Deployment of a Descriptive Big Data Model

Marco Pospiech and Carsten Felden

Abstract Big Data is an emerging research topic. The term remains fuzzy and jeopardizes to become an umbrella term. Straight forward investigations are inhibited since the research field is not well defined, yet. To identify a common understanding, experts have been interviewed. Hereby, the findings are coded and conceptualized until a descriptive Big Data model is developed by using Grounded Theory. This provides the basis for the model's deployment. Here, academic publications and practical implementations marked as Big Data are classified. It becomes evident that Big Data is use-case driven and forms an interdisciplinary research field. Even not all papers belong to this research field. The findings become confirmed by the practical implementations. The chapter contributes to the intensive discussion about the term Big Data in illustrating the underlying area of discourse. A classification to set the research area apart from others can be achieved to support a goal oriented research in future.

Keywords Big Data • Grounded theory • Interview • Qualitative research

1 Introduction

The amount of various business data is growing exponential while their accompanied storing and processing in a traditional way makes a task fulfillment complicated. In this context, the term Big Data occurs increasingly in scientific discussions and publications [1]. A common definition belongs to Gartner: "*Big data* is high-volume, high-velocity, and high-variety information assets that demand cost-effective innovative forms of information processing for enhanced insight and decision making" [2], but considering others, existing definitions are diverse. For example, Bizer et al. [3] follow a semantic perspective and whereby others focusing on processing huge amounts of data [4]. As a consequence, two major issues are conspicuous. On the one hand side, the positioning of goal oriented publications to solve business administration related problems within a Big Data context is

M. Pospiech (✉) • C. Felden
Technische Universität Bergakademie, Freiberg, Germany
e-mail: marco.pospiech@bwl.tu-freiberg.de; carsten.felden@bwl.tu-freiberg.de

J.H. Mayer, R. Quick (eds.), *Business Intelligence for New-Generation Managers*,
DOI 10.1007/978-3-319-15696-5_7

impeded [1]. On the other hand side, already established research areas such as High Performance Computing (HPC) tend to recoin themselves as Big Data to obtain more attention [5]. Thus, the possible research field of Big Data is fuzzy and the term seems to be only used as marketing buzzword. So, the demand of a theoretical base has already been stated since relevant topics and related theories are unknown [1]. In this context, the work isolates the inherent nature of the Big Data topic and to deduce a descriptive model as first step. It is the chapter 's goal, to illustrate the applicability and deployment for practice and research. Thus, research areas are stronger defined to set Big Data apart from and business administrated issues can be addressed.

A quantitative proof of the fuzzy term Big Data seems to be not possible since underlying factors are unknown. In addition, no research has been published, yet, which addresses a theoretical reprocessing of this topic. It is helpful that already Miles and Huberman [6] indicate that qualitative methods are useful in an early state of research to gain a professional perspective based on longstanding experience. In this context, we accomplished expert interviews to gain the consent understanding [7]. Results are coded and conceptualized by using Grounded Theory [8, 9]. A Big Data description model is derived that represents expert opinions. The resulting categories can be used as patterns for future Big Data research and practical implementations to define the research area apart from others. Hereby, the chapter contributes to the intensive discussion about the term Big Data in illustrating the underlying area of discourse.

The papers's concourse is as follows: Sect. 2 provides the research design. Hereby, expert interviews and Grounded Theory as method are briefly shown to be able to create an appropriate basis. The proceeding of methods is illustrated, thus, a rigor research design can be ensured. Section 3 provides the descriptive model. Single categorizes are presented and discussed. The model is used in Sect. 4. Hereby, several Big Data papers and implementations are classified whether they belong to Big Data or not. So the model application and its term specification of Big Data get evident. The chapter is summarized in Sect. 5, where implications and further research topics are highlighted.

2 Research Design

The research design is subdivided into four stages (see Fig. 1). After stating the research goal and problem formulation, we conduct a literature review [10] in a second step to arrange the existing knowledge about Big Data models. Academic databases like *IEEE xplore*, *AIS library*, *ScienceDirect*, and *EBSCOhost Web* were used and relevant papers were identified through the search items *Big Data Theory* and *Big Data Model* in keywords, title, and abstract. In fact, no article was identified. Thus, no published research exists that addresses a specification of the term Big Data through a theoretical approach. Due to this reason, we decided to use expert interviews to obtain initial insights. Expert interviews are based on different phases.

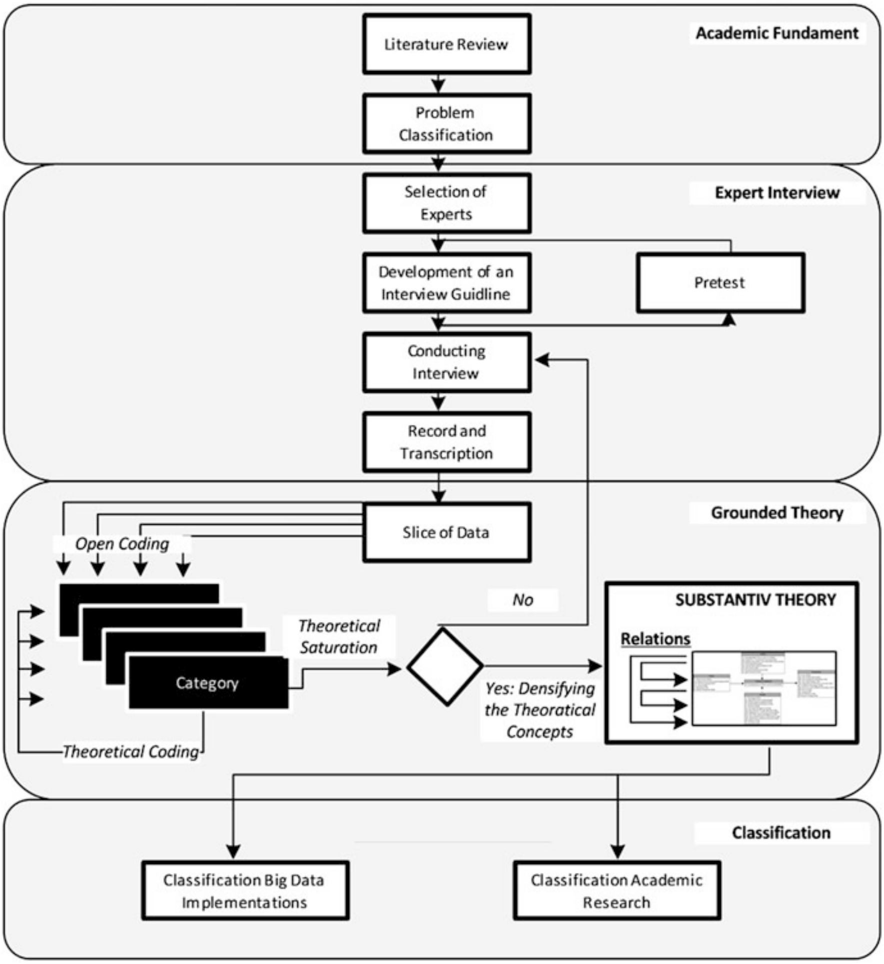

Fig. 1 Research design (Based on [8, 9])

They address the problem and eligible experts as well as a guideline based interview conducting and evaluation [7]. Within the execution, 20 experts from international acting IT companies were interviewed. To ensure the expert's competence, a minimum of 1 year Big Data experience was assumed.

The telephone interviews varied between 30 and 60 min and were conducted between July and December 2012. Comparability of the results is ensured through interview guidelines. Thereby, pretests were conducted. As a result, characteristics, key drivers, definitions, distinctions, potentials, use-cases, issues, strategies of Big Data, and expert statistics were gathered. All interviews were recorded and transcribed.

Grounded Theory [8] collects and analyzes qualitative data (like expert interviews) and is accepted and widespread in information systems research [11].

The method aims the generation of conceptual properties, categories, and relationships through a combination of inductive, deductive, and abductive reasoning in iterative cycles. The usage in an early research status to describe a phenomenon is often applied [12]. Big Data belongs to a relative new research area without existing theoretical fundament, yet. Thus, Grounded Theory is used here since the generation and discovering of concepts and inherent relationships relies to the strengths of the method [9].

The analysis in Grounded Theory consists of four steps [9]. Within the open coding, so called *slices of data* (transcribed interviews) are broken down into several categories. These categories are described in terms of their properties. Thereby, several categories were assigned to the transcribed interview passages, thus similar statements get apparent. This is followed by the theoretical coding, where relations between categories are established. Thereby, new *slices of data* are added and categorized until a theoretical saturation arises and no novel category emerges. Three researches were involved during the model development until all agreed in a theoretical saturation. The result is a substantive approach, which is applicable to the emerging particular area. It is recommended to use existing coding schemata within the theoretical coding, because the negation of them leads usually to confused and unclear theoretically coding [9]. An often used one belongs to [13], where the phenomenon (Big Data) belongs to the center. The original reasons of emergence, the context, possible strategies to face the phenomenon, and consequences are put into relation. To allow a common understanding, Strauss and Corbin [13] defined the categories as follows:

- Phenomenon: The central idea, event, happening, incident about which a set of actions or interactions are directed at managing, handling, or to which the set of actions is related.
- Causal Conditions: The term refers to the events or incidents that lead to the occurrence or development of a phenomenon. Also look at its specific properties and dimensions, which in fact should be explainable by looking back at the specific dimensions of the causal conditions.
- Context: Represents the particular set of conditions within which the action/interactional strategies are taken to manage, handle, carry out, and respond to a specific phenomenon.
- Strategy: Action devised to manage, handle, carry out, and respond to a phenomenon under a specific set of perceived conditions and context. Actions have certain properties. First, it is processual, evolving in nature. Thus it can be studied in terms of sequences, or in term of movement, or change over time. Second, action/interaction is purposeful, goal oriented, done for some reason—in response to or to manage a phenomenon.
- Consequences: Outcomes or results of action and interaction. Consequences may be actual or potential, happen in the present or in the future.

As a result, all identified categories were shrunken until a substantive model emerged. Exemplary scientific Big Data publications and practical implementations marked as Big Data are classified to illustrate how the model supports the academic and practical sharpening of the term.

3 Descriptive Model

The five categories being used and their subcategories are described in this chapter and discussed. Numbers in brackets belongs to the amount of experts stating a category. E.g., (10/20) mean out of 20 probates, 10 mentioned this category. Thus, the importance of a component is illustrated. This supports the verification whether a Big Data paper belongs to the research area.

3.1 Phenomenon

In general, Big Data can be seen as such a *phenomenon* described above and emerges through *context* and *causal conditions*. This gets apparent since popular terms volume, velocity, and variety [2] are implied. This leads to the point that Big Data can be understood as circumstance, in which an increased data volume must be processed or/and stored. In fact, *strategies* to face this *phenomenon* will change in time. But they can always be subdivided into a functional and technology part. The *consequences* are resulting from all schema categories and are issue and advantage related. Considering this proposed approach, all mentioned categories are not novel. They are representing own and nowadays mostly well researched disciplines, which emerged by their own causal conditions and context. Based on gained experiences and use-cases of this single disciplines in past emerged the *phenomenon* Big Data. It applies that a combination of the underlying disciplines achieves a higher value as a single discipline can obtain for its own. The combination of these concepts will only be useful, if a specific idea exists, which produces a positive value. Since the underlying disciplines are data-intensive the *phenomenon* Big Data is into place. In this context, Big Data represents an interdisciplinary research area and is use-case driven. Figure 2 shows the describing Big Data model.

3.2 Causal Condition

In fact, most *causal conditions* are well-known and defining Big Data not by itself. The requirements can be summarized as a need for an extensive environmental understanding. Experts stated a particular demand on monitoring, prediction, and decision support as well as a need to explain circumstances in natural sciences. New possibilities are seen through *context* and *strategy,* which enable the achievement of an improved understanding as before. In a dynamic world, this knowledge must be

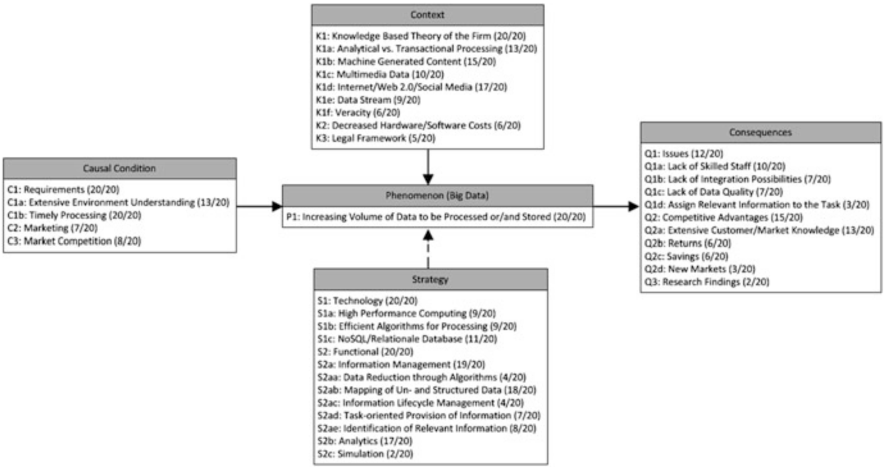
Context
K1: Knowledge Based Theory of the Firm (20/20)
K1a: Analytical vs. Transactional Processing (13/20)
K1b: Machine Generated Content (15/20)
K1c: Multimedia Data (10/20)
K1d: Internet/Web 2.0/Social Media (17/20)
K1e: Data Stream (9/20)
K1f: Veracity (6/20)
K2: Decreased Hardware/Software Costs (6/20)
K3: Legal Framework (5/20)
Causal Condition
C1: Requirements (20/20)
C1a: Extensive Environment Understanding (13/20)
C1b: Timely Processing (20/20)
C2: Marketing (7/20)
C3: Market Competition (8/20)
Phenomenon (Big Data)
P1: Increasing Volume of Data to be Processed or/and Stored (20/20)
Consequences
Q1: Issues (12/20)
Q1a: Lack of Skilled Staff (10/20)
Q1b: Lack of Integration Possibilities (7/20)
Q1c: Lack of Data Quality (7/20)
Q1d: Assign Relevant Information to the Task (3/20)
Q2: Competitive Advantages (15/20)
Q2a: Extensive Customer/Market Knowledge (13/20)
Q2b: Returns (6/20)
Q2c: Savings (6/20)
Q2d: New Markets (3/20)
Q3: Research Findings (2/20)
Strategy
S1: Technology (20/20)
S1a: High Performance Computing (9/20)
S1b: Efficient Algorithms for Processing (9/20)
S1c: NoSQL/Relationale Database (11/20)
S2: Functional (20/20)
S2a: Information Management (19/20)
S2aa: Data Reduction through Algorithms (4/20)
S2ab: Mapping of Un- and Structured Data (18/20)
S2ac: Information Lifecycle Management (4/20)
S2ad: Task-oriented Provision of Information (7/20)
S2ae: Identification of Relevant Information (8/20)
S2b: Analytics (17/20)
S2c: Simulation (2/20)

Fig. 2 Big Data descriptive model

increasingly produced in a timely manner [14]. Hereby, time adequate processing belongs as one part to velocity [2]. Concepts like decision latency or automatization [15] represent core aspects and are often mentioned. Hence, information technology tasks must accomplish user needs in time. Besides requirements are two more causes.

One often applied aspect belongs to marketing. Hereby, experts stated the possibility that Big Data is not novel and only driven by sales persons to maximum their revenue. In this case, Big Data is no new research area and should be ignored by academics. Comparable statements can be found for Business Intelligence (BI), which counts today as one of the most important fields for practice and academic in information systems [16]. This leads to the point that further work has to proof, whether Big Data is solely marketing driven or not.

Another causal condition belongs to market competition and is therefore a drivig force which is known in lots of technology related projects, e.g. Customer Relationship Management (CRM) or Efficient Consumer Response (ECR). It is slightly related to the requirements since it represents the source of them. In dynamic markets, companies have to reduce production cycles, safe costs, identify early trends, react fast, and maximize their profit in a sustainable way. Similar causal conditions can be found for BI in Gluchowski et al. In past, BI was mostly driven by the need of a timely processing and an extended environment understanding to survive in a competitive market. Even a discussion whether BI belongs to a buzzword can be obtained [14]. Thus, the *causal conditions* are common, alone. A more unique Big Data field emerges trough the combination of *context* and *causal conditions*.

3.3 Context

Context describes the circumstances in which Big Data evolved. During our coding, we recognized that most of all context related expert statements were already shown the Knowledge Based Theory (KBT) of the Firm [17]. Here, experts stated: "For the purpose of decision support, different data sources and structures must be consolidated to achieve a most extensive acquisition of information" or "Formats, structures and sources must be integrated to acquire competitive advantages" or "The novelty of Big Data consists besides huge data volume within the usage of heterogeneous sources and formats". In this work, KPT is rather used as concept as a theory and should not interpreted as a mapping mechanism to Grounded Theory. Thus, the KBT of the Firm considers knowledge as unique and most strategically significant resource [17] by focusing on knowledge integration and combination to achieve a competitive advantage [18]. According to [19] information technologies are able to support firms within the KBT since it can be used to synthesize, enhance, and expedite large-scale intra- and inter-firm knowledge management. In this context, an increasing trend is noticeable that integrates various data sources, structures and formats to obtain a competitive advantage. On a practical view, this is also known as variety [2].

However, the KBT arises also through underlying reasons. An expert stated an increasing awareness of analyzing transaction data. Hereby, transactions are no longer the fundamental basis of analyses as we understand it in BI [14] but rather an explanatory variable. Hereby, the way why a transaction appears (e.g. interaction data analysis of Social Media, forum, web shop, GPS tracking, etc. of a customer until the purchase) comes more and more into focus to gain a more extensive environment view.

Another subcategory is seen in machine generated content. It relies to any data consisting discrete events that have been created automatically from a computer application, process or other machines in absents of any human intervention [21]. In this context, experts stated the increasing IT pervasion. Thus, research and practice find themselves in a world of radio-frequency identification (RFID) chips, log files, internet of thinks, GPS tracking (vehicles and mobile phones), and any kind of sensor technology that can be used to gain an extensive environment understanding. In most of the cases, this type of data occurs as stream.

In streaming, a continuous flow of data must be handled. This data relies often to semi-structured time series and implies issues of storing and processing since the permanent transfer produces an enormous data volume. In contrast to static data, only a subset of data can be considered [22]. Experts stated that in most of the cases streaming data must be analyzed in real time. In addition, science application produces stream data where the storing of all features is not possible.

Other key drivers are seen in Internet, Web 2.0, and Social Media. Most of the Big Data *phenomena* occur through the Internet. One aspect belongs to the physical infrastructure, which enables a world spanning interconnection through all participants and consequently an exchange of information and data. Only through this infrastructure, the realization of concepts like cloud, grid, or distributed computing got possible. In addition, it acts as immense information storage for every participant. Besides that, experts stated the fact that the internet evolved as interactive platform where people collaborate and share information, also known as Web 2.0 [23]. Within this movement companies like Google, Yahoo!, Twitter, eBay, Amazon, or Facebook emerged. The rapidly growing content and users forced the companies to develop cost effective and scalable technologies. Here, technologies like MapReduce or Hadoop emerged and are now patterns for further applications [24]. Besides that, Web 2.0 enabled the extensive creation and exchange of various user generated content. User generated content belongs to contributed data, information or media not created by the provider of the web service itself. Examples are hotel ratings, wikis, or videos. This kind of information has rapidly grown during the last years and provides an interesting analysis fundament [20]. One of the most important applications of Web 2.0 belongs to the Social Media [25]. In this context, most of the experts stated that the step into Social Media enables the most exciting possibility to gain a more advanced view about customer and environment as before.

Another aspect belongs to the data type itself. While in past, only structured data went into the analyzed focus, possibilities raised to interpret and gain information automatically from multimedia sources nowadays (image, video, audio, and text). Thus, human generated content gets machine-readable. Prominent examples are seen in text mining or video and image sentiment analysis.

This new types of data and sources implicate another *context*. Whereby traditional data was clean and precise the new ones are rather fuzzy. Text mining or sentiment analyses are always a translation from human content to machine-readable data. During this procedure, information gets lost and diffuse. Even a tracking system with data from vehicle and mobile devices does not work precisely. In addition, combining unknown sources means also unknown data quality. This aspect is already known as veracity [26], but applications have to deal with these uncertainties.

The increasing improvements in storage and processing technologies are also stated as one reason for Big Data. Especially, associated cost savings enables the implementation of scalable systems. Thus, companies and research organizations can fulfill their requirements within the budget limitation. Another aspect belongs to the open source movement. Thus, the generation of content within the internet remains in most of the cases free of cost. In addition, key Big Data technologies are usually open source tools. Here, organizations observing saving possibilities through projects like Hadoop or MangoDB.

In addition, the legal framework represents a relevant *context*. Experts stated that besides all possibilities of Big Data most of all future applications are situated between data privacy restrictions. Thus, it is not legal to combine different sources to gain a broader view over customer and environment, always. Legislation depends on the country and must be proofed in detail. Furthermore, users are developing awareness of data privacy. As a result, countries have to define a framework where users rights are guaranteed and new Big Data concepts are enabled.

3.4 Strategy

As already stated, the inherent nature of Big Data is not defined by *strategy* (shown as punctuate arrow). *Strategy* is necessary to overcome the *phenomenon* but is not an essential component. Hereby, *strategy* is stretched through technology and functional aspects. Technology addresses the *phenomenon* (caused by *context* and *causal conditions*) and is not limited in future. Experts agree in three concepts.

HPC belongs to intensive calculation or storing tasks. The processing of those jobs through personal computer is not appropriate. Thus, concepts like cloud, grid, distributed, and parallel computing are implied and mentioned. The nature of these concepts allows the scaling of needed hardware to fulfill the task in an appropriate manner [27].

Besides high performance hardware, efficient working algorithms were also mentioned. Thus, the time and computation complexity of traditional algorithms suffers. All steps of processing, storing, and analyzing are time relevant in Big Data and need to be supported. One possible solution refers to [28]. In addition, parallel programming models took place into Big Data technologies. Here, Googles MapReduce was often mentioned, where the framework specifies map and reduce functions to execute them in a parallel way [29].

An important part of technologies belongs to the database discussion. Some experts argue relational databases are not able to meet the requirements due the lack of scalability and query performance, where others allude proceedings in relational databases. Often stated NoSQL approaches are key-value, column, or document oriented databases. Besides the discussion between NoSQL and relational databases, the movement to in-memory technologies was also often mentioned. Nevertheless, this kind of technologies enables the computation and storing of high volume data in a proper time. The processing of heterogeneous data sources cannot be solved by just technology.

Thus, the second *strategy* belongs to functional concept. Expert believes that the Big Data *phenomenon* can be tackled through organizational approaches. Hereby, information management (IM) is mentioned. IM aims the efficient utilization of information in respect of the organizational goal. It manages information systems, overarching executive functions and information and communication technologies as well as the handling of information in an economic way. The last one belongs to information logistic, information lifecycle, information demand, information source, information quality, information provision and usages management [30]. During the interviews experts stated several Big Data related aspects to the information management. Hereby, already common approaches like BI were mentioned. But, not all concepts are transferable to Big Data. In future, existing methods must be proofed of Big Data suitability. If not, new methods will emerge.

One core concept belongs to the mapping of structured and unstructured (image, text, audio, video) data. In this context, Big Data application designers must combine various data sources in a useful manner to gain a broader view about reality. Nevertheless, available concepts are rare to guide those activities [1]. Industry experts have to figure out what kind of data sources can support the target activity. In addition, semantic concepts are seen as possible benefit within the combination of various sources.

Another aspect is seen in the task-oriented provision of information. This aspect shares similar aspects to information logistic. Both address the provision of the right information, at the right time, in the right amount and place in an adequate quality. Thereby, information logistics consider only the data flow. The flow of unstructured data is often uncertain. In addition, information logistic neglects the value of information itself [30].

Information lifecycle management is also seen as one possible aspect within Big Data *strategies*. It aims the creation of an equal status between information demand and supply. Hereby, an update of both in an iterative cycle is necessary. Especially within Big Data information management is important, because various data sources and information technology are used whereby quality has to be ensured. Besides that, storage aspects have to be considered. Thus, most of the information loses value in time and a continuous storing remains questionable.

The identification of relevant information can support this aspect. This can be done through management methods as well as through machine learning based techniques. Experts stated that task relevant information must be automatically declared within data sources. This is essential, especially in Big Data.

Another field belongs to the data reduction. As in information identification this can be done through management or mathematical methods. Thereby, any kind of technique that allows a more efficient way of handling data without losing relevant information has to be considered. Possible works can be seen in [31].

17 of 20 experts labeled the analyses of Big Data as core concept. Through the stated *context* and *causal condition* several kinds of analysis techniques evolved and are now considered in Big Data. The most mentioned method belongs to the Knowledge Discovery in Database (KDD) also referred to as Data Mining. It represents the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [32]. There exist several mentioned subcategories of Data Mining. Text Mining transfers unstructured textual data into a machine-readable content, which is subsequently used by machine based learning techniques [33]. Another one belongs to Web Mining. This can be categorized into content, structure, and usage mining [34]. Similar approaches were stated for image, audio and video analyzes. A typical example can be seen by [35]. Even the growing Social Media brings new practices along. One often stated example is seen in social network analyses, where pattern of relationships and interactions between social actors are analyzed to discover underlying structures [36]. Predictive Analytics was often mentioned, which addresses the predictions of the future itself [37]. Certain participants remarked that the nature of Big Data prevents modeling and cleansing. Here, the data must be processed on-the-fly. Another analyses focus

is seen within the detection of relationships between different data sources. Thus, analysts are supported in hypothesizing. Besides the stated analyze concepts, experts highlighted the advanced analysis character of Big Data. Classical Online Analytical Processing (OLAP) cubes as known from BI are not included. Few experts stated the application of advanced visualization techniques. Possible examples of high volume presentation can be obtained from [38].

The final strategy belongs to simulation. Just some experts quoted this fact since it is more seen from an academic perspective. Nevertheless, in most of the cases, simulations handle voluminous data of various inputs [16]. Hereby, the processing time must be acceptable. It is defined as process, which drives a system model with suitable inputs an observing the resultant outputs, whereby simulation one process by another [39]. Examples for Big Data aided work is stated in [40].

3.5 Consequences

According to [13], a *phenomenon* always causes *consequences.* In context of Big Data, they are divided into issues, competitive advantages and research findings. The most stated problem belongs to missing Big Data experts, both technicians as well as functional oriented people. Thus, only a handful of specialists are experienced in new technologies like Hadoop or MapReduce, whereby the demand is huge. Here, universities must adjust their study programs [41]. The requirements on functional experts are enormous. Experts stated that more analytical and domain specific knowledge is necessary as well as an understanding for underlying technical processes to be able to deal with this phenomenon. Especially the functional integration of various data sources is seen as complex, but a technical integration is critical. Within the *phenomenon* data is stored in divers NoSQL approaches, most of the time schema-less. Hereby, the integration of traditional concepts remains difficult. This is aggravated by the fact data is received as stream and processed on-the-fly.

Considering veracity, a huge data quality issue arises. This affects the quality of analytical and simulation results since the kind of data remains fuzzy. Other experts stated quality regarded problems, because data sources are often obtained from third parties. Last but not least, NoSQL is mostly inconsistent.

Another issue is seen in assigning relevant information to the task. Thereby, the spectrum of possible Big Data sources is broad and not any content achieves an advanced understanding of the environment. Hence, functional experts must be supported by organisational as well as by technical methods in order to discovery valuable relationships and items within data sources. So investigations can be performed in a goal oriented way.

The second *consequence* belongs to the competitive advantages. There are closely related to the discussed *causal condition*. Thus, the most important fact is seen in gaining an extensive view about customer and market. Hereby, functional experts are supported through analytics and simulations as well as through organisational methods. Driven by the illustrated *context,* the expectation of new business ideas and markets emerged. Here, experts imagine the combination of various strategies since research enable it at the moment. In addition, the achievement of unsuccessful ambitious aims was also stated. At least, Big Data is not seen as end in itself. Financial advantages must appear. Savings can be obtained by the usage of open source, saved memory, and computational capacity as well as through an increased automatization. Depending on the analytical/simulation focus savings and returns are possible to achieve. This goes along with the KBF theory. Both, users as well as consultants stated this fact.

The final *consequence* addresses the academic application affected by Big Data. Thus, the context and current available strategies allows the exploration of new knowledge. New relationships and insights, especially within natural science, were quoted. Efficient algorithms, simulations, and analytics are highlighted to gain a broader view about the framing environment. Due to the reason that the expert's common background is business, only two experts stated these concepts.

4 Classification of Academic Publications and Implementations

We developed a descriptive model to discover Big Data's inherent nature. The research area of Big Data is interdisciplinary and belongs to the need of storing and processing a huge amount of data, driven by *context* and *causal conditions*. Around the emerging phenomenon various research disciplines has been positioned. Thus, Big Data belongs to applied sciences. Progress in the undergraduate sciences is welcome, but not part of the research field of Big Data itself, e.g. text mining, HPC, etc. To demonstrate the academic model application, we have chosen several papers out of a Big Data state-of-the-art publication randomly [1] and have classified them whether they belong to Big Data or not. The results are illustrated in Table 1. Numbers within the tables are related to the concept abbreviations in Fig. 2. State-of-the-art papers are excluded due to their informing character. Since not each single concept must be fulfilled to be Big Data, we left place for a small discussion part, which should be considered in any further model application. The amount of mentioned times enables a statement how strong the field is related to Big Data.

As illustrated in Table 1, the first three publications belong to Big Data. Hereby most of the concepts are fulfilled. In addition, not all papers [1] are Big Data. The fourth [45] remains to an application and typical strategies like MapReduce were used, but a combination of various data sources is neglected. Thus, the paper is more related to HPC. Furthermore [46], is not focused on a use-case. So the paper supports Big Data research, but relies not on the research area itself.

Similar to Table 1, five practical implementations marked as Big Data are compared and classified in Table 2. The cases are obtained from the leading association for the U.S. technology industry, TechAmerica. Among other aspects,

Table 1 Big Data descriptive model academic classification

Source	Causal condition	Context	Phenomenon	Strategy	Consequence	Big Data	Discussion
[42]	C1; C1a; C1b;	K1; K1c; K1d; K1e; K1f, K2	P1	S1; S1a: S1b; S1c; S2 S2a; S2aa; S2ab; S2b	Q1; Q1c; Q2 Q2c; Q3	Yes	The paper presents a pluggable prototype system to analyze and integrate high volume biometric data (fingerprints, iris image, facial image, voice, DNA) in real time. The system is cloud based underlying systems are HBase and ZooKeeper. Various analyses and deduplication occurs
[43]	C1a; C1b	K1; K1b; K1c; K1d; K1f	P1	S2; S2ae; S2ab; S2b; S2c	Q2; Q2a; Q2b	Yes	Web forum discussions are used to predict stock returns. Here structured data like market return, volatility, or trading volume are combined with unstructured text. Using stepwise regression and sentiment analysis
[44]	C1; C1a; C1b	K1; K1d; K1e; K1f	P1	S2; S2ab; S2b	Q1; Q1c; Q3	Yes	Enormous amount of behavioral data is explored to identify spatiotemporal dynamics of criminal events with the hope of identifying patterns in their aggregation that may be useful to predict and prevent future crimes. Hereby, relationships in both space and time, cross- and auto-correlation measures are combined
[45]	C1; C1a; C1b	K1; K1d; K2		S1: S1a; S1b; S2; S2b	Q2; Q2a; Q2b	No	The article designs a MapReduce based computing model for an option price prediction. Only price data was used as input. The combination of various data sources took no place. Thus, no Big Data publications
[46]	C1; C1b	K1; K1d; K2		S1; S1c		No	The paper proposes the use of single level data stores. Thereby, consistent and durable data structures, that on current hardware, allow programmers to safely exploit the low-latency and non-volatile aspects of new memory technologies are introduced. No application is illustrated only view concepts are fulfilled

Table 2 Big Data descriptive model implementation classification

Implementation	Causal condition	Context	Phenomenon	Strategy	Consequence	Big Data	Discussion
NARA's electronic records archive (ERA)	C1; C1b	K1; K1c; K3		S1; S1c; S2; S2a; S2ad; S2ae	Q1; Q1c; Q1d	No	ERA is designed to archive a variety of records of the U.S. government and worldwide archives. In January 2012, ERA manages about 142 TB of information representing over 7 billion objects (e-mails, images, records, aso.). Several archival and search functions were implemented. The main focus refers rather to store than to process Big Data. No analytical functions are realized. A mapping between un- and structured data does not occur
Wind energy turbine placement & maintenance	C1; C1a; C1b; C3	K1; K1b; K1c; K1e; K1f	P1	S1; S1a; S1b; S1c; S2; S2a; S2aa; S2ab; S2ad; S2ae; S2b; S2c	Q1; Q1c; Q1d; Q2; Q2a; Q2b; Q2c; Q2d; Q3	Yes	Vestas installs an average of one wind turbine every 3 h. Producing wind energy depends greatly on the placement of the turbine. Vestas established a wind placement library, which incorporates data from global weather systems with data collected from existing wind turbines, satellite images, and geographical information, more than 178 parameters in total. Analysis and simulation functions for turbine placement and power production were implemented

TerraEchos perimeter intrusion detection	C1; C1a; C1b	K1; K1a; K1c; K1e; K1f	P1	S1; S1a; S1b; S2; S2aa; S2ab; S2ad; S2b;	Q1; Q1c; Q1d; Q2	Yes	TerraEchos helps organizations protect and monitor critical infrastructure and secure borders. Distinguishing the sound of a whisper from the wind from miles away is a big challenge in order to identify potential risks. The solution continuously consumes and analyzes massive amounts of information-in-motion through HPC. In addition, the system gathers and analyzes information in real-time, maps acoustics and video data
NASA's Human spaceflight imagery collection	C1; C1a	K1; K1c		S2; S2a; S2ac	Q1; Q1b; Q1c; Q3	No	The IRD is responsible for managing a large, complex heterogeneous cyber infrastructure and one of the world's largest imagery archives and provides industry and public with the most historic human spaceflight imagery. The NASA imagery were catalogued and archived in structured and unstructured format. In addition, information lifecycle management was introduced. The main focus refers rather to storing than to processing Big Data. No analytical capabilities were implemented. A mapping between un- and structured data does not occur

(continued)

Table 2 (continued)

Implementation	Causal condition	Context	Phenomenon	Strategy	Consequence	Big Data	Discussion
National weather service (NWS)	C1; C1a; C1b	K1; K1a; K1b; K1c; K1e; K1f; K2	P1	S1; S1a; S1b; S1c; S2; S2a; S2aa; S2ab; S2ac; S2ad; S2ae; S2b; S2c	Q1; Q1c; Q3	Yes	The NWS provides weather, water, and climate data, forecast analyses and real-time warning for the protection of life and property and enhancement of the national economy. Data from satellites, ships, aircrafts, buoys and other sensors are considered. It is used as input for millions of operational models in HPC environments

the foundation analyzes the size and scope of technologies industry-impact on the economy. Thereby, more than 1.200 companies are represented by this organization including SAP, Cisco, Motorola, Apple, SAS and Amazon [47]. Three of five implementations belong to Big Data. Most of the categories are consistent with the findings. In our view, two of the implementations represent archiving cases and are not Big Data. Here, solely the data's volume is seen as major characteristic. Neither a mapping between structured and un-structured data occurs nor a timely processing. Surprisingly, most of the cases did not consider social media and technologies like Hadoop. Information management was always part of the implementation.

5 Conclusion

The chapter addressed the stated issue of a missing descriptive fundament of Big Data to prevent the establishment of a new buzzword. Hereby, expert interviews were conducted, transcribed and used as basis for a Grounded Theory design to obtain the inherent nature. Hereby, emerging concepts where categorized into *causal condition, context, phenomenon, strategy* and *consequences* until a theoretical saturation was achieved. As a result, Big Data is interdisciplinary, *context* and *causal conditions* driven and belongs to applied sciences. The usefulness of the resulting Big Data model was exemplarily tested. In this context, state-of the-art papers [1] and implementations were classified into the model and tested, whether they belong to Big Data or not. In fact, not all of them belong to this topic. Surprisingly, common topics like Hadoop or Social Media are not always part of Big Data implementations.

The resulting model contributes to the specification and classification of Big Data contents. Thus, the area of discourse is illustrated and the positioning of goal oriented publications to solve business administration related issues is assisted. So, Big Data is delimitated from existing research fields such as HPC. Hereby, the insights justify a stronger positioning of Big Data as own research field to motivate further research. Nevertheless, there are limitations. In fact, the descriptive model was an initial step to contribute to the existing discussion and offering a possible consensus about Big Data. Based on the proposed model, a quantitative analysis will follow to clarify if all categories and relationships are significant. In this context, it has to be proven, whether Big Data is marketing driven or not. Another gap arises through the expert selection. Thus, only IT companies were considered. In fact, besides business Big Data applications are also stated in science [10]. In this context, academics must be interviewed as well to gain the opportunity that new concepts might appear.

References

1. Pospiech M, Felden C (2012) Big data – a state-of-the-art. In: Proceedings of AMCIS 2012, pp 1–11
2. Gartner Inc (2013) IT glossary big data. http://www.gartner.com/it-glossary/big-data. Accessed 28 June 2013
3. Bizer C, Boncz P, Brodie M (2011) The meaningful use of big data: four perspectives. SIGMOD 40(4):56–60
4. He Y, Lee R, Huai Y (2011) RCFile: a fast and space-efficient data placement structure in MapReduce-based warehouse systems. In: Proceedings of ICDE 2011, pp 1199–1208
5. Simmhan Y, Barga R, Heasley J (2009) GrayWulf: scalable software architecture for data intensive computing. In: Proceedings of HICSS 2009, pp 1–10
6. Miles M, Huberman A (1994) Qualitative data analysis. Sage, Thousand Oaks
7. Flick U (2009) An introduction to qualitative research. Sage, London
8. Glaser B, Strauss A (1967) The discovery of grounded theory. Aldine Transaction, Chicago
9. Glaser B (1978) Theoretical sensitivity: advances in the methodology of grounded theory. Sociology Press, Mill Valley
10. Cooper H (1998) Synthesizing research: a guide for literature reviews. Sage, Thousand Oaks
11. Hughes J, Jones S (2003) Reflections on the use of grounded theory in interpretive information systems research. In: Proceedings of ECIS 2003, paper 62
12. Hughes J, Wood-Harper T (1999) Systems development as a research act. J Inf Technol 14 (1):83–94
13. Strauss A, Corbin J (1990) Basics of qualitative research: grounded theory procedures and techniques. Sage, Thousand Oaks
14. Gluchowski P, Gabriel R, Dittmar C (2008) Management support systeme und business intelligence. Springer, Berlin
15. Hackathorn R (2012) Current practices in active data warehousing. DM review, white paper
16. Chen H, Chiang R, Storey V (2012) Business intelligence and analytics: from big data to big impact. MIS Q 36(4):1165–1188
17. Grant R (1996) Prospering in dynamically-competitive environments: organizational capability as knowledge integration. Organ Sci 7(4):375–387
18. Barney J, Wright M, David J, Ketchen J (2001) The resource-based view of the firm: ten years after 1991. J Manag 27(6):625–641
19. Alavi M, Leidner D (2001) Review: knowledge management and knowledge management systems. MIS Q 25(1):107–136
20. Krumm J, Davies N, Narayanaswami C (2008) User-generated content. Pervasive Comput 7 (4):10–11
21. Monash C (2010) Three broad categories of data. http://www.dbms2.com/2010/01/17/three-broad-categories-of-data. Accessed 8 July 2013
22. Bitincka L, Ganapathi A, Zhang S (2012) Experiences with workload management in Splunk. In: Proceedings of MBDS 2012, pp 25–30
23. DiNucci D (1999) Fragmented future. Print 53:32–35
24. Borkar V, Carey M, Li C (2012) Inside "big data management". In: Proceedings of EDBT/ICDT 2012, pp 3–14
25. Kaplan A, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of Social Media. Bus Horizons 53(1):59–68
26. IBM Coop (2013) What is big data. http://www-01.ibm.com/software/data/bigdata. Accessed 10 July 2013
27. Sterling T, Stark D (2009) A high-performance computing forecast: partly cloudy. Comput Sci Eng 11(4):42–49
28. Freedman D, Kisilev P (2009) Fast mean shift by compact density representation. In: Proc CVPR recognition 2009, pp 1818–1825

29. Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters. In: Proceedings of OSDI, pp 137–149
30. Krcmar H (2012) Information management. Springer, Berlin
31. Li X, Lillibridge M, Uysal M (2010) Reliability analysis of deduplicated and erasure-coded storage. Proc SIGMETRICS 38(3):4–9
32. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. AAAI Press, Menlo Park, pp 37–54
33. Miner G, Delen D, Fast A, Eider J (2012) Practical text mining and statistical analysis for non-structured text data. Academic, Waltham
34. Wagner L, Van Belle J (2007) Web mining for strategic intelligence: South African experiences and a practical methodology. In: Proceedings of ICDSS 2007, paper 1
35. Lukashevich H, Nowak S, Dunker P (2009) Using one-class SVM outliers detection for verification of collaboratively tagged image training sets. In: Proceedings of ICME 2009, pp 682–685
36. Wasserman S, Faust K (1994) Social network analysis: methods and applications, structural analysis. Cambridge University Press, New York
37. Shmueli G, Koppius O (2011) Predictive analytics in information systems research. MIS Q 35 (3):553–572
38. Balsa Rodriguez M, Gobbetti E, Guitian M (2013) A survey of compressed GPU-based direct volume rendering. In: Proceedings of Eurographics 2013
39. Hartmann S (1996) The world as a process: simulations in the natural and social sciences. In: Hegselmann R, Mueller U, Troitzsch K (eds) Modeling and simulation in the social sciences from the philosophy of science point of view. Kluwer Academic, Dordrecht, pp 77–100
40. Chailan R, Bouchette F, Dumontier C (2012) High performance pre-computing: prototype application to a coastal flooding decision tool. Knowledge and systems engineering (KSE). In: Proceedings of KSE 2012, pp 195–202
41. Buhl H, Röglinger M, Moser F, Heidemann J (2013) Big data – a fashionable topic with(out) sustainable relevance for research and practice? Bus Inf Syst Eng 5(2):65–69
42. Kohlwey E, Sussman A, Trost J (2011) Leveraging the cloud for big data biometrics. In: Proceedings of world congress services 2011, pp 597–601
43. Zimbra D, Chen H (2011) Stakeholder approach to stock prediction using finance social media. In: Chen H (ed) Intelligent systems smart market and money. IEEE, Washington, DC, pp 88–92
44. Toole J, Eagle N, Plotkin J (2011) Spatiotemporal correlations in criminal offense records. ACM Trans Intell Syst Technol 2(4), article 38
45. Venkataraman S, Tolia N, Ranganathan P (2011) Consistent and durable data structures for non-volatile byte- addressable memory. In: Proceedings of USENIX 2011, pp 1–15
46. Zhang Y, Gong B, Hui Liu Y (2011) Parallel option pricing with BSDEs method on MapReduce. In: Proceedings of ICCRD, pp 289–293
47. TechAmerica, Foundation Big Data Commission (2013) http://www.techamericafoundation.org/bigdata. Accessed 28 Oct 2013