2014 47th Hawaii International Conference on System Science

Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments

Katharina Ebner
EBS Business School,
Wiesbaden, Germany
katharina.ebner@ebs.edu

Thilo Bühnen
EBS Business School,
Wiesbaden, Germany
thilo.buehnen@ebs.edu

Nils Urbach
EBS Business School,
Wiesbaden, Germany
nils.urbach@ebs.edu

Abstract

Businesses increasingly attempt to learn more about their customers, suppliers, and operations by using millions of networked sensors integrated, for example, in mobile phones, cashier systems, automobiles, or weather stations. This development raises the question of how companies manage to cope with these ever-increasing amounts of data, referred to as Big Data. Consequently, the aim of this paper is to identify different Big Data strategies a company may implement and provide a set of organizational contingency factors that influence strategy choice. In order to do so, we reviewed existing literature in the fields of Big Data analytics, data warehousing, and business intelligence and synthesized our findings into a contingency matrix that may support practitioners in choosing a suitable Big Data approach. We find that while every strategy can be beneficial under certain corporate circumstances, the hybrid approach – a combination of traditional relational database structures and MapReduce techniques – is the strategy most often valuable for companies pursuing Big Data analytics.

1. Introduction

By intelligently using the information in and around them, organizations are able to improve their decision-making and better realize their objectives [1, 2]. Some authors even claim that organizations may lose competitiveness by not systematically analyzing the available information [3]. However, to obtain the desired insights, data need to be sourced, stored, and analyzed [4, 5]. During the past years, accessing and processing the collected, voluminous, and heterogeneous amounts of data has become increasingly time consuming and complex [6]. With a total of 1.8 zettabyte in 2011, the amount of generated data has not yet reached its climax: as expected by IDC, a global provider of IT market intelligence, the total amount of data collected until the end of 2012 is estimated to be 1.48 times the amount of data collected in previous years, with more than 90% of this data being unstructured [7]. Businesses increasingly use these data masses provided by millions of networked sensors in mobile phones, cashier systems, automobiles, or weather stations to learn more about their customers, suppliers, and operations [8]. For instance, in a recent survey, half of the respondents stressed the importance of analytics in their company and more than 20% claimed to be under pressure to improve their business analytics [2]. This development raises the question of how companies manage to cope with the characteristics of the ever-increasing amount of data, referred to as Big Data. The aim of this paper is to provide a set of organizational contingency factors that influence different Big Data strategies organizations may implement. In order to do so, we reviewed existing literature to identify different Big Data strategies as well as contingency factors and synthesized both into a contingency matrix that may support practitioners in choosing a suitable Big Data strategy for their specific context.

The paper is structured as follows. In Section 2, we present our definition of Big Data and the relevant background that informed our research. In Section 3, we outline our method of inquiry, before we proceed with a presentation of our findings in Section 4. The synthesis of the review results is presented and discussed in Section 5.

2. Background

The characteristics of Big Data were first described in 2001, when Laney [4] identified three key attributes of large data amounts: high variety, volume, and velocity. To date, these attributes have become the defining characteristics of Big Data. However, contemporary authors and business specialists enlarged these defining characteristics with further aspects such as dedicated storage, management, and analysis techniques [8, 9, 10].

978-1-4799-2504-9/14 $31.00 © 2014 IEEE
DOI 10.1109/HICSS.2014.466

Further amendments to the definition include the addition of a fourth V, veracity, by IBM [11], emphasizing the aspect of data quality. Taking these different extensions of the original definition into account, we define Big Data as *a phenomenon characterized by an ongoing increase in volume, variety, velocity, and veracity of data that requires advanced techniques and technologies to capture, store, distribute, manage, and analyze these data*.

The economic potential of Big Data is as diverse as the data itself and the key driver for organizations to adopt Big Data analytics. There are four Big Data categories organizations can leverage: external structured data such as Global Positioning System (GPS) or credit history data, internal structured data such as Customer Relationship Management (CRM) or inventory data, external unstructured data such as Facebook or Twitter posts, as well as internal unstructured data such as text documents and sensor data [12]. All four categories have specific characteristics that certain Big Data strategies may address better than others. Application fields of Big Data cover a wide range of industries and businesses. Suggestions range from health care (reduction of costs resulting from over- and under-treatment) with a $300 billion annual potential, to the public sector and e-government (more efficiently collection of taxes and service quality improvement from education to unemployment offices) with a $250 billion annual potential, over e-commerce, marketing and merchandising (better understanding of consumers with respect to product and price preferences) with a potential of 60% increase in operating margins [8, 9]. These fields' potentials are unlocked by the application of different Big Data techniques such as crowdsourcing, data fusion and data integration, natural language processing, network analysis, predictive modeling, simulation, and visualization. Thus, the application possibilities and economic potentials of Big Data technologies are enormous and executives should assess whether and how they could make use of these potentials.

3. Methods

Big Data is still a new and emerging field of research. Consequently, our understanding of Big Data's basic constituents is still fragmented. By identifying different Big Data strategies and their facilitating conditions, we hope to contribute to the field's knowledge. We rely on the literature review methodology because of its ability to expose theoretical foundations and uncover research potentials [13]. We followed established guidelines for literature reviews [e.g. 14, 15], however, also included industry reports and best practices. This approach has been recommended for newly emerging research themes [15].

Our review follows a three-staged literature analysis in the fields of Big Data analytics, data warehousing, and business intelligence. First, we searched and analyzed existing knowledge to identify a set of corporate Big Data strategies. We particularly looked at the different categories of Big Data depicted in the previous section (i.e. external or internal (un-)structured data) and analyzed the capabilities of existing data analysis approaches with respect to how well they can handle the various Big Data categories. Second, we systematically analyzed the literature to identify context factors that may influence Big Data strategy choice. To that end, we reviewed different factors affecting architecture choice in traditional data warehousing environments, evaluated them regarding their relevance for Big Data analytics, and summarized them in a concept matrix [14]. On the basis of this concept matrix, three groups of factors have been formed, each influencing the strategy decision differently. Finally, we linked the context factors to different types of strategies. The resulting strategy matrix not only provides guidance for practitioners in choosing a suitable Big Data strategy. It also allows deriving company profiles that are associated with different Big Data strategies.

4. Findings

Organizational executives should ask three questions in order to determine how to deal with Big Data [16]: (1) Do we have a Big Data problem or an IT infrastructure problem?; (2) Are we missing critical information that a Big Data solution will help to capture?; and (3) What are our analytical requirements? All three questions are of particular relevance for Big Data strategy choice. The first two questions aim at deciding whether actively targeting Big Data is the right way to solve an existing problem. If a decision for Big Data is made, executives face the challenge of deciding for the right strategy to implement Big Data analytics. To that end, however, the actual requirements of Big Data analysis as well as organizational contingencies have to be identified and potentially weighted against each other. The result is a Big Data strategy which is most suitable for the current organizational needs. Next, we will therefore first present the four basic strategies to Big Data we have identified, before we proceed with the discussion of the derived contingency factors.

4.1. Big data strategies

The four strategies described in this section are based on the suggestions of different vendors as well as different researchers [e.g. 8, 11, 17, 18, 19]. They can be ranked along a continuous scale from full integration into the organization's IT environment to no integration at all (Figure 1). The first strategy is to rely on traditional relational database management systems (RDBMS) for Big Data analysis. This strategy is located at the left end of the continuum in Figure 1 because no specific Big Data technology is integrated into an organization's current environment. A strategy that involves a somewhat stronger incorporation of Big Data technology into the existing environment is the Big Data Analytics as a Service (BDAaaS) approach: while the organization's IT infrastructure remains untouched, often an additional frontend has to be integrated and managed to upload company-internal data into the cloud and access the analytical capabilities. The remaining two strategies require a considerable integration of Big Data technology into an organization's IT environment. For instance, for MapReduce and distributed file systems (DFS) the integration of a completely new software stack into the current application and infrastructure environment is needed. However, this integration does typically not require the adaptation of other enterprise solutions as this is the case with hybrid approaches. Therefore, the hybrid strategy is located at the right end of the continuum in Figure 1. It should be noted that these four strategies are not exhaustive. There are several other strategies as well, that however build upon these four and represent mixed forms or extensions of the strategies depicted here (for further input on that see, for example [9] or [20]).

Each of the four strategies bears specific challenges, opportunities, and architectural peculiarities that have to be considered when choosing one of them. We will discuss all the strategies next by investigating their capabilities concerning volume, variety, and velocity of data. None of the solutions discussed in practice and research adequately addresses veracity, yet [11]. Accordingly, we dropped this facet in our discussion.

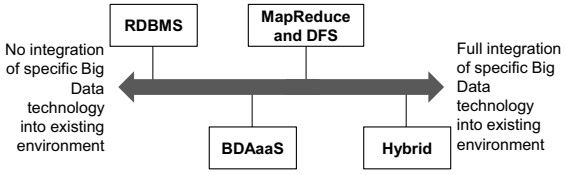

Figure 1. Degree of infrastructure integration of the four presented Big Data strategies

4.1.1. RDBMS. The first and probably most obvious way of dealing with Big Data is by using traditional data warehousing architectures based on standard RDBMS. In this case, data are extracted from various internal and external sources, selected, aggregated, and loaded into a data warehouse. Different business intelligence tools can then be used to analyze and access the data. As volume and velocity of the data to be processed steadily increased since the 1980 [19], most contemporary companies revert to parallelized RDBMS to handle the large amounts of data [21]. Consequently, data are stored on multiple machines, tables are partitioned over the nodes in a cluster, and an application layer allows for accessing the different data portions on the different nodes. The goal of such an architecture is to provide linear speed-up as well as scale-up [21]. This means that twice as much hardware allows for execution of twice as large tasks in the same elapsed time. However, the effort necessary to keep the systems synchronized does not allow for completely linear scale-up or speed-up [22].

A considerable advantage of parallel RDBMS is that they can handle and analyze large volumes of data very fast and stable. For example, Pavlo et al. [19] benchmarked a parallel RDBMS and a MapReduce solution (see section 4.1.2) with each other. Execution of standard data analysis tasks such as selection and joining was significantly faster in the RDBMS than in the MapReduce solution. Hence, looking at the *volume* characteristic of Big Data, RDBMS are a suitable solution.

However, considering the RDBMS capabilities with regard to *variety* and *velocity* of data reveals several problems. As RDBMS are optimized towards data analysis, the loading of data is typically very cumbersome and time-consuming. For instance, Pavlo et al. [19] measured an 11-times longer loading time for the RDBMS than for the MapReduce solution – and they only relied on simple plain text documents. Consequently, processing unstructured input data, if even possible, worsens the loading results of an RDBMS [23]. This problem is even intensified by the fact of velocity; that means data might have to be loaded very frequently. Keeping unstructured, rapidly changing data in an RDMBS is, therefore, way more expensive in terms of processing time than, for example, a MapReduce approach.

Still, many traditional data warehousing and RDBMS vendors are driving further the development of their solutions towards Big Data. For example, HP banks on the usage of massive, parallel SQL databases. They are integrating a large amount of new in-memory analytic functions and new technologies to easily expand or downsize deployments [24].

In summary, (parallel) RDBMS architectures can be a suitable strategy of approaching Big Data, as long as new data is not very frequently loaded into the system and tasks involving a high amount of unstructured data are performed relatively seldom. Also, this approach is worth considering, if the change affinity of an organization is rather low: RDBMS are in fact in place in every modern organization; setup costs are, thus, very low. According to a recent study of BARC, a large European business analysis company, especially the latter is a driving factor for the fact that 89% of the companies in their study revert to RDBMS approaches or plan to do so, while less than 20% apply "pure" Big Data solutions [20].

4.1.2. MapReduce and DFS. The second most often referenced strategy to approach Big Data analysis is the introduction of new systems that use distributed file systems (DFS) and a MapReduce engine. A prominent exponent of such a system is Hadoop (although Hadoop does not exactly follow the MapReduce algorithm) [9]. Hadoop is an open source architecture composed of different engines such as a MapReduce engine and a DFS engine. The data to be analyzed is stored in the distributed file system and then processed using the MapReduce engine. The results are then again stored in the file system and directly streamed to a business intelligence application. In the MapReduce approach, unlike as in RDBMS, small programs are necessary to execute queries [25]. To be more precise, "users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key" [25, p. 107]. These programs are then injected into a distributed processing framework, that decides how many map and reduce instances have to be run on which nodes [19].

Looking at opportunities and drawbacks of the MapReduce strategy, we come to an inverted picture compared to the RDBMS strategy. On the one hand, MapReduce lacks in its processing capabilities: the execution of standard tasks such as select and join can run up to 90 times slower on MapReduce systems than on parallel RDBMSs [9, 19]. Additionally, it takes significantly longer to write a MapReduce program than an SQL query. Therefore, in environments in which large *volumes* of data have to be frequently analyzed or in which analysis objectives (i.e. the underlying queries) change frequently, the time required for programming and waiting for analysis results hinders many organizations from implementing such a system [26].

On the other hand, the use of MapReduce systems allows organizations to perform advanced analysis of unstructured data such as text files or Facebook posts, when traditional business analytics based on RDBMS experience severe problems [9]. Additionally, loading of data into a MapReduce environment only takes tenth of the time needed in RDBMS [19].

To sum up, in cases in which it was too expensive in RDBMS environments to load masses of sensor data, call details, or weblogs into a data warehouse, the MapReduce strategy offers new possibilities for organizations to cope with large data *variety* and *velocity* [26]. However, MapReduce systems are potentially not the best approach in environments with frequently changing inquiry patterns. Companies choosing this approach should also schedule their analyses wisely, because they are very time-consuming. Quick, ad-hoc queries are rather problematic for MapReduce systems. Another facet to be considered before deciding for a MapReduce strategy are setup costs. While the purchasing costs of, for example, Hadoop can be regarded rather low [27], migration, consulting, and training efforts may increase overall costs quickly. Still, however, more improvements and developments are likely to be achieved within the next years. In the last few years, also the big software providers as Oracle, IBM, or Teradata offer pure MapReduce solutions, which they continuously enhance to cope with the rising demand of their customers to perform Big Data analytics.

4.1.3. Hybrid approach. Both aforementioned Big Data strategies, traditional RDBMS as well as MapReduce and DFS, have potentials as well as limitations. Thus, the introduction of a hybrid solution combining the benefits of both approaches seems reasonable and is called for by different authors as well [10, 28]. In general, there are three possible solutions to integrate MapReduce and RDBMS systems [26]. In a MapReduce-dominant approach, the MapReduce technology is extended with relational components leading to efficient processing of large volumes of structured as well as unstructured data, while profiting from typical RDBMS-strengths such as query optimization. However, performance gains in query processing are limited, since the relational components can typically only be integrated on several nodes [18]. RDBMS-dominant hybrids have integrated MapReduce capabilities in their engines to particularly improve processing abilities for unstructured data. Although it has been shown that this approach tends to produce faults under certain circumstances [18], it is the one most often used by commercial vendors [26, 28]. The third and least frequently adopted approach is a loose

coupling of MapReduce systems and RDBMS [28]. The loose-coupled approach might appear most valuable on first sight because it theoretically allows for connecting any MapReduce system with an RDBMS. Lacking interface standardization however causes problems such as complicated data transfer or optimization between both systems [18]. Due to this problem, the benefit of "simply" connecting some systems proves fallacious to some degree leading to the just mentioned low adoption rates.

In summary, regardless of which approach a company chooses, a hybrid strategy allows to efficiently handle Big Data without struggling with some of the problems depicted for the pure RDBMS or MapReduce strategy. Still, performance of hybrid systems does not – at least at the moment – exceed that of uncombined strategies [18]. Rather, such systems process and analyze data with large volume, variety and velocity within acceptable performance and fault boundaries. From a financial point of view, the hybrid strategy approach is more expensive than an uncombined strategy [27]. While the potentially high licensing costs can predominantly be ascribed to the RDBMS involved, often more processing power and storage is needed leading to higher hardware expenses. However, researchers as well as software vendors are currently putting efforts in the development of hybrid solutions that become increasingly performant. Prominent examples are HadoopDB (MapReduce-dominant) [28, 29], the Oracle in-database Hadoop [30], Microsoft, who announced their PolyBase technology in November 2012 [31], or Greenplum [32], who all build up on a traditional RDBMS system (RDBMS-dominant).

4.1.4. Big Data Analytics as a Service. The fourth strategy for dealing with Big Data is to buy in Big Data capabilities. The supply of BDAaaS solutions is rapidly increasing and the variety of vendors is large. Tresata, for instance, has specialized on analyzing banking data [33]. Another example is Cloudera [17], who offers a large variety of BDAaaS solutions for different industries. Therefore, organizations tackling the BDAaaS strategy are likely to find suitable solutions for their specific context.

As the infrastructure for BDAaaS is hosted in the cloud, the costs for BDAaaS in organizations are far more flexible than the costs of implementing in-house Big Data solutions. In addition, the economies of scale and scope realized at the side of the vendor allow them to perform Big Data analytics more efficiently than an average company could do [34]. This is especially interesting for smaller organizations, which often do not have sufficient resources and expertise to realize Big Data analytics cost- and resource-efficient in-house and thus have an interest in utilizing off-the-shelf solutions.

However, BDAaaS has limitations as well. The most critical limitation arises from the data privacy and security discussions that are relevant for all cloud-based services. Especially in operating environments in which sensitive data has to be analyzed in the cloud, an encryption during the data transmission is inevitable. Operating with encrypted data however is yet a problem for most BDAaaS solutions [34]. Further, data policy is already an important issue, when Big Data is analyzed using internally-secured infrastructure. Therefore, shifting such analyses into the cloud may be problematic for some companies, if not impossible. Besides, there is no definite answer, yet, on how to exchange data between cloud providers and company-internal infrastructure [34]. As a result, many cloud providers offer Big Data analytics only for company-external data that can be retrieved without access to internal infrastructures, or they limit their analytics to specific platforms (e.g. Cloudera [17]). Altogether, BDAaaS is particularly valuable for organizations for which the implementation of own Big Data analytics is either too expensive, limited to external data, or not of strategic importance.

4.2. Contingency factors for strategy choice

Choosing an appropriate Big Data strategy, requires consideration of various contingencies such as current business strategy, understandability of Big Data insights to end-users, and structure and complexity of organizational processes [2]. Based on our analysis of the Big Data and data warehouse literature (see Section 3), we identified eight different factors that were grouped into three dimensions: strategic factors, resource factors and operating environment factors (Figure 2). In the following paragraphs, the eight contingency factors will be introduced and analyzed with regard to their specific impact on Big Data strategy choice. Based on this analysis, we derive a contingency matrix that provides decision support for practitioners in Big Data strategy choice.

4.2.1. Strategic factors. The first contingency factor in this group is the *organizational position of Big Data analytic*s. Typically, an organization's preference for a certain IT infrastructure is influenced by their strategic objectives [10, 35]. Data analytics have traditionally played an important role in customer-oriented businesses such as retailing, marketing, or banking [6]. Not surprisingly, many of the companies who are first movers in Big Data

analytics or at least very interested in the technology are also from these industries [20]. With the possibility of making use of Big Data technologies, new differentiation possibilities emerge, and the state of competition in the respective industry sectors

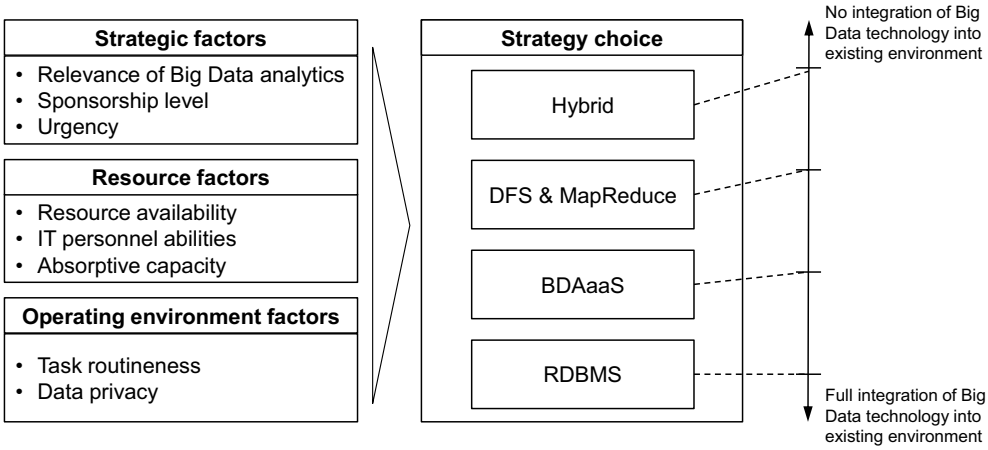

Figure 2. Contingency factors relevant for Big Data strategy choice

might be shifted [8]. Beyond industry specifities, LaValle et al. [2] identified three different Big Data application scenarios (justify actions, guide actions, and prescribe actions) that have direct implications for choosing a Big Data strategy. In line with these authors, we argue that companies who have a strong interest in guiding or prescribing actions will regularly have to analyze external as well as internal data pools. A need for sophisticated extraction and loading capabilities results as they can be found in the MapReduce or also in intelligent hybrid approaches. Furthermore, due to high strategic relevance, setup costs are probably not too relevant for such companies [27].

In organizations in which there is a rather low strategic position assigned to data analysis, facing Big Data with traditional RDBMS approaches can display a worthwhile alternative. Only when the strategic importance of data analysis is on the lower end, a BDAaaS approach should be considered.

As with many other IS investments, the decision for a Big Data strategy also heavily depends on the *sponsorship level* of the planned project within the organization [12, 35]. Sponsorship comprises the allocation of resources for a project, the fighting off of political resistance, and the promotion of the project's benefits within the organization. Accordingly, also in the case of Big Data, sponsorship level from business units as well as upper management is an important factor for strategy choice. Consequently, IT organizations that face a rather low sponsorship for Big Data should revert to BDAaaS approaches or to existing infrastructures such as RDBMS [36].

„Business environments with greater turbulence will experience greater urgency for information systems that provide accurate information" [35, p. 203]. *Urgency*, hence, refers to how timely Big Data analytics are required. Switching from a traditional data warehousing approach (based on RDBMS) to a hybrid or MapReduce approach is time-consuming [19, 29]. On the other hand, keeping or extending an existent RDBMS may yield results earlier than switching to a new system, thus, reducing the time to value [2]. Also employing BDAaaS can display a worthwhile alternative in times of high urgency. However, considering the large amount of different BDAaaS providers on the market, choosing the "right" approach (e.g. in terms of future costs) might be very time-consuming.

4.2.2. Resource factors. Resource factors refer to the resource constraints of an organization when deciding for a strategy to tackle Big Data. The first factor in this group is *resource availability* in terms of money and human resources [12, 36]: the amount of resources available for implementation impacts strategy choice. While large organizations can afford implementing more complex and comprehensive solutions such as hybrid approaches [5], small companies and start-ups may want to keep costs low or at least flexible [27]. Consequently, they might think about a MapReduce or BDAaaS approach. This is also fostered by the fact that startups in contrast to large companies have not a grown infrastructure; introducing completely new solution is not a big issue for them [10, 27].

While organizational size is often positively correlated with the availability of resources [5, 37], it can be argued that start-ups, although they have limited budget and employees, often have knowledgeable staff who are willing to implement technological-edge solutions. Ravichandran [38] also showed that the adoption and development of complex technologies is impeded by knowledge barriers. Therefore, the *abilities of the IT personnel* are an important contingency factor for Big Data strategy choice. More experienced and skilled teams will be able to implement systems of higher complexity such as MapReduce or hybrid approaches, that require deep knowledge from the company's IT personnel not only while implementation, but also during operation [5, 9]. One possibility for creating sufficient expertise within the organization is training the existent staff. Another way is hiring additional staff. Both possibilities, however, require activation of financial resources. In cases in which this is not possible, an RDBMS or BDAaaS strategy should be pursued.

The third contingency factor we identified is *absorptive capacity*. Absorptive capacity is the "ability of key organizational members to utilize available or preexisting knowledge" [2, 37]. It has been shown to be a key driver for an organization's innovativeness [39], particularly concerning the use of analytics to improve the business [2]. As the potential of Big Data is rather diverse and exceeds the possibilities of traditional data warehousing systems, absorptive capacity influences the choice of a Big Data strategy. If employees do not understand the value of an innovative IT system, they will rarely make use of it [37]. Therefore, introducing a MapReduce system with its completely new opportunities might be difficult for companies with low absorptive capacity. Such companies should rather use traditional RDBMS or a BDAaaS strategy to tackle Big Data analysis.

4.2.3 Operating environment factors. These factors describe the circumstances of daily operations. The first factor in this group is *task routineness*. Routine tasks are structured and repetitive, whereas non-routine tasks occur in ad hoc situations and require unique solutions [40]. The more routine tasks are in a company, the less processing capacity in data analytics is needed [35], because optimized queries can be re-used many times. Due to the complex programming language it takes rather long to formulate queries on typical Big Data systems based on MapReduce [19]. This fact makes the MapReduce approach less appropriate for environments with low task routineness in which frequent manipulation of queries is highly realistic. Instead, RDBMS or hybrid approaches are very valuable. However, the MapReduce strategy is valuable in environments with high task routineness, because regularly the same queries are run and new data are frequently loaded.

The second contingency factor of the operating environment is *data privacy*. With information from customers being multiplied and shared across the globe, privacy is a crucial issue more than ever [41]. Therefore, policy makers must address the problems in privacy legislation and understand both, merit and danger of Big Data collection [42]. Furthermore, with rising economic interest in Big Data, it becomes increasingly interesting for attackers [8]. Hence, data privacy has not only legal, but also technological implications for a company's Big Data strategy. Where data privacy is of high relevance, a BDAaaS strategy is not suitable because the company has only very limited impact on where the data is stored and processed [34]. Also, choosing a MapReduce strategy could be an issue because security of distributed file systems (which are used in MapReduce strategies) still shows various security weaknesses. Therefore, as of today, utilizing RDBMS-based approaches depicts probably the most secure strategy.

5. Synthesis and discussion

Based on the above discussion, we find that different organizational environments pursue different requirements on a Big Data strategy. To better support practitioners in Big Data strategy choice, we compared the four identified Big Data strategies regarding how well they addressed each of the contingency factors. We have summarized our findings in the contingency matrix in Table 1. For instance and as discussed above, when the relevance of Big Data analytics is high in a company, the MapReduce strategy seems most fruitful (resulting in a "+" assessment) (see Section 4.2.1). However, also a hybrid solution might be valuable in case it follows a MapReduce-dominant approach [18]. If in turn an RDBMS-dominant implementation is chosen, the hybrid strategy is only slightly better than a "pure" RDBMS approach (resulting in a "+/o" assessment).

Several patterns are observable for the different factor categories. For instance, we can see that BDAaaS is negatively associated with nearly all of the contingency factors. By contrast, we find both the MapReduce as well as hybrid strategy being positively associated with most of them. The RDBMS strategy again has turned out to be a suitable approach in nearly all situations as long as expectations towards Big Data processing performance are not too high. In cases in which

performance plays a more important role, hybrid approaches comprise a workable trade-off between costs, processing, and analysis performance. We therefore argue that the RDBMS strategy might be a valuable option for companies who want to explore the potential of Big Data analytics. As soon as requirements towards Big Data processing increase, switching to a hybrid or MapReduce strategy should be considered. This finding is consistent with Lavalle et al. [2], who suggest introducing Big Data stepwise: starting with existing RDBMS, companies should consider more sophisticated approaches such as hybrid or MapReduce systems when sufficient experience with Big Data analytics has been gained and strategic potentials can be clearly articulated.

Looking at specific industries and company structures, several insights can be derived. For instance, in manufacturing companies ERP systems and relational databases are traditionally highly relevant. Hence, the IT employees are adept at using and implementing such systems. Additionally, the manufacturing systems in modern factories produce large amounts of heterogeneous data that call for advanced analytical techniques, which must allow for frequent and fast execution of standardized and ad-hoc queries (e.g. to quickly determine production barriers or disturbances). Considering these analytical requirements together with the specific RDBMS abilities of the IT employees, we find that companies of the production sector would benefit from an RDBMS strategy most. In turn, in retailing and marketing companies, for which data analytics have a high strategic relevance, IT personnel is typically well trained in data analytics, and the data to be analyzed changes frequently, a MapReduce approach is very promising. However, looking at the operating environment, issues of data privacy are typically of high importance as well. Hence, a hybrid strategy might be an option worth considering, especially in cases in which existing organizational data warehouse infrastructures have to be reused.

Table 1. Impact of identified contingency factors on Big Data strategy choice

	RDBMS	BDAaaS	MapReduce	Hybrid
Strategic factors				
Relevance of Big Data analytics	o	-	+	o/+
Sponsorship level	o/-	-	+	o/+
Urgency	+	o	-	o
Resource factors				
Resource availability	o	-	o	+
IT personnel abilities	o	-	+	+
Absorptive capacity	-	-	+	o/+
Operating environment factors				
Task routineness	o	o	+	o/-
Data privacy	+	-	o/-	o/+

o = Neutral + = High - = Low

The BDAaaS approach is (as many other cloud-based applications are) most useful for smaller usage scenarios. For instance, a company which might want to explore the potentials of Big Data analytics may perform a pilot with a BDAaaS approach before introducing extensive in-house solutions that also perform Big Data analyses on internal data. Also small or family businesses and start-ups should consider such an approach because it offers higher flexibility than an internal solution.

Altogether, we conclude that the combination of traditional RDBMSs with DFS and MapReduce technologies is a workable approach in many Big Data scenarios. This finding might not be surprising and is also underlined by the rising attention hybrid approaches receive in various publications [e.g. 18, 29, 43]. However, it should be noted that hybrid systems are still far from reaching performance of uncombined approaches (see also Section 4.1.3). In particular, the appropriateness of a hybrid solution is highly dependent upon whether a MapReduce- or RDBMS-dominant implementation was used. For instance, to date many RDBMS-dominant approaches tend to produce erroneous results under certain circumstances [18, 19]. Hence, choosing a hybrid without deep understanding of its structural

characteristics may lead to unsatisfactory results and should be avoided.

6. Conclusion

In this paper, we aimed at providing guidance for companies on how to approach the phenomenon of Big Data. Based on a review of existing scientific as well as practitioner literature, we identified four Big Data strategies and discussed them regarding contingencies influencing strategy choice. The eight respective contingency factors can be grouped into three dimensions, namely strategy, resources, and operating environment. Although other authors have already discussed context factors that might influence Big Data strategy choice [e.g. 36, 44], a structured analysis of such contingency factors has not been performed so far. We therefore contribute to the still limited research on Big Data by providing a basis for future discussions on the adequacy and success of various Big Data strategies for differing corporate environments. As illustrated by the analysis of the opportunities and challenges of Big Data in this paper, organizational decision makers need to start thinking about whether and how to facilitate Big Data analytics. They therefore benefit from our research by gaining a better understanding of the different facets they should consider before deciding on Big Data solution investments.

Bearing these contributions in mind, our research also has limitations. One limitation relates to the literature search and selection which might be biased by subjective interpretations and preferences. Beyond that, due to scarcity of scientific research on Big Data we also relied on practitioner literature, for which objectivity of findings cannot always be assured. This scarcity clearly calls for more research, especially as most research focused on technological aspects, while only a very limited number also deals with organizational [9] or societal aspects [41, 42]. Thus, we believe that future research should make efforts to understand the impact Big Data may have on corporate environments as well as on society. For instance, expert interviews or exploratory case studies could be used to confirm, refine, and extend or contingency matrix.

7. References

[1] Park, Y.-T., "An Empirical Investigation of the Effects of Data Warehousing on Decision Performance", Information & Management, 43(1), 2006, pp. 51-61.

[2] Lavalle, S., Lesser, E., Shockley, R., Hopkins, M.S., and Kruschwitz, N., "Big Data, Analytics and the Path from Insights to Value", Sloan Management Review, 52(2), 2011, pp. 20-31.

[3] Argyris, C., and Schön, D.A., Organizational Learning: A Theory of Action Perspective, Addison-Wesley Pub. Co., 1978.

[4] Laney, D., 3d Data Management: Controlling Data Volume, Velocity, and Variety, Stanford, 2001.

[5] Almeida, F., and Calistru, C., "The Main Challenges and Issues of Big Data Management", International Journal of Research Studies in Computing, 2(1), 2013, pp. 11-20.

[6] Ang, J., and Teo, T.S.H., "Management Issues in Data Warehousing: Insights from the Housing and Development Board", Decision Support Systems, 29(1), 2000, pp. 11-20.

[7] Shirer, M., and Murray, P., Idc Predicts: 2012 Will Be the Year of Mobile and Cloud Platform Wars as It Vendors Vie for Leadership While the Industry Redefines Itself, IDC Analyze, 2011.

[8] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Hung Byers, A., "Big Data: The Next Frontier for Innovation, Competition, and Productivity", in (Editor, 'ed.'^'eds.'): Book Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011

[9] Chen, H., Chiang, R.H.L., and Storey, V.C., "Business Intelligence and Analytics: From Big Data to Big Impact", MIS Quarterly, 36(4), 2012, pp. 1165-1188.

[10] Dumbill, E., What Is Big Data? An Introduction to the Big Data Landscape., Strata, 2012.

[11] Morgan, T., Ibm Global Technology Outlook 2012, Warwick, 2012.

[12] Nair, R., and Narayanan, A., Getting Results from Big Data - a Capabilities-Driven Approach to the Strategic Use of Unstructured Information, Booz & Company, 2012.

[13] Viering, G., Legner, C., and Ahlemann, F., "The (Lacking) Business Perspective on Soa – Critical Themes in Soa Research", 9. Internationale Tagung Wirtschaftsinformatik, 2009, pp. 45-54.

[14] Webster, J., and Watson, R.T., "Analyzing the Past to Prepare for the Future: Writing a Literature Review", MIS Quarterly, 26(2), 2002, pp. xiii-xxiii.

[15] Cooper, H.M., Synthesizing Research - a Guide for Literature Reviews, Sage Publications, 3rd edn, 1998.

[16] Stoller, J., "Assessing the Big Data Challenge", CMA Magazine (1926-4550), 86(2), 2012, pp. 18-19.

[17] Cloudera, The Platform for Big Data and the Leading Solution for Apache Hadoop in the Enterprise, http://www.cloudera.com/content/cloudera/en/home.html, 2013.

[18] Gruska, N., and Martin, P., "Integrating Mapreduce and Rdbmss", Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research, 2010, pp. 212-223.

[19] Pavlo, A., Paulson, E., Rasin, A., Abadi, D.J., Dewitt, D.J., Madden, S., and Stonebraker, M., "A Comparison of Approaches to Large-Scale Data Analysis", Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, 2009, pp. 165-178.

[20] Bange, C., Grosser, T., and Janoschek, N., "Big Data Survey Europe - Nutzung, Technologie Und Budgets Europäischer Best Practice Unternehmen", in (Editor, 'ed.'^'eds.'): Book Big Data Survey Europe - Nutzung, Technologie Und Budgets Europäischer Best Practice Unternehmen, Business Application Research Center (BARC), Würzburg, 2013

[21] Dewitt, D., and Gray, J., "Parallel Database Systems: The Future of High Performance Database Systems", Commun. ACM, 35(6), 1992, pp. 85-98.

[22] Oracle, Parallel Processing & Parallel Databases, Oracle Corporation, 1997.

[23] Jacobs, A., "The Pathologies of Big Data", Commun. ACM, 52(8), 2009, pp. 36-44.

[24] Kelly, J., Vellante, D., and Floyer, D., Big Data Market Size and Vendor Revenues, Wikibon, 2013.

[25] Dean, J., and Ghemawat, S., "Mapreduce: Simplified Data Processing on Large Clusters", Commun. ACM, 51(1), 2008, pp. 107-113.

[26] Xu, Y., Kostamaa, P., and Gao, L., "Integrating Hadoop and Parallel Dbms", Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 2010, pp. 969-974.

[27] Proffitt, B., Hadoop Vs. An Rdbms: How Much (Less) Would You Pay?, ITworld, 2012.

[28] Stonebraker, M., Abadi, D., Dewitt, D.J., Madden, S., Paulson, E., Pavlo, A., and Rasin, A., "Mapreduce and Parallel Dbmss: Friends or Foes?", Commun. ACM, 53(1), 2010, pp. 64-71.

[29] Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D., Silberschatz, A., and Rasin, A., "Hadoopdb: An Architectural Hybrid of Mapreduce and Dbms Technologies for Analytical Workloads", Proc. VLDB Endow., 2(1), 2009, pp. 922-933.

[30] Sun, H., and Heller, P., Oracle Information Architecture: An Architect's Guide to Big Data., Oracle, Redwood Shores, 2012.

[31] Lab, G.S., Polybase, Microsoft, Madison, 2012.

[32] Greenplum, A Unified Engine for Rdbms and Mapreduce, www.greenplum.com/technology/mapreduce/, 2013.

[33] Kelly, J., Tresata Goes Deep on Big Data for Banking, Silicon Angle, 2012.

[34] Demirkan, H., and Delen, D., "Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud", Decision Support Systems, 55(1), 2013, pp. 412-421.

[35] Ariyachandra, T., and Watson, H., "Key Organizational Factors in Data Warehouse Architecture Selection", Decis. Support Syst., 49(2), 2010, pp. 200-212.

[36] Russom, P., Big Data Analytics, TDWI Research, 2011.

[37] Ramamurthy, K., Sen, A., and Sinha, A.P., "An Empirical Investigation of the Key Determinants of Data Warehouse Adoption", Decision Support Systems, 44(4), 2008, pp. 817-841.

[38] Ravichandran, T., "Innovation Assimilation in the Presence of Knowledge Barriers, Technology Uncertainty and Adoption Risks", Academy of Management Proceedings, 2001(1), 2001, pp. C1-C6.

[39] Sambamurthy, V., and Zmud, R.W., "Arrangements for Information Technology Governance: A Theory of Multiple Contingencies", MIS Quarterly, 23(2), 1999, pp. 261-290.

[40] Goodhue, D.L., and Thompson, R.L., "Task-Technology Fit and Individual Performance", MIS Quarterly, 19(2), 1995, pp. 213-236.

[41] Tene, O., and Polonetsky, J., "Privacy in the Age of Big Data", Stanford Law Review OnlineFebruary), 2013,

[42] Craig, T., and Ludloff, M.E., Privacy and Big Data, O'Reilly, Sebastopool, 2011.

[43] Su, X., and Swart, G., "Oracle in-Database Hadoop: When Mapreduce Meets Rdbms", Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012, pp. 779-790.

[44] Mcafee, A., and Brynjolfsson, E., "Big Data: The Management Revolution. (Cover Story)", Harvard Business Review, 90(10), 2012, pp. 60-68.