Towards A Big Data Theory Model

Marco Pospiech
Department of Management and Information
TU Bergakademie Freiberg
Freiberg, Germany
Marco.Pospiech@bwl.tu-freiberg.de

Carsten Felden
Department of Management and Information
TU Bergakademie Freiberg
Freiberg, Germany
Carsten.Felden@bwl.tu-freiberg.de

Abstract— **Big Data is an emerging research topic. The term remains fuzzy and is seen as an umbrella term. Origin, composition, possible strategies, and outcomes are uncertain. Thus, the positioning of publications addressing business administrated issues related to Big Data is impeded. From a practitioner's point of view, the ability to communicate a value proposition is impeded due to the difficulty in scoping the intended artifact and the interpretation of arisen company results. So, underlying relationships and concepts have to be described. The missing theoretical fundament of Big Data has been stated in literature. While some publications actually address this need, the majority of them remain methodically weak. In a previous study we deduced an initial qualitative Big Data theory model based on expert interviews and grounded theory. It is this paper's goal to verify the given model in a quantitative way and test it through structural equation modeling. Thereby, hypothesis are deduced and Big Data indicators presented. As a result, a Big Data theory model arises. All hypotheses of our research model are significant, and the study makes three principal contributions to the scientific discussion about Big Data. First, it unveils the underlying characteristics of Big Data. Second, we show the addressability of Big Data through strategies. Hereby, possible strategies to address Big Data are highlighted. Third, we found evidence that positive outcomes like return of investments through Big Data are possible. Thereby, the latter two aspects are of major interest for practice. The presented work contributes to the scientific discussion and supports a development of this domain.**

Keywords-Big Data; PLS; SEM; Grounded Theory; Definition

I. Introduction

The term Big Data occurs increasingly in scientific and practical discussions [1]. One existing definition comes from Gartner [2], here Big Data belongs to "high-volume, high-velocity, and high-variety information assets that demand cost-effective innovative forms of information processing for enhanced insight and decision-making". Bizer et al. [3] are linking semantic aspects whereby others addressing only the processing of huge amounts of data [4]. One can have the impression that more and more concepts and technologies will be associated with Big Data daily. Here, established research areas such as High Performance Computing (HPC) tend to redefine themselves as Big Data to obtain more attention [5]. As a consequence, the positioning of publications addressing business administrated issues related to Big Data is impeded without concerning the aspect that the research field remains fuzzy, which is to be seen by heterogeneous definitions. [5] From a practitioner's point of view, the benefit of Big Data solutions or existing obstacles that need to be addressed by a Big Data project remain unclear. Even the ability to communicate a value proposition is impeded due to the difficulty in scoping the intended artifact and the interpretation of arisen company results, because Big Data remains fuzzy. [6] The missing theoretical fundament of Big Data has been stated, because relevant topics and related theories are unknown [6; 5]. To overcome this obstacle, it is this paper's goal to introduce a Big Data theory model to describe the current relationships and concepts.

Nowadays, just a few studies are addressing this research goal. In most cases, Big Data characteristics are obtained through argumentative considerations [6; 7]. From a methodical point of view, the evidence of those Big Data characteristics is vague [8]. In a previous study, we performed [9] expert interviews to deduce a descriptive Big Data model as reference for further empirical verification. It represents, to our knowledge, the only Big Data model based on a solid methodology. The model postulates the definition of Big Data through causal conditions and context, whereby Big Data itself leads to strategy and therefore strategies to consequences. This continuative research investigates whether the given descriptive model can be confirmed by a structural equation model (SEM). This verification makes three principal contributions to the scientific discussion about Big Data. First, it unveils the underlying characteristics of Big Data. Second, understanding the theoretical background should lead to the ability to reply to the question whether Big Data can be addressed through strategies or not. Third, the paper seeks to prove whether Big Data strategies lead to consequences for organizations as well as other strategies in different IT research areas [10]. Considering the fuzzy benefits and obstacles of Big Data, especially the latter two aspects are of major interest for practice. Thus, within a well-defined Big Data phenomenon description adequate strategies can be chosen. Having an unique understanding of Big Data knowledge sharing between companies becomes possible. Thus, the research area becomes stronger defined, and business administrated issues can be addressed.

The paper is organized as follows: after a literature review, the initially set up of the descriptive Big Data model is presented, and explained. The model is used to derive our hypotheses, which are proven, by a SEM. To allow a rigid and transparent research, we describe our research methodology and evaluate the model quality with common measures. The results are discussed, and implications are highlighted.

II. STATUS QUO

Conducting a literature review by Cooper [11], we analyzed scientific databases of AIS eLibrary, IEEE Xplore, ACM Digital Library, SpringerLINK, and ScienceDirect until January 2015. We identified relevant papers using search terms shown here in quotation marks: *"Big Data Model", "Big Data Theory", "Big Data Hypothesis", "Big Data Structural Equation Model", "Big Data Partial Least Squares", "Big Data PLS"* and *"Big Data SEM"* in title, keywords and abstract. Since we were analyzing the characteristics of Big Data, the search item *"Big Data Definition"* was also considered. We also conducted a backward search to avoid missing relevant articles [12]. In a first round, 981 initial hits have been obtained. Duplications were removed and abstracts manually analyzed. Interesting papers were fully examined for relevance according to the four-eyes-principle. Because our research scope belongs to articles focusing on a Big Data theory development, only four articles remained. We are aware that some other papers, which are not focusing a theory development might influence this Big Data related theory construction. But, this initial work is concentrating on the usage of existing theory developments and not on the derivation though existing literature.

Wu et al. [7] postulated that Big Data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships. Another work belongs to Loshin [6] and defines Big Data as "applying innovative and cost-effective techniques for solving existing and future business problems whose resource requirements exceed the capabilities of traditional computing environments". The derivation of both theories is based only on argumentation. The methodical weak theory development by argumentation [8] is completed by the missing verification in either case. Considering the literature, it gets apparent how divers are the perspectives and how tough an academic description of Big Data is. Thus, today no other works addressing the need of a Big Data model.

Our first proposed model [13] identified an understanding of Big Data through expert interviews. According to Miles and Huberman [14], expert interviews are suitable in an early state of research to gain a professional perspective. Experts were obtained from internationally acting companies and surveyed through telephone interviews. The samples were coded and conceptualized. Grounded Theory was used, because the generation and discovering of concepts and inherent relationships belongs to the strength of the method [15]. As a result, all identified concepts were assigned to a common coding schema with five categories (Fig. 1). The phenomenon (Big Data) belongs to the middle [15].

The model was applied and tested against existing Big Data publications, whereby the affiliation to the research area was proven. Learning from discussions in conferences and during presentations, our descriptive model was extended as shown in Fig. 1 [9]. Besides that we classified academic publications in practical Big Data implementations by the model, too. The number in brackets belongs to the amount of experts stating the concept. Nevertheless, the study had an initial character. A quantitative analysis has to follow up to be able to clarify, if all categories and relationships will be significant [9]. Thus, a deeper understanding of a phenomenon can be achieved through sequential combination of qualitative and quantitative methods, where findings from the qualitative study empirically informed the later quantitative results. [16]

III. RESEARCH MODEL AND HYPOTHESES

Our model shown in Fig. 1 serves as a basis for the proposed research model, it represents, to our knowledge and besides the prior version [13], the only methodical tested understanding of Big Data. The structure, argumentation and the relationships of the hypotheses are based on a general and accepted schema of [15]. The proposed hypotheses are generally worded, their argumentation is very specific on Big Data to be able to gain new insights from the hypotheses. The following paragraphs presents the hypotheses once we discussed the categories.

Grounded Theory originates from social science to explain, how a situation (phenomenon) emerged. The theory is widely used in IS research to develop substantive theories contributing to the existing knowledge base [17]. *Phenomenon* is defined as central idea, event, happening, incident about which a set of actions or interactions is related. In our perspective, Big Data can be seen as phenomenon, because we can observe effects like growing volume and variety. In addition, the term Big Data was often linked in practice and research to phenomena [18; 19]. First, we have to analyze why Big Data occurs.

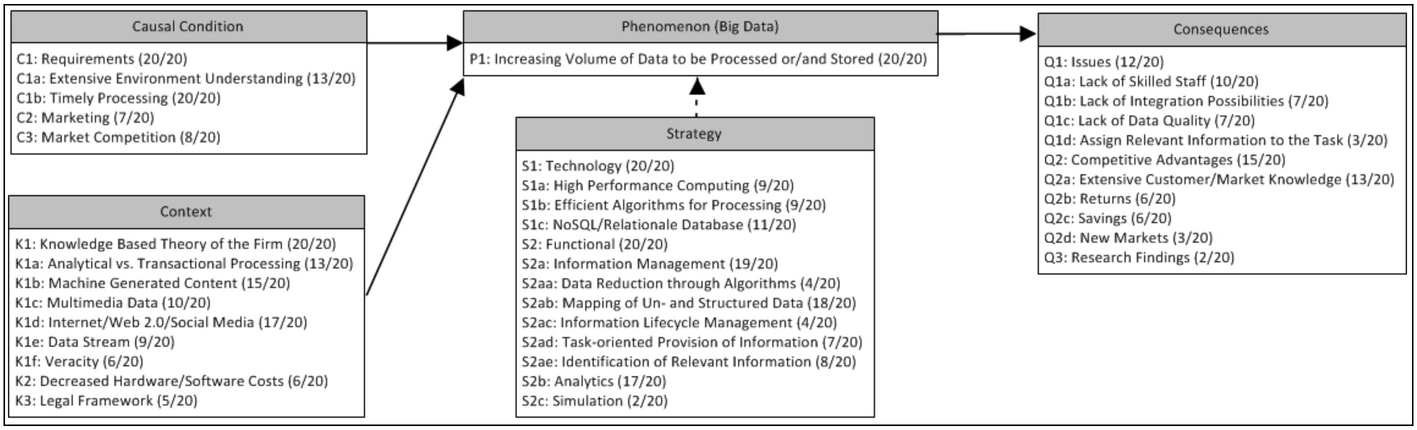

Causal Condition
C1: Requirements (20/20)
C1a: Extensive Environment Understanding (13/20)
C1b: Timely Processing (20/20)
C2: Marketing (7/20)
C3: Market Competition (8/20)
Context
K1: Knowledge Based Theory of the Firm (20/20)
K1a: Analytical vs. Transactional Processing (13/20)
K1b: Machine Generated 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)
Phenomenon (Big Data)
P1: Increasing Volume of Data to be Processed or/and Stored (20/20)
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)
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)

Fig. 1. Big Data Descriptive Model

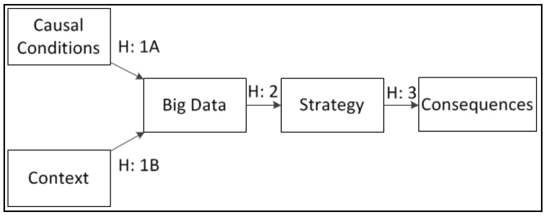


Fig. 2. Big Data Theory Model

The category causal condition (CC) refers to events or incidents that lead to the occurrence or development of a phenomenon [15]. Initial reasons led to the phenomenon, because we recognized a growing volume that has to be processed. [1] have shown that academic and business requirements are drivers of growing data. Thus, the higher the causal conditions of Big Data the higher the phenomenon itself. **Hypothesis 1A:** Causal conditions are positively related to the phenomenon Big Data.

The category context (CO) represents a particular set of properties that pertain to a phenomenon. Properties are locations of events or incidents pertaining to a phenomenon. Thus, context describes the circumstances in which Big Data evolved [1]. Here, we have to observe the environment to identify incidents and locations that pertain to Big Data.One example for a location can be seen in research areas evolving in a parallel manner and pushes the phenomenon Big Data like Social Media [5]. Here, Social Media platforms provide growing user specific content as possible analyzing source. According to [15], the phenomenon is characterized by causal conditions and context. It is our assumption that positive changes in context will lead to more Big Data. Similar, opinions can be found in [1]. **Hypothesis 1B:** Context is positively related to the phenomenon Big Data.

Besides the origin of Big Data, the discussion exists, whether *Big Data* have an effect on the category *strategy* or not. *Strategies* are directed actions to manage, carry out, handle, and respond to a *phenomenon*. *Strategies* are purposeful, goal-oriented, done for some reason, in response to, or to manage a *phenomenon*. [15] In IT, actions like the implementation of technologies or concepts were often addressed to overcome the increasing volume to be processed and/or to be stored [4]. Thus, *strategy* is not a component of the *phenomenon* Big Data itself (shown as punctuate arrow Fig. 1), but it is necessary to address its effects [9]. The absence of any relationship between *strategy* and *Big Data* would question the addressability of Big Data through strategies. **Hypothesis 2:** Big Data is positively related to strategy.

Consequences belongs to outcomes or results of a *strategy*. *Consequences* may be actual or potential, happen in the present or in the future. [15] Within the Big Data domain, *consequences* are often mentioned in form of advantages and issues [20]. Not the *phenomenon*, but rather the *strategies* that manage Big Data determine positive or negative effects [9]. It is of interest, whether an increase within the category *strategy* leads to a positive growth of *consequences*. A possible example could be seen within business analytics. A growing understanding of customers should lead to a higher revenue. Thus, more *strategy* would lead to more *consequences*. **Hypothesis 3:** Strategy is positively related to consequences.

One advantage of the used coding schema [15] belongs to the generalizability to describe various phenomena. We are aware that maybe other concepts like Business Intelligence (BI) would fit into the coding schema (Fig. 1) as well. Nevertheless and at the current state, an initial basis is needed. Further research can extend the presented theory model. Fig. 2 illustrates the proposed research model. Hypothesis and categories (afterwards described as construct) are ordered according to the explanations.

IV. Research Methodology

The paper addresses the introduction of an initial Big Data theory model to describe underlying relationships and concepts. We will evaluate, if relationships between the constructs exist and if they are significant through a SEM. In addition, the belonging of suggested Big Data concepts to constructs is proven, too. We used a partial least squares (PLS) model, because this technique is well suited to study associations between latent variables, if new theoretical ground will be explored and measures are new [22]. The survey (Table 1) was conducted according to the guidelines of [23]. The model was tested under the assumption that the *phenomenon* Big Data can be described by a human consensus, only, because it is determined by its linguistic description [9]. This consensus can be achieved by an experienced crowd. We searched suitable participants within the business networking platform Xing [24]. The platform provides several Big Data interest groups, where a sufficient expertise of the members can be assumed. To ensure quality, only practitioners with at least three years of experience were considered because the underlying model (Fig. 1) is based on the input of IT companies' employees. A pre-test was done by six participants to confirm understandability of our measures. After minor adjustments, the anonymous survey was conducted from May 2014 until August 2014 by dint of an online-survey tool. An introduction of the survey's goal was given and the categories were explained to guarantee a common understanding [23]. All items were requested through a 6-point Likert-type scale anchored on "1 = strongly disagree" and "6= strongly agree", because 6-point Likert-type scales may have a higher reliability than 5-point scales [25].

385 participants joined and 123 completed the survey. This leads to a participant rate of 31.95 %. [1] The professional experience of our participants was high (22.76 % (3-9 years), 40.65 % (10-19 years), and 36.59 % (20 and more years)). The firms of the respondents were located in Europe (82.11 %), America (13.00 %), and Asia (4.88 %), whereby the enterprise size (measured in employees) of the participants was about the same size. 27.64 % of the correspondents were from companies having less than 50 employees, 26.83 % (50-499 employees) and 45.53 % (499 or more employees). Thus, the data set is representative, because our data sample contains opinions of several Big Data interest groups.

[1] According to Chin [26], a sample size should at least exceed ten times the number of indicators of the construct with the most indicators and ten times the largest number of exogenous variables loading on a single endogenous variable. In consideration of the research model, the needed sample size is 120.

A. Measures

The measures of our research model are obtained from the given descriptive model. We followed the model as closely as possible to retain the theoretical underpinnings. Thus, no additional parts are added even if they would have been reasonable. Only minor alignments have been done following the suggestions in conferences and presentations to avoid contextual overlaps. Constructs can be measured in a reflective or formative approach. The distinction is important, because an appropriate specification of the measurement model is necessary to gain meaningful relationships within the structural model. 41 indicators are identified and discussed in the following paragraphs. The identified 41 indicators are discussed in the following sections.

Table 1 shows the applied questionnaire. Constructs can be measured in a reflective or formative approach. The distinction is important, because an appropriate specification of the measurement model is necessary to gain meaningful relationships within the structural model.

We deduced five indicators within the construct *causal conditions (CC)* [27]. The construct is formative, because we were aiming at the identification of reasons that describe the *causal condition* construct. Three of the five indicators belong to the main category requirements [27]. Here, we divided the factor "extensive environment understanding" into *market understanding* (CC1) and *research environment understanding* (CC2), because the need belongs to several aspects. First, enterprises have to gain insights about markets and customers, whereby academics uncover circumstances in sciences. Another requirement is seen in the need of a *timely processing of information* (CC3), because traditional approaches compute usually too long. There is a discussion whether Big Data is *marketing-driven* (CC4) or not [1]. If the factor loading is positive, any further theoretical examinations will be questionable, because Big Data has no novelty and is only precipitated by sales people. The final indicator belongs to *dynamic markets* (CC5). Here, companies have to reduce production cycles, save costs, react fast, and maximize their profits. The increasing pervasion of IT in enterprises to address the market may lead to Big Data.

The second construct belongs to the formative *context (CO)*. *Context* changes will be initiated through the indicator (like multimedia data) and not trough the construct itself. A similar influencing of all measures through the latent variable as in reflective models is doubtful. The first indicator belongs to the Knowledge Based Theory of the Firm (KBT) [27]. Here, knowledge is seen as unique and most *strategically resource* (CO1) by focusing on knowledge integration and combination of several sources to achieve a competitive advantage. KBT is rather used as indicator instead of a theory and it should not be interpreted as a theory mapping. Another context leading to the *phenomenon* of Big Data is the changing analysis focus towards *transactional data processing on the fly* (CO2). Thus, transactions are no longer the fundamental basis of analyses, as we understand it in BI, but rather an explanatory variable. As a result of increasing IT pervasion, more and more *machine generated content* (CO3) occurs in a world of log files, internet of things, GPS, and any kind of sensor technology to gain an extended environmental understanding.

TABLE I. QUESTIONNAIRE

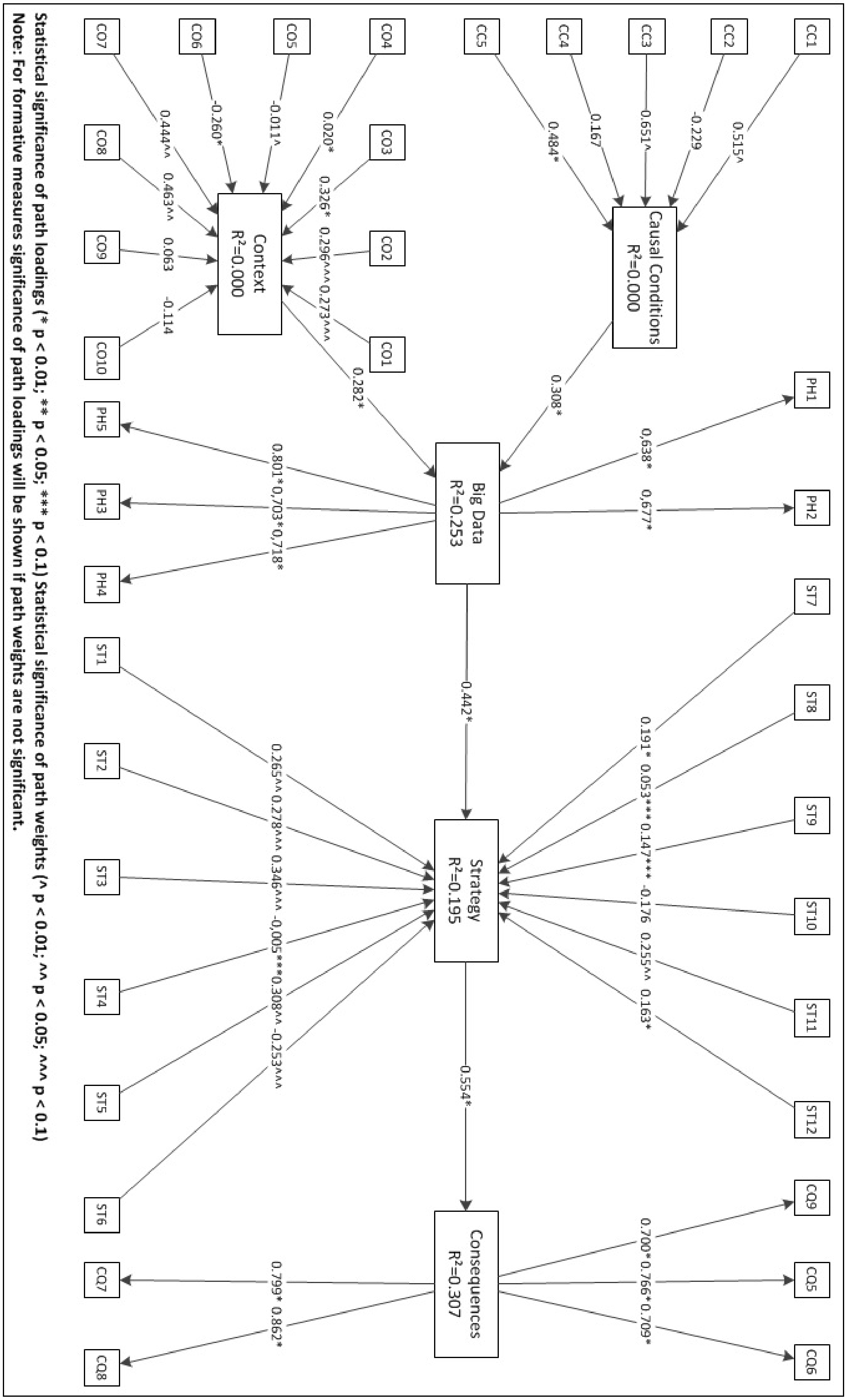
Item	Statement
PH1	An increasing data volume to store is observable
PH2	An increasing data volume to transform is observable
PH3	An increasing data volume to access is observable
PH4	An increasing data volume to visualize is observable
PH5	An increasing data volume to analyze is observable
CC1	Big Data occurs through the need of processing data to allow an market understanding
CC2	Big Data occurs through the need of processing data to allow an research environment understanding
CC3	Big Data occurs through the need of a timely processing, because traditional approaches compute too long
CC4	Big Data is only driven by marketing departments
CC5	Big Data occurs through the existence of dynamic markets
CO1	Big Data emerges since knowledge is seen as unique and strategic resource
CO2	Big Data emerges since transactional data (e.g. ERP) is analyzed on the fly
CO3	Big Data emerges since the amount of machine generated content is growing
CO4	Big Data emerges since the amount of multimedia data is growing
CO5	Big Data emerges since this topic has to consider unknown data quality within the data itself
CO6	Big Data emerges since the internet provides an appropriate infrastructure for data processing
CO7	Big Data emerges since the Web 2.0 forced firms to develop efficient and scalable technologies
CO8	Big Data emerges since social media platforms provide user specific content
CO9	Big Data emerges since the IT costs are decreasing
CO10	Big Data emerges since this topic has to consider legal frameworks
ST1	Cloud computing is an appropriate method to handle Big Data
ST2	Efficient programming models (like MapReduce) are an appropriate method to process Big Data
ST3	Key-Value-oriented databases are an appropriate method to store Big Data
ST4	Document-oriented databases are an appropriate method to store Big Data
ST5	Column-oriented databases are an appropriate method to store Big Data
ST6	Relational databases are an appropriate method to store Big Data
ST7	Streaming is an appropriate method to handle Big Data
ST8	The integration of (un-)structured data is appropriate prerequisite to analyze Big Data
ST9	Task-oriented provision of information is an appropriate method to manage Big Data
ST10	ILM as framework for policies, processes, practices, and tools used to align the business value of information with a effective IT disposition is an appropriate method to manage Big Data
ST11	Simulations are an appropriate method to analyze Big Data
ST12	Analytics (Data-, Text-, and Web Mining, Social-, Image-, Audio- , Video, and Predictive-Analytics, Visualization) are an appropriate method to analyze Big Data
CQ1	Big Data leads to a remarkable lack of skilled staff
CQ2	Big Data leads to a remarkable lack of integration possibilities
CQ3	Big Data leads to a remarkable lack of data quality
CQ4	Big Data leads to issues regarding the assigning of information to the task
CQ5	Big Data leads to an extensive customer/market knowledge
CQ6	Big Data leads to new business models
CQ7	Big Data leads to increasing cost savings
CQ8	Big Data leads to increasing investment returns
CQ9	Big Data leads to new research findings

Besides that, a data type shift comes in to the discussion. Growing *multimedia* (CO4) sources (image, video, audio, and text) must be processed and stored on a large scale. These types of data and sources implicate another measure. Whereby traditional data were clean and precise, the new ones are rather fuzzy. Thus, combining unknown sources means also unexplored *data quality* (CO5). Fig. 1 mentions the context Internet/ Web2.0/ Social Media. We split the fact into three separate measures due to their different meanings. *Internet* (CO6) considers the physical infrastructure which enables a world- spanning interconnection of all participants and consequently an exchange of data. *Web 2.0* (CO7) are interactive platforms where people collaborate and share information. Companies were forced to develop cost-effective and scalable technologies to overcome and analyze the growing content. [28]. Since *Social Media* (CO8) is one of the most important applications within the Web 2.0, we analyze the measure separately [21]. Here, user-specific content creates a possibility to gain a more advanced view about customer and environment than before. Besides that, decreasing *IT costs* (CO9) might be responsible for Big Data. Thus, hardware prices to process and store data are decreasing constantly. In addition, mentioned Big Data technologies like Hadoop are usually open source tools. The finally reflected context is the *legal framework* (CO10). Big Data applications are suited between data privacy. Thus, the combination of sources could be forbidden.

The third construct is the *phenomenon* Big Data (BD) itself. In our previous work [9] the construct *phenomenon* was simple described as an "Increasing Volume of Data to be Processed or/and Stored". Nevertheless, the description goes along with Strauss and Corbin's coding schema [15], but provides less information for suitable PLS indicators. But, a former literature analysis in an earlier research stage has shown that Big Data belongs to a phenomenon occurring within the IT domain [5]. Thus, Big Data is observable in IT tasks. According to the Information Processing Theory, IT belongs to any mechanism that facilitates the gathering of data, the transformation of data into information, the communication and storage of information in the organization. [29] Following this argumentation, we identified *storing* (BD1), *transformation* (BD2), and *accessing/gathering* (BD3) of data where *Big Data* could be observable. Additionally, we assume that the communication of insights about the fields of discourse within *Big Data* is done by *visualization* (BD4) and *analysis* (BD5). In this context, we are able to examine the occurrence of a growing need of processing in an IT domain. To provide more substantiality, the five indicators are going along with the findings of Chen et al. [30]. The operationalization of this construct is reflective because the *phenomenon* influences the measures. A modification within the construct *Big Data* will affect all items, because the items share a common theme.

The construct *strategy* (ST) has a formative operationalization [27]. We examine what kind of concepts belong to Big Data. Thus, the construct is not independent and is constituted by the measures itself. Fig. 1 divides *strategy* into the sub-categories technology and functional [27]. HPC is indicated as first. In, [9] we mentioned concepts like cloud, grid, distributed, and parallel computing. A huge overlap between the concepts grid, distributed, and parallel computing is observable. Thereby, grid and parallel computing can be seen as part of distributed computing [31; 32]. Nevertheless, cloud computing combines the concepts of parallel and distributing computing on top of the service delivery approach [33]. We use *cloud computing* (ST1) as single item to avoid an overlap between all other items. More *efficient programming models* (ST2) considering time and computing complexity to process data in a faster way were mentioned. Here, programming models like MapReduce are imaginable [34]. Another discussion belongs to the database area. Often stated, NoSQL approaches were *key-value-* (ST3), *document-* (ST4) or *column-oriented-* (ST5) *databases* [35]. Besides that, a debate considering the affiliation of *relational databases* (ST6) was observed. One minor adjustment was done according to Fig. 1. After conference discussions, we shift the indicator *streaming applications* (ST7) from the construct *context* to *strategy*. The reason is that streaming applications are a kind of technology and thus, more related to *strategy*. Streaming was developed for specific needs like analytics in operational BI aiming at the processing of a huge amount of data [36]. Thus, it is rather an essential technology than a driver of Big Data. The second sub-category belongs to functional strategies. One major concept is Information Management, which is divided into several sub-items. The *mapping of structured and unstructured data* (ST8) is one of the core concepts. Here, various data sources must be integrated in a useful manner to gain a broader view about the environment. This goes along with the *task-oriented provision of information* (ST9). Thereby, right information, at the right time, in the right amount and place in an adequate quality must be assigned to the right task. *Information lifecycle management (ILM)* (ST10) is also seen as a possible approach of Big Data *strategies*. It comprises policies, processes, practices, and tools to assign time-dependent value to information to be able to facilitate a storage of information according to its value. This includes a deletion at the appropriate point in time [37]. Fig. 1 mentioned the indicator data reduction through algorithms. After discussions, we argue that there is a huge overlap to ILM. One example is IBM's data deduplication method within their ILM solution [38]. ILM always addresses the reduction through the lifecycle concept. In this context, data reduction was not considered. Even the identification of relevant information through management methods or machine learning techniques remains fuzzy. The overlap of data reduction spans from ILM and task-oriented provision to analytics. In this context, we forbear considering this indicator. At least, the items *simulation* (ST11) and *analytics* (ST12) are used to communicate the insights. Within *analytics* manifestations of Data-, and Text Mining, Image-, Audio-, Video-, Predictive-Analytics, and Visualization are possible.

The indicators of *consequences* (CQ) are separated into competitive advantages, research findings, and issues. One of the most cited issues belongs to the missing *Big Data experts* (CQ1). Here, more technical and domain-specific knowledge is necessary. Thereby, the integration of various data sources and technologies like NoSQL is complex and critical (CQ2). Considering the multimedia data type, the analysis of this item remains fuzzy and uncertain [28]. As a result, a *lack of data quality* arises (CQ3). Another issue is seen in *assigning relevant information to the task* (CQ4). Thereby, the area of possible Big Data sources is broad and not every content supports an improved understanding of the domain. Besides those issues, one competitive advantage refers to the *extensive view about customer and market* (CQ5), where strategies like *analytics* or *simulations* may bring insights within the domain.



Statistical significance of path loadings (* $p < 0.01$; ** $p < 0.05$; *** $p < 0.1$) Statistical significance of path weights (^ $p < 0.01$; ^^ $p < 0.05$; ^^^ $p < 0.1$)

Note: For formative measures significance of path loadings will be shown if path weights are not significant.

Fig. 3. Results Big Data Model

This goes along with the question of *new business ideas and markets* (CQ6). Even financial advantages were stated from [30]. *IT savings* (CQ7) can be achievable trough open source tools, saved memory, and computational capacity. Following KBT, *returns* (CQ8) will be possible, if information are seen as strategic resources [28]. The final consequence measure addresses the academic application of Big Data. Here, new insights and relationships (CQ9), in particular within natural science, are imaginable. The construct *consequences* underlies a reflective measurement model. The measures will be the final entity in a causal chain and will rather be influenced through the construct instead of influencing the construct. Such measures have to be reflective [27].

B. *Data Analysis*

Data analysis was conducted as a two-step approach [39] using PLS with the tool SmartPLS 2.0, which was applied in many IS studies [40; 41]. Once the measurement model is analyzed, the relationships of the structural model are evaluated [39].

According to Chin [26], the adequacy of a PLS measurement model with respect to reflective indicators is ensured by examining the individual item reliabilities and by the evaluation of convergent and discriminant validity of the construct´s measures. Within the item reliability, indicator loadings lower than 0.7 must be sequentially eliminated from the model [26]. If the Average Variance Extracted (AVE) of the construct is greater than 0.5, loadings of 0.6 will also be acceptable. All loadings must be significant with at least $p < 0.05$. All measures referring to issues (CQ1; CQ2; CQ3; CQ4) were removed since all indicator loadings were lower than 0.6 [22]. The convergent validity of the different construct measures is analyzed by the Internal Composite Reliability (ICR) of a construct and must be at least 0.7. Furthermore, the convergent validity is investigated through the AVE. The AVE should be greater than 0.5 for all reflective constructs. Both constructs, *Big Data* (ICR=0.834; AVE=0.503) and *consequences* (ICR=0.878; AVE=0.592), are fulfilling the requirements. The final assessment of our measures belongs to the discriminant validity. Here, two measures of goodness are considered. Hair et al. [22] clarify that the squared correlation between any two constructs is smaller than the AVE of each construct. Discriminant validity is given, because the highest correlation ($0.592^2=0.350$) is below 0.503.

The second discriminant validity check belongs to cross-loadings. Thereby, all items must be loaded stronger to the construct they were supposed to measure compared to any other construct. As shown in Table 2, all reflective items (bold font) are stronger loaded for the respective construct. The formative indicators of the model will be considered as valid, if the indicator weights or loadings are significant for at least $p < 0.1$. The first test represents the relative and the second test the absolute impact.

If one weight or loading is significant, empirical support for the indicator's relevance will be assumed [42]. Fig. 3 illustrates the significances of the measures. Nevertheless, the indicators CC2, CC4, CO9, CO10, and ST10 are not significant in both cases.

TABLE II. CROSS-LOADINGS

	BD	**CQ**	**CC**	**CO**	**ST**
BD1	**0.63**	0.28	0.31	0.31	0.32
BD2	**0.67**	0.23	0.28	0.29	0.27
BD3	**0.70**	0.21	0.25	0.31	0.28
BD4	**0.71**	0.23	0.40	0.27	0.31
BD5	**0.80**	0.35	0.26	0.28	0.34
CQ5	0.29	**0.76**	0.38	0.40	0.41
CQ6	0.16	**0.70**	0.17	0.36	0.38
CQ7	0.41	**0.79**	0.38	0.38	0.50
CQ8	0.32	**0.86**	0.37	0.45	0.48
CQ9	0.19	**0.70**	0.07	0.31	0.30

The discriminant validity test takes place comparable to [22]. In contrast to reflective measures, correlations between the constructs should not be squared. To ensure a satisfying level, the construct correlation of the formative constructs with the other model constructs should be lower than 0.9. Because the highest correlation is between *strategy* and *consequences* (0.554), the values are below. In PLS, structural model quality is measured through path significance, determination coefficient (R^2), and predictive relevance (Q^2) [22]. As illustrated in Fig. 3, all hypotheses are significant on a $p < 0.01$ level and have positive loadings. Thus, all hypotheses are accepted.

The explanatory quality of the structural model is measured for endogenous constructs by R^2. Chin [26] defined values of 0.67, 0.33, and 0.19 as substantial, moderate, and weak and values lower than 0.19 as not relevant. As illustrated in Fig. 3, all values fulfil the criteria. Q^2 is only measured for endogenous, reflective constructs and must be greater than 0. The constructs *Big Data* (Q^2=0.128) and *consequences* (Q^2=0.1753) are above the required boundaries. In summary, all model quality requirements are fulfilled, and the model allows the discussion of meaningful and significant assertions.

V. DISCUSSION

All hypotheses of the underlying descriptive model are strongly significant ($p > 0.01$) and all path loadings exceed 0.2 in order to gain meaningful results [43]. Even the predictive Q^2 and R^2 are acceptable. We can assume that *Big Data* is related to *causal conditions* (Hypothesis 1A) and *context* (Hypothesis 1B). An increase in any one of these constructs affects *Big Data* and all inherent measures in a positive manner. The two *causal condition* measures CC1 and CC3 are highly significant and impacting the construct at most. At least, the path loadings for CC5 are significant, but the standard deviation (SD) value of 2.050 expressing inconsistent expert opinions. Eight of ten measures affect the construct *context* significantly, but the indicators CO4 and CO5 have less significant impact. Surprisingly, CO5 affects the construct negatively. Thus, the lower the *data quality* the higher Big Data will emerge. The finding makes sense considering analytics. First, analysts have to understand and structure data before insights can be obtained and data quality concerns are obstructive in any use of analytics. Interesting is the negative path loading of indicator CO6 meaning more *Internet* will reduce the effect of Big Data. Perhaps the effect can be explained through the high SD value of 1.38 within the indicator. CO9 and CO10 are not significant (see Table 3). But, both indicators are part of the descriptive Big Data model and should not be deleted, because this would ignore theoretical underpinnings [22]. Already, the descriptive model (Fig. 1) has shown that these measures are only supported by 30 percent of the interviewed experts, which could explain the missing significance.

The measures of the construct *Big Data* are highly significant and indicate meaningful path loadings. Considering that IT belongs to any mechanism that facilitates the gathering of data, transformation of data into information, communication and storage of information in the organization, we can assume that Big Data should be relevant at any level of IT [29]. The explained variance of $R^2=0.253$ is still weak. Nevertheless, the predictive relevance with $Q^2=0.128$ is high. Hypothesis 2 is also supported. Thus, more Big Data will lead to more strategy. Organizations struggling with at least one of the observed Big Data indicators initialize strategies to address them. The explained variance of $R^2=0.195$ is weak. This could be an indication that not all influencing factors are exposed.

This study confirms eleven significant strategies for Big Data. Only *ILM (ST10)* was not supported. Thus, our study contributes to the practice by identifying strategies to tackle Big Data. Typical NoSQL approaches like ST3 or ST5 are impacting the construct at most. Surprisingly, ST4 is significant, but the path loading is low. The discussion whether relational databases (ST6) are adequate to process Big Data can be affirmed [35]. As expected, ST2 like MapReduce is supported to deal with Big Data. Even the paradigm cloud computing is supported. ST7 is also significant. Thus, all mentioned technologies are supported.

At the moment, Big Data is less addressed by Information Management approaches [5]. Our study indicates evidence that two of three Information Management aspects are appropriate in Big Data. First, ST8 to gain a broader view about the environment. Second, ST9 is supported. Here, our study motivates future research to address these research areas. However, ST10 is not supported. Thus, Big Data will not focus on the deletion of data, if the value of its usage will decrease over time. A possible explanation could be linked to BI or at least to data warehouses [44]. The indicator analytics is highly significant and seems to be appropriate. Even simulation is considered as possible strategy.

We verified a positive relation between *strategy* and *consequences* (Hypothesis 3). This goes along with other findings in IT research domains [10]. The path coefficient is highly significant and impacts *consequences* with a strong loading of 0.554. A growing usage of Big Data *strategies* will cause a rise in *consequences*. The $R^2=0.307$ is still small, but close to moderate [26]. According to the requirements of reflective measures, all issue-related indicators are removed. In this context, we found evidence that only positive outcomes are generated through Big Data *strategies*, yet. This motivates the further development of strategies to increase the outcome of Big Data. Nevertheless, deleting all issue-related indicators contradicts to the findings of literature. The study even provides evidence that positive outcomes through Big Data *strategies* are observable. All indicators are significant on a $p < 0.01$ level. Possible financial outcomes are shown in CQ8 and CQ7. Even the identification of *new business models* is supported. Nevertheless, we found evidence that Big Data *strategies* allow the development of CQ5 and CQ9.

In summary, the study illustrates a significant cause-effect relationship. Thus, if there are increasing exogenous variables (*causal condition* and *context*), the observable indicators (store, transform, access, visualize, analyze) in *Big Data* will grow. This would lead to a growing usage of Big Data related *strategies* and at least to positive outcomes for organizations.

TABLE III. DESCRIPTIVE STATISTICS

Item	Mean	SD	Item	Mean	SD
BD1	5.34	0.72	ST1	4.25	1.31
BD2	4.79	1.00	ST2	5.35	0.58
BD3	5.05	0.60	ST3	4.65	0.95
BD4	4.97	1.01	ST4	4.09	1.58
BD5	5.42	0.63	ST5	4.57	1.02
CC1	5.08	0.76	ST6	3.47	1.92
CC2	4.78	0.91	ST7	4.55	1.15
CC3	4.51	1.82	ST8	4.90	0.95
CC4	2.73	2.05	ST9	4.22	1.21
CC5	4.02	1.73	ST10	4.50	0.98
CO1	4.89	1.03	ST11	4.39	1.36
CO2	4.36	1.28	ST12	5.35	0.55
CO3	5.35	0.66	CQ1	4.63	1.82
CO4	5.12	0.98	CQ2	3.66	1.78
CO7	4.58	1.49	CQ3	4.00	1.95
CO8	4.87	1.06	CQ4	3.98	1.25
CO7	4.58	1.49	CQ5	4.95	1.00
CO8	4.87	1.06	CQ6	5.32	0.57
CO9	3.93	1.63	CQ7	4.17	1.31
CO10	3.93	1.57	CQ8	4.36	1.46
			CQ9	5.18	0.73

VI. CONCLUSION

The paper has addressed an initial theoretical fundament of Big Data. We developed a research model, based on a given descriptive Big Data model, to describe potentially underlying relationships and concepts. Concluding, all hypotheses of our Big Data theory model were positively tested by a PLS model. Thus, this theoretical fundament of Big Data serves as contribution to the scientific discussion, which justifies further research, because Big Data represents a more defined research domain. Furthermore, the study contributes to a clearer understanding of Big Data, which supports the academic and practical discussion, too. Using our model, future applications and research can justify the belonging to Big Data. Now, organizations can easier define Big Data situations and value propositions. Even the transfer of knowledge is simplified, because the model serves as common understanding within the Big Data domain

We identified *causal conditions* and *context* to explain the occurrence of the Big Data phenomenon. Knowing the reasons for occurrence, organizations can adjust their requirements in order to obtain more or less Big Data. According to Information Processing Theory, the presented study found evidence that *Big Data* is observable in all IT mechanism. In addition, we verified a positive relation between Big Data and strategy. Thus, the more *Big Data* is within an organization, the more *strategy* they will use. We even showed significant strategy manifestations. The paper provides possible Big Data *strategies,* which are suitable for the observed increases of *Big Data*. Thus, organizations dealing with Big Data can choose between several options to overcome the phenomenon's effects. We found evidence that Big Data *strategies* lead to positive *consequences*. Organizations considering investments in Big Data related topics find support within the possible value generation in our results. Thus, further research is supported.

Nevertheless, the measured R² of all constructs is low. The next research steps have to identify further constructs which explain the occurrence of Big Data in more detail. Here, the proposed model represents a current state and can be used as a fundament. It is not limited, thus new developments within the area of Big Data could be updated. In addition, we are aware that our model might not cover all related concepts, because we focus on the verification of Tab. 1. In addition, we are aware that all indicators are common for IT concepts like BI, too. Such a generalizability makes them not unique for Big Data. But first and because of the initial status of the model, future research shall identify more specific indicators. Second, the assortment of Big Data *contexts* and *causal conditions* are particular. Even the selected *strategies* are unique within the combination and form an own construct. Thus, our model contributes to a more established Big Data understanding within the scientific discussion, including the awareness that single indicators are common in IT. Also, the applied descriptive Big Data model and our data sample were gained by discussing with practitioners. It is of interest whether a different Big Data understanding between research and practice exists, because this also contributes to the scientific discussion of Big Data and increases the body of knowledge about this topic.

REFERENCES

[1] H. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big Data A Fashionable Topic with (out) Sustainable Relevance for Research and Practice?", BISE, 5, 2, 2013, pp. 65-69.

[2] Gartner. (2014). IT Glossary Big Data [Online]. Available: http://www.gartner.com/it-glossary/big-data/

[3] C. Bizer P. Boncz, and M. Brodie, "The Meaningful Use of Big Data: Four Perspectives", SIGMOD Record, 40, 2014, pp. 56-60.

[4] Y. He, R. Lee, and Y. Huai.., "RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems", ICDE, Hannover, 2011, pp. 1199-1208.

[5] M. Pospiech, and C. Felden, "Big Data – A State-of-the-Art", AMCIS, 2012, Paper 22.

[6] D. Loshin, Big Data Analytics. Waltham: Morgan Kaufmann, 2009.

[7] X. Wu, X. Zhu, and G. Wu, "Data Mining with Big Data", IEEE TKDE, 26, 2014, pp. 97-107.

[8] R. D. Galliers, and F. F. Land, "Choosing an Appropriate Information Systems Research Methodology", Comm. ACM, 30, 1987, pp. 900-902.

[9] M. Pospiech, and C. Felden, "Deployment of A Descriptive Big Data Model", in Business Intelligence for New-Generation Managers, J. Mayer, and R. Quick (eds.), Heidelberg, Springer, 2015, pp. 75-95.

[10] G. Piccoli, and B. Ives, "IT-Dependent Strategic Initiatives and Sustained Competitive Advantage", MISQ, 29, 2005, pp. 747-776.

[11] H. Cooper, Synthesizing Research, Thousand Oaks: Sage, 1998.

[12] J. Webster, and R. Watson, "Analyzing the past to prepare for the future: writing a literature review", MISQ, 26, 2002, pp. 13-23.

[13] M. Pospiech, and C. Felden, "A Descriptive Big Data Model Using Grounded Theory", BDSE, 2013, pp. 878-885.

[14] M. Miles, and A. Huberman, Qualitative Data Analysis. Thousand Oaks, Sage, 1994.

[15] A. Strauss, and J Corbin, Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Thousand Oaks: Sage, 1990.

[16] W Kim H. Chan, and A. Kankanhalli, "What Motivates People to Purchase Digital Items on Virtual Community Websites?" Information Systems Research, 23, 2012, pp. 1232-1245.

[17] F. Niederman, "Using Grounded Theory to Generate Indigenous MIS Theory", AMCIS, San Francisco, 2009.

[18] Gartner. (2014). IT Glossary Big Data [Online]. Available: http://www.gartner.com/it-glossary/big-data/

[19] D. Boyd, and K. Crawford, "Critical Questions for Big Data. Information" Communication & Society, 15, 5, 2012, pp. 662-679.

[20] A. Cuzzocrea, Y. Song, and K. Davis, "Analytics over Large-Scale Multidimensional Data", DOLAP, UK, 2011, pp. 101-103.

[21] A. Kaplan, and M. Haenlein, "Users of the world, unite!", Business Horizons, 53, 2010, pp. 59-68.

[22] J. Hair, C. Ringle, and M. Sarstedt., "PLS-SEM: Indeed a silver bullet", Journal of Marketing Theory and Practice, 19, 2011, pp. 139-152.

[23] E. Babbie, Survey Research Methods, Wadsworth: Cengage, 1990.

[24] Xing 2014. (2014). Xing [Online]. Available: https://www.xing.com/en

[25] R. Chomeya, "Quality of Psychology Test Between Likert Scale 5 and 6 Points", Journal of Social Sciences, 6, 2010, pp. 399-403.

[26] W. W. Chin, "The partial least squares approach for structural equation modeling", in Modern methods for business research, G. A. Maracoulides (eds.), Mahwah, Lawrence Erlbaum, 1998, pp. 295-336.

[27] T. Coltman T. Devinney, and D. Midgley., "Formative versus reflective measurement models", JBR, 61, 2008, pp. 1250-1262.

[28] J. Krumm N. Davies, and C. Narayanaswami, "User-generated content", Pervasive Computing, 7, 2008, pp. 10-11.

[29] J. F. Fairbank G. Labianca, and H. Steensma, "Information Processing Design Choices, Strategy, and Risk Management Performance", Journal of MIS, 23, 2006, pp. 293-319.

[30] H. Chen R. Chiang, and V. Storey, "Business intelligence and analytics: from big data to big impact", MISQ, 36, 4, 2012, pp.1165–1188.

[31] I. Foster Y. Zhao, and I. Raicu., "Cloud Computing and Grid Computing 360-Degree Compared", GCE, Austin, 2008, pp. 1-10.

[32] D. Peleg, Distributed Computing. Philadelphia: SIAM, 2000.

[33] K. Hwang J. Dongarra, and C. Fox., Distributed and Cloud Computing. Waltham: Morgan Kaufmann, 2000.

[34] J. Dean, and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Comm. ACM, 51, 2008, pp. 137-149.

[35] J. Han, E. Haihong, and G. Le, "Survey on NoSQL Database", ICPCA, South Africa, 2011, pp. 363-366.

[36] M. Castellanos C. Gupta, and S. Wang, "Leveraging web streams for contractual situational awareness in operational BI", EDBT, Lausanne, 2010, pp. 1-8.

[37] SNIA 2014. (2014). Information Lifecycle Management [Online]. Available: http://www.snia.org/

[38] M. Ebbers M. Archibald, and C. da Fonseca, (2014). IBM Smarter Data Centers [Online]. Available: http://www.redbooks.ibm.com/

[39] J. Hulland, "Use of partial least squares (PLS) in strategic management research", Strategic Management Journal, 20, 1999, pp. 195-204.

[40] C. Ringle S. Wende, and A. Will (2014). SmartPLS [Online]. Available: http://www.smartpls.de

[41] S. Smith, R. Johnston, and S. Howard, "Putting Yourself in the Picture", ISR, 22, 2011, pp. 640-659.

[42] R. Cenfetelli, and G. Bassellier, "Interpretation of Formative Measurement in Information Systems Research", MISQ, 33, 2009, pp. 689-708.

[43] W. W. Chin, "Issues and Opinion on Structural Equation Modeling", MISQ, 1998, pp. 7-16.

[44] P. Zikopoulos, D. deRoos, and K. Parasuraman., Harness the Power of Big Data - The IBM Big Data Platform. New York: McGraw-Hill, 2012