Empirical Development and Evaluation of a Maturity Model for Big Data Applications

Von der Fakultät Wirtschaftswissenschaften der Leuphana Universität Lüneburg
zur Erlangung des Grades Doktor der Wirtschaftswissenschaften

- Dr. rer. pol -

genehmigte Dissertation von Thomas Hansmann

geboren am 02.09.1985 in Marburg

Eingereicht am:	24.06.2016
Mündliche Verteidigung (Disputation) am:	19.01.2017
Erstbetreuer und Erstgutachter:	Prof. Dr. Peter Niemeyer
Zweitgutachter:	Prof. Dr. Burkhardt Funk
Drittgutachter:	Prof. Dr. Paul Drews

Elektronische Veröffentlichung des Dissertationsvorhabens inkl. einer Zusammenfassung unter dem Titel:

Empirical Development and Evaluation of a Maturity Model for Big Data Applications

Veröffentlichungsjahr: 2017

Veröffentlicht im Onlineangebot der Universitätsbibliothek unter der URL:
http://www.leuphana.de/ub

Contents

Contents i

List of Figures iv

List of Tables v

Abbreviations vii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Statement of the problem . . . 4
- 1.3 Scientific-theoretical and topical classification . . . 5
- 1.4 Applied research methods . . . 9
- 1.5 Organization of the thesis . . . 10

2 Characteristics of Big Data 12
- 2.1 Big Data - Volume, variety, velocity: a first characterization . . . 13
- 2.2 Characterizing dimensions of Big Data . . . 20
- 2.3 Validation and enrichment of the Big Data dimensions using Topic Models 30
  - 2.3.1 Topic Models - Methodological foundation . . . 31
  - 2.3.2 Data selection and preprocessing . . . 34
  - 2.3.3 Analysis on the overall database using Topic Models . . . 34
  - 2.3.4 Analysis on the dimensional level . . . 37
    - 2.3.4.1 IT infrastructure dimension . . . 39
    - 2.3.4.2 Method dimension . . . 40
    - 2.3.4.3 Application dimension . . . 41
  - 2.3.5 Discussion of the results . . . 42
- 2.4 Classification of the results into a generic data analysis process model . . . 44
- 2.5 Distinction between Big Data and Business Intelligence . . . 47
- 2.6 The critical perspective on Big Data . . . 48
- 2.7 Main chapter results . . . 50

3 Maturity Models - Theoretical foundations 52
- 3.1 Reference Models - Definitions . . . 54
- 3.2 Process steps for reference creation . . . 56
  - 3.2.1 Model construction . . . 56
  - 3.2.2 Model application . . . 57

3.3 Maturity Models . . . 58
3.3.1 The concept of Maturity Models . . . 58
3.3.2 Model elements and characteristics . . . 60
3.3.3 Current research . . . 63
3.3.4 A critical perspective on Maturity Models . . . 64
3.3.5 Generalized process models for Maturity Model construction . . . 65
3.3.6 Current research on Maturity Models in the field of Business Intelligence and Big Data . . . 68
3.4 Main chapter results . . . 74

4 Development of the model construction process 75
4.1 Model construction - Theoretical basis . . . 75
4.1.1 Construction model by Bruin et al. [2005] . . . 76
4.1.2 Construction model by Becker et al. [2009] . . . 79
4.1.3 Model comparison and evaluation . . . 81
4.2 Development of the construction model . . . 84
4.2.1 Step 1 - Definition of problem and scope . . . 85
4.2.2 Step 2 - Identification of dimensions . . . 85
4.2.3 Step 3 - Comparison with existing Maturity Models . . . 86
4.2.4 Step 4 - Select design level and methodology . . . 87
4.2.5 Step 5 - Model population . . . 88
4.2.6 Step 6 - Model evaluation . . . 89
4.2.6.1 Model evaluation - Theoretical foundation . . . 89
4.2.6.2 Evaluation against the real world . . . 92
4.2.7 Step 7 - Documentation of the final model . . . 93
4.2.8 Step 8 - Model maintaining . . . 94
4.3 Evaluation of the construction model against the identified research gap . 95
4.3.1 Evaluation against the identified research gap - The principles of Design Science Research . . . 96
4.3.2 Evaluation against the research gap - The principles of general accepted modelling . . . 98
4.4 Main chapter results . . . 100

5 Application of the construction model - Development of the Maturity Model 103
5.1 Definition of problem and scope . . . 104
5.2 Identification of dimensions . . . 106
5.3 Comparison with existing Maturity Models . . . 107
5.4 Select design level and methodology . . . 108
5.5 Model population . . . 111
5.5.1 Model calculation - Theoretical foundation of the Test Theory . . . 112
5.5.2 Development of the questionnaire . . . 120
5.5.3 Data gathering . . . 135
5.5.4 Model calculation - application of the Birnbaum model - description of the initial model . . . 137
5.6 Model evaluation . . . 142
5.6.1 Evaluation of the initial model . . . 142

5.6.2 Evaluation based on the deployment of the fitted Model 148
5.7 Step 7 - Documentation of the final model 165
5.8 Step 8 - Model maintaining 169
5.9 Main chapter results 171

6 Final **174**
6.1 Summary 174
6.2 Limitations 178
6.2.1 Maturity concept based limitations 179
6.2.2 Method based limitations 180
6.3 Outlook and future research 182

A Questionnaire used for the data gathering in construction step 5 - Model population **185**

B Step 6.2 Evaluation based on the deployment of the fitted Model - Evaluation of additional companies **191**

Bibliography **201**

List of Figures

1.1 Interest in Big Data based on Google Search . . . 2
1.2 Number of publications in the field of Big Data research . . . 3
1.3 Reference Frame for Design Science Research . . . 9
1.4 Process steps of the dissertation project . . . 10

2.1 Structure of Chapter 2 . . . 13
2.2 Cycle of data generation . . . 15
2.3 Literature review process . . . 22
2.4 Big Data tag cloud . . . 30
2.5 Application of the Topic Model approach . . . 31
2.6 Description of the Generative process . . . 32
2.7 Current Big Data research in the context of the generic data analysis model 45

4.1 Maturity Model construction process . . . 102

5.1 Example of an Item Characteristic Curve . . . 114
5.2 Exemplary Item Response Theory models . . . 115
5.3 Hierarchy of Dimensions, Topics, and Measurements . . . 121
5.4 Initial model . . . 138
5.5 Fitted model . . . 149
5.6 Number of items per maturity level . . . 151
5.7 Maturity evaluation-relevant aspects from the focus group members point of view . . . 163
5.8 Description of the final Big Data Maturity Model . . . 168

List of Tables

1.1 Characteristics of Behavioural Science and Design Science 6

2.1 Factors influencing the value of data 19
2.2 Characterizations of Big Data 24
2.2 Characterizations of Big Data 25
2.2 Characterizations of Big Data 26
2.3 Characterizations of Big Data from an industry background 27
2.3 Characterizations of Big Data from an industry background 28
2.3 Characterizations of Big Data from an industry background 29
2.4 Results of the Topic Model application on the overall corpus 35
2.5 Results of the Topic Model application on a randomly generated corpus . 37
2.6 Number of publications per dimension after the manual assignment 38
2.7 Results of the Topic Model application on the publications belonging to the IT infrastructure dimension 39
2.8 Results of the Topic Model application on the publications belonging to the method dimension 41
2.9 Results of the Topic Model application on the publications belonging to the application dimension 42

3.1 Maturity model elements 61
3.2 Steps of Maturity Model construction approaches in current research . . . 67
3.3 Analysis of existing Maturity Models in the field of Business Intelligence and analytics 71

4.1 Framework for the analysis of existing Maturity Models 87
4.2 Systematization of evaluation approaches for the evaluation against the identified research gap 95
4.3 Principles of general accepted modelling 99

5.1 Characteristics of the focus group members 104
5.2 Contribution Step 1 - Definition of problem and scope 105
5.3 Contribution Step 2 - Identification of dimensions 107
5.4 Contribution Step 3 - Comparison with existing maturity models 107
5.5 Contribution Step 4 - Select design level and methodology 108
5.6 Contribution Step 5 - Model population 111
5.7 Exemplary result matrix for a test with binary questions 116
5.8 Topics and items in the final questionnaire 122
5.9 Characteristics of the respondents 136

5.10 Contribution Step 6.1 - Evaluation of the initial model 142
5.11 Quantitative analysis of the first model evaluation step 145
5.12 Contribution Step 6.2 - Evaluation based on the deployment of the fitted model 148
5.13 Overview evaluation results step 6.2 156
5.14 Contribution Step 7 - Documentation of the final model 165

B.1 Overview evaluation results step 6.2 192

Abbreviations

BI	Business Intelligence
DQM	Data Quality Management
RDBMS	Relational Database Management System
ICC	Item Characteristic Curve
IRT	Item Response Theory

Chapter 1

Introduction

1.1 Motivation

Since 2000, data generation has been rapidly growing from various sources such as Internet usage, mobile devices, and industrial sensors in manufacturing [Hilbert and López, 2011]. As of 2011, these sources were responsible for a 1.4-fold annual data growth [Manyika et al., 2011]. Furthermore, the storage and processing of the data has become less expensive and facilitated due to technological developments, such as distributed and in-memory databases, running on commodity hardware, and decreasing hardware prices [Armbrust et al., 2010]. The resulting massive influx of data has inspired various notions, with the most popular notion being Big Data. For companies, this trend becomes a major topic of interest. Independent of the industry, the amount of data influences a plurality of processes along the value chain [BMWi, 2014], thus has the potential to change how companies work. Data analysis has changed from being merely just one amongst numerous company-internal topics to being one of the most prioritized and valued focus subjects throughout companies [Accenture and GE, 2015].

The hype of Big Data can be recognized both amongst practitioners as well as in the scientific community. Taking the practitioners dimension, several developments, such as trends in the general interest of specific searches (e.g. referring to the development seen in the Google trend tool for the query "Big Data", that shows an significant increase of

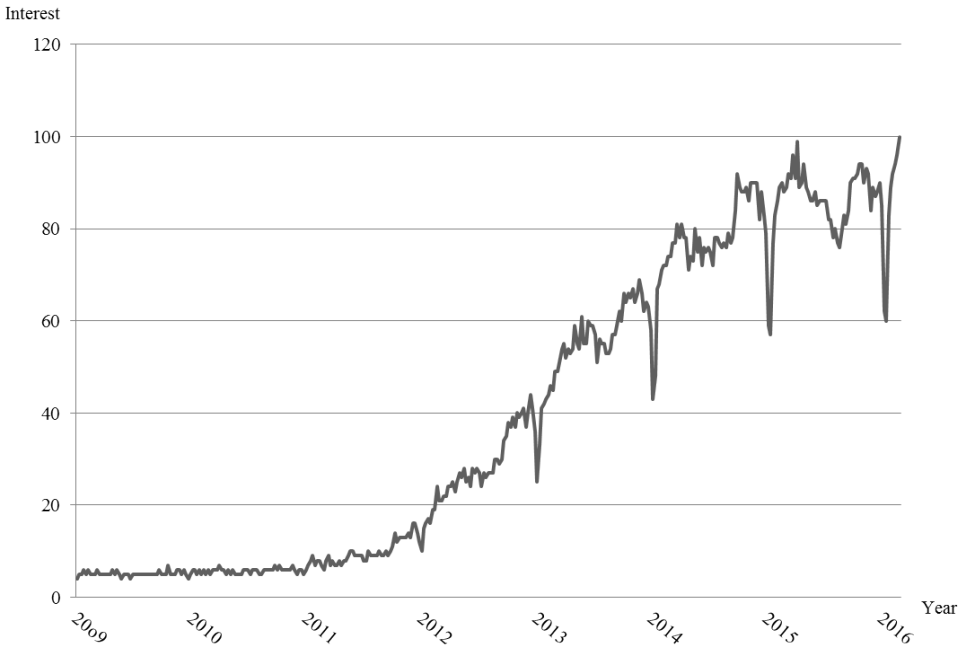


FIGURE 1.1: Interest in Big Data using Google Trend function with the keyword "Big Data"[3]

interest in this topic since the year 2011[1] [2]) or in political initiatives (e.g. multi-million cooperation projects, led by the Federal Ministry for Economic Affairs and Energy for ideas to improve the use of data [Bundesministerium für Wirtschaft und Technologie, 2014]), indicate the increased interest of the public.

Besides the general interest, Big Data is also seen as a competitive advantage and companies feel the need to improve the capabilities in this field. Considering for example the information technology industry and consultancies, the share of business activities connected to this field has been steadily increased. Following the German trade-association *Bitkom*, the world wide turnover generated with Big Data related services and products is expected to rise from 23.6 Billion Euros in 2011 to 166 Billion Euros in 2016 [Weber and Shahd, 2014].

A similar development can be found for the number of scientific publications on Big Data

[1] The Google Trend tool analyzes the number of queries for one keyword in relation to the overall number of keywords and the change over time.

[2] The public and science groups cannot be treated completely separate as a scholar might use Google as a starting point for the research on Big Data as well. Nonetheless, scientific databases such as *Scopus* and *ACM Digital Library* can be utilized similarly as google trends indicators for interest in a certain topic.

[3] The graphic is developed based on the google trend tool. The tool can be accessed via https://www.google.de/trends/

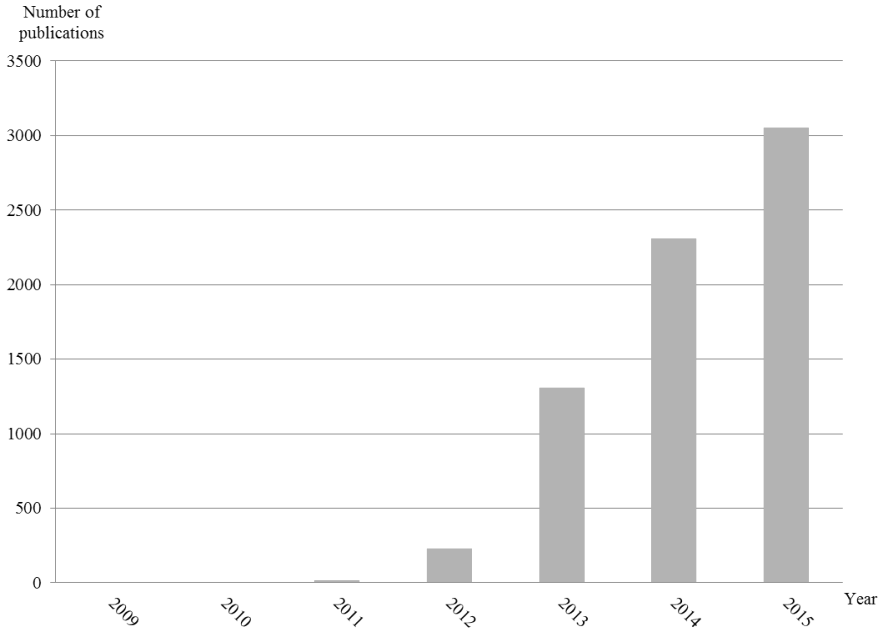


FIGURE 1.2: Number of publications listed in the research database IEEE Xplore for the period 2009 - 2016[4]

as an indicator for the interest of the scientific community. Following the aforementioned significant increase of interest in Big Data since 2011, the course of interest from the scientific community in Big Data, - taking the number of publications from the scientific database IEEE Explore as an example - shows a similar development compared to the public interest in Big Data. The topic of Big Data gained relevance in multiple research disciplines, e.g. computer science [Herodotou et al., 2011], information system research [Agarwal and Dhar, 2014], biology [Marx, 2013], and medicine [Chawla and Davis, 2013]. In addition, formats for discussion increasingly emerged, resulting in the creation of subject specific journals and conferences, such as *Big Data Journal, IEEE Conference on Big Data*, and *International Congress on Big Data.*

Taking this breadth of disciplines and continuous growth of consideration into account, it becomes apparent that despite the partly negative consequences from the increasing orientation towards sheer data [Boyd and Crawford, 2012], following the Gartner Hype Cycle for emerging technologies 2014, [Rivera and van der Meulen, 2014], Big Data has overcome the phase of being a momentary hype, being in a phase of disillusion and on its way to a productivity phase.

[4]The illustration is based on the results from the *IEEE Xplore* database query of the keyword *Big Data* in key words, abstract and title over all disciplines beginning in the year 2009 till 2015.

1.2 Statement of the problem

The relevance of Big Data and this research becomes apparent when looking at its perception by practitioners. Despite its novelty, Big Data is already perceived as a competitive-relevant topic. Following the survey by Accenture and GE [2015], the biggest concern for a company not having implemented a big data strategy properly would be a loss in market share to their competitors. Fostered by the information technology industry's driven marketing, companies increasingly perceive the urge of improving their handling of (Big) data.
As it will be further discussed in Chapter 2, Big Data is not limited to a technological capabilities, which have already been perceived since the 1990s as a competitive factor [Powell and Dent-Micallef, 1997]. Capabilities with relevance for the successful handling and utilizing of Big Data belong to diverse fields such as organization and management. These will be exemplified in the following chapter.

This high perceived relevance in combination with the dynamic development of the topic leads likewise to challenging situation from a company's point of view. With regard to the diversity of existing applications and the early stage of operationalization, only a few best practices for the handling of Big Data exist. Therefore, companies are faced on the one hand with numerous fields of potentially relevant and advantageous possibilities and on the other hand, with an insecurity and uncertainty as to which *capabilities* should be developed in order to utilize the available data in the most successful way possible.

This managerial decision is the starting point for the research in the thesis at hand. The resulting research question is:

How can the analysis of huge data amounts from different sources and with heterogeneous structures be improved?

One approach for the identification of capabilities and their allocation to different levels of professionalism are maturity models. Maturity models describe the capabilities of companies in a specific topic [Paulk et al., 1993]. They belong to the field of design science research and are called artefacts in the language of information system research. In order to identify the need for actions and developments regarding the improvement

of the use of Big Data applications, companies need information about their current maturity and potential future development [Becker et al., 2009]. The characteristics of maturity models will be explained in detail in Chapter 3. In the next section, an overview about types of research in the field of information systems is provided and the research carried out in this thesis is classified.

1.3 Scientific-theoretical and topical classification

The research conducted in the following chapters can be assigned to the field of information system research. Research in this field can be distinguished into Behavioral Science and Design Science.

Mettler [2010] gives an overview about characteristics differentiating the Behavioral Science and the Design Science Paradigm. The characteristics of *Goal, Process of Knowledge Generation*, and *Evaluation of Knowledge* especially explain the classification of the research project in the field of design science (Table 1.1).

With regard to the research question, no underlying theory - as noted for Behavioral Science - is applied to identify and evaluate companies' capabilities in dealing with Big Data. Instead, the resulting maturity model as an artefact can be used to change realities, i.e. the improvement of capabilities based on the maturity evaluation.

Targeting the *process of knowledge generation*, the generalization which is associated with Behavioral science does not fit to the character of maturity. As it will be further explained in Chapter 3, maturity is characterized by its relativity and dynamic. Capabilities associated with a high maturity can decrease in maturity in the course of time as the topic in focus evolves, resulting in an overall improvement of capabilities. Consequently, maturity is associated with a specific point in time and is subject to a high dynamic. Therefore, a generalization of results is hardly possible. Maturity models can rather be understood as approximations to the real world.

In contrast, the iterative approach of Design Science Research fits to the described characteristics, leading to the possibility of a continuous fitting of the developed artefact to changes of the environment.

This approach further applies to the *evaluation of knowledge*. The separation of knowledge generation and application is not possible as during each application of the maturity model, knowledge about needed fittings, resulting from the dynamics described before,

TABLE 1.1: Characteristics of Behavioural Science and Design Science [Mettler, 2010]

	Behavioural Science	Design Science Research
Goal	Description and explanation of realities based on theories	Change of realities based on artefacts
Perception of reality	An ontic reality exists which is responsible for the perception of the subject (realism)	An ontic reality exists; it is linked to a subject which results in a distortion of the findings (relativism)
Evaluation of knowledge	A logical separation of knowledge generation and knowledge application exists. Methodological principals and procedures are supposed to guarantee the quality of the knowledge (positivism)	A logical separation of knowledge generation and knowledge application is not possible/not intended. Little methodological rigour; Firmness of the argumentation defines the goodness of the knowledge (pragmatism)
Construction of knowledge	It is expected that socio-technical connections can be explained based on empirical data (reductionism)	Data are the basis for the artefact construction but they cannot be used to draw one's own conclusions on the holistic context (emergence)
Process of knowledge generation	Gathering, analysis, interpretation, generalization (sequence)	Analysis of the problem and problem formulation, development respective adoption of concepts, evaluation and re-calibration, synthesis (iteration)
Interaction with the subject of analysis	Actions, which influence the subject of analysis should be defaulted (observer)	Possibilities of influence for targeted changes are used actively

is generated. These aspects result in a classification of the model in the field of Design Science.

As it will be demonstrated in Chapter 3, it is the first maturity model in the field of Big Data. Consequently, the goal is to construct a more generalized model, which can be used as a starting point for future industry and application specific maturity models. Therefore, the goal of the thesis based on the research question is to solve a design problem.
The associated, formulated goal of this thesis thus is

The development of an industry-independent maturity model for the field of Big Data

In the course of research, two further goals are pursued:

- A quantitative approach for the model population is applied in the course of the maturity model construction. Based on the results, it is evaluated in how far quantitative approaches, originally used for the population of maturity models for established disciplines, are applicable for topics that contain both novel as well as established aspects, e.g. Big Data.
- The development and testing of a maturity model evaluation process that is supposed to analyze how far the character of maturity as understood in the practical context is correctly represented in the developed model.

In current research so far, those aspects have not been approached in a comprehensive manner dealing with maturity models on Big Data and will be explained in Chapter 3 and 4 in more detail.

The research goals can be taken as a starting point for the classification of this thesis in the different fields of Design Science Research. The research in the field of Design Science Research can be distinguished as well into two groups, Design Science and Design Research (figure 1.3).
Design Science reflects the design research process and aims at creating standards for its rigour. It focuses on considerations regarding the artefact construction and evaluation.

Design Research, in contrast, creates solutions to specific classes of relevant problems by using a rigorous construction and evaluation process. Related research is focused on the development of new artefacts and the adoption of existing ones [Winter, 2008].
The thesis at hand has its primary focus on Design Research as a new artefact - the maturity model - is developed. Additionally, in contrast to existing works, the results also contribute to the field of Design Science as both a new artefact construction process and an artefact evaluation approach are developed. This classification of the output becomes clear, when looking at the classification approach by March and Smith [1995]. Following their work, research in the field of Design Research can lead to four different outputs; *Constructs, Models, Methods*, and *Instances*:

- *Constructs* can be understood as a basic language, a nomenclature, which is used to describe phenomena. They act as a conceptual foundation for the description and problem solving.
- *Models* can be defined as the combination of different constructs. A more detailed discussion about models can be found in Chapter 3.
- *Methods* are used in the problem-solving context as well. They are on a more detailed level in contrast to models as they contain a description of how an improved state can be achieved.
- *Instances* are the transfer of constructs, models, and models into a physical implementation, mostly software, used for the problem solving.

The research of this thesis leads to two outputs: *constructs* and *models*. The latter is the primary focus, as the aspired result of the work is the development of a maturity model for the field of Big Data. In the course of the research, similar to the work by Mettler [2010], an ontology is developed which contains the primal *constructs* of maturity evaluation.
The development of an instance, in this case the software-based implementation of the maturity model for standardized evaluation purposes, is not in the scope of this research.

After the classification of the research, the applied research methods are presented in the next section.

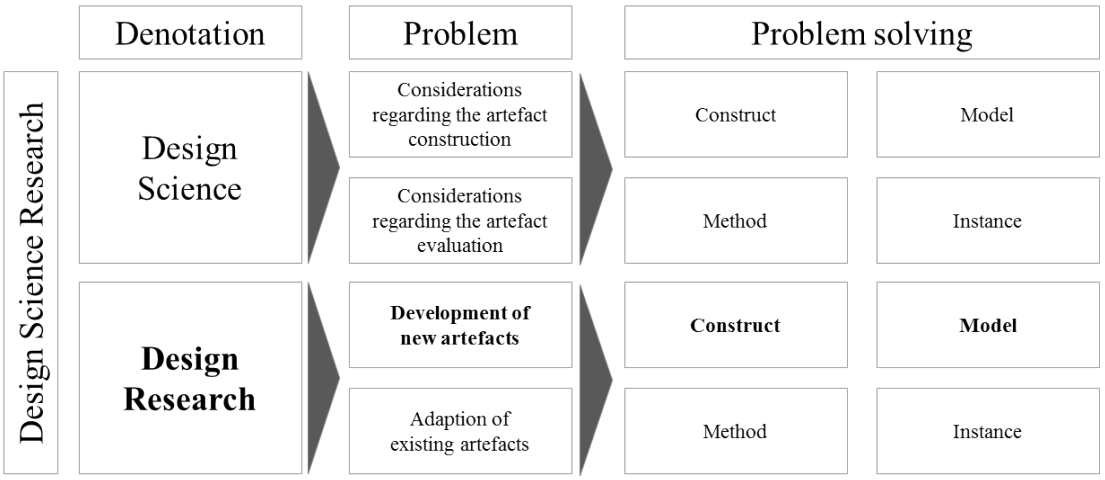


FIGURE 1.3: Reference Frame for Design Science Research along the contained sub-research fields, the targeted problem, and the solving approaches [Winter, 2008; March and Smith, 1995]

1.4 Applied research methods

The thesis at hand pursues a multi-method approach. Three different research methods, all three being quantitative and qualitative, are applied.

Design Science Research acts as a contextual bracket. The maturity model to be developed belongs to the field of reference models and represents an artefact. Therefore the underlying maturity model construction is developed along the Design Science Research principles by Hevner et al. [2004] in order to develop a model with a sound theoretical foundation. Current maturity models have been criticized for lacking a solid theoretical foundation, partly due to the practical-oriented character [De Bruin et al., 2005; McCormack et al., 2009].

In the course of the maturity model construction, both qualitative and quantitative methods are applied. In the beginning of the maturity model construction, quantitative approaches from the field of text mining are applied for a structured literature review aiming at the identification and validation of dimensions that describe Big Data [Blei et al., 2003; Blei and Lafferty, 2009; Chang et al., 2009]. The subsequent model population is carried out based on quantitative approaches from the field of test theory, belonging to the social sciences, based on returned questionnaires answered by participating companies. These quantitative approaches are used to assign the capabilities respective measurements to different maturity levels.

During the course of the research, a focus group is utilized as a qualitative research

method, consisting of industry experts from different consultancies. They support i) the questionnaire development as well as ii) two model evaluation steps.

1.5 Organization of the thesis

The maturity model development starts with a characterization of Big Data in Chapter 2. Based on a structured, quantitative enriched literature review, describing dimensions and characteristics are identified. In order to draw a holistic picture, a critical perspective on Big Data is given.

Chapter 3 contains the theoretical basis of maturity models and their development. It begins with an introduction into models in general, followed by reference models, to which maturity models belong. Afterwards, research on state-of-the-art maturity models is presented. Maturity models from associated fields of Big Data, e.g. Business Intelligence and Management Information Systems are presented and compared regarding the underlying construction model, the construction approach (qualitative/quantitative) and the applied evaluation approach. Based on this comparison, missing aspects with relevance for the Big Data maturity model are identified.

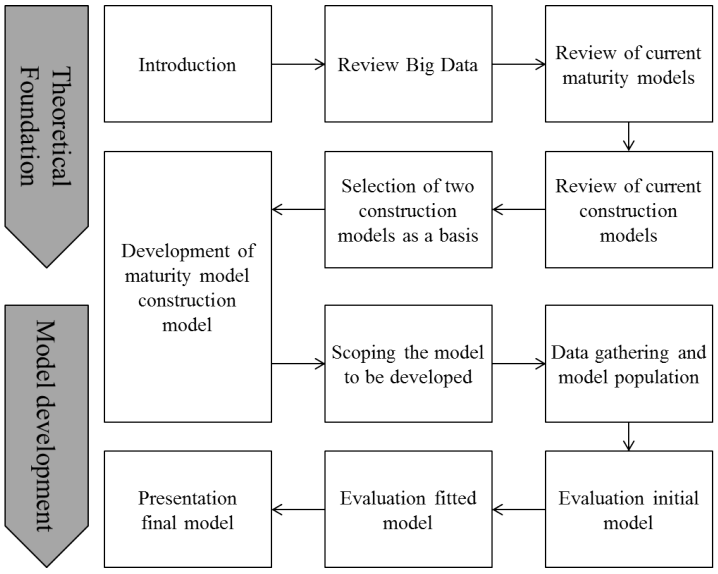


FIGURE 1.4: The dissertation project can be divided in two parts, the setting of the theoretical foundation and the model development, containing as well the model evaluation.

After giving an overview of topic-related maturity models, different construction models for the development of maturity models are discussed. Based on the identified strength and weaknesses of existing construction models, Chapter 4 describes the construction model developed for this thesis, based on two established construction models from De Bruin et al. [2005] and Becker et al. [2009].

Chapter 5 contains the application of the construction model in the field of Big Data. This chapter represents the core of the research and contains the maturity model construction as well as the subsequent evaluation.

The final Chapter 6 summarizes the main findings, describes the limitations of the work, and gives an outlook on potential future research.

Chapter 2

Characteristics of Big Data

Big Data is a subject in different disciplines, indicating its depth and breadth within the practical and scientific discussions. Aspects with relevance for Big Data can be found amongst others in computer science, mathematics, business administration, and the social sciences [Hansmann and Niemeyer, 2014]. Accordingly, the topic of Big Data is attracting increasing attention from the scientific community, which is reflected in the increasing number of i) publications that directly address the notion of Big Data [Chen et al., 2012] [Lynch, 2008], ii) research journals that address solely Big Data, and iii) scientific conferences with a Big Data focus.
Until recently as it will be demonstrated later on in this chapter, publications on Big Data have lacked a clear understanding of the key elements and structure of the topic, which hinders the identification and examination of relevant topics for future research. Consequently, this chapter addresses the following questions:

- Into which dimensions can the concept of Big Data be divided?
- What are the topics for scientific publications within the individual dimensions?

The resulting contribution of this chapter thus is twofold. First, it will delineate a structure for categorizing recent developments in Big Data that is based on analyses of both existing definitions and scientific publications on Big Data. As a result, dimensions and according topics are derived. The identified dimensions will be used as a starting point for the subsequent maturity model development in Chapter 5.

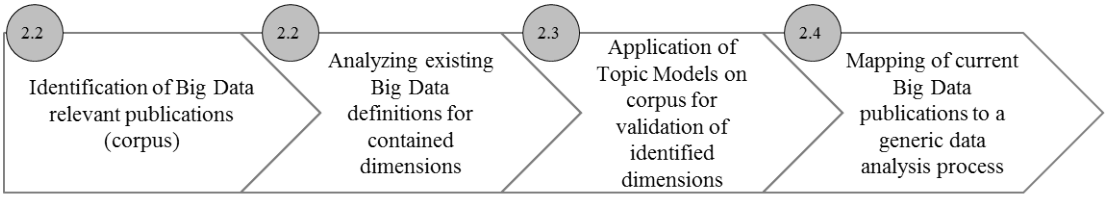


FIGURE 2.1: The literature review consists both of qualitative and quantiative aspects.

Second, the simultaneously identified topics will be used to carve out similarities and differences between Big Data and similar topics such as Business Intelligence in order to point out the existing research gap.

The structure of Chapter 2 (figure 2.1) is the following: The chapter begins with an overview of the publications that define the concept of Big Data to carve out the characterizing dimensions - beginning with the high-level characterization by Laney [2001] (section 2.1). Based on the discussion of existing definitions and meta-studies in section 2.2, in section 2.3, a quantitative literature review approach is used to validate and enrich the identified dimensions. The subsequent mapping of the identified topics to a generic data analysis process (section 2.4) helps to identify those areas, that are the focus areas and white spots in the current research.
Section 2.5 is used to describe overlaps and differences of Big Data and Business Intelligence.
Besides the afore mentioned hype surrounding Big Data, a critical voice arises both from practitioners as well as the scientific community [Boyd and Crawford, 2011]. These critical thoughts are not limited to the legal-related aspect of data security but touch ethical and political aspects as well. Therefore, the chapter ends with a critical perspective on Big Data (2.6).

2.1 Big Data - Volume, variety, velocity: a first characterization

As mentioned above, a continuous increase in the number of publications that address Big Data can be found consistently every year since the early 2000s in scientific databases, such as *Scopus*, culminating in a sharp rise in scientific publications in 2011. Within the existing publications, no common understanding of the notion of Big Data

exists [Madden, 2012]. One characterization approach that has found its way in numerous publications is the one by Laney [2001] which will be used as a first approach to a characterization of Big Data.

This concept inherits an outlining using the so called V's. Despite its lack of a scientific background it will be presented due to its high popularity to provide a comprehensive view.[1]

The initial description using the V's contained three V's in 2001 [Laney, 2001] representing *Volume* (Increasing amount of data available), *Velocity* (Speed of new data generation) and *Variety* (Heterogeneity of available data regarding degree of structure and sources). While Laney did not use the term *Big Data* directly, although he claimed to do so later on, the initial statement was referring instead to the increasing role of data management in the field of e-commerce [Maier, 2013].

In the course of time, these 3 V's have been complemented by *Veracity* (Veracity of the data from different sources) and *Value* (the value which results from the analysis of the data). With regard to its dispersion, these five V's will be taken as a first approach towards providing an initial insight into Big Data and enriched with recent research in the respective field.[2]

Volume

The continuous increase of the available data can be described by a cycle of data generation. Technical advancements such as compression abilities [Armbrust et al., 2010], in combination with decreasing hardware prices, especially the price per stored Megabyte/Gigabyte, facilitate data driven business models and services such as platforms for blogging, social networks and e-commerce [Hilbert and López, 2011]. This results in an increase in the existing data volume. At the same time, the data volume facilitates the development of further data driven business models. The challenges resulting from this influx of data foster further technical advancements, as several Big Data relevant innovations originate from companies faced with massive amounts of data, e.g. the MapReduce framework developed by Google [Lämmel, 2008].

[1] The work by Laney [2001] has not been published in a scientific journal. Nonetheless, it has been cited more than 230 times at the time the work was accessed for this dissertation via Google Scholar.

[2] Besides the 3-5 popular V's, further articles exist, discussing up to 7 V's [van Rijmenam, 2013]. Those are not taken further into account in this research as the V's are taken rather as a first approximation towards Big Data instead of a whole characterization.

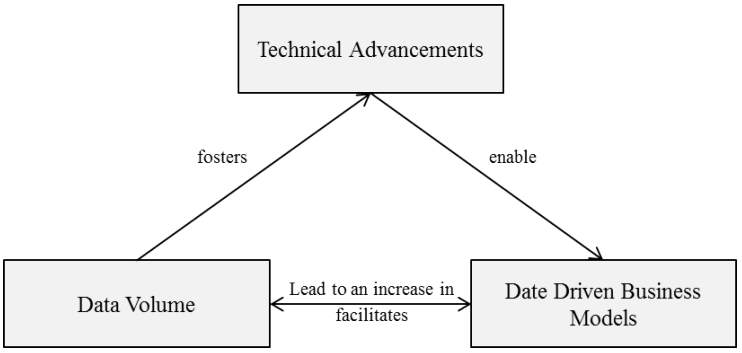


FIGURE 2.2: The cycle of data generation describes the drivers and consequences of the increasing data volume.

Although high volumes of data can be increasingly found in more industries, the information and communication industry has been one of the first industries which has been faced with such a high volume of data. Popular examples of companies dealing with this volume of data are Twitter (more than 12 TB of Tweets per day) or the IT service provider Cisco with a global IP traffic of more than 667 exabytes in 2013 [Kaisler et al., 2013].

Variety

Besides *volume*, the next of the original three V's is *variety*. The variety of data in the Big Data context has its origin in the diversity of available and accessed sources. The type of data in focus of the Big Data discussion can be commonly differentiated by the degree of structure, resulting in a distinction in *structured* and *unstructured* data [Batini et al., 2009].

The notion *Structured Data* refers to those items, which are described by elementary attributes, belonging to a domain. Those attributes are associated with a range of values, mostly statistical data and relational tables. In the context of Big Data, one popular example of structured data is sensor data, generated in the manufacturing environment. Following the elaborations of Batini et al. [2009], unstructured data " [...] *is a generic sequence of symbols, typically coded in natural language.*" With regard to the increasing popularity of social platforms and video platforms, videos and pictures can be seen as a relevant example of unstructured data as well. Currently, no scientifically based estimation for the share of unstructured data in the overall generated data exists; the existing estimations are originated from a practical background and range between 37 % and 90 % [Ziegler and Dittrich, 2007; Grimes, 2008].

Data structure in general can be defined on two levels. The first level targets an database entry (e.g. an extract from Twitter, consisting both of structured data (data and time of the tweet) as well as unstructured data (the tweet, consisting of text and/or a picture). The second level targets the structure of an individual attribute of such a database entry, which can be either structured or unstructured. This results in a two-level data variety, which can lead to data inconsistency and semantic problems [Helland, 2011; Zhang, 2013].

Though tweets from a data source such as Twitter can be transferred in a structure (level one), the understanding of unstructured data in the Big Data discussion is associated with texts and pictures.

Velocity

The aspect of data *volume* and *variety* is accompanied by the speed of new data generation (*velocity*). The dynamic of this growth rate can be shown using the studies published by the IDC, dealing with the progress of digitalization [Gantz et al., 2007]. A sixfold growth of data in four years from 2006 to 2010 was estimated, along with an annual growth rate of 57 % [Gantz et al., 2007]. For the period from 2013 to 2020, a yearly growth rate of 25 % is expected [Turner et al., 2014].[3] Although sensors in production environments are named frequently in this context, the growth drivers are not limited to the manufacturing industry. Wal-Mart is a popular example of a company that faced high data volume and growth, collecting more than 2.5 petabytes per hour, that consist of customer transactions [McAfee and Brynjolfsson, 2012]. From an application-oriented point of view, velocity is targeting both the speed of the data generation as well as the speed of data analysis connected with those data streams [Agrawal et al., 2012].

Comparable to *volume*, scholars are faced with the challenge of velocity as well. One example is the particle accelerator CERN, generating 30 petabytes per year during the different research projects [CERN, 2014].

Although the speed of data generation is currently associated with the hype of Big Data, it goes along with the already existing statements by Moore's regarding growth, specifically his statements concerning the number of Transistors per circuit board that double in a specific amount of time [Schaller, 1997]. Therefore, the named yearly growth rate of data is not higher than expected if the underlying logic of Moore's law is transferred

[3]The forecast of future data volume relies on a set of estimations. Therefore, they can only be taken as an approximation, whose extent can differ amongst different publications and are supposed to only give an idea about the yearly data generation.

to the field of data generation. Instead, one distinctiveness of velocity is the *variety* of the generated data.

Altogether, the changes in *volume, variety,* and *velocity* of generated data lead to multiple demands regarding the infrastructure and methodology for data handling, preparation, and analysis. Based on the described heterogeneity of the data pool with respect to the degree and type of structure, databases that are using a relational schema are mostly not suitable to deal with unstructured data. This accounts especially for the storing and processing of network structures [Stonebraker et al., 2013]. These changing requirements have fostered the Not Only SQL (NoSQL) database movement. NoSQL databases are based on a data scheme that is not necessarily related to the relational scheme, known from SQL databases, and are therefore not able to process data of different structures, e.g. network data or texts.
Additional to the demands with regard to the underlying IT infrastructure, the preprocessing as a preliminary stage of data analysis differs from previously known reporting oriented applications as well. The current focus on data cleaning (the removal of extreme or NULL values, correcting or remove incorrect values, correcting data inconsistency) has been broadened. Unstructured data have to be transferred into a structure that is suitable for further text mining analysis. Although these approaches have been improved within the past years, the pre-processing of human, intentionally generated data, which contain opinions and moods, e.g. customer reviews, is still a fault-prone task [Kaisler et al., 2013].

Veracity

Another aspect of Big Data is aiming at the lack of veracity of the data in focus. This accounts particularly for unstructured, human generated data from company-external sources. The aspect of data veracity is related to the data individual level as well as to the methodological level.
The first one targets the intention and background of users' textual contributions, e.g. in social communities or product review platforms. The human characteristics of self-manifestation, striving for attention, and the will to please may lead to statements that simply do not necessarily represent the actual opinion or sentiment and therefore can distort the analysis [Forestier et al., 2012] [Boyd and Crawford, 2011]. Additionally, spam bots can generate tweets and the like, whose content can falsify the analysis' results of

the corpus [Zikipoulos et al., 2013, 14]. With regard to the high volume of disputable data, whose analysis can be valuable as well, early publications can be found trying to automate the validation of customer profile data [Park et al., 2012]. Additionally, in multiple fields such as product reviews or customer forums, parts of the contributions are intentionally written in order to create a certain image of a product or heighten company reputation. Regarding the potential economic impact, the rule-based identification of those has become its own research field in the past years [Mukherjee et al., 2012].

Veracity in an methodological perspective targets the error-proneness of processing unstructured data, particularly text. Rhetorical figures such as irony or sarcasm complicate the computational linguistics, as well as the use of slang and typos does. Although the research on text processing has gained momentum in the past years, several challenges remain difficult, e.g. suitable stemming approaches and noise identification and reduction [Stavrianou et al., 2007].

Value

Within the field of Business Intelligence, numerous publications can be found about the Business Intelligence value chain, starting from the initial business problem, to the final improved decision making, resulting in the generation of value [Lönnqvist and Pirttimäki, 2006; Brohman and Parent, 2000].

In contrast, less has been written so far on data value in the general Big Data context despite its relevance for investment decisions in areas such as infrastructure or know-how. Therefore, a selection of aspects which influence the value of data in a Big Data context is described (table 2.1).

Although data analysis is recognized as a value-generating topic, its value calculation remains challenging [LaValle et al., 2011]. As described in the last subsection, the heterogeneity of external data sources regarding the access possibilities and the contained texts, messages etc. leads to the aspect of veracity. Amongst others, one primary source of noisy data are social networks, as they can be found in a wide range e.g. in internet stock messages boards [Antweiler and Frank, 2005]. With the increase in relevance of Twitter for trend analysis and sentiment detection, the *original state of data* and the subsequent identification and filtering of noisy data has become a topic of interest as well [Agarwal et al., 2011] [Barbosa and Feng, 2010] [Choi et al., 2012].

Besides noisy data in terms of content, data cleaning is also of interest due to the idiosyncratic writing style that can be often found in social networks [Derczynski et al.,

TABLE 2.1: Factors influencing the value of data

Factor	Description
Original state of data	Degree of structure and share of noisy data influences time and effort needed for data preprocessing [Shankaranarayanan and Cai, 2006]
Operationalizability and sustainability of analysis results	Degree of automatized utilization for managerial decision making and daily business of analysis results [Manyika et al., 2011]
Combination with other datasets	Most of today's best practices in Big Data analysis gain in value because of the gathering and consolidation of datasets from different sources [Mayer-Schönberger and Cukier, 2013, 102-110]
Position in the value chain	The number of application fields of analytic solutions differs between the respective position of the company in the value chain [Manyika et al., 2011]
Accessibility of data	The value of data can decrease with the number of competitors which have access to this data as well
Visualization	Quality of result visualization influences quality of related managerial decision making [LaValle et al., 2011]

2013]. However, one of the drawbacks of unstructured, noisy data is the resulting needed effort for data cleaning and pre-processing, which is time-consuming due to the daily high volume of tweets and messages.

In addition to the need for pre-processing, the aspect of process management gains relevance equally. The degree of integration of analysis processes/results in existing business processes, targeting the *operationalization* and development of a data driven organization, has an influence on the value of data as well [Kiron and Shockley, 2011] [LaValle et al., 2011]. This integration is connected with organizational change towards a data driven organization [Brynjolfsson et al., 2011].

The increasing number of data sources offers the enrichment of existing sources by *combining existing data sets with further data*, such as the customer database by external data, e.g. customer data from social networks or product review pages. The quality of the data matching process is critical for the latter analysis results.

The value of data sources and datasets yet differs between the *company's position on the value chain*. Companies that are closer to the end customer tend to benefit in a first step more from customer/product-centric data such as product reviews. Taking data from blogs about products as an example, e-commerce companies selling this product can benefit more from these data compared to an investment goods company, responsible for producing machines for the fabrication of the product. Consequently, companies closer to the end customer are mostly more experienced in analyzing customer data [Ngai et al., 2009].

The *accessibility of data* and their results in the Big Data context is a more crucial aspect comparable with reporting-oriented Business Intelligence systems, that are primarily based on company-internal data [Lahrmann et al., 2010]. At present, the exposure and access to company-external data and the resulting insights, e.g. market trends, have a decreasing value for a company with an increasing number of competitors using the same data. Consequently, the identification of less popular but meaningful data sources becomes more relevant.

The influence of *visualization* gains in relevance with the increasing heterogeneity of data sources and characteristics [LaValle et al., 2011]. This is reflected by the publications on visualizing unstructured data, e.g. movement profiles [Andrienko and Andrienko, 2012] [Ferreira et al., 2013].

Altogether, although the approach based on several V's is not suitable for the definition of Big Data in this research due to the lack of a scientific foundation and the sole focus on the characteristics of the processed data, it can be valuable in obtaining an initial understanding of Big Data. In order to develop an understanding of Big Data in the scientific community, existing characterizations from relevant publications will be analyzed in the next section.

2.2 Characterizing dimensions of Big Data

The concept of Big Data contains numerous different aspects and no common understanding exists so far as aforementioned in the previous subsection [Madden, 2012].

One approach to create an understanding of the subject in focus is the identification of its describing dimensions. The dimensions are used to structure the subject of Big Data

into different subject areas. Therefore, the goal of the following section is to carve out describing dimensions of Big Data and related topics within the dimensions. In doing so, a three-step approach for a structured literature review is pursued.[4]

1. Existing *meta-studies* in the field of Big Data are analyzed regarding the understanding of Big Data.
2. Based on *existing definitions* of Big Data, describing dimensions are derived.
3. A quantitative literature analysis based on relevant publications surrounding Big Data is carried out in order to validate and enrich the dimensions identified in step two with relevant topics.

Basis for all of the three steps is a structurally identified corpus of Big Data literature. Its underlying process for the identification of relevant Big Data publications as a necessary first step (figure 2.3), will be described subsequently.
As an emerging research field, Big Data generates numerous publications which can be retrieved e.g. via Google Scholar. This search engine can be used for an initial search but not every resulting publication is evaluated in any ranking of scientific publications, and the contribution of their quality cannot be evaluated properly. Therefore, the databases in focus are *IEEE, Scopus, ISI Web of Knowledge, EBSCO, ACM*, and *Springer* to ensure a minimum quality level. For the analysis, papers from scientific journals and conference proceedings from the computer science field, that are published since the year 2000 till 2013 and contain the notion "Big Data" in the title, abstract, or keywords, have been searched, leading to an initial database of 1,322 publications (step 1-3 in figure 2.3).[5] [6]

[4] The following sections about the characteristics of Big Data are based on Hansmann and Niemeyer [2014].
[5] This extended time period has been selected in order to ensure the identification of all publications in the Big Data context since the early beginning, marked by the description of Laney [2001].
[6] The year 2013 as the end date is a consequence as the related publication (Hansmann and Niemeyer [2014]) was the first part of this research, done in the beginning of the PhD studies.

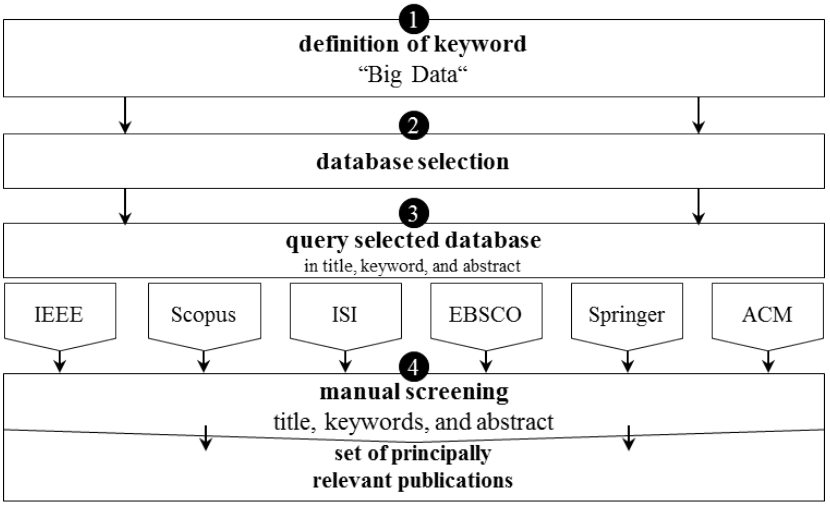


FIGURE 2.3: Literature review process used for the identification of Big Data relevant research publications

The resulting publications have been scanned manually, and papers were removed if they i) belong to conference workshops, ii) are keynote-related paper editorials, or iii) whose content did not belong to the field of Big Data as understood in this research (step 4).[7] Furthermore, duplicates were removed. This selection process resulted in a database of principally relevant publications comprising of 248 documents.

In the beginning of the literature review process, the identified publications have been searched for contained meta studies on Big Data, leading to three publications, Chen et al. [2012]; Pospiech and Felden [2012]; Ward and Barker [2013], which will be presented subsequently.

One of the first *meta studies* of Big Data research has divided the development towards Big Data developments into three phases: *Business Intelligence and Analytics (BI & A) 1.0 - 3.0*, describing key characteristics and relevant topics for each level [Chen et al., 2012]. Following this structure, BI& A 1.0 describes the adoption of *"technologies and applications"* for the collection and processing of primarily structured data from existing in-house systems as Enterprise Resource Planning (ERP) or Customer Relationship Management (CRM), using mostly relational database management systems (RDBMS). The data analysis is grounded on statistical methods from the past three decades. Interestingly, Chen et al. [2012] describe this lowest level as the current industry standard, which points out the existing potential and novelty of Big Data.

[7] Examples are publications, in which the notion *Big Data* appears, but solely to describe a data volume, not using the notion to describe Big Data as an individual topic as understood in this research.

BI & A 2.0 is related to the increasing development of web-based businesses such as e-commerce or social networks. Additionally to the RDBMS from level BI & A 1.0, infrastructure which is capable of storing and processing both unstructured data, as texts and pictures, and high velocity data, as cookie tracking data, has gained admiration.
BI & A 3.0 then describes the increasing relevance of mobile devices such as smart phones or tablets for the analysis and the development of individual user profiles as a basis for the customized offering of services and products.
Pospiech and Felden [2012] review the current literature on Big Data, clustering publications among these dimensions of *data provision* and *data utilization*, combining each with a technical or functional perspective, originating from Gluchowski [2001]. Based on their review, they reveal a focus in current Big Data research on the technical perspective of data provisioning (87 percent of 46 publications). The main topics are dealing with infrastructure architectures, targeting i) the storage of high volumes of data and ii) the performance of data processing. Furthermore, the review reveals a subordinated consideration of the functional data utilization and therefore names the identification of use cases as a recent research gap.
Ward and Barker [2013] in contrast compare definitions of Big Data that are information technology industry driven. Although most of the definitions are related to the product portfolio of the respective company, they contain at least one out of the characteristics of *size* (related to the data volume), *complexity* (related to data variety) and *technology*, which targets the applied infrastructure.
Summing up, existing meta-studies do not come up with an identical characterization of Big Data. Nonetheless, it becomes clear that Big Data is not limited solely to the increasing volume of available data.

Besides these described meta-studies, a number of Big Data characterizations exist that are dedicated to give a more distinct definition of Big Data. These will be analyzed in the next step using a deductive approach. For this step, within the identified 248 documents, characterizations have been identified. Furthermore, a backward search for potential further relevant literature has been carried out, focusing on cited publications, which contain characterizations of Big Data. This process resulted in the identification of the definitions presented in table 2.2.
Based on these identified definitions, dimensions that characterize Big Data have been derived. In doing so, in a first step, similar describing words and phrases, e.g. the words

storage, technology and *database* have been grouped together manually.[8] Afterwards, umbrella terms for each word cluster have been defined, the later dimension name. The identified definitions which will be described subsequently.

TABLE 2.2: List of Big Data characterizations from a research background

Author	Definition	Dimensions
Bizer et al. [2011]	The exploding world of Big Data poses, more than ever, two challenge classes: engineering - efficiently managing data at unimaginable scale; and semantics – finding and meaningfully combining information that is relevant to your concern. (...) In this Big Data World information is unbelievably large in scale, scope, distribution, heterogeneity, and supporting technologies.	Data characteristics
Boyd and Crawford [2012]	We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of: (1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets. (2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims. (3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.	Data characteristics, IT infrastructure, Methods

[8] The manual grouping will be validated in section 2.3 with the automated analysis of the publications using a text mining approach.

TABLE 2.2: List of Big Data characterizations from a research background

Author	Definition	Dimensions
Chen et al. [2012]	(...) data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies.	Data characteristics, IT infrastructure, Methods
Cuzzocrea et al. [2011]	"Big Data" refers to enormous amounts of unstructured data produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from e-government applications to medical information systems, and so forth.	Data characteristics
Diebold [2003]	Recently much good science, weather physical, biological, or social, has been forced to confront - and has often benefited from - the "Big Data" phenomenon. Big data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advantages in data recording and in storage technology.	Data characteristics, IT infrastructure
Jacobs [2009]	Data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time.	Data characteristics, Methods
Madden [2012]	Data that's too big, too fast, or too hard for existing tools to process.	Data characteristics, IT infrastructure

TABLE 2.2: List of Big Data characterizations from a research background

Author	Definition	Dimensions
Manyika et al. [2011]	Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse.	Data characteristics, IT infrastructure, Methods
Ward and Barker [2013]	Big Data is a terminus describing the storage and analysis of large and / or complex data sets using a series of techniques including, but not limited to NoSQL, MapReduce and machine learning.	Data characteristics, IT infrastructure, Methods

The definition by Cuzzocrea et al. [2011] is aimed at the characteristics of the generated data, containing both the amount and structure of the data, complemented with naming exemplary data sources. Bizer et al. [2011] enrich the data characteristics with additional attributes, such as the scope, target, and structure of the data, addressing data heterogeneity in a "Big Data world".

With regard to the data characteristics, Jacobs [2009] focuses solely on the amount of data and adds the aspect of *method*, without giving further details.

Chen et al. [2012] include the aspect of method in terms of analysis as well, and add *IT infrastructure* topics, such as storage and processing purposes. Furthermore, their definition enhances the dimension data characteristics by naming a selection of data sources.

The definition by Madden [2012] incorporates both data characteristics and infrastructure (tools), which is extended by Manyika et al. [2011] with the aspect of method. Both definitions, along with that of Jacobs [2009], emphasize the excessive demand of the current IT infrastructure to handle the changes in the data characteristics.

Those descriptions [Madden, 2012; Manyika et al., 2011; Jacobs, 2009] contrast one of the early definitions from 2003 by Diebold [2003], who states that the availability of the enormous amount of data is a result of the "advantages in recording and storage technology". This suggests a change in the requirements regarding the IT infrastructure,

corresponding with the description of the data generation cycle in figure 2.2.
The definition of Ward and Barker [2013] results from the analysis of existing definitions of Big Data with an industry background, identifying the recurring characteristics of volume and the complexity of the datasets and the technologies; all used for data processing as critical aspects.
The definition by Boyd and Crawford [2012], finally reflects a critical perspective towards Big Data by including exclusively the aspect of mythology, targeting the high expectations regarding data analysis.

In summary, based on the review of existing definitions, three main dimensions of Big Data can be derived within the presented definitions in table 2.2. The named aspects of data characteristics (amount and structure) and sources can be merged into a *Data dimension*. The tools and databases that are required to store and manage data can be combined to an *IT infrastructure dimension*. The data processing for analysis purposes can be embraced into a *Method dimension*. The latter two dimensions are similar to the analysis by Pospiech and Felden [2012].

In order to incorporate also the relevance of Big Data for practitioners, the generated results based on definitions with a scientific background are compared with industry-oriented definitions, shown in table 2.3. The analyzed publications have been identified in a two-step approach. In a first step, the publications of large technology providers (Microsoft, IBM, SAS etc.), offering Big Data applications have been screened for their contained definitions. Second, publications from the three main market research companies in the field of IT and Digitalization (IDC, O'Reilly, Forrester) have been analzyed as well for contained definitions. Again, dimensions are derived from the presented definitions, following the same approach as before.

TABLE 2.3: List of Big Data characterizations from an industry background

Author	Definition	Dimensions
Hopkins [2011]	Big data: techniques and technologies that make handling data at extreme scale economical.	Data characteristics, IT infrastructure

TABLE 2.3: List of Big Data characterizations from an industry background

Author	Definition	Dimensions
Gartner [2015]	Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.	Data characteristics, IT infrastructure, Method
IBM [2011]	Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.	Data characteristics
Carter [2011]	(..) a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis.	Data characteristics, IT infrastructure
Microsoft [2013]	Big data is the term increasingly used to describe the process of applying serious computing power – the latest in machine learning and artificial intelligence – to seriously massive and often highly complex sets of information.	Data characteristics, IT infrastructure, Methods

TABLE 2.3: List of Big Data characterizations from an industry background

Author	Definition	Dimensions
Dumbill [2012]	Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.	Data characteristics, IT infrastructure
SAS [2015a]	Big data is a popular term used to describe the exponential growth, availability and use of information, both structured and unstructured. Ultimately, regardless of the factors involved, we believe that the term big data is relative; it applies (per Gartner's assessment) whenever an organization's ability to handle, store and analyze data exceeds its current capacity.	Data characteristics, IT infrastructure, Methods

The comparison of the definitions from science and practice reveals overlapping contents, indicating a similar basic understanding of Big Data in science and practice. Yet differences can be found in the subordinated consideration of the methodological aspect in the industry driven definitions; most of the IT service providers offer infrastructure related services and considered the data analysis subordinately [Ward and Barker, 2013].

The science-oriented results of this deductive approach - the derived dimensions - are continued to be used as a basis for the further exploration of Big Data in this work.

2.3 Validation and enrichment of the Big Data dimensions using Topic Models

The relevant dimensions of Big Data in terms of *Data, IT infrastructure*, and *Methods* for analysis purposes have been derived in section 2.2 by using a deductive approach. Due to the relevance of knowledge about dimensions and contained topics for the latter development of the maturity model, a comprehensive understanding of Big Data is success-relevant. Additionally, the object in focus, Big Data, is characterized amongst others by its remarkable breadth and depth. Contrary to other emerging fields in computer science in the past, such as Cloud Computing, no common understanding of Big Data exists so far. Therefore, a sole manual analysis of existing definitions as carried out in section 2.2 would not be sufficient, the dimensions have to be validated.

Word clouds as an alternative method for identifying topics in a corpus, focusing solely

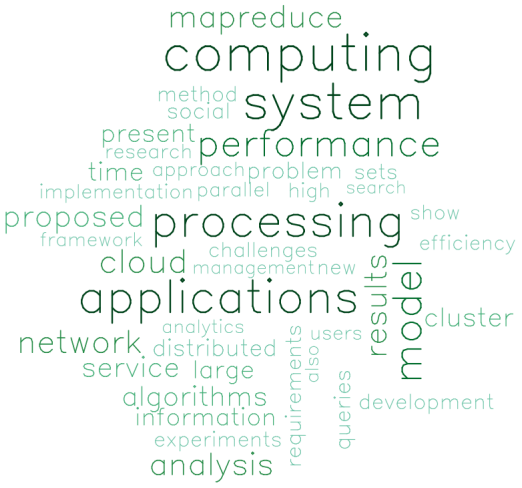


FIGURE 2.4: Tag cloud generated based on the words from the abstracts of the Big Data publications

on the counting of words in order to carve out describing topics, based on Big Data relevant publications would lead to difficulty in interpreting list of words with a decreasing frequency as shown in figure 2.4[9]

[9]The word cloud has been created based on the same corpus as it has been used for the topic model analysis, explained in the next step, using the word cloud package in R [Fellows, 2014]. The more frequent a word appears in a text, the bigger the font size.

Although the word cloud reveals for instance a focus on infrastructure, topics with words such as *system* reduce the information value compared with the topic model approach as they are not set in the context with other words belonging to the same topic. Consequently, word clouds appear rather on online platforms as *flickr* than in scientific literature [Sinclair and Cardew-Hall, 2007] and will thus not be pursued any further.
Alternatively, to validate the afore identified dimensions, the previously described structured literature review following Webster and Watson [2002] is now applied in a second cycle, enhanced by a methodological component - a text mining approach from the field of machine learning. This method is used to *validate* the dimensions carved out from existing definitions and to *enrich* these dimensions with contained topics. Following a two-step approach, first a text mining method is applied on the abstracts of the 248 identified publications, second, the dimensions are enriched by applying the text mining approach on the abstracts of the dimension-specific publications.

Different text mining approaches for the identification of topics in texts exists, four of the most popular ones are i) latent semantic analysis, ii) probabilistic latent semantic analysis, iii) latent Dirichlet allocation, and iv) correlated topic models [Lee et al., 2010]; each of them addressing the weaknesses of its preceding approach. The umbrella term for these approaches is *Topic Models*, and will be described in the following section and applied afterwards.

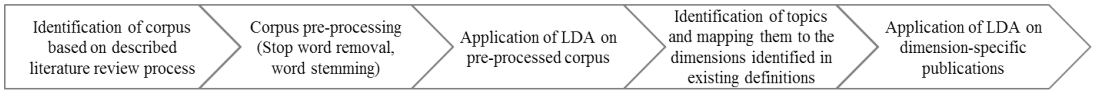


FIGURE 2.5: The validation and enrichment of the identified dimensions of Big Data by applying Topic Models consists of several steps.

2.3.1 Topic Models - Methodological foundation

Topic models are hierarchical probabilistic models that have their origin in the field of machine learning and have been broadly applied, especially in the field of literature analysis [Titov and McDonald, 2008].
Topic models are based on the *generative approach*. The result of the generative process is a document. A document is viewed as a mixture of topics; thus, a document - in

this case the abstract of a Big Data relevant publication - can be represented with a probability distribution over the topics.

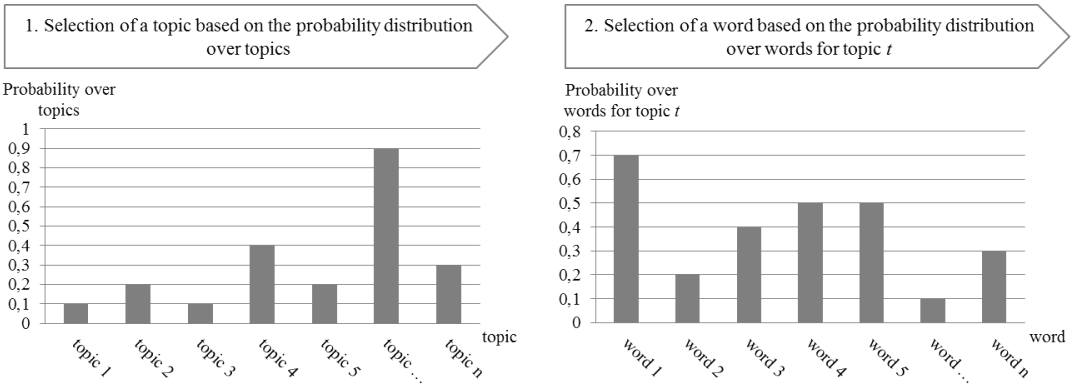


FIGURE 2.6: The Generative Process allows the generation of a document based on the two probability distributions.

A topic is defined over the appearance of certain words together; therefore, individual topics can be represented as probability distributions over words.
Based on that, it is assumed that each document can be generated (therefore *generative approach*) based on the two probability distributions:

1. Based on the probability distribution over topics, a topic is randomly selected.
2. The probability distribution over words, belonging to the selected topic, is taken and a word is selected.

These two process steps can be carried out repeatedly until a document of the favored length is generated.

When applying topic models on unstructured data (text, pictures, videos) [Wang and Mori, 2009], this generative process is reversed to estimate both of the distributions based on the input data - the text which contained topics are supposed to be extracted - with the help of a machine learning technique.
In the following analysis, the focus is on the abstracts of publications related to Big Data. The individual documents are merged into a corpus that is the input for the analysis. Among the different estimation approaches, Latent Dirichlet Allocation (LDA) [Blei et al., 2003] has been successfully applied for similar purposes in terms of literature analysis [Griffiths and Steyvers, 2004]. As it has been proofed as a suitable method, this approach is used in the thesis as well [Wang and Mori, 2009]. This quantitative analysis

represents the enhancement of the classical literature review by Webster and Watson [2002].

Following this approach, in a first step, the two a-priori probability distributions i) over topics and ii) words per topic are defined. In a next step, the distributions are fitted based on the analyzed corpus with the help of LDA. The corpus consists of the abstracts of all publications identified in the literature review, described in the section before (figure 2.3). The subsequent model estimation is carried out based on Collapsed Gibbs sampling [Griffiths and Steyvers, 2004].

Topic model results are represented by lists of the most probable words for each topic. Therefore the interpretation of the topics is at risk of subjectivity. Chang et al. [2009] stated that, "[...] *no quantitative evaluation of the results of topic models exists*". Following the authors' opinions, the existing metrics to evaluate the model fit in terms of the Maximum Likelihood Optimization do not target the explanatory character of the topic models.
Therefore, with regard to the explanatory use of the topic models in this research, the application of topic models is supplemented by the utilization of the word intrusion approach [Chang et al., 2009]. After carrying out the topic model analysis on the corpus in focus, the list with the most probable words (in this case, five to ten) for one topic is complemented by a word that has a low probability for the topic in focus and a high probability for another identified topic. In case the determined topic words make sense together, the identification of the intruder word should be easy for the test taker. For the test arrangement, the identified topics, resulting from the application of the topic models have been checked for the intrusion words by eight scholars from the field of information systems.[10] The results of the word intrusion are measured as model precision (mp). The mp is calculated as follows: number of correct identified intruders divided by the total number of test takers. The results can be found in the respective sections.

[10]The topic lists, manipulated following the idea of the described intrusion approach by adding one intruder word per dimension, have been given to the scholars. A minimum knowledge about information systems was needed in order to allow an interpretation of the identified topics. They have been asked to mark within each topic the word, which they perceive as not fitting to the others words. This process has been carried out individually with each scholar without a discussion between the scholars.

2.3.2 Data selection and preprocessing

Starting point of the application of the topics models is the same literature review process as described in section 2.2.

The subject of analysis are the abstracts of these remaining 248 publications instead of the overall text. It is assumed that the abstract is supposed to hold the key topics of the publications.

The abstracts of these resulting papers have been pre-processed in terms of removing stop words[11] and the words "big" and "data", in order to prevent the subject of the study from becoming part of the analyzed corpus, which would therefore falsify the results due to a high word frequencies in the abstracts. Furthermore, word stemming has been executed via Porter's stemming algorithm [Porter, 1980]. The fitted text corpus has been analyzed using topic models, for which the results will be explained in the following section.[12]

2.3.3 Analysis on the overall database using Topic Models

As shown in table 2.4, the Big Data dimensions that have been derived based on the presented definitions can partly be validated and enriched within scientific publications on Big Data.[13] This result is reflected in the high model precision (mp = 0.76).[14] [15]

In a next step, the words of each topic are discussed and set into context to each other and the potential overarching topic respective dimension.

The words of Topic 1 can be assigned to the IT infrastructure dimension. Particularly, the words *mapreduce* and *hadoop* account directly for the aspect of technologies within the concept of Big Data. The programming framework MapReduce, which was developed by Google, and its open source implementation, Hadoop, have been designed to *process* voluminous data. Both techniques contribute to the rise of distributed, scalable systems within the development of Big Data applications [Dean and Ghemawat, 2008]. The relevance of MapReduce and Hadoop explains the word *parallel* as analysis tasks

[11] Stopwords are connection words like "and" or "then". They have been removed based on the stopword list contained in the R package *tm* [Feinerer and Hornik, 2015]

[12] The analysis has been implemented using the package *lda* in R [Chang, 2015].

[13] If a field is empty, the probability of the word occurring did not differ significantly from the following words, meaning that they are not representative for this topic. The words are listed depending based on the calculated probability for the respective topic in a decreasing order.

[14] The details of the model precision can be found in section 2.3.1.

[15] The displayed number of topics has been determined according to the Harmonic mean method in consideration of the low number of analyzed abstracts [Griffiths and Steyvers, 2004].

TABLE 2.4: Results of the Topic Model application on the overall corpus

Topic 1	Topic 2	Topic 3
mapreduce	algorithm	research
performance	method	search
processing	graph	information
system	experiment	analysis
computing	problem	social
parallel	accuracy	computing
efficiency	parameter	-
cloud	approximate	-
queries	-	-
hadoop	-	-

that can be *computed parallel*, the core of the MapReduce framework. The appearance of *performance* is based on publications, that focus on the performance improvement of a Hadoop cluster for certain analysis purposes [Gu and Gao, 2012] or general performance improvements based on data locality [Hammoud and Sakr, 2011]. The aspect of *efficiency* is related to *performance*, and *queries* is connected with databases in general. Although the last four words cannot be assigned exclusively to Big Data, they are relevant in the MapReduce/Hadoop context. Similar accounts for the word *cloud*, which has gained in relevance in general within the computer science discipline, nonetheless has contributed to the development of Big Data as a scalable storage environment as well [Argawal et al., 2011; Ari et al., 2012; Assunção et al., 2013].

Topic 2 can be identified as the *method* dimension. In addition to the generic, method-related words *algorithm* and *method* itself, the word *graph* brings up a specific type of algorithms that are related to network analysis. Although graph theory is not solely connected with Big Data, it becomes increasingly relevant for social network analysis, which will be explained in further detail in the dimension-level analysis in the next subsection. The fourth word, *experiment* results from the experimental character of several publications in this dimension but hold a rather general character. The same accounts for the remaining words *problem, accuracy, parameter* and *approximate*.

The words of the third topic do not fit completely with the remaining data dimension;

they have a generic character of possible applications in the field of data analysis. The combination of *search, information, analysis, social*, and *computing* indicates the computation-based analysis of a social environment; however, the generic character of the contained words will be analyzed in the next section in more detail (section 2.3.4). Summing the first part of the Topic Model application for the validation of the dimensions up, the comparison of the results of the topic model analysis (database: all abstracts of the identified Big Data publications) and dimensions derived from existing definitions (database: all identified definitions) do have a high degree of overlap.

In a next step, the topic model is applied to a random set of publications from the field of information systems in order to find out how meaningful the results are for the field of Big Data. A randomly generated sample set of 105 publications (earliest publication year 2010) from the included databases used for the literature review process, both from proceedings and journals of the computer science field is analyzed. The results as shown in table 2.5 contain a cross section of computer science related topics on a generic level and therefore do not cover Big Data specific topics. The comparison with the Big Data topics shows that the results of the Big Data corpus in comparison are not only general computer science related topics. One aspect, which is for example missing in the Big Data related publication is Topic 1 from the validation topics. Following words as *programs* and *code*, this topic is focusing on *programming*, which is not represented in the Big Data topics at all. Aspects as MapReduce or Hadoop are missing.
Nonetheless, one has to keep in mind, that scientific disciplines are affected by hypes as well. The hypes within the community have an influence on the topics of publications [Mertens, 2006; Schauer and Schauer, 2009]. Therefore, the analysis of a randomly generated sample of publications do not lead to an all-embracing result, however can be taken as an indicator for general topics.

TABLE 2.5: Results of the Topic Model application on a randomly generated corpus

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
programs	process	algorithms	performance	data
code	management	queries	memories	mined
analysis	development	results	system	behavior
semantics	creation	complexity	caches	image
programming	governance	set	disk	digital
language	services	class	control	research
language	studies	-	monitoring	application
verification	knowledge	-	drivers	analysis
compiler	shared	-	power	techniques

In this section, the topic models have been applied on the overall corpus in order to validate the identified dimensions. In the next section, the topic models will be applied to dimension-specific publications in order to enrich those.

2.3.4 Analysis on the dimensional level

The dimensions described in section 2.2 have been used as a starting point for the following analysis. The identified publications of the overall corpus, resulting from the process described in figure 2.3 have been screened manually as a first step and were assigned to the *Data, IT infrastructure*, and *Method* dimension. This assignment has been carried out by two research assistants (including the author of this work) according to the approach by Webster and Watson [2002], adapted from Salipante et al. [1982]. In doing so, each publication has been assigned to one or more dimensions.

After a first manual run through the publications as a pre-step to the dimensional analysis with topic models, it became clear that i) the derived dimensions do not cover the

recent publications entirely.

Within the analyzed corpus, several papers existed with an focus on the utilization of specific Big Data technologies in an industry context with a subordinated consideration of technological or methodological aspects. The overarching category for those publications has been named *application*. The author chose this notion in accordance with the already established use in two of the analyzed definitions [Chen et al., 2012; Cuzzocrea et al., 2011].

Furthermore, ii) it became clear that an assignment of each paper to only one dimensions is not suitable due to the breadth of contained topics of the recent publications on Big Data. The results of the assignments can be found in table 2.6.

Following the results, recent publications have focused on the infrastructure aspect, followed by methods and applications. There were a total of twelve publications which have been identified manually, targeting data-relevant topics, which explains why this topic did not come up as a separate topic based on the application of the topic models.

TABLE 2.6: Number of publications per dimension after the manual assignment

Dimension	Number of publications
IT infrastructure	112
Method	99
Application	84
Data	12

After the publications have been pre-assigned to the individual dimensions based on a manual process, in the next step, the topic models are applied to the dimension-specific publications except for the data dimension due to its low number of related papers.[16] In the following section, the results are discussed with respect to the extent to which and how they account for the Big Data concept.

[16]The screening of the literature revealed that the topic of data relevant topics, e.g. data quality management, meta data management, data modelling etc. has not been subject yet to publications in the field of Big Data.

TABLE 2.7: Results of the Topic Model application on the publications belonging to the IT infrastructure dimension

Topic 1	Topic 2	Topic 3
cloud	queries	network
computing	database	social
cluster	stores	results
mapreduce	search	latency
processing	analysis	traffic
parallel	research	-
hadoop	index	-
distributed	processing	-
platform	prototype	-
-	framework	-

2.3.4.1 IT infrastructure dimension

The results of the topic model application on the 112 Infrastructure related publications (mp = 0.86) show a distinction between words related to hardware (Topic 1) and software (Topic 2) (table 2.7). *Cloud computing* plays a dominant role within the hardware topic. Although the words *cloud* and *computing* do not account solely for a Big Data infrastructure, the increasing amount of data led to a rise in the cloud applications, and vice versa; therefore, cloud computing, both as a driver and enabling technology, is relevant within a Big Data hardware topic [Argawal et al., 2011]. Furthermore, the *MapReduce* framework in combination with *Hadoop* cluster as a *platform* for the *distributed, parallel processing* of the data play a major role within the field of Big Data-related hardware. This dominance is emphasized as the words come up both in the analysis of the overall corpus as well as based on the dimension-specific corpus.

The words in Topic 2 are representatives for a software-oriented perspective on Big Data, which include *queries, processed* on *databases* or *stores* as subjects of data handling [Feng et al., 2012]. Furthermore, the word combination *search, indexing, analysis*, and *research* are an application/task-oriented perspective on the IT infrastructure. *Prototype* mainly refer to the developed *frameworks* for test scenarios.

In contrast to several words in Topic 1, the words in Topic 2 are not connected explicitly

with Big Data, as the contained words are general infrastructural topics. Topic 3 has lower information value than Topics 1 and 2. Besides the first two words with *network* and *social*, which fit into the named aspect of the (social) network analysis, no remaining words fit into a specific category. The aspect of network analysis will be discussed in the next section.

2.3.4.2 Method dimension

The results of the publications on methods (table 2.8) (mp = 0.71) offer two insights: i) *MapReduce/Hadoop* play a major role in the method-related publications (Topic 2), which is convincing because MapReduce is a methodological approach, Hadoop its implementation. Therefore, whereas in the IT infrastructure section, publications target the development of clusters, the focus in the methods dimension is on the fitting of the MapReduce algorithm to data characteristic-related requirements. One would not cope with the concept of Big Data if it were reduced on MapReduce/Hadoop, but with regard to the open source availability and comprehensive developing community and support, MapReduce/Hadoop have an outstanding position within the Big Data concept [McAfee and Brynjolfsson, 2012]. Furthermore, ii) the results of the methods dimension highlight the aspect of networks. This can be found in Topics 3 and 5, supplemented by the word *graphs*, which demonstrates an increasing relevance for the analysis of *social online networks* and named as a Big Data specific method Chen et al. [2012]. The *study* and *analysis* of *user* behavior targets the analysis of online behavior in social networks and platforms such as Twitter or Facebook, but still are general, research-related words. In contrast to Topics 2, 3, and 5, Topics 1 and 4 do not represent distinctly identifiable Topics. The words in Topic one are generic and do not allow to draw a conclusion to an underlying Topic. Topic 4 holds general methodological words as *algorithm* or *cluster* and point at machine learning based classification, which is not necessarily a Big Data specific topic. Therefore, Topic 1 and 4 have not been considered any further.

TABLE 2.8: Results of the Topic Model application on the publications belonging to the method dimension

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
signal	mapreduce	graph	time	rate
event	processing	search	machine	network
code	hadoop	studies	virtual	effective
resolution	implementation	research	results	service
parameters	computing	results	cluster	systems
-	distributed	users	algorithm	users
-	systems	online	mean	social
-	queries	prediction	classification	temporal
-	efficiency	network	complex	method
-	cloud	engines	method	analysis

2.3.4.3 Application dimension

The resulting topics in table 2.9 (mp = 0.64) cover a broad range of applications in the Big Data context but non distinct identifiable. Topic 1 partly targets the analysis of social networks, which corresponds to topics 3 and 5 from the methods dimension (table 2.8). This finding emphasizes the relevance of the network topic and its analysis within the Big Data context [Chen et al., 2012]. Topic 2 offers words such as *business, challenges*, and *market*, which are generic application-related words that are not distinctive for Big Data; therefore they do not contribute to a clarification of the application dimension. The same accounts for Topic 3. In contrast, Topic 4 represents words such as *storage* and *system*, which is an IT infrastructure aspect within the *application* dimension but again these three words are generic.

TABLE 2.9: Results of the Topic Model application on the publications belonging to the application dimension

Topic 1	Topic 2	Topic 3	Topic 4
social	business	search	model
emerging	challenges	information	cloud
internet	science	time	service
bias	classification	processing	storage
information	speech	results	application
research	research	user	requirements
platforms	-	text	systems
processing	-	emotional	evaluation
user	-	-	-
network	-	-	-

2.3.5 Discussion of the results

In the past sections, Topic Models have been used to enhance a literature review for an emerging topic in order to validate existing dimensions of Big Data and enrich those with topics of current research.

The describing dimensions, derived from existing definitions by applying a deductive approach, could partly be validated based on the analyzed abstracts. Especially the dimensions containing the IT infrastructure respective methodological topic could be retrieved, supported by the highest model precision Within publications related to this dimension, an emphasize in research on MapReduce and its Open Source implementation Hadoop can be found; partly owed to its spread throughout science and practice and its low acquisition costs [Garza et al., 2014].

As could have been expected, an application dimension did not result from the topic model based analysis of the Big Data publications as a consequence of the low number

of publications containing use cases and applications in a Big Data context. They were not associated with specific key words, that would allow an assignment of the related topics to the application dimension. This finding supports the identified research gap by Pospiech and Felden [2012]. Nonetheless, several publications with an application focus could be identified (table 2.6).

A data dimension could not be found in detail within the publications as well, although aspects such as data quality management are critical success factors for Big Data applications [Kwon et al., 2014]. One potential reason for that is simultaneously one limitation of this Topic model approach for the characterization of research fields. Its output depends on the occurrence of a specific term within the title, keywords or abstract. Consequently, publications, which might have a relevance for the topic, such as the role of data quality management of company external data and are related with the topic in focus, but do not contain the phrase *Big Data* will not appear in the corpus, and therefore find no consideration.

This problem might gain in relevance in the future, as the connotation of the phrase *Big Data* could become, for example, increasingly negative due to reasons such as its connection with the media and national surveillance programs as well as its extensive use for marketing purposes. These aspects could lead to a reluctant use by scientists. Furthermore, Big Data relevant aspects could be discussed in nearby fields such as Business Intelligence without being tagged with the keyword "Big Data," and therefore not end up in the analyzed corpus.

A second limitation results from the methodology itself. In the research at hand, the explanatory power of the Topic Models has decreased with an increasing homogeneity of the analyzed corpus. Therefore, the results are more significant for corpuses, which contain publications that cover a broader topic, instead of a narrow focus on one specific field of research. Consequently, this approach is suitable for a first characterization of a research field by delivering interpretable word lists, but can lead to fuzzy results on a more homogeneous corpus in sub-domains.

In the next section, the results will be classified into a generic data analytics process to identify potential blind spots within the Big Data-related research.

2.4 Classification of the results into a generic data analysis process model

By using topic models to analyze the publications on a dimensional level, the dimensions could be enriched by the identification of subtopics in the last sections. Up to this point, an understanding of Big Data was able to be developed, but a more detailed overview about current emphasizes of research is missing. One option presented in this research in order to reveal current emphasizes and blind spots, is to transfer the identified topics into a generic data analysis process, using the individual processes as a second structuring dimension. Additionally, this overview helps to identify potential relevant aspects for the later maturity model development, as the filling of white spots fosters the contribution to research and increases the relevance of the work.

Two popular generic analysis models are the model for *Knowledge Discovery in Databases* (KDD) by Fayyad et al. [1996] and the Cross Industry Standard Process for Data Mining (CRISP-DM) [Shearer et al., 2000]. The CRISP-DM has a business focus which results in phases as *Business Understanding* and *Data Understanding*, that in turn are not a distinct field of research. In contrast, the KDD process steps are on a more generic level and are closer to the topics in research.

Consequently, the model by [Fayyad et al., 1996] has been selected as a basis. It has been augmented by the step *Result visualization* to allow for the increasing complexity of the analysis presentation in contrast to the presentation of reporting-oriented analysis [Liu et al., 2013]. The original step *Transformation* has been included into the *Preprocessing* step, as those two steps hold significant overlaps in the context of Big Data [Assunção et al., 2015]. This fitted model has been augmented with related, exemplary publications from the analyzed corpus, the same from the topic model application as well (figure 2.7).[17] The selection and assignment of the publications has been carried out manually.

Step 1, *Data selection*, is not covered in publications within the analyzed corpus, although the selection of adequate data sources has a major influence on the latter analysis results. This aspect is covered so far only in publications from the field of information requirement analysis [Byrd et al., 1992; Hansmann and Nottorf, 2015].

[17] The exemplary selection represents the main topics within the scanned literature. A assignment of each individual publication is not practicable with regard to the number of publications.

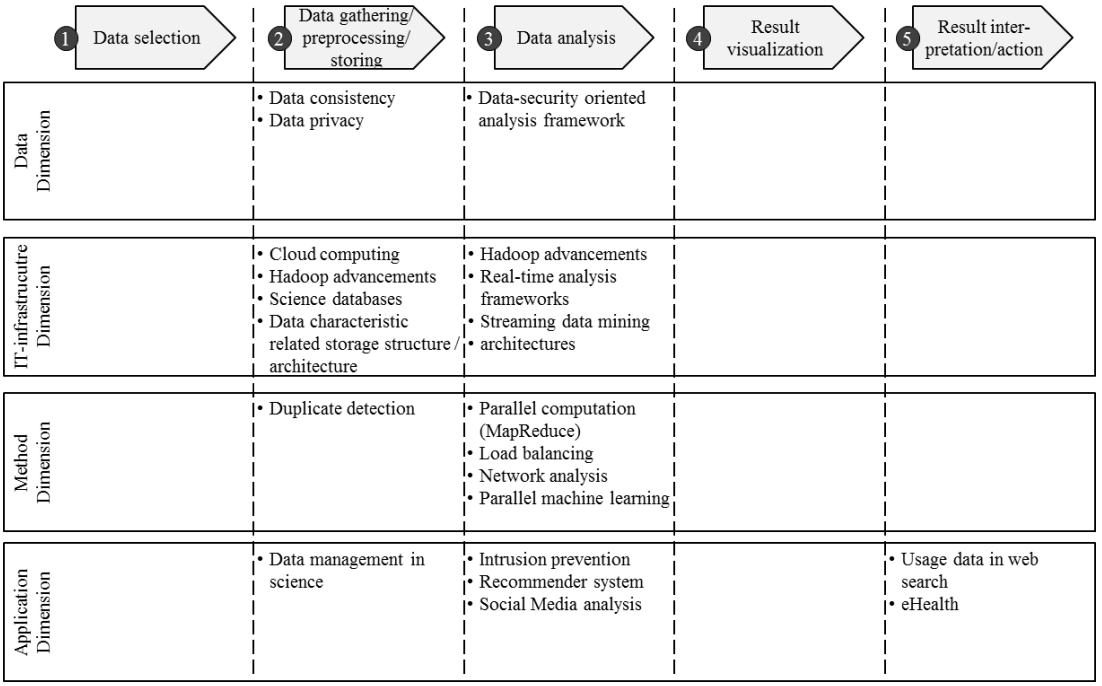


FIGURE 2.7: Current research topics of Big Data along the generic data analysis process model

The publications for the data dimension in Step 2, *Data gathering/preprocessing/storing*, focuses on the aspect of data consistency, when processing data from different sources [Chute, 2012] and algorithm-based data privacy protection, which is of special interest amongst others in the field of patient data [Zhang et al., 2012; Zhu and Li, 2012]. The IT infrastructure within the step 2 inherits, amongst the different Hadoop implementations, the aspect of data import into a Hadoop cluster [Xu et al., 2012]. The methodological focus in step 2 is limited to the method detection of duplicates within databases [Zhe and Zhi-gang, 2010].

Step 3, *Data analysis*, accumulates most of the publications. Data security-oriented analysis frameworks can be found within the data dimension. The IT infrastructure dimension of step 3 is dominated by Hadoop advancements that focus, among other topics, on performance improvements [Kejiang et al., 2012]. According to the number of Hadoop publications in the field of IT infrastructure, the method dimension is dominated by parallel computing, which is primarily based on the MapReduce framework, targeting aspects such as load balancing [Rosas et al., 2012] or energy consumption optimization [Hartog et al., 2012]. In addition to frameworks for parallel computation, the development of analysis frameworks for certain types of data are within the focus

in terms of i) the network analysis, applying graph theory approaches to analyze social networks [Bliss et al., 2012] or ii) real-time analysis for streaming data [Ari et al., 2012]. The first application-oriented publications can be found within the *Data analysis step*; these publications contain algorithms for intrusion detection [Jeong et al., 2012], recommender systems [Han et al., 2012] and social media analysis [Liu et al., 2012].
Publications with *Result interpretation* can be found in the application dimension, targeting the application of usage data from web search [Baeza-Yates and Yoelle, 2012] and ontology-based data access approaches in clinical environments [Curé et al., 2012].

The main findings of the assignment of the publications into the generic data analysis process are

- The current focus is on the *data gathering/preprocessing/storing* and *data analysis*.
- Both topics before the actual analysis (*Data Selection*) as well as afterwards (*Data Visualization and Result interpretation/action*) are considered only sporadically or not at all.

Potential explanation for the dominant focus on the infrastructural and methodological parts within the pre-processing and analysis steps are amongst others, that research in this fields can be carried out without the need for industry partners as performance topics and analysis tasks can be carried out as well in a laboratory environment. In contrast, the visualization and result interpretation/action are sensible topics for companies due to the potential competitive relevance, which can in turn have an influence on the willingness to cooperate with research institutions.
The subordinated consideration of the data dimension-related process steps in the forefront and afterwards of data analysis is surprising, considering that the data selection process as well as the visualization and interpretation play a major role with regard to the many available data sources. These steps are critical for utilizing the analysis results [LaValle et al., 2011] and have been mentioned in one of the presented definitions as well [Chen et al., 2012].[18]

[18] Again, the lack of publications related to step 1 and step 5 does not necessarily imply that no research is carried out on this aspect in a Big Data context; the lack of publications can result from the absence of the term *Big Data* in the key words, title or abstract of the publication.

2.5 Distinction between Big Data and Business Intelligence

With the rise of Big Data, critical opinions over this hype can be heard both from a practitioners perspective [Fox and Do, 2013; Boyd and Crawford, 2012] and research discipline specific perspectives [Barnes, 2013]. Both practitioners and researchers in the field of Business Intelligence are critical as some of them do not acknowledge the novelty of Big Data applications compared with Business Intelligence, which in turn has been in the past years, following the results by Luftman and Ben-Zvi [2010] the most relevant topic for CIOs.

Currently, it is argued that some achievements claimed by Big Data, such as the integration and analysis of social network data or the processing of sensor data, can be found in Business Intelligence applications of the newer generations as well [Buhl et al., 2013; Chen et al., 2012].

Examples are near real-time analysis, which do not follow the classical data flow process [19] that can be executed as well based on complex event processing engines.[20]

Recent Business Intelligence infrastructures have been adopted to meet the challenges of the increasing data amount, applying the MapReduce framework and changing the current ETL process towards an extract, load, transform process in order to overcome this bottleneck of transformation [Dayal and Castellanos, 2009].

These developments are subsumed under the description Business Intelligence 2.0 [Trujillo and Maté, 2012].

Despite those developments, however, most of the currently existing Business Intelligence systems in companies are still static, focused primarily on generating reports based on data from operational systems such as sales data [Negash, 2004]. Therefore, as it will be illustrated in Chapter 3 as well, existing maturity models from the field of Business Intelligence cannot be applied to companies if the focus is on Big Data.

[19] i) Data gathering, ii) extract transform load (ETL), iii) storing and data warehouse, iv) processing on mid-tier servers, e.g. OLAP server or data mining engine and, v) front end application

[20] The event data are directly loaded into an event processing engine with a standing query in order to detect trends, such as those based on streaming data [Chaudhuri et al., 2011], yet without the processing in a data warehouse; ultimately reducing the time needed from the point of data gathering to the data analysis.

Although from the authors point of view, a concluding differentiation between Business Intelligence and Big Data is not possible as also no common understanding of Business Intelligence exists [Gluchowski and Kemper, 2006; Rouhani et al., 2012], this work argues that Big Data is not to be understood as an advancement of Business Intelligence. Rather, it is to be understood as a paradigm respective concept as described by [Lane et al., 2014, P. 46], stating that

> "[]*...the term better reflects a paradigm than a particular technology, method or practice. There are of course, characteristic techniques and tools associated with it but more than the sum of these parts, big data, the paradigm, is a way to of thinking about knowledge through data, and a framework for supporting decision making, rationalizing action, and guiding practice.*"

This broad understanding is consistent with the understanding of Big Data in this thesis, seeing Big Data as a whole concept.[21]

2.6 The critical perspective on Big Data

After an initial phase of high expectations and a one-sided positive appraisal of Big Data, the number of publications and opinions with a critical perspective on Big Data is increasing [Boyd and Crawford, 2012]. The critics and those hesitant to Big Data are on different levels and granularity and come both from scientists as well as from practitioners. In order to structure the different arguments, the critics can be assigned to ***political, legal, ethical***, and ***scientific*** issues. The questions by Boyd and Crawford [2012]

> *"Will data analytics help us understand online communities and political movements? Or will it be used to track protesters and suppress speech?"*

show, that Big Data has both the potential to be applied in a positive way and to be diverted from its intended use.

[21] This definition has not been identified in the course of the structured literature review as it is part of a book and therefore not indexed in scientific databases.

The critical perspective regarding ***political*** and ***legal*** issues has gained momentum amongst others on ground of the publication of documents, describing the extent of the surveillance by the United States National Security Agency [Gellman and Poitras, 2013]. The examination of the breadth and depth of spying revealed extensive surveillance programs by most of the Western intelligence services. Basis for the programs are technologies and data processing techniques, which are used within business-related Big Data applications, as well particularly in the automated processing of text, speech, and other types of unstructured data. Due to a large number of newspaper articles naming Big Data in the context of the surveillance programs, Big Data is perceived increasingly negative [Lyon, 2014].

The discussion about data analysis in a broader context has risen a debate about the ownership of data as well when it comes to public available data from social networks and platforms, both driven from a ***legal*** and ***ethical*** standpoint. Although most of the data such as tweets are freely accessible as determined in the terms of service, it remains unclear from an ethical point-of-view, in how far these data are allowed to be used for analysis purposes, e.g. opinion and sentiment mining by third parties.

From a ***scientific*** point-of-view, Big Data has opened up a new research field, in which scientists from a wide variety of disciplines strive to contribute. At the same time, parts of the research community are critical towards the work of some colleagues, naming

i) a lack of future-time reference

ii) the use of a distorted database

iii) the negative influence of Big Data on the focus and process of research

as main critical points [Gayo-Avello, 2011].

Regarding i), research targeting the prediction of electoral behavior, until today, the existing publications explain the behavior subsequent to the election. No factual prediction has been published so far in the forefront of the election [Gayo-Avello, 2011].

ii) The distorted database is a limitation towards the explanatory power of research but has been named only sporadically in existing publications [Boyd and Crawford, 2012]. Twitter users are representing only an excerpt from the overall population; therefore they are not a representative sub-set. As Twitter does not publish facts about their

users besides basic data such as number of active accounts and share of accounts outside the United States, a social media marketing company has analyzed 35 million Twitter profiles in order to reveal the average Twitter user: a female, 28 years old, a citizen of the United States, and user of an iPhone. Although the 35 Million represent around 15 percent of the active user, it gives a first impression of the background of the tweets. Therefore, statements made on the basis of tweets cannot be generalized on the overall population. This problem is well known in the field of market research but in contrast, the high volume of data used when analyzing social media suggests universal validity. The critics regarding iii), the change of research, is summarized by Anderson [2008], stating in an early publication on huge data volumes that

> *"the petabyte-age [...] forces us to view data mathematically first and establish a context for it later."*

The increasing amount of data available leads to research which puts the emphasize on correlation rather than on causation, although *"Data without a model is just noise"* [Anderson, 2008]. Consequently, the number of models explaining human behavior decreases, although that is one focus of research in the field of social sciences [Boyd and Crawford, 2011].

Summing up the critical perspective on Big Data, the future development of Big Data depends on the belief in the explanatory power of data. Using an excerpt of the characteristics of Big Data by Boyd and Crawford [2012] the explanatory power of data can be understood as *"the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy"*, which will remain both within science and practice.

2.7 Main chapter results

Big Data is one popular topic both in science and practice in which describing characteristics and contained topics are perceived differently. Therefore, the beginning of Chapter 2 offers a first characterization of Big Data. Based on a mixed qualitative and quantitative approach, using recent literature on Big Data, the dimensions *Data, IT*

infrastructure, Method and *Application* have been identified. Topic models have proofed themselves as a suitable approach for the characterization of an emerging research field, although the quality of the results depends on the heterogeneity and number of the documents in the analyzed corpus. Furthermore, relevant publications might be left out because they do not hold the searched keyword, although they might contain relevant information.

Moreover, the assignment of the results to the phases of the generalized data analysis process reveals a under-represented consideration of the steps *Data selection, Result visualization*, and *Result interpretation/action* in the current Big Data research. The emphasis is on research associated with *Data pre-processing* and *Analysis*. The relevance of this comprehensive review of Big Data results, as explained before, from its novelty in combination with the breadth and depth of the contained dimensions and topics. In order to give a comprehensive view on Big Data, critical aspects targeting political, legal, ethical, and scientific issues have been presented.

After the characterization of Big Data, in the next section, the concept of maturity models and the contextual brackets of Design Science Research are presented as a starting point for the subsequent maturity model development.

Chapter 3

Maturity Models - Theoretical foundations

The goal of this dissertation is to develop a model that captures different maturity levels of Big Data. As it will be explained in detail later on, each maturity level consists of capabilities targeting the handling of Big Data relevant processes and activities. The resulting maturity model can be used as a basis for the construction and improvement of an organization as maturity models have proven themselves as a basis for the design of an organization [Carley, 2002].[1] Maturity models belong to the group of reference models, which is a specific type of models studied in information system research.[2] With regard to the relevance of reference models for this research, the aims of the following chapter are

i) Develop an understanding of characteristics of models as well as the construction and application of reference models in general

[1] The understanding of organizations in this work follows [Aldrich, 2008, p. 14], defining organizations as "[...] *goal-directed, boundary-maintaining, activity systems.*" The focus is on the aspect of *activity* - resulting in achievements - as those are primary the ones, which are supposed to be evaluated by applying a maturity model. The change and improvement of those activities in a subsequent step is supposed to increase the maturity. A discussion on the approaches for organization change - *transformation management* and *organizational development* - is not pursued as organizational changes, based on the maturity evaluation, are beyond the scope of this thesis [Inversini, 2005, p. 34ff].

[2] As described in Chapter 1, research in the field of Business Information System Engineering can be distinguished into the design and behavior oriented paradigms. The focus of the work at hand is on the design oriented research.

ii) Give an overview about the elements and characteristics of maturity models as a specific type of reference models

iii) Present the current research both on the maturity model construction models as well as the maturity models in the field of Big Data and nearby fields

In doing so, as a first step, models and reference models as a specific type of models are characterized based on relevant literature.
Second, based on a literature review, existing meta-studies of maturity models are identified and used to carve out the key elements and characteristics of maturity models.
Third, using again the method of a literature review, construction models are compared, followed by the analysis of maturity models from a scientific background with relevance for Big Data in order to carve out contained topics and potential white spots.

Following the characterization of models by [Stachowiak, 1973, p. 131ff], describing features are

- **Reproduction**: A model is a *reference* to an underlying original
- **Reduction**: The *abstraction* from the underlying original leads to the model
- **Pragmatism**: The degree of abstraction and the construction is dependent on the models purpose

The aspect of *Reproduction* is focusing on the underlying original that acts as a starting point for the model development; in this case the underlying original is to be understood as the companies' set of capabilities in dealing with Big Data.

Reduction: It is assumed that all attributes of the underlying original are known but not everything has to be taken necessarily into account for the resulting model. Transferring this to the Big Data maturity model to be developed, the abstraction from the underlying original is carried out by defining specific capabilities that can be used to represent different maturity levels. The selection of items is guided by pragmatism (next aspect) and model purpose.

The feature *pragmatism* is described by the aspects intentionality, temporality, and subjectivity, which can be summarized by asking for whom, when, and what [Stachowiak,

1973, p. 133], which makes this one of high relevance for the later maturity model construction. Both the modeler and the target group in terms of subjectivity and intention have to be taken into account during the modelling process. A Big Data maturity model entails numerous potential target groups, both amongst industries as well as business units. Consequently, the maturity model construction process has to incorporate the identification and selection of the target group.
Furthermore, models describe the underlying original to a point of time and therefore do not claim a permanent validity. This aspect gains in relevance with regard to the dynamic, relative character of maturity as it will be explained in section 3.3. Therefore, a regular update of the model is relevant.

Models that contain a recommendation and universality for a certain context are denoted as reference models [vom Brocke, 2003, p. 31]. As maturity models belong to the field of reference models, types and characteristics of reference models are discussed in more detail in the following section.

3.1 Reference Models - Definitions

Within the field of information systems, reference models have gained in relevance since the 2000's [Fettke and Loos, 2004]. Since reference models are a special type of model, they share parts of the general model characteristics with regard to the assumed *universality* [vom Brocke, 2003, p. 32]. This resulting *universality* under specific conditions is discussed in the scientific community as this would inherit a universal validity of the model [Mettler, 2010, p. 36]. In addition, the aspect of reference has a double meaning. It stands both for the reference towards a subject/system as well as for the reference in terms of a recommendation character of such a model. Therefore, following the definition of [vom Brocke, 2003, p. 34] a reference models as a specialization of models is an

> *"[...] Information model that people develop or use to support the construction of application models in which the relation between reference- and application model is characterized that the subject or content to the reference*

> *model is used for the subject or content for the application model construction as well."*[3]

Thus, in the context of reference models, related model types are *information models* and *application models*, which will be briefly described.

Information models are understood as models that are focused on the information within an organizational system, resulting from the construction process, in which information elements, their connection and dependencies, needed for an information system, are described [Schütte, 1998, p. 63]. One example of information models are Entity-Relationship-Models [Chen, 1976]. Those application models are company-specific models, which i) can contain dynamic aspects and ii) have been developed based on a reference model [Braun, 2009]. Therefore, an application model is a specified reference model.

As reference models are design oriented models, the distinguishing factor between those and natural science oriented models is the focus on utilization rather than truth [March and Smith, 1995, p. 256]. Consequently, reference models can contain aspects of valuation and subjectivity caused by the aspect of pragmatism, named by Stachowiak [1973] [Mettler, 2010, p. 36].

Nonetheless, the reference of reference models can be towards best practice or common practice, which supports its demand for universality [Becker et al., 2002]. The demand for universality has an influence on the dimensions of Big Data, which are taken into account as a basis for the maturity model; the broader the dimensions - describing the maturity object in focus, in this case Big Data - are defined, the more universal the resulting model can be. At the same time, an increasing level of universality can reduce the explanatory power, as the model explanations are too abstract and hold too many generalizations.

After defining reference models, the process for reference creation is explained.

[3] "Ein Referenzmodell (ausführlich: Referenz-Informationsmodell) ein ist Informationsmodell, das Menschen zur Unterstützung der Konstruktion von Anwendungsmodellen enwickeln oder nutzen, wobei die Beziehung zwischen Referenz- und Anwendungsmodell dadurch gekennzeichnet ist, dass Gegenstand oder Inhalt des Referenzmodells bei der Konstruktion des Gegenstands oder Inhalts des Anwendungsmodells wieder verwendet werden."[vom Brocke, 2003, p. 34]; translated by the author of this thesis

3.2 Process steps for reference creation

The existing research in the field of reference models can be categorized into the process of *model construction* and *model application* [Fettke and Loos, 2004].
The first one focuses on steps which have to be undertaken in order to construct a reference model. The latter one deals with the specialization in terms of the application of reference models for the design of information systems in firms. The scoping of the maturity model, which has an influence both on the construction as well as subsequent application will be carried out in Chapter 5. In the thesis at hand, the focus is on the *model construction*, resulting in the maturity model for Big Data.[4]

3.2.1 Model construction

Within existing literature on the model construction process, different meta-models have been identified. Following the meta study of Fettke and Loos [2004], the process of model construction can be classified into four steps: i) problem definition, ii) construction in a narrower sense, iii) evaluation, and iv) maintenance. The named phases are those which can be found throughout the most relevant construction approaches and thus are used for the coming elaborations.[5]
The *problem definition* contains the identification and description of the target state under consideration of the respective subject to be analyzed, the selection of a modelling language, and the utilized modeling conventions. The model *construction in a narrower sense* inherits the conceptualization of the forefront-defined subject, based on the selected modelling language.
The subsequent model *evaluation* analyzes in how far the designed artefact can be used to improve the targeted subject. The model evaluation has been identified as a relevant but subordinately considered phase so far. Although a number of evaluation approaches exists, the related strength and weaknesses, as well as premises, have not yet been analyzed extensively [Fettke and Loos, 2004]. Considering the practical relevance of maturity models, the aspect of evaluation will be one focus area of the maturity model

[4]The Model application can be found as well in the course of the model evaluation but as no instance is developed, it is not the primary focus.
[5]Further Aspects as *communication* appear rarely and therefore are not further considered [Peffers et al., 2007].

construction as described in Chapter 1.[6]
The reference model *maintenance* inherits processes in order to fit the developed model to changes in the environment that affect the model's validity [Hevner, 2007].

Based on the described generic model construction process steps for the development of reference models, numerous construction models for the development of maturity models exist. These will be discussed in section 3.3.5.

3.2.2 Model application

The second step of the reference creation is the *Reference Model application.* Following Schütte [1998], reference model applications can be classified into

- The application for the analysis of the current situation, resulting in an evaluation of the as-is situation
- The description of a company-specific information system

The focus of maturity models is the evaluation of a current status quo [Paulk et al., 1993], the currently existing capabilities in a specific field are in the focus. Consequently, the first type of application is the focus. The application of reference models can be split into four phases, i) problem definition, ii) identification of requirements, iii) search and selection, and iv) construction and application [Fettke and Loos, 2005]
Following the approach by Schütte [1998], the application of reference models for the analysis and improvement of organizations starts with the *existence of a problem*; the deviation between the actual state and a desired state. At this point, the company in focus does not know how to identify and overcome this deviation from the desired state [Bretzke, 1980, p. 33ff].[7]
In a next step, the *causation* for the problem is identified based on the investigation of the company's structure. This structure helps to prioritize which parts of the organization and related processes have to be analyzed in more detail.

[6] The related research goal is the development and application of a maturity model evaluation approach, focusing on the suitable representation of maturity in the developed model.
[7] Based on this statement, it can be discussed in how far maturity models belong to reference models as one main critic of maturity models is the sole focus on the evaluation on status quo; no recommendations for the improvement of the situation are given [Biberoglu and Haddad, 2002].

The current processes act as a basis for the selection of a suitable reference model. Based on that, operations are compared with those given by the reference models. At this point, an in-depth discussion of the processes is needed. Resulting differences are optimized subsequently. The application of reference models is not in the focus of this work as no instance, e.g. software, is developed for the model application. Instead, an application is only carried out during for the model evaluation (Chapter 4) but not for the subsequent improvement of processes etc. as described before.

After describing reference models and the related construction and application processes, maturity models as a subcategory of reference models are described.

3.3 Maturity Models

In the following chapter, the topic of maturity models is described. The approximation follows a multiple-step approach with an increasing focus on maturity models in the context of Big Data.

In a first step, the concept of maturity in general is introduced. Next, research-based maturity models in general are described with their contained elements and characteristics. This outlining of maturity models is followed by the presentation of different approaches for the construction of maturity models. Subsequently, maturity models from the field of information system research, business information system engineering, and in particular for the field of Big Data and associated fields such as Business Intelligence are discussed based on a literature review.

3.3.1 The concept of Maturity Models

The concept of maturity models was published for the first time in 1973 in terms of a staged model for the planning, organizing and controlling of processes related to the management of a company's IT resources [Nolan, 1973; Damsgaard and Scheepers, 1999]. The concept gained attention with the development and publication of the *Capability Maturity Model* (CMM) and its enhanced version *Capability Maturity Model Integrated* [Humphrey, 1988]. The initial CMM has been developed to describe different stages of maturity of software development. Around the CMM, some models have been developed

and improved in the course of time which are accepted by the International Standards Organization, namely BOOTSTRAP and SPICE, as a programming framework respective a framework for the evaluation of business processes [Mettler, 2010, p. 63] [Kuvaja, 1999].

The basic idea of the early maturity models has been the assumption that a high maturity is connected with and also can be measured via the existence or absence of certain process capabilities. Therefore, in the work by Paulk et al. [1993] a maturity level is defined as

> "[...] *a well-defined evolutionary plateau towards achieving a mature software process.*"

Maturity models are used to compare and evaluate different states of capabilities in specific areas in companies. Maturity in this context is defined as *"the state of being complete, perfect of ready"* [Simpson and Weiner, 1989]. Nonetheless, the implicit notion "evolution" by Paulk et al. [1993] can be seen as pointing at the dynamic character of maturity.[8]

Within science and practice, maturity models with two different perspectives can be found [Wendler, 2012]. The less common one is the *life cycle perspective*, which is based on the assumption that an object will pass through different stages on its way of development to a final state of maturity, comparable to the life cycle concept of a product from the market entrance to its decline [Utterback and Abernathy, 1975].

The second maturity model perspective is called *potential performance perspective*. Models with this perspective assume a connection between the maturity level and the related performance, i.e. the higher the maturity of an object, the better its performance. With regard to the investments needed for building up and improving capabilities relevant for Big Data, companies expect a performance improvement in order to advocate the financial effort. With regard to the aspired practical relevance, and the diverse, dynamic field

- a company is not necessarily passing through each stage on its way to maturity and

[8] The influence of this dynamic on the maturity model construction process will be described in more detail in Chapter 4.

- Big Data is closely connected to the aspect of performance as the company's decision making can be improved by capabilities in the field of Big Data.

Therefore, the potential performance perspective is taken.
After describing the concept of maturity, in a next step the elements and characteristics of maturity models are explained.

3.3.2 Model elements and characteristics

Early maturity models have established themselves as helpful instruments for the maturity evaluation of companies [De Bruin et al., 2005]. Existing models come from both a practitioner as well as scientific background and cover diverse topics such as the fields of process management [De Bruin and Rosemann, 2005] and health care [Grindle et al., 2013].
Regardless of the heterogeneity of existing maturity models from different disciplines, a number of characterizing elements can be recognized. In the following subsection, an overview about elements and characteristics of maturity models in general is given. In doing so, based on a literature review, four publications have been identified which describe elements of maturity models as listed in table 3.1.

Lahrmann and Marx [2010] name *i) the maturity concept, ii) dimensions, iii) level, iv) maturity principles* and *v) the assessment* as properties of maturity models.
i) In their view, the maturity concept can target the aspects of *people* (people's knowledge in conducting business-relevant processes [Curtis et al., 2001]), *process* (*"the extent to which a specific process is explicitly defined, managed, measured, controlled, and effective."* [Paulk et al., 1993]), and *object* maturity (*"the respective level of development of a design object"* [Mettler, 2010; Gericke et al., 2006]).
ii) *Dimensions* in the context of maturity models are used to structure the subject of analysis.[10]
iii) The maturity *levels* are used to distinguish and arrange different capabilities in order. For each level, a description is assigned, which contains the main characteristics of the

[9] The order of the elements in the table differs from the one in the text in order to achieve a comparison between the different publications.
[10] A more detailed discussion on the aspect *Dimensions* can be found in Chapter 4 and 5, potential dimensions of Big data have been identified in Chapter 2.

TABLE 3.1: Elements of maturity models identified in literature on the construction and application of maturity models.[9]

Lahrmann and Marx [2010]	De Bruin et al. [2005] Fraser et al. [2002]	Wendler [2012]
Level	Level	Defined set of levels
Assessment	Assessment	Defined criteria for measurement purposes
Dimension	Number of dimensions which offer a problem oriented perspective on the subject	
Maturity concept	Distinct notation for each maturity level e.g. initial, defined, repeatable, managed, optimized	
Maturity principle	Generic description of the characteristics for each maturity level	
	Number of elements that describe each dimension in more detail	

related maturity.

iv) The *maturity concept* targets the scoring of a company, based on the maturity model, which can be either continuous or staged. Continuous maturity models allow a scoring of companies capabilities on different levels for each dimension. Therefore, the level can "*be either weighted sum of the individual scores or the individual levels in different dimensions*" [Fraser et al., 2002]. In contrast, staged maturity models limit the levels to the defined stages and do not allow the specification of situational levels.

v) *Assessment* as the fifth characterizing element targets the approach how the actual assessment of companies' capabilities based on the maturity model is expressed. The description of the result can be either quantitative based on a scale (e.g. Liker-Scale) or qualitative, using descriptions in terms of a couple of sentences [Fraser et al., 2002].

The majority of the identified maturity model elements by Fraser et al. [2002] and De Bruin et al. [2005] are congruent to the one by Lahrmann and Marx [2010] but with partly different notions.

The only difference with regard to content is the level of detail selected to describe each element.

In contrast to the work by Lahrmann and Marx [2010], Fraser et al. [2002] and De Bruin et al. [2005], Wendler [2012] identifies only two basic elements, i) a defined set of levels that represent different stages of development and ii) defined criteria for measurement purposes as conditions, processes and application targets. This broader understanding of maturity models leaves out formal aspects, targeting the description of maturity levels, i.e. the maturity concept and the maturity principle.

After giving an overview about the main elements of maturity models, the elements *levels, assessment,* and *dimensions* as key components, are described in more detail in the next step.

The number of maturity *levels* in a model is between three and six. Less than three levels are too undifferentiated and therefore offer a lack in explanatory power. More than six levels do not allow a distinct differentiation between the single levels [Fraser et al., 2002]. The distinct notation of each level is supposed to hold an evaluating statement that gives a short-term idea of the subject's maturity. The description of the characteristics of each maturity level occurs often in the form of a few sentences, e.g. *"Quantitatively manage organizational growth in workforce capabilities and establish competency-based teams"* as a description of the fourth level of the people capability maturity model [Curtis et al., 1995].

Most of the topics, which have been subject to a maturity model, have a wide scope regarding the content and contain numerous subtopics, e.g. Supply Chain Management [Lockamy and McCormack, 2004] and Business Intelligence [Lahrmann et al., 2010] etc. In order to allow a later *assessment* based on the model, the subject in focus has to be divided into describing *dimensions*, which in turn are dependent amongst others on the model goal and audience [De Bruin et al., 2005]. Dimensions are fractions of the overall subject of analysis. The number of dimensions per maturity model depends on the focus and on the subject of analysis of the model. Each dimension is described to set the boundaries and work as a basis for the further description of elements for each maturity level. These elements have to be explicitly measurable [De Bruin et al., 2005; Fraser et al., 2002].

In summary, a maturity model consists of three to six maturity levels, each described

with a few characterizing sentences and backed up with a number of related measurement per dimension and level.

After describing the elements of maturity models in the last step, the goal is to give an overview about existing construction processes and possible generalization approaches. In section 3.3.3, an overview about research on maturity models in general is given, including a critical perspective on maturity models (section 3.3.4), followed by a summary about construction models (section 3.3.5) and maturity models with relevance for Big Data (section 3.3.6).

3.3.3 Current research

Since the invention of the maturity concept, maturity models are of interest for both scientists as well as practitioners. The interest in the topic results in numerous publications from both groups.

A first overview about the methodological aspects of maturity models in the current research can be found in Wendler [2012], who analyzed 237 maturity models from 22 different domains. The first finding of the literature review by Wendler [2012] is, that despite the arrival of the maturity model concept in different scientific disciplines, the majority of the publications have their origin still in the field of software development and engineering with 89 publications. In contrast, solely three publications belong to the field of BI.[11]

The publication analysis reveals furthermore that i) the number of publications differ significantly between the two main research fields, model construction, and model application.[12].

The majority of the publications is concerned mainly with the *application* of maturity models, aiming at the assessment of processes and organizational structures in different fields such as software development, e-learning , and supply chain management [Lockamy and McCormack, 2004; Marshall and Mitchell, 2004; Neuhauser, 2004]. A smaller share of publications on maturity models is focused on the underlying *construction* process, executing different construction approaches.

[11] This will be reflected in the relatively low number of publications on maturity models, which can be used to extract relevant aspects for a Big Data maturity model.

[12] These two groups are resulting from the origin of the maturity model concept in the field of reference models, which can also be split into model *development* and model *application*. These two groups have already been presented in sections 3.2.2 and 3.2.1

Besides these groups, a class of paper can be identified, which is focused on the transfer and adoption of existing maturity models on new topics and therefore neither develop a complete new model nor are limited to the sole model application [Veldman and Klingenberg, 2009; Drinka and Yen, 2008]. Most of these transfer papers revert to the capability maturity model (CMM) from the field of software development [Paulk et al., 1993].

3.3.4 A critical perspective on Maturity Models

Despite their widespread acceptance, maturity models have been subject to criticism as well, targeting their application and methodological foundations. The application oriented critic targets the lacking explanation of needed changes to reach a higher maturity level as most of the models only serve for the maturity evaluation and do not offer subsequent improvement actions [Kamprath, 2011]. This criticism arises partly from practitioners as a sole maturity evaluation is merely an initial step in the overall process of improvement from a company's perspective.

Further critics target the strong focus on processes, in that processes may lead to a deficient consideration of the workforce's qualification and capabilities [Bach, 1994]. This focus, in turn, leads to a high degree of formalism which then can result in lower innovation capabilities of the employees [Herbsleb and Goldenson, 1996].

In addition, companies that show a congruence with processes of a high maturity level do not necessarily imply success, as the overall company's improvement in performance may not be necessarily achieved only via the specified processes [Montoya-Weiss and Calantone, 1994; Mettler, 2010]. This criticism points out the relevance of the maturity measurement selection during the model construction process.

The methodological oriented critic in contrast is targeting the lack of a sound theoretical foundation of most of the developed models. A literature review on Business Intelligence maturity models from both a scientific and practitioner background by Lahrmann et al. [2010] revealed that only one out of ten models is based on a sound theoretical foundation [Biberoglu and Haddad, 2002]. This lack of a theoretical foundation is referring to both the initial model construction as well as the later model testing for validation. If the model development and validation is based on the input of a few companies or persons, the selection of these can lead to a key informant bias [Lahrmann et al., 2010].

This accounts especially for models populated based on a qualitative approach. Following the critique by Bach [1994] regarding the CMM, in which he states

> *"The CMM has no formal theoretical basis. It's based on the experience of "very knowledgeable people". Hence, the de facto underlying theory seems to be that experts know what they're doing. According to such a principle, any other model based on experiences of other knowledgeable people has equal veracity."*

Consequently, the use of a Delphi study for the model construction does not necessarily lead to a valid model.
Therefore, in the description of the construction process (Chapter 4) and the actual maturity model construction (Chapter 5), the aspects of a methodological based model population and the subsequent model evaluation will be in the focus.

After giving an introduction into research in the field of maturity models in general and related criticism, the following elaborations on current research follow the structure of this thesis. In a first step, the focus is on the research on the model construction. Different construction process models are discussed and compared, followed by the description of existing maturity models from the field of Big Data and Business Intelligence regarding the underlying model construction process and the covered maturity aspects.

3.3.5 Generalized process models for Maturity Model construction

The construction model for the development of the Big Data maturity model is one emphasize in this research. Therefore, different construction models will be described and compared.
An overview about different construction models for the development of maturity models, which are primary modifications of the general design science research process by Peffers et al. [2007] and Hevner et al. [2004], can be found in Steenbergen and Bos [2010]. With regard to the years passed since the publication of the research by Steenbergen and Bos [2010] and the dynamic in this field, this publication is taken as a starting point and is enriched by research papers containing maturity models, which have been published after the one by Steenbergen and Bos [2010] in the year 2010. The publication selection

has been done based on the publication list by Wendler [2012], enriched by models, which have been published after the analyzed period by Wendler [2012], using the same keywords in the databases of Business Source Complete (EBSCO), ACM, Science Direct, Emerald Management, and SpringerLink, following the literature review approach described in Chapter 2, following the approach by Wendler [2012] the key words "maturity model", "capability model", "process improvement model", "maturity grid", "competency model", and "excellence model" have been selected.
Furthermore, publications from the International Conference on Information Systems (ICIS), the European Conference on Information Systems (ECIS), the American Conference On Information Systems (AMCIS), the Hawaii International Conference on System Sciences (HICCS), and the International Conference on Wirtschaftsinformatik (WI) have been taken into account as well as these conference are the leading conferences from the field of Information Systems.
The construction models which that have been identified in the course of the literature review can be found in table 3.2.[13] The common process phases will be described in the following elaborations [Steenbergen and Bos, 2010].

The ***scope*** phase is setting the boundaries of the model to be developed. The subject of the model is defined as well as the latter target group. This decision influences the model content, the further construction process, and the final model description. The decision regarding the trade-off between practical relevance and scientific foundation is made.
In a next step the ***model design*** is defined. The dimensions for the description of the model subject are carved out and the theoretical basis regarding the used methodologies is set. This phase, especially the dimension definition, is influenced by the scope set in the step before.
The subsequent ***develop instrument*** phase contains the description of maturity levels and the identification of related measurements. The order of these two steps depends on the construction approach. In the case a top-down approach is selected, as a first step, the levels are defined followed by the measurement identification. In the case a bottom-up approach is selected, the measurements are identified first and used for the definition of the maturity levels subsequently.

[13] Additional identified publications to the ones presented by Wendler [2012] are the construction approaches by Lahrmann et al. [2011a], Lukman et al. [2011], and Marx et al. [2012].

TABLE 3.2: Steps of the Maturity Model construction approaches in current research mapped along the common process phases known from Reference Model construction

Common process phase	De Bruin et al. [2005]	Mettler and Rohner [2009]	Becker et al. [2009]	Maier et al. [2012]	Lahrmann et al. [2011a]	Lukman et al. [2011]	Marx et al. [2012]
Scope	Scope	Problem identification and Motivation	Problem Definition	Planning	Identify need/new opportunity	-	-
	-	Objectives of the solution	Comparison of existing maturity models	-	Scope	-	-
Design model	Design	Design and development	Determination of development strategy	Development	Design Model	Questionnaire Development/ Data gathering	Questionnaire Development/ Data gathering
	Populate components	-	-	-	-	-	-
Develop instrument	Populate - measurements	-	Iterative maturity model development	-	-	Maturity Model Analysis	Development & Aggregation of domain specific maturity model
	Test	-	Conception of transfer/ evaluation	Evaluation	Evaluate Design	-	-
Implement & Exploit	Deploy	-	Implementation of transfer data	Maintenance	-	-	-
	Maintain	-	Evaluation	Reflect Evolution	-	-	-

During the final ***implement & exploit*** phase, the model developed in the step before is used to evaluate companies' maturity in order to test the model's capabilities and correctness. Furthermore, the model is updated and fitted according to potential dynamics of the topic in focus.

As it can be seen in table 3.2, the listed models differ regarding i) the level of granularity and ii) the emphasis of the individual phases.[14] Additionally, iii) the degree of orientation towards the principles of design science research approach can be taken as a distinguishing factor. Except for the models by De Bruin et al. [2005] and Reyes and Giachetti [2010], the listed construction approaches are similar to the phases of the design science research process [Hevner et al., 2004; Peffers et al., 2007].[15] A more detailed discussion on construction models with relevance for this research will be carried out in Chapter 4.

Up to this point, an introduction into the concept of maturity and maturity models has been given, followed by the description of the main elements of maturity models and existing construction approaches for the maturity model development.
In a next step, existing maturity models with relevance for Big Data - primary from the field of Business Intelligence - are presented in order to carve out the current state-of-the-art as well as to identify white spots that are supposed to be filled with the maturity model to be developed.

3.3.6 Current research on Maturity Models in the field of Business Intelligence and Big Data

Following its origin in the field of Software Development Maturity [Paulk et al., 1993], maturity models are a topic of interest in the context of information system for several years [Raber et al., 2012, 2013a]. Again, the literature review by Wendler [2012], covering the years from 1993 to 2010 is taken as a starting point, using the key words "maturity model", "capability model", "process improvement model", "maturity grid", "competency model", and "excellence model". The searched databases are Business Source Complete

[14] The construction approach used by Raber et al. [2012] is not considered further as it follows the same approach as Lahrmann et al. [2011a].
[15] A discussion of these differences based on the construction models of De Bruin et al. [2005] and Becker et al. [2009] can be found in Chapter 4.

(EBSCO), Science Direct, Emerald Management, and SpringerLink. Furthermore, publications from the International Conference on Information Systems (ICIS), the European Conference on Information Systems (ECIS), the American Conference On Information Systems (AMCIS), the Hawaii International Conference on System Sciences (HICCS), and the International Conference on Wirtschaftsinformatik (WI) have been taken into account as these conferences are commonly seen as leading conferences from the field of Information System research. The considered period of investigation is extend to paper published from 2010 until 2014.

In a next step, the identified maturity models following the approach by Wendler [2012] have been searched for maturity models from fields relevant for this research such as Business Intelligence. The used key words are "Big Data", "Business Intelligence", "Analytic", and "Decision Support". The goal is to identify as well maturity models which may not be developed in the context of Big Data, but contain nonetheless aspects relevant for the maturity model to be developed. This approach follows the argumentation of Chen et al. [2012], describing different development stages of *Business Intelligence and Analytics* in which the highest stage can be understood as Big Data.

The resulting publications have been scanned manually for maturity models, focusing on the topic of Big Data and nearby analytic topics as described above.

After removing non-relevant papers, nine publications additionally to the ones identified by Wendler [2012], mainly from the field of BI could have been identified since 2010 [Lahrmann et al., 2011a,b; Lukman et al., 2011; Marx et al., 2012; Dinter, 2012; Raber et al., 2012, 2013a,b; Brooks et al., 2013], containing both maturity model constructions as well as model applications. Maturity models older than 2010 have not yet been taken into account, as the topic Big Data in fact was not represented in research before 2010. This analysis reveals that currently no maturity model for Big Data with a scientific background exists. This finding demonstrates the relevance of this research. The relatively low number of paper in BI maturity is congruent with the findings in the literature review by Aruldoss et al. [2014], analyzing the current research in the field of BI.

The from the above described process resulting publications can be found in table 3.3, describing the content of the main model development phases of each model in comparison to the generalized process model for maturity model construction. Only those have been considered further, which contain a description of an underlying construction process and which can thus be compared. Construction steps, which are not considered

in the respective publication, are marked with " - ".[16]

Lahrmann et al. [2011a] developed a theoretical, impactful, and research oriented model for the maturity evaluation based on IS theory for the field of Business Intelligence, basing the construction on the model by De Bruin et al. [2005]. Along the IS impact model parts *Deployment, use,* and *impact* the maturity model is developed, analyzing answered questionnaires with structured equation modeling.

The model by Lahrmann et al. [2011a] is based on the construction model by De Bruin et al. [2005] (Chapter 4). Subject in focus is Business Intelligence along the dimensions strategy, organization and processes, IT Support, and quality of service.

The model by Lukman et al. [2011], on BI in Slovenia also implements a quantitative approach, processing questionnaire data, based on the construction process by Becker et al. [2009]. A K-Means cluster algorithm is applied on answered items of the dimensions *technology, information quality,* and *business*. Approaches for the model evaluation/-validation and testing are not described.

Marx et al. [2012] developed a maturity model for corporate performance management, identifying maturity levels for domain-specific dimensions, involving *reporting, planning* and *consolidation*, as well as the generic dimensions *function*, consisting of organization and technology. The three-stage bottom-up construction approach is using a quantitative population approach comparable to Raber et al. [2012], based on the data from 78 companies. The model is the first of its kind, combining Business Intelligence and corporate planning. The final model testing is based on seven case interviews.

The publication by Dinter [2012] is focused on the model application, giving an overview about the current Business Intelligence maturity of companies from German speaking countries, clustered along different industries. Dinter [2012] identifies potential areas of maturity improvement. Maturity in this publication is defined along the dimensions of *functionality, technology*, and *organization* according to Schulze and Dittmar [2006]. The calculation of the model is based on the responses to a questionnaire.[17]

Raber et al. [2013a] develop a BI maturity model based on the data of 71 companies,

[16] As the publications by Lahrmann et al. [2011a,b] respective Raber et al. [2012, 2013a,b] belong together, they are not represented by individual entries in the table. The models by Dinter [2012] and Brooks et al. [2013] are not further pursued as it will explained later on and therefore are not listed in the table as well.

[17] As the underlying maturity model is published by a business consulting group and used partly for marketing purposes, it will not be further considered because on the focus of methodology driven maturity models.

TABLE 3.3: Analysis of existing maturity models in the field of Business Intelligence and analytics along the generalized construction phases *Scope, Design, Populate, Test/Evaluate* and *Maintain* [De Bruin et al., 2005] [a]

Maturiy Model	Scope	Design	Populate	Test/Evaluate	Maintain
Lahrmann et al. [2011a]	BI; Dimensions: Strategy, Organization/Process, IT support	Quantitative bottom-up approach (Rasch Algorithm)	Questionnaire results; 51 companies; cross-industry	-	-
Lukman et al. [2011]	BI in Slovenia; Dimensions: Technology, Information Quality, Business	Quantitative bottom-up approach (K-Means algorithm)	Questionnaire results; 131 companies; cross-industry	-	-
Marx et al. [2012]	Corporate Performance Management Systems; Dimensions: Planning, Reporting, Consolidation, Function, Organisation, Technology	Quantitative bottom-up approach (Rasch Algorithm)	Questionnaire Results; 76 companies	-	-
Raber et al. [2012]	BI; Dimensions: Strategy, Social System, Technical System, Quality, Use/Impact	Quantitative bottom-up approach (Rasch Algorithm)	Questionnaire results; 51 companies; cross-industry	Discussion of final model with three industry experts regarding comprehensiveness, self-assessment, potential BI roadmap	-

[a] Only those publications are listed, that contain a complete maturity model.

following an IS Success approach. The incorporated five dimensions: strategy, social system, technical system, quality, and use/impact, derived from the IS Success concept. The dimension-related elements are identified based on related literature. The population of maturity levels is carried out by using the Rasch algorithm from the field of test theory and agglomerative clustering. This approach will be described in more detail in Chapter 5.

The model developed by Raber et al. [2013a] is applied and tested for the influence of contextual factors, company size and environment by Raber et al. [2013b].

Brooks et al. [2013] develop requirements for a BI maturity model with a Healthcare focus based on existing maturity models. As the publication does not contain a finalized maturity model, it is not further pursued.

Aspects of the presented models will be discussed in detail in section 5.5.2.[18]

Up to this point, the frame for the Big Data maturity model has been set in this chapter by carving out describing elements of maturity models, existing construction approaches as well as existing maturity models with relevance for Big Data. The results of the conducted analysis of the identified publications on Business Intelligence maturity can be summarized as following:

- No Big Data specific maturity model exists
- In contrast, the concept of maturity models can be found in an increasing number of publications from the field of BI
- Existing publications still lack partly in a methodological foundation of the model construction process and the model evaluation [Biberoglu and Haddad, 2002, p. 150], which led to the relatively low number of remaining publications presented in the last section
- Within the group of theoretical based models, the construction process by De Bruin et al. [2005] and Becker et al. [2009] are popular [Lahrmann et al., 2011a]

[18] In addition to the described models, several exist, which contain relevant findings, yet do not have a sound theoretical foundation which is seen as a necessary prerequisite to follow the construction process. The model by Cosic et al. [2012] is one of the few models that develops maturity levels for Business Analytics systems. The resulting maturity model contains aspects that are beyond the description of classical BI applications, e.g. management skills, but Cosic et al. [2012] do not give a characterization of Business Analytics and therefore, the focus remains unclear. The model development process in their paper is based on the construction approach by Becker et al. [2009].

Although Business Intelligence has overlaps with Big Data, the examined models cannot be applied completely as, contrary to the statement by Lahrmann et al. [2010], *"in Business Intelligence systems, data from operational IS is combined with analytical front-ends"*, Big Data applications process data from further sources than that, e.g. social networks or sensor data [McAfee and Brynjolfsson, 2012]. Furthermore in contrast to BI systems, the fields of application for Big Data solutions are more diverse, including product recommendations or predictive maintenance, [Amatriain, 2013; Lee et al., 2013], aiming at a company-wide use of data. With regard to the type of data (unstructured/streaming data), an execution of this analysis within an BI infrastructure would be hard to achieve.

From the authors' point of view, Big Data has a company-wide penetration, seeing it as a main part of corporate decision making on all levels and units, which emphasizes the aspect of a paradigm [Lane et al., 2014, p. 46].

This broader focus becomes obvious when following the elaborations of Davenport et al. [2012], who identify three aspects, that distinguish Big Data from traditional analytics, which can be understood as a part of BI as well.

First, the processed data are increasingly streaming data, originating from sources such as social media, news streams or sensor in production environments. Consequently, the source and structure of the processed data have to be taken into account for the latter model as a measurement for maturity.

Second, companies working with Big Data increasingly hire Data Scientists instead of Data Analysts [Davenport, Thomas H. and Patil, 2012]. The aspect of science emphasizes the statistical part and the model building of the data analysis. Therefore, the organizational units, responsibilities and individuals, who implement the data analysis, do play a role in the Big Data maturity context.

Third, data analysis is moving away from an IT function into the core business. The aim is to use analytical applications throughout the company by the individual employees in order to foster the integration of analysis results in daily decision-making. Therefore, the integration of data analysis into operations and the use of results has to be covered by the model.

As these aspects are not or only partially covered by the presented maturity models, a need for a Big Data maturity model exists.

3.4 Main chapter results

In the past chapter, the concept of maturity and characterizing elements have been explained and set into the context of reference models. In the course of further classification, the framing design science research paradigm, that will be the basis for the model development has been described. By comparing different maturity construction models, key elements of the model construction, have been identified, namely *Scope, Design Model, Develop instrument* and *Implement and Exploit.*

The subsequent analysis of different maturity models from fields such as Business Intelligence revealed that no existing model is focusing on Big Data. Furthermore, a lack of theoretical foundation in terms of i) no underlying theoretical construction model and ii) a lack of evaluation and validation approaches could been identified. Therefore, in the next chapter, a construction model will be developed in order to base the Big Data maturity model on a sound theoretical foundation and avoid the identified weaknesses.

Chapter 4

Development of the model construction process

Up to this point, the two main subjects of this thesis, Big Data and Maturity Models, have been analyzed. Describing dimensions of Big Data are identified based on a qualitative and quantitative, structured literature review (Chapter 2). The concept of maturity models has been explained and set into the context of the reference models and the general design science research, followed by the results of the literature review on maturity models in Big Data relevant fields (Chapter 3). After setting the theoretical frame, the model construction follows in the next two chapters. The construction starts with the development of the construction model. In doing so, two construction approaches will be described, compared, and fitted. This results in the construction model, which will be applied subsequently in Chapter 5 for the development of the Big Data maturity model.

4.1 Model construction - Theoretical basis

As it could be shown in Chapter 3, there exists a large number of maturity models in the field of computer science and information systems research, which differ in their theoretical foundation with regard to the underlying construction model.
Existing maturity models can be categorized into

- a completely new model design,
- models resulting from the enhancement of existing ones,
- a new combined model based on existing ones, and
- models based on the transfer of structure or contents to new domains [Becker et al., 2009].

As no research-based Big Data maturity model exist, a new maturity model will be designed, based on a construction approach.

The construction processes by De Bruin et al. [2005] and Becker et al. [2009] have been selected as a basis due to the following reasons:
First, although the models differ with regard to the level of detail in which the construction is carried out, both models are widely cited and accepted within the scientific community.
Second, both of the utilized construction approaches contain essential aspects that can be used for the construction of a Big Data maturity model, as it will be shown later on. However, none of the models can be used in their original version as they are lacking a sufficient consideration of a quantitative approach for the model population and the model evaluation aspect. These are critical aspects as understood in this work and thus have to be expanded with regard to the research goals described in Chapter 1 and the character of the topic Big Data.

Therefore, the following chapter discusses the development of a new construction model, suitable to the topic of Big Data. In doing so, in a first step, the models by De Bruin et al. [2005] and Becker et al. [2009] will be described and compared subsequently.

4.1.1 Construction model by Bruin et al. [2005]

The model by De Bruin et al. [2005] suggests the development of maturity models along the phases *1) Scope, 2) Design, 3) Populate, 4) Test, 5) Deploy*, and *6) Maintain*.
The first phase, *Scope*, contains the definition of the model's focus, which can be either general or domain-specific. Examples for a general model are the supply chain maturity model by Lockamy and McCormack [2004] and the excellence model by the European

Foundation of Quality Management (EFQM) [EFQM, 2012].[1] General models are neither industry specific nor focused on a specific topic in a certain field [De Bruin et al., 2005].

Contrary, domain specific models as the capability maturity model for software development [Paulk et al., 1993] or the model by Marx et al. [2012] focusing on management control systems, are more specialized on one topic. The second part of the initial *Scope* phase outlines the focus-related stakeholder identification. The stakeholders in focus have an influence both on the later defined items as well as the documentation of the final model.

The subsequent second step - *Design* - is primarily characterized by decisions regarding the application of the constructed model. Crucial aspects are the audience of the final model, the drivers for application, as well as the intended spread of the application throughout the company.

Additionally, the decision about the population approach has to be made. The selected approach depends on the maturity of the domain in focus (which will be explained in detail in the third construction step). In general, the existing population approaches can be split into bottom-up and top-down. Following De Bruin et al. [2005], bottom up is suitable for mature domains, top-down for immature domains.[2]

When using bottom-up approaches, suitable for mature domains, the elements, independently of maturity levels, are defined first. In a second step, the maturity elements are assigned to different maturity levels. In a third step, based on the elements per maturity level, the maturity principles are defined.[3] This procedure offers the possibility of applying quantitative models. This approach is only applied loosely in current model research-driven maturity model constructions.

For younger disciplines, the maturity levels with the related maturity principles are set up first and succeeded by the identification of maturity elements for each maturity level. This procedure is called top-down.

The differentiation into these two population approaches is based on the expectation, that the identification of maturity elements in a first step is hard to achieve due to the lack of experience, both of practitioners as well as scientists in the related field. Vice versa, for a mature discipline it is assumed that potential maturity elements are already

[1] Subject of the model are criteria for the evaluation of the quality management maturity in a company.
[2] The description of these aspects can be found in Chapter 3.
[3] The notions maturity measurements and maturity elements are used synonymously.

well known.
However, in step 2, only the decision for a top-down or bottom-up approach is made - the execution follows in step 3.

After setting the focus and boundaries in the steps 1 and 2, the model *population* as step 3 includes the identification of measurements and maturity indicators. It is thus the most comprehensive step.
Following De Bruin et al. [2005], the methods used for the identification and measurement of maturity depends again on the maturity of the domain. For a more mature domain, a literature-based identification is possible, whereas for younger disciplines, the processing of the results from expert interviews, case studies, or Delphi techniques is recommended. These maturity indicators are identified both on a high level, represented by dimensions that describing the domain (e.g. technological infrastructure as a domain of a business intelligence system), as well as on a low level, regarding indicators within a dimension.

The developed model and the model instruments - used for the model population - are tested for validity, reliability, and generalizability in phase 4, *Test*. The model validation can be segmented into face and content validity. Face validity refers to the quality of the translation of the construct, targeting the accuracy and completeness of the model. Validation techniques are focus groups or interviews. Following De Bruin et al. [2005], the use of different techniques in the population phase can foster the validation as well. Content validity is targeting the completeness of the representation of the topic. This depends on the extend of the literature review carried out in step one and two.
The same techniques are used for the model instrument validation. A focus group is used to validate the survey, which is used for the model assessment. De Bruin et al. [2005] state that a validation of the used instruments lead to a reliable model. Generalizability is achieved by a high volume of deployment in different environments regarding company characteristics.

The aspect of application is connected with step 5, the model *deployment*, which is carried out in a two-step approach. The initial deployment - the maturity evaluation of a company utilizing the developed model - as a first step takes place with collaborators from the model development process, as they are already familiar with the concept of maturity models. In the second step, the model is applied to organizations that have not

been involved in the construction process. The goal of this step is to evaluate up to what degree the model can be applied to companies with different characteristics regarding industry, size, and region. De Bruin et al. [2005] do not describe in detail how the model is supposed to be changed based on the received feedback.

The final, continuous *maintaining* phase comprises of fitting the developed model to the dynamics of the domain. Maturity is understood as a relative characteristic, based on the comparison with other companies' maturity. In the course of time when the knowledge in a domain broadens and deepens, characteristics associated with a high maturity can change. High and low maturity is partly determined by the best and the worst performing companies in the domain analyzed. Consequently, the model maintaining allows for the relative character of the maturity concept. As companies continuously strive to improve their capabilities, the model indicators have to be fitted to these changes in order to keep the maturity model up-to-date. De Bruin et al. [2005] point out the needed resources and partners for the model maintaining, which should be incorporated already in the initial scoping but do not describe the maintaining process in detail.

4.1.2 Construction model by Becker et al. [2009]

The model by Becker et al. [2009], containing eight steps, has been developed based on the criteria by Hevner et al. [2004], describing criteria which should be considered during the development of artefacts to foster a sound scientific foundation.
The model starts with the *problem definition*, containing both the determination of the target domain as well as the target group. Furthermore, the model demand has to be reasoned, describing the underlying model need in detail.
The *problem definition* acts as a basis for the next phase, with the *comparison of existing maturity models* on the same topic. This phase is supposed to identify shortcomings of existing models and to motivate the modification of the own maturity model.
The subsequent third phase represents the determination of the *model development strategy* that can be a complete new model development, the combination of existing models, as well as the transfer of structures and content from existing models to other disciplines, e.g. from software development to quality management.

The main focus of the model is on the fourth phase, the iterative maturity model *development phase*. This phase inherits four sub-phases:

i) *Selecting the design level*: decision and determination of dimensions, which represent the *"fundamental structure of the maturity model"* i.e. the decision between a one-dimensional and multi-dimensional approach regarding the maturity steps and the determination of related individual dimensions as well as belonging attributes

ii) *Selecting the approach*: Selection of an approach for the model population, e.g. Delphi method and creativity techniques

iii) *Designing the model selection*: The actual model population based on the selected approach

iv) *Testing the results*: Focusing on comprehensiveness, consistency, and problem adequacy

The *development phase* is iterative - it is carried out again until the model is evaluated for comprehensiveness, consistency, and problem adequacy successfully.[4]
After the model construction has been completed, the fifth phase, *conception of transfer and evaluation* follows. The resulting model is documented and tools are developed (both document and software-based) to use the constructed maturity model for the maturity assessment of companies. The design depends on the target group, which can consist of practitioners or academics.
The conception is followed by the *implementation of transfer media*, which contains the publishing of the final model, the questionnaires etc. for a company's own self-assessment.
The *evaluation* as the seventh step is supposed to assess the maturity model qualities. The evaluation is carried out by comparing the assessment results from companies with the expectations regarding the distribution of companies amongst the maturity levels. Details on how the comparison is carried out are not given. This step results either in the *model rejection or acceptance*. The acceptance targets the model continuation, which leads again to a re-iteration of the model with regard to the changing conditions

[4] Becker et al. [2009] name domain experts in the context of the construction model description but does not specify, if they are the ones who evaluate the model.

and therefore a continuous need of a model fitting. This reiteration of the construction process can start both with the origin problem definition (step 1) or the conception of transfer and evaluation (step 5). In case the model is rejected, no further steps are carried out. In case the model is supposed to be kept updated, Becker et al. [2009] names a regular validation as necessary without describing this step in further detail.

After describing the models for the construction of maturity models by De Bruin et al. [2005] and Becker et al. [2009], in a next step the models are compared in a first step for matchable phases, followed by the evaluation regarding their potential application for the construction of the Big Data maturity model.

4.1.3 Model comparison and evaluation

For further comparison, the model phases from De Bruin et al. [2005] are contrasted with the phases from the model by Becker et al. [2009] and matchable phases are carved out.[5] De Bruin et al. [2005] initial model *scope* is comparable with the phase *problem definition*. The following *design* phase contains similar aspects as the determination of the *development strategy*. The main *population* phase contains, with the exception of the *result test*, three out of four sub-phases of the *iterative maturity model development* by Becker et al. [2009]. The *result test* phase as the last sub-phase is a stand-alone phase in the De Bruin et al. [2005] model, which also contains aspects of the transfer evaluation and the overall evaluation. In this step, the models differ, as De Bruin et al. [2005] do not explicitly deal with the transfer of the resulting model in a practical context.
The *maintain* phase in De Bruin et al. [2005] can be found in Becker et al. [2009]'s approach as well, in terms of an iterative proceeding after a positive final evaluation, reasoned with the reference that *"maturity models inherently become obsolete because of changing conditions, technological progress or new scientific insights"*.

The comparison reveals that the models contain similar phases and do not show significant differences regarding the associated phases, yet different phases can be identified, described along three main aspects of

- different assumptions

[5] In the following description, the notions *phase* and *step* are used synonymously.

- different emphasizing of phases
- orientation of the model by Becker et al. [2009] on the design science research criteria by Hevner et al. [2004]

which will be explained in detail.

i) Different assumptions

The model by De Bruin et al. [2005] has in the beginning a strong stakeholder focus. The identification of both the model's target audience as well as the related objects that will participate in the construction process are analyzed in detail as it is assumed that differences in this group have to be anticipated for both the initial model construction as well the future model deployment. At the same time, the model lacks in the demonstration of the need for a maturity model in the beginning, contrary to the model by Becker et al. [2009].

The model by Becker et al. [2009] has a focus on model stakeholders, but not as a guiding aspect. The stakeholder focus comes in after the initial model construction in terms of the conception of the transfer and evaluation phase. Following the construction model, Becker et al. [2009] assume that the target audience of the maturity model does not influence the initial model construction and reduces its effects on the transfer and communication of the developed model.

ii) Different emphasizes of phases

Although the model by Becker et al. [2009] emphasizes the documentation of each phase, its description of the content remains partly reduced, compared with the model by De Bruin et al. [2005]. This accounts particularly for the aspect of the initial model construction as well the model validation, which is described in detail by De Bruin et al. [2005]. In contrast, Becker et al. [2009] focusses on the model evaluation instead of validation. De Bruin et al. [2005] give an overview of different approaches for the model population, categorized into top-down and bottom-up approaches, and discuss briefly its respective application condition regarding the maturity of the topic. With regard to the multiplicity of available population approaches, this categorization fosters the identification of a suitable method.

As the evaluation, in how far the understanding of maturity in an practical environment

is represented correctly in the model is one goal of this work, the evaluation step will be emphasized as well in the later Big Data maturity model construction as well.

iii) Orientation towards the design science research principles

The construction process proposed by Becker et al. [2009] is oriented towards the design science research principles by Hevner et al. [2004]. This results, in contrast to the model by De Bruin et al. [2005], in a documentation of each phase, which improves the comprehension of the overall construction process and the understanding of the results from each phase. None of these documentations can be found in the model by De Bruin et al. [2005]. It is limited to the description of possible outcomes of each phase, a documentation is not carried out. Furthermore, the comparison of existing maturity models with the pre-assigned problem definition cannot be found in the model by De Bruin et al. [2005]. Consequently, each model is a new development and therefore cannot benefit from both the findings and weaknesses of existing models. Following this approach, the model developer runs the risk of repeating blemished processes and recreating already existing knowledge.

Summing up, both of the construction approaches contain aspects that can be used for the construction of a Big Data maturity model, but none of the models can be used in their original versions.
In general, the model by Becker et al. [2009] offers two advantages compared with De Bruin et al. [2005]:

- The model follows the design science research approach. Following these criteria helps to avoid criticism regarding the lack of a scientific foundation.
- The more detailed description of Becker allows a better understanding of the different phases and the needed fitting to the construction of the Big Data maturity model.

Nonetheless, as described in the comparison of the two models, Becker et al. [2009] can not be used without any fittings as the order of different phases, e.g. the consideration of stakeholder needs. Therefore, the model by Becker et al. [2009] is used a basis and will be enhanced by several aspects originating from the model by De Bruin et al. [2005].

In the following section, the construction approach used in this thesis will be developed.

4.2 Development of the construction model

As described in Chapter 3 and 4, basis for the development of a maturity model is an underlying construction model which acts as a guidance for latter maturity model construction.

The construction approach is supposed to comply with the following criteria:

- Congruence with the seven principles from design science research by Hevner et al. [2004] in order to reach a problem-solving artefact, which complies with the requirements of a contribution to research.[6]
- The developed model is ought to be *reusable beyond this thesis* and therefore hold a certain level of generality. Consequently, each construction step, the intermediate results and generated artefacts have to be described in a sufficient granularity.[7]
- The model is not supposed to limit the applicable methods for the model population, as De Bruin et al. [2005] showed that the available methods change with the maturity of the subject in focus. This criterion fosters both the demand for a multi methodological construction approach by Hevner et al. [2004] as well as the reuse of the model at a later point in time.
- As carved out in Chapter 2, Big Data contains numerous different topics. Therefore, the scoping has to be emphasized in order to clarify, which dimensions and aspects of Big Data are covered by the maturity model.
- With regard to the one research goal targeting the model evaluation, the construction model has to have one emphasis on the model evaluation, as described in Chapter 1.

Based on this demands, in the next section the construction model is developed. For each step, the goal of the step, potential methodologies and the targeted outcome are

[6] Details of the criteria by Hevner et al. [2004] will explained in section 4.3.

[7] Regarding the demand for generality, Winter [2008] states *"The trade-off between the level of solution generality and the problem scope is addressed by situational artefacts. Since a unique solution is not applicable to many problem situations without generalizations, which diminish its solution power, situational adaptations should be incorporated in order to preserve application value (i.e., solution specificity) while covering a broad problem scope."* Therefore, one goal of the construction process is to allow this trade-off, allowing to develop a maturity model independent of an industry and size.

displayed.

The representation of the construction model is based on flow chart notation to visualize the different loops and give a clear differentiation between sequence and outcome [Hering, 1984].[8] Each step is assigned a number that is used as reference during the construction model description.

4.2.1 Step 1 - Definition of problem and scope

Following Becker et al. [2009], the construction model which is used in this thesis starts with an initial *definition of problem and scope*. This phase contains the formulation of the research goal, based on an existing problem and the related object of investigation. The notation research *goal* instead of question is used, since from the authors point of view, the result of the research process is an artefact rather than an answer of a specific question.

Furthermore, the *target group* of the resulting model (which can be within one company or industry-wide) and the *stakeholders* in the development process, similar to the approach by De Bruin et al. [2005], are defined. Additionally, the need for the Big Data maturity model has to be described. The caused, articulated need for a maturity model fosters future model acceptance and the willingness of companies to participate in the development process.

4.2.2 Step 2 - Identification of dimensions

In the next phase, *dimensions for the object of analysis*, in this case Big Data, are identified.

The dimension identification is emphasized by using an individual phase in contrast to the construction models by Becker et al. [2009] and De Bruin et al. [2005]. In doing to, the construction model is applicable as well for novel topics (e.g. Big Data), for which no common understanding of the subject in focus and associated topics exist. Depending on the model's target group, defined in the step before, methods for the identification of the dimensions can be a structured literature review, both manual as well as generative,

[8] The original symbol for input/output has been changed to a document symbol in order to improve the readability.

in order to develop a sound theoretical understanding of the topic [Webster and Watson, 2002] [Hansmann and Niemeyer, 2014]. Furthermore, the extend of information available on the topic in the literature as well as the already existent practical knowledge influences the methods which can be used for the later model population [De Bruin et al., 2005]. The results can be enriched and validated by conducting expert interviews or discussing the initially identified dimensions with focus groups.[9] This step of the process can be simplified in the case that a rich literature corpus on the respective topic exists and that it contains established descriptions of dimensions of the subject in focus.

4.2.3 Step 3 - Comparison with existing Maturity Models

The next step contains a *comparison with existing maturity models*. The comparison is targeting both the methodological basis of the model construction as well as the model results. With respect to methodology, existing maturity models can be analyzed regarding the theoretical foundation in terms of the existence and design of an underlying construction model and the applied methodologies for the model population. The identification and comparison can be carried out by a structured literature review and by using the literature corpus that has been gathered the step before.[10] A suggested frame for the analysis of the potentially existing maturity models can be found in table 4.1 and has already been used in Chapter 3.

Regarding the content, the comparison is supposed to result in both the identification of aspects which can be transferred to the Big Data maturity model (validation of identified dimensions) as well as the identification of Big Data relevant aspects which are not covered by existing models yet. The identification of white spots fosters the relevance and argumentation of the need for a new model. Referring to the demand for re-utilization of the construction model, with regard to extent of the contribution, this phase will increase in relevance with the growing number of Big Data maturity models available in the future.

[9] The focus group plays a relevant role in this construction model. The members of the focus group with the different backgrounds (e.g. studies, extend and field of practical experience) have an influence on the later outcome, which can be a limitation as it will be described in Chapter 6. Nonetheless, with regard to the practical relevance and orientation of maturity models, a focus group can help to transfer the practical understanding of maturity into the final model.

[10] The author is aware of the potential challenge, that for the topic in focus no maturity model already exists, especially for emerging topics. In this case, the literature review has to be extended to nearby topics as shown in the application of the construction model in Chapter 5.

TABLE 4.1: Framework for the analysis of existing maturity models as part of the Maturity Model construction

Author	*Subject of Analysis*	*Description of underlying construction model*	*Applied (Population) Methods*	*Transferable results*
	Describing the used dimensions	Origin (practice or research) and number of steps	Describing methods and naming related steps in the construction	Description of model elements which can be transferred to the model to be developed

4.2.4 Step 4 - Select design level and methodology

The next phase, *Select design level and methodology*, inherits several sub-phases such as i) the selection of model dimensions based on the results of phase two, and ii) the selection of the population method (top-down/bottom-up) that will be the basis for the model to be developed. The selection of the design level in terms of dimensions depends on the model *scope* defined in the beginning of the model construction. An initial dimension identification can be carried out based on existing literature and expertise knowledge. The identified dimensions are discussed with a focus group, carving out in how far they represent the object in focus from a practitioner's point of view. The early integration of practitioners fosters the overall quality and practical relevance as explained before. With regard to the differences of Big Data between industries, the members of the focus group have to be selected based on the scope and target of the maturity model defined in phase one as well.
The decision regarding the methodology for the population depends on the results from

- the scope, defined in phase one,
- the knowledge in science and practice which could be gathered in phase three, and
- the maturity of the subject in focus.

As explained before, for rather mature disciplines, a bottom up approach is used. The more immature the topic is and the less information that is available from existing

maturity models, the more suitable a top-down approach is [De Bruin et al., 2005]. Both approaches are discussed in detail in the following population phase.
The selected design level, the dimensions, are discussed with the focus group and rejected in the case a wrong level of detail has been chosen. In case of an approval, the model population follows.

4.2.5 Step 5 - Model population

Next follows the *population phase*. The order of the contained sub-phases depends on the selected population method. In case a top down-approach has been selected, the dimension-individual and the aggregated maturity levels can be defined based on Delphi studies, expert interviews, or creativity techniques [De Bruin et al., 2005].
A defined maturity level at this stage is represented by several sentences, describing the core characteristics of each maturity level, both on a dimensional level as well as on an aggregated level. These levels are used as a basis for the further identification of topics and related measurements. A topic is part of a dimension, e.g. the structure of the data warehouse can be a topic of a dimension called *IT infrastructure*. The identified topics are discussed subsequently with the focus group regarding comprehensiveness and explanatory power, followed by the identification of measurements per topic, which allow the measurement of maturity (again using the same methods mentioned at the beginning of the population phase plus relevant literature). Measurements have to target the existence and characteristics of certain processes, roles etc., representing different capabilities and related stages of maturity of one topic.
If a bottom-up approach has been selected, as a first step, the topics and related measurements for the maturity measurement are identified. In a second step, the items are assigned to different levels. Based on the assigned measurements per level, the maturity level is defined. The assignment of measurements can be carried out using qualitative (e.g. focus groups, expert interviews) or quantitative approaches (e.g. calculation of the item difficulties).[11]
With regard to the measurements' relevance to the result of the maturity model, another testing phase is included. In case a quantitative bottom-up approach is pursued, a questionnaire is needed to gather the relevant input data. Therefore, the questionnaire,

[11] A detailed description of the approach used for the assignment of items to different maturity levels can be found in Chapter 5.

containing the literature-based identified items should be discussed with the focus group and pre-tested.[12] The gathered survey data can be processed with quantitative methods from the social science. The goal is to calculate the difficulty of each item. The higher the calculated difficulty, the higher the associated maturity level is. Based on the difficulty, the items are clustered on different maturity levels.
The resulting initial model is evaluated in a next step. As the evaluation is one focus of the thesis, it will be described in more detail in the next section.

4.2.6 Step 6 - Model evaluation

4.2.6.1 Model evaluation - Theoretical foundation

Before describing the evaluation approach in the construction model in detail, the aspects of evaluation and validation as two different topics in information system research and particular in design science research, are described. Based on this theoretical basis, the evaluation approach used within the construction model will be described.

The aspect of validation has been discussed in the past decades increasingly in the field of information system research [Straub Jr., 1989; Boudreau et al., 2001], as the discipline has been confronted with criticism regarding the lack of a solid validation. Following Balci [1998], validation can be described as *"building the right model"*. This contrasts the often mixed verification which is understood as *"building the model right"*.
Within the existing publications on validation from the field of maturity model research, different types of validity can be identified, primarily *construct validity* and *instrument validity*, whereas the first one consists of content and face validity [De Bruin et al., 2005; Boudreau et al., 2001]. Although validation has been increasingly emphasized in recent information system research, the analysis by Boudreau et al. [2001] reveals, that solely 23 % respectively 37 % of 193 publications in the top tier journals from the field of information system research test for construct respective instrument validity.
Validation is aiming at the reproducibility of results based on a described procedure model [Balci, 1998], in the case of maturity models the underlying construction model. From the authors' point of view, in the context of the maturity model construction, a

[12] The notion *measurement* is used synonym to the notion *item* in this work as the measurements are later translated into items in the questionnaire used for the data gathering.

validation of the resulting model is confronted with three major obstacles, making a full validation difficult to achieve:

i) Several process steps of the model construction are carried out in collaboration with members of the focus group and industry experts. As a first step, the results of the evaluation and fitting of the questionnaire during the interaction with the members of a focus group are based on the state of knowledge of the respective person, which may change in the course of time due to an increase in experiences.

ii) With regard to the relative character of maturity, the assignment of certain capabilities to a maturity level changes in the course of time as well, leading to the need of a model maintaining phase in the construction process. Consequently, the model construction process could lead to different results at a later point in time, resulting from an increase in companies' professionalism and technological development.

iii) In case a quantitative bottom-up approach is selected, the data for the model calculation are gathered from surveying companies' current status, regarding certain processes. Again, the course of time has an influence of the companies' capabilities, which in turn influences the data base and therefore may change the resulting model.

Summing up, the change of maturity in the course of time as well as the integration of individuals and companies in the model construction process has an influence on the resulting maturity model. Therefore, the reproducibility of the maturity model can only be achieved under several limitations that in turn reduce the applicability and practical relevance of the model. Consequently, the focus is on the evaluation of the maturity model instead of validation.

Evaluation, in a general sense, is understood as the systematic process applied to the targeted and goal-oriented evaluation of an object [Sanders, 2006, p. 25]. The execution of an evaluation is not only connected with an interest in gaining knowledge. An evaluation can serve furthermore as the documentation of effects.
In the context of design oriented information system research, evaluation is understood as the assessment of the output of the Design Science Research process, which can be

artefacts or IS Design Theories [Venable et al., 2012].

Existing design research processes contain phases focusing explicitly on evaluation instead of validation, e.g. the approaches by Peffers et al. [2007], March and Smith [1995] and Hevner et al. [2004] whose individual research steps can be grouped to the steps of *Build, Evaluate, Theorize,* and *Justify.*

Riege et al. [2009] is supporting the relevance of evaluation in the context of information system research by drawing a connection between evaluation and validation, stating that a constructed, not yet evaluated artefact does not represent a valid research result.

In order to achieve this "valid" research result, both the construction model as well as the results of the model application are evaluated.

Evaluation approaches in the field of design science research can be distinguished based on the evaluation of the research/artefact against the research gap or the real world problem [Bucher et al., 2008; Cleven et al., 2009]:

i) The artefact is evaluated against the identified research gap, the focus is on the evaluation of the accurate construction of the artefact, based on requirements defined before.

ii) The artefact is evaluated against (an expert of) the real world by applying the artefact to the real world problem in focus.

iii) The research gap is evaluated against the real world. This approach plays a subordinate role in the field of information system research and therefore will not be further pursued.

The focus of the thesis at hand is both on

- the evaluation **against the identified research gap** (= evaluation of the construction model itself) and
- the evaluation **against the real world** (= evaluation of the Big Data maturity model as a step of the model construction).

This approach goes beyond a sole focus on the construction process as demanded by Winter [2008].

The evaluation of the construction model against the research gap will be carried out in the end of Chapter 4. The evaluation of the maturity model against the real world as a step in the model construction will be explained in the following section.[13]

4.2.6.2 Evaluation against the real world

For the construction model at hand, a two-step evaluation approach has been developed. First, the initial model, which resulted from the model population phase, will be discussed with the members of the focus group. Subject of this discussion is the distribution of the items amongst the maturity models in case a bottom-up approach has been selected. This step is named *Evaluation of the initial model* (Step 6.1). The goal is to identify how the item difficulty, calculated in the population step before, and the resulting item order are congruent with the item difficulty perceived by the member of the focus group. In other words, the goal is to identify, if the focus group members would assign the items/measurements on the same maturity levels as it has been done during the population step.

At this point, an additional member should complement the focus group, in order to integrate the opinion of a person who has not already influenced the model construction process. Based on the input of the focus group, the model is fitted respectively.

Second, after the incorporation of the feedback, the resulting fitted model is applied and evaluated with the help of the focus group. This step is named as the *Evaluation based on the deployment of the fitted Model* (Step 6.2). The focus group should again be expanded for this step as some members of the focus group have already participated in the construction of the questionnaire as well as in the first model evaluation step (step 6.1), and therefore have already realized an understanding of maturity in the overall model. This previous knowledge can lead to a falsification of the evaluation results.

For the second evaluation step, construction step 6.2, every member of the focus group designates at least one company, he is familiar with, based on consulting projects. By that, it is expected, that he is able to evaluate the companies' capabilities and maturity

[13] With regard to the focus on the evaluation of the model against the real world, an evaluation of the resulting initial model from a statistical point of view based on the item fit [Reise, 1990] values is not in the focus.

in the field of Big Data. The goal is to have a selection of companies in all spectrums, ranging from immature to mature, in order to test the different levels of the maturity model.[14] The fitted model is used to evaluate these companies' Big Data maturity. Concurrent, the focus group member evaluates the same companies, he determined upfront. Therefore, for each selected company in step 6.2, two maturity evaluation exists, the model based and the expert-based.

Potential differences between the results from the model and the maturity evaluation of the industry expert are discussed in the next step in order to investigate if the potential differences result from i) missing or wrong measurements in the model from the experts' point-of-view or ii) the error-prone distribution of measurements on maturity levels.[15]

In the event that the model is rejected after these two evaluation steps, three potential starting points for the necessary model adjustments exist, depending on the expressed criticism during the initial model discussion and the model deployment.[16] In the case that missing measurements in the initial model or during the model deployment are identified, the process step *model population* is carried out again. In case the degree of granularity of the model is criticized, the model construction starts again at step 4, *Select Design Level and Methodology*. If criticism on a higher level targets the dimensions of the model, the *dimension identification* process is carried out again.

4.2.7 Step 7 - Documentation of the final model

In the case that the model has been approved after the second evaluation step (phase 6.2), the individual maturity levels will be defined in terms of several characterizing sentences based on the assigned items per maturity level. In case a top-down approach has been selected, these descriptions already exist but may have to be fitted in case several measurements have been reassigned during the model evaluation. The level of detail depends on the number of measurements that have been assigned to the individual

[14] In case the model has an industry focus, the companies which are used for the second evaluation step should belong to the respective industry as well.

[15] A quantitative model evaluation (e.g. calculation of the infit and outfit values [Marx et al., 2012] has deliberately not further pursued as the *Evaluation of knowledge* in Design Science Research follows the pragmatism (section 1.3), which is in this case the correct representation of the practical understanding of maturity.)

[16] The author has deliberately refrained from setting a numerical maximum difference between the experts and the models maturity evaluation, from which on the model is rated as rejected. If a model is rejected has to be decided from case to case. In this decision, both the measured difference as well as the reasons for this difference has an influence.

maturity levels and dimensions.
In the case a sufficient number of measurements exist per dimension and level, dimension-individual characterizations per level can be carried out.[17] By the end of this step, the final maturity model exists and can be further published.

4.2.8 Step 8 - Model maintaining

The final *maintaining phase* (Step 8) contains a regular model update with regard to the dynamics and the speed of development of the subject in focus, similar to the approaches by De Bruin et al. [2005] and Becker et al. [2009].[18] This dynamic becomes apparent just by looking at the change in volume of data that have been the focus in the past years and the increasing need for capabilities in the identified dimensions to cope with this development (Chapter 3). In contrast, the update frequency for a maturity model in an established discipline will be lower due to the reduced number of relevant developments. More details on the model maintaining are given in Chapter 5.
The maintaining process in this construction model starts again with the beginning of the construction, as the problem definition might have changed due to the dynamic of Big Data, in the time elapsed since the last model development. A start of the maintaining process at a later point in the maturity model would lead to a lack in re-considering the overall problem definition, which may have changed since the beginning of the construction process.

After the construction model (that will be applied in the following Chapter 5) has been developed, it will be evaluated against the identified research gap.

[17] Sufficient in this case means that the explanatory power of the number is sufficient in order to describe the level textual.
[18] Again, the author has deliberately refrained from setting a frequency for the maintaining process. This frequency for the model update depends on the subject in focus. Rather novel topics, e.g. Big Data, carry a need for a more frequent model update in order to keep the model as close to the development in the industry environment.

4.3 Evaluation of the construction model against the identified research gap

For the evaluation of the artefact - the developed construction model - against the identified research gap, the focus is on the correctness of its construction based on requirements defined in the forefront. For this evaluation exist different approaches as listed in table 4.2.

TABLE 4.2: Systematization of evaluation approaches for the evaluation against the identified research gap [Riege et al., 2009]

Evaluation Approach	Exemplary Utilization
Demonstration Example	vom Brocke [2006]; Gehlert [2007]; Klesse [2007]
Prototype Construction	vom Brocke [2006]; Gehlert [2007]; Braun [2007]
Prototype Application	Braun [2007]
Attribute based Comparison	Gehlert [2007]; Braun [2007]; Klesse [2007]
Meta-model based Comparison	Braun [2007]
Simulation	König and Weitzel [2003]
Survey	Gemino and Wand [2003]; Klesse et al. [2005]
Laboratory Experiment	Batra et al. [1990]
Field Experiment	Braun [2007]
Action based Research	Grütter et al. [1998]; Schwinn [2006]

For the evaluation against the research gap in the thesis at hand, an *attribute-based comparison* and the *survey-based comparison* could be used. The other evaluation methods cannot be used primary due to the character of the model to be evaluated. A construction model is not an artefact which can necessarily applied in the practical environment. Therefore, approaches as field experiments or action based research cannot be applied. They are originally used to evaluate resulting artefacts instead of models in terms of construction processes which lead to artefacts.[19]

[19]Reasons why the other approaches cannot be used: *Demonstration Example*: The application of the developed artefact at a fictional company is not suitable as it is a process model which has no company-relation. The *construction or application of a prototype* is not pursued as a software-based implementation of the construction process is not in the focus of this thesis. *Simulations* cannot be applied since the application of the process model cannot be simulated. A *survey* could hardly be used as the respondents might not necessarily be aware of the criteria of correctness. A *laboratory experiment* does not play a role as the creation of laboratory conditions is challenging for construction models, comparable to the *demonstration example*.

Consequently, for the evaluation against the research gap, the attribute based comparison is selected, as already published sets of established attributes for the evaluation exist.

The attribute-based evaluation in this research consists of two parts. As a first step, the developed construction model is tested against the criteria for design science research by Hevner et al. [2004]. Those criteria act as a basis for the model construction and therefore represent the mentioned pre-defined requirements.[20] The results of the evaluation based on the criteria by Hevner et al. [2004] can be found in section 4.3.1.
The second part of the evaluation against the research gap is oriented towards Becker et al. [1995] principles of general accepted modelling. Those principles have been developed to improve the quality of models in the field of information systems. The results of this evaluation can be found in section 4.3.2.

4.3.1 Evaluation against the identified research gap - The principles of Design Science Research

For the first part of the evaluation against the identified research gap, the developed construction model is evaluated based on the principles of design science research by Hevner et al. [2004] in order to test the construction model for its theoretical foundation. The principles have been developed originally in order to act as guidelines for research in Information Systems. The model by Becker et al. [2009] - which has been taken as a basis of the construction model presented in this chapter - is already evaluated based on these principles. Therefore only those aspects of the construction model at hand are re-evaluated, that are different to the model by Becker et al. [2009].

- *Iterative development procedure*

  In the course of the construction, several evolutionary stages of the model have been developed, executing multiple testing and fitting phases, based on the discussion with the members of the focus group, the industry experts. Their potential

[20] Riege et al. [2009] do not mention these criteria in the list of evaluation methods. They are focusing in their systematization of evaluation approaches on the European Information System research instead of the Anglo American understanding of information system research and mention numerous *"significant differences"* between those two fields. At the same time, it is stated that those are not relevant for the aspect of evaluation. Therefore, the approach by Hevner et al. [2004] with an Anglo American background, can be applied in this thesis, which belongs to the European Information System Research.

negative evaluation and the resulting rejection can result in a loop back to an earlier construction phase until the model is approved. This re-iteration assures that only such a model is finally documented, which has been completely approved. In addition, as a potential first iteration already exists at stage four, potential errors are not carried through the whole model construction.

By cooperating with different industry experts from different companies, the danger of a one-sided influence on the model design is reduced [Yousuf, 2007].

- *Model evaluation*

  The model is tested multiple times and evaluated based on the use of a focus group and an industry deployment phase, with both the initial model as well the fitted model being tested. By comparing the model's maturity evaluation of different companies with industry experts' evaluations, the realistic representation and understanding of real-world maturity in the model is evaluated.

- *Multi-methodological procedure*

  Several research methods can be applied in the course of the construction process, such as quantitative methods in terms of text mining for the identification of Big Data dimensions, and quantitative approaches from the field of test theory for the model population. Additionally, qualitative research methods as expert interviews are applied during the model evaluation phases.

- *Identification of problem relevance*

  The different maturity levels of Big Data and related capabilities are of interest to both scientists and practitioners. Maturity models as a sub-form of reference models are developed to support the design of information systems. Therefore, they can have an influence on managerial decision-making, e.g. for the development of organizational structures, budget allocation etc. As this managerial decision making is located at the beginning of a change process, e.g. towards a more data driven company, the ex-ante capability evaluation using a maturity model gains in relevance. Those aspects have to be worked out in the beginning of the model construction as an early justification of the model building.

The remaining aspects by Hevner et al. [2004] (*Comparison with existing maturity models, problem definition, targeted representation of results, and scientific documentation*)

are not affected by the changes made to the model by Becker et al. [2009] and therefore do not have to be re-evaluated.

4.3.2 Evaluation against the research gap - The principles of general accepted modelling

Besides the evaluation based on the principles of design science research (section 4.3.1), the construction processes are tested for coherence with the principles of general accepted modelling. These belong to the the field of business information system engineering, derived by Becker et al. [1995] from the generally accepted accounting principles [Leffson, 1987]. The contained *conventions* act as guidelines during the modelling process, intended to improve the model quality.
The evaluation of (process-) models based on the principles stated above has been carried out numerous times and therefore has been established as a community standard [Schütte, 1998].
The six normative oriented principles (table 4.3) have been developed as a basis for the evaluation of the model construction and hold a "[...] *customer oriented understanding of model quality.*" Customers in this case are individuals that will apply the construction model for the development of a maturity model, which can be both scientists and practitioners.

The six principles cannot be applied completely as the construction model - developed in this chapter - differs in a number of aspects from the understanding of models in the field of business information system engineering:

- *Lack of an overall context that the model has to be placed in*: As the model is neither part of a general company or process model, nor is it supposed to be compared with other construction models, the evaluation aspects targeting ***Comparability*** do not have to be taken further into account. The same accounts for the aspect of ***Profitability***. Model refinements do not have to be evaluated in the forefront regarding the potential resulting benefit, as the model is not applied in a business context with a financial goal. The aspect for profitability becomes more relevant, when the focus is on the model application.

TABLE 4.3: Principles of general accepted modelling [Becker et al., 1995]

Criterion	Description
Correctness	i) Syntactic correctness: Compliance with the rules of the modeling language ii) Semantic correctness: Consensus amongst model users regarding the correctness of the overall model as well as individual parts of the model.
Relevance	The focus is on the modelling of only those circumstances that are relevant for the underlying modelling purpose.
Profitability	Determination of degree of the model refinement depending on the relation between the use and costs of the additional refinement.
Clarity	Aiming at the readability, clearness, and understandability of the model by focusing on appropriate hierarchy, appropriate layout, and receiver-oriented filtering.
Comparability	i) Processes in the real world which are perceived identically should be identically modelled as well. ii) New, company-oriented models should be set-up on similar constructs, using the existing company-oriented models in order to foster the meta model transformation.
Systematic Structure	Identical objects, utilized in different models (data model, process model, etc.), should be used correspondingly in order to achieve consistency of the overall model.

- *Overall context*: The ***Systematic Structure*** is not relevant as the construction model is not part of an overall model, therefore a consistent use of notations between different models does not have to be pursued.

After eliminating those aspects due to the different type of model in focus, the remaining aspects are ***Correctness, Clarity,*** and ***Relevance***.[21] These aspects contribute to an understandable description of the individual process steps including the resulting artefacts. The demand of each criterion and how it is fulfilled by the developed construction model will be explained in the next step:

The aspect of ***Correctness*** is targeting the syntactic and semantic correctness. Although no underlying modelling language in terms of e.g. EPKs exists, the description

[21] In contrast to the work by Mettler [2010], the construction process model instead of the resulting maturity model is evaluated in the context of the evaluation against the identified research gap. The author is convinced that the construction model, despite its differences as explained before, can be understood as a process model as targeted by the principles of general accepted modelling. This does not account for the resulting maturity model, as it does not follow a process logic.

of the construction process is based, comparable to Becker et al. [2009], on the notation DIN 66001, fostering the overall correctness of the model.
The aspect of ***Clarity*** is contributing as well to the understandability of the model. A sufficient readability, clearness, and understandability have been achieved, as each step has a distinct name and the related actions for each step are documented, containing both the needed input and the potential output as shown in figure 4.1.
Relevance is focusing on the steps, which are taken into account for the model. The demand is that only those aspects should be integrated in the model, which are relevant for the underlying modeling purpose.
The demand is fulfilled, as in case one of the construction steps would have been left out during the construction, the construction of the maturity model would not be possible. In order to clarify the relevance, each step and its contribution to the model development has been explained in detail in this chapter.

Altogether, the presented construction model follows the principles of design science research and follows the principles of general accepted modelling. The model has been evaluated successfully against the identified research gap and tested for the correctness of construction. Therefore, the construction model represents a contribution to design science research [Hevner et al., 2004] and will be applied for the model construction in Chapter 5.

4.4 Main chapter results

A significant number of existing maturity models lack a theoretical basis, targeting i) the overall model construction process, ii) the methods applied for the model population, and iii) the evaluation of the resulting maturity model (Chapter 3). This is reflected in a generalized maturity model critic, targeting the lack of a sound theoretical foundation [Biberoglu and Haddad, 2002].
As a first step towards the development of a maturity model construction model with an appropriate theoretical foundation, different construction approaches have been compared. The approach by Becker et al. [2009] has been selected as a basis for the development of the construction approach applied for the construction of the Big Data maturity

model and enriched by several aspects from the construction model by De Bruin et al. [2005]. As both models do not completely fulfill the demands based on the research goals (Chapter 1 with regard to i) the object in focus and ii) the emphasis of the model evaluation), the fitted construction model emphasizes the description of the maturity object in focus and the evaluation of the developed artefact. The governing aspect for the construction approach has been the correspondence with the principles of design science research by Hevner et al. [2004], resulting in a sound theoretical foundation of the resulting model.

Another emphasis within the construction model is on the evaluation of the resulting maturity model. Consequently, a two-step evaluation approach has been developed.

In the first step, the model construction process was evaluated against the identified research gap based on the principles of design science research in order to comply with "[]*...the requirements for effective design-science research.*" [Hevner et al., 2004]. Additionally, the model has been evaluated successfully based on the principles of general accepted modelling [Becker et al., 1995]. This first evaluation step targets formal and theoretical aspects.

The second part of the evaluation is focusing of the evaluation of the maturity model to be developed against the real world. This is carried out during the actual model construction (step 6.1 & step 6.2). By i) discussing the initial model with the industry experts of the focus group and ii) applying the fitted model to companies and comparing the results with the industry experts maturity assessment, a correct representation of maturity in the practical context is supposed to be reached.

After describing the concept of Big Data (Chapter 2) and giving an overview about existing maturity models in the field of Big Data (Chapter 3), the development path towards a maturity model - the construction model - has been described in Chapter 4. This developed construction model will be applied in the following Chapter 5 in order to develop the Big Data maturity model.

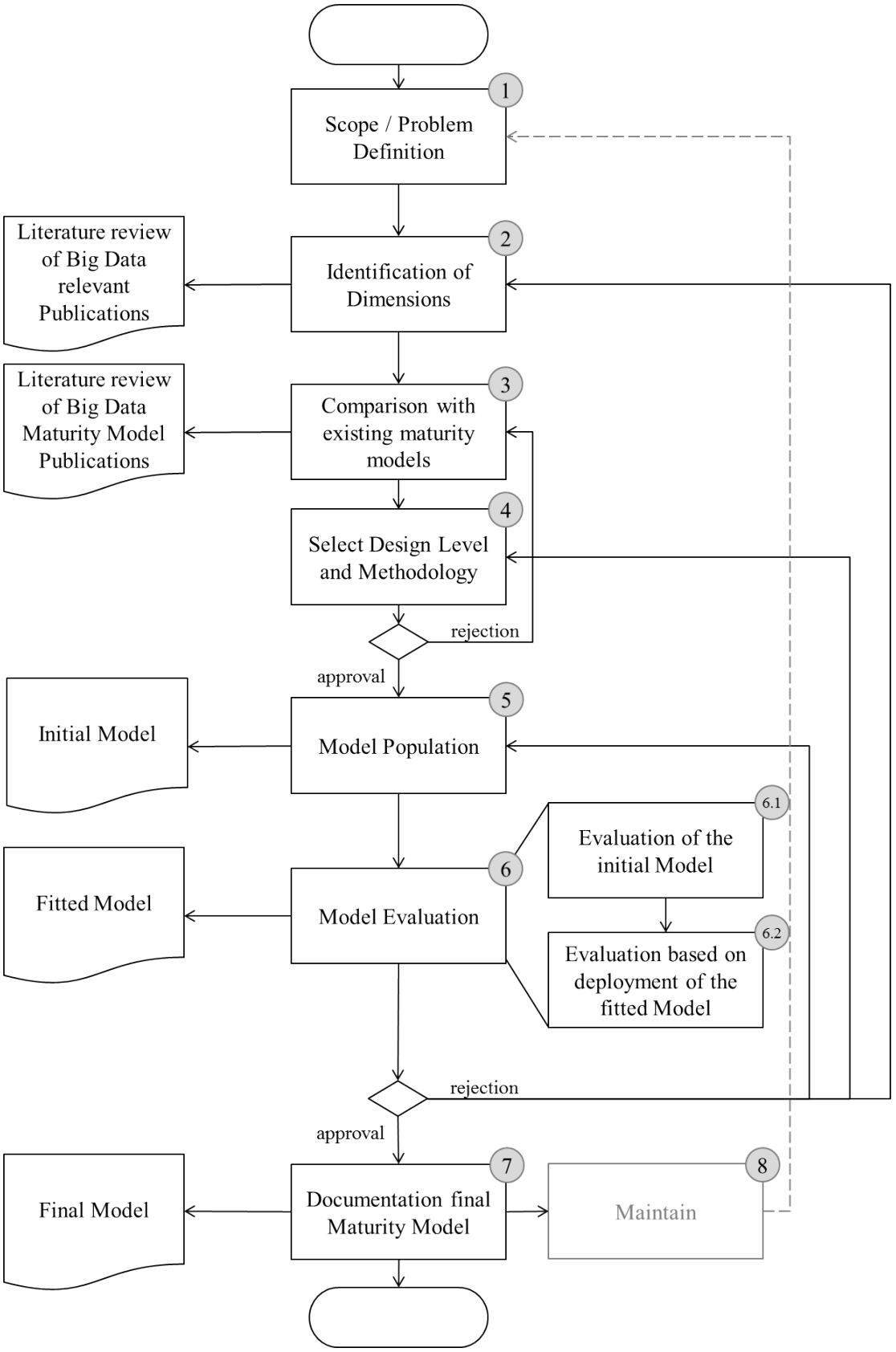


FIGURE 4.1: Developed Maturity Model construction process for the Big Data Maturity Model construction in this thesis

Chapter 5

Application of the construction model - Development of the Maturity Model

In the last section, a construction model incorporating the goals described in Chapter 1 has been developed. The model based on the work by Becker et al. [2009] and enriched by aspects from De Bruin et al. [2005] has been evaluated against the identified research gap in order to ensure that its application results in an outcome - the Big Data maturity model - with a sound theoretical foundation.
In the following chapter, the developed construction model is applied in order to construct the Big Data maturity model. Each subsection describes one development step.

As it has already described in Chapter 4, a focus group plays an essential role throughout the model development process in this thesis. Therefore the group members' characteristics are described in the forefront of the maturity model construction. The focus group consists of six industry experts and has been enhanced in the course of the maturity model development. Table 5.1 provides an overview about the members' backgrounds and the phases they contributed to.
One potential bias that can occur in case a focus group is participating in multiple construction steps within the construction process for one maturity model is the phenomena of a self-fulfilling prophecy. Members of the focus group, who have already

TABLE 5.1: Characteristics of the focus group members and the construction steps they have participated in

Expert	Occupation	Years of Experience	Phases participated in
No. 1	CEO IT Service Provider	20	2
No. 2	Senior IT Consultant	8	2, 4, 6.1, 6.2
No. 3	Senior IT Consultant	8	2, 4, 6.1, 6.2
No. 4	Senior IT Consultant	12	2, 4, 6.1, 6.2
No. 5	VP IT Service Provider	14	6.1, 6.2
No. 6	Senior IT Consultant	7	6.2

participated in earlier construction steps have gained previous knowledge. This can lead to a bias during the model evaluation, as they are already familiar with the topics and measurements in the focus and have influenced their distribution among the maturity levels. Thus, as it is a critical factor to avoid this, for both evaluation steps (6.1 and 6.2), the focus group is enhanced by one person per step in order to integrate a further perspective and knowledge regarding Big Data in the model construction process.
The focus group consists of consultants with cross-industrial experiences. This background supports the elimination of a potential model bias resulting from an unequal distributed knowledge throughout the focus group as already mentioned.[1]

For a better guidance, at the beginning of each of the following subsections, a summary, containing the goal, method, and result of the described construction step can be found.

5.1 Definition of problem and scope

Problem Definition

As described in Chapter 3, based on the existing maturity models from the field of Business Intelligence, the context of Big Data inherits a multiplicity of challenges for companies in different fields, amongst others the development and fitting of company-internal processes to data analysis application, the integration of analysis results into

[1] As all of the focus group members are male, the language in the context of the focus group is not gender-neutral.

TABLE 5.2: Contribution Step 1 - Definition of problem and scope

Goal	*Method*	*Result*
Identification of the problem the maturity model is designed to solve; Definition of the model scope in order to develop a model with a practical relevance, setting the frame, carve out the models' target group	Literature review; Discussions with Focus group members	Problem: Companies struggle in identifying and building up capabilities needed to deal successfully with the massive influx of data; Scope: The model is supposed to be cross-industrial without a focus on specific applications or departments; Target group: Model is of interest both for scientists as well as practitioners from different industries

existing workflows, the measuring of success of applied analysis techniques, and the implementation of data quality management [Bizer et al., 2011; George et al., 2014; Kwon et al., 2014].

With the increasing relevance of data as a production factor, data analysis becomes a competitive factor [McAfee and Brynjolfsson, 2012]. Consequently, companies increasingly feel the need to build up capabilities in order to handle and utilize the massive influx of data. At the same time, companies are insecure regarding their current capabilities and those they have to build and improve due to the heterogeneity of potential fields of application, the marketing-driven selling of technological solution, etc. An evaluation of companies' capabilities can be used as a basis for the identification of potential improvements in order to gain a competitive advantage [Fisher, 2004]. One instrument for this evaluation is a situational artefact, the maturity model, [Winter, 2008]. As maturity models have proofed themselves both in science and practice as suitable for the evaluation of capabilities, they are selected as the appropriate tool.

Scope

Next, the *scope* of the model to be developed is defined. Following the differentiation by De Bruin et al. [2005] the model is a *general model*. It has no focus on data analysis in a specific application context, neither on an industry department or company size. That means that the resulting model can be deployed at every company, independent of its size, industry, or purpose.

The intended Scope is to cover different degrees of maturity respective capabilities of companies from different industries that are potentially confronted with Big Data. The selected level of abstraction of the model is needed due to the partly novel character of the field of Big Data (see Chapter 2). At the same time, such an broad level offers great potential for the subsequent maturity model application. Despite the existing overlaps of Business Intelligence and Big Data (Chapter 2 and 3), the goal is to incorporate aspects that are Big Data specific, as it will be exemplified in section 5.5.

Additionally, the broad focus, resigning an application or industry focus is potentially fostering the interest of a greater number of companies and industry experts to participate in the model construction process. This interest is relevant, as the model to be developed represents a new territory. A discourse with a wide range of practitioners is seen as critical to evaluate and improve the initial model, fostering the overall relevance and rigour.

Following the developed construction model in the previous section, the first phase (Problem definition and Scope) also has to indicate the definition of the *target group*. The following model construction and the resulting maturity model have two different target groups. The construction process is primarily of interest for the scientific community as it is methodological driven. Its description is not of value for companies as it does not hold any insights for the evaluation of their capabilities in the field of Big Data. The resulting model however is of interest both for i) practitioners, especially consultants and decision makers from relevant fields, such as organizational development, IT, BI and similar, as it can be used for the as-is evaluation of companies in order to derive potential fields for improvement. Additionally, it is of interest for ii) scientists as it is a contribution to the field of design science research and Business information system engineering.

5.2 Identification of dimensions

Criticism of recent maturity models targets the inaccurate use of describing dimensions in order to provide the user with a deeper understanding of the subject in focus [Becker et al., 2009; Mettler and Rohner, 2009]. Thus, this step gains in relevance with regard to the depth and breadth of the topic Big Data. As mentioned in section 2.2, Big Data can be structured along the dimensions Data, IT infrastructure, Methods, and Application.

TABLE 5.3: Contribution Step 2 - Identification of dimensions

Goal	*Method*	*Result*
Identification of dimensions that can be used to structure the topic of Big Data; Dimensions act as a basis for the advancement of the model construction	Structured literature review based on publications in the field of Big Data	Big Data can be structured along the dimensions of Data, Infrastructure, Method, and Application [Hansmann and Niemeyer, 2014]

These dimensions are taken as a starting point and will be discussed with the focus group in construction step number four, the selection of the design level.[2] [3]

5.3 Comparison with existing Maturity Models

TABLE 5.4: Contribution Step 3 - Comparison with existing maturity models

Goal	*Method*	*Result*
Identify already covered aspects and shortcomings of maturity models in the field of Big Data in order to derive aspects for the model to be developed	Structured Literature Review	Relevant aspects can be found in maturity models from the field of BI, DQM, and Corporate performance Management

The shortcomings of recent maturity models with relevance for Big Data, primarily from the field of BI, have been discussed in detail in Chapter 3. Although existing models are covering several aspects of relevance for Big Data as well, such as the distribution of analysis results, other aspects are not yet covered, compared with the characterization of Big Data (Chapter 2). This results from the partly different focus of BI and Big Data. In some cases BI is rather perceived as a support function, in contrast Big Data can be understand as a comprehensive concept, applied throughout the company and moreover influencing companies' work [Manyika et al., 2011; Hansmann and Niemeyer, 2014; Lane

[2] As the in-depth discussion of describing dimension has already been conducted in Chapter 3, this step is limited to the brief naming of potential dimensions.

[3] Even though thoroughly analyzed, the discussion with the focus group revealed some necessary adoptions, which will be explained in construction step 4.

et al., 2014]. Examples of missing topics can be found in the data dimension - the origin and source of the analyzed data. This aspect targets the multiplicity of available data sources as a characterizing element of Big Data. Furthermore, organizational aspects regarding the integration of analysis results in decision-making processes or the type of the data analysis task also have not been taken into account yet.
These missing aspects will be considered during the subsequent identification of potential topics and related items/measurements, supported by the input from the focus group members based on their project and industry experience.

5.4 Select design level and methodology

Table 5.5: Contribution Step 4 - Select design level and methodology

Goal	*Method*	*Result*
Definition of the level of detail of the maturity model and the model focus; Selection of the population approach (top-down or bottom-up) and the applied methods (quantitative, qualitative, mixed)	Discussion with the focus group	The model will be developed along the dimensions organization and data; A bottom-up approach is selected for the model population, applying quantitative methods from the field of social sciences and discussions with the focus group/expert interviews

Definition of the design level

The starting point for the definition of the design level are the dimensions identified in construction step two, *Data, IT infrastructure, Method* and *Application*. The aim is to test these dimensions with regards to their congruence for the demand of an application-independent model. Therefore, they have been discussed with the focus group in the forefront. For the *Data dimension*, being identified as a central aspect of Big Data, two main aspects have been raised during the discussion: the source and structure of the data as well as adjunctive capabilities. The main characteristics of data, source and structure can be seen as highly application-dependent in a first step, as they are usually gathered or processed for a certain purpose. However, considering the processes behind to collect, assess and use the various, increasingly external sources as well as

diversified structures, insightful conclusions about a companies' maturity in this field can be derived, independent of the actual application. As up to 80 % of the overall data volume are unstructured data [Ziegler and Dittrich, 2007; Grimes, 2008], companies need to develop appropriate skills and ways of handling. Irrespective of the data's later use or origin deployment, associated processes and capabilities have to be implemented, such as for the aspect of data quality management. Being a success-critical function in the Big Data context [Kwon et al., 2014], the meaningful identification, evaluation and combination of data gains in relevance. Additionally processes to identify and process relevant data sources and implement appropriate solutions have to be established.
Thus, even though data is most likely gathered or used for a certain purpose or application, the data dimension has been clearly reaffirmed by the focus group as being congruent for an application-independent model. The knowledge of which data a company gathers and in which way these are processed, provides insights about the companies' analytical processes and capabilities and thus allows the identification of numerous further adjunctive topics.

In the context of the processes and utilization around data quality management, a second potential dimension was raised during the discussion with the focus group, containing aspects regarding standardization of analytic-related processes. This aspect is closely related with governance, which plays a major role in the field of BI [Zimmer et al., 2012]. In contrast, it can only be found loosely in recent publications in the Big Data context [Tallon, 2013; Soares, 2012], therefore it could not be identified in the research by Hansmann and Niemeyer [2014]. One potential reasons for the low number of publications is the early stage of the topic Big Data and the lack of further published experiences with standardization [Kaisler et al., 2013].
However, in order to incorporate the aspect of processes and standardization, comparable to different Business Intelligence maturity models [Dinter, 2012], an *organization dimension* is used as a second dimension, as the aspect of organization is perceived as one major aspect within the overall Big Data concept [McAfee and Brynjolfsson, 2012]. The focus group members argued that especially organization and related process topics represent the challenges that are posed in Big Data projects, such as the integration of analysis results in the related business processes.

As indicated in Chapter 2 and section 5.2, the *application* dimension does not fully

coincide with the goal of an application-independent model, following the results of the discussion with the focus group. It has been pointed out that the diversity of the potential applications and the early state of Big Data does not allow the identification of applications and related measurements, which can be generalized for the maturity model. Consequently, the *application dimension* has not been further pursued.

Furthermore, the application-dependency of the dimensions *IT infrastructure* and *Methods* does not allow to further pursue these dimensions either. Depending on the application purpose, different hard- and software solutions can be found within companies, such as Hadoop. The increasing dissemination of this framework, fostered by its distribution as an Open Source software for the parallel processing of data, is linked with the overall rise of Big Data [Argawal et al., 2011]. Nonetheless, its existence cannot be taken as an indicator for maturity as its application does not necessarily stand for its productive use.[4]

The argument against the method dimension follows the same logic. Based on the structure of the data to be analyzed (structured, unstructured) and the type of analysis (e.g. monitoring or prediction), different sets of methods are available. Not all of the resulting potentially applicable methods can be covered by one maturity model. Additionally, mathematical methods and their applications are subject to hypes. As the interest in different methods changes over time, its application in companies can change respectively [Bennett and Campbell, 2000]. Therefore this aspect is not suitable for the measurement of maturity as well in the context of a first, high-level maturity model.

Summing up, the conducted discussion led to the adoption and selection of the dimensions *data* and *organization* as a basis.

Definition of methodology

Following the work from De Bruin et al. [2005] the selection of the population method depends on the maturity of the topic in focus, proposing the application of top-down approaches for young, immature disciplines and vice-versa, the application of bottom-up

[4] One member of the focus group gave an advancing argument. Taking an e-commerce company that analyzes unstructured data as an example, he argued that the e-commerce company may have a different IT infrastructure compared with an investment good company that primarily deals with structured sensor data in a manufacturing environment. Although both companies may have the same maturity level with respect to Big Data, they yet differ in the applied infrastructure. Therefore, the abstraction level of infrastructure would need to be on such a high level that this dimension would not lead to any further insights.

TABLE 5.6: Contribution Step 5 - Model population

Goal	*Method*	*Result*
Identification of measurements, covering topics that belong to the *organization* and *data dimension*; assignment of items to different maturity levels	Method from the field of Item Response Theory (fitted Birnbaum model), Data gathering based on a questionnaire, Discussion with the focus group	17 topics with 3-5 items are identified; the majority of the items belong to the organization dimension; the items are assigned based on the calculated item difficulty to the individual maturity levels, resulting in an initial maturity model

approaches for older, mature disciplines. As described in Chapter 2, Big Data holds both novel as well as established aspects. Therefore, a distinct selection of the top-down or bottom-up approach is not possible yet. With regard to the research goal (Chapter 1 - the evaluation in how far a quantitative approach can be selected for the model construction with both novel and established aspects - a quantitative bottom-up approach is selected. Accordingly, as a first step, items that can be used for the measurement of maturity are identified, followed by the assignment of items to the maturity levels.

5.5 Model population

The construction step five, *Model Population*, is of high relevance for the overall model construction. As it contains numerous tasks, an overview about the contained sub-steps and described content is given first.

i) As a quantitative bottom-up approach has been selected, in a first step the theoretical background of Test Theory and its potential approaches for the calculation of the item difficulty are explained.

ii) The basis for the calculation of the item difficulty are answered questionnaires by companies. Therefore, the topics and items contained in the questionnaire are explained as well as the characteristics of the gathered data base.

iii) Based on the gathered data, the initial maturity model, consisting of maturity levels with associated items, is calculated and described.

5.5.1 Model calculation - Theoretical foundation of the Test Theory

As a bottom-up approach has been selected, in a first step the measurements are identified and in a second step assigned to the different maturity levels. The assignment is carried out based on the associated maturity. Therefore, the goal is to carve out the maturity of each item, which is used to bring the items in an order, representing an increasing/decreasing maturity. This maturity is calculated, as it will be described in this section, based on survey results. The survey targets the existence or absence of certain processes etc. with relevance for Big Data in companies.[5]
One possibility to prioritize the items is based on the individual *item difficulty*. This value can be derived with a class of methods from the field of test theory, belonging to the social sciences. In order to foster an understanding of the applied methodology, the different approaches, belonging to the test theory will be introduced in the next section and applied on the gathered data subsequently in section 5.5.2.

In general, test theory describes the relation between the attribute that is supposed to be measured with a test and the actual test behavior [Rost, 2004]. The test theory addresses measurement problems which can be associated with test development, test-score equating, and the identification of biased test items [Mislevy, 1982].
The aspect of measurement problems is subject to the test theory approaches, analyzing in how far the questions of a questionnaire are measuring only the intended abilities. The goal is to eliminate factors, which influence the probability of a correct test answering, that are at the same time not connected with the ability that is supposed to be tested. One example is the formulation of mathematical questions in high school tests. In the event that both native and non-native speakers are taking the test, more

[5] A more detailed description of the content of the questionnaire is set aside at this point with regard to the reading fluency. The detailed description of the questionnaires role can be found in later in this section.

complex formulations and the resulting problems of understanding can influence the solving probability (although both groups taking the test have the same mathematical abilities). One indicator that is used for the analysis of the correct measurement is the item difficulty, which will be explained in the course of this section.

The test theory consists of two groups, classical test theory and probabilistic test theory, also known as item response theory (IRT).
The IRT was able to solve different measurement problems [Becker, 2004] [Rost, 1999]. Despite the predominated application of methods from the field of the classical test theory in the clinical environment, the item response theory has gained extensive attention within the scientific community [Becker, 2004], leading to initial approaches that combine classical test theory and IRT [Verstralen et al., 2001].
The classical test theory will not be further pursued in this work as:

i) The approaches from the field of test theory do not calculate the item characteristic curve. Therefore, the item difficulty values cannot be retrieved, although needed for the identification of the related maturity as described before.

ii) The statistics calculated based on the classical test theory (item/test/person statistics) are sample dependent [Embretson and Linacre, 1996]. Consequently, the calculated values cannot be transferred to other samples. This characteristic does not fit with the described design science research demand for a generalizability of the artefact, respective model [Gregor and Hevner, 2013]. Therefore, the application of the classical test theory would result in a maturity model, which is only valid for those companies that have contributed to the underlying database in terms of answered questionnaires. In contrast, the item and person parameter, calculated based on item response models are sample independent [Hambleton et al., 1991, p. 18].

iii) The measurement of models belonging to the IRT is indirect. The observable test behavior and related test score is based on a non-observable characteristic, a latent trait, which in turn influences the test behavior. This assumption goes along with the bottom-up maturity model construction, as each question regarding the existence of certain processes is supposed to represent a certain level of maturity.

Consequently, the answering behavior in the questionnaire is steered by the latent trait, the companies' capabilities, representing a certain level of maturity. In contrast, the measurements of models belonging to the Classical Test Theory are direct. The concept of the latent trait does not exist, therefore models belonging to Classical Test Theory cannot be used and models from the IRT are in the focus.

In the next step, different approaches of the IRT are discussed to derive the approach which will be further pursued in this thesis. As the used method from the field of IRT has a major influence at least on the initial model, resulting from the model population (construction step no. 5), the different potential approaches have to be discussed before, focusing on the different model-specific underlying assumptions.

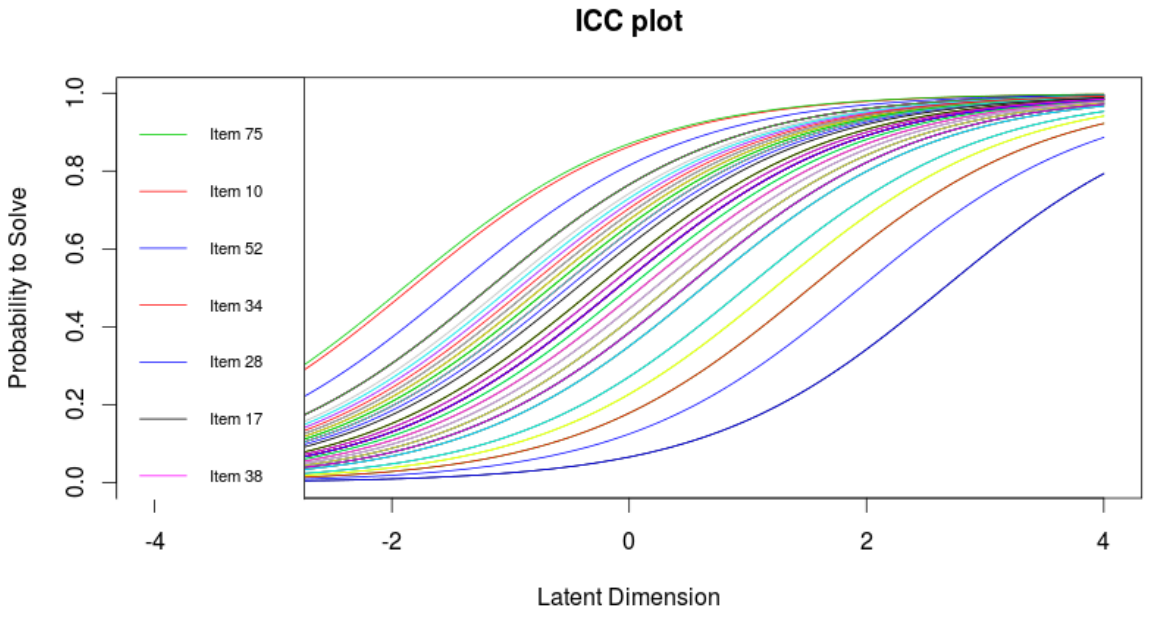


FIGURE 5.1: Example of an Item Characteristic Curve

In order to understand the different models in detail, the item characteristic curve as a basis of the IRT will be explained upfront.
The Item Characteristic Curve (ICC) describes the relation between a person's capability, in this case denoted with *"latent dimension"* and the probability to solve a question, depending on the person's capability and the item difficulty. This curve is calculated for each item individually, following the shape of a logistic function. The slope of the

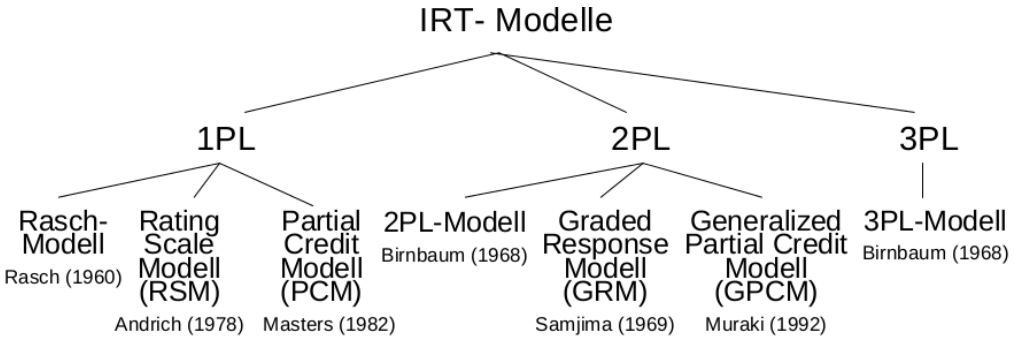


FIGURE 5.2: Exemplary Item Response Theory models clustered depending on the number of parameters taken into account [Becker, 2004, p. 52]

ICC in its middle region allows for preliminary conclusions regarding the discriminative accuracy of a question. The steeper the slope, the higher the discriminative accuracy of a question, meaning that a slight deviation of a person's ability (abscissa) results in a significant change in a person's probability to solve the question.
The numerous models belonging to the IRT can be structured based on the number of parameters, which are specified in the item response function as shown in figure 5.2.

One parametric models (1PL) use a single location parameter to describe the position of the item on the latent trait. Popular ones are the Rasch model, the Rating Scale Model and the Partial Credit Model. Within each category, the models can be differentiated with regard to the type of used scale, either dichotomous or polytomous, based on the answering possibilities.
Two parametric models (2PL), assume that the slope of the item characteristic curve is not identical for every item and therefore add a slope parameter, that allows the calculation of an individual slope for each item. Consequently, differences in the discriminatory power between the different items are measurable.
Models belonging to the *three parametric* (3PL) category contain a third parameter, incorporating the aspect of guessing. The basic assumption is that if a person has a low to zero capability respective knowledge in a specific field, he can still solve a question correctly by guessing, leading to a solving probability > 0.

The starting point for the explanation of the different IRT approaches will be the 1 PL Rasch model that has already been applied for the construction of maturity models [Marx et al., 2012].

Rasch model

The Rasch model, developed in the year 1960 by the Danish mathematician Georg Rasch, is one of the most popular IRT models besides the Birnbaum model. Is has reached prominent use, e.g. for the design of the PISA 2006 study [Prenzel et al., 2007]. The initial Rasch model has been used to analyze attitude or performance tests. Questions can be answered with 1 (agreement respective correct answer) or 0 (rejection respective incorrect answer). An exemplary result matrix for a test with four questions, taken by five people, can be found in table 5.7, with persons 1 and 2 as the best test takers.[6]

TABLE 5.7: Exemplary result matrix for a test with binary questions

	Question				
Person	1	2	3	4	Columnbind
1	1	0	1	1	3
2	1	1	0	1	3
3	0	1	0	1	2
4	1	0	1	0	2
5	0	0	1	1	2
Rowbind	3	2	3	4	

The Rasch model determines the solving probability of a question, depending on the person's capability, denoted with θ for person $_i$ and the question difficulty β for question $_j$. U_{ij} represent the unknown outcome before the person $_i$ answers question $_j$. $U_{ij} = 1$ stands for a correct answer.

The resulting Rasch model equation is

$$P(U_{ij} = 1|\theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}} \tag{5.1}$$

Based on this formula, every item characteristic curve calculated, using the Rasch model, has the same slope. Therefore, for each question, the same increase in a person's capability results in the same increase of the solving probability - each question has the same discriminative accuracy.

[6]Statements regarding the best test taker can only be made when the question can be answered only correct/incorrect and the question does not target the agreement or disagreement with a statement.

Two central assumptions of the Rasch model's test respective questionnaire are *sufficient statistics* and *local stochastic independence* [Becker, 2004, p. 45]. The demand for the existence of sufficient statistics in the Rasch model states, that for each unknown parameter (θ_i and β_j), all relevant information must be available. Relevant information in this case denominates the column sums for the number of solved questions per person i and the number of correct answers per question j (as is evident in in table 5.7).

Contrary to the demand for *sufficient statistics* as just described, the assumption of *local stochastic independence* cannot be kept in total when using the Rasch algorithm for the maturity model construction. Stochastic independence in this case means that the answering of one question does not influence the probability of the answering behavior for another question in the questionnaire.

The potential questions regarding the source and structure of data that are analyzed are taken as an example. These data characteristics may have an influence on the used application to carry out the analysis tasks. In case a company includes both structured and unstructured data from company-internal and external sources, the response behavior of different questions e.g. from the field of data quality management would be influenced - the company is likely to have a stronger focus on data quality management with regard to potential errors and noise, especially in those company-external data. By that, the response behavior is not necessarily - but can be - influenced. Therefore, the questionnaire has to be designed during the population step (construction step 5) for the data gathering in a way to prevent an answer from excluding another answer and reduce the mutual influence.

The Rasch model has been applied numerous times for the development of maturity models in the field of information systems, e.g Lahrmann et al. [2011a]; Marx et al. [2012].

One prominent publication modifies the Rasch approach by calculating in the forefront of the model population a delta, consisting of the perceived relevance of an item and the expected costs of its implementation [Marx et al., 2012]. This approach is based on the assumption that the employee answering the questionnaire is able to estimate the effort needed to implement an item, e.g. a certain process or methodology. This approach has not been selected. In contrast to the subject in focus by [Marx et al., 2012], Management Control Systems, Big Data holds partly novel aspects. Consequently, a reasonable

estimation of an implementation effect and related costs by the respondent is not possible for every contained item due to the lack of experience, such as the existence of a Big Data strategy. Therefore, the questionnaire used in this research project is focusing solely on the existence of certain processes etc. in order to reduce the bias by incorrect estimations.

Besides the Rasch model, other 1PL models are commonly used, such as the *Rating Scale Model* and the *Partial Credit Model*. In contrast to the binary Rasch Model (two answering possibilities: "solved, not solved" respective "agrees, does not agree"), those two models belonging to 1PL allow the processing of data from different scales, e.g. "agree totally" up to "total disagreement". Yet, as the focus is on a binary input (a characteristic exists: yes/no), they will not be considered further.

Birnbaum Model

An advancement of the *Rasch model* has been developed by Birnbaum [1968], relaxing one of the main characteristics of the Rasch model regarding the identical slope of each Item Characteristic curve. This approach is implemented by adding a flexibility parameter δ for question j into the already known equation. For this 2PL model, the resulting equation is

$$P(U_{ij} = 1|\theta_i, \beta_j, \delta_j) = \frac{e^{\delta_j(\theta_i - \beta_i)}}{1 + e^{\delta_i(\theta_i - \beta_i)}} \tag{5.2}$$

This relaxation of the demand for the models' identical slope, led to fast acceptance and application of the Birnbaum model, as for some tests it is more realistic to have questions with different discriminative accuracies.
Further 2PL models are the Graded Response Model and the Generalized Partial Credit Model, which are similar to the Birnbaum model, but hold different assumptions.

The 2PL model can be complemented by a third parameter, leading to 3PL models. This third parameter is used for the determination of the intercept of the item characteristic curve. Normally, the curve starts at zero, containing the possibility that a person, with zero abilities relevant for the asked question taking the test, has a probability of zero to solve a question. One might argue that a test taker always has a certain probability to solve a question as, even without any relevant knowledge, he can select an answer based

on guessing, which might lead to a correct answer. Consequently, the already known equation from Birnbaums 2PL model is complemented by the parameter γ, describing the starting point of the Item Characteristic Curve for question j.

$$P(U_{ij} = 1|\theta_i, \beta_j, \delta_j, \gamma_j) = \gamma_j + (1 - \gamma_j)\frac{e^{\delta_j(\theta_i - \beta_i)}}{1 + e^{\delta_i(\theta_i - \beta_i)}} \tag{5.3}$$

A modified 3PL Birnbaum model has been selected for the item difficulty calculation, i) incorporating the guessing parameter but ii) assuming an identical slope for each ICC.

$$P(U_{ij} = 1|\theta_i, \beta_j, \gamma_j) = \gamma_j + (1 - \gamma_j)\frac{e^{(\theta_i - \beta_i)}}{1 + e^{(\theta_i - \beta_i)}} \tag{5.4}$$

Regarding i): The guessing parameter takes into account the potential insecurity of the respondents. With regard to the partly novel aspects of Big Data, the answering behavior might be influenced apart from the actual capabilities by an optimistic attitude towards the companies' capabilities. Optimistic in this case means that even if an analytical process, application, etc. has not been implemented and solely tested, based on a proof of concept, the test taker may answer for this capability with "yes".
Furthermore, a respondent might not always be aware of details regarding specific questions and thus might answer accordingly to how he interprets the question.
Regarding ii): Incorporating different slopes would lead to differences in the discriminative power of the maturity levels. In other words, some capabilities would be harder to achieve than others, which would add another aspect for ranking besides the item difficulty. This would result in a weighting of the different maturity levels.
In contrast, the same slope for every ICCs allows the grouping of items on maturity levels and comparing different maturity levels only based on the item difficulty value.
The calculation of the item difficulty in the next section is carried out based on data from the answered questionnaires. In this case being the companies' respective employees who have completed the questionnaire. As said before, the questionnaire contains questions regarding the existence/absence of processes, objects etc. in order to measure the existence of different Big Data relevant capabilities. The calculated item difficulty is used to prioritize the items and bring them into an order that in turn, is supposed

to represent the associated maturity. The application result of the described Birnbaum model is illustrated in the following section 5.5.2.

5.5.2 Development of the questionnaire

The development of the questionnaire is the starting point for the model population, as a bottom-up approach has been selected. The questionnaire is used to gather the data needed for the following model population. In order to develop the questionnaire, the topics and related measurements have to be defined. Therefore, in this step, the later maturity model content is defined.
The questionnaire has to comply with the following two demands:

i) The identified topics with relevance for Big Data are supposed to contain aspects that are not yet covered by existing maturity models from the field of BI, but at the same time, do not focus on specific application scenarios as described in section 5.4.

ii) The second demand results from the application of the test theory, holding the need for a questionnaire whose questions are independent from each other. The probability of one response must not be dependent on the answering of another question.

The topics and items are carved out based on the identified maturity models from the field of BI, relevant literature, and enhanced by the input from the focus group.
The questionnaire targets different *topics*, referring to the *dimensions* Data and Organization, selected in construction step four. A topic is targeting a specific field, e.g. the data quality management or success control of data analytics. For each topic, several *measurements* have been defined. These measurements are targeting different occurrences of one topic, e.g. a manual data quality management, or an automated data quality management. As these measurements are transferred into the questionnaire, they are called *items* likewise. Following the selected methodology as described before, each item in the questionnaire is targeting the existence or absence of a process, object etc.

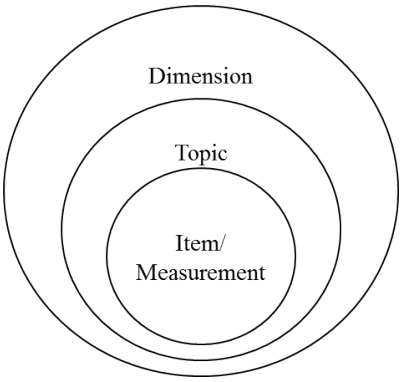


FIGURE 5.3: Hierarchy of Dimensions, Topics, and Measurements

To design the questionnaire, some pre-work has been carried out before. In a first step, topics and related measurements have been identified and transferred into a questionnaire structure. The identification has been carried out based on existing literature belonging to the field of Big Data and maturity models from nearby fields as described in Chapter 3.

This draft has been discussed in a next step with the focus group regarding comprehensibility and understandability. The input of the focus group at this stage is needed as several aspects of Big Data are not yet covered by recent literature due to the novelty of the topic. This accounts as well for the identification of measurements.

This conducted discussion resulted in a questionnaire that has been pretested with the focus group members for clarity and understandability, consisting of 17 topics with three to five items, each assumed to represent a different difficulty (table 5.8). The resulting topics and items as the basis for the questionnaire design will be presented in the next section.

The final questionnaire consists of two parts:

i) Organization dimension related questions

ii) Data dimension related questions

Part 1: Organization dimension

With regard to the overlap with the field of Business Intelligence, the questions of the organization dimension are partly related to existing maturity models for BI applications

TABLE 5.8: The final questionnaire, developed based on existing research-based maturity models from nearby fields and the input of the focus group consists of 17 different topics. For each topic, three to six characteristics have been identified, each associated with a different degree of difficulty. O for Organization and D for Data in brackets describe the belonging of the topic to the respective dimension

Topic	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Existence and Reach of an Analysis Strategy (O)	No Strategy exists	Department-wide	Cross-department	Division-wide	Company-wide	
Existence and Status of a Data Analysis Project (O)	No project exist, no project planned	No Project exists but planned within the next 12 month	Project is planned	Analysis phase of the analysis relevant project is completed	Project is currently implemented	Project is completed
Sponsor of Data Analysis Project (O)	Decentralized IT department sponsor	Central IT department sponsor	Central department management sponsor	BI Department sponsor	Top Management sponsor	
Cost Control of Data Analysis (O)	No cost-benefit calculation	Project-based cost-benefit calculation	Use-oriented cost-benefit calculation	Success-oriented cost-benefit calculation		
Implementation of Data Analysis (O)	Ad-hoc analysis (e.g. spreadsheet based)	Static reports	Integration of analytical front-ends on reports	Use of analysis software	Flexible, pro-active analysis solution	
Frequency of Data Analysis (O)	Less than weekly	Weekly	Daily	Continuously (real-time)		
Type of Data Analysis (O)	Monitoring	Classification	Exploration	Prediction		
Result provisioning (O)	Printed/digital (.pdf, .xls)	Department-wide online portal	Company-wide online portal	On mobile devices		
Result processing (O)	Manual processing without a standardized process	Manual processing based on a standardized process	Automated processing			
Usergroup I&II (O)	Single Data Analyst	Key User	Individual Analytics Department	Middle Management	Company-wide	
Success Control (O)	No success control exists	Sporadic user meetings for success control	Regular user meetings for success control	Standardized, irregular success control	Standardized, regular success control	
Process model of Data Analysis (O)	Informal data analysis processes / no standardization	Established data analysis processes based on routines	Standardized, documented data analysis processes on department level	Standardized, documented, and controlled data analysis processes on department level	Mandatory, company wide data analysis processes and controls	
Identification of new Data Sources (D)	Focus on already processed data	Irregular screening for further, company-internally available data	Irregular screening for further, company-internally and externally available data	Regular screening for further, company-internally and externally available data	Development of a data landscape for company-internal and external data	Development of a data landscape for company-internally and externally available data and evaluation of potential data sources
Source and Structure of processed Data (D)	Analysis of internal, structured data	Analysis of internal structured and unstructured data	Analysis of internal structured/unstructured + external structured data	Analysis of internal/external structured/unstructured data		
Combination of Data Sources (D)	No combination	Manual, sporadic	Manual, regular	Automated, standardized		
Data Quality Management (D)	Manual ad-hoc	Defined DQM roles	Defined DQM processes	Automated DQM	DQM Team	

[Lahrmann et al., 2011b; Dinter, 2012]. The novelty of the research at hand however is ensured through two aspects:

- The fitting and updating of the answer options of related topics from nearby maturity models to the Big Data context with related measurements (i.e. items targeting further capabilities of an already covered topic). In doing so, potentially relevant topics from existing models are further incorporated, but fitted to the further development of the Big Data environment. This fitting procedure is aimed to incorporate the matter of partly existing overlaps of BI and Big Data.
- The identification and integration of new topics and related items in the questionnaire, that are clearly relevant for and associated with Big Data and have not been questioned so far in the BI context.

Therefore, in the next section, both the identified topic as well as the referring measurements are described. Additionally, it is carved out which topics and measurements go beyond aspects that have been already covered by existing maturity models.

Data analysis strategy

A strategy in the Big Data context is understood a description for the goals and intention pursued with data analysis for the companies' processes and success. It is an indicator for the perceived relevance of Big Data, as a strategy formulation needs a debate on the future direction of the topic in focus and additionally puts more emphasize on it [McAfee and Brynjolfsson, 2012].

The existence of a strategy supports informed investments decision and supports the avoidance of over-exceeding budget which is of special interest with regard to the needed investments to develop relevant Big Data capabilities [Luftman et al., 2015]. The topic of strategy can be found in the field of BI as well [Lahrmann et al., 2011b], but has increased in relevance with regard to the role and perceived relevance of Big Data for the overall companies' success. Big Data has the potential to open up new revenue streams by enhancing the existing business model or generating new ones. The decision in how far the utilization of Big Data within the company is pursued, is reflected in such a strategy. Therefore, a Big Data strategy is supposed to be more closely related with the overall business, determining the role of the utilization of data in the company.

With regard to the novelty of the topic Big Data, the focus is on the scope of the strategy

instead of its update cycle, which is in the focus for more established topics [Lahrmann et al., 2011a].

The response options cover the spectrum from the lack of a Big Data strategy to a company-wide strategy, representing different scopes of the strategy. The company-wide strategy is associated with a far-reaching penetration of analytics throughout the company. Response options, targeting a potential link between the overall IT strategy and the Big Data strategy as it can be found in Dinter [2012] have not been included, as Big Data is understood in this research as an overall paradigm, which has overlaps with corporate IT but is not solely part of it.

Data analysis project

The existence of an analytic-relevant project can be used as a measure for maturity, as the implementation, improvement etc. of analysis relevant processes points towards a recognized relevance of the Big Data topic for the company. This topic carries some novel aspects. The aspect of a Big Data project has been raised by a member of the focus group during the discussion of the initial draft of the questionnaire and represents a new topic, as it has not been subject to recent maturity models, presented in section 5.3.

It is assumed that the implementation of processes, tools etc. in the Big Data context are carried out normally in the course of a project, as these enhancements are not part of the regular business. Additionally, each project carries the need for budget and the investment of workforce, whose existence speaks for at least a minimum believe into the benefits of analytics for the business.

The response options cover, besides the absence of any projects, every stage of a project from the planning stages to the running project that is in progress up to the already completed project.

Data analysis project sponsor

The question regarding the project sponsor is related to the question for a Big Data strategy and follows the same logic. The more significant the influence of a project on the company is, the higher the project will be located in the company structure. This aspect can be found in recent BI maturity models [Lahrmann et al., 2011a]. A sponsor located in the upper parts of the organisation stands for the management commitment. The more comprehensive the project becomes, the more departments are

involved. Therefore, again the response options range from a de-central sponsor up to a sponsor on management level.

Cost Control

During the past years, the investments into information systems, especially those with an analytical focus, have risen and are expected to rise further as a result of i) the further spread of these systems throughout the company, independent of the industry and ii) the need of more sophisticated systems, tools etc. for the roll-out of analysis tasks into the productive environment [Capgemini, 2015]. Potential investments areas are amongst others purchasing and running of hard- and software, cloud computing environments as well as the generation and improvement of the workforce capabilities, such as software training.

With regard to the increasing spread and heterogeneity of analysis applications, the cost control complexity increases. This connection can be found in a recent BI maturity model by Raber et al. [2013a]. Consequently, the analysis of costs results in a need for the examination of data analysis on detailed level, in order to identify the individual cost driver and allocate the costs correctly. This aspect is more in focus for already productively running IT systems. The answering options hence cover the range from the absence of cost control to a comprehensive cost control, incorporating the success that was created due to the made investments. Specific investment areas have not been carved out.

Implementation of data analysis

Another aspect which has been stressed during the discussion with the focus group is the aspect of the implementation of the data analysis. Implementation in the context of data analysis is targeting the technical system in terms of software application/tool that is used to carry out data analysis related tasks.

This topic of implementation has been integrated despite its close link to the infrastructural respective application dimension, as the items are not exclusively applicable for one specific tool. Instead, they represent a general degree of professionalism.

This topic targets those capabilities, that are associated with a different type of software or tool, that is used in a company. More specific analysis applications speak for a higher perceived relevance of the analysis topic and are related with higher investments into technology. Furthermore, the analysis of unstructured data (texts, pictures etc.) poses

high demands for capabilities to the tools users. Consequently, the tool in usage allows conclusions regarding the companies' capabilities and activities in the field of data analysis. More sophisticated solutions have advantages regarding the volume and the structure of the processed data as well as the contained statistical methods.

The response options range from the usage of *spreadsheets* as analysis tools[7] up to *flexible, pro-active analysis solutions*, which can be used for explorative approaches.[8] The response options *Use of Analysis Software* and *Flexible, pro-active analysis solution* go beyond the aspects covered by existing maturity models in the field of Business Intelligence and therefore represent an enhancement of current research.

Frequency of data analysis

This aspect targets the frequency, in which analysis are carried out on (parts of) the data pool. A higher frequency is associated with a deeper integration of data analysis in the business processes, focusing on a productive use of analytics as a core business function and allowing a faster reaction of the company. Potential applications for these can be the fitting of demand forecasting, management of external risks (e.g. environmental risks), reputation management, or the analysis of manufacturing streaming data for quality management purposes.

The incorporation of the *frequency* topic represents an enhancement of current BI maturity models as BI oriented reportings are generated with a lower frequency compared for example with tracking applications in the field of supply chain monitoring or customer tracking in online marketing as recent Big Data applications. For reporting purposes, e.g. sales reporting in wholesale, a daily update and analysis frequency is in most cases sufficient, whereas for monitoring purposes in production environments or the tracking of sentiments towards brands or companies, based on media with a high update rate such as Twitter, the continuous analysis increases the explanatory power.

The response options range from "less frequent than weekly" up to "continuously".

Type of data analysis

As described in Chapter 2, the diversity of data analysis applications has increased in the past years. The focus of Business Intelligence applications is on reporting, the calculation of key figures is carried out using primarily on transactional data, generated

[7] At this point, the focus is on the use of the spreadsheet as the analytical application, not as a front-end for underlying analytical applications.

[8] One example for those functionalities are analytic workbenches like SAS Visual Analytics [SAS, 2015b].

based on ERP systems as a consolidation with further unstructured, potentially external data sources does not go along with a classical data warehouse structure [Trujillo and Maté, 2012].

In the course of time, the increasing data depth (number of attributes per entry, e.g. on customer level) and data breadth (time period since the attributes are stored) allow further types of data analysis. Those types differ - amongst others - in the applied methods (e.g. supervised/unsupervised learning approaches) or the focus (e.g. analysis of real-time data/analysis of data for prediction purposes). As these different types of analysis demand for more sophisticated methodologies, infrastructure and integration into the business and decision making process, they can be taken as a measurement for maturity.

The response options range from reporting (aggregation of sales data etc.), exploratory analysis (focusing on pattern recognition, e.g. in customer data, without focusing on specific attributes, e.g. sales or potential), to predictive analysis (forecasting the development of sales, trends etc.).

Another aspect that has been discussed during the questionnaire development are prescriptive analytics. This aspect however has not been further considered. Based on the discussion with the focus group and following the argumentation that methods and tools - associated with prescription - have not yet arrived in practice. The term lacks, in contrast to the ones mentioned before (exploration etc.) a clear understanding.

Result provisioning

Result provisioning focusses on the possibility how analysis results are made accessible for end users from the business side.[9] [10] This topic, already existing in the context of Business Intelligence, gains in relevance as a sufficient information exchange between the person generating the analysis results (Data Scientist) and the one using the insights for decision making (business person) tends to be less often existing in contrast to Business Intelligence [Rajaram, 2013]. At the same time, the availability of analysis results has an influence on the integration in employees' decision-making, e.g. forecasted market developments in the product development process [Marx et al., 2012].

[9] Analysis results in this context are understood as the information, resulting from the application of a mathematical operation (as described in the topic *Type of Data Analysis*) on data.

[10] In this context, the business side/end user targets the employees which is the decision maker, e.g. a marketing manager, deciding about the allocation of the marketing budget on different channels.

With regard to the needed statistical background for the analysis task, current applications in the field of Big Data rather focus on a Data Scientist as an user instead of a person from the business. Additionally, the provisioning of analysis becomes more relevant with regard to the increasing number of employees, working with analysis applications combined with an increased employees' mobility.

The options cover the range from rather inflexible solutions, such as *printing the results* or storing them on a department wide platform, which leads in turn to silo structure up to the highest mobility, offered through the *provisioning of the results on mobile devices* such as tablets or smartphones [Trujillo and Maté, 2012].

Result processing

Result processing is focusing on the further use of analysis results for decision making. Examples for analysis results in this case are customer values, predicted demand quantities, or times of expected machine failures. Following the idea of a data driven company, an automated processing and integration of analysis results is supposed to support decision-making on a wide basis [Provost and Fawcett, 2013]. An early example for analysis results integration is the use of counter and marketing data for the disposition planning in production environments, proving its positive influence on the performance [Mukhopadhyay et al., 1995]. In this case, the analysis is limited to the calculation of remaining stock. Nonetheless, the following business process - the order - is triggered automatically.

Numerous currently dominating applications in the field of Big Data, e.g. sales forecasting, customer segmentation, and marketing mix analysis hold the potential for an automated processing of the analysis results in order to facilitate the subsequent decision making, e.g. the targeted display of advertisement or profiled product suggestions.

The degree of standardization and integration of analysis results in business processes and decision-making allows conclusions concerning in how far data analysis is anticipated as a part of the daily decision making. Therefore, this question can be used as a measurement for maturity. This topic has not been explored in current maturity models, but was brought up by the members of the focus group, advocating an understanding of data analysis as an essential part of daily business processes.

The manual, un-standardized processing of analysis results is expected to be the lowest level of maturity and points towards a gap between analysis and business processes,

e.g. the utilization of reported marketing effects from past sales as the basis for decision making in order to optimize future pricing. A manual standardized processing, e.g. a customer segmentation could be the transfer of analysis results in pre-defined spreadsheets or Customer Relationship Management systems for pricing decisions. The automated processing of results describes the automated derivation of action, triggering processes that depend on the analysis results.[11]

Usergroup I & II

The next two questions have been grouped as they contain the same answering options. Both questions target the persons involved in the analysis process.

The questions regarding Usergroup I asks for the persons, who are responsible for the development of analysis, defining the reports etc.

Usergroup II describes the persons working with the analysis results. Therefore, usergroup II is linked to the question regarding the *result provisioning*, explaining how the results of the analysis are presented to the persons, belonging to usergroup II. A close tie between the employees using the analysis results and those defining the needed reports and analysis can improve the business understanding on the modellers side, resulting in an improved result quality from the analysis users perspective. This striving for closeness can lead up to the point, in which a business user is able to define its own reportings and carry out analysis tasks on its own. That aspect is known in the Business Intelligence context as self-service Business Intelligence [Abelló et al., 2013]. The respective term self-service analytics has already been introduced as well [Kridel and Dolk, 2013].

The extend in how far the business functions can work with data analysis on their own depends amongst other on the used software tool. Approaches as IBM's Watson technologies, being able to process natural language and transform spoken questions into analysis queries, foster the use by the business [IBM Systems and Technology, 2011]. It is assumed that an increased use of such tools by the business goes along with investments into training etc. in order to enable the business to benefit from the applications. This in turn allows to draw conclusions regarding the perceived relevance of analytics in the company.

The response options are the same for both questions and cover the range from one single person in a company responsible for developing/using analytics, over a key user

[11] One exception is the strategic decision-making in which automated analysis result implementation is not possible, resulting from the informal character of the strategizing process [Whittington, 2003].

per department, up to a analytics development/usage throughout the company [Raber et al., 2013a].

Success control

Success control has already been touched upon with the topic of *cost control*. Comparable to the result processing, for analysis tasks whose results are integrated into strategic decision making with longer time horizons, a success measurement is a challenging task. Success measurement in this work is understood as the measurement of the impact resulting from the use of analytics application, e.g. cost reduction in the production environment due to improved quality management, or increasing sales due to improved marketing actions.

Success control and cost control gain in relevance in the Big Data context as data are increasingly perceived as a factor of production, whose use should be managed and optimized [Capgemini, 2012]. However, success control in the context of Big Data raises one problem: On one hand, the integration of predicted future trends into decision-making and, based on that, the derivation of concrete actions, such as sales forecasting and needed stocks, become increasingly important. On the other hand, with increasing time horizons and granularity of actions, which are supposed to be predicted (e.g. market development scenarios for the strategy development), the precision of the prediction becomes partly less measurable due to its qualitative character and complexity. Nonetheless, the existence of defined processes and time frames for success control can be taken as an indicator for maturity as it measures in how far the company is concerned with this aspect. It has not been taken into account in current maturity models.

The measurement of success in this context targets the existence of measurement related processes instead of specific key performance indicators. Consequently, the response options cover different routines that can be applied to the success control, ranging from the lack of a success control up to a regular, standardized success control. The formulation of items based on processes in the context of success control allows a data gathering that is independent of the company's individual utilization.

Standardization of analysis process

Standardization as shown in the topic *Success Control* is equally relevant for the analysis process itself. The target of this topic is to identify how the execution of data analysis tasks is based on (standardized) processes or routines and in the case that

standardization exists, the extent and level of definition. Although data analysis is supposed to be a flexible task, standardized process models for each field of application could foster the increasing use and integration of analysis into day-to-day business, as it helps to understand the processing of the analysis tasks, which is relevant as well for its further improvement. Potential process models for the analysis have been discussed in terms of the CRISP-DM [Shearer et al., 2000] and the model by Fayyad et al. [1996]. Additionally, the structuring of the data analysis on a process level shows a potential increased maturity, as data analysis is supposed to be transferred in an productive state as part of existing business processes. Process relevant items can be found in the model by Lahrmann et al. [2011a] in terms of defined governance processes and also in recent literature on Big Data governance [Tallon, 2013].
Again, the response options range from the absence of any process definitions up to the company-wide standardized and controlled processes.

Part 2: Data dimension

The second dimension selected for the maturity model is the data dimension. The contained topics are focusing on three main drivers of Big Data in a data context as described in Chapter 2, namely

- Data volume,
- Data structure, and
- Data source.

Additionally, the aspect of Data Quality Management has been identified as one challenging aspect in current Big Data applications [Accenture and GE, 2015]. This data dimension is a distinguishing factor to similar research - *Data* as an individual dimension is not subject to current BI maturity models.

Identification of new data sources

The first question of the data dimension is targeting those processes that are related to the identification of new data sources for analysis purposes. The identification of data sources plays an increasingly important role within the overall analysis context [Hansmann and Nottorf, 2015]. Its increased relevance is fostered by the increasing number and

heterogeneity of available data sources, especially of company-external sources. With regard to the static character of BI, the search for and evaluation of new data sources has played a subordinate role so far [Hansmann and Nottorf, 2015]. It is assumed that the perceived relevance of data analysis for a company has an influence on the structure and comprehensiveness of the processes for the identification of data.

With regard to the novel character of Big Data and the expected heterogeneity of companies' capabilities in this field, the response options are selected to mirror the differences between a structured process model for the identification of data in an analysis context and the information requirement analysis, originating from the field of data warehousing and BI, focusing primarily on company-internal data [Hansmann and Nottorf, 2015].

The options range from a focus on already-existing data in the data warehouse up to the regular generation of a data landscape, including a pre-evaluation of data. The pre-evaluation has gained in relevance within the Big Data discussion. With the increasing number of data sources available, the selection of the most informative data sources out of the overall identified data sources becomes increasingly relevant, especially with regard to the needed pre-processing effort for unstructured data [Assunção et al., 2013]. Pre-processing in this case is not limited to Data Quality Management (DQM). It contains as well the aspect of data matching, bringing person-centric data from different sources together. This aspect will be targeted in the topic *Combination of data source*.

Source and structure of processed data

Comparable to the questions of the *type of analysis* and its *frequency*, the selected data itself are equally influenced by the goal of analysis as well. Nonetheless, within most of the companies, irrespective of their industry and size, use cases for the analysis of data, both from internal and external sources as well as structured and unstructured can be identified.

The sources and structure of the processed data allow conclusions regarding existing capabilities both in the field of data pre-processing and analysis. Unstructured data, especially human generated ones (e.g. product reviews) tend to hold noise as it will be explained in detail in the topic *Data Quality Management*.[12] Therefore, the pre-processing is more demanding.

Furthermore, unstructured data require a different set of mathematical methods for the analysis, which are not covered necessarily by existing data analysis applications due to

[12]Further elaborations on this topic can be found in section 2.1

focus on numerical data in the past.

The response options cover again a range from company-internal structured data to both structured and unstructured data from in- and outside the company.[13] Yet, the answering option *poly-structured data*, has not been included as it does not hold sufficient discriminatory power with regard to its heterogeneous understanding in practice. A dataset, containing a tweet, a user ID, and a time stamp for example, could be seen as poly-structured, although the understanding of analyzing Twitter data is most commonly understood as the analysis of unstructured data, which is the view that has been followed in this work.

Combination of data sources

After the processed data sources have been identified, the next question targets the potential merge of different data sources. The enhancement of existing data sources accounts especially for the marketing context with regard to the use of company external data, e.g. the enrichment of the customers' data through their activity in social networks in the setting of customer profiling e.g. for product recommendation purposes [Han et al., 2012; Amatriain, 2013]. This can contain as well the purchasing and processing of specific data sets from service providers, e.g. socio-economic data from external market research institutes.

This aspect has already been triggered by the topic before, the *Source and Structure of processed data*. The extend of automation of the data sources combination allows conclusions regarding the integration of analysis tasks in the existing business processes. An automated combination of different data sources is assumed to allow a more frequent update of the consolidated data set compared with manual approaches. A frequent update is seen as crucial for different analysis tasks, e.g. customer targeting and product recommendations.

The answering options range from the lack of a combination of sources to an automated combination.

Data Quality Management

Data quality is perceived by companies as one major challenge for the successful application of BI [Lahrmann et al., 2011a,b]. This accounts as well for Big Data [Kwon et al., 2014; Zhang, 2013]. DQM as a core topic of Big Data becomes increasingly important

[13] For each response option, an example data source/type is given in the questionnaire.

due to the developments named before (variety in data volume, structure, and sources). Each data source has to be evaluated in the forefront of the analysis in order to gain insights about possible quality issues, e.g. missing values or inconsistency. Therefore, with an increasing number of sources, especially those from outside the company, where the data generation and thus the data quality cannot be influenced, the needed effort and importance of data quality management increases. This accounts especially for human-generated data that are usually generated based on a certain intention. They often contain, in contrast to machine generated data (e.g. transactional data from ERP systems) opinions and sentiments, which need a check for plausibility.
In order to remain application-independent, the focus of the response options is on the underlying process of the data quality management instead of its actual execution, which would rather target specific data curation approaches. It ranges from a manual DQM to a completely automated DQM approach.

Up to this point, the topics and related items for the two dimensions, *Organization* and *Data* have been described. Besides those aspects, complementing information regarding the company and respondent are additionally aimed to be gathered: i) the belonging industry, ii) number of employees and iii) the organizational level of the person answering the questionnaire. These information are supposed to support the subsequent review, in order to analyze in how far the results of the model population are based on an equally distributed database regarding the company characteristics, to foster the generalizability of the results.
For aspect i), a significant abundance of one industry within the participating companies may influence the item distribution of the initial model. The company's size - aspect ii - allows amongst others to draw conclusions regarding the number of departments and potential fields of data analysis application. Larger corporations may have more financial funds - in contrast to smaller companies - to invest into the development of analytical companies.
The information about the organizational level of the person, aspect iii), holds its explanatory power in combination with the information about the number of employees. Employees from the management level, especially of larger companies might lack the detailed information regarding specific process-oriented questions. Furthermore, the perception of items targeting the big picture, such as the existence and spread of a Big

Data strategy might differ between the lower and top management level in larger corporations. Therefore, this information can be used later on, in order to evaluate the explanatory power of the initial model. Ideally, the companies are distributed equally among all of the three different characteristics.

Preliminaries with relevance for the questionnaire

As described before, one demand with influence for the questionnaire to be developed is *local stochastic independence*. This first demand is fulfilled as none of the provided answering options for the different topics are mutually exclusive. Each combination of responses is logically possible.

The second demand concerns the *homogeneity* of items, determining that all items are measuring the same capability, in this case maturity.

Related to this demand is the uni-dimensionality, the third demand, requesting that the answering is only influenced by one single latent trait, the Big Data capabilities. Those three demands are fulfilled by developing a questionnaire, which questions are formulated clearly, are furthermore clearly delineated, and that ensures that each question is directly related to the company's maturity with regard to data analysis.[14]

Up to this point in the construction step five, the topics and related measurements for each topic have been identified based on relevant maturity models and literature and on discussions with the focus group for validation subsequently. For each topic, the different measurements have been listed and the respondent was asked to tick in the provided questionnaire, which process etc. are implemented in his or her company.

As explained, the questionnaire has been pre-tested by the members of the focus group regarding the ease of use. After changes targeting formulations have been made, the final questionnaire has been used in a next step to gather the data, which are needed for the further maturity model development.

5.5.3 Data gathering

Based on the research goal, to develop a cross-industrial, cross-application model irrespective of a company's size, every company with the potential to utilize data analysis could participate in the survey, independent of its size or industry. The questionnaire

[14] Consequently, questions regarding the company's characteristics, e.g. number of employees or industry, are not taken into account for the model calculation.

TABLE 5.9: Characteristics of the respondents

Industry	Number	%
Manufacturing sector	16	23
Information and communication	15	21
Retail	6	8
Services	12	17
Banking and Insurance	10	14
Energy	2	3
Healthcare	10	14
Total	71	100

has been published in both a .pdf-format as well as in an online format, based on the open-source software LimeSurvey.[15] This software enables in the course of the data gathering an analysis of the number of viewed but not completed questionnaires. This information can be used to calculate the return rate. Furthermore, the software can identify at which question the answering of the questionnaire has been stopped.

The questionnaire has been distributed via four groups in the business network Xing (www.xing.de) by posting the link to the survey in the groups and explaining the purpose of the survey.[16]

The groups have been identified by using the search query for *"Business Intelligence"* and *"Big Data"* on xing.com. Groups belonging to Business Intelligence have been integrated, as a analysis of the discussed topics based on the used headlines has revealed that Big Data relevant topics (e.g. increasing data volume, analysis of unstructured data etc.) are discussed as well next to Business Intelligence topics in these groups, indicating the existing overlaps between the topic of Big Data and Business Intelligence.[17]

[15] https://www.limesurvey.org/en/

[16] https://www.xing.com/communities/groups/business-intelligence-47c2-1068645;
https://www.xing.com/communities/groups/business-intelligence-und-big-data-nervensystem-fuer-wirtschaft-und-gesellschaft-47c2-1069429;
https://www.xing.com/communities/groups/big-data-creating-value-from-data-47c2-1068031;
https://www.xing.com/communities/groups/big-data-47c2-1063962

[17] The selection of the groups can be seen as an limitation as only those persons become members of interest groups, who are already interested in those topics, which can have an influence on the characteristics of the surveys' participants. Nonetheless, as it will be shown in table 5.9, the gathered data set is not dominated by a certain type of company.

The questionnaire has been distributed in the period from March to June in 2014. In each group, the link has been reposted four weeks after the first posting in order to increase the number of participants. Besides the questionnaire link, a brief description of the surveys' purpose and contact persons have been given. 157 people clicked on the survey link, 65 people completed the questionnaire.[18] Additionally, 6 persons preferred to complete the paper-based questionnaire, leading to a total of 71 completed questionnaires. These have been used as a data basis for the further model construction.

A summary of the characteristics of participating companies can be found in table 5.9. Except for the industries energy and retail, the respondents are distributed relatively equal along the industries.

As described before, a potential limitation results from the already existing interest of the survey participants into this topic. This aspect can be found as well in similar research, e.g. the model by Lahrmann et al. [2011a], who collected their data during congresses on Business Intelligence. The participants of such events, being used to gather an adequate sample, can be seen as having an already existing interest, comparable to the members of the Big Data fora on Xing. However, by distributing the survey link on multiple fora, containing both Business Intelligence as well as Big Data topics, the gathered data base is even more multifaceted and as random as possible in this research context. Furthermore, the sample size of 71 companies exceeds already accepted publications [Lahrmann et al., 2011a].

However, generally occurring biases, resulting from the potential characteristics of the respondents that influence his answering behaviour, irrespective of the context, are still given.[19]

5.5.4 Model calculation - application of the Birnbaum model - description of the initial model

After the data have been collected, the subsequent calculation of the item difficulty in order to identify a maturity of each item has been conducted. Basis for the calculation of the item difficulty are the coded survey results. All topics and belonging items have been transferred into an excel file, one column for each item. The rows represent

[18] Data regarding the number of people who read the article at Xing are not available.

[19] Nonetheless, a potential lack of randomness is not a difficulty, as the IRT is sample-independent (Section 5.5.1).

	Organization Dimension	Data Dimension
6	▪ Use-oriented cost-benefit calculation ▪ Automated result processing ▪ Usergroup II – Middle management ▪ Standardized, irregular success control ▪ Department-wide strategy ▪ Usergroup I – Individual Analytics Department	▪ Development of a data landscape and evaluation of potential data sources ▪ Automated, standardized combination of data sources ▪ Irregular, manual combination of data sources
5	▪ BI department sponsor ▪ Standardized, documented, and controlled data analysis processes on department level ▪ Usergroup I – Company-wide ▪ Analysis are carried out continuously ▪ Regular user meetings for success control ▪ Result provisioning on mobile devices ▪ Integration of different analysis front-ends on reports	▪ Defined DQM roles ▪ Automated DQM ▪ DQM team ▪ Defined DQM processes
4	▪ Top Management sponsor ▪ Project-based cost-benefit calculation ▪ Mandatory, company-wide analysis processes and controls ▪ Analysis phase of the analysis relevant project is completed ▪ Project is currently implemented ▪ Usergroup II – Company-wide ▪ Manual result processing based on a standardized process ▪ Company-wide analysis strategy ▪ Result processing on a company-wide online platform ▪ Success-oriented cost-benefit calculation ▪ Usergroup I – Key User ▪ Analysis for classification ▪ Analysis for exploration ▪ Standardized, documented data analysis processes on department-level ▪ Standardized, regular success control ▪ Decentralized IT department sponsor ▪ Flexible, pro-active analysis solution ▪ Analysis for prediction ▪ Cross-department strategy ▪ Central IT department sponsor ▪ Individual analysis software ▪ Usergroup II – Individual Analytics department ▪ Analysis are carried out daily	▪ Analysis of internal structured/unstructured + external structured data ▪ Manual, regular combination of data sources ▪ Regular screening for further, company-internally and externally available data ▪ Irregular screening for further, company-internally available data ▪ Analysis of internal/external structured/unstructured data ▪ Analysis of company internal structured and unstructured data ▪ Development of a data landscape for company-internal and external data
3		▪ Manual ad-hoc DQM
2	▪ Sporadic user meetings for success control ▪ Analysis for monitoring ▪ Established data analysis processes based on routines ▪ Analysis are carried out less than weekly ▪ Usergroup II - Single data analyst ▪ No strategy exists ▪ Ad-hoc analysis (e.g. spreadsheet based) ▪ Central department management sponsor ▪ Manual result processing without a standardized process ▪ Static reports ▪ Printed/digital (.pdf, .xls) result provisioning ▪ Analysis are carried out weekly ▪ Usergroup I - Single data analyst	▪ Analysis of internal, structured data ▪ Focus on already processed data
1	▪ No cost-benefit calculation ▪ Informal data analysis processes / no standardization ▪ No success control exists ▪ No Project exists but planned within the next 12 month ▪ Project is planned ▪ No project exists, no project planned ▪ Project is completed ▪ Division-wide strategy ▪ Usergroup II – Key User ▪ Result provisioning on department-wide online platform ▪ Usergroup I – Middle Management	▪ No combination of data sources

FIGURE 5.4: Initial model resulting from the clustering of the items based on the calculated individual item difficulty

the companies, which have completed the questionnaire. For each item that has been ticked, a *"1"* has been marked down, for non-marked a *"0"*.[20] The resulting matrix has been used in a first step to calculate the item difficulties for each item, using the fitted Birnbaum model algorithm (Section 5.5.1). The calculation was carried out using the ltm package for the statistic software R [Rizopoulos, 2006].

In a second step, the items have been clustered based on the item difficulty value, using a ward clustering with the R build-in package *stats*. Each cluster represents a maturity level, consisting of different items with a similar item difficulty. The number of cluster equals the number of the maturity level. In this research - in contrast to the majority of the publications - the items have been clustered on six instead of five levels. This higher number has been chosen in order to represent the broad set of measurements, covering both very immature aspects up to capabilities, expected to be associated with a very high maturity. This manual determination of the number of maturity levels can be found similar in Raber et al. [2013a].[21]

The results of this clustering, the *initial model*, can be found in figure 5.4. The unequal distribution of the number of items per maturity level is a result of the applied method. Comparable to the models by Marx et al. [2012]; Lahrmann et al. [2011a]; Raber et al. [2012], the model tend to have an emphasize around the middle maturity levels, as it will be explained in detail later in this section.

Furthermore, not for every topic, the same number of measurements could have been defined. The number ranges between three to six. Therefore, not necessarily one item of each topic can be found per maturity level.[22]

The initial model was tested regarding consistency on an intra- and inter-maturity level basis in a next step. The test on an intra-maturity level is intended to analyze if two or more items are assigned to one maturity level, which contradict each other, e.g. *no data analysis strategy exists* and *a companywide analysis strategy exists*.

The analysis on an inter-maturity basis is supposed to identify if two or more items of

[20] Those items, which have been not ticked at least one are removed from the list, as they do not contribute to the calculation of the difficulty. For the model at hand, this has been the item "Irregular screening for further, company-internally and externally available data".

[21] In contrast to the model by Marx et al. [2012], no dimension specific models have been calculated. This means that the item difficulties have been calculated based on the overall responses, no dimension-individual data sets, that would contain only the responses to topics belonging to one of the two dimensions have been created. The low number of items for the *data dimension* does not allow for results with a sufficient explanatory power for this dimension. Therefore, the items per maturity level have been assigned manually to the belonging dimension.

[22] The effect of this on the model evaluation (construction step 6) will be explained later in this section.

the same topic are obviously distributed in an improper form on the different levels. Using the example of the data analysis strategy again for the inter-maturity level analysis, the distribution of the item *no data analysis strategy exists* on level six as the highest maturity level and *a companywide analysis strategy exists* on level one as the lowest level should be identified as a potential error. The goal of this intra- and inter-maturity level analysis is to test, if the quantitative approach leads to suitable initial results without a need for a further model fitting.

This analysis does not represent an evaluation. This approach has an explanatory, descriptive character and helps to interpret the initial model. The goal is not to carry out an detailed discussion of each item but to provide a first overview of the initial, yet to be fitted results for each level. The actual evaluation begins with the discussion of the initial model with the focus group, described in section 5.6.1 (construction step 6.1).

Level 1

The co-existence of the items *absence of analysis relevant projects* and *already finished projects* represent very different states with regard to the role of data analysis in a company and do not fit into one maturity level (intra-maturity level). Furthermore, the *existence of division-wide analysis strategy*, expected to represent a higher maturity, does not fit at a first glance with other items on this lowest level as *no cost-benefit calculation exists* or the *lack of a success control*.

Level 2

On maturity level 2, no analysis strategy exists, which does not fit at a first sight with the division wide analysis strategy on level 1, as one would expect that with an increasing maturity, the scope of the strategy broadens (inter-maturity level). Additionally, the presentation of analysis results in a digital (pdf-file) or printed format does not correspond necessarily with the distribution of analysis results via a department wide online portal on level one.

Level 3

Maturity level 3 is represented by only one single item, the manual, ad-hoc data quality management. This appears to be suitable on a first glance, as the other items, related to DQM are expected to be associated with a higher difficulty, are assigned to higher maturity levels.

Level 4

No potential errors could be found on this stage. What has been noticeable yet is the great accumulation of items regarding the purpose of data analysis in terms of *classification, exploration*, and *prediction*. Especially the last one could be expected to represent a higher maturity with regard to the needed statistical capabilities of the analyst. Regarding the data dimension, the accumulation of items regarding the source and structure of the processed data is remarkable.

Level 5

Level 5 contains unexpected items as well, as the *Data analysis is based on department wide processes and controls*, although the data analysis based on company-wide processes can be found on level four already.[23] Comparable to level 4, the data dimension contains an aggregation of items regarding the underlying processes of the Data Quality Management, equally representing different stages of maturity, e.g. the differing requirements between *Defined roles for Data Quality Management* and *Automated Data Quality Management*.

Level 6

On level six, the *department-wide analysis strategy* seems not to fit on a first glance, as the *company-wide analysis strategy*, expected to represent the highest maturity within this topic, has been assigned to level four already.[24] The item *success control of the analytical application is carried out irregularly but based on standardized processes* is also assigned potentially wrong as well as the item *regular success control based on standardized processes* - assumed to be the optimal approach - can both be found on level 4.
Regarding the data dimension, the *irregular, manual combination of different data sources* does not necessarily fit with the highest maturity and is in contrast to the automated, standardized combination of data sources, which can be found on level six as well.

Altogether, within the initial model, only a limited number of potential intra- and inter-maturity level errors could be found.
In the next step, the model will be discussed with the members of the focus group (the

[23] One reason can be that with an increasing size of a company, overarching processes can be challenging to establish.
[24] The reason behind can be the same as described for the standardization of the analysis processes. Due to potential different industries etc. within one company, a strategy on a lower level can appear to me more valuable.

participating members can be found in table 5.1) regarding the item distribution. This discussion will contain a more detailed interpretation of the items from a practitioner's point of view. Based on the discussion, the model is adjusted accordingly, resulting in the *fitted model*.

5.6 Model evaluation

5.6.1 Evaluation of the initial model

TABLE 5.10: Contribution Step 6.1 - Evaluation of the initial model

Goal	*Method*	*Result*
The evaluation in how far the item distribution of the initial model is suitable from the focus group point-of-view; The analysis of the influence of the topics novelty on the extent of the re-assignment of the belonging items.	Focus Group Discussion	Several items of the initial model had to be re-assigned to other maturity levels; The extent of the re-assignment differs amongst the different topics and does not depend on the topic's novelty.

Up to this point, the following actions have been carried out: The model scope has been defined (Big Data), followed by the identification of dimensions, which are used to structure the subject in focus (*Organization* and *Data*). Along those dimensions, different topics and belonging measurements have been identified and transferred into a questionnaire. This questionnaire has been used to carry out a survey, returned by 71 companies, which of the identified measurements respective items they fulfil. These data have been used to calculate for each item the item difficulty value β_j as described in section 5.5. Based on the calculated values, the items have been clustered on six maturity level - the *initial model*.

This *initial model* is the starting point for the model evaluation. It has been discussed and evaluated individually in the next step with the members of the focus group. At this point, the focus group has been enhanced by another expert for the model evaluation who so far has not been involved in the maturity development process. This enhancement is made in order to improve the model evaluation. The procedure is supposed to test,

in how far people that are confronted with the initial model - without knowing and having influenced the underlying questionnaire - evaluate the overall model as well as the individual items. Potential differences between existing group members and new members could speak for the already targeted self-fulfilling prophecy. Persons, who have already participated in the construction process might recognize their own aspect from the steps before, e.g. the selection of topics and belonging items. At the same time, the exchange of all focus group members would lead to a loss of a sufficient continuity on the construction process. Therefore, the author of this thesis has decided to carry out the evaluation utilizing a mixed focus group, consisting both of a new as well as established members.

The model discussions have been executed through one-to-one interviews via telephone and a shared desktop application. As a first step, the initial model has been presented by the author to the respective expert. Additionally, the underlying process model has been described briefly to the one who has not participated in the model construction process.[25]

Each item has been analyzed individually with the focus group members, discussing if it can remain on the assigned maturity level and, if not, to which level it should be re-assigned and why.[26] The individual evaluations have been documented, consolidated, and the items have been re-assigned accordingly. Re-assignment in this case means the assignment of an item to another maturity level than originally assigned to, in the initial model, based on the results from the application of the fitted Birnbaum model.

Table 5.11 provides an overview about the changes that have been made to the items in the initial model in consultation with the focus group. The values represent the average of change per topic. The maximum change is five levels.

Both the i) extent and the ii) amplitude of the changes by the focus group have been analyzed by calculating the deviation and squared deviation.[27] For i) extent, each item of a topic has been multiplied by the strength of its change (1 - 5) for each topic, divided

[25] Details on the focus group members participating in this step can be found in table 5.1.

[26] An overview containing the initial model, the re-assignments (if available), and the comments for each item made by the members of the focus group can be found in the file *evaluation-initial-model.xlsx* on the enclosed usb stick.

[27] The maximum results from the six maturity levels, as an item can be changed at maximum five maturity levels, e.g. from level one to six

by the number of items of the respective topic.[28] The calculated values for each topic describe the extent of reassignments by the focus group.

For ii), the squared deviation is calculated the same way yet each extend (1-5) of change is squared (amplitude). This leads to greater emphasis of higher reassignments. This second analysis is carried out in order to identify potential topics, which have undergone only a few changes, but with a very high extend. This high extend would point at a strong error-prone item assignment, based on the results from the Birnbaum model.

The results of this analysis (extend and amplitude) reveal two main insights: first, at least 1 item per topic has been changed after the discussion with the focus group. Second, the extend of changes differs considerably between the topics, from a change of one item per topic (e.g. *Frequency of data analysis*) up to the reassignment of every item of a topic (e.g. *Distribution of analysis results*). It can be shown that the quantitative approach does not lead to entirely inaccurate results, yet cannot be used without a further fitting process.

Taking these insights, an analysis of the discussed reassignments is carried out subsequently in order to

- understand the underlying reasons for the proposed fittings carried out by the members of the focus group, and to
- use the gathered information to evaluate in how far the bottom-up approach can be used for the maturity model construction in a field that has both novel and established aspects.

Model adjustments

Out of the overall topics (table 5.11), both one topic with a high (*Data analysis strategy*) and a low need (*Identification of Data Sources*) for reassignment and the belonging argumentation by the focus group are presented exemplary. The selection was made, as a complete presentation of all items would include a high level of redundancy and thus unnecessary expand the discussion.[29]

Big Data strategy

All strategy related items have been reassigned by at least one maturity level, following

[28] Items, that have not been reassigned are not taken into account, as they are eliminated due to the multiplication with 0 (no change).

[29] The argumentations for each individual topic can be found in the Appendix.

TABLE 5.11: Analysis of the focus group model fitting during the first evaluation step - The first number in the brackets represents the number of items per topics, which have been reassigned, the second the overall number of items per topic.

Topic	Normal Deviation	Squared Deviation
Combination of Data Sources (3|4)	1.8	5.3
Data Analysis strategy (5|5)	1.6	4.8
Distribution of analysis results (4|4)	1.5	4.5
Data analysis project sponsor (4|5)	1.0	5.0
Data analysis project status (4|6)	0.8	2.2
Success Control of Analysis (3|5)	0.8	1.6
Cost Control of Analysis (3|4)	0.8	2.3
Structure and Source of processed Data (2|4)	0.8	1.3
Processing of Analysis Results (2|3)	0.7	1.3
User Group I - Recipient of Reports/-Analysis Results (3|5)	0.6	1.8
User Group II - Definition of Reports/-Analysis (3|5)	0.6	1.0
Purpose of Analysis (2|4)	0.5	1.0
Data analysis process (2|5)	0.4	0.8
Applied Analysis tools (2|5)	0.4	0.8
Identification of Data Sources (3|5)	0.3	0.7
Frequency of analysis (1|4)	0.3	0.3
Data Quality Management (1|5)	0.2	0.2
Average	0.77	2.05

the idea that the larger the reach of a strategy and the affected departments, the more mature the company is; a company-wide data analysis strategy proves a rather holistic and thus a more advanced approach. During the discussion, the influence of the company size on the scope of the Big Data strategy has been discussed. Especially smaller companies with a focus on technology may be very mature regarding their processes but lack a strategy due to their size or a lower perceived need. At the same time, the definition of a company-wide analysis strategy might not be suitable for a multi-national corporation due to its country-specific attributes and directions.[30]

[30] Another indication for the novelty of the topic data analysis strategy is the drop-out rate, as 91% of the respondents, which did not complete the questionnaire, cancelled the answering at the strategy

Identification of data sources

The next topic is the underlying process regarding the search for new data sources, targeting Big Data specific capabilities. Again, the item *focusing on existing data within the data warehouse* - expected to represent the lowest maturity - has not been reassigned on maturity level one. The focus group agreed that the existence of a dedicated data warehouse speaks for the availability of a basic structure for data analysis. Consequently, this item has been reassigned to level two. The discussion points out the difficulties connected with the development of a maturity model covering the range of completely immature to highly mature companies for a broad topic like Big Data.

The group also agreed on the high maturity of the two items regarding the development of a data landscape (*Development of a data landscape* and *Development of a data landscape and pre-evaluation of data sources*). Following the experience of the focus group members, those capabilities exist only in a very few companies. Its rare existence is influenced, amongst others, by the effort needed due to the high number and heterogeneity of data sources, especially for larger corporations. At the same time, larger corporations are the ones that might benefit more from a data landscape due to the diversity of existing data sources. Additionally, the lack of an appropriate software tool to display the structure of and the connections between the different data sources also hinders the more frequent development of data landscapes.

During the discussion of the proposed re-assignments with the focus group, the aspect of the overall market maturity has been pointed out. Currently, only a very few companies are associated with higher maturity levels, especially with level six. These rare practical examples for those capabilities in turn hardens the assignment of items to the suitable maturity level for the members of the focus group as only limited experience exists with companies and their related characteristics on that level.[31] This aspect points out the relevance of incorporating experienced members in the focus group to gain a broad knowledge base with numerous different companies and related capabilities. This

question. This might be influenced as well by the positioning of the questions as the first question of the questionnaire.

[31] One aspect that has been mentioned by different members of the focus group is the potential influence of company characteristics on the assignment of the items. The aspect of strategy can be used as an example. In most cases, a strategy on divisional level or company-wide represents a highly perceived relevance of the topic data analysis on the management level and therefore a high maturity. This statement does not necessarily account for smaller companies, where the data analysis strategy is defined on a management level due to the rather small size of the company instead of the perceived relevance.

is given in the research at hand as the average industry experience for the members is more than eleven years.

The quantitative results from table 5.11 in combination with the argumentation from the focus group allow a first evaluation, in how far a quantitative, bottom-up approach can be applied for maturity model construction in a domain, containing both novel as well as established aspects.[32]

i) At least one item per topic had to be reassigned, in average 0.77/2.05 levels (normal deviation/squared deviation).

ii) The difference between the normal and the squared deviation shows, that a few topics have been reassigned more extensively than others - therefore, no general statements about the need for reassignment with regard to the model evaluation step can be made.

iii) On a more detailed level, the quantitative analysis of the reassignment (table 5.11) shows, items with a novel character - those which are mostly characterizing for Big Data - have not been reassigned more extensively compared with items that can be characterized as more established.

iv) Both in the initial as well as the fitted model, the number of items per maturity level is not equal along the six levels. Not every topic is represented on each maturity level with one item.[33]

Altogether, as the novel topics did not show an increased need for reassignment compared with established topics, it can be stated, that in general the quantitative bottom-up approach could be applied successfully. The discussions with the focus group along the individual items in an application-context, enhanced with anecdotic explanations allowed a deeper understanding of the items' relevance and its appearance in the practical context. In particular for a novel topic like Big Data, the comprehensive evaluation of the initial model has proofed itself as relevant and necessary.

[32] A concluding evaluation is not possible as the statements can only be made based on the fitting results gathered during this construction process. The results of further construction processes of models for objects with a mixed maturity, are needed to improve the explanatory power.

[33] This aspect is of relevance for the calculation of a companies' maturity level in construction step 6.2., where it will be discussed in more detail.

The evaluation processes described in this section - the evaluation of the initial model in consultation with the focus group - resulted in the *fitted model* (figure 5.5). The evaluation of the fitted model based on the deployment will be discussed in the next section. Up to this point, the *fitted model* does not represent a complete maturity model as it lacks the textual description of the different maturity levels.

5.6.2 Evaluation based on the deployment of the fitted Model

TABLE 5.12: Contribution Step 6.2 - Evaluation based on the deployment of the fitted model

Goal	*Method*	*Result*
The evaluation in how far the fitted model represents the understanding of maturity in a practical context.	Discussion with the Focus Group, Qualitative Analysis	The maturity evaluation of eleven companies based on the developed model shows a high agreement with the maturity evaluation, carried out by the members of the focus group, of the same companies; the model represents the practical understanding of maturity correctly.

After the initial model has been fitted based on the input from the focus group, representing the first phase of the model evaluation (step 6.1), the fitted item distribution can be found in figure 5.5, describing the *fitted model*.
Based on this fitted model, the second step of the evaluation (Construction Step 6.2) will be carried out in this section. The result will be the final model, described in construction step 7 (section 5.7).

The fitted model is evaluated by applying it to different companies and comparing the resulting maturity with the industry experts' maturity evaluation of the same company. The basic requirement is a personal, comprehensive knowledge by the respective expert with the company to be evaluated. In doing to, the focus group members contribution is twofold:

	Organization Dimension	Data Dimension
6	▪ Result provisioning on mobile devices	▪ Development of a data landscape and evaluation of potential data sources ▪ Automated DQM ▪ Analysis of internal/external structured/unstructured data
5	▪ Use-oriented cost-benefit calculation ▪ Automated result processing ▪ Usergroup II – Middle Management ▪ Standardized, documented, and controlled data analysis processes on department level ▪ Usergroup I – Company-wide ▪ Analysis are carried out continuously (real-time) ▪ Top Management sponsor ▪ Project-based cost-benefit calculation ▪ Mandatory, company wide data analysis processes and controls ▪ Usergroup II – Company-wide ▪ Company-wide strategy ▪ Result provisioning on company-wide online platform ▪ Success-oriented cost-benefit calculation ▪ Flexible, pro-active analysis solution ▪ Analysis for prediction ▪ Cross-department strategy	▪ Defined DQM roles ▪ DQM team ▪ Defined DQM processes ▪ Analysis of internal structured/unstructured + external structured data ▪ Development of a data landscape for company-internal and external data
4	▪ Automated, standardized combination of data sources ▪ Usergroup I- Individual Analytics Department ▪ Standardized, irregular success control ▪ Department-wide strategy ▪ BI department sponsor ▪ Regular user meetings for success control ▪ Integration of analytical front-ends on reports ▪ Analysis for classification ▪ Analysis for exploration ▪ Standardized, documented data analysis processes on department level ▪ Standardized, regular success control ▪ Central IT department sponsor ▪ Use of analysis software ▪ Usergroup II – Individual analytics department ▪ Analysis are carried out daily ▪ Division-wide strategy ▪ Result provisioning on department-wide online platform	▪ Regular screening for further, company-internally and externally available data ▪ Irregular screening for further, company-internally available data ▪ Analysis of internal structured and unstructured data
3	▪ Analysis phase of the analysis relevant project is completed ▪ Project is currently implemented ▪ Manual result processing based on a standardized process ▪ Usergroup I – Key User ▪ Decentralized IT department sponsor ▪ Sporadic user meetings for success control ▪ Focus on already processed data ▪ Analysis for monitoring ▪ Established data analysis processes based on routines ▪ Central department management sponsor ▪ Project is completed ▪ Usergroup I – Middle Management	▪ Manual ad-hoc DQM ▪ Manual, regular combination of data sources
2	▪ Usergroup II - Single Data Analyst ▪ Ad-hoc analysis (e.g. spreadsheet based) ▪ Manual result processing without a standardized process ▪ Static reports ▪ Analysis are carried out weekly ▪ Usergroup I - Single Data Analyst ▪ No Project exists but planned within the next 12 month ▪ Usergroup II – Key User	▪ Analysis of internal, structured data ▪ Manual, sporadic combination of data sources
1	▪ Analysis are carried out less than weekly ▪ No strategy exists ▪ Printed/digital (.pdf, .xls) result provisioning ▪ No cost-benefit calculation ▪ Informal data analysis processes / no standardization ▪ No success control exists ▪ Project is planned ▪ No project exist, no project planned	▪ No combination of data sources

FIGURE 5.5: The fitted model resulting from the second evaluation step

i) Each member answers the already known questionnaire from the population phase (Step 5) for a number of companies he is familiar with. This dataset is the input for the maturity assessment of the companies based on the fitted model.[34] [35]The only demand for the company selection by the focus group members has been a deeper knowledge of its specific Big Data relevant capabilities. Therefore, the last project with the respective company was not allowed to be longer ago than one year. By doing so, it was ensured that the member of the focus group has gained substantive, recent knowledge about current practices regarding processes, capabilities etc. in those companies. For each of those companies to be evaluated, the items in the questionnaire that describe the status quo have been ticked either digitally or in a printed version by the respective focus group member, who has proposed the company. Based on the number of items ticked per maturity level, the overall company's maturity is assessed.

ii) Additionally, the member is asked to evaluate the same companies' maturity on a scale from one to six regarding the existing capabilities in the field of Big Data and explain the driving aspects for his assessment. At this point, the members do not know how the fitted model looks like to make sure, that the evaluation is carried out based solely on the expert's practical experience. This experience contains the insights into different companies during consulting projects. Based on these learnings, the experts do have a personal understanding of different levels of maturity, based on the capabilities and can compare different companies.

One challenge that occurs on the way from an answered questionnaire to a calculated maturity level results from the unequal number of items per maturity level - resulting from the analysis of the quantitative bottom-up approach - comparable to the work by Marx et al. [2010]. Therefore, different potential approaches in order to deal with this

[34]Different ways of calculating a maturity based on the data from the answered questionnaire will be explained later in this section.

[35]Due to the anonymous data gathering for the model population, it cannot be ruled out that companies, which are selected by the members of the focus group for this second evaluation phase, are already part of the dataset gathered during the population phase (Construction step 5). Nonetheless, this has no further influence on the overall construction process as the company data in the second evaluation step do not influence directly the model construction; they are solely used for the individual maturity assessment. Instead, the argumentation of the focus group member for his evaluation is influencing the evaluation.

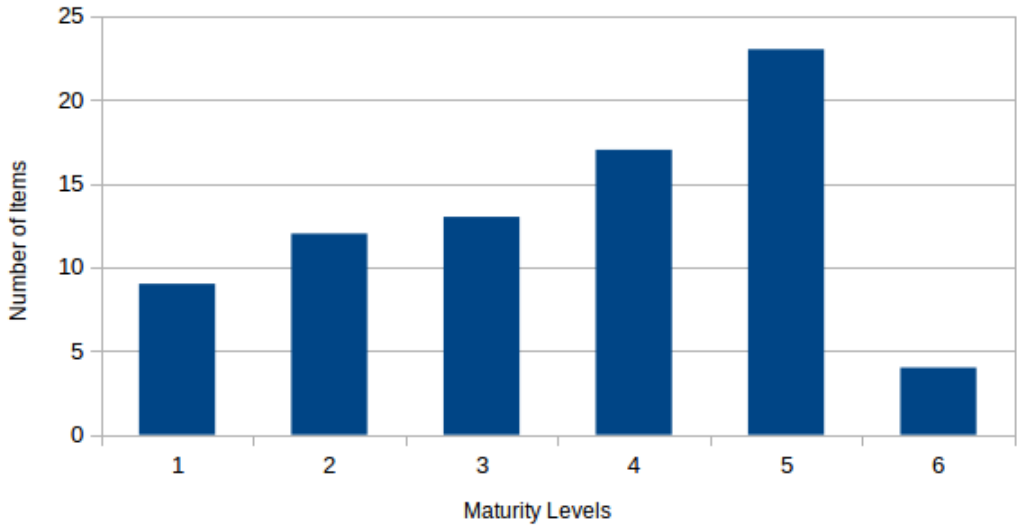


FIGURE 5.6: The number of items per maturity level of the fitted model

inequality are presented before continuing with the description of the actual assessment (construction step 6.2).

Utility Analysis

One approach to define the companies' capability based on the filled questionnaire is a utility analysis. The product of the number of items checked per maturity level i, multiplied with the respective maturity level m, is divided by the number of overall topics (except the question regarding company characteristics) within the questionnaire, used for the maturity model development, in this case 17. The resulting value is the calculated maturity level for the respective company.

This procedure does not take the differing number of available items for each maturity level into account. A company is more likely to end up on maturity level four and five as these levels hold the majority of the items and also hold partly more than one item per topic - the main challenge as described before.[36] Due to the unequal distribution of items in the model, the *utility analysis* based approach seems not suitable. As not each of the 17 topics is represented with at least one item on level six, this level cannot be achieved when calculating the companies' maturity based on a utility analysis. That obstacle leads to a situation in which a company that is fulfilling all items belonging to level six

[36] Asking the focus group members to assign only one item of a topic to each maturity level is not suitable, as not every topic might be relevant for each maturity level. Vice versa, two or more items of the same topic can be relevant for the same maturity level.

(in this research covering four different topics), would nonetheless end up on level five as only a part of the overall topics are represented on level six, the remaining ones can be found on lower maturity levels. Consequently, a second approach is presented, which takes the different number of items per maturity level into account.

Normalized utility analysis

One possibility to deal with the challenge of the unequal number of topics per maturity level is to enhance the calculation with a normalization towards the optimal result, in this case maturity level six.

In doing so, a scaling factor is calculated, again based on the number of questions in the questionnaire, $n_q = 17$, that represents the number of topics. In case one item of each topic would be represented on each of the six maturity levels, level six would hold 17 items, representing the 17 topics.[37] As level six does not hold an item of every topic, the optimal result for a company could be that every item on level six (4 items) and five (13 items) are marked by a respondent, as this represents the best possible result.[38] [39] Therefore, the resulting scaling factor is calculated based on those answers, which results in the highest maturity:

i) Starting with the highest maturity level, for each level the number of different contained topics, representing the maximum number of items that can be ticked on this level, is multiplied with the respective maturity level.

ii) The result is divided by the overall number of available items on this level.[40]

iii) The results for each maturity level are added up.

iv) The resulting sum represents the scaling factor.

In the fitted maturity model, at least one item for the 17 covered topics can be found on level six and five. Therefore, the calculation for the remaining maturity levels would results in zero, as for the remaining maturity levels four to one, no further items can be

[37] In this scenario, the application of the utility based analysis would lead to a calculated maturity of six in case every item on level six is correct for a company.

[38] The number 13 represents the 13 remaining topics, that are represented with at least one item on level five.

[39] A differentiation between the two dimensions is not pursued due to the relatively low number of data dimensions related item. Consequently, a dimension-specific maturity calculation is not carried out.

[40] Especially for maturity level four and five, the number of topics per level and the number of items per level differ, as several topics are represented on the level with more than one item.

ticked, because the 17 topics are already covered by level six and five.
For the scaled maturity calculation, each maturity level (1-6) is multiplied with the number of items checked on the respective level, divided by the number of available items on this level. This procedure is carried out for each level and the results are added up. The resulting number is divided by the scaling factor and multiplied with the overall number of available maturity levels.

Both described methods - utility based analysis and normalized utility analysis - are potential approaches to calculate a distinct value for a companies' maturity. However, both of them have not been discussed so far in the scientific community and need - before being applied - a further discussion of their suitability based on their application for different maturity levels. Nonetheless, the approaches have been presented at this point in order to carve out the challenges and potential solution approaches for the maturity level calculation in the context of this research.

Besides the *unequal distribution of items* along the different maturity levels, a second aspect is challenging the calculation of maturity levels during the model evaluation (construction step 6.2): the *different scales of measurement* that are used during the maturity model population and evaluation (section 5.5 and 5.6). As this aspect is of relevance for the following elaborations on the application of the maturity model, it will be discussed at this point in research instead of in the limitation section in Chapter 6.
In the initial calculation of the item difficulty parameter for each item individually, a ratio scale was used, containing an absolute zero and ranging from 0 to 1. When clustering the results on the different maturity levels, the measurement scale was changed to an ordinal scale, ranging the items from one to six, but without the possibility to explain the differences between two maturity levels mathematically. A difference between the levels based on the individual differences of the item difficulty of each item could only be carried out for the initial model.
As the focus group changed the order of the items, the sum of the item difficulty cannot be taken as a measure to determine the difference between two maturity levels. Consequently, the "breadth" of maturity levels differ within the model, which in turn complicates the interpretation of the results from the calculation approaches described before.

With regard to the described challenges, resulting from

i) the unequal distribution of items on the different maturity levels and

ii) the different scales used during the maturity model construction,

none of the described approaches for the calculation of a companies' maturity (utility analysis and the incorporation of a normalization factor) have been pursued further. The hence resulting values would have been most likely imprecise and would have not supported a sound model evaluation.

Instead, the number of the fulfilled items per maturity level, based on the answered questionnaire, has been compared with the personal assessment from the focus group member.

The process has been carried out as described in the beginning of the model construction step 6.2. In a first step, the questionnaire (section 5.5.2) has been completed by the industry expert for companies for which he is familiar with Big Data relevant capabilities, based on past experiences from consulting projects.[41] Those data from the answered questionnaire have been used to calculate the degree of fulfilment per maturity level: the number of ticked items per maturity level divided by the number of topics per maturity level. The resulting number shows, in how far a company fulfils the capabilities, associated with the individual maturity levels.

In a next step, the expert has been asked to assess the maturity for Big Data of the respective company on a scale from 1 - 6 and additionally to describe the reasons behind his estimation. The experts' maturity assessment has then been compared with the degrees of fulfilment per maturity level that have been generated from the model application. The goal was to identify which arguments have influenced the experts' evaluation and which of those have not been considered in the constructed maturity model or are perceived as being assigned to the wrong maturity levels.

The advantage of this approach is twofold. First, it allows avoiding the weaknesses of the solely quantitative calculation of a maturity level based on the answered questionnaire as described before (*unequal distribution of items/different scales*).

Second, by discussing the focus group members' assessment, the guiding aspects from a

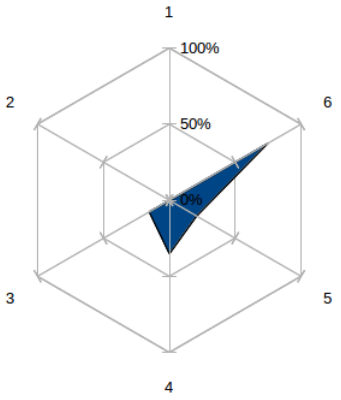
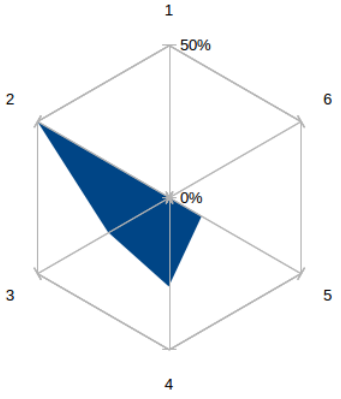
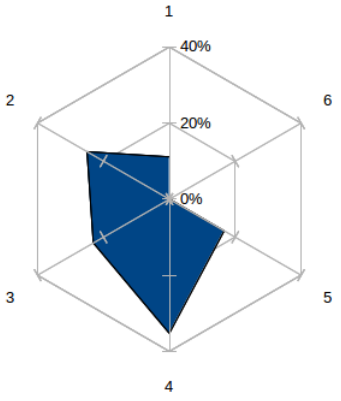
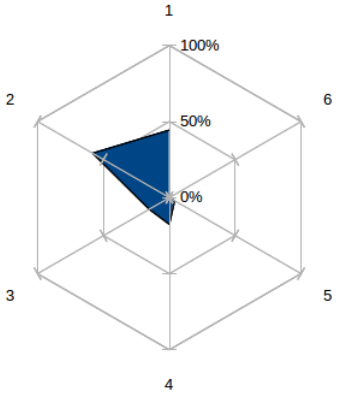
[41] In the forefront, the industry experts, who have not participated in the model development so far, received a short introduction into the concept of maturity and the questionnaire without pointing out the different levels of maturity, which are associated with the different items.

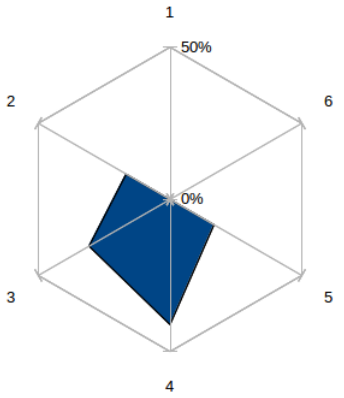
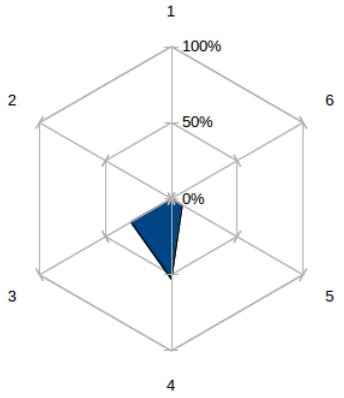
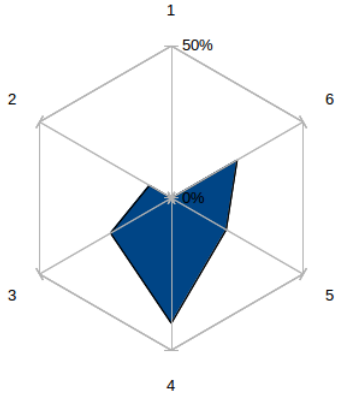
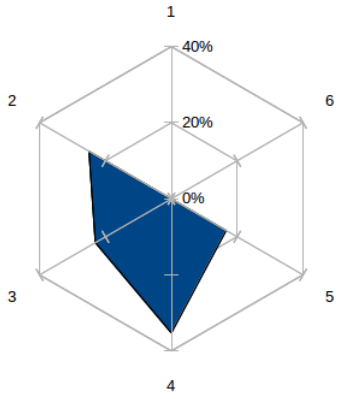
practitioners' point of view can be identified and included in the fitted model. This allows an optimal representation of the practical understanding of maturity in the model, fostering at the same time the relevance of the final model.

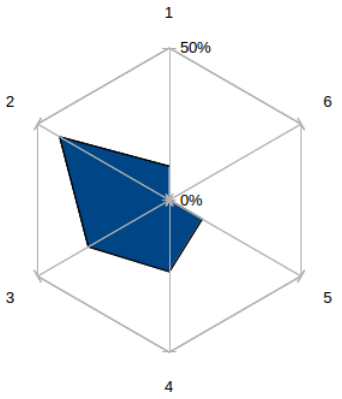
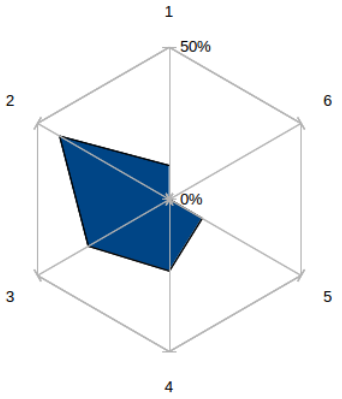
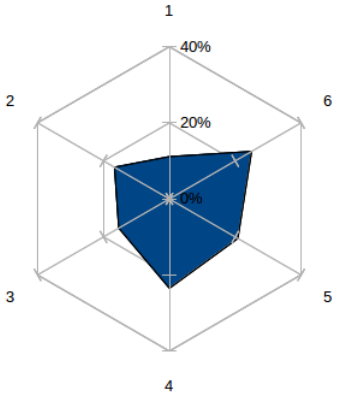
The evaluation of the fitted model has been carried out based on eleven companies, proposed by the members of the focus group. In the following description, with regard to the scope of this work, a selection out of those companies is described in detail. Those companies, whose evaluation resulted in the identification of further aspects, that have not been covered by the model yet have been selected (companies 5 & 8).[42]
The companies evaluation is distinguished in two parts: i) Those whose evaluation has been carried out by an expert that has already been part of the focus group before (companies 1-5) and ii) those, that have been evaluated by a new member in order to carve out potential differences (companies 6-11). For each part a detailed company evaluation is presented. Within this description, for each company, the characteristics (industry, number of employees), the estimation of the industry expert (ML_{expert}) and a radar chart, describing the degree of fulfilment per maturity level based on the fitted model are described. Each of the axis of the radar charts stands for one maturity level and shows the degree of fulfilment.[43] An overview about the comparison of the eleven companies is given in table B.1. The analysis of the aspects, which have driven the experts' evaluation, distinguished into those aspects, which are already covered by the model and those, which have been newly identified is given in figure 5.7.

[42]The in-depth discussion of the remaining companies can be found in the appendix.
[43]With regard to readability, the maximum scale level is not always 100%. It is fitted depending on the maximum degree of fulfilment.

TABLE 5.13: Overview evaluation results step 6.2

No.	Industry	Employees	Expert evaluation	Degree of fulfilment
1	Service Industry	500 - 1000	4-5	
2	Retail	1000-5000	2-3	
3	Service Industry	1-100	3	
4	Banking/ Insurance	1-100	1-2	

No.	Industry	Employees	Expert evaluation	Degree of fulfilment
5	Telecom-munication	1000-5000	4	
6	Banking/ Insurance	100-500	3-4	
7	Telecom-munication	>50,000	4	
8	Banking/ Insurance	1000-5000	3-4	

No.	Industry	Employees	Expert evaluation	Degree of fulfilment
9	Retail	1000-5000	3	
10	Banking / Insurance	5000-10000	2-3	
11	Banking / Insurance	1000-5000	4	

Company five represents a company that has already reached a higher maturity, but is not yet able to move up to the top level, which can be found similarly in numerous other companies, following the experts' opinion. The maturity of the basic analytical task in terms of reporting and classical data mining (e.g. next best offer, churn prevention) is on a high level. The related processes are already partly automated and the results

are processed, followed by a distribution on a department wide platform and further handled by different sub-departments as well. There exists an independent analytics department, defining the existing analysis and working with the analysis software applications. Nonetheless, the recent processing focus is on data already stored in the data warehouse, primarily gathered based on the customer relationship management software. Following the expert's evaluation, the company is already aware of further potentials in the field of data analysis; therefore several analysis relevant projects are currently conducted, amongst others the combination of company internal and external data and the identification of further potential fields of analysis. However, a recent problem that slows down the positive development is the lack of coordination in the different projects, despite existing overlaps.

An indicator for the highly perceived relevance of the data analysis is the existence of a management-oriented, department wide defined data analysis strategy, whose implementation yet is a problem area. Existing approaches have an explorative character, testing new applications and thereby leading recently to numerous isolated applications. A generalization and integration is planned but an agreement on one tool could not been achieved so far due to the perceived insecurity of the decision makers regarding the actual use of further applications. This problem is not tackled further internally, a success control is only based on irregular user meetings. The expert compared the recent situation of the company, which can be found similarly in numerous other companies, with the chicken-and-egg problem. Without an initial implementation of the software and definition of respective processes, the evaluation of the future benefits is difficult, although the potential use is supposed to act as decision-making criteria.

Despite the numerous integrated data sources, the DQM is still carried out manually. Currently, the driver of the analysis movement is the BI department, perceiving in turn the IT department rather as a deliverer of the data and the related infrastructure, although the IT department is currently one primary Big Data project sponsor. At the same time, they are the only department with a sufficient knowledge for setting-up the necessary future infrastructure. However, the IT department lacks in knowledge how to evaluate the analysis-relevant business processes. Altogether, the company in focus has already professionalized the field of data analysis but will take several years to further develop, primarily slowed down due to organizational challenges. The challenging role of organisational aspects proofs the relevance of the *organisation dimension* in the context

of Big Data, although it is not yet present in recent research as shown in Chapter 2. The highest degree of fulfilment for company five can be found in level four, which is coherent with the experts evaluation.

Company eight already has extensive experience in the field of data analysis. In the past, the analytics department has developed a comprehensive reporting structure, offering insights into product and customer key figures (e.g. value contribution per product/customer), processing both company internal (structured/unstructured) as well as company external (structured) data. More comprehensive projects have already started, focusing on infrastructural aspects such as the implementation of new analysis software applications. These projects aim at creating a homogeneous application landscape throughout the company: a "Big Picture" is drawn, supported by the definition of a department wide data analysis strategy and the department-wide distribution of analysis results. Furthermore, a project-based cost-benefit calculation is carried out.
Yet, the recent challenges are located in the fields of i) knowledge and ii) data handling. Aspect i) results from the recent ambitions of the marketing department to develop individual models, e.g. for the prediction of the acquisition of new customers, based on matters such as marketing expenditures. As no further employee training has been carried out so far, the employees in the marketing department are overstrained, both with the handling of the software as well as with the methodological aspects. The resulting high number of support requests for the Business Intelligence department slows down their performance due to the small number of employees with the relevant knowledge. As a result, the marketing department is not able to demonstrate the value contribution of their projects. A second knowledge aspect targets the management that is interested in data analysis but is not yet aware of the full potential. The interest results partly from a perceived external pressure, as competitors increasingly focus on this topic.
The infrastructure related aspect ii) targets the separated data management of data from the operating systems and the web server data (click stream data from the company's homepage) in two different data warehouses. Those data are not integrated in recent analysis tasks (e.g. the customers reaction to a newsletter). Comparable to most of the other companies that have been evaluated, the data quality management is carried out manually. Potential errors are identified based on questionable reporting results which then are corrected manually.
During the interview with the industry expert, a potential enhancement of the item

set for the maintaining process has been mentioned, pursuing the coordination of different analysis-relevant projects. Especially when several projects are carried out in a parallel fashion, coordination would foster the overall project's success in avoiding the same repeated efforts. Consequently, the topic of project management with a focus on coordination would allow assumptions on the maturity and thus could be integrated in the next model.

Nonetheless, the model-based assessment is close to the experts' one, with the highest degree of fulfilment for the maturity levels 3 and 4, comparable with the experts assessment of 3-4.

Main findings from the model evaluation

In this step, the maturity of eleven companies has been evaluated using the constructed Big Data maturity model; two have been described in detail. The results have been compared with the maturity assessment of the same company conducted by an industry expert. As shown in table B.1, the experts' evaluation and the related model-based assessment show a high agreement.[44] It appears that the fitted model, resulting from step 6.1. is already representing the practical understanding of Big Data maturity correctly. This in turn supports the applied methods, both the use of a focus group in general in the course of the model construction, as well as the selection of focus group members with a broad knowledge basis, allowing the targeted development of an application-independent, cross-industrial model.

Altogether, the application of the model results in similar maturity evaluations as those made by the industry experts, no major deviations between the two maturity estimations have been found. Figure 5.7 shows which aspects with relevance for the maturity assessment from the experts point of view have been covered by the model and which are missing.[45] Without any influence on the selection of the companies for construction step 6.2, the maturity-influencing aspects named by the focus group members during the company-individual discussion are covering a broad range, both organizational, as well as data, technological and methodological topics.

Existing differences between the model's maturity evaluation and the experts' evaluation can be explained partly by the experts' integration of the aspect "future potential" in the

[44] By comparing the experts' assessment and the model-based assessment (using the degree of fulfilment), it can be shown that the differences between the model based assessment and the experts' assessment are not higher for those experts, which have been added to the group for construction step 6.2.

[45] The aspects are distinguished by the belonging dimensions Data (D) and Organization (O).

field of Big Data, which is not covered by the current model, yet influencing the expert's overall evaluation. In contrast, the model evaluated the as-is situation. For multiple companies, the experts stated that companies that already have reached a higher level of maturity could be more professional based on the existing resources, e.g. infrastructure and workforce, and therefore the experts tended to evaluate those companies lower than the model did.

One further aspect, which has an influence on potential differences between the model outcome and the experts evaluation targets the emphasize of individual aspects by the expert. Taking the organizational location of analytics described for company nine as an example, this aspect has dominated the experts evaluation, resulting in an average evaluation based on a spill-over effect (Appendix B). This effect cannot be eliminated completely as it is an inevitable part of the integration of the experts person opinions [Calder, 1977].

In the course of the model evaluation in construction step 6.2, two topics could be identified, that were not covered by the existing model so far: the topic of i) *knowledge management* and ii) *project coordination*. Targeting i), one goal for a future maturity model could be to carve out if and to which extend a structured knowledge management approach for Big Data relevant capabilities exists, allowing the re-use of already generated knowledge from executed projects.

Topic ii) helps to identify the extend and scope of a project management for analytic relevant projects, which again supports the aspect of re-use, for example the company-wide publishing of already consolidated databases, relevant for different projects. In addition, the extend of the project management allows to draw conclusions in how far a company is interested in the *Big Picture* of analytics, combining the insights from different projects.

Evaluation interpretation

After the description of the results from construction step 6.2, the gained insights are discussed with regard to the explanatory power. The high agreement between the model and the experts' based maturity evaluation gains in relevance considering that the consultant and the model do have different horizons regarding the maturity evaluation. The model covers two out of four dimensions (*Data* and *Organization*), that characterize Big Data. As the measurements are targeting only data and organizational aspects, the

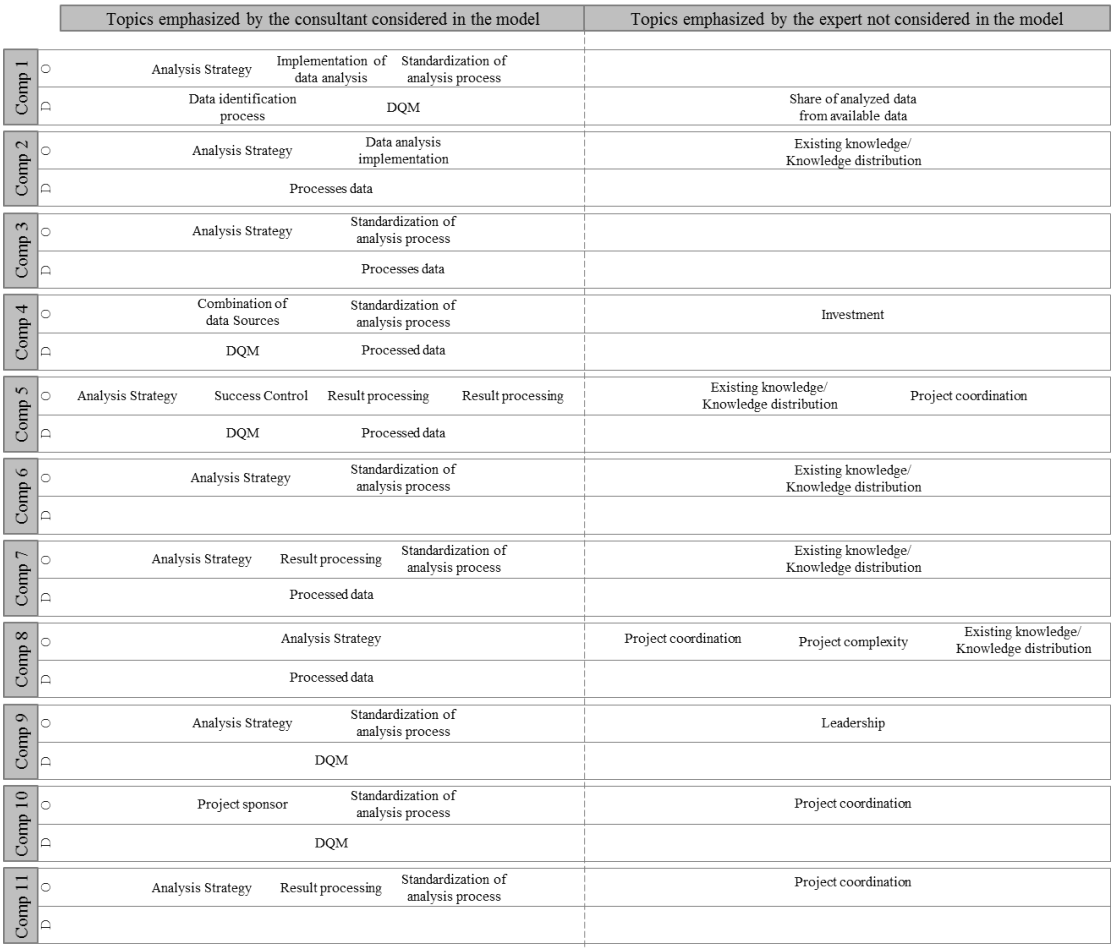


FIGURE 5.7: Maturity evaluation-relevant aspects from the focus group members point of view and their coverage by the developed maturity model

maturity evaluation is referring to these two dimensions.

In contrast, the industry expert has a holistic view on the company, including all four dimensions in his evaluation, which became obvious during the interviews. Although a high number of organizational aspects have been influenced the experts evaluation, infrastructural aspects were mentioned as well, e.g. the data warehouse structure or the applied analysis software.[46] Keeping this different basis of evaluation in mind, the relatively low differences between the model and the expert's evaluation could imply that the maturity of organizational and data topics can be used as an indicator for the overall company's maturity. Taking company six as an example, the respective expert described a highly developed infrastructure but due to a lack of process standardization etc., the available potential of the infrastructure is not sufficiently exploited, leading

[46]The experts have not been asked to limit their evaluation to the two dimensions of the model as a distinct differentiation is hard to achieve for the experts.

in turn to a lower maturity evaluation of the consultant. This aspect holds potential for future research targeting the relation and interaction of the maturity of individual dimensions, which will be described in detail in Chapter 6.

A challenge of the conducted evaluation approach, described in this section, results from the integration of consultants as members of the focus group into the evaluation process. The mandating of a business consultancy to develop the field of Big Data allows conclusions regarding the company's size and/or professionalism. The companies selected by the members of the focus group had at least already basic capabilities in the field of data analysis and tend to be on a higher maturity level than level one. Therefore, the evaluation of the model for its correct representation of maturity level one is covered primary by the first evaluation step, as no companies on level one have been assessed during the second evaluation step. At the same time, however, the input from consultants is valuable and needed, as they have insights into numerous companies, which allows them a sound comparison and benchmarking of the capabilities. Therefore, the aspects of potential biases mentioned are a general challenge when integrating consultants into a maturity model evaluation process and not distinct for this research. The sane accounts for level six as no companies on that level have been assessed.

Based on the results of the model evaluation in construction step 6.2, no further fittings of the model are needed. The results of the focus group members' evaluation of eleven companies have been congruent with the assessment based on the deployment of the model, the majority of the maturity-relevant aspects from the focus group members point of view have been represented in the model as well. This correct representation of the practical understanding of maturity in the developed model resulted in the model approval by the focus group. The identified topics for potential enhancements are discussed in Chapter 6 in the outlook.

Altogether, after the approval of the model by the focus group members, the *fitted model* has been transferred into its final state, the *final model* as explained in the next section. Referring to the developed construction approach, in the event that the industry experts would have rejected the model, the loop back would be able to lead to three different process steps in the construction model, depending on the experts' feedback.[47]

[47] Potential starting points are *Identification of dimension* in the case that industry experts which have not participated yet in the model construction process miss a relevant dimension, the same accounts for starting point *select design level*. The third starting point is the *model population*, relevant in case

As the next construction step, the approved *fitted model* has been documented. This transfer of the item list into a textual description with a few sentences per maturity level results in the *final Big Data maturity model*.

5.7 Step 7 - Documentation of the final model

TABLE 5.14: Contribution Step 7 - Documentation of the final model

Goal	*Method*	*Result*
Characterization of the individual maturity levels based on the items assigned to the respective maturity level.	Induction	For each maturity level, precisely describing sentences are defined, holding the main aspects that are associated with the respective maturity.

After the fitted model has been evaluated successfully, the items per level are used as a basis for the documentation of the final model. In a first step, the dimension-individual maturity levels are described.[48] Second, the two dimensions are aggregated and the overall maturity levels are characterized as shown in figure 5.8.

Data Dimension Maturity

Level 1: Companies associated with maturity level one do not combine different data sources, therefore the existing analysis cannot be carried out on a consolidated database.

Level 2: Companies on level 2 limit their data scope on company internal data. The existing data sources are combined irregularly and manually. Potential further data for analysis purposes are not considered.

Level 3: From level 3 on, Data quality management becomes a relevant topic for organizations, yet they still follow a manual, thus less advanced approach. Similar, the data in focus from different sources on this level are combined on a regular basis, yet still in a manual manner. The related databases are updated frequently on a weekly basis.

Level 4: Companies on level 4 are aware of the diversity and dynamic of data available. Sources and types of data increasingly become a topic of interest; additionally

the model testing reveals that multiple maturity-relevant topics of the existing dimensions are missing in the model.

[48] Due to the relatively low number of items for the data dimension, the description are shorter compared with the ones from the organization dimension.

unstructured data are processed as well. The company is regularly searching for further potential data sources in- and outside the company. With the increasing number of used data sources, the combination of different sources is carried out automatically.

Level 5: The awareness of the relevance of data multiplicity within companies on level 5 leads to standardization approaches. The identification of further data is based on a process model, resulting in the development of a data landscape, which covers both a company internal- and external perspective. With the increasing number and diversity of the data analyzed, the aspect of data quality management becomes a major topic, resulting in the occupation of Data Quality Management teams.

Level 6: Within companies on level 6, the potential of data analysis - processing company internal and external data with different degrees of structure - is fully recognized. In order to identify the numerous potential data sources, a pre-evaluation of data regarding the potential benefits of analysis is carried out. DQM is carried out automatically owing for the heterogeneity of the processes data.

Organization Dimension Maturity

Level 1: Data analysis as a dedicated process is only a side issue for companies associated with level one. No standardization in the field of data analysis is carried out, the existing processes are informal. As no overall data analysis strategy exists, the existing actions and Big Data relevant projects are not set into an overall context. The static format of the distributed analysis results (.pdf files or printed documents) hinders an integration of analysis results in existing business processes.

Level 2: Maturity level 2 is characterized by initial movements towards a professionalism of data analysis tasks. At least one analysis relevant projects are planned for the upcoming 12 months. Furthermore, an irregular success control of the data analysis tasks is set up. Despite the increasing interest and professionalism, the aspect of data analysis is still focused on one single person in terms of an individual data analyst. The results are used manually without integration in existing processes. This goes along with the use of spreadsheets as a tool to carry out analytical tasks.

Level 3: In organizations associated with maturity level 3, the aspect of analytics has caught the attention of the middle management and increases in relevance throughout the organization, as at least Key Users, responsible for the definition of analysis tasks for each department, can be found. The analysis is yet limited to the monitoring of key figures. Despite the increasing relevance, the aspect of standardization on a process

level still plays a subordinate role within the organization. The data analysis is carried out based on established processes or routines.

Level 4: Level 4 is characterized by the aspect of increasing professionalism and, especially in contrast to level 3, by the aspect of standardization. Fostered by a high level management sponsoring, both the analysis as well as the subsequent success control are standardized. Existing applications of data analysis are becoming more diverse, including classification and exploration. Moreover they are following a strategy on divisional or cross-divisional level. The analytics department is responsible for the design of the analysis tasks and reports. In addition to analysis processes and control, the result distribution is also standardized based on a department-wide platform, targeting the aspect of knowledge management.

Level 5: Level 5 is characterized by a penetration of data analysis applications throughout the whole organization. The objectives are defined and communicated based on a department or company-wide analysis strategy. This includes a change of the application landscape towards flexible, pro-active analysis solutions. The integration of analysis results into existing processes is automated, allowing for the continuous processing of analysis tasks. The topic of success and cost control becomes relevant with the increasing use of analysis applications throughout the company.

Level 6: Companies associated with level six use applications that allow to make the analysis results available on mobile devices.[49]

After the individual dimensions have been described, the overall final model has been carved out. Basis for this description has been the fitted model, consisting of a number of items per maturity level and dimension as presented in figure 5.5. The key characteristics of each level, both of the data and organization dimension have been used to formulate an aggregated, self-explaining characterization for each level. The resulting model can be found in figure 5.8, representing the goal of this thesis.

[49] As described before, the low number of measurements on this level is a result from the bottom-up approach as i) the method leads to an accumulation of items on the middle levels and ii) the challenge of identifying maturity measurements of higher levels for a rather novel topic [Raber et al., 2012]. Potential enrichments will be discussed in Chapter 6.

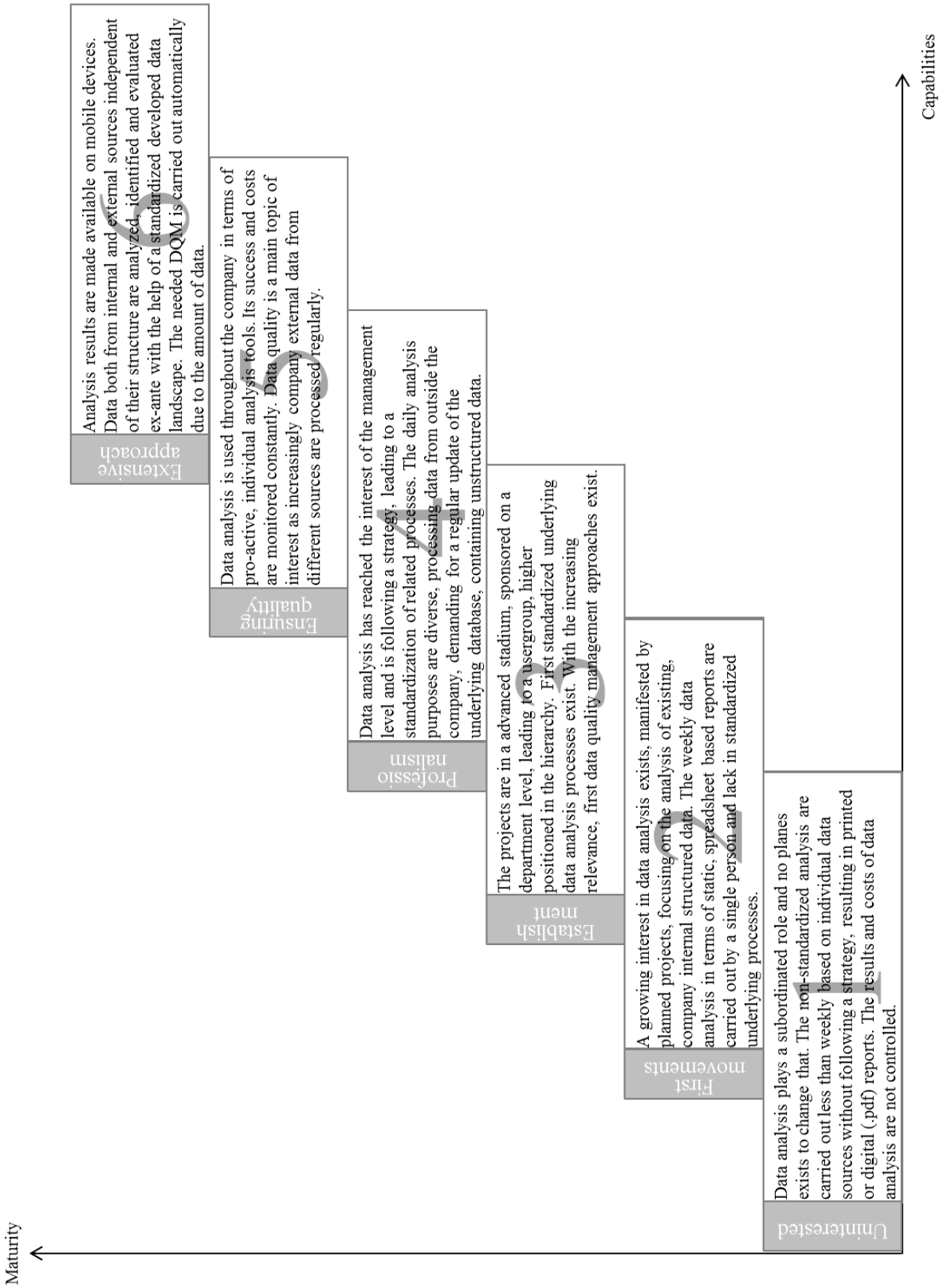


FIGURE 5.8: Description of the final Big Data Maturity Model

5.8 Step 8 - Model maintaining

The model maintaining process follows the iterative character of the design science research paradigm. The goal of a maintaining phase is to keep the *final model* and therefore the item list (*fitted model*) up to date. The actual execution of the maintaining process would lead to the beginning of the model construction process, re-starting with the *definition of the scope and the problem.* As this process has already been carried out in the course of the model construction (Chapter 5) and with regard to the limited time of the research project, the execution of the maintenance phase was not executed.[50] Nonetheless, with regard to the relevance of a regular updating of the model due to the dynamic environment in the context of Big Data, its potential execution is explained. Taking the practical relevance of the model in combination with its origin in research into account and incorporating the defined stakeholders (construction step 1), both the

- maintaining of the model as an artefact in a scientific context as well as
- the maintaining of the model as an evaluation instrument for (company-internal) consultants and other practitioners in a practical environment[51]

are explained along the aspects of *Scope* and *Topics and Items*. Those two areas represent the main parts, which are affected during a model maintaining, as they are influenced the most by the dynamics of the underlying subject in focus.

Scope

With regard to the novelty of the Big Data concept and the low number of published use cases and best practices, the model developed in this thesis is a general model, detached from industry or application specific aspects. With the expected increasing number of publications on maturity in the field of Big Data, the scope in a future maintaining process within the *research context* can be steered towards remaining research gaps, e.g. specific industries, departments, company sizes, etc.

Potential changes in the scope can hold the need for a re-consideration of the describing Dimensions. Those aspects will be further elaborated in the outlook in Chapter 6.

[50] Therefore, no box with goal, method, and result exists.

[51] The focus on the consultant in a practical environment results from the high popularity of maturity models as tools in the consultancy business as an initial part of projects to evaluate the status quo. Nonetheless, the application of maturity models is of course not limited to consultants and can be used company-internally as well.

In a *practical context*, the scope is influenced by the target group with regard to the role of a maturity model in an industry context, e.g. as a potential consultancies marketing tool. In this context, a broader focus can be more suitable for consultancies and allowing regular re-usability, alternatively focusing on a specific field of application, e.g. marketing.

Topics and Items

One element of maturity is its relative character. With regard to the dynamics in the field of Big Data, measurements that belong to a higher maturity level in the final model will decrease in maturity with time due to technological and organizational developments and might become standard requirements. Examples could be the increasing share of companies defining and communicating a Big Data strategy, the increasing analysis of unstructured data and the processing of company-external data. Therefore, the item set has to be fitted during each maintaining process.

In a *research context*, based on the experience from the construction process described in the thesis at hand, the member of the focus group are suggested to be exchanged - but only partly - with each maintaining process. The goal is to gather neutral input from experts that have not yet been in contact with the maturity model. At the same time, members that have already participated in previous construction processes contribute by describing experienced changes in the practical environment since the last model construction. Additionally, with an increasing number of publications, the identification of topics and measurements can be increasing carried out with the help of relevant literature.

In contrast, in a *practical context*, the focus is more on the update on the topics and measurements, instead on the underlying methodology for the construction of the model itself. This step benefits from an industry expert's practical experience and knowledge. The set of measurements can be fitted based on insights from current projects amongst different companies, independent of existing maturity models or comprehensive underlying literature research.

From the authors' point of view, the update frequency for the maintaining both in a research and practical context, is steered primary by the subject in focus and limited by the setting of the model. Novel topics in a dynamic environment, e.g. Big Data, hold the need for a higher update frequency, compared with established topics in order to

keep the model up to date and relevant. Nonetheless, information about temporal gaps cannot be found in closely related publications dealing with maturity models in the field of Business Intelligence or management control systems as well.

The practical oriented *business intelligence Maturity Audit* [Dittmar et al., 2013] is published in a two-year rhythm by the consulting company Sopra Steria/Steria Mummert. This period appears from the author's point of view as being too long due to the dynamics in the field of Big Data, considering the increasing number of publications, applications, and speed of development. Yet, with the higher frequency requirements as stated above, a yearly update of the model is recommended as an approximate guidance.

5.9 Main chapter results

In the past chapter, the developed construction model, described in Chapter 4 has been deployed. The dimensions *Data* and *Organization* have been identified as a basis for the model application. A bottom-up approach has then been applied for the model population. As a first step, a questionnaire has been developed, containing topics and related measurements with a focus on analysis process standardization, analysis strategies etc. that targets different capabilities. It has been discussed with industry experts, and pre-tested. The majority of the identified 17 topics belong to the organization dimension.

The resulting questionnaire has been answered by 71 companies from different industries and sizes.

In the next step, an algorithm from the IRT, a fitted Birnbaum approach, has been applied on the gathered data in order to calculate the item difficulty. The difficulty was used as a vehicle to determine the maturity of each item. The higher the difficulty value was, the more mature the associated capability was rated. Based on the difficulty, the items have been subsequently assigned to six maturity levels using a ward clustering.

The resulting *initial model* has been evaluated in a two-step approach. In a first step, it has been discussed with industry experts and the items have been reassigned accordingly. The resulting *fitted model* has been deployed at several companies out of the industry experts environment and the resulting model-based maturity assessment has been compared with the maturity evaluation of the same company by the respective industry expert. Based on interviews with the focus group members, reasons for

potential differences between the model based maturity evaluation and the industry expert evaluation have been identified (figure 5.7). Finally, the resulting model has been documented and the maintaining process has been described.

In order to give a structured overview about the insights of the chapter with relevance for the goals of this research, the main findings of the model population and evaluation will be described more thoroughly. The elaborations on *model population* focus on the findings from construction step 5, with an emphasize on the identified topics and the applied methodology. *Model evaluation* (section 5.6.1 and 5.6.2) focusses both on the findings from the actual evaluation process, targeting reasons for the needed model fittings as well as an assessment of the developed evaluation approach.

Model population

Although Big Data has become a popular topic in the overall discussion, both in science and practice, the initial identification of items that are beyond BI and specific for Big Data - based on the literature - has proven challenging, as only a few capability-relevant publications exist. The discussion of the resulting questionnaire with the focus group accordingly led to several further topics and measurements, such as the process for the identification of further data sources. The identified topics can be distinguished into i) completely new topics and items (e.g. the process for the identification of new data sources), which have been identified, and ii) novel, more advanced items for topics, that have already been used in existing maturity models (e.g. the result processing).

The contribution to research of the population step results from the identification of several topics, described in the section of the questionnaire development, which are characteristic for Big Data and have not been considered by existing models from closely related fields as well as the identification of new measurements for topics already existing in maturity models from nearby fields.

Model evaluation

The model evaluation showed that a quantitative approach can be applied for the assignment of items on different maturity levels in the developed model construction context, but hold the need for a further fitting, as the initial results in the thesis at hand did not represent in a first step an already suitable model. Every item had to be reassigned by

an average of 0.77 levels.

However, although all items had to be reassigned, with regard to the maximum possible reassignment of five (from level one to level six and vice versa), the application of the quantitative approach leads to suitable results. It could be shown that the extent of reassignment is not linked with the novelty of the topic in focus. Both measurements related to established topics (e.g. *Combination of data sources* known from the field of BI), and novel topics (e.g. *Big Data strategy*) underwent similar reassignments. Similarly, within the group of topics with a low extend of reassignment, both novel as well as established topics could have been found.

The model could be successfully deployed in an industry context as the second part of the evaluation. A high agreement between the model based evaluation and the experts evaluation of the companies' maturity could have been found. Following these findings, it can be concluded that the model represents the practical understanding of maturity adequately.

With this summary of the Chapter 5, the construction process is completed. The resulting Big Data maturity model can be found in figure 5.8. The list of according measurements per maturity level can be found in figure 5.5.

In the final chapter, a conclusion of the thesis and the research-related limitations is given, followed by the concluding outlook for further potential research in the field of Big Data maturity models.

Chapter 6

Final

In the beginning of this research, the guiding research question - searching for possibilities to improve the analysis of huge data amounts from different sources and with heterogeneous structures can be improved - has been described. From the general introduction to specific aspects for the relevant questions, the chapters 2 to 5 have been used to fulfil the research goal, the development of an industry-independent maturity model for the field of Big Data.
In order to recapitulate the work, the contribution of each chapter is described and the results are compared with the set goals in the beginning of the thesis. This summary is followed by the discussion of potential limitations, arising from the concept of maturity itself as well as the method-based limitations. The final part of this chapter gives an outlook of the potential further development of maturity models and Big Data in the research context and describes potential future research fields.

6.1 Summary

The beginning of Chapter 2 offers a first characterization of Big Data. Based on a mixed qualitative and quantitative approach, using recent literature on Big Data, *Data, IT infrastructure, Method* and *Application* have been identified as describing dimensions.
The assignment of recent publications on Big Data to the phases of a generalized data analysis process - aiming at the identification of current emphasizes and white spots in research on Big Data - revealed an under-represented consideration of the steps *data*

selection, result visualization, and *result interpretation/action* in the current Big Data research. The emphasis is on research associated with data pre-processing and analysis, driven by infrastructural-related topics.

Besides the description of dimensions and relevant topics, a comprehensive overview has been given by presenting a critical perspective on Big Data, targeting political, legal, ethical, and scientific issues.

In the following Chapter 3 the concept of maturity models and the contextual brackets of Design Science Research have been presented as a starting point for the subsequent maturity model development.

By comparing different maturity construction models, key elements of the model construction have been identified, namely *Scope, Design model, Develop instrument* and *Implement and Exploit*. The analysis of different maturity models in the data analysis context from fields such as Business Intelligence revealed that no existing model is solely focusing on Big Data. Furthermore, for those models that partly address Big Data relevant topics, a lack of theoretical foundation in terms of i) no underlying theoretical construction model and ii) a lack of evaluation and validation approaches could be identified.

Based on the described characteristics of maturity models and the related construction models (Chapter 3), Chapter 4 contains the development of the construction model used in this thesis.

In doing so, the construction approach by Becker et al. [2009] has been selected as a basis and enriched by several aspects from the construction model by De Bruin et al. [2005]. As both models do not consider sufficiently i) the specialty of the object in focus (novel, yet undefined topic with a need for clarification) and ii) the emphasis of the model evaluation (ensuring that the model represents the practical understanding of maturity in the practical context) - described in Chapter 1 and 4, the developed construction model that is applied in this research, emphasizes the description of the maturity object in focus, Big Data, and the evaluation of the developed artefact, the maturity model. A two-step evaluation approach has been developed, evaluating both the construction model itself (evaluation against the identified research gap) as well as the resulting maturity model (evaluation against the real world).

The governing aspect for the construction approach has been the correspondence with the principles of design science research by Hevner et al. [2004] and the principles of

correct modelling [Becker et al., 1995], resulting in a sound theoretical foundation of the resulting construction model.

Chapter 5 contains the deployment of the developed construction model. Out of the four dimensions, identified during the literature review in Chapter 2, the dimensions *Data* has been identified and complemented with the dimension *Organization* as a basis for the model development. As a bottom-up approach has been used, in a first step, a questionnaire, containing measurements that target different capabilities of Big Data relevant topics, has been developed, discussed with industry experts, and pre-tested.

One finding of the developed questionnaire has been, that the identified measurements partly represent an enhancement of existing maturity models, e.g. from the field of BI. Both new topics e.g. the process for the identification and evaluation of new data sources, as well as new items for existing topics, e.g. the result processing have been identified as relevant in the field of Big Data.

The resulting questionnaire has been answered by 71 companies from different industries and sizes. In the next step, for each item the related item difficulty has been calculated, using a statistical approach from the field of the IRT. This difficulty has been used as a vehicle to determine the maturity of each item, assuming that the higher the calculated difficulty, the more mature the associated capability is expected to be. Based on the difficulty, the items have been subsequently assigned to six maturity levels using a ward clustering.

The resulting *initial model* has been evaluated in a two-step approach. In a first step, it has been discussed regarding the correct assignment of the measurements to levels with the members of the focus group and the measurements have been reassigned accordingly. As the second part of the evaluation, the resulting *fitted model* has been deployed at eleven companies and the resulting model-based maturity assessment has been compared with the maturity evaluation of an industry expert. Based on interviews with the focus group members, reasons for potential differences between the model based maturity evaluation and the industry expert evaluation have been identified, incorporated and translated into the *final model*.

Following the description of the final model, the maintaining process has been presented. With regard to the different needs, the execution of the maintaining has been described, both in an industry as well as research context.

After describing the course of research, the fulfilment of the overarching goal, the

- Development of a maturity model for the field of Big Data

and the subordinated goals

- Using a quantitative bottom-up approach for the model population and the
- Development of a model evaluation process,

defined in Chapter 1, are discussed regarding their fulfilment.

1. Development of a maturity model
An indicator for the success of this research project is the fulfilment of the research goals. By developing the model as described in Chapter 5, the overarching goal of this thesis as stated in Chapter 1, has been accomplished. Contrary to the criticism of most of the existing maturity models, the presented model has been developed on a sound theoretical foundation, based on a design science oriented construction model, applying a quantitative bottom-up population approach. By deploying the model at eleven companies it could be shown that the constructed model can be used to address the problems described in Chapter 1; the companies' uncertainty which capabilities should be developed, in order to improve the handling of Big Data. The evaluation of companies' capabilities can be taken as a starting point for an improvement of abilities in the field of Big Data.

2. Using a quantitative bottom-up approach for the model population
The next goal, the testing in how far a quantitative approach can be applied in a field, which contains both novel and established aspects, could also be met. It could be shown that one challenge, when applying a quantitative bottom-up approach for the model construction for a novel topic like Big Data, is the identification of measurements. Due to the absence of existing maturity models for Big Data and the low number of relevant publications, the input of a focus group is needed in order to gain a sufficient practical relevance.

3. Development of model evaluation process
The development and application of a suited maturity model evaluation that is supposed to i) contribute to the theoretical foundation and ii) prove the practical relevance and applicability of the model as the third research goal could be accomplished as well.

Concerning the first part of the model evaluation, it could be shown that a quantitative approach can be applied for the assignment of items on different maturity levels in the developed model construction context. At the same time, the discussion of the initial model with the focus group members - as the first step of the evaluation against the real world - yet revealed the need for a further fitting.
Every item had to be reassigned by the industry experts by an average of 0.77 levels. This low reassignment shows that although all items had to be reassigned, the application of the quantitative approach leads to suitable results. It could be shown that the extent of reassignment is not linked with the novelty of the topic in focus. Both items related to the established topic known from the field of BI, and novel topics underwent similar reassignments. Similarly, within the group of topics with a low/high extend of reassignment, both novel as well as established topics could been found.

The model could be successfully deployed in an industry context as the second part of the evaluation. In this context - the identification of companies, covering the whole range from maturity level one to maturity level six in order to achieve a full-range model evaluation - has proofed itself as demanding for a novel topic like Big Data. Nonetheless, a high agreement between the model based evaluation and the expert's evaluation of companies maturity could be found. This can be interpreted as an adequate representation of the practical understanding of maturity by the developed model.
The positive model evaluation gains in weight when looking at the different scopes of the two maturity evaluations. The model is focusing on the dimensions *Organisation* and *Data*, whereas the expert is evaluating the companies' capabilities along all Big Data dimensions. Despite these differences, the comparison shows a high agreement between the model and the experts' based assessment, indicating a broad coverage and relevance of the selected dimensions.

6.2 Limitations

The thesis at hand contains several limitations, which can be predominantly divided in those, that result from i) the concept of maturity models and those that result from ii) the applied methods in the course of this research.

6.2.1 Maturity concept based limitations

One major limitation concerns the relation between maturity and performance. A higher maturity, although perceived as a desirable state, does not necessarily lead to an improvement of a company's performance. An insufficient operationalization can override a high maturity. Consequently, a correlation between organizational performance and maturity cannot be drawn directly, although several topics of the developed model focus on the operationalization, e.g. the definition and standardization of analysis processes. Therefore, the relevance and explanatory power of the developed maturity model can be fostered by examining an improved performance on a higher maturity level. A longitudinal study, analyzing different companies on different maturity levels and comparing financial performance indicators in the course of time could help to close this gap.
The aspect of performance is of special interest for Big Data with regard to the volume of potential investments and the needed organizational changes. Due to the novelty of this topic, to the best of the authors' knowledge, no research in this field exists so far.
A second limitation results from the dynamic character of the topic Big Data. Due to the speed of development e.g. regarding the costs and the scope of analytics applications as well as the increasing professionalism in the utilization of large amounts of data with different degrees of structure, the model holds a high need for a regular update. With regard to the needed time for the model development in a research context, especially the needed time for data gathering and the evaluation process, it can be partly outdated when the construction process is completed. Therefore, the maintaining process plays a critical role in order to ensure a permanent model relevance.

Another potential limitation comes with the selected level of abstraction, being a core decision of each maturity model construction process. The maturity model holds a certain level of abstraction due to the generalization regarding industry, company size, and application. On the one hand, this limits the explanatory power, as the contained items are selected to be regained in the majority of the companies. With regard to the early stage of the topic Big Data, the trade-off has been made towards a general model that is intended to act as a starting point for further, more dimension specific maturity models in the Big Data context. With the increasing spread of Big Data related applications, it is assumed that with each maintain process, the granularity of the models measurements can be improved.

6.2.2 Method based limitations

Within the contextual brackets, the Design Science Research, both qualitative and quantitative methods have been applied: mainly i) text mining, ii) focus group, and iii) IRT. The applied methods will be evaluated regarding the resulting limitations.

Text Mining

One limitation resulting from the applied methodology can be found in Chapter 2 during the identification and validation of dimensions on Big Data. The limitation targets the influence of the breadth of the analyzed topic and the results of the topic models. It could be shown that the explanatory power of the approach decreased when the analyzed corpus contained only dimension-specific publications, e.g. solely technology-oriented. The results based on the broader corpus of Big Data publications, including all dimensions, led to better results. These differences in explanatory power are represented in the different values for the model precision. Therefore, the results from the topic model application on the specific dimensions have not been processed further.

Additionally, the corpus with 247 analyzed abstracts is relatively low compared with other publications using the topic model approach on established topics, owed to the novel character of Big Data. With the increasing number of publications regarding Big Data, the explanatory power of the topic model applications may increase. As the focus of the work at hand is on the phenomena of Big Data, further search terms have been left out intentionally.

Nonetheless, for future research, the focus on the search term *Big Data* carries the risk that publications, which are relevant for the topic but marked with another tag, e.g. the rising notion of *Advanced Analytics* are left out. Furthermore, relevant topics, e.g. organizational aspects might not appear in the analysis, as publications on that topic have not been published yet.

Focus Group

The input of the focus group depends amongst others on the members' academic backgrounds and practical experience. Therefore, the composition of the focus group has an influence on the overall results, as the focus group members have an influence on the resulting maturity model in terms of the contained topics and measurements and their assignment to different maturity levels. This limitation could partly be solved.

Although a random set of focus group members is hard to be achieved, by increasing the number of the members of the focus group with a heterogeneous background, regarding to the aspects mentioned above, a broad knowledge base could be reached. Yet, a higher number of new focus group members for each construction step could contribute to the broader knowledge base.

Item Response Theory

The potential limitations that go along with the application of the IRT have already been discussed in the model population step (Chapter 5). Besides that, the basic concept of using the item difficulty as a vehicle to determine the maturity of each measurement, contained in the questionnaire can be seen critical, as it leads to problems concerning the different scales. As it has been described in construction step 5 (Chapter 5), the continuous numbers per item (valuation 0-1) can be seen in conflict with the maturity concept, limited to absolute numbers, that range in the thesis at hand from 1 to 6. Consequently, the delta of the item difficulty between the item with the highest item difficulty value and the one with the lowest difficulty of one maturity level is not necessarily equal along all maturity levels. This limitation loses in relevance as, resulting from the re-organization of the items based on the discussion with the focus group, the final order of the items is not following the initially calculated difficulty values.

As the approach from the IRT has been applied on a data set resulting from answered questionnaires, the *Characteristics of the Respondents* pose an influence and potential limitation. The responses do not necessarily reflect the actual situation of a company, as the respondent may not have a sufficient overview, e.g. about existing projects. This accounts especially for larger companies. Consequently, the background and knowledge of the respondents have a major influence on the quality of the initial model. This potential bias is not specific for this thesis, it is rather a general problem of data gathering based on un-controlled answering of questionnaires and can be found in similar works as well [Lahrmann et al., 2011b]. By distributing the questionnaire in different fora, allowing an anonymous answering without a personal approach, the sample has been compiled as random as possible.

For future research, personal interviews, following a random sampling of companies and interviewees for the data gathering could be used in order to gain a better insight into

the company practices and allow to interpret the responses.[1]

6.3 Outlook and future research

The interest in maturity models in the field of Big Data as a tool for the evaluation of companies' capabilities is supposed to increase in the coming years. Based on the increasing utilization of data in different contexts throughout different industries, analytics will be increasingly perceived as a potential competitive advantage [McAfee and Brynjolfsson, 2012; Buhl et al., 2013]. The presented maturity model is the first published model for the evaluation of capabilities in the field of Big Data with a research background. It can be taken as a starting point for different directions of future research questions in the Big Data maturity context, that will be discussed subsequently. Starting from the described limitation - the level of abstraction - an overview about the potential directions is described.

The scope of the maturity model has been defined in the beginning of the model construction. With regard to the novelty of the topic Big Data, a broad scope has been selected to develop a model that allows a holistic view on capabilities in the Big Data context. Coming from this holistic model, potential further research directions can be structured along two perspectives - *contextual aspects* and *Big Data Dimensions*.

The contextual aspects target *size, industry*, and *field of application*. The different focuses can be integrated in the model during the maintaining process in phase one, the definition of the model scope. These can be seen as the framing conditions, which define the context, for which the model is developed.
The Big Data Dimensions - *Data, Method, IT infrastructure* and *Organisation* - enable a further scoping, allowing a drill down into dimension-specific capabilities.
The combination of those two perspectives, contextual aspects and the Big Data dimensions, allow a structured demonstration of the fields for potential further research.[2]

[1] Although the interview character can lead in turn to a bias of the respondents' behaviour as well [Newman et al., 2002].
[2] As the dimensions have already been explained in Chapter 2, the following description is narrowed to the contextual aspects.

Size

A potential direction could be the development of maturity models, that focus on companies with a specific size, e.g. small-medium enterprises, followed by the comparison of a size-specific and a size-independent maturity model. Although the increasing digitalization affects companies of all sizes, it could be more challenging for smaller companies, due to budget and knowledge constraints, to build up relevant capabilities in the field of Big Data. A size-specific maturity model would allow for the comparison between the capabilities of companies with similar sizes, in turn, drawing a more specific picture.

Industry

As described in Chapter 5, the industry of a company and the history based experiences may influence its maturity. Companies belonging to the finance sector could in general possess a higher maturity as they have constantly been confronted with and were utilizing larger amounts of data compared with manufacturing businesses. Therefore, the development of an industry specific maturity model and the comparison with a cross-industrial model can reveal industry relevant capabilities and characteristics.

Field of Application

For the developed model, the goal was to identify topics and items that can be used for a company's maturity evaluation independent of the underlying application. This resulted in a majority of topics belonging to the organization dimension. With an increasing integration of analytics throughout companies, the number of applications is supposed to increase as well. Therefore, the development of application-specific models, e.g. for the field of marketing or production, allows the identification of application-relevant topics. This focus may allow the integration of further dimensions, e.g. the infrastructure into the model.

Another aspect targets a cross-dimensional facet. As described in the section model evaluation (section 5.6), the maturity of one aspect may influence the overall maturity of a company. Therefore, the uni- and bilateral dependencies of different topics hold potential for further research. Investigations on this field can allow carving out aspects, which are dominating the overall maturity evaluation of a company.

For each of the three described aspects it can be decided, if a general perspective (without a drill-down) or a specific perspective is pursued. The combination of different aspects

allows numerous future research fields, contributing to the relevant topic of maturity models in the field of Big Data.

Summing the future of maturity models up, as described, the increasing relevance of Big Data for companies is going to have a positive influence on the topic of maturity models in the analytics context. Both from a research perspective with regard to the increasing number of potential dimension-specific maturity model research topics, as well as from a practical respective companies' perspective, focusing on the need for tools to evaluate the capabilities in the field of analytics, the already high need for Big Data maturity models will increase. Hevner et al. [2004] sums this understanding up by stating

> *"The goal of behavioral-science research is truth.*
> *The goal of design-science research is utility."*

which has been used as a guidance for the development of the maturity model in this thesis.

Appendix A

Questionnaire used for the data gathering in construction step 5 - Model population

Fragebogen Forschungsprojekt Big Data Reifegradmodell

Sehr geehrte Damen und Herren,

vielen Dank für Ihre Unterstützung unseres Forschungsvorhabens.
Ziel der Befragung ist die

Entwicklung und empirische Validierung eines Reifegradmodells
zur Bewertung der Fähigkeiten von Unternehmen
im Umgang mit großen Datenmengen (Big Data)

Ein Reifegradmodell dient der Bewertung von unternehmensseitigen Fähigkeiten sowie dem Aufzeigen von Optimierungspotentialen.
Das auf Basis der Befragungsergebnisse entwickelte Modell wird im Anschluss praxis-orientiert aufbereitet und den teilnehmenden Unternehmen zur Verfügung gestellt, so dass es für Benchmarking-Zwecke verwendet werden kann. Bei der nachfolgenden Reifegradmodellanwendung unterstützen wir Sie gerne.

Bitte füllen Sie den Fragebogen so aus, dass Sie die für Sie zutreffende(n) Antwort(en)/Ausprägungsmöglichkeit(en) mit einem **x** markieren (Dauer: ca. 10 Minuten). Dabei wird Big Data mit einem Fokus auf Datenanalyse betrachtet, weshalb die Begriffe im Folgenden synonym verwendet werden. Unter Datenanalyse verstehen wir sämtliche Tätigkeiten, welche im Zusammenhang mit dem Erstellen von Reports/Berichten und sonstigen Datenauswertungen stehen.
Die von Ihnen gemachten Angaben werden selbstverständlich anonymisiert verarbeitet. Aufgrund der Ergebnisdarstellung in aggregierter Form sind Rückschlüsse auf einzelne Unternehmen nicht möglich.

Bitte speichern Sie den ausgefüllten Fragebogen und senden ihn per Antwortfunktion zurück oder an die folgende E-Mail Adresse:

thomas.hansmann@leuphana.de

Alternativ können Sie den Fragebogen auch faxen.
Vielen Dank!

Bei Rückfragen stehen wir Ihnen selbstverständlich jederzeit zur Verfügung. Falls gewünscht, kann der Fragebogen auch gemeinsam mit Ihnen im Rahmen eines Telefonats oder persönlichen Gesprächs durchgegangen und ausgefüllt werden.

Prof. Dr. Peter Niemeyer und Thomas Hansmann
Institut für elektronische Geschäftsprozesse
Leuphana Universität Lüneburg
Scharnhorststr. 1
21335 Lüneburg
Fon: 04131-677-1664
Fax: 04131-677-1749

Fragebogen Forschungsprojekt Big Data Reifegradmodell

Bitte nennen Sie die Branche, in der Ihr Unternehmen tätig ist:							
☐ Produzierendes Gewerbe	☐ Informations- und Kommunikationsbranche	☐ Handel	☐ Dienstleistungen	☐ Banken & Versicherungen	☐ Energie	☐ Gesundheitswesen	☐ Sonstiges

Bitte nennen Sie die Mitarbeiteranzahl Ihres Unternehmens:						
☐ 0-100	☐ 100-500	☐ 500-1000	☐ 1000-5000	☐ 5000-10.000	☐ 10.000-50.000	☐ > 50.000

Bitte nennen Sie Ihre Position im Unternehmen:			
☐ Geschäftsleitung/ Vorstand	☐ Bereichsleitung/ Abteilungsleitung	☐ Teamleitung/ Projektleitung	☐ Mitarbeiter(in)

Block 1 – Management und Organisation

Datenanalyse-Strategie Bitte nennen Sie, sofern vorhanden, das Level, auf dem die Strategie für Datenanalyse definiert/vorhanden und kommuniziert ist				
☐ Keine Strategie	☐ Bereichsweit (z. B. Produktmarketing)	☐ Bereichsübergreifend (Gesamtes Marketing)	☐ Unternehmenssparte	☐ Unternehmensweit

Datenanalyse-Projekt/Initiative Bitte beschreiben Sie, in wie weit Sie derzeit ein eigenständiges Projekt im Bereich Analytik/Big Data verfolgenden					
☐ Kein Projekt, auch nicht geplant	☐ Kein Projekt, jedoch fest in den nächsten 12 Monaten geplant	☐ Derzeit in der Projektplanungsphase	☐ Analysephase abgeschlossen	☐ Projekt wird derzeit durchgeführt	☐ Projekt abgeschlossen

Datenanalyse-Projektsponsoren Für den Fall, dass Sie derzeit ein Analytik-Projekt haben, nennen Sie bitte den Projektsponsor / unternehmensinternen Auftraggeber (ansonsten weiter mit der nächsten Frage)				
☐ Mehrere dezentrale Sponsoren aus der IT	☐ Zentraler Sponsor aus der IT	☐ Zentraler Sponsor aus dem Management Bereich	☐ Analytik/BI Abteilung Sponsor	☐ Geschäftsleitung Sponsor

Fragebogen Forschungsprojekt Big Data Reifegradmodell

Kostenkontrolle der Datenanalyse Bitte markieren Sie, wie die derzeitigen Analyselösungen auf Wirtschaftlichkeit überprüft werden			
☐ Keine Kosten-Nutzen-Rechnung	☐ Projektbasierte Kosten-Nutzen-Rechnung	☐ Verwendungsorientierte Berechnung	☐ Erfolgsorientierte Kalkulation (Gesamtsicht)

Derzeitige Umsetzung der Datenanalyse Bitte markieren Sie, auf Basis welcher Systeme und Anwendungen derzeit Analysen durchgeführt werden				
☐ Ad-hoc Analysen (z.B. spreadsheet-basiert mit MS Excel)	☐ Statische Reports (vorgefertigte Reports)	☐ Integration verschiedener Analyse-Frontends für Aufbereitung statischer Reports	☐ Nutzung Analysesoftware (Anwendung verschiedener Algorithmen auf Daten des Data Warehouse)	☐ Flexible, proaktive Analytiklösung (vereinfachte Anwendung verschiedener Methodiken auf Daten verschiedener Quellen etc.)

Häufigkeit der Analysen Bitte markieren Sie, wie häufig Analysen auf Ihrem Datenbestand durchgeführt werden			
☐ seltener als wöchentlich	☐ wöchentlich	☐ täglich	☐ laufend (Echtzeit)

Verwendung der Analysen Bitte markieren Sie, für welche Zwecke Ihr Unternehmen Analyseergebnisse verwendet			
☐ Monitoring von einzelnen Größen (z. B. Sensordaten, Abverkaufszahlen etc.) Zeitverlauf	☐ Klassifikation (Kundenpotentiale, Abwanderungsanalysen, etc.)	☐ Explorativ (Verwendung verschiedener Analysemethoden auf einen Datensatz)	☐ Prädiktiv (Zukunftsorientierte Aussagen, z.B. Strategieentwicklung

Bereitstellung von Analysen / Reports Bitte markieren Sie, auf welchem Wege derzeit Reports den Mitarbeitern bereit gestellt werden			
☐ Reports werden ausgedruckt / digital (z.B. in pdf./xls-Form) zur Verfügung gestellt	☐ Standardberichte werden auf einem abteilungsweiten Informationsportal zur Verfügung gestellt	☐ Standardberichte werden auf einem unternehmensweiten Informationsportal zur Verfügung gestellt	☐ Standardberichte werden auf mobilen Endgeräten bereit gestellt (Smartphone/Tablet)

Weiterverwendung der Analyseergebnisse Bitte markieren Sie, wie Ihr Unternehmen Analysergebnisse in der Folge weiter verwenden		
☐ Manuelle Weiterverwendung ohne Integration in definierte Entscheidungsprozesse	☐ Standardisierte, manuelle Weiterverwendung der Ergebnisse in definierten Entscheidungsprozessen	☐ Standardisierte, automatisierte Weiterverwendung

Fragebogen Forschungsprojekt Big Data Reifegradmodell

Nutzergruppe I Bitte markieren Sie, welche Mitarbeiter in Ihrem Unternehmen softwareseitig mit Analytik-Lösungen arbeiten (entwickeln von Methoden, erstellen von Analysen)				
☐ Einzelner Datenanalyst	☐ Key User/ Datenanalysebe-auftragter je Abteilung	☐ Eigenständige Datenanalyse-Abteilung	☐ Mittleres Management	☐ Alle Unternehmensbereiche

Nutzergruppe II Bitte markieren Sie, welche Mitarbeiter in Ihrem Unternehmen mit den Ergebnissen der Analytik-Lösungen arbeiten				
☐ Einzelner Datenanalyst	☐ Key User/ Datenanalysebe-auftragter je Abteilung	☐ Eigenständige Datenanalyse-Abteilung	☐ Mittleres Management	☐ Alle Unternehmensbereiche

Erfolgskontrolle der Datenanalyse Bitte markieren Sie, in welchem Umfang Ihr Unternehmen derzeit den Erfolg der Datenanalyseanwendungen kontrolliert				
☐ Kein Erfolgskontrolle	☐ Sporadische Nutzermeetings	☐ Regelmäßige Nutzermeetings	☐ Standardisierte, unregelmäßige Erfolgskontrolle	☐ Standardisierte, regelmäßige Erfolgskontrolle

Prozessmodell der Datenanalyse Bitte markieren Sie, wie weit die derzeitige Analyselösung auf standardisierten, dokumentierten Prozessen aufsetzt				
☐ Informelle Prozesse / keine Standardisierung	☐ Erste etablierte Prozesse auf Basis von Routineabläufen	☐ Standardisierte, dokumentierte Prozesse auf Abteilungsebene	☐ Standardisierte, dokumentierte Prozesse & Kontrollen auf Abteilungsebene	☐ Verpflichtende, unternehmensweite Prozesse und Kontrolle

Block 2 - Datenmanagement

Vorgehensmodell Datensammlung Bitte markieren Sie, wie Ihr Unternehmen bei der Identifikation von Daten für Analyseanwendungen vorgeht					
☐ Fokussierung auf bestehende Daten aus Datawarehouse	☐ Unregelmäßiges Screening nach weiteren, unternehmensinter n verfügbaren Daten	☐ Unregelmäßiges Screening nach weiteren, unternehmensin- und extern verfügbaren Daten	☐ Regelmäßiges Screening nach weiteren, unternehmensin- und extern verfügbaren Daten	☐ Standardisiertes Vorgehensmodell für die regelmäßige Erstellung einer Übersichtskarte über unternehmensin- und extern verfügbare Daten	☐ Erstellung Übersichtskarte sowie zusätzliche Vorab-Evaluierung von Daten bzgl. ihrer potentiellen Nutzenstiftung (für konkrete Anwendungsfälle)

Fragebogen Forschungsprojekt Big Data Reifegradmodell

Quellen der verwendeten Daten Bitte markieren Sie, aus welchen Quellen in Ihrem Unternehmen Daten für weiterführende Analysen geschöpft werden			
☐ Interne, strukturierte Daten (z.B. Absatzzahlen)	☐ Interne, strukturierte und unstrukturierte Daten (z. B. Absatzzahlen und Außendienstmitarbeiterberichte)	☐ Interne, strukturierte und unstrukturierte Daten sowie externe, strukturierte Daten (z. B. quantitative Marktforschungsdaten)	☐ Interne, strukturierte und unstrukturierte Daten sowie externe, strukturierte und unstrukturierte Daten (z. B. Soziale Netzwerke, qualitative Marktforschung)

Zusammenführung verschiedener Datenquellen Bitte markieren Sie, in wie weit derzeit die Daten aus verschiedenen Quellen zusammengeführt werden			
☐ Keine Zusammenführung	☐ Sporadische, manuelle Zusammenführung, wird nur bei Bedarf für spezifische Analysefragestellung durchgeführt	☐ Manuelle Zusammenführung, regelmäßige Integration neuer Datenquellen, unabhängig individueller Analysefragestellungen	☐ Automatisierung, standardisierte Datenzusammenführung

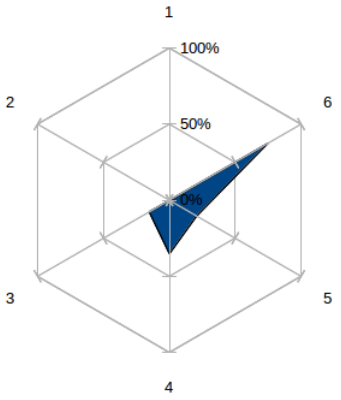
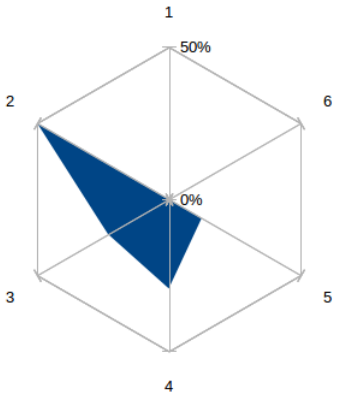
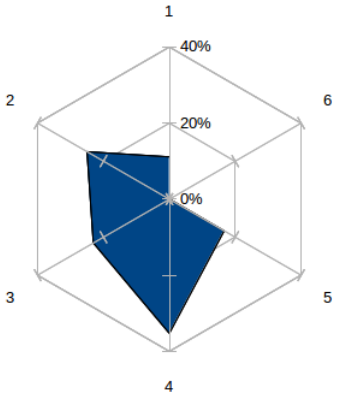
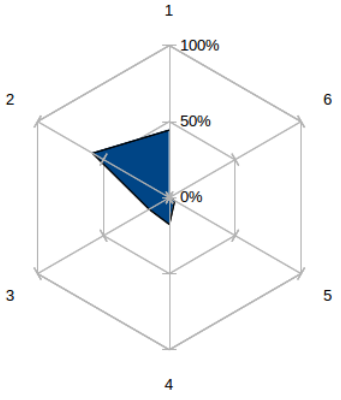
Durchführung Datenqualitätsmanagement (DQM) Bitte markieren Sie, wie das DQM im Vorfeld der Datenanalyse durchgeführt wird				
☐ Manuelles, ad-hoc DQM	☐ Definierte DQM Rollen	☐ Definierte DQM Prozesse	☐ Automatisiertes DQM	☐ Zuständiges DQM-Team

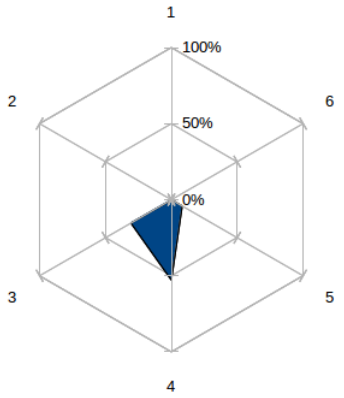
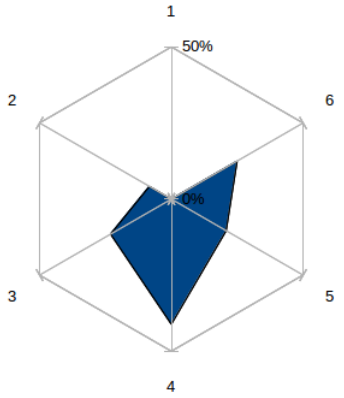
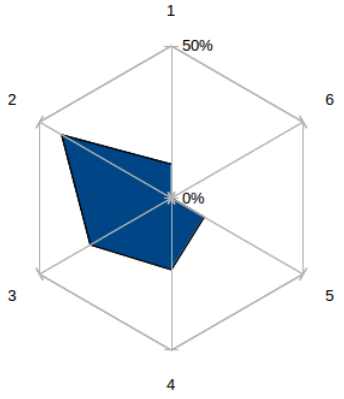
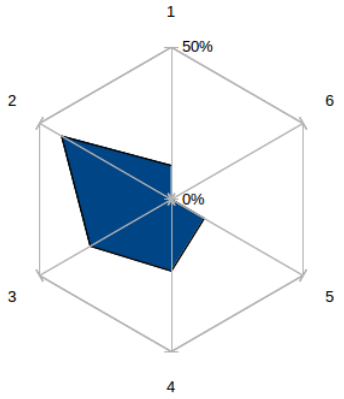
Appendix B

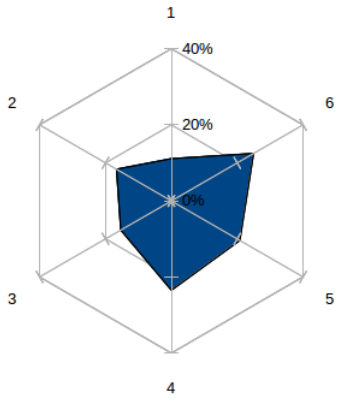
Step 6.2 Evaluation based on the deployment of the fitted Model - Evaluation of additional companies

In section 5.6.2, the second evaluation step for the evaluation against the real world has been explained. Besides the two exemplary described ones, nine more companies have been evaluated. The aspects driving the experts assessment can be found subsequently.

TABLE B.1: Overview evaluation results step 6.2

No.	Industry	Employees	Expert evaluation	Degree of fulfilment
1	Service Industry	500 - 1000	4-5	
2	Retail	1000-5000	2-3	
3	Service Industry	1-100	3	
4	Banking / Insurance	1-100	1-2	

No.	Industry	Employees	Expert evaluation	Degree of fulfilment
6	Banking/ Insurance	100-500	3-4	
7	Telecom-munication	>50,000	4	
9	Retail	1000-5000	3	
10	Banking / Insurance	5000-10000	2-3	

No.	Industry	Employees	Expert evaluation	Degree of fulfilment
11	Banking / Insurance	1000-5000	4	

Company 1

Company one is a service company, offering the design and operation of bonus programs to its customers, e.g. airlines or retail companies. Following the expert's opinion, the company is one of the most mature companies he is aware of. Data analysis is centralized within the company due to the company's organic growth, keeping the data analysis centralized for national affiliates as well. Another indicator for a high maturity is the implemented automated data quality management.

The company is open-minded towards innovations in the field of data analysis, e.g. the availability of analysis results on mobile devices has been enforced by the management. Comprehensive investments have been undertaken in recent years in data analysis applications and infrastructure, primarily driven by demands by the marketing and sales department.

The company is continuously screening the company-internal data universe and its environment for further available data sources. Processes in the field of data analysis are standardized on a company-wide level and controlled. Altogether, data and their analysis are one of the company's primary resources leading to a maturity evaluation of four to five. The reason for the high but not highest evaluation is that these amounts of data stored are not utilized for analysis purposes completely. Only parts of the available data sources are further processed, leaving potential further insights aside.

The model based evaluation has its highest fulfilment on level six as three out of four

items on level six are fulfilled, resulting in a blurred representation of the actual distribution of maturity.

Company 2

Company two is an example of a corporation whose management is aware of the potential that data analysis holds. This awareness resulted in the definition of a cross-divisional analysis strategy and the use of analysis applications throughout the company. Several initiatives have been started but only a few applications are up and running, primarily in the field of analytical online marketing, e.g. customer retargeting. The analysis are primarily static, pre-defined reports. The subsequent result processing is not standardized and carried out manually. External data are not processed, the focus is on internal, structured data already stored in the data warehouse. One problem area is a lack of analytical relevant knowledge. The pressure to build know-how in the field of analysis it not yet sufficient as analysis are carried out only on a weekly basis. Following the experts' opinion, the described company will reach a higher state of maturity in the next few years under the condition that the management increases the pressure in terms of definition of further analysis relevant projects and at the same time, increases the investment in the development of data analysis knowledge.

The model-based evaluation for company two is closer to the experts evaluation compared with company number one. Level two holds the highest degree of fulfilment, followed by four, three and five. This reveals one weakness of the sole graphical representation of maturity. Due to the unequal distribution of topics and related items/measurements on the maturity levels - the majority is on levels four and five - in most model applications these levels show a certain degree of fulfilment. In contrast, if the maturity level is calculated, resulting in a single number, these outlier are reduced, as the average over the individual degrees of fulfilment is calculated.

Company 3

The third company offers IP TV solutions, similar to apple TV. Comparable to company number two, several initiatives have already been started, driven by the management level, which has also fostered the company-wide definition of a analysis strategy. The comprehensive amounts of data are only partly analyzed using an individual data analysis software and the results are processed manually, hampering the operationalization of analysis potentials. The TV log files, for example, are analyzed and used for time-delayed

next best offer applications but not for real time targeting personalized advertisement purposes. Another aspect determining the middle maturity is the lack of standardization, both for processes regarding the data analysis itself as well as the success control, and the centering of the report definition on one person, although the whole company is working with the analysis application. Furthermore, the lack of structures and processes currently hinders the company from reaching a higher level of maturity from the experts point of view.

As mentioned before, the explanatory power of items can differ, depending on the company in focus. In this case, the industry expert emphasized that the company-wide analysis strategy, as well as the project sponsoring by the management, is rather owed to the relatively small size of the company than the highly perceived relevance of the aspect of Big Data.

The model application results in the highest degree of fulfilment for level four, followed by two and three, speaking for, comparable to the experts evaluation, an average maturity of three.

Company 4

Company four belongs to the field of Banking/Insurance and can be seen as an example of a company that is at a very early stage of analytics. Although the management is project sponsor, the focus is on internal structured data, which are already stored in the data warehouse. No screening for further data sources exists. The data preprocessing contains an individual, manual gathering of data from different internal sources, which are processed manually in terms of combination, data quality management, and structuring. In the industry expert's opinion, one reason for the low maturity that simultaneously also hinders a short-term improvement of the maturity, is the lack of defined responsibilities for data analysis. Although the company has an individual analysis software (QlikView), the analysis are mostly ad-hoc and spreadsheet-based and carried out less often than on a weekly basis. Summarizing, the company has multifaceted fields of application with regard to the potential available data sources as well as the numerous nearby analysis tasks. Those are not pursued due to i) the lack of standardization and automation as well as ii) the lack of investments in knowledge and standardization, which are at this stage necessary to move to a next maturity level.

Again, the model based assessment os close to the experts one, although the aspect of responsibility is not covered by the developed maturity model.

Company 6

Company six is an example of the relevance and influence of the workforce and their knowledge on the overall success of a Big Data application. The organization is staffed with a comprehensive IT infrastructure for analysis purposes, containing both sufficient storage capacities as well as analysis front-ends. The further use and development of analysis applications is impeded by the concentration of the relevant knowledge on two individuals within the whole organization, which in turn only results in limited capabilities in the analysis field. The company's management is aware of this situation but has not yet reacted. The two employees responsible for data analysis are lacking in an understanding of the needs of the departments regarding reports and analysis as well as the integration of results in the business processes. At the same time, the existing knowledge is distributed unequally between the departments. A majority of the employees (users of the analysis applications) have only very limited knowledge regarding the available applications which also accounts for analysis tasks set up in Microsoft Excel. Consequently, only a small share of the stored data is processed.

During the discussion with the industry expert, evaluating company six, one potential enhancement of the item set has been identified. The developed maturity model does not measure and evaluates the centralization of analysis relevant knowledge within the department or company. A company that concentrates its knowledge on a small group of individuals faces a short-term loss of knowledge in the event that these employees leave the company. Alternative items in the model that partly allow statements about these aspects are those regarding the process standardization and documentation. Although this does not compensate the loss of knowledge carriers, standardized and documented processes can help to absorb such loss [Szulanski, 1996]. Nonetheless, the relevance of knowledge as a critical resource in the data analysis context will increase due to the lack of sufficient qualified workforce [Manyika et al., 2011].[1] Again, the model-based results with the highest fulfilment on level 3 and 4 are coherent with the assessment by the expert.

Company 7

[1] An emphasized consideration of knowledge and its diffusion could be an advancement of further models, as it will be explained in more detail in the outlook (Chapter 6). It could not be further considered in the current model due to i) the limited time of the thesis and ii) as it has not been mentioned by the focus group as an aspect, which absence reduces the value of the model significantly.

Company seven is another example of a higher maturity as several analysis-relevant processes are already standardized and data analysis has already become a strategic topic. The analysis department has defined numerous analysis and is continuously working on further analysis tasks, although such tasks are not necessarily always based on clearly defined targets.

The company's development towards a stronger Big Data orientation is slowed down by i) a lack of knowledge in the central IT and ii) a lack of managerial decision-making and operationalization of analytics on department level, comparable to company number five.

Regarding i) the IT is unable to cope with the requirements, set by the analysis department. Therefore, the employees in the IT department try to postpone the requests in order to avoid an obvious excessive demand. As a consequence, the analysis department has started building their own IT department to reduce the dependence on the centralized IT department. The second obstacle is the lack of decision-making by the department management regarding matters such as investments. This affects the processing and operationalization of analysis results as they are not yet integrated in existing business processes. In turn, after the initial analysis is carried out, no additional insights and values are generated.

The two issues lead in addition to a third problem. Both the sporadic implementation of employees' ideas by the IT department and the lack of operationalization result in the employees' lack of motivation. Altogether, this negatively affects the overall department performance.

The highest degree of fulfilment can be found for level four, which goes along with the experts' assessment.

Company 9

In contrast to company number eight, described in construction step 6.2, company number nine does not have a knowledge problem. The employees in the BI department are well experienced and have extensive knowledge in the field of data analysis, focusing on company internal, structured data. The relevance and role of data analysis is formulated in a company-wide strategy. Despite the awareness of analysis on the management level, the company lacks of sufficient standardization, both regarding the analysis processes itself as well as the success control, the DQM and the identification for new data sources.

The major topic slowing down the company's development, leading to a middle maturity level despite the highly capable workforce, is the organizational structure and the management approach. The management of the BI department steers the employees project work based on a so called championship challenge. [2]

The analysts do not share their knowledge and the BI department has a higher fluctuation compared with other departments of the company. The model application results in an evaluation with the highest fulfilment for level two and three, close to the expert's evaluation. The industry expert mentioned one situation, in which the only the score code of a model exists, the underlying process code is not available or understandable any more (several thousand lines of code without comments) and therefore the results are not **observable** as well. Another example is a larger excel-based analysis tool, which starts a macro, which in turn starts a macro and so on. Again, the employee in charge has left the company, therefore the individual model steps are not observable.

Company 10

Company ten is another example for a company which has started its development towards a higher data orientation but lacks in a holistic approach and coordination of the different initiatives. As a consequence, in a rather explorative-driven approach different isolated applications are implemented, leading to data silos. This fragmented development is partly owed by the geographical spread of the different company parts, leading to a need for intensive coordination amongst the individual countries.

Besides the geographical aspect, the second critical aspect is the responsibility for the analysis relevant projects in the IT department, as no central BI department exists. The IT department is responsible for analysis related tasks e.g. adding further data sources or changing reporting structures. As a consequence, a significant share of the IT department's capacity takes the handling of change requests from application users.

A third aspect results from the lack of standardization. Currently, the company lacks in a data governance, leading to a suboptimal meta data management. It exists no common

[2] For every scoring model building task, two colleagues compete against each other. The model with a higher precision will be implemented. When the model maintaining is carried out at a later point in time, the winning colleague is challenged by another colleague. This system leads to a highly competitive environment amongst the colleagues of the BI department. Consequently, the employees preprocess the data on their own (therefore every analyst works with a different database) and do not give variables descriptive denominations, using hashes instead in order to prevent their competitors from learning based on their results. As a result, if an employee leaves the company, the colleagues are not necessarily able to understand the developed models. This problem is currently tried to be solved by a project for the development of a homogeneous database.

understanding along the different subsidiaries of main aspects, e.g. the definition of "new customer", "turnover" etc. Therefore, the company focusses primary at the re-organization of the reporting-oriented processes to reduce the "uncontrolled growth" as named by the expert.

Although the three named critical aspects are not covered by the model, the resulting evaluation is the same as the experts' one.

Company 11

Company eleven has set up a BI competence center resulting from a high management awareness, which has been both staffed with sufficient manpower as well as a budget for analytic-relevant projects. A company-wide analysis strategy has been defined and used for the development and coordination of individual initiatives based on a road map. Therefore, the company is aware of the next steps to improve the capabilities in the field of Big Data.

Currently missing capabilities are in the field of operationalization of the analysis results as no standardized process exists for the further processing of results and reports. The company is aware of the related potential but the needed organizational changes have been pushed to the end of the Big Data roadmap.

During the interview with the consultant, it became clear again that the aspect of coordination of different projects is another relevant aspect. As a few items on level six are fulfilled, the model evaluation differs from the expert's assessment, whose evaluation is steered by the lacking operationalization of analysis results. This raises the challenge, that some topics have a higher perceived relevance from an experts point of view for the maturity assessment than others.

Bibliography

[Abelló et al. 2013] Abelló, Alberto ; Darmont, Jérôme ; Etcheverry, Lorena ; Golfarelli, Matteo ; Mazón, Jose-Norberto ; Naumann, Felix ; Pedersen, Torben ; Rizzi, Stefano B. ; Trujillo, Juan ; Vassiliadis, Panos ; Vossen, Gottfried: Fusion Cubes: towards self-service Business Intelligence. In: *International Journal of Data Warehousing and Mining* 9 (2013), No. 2, P. 66–88

[Accenture and GE 2015] Accenture ; GE: Industrial Internet Insights Report / Accenture, GE. URL `https://www.accenture.com/t20150523T024822{_}{_}w{_}{_}/us-en/{_}acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Dualpub{_}2/Accenture-Industrial-Internet-Changing-Competitive-Landscape-Industries.pdf`, 2015. – Research report

[Agarwal et al. 2011] Agarwal, Apoorv ; Xie, Boyi ; Vovsha, Ilia: Sentiment analysis of twitter data. In: *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, Association for Computational Linguistics, 2011, P. 30–38

[Agarwal and Dhar 2014] Agarwal, Ritu ; Dhar, Vasant: Editorial — Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. In: *Information Systems Research* 25 (2014), No. 3, P. 443–448

[Agrawal et al. 2012] Agrawal, Divyakant ; Bernstein, Philip ; Bertino, Elisa ; Davidson, Susan ; Dayal, Umeshwar ; Franklin, Michael ; Gehrke, Johannes ; Haas, Laura ; Halevy, Alon ; Han, Jiawei ; Jagadish, H V. ; Labrinidis, Alexandros ; Madden, Sam ; Papakonstantinou, Yannis ; Patel, Jingnesh M. ; Ramakrishnan, Raghu ; Ross, Kenneth ; Shahabi, Cyrus ; Suciu, Dan ; Vaithyanathan, Shiv ; Widom, Jennifer: *Challenges and Opportunities with Big Data: A community*

whitepaper developed by leading researchers across the United States. 2012. – URL http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf

[Aldrich 2008] Aldrich, Howard: *Organizations and Environments*. Stanford University Press, 2008. – 384 P. – ISBN 0804758298

[Amatriain 2013] Amatriain, Xavier: Mining large streams of user data for personalized recommendations. In: *ACM SIGKDD Explorations Newsletter* 14 (2013), No. 2, P. 37

[Anderson 2008] Anderson, Chris: The end of theory: The data deluge makes the scientific method obsolete. In: *Wired Magazin* (2008), aug, P. 8–10

[Andrienko and Andrienko 2012] Andrienko, Natalia ; Andrienko, Gennady: Visual analytics of movement: An overview of methods, tools and procedures. In: *Information Visualization* 12 (2012), sep, No. 1, P. 3–24

[Antweiler and Frank 2005] Antweiler, Werner ; Frank, Murray Z.: Is all that talk just noise? The information content of internet stock message boards. In: *The Journal of Finance* 59 (2005), No. 3, P. 1259–1294

[Argawal et al. 2011] Argawal, Divyakant ; Abbadi, Amr E. ; Das, Sudipto: Big Data and Cloud Computing : Current State and Future Opportunities. In: *Proceedings of the 14th International Conference on Extending Database Technology*. New York, New York, USA : ACM, 2011, P. 530–533

[Ari et al. 2012] Ari, I ; Olmezogullari, E ; Celebi, O F.: Data stream analytics and mining in the cloud. In: *IEEE 4th International Conference on Cloud Computing Technology and Science*, 2012, P. 857–862

[Armbrust et al. 2010] Armbrust, Michael ; Stoica, Ion ; Zaharia, Matei ; Fox, Armando ; Griffith, Rean ; Joseph, Anthony D. ; Katz, Randy ; Konwinski, Andy ; Lee, Gunho ; Patterson, David ; Rabkin, Ariel: A view of cloud computing. In: *Communications of the ACM* 53 (2010), No. 4, P. 50–58

[Aruldoss et al. 2014] Aruldoss, Martin ; Lakshmi Travis, Miranda ; Prasanna Venkatesan, V: A survey on recent research in business intelligence. In: *Journal of Enterprise Information Management* 27 (2014), No. 6, P. 831–866

[Assunção et al. 2013] Assunção, Marcos D. ; Calheiros, Rodrigo N. ; Bianchi, Silvia ; Netto, Marco A. S. ; Buyya, Rajkumar: *Big Data Computing and Clouds: Challenges, Solutions, and Future Directions.* 2013. – URL `http://arxiv.org/abs/1312.4722`

[Assunção et al. 2015] Assunção, Marcos D. ; Calheiros, Rodrigo N. ; Bianchi, Silvia ; Netto, Marco A. S. ; Buyya, Rajkumar: Big Data computing and clouds: Trends and future directions. In: *Journal of Parallel and Distributed Computing* 79-80 (2015), P. 3–15

[Bach 1994] Bach, James: The Immaturity of CMM. In: *American Programmer* 7 (1994), No. 9, P. 13–18

[Baeza-Yates and Yoelle 2012] Baeza-Yates, Ricardo ; Yoelle, Maarek: Usage data in web search: benefits and limitations. In: *Lecture Notes in Computer Science - Scientific and Statistical Database Management.* Berlin Heidelberg : Springer, 2012, P. 495–506

[Balci 1998] Balci, Osman: Verification, validation, and accreditation. In: *Proceedings of the 1998 Winter Simulation Conference*, IEEE Computer Society Press, 1998, P. 41–48

[Barbosa and Feng 2010] Barbosa, Luciano ; Feng, Junlang: Robust sentiment detection on twitter from biased and noisy data. In: *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 2010, P. 36–44

[Barnes 2013] Barnes, Trevor J.: Big data, little history. In: *Dialogues in Human Geography* 3 (2013), No. 3, P. 297–302

[Batini et al. 2009] Batini, Carlo ; Cappiello, Cinzia ; Francalanci, Chiara ; Maurino, Andrea: Methodologies for data quality assessment and improvement. In: *ACM Computing Surveys (CSUR)* 41 (2009), No. 3, P. 16

[Batra et al. 1990] Batra, Dinesh ; Hoffler, Jeffrey a. ; Bostrom, Robert P.: Comparing representations with relational and EER models. In: *Communications of the ACM* 33 (1990), No. 2, P. 126–139

[Becker 2004] Becker, Janine: *Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT)*, Free University Berlin, PhD thesis, 2004. – 217 P

[Becker et al. 2002] Becker, Jörg ; Algermissen, Lars ; Delfmann, Patrick ; Knackstedt, Ralf: Referenzmodellierung. In: *Das Wirtschaftsstudium* 30 (2002), No. 11, P. 1392–1395

[Becker et al. 2009] Becker, Jörg ; Knackstedt, Ralf ; Pöppelbuss, Jens: Developing Maturity Models for IT Management. In: *Business & Information Systems Engineering* 1 (2009), No. 3, P. 213–222

[Becker et al. 1995] Becker, Jörg ; Rosemann, Michael ; Schütte, Reinhard: Grundsätze ordnungsgemäßer Modellierung. In: *Wirtschaftsinformatik* 37 (1995), No. 5, P. 435–445

[Bennett and Campbell 2000] Bennett, Kristin P. ; Campbell, Colin: Support vector machines: hype or hallelujah? In: *ACM SIGKDD Explorations Newsletter* 2 (2000), No. 2, P. 1–13

[Biberoglu and Haddad 2002] Biberoglu, E ; Haddad, H: A survey of industrial experiences with CMM and the teaching of CMM practices. In: *Journal of Computing Sciences in Colleges* 18 (2002), No. 2, P. 143–152

[Birnbaum 1968] Birnbaum, Alan: Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M. (Publisher) ; Novick, M.R. (Publisher): *Statistical Theories of Mental Test Scores*. Reading, USA : Addison-Wesley, 1968, P. 395–479

[Bizer et al. 2011] Bizer, Christian ; Boncz, Peter ; Brodie, Michael L. ; Erling, Orri: The meaninful Use of Big Data: Four Perspectives. In: *SIGMOD* 40 (2011), No. 4, P. 56–60

[Blei et al. 2003] Blei, D M. ; Ng, A Y. ; Jordan, M I.: Latent Dirichlet Allocation. In: *The Journal of Machine Learning Research* 3 (2003), P. 993–1022

[Blei and Lafferty 2009] Blei, David M. ; Lafferty, John D.: Topic models. In: Srivastava, A (Publisher) ; Sahami, M (Publisher): *Text Mining Theory*

and Applications Vol.. 3. ACM Press, 2009, Chap. 4, P. 113–120. – URL http://portal.acm.org/citation.cfm?doid=1143844.1143859. – ISBN 1420059459

[Bliss et al. 2012] Bliss, Catherine A. ; Kloumann, Isabel M. ; Harris, Kameron D. ; Danforth, Christopher M. ; Dodds, Peter S.: Twitter reciprocal reply networks exhibit assortativity with respect to happiness. In: *Journal of Computational Science* 3 (2012), No. 3, P. 388–397

[BMWi 2014] BMWi: *Digitale Agenda 2014 – 2017*. 2014. – URL https://www.digitale-agenda.de

[Boudreau et al. 2001] Boudreau, Marie-Claude ; Gefen, David ; Straub, Detmar W.: Validation in Information Systems Research: A State-of-the-Art Assessment. In: *MIS Quarterly* 25 (2001), No. 1, P. 5–12

[Boyd and Crawford 2011] Boyd, Danah ; Crawford, Kate: Six Provocations for Big Data. In: *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011, P. 1–17

[Boyd and Crawford 2012] Boyd, Danah ; Crawford, Kate: Critical Questions for Big Data. In: *Information, Communication & Society* 15 (2012), No. 5, P. 662–679

[Braun 2007] Braun, Christian: *Modellierung der Unternehmensarchitektur : Weiterentwicklung einer bestehenden Methode und deren Abbildung in einem Meta-Modellierungswerkzeug*. Berlin : Logos Verlag, 2007. – 303 P. – ISBN 3832514716

[Braun 2009] Braun, Robert: *Referenzmodellierung: Grundlegung und Evaluation der Technik des Modell-Konfigurationsmanagements*. Logos Verlag, 2009. – 222 P. – ISBN 3832518940

[Bretzke 1980] Bretzke, Wolf R.: *Der Problembezug von Entscheidungsmodellen Einheit der Gesellschaftswissenschaften*. Tübingen : Mohr Siebeck, 1980. – 280 P

[vom Brocke 2003] Brocke, Jan vom: *Referenzmodellierung: Gestaltung und Verteilung von Konstruktionsprozessen*. Jan vom Brocke, 2003. – 424 P. – URL http://books.google.com/books?id=5wV2Ol2Cw4gC{&}pgis=1

[vom Brocke 2006] Brocke, Jan vom: *Serviceorientiertes Prozesscontrolling. Gestaltung von Organisations- und Informationssystemen bei Serviceorientierten Architekturen*, University Münster, Monograph, 2006

[Brohman and Parent 2000] Brohman, M K. ; Parent, Michael: The business intelligence value chain: Data-driven decision support in a data warehouse environment: An exploratory study. In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, IEEE, 2000, P. 1–10

[Brooks et al. 2013] Brooks, Patti ; El-Gayar, Omar ; Sarnikar, Surendra: Towards a Business Intelligence Maturity Model for Healthcare. In: *Proceedings of the 2013 46th Hawaii International Conference on System Sciences*. Washington, DC, USA : IEEE Computer Society, 2013 (HICSS '13), P. 3807–3816

[Brynjolfsson et al. 2011] Brynjolfsson, Erik ; Hitt, Lorin M. ; Kim, Heekyung H.: *Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?* 2011. – URL `http://www.ssrn.com/abstract=1819486`

[Bucher et al. 2008] Bucher, Tobias ; Riege, Christian ; Saat, Jan: Evaluation in der gestaltungsorientierten Wirtschaftsinformatik-Systematisierung nach Erkenntnisziel und Gestaltungsziel. In: *Wissenschaftstheorie und gestaltungsorientierte Wirtschaftsinformatik. Arbeitsbericht* (2008), No. 120, P. 69–86

[Buhl et al. 2013] Buhl, Hans U. ; Röglinger, Maximilian ; Moser, Florian ; Heidemann, Julia: Big data: A fashionable topic with(out) sustainable relevance for research and practice? In: *Business and Information Systems Engineering* 5 (2013), No. 2, P. 65–69

[Bundesministerium für Wirtschaft und Technologie 2014] Bundesministerium für Wirtschaft und Technologie: Smart Data – Innovationen aus Daten / Bundesministerium für Wirtschaft und Technologie. Berlin, 2014. – Research report

[Byrd et al. 1992] Byrd, Terry A. ; Cossick, Kathy L. ; Zmud, Robert W.: A synthesis of research on requirements analysis and knowledge acquisition techniques. In: *MIS Quarterly* 16 (1992), No. 1, P. 117–138

[Calder 1977] Calder, Bobby J.: Focus Groups and the Nature of Qualitative Marketing Research. In: *Journal of Marketing Research* 14 (1977), No. 4, P. 353–364

[Capgemini 2012] Capgemini: The deciding factor: big data & decision making / Capgemini. URL `https://www.uk.capgemini.com/resource-file-access/resource/pdf/`

The{_}Deciding{_}Factor{_}{_}Big{_}Data{_}{_}{_}Decision{_}Making.pdf, 2012. – Research report

[Capgemini 2015] Capgemini: Big & Fast Data: The Rise of Insight-Driven Business / Capgemini. URL https://www.capgemini.com/resource-file-access/resource/pdf/big{_}fast{_}data{_}the{_}rise{_}of{_}insight-driven{_}business-report.pdf, 2015. – Research report

[Carley 2002] Carley, Kathleen M.: Computational organizational science and organizational engineering. In: *Simulation Modelling Practice and Theory* 10 (2002), No. 5-7, P. 253–269

[Carter 2011] Carter, Philip: Big Data Analytics: Future Architectures , Skills and Roadmaps for the CIO / IDC. 2011. – Research report

[CERN 2014] CERN: *Computing*. 2014. – URL http://home.cern/about/computing. – Date Accessed: 2014-02-21

[Chang 2015] Chang, Jonathan: *lda: Collapsed Gibbs Sampling Methods for Topic Models*. 2015. – URL http://cran.r-project.org/package=lda

[Chang et al. 2009] Chang, Jonathan ; Boyd-Graber, Jordan ; Gerrish, Sean ; Wang, Chong ; Blei, David M.: Reading tea leaves: How humans interpret topic models. In: *Advances in Neural Information Processing Systems 22*. 2009, P. 288–296

[Chaudhuri et al. 2011] Chaudhuri, Surajit ; Dayal, Umeshwar ; Narasayya, Vivek: An Overview of Business Intelligence Technology. In: *Communications of the ACM* 54 (2011), No. 8, P. 88

[Chawla and Davis 2013] Chawla, Nitesh V. ; Davis, Darcy A.: Bringing big data to personalized healthcare: A patient-centered framework. In: *Journal of General Internal Medicine* 28 (2013), No. 3, P. 660–665

[Chen et al. 2012] Chen, Hsinchun ; Chiang, R H L. ; Storey, V C.: Business intelligence and analytics: From big data to big impact. In: *MIS Quarterly* 36 (2012), No. 4, P. 1–24

[Chen 1976] Chen, Peter Pin-Shan: The entity-relationship model—toward a unified view of data. In: *ACM Transactions on Database Systems* 1 (1976), No. 1, P. 9–36

[Choi et al. 2012] Choi, Jaeho ; Croft, W B. ; Kim, Jin Y.: Quality models for microblog retrieval. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge management - CIKM '12*. New York, New York, USA : ACM Press, 2012, P. 1834–1838

[Chute 2012] Chute, Christopher G.: Obstacles and options for big-data applications in biomedicine: The role of standards and normalizations. In: *International Conference on Bioinformatics and Biomedicine*, IEEE, 2012, P. 1–1

[Cleven et al. 2009] Cleven, Anne ; Gubler, Philipp ; Hüner, Kai M.: Design alternatives for the evaluation of design science research artifacts. In: *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 2009

[Cosic et al. 2012] Cosic, Ranko ; Shanks, Graeme ; Maynard, Sean: Towards a business analytics capability maturity model. In: *ACIS 2012 : Location, location, location : Proceedings of the 23rd Australasian Conference on Information Systems 2012*, ACIS, 2012, P. 1–11

[Curé et al. 2012] Curé, Olivier ; Kerdjoudj, Fadhela ; Faye, David C. ; Le Duc, Chan ; Lamolle, Myriam: On The Potential Integration of an Ontology-Based Data Access Approach in NoSQL Stores. In: *Third International Conference on Emerging Intelligent Data and Web Technologies*, 2012, P. 166–173

[Curtis et al. 1995] Curtis, Bill ; Hefley, William E. ; Miller, Sally: Overview of the People Capability Maturity Model. / Carnegie-Mellon University Software Engineering Institute. URL `http://oai.dtic.mil/oai/oai?verb=getRecord{&}metadataPrefix=html{&}identifier=ADA301167`, 1995. – Research report

[Curtis et al. 2001] Curtis, Bill ; Hefley, William W. ; Miller, Sally A.: *The People Capability Maturity Model: Guidelines for Improving the Workforce*. Boston, USA : Addison-Wesley Educational Publishers, 2001. – 624 P. – URL `http://www.amazon.de/The-People-Capability-Maturity-Model/dp/0201604450`

[Cuzzocrea et al. 2011] Cuzzocrea, Alfredo ; SONG, Il-Yeol ; Davis, Karen C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: *DOLAP '11 Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. New York, New York, USA : ACM, 2011, P. 101–103

[Damsgaard and Scheepers 1999] Damsgaard, Jan ; Scheepers, Rens: A stage model of intranet technology implementation and management. In: *Proceedings of the Seventh European Conference on Information Systems, ECIS 1999, Copenhagen, 1999*, Copenhagen Business School, 1999, P. 100–116

[Davenport et al. 2012] Davenport, Thomas H. ; Barth, Paul ; Bean, Randy: How 'Big Data' Is Different. In: *MIT Sloan Management Review* 54 (2012), oct, No. 1, P. 22–24

[Davenport, Thomas H. and Patil 2012] Davenport, Thomas H. and Patil, DJ: Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. In: *Harvard Business Review* (2012), P. 70

[Dayal and Castellanos 2009] Dayal, Umeshwar ; Castellanos, Malu: Data integration flows for business intelligence. In: *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, ACM, 2009, P. 1–11

[De Bruin et al. 2005] De Bruin, Tonia ; Freeze, Ronald ; Kaulkarni, Uday ; Rosemann, Michael: Understanding the main phases of developing a maturity assessment model. In: *16th Australasian Conference on Information Systems*, 2005

[De Bruin and Rosemann 2005] De Bruin, Tonia ; Rosemann, Michael: Towards a Business Process Management Maturity Model. In: Bartmann, D (Publisher) ; Rajola, F (Publisher) ; Kallinikos, J (Publisher) ; Avison, D (Publisher) ; Winter, R (Publisher) ; Ein-Dor, P (Publisher) ; Becker, J (Publisher) ; Bodendorf, F (Publisher) ; Weinhardt, C (Publisher): *ECIS 2005 Proceedings of the Thirteenth European Conference on Information Systems*, Verlag and the London School of Economics, 2005, P. 1–12

[Dean and Ghemawat 2008] Dean, Jeffrey ; Ghemawat, Sanjay: MapReduce : Simplified Data Processing on Large Clusters. In: *Communications of the ACM* 51 (2008), No. 1, P. 107–113

[Derczynski et al. 2013] Derczynski, Leon ; Ritter, Alan ; Clark, Sam ; Bontcheva, Kalina: Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP, 2013, P. 198–206

[Diebold 2003] Diebold, Francis X.: "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. In: *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Volume III*. 2003, P. 115–122

[Dinter 2012] Dinter, Barabara: The Maturing of a Business Intelligence Maturity Model. In: *Americas Conference on Information Systems AMCIS*, AMCIS, 2012

[Dittmar et al. 2013] Dittmar, Carsten ; Ossendoth, Volker ; Schulze, Klaus-Dieter: Business Intelligence - Status quo in Europa / Steria Mummert. URL `http://www.bi.soprasteria.de/bi-news-infos/europaeische-bima-studie-2012-13`, 2013. – Research report

[Drinka and Yen 2008] Drinka, Dennis ; Yen, Minnie Yi-Miin: Controlling Curriculum Redesign with a Process Improvement Model -. In: *Journal of Information Systems Education* 19 (2008), No. 3, P. 331–342

[Dumbill 2012] Dumbill, Edd: *What is Big Data?* 2012. – URL `https://www.oreilly.com/ideas/what-is-big-data`. – Date Accessed: 2016-03-12

[EFQM 2012] EFQM: *Model Criteria | EFQM*. 2012. – URL `http://www.efqm.org/efqm-model/model-criteria`. – Date Accessed: 2015-08-11

[Embretson and Linacre 1996] Embretson, Susan E. ; Linacre, John M.: The new rules of measurement. In: *Psychological Assessment* 8 (1996), No. 4, P. 341–349

[Fayyad et al. 1996] Fayyad, Usama M. ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: Knowledge discovery and data mining: Towards a unifying framework. In: *2nd International Conference on Knowledge Discovery and Data Mining*, 1996, P. 82–88

[Feinerer and Hornik 2015] Feinerer, Ingo ; Hornik, Kurt: *tm: A framework for text mining applications within R*. 2015. – URL `http://tm.r-forge.r-project.org/`

[Fellows 2014] Fellows, Ian: *wordcloud: Word Clouds*. 2014. – URL `http://cran.r-project.org/package=wordcloud`

[Feng et al. 2012] Feng, Zhu ; Jie, Liu ; Lijie, Xu: A Fast and High Throughput SQL Query System for Big Data. In: *Lecture Notes in Computer Science* 7651 (2012), P. 783–788

[Ferreira et al. 2013] Ferreira, Nivan ; Poco, Jorge ; Vo, Huy T. ; Freire, Juliana ; Silva, Cláudio T: Visual exploration of big spatio-temporal urban data: a study of New York City taxi trips. In: *IEEE transactions on visualization and computer graphics* 19 (2013), No. 12, P. 2149–2158

[Fettke and Loos 2004] Fettke, Peter ; Loos, Peter: Referenzmodellierungsforschung. In: *Wirtschaftsinformatik* 46 (2004), P. 331–340

[Fettke and Loos 2005] Fettke, Peter ; Loos, Peter: Der Beitrag der Referenzmodellierung zum Business Engineering. In: *HMD - Praxis der Wirtschaftsinformatik* 241 (2005), P. 18–26

[Fisher 2004] Fisher, David M.: The business process maturity model: a practical approach for identifying opportunities for optimization. In: *Business Process Trends* 9 (2004), No. 4, P. 13

[Forestier et al. 2012] Forestier, Mathilde ; Stavrianou, Anna ; Velcin, Julien ; Zighed, Djamel A.: Roles in social networks: Methodologies and research issues. In: *Web Intelligence and Agent Systems* 10 (2012), No. 1, P. 117–133

[Fox and Do 2013] Fox, Stephen ; Do, Tuan: Getting real about Big Data: applying critical realism to analyse Big Data hype. In: *International Journal of Managing Projects in Business* 6 (2013), sep, No. 4, P. 739–760

[Fraser et al. 2002] Fraser, Peter ; Moultrie, James ; Gregory, Mike: The use of maturity models/grids as a tool in assessing product development capability. In: *IEEE International Engineering Management Conference* Vol.. 1, IEEE, 2002, P. 244–249

[Gantz et al. 2007] Gantz, John F. ; Mcarthur, John ; Minton, Stephen: The Expanding Digital Universe / International Data Corporation. URL http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf, 2007. – Research report

[Gartner 2015] Gartner: *Gartner IT Glossary.* 2015. – URL http://www.gartner.com/it-glossary/big-data/. – Date Accessed: 2016-03-12

[Garza et al. 2014] Garza, Paolo ; Margara, Paolo ; Nepote, Nicolo ; Grimaudo, Luigi ; Piccolo, Elio: Hadoop on a low-budget general purpose hpc cluster in academia. In: *Advances in Intelligent Systems and Computing* 241 (2014), P. 187–192

[Gayo-Avello 2011] Gayo-Avello, Daniel: Dont Turn Social Media Into Another Literary Digest Poll. In: *Communications of the ACM* 54 (2011), No. 10, P. 121–128

[Gehlert 2007] Gehlert, Andreas: *Migration fachkonzeptueller Modelle*, Technical University Dresden, PhD thesis, 2007. – 392 P

[Gellman and Poitras 2013] Gellman, Barton ; Poitras, Laura: *U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program.* 2013. – URL https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497{_}story.html

[Gemino and Wand 2003] Gemino, Andrew ; Wand, Yair: Conceptual Modeling and System Architecting - Evaluating Modeling Techniques Based on Models of Learning. In: *Communications of the ACM* 46 (2003), No. 10, P. 79–84

[George et al. 2014] George, Gerard ; Haas, Martine R. ; Pentland, Alex: From the editors: Big data and management. In: *Academy of Management Journal* 57 (2014), No. 2, P. 321–326

[Gericke et al. 2006] Gericke, Anke ; Rohner, Peter ; Winter, Robert: Networkability in the Health Care Sector - Necessity, Measurement and Systematic Development as the Prerequisites for Increasing the Operational Efficiency of Administrative Processes. In: *Proceedings of the 17th Australasian Conference on Information Systems*, 2006

[Gluchowski 2001] Gluchowski, Peter: Business Intelligence. In: *HMD Praxis der Wirtschaftsinformatik* 222 (2001), P. 5–15

[Gluchowski and Kemper 2006] Gluchowski, Peter ; Kemper, Hans-Georg: Quo Vadis Business Intelligence. In: *Zeitschrift BI-Spektrum* 1 (2006), P. 12–19

[Gregor and Hevner 2013] Gregor, Shirley ; Hevner, Alan R.: Positioning and Presenting Design Science Research for Maximum Impact. In: *MIS Quarterly* 37 (2013), No. 2, P. 337–355

[Griffiths and Steyvers 2004] Griffiths, T L. ; Steyvers, Mark: Finding scientific topics. In: *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), apr, P. 5228–5235

[Grimes 2008] Grimes, Seth: *Unstructured Data and the 80 Percent Rule*. 2008. – URL `http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/`. – Date Accessed: 2016-02-23

[Grindle et al. 2013] Grindle, Mark ; Kavathekar, Jitendra ; Wan, Dadong: A new era for the healthcare industry / Accenture. 2013. – Research report

[Grütter et al. 1998] Grütter, Rolf ; Geyer, Georg ; Schmid, Beat: Konzeption einer ELIAS-Applikationsarchitektur für das klinische Studiendatenbanksystem der L.A.B. Neu-Ulm. In: *Wirtschaftsinformatik* 40 (1998), No. 4, P. 291–300

[Gu and Gao 2012] Gu, Chunhao ; Gao, Yang: A Content-Based Image Retrieval System Based on Hadoop and Lucene. In: *Second International Conference on Cloud and Green Computing*, 2012, P. 684–687

[Hambleton et al. 1991] Hambleton, Ronald K. ; Swaminathan, Hariharan ; Rogers, H. J.: *Fundamentals of item response theory. Measurement methods for the social sciences series, Vol. 2.* Thousand Oaks, USA : Sage Publications, 1991

[Hammoud and Sakr 2011] Hammoud, Mohammad ; Sakr, Majd F.: Locality-Aware Reduce Task Scheduling for MapReduce. In: *2011 IEEE Third International Conference on Cloud Computing Technology and Science*, IEEE, 2011, P. 570–576

[Han et al. 2012] Han, Xiaoyue ; Tian, Lianhua ; Yoon, Minjoo ; Lee, Minsoo: A Big Data Model Supporting Information Recommendation in Social Networks. In: *Second International Conference on Cloud and Green Computing*, 2012, P. 810–813

[Hansmann and Niemeyer 2014] Hansmann, Thomas ; Niemeyer: Big Data - Characterizing an Emerging Research Field Using Topic Models. In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2014, P. 43–51

[Hansmann and Nottorf 2015] Hansmann, Thomas ; Nottorf, Florian: Decision Support in the Field of Online Marketing - Development of a Data Landscape. In: Obaidat, Mohammad S. (Publisher) ; Holzinger, Andreas (Publisher) ; Filipe, Joaquim (Publisher): *E-Business and Telecommunications* Vol.. 554. Cham, Switzerland : Springer International Publishing, 2015, P. 76–95. – ISBN 978-3-319-25914-7

[Hartog et al. 2012] Hartog, Jessica ; Fadika, Zacharia ; Dede, Elif ; Govindaraju, Madhusudhan: Configuring a MapReduce Framework for Dynamic and Efficient Energy Adaptation. In: *IEEE 5th International Conference on Cloud Computing*, 2012, P. 914–921

[Helland 2011] Helland, Pat: If You Have Too Much Data, Then 'Good Enough' Is Good Enough. In: *Communications of the ACM* 54 (2011), No. 6, P. 40–47

[Herbsleb and Goldenson 1996] Herbsleb, James D. ; Goldenson, Dennis R.: A systematic survey of CMM experience and results. In: *Proceedings of IEEE 18th International Conference on Software Engineering*, IEEE Comput. Soc. Press, 1996, P. 323–330

[Hering 1984] Hering, Ekbert: Programmablaufplan nach DIN 66001. In: *Software-Engineering*. Wiesbaden : Vieweg+Teubner Verlag, 1984, P. 26–34. – ISBN 978-3-528-04284-4

[Herodotou et al. 2011] Herodotou, Herodotos ; Lim, Harold ; Luo, Gang ; Borisov, Nedyalko ; Dong, Liang: Starfish : A Self-tuning System for Big Data Analytics. In: *5th Biennial Conference on Innovative Data Systems Research (CIDR '11)* Vol.. 11, CIDR, 2011, P. 261–272

[Hevner 2007] Hevner, Alan R.: A Three Cycle View of Design Science Research. In: *Scandinavian Journal of Information Systems* 19 (2007), No. 2, P. 87–92

[Hevner et al. 2004] Hevner, Alan R. ; March, Salvatore T. ; Park, Jinsoo ; Ram, Sudha: Design science in information systems research. In: *MIS Quarterly* 28 (2004), No. 1, P. 75–105

[Hilbert and López 2011] Hilbert, Martin ; López, Priscila: The world's technological capacity to store, communicate, and compute information. In: *Science* 332 (2011), apr, No. 6025, P. 60–65

[Hopkins 2011] Hopkins, Brian: *Big Data, Brewer, And A Couple Of Webinars.* 2011. – URL `http://blogs.forrester.com/brian{_}hopkins/11-08-29-big{_}data{_}brewer{_}and{_}a{_}couple{_}of{_}webinars`. – Date Accessed: 2012-07-01

[Humphrey 1988] Humphrey, Watts S.: Characterizing the software process: a maturity framework. In: *IEEE Software* 5 (1988), No. 2, P. 73–79

[IBM 2011] IBM: *IBM - What is Big Data?* 2011. – URL `http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html`. – Date Accessed: 2015-08-29

[IBM Systems and Technology 2011] IBM Systems and Technology: *Watson – A System Designed for Answers.* 2011. – URL `http://www-03.ibm.com/innovation/us/engines/assets/9442{_}Watson{_}A{_}System{_}White{_}Paper{_}POW03061-USEN-00{_}Final{_}Feb10{_}11.pdf`

[Inversini 2005] Inversini, Simone: *Wirkungsvolles Change Management in Abhängigkeit von situativen Anforderungen. Organisationale Veränderungsprozesse im Spannungsfeld von betrieblichen Voraussetzungen und Umweltanforderungen unter Berücksichtigung der wirtschaftlichen, organisationsbezoge*, University Potsdam, PhD Thesis, 2005

[Jacobs 2009] Jacobs, Adam: The pathologies of big data. In: *Communications of the ACM* 52 (2009), aug, No. 8, P. 36

[Jeong et al. 2012] Jeong, Hae-Duck J. ; WooSeok, Hyun ; Jiyoung, Lim ; Ilsun, You: Anomaly Teletraffic Intrusion Detection Systems on Hadoop-Based Platforms: A Survey of Some Problems and Solutions. In: *15th International Conference on Network-Based Information Systems*, 2012, P. 766–770

[Kaisler et al. 2013] Kaisler, Stephen ; Armour, Frank ; Espinosa, J A. ; Money, William: Big Data: Issues and Challenges Moving Forward. In: *2013 46th Hawaii International Conference on System Sciences, HICCS '13*, Ieee, 2013, P. 995–1004

[Kamprath 2011] Kamprath, Nora: Einsatz von Reifegradmodellen im Prozessmanagement. In: *HMD Praxis der Wirtschaftsinformatik* 48 (2011), No. 282, P. 93–102

[Kejiang et al. 2012] Kejiang, Ye ; Xiaohong, Jiang ; Yanzhang, He ; Xiang, Li ; Haiming, Yan ; Peng, Huang: vHadoop: A Scalable Hadoop Virtual Cluster Platform for MapReduce-Based Parallel Machine Learning with Performance Consideration. In: *IEEE International Conference On Cluster Computing Workshops*, 2012, P. 152–160

[Kiron and Shockley 2011] Kiron, David ; Shockley, Rebecca: Creating business value with analytics. In: *MIT Sloan Management Review* 53 (2011), No. 1, P. 57–63

[Klesse 2007] Klesse, Mario: *Leistungsverrechnung im Data Warehousing : Entwicklung einer Methode*, University of St. Gallen, PhD Thesis, 2007. – 451 P

[Klesse et al. 2005] Klesse, Mario ; Wortmann, Felix ; Schelp, Joachim: Erfolgsfaktoren der Applikationsintegration. In: *Wirtschaftsinformatik* 47 (2005), No. 4, P. 259–267

[König and Weitzel 2003] König, Wolfgang ; Weitzel, Tim: Netzeffekte im E-Business. In: Uhr, Wolfgang (Publisher) ; Esswein, Werner (Publisher) ; Schoop, Eric (Publisher): *Wirtschaftsinformatik 2003/Band I.* Heidelberg : Physica-Verlag HD, 2003, P. 9–33. – ISBN 978-3-642-63266-2

[Kridel and Dolk 2013] Kridel, Don ; Dolk, Daniel: Automated self-service modeling: predictive analytics as a service. In: *Information Systems and e-Business Management* 11 (2013), No. 1, P. 119–140

[Kuvaja 1999] Kuvaja, Pasi: Bootstrap 3.0—a spice1 conformant software process assessment methodology. In: *Software Quality Journal* 8 (1999), P. 7–19

[Kwon et al. 2014] Kwon, Ohbyung ; Lee, Namyeon ; Shin, Bongsik: Data quality management, data usage experience and acquisition intention of big data analytics. In: *International Journal of Information Management* 34 (2014), No. 3, P. 387–394

[Lahrmann and Marx 2010] Lahrmann, Gerrit ; Marx, Frederik: Systematization of Maturity Model Extensions. In: Winter, Robert (Publisher) ; Zhao, J L. (Publisher) ; Aier, Stephan (Publisher): *Proceedings of DESRIST 2010*. Berlin, Heidelberg : Springer, 2010, P. 522–525

[Lahrmann et al. 2011a] Lahrmann, Gerrit ; Marx, Frederik ; Mettler, Tobias ; Winter, Robert ; Wortmann, Felix: Inductive Design of Maturity Models: Applying the Rasch Algorithm for Design Science Research. In: Jain, Hemant (Publisher) ; Sinha, Atish P. (Publisher) ; Vitharana, Padmal (Publisher): *Service-Oriented Perspectives in Design Science Research* Vol.. 6629. Berlin, Heidelberg : Springer, 2011, P. 176–191

[Lahrmann et al. 2010] Lahrmann, Gerrit ; Marx, Frederik ; Winter, Robert ; Wortmann, Felix: Business Intelligence Maturity Models: An Overview. In: *VII Conference of the Italian Chapter of AIS*. Naples : AIS, 2010

[Lahrmann et al. 2011b] Lahrmann, Gerrit ; Marx, Frederik ; Winter, Robert ; Wortmann, Felix: Business intelligence maturity: Development and evaluation of a theoretical model. In: *System Sciences (HICSS), 2011 44th Hawaii International Conference on* IEEE (Veranst.), 2011, P. 1–10

[Lämmel 2008] Lämmel, Ralf: Google's MapReduce Programming Model — Revisited. In: *Science of Computer Programming* 70 (2008), No. 1, P. 1–30

[Lane et al. 2014] Lane, Julia ; Stodden, Victoria ; Bender, Stefan ; Nissenbaum, Helen: *Privacy, Big Data, and the Public Good*. Cambridge : Cambridge University Press, 2014

[Laney 2001] Laney, Doug: 3D Data Management: Controlling Data Volume, Velocity, and Variety / META Group. Stamford, CT, USA, 2001 (February 2001). – Research report

[LaValle et al. 2011] LaValle, S ; Lesser, Eric ; Shockley, Rebecca: Big Data, Analytics and the Path From Insights to Value. In: *MIT Sloan Management Review* 52 (2011), No. 2, P. 21–31

[Lee et al. 2013] Lee, Jay ; Lapira, Edzel ; Bagheri, Behrad ; Kao, Hung-an: Recent advances and trends in predictive manufacturing systems in big data environment. In: *Manufacturing Letters* 1 (2013), No. 1, P. 38–41

[Lee et al. 2010] Lee, Sangno ; Baker, Jeff ; Song, Jaeki ; Wetherbe, James C.: An Empirical Comparison of Four Text Mining Methods. In: *2010 43rd Hawaii International Conference on System Sciences*, IEEE, 2010, P. 1–10

[Leffson 1987] Leffson, Ulrich: *Die Grundsätze ordnungsmässiger Buchführung.* Düsseldorf : IDW-Verlag, 1987

[Liu et al. 2012] Liu, Chunyan ; Zhu, Conghui ; Zhao, Tiejun ; Zheng, Dequan: Extracting Main Content of a Topic on Online Social Network by Multi-document Summarization. In: *8th International Conference on Computational Intelligence and Security*, 2012, P. 52–55

[Liu et al. 2013] Liu, Zhicheng ; Jiang, Biye ; Heer, Jeffrey: ImMens: Real-time visual querying of big data. In: *Computer Graphics Forum* 32 (2013), No. 3 PART4, P. 421–430

[Lockamy and McCormack 2004] Lockamy, Archie ; McCormack, Kevin: The development of a supply chain management process maturity model using the concepts of business process orientation. In: *Supply Chain Management: An International Journal* 9 (2004), No. 4, P. 272–278

[Lönnqvist and Pirttimäki 2006] Lönnqvist, Antti ; Pirttimäki, Virti: The measurement of business intelligence. In: *Information Systems Management* 23 (2006), No. 1, P. 32–40

[Luftman and Ben-Zvi 2010] Luftman, Jerry ; Ben-Zvi, Tal: Key Issues for IT Executives 2009: Difficult Economy's Impact on IT. In: *MIS Quarterly Executive* 9 (2010), No. 1, P. 203–213

[Luftman et al. 2015] Luftman, Jerry ; Lyytinen, Kalle ; Ben-Zvi, Tal: Enhancing the measurement of information technology (IT) business alignment and its influence on company performance. In: *Journal of Information Technology* (2015), P. 1–21

[Lukman et al. 2011] Lukman, Tomaž ; Hackney, Ray ; Popovič, Aleš ; Jaklič, Jurij ; Irani, Zahir: Business Intelligence Maturity: The Economic Transitional

Context Within Slovenia. In: *Information Systems Management* 28 (2011), No. 3, P. 211–222

[Lynch 2008] Lynch, Clifford: How do your data grow ? In: *Nature* 455 (2008), No. September, P. 28–29

[Lyon 2014] Lyon, David: Surveillance, Snowden, and Big Data: Capacities, consequences, critique. In: *Big Data & Society* 1 (2014), No. 2, P. 1–13

[Madden 2012] Madden, Sam: From Databases to Big Data. In: *IEEE Computing* 16 (2012), No. 3, P. 4–6

[Maier et al. 2012] Maier, Anja M. ; Moultrie, James ; Clarkson, P. J.: Assessing Organizational Capabilities: Reviewing and Guiding the Development of Maturity Grids. In: *IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT* 59 (2012), No. 1, P. 138–159

[Maier 2013] Maier, Markus: *Towards a Big Data Reference Architecture*, Eindhoven University of Technology, Monograph, 2013

[Manyika et al. 2011] Manyika, James ; Chui, Michael ; Brown, Brad ; Bughin, Jacques ; Dobbs, Richard ; Roxburgh, Charles ; Byers, Angela H.: Big data : The next frontier for innovation , competition , and productivity / McKinsey Global Institute. 2011 (June). – Research report

[March and Smith 1995] March, Salvatore T. ; Smith, Gerald F.: Design and natural science research on information technology. In: *Decision Support Systems* 15 (1995), No. 4, P. 251–266

[Marshall and Mitchell 2004] Marshall, Stephen ; Mitchell, Geoff: Applying SPICE to e-learning: an e-learning maturity model? (2004), P. 185–191

[Marx et al. 2010] Marx, Frederik ; Lahrmann, Gerrit ; Winter, Robert: Aligning corporate planning and business intelligence: a combined maturity model. In: *VII Conference of the Italian Chapter of AIS (itAIS 2010)*, 2010, P. 1–10

[Marx et al. 2012] Marx, Frederik ; Wortmann, Felix ; Mayer, Jörg H: Ein Reifegradmodell für Unternehmenssteuerungssysteme. In: *Wirtschaftsinformatik* 54 (2012), No. 4, P. 189–204

[Marx 2013] Marx, Vivien: Biology: The big challenges of big data. In: *Nature* 498 (2013), No. 7453, P. 255–260

[Mayer-Schönberger and Cukier 2013] Mayer-Schönberger, Viktor ; Cukier, Kenneth: *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Boston, USA : Houghton Mifflin Harcourt, 2013

[McAfee and Brynjolfsson 2012] McAfee, Andrew ; Brynjolfsson, Erik: Big data: the management revolution. In: *Harvard business review* 90 (2012), No. 10, P. 3–9

[McCormack et al. 2009] McCormack, Kevin ; Willems, Jurgen ; Bergh, Joachim van den ; Deschoolmeester, Dirk ; Willaert, Peter ; Indihar Štemberger, Mojca ; Škrinjar, Rok ; Trkman, Peter ; Bronzo Ladeira, Marcelo ; Paulo Valadares de Oliveira, Marcos ; Bosilj Vuksic, Vesna ; Vlahovic, Nikola: A global investigation of key turning points in business process maturity. In: *Business Process Management Journal* 15 (2009), No. 5, P. 792–815

[Mertens 2006] Mertens, Peter: Moden und Nachhaltigkeit in der Wirtschaftsinformatik. In: *HMD-Praxis der Wirtschaftsinformatik* 250 (2006), No. 1, P. 109–118

[Mettler 2010] Mettler, Tobias: *Supply Management im Krankenhaus - Konstruktion und Evaluation eines konfigurierbaren Reifegradmodells zur zielgerichteten Gestaltung*, St. Gallen, PhD Thesis, 2010

[Mettler and Rohner 2009] Mettler, Tobias ; Rohner, Peter: Situational maturity models as instrumental artifacts for organizational design. In: *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*. New York, New York, USA : ACM Press, 2009, P. 22:1–22:9

[Microsoft 2013] Microsoft: *The Big Bang: How the Big Data Explosion Is Changing the World*. 2013. – URL `http://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/$\delimiter"026E30F$n`. – Date Accessed: 2016-03-12

[Mislevy 1982] Mislevy, Robert J.: Foundations of a new Test Theory. In: *ETS Research Report Series* 1982 (1982), No. 2, P. i–32

[Montoya-Weiss and Calantone 1994] Montoya-Weiss, Mitzi M. ; Calantone, Roger: Determinants of new product performance: A review and meta-analysis. In: *Journal of Product Innovation Management* 11 (1994), No. 5, P. 397–417

[Mukherjee et al. 2012] Mukherjee, Arjun ; Liu, Bing ; Glance, Natalie: Spotting fake reviewer groups in consumer reviews. In: *Proceedings of the 21st international conference on World Wide Web - WWW '12*. New York, New York, USA : ACM Press, 2012, P. 191–200

[Mukhopadhyay et al. 1995] Mukhopadhyay, Tridas ; Kekre, Sunder ; Kalathur, Suresh: Business Value of Information Technology: A Study of Electronic Data Interchange. In: *MIS Quartely* 19 (1995), No. 2, P. 137–156

[Negash 2004] Negash, Solomon: Business intelligence. In: *Communications of the Association for Information Systems* 13 (2004), P. Article 15

[Neuhauser 2004] Neuhauser, Charlotte: A maturity model: Does it provide a path for online course design. In: *The Journal of Interactive Online Learning* 3 (2004), No. 1, P. 1–17

[Newman et al. 2002] Newman, Jessica C. ; Des Jarlais, Don C. ; Turner, Charles F. ; Gribble, Jay ; Cooley, Phillip ; Paone, Denise: The Differential Effects of Face-to-Face and Computer Interview Modes. In: *American Journal of Public Health* 92 (2002), feb, No. 2, P. 294–297

[Ngai et al. 2009] Ngai, Eric W. ; Xiu, Li ; Chau, Dorothy C.: Application of data mining techniques in customer relationship management: A literature review and classification. In: *Expert Systems with Applications* 36 (2009), No. 2, P. 2592–2602

[Nolan 1973] Nolan, Richard L.: Managing the computer resource: a stage hypothesis. In: *Communications of the ACM* 16 (1973), No. 7, P. 399–405

[Park et al. 2012] Park, Sung-Hyuk ; Huh, Soon-Young ; Oh, Wonseok ; Han, Sang P.: A social network-based inference model for validating customer profile data. In: *MIS Quarterly* 36 (2012), No. 4, P. 1217–1237

[Paulk et al. 1993] Paulk, M C. ; Weber, C V. ; Garcia, S M. ; Curtis, Bill: Capability Maturity Model for Software, Version 1.1. In: *Software* 10 (1993), No. 4, P. 18–27

[Peffers et al. 2007] Peffers, Ken ; Tuunanen, Tuure ; Rothenberger, Marcus A. ; Chatterjee, Samir: A Design Science Research Methodology for Information Systems Research. In: *Journal of Management Information Systems* 24 (2007), No. 3, P. 45–77

[Porter 1980] Porter, M F.: An algorithm for suffix stripping. In: *Program* 14 (1980), No. 3, P. 130–137

[Pospiech and Felden 2012] Pospiech, M ; Felden, C: Big Data–A State-of-the-Art. In: *Americas Conference on Information Systems AMCIS 2012*, 2012, P. Paper 22

[Powell and Dent-Micallef 1997] Powell, Thomas ; Dent-Micallef, Anne: Information Technology as Competitive Advantage: The Role of Human, Business and Technology Resources. In: *Strategic Management Journal* 18 (1997), No. 5, P. 375–405

[Prenzel et al. 2007] Prenzel, Manfred ; Carstensen, Claus H. ; Frey, Andreas ; Drechsel, Barbara ; Rönnebeck, Silke: PISA 2006 – Eine Einführung in die Studie. In: Prenzel, Manfred (Publisher) ; Artelt, Cordula (Publisher) ; Baumert, Jürgen (Publisher) ; Blum, Werner (Publisher) ; Hammann, Marcus (Publisher) ; Klieme, Eckhard (Publisher) ; Pekrun, Reinhard (Publisher): *PISA 2006 : Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster, Germany : Waxmann, 2007, Chap. 2, P. 31–60

[Provost and Fawcett 2013] Provost, Foster ; Fawcett, Tom: Data Science and its Relationship to Big Data and Data-Driven Decision Making. In: *Data Science and Big Data* 1 (2013), No. 1, P. 51–59

[Raber et al. 2012] Raber, David ; Winter, Robert ; Wortmann, Felix: Using Quantitative Analyses to Construct a Capability Maturity Model for Business Intelligence. In: *2012 45th Hawaii International Conference on System Sciences*, URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6149408`, 2012, P. 4219–4228

[Raber et al. 2013a] Raber, David ; Wortmann, Felix ; Winter, Robert: Situational Business Intelligence Maturity Models: An Exploratory Analysis. In: *2013 46th Hawaii International Conference on System Sciences*, URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6480304`, 2013, P. 3797–3806

[Raber et al. 2013b] Raber, David ; Wortmann, Felix ; Winter, Robert: Towards The Measurement Of Business Intelligence Maturity. In: *Proceedings of the 21st European Conference on Information Systems (ECIS 2013)*, 2013

[Rajaram 2013] Rajaram, Dhiraj: Why some data scientists should really be called decision scientists. In: *Analytics Magazine* (2013), No. October

[Reise 1990] Reise, Steven P.: A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT. In: *Applied Psychological Measurement* 14 (1990), No. 2, P. 127–137

[Reyes and Giachetti 2010] Reyes, Heriberto G. ; Giachetti, Ronald: Using experts to develop a supply chain maturity model in Mexico. In: *Supply Chain Management: An International Journal* 15 (2010), No. 6, P. 415–424

[Riege et al. 2009] Riege, Christian ; Saat, Jan ; Bucher, Tobias: Systematisierung von Evaluationsmethoden in der gestaltungsorientierten Wirtschaftsinformatik. In: *Wissenschaftstheorie und gestaltungsorientierte Wirtschaftsinformatik*. Heidelberg : Physika, 2009, P. 69–86

[van Rijmenam 2013] Rijmenam, Mark van: *Why The 3V's Are Not Sufficient To Describe Big Data*. 2013. – URL `https://datafloq.com/read/3vs-sufficient-describe-big-data/166`. – Date Accessed: 2014-09-17

[Rivera and van der Meulen 2014] Rivera, Janessa ; Meulen, Rob van der: *Gartner's 2014 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business*. 2014. – URL `http://www.gartner.com/newsroom/id/2819918`. – Date Accessed: 2015-07-05

[Rizopoulos 2006] Rizopoulos, Dimitris: ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. In: *Journal of Statistical Software* 17 (2006), No. 5, P. 1–25

[Rosas et al. 2012] Rosas, Claudia ; Sikora, Anna ; Jorba, Josep ; Moreno, Andreu ; César, Eduardo: Improving Performance on Data-Intensive Applications Using a Load Balancing Methodology Based on Divisible Load Theory. In: *International Journal of Parallel Programming* (2012), P. 1–25

[Rost 1999] Rost, Jürgen: Was ist aus dem Rasch-Modell geworden? In: *Psychologische Rundschau* 50 (1999), No. 3, P. 140–156

[Rost 2004] Rost, Jürgen: *Testtheorie - Testkonstruktion.* Bern : Huber, 2004

[Rouhani et al. 2012] Rouhani, Saeed ; Asgari, Sara ; Mirhosseini, Seyed V.: Review Study: Business Intelligence Concepts and Approaches. In: *American Journal of Scientific Research* 50 (2012), No. 50, P. 62–75

[Salipante et al. 1982] Salipante, Paul ; Notz, William ; Bigelow, John: A matrix approach to literature reviews. In: *Research in organizational behavior* 4 (1982), P. 321–348

[Sanders 2006] Sanders, James R.: *Handbuch der Evaluationsstandards - Die Standards des "Joint Committee on Standards for Educational Evaluation".* 3rd. Wiesbaden, Germany : VS Verlag für Sozialwissenschaften, 2006. – 362 P

[SAS 2015a] SAS: *Big Data - What it is and why it matters.* 2015. – URL `http://www.sas.com/en{_}us/insights/big-data/what-is-big-data.html`. – Date Accessed: 2016-03-02

[SAS 2015b] SAS: *SAS Visual Analytics - Fact Sheet.* 2015. – URL `http://www.sas.com/content/dam/SAS/en{_}us/doc/factsheet/sas-visual-analytics-105682.pdf`. – Date Accessed: 2016-03-22

[Schaller 1997] Schaller, Robert R.: Moore's law: past, present and future. In: *Spectrum, IEEE* 34 (1997), No. 6, P. 52–59

[Schauer and Schauer 2009] Schauer, Hanno ; Schauer, Carola: Moden in der Wirtschaftsinformatik: Wissenschaftstheoretische und wissenschaftspraktische Überlegungen zu einer von Hypes geprägten Disziplin. In: *Wirtschaftsinformatik (1)*, 2009, P. 431–440

[Schulze and Dittmar 2006] Schulze, Klaus-Dieter ; Dittmar, Carsten: *Business Intelligence Reifegradmodelle.* Chap. Business I, P. 71–87. In: Chamoni, Peter (Publisher) ; Gluchowski, Peter (Publisher): *Analytische Informationssysteme.* Berlin, Heidelberg : Springer Berlin Heidelberg, 2006

[Schütte 1998] Schütte, Reinhard: *Grundsätze ordnungsmäßiger Referenzmodellierung.: Konstruktion konfigurations- und anpassungsorientierter Modelle.* 1998

[Schwinn 2006] Schwinn, Alexander: *Entwicklung einer Methode zur Gestaltung von Integrationsarchitekturen für Informationssysteme*, University of St. Gallen, PhD Thesis, 2006

[Shankaranarayanan and Cai 2006] Shankaranarayanan, Ganesan ; Cai, Yu: Supporting data quality management in decision-making. In: *Decision Support Systems* 42 (2006), No. 1, P. 302–317

[Shearer et al. 2000] Shearer, Colin ; Watson, Hugh J. ; Grecich, Daryl G. ; Moss, Larissa ; Adelman, Sid ; Hammer, Katherine ; Herdlein, Stacey a.: The CRISP-DM model: The New Blueprint for Data Mining. In: *Journal of Data Warehousing* 5 (2000), No. 4, P. 13–22

[Simpson and Weiner 1989] Simpson, John ; Weiner, Edmund: *The Oxford English Dictionary - Hardback - John Simpson, Edmund Weiner -*. 2nd. Oxford; New York : Oxford : Clarendon Press ; Oxford ; New York : Oxford University Press, 1989

[Sinclair and Cardew-Hall 2007] Sinclair, J ; Cardew-Hall, M: The folksonomy tag cloud: when is it useful? In: *Journal of Information Science* 34 (2007), No. 1, P. 15–29

[Soares 2012] Soares, Sunil: *Big Data Governance: An Emerging Imperative.* MC Press, 2012. – ISBN 1583473777

[Stachowiak 1973] Stachowiak, Herbert: *Allgemeine Modelltheorie.* Wien : Springer, 1973

[Stavrianou et al. 2007] Stavrianou, Anna ; Andritsos, Periklis ; Nicoloyannis, Nicolas: Overview and semantic issues of text mining. In: *ACM SIGMOD Record* 36 (2007), sep, No. 3, P. 23

[Steenbergen and Bos 2010] Steenbergen, Marlies V. ; Bos, Rik: The Design of Focus Area Maturity Models. In: *Global Perspectives on Design Science Research* Vol.. 6105. Berlin, Heidelberg : Springer, 2010, P. 317–332

[Stonebraker et al. 2013] Stonebraker, Michael ; Madden, Sam ; Dubey, Pradeep: Intel "big data" science and technology center vision and execution plan. In: *ACM SIGMOD Record* 42 (2013), No. 1, P. 44–49

[Straub Jr. 1989] Straub Jr., D etmar W.: Validating research instruments in MIS. In: *MIS Quarterly* 13 (1989), No. 2, P. 147–166

[Szulanski 1996] Szulanski, Gabriel: Exploring Internal Stickiness: Impediments to the Transfer of Best Practice Within the Firm. In: *Strategic Management Journal* 17 (1996), No. Winter, P. 27–43

[Tallon 2013] Tallon, Paul P.: Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. In: *Computer* 46 (2013), No. 6, P. 32–38

[Titov and McDonald 2008] Titov, Ivan ; McDonald, Ryan: Modeling online reviews with multi-grain topic models. In: *Proceeding of the 17th international conference on World Wide Web WWW 08*, 2008, P. 1–15

[Trujillo and Maté 2012] Trujillo, Juan ; Maté, Alejandro: Business Intelligence 2.0: A General Overview. In: Aufaure, Marie-Aude (Publisher) ; Zimányi, Esteban (Publisher): *Business Intelligence* Vol.. 96. Berlin, Heidelberg : Springer, 2012, Chap. 5, P. 98–116. – ISBN 978-3-642-27357-5

[Turner et al. 2014] Turner, Vernon ; Gantz, John F. ; Reinsel, David ; Minton, Stephen: The Digital Universe of Opportunities: Rich Data and Increasing Value of the Internet of Things / IDC. 2014. – Research report

[Utterback and Abernathy 1975] Utterback, James M. ; Abernathy, William J.: A dynamic model of process and product innovation. In: *Omega* 3 (1975), No. 6, P. 639–656

[Veldman and Klingenberg 2009] Veldman, Jasper ; Klingenberg, Warse: Applicability of the capability maturity model for engineer-to-order firms. In: *International Journal of Technology Management* 48 (2009), No. 2, P. 219–239

[Venable et al. 2012] Venable, John ; Pries-Heje, Jan ; Baskerville, Richard: A comprehensive framework for evaluation in design science research. In: Peffers, Ken (Publisher) ; Rothenberger, Marcus (Publisher) ; Kuechler, Bill (Publisher):

DESRIST'12 Proceedings of the 7th international conference on Design Science Research in Information Systems: advances in theory and practice. Berlin, Heidelberg : Springer, may 2012 (Lecture Notes in Computer Science), P. 423–438

[Verstralen et al. 2001] Verstralen, Huub ; Bechger, Timo ; Maris, Gunter: *The Combined Use of Classical Test Theory and Item Response Theory*. 2001. – URL `http://www.cito.nl/pok/poc/eindfr.htm`

[Wang and Mori 2009] Wang, Yang ; Mori, Greg: Human action recognition by semilatent topic models. In: *IEEE transactions on pattern analysis and machine intelligence* 31 (2009), No. 10, P. 1762–1774

[Ward and Barker 2013] Ward, Jonathan S. ; Barker, Adam: *Undefined By Data: A Survey of Big Data Definitions*. 2013. – URL `http://arxiv.org/abs/1309.5821`

[Weber and Shahd 2014] Weber, Mathias ; Shahd, Maurice: *Weltmarkt für Big Data wächst rasant*. 2014. – URL `https://www.bitkom.org/Presse/Presseinformation/Weltmarkt-fuer-Big-Data-waechst-rasant.html`. – Date Accessed: 2016-02-24

[Webster and Watson 2002] Webster, Jane ; Watson, Richard T.: Analyzing the past to prepare for the future: Writing a literature review. In: *MIS Quarterly* 26 (2002), No. 2, P. 13–23

[Wendler 2012] Wendler, Roy: The maturity of maturity model research: A systematic mapping study. In: *Information and Software Technology* 54 (2012), No. 12, P. 1317–1339

[Whittington 2003] Whittington, Richard: The Work of Strategizing and Organizing: For a Practice Perspective. In: *Strategic Organization* 1 (2003), No. 1, P. 117–125

[Winter 2008] Winter, Robert: Design science research in Europe. In: *European Journal of Information Systems* 17 (2008), No. 5, P. 470–475

[Xu et al. 2012] Xu, Weijia ; Luo, Wei ; Woodward, Nicholas: Analysis and Optimization of Data Import with Hadoop. In: *26th International Parallel and Distributed Processing Symposium*, 2012, P. 1058–1066

[Yousuf 2007] Yousuf, Muhammad I.: Using experts' opinions through Delphi technique. In: *Practical Assessment, Research & Evaluation* 12 (2007), No. 4

[Zhang 2013] Zhang, Du: Inconsistencies in big data. In: *2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, 2013, P. 61–67

[Zhang et al. 2012] Zhang, Xuyun ; Liu, Chang ; Nepal, S ; Dou, Wanchun ; Chen, Jinjun: Privacy-Preserving Layer over MapReduce on Cloud. In: *Second International Conference on Cloud and Green Computing*, 2012, P. 304–310

[Zhe and Zhi-gang 2010] Zhe, Liu ; Zhi-gang, Zhao: An Algorithm of Detection Duplicate Information Based on Segment. In: *2010 International Conference on Computational Aspects of Social Networks*, IEEE, sep 2010, P. 156–159

[Zhu and Li 2012] Zhu, Qing ; Li, Ning: Privacy Protecting by Multiattribute Clustering in Data-Intensive Service. In: *11th international Conference on Trust, Security and Privacy in Computing and Communications*, 2012, P. 1273–1278

[Ziegler and Dittrich 2007] Ziegler, Patrick ; Dittrich, Klaus R.: Data integration - problems, approaches, and perspectives. In: *Conceptual Modelling in Information Systems Engineering*. Springer Berlin Heidelberg, 2007, P. 39–58

[Zikipoulos et al. 2013] Zikipoulos, Paul ; Deroos, Dirk ; Parasuraman, Krishnan ; Deusch, Thomas ; Corrigan, David ; Giles, James: *Harness the Power of Big Data*. New York, New York, USA : McGrawHill, 2013. – ISBN 9788578110796

[Zimmer et al. 2012] Zimmer, Michael ; Baars, Henning ; Kemper, Hans G.: The impact of agility requirements on Business Intelligence architectures. In: *45th Hawaii International Conference on System Science (HICSS)*, 2012, P. 4189–4198