production. In the fourth stage, AI creates a significant impact, processes are automated, and productivity is improved. The highest maturity level describes a synergy stage between human intelligence and AI, resulting in augmented intelligence.

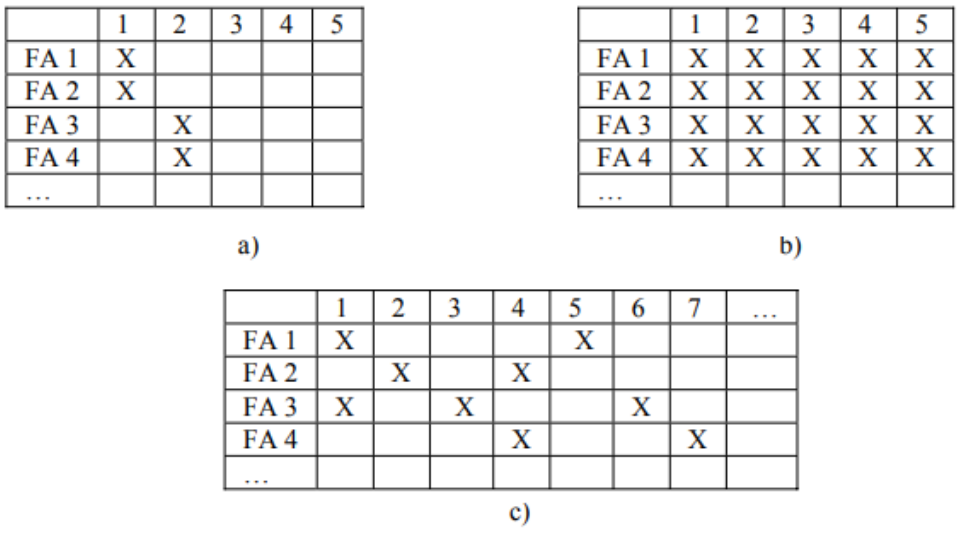
Level	AIMM	AMM	Gartner AIM	Ovum AIM
1a	Initial	Non-algorithmic		AI Novice
1b			Planning	
2	Assessing		Experimentation	AI-ready
3	Determined	Semi-automated	Stabilization	
4	Managed	Automated	Expansion	AI proficient
5	Optimized	Super intelligent	Transformation	AI advanced

Table 12: Model maturity level mapping

2.5 Maturity Model Development

2.5.1 Maturity Model Types

The first concept of a maturity model within information systems originated in 1974 [70]. Five years later, the original four-level model was extended to a six-level model, and the notion of gradually improving business processes by using these stages started. Since then, numerous maturity models have been developed. A maturity model is an assessment tool for specific areas of interest, which measures the degree of sophistication at which activities within this area are executed. Common for all maturity models is a set of interest areas: focus areas and a defined scale of maturity. The main differences between these models are in the type and structure. The three basic types are: staged fixed-level, continuous fixed-level, and focus area models. The difference between these types is illustrated in Figure 7. Staged fixed-level models ('a' in Figure 7) have fixed and generic maturity levels, generally five. Each level has several focus areas associated with that level, and all those focus areas need to be satisfied to reach that maturity level. A continuous fixed-level model ('b' in Figure 7) generally has five levels as well, but each focus area has its own maturity level. A focus area maturity model ('c' in Figure 7) has several specific maturity levels per focus area, not limited to five. The overall maturity of an organization is the combination of the maturity levels of all focus areas [71].



	1	2	3	4	5
FA 1	X				
FA 2	X				
FA 3		X			
FA 4		X			
…					

a)

	1	2	3	4	5
FA 1	X	X	X	X	X
FA 2	X	X	X	X	X
FA 3	X	X	X	X	X
FA 4	X	X	X	X	X
…					

b)

	1	2	3	4	5	6	7	…
FA 1	X				X			
FA 2		X		X				
FA 3	X		X			X		
FA 4				X			X	
…								

c)

Figure 7: Three Types of Maturity Models: Staged Fixed-Level (a), Continuous Fixed-Level (b), Focus Area (c), source [71]

Various IT-related maturity models exist. Mettler & Rohner (2009) found 135 different maturity models related to information systems. De Bruin et al. (2005) identified 150 maturity models related to IT management. A recent systematic literature review identified the methodologies, methods, and guidelines used by the academic community to develop IT maturity models [73]. It is essential to define the difference between these concepts when analyzing maturity model development. The methodology refers to a set of steps to conduct any type of research. The method is the tool that researchers use to gather data in order to complete those steps. Guidelines describe the steps that are necessary to develop a maturity model.

2.5.2 Methodologies

The 2020 systematic literature review finds that the majority of researchers are using their own methodologies to design maturity models [73]. These 'ad hoc' methodologies make up almost half of all studies. The other 37% of the studies did not specify any methodology. These numbers indicates a lack of good practices in the research development in this domain. Established methodologies such as design science research (DSR) and action research only make up 15% and 3%, respectively.

In recent years, the adoption of design science methodology within maturity model development is growing [73]. Design science is the design and investigation of artifacts in context [20]. The artifact is designed to improve a problem within the context. The maturity model is the artifact, and the context is the set of capabilities it aims to improve. The validation method in design science uses a model of the real-world context, which simulates realistic conditions.

Action research is an approach where a researcher collaborates with a practitioner to solve a real-world problem [20]. Hult and Lennung [74] provide a more detailed definition of information systems action research, based on six characteristics. Action research; (1) aims at understanding an immediate situation; (2) simultaneously assists in practical problem solving and expands scientific knowledge; (3) is performed collaboratively and enhances the competencies of the respective authors; (4) uses data feedback in a cyclical process; (5) is primarily applicable for the understanding of change processes in social systems and (6) is undertaken within a mutually acceptable ethical framework. Within the context of maturity model development, a model under development would be applied to solve a problem for a client while simultaneously researching the social impact and contributing scientific knowledge.

While DSR and action research have similarities, they differ significantly [75]. DSR per definition includes the design of an artifact, while action research does not. Within action research, the artifact under development is applied to a real-world problem, while in DSR a model of the artifact and context can be used. Action research collaborates with practitioners and focuses on researching the social context, while this is not necessarily the case with DSR. Despite the differences, these two methodologies can complement each other, especially in the validation of the artifact. Using a model of a real-world context in DSR is considered less robust than action research. However, performing simulations with a model requires less time than solving a real-world problem, making it easier to perform multiple experiments and generate a stable result. While action research provides a more realistic validation of the artifact, DSR provides a suitable methodology for the early development of maturity models.

2.5.3 Research Methods

The 2020 systematic literature review identified twelve different methods employed in the development of maturity models, with almost half of the articles combining multiple methods [73]. The most popular method is the literature review in 68% of studies, followed by interviews (20%), case study (7%), focus group (6%), and surveys (5%). Other methods are less popular and only used in one or two studies.

The following methods are used in maturity model development:

- Literature review: method of identifying, evaluating, and synthesizing the existing work produced by researchers, scholars, and practitioners [37].
- Interview: qualitative research technique that involves asking in-depth questions with exerts to collect data.
- Case study: in-depth research of a single instance to explore causes of underlying principles.
- Focus group: a technique where a group discusses a specific topic, aiming to synthesize personal experiences and perceptions through moderated interaction [76].
- Survey: data collection method using written response to questions.
- Delphi: iterative method to collect and distill the judgments of experts using a series of questionnaires interspersed with feedback [77].
- Content analysis: a method where concepts found in the interpretative data analysis are added to a conceptual research framework [20].
- Workshop: method using domain-related cases in a workshop format
- Entropy: a method that uses weights to evaluate results. The weights are determined according to the influence of the relative change [78].
- Card sorting: a method where participants assign categories to sub-components, optionally with hierarchy.
- Meta-synthesis: method to integrate findings from multiple qualitative studies [79].
- Analytical hierarchy process: a method where alternatives are analyzed by hierarchical (sub)criteria [78].

2.5.4 Guidelines for Developing Maturity Models

Guidelines provide a description of the steps described in a methodology for developing maturity models. Many researchers have presented their own guidelines or synthesized them from literature. At least 14 guidelines for maturity model development have been identified [73],[80]. While most researchers use their own guidelines, the adoption of maturity model development guidelines in the scientific community has increased over the past 15 years [73]. Based on citation count, the models of Becker et al. [15] and de Bruin et al. [16] are the most popular. In addition, the less-cited models of Mettler and Rohner [72], Maier et al. [81] and van Steenbergen [82] are considered as well, as they present different approaches that include organizational characteristics, a grid maturity scale, and specific steps for focus area model development.

In 2005, de Bruin et al. concluded that there is little documentation on developing a maturity model that is theoretically sound, rigorously tested, and widely accepted [16]. De Bruin et al. were the first to generalize the phases of maturity model development into six general phases [16]:

1. **Scope**: determine model focus and stakeholders
2. **Design**: develop the model architecture, audience, and application
3. **Populate**: determine domain components and measurement methods
4. **Test**: analyze the relevance and rigor of the model and measurement instruments
5. **Deploy**: make the model available for usage
6. **Maintain**: keep a repository to support model evolution and development

A 2020 systematic literature review identified that the guidelines by Becker et al. [15] and Hevner et al. [83] are the most widely accepted, with 10% and 5% adoption [73]. The guidelines by Hevner et al. outline design science research within information system research and are not specific for maturity model development. Becker et al. build on top of this knowledge by proposing guidelines specific to maturity model development. Becker et al.'s procedure model for developing maturity models consists of seven steps:

1. **Problem definition:** design science aims to develop a problem-solving artifact; therefore, the first step is to define the problem the model aims to solve.
2. **Comparison of existing maturity models:** compare existing models to determine the design strategy.
3. **Determination of development strategy:** construct a new model, combine existing models, or transfer structures or content from existing models to a new context.
4. **Iterative maturity model development:** the development of the model itself, in four iterative steps: select level design, select approach, design model section, and test result.
5. **Conception of transfer and evaluation:** transfer results to the academic and practitioner community.
6. **Implementation of transfer data:** make the maturity model accessible to the target audience.
7. **Evaluation:** test the designed model for comprehensiveness, consistency, and problem adequacy. If insufficient, another development phase is entered.

Mettler and Rohner [72] present guidelines for designing situational maturity models. Situational models are designed to include the situativity in organizational design. These guidelines can assure a better fit for a specific organization but limits the generalizability of the model and comparability between organizations. While the authors do not explicitly define guidelines, they present them according to an example case:

1. **Problem identification and motivation**
2. **Objectives of the solution**
3. **Design and development:** consist of basic maturity model design, specification of the maturity levels, configuration of parameters, and proof of concept.

Maier et al. [81] propose guidelines for developing maturity grids. According to the authors, maturity grids differ from maturity models in their work orientation, models of assessment, and intent. Maturity models define specific processes, while maturity grids define general characteristics of what any process should look like. Maturity models generally use surveys with Likert scales or binary yes/no questionnaires in their assessment, while maturity grids are structured around a matrix or grid. In their intent, maturity grids tend to be less complex and formal than maturity models. The development guidelines consist of four steps and corresponding decision points [81]:

1. **Planning:** Specify audience, aim, scope, and success criteria
2. **Development:** Select process areas, maturity levels, formulate cell text and define administration mechanism
3. **Evaluation:** Validate and verify the model
4. **Maintenance:** Check benchmark, maintain results database, document and communicate development process and results

Van Steenbergen et al. [82] present guidelines for the design of focus area maturity models. Focus area models do not use the generic maturity levels that fixed-level models use but are designed to enable incremental improvement in specific functional domains. Each focus area has its own number and type of maturity level. The development consists of ten steps:

1. **Identify the scope and functional domain:** Determine scope and domain to ensure a useful model
2. **Determine focus area:** Determine focus areas within the domain from literature, focus groups, or case studies
3. **Determine capabilities:** Develop a rationale of how to focus areas can be incrementally improved
4. **Determine dependencies:** Identify dependencies between capabilities within and outside the focus area
5. **Position capabilities in a matrix:** Capabilities that are dependent on others are placed further to the right. Independent capabilities may be placed on the same scale
6. **Develop assessment instrument:** Formulate control questions for maturity assessment
7. **Define improvement actions:** present suggestions for incremental improvement
8. **Implement the maturity model:** Perform surveys, interviews, or workshops to collect assessments
9. **Improve Matrix iteratively:** Evaluate how the model assists in improvement
10. **Communicate results:** Communicate model to practitioners and the scientific community

Table 13 presents an overview of the identified maturity model development guidelines according to the classification by van Steenbergen [78].

Common phase	**De Bruin et al.** [16]	**Becker et al.** [15]	**Mettler and Rohner** [72]	**Maier et al.** [81]	**Van Steenbergen et al.** [82]
Scope	-Scope	- Problem definition - Comparison of existing models	-Problem identification & motivation -Objectives of the solution	-Planning	-Identify scope and functional domain
Design model	-Design -Populate components	-Determination of development strategy -Iterative model development	-Design and development	-Development	Determine: -Focus area -Capabilities -Dependencies -Position capabilities
Develop instrument	-Populate measurements -Test	-Conception of transfer and evaluation		-Evaluation -Maintenance	-Develop assessment instrument -Define improvement actions
Implement & exploit	-Deploy -Maintain	-Implementation of transfer data -Evaluation			-Implement MM -Improve Matrix iteratively -Communicate results

Table 13: Development guidelines overview

3. Design and Development

This Chapter covers the maturity model design and development. Section 3.1 outlines the development strategy for the ADM Maturity model. Section 3.2 describes the execution of the development strategy. Section 3.3 presents the capabilities, sub capabilities, and processes of the model. Section 3.4 summarizes the result by presenting the ADM Maturity Model Assessment Tool.

3.1 Mixed-Method Development Strategy

The present research applies the design science methodology. Within this methodology, the decision is made to apply the guidelines of Becker et al. [15]. The reasons for this choice are the following: (1) the guidelines are based on DSR, in line with Hevner et al. [83] (2) have the highest adoption in the scientific community, and (3) explicitly include the comparison of existing maturity models. The procedure model for the present research can be found in Figure 8. The procedure model is executed using a mixed-method strategy, consisting of a systematic literature review, metamodel analysis and synthesis, expert interviews, and market research. This section describes each step of the procedure model and methods used.

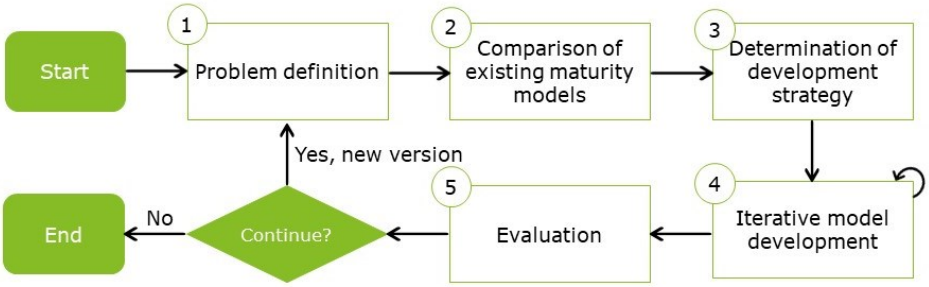


Figure 8: Procedure model for guidelines based on Becker et al. [15]

The first step is the *problem definition*. In this phase, the research problem is identified, the relevance established, the stakeholders are identified, and the research goals are formulated. The research problem is formulated based on a literature review, market analysis, and expert interviews. The interviews are additionally used to establish relevance and demand for the ADM maturity model. The problem definition is presented in Section 1.1.

The second step is the *comparison of existing maturity models*. For this, a systematic literature review is performed using the approach of Kitchenham [37] to identify maturity models for data management and artificial intelligence. The comparison of existing maturity models is presented in Section 2.3-2.4.

The third step covers the *determination of the development strategy*. The basic strategies to developing a maturity model are: design a completely new one, enhance existing models, combining several models into a new one, and the transformation of structure or content from existing models to a new domain [18]. The choice is made to combine existing maturity models, for which the motivation can be found in Section 3.1.1.

The fourth step includes the *iterative model development*, containing the phases of select the level design, select approach, design model, and test result. The choice is made to use a metamodeling approach. The metamodel for each maturity model for data management and artificial intelligence is modeled and used to compare similar constructs. The metamodel approach is detailed in Section

3.1.2 and the systematic comparison is presented in Section 3.1.3. The result of the metamodel comparison is an overview of all the processes and capabilities of the maturity models from the literature. Expert interviews and market research are conducted, as presented in Section 3.1.4, to identify capabilities and processes that are relevant for augmented data management. By analyzing the market research and expert interviews as presented in Section 3.1.5, the maturity model is constructed as described in Section 3.1.6.

The final step combines the last three steps of Becker et al. [15] into one *evaluation* phase. The evaluation phase consists of expert interviews to validate the model and case studies to demonstrate the applicability of the model. This mixed-method validation strategy is detailed in Section 4.1. Table 14 provides an overview of these procedure steps and the corresponding sections of the thesis.

Procedure Model Step	Section
1 Problem Definition	Section 1.1
2 Comparison of models	Section 2.3-2.4
3 Determination of strategy	Section 3.1
4 Iterative development	Section 3.2-3.3
5 Evaluation	Chapter 4

Table 14: Procedure Model Steps and Corresponding Sections

3.1.1 Combining Existing Models

The systematic literature review resulted in an overview of data management and artificial intelligence maturity models in Sections 2.3 and 2.4. Based on the overview of all the model structure, assessment, support, and content, there are strong indications that a combination of existing maturity models is the best strategy. All data management models consist of a similar set of five or six maturity levels set of similar (sub) capabilities. Some of these data management models are widely known and used. Building on top of these models enhances the potential usability and ease of use. As augmented data management seeks to improve current data management processes, these capabilities provide an excellent starting point. The maturity levels of existing data management models can be extended with knowledge from artificial intelligence maturity models to reflect augmentation. Therefore, the choice is made to combine existing maturity models. As all the existing models are continuous-fixed level, the choice is made to use this model type.

For defining maturity stages, a top-down or bottom-up approach can be used. With a top-down approach, the maturity stage definition is proposed first, followed by determining the measures that fit the definition. In contrast, a bottom-up approach starts with the requirements and measures and defines the maturity stage based on those. A top-down approach works well in a new domain where little is known on maturity indicators and measurements, while a bottom-up approach is better suited within more developed domains [16]. Due to the novelty of augmented data management, the choice is made to adopt a top-down approach for defining maturity stages.

3.1.2 Metamodel Approach

The design strategy is to combine existing maturity models. A metamodel approach is used to identify similar constructs of each maturity model to systematically compare and synthesize them, as presented in Section 3.1.3. The metamodeling approach is essential, as each maturity model uses different definitions and structures. The metamodel is constructed from the content diagram and description in the publication of each model. Figure 9 displays the metamodels for the four AI maturity models, and Figure 10 displays the seven data management maturity models.

By analyzing the metamodels, similar constructs can be identified across maturity levels. Figure 9 presents the metamodels for the AI maturity models. Each row represents similar constructs in terms of granularity and hierarchical structure. The top row presents the name of the four AI maturity models. The second row presents the common construct 'maturity model', named 'dimension' in AIMM and AMM, 'capability' in Gartner AIM, and 'pillar' in Ovum AIM. The third row presents the construct 'maturity level', which has the same name in every model. The final row presents 'assessment' constructs, which AMM and Gartner AIM define as 'indicator' and Ovum AIM as 'attribute'. The capabilities and maturity levels of each maturity model are compared using the systematic comparison method, as presented in Section 3.1.3.

Figure 10 displays the metamodels for the DM maturity models. From top to bottom, the rows present the following common constructs: model name, group, capability, maturity level, and assessment. The data management maturity models are more diverse in terms of architecture, definitions, and granularity. Based on the metamodel analysis, the constructs are divided into capabilities and sub-capabilities of the same granularity. In each maturity model, the 'higher level' capability is defined as Focus area (MD3M), Capability (DCAM), Maturity component (Stanford), Process area (CMMI), Category (IBM DGMM), knowledge area (DAMA DMBOK), and building block (Garter EIMM). The 'lower level' sub-capabilities are defined as capability (MD3M), sub-capability (DCAM), Related processes (CMMI), sub-steps (IBM DGMM) Activity (DAMA DMBOK). The capabilities, sub-capabilities, and maturity levels of each maturity model are compared using the systematic comparison method, as presented in Section 3.1.3.

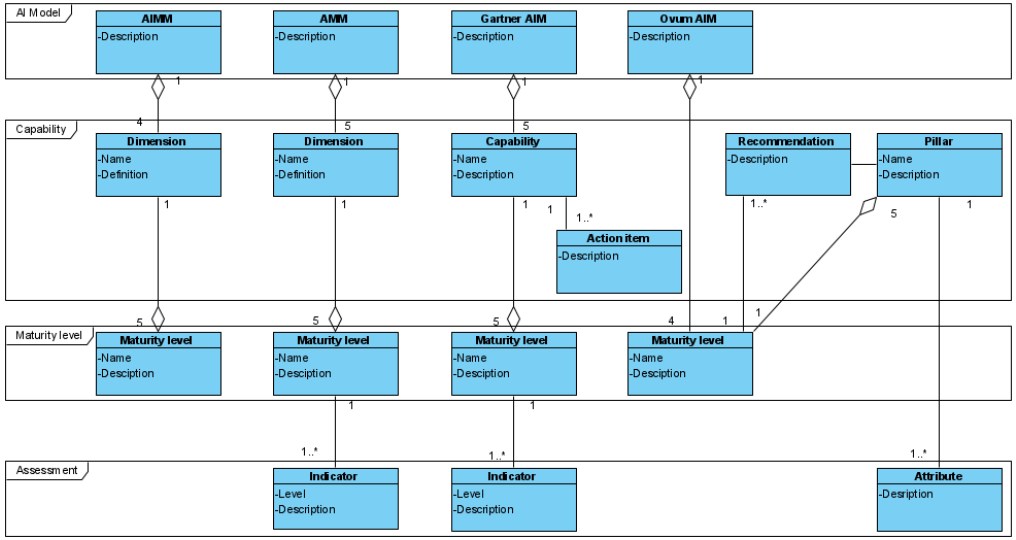


Figure 9: Metamodels for AI Maturity Models

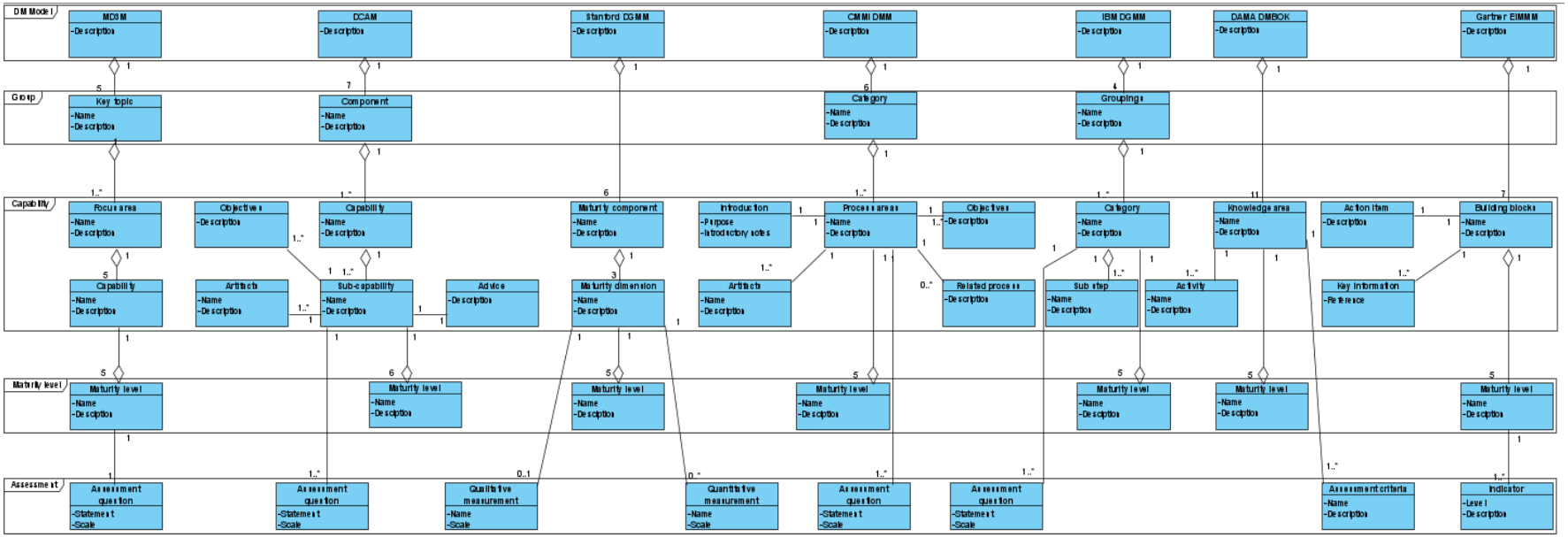


Figure 10: Metamodels of Data Management Maturity Models

3.1.3 Systematic Metamodel Comparison

In the metamodel analysis, similar constructs from different maturity models are identified in order to compare them. The method by Lautenschutz et al. [84] is used to systematically compare these constructs. The method consists of four steps:

1. Conduct a literature review of related maturity models, as presented in Section 2.3-2.4.
2. Use the construct diagram of each model to construct its metamodel. The metamodels are used to identify and compare similar constructs of each model, as presented in Section 3.1.2.
3. Select the pivot model for the comparison. The result of the comparison does not depend on the pivot model; it merely determines the presentation.
4. Perform a systematic comparison of constructs of each model with the pivot model. Equivalent constructs can be mapped to each other or presented as a unique construct.

The systematic comparison (step 4) is the most crucial step of the method. Metamodel analysis is used to compare constructs of different maturity models that do not necessarily have the same name or description [85]. The analysis is done using either deductive or inductive reasoning. Deductive reasoning means a thinking process starting from general constructs to more detailed ones, i.e., by splitting the general construct into smaller constructs and trying to compare those. Inductive reasoning works from detailed to general by grouping multiple smaller constructs into a group or category which can be matched to other constructs. For example, DCAM presents the capability 'baseline data quality', DAMA DMBOK present the capabilities 'define scope of initial assessment' and 'perform initial data quality assessment'. By using inductive reasoning, it can be argued that both capabilities by DAMA DMBOK are sub capabilities of the umbrella category 'baseline data quality'. On the contrary, using deductive reasoning, it can be argued that 'baseline data quality' could be split into the more detailed capabilities described by DAMA DMBOK.

During the present research, multiple matrices are constructed by comparing the model's maturity levels and capabilities. Figure 11 illustrates this concept. The matrix columns consist of the different maturity models, while the first column is the pivot model. The rows present the compared constructs. One table is constructed for the maturity levels and one for the capabilities. Each construct, illustrated in Figure 11 as a green cell, has a description used for the qualitative content analysis and as a rationale to substantiate its mapping relative to the pivot construct. Constructs that are on the same row are equivalent. From the simplified example in Figure 11 it can be seen that model B does not have a construct that maps to the first construct of the pivot model (empty cell), but model C does (green cell).

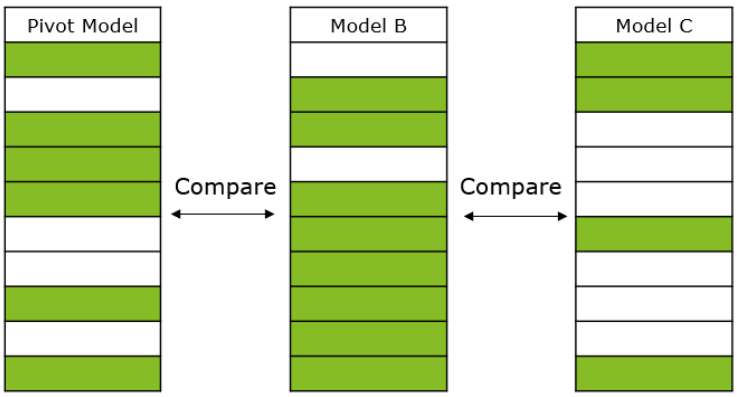


Figure 11: Visualization of the Comparison Table and Systematic Metamodel Comparison

The comparison matrix creates an overview of all constructs of the maturity models and whether they overlap with other models. The DAMA DMBOK model is used as the pivot model for comparing the data management models as it covers all the selected capabilities. For comparing AI maturity models, the pivot model is AMM. The result of the comparison does not depend on the pivot model but choosing the most comprehensive model facilitates a structured comparison. The result of the comparison is an overview of all activities related to the selected capabilities while additionally displaying overlap between models. The overview of the systematic comparison can be found in Appendix B.

3.1.4 Expert Interviews and Market Research

The comparison table for the selected data management capabilities provides an overview of all associated sub-capabilities and processes. Seven interviews are conducted with industry experts to identify which capability activities are relevant for augmented data management and compose the ADM maturity model. The industry experts are selected based on the criteria that they have at least five years' experience with both artificial intelligence and data management.

The goals of the interviews are (1) to confirm the selection of capabilities as having the highest potential and priority to be augmented, (2) to confirm the subfields of AI: ML, NLP, Expert systems, Vision recognition, Speech recognition, Planning and Robotics, and (3) to identify which processes can be augmented with AI. The following questions are used for the semi-structured interviews to establish these goals:

1. How to leverage AI technology to augment data management capabilities?
2. Do you agree that the main subfields of AI are ML, NLP, Expert systems, Vision recognition, Speech recognition, Planning, and Robotics? Would you add/remove some?
3. Do you think metadata management, master data management, data integration, data quality, and database management have the largest potential for AI augmentation?
4. How can AI be leveraged within those capabilities? What are the current and future applications?
5. Are you familiar with maturity models? Do you think it could be useful to have a maturity model for augmented data management?

In addition to expert interviews, market research is conducted on industry-leading software vendors that claim to incorporate augmented data management capabilities into their tools. These software vendors are selected from market reports from Gartner [86] and by asking the experts during the interviews. By combining the interviews and the market research, a comprehensive overview can be constructed of augmented data management processes; Section 3.1.6 explains this process in more detail. Table 17 in Section 3.2.5 presents a list of all interviews and market research sources that are used for the construction of the model.

3.1.5 Qualitative Data Analysis

The expert interviews are transcribed for analysis. To extract grounded and valuable insight from the data, the qualitative data analysis guide by Dey is used [87]. The full transcripts of the expert interviews can be found in Appendix C. The analysis approach can be summarized in three steps:

1. **Reading and Annotating**: Read the transcriptions, highlight, and annotate sections relevant to the maturity model, the open questions, or the evaluation criteria.
2. **Categorizing Data**: Label the data with categories. The list of categories can be found in each transcription Appendix.
3. **Corroborating Evidence**: Data is combined based on the categories and content, e.g., the (sub) capability or process that the comment targets.

To increase the support and validity of the findings, the qualitative content analysis is peer debriefed. Peer debriefing is the critical analysis of the interpretation process by independent by scientific peers [20]. The choice for this validation method is made because the peer can critically evaluate the interpretation process without being an expert on AI and data management. The peer has experience in qualitative content analysis and was presented with the transcripts of the expert interviews, the interview questions, the labels, and the decisions made during the interpretation process. After critical evaluation, the peer noted that the transcripts often literally contained the labels and that the interpretation process was straightforward and valid.

3.1.6 Constructing the ADM Maturity Model

The ADM maturity levels are synthesized from the systematic comparison of the maturity levels from the AI and data management maturity models. The synthesis is done along two axes, one for data management maturity and one for AI maturity.

The ADM capabilities are constructed by combining the systematic comparison with the expert interviews and market research. Figure 12 illustrates and simplifies this process. On the left, all models are systematically compared to identify all related data management processes for the five selected capabilities from literature. The interviews and market research identify which processes can be augmented by leveraging AI; these processes compose the draft ADM maturity model. The draft version is evaluated and refined in order to present the final version.

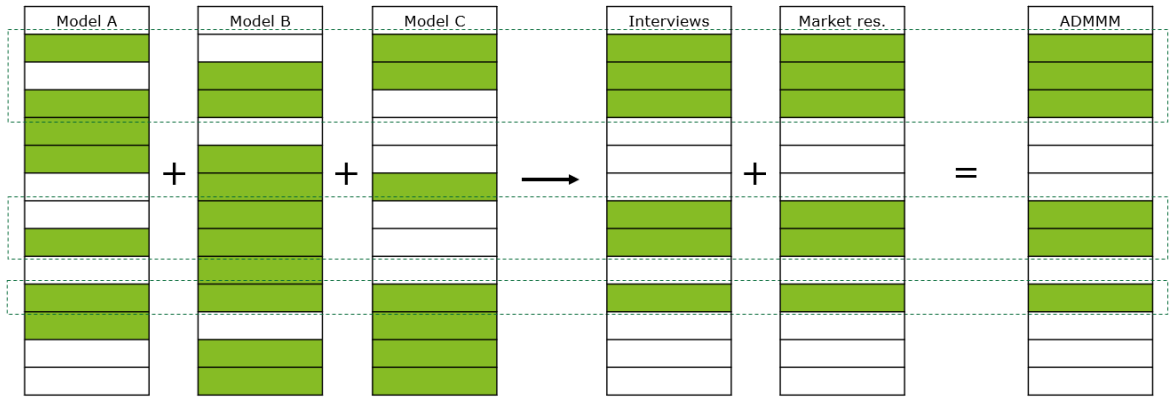


Figure 12: Construction of ADM Capabilities Simplified

3.2 ADM Maturity Model Version 1.0

This section describes the development of the first version of the ADM Maturity Model. Section 3.2.1 compares the maturity levels of existing models. Section 3.2.2 presents the maturity levels for the ADM Maturity Model. Section 3.2.3 compares and synthesizes the capabilities of existing data management models. Section 3.2.5 presents the process of selecting sub-capabilities and processes, which are presented in Section 3.3.

3.2.1 Synthesizing Maturity Levels

Table 15 presents the maturity levels for the AI models and the mapping to the pivot model AIMM. Table 16 presents the maturity levels for the DM models and the mapping to the pivot model DAMA DMBOK. All maturity levels are either mapped onto a level from the pivot model or presented as a new level, using the systematic comparison method. An example from Table 15: the AIMM does not describe a maturity level in which an organization has not taken any proactive steps in leveraging AI. The lowest maturity level of AMM describes such a state, and therefore level 0 is introduced. The lowest maturity level of GAIM and OAIM both describe a state where no proactive steps in AI have been taken, so they are mapped to level 0 accordingly. The description of all maturity levels and the rationale for their mapping can be found in Appendix B.

Lvl	AIMM	AMM	GAIM	OAIM
0		Non-algorithmic	Planning	AI Novice
1	Initial	-	Experimentation	-
2	Assessing	-	-	AI Ready
3	Determined	Semi-automated	Stabilization	-
4	Managed	Automated	Expansion	AI Proficient
5	Optimized	Super intelligent	Transformation	AI Advanced

Table 15: Maturity levels of AI models

Lvl	DAMA	DCAM	Stanford	Gartner	CMMI DMM	IBM	MD3M
0	No Capability	Not initiated	-	Aware	-	-	
1	Initial / Ad Hoc	Conceptual	Initial	Reactive	Performed	Initial	Initial
2	Repeatable	Developmental	Managed	Proactive	Managed	Managed	Repeatable
3	Defined	Defined	Defined	-	Defined	Defined	Defined process
4	Managed	Capability achieved	Quant. managed	Managed	Measured	Quant. managed	Managed & measurable
5	Optimization	Capability enhanced	Optimizing	Optimized	Optimized	Optimizing	Optimized

Table 16: Maturity levels of DM model

3.2.2 ADM Maturity Model Levels

As identified in Section 3.2.1 the maturity levels are synthesized for the ADM Maturity Model and resulted in two maturity axes: one for data management process maturity and one for augmentation maturity. The initial idea was to combine the levels from both AI and DM maturity models into one scale for augmented data management. Combining these scales was not feasible, as process maturity and the degree of augmentation are fundamentally different concepts; one process can score high on data management maturity while scoring low on augmentation. As data management maturity models are being applied in practice, and the ADM Maturity Model intends to complement these models, the choice was made to include two scales with separate maturity levels. Figure 13 presents the two maturity axes and a short description of each level. The vertical scale is the data management maturity scale, the horizontal scale the augmentation maturity scale. The two scales are applied independently to assess the same (sub) capabilities and processes.

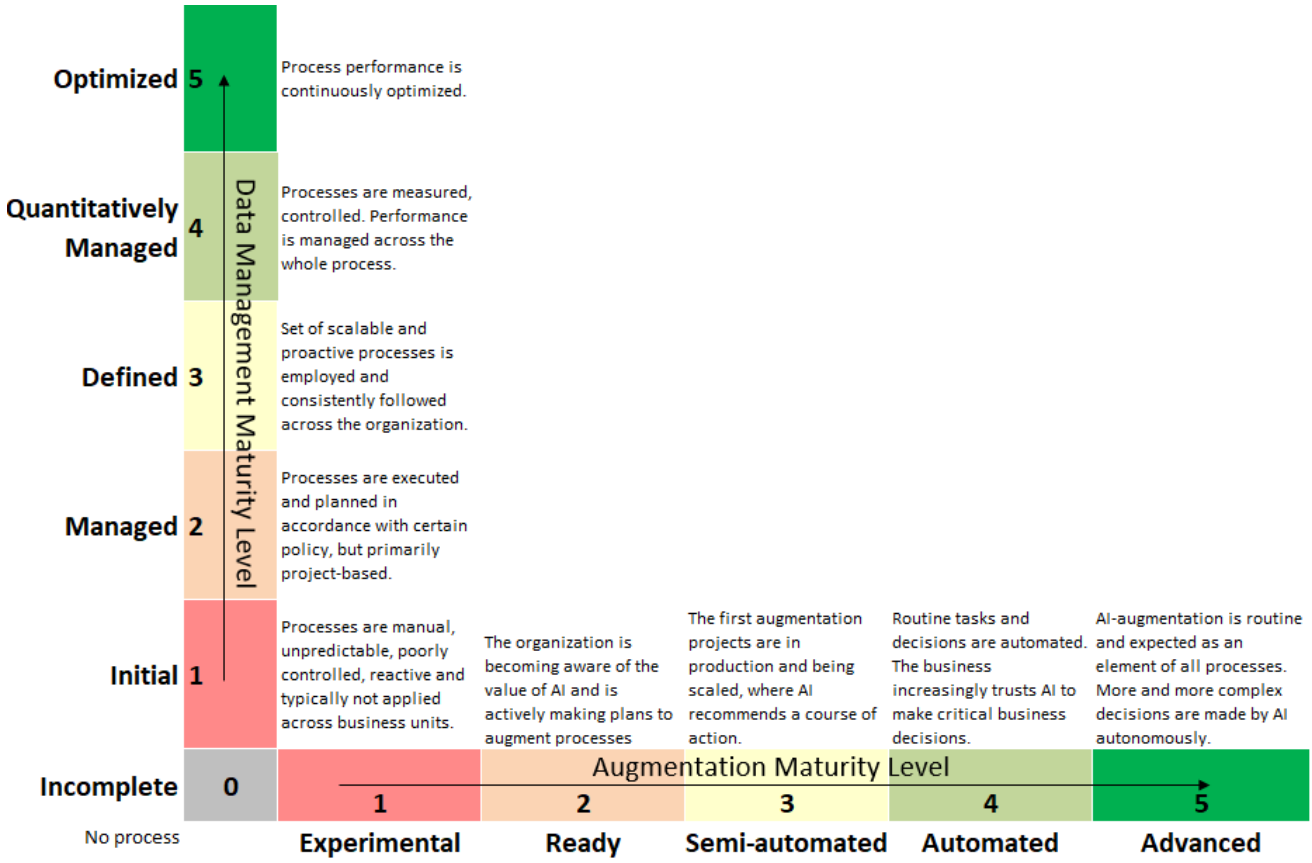


Figure 13: ADM Maturity Axis

3.2.3 Selecting Capabilities

The seven data management models combined contained a total of 32 capabilities, which are presented in Section 2.3.4. Some of these capabilities are quite detailed, others are universal and were mentioned in multiple models. A selection is made of five universal capabilities with the largest potential to be augmented to scope the research. The other capabilities still have the potential to be augmented; the selection is made with the sole purpose of scoping the research. The selection is made based on three criteria that indicate a high potential and priority for the capability to be augmented:

- **The amount of data**: Data is a prerequisite to apply AI and ML. Data heavy processes have a large potential to be optimized by AI and ML [86].
- **Literary consensus**: Capabilities that are present in multiple models can be seen as 'core elements' of data management.
- **The amount of manual work**: By augmenting time-intensive tasks, more data can be analyzed, while data professionals can focus on other value-adding activities.

Using these criteria, the selection is based on the following observations from the systematic literature review and market research: Data quality, metadata management, and master data management are amongst the most mentioned capabilities from literature, as resulted from the literature review in Section 2.3.4. These capabilities are amongst the most data-heavy. Informatica, one of the leading developers of Enterprise Cloud Data Management tools, incorporates AI in its platform to enhance data quality, data integration, metadata management, and master data management [88]. A survey by Gartner on Data and Analytics trends in 2018 revealed tasks with the highest priorities to automate [89]. Data integration is ranked highest with 49%. Data ingestion (29%) tasks closely overlap with data integration and database management capabilities. Data preparation and cleansing (37%) tasks are part of data quality. IBM and 451 Research identify database management as a field where AI has a high potential in automating tasks and improving operational performance [90]. To summarize, the five selected capabilities are metadata management, data integration, master data management, data quality, and database management.

The capabilities that are not included are still crucial to data management as a whole and can still have the potential to be augmented. For example, data governance is an important capability yet has a low potential to be augmented as it covers mostly social processes. Data management processes can be optimized and augmented; however, if no one adheres to the policies and processes are not governed, the outcome is still bad. Some data management maturity models mention different capabilities that are included as sub-capabilities within the selection—for example, the capabilities of data modeling, risk management, and compliance. Data modeling is included as part of metadata management, master data management, and database management. Risk management and compliance are partly covered in data integration, metadata management, and master data management.

3.2.4 Synthesizing Capabilities

For the ADM Maturity Model, five capabilities were selected in Section 3.2.3. These capabilities are first synthesized from existing data management models by using the systematic metamodel comparison method. This comparison results in one overview of all the relevant sub-capabilities and processes. The overview consists of distinct 46 sub-capabilities. The overview including descriptions and mapping rationale, can be found in Appendix B. Using this overview, a selection can be made of sub-capabilities and processes that are relevant to the ADM Maturity model, which is covered in Section 3.2.5 and 3.3.

The rest of this section summarizes the main sub-capabilities and processes from all data management models associated with the five selected capabilities: data quality, metadata management, data integration, master data management, and database management.

Data Quality

Data quality is a set of practices concerning the planning, implementation, and control of activities that aim at maintaining a high quality of information in order to meet the requirements of data consumers. This process applies to the acquisition of data, the implementation of advanced data processes, and the effective distribution of data. Because of this, data quality is strongly related to other capabilities. The definition of data quality depends on the dimensions considered and must be suitable for the organization's business goals. The following presents a set of dimensions with general agreement on the definition and measurement approaches [4]:

- Accuracy: the degree of representing real-life entities, measured in comparison with a verified accurate source.
- Completeness: the proportion of data stored against the potential for 100%
- Consistency: the absence of difference when comparing two or more representations of a thing against a definition.
- Integrity: a combination of completeness, accuracy, and consistency
- Reasonability: the degree data patterns match expectations, can be measured by comparing to benchmarks or historical data
- Timeliness: The degree to which data represent reality from the required point in time can be measured by latency.
- Uniqueness: No entity instance will be recorded more than once within the dataset, can be measured by testing against the key structure.
- Validity: Data is valid if it conforms to the syntax (format, type, range) of its definition and measured against the syntax itself.

A variety of tools is associated with managing data quality. Data profiling tools produce generic statistics that enable quality assessments and can be used for monitoring. Profiling tools augmented with visualization capabilities serve the data discovery process. Data querying tools make it easier to request large data sets for more in-depth analysis. Modeling and ETL tools are used to create data processes and therefore have a direct impact on the data quality. Data quality rule templates guide data expectations. Formulating rules that bridge business and technical terms helps in providing complete and quality data. Metadata repositories can be used to this extend, as definitions of high-quality data are a valuable kind of metadata.

The following sub-capabilities and processes are related to data quality management [4], [43], [51], [64]:

- Establish data quality policies:
    - Define high-quality data: determine fit that fulfills business requirements for all stakeholders
    - Identify critical data: focus on the most important data
    - Establish awareness: communicate the benefits of data quality and consequences of bad quality.
- Baseline data quality: measure data objectively to understand data content and relationships
- Data profiling: develop an understanding of the content, quality, and rules of the specified data set
- Build the business case
    - Identify and prioritize potential improvements

    - Define goals
- Cleanse the data: define mechanisms, rules, processes, and methods to validate and correct data according to predefined business rules
- Monitor the data quality over time: Incorporate a systematic approach to measure and evaluate data quality according to processes, techniques and against data quality rules

Metadata Management

Metadata describes the information about technical and business processes, data rules and constraints, and logical and physical structures. It describes the data itself and the relationships between the data and concepts. Metadata management structures the planning, implementation, and control to establish high quality and integrated data [4]. Metadata helps an organization understand its data, its systems, and its workflows. It enables data quality assessment and is integral to the management of databases and other applications. A metadata repository acts as a catalog of data within the whole organization. Metadata repository management tools can be used to find available data in one central location. Data can either be manually entered or extracted from sources via connectors. The tool itself simultaneously acts as a metadata source, as it can exchange metadata from other repositories [4].

The following sub-capabilities and processes are related to metadata management: [4], [64]:

- Create Metamodel: create a data model for the metadata repository.
- Apply Metadata standards: metadata should be monitored to comply with quality and security standards and enable data exchange. Standards include naming conventions, custom attributions, security, visibility, and processing documentation.
- Manage metadata stores: this involves monitoring, responding to reports, warnings, job logs, and resolving various issues in the repository environment. Monitoring operational data ensures that issues are resolved, such as failing operations.
- Create and maintain metadata: central for all sub-activities is to assign accountability, set and enforce standards, and create feedback loops for continuous improvement.
    - Ensure data lineage: create technical metadata to make an audit trail for data movement, where does the data come from, where does it go, how is it transformed
    - Integrate metadata: Collect and integrate metadata from diverse sources to ensure knowledge about the similarities and differences in the organization’s data.
    - Merge business metadata from the business glossary with technical metadata from the data dictionary to bridge business and technical teams.
    - Distribute and deliver metadata: Provide standard ways to make metadata accessible from a centralized repository to metadata consumers (people and systems) via intranet sites, reports, data warehouses, modeling tools, and APIs.
- Query, report, and analyze metadata: use metadata in BI, business decisions, and business semantics. Conduct impact analysis: map dependencies and impact on data flow

Data Integration

Data integration describes processes that are related to the movement and consolidation within and between data stores, applications, and consolidations [4]. Since large companies often have hundreds of databases and applications, managing the process of data movement is key. This capability focuses on the data flow processes, whereas database management focuses on related technical activities.

Several tools can be used for data integration. The primary tool is a data transformation engine/ETL tool for designing data transformations. ETL processes can be done virtually by using a Data virtualization server. Enterprise Service Bus middleware can be used to implement near real-time messaging between heterogeneous data stores, applications, and servers. Data profiling tools can be used to perform statistical analysis on the format, completeness, consistency, validity, and structure of the data. Metadata repositories can be used to identify data sources to integrate and to document the technical structure and business meaning integrations [4].

The following sub-capabilities and processes are related to data integration [4]:

- Define data integration and lifecycle requirements: understand the business objectives and the data required to meet those objectives.
- Perform data discovery: identify potential sources and high-level assessment of suitability. The metadata repository is an important source.
- Design integration solutions:
    - Select interaction model: hub-and-spoke, point-to-point or publish-subscribe
    - Design data services: create or re-use existing integration flows
    - Data profiling: understanding content and structure for mapping
    - Map source to targets: transformations and technical format
- Develop data integration solutions:
    - Develop data services: often tools or vendor suites
    - Develop data flows: integration or ETL data flow tools
    - Develop a publication approach: event-driven or periodically
    - Develop complex event flows: process real-time data and define triggers to execute action in response to signals or predicted data.
    - Maintain metadata: document data structures of source, target, and stage systems
    - Document data lineage: create metadata to document integrated dataflows.
- Monitor: automated or human monitoring for issues that trigger alerts or an automated response

Master Data Management

Master data is data that has a common definition across an organization and provides the context for business activity data. Master Data Management (MDM) entails control over Master Data values and identifiers that enable consistent use across systems of the most accurate and timely data about essential business entities. It includes the details of internal and external objects involved in business transactions, such as customers, products, vendors, and controlled domains. The goals of MDM include ensuring the availability of accurate, current values while reducing risks associated with ambiguous identifiers. Ambiguous those identified with more than one instance of an entity and those that refer to more than one entity. DAMA DMBOK distinguishes reference data in addition, which is data that is used to characterize other data, such as reference lists, code, and description tables. Master Data Management can be implemented through data integration tools, data remediation tools, operational data stores, data sharing hubs, or specialized MDM applications [4].

The following sub-capabilities and processes are related to master data management [4], [43], [64]:

- Define MDM drivers and requirements: define application-wide requirements
- Evaluate and assess data sources: understand the structure of application data, understand the quality of the data, identify the disparity

- Define architectural approach: the number of source systems integrated and sharing approach (one/multiple hubs)
- Model master data: define a logical model of subject areas within the data-sharing hub
- Define stewardship for master data: define ownership of data, create a console for monitoring and manual interventions
- Establish governance policies: enforce the use of master data by systems and people
- Ensure maintenance: establish a single point of truth
    - Manage potential overlay tasks: manage overlays between records with different data
    - Match duplicate suspects: identify records that point to the same entity

Database Management

Database management refers to the technical design, implementation, and support of stored data to maximize its value. Database management is responsible for the technical operations regarding inaccuracy and consistency of data over the entire lifecycle: acquisition, migration, retention, expiration, and disposal. All databases share common processes. Archiving is the process of moving data to low-cost and low-performance storage. Capacity and growth projections deal with balancing storing capacity. Change data capture refers to the process of detecting changes in the data. These changes can either be tracked via versioning or by logging changes. Purging is the process of completely removing data beyond recovering, which can speed up the system but also provides a risk when misused. Resilience and recovery refer to how tolerant a system is to error conditions. Retention planning includes the timeframe that data is kept available. Data modeling tools allow the generation of database definition language. Database monitoring tools automate monitoring of key metrics such as capacity, availability, cache performance, and user statistics. It can alert administrators when issues arise. Database management tools have a function for configuration, installation of patches, backup and restore, database cloning, test management, and data cleanup routines.

The following sub-capabilities and processes are associated with database management [4]:

- Define storage requirements: Establish file storage systems, capacity growth projections, data retention, and purge period
- Identify usage patterns: predict peaks and valleys and take advantage
- Manage access authorization to different files, digital and physical
- Plan for business continuity: make backups, recover data
- Develop database instances: installing and updating DBMS software, maintaining environment installations, installing, and administering related data technology.
- Mange database performance:
    - Set performance service levels
    - Manage database availability: related to manageability, recoverability, reliability, serviceability
    - Manage execution: plan tasks efficiently and respond to issues
- Manage data migration: automated and manual data remediation

3.2.5 Selecting Sub Capabilities and Processes

To select relevant sub-capabilities and processes from the overview in Section 3.2.3, expert interviews and market research is conducted. Table 17 provides an overview of all references from the interviews and market research used to select processes. Section 3.2.6 and 3.2.7 introduce the expert interviews and market research, Section 3.3 presents the selected sub capabilities, processes, and references used. The interview transcriptions can be found in Appendix C; a more detailed overview of the market research can be found in Appendix D.

Reference	Type	Position/Product	Date consulted
1	Interview	Manager Artificial Intelligence	13-05-2020
2	Interview	Senior consultant analytics & cognitive	14-05-2020
3	Interview	Senior consultant analytics & cognitive	15-05-2020
4	Interview	Senior manager AI strategy	15-05-2020
5	Interview	Senior manager Oracle EDM	18-05-2020
6	Interview	Assistant professor	26-05-2020
7	Interview	Manager EDM	26-05-2020
8	Slides [91]	Deloitte Digital FTE	23-07-2020
9	Slides [92]	Deloitte CongiSteward	23-07-2020
10	Whitepaper [88]	Informatica CLAIRE	27-07-2020
11	Journal Article [93]	SnowFlake Cloud Data Warehouse	27-07-2020
12	Website [94]	Alteryx Self-Service Data Analytics Platform	28-07-2020
13	Whitepaper[95]	Oracle Autonomous Database	28-07-2020

Table 17: Reference Table for Constructing ADM Capabilities

3.2.6 Expert Interviews

A total of seven expert interviews are conducted in May 2020 with six managers and one assistant professor. Each expert has between 5- and 10-years' experience with AI and data management within the industry. The first seven references in Table 17 provide an overview of the interviews conducted. During the interviews, the experts were asked to comment on the capability selection, the AI subfields, and identify which and how data management processes can be augmented with AI.

All experts agreed with the proposed subfields of AI: ML, NLP, Expert systems, Vision recognition, Speech recognition, Planning, and Robotics. All interviewees agreed on the selection of capabilities, which indicates that these indeed have the largest potential and priority to be augmented. However, as all interviewees were specialized in a selection of the five capabilities rather than all capabilities, this is not conclusive.

3.2.7 Market Research

A total of six sources are consulted for the market research on augmented data management tools. These sources are presented in Table 17, reference 8 to 13. All the tools were identified as industry-leading during the expert interviews. The market research goal is to provide insight into tools that enable to augment data management processes. The following section provides a brief description of each tool; a more detailed description can be found in Appendix D.

Deloitte DFTE [91]

Digital FTE's are a set of tools designed to augment the human workforce during projects by executing repetitive and rule-based activities without human intervention. These accelerators work

with minimal input and aim to reduce operational costs by automating tasks and eliminating human errors. This enables humans to focus on more complex and value-adding tasks.

Each DFTE tool performs a specific task. A selection of tools covers tasks related to data management: performing data validation checks, generate business and technical metadata, automate the conversion of logs to structured reports, automate data lineage, mapping source to target data models, migrating data, and configuring MDM hubs.

Deloitte CogniSteward [92]

Deloitte CogniSteward is an advanced data management self-service tool that augments manual, costly, and time-consuming data steward activities related to data quality, metadata management, and master data management. The solution can either be used as an accelerator during a project, or can be part of the deliverable where it is continuously being used by clients to handle complex and vast amounts of data.

Informatica [88]

Informatica is a data integration and data management software company. Gartner identifies Informatica as a market leader in data integration, data quality, metadata management, and master data management tools. Informatica introduced AI/ML functionalities under the name CLAIRE, which supports various solutions on their intelligent data platform. The platform offers the modules Data Catalog; Data Engineering; Data Integration; Data Quality & Governance; Data Privacy; iPaaS: Data, API & Application integration and Master Data Management

Snowflake [93]

Snowflake is a cloud-based data platform based on data warehouse automation provided as Software-as-a-Service. Snowflake differentiates itself from traditional data warehouse solutions or big data platforms by a unique architecture and service execution designed for the cloud. As Snowflake provides a service, database management is completely outsourced. Snowflake leverages the cloud to provide scalable storage and computing capacity as well as an optimized execution of both. While Snowflake is not transparent in the underlying technologies, it can be presumed that AI is leveraged in scaling resources and optimizing queries.

Alteryx [94]

Alteryx is a software tool that aims to make advanced analytics accessible to data analysts by combining data preparation, data integration, and analytics into one no-code platform. Alteryx augments time-consuming and manual data management and analytics activities using drag and drop tools. The platform offers seven products: Analytics Hub, Designer, Server, Connect, Promote, Intelligence Suite, and Datasets.

Oracle Autonomous Database [95]

Oracle Autonomous Database combines the flexibility of the cloud with the power of machine learning to deliver data management as a service. The goal is to minimize manual intervention and human errors within database management and ensure data safety and optimal performance. These automation functionalities allows IT staff to focus on higher-value activities while saving costs on repetitive and time-consuming tasks. Autonomous databases achieve this by being self-driving, self-securing, and self-repairing.

3.3 ADM Maturity Model Capabilities

This section presents the capabilities, sub-capabilities and processes of the first version of the ADM Maturity Model. Each capability is presented in a table, which also points to the references from the expert interviews and market research as presented in Table 17.

Data Quality

Table 18 presents an overview of the sub-capabilities, processes, and references related to augmented data quality.

Sub-Capability	Process	Reference
Assess DQ	Assess large datasets to generate a data quality score based on statistics and data rules.	**1,8,9,10**
	Specify data quality dimensions and data validation rules manually or reverse engineer from the data itself.	**4,10**
Data Profiling	Profile large and complex datasets by parsing the data and recognizing data types, structures, metadata, data categories (e.g. email, address) and generate basic statistics	**1,8,9,10**
Data Cleansing	Categorize data quality issues	**4**
	Suggest actions for data cleansing and standardization	**3,4,5,10**
	Learn from manual data cleansing to generate suggestions for similar DQ issues or perform cleansing autonomously	**3,4,5,9**
Monitor DQ	Perform ongoing data quality/validation checks on data pipelines	**6,7,9**
	Detect anomalies by significant differences between actual data and expected values from historical data	**6,7,9,10**

Table 18: ADM³ v1.0 Data Quality Overview

Metadata Management

Table 19 presents an overview of the sub-capabilities, processes, and references related to augmented metadata management.

Sub-Capability	Process	Reference
Define Metadata Architecture	Generate metamodels based on the data.	**1,7**
	Reverse-engineer (meta)data rules based on datasets.	**4**
	Rationalize data dictionaries through industry taxonomies, folksonomies and client specific taxonomies	**9**
Create and Maintain Metadata	Generate metadata from structured data: identify topics, taxonomies (recognize names, address, email, etc.) and generate keywords.	**2,8,9,10**
	Extract attributes from unstructured data (text, images, video) to generate metadata.	**2,8,9,10**
	Catalog and index this metadata in a repository	**9**
	Merge business and technical data based on similarity and relation.	**10,12**
	Generate end-to-end data lineage by creating metadata that tracks data flows and transformations.	**4**
Analyze Metadata	Convert the technical session logs to structured reports to check if data is moved or mapped as expected.	**7,8**

	Validate file metadata, highlight potential mismatches and missing data and restructure metadata automatically or request manual input.	**7,8**

Table 19: ADM³ v1.0 Metadata Management Overview

Data Integration

Table 20 presents an overview of the sub-capabilities, processes, and references related to augmented data integration.

Sub-Capability	Process	Reference
Data Discovery	Recommend data during discovery, based on relationships and similarity in data.	**10**
Ensure Data Lineage	Reverse engineer data lineage from code or from metadata.	**2,3,8,9,10**
	Perform impact analysis	**10**
	Perform root-cause analysis	**10**
Design DI	Model data for structured data, discovering attributes and structure	**1**
	Model data for unstructured data, discovering attributes and structure	**1,9**
	Map source data model to target data model	**1,2,5,6,8**
Develop DI	Validate file metadata, highlight potential mismatches and missing data and restructure metadata automatically or request manual input.	**7,8**
	Ingest, validate and transform data to the target structure based on existing mappings and user interaction	**1,2,5,6,8,10**

Table 20: ADM³ v1.0 Data Integration Overview

Master Data Management

Table 21 presents an overview of the sub-capabilities, processes, and references related to augmented master data management.

Sub-Capability	Process	Reference
Evaluate and Assess Data Sources	Generate data models and linkages: identify relationships between columns across different sources.	**9,10**
Model Master Data	Generate a master data model by recognizing entities and hierarchical structure (bottom-up)	**10**
	Detect and map data entities onto a predefined master data model (top-down)	**10**
	Configure MDM hub as per data model	**8**
Define Stewardship and Maintenance Process	Identify duplicate data based on clustering and blocking attributes.	**1,3,6,7,9,10**
	Learn resolution & reconciliation rules from human interventions to generate recommendations or autonomously establish a single point of truth.	**1,9**

Table 21: ADM³ v1.0 Master Data Management Overview

Database Management

Table 22 presents an overview of the sub-capabilities, processes, and references related to augmented database management.

Sub-Capability	Process	Reference
Manage Database Performance	Forecast computational demand to scale computational resources or schedule jobs based on available resources to meet performance criteria.	**5,6,10,11**
	Optimize queries for better response time.	**6,10,11**
	Monitor database logs to detect anomalies in run jobs to trigger automated fault recovery, threat detection or request manual interaction.	**5,10,12,13**
	Provision, manage, monitor, backup, recover and tune databases.	**5,13**

Table 22: ADM³ v1.0 Database Management Overview

3.4 Result of the Development Phase

In order to operationalize the ADM Maturity Model, an Excel assessment tool is made. The goal of the assessment tool is to guide data management consultants in performing a maturity assessment in line with the requirements presented in Section 1.2.4. The first version of the assessment tool was used in the evaluation stage of the present research described in the next Chapter.

The first version of the tool consisted of 9 different tabs that represent consecutive steps in the assessment protocol:

1. **Introduction**: The first tab includes instructions on how to use the model, information on the background and development of the model, and an overview of the capabilities and maturity levels.
2. **Maturity Levels**: The second tab presents the two maturity axes. The participant is asked to carefully read and understand the different maturity levels.
3. **Assess Capability Processes**: The following five tabs each cover one capability, with the sub-capabilities and processes as presented in Section 3.3. The participant is asked to relate the process description to a maturity level that represents the organization's current processes, both for data management process maturity and augmentation maturity. The participant is also asked to set a desired maturity. Figure 14 shows the data quality tab of the assessment tool, where the current and desired maturity can be specified.
4. **Results**: The overall maturity is automatically calculated for all capabilities and sub-capabilities and displayed in the results tab. The results tab displays the current and desired maturity and allows to identify gaps, which serves as a starting point to construct an improvement roadmap for selected (sub) capabilities. Figure 15 displays the results tab.
5. **Evaluation** (draft only): During this research, the participant is asked to fill in an evaluation form that quantifies various metrics regarding the maturity levels, processes, understandability, ease of use, usefulness, and practicality. The participant is also asked open questions on how to improve those aspects of the maturity model. Chapter 4 goes into more detail on evaluating and improving the draft version of the model and assessment tool.

Capability 1: Data Quality

"The planning, implementation and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and

Sub-capability	Processes	DM	Current	Desired	ADM	Current	Desired
Assess DQ	Assess large datasets to generate a data quality score based on statistics and data rules.	0	UNDEFINED	0	0	UNDEFINED	0
	Specify data quality dimensions and data validation rules manually or reverse-engineer from the data itself.	0	UNDEFINED	0	0	UNDEFINED	0
Data Profiling	Profile large and complex datasets by parsing the data and recognizing data types, structures, metadata, data categories (e.g. email, address) and generate basic statistics	0	UNDEFINED	0	0	UNDEFINED	0
Data Cleansing	Categorize data quality issues	0	UNDEFINED	0	0	UNDEFINED	0
	Suggest actions for data cleansing and standardization	0	UNDEFINED	0	0	UNDEFINED	0
	Learn from manual data cleansing to generate suggestions for similar DQ issues or perform cleansing autonomously	0	UNDEFINED	0	0	UNDEFINED	0
Monitor DQ	Perform ongoing data quality/validation checks on data pipelines	0	UNDEFINED	0	0	UNDEFINED	0
	Detect anomalies by significant differences between actual data and expected values from historical data	0	UNDEFINED	0	0	UNDEFINED	0
Data Quality	Overall maturity	0	UNDEFINED	0	0	UNDEFINED	0

LEVEL	DM MATURITY	DESCRIPTION	ADM MATURITY	DESCRIPTION
0	Undefined	No formal processes	Undefined	No formal processes
1	Initial	Processes are unpredictable, poorly controlled, reactive and typically not applied across business units.	Experimental	Processes are unpredictable, poorly controlled, reactive and typically not applied across business units.
2	Managed	Processes are executed and planned in accordance with certain policy, but primarily project-based.	Ready	The organization is becoming aware of the value of AI and is actively making plans to augment processes
3	Defined	Set of standard proactive processes is employed and consistently followed across the organization.	Semi-automated	The first augmentation projects are in production and being scaled, where AI recommends a course of action.
4	Quantitatively Managed	Processes are measured and controlled. Performance is managed across the whole process.	Automated	Routine tasks and decisions are automated. The business increasingly trusts AI to make critical business decisions.
5	Optimized	Process performance is continuously optimized.	Advanced	AI-augmentation is routine and expected as an element of all processes. More and more complex decisions are made by AI autonomously.

Figure 14: Maturity Assessment Tool v1.0 Tab for Data Quality

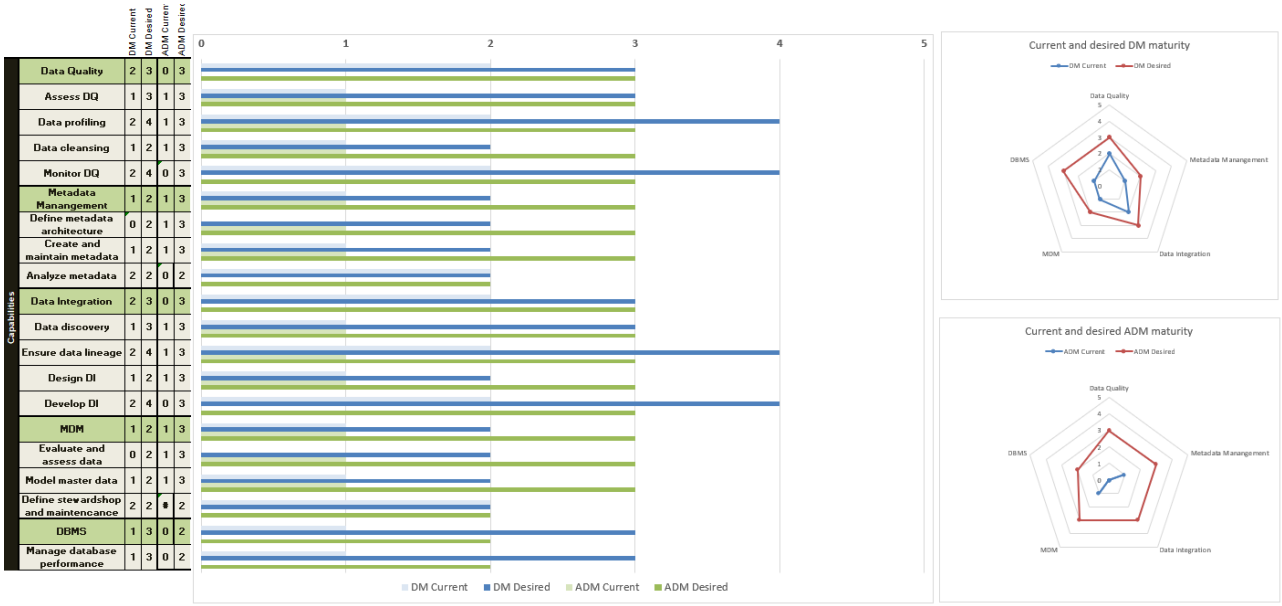


Figure 15: Maturity Assessment Tool v1.0 Results Tab

4. Evaluation and Refinement

This chapter covers the evaluation of the first version of the ADM Maturity Model and presents changes towards the second version. Section 4.1 presents the mixed-method validation strategy. Section 4.2 presents the evaluation through expert interviews. Section 4.3 presents the second version of the ADM Maturity Model and motivates the changes that resulted from the expert interviews. Section 4.4 presents the evaluation through multiple case studies.

4.1 Mixed-Method Validation Strategy

The mixed-method validation strategy of the present research follows the Framework for Evaluation in Design Science (FEDS) [96] evaluation design process, consisting of four steps: (1) explicate the goals of the evaluation, (2) choose the evaluation strategy or strategies, (3) determine the properties to evaluate, and (4) design the individual evaluation episode(s). The remainder of this section describes these steps in more detail.

The goal of validation research in design science is to predict the effects of applying the artifact in a real-world problem context [20]. In the context of the present research, the goal is to predict the effect of applying the ADM Maturity Model by data management consultants to perform a maturity assessment within an organization. We choose to apply the human risk & effectiveness, as this is the most suitable for a socio-technical artifact with uncertainties about social and use issues and a need to establish effectiveness in real use [96]. The properties to evaluate are the evaluation criteria by Salah et al. [24] as presented in Section 1.2.4.

The validation model consists of the model of the artifact, interacting with a model of the problem context. The model of the artifact is the ADM Maturity Model. The context is modeled in two validation episodes concerning the stakeholders that interact with the maturity model: expert interviews with data management consultants and case studies with participating organizations. Figure 16 places both evaluation episodes on the two-dimensional characterization of the FEDS [96].

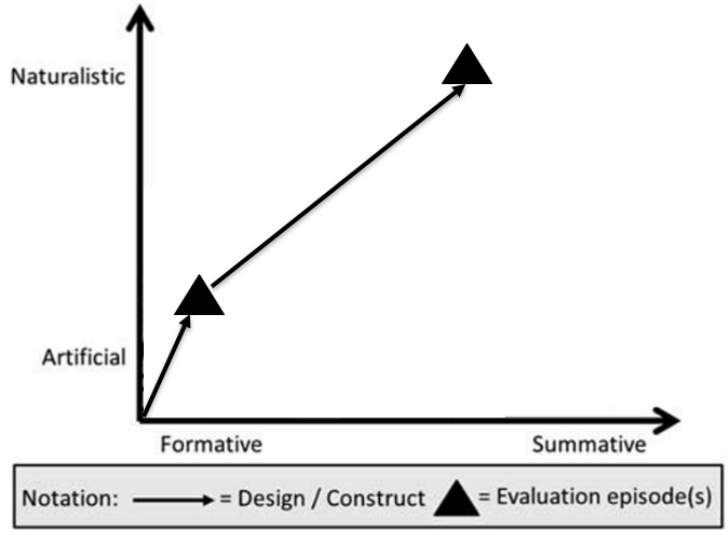


Figure 16: Evaluation Episodes, Based on [96]

The first validation episode consists of expert interviews with data management consultants. This stakeholder group is the intended user of the ADM Maturity Model and has experience with performing maturity assessments at multiple organizations. The purpose of this evaluation is formative, as the goal is to improve the maturity model. The evaluation paradigm is artificial, as the model of the context is the data management consultant visualizing applying the maturity model in a real-world context and providing their opinion on the evaluation criteria based on that. Multiple interviews with different data management consultants are performed to improve the validity and generalizability of the findings.

The second validation episode consists of case studies with organizations that want to improve their data management capabilities. This stakeholder group is the functional beneficiary of the ADM Maturity Model. For this validation, the ADM Maturity Model is improved based on the evaluation in the previous episode. The purpose of this evaluation is more summative, as it aims to conclude the research. The paradigm is naturalistic, as it explores the performance of the maturity model by assessing the real-world processes of the participating organization. Multiple case studies at different organizations are performed to improve the validity and generalizability of the findings.

4.2 Expert Interviews

4.2.1 Interview Participants

The first validation episode consists of expert interviews with data management consultants, the intended users of the ADM Maturity Model. As they are the intended user, their opinion on the understandability, ease of use, usefulness, and practicality is essential. Table 23 presents the participants of the expert evaluation. The participants in bold participated in the first round of interviews and are experts in AI and data management. Partaking from this group is essential, as they can validate that the maturity model is built on their expertise, which is also known as member checking [20].

Participants are asked to only comment on aspects they consider themselves experts in. For example, less experienced consultants are asked to focus on understandability and ease of use. Simultaneously, more experienced consultants are also invited to focus on the usefulness, practicality, and content of the model. Each participant is given a reference number, which is used to substantiate changes in the model pas presented in Section 4.3, based on their statements during the interview, transcribed in Appendix E.

Reference	Country	Position	Date Conducted
1	Netherlands	Consultant Enterprise Data Management	12-08-2020
2	Netherlands	Manager Enterprise Data Management	27-08-2020
3	Netherlands	Consultant Enterprise Data Management	14-08-2020
4	Netherlands	Analyst Enterprise Data Management	31-08-2020
5	Netherlands	Senior Consultant Enterprise Data Management	31-08-2020
6	Netherlands	**Senior Manager Oracle EDM**	31-08-2020
7	Netherlands	Analyst Enterprise Data Management	01-09-2020
8	Netherlands	Consultant Enterprise Data Management	03-09-2020
9	UK	**Senior Consultant Analytics & Cognitive**	03-09-2020
10	Netherlands	**Manager Enterprise Data Management**	02-09-2020
11	Netherlands	Manager Enterprise Data Management	10-08-2020

Table 23: Interview Participants

4.2.2 Interview Protocol

To critically evaluate every aspect of the model, the evaluation template by Salah et al. is used [24]. This template includes a set of evaluation criteria for the maturity levels, processes, and the use of maturity model itself. The template proposes a set of statements to score these criteria on a 5-point Likert scale and a set of open questions to identify potential improvements.

The interview is structured by using the ADM Maturity Assessment Tool and used the following protocol:

1. **Preparation**: The interview participant received the ADM Maturity Assessment Tool in advance and was asked to read the tool and evaluation criteria briefly. The participant was asked to evaluate the tool as if he/she had to apply it based on the evaluation criteria.
2. **Evaluating tabs**: During the interview, every tab of the tool was discussed with the participant in chronological order: Introduction, Maturity Levels, Data Quality, Metadata Management, Data Integration, Master Data Management, Database Management, Results, and Evaluation. The participant was asked to provide feedback on every aspect of the maturity model and tool.
3. **Evaluating processes**: For every capability tab, the participant was asked to answer questions 3, 4, and 5 from the evaluation template. The open questions from the evaluation template are presented at the end of this section.
4. **Rate evaluation criteria**: At the Evaluation tab, the participant was asked to rate a set of statements regarding the evaluation criteria on a 5-point Likert scale. The evaluation statements are presented in Section 4.2.4
5. **Open questions**: The participant was asked to answer the other open questions from the evaluation template, provided that these had not yet been covered when discussing the individual tabs.

The template proposes the following open questions:

> Q1. Would you add any maturity levels? If so, please explain what and why?
>
> Q2. Would you update the maturity level description? If so, please explain what and why?
>
> Q3. Would you add any processes or practices? If so, please explain what and why?
>
> Q4. Would you remove any of the processes or practices? If so, explain what and why?
>
> Q5. Would you redefine/update any of the processes or practices? If so, please explain what and why?
>
> Q6. Would you suggest any updates or improvements related to the scoring scheme? If so, please explain what and why?
>
> Q7. Would you suggest any updates or improvements related to the assessment guidelines? If so, please explain what and why?
>
> Q8. Would you like to elaborate on any of your answers?
>
> Q9. Could the model be made more useful? How?
>
> Q10. Could the model be made more practical? How?

4.2.3 Qualitative Data Analysis

The interview transcripts are analyzed using the approach described in Section 3.1.5: reading and annotating, categorizing data, and corroborating evidence. The initial set of categories consists of the evaluation criteria. For relevant data outside of these criteria, additional categories are created. The expert interviews and case study transcripts and analysis findings are also peer debriefed to increase validity. The peer noted that the interview contained questions specifically tailored at the evaluation criteria, and therefore the labeling and interpretations are straightforward and valid.

4.2.4 Evaluation Criteria

During the third step of the interview protocol, the participants are asked to rate a set of statements regarding the evaluation criteria of the model. Table 24 presents these statements, along with the results from 11 interviews. The lowest score, median, and average scores are presented to reflect the distribution.

Criteria (N=11)	*Min*	*Med.*	*Avg.*
Maturity Levels			
The maturity levels are sufficient to represent all maturation stages of the domain (*Sufficiency*)	4	5	**4.5**
There is no overlap detected between descriptions of maturity levels (*Accuracy*)	2	4	**4.2**
Processes and Practices			
The processes and practices are relevant to the domain (*Relevance*)	3	5	**4.5**
Processes and practices cover all aspects impacting/ involved in the domain (*Comprehensiveness*)	3	4	**4.3**
Processes and practices are clearly distinct (*Mutual Exclusion*)	3	4	**4.1**
Processes and practices are correctly assigned to their respective maturity level (*Accuracy*)	4	5	**4.8**
Maturity Model			
Understandability			
The maturity levels are understandable	4	5	**4.8**
The assessment guidelines are understandable	4	5	**4.8**
The documentation is understandable	4	5	**4.8**
Ease of Use			
The scoring scheme is easy to use	4	5	**4.7**
The assessment guidelines are easy to use	4	5	**4.9**
The documentation is easy to use	4	5	**4.9**
Usefulness and Practicality			
The maturity model is useful for conducting assessments	4	5	**4.8**
The maturity model is practical for use in industry	4	4.5	**4.5**

Table 24: Evaluation Criteria Scores from Interviews (N=11)

4.3 ADM Maturity Model Version 1.1

This section presents the final version of the model and the improvements made based on expert evaluation. Section 4.3.1 to 4.3.9 present version 1.1 of the ADM Maturity Model through the topics of the Maturity assessment tool, Section 4.3.10 and 4.3.11 propose recommendations following the assessment. Statements from the expert evaluation substantiate each change in the model. The references point to one of the experts from Table 23 and the labeled statements can be found in Appendix E. The detailed list of changes and motivation can be found in Appendix F. The focus of these changes is to improve the model; positive and confirmative feedback is not mentioned explicitly and is only reflected in the evaluation scores in Table 24.

4.3.1 Introduction

The first tab with the introduction to the model and the assessment tool can be found in Figure 17. The instruction describes the course of an assessment: Assess each process, automatically calculate the maturity level of each (sub) capability, fill out the evaluation form, and present results after analysis. The second section presents the background and development of the model. The bottom half of the tab presents an overview of the capabilities, sub capabilities, and maturity levels.

The background is extended with the following sentences: '*Augmented data management is the application of AI to enhance or automate data management processes and decisions, which greatly enhances the speed, efficiency, and capacity to manage data beyond human ability*'. This addition improves the understandability and relevance of the model, as it underlines the motivation to apply augmented data management (interview 8).

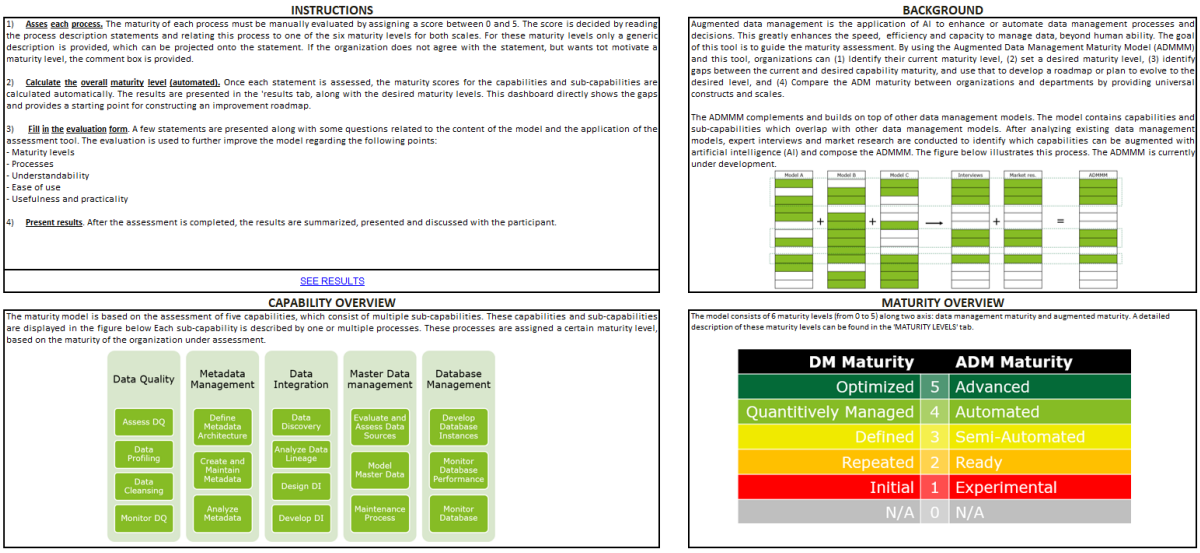


Figure 17: Introduction Tab of ADM Maturity Assessment Tool v1.1

4.3.2 Maturity Levels

When presenting the maturity levels, a definition of augmented data management is introduced. In the first version of the model, the following definition was used: *'Augmented data management is the human-centered application of artificial intelligence to enhance data management capabilities.'* The definition caused some confusion on the meaning of augmentation (interview 2,4) and AI (interview 1,3,4). This confusion sparked the discussion on how detailed the definition of augmentation and AI needs to be for a participant to participate in the assessment without making it unnecessarily complicated.

The capability's processes describe 'what' is done, not 'how' it is done. Therefore, a participant does not need to have in-depth knowledge on 'how' AI works, yet needs to understand the definition of augmented data management. The maturity levels for augmentation describe the different stages of augmentation. On the lowest level, the process is being executed manually. On the highest level, the process is being executed by AI. 'What' AI executes can be divided into tasks and decisions. Manual tasks can be automated, and decisions can be supported by or made by AI autonomously. The definition was changed to reflect these functionalities:

'Augmented data management is the application of AI to enhance or automate data management processes and decisions.'

Figure 18 displays the maturity levels and their description. A detailed overview of the changes and motivation can be found in Appendix F. Maturity level 0 changed from 'incomplete' to N/A to improve sufficiency. Each maturity scale now has its own description of level 1 to improve understandability. Within the description for maturity level 3, 'recommendations' is added to improve the sufficiency. Also, governance is added to the process description to improve the comprehensiveness.

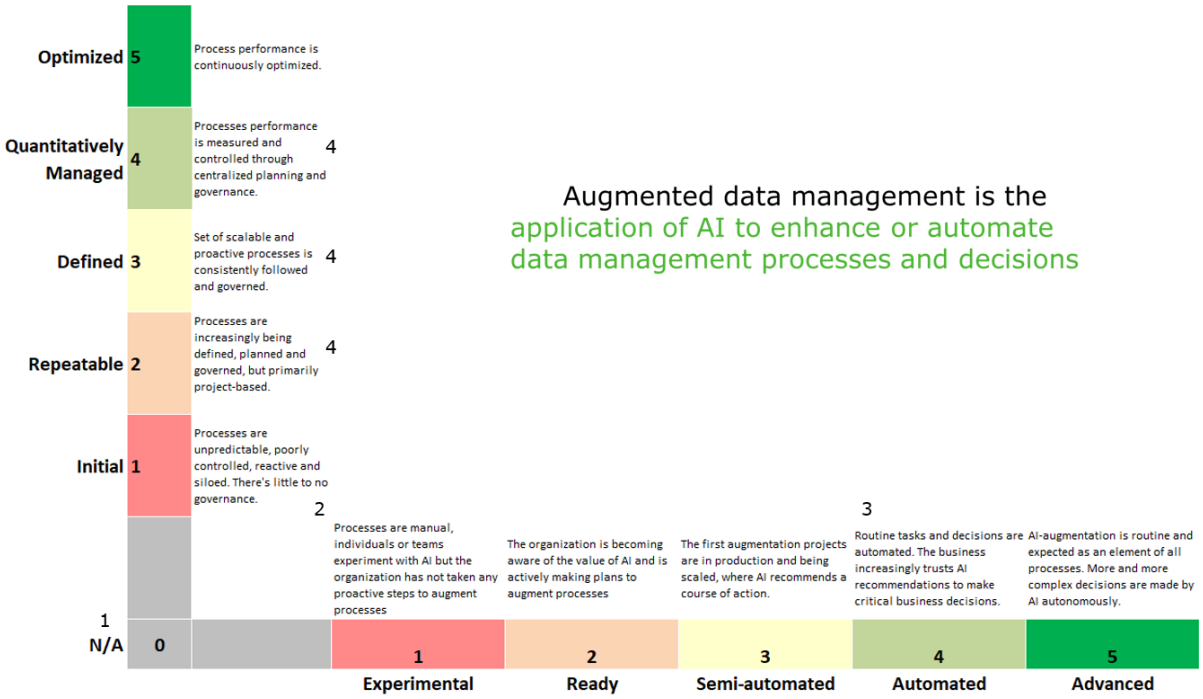


Figure 18: Maturity Levels and ADM Definition

4.3.3 Data Quality

Table 25 presents the overview of the data quality sub capabilities and processes. Changes are indicated by bold text and numbered between brackets, and the final column refers to an expert interview. A detailed overview of the changes and motivation can be found in Appendix F.

Sub-capability	Process	References
Assess DQ	Specify data quality dimensions**/requirements (1)** and data validation rules **(2) (3)**	**2,4,5,9**
	Assess **(**large**) (4)** datasets to generate a data quality score based on statistics and data rules **(2)**	**2,5,6,9**
Data Profiling	Profile **(**large and complex**)** datasets by parsing the data and recognizing data types, structures, metadata, data categories (e.g. email, address) and generate basic statistics **(4)**	**5,6**
Data Cleansing	Categorize data quality issues	
	Suggest actions for data cleansing and standardization	
	Learn from manual data cleansing to suggest **and/or** perform cleansing of similar DQ issues **(5)**	**5**
Monitor DQ	Perform ongoing data quality/validation checks on data pipelines **and/or data mutations (6)**	**5,9**
	Detect anomalies **(i.e. significant differences) (7)** between actual data and expected values from historical data	**5**

Table 25: ADM³ v1.1 Data Quality Overview and Changes

The understandability of the capability is improved by switching the first and second processes into a logical order, by adding /requirements, and by placing 'significant differences' between brackets. Adjectives such as 'large' and 'complex' and terms like 'reverse-engineer' are removed to improve process accuracy. Such wording implies a particular execution of processes, while these are meant to be general. 'Or' is changed to 'and/or' to improve the process accuracy as well. 'Data mutations' is added to cover all monitoring processes and therefore improve the comprehensiveness.

4.3.4 Metadata Management

Table 26 presents the overview of the data quality sub capabilities and processes. Changes are indicated by bold text and numbered between brackets, and the final column refers to an expert interview. A detailed overview of the changes and motivation can be found in Appendix F.

Sub-capability	Process	References
Define Metadata Architecture	Generate metamodels based on the data	
	Specify (1) metadata (2) rules based on existing datasets	**2, 5,7**
	Rationalize metadata repositories **(e.g. data dictionaries, business glossary, data catalog) (3)** through industry and client specific taxonomies **(4)**	**5,7,9**
Create and Maintain Metadata	Generate metadata from structured data: identify topics, taxonomies (recognize names, address, email, etc.)	
	Generate metadata from unstructured data (5): e.g. text, images, video	**5**
	Catalog and index metadata in a repository **(e.g. data catalog) (3)**	**5,7,9**

	Merge business and technical **meta**data (6) based on similarity and relation	**5,7**
	Create data lineage metadata by tracking data flows and transformations (7)	**4,8**
Analyze Metadata	**Highlight potential mismatches and/or missing metadata and resolve accordingly (8)**	**3,4,5**
	Analyze data flows and transformation logs to check whether data is moved or mapped as expected (9)	**5,7,8**
	Analyze metadata in order to generate insights e.g. into data usage and patterns (10)	**5,7**

Table 26 ADM³ v1.1 Metadata Management Overview and Changes

The order of processes is changed to follow a logical order. The mutual exclusion is improved by explicitly mentioning metadata instead of (meta) data. The understandability is improved by removing unnecessary or unknown terms such as 'folksonomies' and 'technical session logs' and by unifying definitions. 'Data dictionaries, business glossary, and data catalog' and 'analyze metadata in order to generate insights' are added to improve process comprehensiveness. 'Reverse-engineer' is replaced by 'specify' to improve process accuracy.

4.3.5 Data Integration

Table 27 presents the overview of the data quality sub capabilities and processes. Changes are indicated by bold text and numbered between brackets, and the final column refers to an expert interview. A detailed overview of the changes and motivation can be found in Appendix F.

Sub-capability	Process	Reference
Data Discovery	Recommend **additional and/or alternative (1)** datasets during discovery, based on relationships and similarity in data	**5**
Analyze (2) Data Lineage	Generate end-to-end data lineage **e.g. from code, metadata and/or data integrations (3)**	**5,6**
	Perform impact/risk analysis **to identify system and data dependencies (4)**	**4,9**
	Perform root-cause analysis	
Design DI	Model data for structured data, discovering attributes and structure	
	Model data for unstructured data, discovering attributes and structure	
	Map source data model to target data model	
Develop DI	**Develop data flows: ingest, validate and transform data to the target structure (5)**	**5,6,7**
	Develop data flows based on existing integrations and user interaction (6)	**5,6**

Table 27: ADM³ v1.1 Data Integration Overview and Changes

The comprehensiveness is improved by adding 'and/or' and by removing 'reverse-engineer' to cover all relevant processes. Some processes, such as 'analyze data lineage' and 'develop data flows', are reformulated to cover the relevant processes and exclude others. For risk/impact analysis, an example is given to improve the understandability and underline the difference with root-cause analysis.

4.3.6 Master Data Management

Table 28 presents the overview of the data quality sub capabilities and processes. Changes are indicated by bold text and numbered between brackets, and the final column refers to an expert interview. A detailed overview of the changes and motivation can be found in Appendix F.

Sub-capability	Process	Reference
Evaluate and Assess Data Sources	**Scan data sources and identify potential master data entities (1)**	**5**
	Identify relationships and overlap between datasets (2)	**5**
	Generate a trust score for potential master data (3)	**9**
Model Master Data	Generate a master data model by recognizing entities and hierarchical structure (bottom-up)	
	Detect and map data entities onto a predefined master data model (top-down)	
	Configure **master data management hub/tool (4)** as per data model	**2,4,5**
Maintenance Process (7)	**Learn match and merge rules from human labeling of master data (5)**	**5,6,8,9**
	Match and/or merge duplicate data to establish a single point of truth (6)	**8,9**

Table 28: ADM³ v1.1 Master Data Management Overview and Changes

The process comprehensiveness is improved by adding three processes that were identified as missing and by spitting a process. To improve understandability, 'clustering and blocking attributes' is simplified into 'match and merge rules', and '\tools' is added to another process. The last sub capability is renamed only to comprehend the maintenance processes

4.3.7 Database Management

Table 29 presents the overview of the data quality sub capabilities and processes. Changes are indicated by bold text and numbered between brackets, and the final column refers to an expert interview. A detailed overview of the changes and motivation can be found in Appendix F.

Sub-capability	Process	Reference
Develop Database Instances (1)	**Recognize data model to construct logical and physical database design (1)**	**2,3,4**
Manage Database Performance	Optimize queries for better response time	
	Provision, manage, patch, backup, recover and tune databases	
	Forecast computational demand and scale resources or (re)schedule jobs to meet performance criteria (2)	**3**
Monitor Database (3)	**Optimize storage performance and costs by monitoring data usage (4)**	**2,7,8**
	Detect anomalies in database logs for threat detection and/or fault recovery	**4,8**

Table 29: ADM³v1.1 Database Management Overview and Changes

Two processes are added to improve the comprehensiveness. The understandability is improved by simplifying the formulation of one process and introducing a new sub capability 'Monitor Database', to align with other capabilities that also monitor sub capabilities.

4.3.8 Universal Capabilities

Multiple participants noted an overlap between the processes of different capabilities, which might give some confusion (interview 3,4,5,6,7,8). One proposed solution is to mention the overlap of those processes. Simultaneously, this could make it more complicated, as you would refer to processes that have not yet been covered. Some participants suggested to include universal capabilities (interview 6,7,8) by combining processes from all capabilities. These universal capabilities add another dimension to the model, as their maturity influences the maturity of all other capabilities.

Sub-capability	Process
Specify Data Rules	Specify data quality dimensions/requirements and data validation rules (DQ)
	Specify metadata rules based on existing datasets (MD)
Data Modelling	Generate metamodels based on the data (MD)
	Model data for structured data, discovering attributes and structure (DI)
	Model data for unstructured data, discovering attributes and structure (DI)
	Generate a master data model by recognizing entities and hierarchical structure (bottom-up) (MDM)
	Recognize data model to construct logical and physical database design (DBMS)
Validation Checks	Assess (large) datasets to generate a data quality score based on statistics and data rules (DQ)
	Develop data flows: ingest, validate and transform data to the target structure (DI)
Similarity Identification	Merge business and technical metadata based on similarity and relation (MD)
	Recommend additional and/or alternative datasets during discovery, based on relationships and similarity in data (DI)
	Identify relationships and overlap between datasets (MDM)
Monitoring	Perform ongoing data quality/validation checks on data pipelines and/or data mutations (DQ)
	Detect anomalies (i.e. significant differences) between actual data and expected values from historical data (DQ)
	Analyze data flows and transformation logs to check whether data is moved or mapped as expected (MD)
	Optimize storage performance and costs by monitoring data usage (DBMS)
	Detect anomalies in database logs for threat detection and/or fault recovery (DBMS)

Table 30: ADM³ v1.1 Universal Capabilities Overview

4.3.9 Results

The results of the maturity assessment are displayed, as shown in Figure 19. There are two important changes in calculating the maturity level of the (sub) capability. First, the maturity scores are rounded at one decimal when calculating the average score. This calculation gives a more accurate representation of the real maturity and gap between the current and desired score. Second, the maturity level of the whole capability is now the average of all the sub-capabilities instead of the average of all processes. All sub capabilities are equally important; this is now reflected in the results (interview 1).

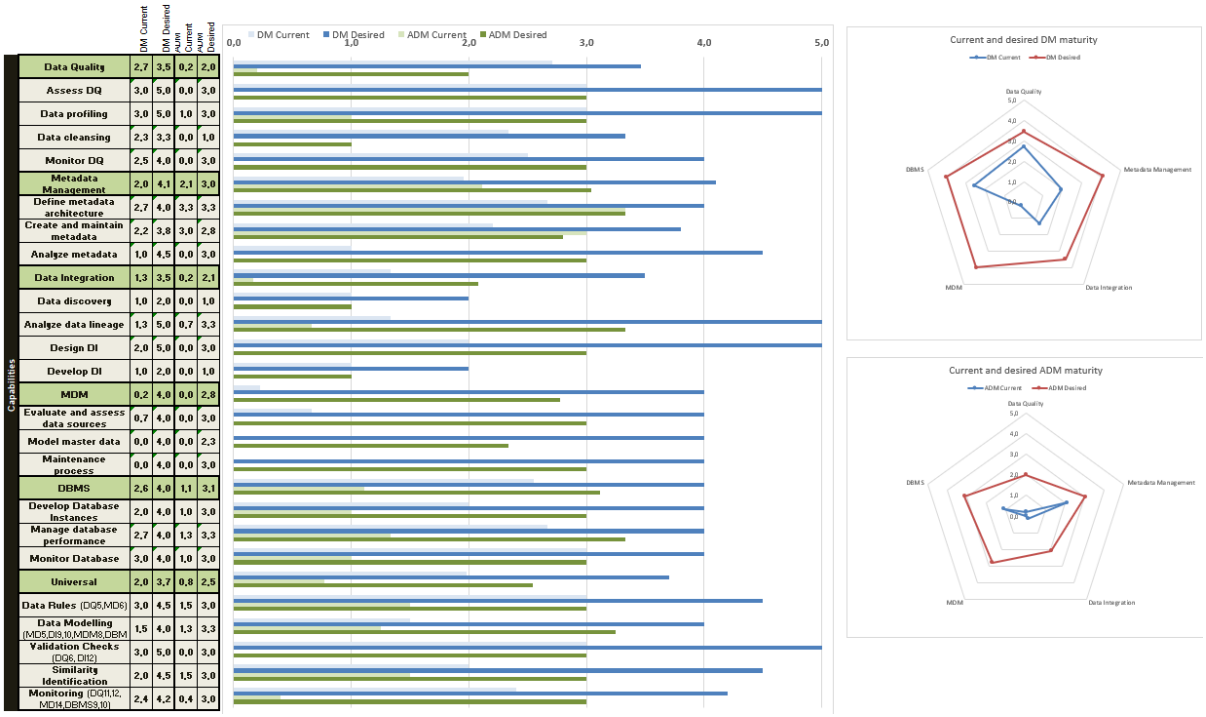


Figure 19: ADM Maturity Assessment Tool 1.0 Results Tab

4.3.10 Improving Capabilities

The maturity assessment is the first step in improving data management capabilities. Multiple participants indicated that it would be useful to provide guidelines on constructing an improvement roadmap, based on the assessment (interview 1,6,7). Following this advice, the choice is made to present general and high-level recommendations on improving capabilities in Section 4.3.10 and constructing an improvement roadmap in Section 4.3.11. These recommendations can be followed by the assessor (researcher, data management consultant) or the organization being assessed.

The gaps identified in the current and desired maturity provide improvement opportunities. The maturity assessment processes describe what needs to be done yet does not prescribe how to do it. The 'how' is dependent on the organization itself and cannot be captured in a universal model. Therefore, the maturity model can only be used to outline what needs to be done. Section 4.3.11 goes into detail on defining the 'how' and constructing an improvement roadmap. The section below outlines the difference between maturity levels and general steps on how to bridge this difference.

Steps to improve Data Management Process Maturity:

From 0 to 1: Identify experts within the business unit that can perform the process with limited tools when problems arise. Champions/heroes can be assigned within the business unit.

From 1 to 2: Create awareness around the basic data management topics. Define some roles, responsibilities, and processes within the business unit and/or during projects. Involve relevant stakeholders and introduce a consistent toolset.

From 2 to 3: Define organization-wide roles, scalable processes, and verify those with stakeholders. Implement policies, standards, and checks to ensure adherence.

From 3 to 4: Define process metrics/KPI's to track process performance. Ensure centralized planning and governance.

From 4 to 5: Leverage the collected metrics/KPI's to improve processes.

Steps to improve Augmentation Maturity:

From 0 to 1: Identify AI-experts and draft the first use cases for augmentation.

From 1 to 2: Ensure organizational support, data availability, and plan the first augmentation pilots. Prove the potential value of AI by presenting a proof of concept.

From 2 to 3: Make tactical investments to enable the relevant skills, technology, and data storage to start realizing the plans and scaling up the pilots. Establish a center of excellence to share AI experts, best practices, and technology throughout the organization.

From 3 to 4: Augment an end-to-end process. Consider employing AI as a value driver for all projects regarding process improvement.

From 4 to 5: Expand augmentation and succeed with high-risk/high-return use cases.

4.3.11 Improvement Roadmap

To identify how organizations can close the gap between the current and desired maturity, an improvement roadmap is constructed. Constructing a roadmap is the process of determining the actions, steps, and resources needed to implement improvement. While the actual implementation of these improvements is outside of the scope of a maturity model, the model can assist in the development of the improvement roadmap. Based on IBM DGMM [64] and DAMA DMBOK [4], the ADM Maturity Model proposes the following steps in building a roadmap:

1. **Summarize the assessment findings and define scope:** The results tab of the assessment tool presents an overview of the current and desired maturity of all (sub) capabilities. The summary focuses on the gaps between the two. A low maturity score is not necessarily bad, as some processes might not be of importance or do not occur at all. Next to the gaps, current strengths, transcendent capabilities, and other notable results can be highlighted. The scope of the improvement roadmap should clearly be defined before identifying potential improvements. By defining the scope, efforts can be focused on the most important (sub) capabilities and processes.
2. **List the key people, process, and technology initiatives necessary to bridge the gap:** For each (sub) capability, key improvement initiatives regarding people, process, and technology are listed. These initiatives need to be devised in collaboration with all relevant stakeholders. Examples of such initiatives are assigning specific roles and tasks, defining metrics to monitor process performance, and implementing new IT systems. Initiatives can be rated on expected risks, costs, and benefits. For large and important initiatives, a more detailed business case is recommended.
3. **Prioritize key initiatives and construct a timeline**: The identified initiatives must first be prioritized based on risk, costs, benefits, or other factors important to the organization. A timeline is constructed for implementing these initiatives, including oversight activities. Figure 20 presents an example timeline with the key initiative to implement data governance and master data management.

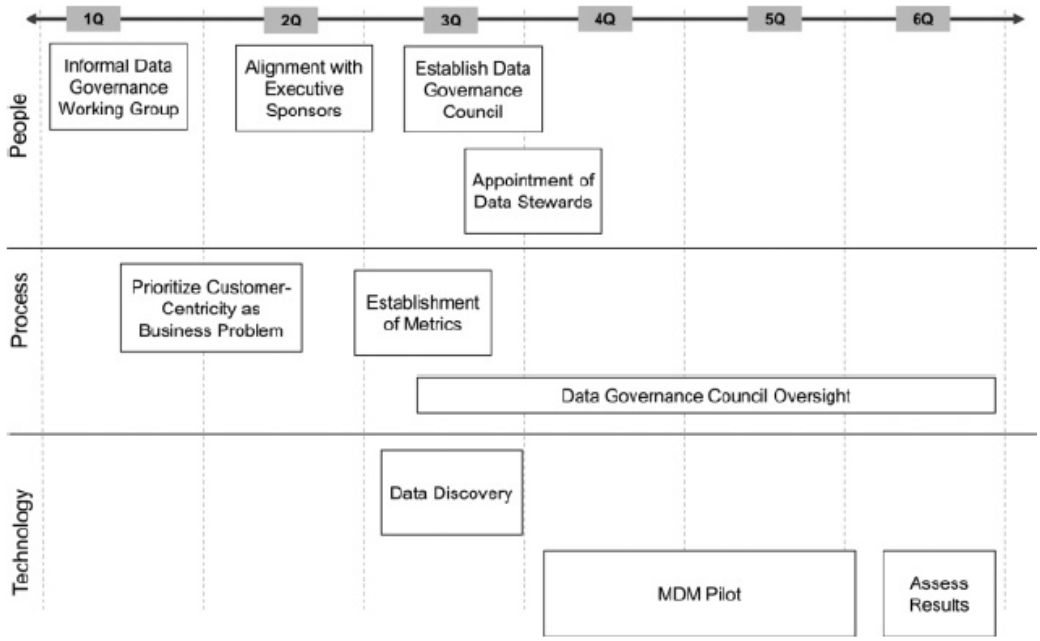


Figure 20: Example Timeline for Implementing Master Data Management, source [64]

4.3.12 Other Expert Feedback

Next to the feedback that motivated changes in the model as presented in the previous section, this section presents relevant feedback that did not directly result in changes. In part, this is feedback that did not find support amongst the other participants. For example, participant 2 motivated that data profiling should be part of the sub-capability assessing data quality, while participants 6, 8, and 9 indicated that they consider it a separate sub-capability. Other feedback lead to a small discussion during the interview. For example, participant 6 questioned the difference in the process for handling structured and unstructured data. After indicating the difference in approach for applying AI, the participant agreed those are separate processes.

Other feedback that has not led to changes can still be considered for future versions of the model. Participant 11 suggested that during the assessments, participants could suggest missing processes and/or sub-capabilities. Participants 2 and 4 indicated that it could help provide an example of processes at maturity level 1 and 5, so it would be easier to relate. Participant 5 suggested that having an online tool would be better than Excel. Participants 10 and 11 indicated that governance and adherence to policies should be part of the model as well. In theory, a process could be at the highest maturity level and still yield no benefits if it is not governed. Besides, those participants indicated that there are options to make the model more objective, for example, by quantifying a percentage of processes or by calculating deviations from a reference model.

4.4 Case studies

4.4.1 Case Study Participants

The second validation episode consists of case studies with organizations that want to improve their data management capabilities, the functional beneficiary of the ADM Maturity Model. These case organizations are selected to represent typical Deloitte Enterprise Data Management customers: large enterprises with vast amounts of business-critical data. The organizations that participated in the case studies are a health insurer, a bank, and a (non-health) insurer. Each case study consisted of one or multiple assessment interviews and a follow-up meeting to discuss the results. Table 31 presents an overview of all case study meetings and participants. Section 4.4.2 presents the case study protocol.

Meeting	Case	Position	Date
Assessment	1: Health Insurer	Corporate Data Steward	07-09-2020
Assessment	1: Health Insurer	Data Quality Expert	08-09-2020
Assessment	1: Health Insurer	Data Quality Expert	10-09-2020
Assessment	2: Bank	Business Consultant Data Management	11-09-2020
Assessment	3: Insurer	Head of Data Management	11-09-2020
Follow-up	1: Health Insurer	All Participants	22-09-2020
Follow-up	3: Insurer	Single Participant	25-09-2020
Follow-up	2: Bank	Single Participant	28-09-2020

Table 31: Overview of All Case Study Meetings

4.4.2 Case Study Protocol

During the case study, the researcher applied the ADM Maturity Model through the assessment tool, as a data management consultant would. The goal of the case study is threefold: Primarily to test the functional requirements (Section 1.2.4): that the maturity model can be used to assess the current processes of an organization and can be used to formulate improvements. Secondary, to evaluate the non-functional requirements or evaluation criteria as presented in Section 4.2.4. Third, to evaluate the recommendations on improving capabilities and constructing an improvement roadmap. The following protocol was used for the application of the ADM Maturity Model Assessment Tool:

1. **Selection:** Participants are selected for knowing the relevant data management capabilities of their organization. If participants lack knowledge of one or multiple capabilities, they were asked to refer to colleagues that can fill this gap.
2. **Preparation:** Each participant received the assessment tool in advance and was recommended to read it as preparation.
3. **Assessing Processes:** The participant was introduced to the model, its capabilities, and maturity levels. The participant was then asked to assess each process on both data management and augmented data management maturity.
4. **Evaluating Processes**: For every capability tab, the participant was asked to answer questions 3, 4, and 5 from the evaluation template.
5. **Rate Evaluation Criteria**: At the Evaluation tab, the participant was asked to rate a set of statements regarding the evaluation criteria on a 5-point Likert scale. The evaluation statements correspond to the ones from the expert interviews and are presented in Section 4.4.6.

6. **Open Questions**: The participant was asked to answer the other open questions from the evaluation template, provided that these had not yet been covered when discussing the individual tabs.
7. **Return Results:** Following the interview, the participant received the completed assessment as validation.
8. **Follow-up:** Two weeks after the assessment, a follow-up meeting was planned to discuss the results and provide the general recommendations, as presented in Section 4.3.10 and 4.3.11.
9. **Evaluate Recommendations:** At the end of the follow-up meeting, the participant was asked to rate the recommendations on the maturity model evaluation criteria: understandability, ease of use, usability, and practicality

4.4.3 Case 1: Health Insurer

The first case study organization is a Dutch health insurer. From this organization, one Corporate Data Steward and two Data Quality Experts participated in the assessment interviews. The interview was conducted via video conference and correctly followed the case study protocol. Table 32 presents an overview of the data management process maturity. Table 33 presents the results of the augmented data management maturity. The results are split per participant, and an average per capability is calculated. The result of all sub capabilities can be found in Appendix H.

In general, the results per participant differ less than one maturity level. Larger deviations are often caused by participants skipping processes, especially in data integration and database management. On average, data quality is the most mature capability. Within data quality, data profiling and assessing data quality are the highest at 2.3 (DM) and 2.7 (ADM), with the desired maturity of level 4 for both. Data profiling scores the highest augmented maturity at 2.7 due to the use of data profiling tools. Metadata is the capability with the highest ADM maturity score, surpassing DM maturity on two of three sub capabilities. The case organization indicated that they just introduced an AI-powered data catalog, which they currently do not use to its full potential. Metadata analysis is the most immature sub capability with the largest gap to the desired maturity and therefore provides a large improvement opportunity for both DM and ADM. Master data management scores low in general, with 'evaluate and assess data sources' as the highest sub capability at 0.9 (DM) and 1.3 (ADM) due to some functionalities in the data catalog. Within universal capabilities, 'data rules' is the most mature. The other universal capabilities; validation checks, similarity identification, and monitoring score low. Improving these universal capabilities would result in a higher maturity in all other capabilities.

Participant	**Data Quality**		**Metadata mgmt.**		**Data Integration**		**Master data mgmt.**		**DBMS.**		**Universal Capabilities**	
1	2,2	3,3	1,5	2,9	1,7	3,2	1,0	2,8	-	-	1,5	2,6
2	2,7	3,5	2,0	4,1	1,7	5,0	0,2	4,0	2,6	4,0	2,0	3,7
3	1,4	3,5	1,0	3,8	0,2	3,3	0,7	3,6	-	-	0,9	3,3
Avg. DM/ADM (as-is | to-be)	**2,1**	**3,4**	**1,5**	**3,6**	**1,2**	**3,8**	**0,6**	**3,5**	**2,6**	**4,0**	**1,5**	**3,2**

Table 32: Assessment Results DM Maturity of Case 1

Participant	Data Quality		Metadata mgmt.		Data Integration		Master data mgmt.		DBMS.		Universal Capabilities	
1	1,2	3,3	1,4	3,4	2,2	4,0	0,7	2,7	-	-	0,9	2,6
2	0,2	2,0	2,1	3,0	0,3	3,2	0,0	2,8	1,1	3,1	0,8	2,5
3	1,3	3,1	1,7	3,9	1,0	3,7	0,7	3,0	-	-	1,2	3,5
Avg. DM/ADM (as-is | to-be)	**0,9**	**2,8**	**1,8**	**3,4**	**1,2**	**3,6**	**0,4**	**2,8**	**1,1**	**3,1**	**1,0**	**2,9**

Table 33 Assessment Result ADM Maturity of Case 1

4.4.4 Case 2: Bank

The second case study organization is an international bank headquartered in the Netherlands. The interview participant is a Business Consultant with a focus on data management. The interview was conducted via video conference and followed the case study protocol. Table 34 presents the assessment results of both DM and ADM maturity. The result of all sub capabilities can be found in Appendix H.

The largest gap to the desired state within data quality is within data cleansing at 1 level for DM and 1.7 level for ADM. Monitor data quality scores the highest at level 3 for both DM and ADM. Metadata is the highest scored capability, also compared to the other case organizations. Analyze metadata lacks behind at level 1 for both DM and ADM, while the desired maturity is at level 2.5 and 2 respectively. Data integration and master data management score relatively low, with no to little ADM. The gap for all DM and ADM sub capabilities is relatively small, around 1 level, which might provide short term improvement opportunities. The universal capabilities score relatively high, which is reflected in the relatively high maturity levels compared to the other case organizations. These universal capabilities can also be leveraged in data integration and master data management. Mature universal capabilities also pave the way for augmentation.

Capability	DM		ADM	
Data Quality	1,9	2,7	1,4	2,3
Metadata Management	2,1	3,3	1,8	2,5
Data Integration	1,6	2,6	0,5	1,3
MDM	1,3	2,3	0,0	0,8
DBMS	2,3	3,3	0,3	1,3
Universal	2,0	2,8	1,1	1,9

Table 34: Assessment Result of Case 2

4.4.5 Case 3: Insurer

The third case study organization is an international insurance and financial services provider headquartered in the Netherlands. The interview was conducted via video conference with the Head of Data Management and followed the case study protocol. Table 35 presents the assessment results of both DM and ADM maturity. The result of all sub capabilities can be found in Appendix H.

In general, all capabilities have a gap of 1.5 level for DM and less than 1 level for ADM. Data quality has the highest maturity for DM at level 1.7 and has the highest desired maturity at 3.6, which indicates the importance and desire for improvement. Within metadata management, 'Create and Maintain Metadata' is the only sub capability with ADM at level 1.5, while the desired maturity is only 0.5. Analyze Metadata is the only sub capability within metadata management with a desired

ADM at level 1. Data Integration and Master Data Management are the only capabilities with a gap in current and desired ADM maturity, which indicates a desire for experimentation projects. The Database capability was not scored, as the participant stated that it is outside of his role. There is a desire for improvement in the universal DM and some ADM capabilities, which would improve the maturity of all other capabilities.

Capability	DM		ADM	
Data Quality	1,7	3,6	0,3	0,3
Metadata Management	1,1	2,7	0,5	0,5
Data Integration	1,2	2,5	0,0	0,7
MDM	1,1	2,5	0,0	0,9
DBMS				
Universal	1,3	2,7	0,1	0,5

Table 35: Assessment Result of Case 3

4.4.6 Evaluation of Maturity Model and Assessment Tool

During the fourth, fifth, and sixth step of the case study protocol, the participants are asked to rate a set of evaluation statements and to respond to the open questions from the evaluation template.

Two evaluation points recurred during the application of the model and at the open questions. The first is that all participants needed an additional explanation of at least one process. Having received the requested feedback, none of the participants named this as a disadvantage of the model. This type of interaction is expected, as the maturity model is intended to be used within expert assisted assessments. The second point covers the participant's profile. None of the participants was able to score every process of the maturity model. Many indicated that processes related to data integration and database management are outside of their function.

One participant noted that it would be valuable also to include data governance in the assessment. Another mentioned that it is unclear whether the most or least mature process should be used to rate the statements. Another participant noted that the model could be improved by making it more straightforward and the processes more mutually exclusive. Table 36 repeats the evaluation statements and presents the scores from the case study participants.

Criteria (N=4)	*Min*	*Med.*	*Avg.*
Maturity Levels			
The maturity levels are sufficient to represent, all maturation stages of the domain (*Sufficiency*)	3	4,5	**4,3**
There is no overlap detected between descriptions of maturity levels (*Accuracy*)	5	5	**5,0**
Processes and Practices			
The processes and practices are relevant to the domain (*Relevance*)	4	5	**4,8**
Processes and practices cover all aspects impacting/ involved in the domain (*Comprehensiveness*)	4	5	**4,8**
Processes and practices are clearly distinct (*Mutual Exclusion*)	2	4,5	**4,0**
Processes and practices are correctly assigned to their respective maturity level (*Accuracy*)	3	5	**4,5**
Maturity Model			

Understandability			
The maturity levels are understandable	4	5	**4,8**
The assessment guidelines are understandable	4	5	**4,8**
The documentation is understandable	4	5	**4,8**
Ease of Use			
The scoring scheme is easy to use	4	5	**4,8**
The assessment guidelines are easy to use	5	5	**5,0**
The documentation is easy to use	4	5	**4,8**
Usefulness and Practicality			
The maturity model is useful for conducting assessments	3	4,5	**4,3**
The maturity model is practical for use in industry	4	4,5	**4,5**

Table 36: Evaluation Criteria Scores from the Case Studies (N=4)

4.4.7 Evaluation of Recommendations

During the follow-up meeting, the result of the assessment was discussed with the participants. The results consist of the assessment scores as presented in Section 4.4.3, 4.4.4 and 4.4.5 and the recommendations on improving capabilities from Section 4.3.10 and the guidelines on constructing a roadmap from Section 4.3.11. During the final step of the case study protocol, the participants are asked to comment on the follow-up meeting and rate a set of statements regarding the evaluation criteria.

All participants indicated that the assessment was useful. Common feedback was that it is interesting to look at the different capabilities on a high level, compare with the other case studies, and revise how AI can play a role within data management. The improvement steps are perceived as easy to use, useful, and practical. The roadmap steps were considered too high level to be practical, limiting their usefulness and ease of use. Multiple participants indicated that these roadmap steps could be improved on the level of detail and cohesion with the improvement steps. One participant stated that it would be valuable if the improvement roadmap incorporates an agile way of working. Table 37 presents the evaluation statements and presents the score for the case study participants.

Criteria (N=5)	*Min*	*Med.*	*Avg.*
Maturity Model			
Understandability			
The improvement steps are understandable	5	5	**5**
The roadmap steps are understandable	3	4	**4**
Ease of Use			
The improvement steps are easy to use	4	4	**4**
The roadmap steps are easy to use	3	3.5	**3.5**
Usefulness and Practicality			
Improvement steps are useful for formulating improvement initiatives	4	4	**4**
The improvement steps are practical for formulating improvement initiatives	3	3	**3**
The roadmap steps are useful	4	4	**4**
The roadmap steps are practical	2	3	**2.8**

Table 37: Recommendation Evaluation Criteria Scores from the Case Studies (N=5)

5. Conclusion

This chapter concludes the research by answering the research questions that were set out at the beginning of the master thesis. The main research question was:

What constitutes a maturity model for Augmented Data Management that allows organizations to assess and improve their Data Management operations by leveraging AI?

The main research question was split up into four sub-questions, which are covered in Section 5.1-5.4. Section 5.5 concludes the research by answering the main research question. Section 5.6 discusses the contribution to research and Section 5.6 of the contribution to practice.

5.1 Augmented Data Management

The first sub-question is:

1. **How can Artificial Intelligence be leveraged to Augment Data Management capabilities?**
    a. What is Augmented Data Management?
    b. What is Artificial Intelligence?

Augmented data management is defined as the human-centered application of artificial intelligence to enhance data management capabilities. Computer systems perform tasks and decisions that are otherwise performed by humans. By dividing tasks and decisions over human and artificial intelligence, both can focus on the most value-adding tasks. Computers are inherently good at performing delineated and repeated activities, with a scalable and accurate execution. This functionality is incredibly valuable when processing large and complex data. Humans are inherently good at creative problem solving, which can be complemented by artificial intelligence.

Artificial intelligence refers to human intelligence in machines and consist of the following subfields: *Machine learning* algorithms learn to perform tasks without explicit instructions. *Natural language processing* enables a computer to understand and process human language. *Expert systems* incorporate knowledge gained from human experts into information systems. *Vison recognition* enables computers to understand and process visual images. *Speech recognition* enables a computer to process and produce spoken language. *Planning* allows computers to define an optimal sequence of actions. *Robotics* allows physical machines to perform actions automatically.

5.2 Existing Maturity Models

The second sub-question is:

2. **Which Data Management and artificial intelligence maturity models are available in current literature?**
    a. What does a Data Management model consist of, according to published literature?
    b. What does an Artificial Intelligence maturity model consist of, according to published literature?
    c. What are Data Management capabilities included in the reported models in the literature?

A systematic literature review was performed to identify all maturity models on data management and artificial intelligence in published literature. The search resulted in seven models for data management and four models for artificial intelligence. All models showed a relatively similar structure with four to six maturity levels and a set of (sub) capabilities and processes. All data management models contained an assessment method, while this was less common for AI models.

All seven data management maturity models combined present 32 capabilities, of which 13 overlap between two or more models. These include data quality, data architecture, data governance, stewardship, metadata management, master data management, data strategy, data storage and operations, information life cycle management, organizational structures, awareness, security, and technology infrastructure.

5.3 Maturity Model Development

The third sub-question is:

3. **How to design a maturity model for Augmented Data Management?**
    a. What are the maturity models' goals and requirements?
    b. Which method can be used to design a maturity model?
    c. Which method can be used to evaluate and validate a maturity model?

The functional requirements of the maturity model are that it must be able to assess the current state of capabilities and that improvement measures can be derived from the application of the model. The non-functional requirements were formulated as design goals and used as evaluation criteria. These criteria include maturity level *sufficiency* and *accuracy*, *process relevance*, *comprehensiveness*, and *mutual exclusion*, maturity model *understandability*, *ease of use, usefulness, and practicality*.

Various methodologies, methods, and guidelines exist on maturity model design and validation. Design science and action research are the most used established methodologies, while 'ad hoc' methodologies are the most popular. The most used methods are literature review, interviews, case studies, focus groups, and surveys. Other methods include Delphi studies, workshops, and various analytical methods. Different authors propose guidelines, based on methodologies, that detail exact steps for developing a maturity model. The authors that introduced the most adopted and most-cited guidelines are Becker et al. [15], de Bruin et al. [16] and Hevner et al. [83].

For the ADM Maturity Model design, a decision is made to apply the guidelines of Becker et al. because they are based on design science research and have the highest adoption in the scientific community. The development strategy is to combine existing data management and artificial intelligence maturity models. The maturity levels, (sub) capabilities, and processes are systematically compared and synthesized. Sub capabilities and processes that can be augmented are identified by performing interviews with AI and data management experts. The first version of the maturity model is evaluated and validated in expert interviews with the main stakeholders. The improved and final version of the model is validated in multiple case studies.

5.4 ADM Maturity Model

The fourth sub-question is:

4. **What constitutes the ADM maturity model?**
    a. Which maturity levels and definitions can be distinguished?
    b. Which capabilities can be distinguished?
    c. How to perform a maturity assessment?

The ADM Maturity Model consists of two maturity axes with five maturity levels. One axis is for data management process maturity and consists of the levels *initial, repeatable, defined, quantitatively managed, and optimized*. The other axis covers augment data management maturity and consists of the levels *experimental, ready, semi-automated, automated, and advanced.*

The ADM Maturity Model consists of five capabilities: *data quality, metadata management, data integration, master data management, and database management*. Each capability consists of multiple sub-capabilities and processes that are relevant to the respective domain. Next to these domain-specific capabilities, the model consists of the following universal capabilities: *define data rules, data modeling, validation checks, similarity identification, and monitoring*.

The ADM Maturity Model can be operationalized with the assessment tool, which can be used to structure an expert assisted maturity assessment. The assessment tool introduces the maturity model, the maturity levels, and can be used to assess the (sub) capabilities and processes. The tool presents the assessment results, which can be used to improve capabilities and construct a roadmap as outlined in the present research.

5.5 Main Research Question

The main research question is:

What constitutes a maturity model for Augmented Data Management that allows organizations to assess and improve their Data Management operations by leveraging AI?

Based on the results obtained in our research, it can be concluded that the ADM Maturity Model, as presented in this thesis, fulfills all requirements to be used by organizations to assess and improve their data management operations by leveraging AI. The ADM Maturity Model fulfills the functional requirements, as demonstrated in the case study. Using the maturity model and corresponding assessment, the current and desired maturity of three case organizations was assessed. The assessment results were used to provide recommendations on improving capabilities and constructing a roadmap to implement improvement initiatives.

The ADM Maturity Model was evaluated by the intended user stakeholder, data management consultants during expert interviews, and by the functional beneficiary stakeholder, organizations that want to improve their data management capabilities during the case studies. During the evaluation, these stakeholders rated the criteria from 1 (strongly disagree) to 5 (strongly agree). The average and median score for every criterion is between 4 and 5 for both stakeholder groups. Based on these scores, it can be concluded that the ADM^3 consists of sufficient and accurate maturity levels, (2) the processes and capabilities are relevant, comprehensive, mutually exclusive and accurate and (3) the model itself is understandable, easy to use, useful and practical

The same evaluation criteria are used to evaluate the recommendations on improving capabilities and constructing an improvement roadmap. Based on these scores, it can be concluded that these recommendations on improving capabilities are understandable, easy to use, and useful. It can also be concluded that the recommendations on constructing a roadmap are understandable and easy to use. The evaluation criteria for the practicality of the recommendations are assessed as neutral.

5.6 Contribution to Practice

The contribution of the present research to practice is twofold. First, this thesis introduced a maturity model and assessment tool that can be used to assess current data management capabilities and improvement opportunities. Second, this thesis provided an evaluation with practitioners, data management consultants and organizations, which indicated that the proposed maturity model and assessment tool are promising.

Practitioners can directly use the ADM Maturity Model and the corresponding assessment tool. The model presents a set of capabilities, sub capabilities, and processes that can be augmented and a set of maturity levels to assess those processes. The assessment tool can be used to guide and perform the assessment. Following the assessment, this thesis outlines general recommendations on improving capabilities and constructing a roadmap. The ADM Maturity Model is compatible with other data management maturity assessments. Practitioners commonly perform maturity assessments, and the ADM Maturity Model can be used to complement these assessments. The comparison matrices can be used to identify how this model relates to other data management maturity models.

The evaluation of the maturity model and assessment tool is promising. Data management consultants, the intended users of the model, assigned high scores to all the evaluation criteria. Participating organizations, the intended beneficiaries of the model, also assigned high scores for the same evaluation criteria during the case studies. These high scores are a strong indication that the ADM Maturity Model is easy to use, useful, and practical in a real-world application.

5.7 Contribution to Research

The contribution of the present research to the scientific body of knowledge is twofold. First, the thesis presented and demonstrated a maturity model development approach that combines existing frameworks and methodologies. Secondly, the thesis introduced the first maturity model for augmented data management.

During the author's background research on maturity model development, a systematic literature review identified the research methodologies, methods, and guidelines used by the academic community to develop IT maturity models [73]. Almost half of the 109 articles used an 'ad hoc' methodology, and a third did not specify a methodology at all. The same literature review revealed roughly the same division for 'ad hoc' and non-specified development guidelines. These numbers are surprising as building a cumulative tradition is commonly seen as a requirement for a coherent research field, as argued by Keen [97]. If maturity model research wants to become a mature research (sub) field, building on top of current knowledge is a high priority requirement.

The ADM Maturity Model was developed by leveraging existing and peer-reviewed methodologies, guidelines, and methods. These include the Design Science Research Methodology[20], the guidelines by Becker et al. [15], the systematic literature review method by Kitchenham [37], the metamodeling analysis and comparison method by Lautenschutz et al. [84], the qualitative content analysis guide by Dey [87], the evaluation template for maturity models by Salah et al. [24] and the

Framework for Evaluation in Design Science [96]. In addition to using existing methodologies, the present research also builds on top of existing maturity models. Instead of creating a new model, this thesis compares existing models and developed a complementary model. This thesis contributes to the coherence of the research field by demonstrating an approach that builds on top of existing knowledge.

As part of reflecting on the contribution of this thesis to research, we attempted to position the present work in terms of theory-development efforts. Following the classification by Gregor [98], the present research is of type II: Explanation. The contributions of this work are in presenting what augmented data management is, how AI can be leveraged in data management capabilities, why augmentation is a solution to limitations of current practices, which processes can be augmented, and how capabilities can be improved. In doing so, the theory is more elaborate than type I (according to Gregor's classification [98]). While it is predicted that augmented data management will be adopted rapidly in the future, this thesis's research method focuses on explaining current applications. Therefore, this study is not of type III: Prediction. The research framework, as presented in this thesis, can be applied to design new maturity models. However, it is not the main contribution as in type V research.

The second contribution is the development of a novel maturity model. Artificial intelligence and data management are both increasingly popular research fields, which resulted in the development of various maturity models in either field. This thesis presents the first maturity model that combines both fields within augmented data management. ADM has received tremendous interest from the industry in recent years and -through this research- is now also thoroughly scientifically researched for the first time.

6. Discussion

This chapter discusses the findings and limitations of the project. Section 6.1 reflects on the research methodology. Section 6.2 reflects on the model itself and its development. Section 6.3 presents the implications for practice. Section 6.4 presents the impactions for research, and Section 6.5 closes the chapter by discussing research limitations and future work.

6.1 Reflection on the Chosen Research Methodology

The present research set out to leverage existing methodologies, methods, and guidelines to develop a maturity model. This approach provides a solid foundation and validated approach for every step of the development and evaluation process. Because of this, the focus is more on the correct and valid execution of the methodologies and methods rather than substantiating and defending a self-introduced 'ad hoc methodology'.

The ADM Maturity Model was developed following the design science research methodology and the guidelines by Becker et al.[15]. The methodology aimed to evaluate the maturity model in a model of the context with regard to the stakeholder goals. This methodology proved to be suitable, as it results in strong indications about the effects in the real-world context.

The present research used a top-down approach to construct the maturity model. Within a top-down approach, the maturity stage definition is proposed first, followed by determining the measures that fit the definition. A bottom-up approach works vice versa and starts with collecting measures from practice. As augmented data management is a novel field, mature organizations are scarce and hence the top-down approach. This approach is a limitation of the present research, as a bottom-up approach ensures that the defined measures appear in practice. As the field of ADM matures, it would be more feasible to apply a bottom-up approach, such as a Delphi study. It would be interesting to reproduce this study once more mature organizations in terms of ADM are identified.

The present research leveraged a mixed-method development strategy to complement the shortcomings of the individual methods. The systematic literature review ensures that the research utilizes and builds on top of current knowledge. The metamodel analysis provides a thorough analysis of current AI and data management maturity models and simultaneously positions the ADM Maturity Model to complement them. However, solely reviewing current literature lacks practical relevance. Novel knowledge is gathered from practice by performing expert interviews to identify existing applications of augmented data management. We believe that this mixed-method development approach provides a solid foundation by combining current literature and empirical research.

The resulting maturity model is evaluated using a mixed-method validation strategy. Expert evaluations provide a relatively simple method to simulate conditions of practice and gather feedback to improve the model. However, this method only included one stakeholder group and is artificial. Case studies provide a more naturalistic method and simulate the conditions of practice for functional beneficiaries by applying the model during a maturity assessment at the case organizations. We believe that the expert evaluation and case study method complement each other's shortcomings and provide a valid evaluation strategy.

The case studies demonstrated the applicability of the model yet revealed that ADM often is very immature. The lack of maturity revealed the main drawback of choosing the case study research

method in a relatively new field: the case organizations were approached to participate and are not (yet) committed to augmented data management. In the optimal real-world context, an organization would apply the model when they experience the limitations of current data management practices. In that case, the model is expected to be the most useful. Despite the limitations on the motivation of participating organizations, the author thinks that case studies remain a valid choice as they simulate a real-world context. Even if an organization has not yet implemented AI technology, assessing their processes is both valuable to the organization and for evaluating the model.

The used methodology seemed optimal at the start of the research, considering the available time and resources. Technical action research could have been a better alternative, as the maturity model would be applied in a real-world context, and its effects in practice can be studied [20]. However, technical action research requires a dedicated sponsor, a suitable participating organization, and presumably considerably more time. Due to the novelty of the maturity model, it would be hard to find a sponsor and participating organization, as they have to invest resources. Therefore, design science research was the best viable methodology. Applying the ADM Maturity Model in technical action research would be the next step in validating the model and its effects. Such a research can be designed with a sponsor such as Deloitte, with multiple experienced data management consultants on board that would apply the model during a maturity assessment as part of a data management improvement project at a client organization. The researcher can then observe the effects and validate the model in a real-world context.

6.2 ADM Maturity Model Reflection

The present research resulted in the ADM Maturity Model, which was evaluated by eleven experts and applied in three case studies with five participants. During both rounds, the maturity model received a high rating for all evaluation criteria. Each criterion was rated on average between 4.2 and 5.0 on a 5-point scale. These scores are an indication that the maturity model fulfills the design goals of the present research. However, when critically reflecting on the model, a few points for discussion arise.

One recurring discussion point was the question about which processes should be leading when assessing maturity. The model mentions ‘processes’ and not individual processes, which opens up the question whether an average score should be used, or the score for the top-level process or the one process that is lagging. The argument for using the top-notch process is that this reflects the ability of the organization; if they manage to execute one process, they can transfer this knowledge to the other processes as well. The argument to go for the least mature process is that this process can limit the whole capability as the weakest link. Scoring an average also is sub-optimal as it captures neither. Based on these arguments, we think that it would be best to coordinate this choice with the participating organization.

Another major point of discussion is the concept of automation versus artificial intelligence, which caused some confusion. These two terms are inherently different as automation refers to streamlining repetitive tasks, while artificial intelligence is about emulating human intelligence and decisions. The distinction between the two can be hard to identify, as extensive rule-based automation can come across as AI. Similarly, AI can be used to automate simple tasks. The user often uses the system without being aware of the technology behind it. An example of such a system is Google Translate. Before late 2017, Google Translate used rule-based automation to translate text. At the end of 2017, Google replaced this system with AI, which reportedly reduced the code from 500,000 lines to 500 [99]. At the start, users might not have noted a difference in translation quality,

while the efficiency changed drastically. This example illustrates the difficulty in assessing whether a process truly leverages AI. Therefore, in this work, we deliberately chose to include both 'automation' and 'decision making' in the definition of augmented data management and the maturity levels to limit the discussion. This choice is motivated by the fact that the model processes describe 'what is done' and not 'how it is done'. Similarly, augmented data management focuses on 'what is being done for you' and not 'how it is being done'. The 'what' is in automating tasks and decisions, regardless of the 'how' being automation or AI. The stakeholders benefit either way.

A discussion point of the model is the subjectivity in assessing processes. We must note that this is a common characteristic of maturity models and only becomes a limitation if participants cannot align their views. One benefit of a maturity assessment is connecting participants with different viewpoints and discussing current capabilities and planning for future capabilities. The difference and subjectivity are demonstrated in the first case study, where three participants assessed the same processes at the same organization. The participants rarely assigned the same maturity level. The deviation is often limited to less than one maturity level. The model could be made more objective, which presumably would make it easier to use, which would increase the focus on improvements.

A set of limitations is associated with maturity models in general [15], [39]: they are over-simplistic compared to reality, they focus on a single path to maturity while neglecting alternative approaches, the applicability may be constrained by internal (technology, supplier relations) and external factors (market conditions). During this research, we were able to overcome some common limitations of maturity models. A lack of fundamentals was avoided by presenting a literature review on AI. Unnecessary duplicate models were avoided by comparing existing maturity models, and a lack of transparent development method was avoided by introducing the multi-method development and evaluation strategy.

Another limitation of the model is traceable to the participant profile. The case studies showed that none of the participants were able to score all the processes. Just as data management models consider different capabilities to be relevant, data management functions also differ at organizations. To assess all ADM Maturity Model processes and capabilities, the participants need to come from multiple disciplines, professional backgrounds, and different organizational roles such as data engineers, data scientists, data quality experts, data stewards, and data managers. It might be difficult in an organization to form such a multidisciplinary team with that many and diverse experts due to time pressure or resource limitations.

A final limitation is traceable to the chosen data management capabilities. As already indicated, the choice is made to scope the research and select *five* capabilities, where AI has the largest potential. This selection results in the model not covering the whole data management field. Capabilities with a low potential for AI, such as data governance, are also vital to data management. In turn, some capabilities are not covered in the model, while they can still have the potential to be augmented. The processes related to these capabilities frequently change due to new technologies and innovations. Disrupting innovations might even influence whole capabilities. For example, the semantic web or linked open data can fundamentally impact metadata management and data integration, making some processes obsolete [100]. To keep the proposed maturity model viable and valuable, we think that every few years, it would be good to be revised so that organizations assure the most recent developments in augmented data management are covered accordingly.

6.3 Implications for Practice

As a result of the present research, practitioners can now perform maturity assessments that incorporate augmented data management processes and maturity levels. Business and IT are often considered separate fields. Practitioners with technical knowledge apply AI to data management challenges while the business is unaware of these applications. Similarly, practitioners within the data management field are often not aware of the potential of artificial intelligence to add value from a business perspective. The ADM Maturity Model bridges business and IT capabilities into a single maturity model.

In terms of maturity, it is expected that a certain level of data management maturity is required before applying augmented data management effectively. It can be reasoned that it would be hard to augment processes that are not defined within an organization, limiting the applicability of the augmented part of the model. Meanwhile, it provides insight into which capabilities are not (yet) ready to be augmented, which is valuable and can be used to prioritize initiatives. Therefore, we think that the model can be used within organizations that are ready to implement or improve augmented data management, as well as less mature organizations that are in the preparatory phase.

What would take organizations beyond the one in which this project took place to use the proposed maturity model? It is our understanding that any data management practitioner can likely apply the model. During the evaluation, all data management consultants agreed that the maturity model was understandable and easy to use, regardless of their experience and knowledge of AI. Furthermore, the market research revealed that there are many tools available that already leverage AI. Therefore, only knowledge and experience in system implementation is necessary to introduce augmented data management. Besides, we noticed that organizations that already have tools with AI functionalities are sometimes unaware of these functionalities or their potential. Likewise, the maturity model can be applied to get the most out of current tooling. Altogether, we think that the ADM maturity model has versatile use cases for variousdegrees of organizational maturity and practitioner experience.

6.4 Implications for Research

The findings within the present research might impact future research and lead to new potential research questions. The first set of research questions focusses on future research on the current ADM Maturity Model:

- What is the relationship between data management maturity and augmented maturity?
- Does a high maturity level for augmented data management lead to a higher process maturity level?
- Does a high maturity level for data management maturity lead to a higher augmented maturity level?
- How does the ADM Maturity Model hold up in practice (action research)?
- Which level of data management maturity is required before implementing augmented data management?
- Can the ADM Maturity Model be applied as a self-assessment tool?
- Which objective measures can be identified to assess the processes of the ADM Maturity Model?

Answering these research questions will provide empirical evidence that could help practitioners reason about the suitability of the maturity model to their own organizational context. The presence of more empirical data is also vital to the understanding of the implications of a maturity level for an organization

Second, the following set of research questions focusses on augmented data management as a whole:

- Which other capabilities can be augmented?
- What is the effect of augmented data management in terms of efficiency?
- What are the limits of the current practices? At what point is augmentation necessary or viable?
- Can the maturity scale be applied to other domains, i.e., augmented analytics?
- What is the effect of augmented data management on analytics?

Answering these research questions will provide empirical evidence on the full potential, benefits, and impact of augmented data management. The prospect based on current literature is promising; more empirical data can reveal whether this can be realized in practice.

6.5 Research Limitations and Future Work

The ADM maturity model is developed through a literature review, expert interviews, and case studies. Limitations inherent to these methods are the risk of overlooking specific sources, biased or influenced experts and case study participants, and misinterpreting qualitative data [20].

The systematic literature review on maturity models for data management and artificial intelligence aimed to be as inclusive and thorough as possible. Various data sources were consulted to the point that additional sources seemed not to yield additional results. Despite this, it is possible that certain sources have been overlooked or additional research has been published that would have influenced the present research.

The experts that participated in the evaluation and case study participants might be biased or (unconsciously) influenced during the research. Some of the experts are part of the Enterprise Data Management team at Deloitte, which facilitates the present research. While on average, these experts tend to score the evaluation criteria statements lower than the other participants, they do have an interest in the research, which might influence them. This effect might lead to a more critical and lower evaluation, as these participants can benefit from the resulting maturity model. Another influencing factor might be that experts and participants prefer to be kind and therefore moderate their criticism.

The researcher might be biased towards positive results, which could be reflected in evaluation questions and criteria that are formulated or framed to persuade positive feedback. Throughout the research process, the researcher was conscious of this and aimed to minimize this effect by using a standard template for evaluation criteria and questions. The researcher could also be biased in interpreting the results. A second independent researcher was asked to verify the qualitative content analysis from the interview transcripts by peer debriefing to avoid this bias.

To claim the generalizability of the maturity model and research, it must be argued that the findings of the sample population can be extended to the population at large [20]. The sample population for

the expert interviews consists of data management consultants with experience in maturity assessment at organizations in the Netherlands. The sample population of the case study consists of organizations in the Netherlands. During both evaluations, participants indicated that the ADM Maturity Model is easy to use, useful and practical, supporting the claim that these effects are generalizable for organizations within the Netherlands. It can also be argued that these effects are generalizable for organizations worldwide. Some of the case study organizations, and the organizations that the consultants have experience with, operate internationally, which indicates that other branches might have similar processes. One primary argument supporting this claim is that the ADM Maturity Model processes are also present in widely used data management models such as DAMA DMBOK, which has many certified practitioners worldwide. Not all organizations execute all processes or have all capabilities, yet they can be expected to cover some. Therefore, it can be argued that (a selection of) processes of the ADM Maturity Model can be generalized to most organizations that do some form of data management.

The ADM Maturity Model does introduce a new augmentation maturity scale synthesized from lesser-known AI maturity models from and recommendations on improving capabilities and constructing a roadmap. These augmented constructs of the model are more useful for applying to organizations that have somewhat mature data management processes and have the technical capabilities to apply AI. The expert interviews revealed that there are successful implementations of augmented data management worldwide. The case studies showed that the model can also be applied to international organizations regardless of their augmentation maturity. However, less mature organizations seemed to focus more on improving data management in general than augmented data management. While it can be useful for organizations to know they score low in data management, having some maturity seems to improve the usefulness of applying the augmentation scale.

Future work should first focus on improving the current limitations of the model. First of all, the model could be made more objective by quantifying measures. For example, by asking participants to indicate a percentage of processes that adheres to a certain standard or by assigning a particular maturity score to objective constructs, such as architecture and data types. Second, the protocol around the selection of assessment participants should be improved to make it more multidisciplinary, so it covers all capabilities. Future work could also focus on adding additional capabilities, such as data governance, to make the model more comprehensive. To include more capabilities the same development procedure, as used in the present research, can be used. The maturity model should also be applied to more organizations, preferably during action research at an organization that wants to implement or improve its augmented data management. Within such research, the maturity model can be studied and validated in practice. At last, future work should revise the model and keep it up to date, as artificial intelligence and data management are fast-changing fields.

7. Bibliography

[1] M. Fleckenstein and L. Fellows, *Modern data strategy*. 2018.

[2] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014, doi: 10.1016/j.ins.2014.01.015.

[3] M. M. C. Francisco, S. N. Alves-Souza, E. G. L. Campos, and L. S. De Souza, "Total data quality management and total information quality management applied to costumer relationship management," *ACM Int. Conf. Proceeding Ser.*, no. 1, pp. 40–45, 2017, doi: 10.1145/3149572.3149575.

[4] DAMA, *DAMA-DMBok: Data management body of knowledge (2nd edition)*. 2017.

[5] G. H. Kim, S. Trimi, and J. H. Chung, "Big-data applications in the government sector," *Commun. ACM*, vol. 57, no. 3, pp. 78–85, 2014, doi: 10.1145/2500873.

[6] D. Laney, "Gartner ' s Enterprise Information Management Maturity Model," no. October, 2018.

[7] Janine S. Hiller and F. Bélanger, "Privacy Strategies for Electronic Government," *E-Government*, pp. 162–198, 2001, doi: 10.1.1.470.2867.

[8] R. Sallam *et al.*, "Top 10 Data and Analytics Technology Trends That Will Change Your Business," 2019.

[9] Cisco, "Cisco Annual Internet Report (2018–2023)," *Cisco*, pp. 1–41, 2020.

[10] S. Sicular and D. Aron, "Leverage Augmented Intelligence to Win With," 2019.

[11] S. J. Russell and P. Norvig, *Artificial Intelligence A Modern Approach*. 2003.

[12] C. Bishop, "Pattern Recognition and Machine Learning," *J. Electron. Imaging*, vol. 16, no. 4, Jan. 2006, doi: 10.1117/1.2819119.

[13] J. Hare, C. Idoine, and P. Krensky, "How Augmented Machine Learning Is Democratizing Data Science," 2019.

[14] NewVantage Partners, "Big Data and AI Executive Survey 2020," pp. 1–16, 2019.

[15] J. Becker, R. Knackstedt, and J. Pöppelbuß, "Developing Maturity Models for IT Management," *Bus. Inf. Syst. Eng.*, vol. 1, no. 3, pp. 213–222, 2009, doi: 10.1007/s12599-009-0044-5.

[16] T. de Bruin, M. Rosemann, R. Freeze, and U. Kulkarni, "Understanding the main phases of developing a maturity assessment model," *ACIS 2005 Proc. - 16th Australas. Conf. Inf. Syst.*, no. January, 2005.

[17] M. Röglinger, J. Pöppelbuß, and J. Becker, "Maturity models in business process management," *Bus. Process Manag. J.*, vol. 18, no. 2, pp. 328–346, 2012, doi: 10.1108/14637151211225225.

[18] D. Proença and J. Borbinha, "Maturity Models for Data and Information Management," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11057 LNCS, no. Cmmi, 2018, pp. 81–93.

[19] M. van Steenbergen, R. Bos, S. Brinkkemper, I. van de Weerd, and W. Bekkers, "Improving IS Functions Step by Step: the Use of Focus Area Maturity Models," *Scand. J. Inf. Syst.*, vol. 25, no. 2, p. 2, 2013.

[20] R. J. Wieringa, *Design science methodology: For information systems and software engineering*. 2014.

[21] I. F. Alexander, "A Taxonomy of Stakeholders: Human Roles in System Development," *Int. J. Technol. Hum. Interact.*, vol. 1, no. 1, pp. 23–59, 2005, doi: 10.4018/jthi.2005010102.

[22] T. Mettler, P. Rohner, and R. Winter, "Towards a Classification of Maturity Models in Information

Systems," in *Management of the Interconnected World*, Heidelberg: Physica-Verlag HD, 2010, pp. 333–340.

[23] T. Mettler, "A Design Science Research Perspective on Maturity Models in Information Systems," vol. 41, no. 0, 2009.

[24] D. Salah, R. Paige, and P. Cairns, "An evaluation template for expert review of maturity models," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8892, pp. 318–321, 2014, doi: 10.1007/978-3-319-13835-0_31.

[25] T. Mettler, "Maturity assessment models: a design science research approach," *Int. J. Soc. Syst. Sci.*, vol. 3, no. 1/2, p. 81, 2011, doi: 10.1504/ijsss.2011.038934.

[26] M. Röglinger and J. Pöppelbuß, "What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management," *19th Eur. Conf. Inf. Syst. ECIS 2011*, vol. 4801, 2011.

[27] M. Mills, "Artificial Intelligence in Law : the State of Play 2016," *Thomst. Reuters, Leg. Exec. Inst.*, p. 6, 2016.

[28] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, no. 3–4, pp. 197–387, 2013, doi: 10.1561/2000000039.

[29] M. Ivanović and M. Radovanović, "Modern machine learning techniques and their applications," *Electron. Commun. Networks IV - Proc. 4th Int. Conf. Electron. Commun. Networks, CECNet2014*, vol. 1, pp. 833–846, 2015, doi: 10.1201/b18592-153.

[30] J. T. Kwok, Z.-H. Zhou, and L. Xu, "Machine Learning," in *Springer Handbook of Computational Intelligence*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 495–522.

[31] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of The Art, Current Trends and Challenges," no. Figure 1, 2017.

[32] D. Waterman, *A guide to expert systems*. Addison-Wesley Longman Publishing Co., Inc.75 Arlington Street, Suite 300 Boston, MAUnited States, 1985.

[33] J. Liebowitz, "Expert systems: A short introduction," *Eng. Fract. Mech.*, vol. 50, no. 5–6, pp. 601–607, Mar. 1995, doi: 10.1016/0013-7944(94)E0047-K.

[34] J. VijiPriya, J. Ashok, and S. Suppiah, "A Review on Significance of Sub Fields in Artificial Intelligence," *Int. J. Latest Trends Eng. Technol.*, vol. 6, no. 3, pp. 542–548, 2016.

[35] K. R. Chowdhary, *Fundamentals of Artificial Intelligence*. Springer India, 2020.

[36] M. McMahon, D. Mumper, M. Ihaza, and D. Farrar, "How smart is your manufacturing? Build smarter with AI," *Proc. - Int. Comput. Softw. Appl. Conf.*, vol. 2, pp. 55–60, 2019, doi: 10.1109/COMPSAC.2019.10183.

[37] S. C. B. Kitchenham, "Guidelines for performing Systematic Literature Reviews in Software Engineering," *EBSE Tech. Rep.*, vol. 1, 2007.

[38] D. Arnott and G. Pervan, "Eight key issues for the decision support systems discipline," *Decis. Support Syst.*, vol. 44, no. 3, pp. 657–672, 2008, doi: 10.1016/j.dss.2007.09.003.

[39] D. Proença and J. Borbinha, "Maturity Models for Information Systems - A State of the Art," *Procedia Comput. Sci.*, vol. 100, no. 2, pp. 1042–1049, 2016, doi: 10.1016/j.procs.2016.09.279.

[40] M. Khoshgoftar and O. Osman, "Comparison of maturity models," *Proc. - 2009 2nd IEEE Int. Conf. Comput. Sci. Inf. Technol. ICCSIT 2009*, pp. 297–301, 2009, doi: 10.1109/ICCSIT.2009.5234402.

[41] D. Krismawati, Y. Ruldeviyani, and R. Rusli, "Master data management maturity model: A case study at statistics business register in statistics Indonesia," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 931–936, 2019, doi: 10.1109/ICOIACT46704.2019.8938482.

[42] F. G. Pratama, S. Astana, S. B. Yudhoatmojo, and A. N. Hidayanto, “Master Data Management Maturity Assessment: A Case Study of Organization in Ministry of Education and Culture,” *2018 Int. Conf. Comput. Control. Informatics its Appl. Recent Challenges Mach. Learn. Comput. Appl. IC3INA 2018 - Proceeding*, no. Mdm, pp. 1–6, 2019, doi: 10.1109/IC3INA.2018.8629524.

[43] M. Spruit and K. Pietzka, “MD3M: The master data management maturity model,” *Comput. Human Behav.*, vol. 51, pp. 1068–1076, 2015, doi: 10.1016/j.chb.2014.09.030.

[44] A. R. A, “Master Data Management Maturity Assessment : A Case Study of a Pasar Rebo Public Hospital,” *Int. J. Emerg. Trends Eng. Res.*, vol. 7, no. 5, pp. 15–20, Jun. 2019, doi: 10.30534/ijeter/2019/02752019.

[45] R. Iqbal, P. Yuda, W. Aditya, A. N. Hidayanto, P. Wuri Handayani, and N. C. Harahap, “Master Data Management Maturity Assessment: Case Study of XYZ Company,” in *2019 2nd International Conference on Applied Information Technology and Innovation (ICAITI)*, 2019, pp. 133–139, doi: 10.1109/ICAITI48442.2019.8982123.

[46] F. Sanchez-Puchol and J. A. Pastor-Collado, “Focus area maturity models: A comparative review,” *Lect. Notes Bus. Inf. Process.*, vol. 299, pp. 531–544, 2017, doi: 10.1007/978-3-319-65930-5_42.

[47] N. Qodarsih, S. B. Yudhoatmojo, and A. N. Hidayanto, “Master Data Management Maturity Assessment: A Case Study in the Supreme Court of the Republic of Indonesia,” *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–7, 2019, doi: 10.1109/CITSM.2018.8674373.

[48] Y. Baolong, W. Hong, and Z. Haodong, “Research and application of data management based on Data Management Maturity Model (DMM),” *ACM Int. Conf. Proceeding Ser.*, no. Dmm, pp. 157–160, 2018, doi: 10.1145/3195106.3195177.

[49] Enterprise Data Management Council, “Data Management Capability Assessment Model (DCAM),” 2014.

[50] M. A. Thomas, J. Cipolla, B. Lambert, and L. Carter, “Data management maturity assessment of public sector agencies,” *Gov. Inf. Q.*, vol. 36, no. 4, p. 101401, 2019, doi: 10.1016/j.giq.2019.101401.

[51] CMMI, “Data Management (DMM) Model Data Management,” 2019.

[52] T. Hulme, “Information governance: Sharing the IBM approach,” *Bus. Inf. Rev.*, vol. 29, no. 2, pp. 99–104, 2012, doi: 10.1177/0266382112449221.

[53] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, “A systematic literature review of data governance and cloud data governance,” *Pers. Ubiquitous Comput.*, vol. 23, no. 5–6, pp. 839–859, 2019, doi: 10.1007/s00779-017-1104-3.

[54] D. Heredia-Vizcaíno and W. Nieto, “A Governing Framework for Data-Driven Small Organizations in Colombia,” in *Advances in Intelligent Systems and Computing (special issue from WorldCIST 2019 – 7th World Conference on Information Systems and Technologies)*, vol. 1, no. 1, 2019, pp. 622–629.

[55] M. Chessel, “IBM Information Governance Model,” *Int. J. Inf. Manage.*, 2010.

[56] D. H. Kurniawan, Y. Ruldeviyani, M. R. Adrian, S. Handayani, M. R. Pohan, and T. Rani Khairunnisa, “Data Governance Maturity Assessment: A Case Study in IT Bureau of Audit Board,” *Proc. 2019 Int. Conf. Inf. Manag. Technol. ICIMTech 2019*, vol. 1, no. August, pp. 629–634, 2019, doi: 10.1109/ICIMTech.2019.8843742.

[57] Dimas Agung Saputra ; Dika Handika ; Yova Ruldeviyani, “Data Governance Maturity Model (DGM2) Assessment in Organization Transformation of Digital Telecommunication Company: Case Study of PT Telekomunikasi Indonesia,” *ICACSIS*, no. 1, p. 43, 2018, doi: 10.1017/CBO9781107415324.004.

[58] A. Grillenberger and R. Romeike, “Key concepts of data management ⇓ an empirical approach,” *ACM Int. Conf. Proceeding Ser.*, pp. 30–39, 2017, doi: 10.1145/3141880.3141886.

[59] G. B. de Figueiredo, J. L. R. Moreira, K. de Faria Cordeiro, and M. L. M. Campos, “Aligning DMBOK and

Open Government with the FAIR Data Principles," vol. 1, no. October, 2019, pp. 13–22.

[60] A. A. Arman, G. Ramadhan, and M. Fajrin, "Design of data management guideline for open data implementation: (Case study in indonesia)," *ACM Int. Conf. Proceeding Ser.*, vol. 2015-Novem, pp. 17–23, 2015, doi: 10.1145/2846012.2846024.

[61] O. Barrenechea, A. Mendieta, J. Armas, and J. M. Madrid, "Data governance reference model to streamline the supply chain process in SMEs," *Proc. 2019 IEEE 26th Int. Conf. Electron. Electr. Eng. Comput. INTERCON 2019*, pp. 1–4, 2019, doi: 10.1109/INTERCON.2019.8853634.

[62] D. Stock, F. Wortmann, and J. H. Mayer, "Use cases for business metadata - A viewpoint-based approach to structuring and prioritizing business needs," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6105 LNCS, pp. 546–549, 2010, doi: 10.1007/978-3-642-13335-0_42.

[63] EDM Council, "Exploring DCAM." [Online]. Available: https://edmcouncil.org/page/exploringdcam. [Accessed: 17-Mar-2020].

[64] S. Soares, *The IBM Data Governance Unified Process*. 2010.

[65] Irina Steenbeek, *The Data Management Toolkit*. Independently published, 2018.

[66] S. Alsheibani, C. Messom, and Y. Cheung, "Towards an Artificial Intelligence Maturity Model : From Science Fiction to Business Facts," *Twenty-Third Pacific Asia Conf. Inf. Syst.*, no. 2018, 2019.

[67] P. Gentsch, *AI in Marketing, Sales and Service*. 2019.

[68] J. Siekmann and W. Wahlster, "Artificial Intelligence Maturity Model," no. March. p. 20, 2020.

[69] T. Pringle and E. Zoller, "How to Achieve AI Maturity and Why It Matters An AI maturity assessment model and road map for CSPs," *Ovum*, 2018.

[70] J. V. Carvalho, Á. Rocha, and A. Abreu, "HISMM - Hospital Information System Maturity Model: A Synthesis," no. 222, 2017, pp. 189–200.

[71] M. Van Steenbergen, M. Van Den Berg, and S. Brinkkemper, "A balanced approach to developing the enterprise architecture practice," *Lect. Notes Bus. Inf. Process.*, vol. 12 LNBIP, pp. 240–253, 2008, doi: 10.1007/978-3-540-88710-2_19.

[72] T. Mettler and P. Rohner, "Situational maturity models as instrumental artifacts for organizational design," *Proc. 4th Int. Conf. Des. Sci. Res. Inf. Syst. Technol. DESRIST '09*, no. April 2014, 2009, doi: 10.1145/1555619.1555649.

[73] R. Pereira and J. Serrano, "A review of methods used on IT maturity models development: A systematic literature review and a critical analysis," *J. Inf. Technol.*, 2020, doi: 10.1177/0268396219886874.

[74] M. Hult and S. Lennung, "Towards a Definition of Action Research: a Note and Bibliography," *J. Manag. Stud.*, vol. 17, no. 2, pp. 241–250, 1980, doi: 10.1111/j.1467-6486.1980.tb00087.x.

[75] J. Iivari and J. Venable, "Action research and design science research - Seemingly similar but decisively dissimilar," *17th Eur. Conf. Inf. Syst. ECIS 2009*, no. June 2014, 2009.

[76] T. O.Nyumba, K. Wilson, C. J. Derrick, and N. Mukherjee, "The use of focus group discussion methodology: Insights from two decades of application in conservation," *Methods Ecol. Evol.*, vol. 9, no. 1, pp. 20–32, 2018, doi: 10.1111/2041-210X.12860.

[77] G. J. Skulmoski, F. T. Hartman, and J. Krahn, "The Delphi Method for Graduate Research," *J. Inf. Technol. Educ. Res.*, vol. 6, pp. 001–021, 2007, doi: 10.28945/199.

[78] L. Bai, H. Wang, N. Huang, Q. Du, and Y. Huang, "An environmental management maturity model of construction programs using the AHP-entropy approach," *Int. J. Environ. Res. Public Health*, vol. 15, no. 7, 2018, doi: 10.3390/ijerph15071317.

[79] M. Cahill, K. Robinson, J. Pettigrew, R. Galvin, and M. Stanley, “Qualitative synthesis: A guide to conducting a meta-ethnography,” *Br. J. Occup. Ther.*, vol. 81, no. 3, pp. 129–137, 2018, doi: 10.1177/0308022617745016.

[80] G. A. García-Mireles, M. Á. Moraga, and F. García, “Development of maturity models: A systematic literature review,” *IET Semin. Dig.*, vol. 2012, no. 1, pp. 279–283, 2012, doi: 10.1049/ic.2012.0036.

[81] A. M. Maier, J. Moultrie, and P. J. Clarkson, “Assessing organizational capabilities: Reviewing and guiding the development of maturity grids,” *IEEE Trans. Eng. Manag.*, vol. 59, no. 1, pp. 138–159, 2012, doi: 10.1109/TEM.2010.2077289.

[82] M. Van Steenbergen, R. Bos, S. Brinkkemper, I. Van De Weerd, and W. Bekkers, “The design of focus area maturity models,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6105 LNCS, pp. 317–332, 2010, doi: 10.1007/978-3-642-13335-0_22.

[83] Hevner, March, Park, and Ram, “Design Science in Information Systems Research,” *MIS Q.*, vol. 28, no. 1, p. 75, 2004, doi: 10.2307/25148625.

[84] D. Lautenschutz, S. España, A. Hankel, S. Overbeek, and P. Lago, “A Comparative Analysis of Green ICT Maturity Models,” vol. 52, no. i, pp. 153–137, 2018, doi: 10.29007/5hgz.

[85] S. Elo and H. Kyngäs, “The qualitative content analysis process,” *J. Adv. Nurs.*, vol. 62, no. 1, pp. 107–115, 2008, doi: 10.1111/j.1365-2648.2007.04569.x.

[86] D. Feinberg and R. Adam, “Hype Cycle for Data Management 2019,” *Gart. Res. Rep.*, no. July, 2019.

[87] I. Dey, *Qualitative data analysis: A user-friendly guide for social scientists*. 2003.

[88] Informatica, “Artificial Intelligence for the Data-Driven Intelligent Enterprise,” 2020.

[89] M. Beyer, “Future of Data Management, 2019 Edition,” in *Gartner Data & Analytics Summit*, 2019.

[90] IBM and 451 Research, “AI and Data Management,” 2019.

[91] Deloitte Development LLC, “Digital FTE - Profiles.” 2019.

[92] Deloitte Development LLC, “CogniSteward and Other Assets Overview.” 2018.

[93] B. Dageville *et al.*, “The snowflake elastic data warehouse,” *Proc. ACM SIGMOD Int. Conf. Manag. Data*, vol. 26-June-20, pp. 215–226, 2016, doi: 10.1145/2882903.2903741.

[94] Alteryx, “Alteryx Analytic Process Automation Platform,” 2020. [Online]. Available: https://www.alteryx.com/products/apa-platform. [Accessed: 28-Jul-2020].

[95] Oracle Corporation, “Oracle Autonomous Database,” no. September, p. 1, 2018.

[96] J. Venable, J. Pries-Heje, and R. Baskerville, “FEDS: A Framework for Evaluation in Design Science Research,” *Eur. J. Inf. Syst.*, vol. 25, no. 1, pp. 77–89, 2016, doi: 10.1057/ejis.2014.36.

[97] P. G. W. Keen, “Association for Information Systems AIS Electronic Library (AISeL) MIS RESEARCH: REFERENCE DISCIPLINES AND A CUMULATIVE TRADITION,” *Proc. First Int. Conf. Inf. Syst.*, pp. 9–18, 1980.

[98] S. Gregor, “The nature of theory in Information Systems,” *MIS Q. Manag. Inf. Syst.*, vol. 30, no. 3, pp. 611–642, 2006, doi: 10.2307/25148742.

[99] J. Clark, “The power of AI, seen via Google translate,” 2017. [Online]. Available: https://jack-clark.net/2017/10/09/import-ai-63-google-shrinks-language-translation-code-from-500000-to-500-lines-with-ai-only-25-of-surveyed-people-believe-automationbetter-jobs/.

[100] O. Berneers Lee, T., Hendler, J., & Lassila, “The Semantic Web: A new form of Web content that is meaningful to computers will unleash the revolution of new possibilities.,” *Sci. Am.*, vol. 5, no. 284, pp. 34–43, 2001, doi: 10.1038/scientificamerican0501-34.

8. Appendix

A. Description of Data Management Literature

Model	Type	Method	Validation
MD3M [43]	DM Model	Literature research	Academic validity in the development process by building on peers and a single case study
MD3M [42], [61], [41], [44], [45]	Case study	Literature search for model: apply MD3M	Measure master data maturity in external case study using MD3M
MD3M [46]	Description of type/class	Literature search for FAMM, present characteristics	Validity through systematic method and referencing published models
DCAM [49]	DM Model	Synthesis of best practices of EDM council members into DCAM	Not validated, purely based on experience of unidentified professionals
DCAM, CMMI, IBM [48]	Conceptual overview	Synthesize DM models	Assess DM maturity in external case study
CMMI DMM [51]	DM Model	Synthesis of best practices	Not validated, build on top of established CMM
CMMI DMM [50]	Case study	Apply DMM	Case study at 15 government agencies
DMBOK, CMMI DMM [1]	Opinion and personal experience	Combine literature and personal experience	Reference to established publications, written by industry experts (EDM Council)
Gartner, Stanford, CMMI DMM [18]	Conceptual overview	Literature research	Reference to established publications, presenting an overview
IBM [64]	DM Model	Synthesis of best practices	Written by industry experts (IBM)
IBM [52]	Opinion and personal experience	Justify proposition by personal experience	Personal experience (IBM) talking to CIO's, CTO's and Chief Architects
IBM, DMBOK [53]	Conceptual overview	Systematic literature review	Reference to established publications, presenting an overview
Stanford [57] [56]	Case study	Apply Stanford DGMM	Case study at external organization
Gartner [6]	DM Model	Synthesize best practices from experience	Based on dozens of workshops and thousands of analyst interactions (Gartner)
DMBOK [62]	Opinion and example	Using aggregation in order to develop a classification for metadata use cases	Case study
DMBOK [60]	Theory	Modify DMBOK to fit Open Data characteristics	Present mapping between DMBOK and Open Data principles, no validation
DMBOK [58]	Survey	Literature review to determine central principles of DM	Semi-structured face-to-face interview with an internationally renowned professor on data management
DMBOK [4]	DM Model	Synthesize best practices	Based on professional experience, references to many other published models
DMBOK [59]	Theory	Align DMBOK with Open Government and FAIR principles	No validation, builds on DMBOK
DMBOK [61]	Case study	Propose data governance reference model using DMBOK	Case study at external company

Table 38: Overview of Data Management Maturity Model Literature

B. Systematic Comparison

B1: Data Management Maturity Levels

Stanford DGMM	Description	Rationale	Gartner	Description	Rationale	CMMI DMM	Description	Rationale	IBM DGMM	Rationale	MD3M	Rationale	Synthesis DM lvls	Description	Short
-	-	-	**Aware**	Level 1 organizations are typically in the lower 10% of those we advise. They are generally aware of key issues and challenges, but lack the budget, resources and/or leadership to make any meaningful advances in EIM.	Aware of issues, but no formal data management processes in place.								**Incomplete**	No organized data management practices or formal enterprise processes for managing data. Organizations are aware of key issues and challenged but lack the budget, resources and/or leadership to make any meaningful advances.	
Initial	Process unpredictable, poorly controlled and reactive	1:1 mapping	**Reactive**	Level 2 organizations represent approximately 30% of those we come across. They generally operate in a reactive application centric mode, waiting until information-related problems manifest in significant business losses or lack of competitiveness before	Ad hoc = reactive	**Performed**	Processes are performed ad hoc, primarily at the project level. Processes are typically not applied across business areas.	1:1 mapping	**Initial**	1:1 mapping	**Initial**	1:1 mapping	**Initial**	General-purpose data management using a limited tool set, with little or no governance. Organizations operate in a reactive application-centric mode, waiting until information-related problems manifest in significant business losses or lack of competitiveness before addressing them. Data handling is highly reliant on a few experts. Roles and responsibilities are defined within silos. Each data owner receives, generates, and sends data autonomously. Controls, if they exist, are applied inconsistently. Solutions for managing data are limited. Data quality issues are pervasive but not addressed. Infrastructure supports	Processes are (manual) unpredictable, poorly controlled, reactive and typically not applied across business units.
Managed	Process characterized for projects and is managable	1:1 mapping	**Proactive**	Level 3 organizations represent the approximate 40% of those that are more or less mainstream today in terms of their information-related capabilities. They have become more proactive in addressing certain areas of information management, and have started to put the "enterprise" in enterprise information management. Some programs are operational and effective, but there is little leverage or	Process in place, but no organizational alignment	**Managed**	Processes are planned and executed in accordance with policy; employ skilled people with adequate resources to produce controlled outputs; involve relevant stakeholders; are monitored and controlled and evaluated for adherence to the defined process.	1:1 mapping	**Managed**	1:1 mapping	**Repeatable**	1:1 mapping	**Managed**	Emergence of consistent tools and role definition to support process execution. Processes are planned and executed in accordance with policy; employ skilled people with adequate resources to produce controlled outputs; involve relevant stakeholders; are monitored and controlled and evaluated for adherence to the defined process. The organization is getting aware of data quality issues and concepts. Concepts of Master and Reference Data begin to be recognized. Some programs are operational and effective, but there is little leverage or alignment across programs and investments. This might be a suitable entry point for ADM: emergence of support processes and increased awareness. Good timing to identify first ADM usecases and ensure readyness.	Processes are executed and planned in accordance with certain policy, but primarily project-based.
Defined	Process characterized for the organization and is pro-active	1:1 mapping				**Defined**	Set of standard processes is employed and consistently followed. Processes to meet specific needs are tailored from the set of standard processes according to the organization's guidelines.	1:1 mapping	**Defined**	1:1 mapping	**Defined process**	1:1 mapping	**Defined**	Introduction and institutionalization of scalable data management processes and a view of DM as an organizational enabler. Data management capabilities established and verified by stakeholders: roles and responsibilities structured, policy and standards implemented, glossaries and identifiers established, sustainable funding. Characteristics include the replication of data across an organization with some controls in place and a general increase in overall data quality, along with more formal process definition leads to a signifiicant reduction in manual intervention. This, along with a centralized design process, means that process outcomes are more predictable.	Set of scalable and proactive processes is employed and consistently followed across the organization.
Quantitatively managed	Process quantitatively measured and controlled	1:1 mapping	**Managed**	Level 4 organizations represent approximately 15% of those that are clear leaders in their industry with respect to managing and leveraging information across more than two programs. They take a decidedly managed approach to information	1:1 mapping	**Measured**	Process metrics have been defined and are used for data management. These include management of variance, prediction, and analysis using statistical and other quantitative techniques. Process performance is managed across the life of the process.	1:1 mapping	**Quantitatively managed**	1:1 mapping	**Managed & measurable**	1:1 mapping	**Quantitatively managed**	Standardized tools for data management from desktop to infrastructure, coupled with a well-formed centralized planning and governance function. Process metrics have been defined and are used for data management. These include management of variance, prediction, and analysis using statistical and other quantitative techniques. Process performance is managed across the life of the process.Expressions of this level are a measurable increase in data quality and organization-wide capabilities such as end-to-end data	Processes are measured, controlled. Performance is managed across the whole process.
Optimizing	Focus on continuous process improvement	1:1 mapping	**Optimized**	Level 5 organizations are few and far between, representing fewer than 5% of those throughout the world. They are typically model organizations frequently cited for their EIM excellence that have optimized many (if not most aspects of acquiring,	1:1 mapping	**Optimized**	Process performance is optimized through applying Level 4 analysis for target identification of improvement opportunities. Best practices are shared with peers and industry.	1:1 mapping	**Optimizing**	1:1 mapping	**Optimized**	1:1 mapping	**Optimized**	Process performance is optimized through applying Level 4 analysis for target identification of improvement opportunities. When data management practices are optimized, they are highly predictable, due to process automation and technology change management. Tools enable a view data across processes. The proliferation of data is controlled to prevent needless duplication. Well-understood metrics are used to manage and measure data quality and processes. Best practices are shared with peers and industry.	Process performance is continuously optimized.

B2: AI Maturity Model Levels

Lvl	AIMM	Description	AMM	Description	Rationale	GAIM	Description	Rationale	OAIM	Description	Rationale	Synthesis AI level	Description	Short v0.7
0			**Non-algorithmic**	No decided AI strategy, decisions made solely by humans. Data is not regarded as critical to success	No AI practices or plans exist, which describes a maturity below initial	**Planning**	Internal conversations about AI are ad hoc and speculative. Only a very small minority have experimented or piloted projects. At this stage, the first focus in quiries with Gartner are placed, with topics such as "What is artificial intelligence?" and "How can AI benefit my line of business?" To evolve to the next stage, identify three to six use cases.	Even though some AI experience is mentioned, this stage is to immature to be at the "initial stage", because in this phase AI is actually used by individuals or teams.	**AI Novice**	AI Novices are the most immature phase, where the CSP has not taken proactive steps on the AI journey and at best is in "assessment mode." The CSP will not be in a position to take advantage of the opportunities afforded by AI capabilities, hindered by the lack of a cohesive strategy, limited organizational alignment, and insufficient data availability. First use cases need to be drafted to move to the next stage.	No AI practices exist, combined with a lack of data. Future plans focus on assessing AI technology.	**Planning/novice**	Non-augmented (the most immature phase, where the organization has not taken proactive steps on the AI journey and internal conversations about AI are ad hoc and speculative. The organization will not be in a position to take advantage of the opportunities afforded by AI capabilities, hindered by the lack of a cohesive strategy, limited organizational alignment, and insufficient data availability. First use cases need to be drafted to move to the next stage.	
1	**Initial**	Organizations associated with maturity level one - the capability exists but there is a lack of organizational knowledge. AI responsibilities are decentralized and have no dedicated unit for AI. It is essentially used by individual s' function or team without the close assurance of the organization about the actual usage. Since there is lack of organizational awareness about the usage, it cannot be sufficiently measured and control by the organizational IT. Therefore, there is no AI related governance or regular principles of operation which are extended to AI services by the users themselves.				**Experimentation**	Initial proof-of-concept (POC) project plans are drafted and may be in pilot. An informal community of practice may have met several times for knowledge sharing and early standardization efforts. Aim to prove AI value (not technology) to evolve to the next stage.	This stage describes the first AI uses, which correspond to the initial phase.				**Initial/experimental/planning**	The capability exists but there is a lack organizational knowledge. AI is essentially used by individuals or teams without organizational awareness. Therefore, there is no AI related governance or regular principles of operation. Initial POC projects are starting with the aim to prove AI value in order to evolve to the next stage.	Processes are (manual) unpredictable, poorly controlled, reactive and typically not applied across business units.
2	**Assessing**	Here, the capability is well-developed and the organization has decided to move towards AI - with some initial steps in that direction. The AI substructure is already functioning centrally and basic capabilities such as ad hoc analyzes are provided. Organization initial AI strategy; for each AI application, and defined value propositions are one of the main drivers of AI adoption. However, decentralized solutions still exist and the organization is faced with AI restrictions/issues.							**AI Ready**	CSPs that are AI Ready are in a suitable position to start their AI journey: they are sufficiently prepared in terms of strategy, organizational set-up, and data availability to move forward and implement AI technology and solutions in defined operational scenarios. Such CSPs must take the next step by making tactical investments to enable the relevant skills, technology, and data to start realizing these plans.	Initial strategy, organizational support and available data allow for making plans for AI adoption. However, the organization is not determined yet..	**Assessing/ready**	The capability is well-developed and the organization is ready to start their AI journey in terms of strategy, organizational set-up and data availability. An initial strategy for each AI application exists and a value proposition is one of the main drivers. However, decentralized solutions still exist and the organization is faced with AI restrictions/issues The next step is to make tactical investments to enable the relevant skills, technology, and data to start realizing these plans.	The organization is becoming aware of the value of AI-augmentation and includes it in the strategy.
3	**Determined**	Based on the knowledge obtained in level 2, the organization becomes more conscious of inherent risks and opportunities for AI. Organizations that achieve this level have an AI strategy focused on technology and tools. The organization has standard operating procedures that cover AI scenarios; change management is introduced. The organization also employs AI talent and resources are provided. This level has strong top management which influences the challenge of aligning the AI with organizational goals; IT capabilities can be addressed	**Semi-automated**	Rudimentary AI strategy, algorithms recommend courses of action, humans make final decisions. Data is regarded as business-relevant	Both describe a basic AI strategy and adoption of AI in operations.	**Stabilization**	The first AI projects are in production. An executive sponsor exists. Budget for AI projects is available and protected by executive management. A COE has been instantiated, making AI experts, best practices and technology available for projects. Scaling AI pilots in production is the major focus. To evolve to the next stage, develop an end-to-end process for development and deployment of AI, and stabilize your platform for further AI expansion and governance.	Both describe a basic AI strategy and management support for adoption of AI in operations.				**Determined/semi-automated/stable**	Organizations that achieve this level have an AI strategy focused on technology and tools. The first AI projects are in production. An executive sponsor exists. Budget for AI projects is available and protected by executive management. A COE has been instantiated, making AI experts, best practices and technology available for projects. Scaling AI pilots in production is the major focus. To evolve to the next stage, develop an end-to-end process for development and deployment of AI, and stabilize your platform for further AI expansion and governance.	The first AI-augmentation projects are in production with a focus of scaling these projects
4	**Managed**	Organization capability is very well developed. In terms of achieving the primary goal for this stage there is a well-defined value proposition and full top management support. Regarding the data dimension, the increase of data regarding the source and structure. Additionally, appropriate data science exists to make critical business decisions using AI	**Automated**	Decided AI strategy, introduction of CDO. Routine decisions are made and executed by algorithms. Data science is to make decisions using AI, which could imply that routine decisions are automated. Both describe managed AI processes	Management support results in a decided strategy. Data science is to make decisions using AI, which could imply that routine decisions are automated. Both describe managed AI processes	**Expansion**	All new digital projects, including process overhauls for optimization, consider employing AI as a way of delivering value. New products and services have embedded AI. Accountability around each project is in place. Business units trust AI techniques and are ready to use them. IT DevOps/MLOps professionals are mainly responsible for maintenance and update of AI solutions. To evolve to the next stage, expand data sources and proceed with high risk/high-return use cases.	AI adoption plans and practices are shared across the organization and therefore are managed processes	**AI proficient**	CSPs at this phase in their AI development have a reasonable degree of experience and an understanding of how they want to move forward with AI. But there are still gaps and limitations in their strategy roadmap, data capabilities, and technology resources. These shortfalls affect the range and depth of AI-powered operational scenarios they can address, which ultimately means missed opportunities.	AI is implemented in a variety of processes throughout the organization, these are managed processes. Still some gaps exist to optimize AI usage.	**Managed/automated/expansion/proficient**	All new digital projects, including process overhauls for optimization, consider employing AI as a value driver and source of competitive advantage. New products and services have embedded AI. Accountability around each project is in place. Business units trust AI techniques and are ready to use them to make critical business decisions. Routine decisions are made and executed by algorithms. To evolve to the next stage, expand data sources and proceed with high-risk/high-return use cases.	Routine tasks and decisions are automated. The business increasingly trusts AI to make critical business decisions.
5	**Optimized**	Organizations are at the final level of AI maturity. They are realizing the full role; responsibilities and accountability are clearly defined within each AI project. Data structure is flexible and pro-active to achieve business impact. Altogether, the initial model will be discussed with the AI expert regarding the potential maturity level errors could be found.	**Super Intelligent**	Decisions made by AI autonomously	Both describe a state where AI is utilized to its maximum potential.	**Transformation**	AI is routine and expected as an element of performing all business processes. All workers in process and application design are trained in and understand AI's strengths and weaknesses. AI-powered applications interact productively within the organization and across the business ecosystem.	AI is utalized to its maximum potential (optimized), which results in business process transformation	**AI Advanced**	CSPs with an AI Advanced score have achieved a good level of AI maturity and are ahead of other CSPs in the AI journey. These CSPs have AI expertise and experience, with a proven track record in AI-powered use cases. They are typically more advanced in digital transformation and are often larger organizations with more resources at their disposal to invest in AI. But CSPs that are AI Advanced cannot be complacent and must ensure they keep ahead of new developments in AI and the potential impact on their business (both positive and negative).	Both describe a state with a proven track record and where the organization strives to continuously optimize processes.	**Optimized/advanced**	AI is routine and expected as an element of performing all business processes, decisions are even made by AI autonomously. The organization has AI expertise and experience, with a proven track record in AI-powered use cases. The organization keeps ahead of new developments in AI and the potential impact on their business.	AI-augmentation is routine and expected as an element of all processes. More and more complex decisions are made by AI autonomously.

B3: ADM Levels

Level		DM levels			AI levels		ADM Levels	ADM Level description
0	Incomplete	**No organized data management practices** or formal enterprise processes for managing data. Organizations are aware of key issues and challenged but lack the budget, resources and/or leadership to make any meaningful advances.	No organized data management practices	Planning/novice	The most immature phase, where the organization has **not taken proactive steps on the AI journey** and internal conversations about AI are ad hoc and speculative. The organization will not be in a position to take advantage of the opportunities offered by AI capabilities, hindered by the lack of a cohesive strategy, limited organizational alignment, and insufficient data availability. First use cases need to be drafted to move to the next stage.	The organization has taken no proactive steps in augmenting processes	Undefined	The organization has no organized data management practices or formal processes. The organization is aware of the key issues and challenges but is not in a position to take advantage of augmented data management due to the lack of strategy, organizational alignment, data and resources. To move to the next maturity level, the first use cases need to be drafted.
1	Initial	General-purpose data management using a limited tool set, with little or no governance. Organizations operate in a **reactive application-centric mode**, waiting until information-related problems manifest in significant business losses or lack of competitiveness before addressing them. Data handling is highly reliant on a few experts. Roles and responsibilities are **defined within silos**. Each data owner receives, generates, and sends data autonomously. Controls, if they exist, are applied **inconsistently**. Solutions for managing data are limited. Data quality issues are pervasive but not addressed. Infrastructure supports are at the business unit level.	Processes are unpredictable, poorly controlled, reactive and typically not applied across business units.	Initial/experimental /planning	The capability exists but there is a lack organizational knowledge. **AI is essentially used by individuals or teams without organizational awareness**. Therefore, there is no AI related governance or regular principles of operation. Initial POC projects are starting with the aim to prove AI value in order to evolve to the next stage.	AI-augmentation might be leveraged by individuals or teams without organizational awareness and is therefore not governed, measurred and controlled on organization level.	Initial	General-purpose data management using a limited tool set, with little or no governance. Organizations operate in a reactive application-centric mode, waiting until information-related problems manifest in significant business losses or lack of competitiveness before addressing them. Data handling is highly reliant on a few experts. Roles and responsibilities are defined within silos. Each data owner receives, generates, and sends data autonomously. Controls, if they exist, are applied inconsistently. Solutions for managing data are limited. Data quality issues are pervasive but not addressed. Infrastructure supports are at the business unit level. AI is used by individuals or teams, without organizational awareness. To move to the next stage, the value of general data managment needs to be proven as a prerquisite for augmented data management
2	Managed	Emergence of consistent tools and role definition to support process execution. **Processes are planned and executed in accordance with policy**; employ skilled people with adequate resources to produce controlled outputs; involve relevant stakeholders; are monitored and controlled and evaluated for adherence to the defined process. The organization is getting aware of data quality issues and concepts. Concepts of Master and Reference Data begin to be recognized. Some programs are operational and effective, but there is **little leverage or alignment across programs** and investments.	Processes are executed and planned in accordance with certain policy, but primarily project-based.	Assessing/ready	The capability is well-developed and the organization is ready to start their AI journey in terms of strategy, organizational set-up and data availability. **An initial strategy for each AI application exists and the value proposition is one of the main drivers**. However, decentralized solutions still exist and the organization is faced with AI restrictions/ issues The next step is to make tactical investments to enable the relevant skills, technology, and data to start realizing these plans.	The organization is becomming aware of the value of AI-augmentation and includes it in the strategy.	Managed & ready	Some data management programs are operational and effective, tools are consistently used and roles are defined to support process execution. These processes are managed by skilled people with adequate resources to result in controlled outputs. The organization is starting to recognize importance of concepts such as data quality, master data and reference data. Some decentral applications may effectively leverage AI/automation, but organizational alignment is lacking. **The organization is aware of the benefits of data management and the potential of AI and is ready to make plans to augment data management**. The next step is to make tactical investments to enable relevant skills, technology and data to realize these plans.
3	Defined	Introduction and institutionalization of scalable data management processes and a view of DM as an organizational enabler. Data management capabilities established and verified by stakeholders: roles and responsibilities structured, policy and standards implemented, glossaries and identifiers established, sustainable funding. Characteristics include the replication of data across an organization with some controls in place and a general increase in overall data quality, along with **more formal process definition** leads to a significant reduction in manual intervention. This, along with a **centralized design process**, means that process outcomes are more **predictable**.	Set of scalable and proactive processes is employed and consistently followed across the organization.	Determined/semi-automated /stable	Organizations that achieve this level have an AI strategy focused on technology and tools. The first AI projects are in production. An executive sponsor exists. Budget for AI projects is available and protected by executive management. A COE has been instantiated, making AI experts, best practices and technology available for projects. **Scaling AI pilots in production is the major focus**. To evolve to the next stage, develop an end-to-end process for development and deployment of AI, and stabilize your platform for further AI expansion and governance.	The first AI-augmentation projects are in production with a focus of scaling those projects	Defined & semi-automated	Data management processes are defined, scalable and seen as an organizational enabler. Related capabilities are established and verified by stakeholders: roles and responsibilities structured, policy and standards implemented, glossaries and identifiers established, sustainable funding. Attributes of this level include the replication of data across an organization with certain controls set up and a general increment in data quality. **Processes are formally defined and augmentation pilots generate AI supported recommendations, while processes are primarily manual**. To evolve to the next stage, augment an end-to-end data management process and stabilize the platform for further expansion and governance.
4	Quantitatively managed	Standardized tools for data management from desktop to infrastructure, coupled with a well-formed centralized planning and governance function. Process metrics have been defined and are used for data management. These include management of variance, prediction, and analysis using statistical and other quantitative techniques. Process performance is managed across the life of the process. Expressions of this level are a measurable increase in data quality and organization-wide capabilities such as end-to-end data audits.	Processes are measured, controlled. Performance is managed across the whole process.	Managed/automated /expansion/proficient	All new digital projects, including process overhauls for optimization, consider employing AI as value drivers and source of competitive advantage. New products and services have embedded AI. Accountability around each project is in place. **Business units trust AI techniques and are ready to use them to make critical business decisions**. **Routine decisions are made and executed by algorithms.** To evolve to the next stage, expand data sources and succeed with high-risk/high-return use cases.	Routine tasks and decisions are automated. The business increasingly trusts AI to make critical business decisions.	Quantitativrely Managed & automated	Data management processes performance are managed, centrally planned and governed. Standardized tools for augmented data management are used from desktop to infrastructure. **Routine tasks are automated by AI. Business units trust AI techniques and are ready to use them to make critical business decisions.** Process metrics have been defined, such as variance, prediction and analysis using statistical and other qualitative techniques. Attributes of this level are a measurable increase in data quality and organization-wide capabilities such as end-to-end data audits. To evolve to the next stage, expand data sources and succeed with high-risk/high-return augmentation cases.
5	Optimized	Process performance is optimized through applying Level 4 analysis for target identification of improvement opportunities. When data management practices are optimized, they are highly predictable, due to process automation and technology change management. Tools enable a view data across processes. The proliferation of data is controlled to prevent needless duplication. Well-understood metrics are used to manage and measure data quality and processes. Best practices are shared with peers and industry.	Process performance is continuously optimized.	Optimized/advanced	AI is routine and expected as an element of performing all business processes, **decisions are even made by AI autonomously**. The organization has AI expertise and experience, with a proven track record in AI-powered use cases. The organization keeps ahead of new developments in AI and the potential impact on their business.	AI-augmentation is routine and expected as an element of all processes. More and more complex decisions are made by AI autonomously.	Optimized & advanced	Process performance is optimized through applying Level 4 analysis for target identification of improvement opportunities. **Processes are continuously being augmented to the point where complex decisions are made by AI autonomously**. Tools enable a view of data across processes, where the proliferation of data is controlled to prevent needless duplication. Well-understood metrics are used to manage and measure data quality and continuously improve processes. Best practices are shared with peers and industry.

B4: Data Management Capabilities

-File has too many cells to include in the appendix, Excel file is available on request.-

B5: ADM Capabilities

Capability	Ref.	Reference description	Ref.	Reference description	ADM Capability description	ADM statements
Augmented Data Qualtig						
Assess DQ	1,4	assess large datasets on data quality dimensions using statistics (1) Reverse-engineer business rules from datasets (4)	8,9,10	Perfom validation checks on customizable set of data rules (8) and generate score (9,10) Reverse-engineer and suggest DQ rules (10) Automates end-to-end DQ validation (8)	Assess large datasets to generate a data quality score based on statistics and data rules. Data quality dimensions and data rules can be manually specified or reverse-engineered from the data itself to automate the DQ assessment from end-to-end.	1. Assess large datasets to generate a data quality score based on statistics and data validation rules. 2. Specify data quality dimensions and data validation rules manually or reverse-engineer from the data itself.
Data Profiling	1	automated profiling of datasets by parsing the data to identify the structure and data types (1)	8,9,10	Compare table data, structure and metadata (8) Show basic statistics and automatically identify data categories using NER and ML (9) Parse complex files using NER and NLU (10)	Automatically profile large and complex datasets by parsing the data and recognizing data types, structures, metadata, taxonomies (e.g. email, address) and generating basic statistics.	1. Profile large and complex datasets by parsing the data and recognizing data types, structures, metadata, taxonomy (e.g. email, address) and generating basic statistics
Data cleansing	3,4,5	Automated categorization of DQ issues (4) Learn from manual data cleansing by data steward to provide suggestions and eventually automate data cleansing (3,4,5)	9, 10	Data quality suggestions for cleansing and standardization (10) Supervised learning capabilities can be used to expedite cleansing and standardization activities reducing the human hours spent on these activities (9)	Automatically categorize data quality issues and provide suggestions for data cleansing and standerization. Supervised learning is applied by learning from manual data cleansing to generate suggestions for similar DQ issues or perfom cleansing autonomously.	1. Categorize data quality issues 2. Suggest actions for data cleansing and standarization 3. Learn from manual data cleansing to generate suggestions for similar DQ issues or perform cleansing autonomously
Monitor DQ	6,7	automated data quality checks to detect anomality's from the baseline (6,7). Unexpected deviations can trigger automated data cleansing or request manual intervention	9, 10	Uses ML and predictive modelling to analyze timeseries data. Shows the actual versus the predicted result, significant differences are marked (9) Anomaly detection (10)	Historical data is used to predict expected data values and attributes. Significant differences between the actual data and prediction signals anomalies in data quality, which triggers a request for manual intervention or automated data cleansing.	1.Perform ongoing data quality/validation checks on data pipelines 2. Detect anomalies by significant differences between actual data and expected values from historical data.
Augmented Metadata management						
Define metadata architectu	1,4,7	Automated metamodel generation (1,7) Reverse-engineer business rules from datasets (4)	9	Rationalize data dictionaries through industry taxonomies, folksonomies and client specific taxonomies (9)	Automatically generate metamodels based on data. Reverse-engineer (meta)data rules based on datasets. Rationalize data dictionaries through industry taxonomies, folksonomies and client specific taxonomies.	1. Generate metamodels based on the data. 2. Reverse-engineer (meta)data rules based on datasets. 3. Rationalize data dictionaries through industry taxonomies, folksonomies and client specific
Create and maintain metada	2,4	Automatically generate metadata: topics, categories, keywords, names (2) Extract meaning from speech and video: recognize people, objects, topics Automate end-to-end data lineage, including visualization (2,4)	8,9,10	Generate business & technical metadata (8) Catalog and index metadata, create tags (9) Automate data taxonomy, recognize name, email etc (9,10) Automatically extract attributes/metadata from text and images (9,10) Merge technical and business terms into a repository (12), provide recommendations (10)	Automatically generate metadata from structured data: identify topics, taxonomies (recognize names, address, email, etc) and generate keywords. Extract attributes from unstructured data (text, images, video) to generate metadata. Catalog and index this metadata in a repository that merges business and technical data based on similarity and relation. Automate end-to-end data lineage by creating metadata that tracks data flows and transformations.	1. Generate metadata from structured data: identify topics, taxonomies (recognize names, address, email, etc) and generate keywords. 2. Extract attributes from unstructured data (text, images, video) to generate metadata. 3. Catalog and index this metadata in a repository 4. Merge business and technical data based on similarity and relation. 5. Generate end-to-end data lineage by creating metadata that tracks data flows and transformations.
Analyze metadata	7	Detect outliers from logs and generate a response scheduling failed operations or request manual input (7)	8	Automates the conversion of technical session logs to structured reports (8) Highlight metamodel mismatches and generate reports (8) Validate file metadata & re-structure it (8)	Automate the conversion of technical session logs to strucutred reports to check if data is moved or mapped as expected. Validate file metadata, highlight potential mismatches and missing data and restructure metadata automatically or request manual input.	1. Convert the technical session logs to strucutred reports to check if data is moved or mapped as expected. 2. Validate file metadata, highlight potential mismatches and missing data and restructure metadata automatically or request manual input.
Augmented Data Integration						
Data discovery			10	Identify relationships between datasets, recommend data (10)	Recommend data during discovery, based on relationships and similarity in	1. Recommend data during discovery, based on relationships and similarity in data.
Ensure data lineage	2,4	Automate end-to-end data lineage, including visualization (2,4)	8,9,10	Automate data lineage capture (8), also visualize in ER diagram or table (9,10), perform risk/impact analysis (10) Reverse engineer data lineage from code (8)	Automate end-to-end data lineage and visualize flows in ER diagram. Data lineage can be reverse-engineerd from code or constructed from metadata.	1. Reverse engineer data lineage from code or from metadata. 2. Perform impact analysis 3. Perform root-cause analysis
Design DII	1,2,5,6	Automated data modelling Automate ETL flow creation: schema mapping from source to target (1,2,5,6)	8,9,11	Mapping source to target for DI (8) Ingest data from unstructured sources (9), discover attributes and structure for optimal storage (11)	Automate data modelling and map from source to target for structured and unstructured data discovering attributes and structure.	1. Model data for structured data, discovering attributes and structure 2. Model data for unstructured data, discovering attributes and structure 3. Map source data model to target data model
Develop DII solutions	1,2,5,6	Automate ETL flow creation: schema mapping from source to target (1,2,5,6) Automate data integration in low-code platforms and code completion (1) Supplement incomplete data (6)	8, 10,12	Automatically migrate data from one database, data warehouse or server to another (8) Automate data ingestion pipelines (8) Automatic data ingestion, validation and transformation (8) Automatic data pipeline creation (8) Auotomaticlly integrate new data by learning from existing mappings and user interaction (10) Create low-code DI solutions (12)	Automate data integration completely or partially in low-code tools that map source to target schemas. Data is automatically ingested, validated and transformed to the target structure, based on exisitng pipelines and manual integrations.	1. Design data integration flows that use standard templates to map source to target schemas. 2. Ingest, validate and transform data to the target structure based on existing mappings and user interaction
Augmented Master Data Management						
Evaluate and assess data sources			9,10	Generate data models and linkages: identify relationships between columns across different sources (9,10)	Generate data models and linkages: identify relationships between columns across different sources.	1 Generate data models and linkages: identify relationships between columns across different sources.
Model Master data	2,3	Create a single view of customer/product from multiple sources (3) Automated reference data management, e.g. international tax codes (2)	8, 10	Automate MDM hub configuration as per data model (8) Entity and structure discovery: map data domains into hierarchical business etities (10) Detect and map master data entities on the MD model (10)	Model master data bottom up: recognize entities and hierarchical structure to create a single view of customer/product. Or top down: detect and map master data entities onto a master data model. Automate MDM hub configuration	1. Generate a master data model by recognizing entities and hierarchical structure (bottom-up) 2. Detect and map data entities onto a predefined master data model (top-down) 3. Configure MDM hub as per data model
Define stewardship and maintenance process	1,3,6,7	Identifying duplicate data and analyze potential matches (1,3,6,7) Learn from data steward labeling of master data to generate recommendations and automated master data identification Identification of data attributes and overlap between systems (1)	9, 10	Match data records based on clustering and blokcing attributes (9,10) Learn resolution & reconciliation rules from human intervention (9)	Identify duplicate data based on clustering and blokcing attributes. Learn resultion & reconciliation rules from human interventions to generate recommendations or autonomously establish a single point of truth.	1. Identify duplicate data based on clustering and blokcing attributes. 2. Learn resultion & reconciliation rules from human interventions to generate recommendations or autonomously establish a single point of truth.
Augmented Database Management						
Manage database performa	5,6	Use usage forecast to scale computational resources and balance computational load (5,6) Assisted query optimizer for faster response times (6) Automated fault recovery (5) Monitor for anomality detection from database logs which can trigger an automated response or request DB administrator interaction (5) Automate patching, configuration, release management and scheduled downtime (5)	10,11,12,13	Query optimizer (10,11) Anomaly detection in run jobs (10), fault recovery (12,13), threat detection (12) Predict and scale resources to meet performance criteria (10,11) Handle external system issues autonomously, e.g. by adding cloud resources (10) Automated provisioning, management, montoring, backup, recovery and tuning (13)	Forecast computational demand to scale computational resources to meet performance criterea. Optimize queries for better response time and schedule queries to match available resources. Monitor database logs to detect anomilies in run jobs to trigger automated fault recovery, threat detection or request manual interaction. Automated provisioning, management, monitoring, backup, recovery and tuning of databases.	1. Forecast computational demand to scale computational resources or schedule jobs based on available resources to meet performance criterea. 2. Optimize queries for better response time. 3. Monitor database logs to detect anomilies in run jobs to trigger automated fault recovery, threat detection or request manual interaction. 4. Provision, manage, monitor, backup, recover and tune databases.

B6:AI Capabilities

Dimension	AIMM	Description	AMM	Description	Rationale	GAIM	Description	Rationale	OAIM	Description	Rationale
Strategy			**Strategy**	-	1:1 mapping	**Vision & strategy**	-	1:1 mapping	**Strategy**	The strategy pillar examines the state and nature of a CSP's plan of action and road map to support AI.	Pivot
Data	**Data structure**	Data refer to containing both the amount and structure of the data to getting AI systems to work by enabling high-velocity capture, discovery or analysis.	**Data**	-	1:1 mapping				**Data**	This pillar assesses the state and availability of a CSP's data assets and its analytics capabilities, as these are crucial for a successful AI deployment.	1:1 mapping
Organization	**Organization**	Describes business characteristics and resources that might influence AI process such as firm size, managerial structure, decision-making and	**Organization/people**	-	1:1 mapping	**Organization & Governance**	-	1:1 mapping	**Organization**	This pillar examines how culturally and organizationally ready a CSP is to support AI and its effects on business transformation.	1:1 mapping
People	**People**	People refers to all those individuals within an organisation to create artificial intelligence technologies.		-							
Technology	**AI functions**	AI functionality refers to the tools and technologies that are required to handling AI at scale	**Analytics**	-	Maturity description mentions tools and techniques	**Technologies employed**	-	Indicators mention tools and techniques	**Technology**	This pillar explores and assesses the different AI technologies and capabilities being leveraged by the CSP, and how the CSP has gone about implementing AI solutions.	1:1 mapping
Operations			**Decisions**	-	Maturity description mentions operational tasks and processes	**AI Usage**	-	Indicators mention operational processes	**Operations**	This pillar assesses where and how CSPs are implementing AI across four core operational elements: customer support, sales and marketing engagement, networks, and fraud detection management. The associated questions explore a range of potential use case scenarios, in both a B2C and B2B context.	Pivot
Budget						**Budget & measures**	-				

C. Transcripts of Expert Interviews

	Permission to record is asked before start.
	
	General Questions:
	• How to leverage AI technology: ML, NLP, Expert systems, Vision recognition, Speech recognition, Planning and robotics
	  ○ Do you agree that these are the main AI sub fields, of would you add/remove some?
	• To augment data management capabilities: metadata management, master data management, data integration, data quality, database management etc.
	  ○ Do you agree that these capabilities have the largest potential to leverage AI?
	  ○ How is AI leveraged within these capabilities and what would be future applications?
	• Are you familiar with maturity models? Think it could be useful for AI in DM?
	• Want to be updated on my research and/or participate in future interviews?
	
	1[redacted] – 13-05-2020, Deloitte Canada
	
	1: I work at Deloitte OMNIA, which is a department specialized in AI. Around 400 people work in OMNIA, with around 200 based out of data management and analytics.

DI	1**:** In my experience, around 80% of AI is data preparation, which is not all done by data scientists and machine learning experts. There are many tools available for data preparation, but often we build custom solutions.
MDM/DQ/DI	1; For complex data modelling, machine learning is used.
MDM	1: There's one case of a large airline company which had no data governance. They started a loyalty program, with data spread over more than 25 systems. Customer login platform was from SAP. Customer data platform from another data platform. These two companies had a clash on who owns the customer data. What we did was build a customized data governance framework. We distinguish declarative customer data and enhanced customer data. For customer data SAP was the system of record, for enhanced data in the other systems. We identified the data attributes: 1700+ and showed the overlap between systems. Then we thought: if this could be automated that would be good enough.

Example	1: We put up an automation proposal for them, but they didn't have the budget. They rather do it with more people. Now they handed it over to another party who is going to automate it. We did master/reference data management, so that it is clear where what data is and who owns it D: Did you use any DM framework? 1: Yes we have a general reference framework, but we build one that suits the client
DQ	1: In another case we worked for a large transportation company. Data coming from ERP, CRM and reporting streams. Master data, reference data, transactional data, informational data. We put up a proposal for them to first we do a profiling, data quality check, for each of those sets. Then we are going to create a data strategy for the company. Then you can build your platform to automate it. D: Do you use AI in data profiling?
DQ	1: We did data migration at large airline company. Another company failed, so they came to Deloitte. Using AI concepts, they build up tools which profile the data, map the data. They were able to map 80% automated, only using a few 'human' business rules D: Do you use AI in data quality?
DQ	1: When you set up a data quality platform for any company, the system provides a low-cost profiling tool. It will display scores for missing values and correct values. Then business sets a benchmark for correctness. This might not always be right: for example, one system put everything in one line so 100% filled but not correct.
Example	1: When you talk about MDM, Informatica, Oracle, all provide you with some kind of intelligent framework. For example about the airline. This platform had a customer relation module build inside, where all customer data is captured. If someone books a flight with only a name and nothing else, it will try to keep merging it with other data. If the data is complemented, it will link the different data sources.
Example	1: Data integration is almost drag and drop nowadays. If you use tools like Alterix, cloud providers like Azure also provide a lot of services in a low-code platform.

DI	1: There's a lot of AI in conversion/migration: take source, try to map it to target. Deloitte Omnia can automate 60-70% of the activities. In this case they were migrating the mainframe into 28 different systems. This is a very complex task for a human, but east for AI. D: Do you have any experience with maturity assessments? 1: Must be 2/3 days ago, with the Deloitte framework. D: Do you think it would be a good addition to develop a maturity model for augmented data management?
Demand	1: Surely. However, maturity models change quite quickly. Something from 2017 might not be bad now. **2: [redacted] – 14-05-2020, +- 5 years at Deloitte** 2: I'm a Senior manager at analytics and cognitive from a more technical perspective. I have experience with classic data management, but also innovative projects with NLP, chatbots, RPA. So I've done projects where these two come together D: Do you use a specific DM framework? 2: Also DAMA DMBOK D: Which capabilities do you think AI can play a role?
MD/DI	2: Data lineage has a large potential for AI, can be seen as metadata management
MDM	2: Within MDM: deduplication is perfect for AI, NLP can be used.
DI	2: Data integration: creation of ETL can be automated
DQ	2: DQ might be the most important one: data profiling, creating analogies
Example	2: AI functionalities can be custom built in Python or in tools like Informatica, Microsoft SSIS. These tools will integrate AI to a level where you don't notice but it is an integral part of the solution

	2: Customers don't use AI functionalities. The focus is identifying duplicates. I think they use the Stanford library for NLP, where entity extraction is improved. D: Do you see voice commands growing in this field? 2: Deloitte created multiple chatbots. In general, this is upcoming like with more and more voice-controlled devices. I think web apps will stay as an interface. D: How about automated metadata?
MD	2: Yes. Microsoft sharepoint already automatically generates tags. It looks in the text to find common words and duplicates to generate metadata. D: Also for audio and video?
MD	2: I think this is technically feasible**,** but the privacy aspect might be harder. Even in teams it is possible to generate live captions in English. Functionalities like this will be integrated, where it is not explicitly mentioned D: Have you seen automatic integration and mapping?
DI	2: Yes, I've seen it promoted but I haven't seen it in practice. It is quite time intensive to program these mappings D: Have you seen other solutions that reduced manual work?
DI/MDM	2: We've created a custom solution for tax-codes. In order to have international trade, every product has a specific tax-code. Previously, there would be people that managed this catalog and assigned the right codes. Now, this is completely automated. This is an example of reference/metadata D: Do you see chances in other areas aswell? 2: Data governance, Colibra for example, Qlick Sense, Qlick view. Colibra incorporates data quality rules, also allows to map datasets like the tax-code example. D: Are you familiar with maturity models?

	2: Yes, IDO has a maturity assessment. D: Do you think it is useful to develop a maturity model for ADM?
Demand	2: Yes. Some technologies are included in IDO, from RPA, to machine learning to generic all-purpose AI
	3: [redacted] – 15-05-2020 2,5 years at Deloitte 3: I'm a Senior consultant analytics and cognitive. 5+ experience in data management. Focus on MDM, Data quality, Informatica. I'm in the Informatica expert community of practice for Deloitte UK.
MDM	3 :There's A lot of focus on MDM: 'single view of customer'. Informatica leads the market in MDM, DG, DQ, DI
Example	3: The closest I've seen where AI is applied to DM was this accelerator called Cognitive Data Steward**.** Deloitte US has developed this accelerator, which is probably one of the highest selling in A&C. They use an open source platform to automate steward activities. I know a case at a Oil & Gas company, where this was implemented. I can direct you to the chief architect of this program. D: Do you use DAMA DMBOK? 3: Every member firm has their own solutions, but I believe it is quite similar
General	3: Informatica primary does Data integration, DQ, MDM, Meta. The overall platform is CLAIRE. One of the examples is the 360 degree customer data view. I could send you some material on Claire 3: A lot of tools are business friendly, user friendly, not a lot of hand coding that you need to do. There are other tools in the market as well, one specific in Houston that had a lot of NLP, ML like applications for DM. They had specific tools that automate DM activities, but it was largely focused on the oil & gas industry. I can refer you to some people and send some more material. D: Have you seen outlier detection being applied?
Example	3: Tools like informatica might come with AI empowered functionalities, but it is black box**.** So you wouldn't know about these functionalities that might be there. Deloitte is implementing these tools and gathering

	requirements, but is not on the back-end. Other tools are Reltio, Stebo, IBM with Watson, SAP MDG
MDM	3: Within MDM find duplicates using matching rules. If you want to integrate/merge these records, you can specify some business rules, for example trust scores, and the system can take over and do the rest.
	D: Have you also seen the use of reference directories?
DI	3: Yes, they do suggestions. Also from additional sources like social media, to enhance your customer data.
	D: Do you see any manual activities where potential is?
DQ	3: A lot of data quality work. It kind of goes hand in hand with the data stewardship, was they do a lot of manual work**.** I'm not sure if I have access, but I can refer you to people working with Cognitive Steward.
	D: Are you Familiar with MM?
	3: Yes.
	D: ADM MM good addition?
Demand	3: Yes! A lot of our clients are non-technology. Traditionally they didn't have any data teams in place. Now they have and are heavily involved in creating custom solutions. If you include ADM in a MM, it will help evaluate clients technology stack and what kind of technology they are developing
Demand/ example	3: Every company wants to be the Google or Amazon of their industry. Everyone wants to build their own market place and own custom set of tools. That's why some customers are building a custom solution instead of something out of the box like Informatica.
	3: I'll ty to connect you to some folks from the US, as they are a bit ahead of the rest of the world. Put more money and effort in
	4: [redacted] 15-05-2020 9
	4: I'm a Senior manager A&C, 9 years at Deloitte, 3 years in AI strategy
	4: Everything needs to be AI, while in practice many organizations are not ready. For the last years I'm looking on how to bridge the AI capabilities that vendors present and practice. Why not? People are

Demand	hesitant, infrastructure needs to be ready, sales wants to claim AI while there is not much AI
General	4: a Maturity model needs to go from: AI readiness -> foundations -> first use case -> full scale. What was top of the line 10 years ago, is basic now. Maybe add extra step above 5: increased automation 4: I think Dutch colleagues created an online questionnaire for maturity assessment for the IDO framework. There are many maturity models for individual frameworks
Example	4: Informatica is market leader in data management tools. For Claire they have a roadmap for each capability and use cases. Informatica has high R&D budgets, so with there you can se where the market is going. Data steward is assisted by Claire in categorizing DQ issues, comparison of terms
Example	4**:** Within Deloitte we have Digital FTE's, which are also related to data managements
DQ	4: First step of AI is often the automation of small sub tasks, then spreads to more complex tasks. The start phase is 'supervised', so not automated but suggested. The end decisions are still at the data stewards. These data stewards can than identify in which tasks and which domain the AI is always right. Then this task can be automated D: Do you built custom solutions or use software vendors?
Example/DQ	4: We're also building solutions our self. When you're defining metadata, you use business rules. For example for data fields like date format. In large organizations, these business rules might not be documented. Our solution reverse-engineers these business rules, to automatically generate metadata. These rules can then be applied or used to identify mistakes
Demand	4: More and more companies are going to a CAP architecture. This is a streaming architecture, event-driven. Most data lakes die because of a lack of metadata: going from a data lake to a data swamp. The challenge is to have a transparent view of the contents of a data lake, this is where metadata comes in. Normally, you would be responsible to add this metadata if you want to add data to the data lake, but as the data is being streamed this is not possible to do manually. So if you want to go

	to a streaming cap architecture, AI-driven/automatic metadata is the only viable option.
	5: [redacted] 18-05-2020 5: I'm 9.5 years at Deloitte. Experience in Oracle implementations, later on a focus on data migration. Last 5/6 years responsible for data migration projects and Informatica MDM. Within Oracle EDM practice with 20/25 colleagues, also collaborations with EA and A&C. 5: Recent years we've built a lot of capabilities around data management, from data governance as well as technical Informatica implementation. There have been collaborations between A&C and DM, but in the Dutch Deloitte department we haven't seen much collaboration, but it is slowly coming.
Example	5: Informatica and Oracle have embedded AI technology within their products. I think it is more 'AI light', more like chatbots and RPA. This is hidden for the user. In the coming years I think we will do more of those AI empowered tools, implementations at customers. But for us there is not a specific ask D: Maybe customers are not aware that they use AI?
Demand	5: Correct. Customers ask us to implement a new MDM solution and AI might be part of that. Other firms, for example the US have seen more of those project with an emphasis on AI.
DI	5: There's this Data lineage tool: Axon. If you look for certain data, it will analyze where it comes from, find related data and gives a ranking for usability. This is a piece of AI. I think that we can improve awareness here a bit, so that we can communicate this to customers. 5: Oracle has enterprise data management , back in the days there was customer data hub and product data hub, but this is combined into one now. Whether there is AI implemented, I'm not sure, and if there is it might be hidden.
DQ/Demand	5: In my opinion, one of the low hanging fruits would be data quality. Within DQ it is relatively simple to add value and processes require a lot of manual work. For example, Informatica IDQ has out of the box functionalities, but the corrections are still assigned to manual work. These are typical processes where AI can learn from a human data steward and then perform these actions itself. AI is probably present in the triggers itself.
MDM/Demand	5: To put into perspective: customers often have their own ways of working (legacy) with data related as customers, products, vendors, the three most important MDM constructs. Customers come to Deloitte that they want to

	switch to a new way of working, by utilizing products like Informatica. In the transition from old to new, there are a lot of manual transformations. I think there is a lot of automation potential for AI. The most implemented Informatica products: IDQ, 360 customer, 360 vendor, 360 product, MDM multi domain D: Within Database management, do you see load balancing?
DBMS	5: Yes, Oracle has autonomous database. Oracle's promise is that many tasks, such as patching and scheduled downtime will be performed automatically. This means that you need less DBA's and faults are automatically recovered. D: Do you seeChatbot/NLP interface?
General	5: Yes I think Axon and EDC has this. In the Netherlands we didn't implement this yet. Pieter Hens from the Belgian office has more experience in this. D: Are you familiar with data management models? 5: We use a custom Deloitte DM framework which is similar to DAMA DMBOK. This also includes a MDM roadmap, based on maturity. Often we do an initial MDM maturity assessment D: Do you think it is useful to develop a maturity model to augmented data management?
Demand	5: I could 'call, raise'. I've neem saying this for years. MDM is important and fundamental to organizations. But the combination with AI and possibly analytics would make this more attractive. If you can combine MDM with AI and analytics (example customer 360 insights), you can have an incredible offering for clients. **6: [redacted] 26-05-2020 – Assistant professor UvA** 6: I have a broad definition of the data management field, whenever you need to store or use data. It encompasses nearly everything in computer science. From a business perspective this is mostly transactional databases where you basically store what's happening in your company and you make sure you don't lose data and analytical databases that are used for reporting, analysis and planning. Also for AI you need an approach to store you data, so that's also an angle D:Do you see AI in DBMS? 6: The main advantages of databases is that they are simple to use. The majority that works with databases doesn't know how it actually works, they don't need to know. You only need SQL or SQL tool and the database will figure

DBMS	out the rest. DBs have a lot of internal components, like the query optimizer to give you a fast answer without using many resources. That is an area where a lot of researching in optimizing the query optimizer using AI. Another trend is to move DB infrastructure to the cloud. From a business perspective it's mostly cost savings. From the cloud service provider there's a lot of research into using ML for optimizing the cloud deployment of DB. For example, DB's are hard to configure. Usually this is done by a DBA, it is often hard to find DBA's and they can do a limited amount of DBs. A big part of research is to use ML to automatically tune these DBs, monitor their performance and automatically optimize them. D: Do you also see AI for configuration or load balancing?
DBMS	6: This is mostly for the configuration of the DB. Another maybe more researched area about the cloud is that you can scale your computational resources, you can add more machines. That's also an important area where a lot of ML or forecasting techniques are explored. D: Is that from the CSP or consumer side? 6: That CSP, but also other startup companies like Snowflake that are not CSP, but build/design DB's that are really made for the cloud. D: What's the difference?
DBMS	6: I think their main difference is that they automate all the scaling decisions. That is really hard to do, if you would manually do it you would need to constantly monitor the load. If it's read only it's simple, but when editing data this gets more difficult. I think they automated this for their customers to be cheaper and faster than their competitors, mostly cost savings. Especially compared to traditional databases, like oracle, where you buy. D: How aboutAutonomous DB? 6: There's decades of research on ADB's. I think the main direction is that you don't need to manually tune the DB. Check out the video by Andy Pavlo from Peloton D: Do you see AI in data quality? 6: In traditional data infrastructures you had a data warehouse and before storing the data, you would come up with a schema for the data with some integrity constraints, everything was very much defined. The problem is that with the rise of big data and ML techniques, the paradigm to work with data has changed. It's more like: we need to collect all this data and then hire some smart people to figure out how the data could be useful for the business. The problem with that is that you don't know what data you need before you

DI	explore it and you also don't have a model for the data. So often people just store a lot of data and not in a DB because its hard to come up with schema's and its relatively expensive. The problem is that if you collect the data and store it in a cheap cloud file system, it is easy to get DQ problems. What usually happens in companies is that people work with this data and at some point in time something goes wrong. People are alerted, it gets fixed and some rule is introduced to prevent it from happening again. That's very reactive. A large part of my recent research is to automate this. In software engineering it is common to build in tests while developing to make sure it works correctly. We're making something similar for data quality. We're building a library that uses checks with basic statistics on data pipelines. People were still writing this.
Monitoring	The next line of research is to automate this: if I regularly use a data stream and nothing crashes, I can use this data stream as a baseline to detect anomalies D: Are these checks at data ingestion or before usage?
Monitoring	6: Usually when you collect data you don't have a baseline so you don't know the value. If the data looks valuable you will store it in a data warehouse, but then you don't need these checks because you can do it in the data warehouse. But people very skip the data warehouse, because it expensive and takes a lot of effort to create and maintain a data schema, so people use different tools for that now. Tools like apache spark, python pandas, which is quite flexible and ad hoc. We've come up with checks that fit in this new architecture pattern. This library runs statistics on dataframes in a simple and
DQ	effective way. The next step will be that you automatically generate the quality rules based on the data, but that is very difficult. D: Does it also provide suggestions?
DI	6: This library doesn't, but I've worked in projects that did. One important problem is that you can have incomplete data. In many cases there's come downstream system that cannot handle incomplete data and you have to come up with a solution to fill this in, this can be done by ML D: How about Data lineage?
DI	6: Important and difficult topic. Nowadays, a lot of decisions are (semi) automated, for example loan applications, you should understand these automated processes. We're looking into libraries like pandas or scikitlearn and instrument these libraries and try to automatically record what they are doing to pieces of the data and use techniques from DB's to track lineage. Some open source applications are openDB and MLFlow.
MDM	6: Mike Stonebreaker is saying data integration is the biggest unsolved problem, so that a good area to look into. There are many individual problems. For example schema mapping when schema's are not 100% alike, finding a

	common schema. Another problem is entity matching, where two DB's refer to the same entity, duplicate detection.
General	D: Are you familiar with maturity models? Not really. There's a great guide by Google: rule for machine learning, talks about the different phases that ML projects go through at Google and also talk a lot about the infrastructure and reliability of data that you need. They call it phase 0. It takes a lot of time to get the infrastructure right and I think that is the differentiating factor.
	7: [redacted] – 26-05-2020 – Manager Deloitte EDM D: Have you seen AI in Metadata?
MD	7: Not seen it in reality. For all DG and Metadata project I've done, we created the metadata ourselves. This can be automatically generated, but standards should be supported and validated by the business. In sectors where there are standards enforced, like the banking industry, these standards are no-brainers so in those kind of things AI can assist a lot.
	D: It is all manual now?
MD/Demand	7: Yes, in my projects. Actually nobody cares about metadata. It's a difficult discussion with business because metadata is very difficult to substantiate how good and clean metadata will help you increase revenue. This can be substantiated by the time stewards spend on working with crap data and time is money. At the end of the day metadata always lags behind in the mind of the business and nobody wants to invest a lot. So, for your thesis its important that you differentiate theory and pragmatic standpoint and give a solution. You can say that AI can help in these things, but you have to understand that AI is getting mature and data management is still not mature after 20 years. Now people are starting to realize the potential of data. In you thesis you can mention how AI can help in an area and this is the way we propose the business to approach is via AI, because they don't want to spend a lot of money on metadata. That's why everything is done now manually, they don't want to spend a lot of money, but they're still spending money and that's where AI could help.
	D: I think people (customers) that come to Deloitte already see the added value and want to improve?
General/ Demand	7: The approach (motivation) per industry differs. People in finance don't want to improve their data management, unless it's a regulation. So when they want to improve, it's more defensive. In other industries. Like energy or pharmaceutical. They have a more offensive approach, when you need data to optimize your production. Industries differ in their appetite to improve in data management and spend on AI.

Demand	7: I think in many cases AI can be cheaper. Those are the things you have to find out. Like the metadata example, people think its cheaper to do it manually than using AI, buy they end up spending more money and longer lead times. Without data dictionary and metadata repository. D: Other areas where you see aI?
MDM	7: Lets take MDM, I give you one example where it will work and one where it won't. When you have data from multiple sources and the system has to assess if it's the same record or not. To do that you have tolerances for a match, no match and potential match. Current MDM solutions can distinguish between match and no match. AI can play a role when there's a potential match. Previously, these potential matches had to be manually reviewed by data stewards. This can be automated by AI, but it needs a lot of data and logic. Where it will not work is in FSI, when you compare incoming stock prices. This is still done manually because banking relies on more on people so that they don't lose revenue. But, if AI can be proved to be as good as people, then AI can help. D: Cogni stewards?
MDM	7: Yes good fit. Example with assigning trust values to sources. This is already done in hierarchical systems or survivorship model (MDM): decision threes to categorize. Also in this example, the no match and match are easy to decide but for potential matches there's still a lot of repetitive manual work that can be automated by AI. D: AI in Data quality?
DQ	7: Data quality is very vague and every organization has their own definition. That makes it hard to design the AI engine, but of course AI can help. For every industry you can have a set of predefined DQ KPI's, maybe the presentation can be modified. The system can do automated checks. In a recent project they only did random checks, but AI can do a scan in a more massive and exhaustive way that a human cannot do. If it can generate data quality reports periodically, that can be a huge asset for an organization.
DQ	7: The data quality you are talking about is mostly from literature. Most large organizations often don't have measures for data quality and its really an issue. They are not plugged into data quality tools, because they think it is not of enough importance to spend money there. They spend money on DM and don't want to spend on different facets of DM.
Demand	7: Now, data as an asset, as in that data is more expensive than gold now. That is a realization that has started a couple of years now. In the coming years we will see more challenges, the world will change and we will rely more on data. D: Do you think it is a good addition to have a maturity model for ADM?

Demand	7: It will be of course an important addition. But, an addition also means additional cost. If we do a pure play MM project and then you add this as a step for each component within DAMA DMBOK, then you're looking at more cost. If it's cost-effecitve then sure, it's the future. There should not be a MM assessment without AI. What you say makes sense, but they need to know whether spending on AI will change their business processes or data dramatically. I think we should include that in the assessment. D: Do you have any other requirements? 7: Its also political. AI is not generally accepted yet and the perception is that people will lose jobs. It might not be true, but it's the perception. In a previous project people lost their jobs and it was a large issue. Offering an out of the box solution might be something that Deloitte doesn't want to do as it might scare clients. The MM should be a plus, an add-on, not built into every component.
General	

Table C1: Transcripts of Expert Interviews

Labels

DQ	Relevant for data quality
MD	Relevant for metadata management
DI	Relevant for data integration
MDM	Relevant for master data management
DBMS	Relevant for database mangement
Example	Example of (augmented) data management projects or tools
Monitoring	Relevant for monitoring of data in general

Table C2: Lables for Qualitative Content Analysis

D. Market Research Extended

Deloitte DFTE [91]

Digital FTE's are a set of tools designed to augment the human workforce during projects by executing repetitive and rule based activities without human intervention. These accelerators work with minimal input and aim to reduce operational costs by automating tasks and eliminating human errors. This enables humans to focus om more complex and value adding tasks. The following DFTE's are related to data management:

Ready for pilot:

- Cloud Service Migrator: automatically migrate op-premise database to Oracle cloud service.
- Local ADW Migrator: automatically migrate on-premise server to autonomous data warehouse.

- Cloud ADW Migrator: automatically migrate Oracle database cloud service to autonomous database warehouse
- LogEasy: automates the conversion of technical Informatica's session logs to structured reports.
- Informatica Metadata Validator: highlights metadata mismatches between the source and target schemas and creates custom reports.
- Azure Ingestor: Automates data ingestion pipelines for structured files, perform data quality validations and ingest to data lake using Azure Databricks.
- Data Health Inspector: performs validation on client data against a customizable set of data rules.
- Automate Data Integrate: automates Oracle Data Integrator interface creation (mapping source to target).
- Data Health Advocate: compares table data, structure and metadata between two different databases and provides a validation summary.
- Data Quality Accelerator: Automates the end-to-end processes of Informatica data quality validations.
- SFCloudCast : Performs standard Salesforce object data loading from source to target through automatic data ingestion, validation and transformation
- Infa MDM Configurator: automates Informatica MDM hub configuration as per the customized data model, match-merge process and trust setting.
- Oracle ODI Doc Builder: automates Oracle Data Integrator project documentation and data lineage capture
- SAP2GC Data Ingest: Automating data pipeline creation between SAP HANA and Google Big query

Live:

- D-Modeler: Generates business & technical metadata and domain KPIs from source systems
- UNDIAL RE: Reverse engineers data lineage from code, generating source to target mappings and visualization capabilities.
- D-Ingest: Provides a set of meta data driven data ingestion pipelines for structured and semi-structured files, including file validation and destination routing.
- Botomatica: Metadata driven tool to automate simple data ingestion.
- D-Lineator: automation tool for generating simple pass through data ingestion between multiple systems.
- AutoTest: validates file metadata & re-structures it

Deloitte CogniSteward [92] **(DQ, MM, MDM)**

Deloitte CorniSteward is an advanced data management self-service tool that augments manual, costly and time consuming data steward activities. The solution can either be used as an accelerator during a project, or can be part of the deliverable where it is continuously being used by clients to handle complex and vast amounts of data. The CogniSteward provides the following features:

Data Ingestion:

- Ingest metadata and data from Structured Sources
- Extract data from unstructured sources, such as reports, e-mail, scanned documents and images

Data Quality:

- Data Profiling: show basic statistics and data types. Data categories are automatically identified using name entity recognition and machine learning.
- DQ assessment: specify DQ rules on columns, categories and technicalities (e.g. value must not be blank) and generate a score for these checks.
- Data Cleansing: Supervised learning capabilities can be used to expedite cleansing and standardization activities reducing the human hours spent on these activities
- Monitor DQ: uses ML and predictive modelling to analyze timeseries data. It shows basic statistics for the dataset. It shows the actual versus the predicted result, significant differences are marked.

Metadata management:

- Rationalize data dictionaries: Contextualize through Industry Taxonomies, Folksonomies (Tagging, Crowdsourcing) and Bespoke Taxonomies (client specific)
- Catalog and index Metadata: automatically generate metadata tags and generate a catalog, increasing the findability.
- Dynamically generate data models and linkages: identify relationships between columns across different sources, proving a unified view of the enterprise data.
- Dynamically generate data lineage: set similarity thresholds for the header and content, produce a similarity table and visualize data linkage in an ER diagram .
- Taxonomy: uses NLP and DL to automate categorization. It uses a dataset with known categories to predict a category for a new data set. Only below a certain confidence level, manual review is needed
- Attribution: automates attribute extraction by leveraging NLP and image processing. Based on a learning data set, CogniSteward extracts attributes from the product descriptions and images and predicts attributes for a new data set. Attributes with a low confidence level can be manually reviewed and corrected.

Master data management:

- Data Linkage: identify relationships between columns across different sources, proving a unified view of the enterprise data. A threshold can be set for similarity in header and content, which produces a table with matching scores and an ER diagram.
- Match data records: (Entity resolution & Reconciliation)
- Deduplication: match data records based on thresholds for clustering and blocking attributes. Utilizes supervised learning from human correction to learn resolution & reconciliation rules.

Informatica [88] **(MDM, Meta, DQ)**

Informatica is a data integration and data management software company. Gartner identifies Informatica as a market leader in data integration, data quality, master data management and metadata management tools. Informatica introduced AI/ML functionalities under the name CLAIRE, which supports various solutions on their intelligent data platform.

- Data Cataloging: scan and catalog datasets for data discovery.

    - Relationship discovery: Machine learning is used to automatically identify relationships between datasets by identifying primary keys, unique keys and joins across datasets.
    - Data similarity: Similar data is identified across datasets by using clustering algorithms.
    - Domain discovery: data fields are automatically classified and given semantic lables such as phone number, first name, email and company name
    - Entity discovery: Data domains are combined into hierarchical business entities. For example: order file has a customer, which has an address that consists of a street, city, state and zip.
- Analytics: streamline data preparation for analytics
    - Synthesis matching: leverages ML and NLP to discover non-obvious relationships based on contextual attributes. For example by identifying household relationships based on web sessions and customer support interaction.
    - Optimized at-scale processes: Cost-based optimization to change the join order in a data pipeline for optimal performance.
    - Join-column recommendations: Suggest join columns when the user combines multiple datasets.
    - Apache Zeppelin recommendations: Automatically suggest visualisations
    - Data recommendations : Provides users with suggestions for better-ranked and similar datasets to complement or substitute other datasets.
    - Structure discovery: Automate file injection and parse complex files using NER (name entity recognition) and NLU (natural language understanding) to discovery and visualize the structure.
- Data governance and compliance
    - Automatic DQ enrichment: Create metadata labels to classify unstructured text by using NLP and NER.
    - Associate business terms with physical datasets: Recommend relevant data elements to be linked with business terms.
    - Assess DQ: Automatically execute DQ assessments based on DQ rules for various dimensions.
    - ML/NLP assisted DQ rules: Automatically reverse-engineer and suggest DQ rules based on the dataset.
- Data Privacy and Protection: Identify and control sensitive data.
    - Subject registry identity mapping: Identify correlation to sensitive data for privacy compliance.
    - Sensitive data mapping and movement: Visualize data lineage for sensitive data and monitor possible compliance violations.
    - Risk simulation plans: Evaluate protection techniques and calculates a risk score and impact for data stores.
    - Anomaly detection: Identify unusual behavior by using statistical and ML approaches on a multidimensional model of user activities.
- DataOps
    - Predictive analytics: Auto-scale of data management runtime resources.
    - Anomaly detection in run jobs: Automatically detect anomalies related to Informatica jobs and data processing.
- Future capabilities:

- Self-integration: Automatically integrate new data automatically by learning from existing mappings and user interaction.
- Development assistance: Provide recommendations for auto-completion, templates, security, data cleansing and automatic performance optimizations.
- Auto-mapping: Detect and map master data entities to the master data model.
- Self-heal: Handle external system issues. For example by adding additional cloud resources.
- Self-tune: Predict and adjust schedules or compute resources to meet performance criteria.
- Self-secure: automatically detect data and mask it before it leaves a secure region.

Snowflake [93] **(DBMS)**

Snowflake is a cloud-based data platform based on data warehouse automation provided as Software-as-a-Service. Snowflake differs itself from traditional data warehouse solutions or big data platforms by a unique architecture and service execution designed for the cloud. Traditional solutions either have a shared disk architecture or a shared nothing architecture. Within a shared disk architecture, multiple servers access the same database. The scalability of this architecture is limited to the database and network performance. Shared nothing architectures emerged as a solution to this bottleneck, where each server has its own data storage. By moving the data close to the processing nodes, bandwidth and network latency problems are solved. This requires a balance between storage and processing capacity, often resulting in an underutilization. Data also needs to be shuffled between nodes which adds overhead. Snowflake presents a multi-cluster shared data architecture, with a loosely coupled storage, computing and service layer. Data storage resides on a simple cloud storage provider. The computing layer consists of elastic processing clusters. The service layer manages the clusters, queries, transactions and metadata. This architecture allows Snowflake to be more flexible in scaling and pooling computing and storage resources and allows customers to pay per use instead of per resource. Resources are automatically scaled to match the load without manual intervention.

A data warehouse stores structured data in a relational database that can be used for reporting and data analysis, often using SQL. Traditional data warehouses are not optimized for semi-structured data as it does not adhere to a fixed schema. To process semi-structured data, big data solutions like Hadoop are used. Snowflake supports both structured and semi-structured data. The processing and storage of semi-structured data is augmented. When semi-structured data is loaded, it automatically discovers the attributes and structure and optimizes the storage. Repeated attributes and structure characteristics are stored separately for better compression and fast access. Statistics about these attributes are stored in the metadata repository for query optimization.

As Snowflake provides a service, database management is completely outsourced. Snowflake leverages the cloud to provide scalable storage and computing capacity as well as an optimized execution of both. While Snowflake is not transparent in the underlying technologies, it can be presumed that AI is leveraged in scaling resources and optimizing queries.

Alteryx [94] **(DI)**

Alteryx is a software tool that aims to make advanced analytics accessible to data analysts by combining data preparation, data integration and analytics into one no-code platform. Alteryx augments time-consuming and manual data management and analytics activities using drag and

drop tools. The platform offers seven products: Analytics Hub, Designer, Server, Connect, Promote, Intelligence Suite and Datasets.

The Analytics Hub has data catalog capabilities that allow users to discover available data within the organization. The Designer tool allows to create ETL & ELT data pipelines to profile, prepare and integrate data sources. The Server tool can be used to manage and schedule workflows. Built-in automatic recovery and fault-tolerant capabilities ensure no downtime. Connect is a data discovery tool. It functions as an metadata repository to catalog all data assets. These data assets can be linked with business terms from the business glossary, which enables data discovery. Promote allows data scientists to build, manage and deploy predictive models. Intelligence Suite is a augmented analytics tools, which enables users to generate insights from semi-structured and unstructured data with code-free machine learning. The Datasets tool partners with data brokers to enrich the user's data with location and business insights.

Oracle Autonomous Database [95] **(DBMS)**

Oracle Autonomous Database combines the flexibility of cloud with the power of machine learning to deliver data management as a service. The goal is to minimize manual intervention and human errors within database management and ensure data safety and optimal performance. This allows IT staff to focus on higher value activities, while saving costs on repetitive and time-consuming tasks. Autonomous databases achieve this within 3 primary categories: self-driving, self-securing and self-repairing.

- Self-driving: The Autonomous Database automates database and infrastructure provisioning, management, monitoring, backup, recovery and tuning.

- Self-securing: The Autonomous Database is more secure than a manually operated database because it automatically protects itself from internal and external vulnerabilities and attacks. The Oracle Cloud provides continuous threat detection, while the Autonomous Database automatically applies all security patches online. This preventative approach is critical because 85% of security breaches today occur after a Common Vulnerability and Exposure alert has been issued.

- Self-repairing: The Autonomous Database provides preventative protection against all unplanned and planned downtime – and rapid, automatic recovery from outages without downtime. The Autonomous Health Framework leverages AI by integrating multiple areas of diagnostics and enabling analysis and action to be taken at runtime to minimize operational disruption.

E. Transcript of Expert Evaluation

Understan-dability	**1: [redacted] 12-08-2020 16:30** **Maturity levels** 1: What I'm realizing, is that you're going to interview people that probably are familiar with the horizontal axis, the data management maturity, but not with the horizontal axis, AI. What you could do is to include a bit more on the applications of AI. D: What I was thinking is that interviewees don't necessarily need to know everything on AI, but they need to know that part of the process is done by AI, so it can be automated or support manual tasks.

	1: You can always say that, I would do that for sure. If someone wants to know more about AI, you can include something. But I don't know if people need that. What you just said, that something that you can always include. **Data Quality** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved. 1: The overall maturity score, do you calculate that from the processes or sub-capabilities? D: The processes.
Scoring schema	1: With that, you claim that sub-capabilities with more processes are more important than capabilities with less processes. I would say it should be weighted in the overall maturity. The sub capability should be equally important.
Relevance/ compreh.	1: I have already seen all the processes from the sub capabilities and provided feedback earlier on. I see that you've incorporated this feedback so I don't need to go through them. They all look all right.
	Results 1: Some capabilities and sub capabilities have a lot of processes, other have less. Do you incorporate that in some way? D: No, but in the results tab, you can see every individual sub-capability. The idea is to look at those sub-capabilities and identify improvements on sub-capability level. Because you look at this level, I don't think that you need to take that into account. 1: This looks great. Maybe you can make the line [for the spider diagram] more clear? D: When you fill in all processes, it turns into a spider diagram, so that will increase the visibility. 1: It's all clear, it looks really good.
Level sufficiency	**Evaluation** 1 [Question 1]: No, I think you should keep it at five. You build upon DAMA DMBOK and other models, and they have five levels. If you add another scale on augmentation, you shouldn't deviate and just use five. Other than that, I think those are enough to score the current situation.
Level sufficiency	C [Question 2]: [In the previous version] you had 'strategy' in the description and also manual between brackets, I thought that strategy was already quite sophisticated and that you should remove the brackets around manual. Other than that I wouldn't change anything.
Processes compreh.	1 [Question 3]: No, to me this sounds like it covers everything. But I'm not involved with AI on a daily base, to I can imagine that due to this lack of knowledge I could miss something.
Process Relevance	1 [Question 4]: No, I think everything is relevant for the capability that it is in.

	1 [Question 5]: No, the same case as in the previous question. I've looked at it before, so I would have suggested changes then.
Scoring schema	1 [Question 6]: No I wouldn't. The only thing I would do is to change the way the averages are calculated, like we talked about before.
	1 [Question 7]: No
	1 [Question 8]: No
Usefulness, practicality	1 [Question 9]: What I can imagine is that an example could be an addition. But I'm not sure how and when. This is about useful.. I don't know. Maybe that if you have clients that are clearly looking for an improvement roadmap, how to get from a to b. You have the descriptions of these levels that say something about those levels and what you need to do. But I can imagine as a organization you still have some questions. Good to have this insight, but what now.
	D: How would you do that? How to kickstart the roadmap?
Roadmap	1: That's also step two. This is step one, to get insights. If you look at it from a Deloitte perspective. First you make an offer to do an assessment to see where they are at. The second step is to get from level two to four in terms of time, resources, specific activities, priorities etc. I don't think that its necessary to include it. But it could make it more attractive to say something about it, like a teaser. I don't know how. I can imagine that it improves the usefulness, but at the same time I'm not sure on how to achieve that. However, I think that is outside of the scope of a maturity assessment.
Roadmap	1: What you could do. You describe the levels, but you don't describe what is needed to cross from one level to another. What you could do is to draw an arrow from one level to another, with under it some generic activities that you need to do to go to the next level. I think that you can describe that in a generic sense.
	1 [Question 10]: What we just talked about contributes to both the usefulness and practicality. Also, the less generic you are and the more specific in describing the actions needed to go to the next level, the more practical you make it.
Roadmap	1: Let me show you something. This is another assessment, where you have to choose between four answers. This results in a level, one to five. If a specific answer is given, there is a specific result that elaborates on the consequence for this level.
	D: Thank you, good to see this example.
	1: In general it looks good, clear, nice Excel, structured. The graph and spider graph are good applications.
	2: [redacted] 27-08-2020 15:30-16:30 **Introduction, maturity levels:**
Understan.	2: The definition of ADM was not that clear that I would get it immediately. A concrete use-case would have helped to understand 'human centered application of AI.

Process Compr.	2: I'm missing data governance: sponsorship, vision, plans that an organization has with the application of AI. I don't think that you can measure that along the process levels, but it does give an idea about the organization and how they think about how AI can help them. If we do a data management project at an organization, we often start with data governance and data quality. So if you want to do something with augmented data management, I would expect those two. I do get that it doesn't fit in the model as it is currently, where the focus is on automation. I would still recommend to include data governance as an important point. **DQ:** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved.
Process ME	2: I think there's an overlap at assess data quality and data profiling, why is it separate? D: Some models mentioned it as a separate sub capability, it could also be under assess data quality.
Understan.	2: For asses, you're going to look at your dataset and assign a score. I miss a process before where you set data quality requirements that you use to assign a score. If that is what you define as data quality dimensions, it could make sense to put that before assigning a score. On the other side, it can make sense to assess first to then extract your data quality requirements. D: I do mean requirements with dimensions.
Understan.	2: Then it could make it more clear if use dimensions/requirements. **Metadata:** D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved.
Compr.	2: My focus is on data governance and data quality, the other capabilities are a bit more outside. Based on DAMA DMBOK I see the usual key words and didn't see anything unusual.
Process Accuracy	2: One thing that stood out it that you use the word 'reverse-engineer'. At DQ you mention specify, here reverse-engineer, that already sounds like something automated. D: Do you think it is good to use the same word? 2: At one you mention both specify and reverse-engineer, at another you only mention reverse-engineer, I don't get why. For consistency, I would use the same word choice. 2: You make the distinction between structured and unstructured data. Don't you think that would make it more complex? Is there a difference in approach? D: From an AI perspective yes, as for unstructured data you first have to recognize the structure before you can extract certain data fields. That would make it more complex to augment it in my opinion. 2: Yes I agree.

Understan.	**Data Integration** 2: Seems logical. D: here's also a reverse-engineer? 2: Yes, but data lineage is more complex than specifying data quality attributes. I think that data lineage always is tool-based and manual data lineage is virtually impossible. **Master Data Management** 2: I was wondering whether you had to configure an MDM hub as per data model and I was also wondering whether the meaning is clear to everyone.
Process Compreh.	**Database management** 2: Performance I get. But, shouldn't you also mention something about creating and configuring databases. What I mean by that is that depending on your use-case there are different ways to configure you database. In DAMA DMBOK this is under 'data storage', manage database technology. I miss the step where you determine the ideal database configuration. D: In my understanding, managing database technology covered mostly hardware. 2: If I read it, I understand that you think that. The definition that they give is 'the design, implementation and support of stored data. I still mis some design database and implement database, manage performance is aftercare.
Process Compreh.	2: For example, there are multiple configuration options that each have their own performance and price tag. You only want to use the fast technology for critical business processes, but not for a department that is experimenting. I think that there's a gap there. For that knowledge you should focus on the more technical people, maybe Oracle team or [NAME].
Scoring schema	**Results** 2: I did maturity assessments in the past and I recognize this as a possible visualization option. **Evaluation** 2: At a previous maturity assessment we specified all the processes at all levels. The model wasn't that good because it was very specific. You model is quite generic which would make it harder to fill. One might have a few aspects in level 4 but also a few in 3. With the specific levels we also had this issue. What I learned and proved helpful at the [NAME] and [NAME] assessment is that you specify level 1 and level 5. D: Did you specify if on process, sub-capability or capability level? 2: On process level [Note: after reviewing the documents mentioned, the description was on capability level but mentions multiple processes of that capability]. 2: Level 1 and 5 were clear, but the discussion was around the levels 2, 3 and 4. Instead of that discussion, it is more valuable to use set a target and use that to have a conversation.

Understan.	2: It's confusing that level 1 is shared across the two scales. Level 2-5 match for augmentation, level 1 not. 2: Using two scales might make it more complex. However, I realize that combing these two would not work. 2: Maturity levels: what stood out is 4 and 5 'critical business decisions' and 'more complex decisions'. Which one is the superlative? I get the 'business to trust'. There seems like an overlap, is complex more than critical? D: The idea is that at level 4 the business trusts AI recommendations to make decisions. 2: It seems like AI is already making critical business decisions at level 4, instead of AI recommendations being used. 'uses input from AI', 'AI recommendations'.
Level accuracy	**Open questions** 2 [Question 7]: Did you think about data warehousing and BI? D: Yes, but decided to leave out of scope as it is such a large field.
Usefulness	2 [Question 10]: I think I mention a lot of things that cover that already. Certain choices have been made and everyone has a different opinion. I think that the model is useful and yes, it can be improved in practice. What you'll see it that you will tweak the model for the customer because certain cases are more specific. I think the model can be applied just fine.
	3: [redacted] 14-08-2020 13:38 **Introduction** 3: What is the scope of this tool? Does it just focus on the assessment or also the improvements?
Roadmap	D: Just the assessment, but the model is intended to also be used for constructing an improvement roadmap. 3: You could be more specific about that.
Level Accuracy	**Maturity Levels** 3: When you're talking about a mature state of AI assessment you say that people trust AI to make critical business decisions. In a real case it would be rather hard to do so, particularly given the maturity of AI at the moment. (…) D: Increased trust means that increasingly more complex tasks are performed autonomously. 3: That means that they probably implemented AI in a better way, therefore they would trust the results and the model itself can provide better insights. I guess the only concern I have, because the text was intended to be short, it might be a bit ambiguous when I look at it. Because when you say that trust of AI already exists similarly might not become what we see in a real case. Maybe some challenges would

	apply there. Probably changing the framing over the increase on AI maturity level, gaining confidence from stakeholders would help some. 3: Which is good I think, is having the 0-5 standards from the standards like DAMA DMOK. It would give people a consistent feeling for your model.
Level sufficiency	**DQ** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved. 3: I think the [scoring] is clear and quite familiar in terms of how maturity models work.
Scoring schema	3: For project wise, we haven't done much AI. We do project automation, but not AI. Reading the first process (…) that is something where you don't really need a machine to learn. I struggle a little bit now, thinking beyond process automation, what role AI plays within 0-5 maturity level. What if a company does not have AI in place, while they have automation itself which allows them to generate a DQ score. How much would you rate them?
Level Sufficiency	D: That's a good one. My model covers more the front end of processes. It's more on what is done than how it's being done. In my model it would be possible to augment processes without AI. 3: One thing that is good that you have the intended score there. Some companies that find AI very valuable will not necessarily put AI in processes. Particularly when it comes to setting up a pipeline, making sure that the data is in place, because lot of them are very much business rule specific and requires business rules, require compliance and stuff. So letting ML is an effort that they don't want to pay. Having that intended score in place would help them to get clear where they focus on and then understand the variety of what the company actually needs and how the model could cater. (…)
Scoring schema	3: Normally when you see a maturity model you would see like 20 rules (processes). When I was thinking about my previous questions, I realized that the applicability of AI is small in that case and I think that level of detail is good enough. **Metadata** D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved. 3: Not my sweet spot
Process ME	**Data Integration** 3: I think it looks good in general, like the process are well-thought. I'm not sure if validate metadata is key for metadata or for data quality in general. It seems more like a DQ perspective rather than a DI perspective, obviously they are interrelated for sure. In general for me, when talking about DI you would not put that specific element in here. Then it could be a bit repetitive. 3: When you're talking about performing impact analysis and root-cause analysis, this is more on the data lineage side. And I'm not 100% sure if it's what I understand of DI. It's just a personal preference, where do I define scope.

	D: Where would you put data lineage? 3: For me it's a separate issue. For me DI, its more on the process where you want to make sure that data is integrated well. Data lineage is more on process operational wise, when things go wrong you want to track where the problem is. And this would have more implications on other processes as well. So there's a difference in where people consider this business critical.
Process ME	D: Do you think it is good to mention which processes overlap? 3: I think so. Also it is unavoidable that topics are intertwined.
Process accuracy	**MDM** 3: When you're talking about data models, that's also something you mention in metadata. From what I'm reading, the rest is good. **DBMS** 3: When I'm reading the first statement: Is database management closely related to infrastructure management? D: These processes are most likely to be performed in the cloud. 3: It is. If you're talking about cloud and about scheduling jobs, that becomes a process of doing DI. I'll give you an example based on our current project. We do resource allocation and scale up on demand. Those are in each steps of DM or AI job from end-to-end, whenever they are jobs related. So, specifically mentioning it here without mentioning it otherwise was something I was confused upon and looks very cloud specific, resource specific, infrastructure specific to me. So talking about AI, this would become other relevant topics rather than data management. Other than talking about databases, we're also talking about infra, security, networking, compliance, but that's not the key mention of you model. Putting this in database management looks relevant but not like the key thing.
Process relevance	D: Do you also think that of other statements? 3: No not really. If you're talking about queries, if things happen in SQL, that is database related. Jobs are more applicable in every single level, where queries are more happening on the database level. 3: Provisioning and managing are also very much infrastructure related. If you want to do the turnover or recover of your database you probably just need to set up your recover plan on your own system itself, rather than you configure whatever database that you are using. So it's more cloud specific than database related.
Process relevance	3: Recover and tune database; the general logic behind it is that companies need to allocate resources in different locations to make sure people can access it in real time or any time they prefer. If that is the case it is on networking and resource allocation rather than database allocation. So in that, it is rather 'other' than database.
Process relevance	

Process accuracy	3: Database logs: fault recovery, threat detection, manual intervention can be applicable in two ways. This is also applicable with DI. When you want to move one database to another when there is a fail over or if there's an error or bug. The level of problem would not necessarily happen on the database itself or how you configure SQL queries, but more on how do you do the integration from one data center/warehouse to another. This has overlap with data integration, maybe you can mention this as well.
Process comperh.	3: When you are mentioning data models, this is also relevant for database management as well. Because usually you design your model with SQL with your database. That model itself can happen on the database. The physical level of a database also requires you to have a logical representation of the data model before you implement it in the database itself. So that is also important to have. D: Something like recognizing the data model and automatically store it efficiently? 3: Yes 3: There are also some other relevant things in terms of the quality that you can also mention here as an overlap. 3: Why not data reporting, data dashboarding? D: That was out of scope, as it is such a large field.
	[redacted] **31-08-2020 10:00-11:00**
Understan.	4: The Introduction tab looks good, looks tight. 4: [the maturity levels] For me it was clearly different for me, as I don't have experience with ADM. When I go through the tabs afterwards, I had to search what exactly covers ADM and the augmentation maturity levels. Does ML come into play? I see automation. For me, more explanation around the concept and which parameters are associated would make it more clear. I don't know whether the consultant is supposed to have this knowledge, and the model is purely to do an assessment at the customer, or whether there needs to be more clarity on when which processes have a maturity level.
Understan.	4: If it think of automation, I don't think of AI. When I think of AI, I think about decisions. In which AI makes recommendations or makes the decision. The automation of processes is not directly AI for me. Other than that, the levels are clear and the description is also clear. D: I agree it would be good to clarify the definition.
Format	**DQ** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved. 4: The first thing that I see is that the quote is folded.

Understan.	4: What might have an added value is to add a column that shows an example. I mean one example for the maturity levels. That might give more clarity and context.
Understan.	4: I think that most [processes] are clear. I was wondering what you mean with reverse-engineering? D: That's when you're looking at the data to extract rules. For example, if you see that in 90% of the cases a value is not null, you can suggest that as a data rule.
Compreh. Process ME	4: I think it covers most of it. I was wondering whether there is a logical order for the capabilities and processes. For me it would make sense to distinguish capabilities that are needed to the development teams, and capabilities that have more to do with operations. Monitoring for example, it typical for operations, I think you can make a distinction there. If you talk about validations etc. that's also testing, testing validation if data matches. That could also be combined.
Format	4: For some processes you use points, and others you do not. Please be consequent. I've tried the scoring schema and that works all right
	MD D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved.
Understan.	4: What is the difference between the fourth and the first and the third? D: The first is about creating a general metamodel for a whole dataset. The third is about proposing a certain metamodel architecture based on a taxonomy. The fourth is about creating the actual metadata, so more on an individual base.
Process Relevance	4: Why data lineage here and not in integration? D: This is about the creation of metadata to be able to track data lineage. 4: I get it, we could check if it might be better suited at DI. I do get that this is specific for metadata, where DL is the result of creating this metadata. D: Do you think that indicating overlap would make it more clear?
Process ME	4: At this point you didn't talked about DI yet. So I think it is best that the distinction is clear enough so you don't need that. I think it would cause more confusion. Of course there's overlap, I don't think you need to indicate that. 4: I wrote down validating metadata, which I see over here. 4: Is there data mapping here? D: No, that is also in data integration.

Understan.	4: The added value on row 3 [capability definition], I think that that is important to have. You could also include it in the instructions, but I understand that you might want to keep your template and that it's a lot of extra information. But it's good to have a clear definition.
	DI
Process Relevance	4: About the difference between metadata and DI metadata validation. At metadata you're creating an overview. You're generating a report, monitoring the logs and identify where it goes wrong. For me that fits better at DI, monitoring the end-to-end process. Normally, if I look at my current projects, there are the integration logs. When you're validating where the potential mismatches are, than that would be more metadata related as it is more in the design. When you're monitoring, it's a process that happens afterwards where you check whether the integration is correct.
Process ME	4: Perform impact analysis and root-cause analysis, can that be interpreted differently? Is that about end-to-end data lineage or certain parts. D: Impact analysis is more about discovering dependencies. Root cause is when there is a system failure. 4: than it is clear.
	MDM
Format	4: This one is clear for me. One thing is that you shorten MDM, please be consequent. Use a abbreviation or don't, but introduce it. Also you use the & mark, is that necessary or can you just write it? 4: Processes look good
	DBMS
Process compreh.	4: This is more operations, which stood out. Not that it matters. It's only one sub-capability? I don't know if you missed one? D: What would you miss? 4: No specifics. Here you're looking at the performance of the DB, but not the development. There are something that you can take into account when developing, for example build to scale. Don't you miss a piece setting up databases? And which processes are involved into that?
Format	4: Monitoring DB logs is also here. At DQ you call it monitoring, at metadata you name it analyze. I don't know if you want to mention this as a separate capability? Like you do at DQ. The creation of logs is done to monitor them. So, maybe be more consequent in the definition? I do get that you want to follow the terms from literature. 4: Capabilities that I'm missing it data security and privacy, for example the usage of encryption and pseudofiction, including privacy in data is relevant. I'm not sure whether AI can play a role there. D: At the start of my research I scoped it to these five capabilities, where I think there is the largest potential for AI. There are other areas where AI can play a role, but it is outside of the scope of my research.

Process compreh.	4: I get it, but I'll say them anyway. Data governance, data sharing, data analytics. It's just some suggestions. I know for example that security is a thing when designing data and metadata or data integration. That it can be a sub capability just like monitoring. D: How would you put data governance in this model? 4: The first question is there is the ownership of the data. But also: how do you make sure that standardization of the data is widely used and who's responsible for that. It's more about control, but it could also be about importance. Where is it most important to adhere to the standard, where do we need extra checks? Maybe AI can play a role here, where these checks should be, which would be a validation factor for DQ. D: I see that it is important, but where do you see the role for AI in data governance? 4: I see here that you trust AI to make critical business decisions, I don't know if you can apply that here. The way that's is described leaves the questio9: where's the line? Is it purely operational? What are the business decisions in the operational aspect? Eventually, these human choices is where AI can have the added value.
	Results
Format	4: The only thing that stood out was that the color difference between current and desired for ADM is not very large.
Usefulness	4: It's really nice and a good way to make an assessment like this. With such outcomes, where you have one overview that you can also share with the customer.
	5: [redacted] 02-09-2020 11:00
	Introduction 5: Evaluation is something you do not regularly do in an assessment? Only for us right? D: Yes, only for the exert evaluation.
	DQ D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved.
Process accuracy	5: Why do you mention manually? Then you're proposing a judgement? We're talking about the process itself, so it doesn't matter whether it is manual of automated. It's about when you're scoring data quality, one part is determining what you're measuring and the other is actually measuring it. Only when you're combining it with your scale, than it is interesting to look which score I would assign. D: It would be to illustrate how this would look like, but I do think that it's also present in the scale.
Understan.	5: Wouldn't you switch the first two? So that you first specify the requirements and then score the data using those requirements.

	D: Good one.
Understan.	5: Also ‘or reverse-engineer’, I think it is important to keep it simple and if you have an ‘or’ statement in your process, you don’t know what to score. If you keep it simple: determine what you want to measure and measure. Than I would just say specify (..) rules. In that you keep it simple and make it easier for the person who’s filling it in to do so.
Process accuracy	5: Do we need something subjective such as ‘large’ and ‘complex’. Isn’t it interesting if they apply profiling at all? I get why it’s there, if it’s simple data it’s of less added value. But here, simply profile data wouldn’t that be better?
Process accuracy	D: I think that’s where the potential value for AI is, in small datasets you might be better of by doing it by hand.
	5: All right fine.
Understan.	5: Basic statistics. I would put a mark here, if more people question what you mean with basic statistics. Eventually, the most important thing of such a tool is that there should be no discussion on the processes. The more discussion on the process, the less value the scoring has.
Understan.	5: What is standardization here? Maybe it is standardized cleansing?
Understan.	5: Would it be an option to avoid the ‘or’ statement?
	D: I could remove ‘or perform autonomously’
	5: They are two different thing, both interesting to have.
	D: Maybe ‘suggestions for’ between brackets?
Understan	5: Would be a good one. You could also do to perform and/or suggest cleansing. Or suggest and/or perform.
	5: Perform ongoing DQ checks. It’s always hard that there’s an overlap. My first question would be: how is this different than assessment. At once there’s a new construct: pipelines.
	D: The difference is that it’s on ongoing data. Once you have an initial score and cleansed your data, this is meant to monitor the changes.
Process compreh.	5: Isn’t it data pipelines and/or data mutations? So in case something changes, the checks are performed.
Understan.	5: It’s important that you use the same terms consistent, so people can clearly see the difference between sub processes.
	5: Is it by significant differences or in case of? I would say, detect anomalies in case of difference between actual data and historical data. Or does expected values add something?

	D: Expected value is different than historical data. The expected value might differ from historical data by multiple variables, but is calculated from historical data. 5: Maybe you can say anomalies i.e. significant differences? Yes, that would be good.
Understan.	5: Other than that, it looks complete. D: Would you keep data profiling as a separate sub capability or would you say it is part of assessing data quality? 5: I see DQ assessment as: identify what you are going to assess, than determine rules, than determine metrics, implement that, monitor it and fix is. You focus more in AI context, that doesn't have to be the same. I see profiling definitely as part of assessing. It would make sense to combine it. Than you clearly have all steps before you're actually cleaning the data. Profiling is a form of measuring. **MD** D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved. 5: Extract attributes? Why not generate metadata?
Understan.	D: That's better
Process compreh.	5: Definition wise, often if we talk about catalog, we're talking about the search function. Could you add e.g. data catalog? If we talk metadata, we often talk about three important products: business glossary, for data definitions from a business perspective, second is a data dictionary, more technical metadata like source, physical name, and the final one is the data catalog, which provides the search function. I would leave 'this', because those are separate processes.
Understan.	5: Merging business and technical data? Are we talking about metadata? It's like vertical data lineage, where you couple a business term to a technical term in a database. D: Yes, metadata
Understan.	5: You have horizontal and vertical data lineage. You could mention this. Horizontal data lineage demonstrates the path along which data flows from creation in the source system to the destination. Vertical data lineage shows a path by linking items in different data models e.g. logical, conceptual and physical. Generating metadata in your case is horizontal. Merging business and technical metadata looks like vertical data lineage. An expert on data lineage would really look for the difference, you could add it as e.g. You can check whether that makes it too complex.
Process accuracy / ME	5: I would change the second one in specify metadata rules. Is reverse-engineer important here? Otherwise it is similar to the discussion that we had on 'manual'. If you talk about specify, you would give the respondent the change to fill in themselves whether it was manual or automatically reverse-engineered. D: That a good one, would you keep 'based on datasets'?

	5: Yes, that helps.
Process compreh.	5: Rationalize data dictionaries. Here again, we have a data dictionary, business glossary, data catalog, the three core products of metadata management. Here you're talking about data dictionaries. If you say, data repositories, e.g. those three, than you cover both.
Process compreh.	5: Analyze metadata, that's a very interesting one. What I miss if I think about analyze metadata, is analyzing in a way that creates value. If you're talking about data and analytics, I also talk about metadata analytics. It is interesting to see which customer buys something and at what time. By using AI, you can generate insights from metadata. For example by analyzing the timestamp. Analyze metadata in order to create valuable insights.
Understan.	5: Convert technical session logs.. Is a bit long. I find technical session logs a concept that is too difficult to present without discussion. I think it is better to find something else. Maybe add convert? What you want to do is monitor data flows and transformations. I would make it Analyze data flows and transformation logs to check whether data is moved or mapped as expected.
Format	5: Small thing: sometime you use a dot, sometimes you don't, get that straight.
Format	5: For the last one, I would remove 'file' for consistency. What you're doing is analyzing metadata. I would make it Highlight potential mismatches and/or missing metadata and resolve accordingly.
Understan.	5: If we're looking at the order, I would say 3 first, than the second and then 1. Than you go from simple to more complex tasks.
	DI 5: What do you mean with data discovery?
	D: It's when you're designing a data integration. The AI can then recommend additional datasets. Like, when you're using customer data it can recommend to also use sales data. Or it can recommend a similar dataset with a higher quality score.
Process compreh.	5: Maybe we can make it recommend additional and/or alternative datasets. That it's clear that it is something else.
Process accuracy	5: At MD you mention generate data lineage, here you mention reverse-engineer. Do you need to make a difference? Metadata management vs. data integration? This could be a semantic discussion, but I think for you it is best to make it as simple and accurate as possible.
Process accuracy	5: We changed reverse-engineer twice already. Here again, it suggests that is happens automatically while you could also do it manually. Is it important to mention reverse-engineer?
	D: No, but I do think that this is the process that is least likely to be done manually.
	5: You're mostly scaping here. Everything has been done and you're using a metadata management tool to automatically extract this type of data.

	D: Maybe change into end-to-end in DI and create data lineage MD?
	5: Yes and is it necessary to mention from code and from metadata.
	D: Not necessary, but I think those are the two ways to can construct data lineage.
	5: Validate file metadata isn't that the same as in MD?
Process ME	D: I'm not sure why they are both in here.. Maybe they are both relevant or maybe it was by accident.
Process ME	5: I wouldn't include it in both. You already have quite some. For a thesis it might be, the more the better, but [less processes would make it more] practical. I wouldn't be surprised if you only had 5 or 6 processes per capability.
	5: Is the last one different than mapping source to target? You could leave it there any maybe change it a bit so it is clear what the difference is.
Process ME	D: This is more about monitoring mismatches on existing data integrations, the mapping happens when designing the integration.
	MDM 5: I'm missing a process there you search in which master data sources actually is master data. The sub-capability refers to it, but the process not really.
Process compreh.	D: I mean that you search for similar columns across different sources.
	5: That's perfect, but that's not how I read it. So generate data model equals scanning all sources and searching for customer? It would make it more clear if you would separate the two. We're talking about the same thing, but you see generate data models as similar as scanning sources, I think that is farfetched. I would rather have a process, I don't know how you want to word it, but scan everything and identify master data. [..] I would say: scan data sources and identify potential master data entities.
Process comreh.	K; Identify linkages.. The first one leads to a selection of results and at two you're combining thing. We have a list with five data sources, and two include client data.
	5: One thing is consistency, data fields, data entities, data columns, sometimes it's the same, so then try to use the same definition.
Format	D: What would you say is the best definition?
	5: I think entities is equal to fields and that columns is one level higher. So, if you're talking about one field or multiple.
Format	5: What is a data hub?
Undertan.	D: Something like informatica.

Understan.	5: Aah I get it, so configure your tool based on your data model. Maybe I'm not into tooling. Maybe include /tooling? 5: Identify duplicate data? Is that not also data quality or is it specific master data? D: It's relevant in both, but here you are identifying duplicates to determine what the master data is. 5: The final one is too long and has an OR. For me this is more master data quality instead of stewardship, which is more about roles. D: I used the same term as DAMA DMBOK, there its also combined. 5: Believe me, we both can find a lot of mistakes in DAMA DMBOK. I don't get that you want to change one sub-capability, then you can keep it. Else, I would define it more like MDM quality.
Understan. Format	5: Do you need establish a single point of truth? I think it's a bit farfetched. To generate recommendations is possible. I would also say identify duplicate master data. Also you're introducing two new concepts, clustering and blocking attributes. Possibly they have a value, if not, I would keep it simple and use the same wording as before. D: Clustering and blocking attributes are the data fields that you use to identify or exclude duplicates. For example, if you look at Deloitte employees, you want to exclude Deloitte as their workplace, while you want to look for overlap in last name. 5: Than it is fine, but an explanation is needed.
	DBMS 5: Quite technical, so hard for me to say something about it. Than it is important to have a respondent profile.
Levels accuracy	**Evaluation** 5: The maturity levels from 2,3,4,5 I get. But 0 and 1. Is the description about your processes? Or are you also drawing a conclusion, because the processes are manual, unpredictable and poorly controlled, you can't leverage AI at all. Than I would also add that the processes are not suitable to be augmented. Processes are not ready, e.g. unpredictable, poorly controlled. What I want to shield you from is value judging their processes, the value should be in combination with AI. So the processes are not good enough from AI. 5: The second level is quite some step. I think it is quite something if an organization sees the value of AI.
Maturity accuracy Level sufficiency	5: Do you need a level 0? If you see a difference between 0 and 1, go for it. But how I see it, it is either not applicable. Its easier to have a file point scale, now you have a 6 point scale where there is no description at 0. If you remove it, I don't you miss something. 5: [Q10]: Nice to have for the future: an online tool instead of Excel
	6: [redacted] 31-08-2020 15:00

Future work	6: You're talking about augmented data management, what do you mean with augmented data management?
Understan.	D: I have the definition here on the next page. D: I also use two axis, one is the CMMI/DAMA DMBOK axis which is quite standard. The second axis is my main contribution and describes the augmented maturity. I separated the two to make it more clear. D: Looks wise to me.
Usefulness	**DQ** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved. 6: Maybe it's a bit too early, but I think it is interesting to apply this at one of my customers, its relevant for them. We're doing an MDM program and they are looking at AI, RPA and test automation as topics. But the combination between AI and DM is not covered yet. So that would be interesting and this could be an eye opener for them to develop their IT strategy, of which data and operational efficiency are also part. 6: Assess large datasets to generate DQ. I would put large between brackets, it doesn't have to matter whether those are large or smaller datasets. D: That's where I think that the added value of AI is the largest.
Process accuracy	6: OK, so it wouldn't be possible to couple the two and use AI for smaller datasets? What I would do then is make it more specific, a certain range, because now you're generating more questions. 6: Data cleansing, in my experience, has the most to gain from AI. 6: I notice that the questions are formulated with an end-goal to AI. That is on purpose I assume and I also get it. I think that this is quite clear. Again, I would get a closer look at words like complex and large, to provide some more guidelines or would specify it more with a range for example.
Process accuracy	6: I don't see transformations here. Where data management is also important is for data migrations and then transformations are important, where AI could play an important role. D: Do you see data profiling as a separate sub capability or combined with assessing data quality? 6: No, I see it as separate. At assessing you're also talking about the definition of data quality and with data profiling you're looking to make a large dataset a bit smaller.
Process ME	**MD** D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved. 6: Looks fine, missing some data modelling.

	D: That is included at master data management.
	DI
	6: I'm thinking whether from code and from metadata are enough to map your data lineage. Often you see that many things are coupled with point-to-point integrations. Sometimes there's a data hub in-between. But if you look in the systems, like it says 'code' that might not be enough.
Process compreh.	D: What other method would you add?
	6: Look at integrations, so the interfaces between the systems, some sort of service bus or integration system.
	6: What is the difference between structured and unstructured?
Process compreh.	D: When applying AI it is a different method. For unstructured data you first have to extract the data from video or text files, this is an extra step.
Understan.	6: I'm looking at the last two, can you explain?
	D: The last one is about learning from human integrations to perform other integrations autonomously. The other one is about monitoring integrations and identifying mismatches of existing integrations.
Understan.	6: I'm assuming you're going to be present at the case study, but you can look on how you can make these two more clear without voiceover.
	D: How would you advise to make it more clear?
	6: The easiest is to present an example.
	MDM
	6: Generate master model… very valuable if you can do that.
	6: Clustering and blocking attributes?
	D: Those are the attributes that you use to combine results or exclude certain data fields.
Understan.	6: We did a deduplication project where 90% overlap would be identified as potential duplicate. That's a very valuable one, if you can use AI for that than you make a lot of people happy.
Usefulness	D: For configuring an MDM hub, I'm looking for a better explanation, how would you improve that?
	6: The first thing is what is an MDM hub? I define it as a separate system where you maintain your master data, whether it is a coordination point or single source of truth. As per data model? That means that you already have a data model for this hub adapts to the model.
	DBMS

	6: These are typical DBA processes. In a lot of cases you see that roles, except for the manager, dependent on the tools that the organization uses, will not deal with these processes. Often there's an DBA behind it. But those are valid processes, also for performance. Maybe you have heard from the Oracle Autonomous DB? You could use it as a reference.
Process relevance	6: Other than that I have nothing to add. Something that I miss is a piece of security. Maybe it is also under the third process.
	Evaluation 6: [Question 1]: I think these levels are fine. DAMA DMBOK also uses five levels right? Than I would also use that.
Level sufficiency Level accuracy	6: [Question 2]: The description for 1, there's not AI but you still call it experimental. For 2 and 3 the title might sound more fancy than the description. Awareness and making plans is a step before being ready if you ask me. Recommend course of action of that semi-automated? [What would be ready?] I would 'the organization is aware of the value of AI and has developed plans to augment.' In my feeling this would take it one step further. 6: Maybe good to check with the A&C colleagues of they have some definitions? 6 [Question 6]: No, it's to early to say something about this. You can evaluate this in case studies first. 6 [Question 7]: Maybe an added value would be if you could combine some topics in some way. So that you combine a topic from MD with MDM, DQ with DI, some overlapping topics.
Process ME	6: What could also be interesting is that if you have certain results, an improvement roadmap needs to be constructed. How would you approach that? What are the different aspects on how to do that. D: I'm looking into that.
Roadmap	6: What would make it interesting, is that if we would apply the model for a customer, it would result in a report. What you do then, that is more interesting, because it probably results in a project or program, where you are going to implement these things and provide recommendations. For now maybe not very important, but for the future it is. It is very nice that you not only present the results but can also elaborate on what you can do.
Roadmap	6: Let me know if you're ready. I think we can apply this at a few customers. Such a data management assessment, we do in 4-6 weeks. It includes a end report, recommendations and a roadmap on how to implement. This is the bridge with AI and that is interesting. We could include this as a component in such an assessment. So let me know when you're ready and we'll see how we can apply this.
Usefulness /practical	**7: [redacted]1-09-2020 14:00** 7: About the term AI, what is your definition of AI? When do you think a organization leverages AI? D: Mostly in terms of automation and manual work required for processes
Understan.	7: In step 4 I see automation and AI, it might be good to have a good distinction between the two.

Level accuracy	D: I agree, however, it is hard to distinguish on the front end. Still stings are done for you so that might also be a form of augmenting 7: For a lot of companies RPA is the level that they're at. **DQ** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved. 7: Is it DAMA DMBOK? Let's see. Every process needs to be scored on both scales? D: Yes, those two. 7: I think the processes and sub-capabilities are quite clear. I think that if you look at assessing DQ, it is partially automated based on statistics and data rules, but ML is not yet used in most organizations. 7: At DQ, there's always a large part on root-cause analysis when things go wrong. Metadata also plays a large role in that. I don't see that here. I think augmentation can play a role there, so that you can see there the data comes from. I know tools like Octopai and Informatica have tools that can map your whole data stream. D: Good that you mention that, these processes are included in metadata management and data integration. 7: I think that it is good that you have detection of anomalies, as that happens for example at banks. Machine learning is really applied there to flag strange numbers, so I see it happening that augmentation is used for that.
Relevance	**MD** D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved. 7: This is really extensive. What comes to mind is that within a lot of companies have a different definition for a data attribute, like a client, customer, counter party and all departments call it that but they have the same meaning and that should be mapped. I think I see that at rationalize data dictionaries and at merging business and technical metadata. Do you mean that as well?
Process compreh.	D: Yes. 7: It's good that you have lineage in there. 7: Aggregation of metadata is done in the catalog. Specifying metadata is done in glossaries.
Relevance	7: I see that at defining metadata architecture that you're directly talking about reverse-engineer and rationalization. If you talk about the capability, isn't it also about specifying itself? Does it have to be automated?

Process compreh. Process accuracy	D: I agree, I'm thinking of downplaying the processes, as automation is already in the maturity levels itself. 7: Metadata analytics, maybe I can add something here. Managing metadata is different than getting value from analytics. This can be discussed but during my thesis we found out that simple metadata management can bring a lot of value. Basically what they did is based on metadata analyze which datasets exists. They matched datasets and found out that two departments are buying the same datasets. By profiling metadata they could compare which datasets are the same or similar. So analyzing the metadata itself is a capability I would say. Something like profile dataset metadata and match similar sets.
Process compreh.	7: What I've also seen is that an organization made a lot of BI reports. Daily they would produce 10.000 reports to all sorts of mailboxes and users, while they had no idea who used these reports. That provides quite a strain on the infrastructure, as it needs computational power. These guys then applied metadata management. You have technical metadata, business metadata and operational metadata. Based on operational metadata they started looking if these reports are really being used and opened and they managed to remove 4000 reports. I would say something like 'analyze metadata to do whatever you want haha'.
Process compreh.	7: Technical session logs is quite specific. I would also add a generic process for analyzing metadata.
	DI 7: Data discovery also involves metadata I think.
Process accuracy	7: Good that you have Impact and root cause analysis. Also good that you separate them.
	7: I think you cover everything. The largest work is in design DI and map source to target model.
Process Relevance / ME	7: What do you mean with validate file metadata? D: Its about monitoring integrations and checking whether the metadata is created as expected. Those are constant checks on existing integrations.
Understan.	**MDM** 7: I see bottom-up and top-down, that is good.
Process relevance	7: What I would add is detecting the golden source, how do determine that? For a lot of organization that's quite something. Maybe it is part of detect master data model, but the identification is a process on its own.
	DBMS 7: The first one is very important. If augmentation could help that would be good.
	7: You could also look at redundancy. AI is used there for sure.

Process comprh. Process compreh.	7: What I would also would be interesting to see whether augmentation can be used to determine storage types. You have all types of databases. I know that CSP's already provide such services for customers. If you have data that is used often, you want it on servers that have a faster response time than back-up data that is never used. Cost optimization of data storage. **Results** 7: The best would be that you can have an advise based on the results, but that would differ per customer.
Roadmap	7 [Q1]: I believe 5 levels is sufficient for a practical assessment. 7 [Q4]: In data management, nearly everything is interrelated. Making everything completely MECE would be close to impossible 7: [Q6]: No, but please understand that in practice, the scoring would be open for interpretation and discussion. If you let 10 employees score their company, you will probably get 10 unique results. 7 [Q7]: The guidelines are clear to me. 7 [9]: Some processes are very in-depth (e.g. Rationalize data dictionaries through industry taxonomies, folksonomies and client specific taxonomies). Such processes could be described on a higher level in my opinion. For example: The creation of data dictionaries using taxonomies... etc. Rationalization would then be a higher level of maturity. 7 [10]: Practical examples always help clients, also in their thinking about the different levels of maturity. However, this is something that grows as you use the model. I think these are not required for this stage of scientific work.
Understan.	**8: [redacted] 03-09-2020 09:30** 8: What improves about a data management capability when you apply augmentation? The question that I had when reading through the model is, there's a lot of processes, manual processes being automated. But I don't see the data management capabilities that you mention in the definition in the scoring model. The maturity levels themselves mention processes in general, not data management processes.
Level sufficiency	D: True, those are generic levels that need to be related to the data management specific processes. As you have to relate them, it becomes clear that those are data management processes. 8: Did you think about making it specific data management levels? D: Yes but didn't choose to do so. 8: You can think of naming it data management processes, than it's still generic but scoped. If you mention processes, I can think of business processes and very specific data management processes and if I look in the capabilities I think that it is the latter. D: That would be an option. But that would make every description longer.

	8: It becomes more clear when you get to the processes, but if you go through the first two tabs it's still not clear. It's something minor, I do get the setup and that looks fine. D: Do you think there's enough distinction between levels or is there some overlap?
Level accuracy	8: No I think it is clear.
Process ME	**DQ** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved.
Process Relevance	8: I think that those items make sense. They are specific, which I like as I can relate it to an organization or a person that works there to determine whether they do it or not. I see some assumptions that organizations have some data rules as thresholds for the data quality. So you can see that there is a certain data management maturity required to answer these questions correctly. Some are quite detailed, for example data pipelines. I don't think there are many organizations that get this concept, let alone monitor the flow of data in a pipeline. For that you have the different levels, so that's fine. D: Do you think data profiling should be part of assessing data quality or separate? 8: I think it is different
Process ME	**MD** D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved. 8: Everything looks familiar. 8: Data lineage is a good one for that category. The tracking itself of data flows and transformations is more analyze metadata in my opinion. Especially related to the first item. 8: I would rename the generate end-to-end as it is confusing in regard to data flows and transformations. With the voice-over it is clear that it is about creating metadata that helps create the lineage later on. Other than that this looks good.
Process relevance	**DI** 8: Impact and root cause how does it relate data lineage? For me it is more related to the analysis of data lineage. Ensure data lineage is more about making sure that you can track the data, this is more in the case something goes wrong and you want to make an analysis. I would rename the sub capability to reflect analyzing data lineage. 8: Data discovery is a good one. 8: Did you think about to include data modelling?
Process compreh.	D: Yes, I think about making a universal capabilities, such as modeling.

Process Relevance	8: Exactly. 8: Design for me is more about modelling. Develop is also about the integration where you ingest and validate the data based on the structure that you defined.
Process compreh.	**MDM** 8: Also here you see data modelling return. 8: Why did you choose for stewardship in this sub-capability? I see the maintenance process, but not stewardship. This is more confusing as it is more related to the governance perspective. Maintenance processes correspond to the processes that are there. D: Yes it is only maintenance, but I used the same definition from DAMA DMBOK.
Process compreh.	**DBMS** 8: The first and the third is more about monitoring. That depends on how you define it, but in the other capabilities you had a separate monitoring sub capability. I don't know whether you want to split it, but you could look at it. 8: There's a lot in the items. I don't think that is bad, because those are quite standard processes to manage the data within the organization.
Process comreh.	D: Would you add any processes? P Not directly. If you would add storage cost optimization, this could also be part of monitoring as well.
Process compreh.	**Results/eval** 8: What you could think about is maybe add a bit on the added value of augmented data management in the background. Which type of business problems or use cases would AI help. You might not have the full space, but it would be good to add 2 or 3 sentences. It's also dependent whether you get additional documentation or just this document.
Understan.	8: The [documentation] is understandable and easy to use. It gives enough indication on how to interpret and use the model.
Ease of use Usefulness	8: For a self-assessment you might miss context. Some processes are quite detailed and might be harder to recognize by the employee of an organization, but it is definitely very useful. I think that the voice over and the joining consultant would really deliver value. I think the rise of AI is growing and how you can use this to improve your data management, is something that is increasingly being asked within the industry and by organizations in the future. D: Did you encounter organizations that advance in that state? 8: Yes, in monitoring of data. Not necessary with AI, but with different tools that monitor the data that flows through their systems and processes to define actions. They are working to create their own datasets and apply AI, but that is more in the future.

	8: Maybe start every process with a verb? **9: [redacted] 03-09-2020 15:00** **DQ**
Format	D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved.
	9: In order to generate this quality score, do we have a process assumed where we define the metrics or KPI's or business rules that are needed to generate the score? D: Yes, that's the second process. I will switch the two.
Understan.	9: Maybe add point 2 to include some business outcomes or business rules in it. D: Do you mean to add some process that tracks the influence of data quality on business outcomes?
Process compreh.	9: Yes, correctly. 9: Data profiling.. you got pretty much everything there.
Process Compreh.	9: Even for data cleansing, if you're talking about data cleansing rules, you could differentiate whether that is applicable to a local dataset or global dataset. You'll find a couple of tools in the market where in terms of data cleansing, it can start of as something identified in a business unit and that can be scaled up for other global purpose as well. There's aspect in scalability in term of AI in order to scale those rules and apply to other datasets.
Process compreh.	9: In the current state, a lot of this is currently done manually. 9: Are those checks just done on pipelines or also on data lakes and actual MDM kind of tools? D: I mean for all mutations. 9: This one can be not only data pipelines, but also data that is stored as well. Usually what happens is that you have a MDM tool implemented, with data cataloging as well as a lot of data integration from different source systems to target systems. You can monitor the quality of data in the data pipelines but you also need to monitor you master data which is centrally stores as well when any end user is modifying it.
Process compreh.	D: Do you see data profiling separate from assessing data quality? 9: I think it should definitely be a separate one. **MD**
Process ME	D: Again, the same questions for this capability, do you miss a process, think one is redundant or the description should be improved.

	9: I'd also call out business glossaries as well. D: I was thinking about extending the definition. 9: Correct. And it's definitely popular in the market right now, a lot of companies are doing data cataloging now.
Process compreh.	9: What is folksonomies?
	D: Terms that are known to the general public.
Understan.	9: One more thing you can add: when we talk about metadata, we talk a lot about definition, scope and boundaries. (..) when we are defining the architecture along with the dictionary we're going beyond the definition and descriptions. We're trying to identify who are the owners, what's the extend of each definition, where does the ownership end and begin for the next business or function. Scope boundaries and ownership are three aspects for moving beyond that definitions from a catalog perspective.
Process compreh.	**DI** 9: Is there something in here that handles trust from the sources or decay? Meaning, I you're getting customer data over a period of time from let's say two different CRM systems. But eventually, the capability has defined that CRM A has more trusted customer data from a downstream user perspective. Is there a capability or a process that defines that trust between different sources.
Process compreh.	9: If you're trying to incorporate that you can also think about data decay. It's a rule that it's going to be valid for the next four years, after that it should not be trusted compared to other sources. From an MDM perspective this is called decay. You can think of any process that can incorporate that kind of AI component to this particular decay issue.
Process compreh.	9: Maybe for impact and root cause analysis you can add some wording, some sort of description. I think that will help analyze the maturity of that process.
Understan.	9: Map source to target, you will see a lot of tools in the market that will do that to some extent. They pick up keywords and stuff like that, first name will be first name, so automated mapping between source and target.
	9: Do you want to include something on differentiating on real-time vs. batch?
	D: Maybe at database management.
Process compreh.	9: When you say develop DI, it's basically implementing the integration. When you do the development you also need to think about how this can be done and how AI can help on figuring out whether a real time integration is needed. What they figured out is that the actual frequency of updating this data was very low, so there was no point in updating this real time.
	D: Something like monitoring DI?

	9: Yes. **MDM**
Process compreh.	9: This is where the whole trust system, trust scores come in as well. Usually, from an MDM tool perspective there's this back-end configuration to basically data mine trust levels between the different sources of master data and also decay rules for master data. If you're trying to model the master data you can add this process or at assess data sources. While you assess data sources you can also assess from a trust and decay perspective.
Missing process	D: How do you normally assign a trust score?
	9: Usually based on business efforts. I used to work with business and functional stakeholders to ask them question on would you trust this data from system A more than system B. That was completely done manual, because we needed business inputs from it. But if you think from an AI or automation perspective, arguably there can be a capability which does a lot of back-end modelling or algorithms to figure out based on last 10,000 observations, that the phone number from system B was more trusted in the downstream analysis, so let's pick system B as the right source.
	9: What is Clustering and blocking?
	D: The matching and excluding variables to determine duplicates.
Understan.	9: One thing would be talking about match and merge rules. One of the core components is match and merge, so you can include a process which identifies duplicate data and how to do you identify, basically data mine, your match and merge rules.
	D: Maybe that is similar to resolution and reconciliation?
Process compreh.	9: It's more on the merge rules. You can both terms interchangeably.
Understan.	9: Point 10 is one of the key ones. Especially when you're talking about augmented data management. One of the prime activities that the data steward keeps on doing is resolving duplicate records. You can either split this process up in two or highlight that this is going to be one of the key ones from a stewardship perspective. One for learn rules and one for establish a single point of truth.
Process compreh. /ME	9: You need match and merge to deduplicate your data. At the same time you need to govern the data that is fed into the system by the users. So think of any create or update or delete or block kind of process for master data and whether there's some AI capability that can be used for the creation of master data or update of master data. (also requirements for attributes to create).That's definitely one aspect where you would get a lot of AI related capabilities for the setup and maintenance of master data.
Process compreh.	**DBMS** 9: I'm not too close to the technical side, but I think you have everything here. I don't see anything that I can add.
	Result/Eval 9: Result page looks good, captures everything. I've seen those formats before.

	9 [Question 1]: No I think this is good 9 [Question 2]: Same 9 [Question 6]: I like the spider diagram. Also in other models, this is often what we go for. **10: [redacted] 02-09-2020 14:00** 10: Do you use any reference framework to come to this model? D: I looked at many different data management maturity models and also AI maturity models. **DQ** D: I want to ask you to read all the processes and comment whether you miss one, think one is redundant or the description should be improved. 10: So what are the capabilities? Do you have that listed? D: Yes it is here [Introduction]. 10: So basically data governance is a part of DQ and Metadata management together? D: Data governance is not mentioned as a separate capability. Mainly because I looked at which processes that AI can assist in. I figured that data governance that is mostly human, so making the policy and assigning tasks. I do think it is important, but then it should be more in the recommendations. 10: I'm trying to think how this would work in practice. So basically you're doing an assessment of the current state and future state. How would you quantify the current state of maturity? D: You would look at the process statement and then relate those to the two maturity scale. So for example a process might be level 3, and has a desired level 4. Same holds for the augmentation maturity. This will later be the start for constructing a roadmap on how to get to a higher maturity. 10: I'm seeing the slide, it makes sense. It's very similar to the normal maturity levels that we use, the CMMI-like tooling. D: It's really meant as an addition.
Scoring schema	10: I'm looking into ways to make your models objective. The scoring right now is more subjective. D: Can you elaborate a bit more? 10: Basically, when you're rating the processes your lacking some rules. For example, 'a set of standard processes' what is standard? It's not defined. The problem with data quality is that.. You can have processes that are defined for the whole organization, but still the data is bad. Because it is produced in the wrong way in the first place. When you have rules, this is what I mean, I would be happy that the

Future work	statements are more granular, more objective. For example: central process, no peripheral process at all. That would be very objective to me. 10: I can imagine this is very difficult to make it more objective. What you've done is great work. I'm speaking from experience, data quality and data governance, making it more objective is quite difficult. D: It's a choice I made. Other models that I've seen had a more detailed description for every process for every maturity level. I think that that would make it more complex. I chose to make general descriptions that can be liked to the maturity levels. 10: There's no need to make it more complex, the model looks fine and acceptable. That's very important, the more complex it gets the less attractive and acceptable it is for the client. So this is directly usable and acceptable and it's a great model. Only thing is, go back and think if you can make those descriptions, I don't want you to make any changes to the model, think back if the descriptions can be more watertight. There shouldn't be any loopholes. Either its two or not two, either its three or not. If you can't do it, that's fine, but I'm being overcritical.
Usability /future work	D: more on a process level: do you think one is missing, one is redundant? 10: No it looks good. **MD** 10: What I think is data governance is an essential enabler of data quality. If there's a lack of data governance, eventually the data will be bad. You need to have an angle here somewhere, there has to be a process that you mention and assess the maturity upon, is what data governance processes are there that ensures that the organization adheres to the metadata.
Process compreh.	D: So, add a process on how much the organization follows metadata governance? 10: Or if you want to avoid mentioning data governance: the extent to which metadata driven modelling is followed. Or processes that make sure that data is created and maintained as following the metadata. Only having metadata is not enough. You have to have a process that ensures that the metadata is being followed or used. You also need to assess the maturity of that process as well. You can also tie it to
Process compreh.	percentages, if 10% adheres to the rules you're at level 1 for example. D: The way I constructed the model is to have processes where AI can play a role, this would be a separate process where AI is out of the picture. 10: Just as you have analyze metadata, you could have a process analyze the adherence to the metadata. Just think about it, you don't have to do it now. D: I do think it is interesting.
Process compreh.	10: Here, let me show you something. This is another model on data governance, which was the name for data management at that time. (…). In this maturity model we would compare the governance framework of the organization with five of six governance standard models that we had. Than we would see the deviation from the standard model and there we would assess the maturity. (…) Think in these

Future work	lines. If you can make it granular to this level, than the ambiguity in assessing any aspect of data management would become lesser. 10: When it comes to AI now. You can have a set of standard processes and you can measure the mean deviation or the percentage deviation from those standard processes in every organization. And based on the percentage value you can tie it to a number of maturity. D: Do you think it is still possible to have universal levels?
Future work	10: Yeah, why not? You just need a definition around the levels in terms of numbers. Your description is now in verbal English, we need some numbers or percentages around it. **MDM** 10: Let me give you an example. Generate a master data model by recognizing entities and hierarchical structure. Now you need to define the levels. This is an arbitrary example I'm giving you. Basically, MDM hubs have four different architectural styles: registry, consolidated, co-existence and transactional. Now, for example, if this data model does not comply to one of those four, then its certainly not industry standard. Those four styles have their own maturity level, if it's registry style it is one, etc. You can use these styles to generate a maturity score. This is a way you can make it more objective. If you don't have
Future work	the experience it's hard, but most people with experience can link such formats to a certain maturity level. Just think about it. It's not necessary, but it would make you model even better. D: I do think there's a lot to gain there and this is valuable feedback. But to do this for every process would really require a lot of work. One goal of the evaluation meetings as we're having right now is to look how water tight the processes and maturity levels are formulated, so that is something I can incorporate. **DBMS** 10: This is a different side of data management, which I haven't done in my life. In other organizations all the DBMS guys did this. It looks okay **11: [redacted] 10-08-2020 14:00** Note: No audio was recorded. The participant is asked to briefly comment in writing on the statements and evaluation questions. 11 [Q1]: No 11 [Q2]: No 11 [Q3]: No, I think the processes presented are exemplary for the capabilities. 11 [Q4]: Only if you learn during test runs that you're not keeping the interview under an hour but I think in our previous session it turned out that this was not the case. 11 [Q5]: As mentioned Master Data Management is a different slice of the data instead of a true different capability in my perspective, however this is logical in your model from a use case perspective (and also

Process accuracy	as you're using DMBOK as a basis). No change, just something to be aware of (and to put into the discussion of your final thesis). 11 [Q6]: Difficulty is that it always mentions 'processes' and not individual processes so the question arises if I should score an average for all processes or for that one process that is completely top-notch or the one process that is lagging behind. Not something you can easily resolve as it is inherent to the CMMI scoring model upon which you are building, option would be to score in percentages ('What percentage of processes are in the this maturity level?') but that would create a lot of additional work for which I currently don't know if you have time. I would recommend to put this in the discussion of your thesis.
Future work	11 [Q7]: If any I'd separate the tables on DM and ADM maturity for ease of reading. Or have a single white column between them to visually separate them.
Level accuracy	11 [Q9]: Maybe add an empty line for a process to be added by the assesor if they for example have an MD-process that they'd want to showcase in terms of current or desired maturity.
Process compreh	

Table E1: Transcripts of the Expert Evaluation

Levels sufficiency	Whether the levels are sufficient to represent all maturation stages of the domain
Levels accuracy (overlap):	Whether there is overlap between the maturity levels
Process Relevance	Whether processes are relevant to the domain
Process Comprehensiveness	All aspects covered, missing processes or (sub) capabilities, suggestions for improving definitions so they cover all aspects
Process ME mutual exclusion	Whether processes are distinct
Process accuracy	Whether process are formulated uniformly, they can be translated to every maturity level.
Understandability	Suggestions for improving definitions, clarifying, giving examples, logical order.
Ease of Use	Remarks regarding ease of use
Usefulness and Practicality	Remarks regarding usefulness and practicality
Scoring schema	Remarks about the scoring schema
Roadmap	Remarks on improving capabilities or actions following the assessment
Format:	Small things format related, for example cell size, punctuation, consistent word choice
Future work	Recommendations on future work

Table E2: Labels for Expert Evaluation

F. Changes After Expert Evaluation

Maturity Model (Figure 18):

1: Maturity level 0 changed from ‘incomplete’ to N/A to improve sufficiency. The ‘incomplete’ level had no description and did not add any value (interview 5). By renaming it, it is clear that organizations only score level 0 if the process is not applicable to them.

2: Each maturity scale has their own description of level 1 to improve understandability. In the previous version, there was a shared level 1 which caused confusion (interview 2,5).

3: Within the description for maturity level 3 ‘recommendations’ is added to improve sufficiency. The previous description implied that AI would make decisions, whereas in this stage AI provides recommendations to make decisions (interview 2).

4: Governance is added to the process description to improve comprehensiveness. Multiple participants indicated missing data governance within the model (interview 2,4,10). While data governance is not one of the selected capabilities it is still incorporated within the process maturity, which is now reflected in the description.

Data Quality (Table 25)

1: ‘/requirements’ is added to improve understandability, as some organizations use the term data quality requirements instead of data quality dimensions.

2: Switch order of second and first process to improve the understandability. It is logical to first specify the data quality dimensions before generating a data quality score.

3: ‘Reverse-engineer data quality rules’ is removed to improve process accuracy. Reverse-engineer implies that the process has a high augmented maturity, while the process description should also fit lower maturity level descriptions.

4: Put adjectives ‘large’ and ‘complex’ between brackets to improve process accuracy. AI is expected to specifically be of added value for complex and large datasets, but processes exist for smaller datasets as well. It also improves the understandability, as it avoids confusion on the definition of large and complex.

5: Changed from ‘or’ to ‘and/or’. The process covers the ability to learn from manual data cleansing, the and/or refers to two different outcomes on low and high maturity. The process is now universal for all maturity levels, which improves the accuracy.

6: ‘Data mutations’ is added to improve the process comprehensiveness. The underlying process was intended to continuously monitor new data and mutations to a dataset that is already cleansed.

7: ‘Significant differences’ is put between brackets. In this context ‘anomalies’ are ‘significant differences’, the process is reformulated to reflect this and improve the understandability.

Metadata (Table 26)

1: ‘Reverse-engineer’ is replaced by ‘specify’ to improve process accuracy. Reverse-engineer implies that the process has a high augmented maturity, while the process description should also fit lower maturity level descriptions

2: ‘(Meta)data’ is replaced by ‘metadata’ to improve process mutual exclusion. This process specifically covers metadata rules, the process now reflects this.

3: Added ‘data dictionaries, business glossary and data catalog’ to improve process comprehensiveness, as those are the three main types of metadata repositories.

4: Removed ‘folksonomies’ to improve understandability. Some experts did not recognize the term as being typical for metadata management.

5: From ‘extract attributes to generate metadata’ to ‘generate metadata’: Process is rephrased to be more in line with the previous process to improve understandability.

6: Added ‘meta’ to improve process mutual exclusion, as the process specifically covers metadata.

7: Changed ‘Generate end-to-end data lineage…’ to create data lineage metadata to improve process relevance. This process merely covers the creation of metadata to be able to generate data lineage. Data lineage itself is part of data integration.

8-9-10: Changed order of processes from simple to complex in order to improve understandability and ease of use.

8: Changed from ‘validate file metadata’ to improve mutual exclusion, as validation checks are also part of data quality. The process is now more tailored towards metadata management.

9: Changed ‘convert technical session logs’ to ‘analyze data flows and transformation logs’ to improve understandability and process comprehensiveness. Technical session logs was perceived as something technical and specific, the description is more general now.

10: Process added to improve process comprehensiveness. A process was missing that covers getting insight from metadata, other than the checks mentioned.

Data Integration (Table 27)

1: Added ‘additional and/or alternative’ to improve comprehensiveness and understandability. The process description now specifies the intention behind recommending datasets.

2: Changed the sub-capability from ‘ensure’ to ‘analyze’ data lineage to improve relevance. The processes describe the analysis of data lineage rather than ensuring.

3: Removed ‘reverse-engineer’ for process accuracy. Improve comprehensiveness by including integrations, as these can also be used to analyze data lineage.

4: Added ‘to identify system and data dependencies’ to improve mutual exclusion and understandability. There was some confusion about the difference between impact/risk analysis and root-cause analysis, which is reduced by clarifying the motivation behind impact/risk analysis.

5: Changed from ‘validate file metadata’ for mutual exclusion. Validating metadata had a lot of overlap with metadata processes. The emphasis is now more on the transformation of data for better understandability.

6: Changed from ‘Ingest, validate and transform data’ to improve relevance and mutual exclusion. Validating data is associated with data quality, while the focus of this process is on learning from

existing integrations and user interaction. Therefore, the process is simplified to ‘develop data flows’.

Master Data Management (Table 28)

1: Added process for comprehensiveness. There was a process missing for identifying data sources.

2: ‘Generate data models and linkages’ process split into two to improve mutual exclusion. Generating metamodels was intended to reflect scanning data sources, but has become obsolete by the new process. This process now solely covers the linkages between datasets.

3: Added process comprehensiveness. A process was missing for generating a trust score to identify master data.

4: Improve uniformity by using master data management instead of the abbreviation MDM. Also add ‘/tool’ to improve understandability.

5: Improve understandability by renaming ‘clustering and blocking attributes’ to , ‘ match and merge’, as this was a more common term within master data management.

6: Added process to improve comprehensiveness. A process was missing where the single point of truth/golden record is identified.

7: Remove ‘and define stewardship’ in the sub-capability description to improve comprehensiveness. The processes in this sub-capability only cover maintenance processes.

Database Management (Table 29)

1: Added process to improve comprehensiveness. A process was missing for developing database instances.

2: Simplified description to improve understandability.

3: Improved understandability by introducing a new sub-capability. Other capabilities have a monitoring sub-capacity, while database management also covered process related to monitoring.

4: Added process to improve comprehensiveness. A process was missing for optimizing storage performance and cost.

G. Transcript of Case Study Evaluation

	Note: all transcriptions only cover comments about the maturity model and evaluation criteria. The results of the maturity assessment can be found in the thesis, any details provided during the interviews are undisclosed. **1: Assessment Case 1: Health Insurer Corporate Data Steward 07-09-2020** DQ [Rating Processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Compreh.	1: No, it's fine like this. I interpreted monitoring data differently, within the organization we interpret it differently. It's not on the data itself, but the data quality itself, of a dataset. D: Can you elaborate a bit more? 1: It's more the governance around it. MD [Rating processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Understan.	1: I needed an explanation a couple of times. Other than that it's fine. DI [Rating processes] 1: I'll skip the last two sub capabilities, you can better ask my colleagues. D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 1: I think it is fine like this MDM [Rating processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Compreh.	1: No it's fine DBMS: 1: I think for these processes you can better go to another colleague than me.

	D: Do you want to go through the evaluation now or should I e-mail it to you? 1: You can e-mail it to me and then I will fill it in and send back. **2: Assessment Case 1: Health Insurer Data Quality Expert 08-09-2020** DQ [Rating Processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Compreh.	2: No. It is hard if you're young in your data management, to score this. MD [Rating processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 2: Not really, nothing I can come up with. DI [Rating processes, skipping some] 2: For this capability I feel like I'm not the right target group. Its not that your questions are wrong, I'm just not the right one to respond. MDM [Rating processes] DBMS: [Rating processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Participant Profile	2: No. I can answer this from my old position. My feedback is mainly that you have to check who you're asking. I think if I ask the administrator that just started, would score different. I had that position three years ago. 2: It differs who you talk to and what answer you will get. You could look into assigning a certain weight to people, so that my answers for DI for example weigh less. 2 on [Question 1-8]: Nothing to add
Compreh.	

	2 on [Question 9]: The model focusses on the technical aspect of data management and not on governance. A integration of the two would make the model more useful. 2 on [Question 10]: No **3: Assessment Case 1: Health Insurer Data Quality Expert 10-09-2020** DQ [Rating Processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 3: No, but that is due to our current state I think. You can see that we don't score so high. I think you included enough, I can't add anything. I think along the road, when your maturity increases you can start to add something. I think with a little assistance, everything is clear as well. MD [Rating processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Understand.	3: I get most of them, especially after your comments. So that's clear. DI [Rating processes, skipping some] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 3: No, you have data discovery, data lineage and impact analysis. I don't think there's more to it. MDM [Rating processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Understan.	3: No, the ones that were unclear, were the ones that I was confused about myself, like the bottom up top down one. Other than that I don't miss anything. DBMS [Skipping everything]

Level accuracy	Evaluation 3 on [Question 1]: The usage of the model, if you want to keep it simple, you should keep it at this. Content-wise its fine, so that's good. 3 on [Question 2]: You use the terms AI and automate. As long as it is clear that you can apply both, because they are different activities. That could be the only thing. Level 4 is automated, while you mention AI. I can imagine people can get confused. Maybe in the future you need to change the descriptions, but for the time being this is not the case. 3 on [Question 6]: No 3 on [Question 7]: No, that was clear to me
Usefulness	3 on [Question 9 and 10]: I don't think so. The model itself is fine I think. It's more about in which phase the organization is on data management maturity and are they ready to look at next steps. What I said before, it is good to look at automation and AI, but if you're at basis level, than you need to question yourself if you ready for this model, but you already explained that. I think that you need a certain foundation to apply this model.
Participant profile	3: It's also about who you're asking. Some things I wasn't able to answer because it's more technical, in our case that would be at data engineering. Every organization has arranged its data management different. I'm mostly business, but not so much on technology. So that another tip, to look at who you're asking. I don't think there's a function that covers all. So maybe its valuable to do these sessions multidisciplinary. I also think that the value for the participants will be higher. I often see that if you discuss data management aspects with people from business, architects and data engineers, that conversation often is very valuable. 3 on [Question 8]: No
Scoring schema	**4: Assessment Case 2: Bank Business Consultant Data Management 11-09-2020** DQ [Rating Processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 4: I notice that there can be large differences per business line. There can be a large difference in specific processes if you talk about the company as a whole. In one chain it might be very high, while another chain can be very low. That weighing is not incorporated. This is an overall view by me, but than you miss the details per business line and I think that there those maturity levels come in. Per business line you can decide how much time and effort you want to spend on improving the maturity, how fast you can go and which business value you can

Scoring schema	generate. I think it can be interesting in practice. This provides a overall score, which does not give a score per business line. D: I agree, it would be interesting to apply this model in different teams. Also, I see a capability as the ability to perform certain processes. So if you rate the highest one, you know the capability is there and can be transferred to less maturity processes. 4: Yes, sure. But my experience is that the weakest process in the chain limits the ability to go to AI. It seems that you can only apply AI when you reach level 2 or 3,
Compreh.	
Level sufficiency	

	before the manual input it too high. The system as a whole it not ready then, that's my opinion. MD [Rating processes]
Compreh.	D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 4: They are more extensive than I expected. Creating metadata, instead of doing it all manually, it could be interesting to do everything in bulk automatically. I think that is a good sub capability to mention. About metadata architecture, these are the process steps, but we also have some part that covers which organizational structure you need to take decisions, I miss that.
Process mutual excl.	D: Yes, my model focusses more on processes where AI can assist. I do think there are processes like data governance and organization architecture that are also important at those capabilities. DI [Rating processes, skipping some] D (Question 3,4,5): Based on time, let's continue MDM, DBMS [Rating processes] 4 on [Question 1]: In practice we see that it is increasingly important to know the details. I think that a maturity scale with two layers (12 levels) would be better.
Understan.	**5: Assessment Case 3: Insurer Head of Data Management 11-09-2020** DQ [Rating Processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 5: No, if you look at the sub-capabilities it follows the main topics of DMBOK MD [Rating Processes]
Process compreh.	D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 5: I wouldn't remove one. Some have overlap. I don't think you can get it MECE. I think you cross checked it with DMBOK?

	D: Indeed, some have an overlap but I do try to make distinct enough for each capability. 5: I do think its fine, these are the main steps. Also when you look at the sub capabilities it is fine. We use different sub capabilities. This is more grouped on process steps and we group more on domain, but it covers the same. DI [Rating Processes] D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague? 5: Not directly. What could help is an example column. That would help to understand the context. D: Would you say an example for level 1 and example for level 5?
Process mutual excl.	5: Indeed, now I have to imagine it in my head and relate the maturity levels. MDM
Ease of use	[Rating Processes]
Ease of use	D (Question 3,4,5): Do you miss any process? Do you think one is redundant or a description is vague?
Mutual exclusion	5: What I might miss here, is that if you look at DMBOK MDM. You have three functions: the match functionality, the DQ functionality and the golden record and propagation functionality. The last one I'm still missing a bit. Matching is here, DQ maybe a bit less, but it is in DQ itself. You could also add propagation of master data, maybe also as part of data integration. D: That might be included in 'single point of truth'. Maybe I could add /golden record? 5: What I mean is that if you generated a single point of truth, then you want to link all the downstream systems to this golden record, instead of point-to-point to the local source systems. It's more about pushing this golden record. DBMS 5: This is more towards IT, that would be mostly guessing for me.
Scoring schema	D: We can skip processes if you can't score them.

	5: At my previous employer, all these capabilities were within one unit: analytics, engineering and data management. At my current organization this is more split. Evaluation 5 on [Question 5]: I think we can make it more MECE. 5 on [Question 9]: I get what you're trying to do, but in my head I'm constantly switching between two levels. 5 on [Question 10]: I would make it strongly agree if we could make the model more simple. D: Where do you think I can improve the model? 5: I think we can make the processes more MECE, I think they are comprehensive. Also make the maturity scores more specific for the data management processes. The third is to make the model more simple, now you have sort of a matrix but we could think about making it more simple.
Particpant profile	D: I tried to combine the levels, but that is impossible because of the difference. It could be an option to only do the augmentation scale and not data management maturity. 5: I agree, because it is not strictly necessary. D: I thought that it could also make it easier, as participants often work with process maturity levels and by letting them score this scale first they can focus on the other level. Did you experience something like that? 5: The challenge was mostly that you switch from as to 'to be', so that makes it more complex.
Usefulness Usefulness	**1-3: Follow-up Case 1: Health Insurer All Participants 22-09-2020** D: What did you think of the presentation of the results and recommendations? 1: When I look at the example timeline, I think that we already did some of those steps. But if I remember correctly, I didn't see these processes in the process questions. So I'm thinking whether the recommendations match with the questions that were asked, I think we need to dive into this and see what it can mean for us. D: Yes, the example deadline is an example of the difference between 'what' happens and 'how' it happens. If you implement a council and tooling, from the timeline, it is expected that some of the processes from the assessment are

	performed. It is very dependent on the organization which initiatives occur, it is merely an example. 2: I think it is quite a lot. I don't know yet what to do with it as it is a lot. The other is that we couldn't fill in every process, but that doesn't give the complete picture. You can ask us about data integration, but you could also include them. I think such an assessment works better if you include multiple disciplines and maybe assign a weighing. An average score might not be representative for the organization. D: I agree. Data integration and database management might be a bit outside of your scope. I hope that the capabilities that are within your scope, data quality, metadata management, that the results and steps can be used to come to some improvements in that area. 1: I think it is definitely helpful. 3: It could also be of value to ask other departments what they think about metadata, what their vision is, then you might not always get a well informed answer. But still is would be interesting to weigh this in to generate a wide example.
Usefulness	D: All right. I will send you the slides of all the results so that you can analyze them and hopefully can use them to identify improvement initiatives. I also want to ask
Usefulness	
Practicality	

Practicaloty	you to rate a few evaluation criteria based on the improvement steps and roadmap steps. [Rate Evaluation criteria] 2 on ease of use: I don't know yet. I can't say I can't use them, so I'm neutral. 3 on ease of use: Yes they are usable, I get what it says, on the other side it's no because we still have to experience it.
Usefulness	1 on ease of use: I think these steps are easy to use, they are understandable so you can implement it. But, the content is different. 2 on usefulness: I think they are useful, but not yet how useful. 1 on usefulness: What my doubt is, is that we've been on data management for a while now. If you're just new I think it is so high level that you can't work with it. I think I score a three. 3 on practicality: If I need to apply those improvement steps now, I wouldn't know what to do, so it's a two for me **4: Follow-up Case 3: Insurer Single Participant 25-09-2020** D: Do you apply road mapping currently? 4: Certainly. Not on augmentation, that's where you trigger me. For data governance, data quality, metadata management we have roadmaps, in EPIC format for the following year. We focus less on how we can apply AI. With that lens we could look at the roadmap for next year. 4: Interesting to see how we compare to the other use cases. I think there's always a difference due to interpretation. Somehow maturity models remain subjective. D: Do you have any questions or remarks regarding the slides that I just showed you? 4: No, not directly. I enjoyed it to participate and to see how we compare to other case organizations, namely on the augmentation. Also useful for me to consider

	the promise of AI within data management. You could wait or already look at how you can apply it. Also this is a topic that's not covered by DMBOK at the moment. D: At last I want to ask you to rate these statements, like we did before. It's about the understandability, ease of use, practicality and usefulness. 4 on ease of use: four, the only reason I'm not giving it a five, is that you could specify more structure. You could give more details on the improvement steps and maybe add how it would fit in an agile way of working. In our roadmaps we have some 'one of' things that, single changes that are covered in the agile process. Other processes are ongoing, data quality issue management for example. The principle is ongoing, but large changes show up in the backlog. 4 on usefulness improvement steps/roadmap: four, I think they are fine. You should detail them a bit more. It is now described as an introduction. 4 on practicality improvement steps/roadmap: a three, I think as a construct a good starting point to define initiatives. Maybe you could come up with guidelines that would structure it a bit more, also like the agile piece. How to go to execution, it is a bit high level now. **5: Follow-up Case 2: Bank Single Participant 28-09-2020** D: What did you think of the slides that we just went through? 5: It is very valuable to give attention to all the capabilities and the steps that you can take to improve them. Also to have an overall view, at a certain point you're working on one capability in a silo. D: At last I want to ask you to rate these statements about the understandability, ease of use, practicality and usefulness. 5 on understandability improvement steps: They are understandable, but they miss cohesion. In the foundation they are ok, but they don't really connect. I don't know how they relate to your assignment. D: It's something extra that I wanted to give to the participants, the main value is in the model and the assessment. If the assessment would be more thorough, these steps would be in more detail. 5 on ease of use for roadmap: It's hard to apply these roadmaps. How do you make it possible to let everyone understand what needs to be done and how important it is. What is miss in these steps is a step where you have a main vision. Without that, you don't get much value from the 'who', why would you do all these complicated things? To provide a motivation. D: how would you name this step?

	5: Define a mission or vision statement. The phenomena of AI is that it has potential and can add value, but it can prove itself more if it an realize a vision. D: Thank you, do you have any other questions or comments? 5: Like I said, it's very nice to go over all capabilities on a high level. Than you see that it can help to identify where you are now and where you want to go. I think it is good to have this as a KPI. Certainly if those large gaps, if you can emphasize them visually. If you can work with colors or size to show priority, that would really help to get the message across. That's what I really like about the improvement steps.

Table G1: Transcription of Case Study Evaluation

Comprehensiveness	All aspects covered, missing processes or (sub) capabilities, suggestions for improving definitions so they cover all aspects
Ease of use	Remarks regarding ease of use
Level accuracy	Whether there is overlap between the maturity levels
Level sufficiency	Whether the levels are sufficient to represent all maturation stages of the domain
Participant profile	Regarding the selection of participants
Practicality	Remarks regarding practicality
Process mutual exclusion	Whether processes are distinct
Scoring schema	Remarks about the scoring schema
Understandability	Suggestions for improving definitions, clarifying, giving examples, logical order.
Usefulness	Remarks regarding usefulness

Table G2: Labels for Case Study Evaluation

H. Result of Case Studies Maturity Assessment

	Case 1: Health Insurer				Case 2: Non-health insurer				Case 3: Bank			
	DM		ADM		DM		ADM		DM		ADM	
Data Quality	2,1	3,4	0,9	2,8	1,7	3,6	0,3	0,3	1,9	2,7	1,4	2,3
Assess DQ	2,3	4,3	0,3	3,5	1,0	3,0	0,0	0,0	2,0	3,0	0,0	1,0
Data profiling	2,7	4,3	2,7	4,0	2,0	4,0	0,0	0,0	1,0	2,0	1,0	2,0
Data cleansing	1,8	3,6	0,1	1,9	1,7	3,3	0,0	0,3	1,3	2,3	1,0	2,7
Monitor DQ	1,5	3,7	0,0	3,3	2,0	4,0	1,0	1,0	3,0	4,0	3,0	4,0
Metadata Management	1,5	3,6	1,8	3,4	1,1	2,7	0,5	0,5	2,1	3,3	1,8	2,5
Define metadata architecture	1,9	3,7	2,4	3,8	1,0	2,0	0,0	0,0	3,0	4,0	2,0	3,0
Create and maintain metadata	1,7	3,5	2,7	3,5	1,2	3,0	1,5	0,6	2,3	3,3	2,4	2,5
Analyze metadata	0,8	3,7	0,2	3,0	1,0	3,0	0,0	1,0	1,0	2,5	1,0	2,0
Data Integration	1,2	3,8	1,2	3,6	1,2	2,5	0,0	0,7	1,6	2,6	0,5	1,3
Data discovery	0,5	3,0	2,0	4,0	1,0	3,0	0,0	1,0	1,0	2,0	0,0	1,0
Analyze data lineage	1,3	4,0	1,0	3,6	1,0	2,3	0,0	1,0	1,5	2,5	1,0	1,0
Design DI	2,0	5,0	0,0	3,0	1,3	2,3	0,0	0,3	2,0	3,0	0,0	1,0
Develop DI					1,5	2,5	0,0	0,5	2,0	3,0	1,0	2,0
MDM	0,6	3,5	0,4	2,8	1,1	2,5	0,0	0,9	1,3	2,3	0,0	0,8
valuate and assess data sources	0,9	3,7	1,3	3,0	1,0	2,3	0,0	1,0	0,7	1,3	0,0	0,3
Model master data	0,7	3,6	0,0	2,4	1,3	2,7	0,0	0,7	1,7	3,0	0,0	1,0
Maintenance process	0,3	3,2	0,0	3,0	1,0	2,5	0,0	1,0	1,5	2,5	0,0	1,0
DBMS	2,6	4,0	1,1	3,1					2,3	3,3	0,3	1,3
Develop Database Instances	2,0	4,0	1,0	3,0					2,0	3,0	0,0	1,0
Manage database performance	2,7	4,0	1,3	3,3					2,7	3,7	0,7	1,7
Monitor Database	3,0	4,0	1,0	3,0								
Universal	1,5	3,2	1,0	2,9	1,3	2,7	0,1	0,5	2,0	2,8	1,1	1,9
Data Rules (DQ5,MD6)	2,0	3,8	1,0	3,5	1,0	2,0	0,0	0,0	2,0	3,0	1,0	2,0
Data Modelling (MD5,DI9,10,MDM8,DBMS5)	1,1	3,1	1,1	2,8	1,3	2,3	0,0	0,3	2,3	3,3	0,5	1,5
Validation Checks (DQ6, DI12)	1,8	3,7	0,3	3,0	1,5	3,0	0,0	0,0	2,5	3,5	0,5	1,5
Similarity Identification (MD11,DI5,MDM6)	1,2	3,7	1,8	3,6	1,0	2,7	0,0	1,0	1,0	2,0	0,3	1,3
Monitoring (DQ11,12, MD14,DBMS9,10)	1,4	3,6	0,1	3,2	1,7	3,3	0,3	1,0	2,3	3,3	2,0	3,0