Big Data – A State-of-the-Art

Marco Pospiech
Chair of Information Systems, Technical University Freiberg, Freiberg, Saxony, Germany., marco.pospiech@bwl.tu-freiberg.de

Carsten Felden
Chair of Information Systems, Technical University Freiberg, Freiberg, Saxony, Germany., carsten.felden@bwl.tu-freiberg.de

Big Data – A State-of-the-Art

Marco Pospiech
TU Bergakademie Freiberg
marco.pospiech@bwl.tu-freiberg.de

Carsten Felden
TU Bergakademie Freiberg
carsten.felden@bwl.tu-freiberg.de

ABSTRACT

The term *Big Data* has an increased and tautological occurrence in scientific publications. It is of interest how and whether the data provisioning is able to support enterprises in the handling and value creation of this emerging issue. Considering the amount of growing publication and the fuzzy nature of this term, an overview is requested to avoid duplications to gain relevant findings and to identify potential research gaps. To address this issue, a general literature review is accomplished, which extrapolates and clusters discussed research fields and potential gaps. It becomes apparent that a huge part of the research is technical driven. Moreover, no identified paper addresses the research area of functional data provisioning. This initiates further investigations to discuss whether Big Data itself negate such intention or research has missed it and improvements regarding Big Data are possible.

Keywords

Big Data, State-of-the-Art, Decision Support, Cloud Computing

INTRODUCTION

The long lasting discussion of an information supply affects the discussion of information flooding of decision makers in enterprises, too. *Big Data* seems to be another *brick in the wall* of this discussion. The amount of business data is globally growing exponential while their accompanied tasks of storing and processing are not able to follow anymore (Reddi et al., 2011). In this context, the vague and fuzzy term *Big Data* occurs increasingly in scientific discussions, (see Figure 1), but varies in research disciplines and definitions (Cuzzocrea et al., 2011; Bizer et al., 2011). It is the paper's goal to gain insight into the scientific discussion of the interdisciplinary term to address potential research gaps in the area of data provisioning. Hereby, findings for further design, development, and application of information and communication system can be obtained. Thus, the paper clusters Big Data relevant research to gain an overview on current discussions and research gaps.

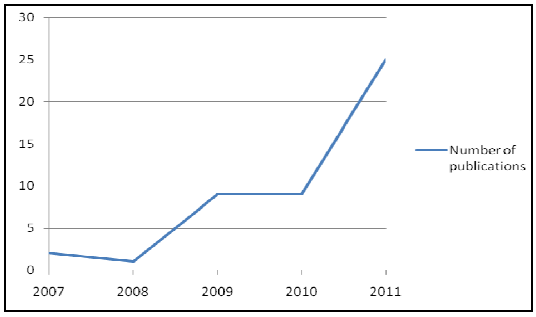

Figure 1. Number of evaluated Publications

According to Gartner, 85 percent of all enterprise infrastructures will be overwhelmed until 2015 by Big Data (Casonato et al., 2011). In this context, Big Data related issues can be addressed from a functional as well as from a technical perspective. Here, traditional database management, storage, and analytical capabilities do not provide acceptable performance according to existing business requirements. (Jacobs, 2009) But, the increasing demand in availability, variety, and complexity of new data sources requests advanced data management concepts (Casonato et al., 2011). But in fact, the challenge is not novel. During the 1970s, information concepts like Management Information System (MIS) failed due the emerging information flood (Koreimann, 1971). Such an issue was addressed by approaches like the Data Warehouse (DWH) concept during the

1990s. But now, Big Data refers to a challenge of an unfavorable ratio between available data and current information technologies or concepts. In consequence, there is a need for research and based on this the development of further solutions to address the demonstrated issue. Information lifecycle management (Robek et al., 1996), lean (Hicks, 2007) or value oriented information management (Sankar, 1998) are prominent examples. This approaches or in particular the demand and supply of information and the inherent value to gain an appropriate and task oriented reduction of data, by avoiding a loss of knowledge can be understood as functional data provisioning (Figure 2). Thus, it is the question of research, whether Big Data is addressed by both, functional and technical data provisioning aspects, or not. An unbalanced ratio can indicate a potential research gaps. A State-of-the-Art about the scientific Big Data literature seems to be appropriate to gain an overview about Big Data research fields to avoid duplication, to consider relevant findings, and to identify potential research gaps in favor of further activities.

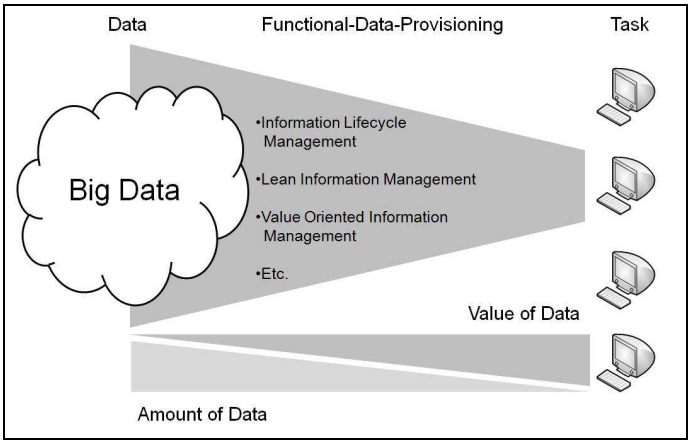

Figure 2. Blueprint Functional Data Provisioning

The concourse of the paper is as follows: the following section introduces the research design followed by a literature analysis. Hereby, the identified papers get evaluated and categorized to determine specific research fields and further research perspectives. Finally, the paper is summarized to discuss conclusions.

RESEARCH METHOD AND RELATED WORK

The first step of any research is the identification of already existing examinations with similar scientific aims. In this context, three State-of-the-Art studies (Agrawal et al., Cuzzocrea et al., 2011; Agrawal et al., 2010) are identified within the literature. Agrawal et al. (2011, 2010) provide a State-of-the-Art for scalable database management systems by discussing updates in heavy applications as well as in decision support systems for analytics in Big Data cloud infrastructures. Even Cuzzocrea et al. (2011) discuss the current State-of-the-Art of Big Data analytics and provide research questions. But, no paper follows a rigorous research method, thus the results are not comprehensible.

Existing scientific publications are analyzed to execute an academic State-of-the-Art. According to Cooper, a literature review represents a research method and is defined as a number of primary examinations with the similar questions and do not provide own new findings concerning to the research field. Hereby, the method scopes in describing, consolidating, integration, and analyzing results of primary investigations (Cooper, 1998). Moreover, a literature review can be described as a process and contains five phases (see Figure 3).

The *problem formulation* defines the academic goal of the review itself: research questions and hypothesizes are formulated. Suitable literature according to the problem formulation has to be identified during the second phase. In addition, the acquired literature will be proofed of relevance and classified within the *literature evaluation*. The results are analyzed and interpreted. Thus, results gets inspected and evaluated in respect to the addressed problem formulation. Finally, the generated results are presented in an appropriate manner. (Cooper, 1998)

LITERATURE ANALYSES IN RESEARCH FIELD BIG DATA

There are valuable insights from the various phases of the review process. This is the basis for further investigations in succeeding investigations. Here, the investigation follows the proceeding shown in Figure 3.

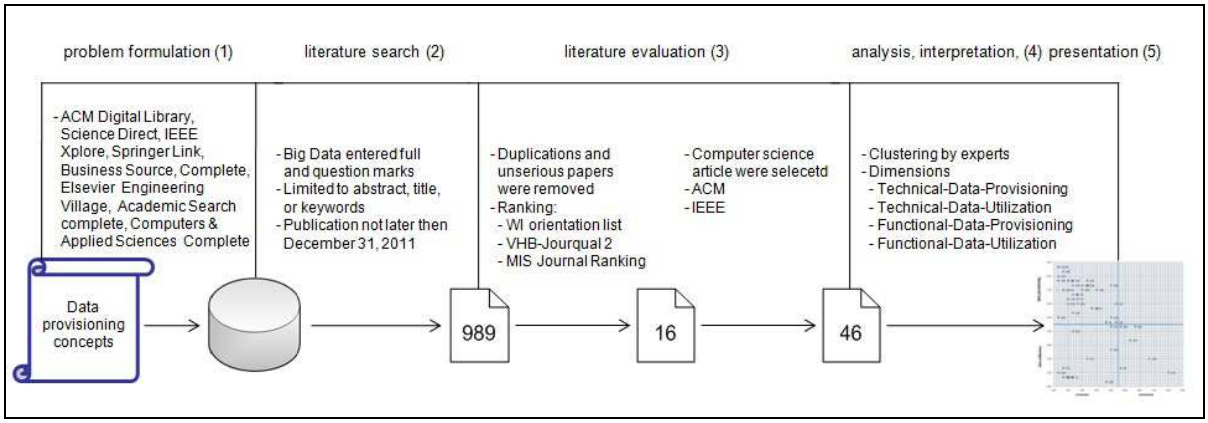

Figure 3. Process of the Literature Review

Literature search and evaluation

Figure 3 depicts the usage of several academic databases to identify relevant papers. To ensure comparability, the search item *Big Data* was requested in quotation marks and limited to abstract, title, or keywords. As a result, 989 hits regarding the search term have been obtained.

The literature evaluation serves in favor of proofing and securing quality and relevance. In an inital step, duplications and unserious papers were removed according to a four-eyes-principle. To address the stated problem statement, ranked papers were considered, only, to benefit from the basis of a peer reviewed process. To guarantee the domain relevance of the publications, well accepted rankings were used (AIS, 2007; VHB, Ulrich et al., 2008) and led to an amount of 16 papers. It became obvious that the biggest subset refers to computer science literature. To address this impact, computer science papers were considered as well. As a result, all peer reviewed papers provided by the two largest computer science associations ACM and IEEE were selected to remain quality, since the discipline itself has no general accepted journal ranking available. In this context, another 30 papers were embedded into the investigation, which led to 46 papers (listed in Table 1).

no.	reference	summary
1	(Wu et al., 2007)	The paper improves the decision tree algorithm efficiency by using the Sprint algorithm
2	(Wang, G. and Hu, F., 2007)	The article presents a attribute reduction algorithm that runs in a lower time complexity
3	(Zhang et al., 2008)	MapReduce in financial market to achieve better performance in option price predictions
4	(Cohen et al., 2009)	Philosophy and techniques of Magnetic, Agile, Deep data analysis in BI gets introduced
5	(Jacobs, 2009)	The work addresses underlying technical issues of analyzing Big Data
6	(Yan and Pi, 2009)	The article extends the k-means clustering algorithm by fuzzy logic
7	(Simmhan et al., 2009)	Presents a scalable software architecture, called GrayWulf for intensive cloud computing
8	(Reisser and Priebe, 2009)	Contains a metadata framework that administrate complex Big Data processes
9	(Panchaksharaiah, 2009)	The paper shows relational database techniques to overcome Big Data
10	(Freedman and Kisilev, 2009)	Introducing a new technique that segments videos or images in a less complexes manner
11	(Lukashevich et al., 2009)	Proposes an improved image classification by using a one class support vector machine

12	(Bower et al., 2009)	Use case in clinical epilepsy analysis; Multi scale annotation format gets introduced
13	(Agrawal et al., 2010)	The paper contains a State-of-the-Art of Big Data and cloud computing
14	(Eagle, 2010)	Studying how complex human behavior data can be used to predict risky behavior
15	(Alvaro et al., 2010	Presenting overlog (a declarative logical language) to implement an analytical stack
16	(Amer-Yahia et al., 2010)	Discussing impact of social networks to consumer information
17	(Bicking and Wimmer, 2010)	Text Mining techniques get compared that support analysis in government applications
18	(Li et al., 2010)	The article presents a reliable deduplicated key-value store to achieve space efficiency
19	(Foumier et al., 2010)	Use case in e-learning. Here, data gets analyzed to aim insights in learning networks
20	(Stonebraker et al., 2010)	Compares and quotes possible improvements of parallel databases and MapReduce
21	(Zhou et al., 2010)	Provides a keyword search MapReduce implementation for large-scale XML data
22	(Cuzzocrea et al., 2011)	Presents a State-of-the-Art of analytical Big Data technologies and cites current issues
23	(Agrawal et al., 2011)	State-of-the-art of Big Data; Article focus on scalable database management systems
24	(Khan and Hornbæk, 2011)	Debates non-volatile memory technology to improve storage performance
25	(Venkataraman et al., 2011)	The article presents a attribute reduction algorithm that runs in a lower time complexity
26	(Toole et al., 2011)	Use case in criminal analysis; Data mining methods detecting spatiotemporal patters
27	(Meijer, 2011)	Introduce LINQ as generalization of the relational algebra and SQL as basis for Big Data
28	(Li et al, 2011)	Presents DrayadLINQ, a declarative, data-centric language that addresses Big Data
29	(Huai et al., 2011)	Presents a model that abstracts critical behavior and interactions for Big Data analytics
30	(Bajda-Pawlikowski et al., 2011)	Paper compares commercial DBMS, Hadoop, PostgreSQL, VectorWise
31	(Dai et al., 2011)	Introduces HiTune, a scalable, lightweight and extensible Hadoop performance analyzer
32	(Das et al., 2011)	Paper loads a Predictive Model Markup Language defined data mining model in distributed databases segments
33	(Reddi, et al., 2011)	Analyzes energy-efficiency of CPU approaches in Big Data environments
34	(Silberstein, et al., 2011)	Article analyzes performance of feed following and provides a view selection approach
35	(Bizer et al., 2011)	Discuses different challenges regarding to Big Data and semantic technologies
36	(Cramer et al., 2011)	The article highlights social research possibilities offered by mobile devices
37	(Alexander et al, 2011)	Discuss issues and changes of Big Data processing in scientific and engendering tasks.
38	(Satzger et al., 2011)	Shows issues of data management systems; Introduce a Big Data streaming technology

39	(Han et al., 2011)	The paper presents a current state-of-the-art of NoSQL technology
40	(Chen, 2011)	Big Data in the financial market to predict further stock markets based on Data Mining
41	(Junwei et al., 2011)	Addresses the overload of message handling due MapReduce framework; Suggest a message oriented approach
42	(Qin et al., 2011)	Paper introduce a parallelizable iterative-style prediction based Percentile algorithm
43	(Kaushik et al., 2011)	Presents GreenHDFS; an energy-conserving variant of the Hadoop distributed file system that uses a supervised machine learning technique to identify rare used data
44	(He et al., 2011)	Introduce a placement structure called RCFile and its implementation in Hadoop
45	(Lee et al., 2011)	Proposes a system called YSmart, a correlation aware SQL-to-MapReduce translator
46	(Kohlwey et al., 2011)	Presents a prototype system for searching in cloud-scale biometric data

Table 1. Summary Selected Publications

Literature analysis and interpretation

The topic *Big Data* has an increased occurance within the scientific discussion. In this context, a significant raise of publications can be observed from 2010 to 2011. Almost 70 percent of all ranked papers were published in the recent two years. Hence, the relevance of the topic in the discipline can be assumed.

The tautological usage of the term *Big Data* demands a categorization of papers based on similarities. Here, independent research fields can be explored in order to be able to meet the research aim. In this context, the two dimensional framework of Gluchowski (2001) is chosen. This approach allows a positioning of diverse concepts and applications concerning to a specific research field. According to him, the system is stretched in the vertical axe by data processing forms and in the horizontal perspective by orientation perspectives (Figure 3). The range of technology elements reaches from data driven components to application specific presentation and analytical forms. Here, the upper section contains mainly data storage and provision approaches, whereas the lower section addresses methodic components and aspects where offered data is utilized. In consequence, the left side corresponds more to technique relevant attributes as the right one. Clusters positioned in the middle addressing several dimensions. Regarding to the research question and the chosen approach of Gluchowski, the data utilization cluster will be described more briefly to maintain completeness and rigorousness. Thus, all publication gets clustered regarding the inherent characteristics. In this context, a classification of current research fields can be achieved.

To enable a clustering of the research fields, experts (researchers, practitioners and software vendors) were asked to sort all relevant publications by a triple blind review process. Thus, the heterogeneity of the experts perspective ensures an acceptable categorization and an explorative consent within the community. Hereby, each expert assigned a specific publication position within of the framework by using a Likert scale (Likert, 1932). According to Dawes (2008), 10-point-scales are common and meaningful. In this context, two 10-point-scales were used where the responds has to choose between zero (paper refers totally to a technical/data utilization perspective) and nine (paper refers totally to a functional/data provisioning perspective). The results are aggregated and means are calculated. That allows a refined differentiation among the identified objects. The results are presented in Figure 4 (occurring numbers within the brackets refer to the references in Table 1) and are successively refined in the following sections. Hereby, key articles are highlighted and further research exposed.

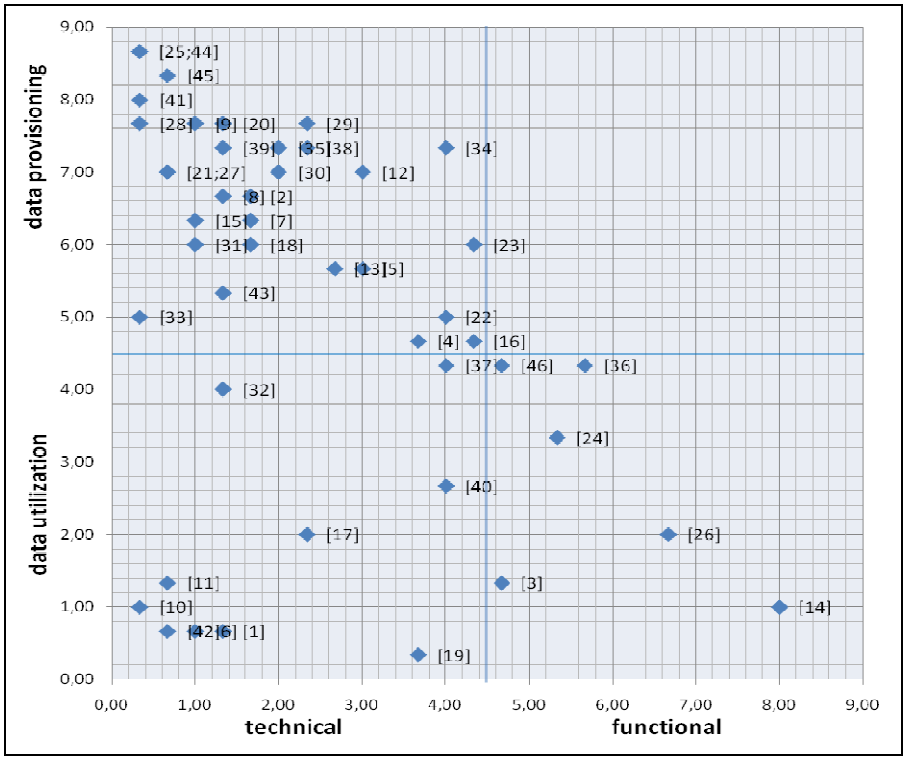

Figure 4. Clustering of Big Data Research Fields

Technical-Data-Provisioning

The biggest cluster is within the data provisioning and technical perspective. All these articles discuss how the increasing amount of data can be recorded, stored, and preceded to meet performance requirements. Surprisingly, more than 87 percent of all Information System papers are categorized into the technical-data-provisioning cluster. Thus, an unbalanced ratio occurs in the discipline. In contrast, the technical driven computer science research treats also functional aspects. The literature describes the Big Data itself as voluminous, large in scale and scope, distributed, and various (Junwei et al., Bizer et al. and Han et al., 2011). As a result, Big Data environments have to consider those inherent requirements (He et al. 2011; Simmhan, Reisser and Priebe, 2009)

To overcome those characteristics, one research field discuses several database management systems. Here, a vast majority consider that relational databases are not able to meet the requirements due the lack of scalability and query performance, (Han et al, 2011) whereas others allude proceedings in relational databases (Panchaksharaiah, 2009). Other authors discuss parallel working databases as a powerful performance alternative (Stonebraker et al., 2010). In addition, Han et al. (2011) compare several NoSQL approaches like key value, column oriented and document based databases, whereas Agrawal et al. (2011) highlight distributed databases, if a huge amount of data has to be analyzed. At least Venkataraman et al. (2011) discuss the storage hardware itself and compare several memory technologies like Non-Volatile Byte-Addressable and Dynamic-Random-Access-Memory (Venkataraman, 2011). Another research field considers the placement structure. Here, RCFile supports fast loading, fast query processing, efficient storage space utilization and adaptivity provides a performance efficient alternative to row-, column-, and hybrid-stores. In this context, RCFile enjoys a widely distribution. It is used by Facebook and adopted by several high level languages. (He et al., 2011) Furthermore, efforts can be observed where the amount of data is reduced by algorithms. Here, Wang and Hu (2007) suggest a time complex less attribute reduction algorithms, whereas Li et al. (2010) provide a reliable reduplication approach.

An identified key driver of Big Data is cloud computing. (Agrawal et al., 2010, 2011; Dai et al., 2011) One aspect is the distributed computing of large data sets. In this context, Googles MapReduce represents the center of research and is quoted in almost all publications. The framework specifies a map function and a reduce function. Hereby, the approach splits a large problem into small pieces and assigns it to appropriate pairs (key-value). These smaller tasks will be automatically preceded in an efficient parallel way and finally consolidated by reduced functions. (Dean et al., 2004) The improvement of

MapReduce reflects another research field. For instance, Junwei et al. (2011) introduce a message oriented approach to overcome the traffic overload in MapReduces. In addition, other efforts are fruitful to merge parallel databases and MapReduce. Due MapReduce performs well for complex analytical and extract-transform-load (ETL) tasks, whereas parallel databases are excellent in querying large data sets. (Stonebraker et al., 2010) Another discussion refers to Apaches Hadoop framework (Cuzzocrea et al., 2011). Hadoop runs on MapReduce tasks and holds several components like the key value oriented database HBase, Hadoops Distributed File System (HDFS), as well as the SQL oriented language Hive/HiveQL, the high level language Pig/PigLatin and their translators. (Bajda-Pawlikowski et al., Lee et al., 2011) In this context, Hadoop's Distributed File System (HDFS) represents the underlying file system and is object of intensive research. Thus, projects like HadoopDB or Hadapt received obviously performance advantages (Bajda-Pawlikowski et al., 2011). Lee at al. (2011) improve Hive and Pig by YSmart, an advanced SQL-to-MapReduce translator that provides an evident performance increase during complex queries (Lee et al., 2011). Dryad represents Microsoft's runtime pendant for the execution of parallel data applications. (Isard et al., 2007) Thereby, it formulates program codes as directed acyclic graph, where data is flowing between operations. In this context, DryadLINQ represents a declarative, data flow programming language for Dryad and can be translated in any .NET language. (Toole et al., 2011)

Another research fields refers to BI. Here, BI and the underlying DWH have to tackle the massive acquisition, storage, and analyzes of Big Data in distributed systems due to support enterprises with valuable information. Here, Hadoop and Hive are two known implementations of ETL and the analytical layer (Cuzzocrea et al., 2011). Cohen et al. (2009) quote initial requirements for BI in Big Data, because traditional solutions are not aiming this topic. In this context, Bajda-Pawlikowski et al. (2011) beat the performance of commercial DWH infrastructures by the column-oriented Hadoop extension, Hadapt.

It becomes apparent that Big Data is technical driven, yet. Thus, researches have to design and observe enterprise applications that deal in Big Data. Hence, technical issues and requirements become evident. Here, an important field will be the DWH and the implications and advantages that can be achieved due Big Data technology. In this context, issues occurring in context of multidimensional structures have to build up on top of distributed file structures or missing capabilities of query languages like HiveQL to support complex analytics. In addition, integration concepts are needed due heterogeneous data sources and potential combination efforts of NoSQL and relational databases (Cuzzocrea et al., 2011).

Technical-Data-Utilization

Big Data characteristics affect not only the data provision, but also their utilization. Thus, the computation and time complexity suffers by current algorithms and technology (Cohen et al., 2009). To overcome the obstacle of huge amount of data, parallel processing of analytical tasks is well discussed in literature. (Freedman and Kisilev, 2009; Wu et al., 2007) The MapReduce framework seems promising and obtained attention. Here, for instance machine learning algorithms or statistical methods like ordinary least squares or conjugate gradient are already successfully developed. (Cohen et al., 2009) Thus, Yan and Pi (2009) introduces a parallelizable clustering algorithms and Freedman and Kisilev (2009) a segmentation model, whereas Wu et al. (2007) proposes a decision tree. In addition, there are approaches, which use text mining techniques (Chen, 2011). In this context, the term *deep analyzes* as new direction of data analysis driven by complex statistical analysis and sophisticated machine learning techniques arises more and more in the literature. (Agrawal et al., 2010, 2011; Cohen et al., 2009) The term data mining occurs in publications more often in order to achieve valuable insights by identifying data patterns (Wu et al, 2007; Zhang et al., 2008; Bicking and Wimmer, 2010; Bizer et al. and Toole et al., 2011).

Especially, data mining or deep analytics are faced by a trade-off. Thus, too many data will influence the information value due unacceptable processing time. In this context, the approach of Beekmann (2003) should be observed. Thus, the work reduces the amount of considered analysis data by sample test. Here, analyzes of sample data will almost contain the same value as if the entirely data had been taken. Consequently, the storage space and computing effort gets reduced. But accordingly, further work is necessary. In addition, issues occurring due the law of addition. Therefore, not all algorithms can be processed in a parallel way (e. g. median) due to the reason that valuable information gets lost during processing. In an initial step, Qin et al. (2011) proposes an approach where missing data is successfully estimated. Thus, similar concepts are beneficial. Another research field is expected in the adequate visualization of Big Data. In this context, specifications and solutions are not yet addressed.

Functional-Data-Provisioning

No paper covers a functional and data provisioning perspective. It remains questionable why existing approaches are not linked to the Big Data discussion to overcome such an information flood. An improved information management is requested that obtain an advanced information demand analysis and provide better information supply. Consequently, the data itself and the required processing time can be reduced. Potential concepts like information lifecycle management (Robek et al., 1996)

or lean information management (Hicks, 2007) have to come into the discussion to overcome the information wave. Here, the value of the information itself is to scrutinize. Further investigations of suitable approaches and Big Data have to be done. If existing concepts are not sufficient, researchers have to detect issues, hindering factors and further requirements to design an appropriate concept. In consequence, the question has to be answered weather and how Big Data can be processed and reduced in a task oriented way, by avoiding a loss of knowledge.

Functional-Data-Utilization

The usage of Big Data enjoys a wide distribution in various disciplines. Here, Big Data occurs in social sciences (Amer-Yahia et al., 2010), e-government (Bicking and Wimmer, 2010), finance (Zhang et al., 2008), medicine (Bower et al., 2009), behavioral science (Eagle, 2010), ubiquitous computing (Bizer et al., 2011), construction (Khan and Hornbæk,, 2011), crime analysis (Toole et al., 2011), bioinformatics (Yan and Pi, 2009), climate science (Alexander et al. 2011), astrology (Freedman and Kisilev, 2009) and BI (Cuzzocrea et al., 2011). For the first time, Zhang et al. (2011) enables a performance acceptable option price prediction by using the Monte Carlo method, whereas Toole et al. (2011) discovers spatio-temporal crime patters by segmentations. Thus, it appears that Big Data affects various disciplines. Thereby, it is the task of further research to identify potential use cases. Moreover, Big Data and BI applications require more research due the quoted technologies are promising to overcome volume, performance and heterogeneity issues (Cuzzocrea et al., 2011).

CONCLUSION

The topic has an increased occurrence within the scientific discussion. The work supports the discussion about Big Date by providing a structured overview. Here, the paper identifies current research fields and scrutinizes, whether data provisioning driven concepts are affecting Big Data. It gets obvious that the topic is technical driven, yet, and conducted by performance aspects. Especially, the technical data provisioning part gained a high consideration. In addition, methods like data mining or deep analytics are discussed in utilization and furthermore, use cases are presented. Surprisingly, not a single publication considers functional data provisioning concepts.

In this context, there are missing investigations about this topic in an overall perspective. This paper uncovers a relevant research gap that tackles the challenge not merely by costly technologies. An adequate functional data provisioning out of the Big Data has to support the prevention of flooding users with information. Hence, potential concepts like information lifecycle management or lean information management can support a task and user oriented reduction of the huge amount of data and disclose the inherent value. These concepts implicate methods that enable a continuously determination of an information demand and supply. For instance, lean information management follows a pull principle and is driven by a value paradigm. Thus, the customer causes the data recording. This leads to the point that no data is stored senseless and the user itself regularizes the supply and demand. Nevertheless, such an approach requires further research to be able to offer appropriate solutions.

The work does not deliverer a definition or related theories. Thus, former articles that affecting Big Data relevant topics but missing the term itself were not taken into account, yet. Further investigations have to treat these gaps. Additionally work will investigate, if the nature of Big Data neglects the intention of functional data provisioning or if no one has addressed it, yet. In the latter case, existing approaches have to be identified and proofed against suitability and value contribution in order to solve the challenge to get task and user oriented data out of the Big Data. This leads to the demand that relevant requirements of a new approach have to be identified. Thus, users suffering on an information flood will obtain a scoped overview and will be able to use information as a competitive advantage referring to their assigned tasks.

REFERENCES

1. AIS (2007) MIS Journal Ranking List, Retrieved 2012-02-21, from: http://ais.affiniscape.com/displaycommon.cfm?an=1&subarticlenbr=432.
2. Agrawal, D., Das, S. and Abbadi, A. (2011) Big Data and Cloud Computing: Current State and Future Opportunities, *Proceedings of the EDBT*, ACM, March 22-24, Sweden, 530-533
3. Agrawal, D., Das, S. and Abbadi, A. (2010) Big Data and Cloud Computing, *Proceedings of the VLDB*, ACM, September 13-17, Singapore, 1647-1648.
4. Alexander, F., Hoisie, A. and Szalay A. (2011) Big Data Guest Editors Introduction, *Computing in Science & Engineering*, IEEE, 2011, 10-13.
5. Alvaro, P., Condie, T. and Conway, N. (2010) BOOM Analytics, *Proceedings of the EuroSys*, ACM, April 13-16, France, 223-236.

6. Amer-Yahia, S., Doan, A. and Kleinberg, J. (2010) Crowds, Clouds, and Algorithms, *Proceedings of the SIGMOD 2010*, ACM, June 6-10, USA, 1259.
7. Bajda-Pawlikowski, K., Abadi, D. and Silberschatz, A. (2011) Efficient Processing of Data Warehousing Queries in a Split Execution Environment, *Proceedings of the SIGMOD*, ACM, June 12-16, Greece, 1165-1176.
8. Beekmann, F. (2003) Stichprobenbasierte Assoziationsanalyse im Rahmen des Knowledge Discovery in Databases, DUV, Wiesbaden.
9. Bicking, M. and Wimmer, M. (2010) Need for Computer-Assisted Qualitative Data Analysis in the Strategic Planning of E-Government Research, *Proceedings of the Annual International Conference on Digital Government Research*, ACM, May 17-20, Mexico, 153-162.
10. Bizer, C., Boncz, P. and Brodie, M. (2011) The Meaningful Use of Big Data: Four Perspectives, in *SIGMOD*, 40, 4, ACM, December 2011, USA, 56-60.
11. Bower, M., Stead, M., and Brinkmann, B. (2009), Metadata and Annotations for Multi-scale Electrophysiological Data, *Proceedings of EMBS*, IEEE, September 3-6, USA, 2811-2814.
12. Casonato, R., Lapkin, A., Beyer, M. and Genovese Y. (2011) Information Management in the 21stCentury, Gartner Research, White paper.
13. Chen, H. (2011) Smart Market and Money, *IEEE Intelligent Systems*, IEEE, November-December 2011, USA, 82-96.
14. Cohen, J., Dolan, B. and Dunlap, M. (2009) MAD Skills, *Proceedings of the VLDB*, ACM, 2, August 24-28, France, 1481-1492.
15. Cooper, H. M. (1998) Synthesizing Research: A Guide for Literature Reviews, Sage Pubn, Thousand Oaks.
16. Cramer, H., Rost, M. and Shamma, D. (2011) Using App Stores, Wide Distribution Channels and Big Data in UbiComp Research, *Proceedings of the UbiComp*, ACM, September 17-21, China, 619-620.
17. Cuzzocrea, A., Song, Y. and Davis, K. (2011) Analytics over Large-Scale Multidimensional Data, *Proceedings of the DOLAP*, October 28, UK, ACM, 101-103.
18. Dai, J., Huang, J. and Huang, S. (2011) HiTune: Dataflow-Based Performance Analysis for Big Data Cloud, *Proceedings of the USENIXATC*, USENIX Association Berkeley, USA.
19. Das, K., Fratkin, E. and Gorajek, A. (2011) Massively Parallel In-Database Predictions using PMML, *Proceedings of the workshop on PMML*, ACM, August 21-24, USA, 22-27.
20. Dawes, J. (2008) Do data characteristics change according to the number of scale points used?, *International Journal of Market Research*, 50, 1, 61-77.
21. Dean, J., Ghemawat, S. and Inc, G. (2004) MapReduce: Simplified Data Processing on Large Clusters, *Proceedings of the OSDI*, 6, 137-149.
22. Eagle, N. (2010) Big Data, Global Development, and Complex Social Systems, *Proceedings of the SIGSOFT*, ACM, November 7-11, USA, 3.
23. Gantz, J., Chute, C. and Manfrediz, A. (2008) The Diverse and Exploding Digital Universe. IDC.
24. Gluchowski, P. (2001) Business Intelligence, *HMD – Praxis der Wirtschaftsinformatik*, 38, 222, 5-15.
25. Han, J., Haihong, E. and Le, G. (2011) Survey on NoSQL Database, *Proceedings of the ICPCA*, IEEE, October 26-28, South Africa, 363-366.
26. He, Y., Lee, R. and Huai, Y. (2011) RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems, *Proceedings of the ICDE*, IEEE, April 11-16, Germany, 1199-1208.
27. Hicks, B. (2007) Lean information management, *International Journal of Information Management*, 27, 4, 233-249.
28. Huai, Y., Lee, R. and Zhang, S. (2011) DOT: A Matrix Model for Analyzing, Optimizing and Deploying Software for Big Data Analytics in Distributed Systems, *Proceedings of SOCC*, ACM, October 27-28, Portugal.
29. Foumier, H., Kop, R. and Sitlia, H. (2011) The Value of Learning Analytics to Networked Learning on a Personal Learning Environment, *Proceedings of the International Conference on Learning Analytics and Knowledge*, ACM, February 27-March 1, Canada, 104-109
30. Frank, U., Heinzl, A. and Schoder, D. (2008) WI-Orientierungslisten, *Wirtschaftsinformatik*, 50, 2, 155-163.

31. Freedman, D. and Kisilev, P. (2009) Fast Mean Shift by Compact Density Representation, *Proceedings of the CVPR Recognition*, IEEE, June 20-25, USA, 1818-1825.
32. Isard, M., M. Budiu and Yu. Y. (2007) Dryad: distributed data-parallel programs from sequential building blocks, *Proceedings of SIGOPS*, Portugal, ACM, 59-72.
33. Jacobs, A. (2009) The Pathologies of Big Data, *Communications of the ACM, 52, 8*, ACM, August 2009, USA, 1-12.
34. Junwei, G., Jiang, X. and Fang, Y. (2011) Improvement of the MapReduce Model Based on Message Middleware Oriented Data Intensive Computing, *Proceedings of the CIS*, IEEE, December 3-4, China, 86-89.
35. Khan, A. and Hornbæk, K. (2011) Big Data from the Built Environment, *Proceedings of the LARGE*, ACM, September 18, China, 29-32.
36. Kaushik, R. T., Abdelzaher, T. and Egashira, R. (2011) Predictive Data and Energy Management in GreenHDFS, *Proceedings of the IGCC*, IEEE, July 25-28, USA, 1-9.
37. Kohlwey, E., Sussman, A. and Trost, J. (2011) Leveraging the Cloud for Big Data Biometrics, *Proceedings of the SERVICES*, IEEE, July 4-9, USA, 597-601.
38. Koreimann, D. (1971) Methoden und Organisation von Management-Information-Systemen, Gruyter Verlag, Berlin.
39. Lee, R., Luo, T. and Huai, Y. (2011) YSmart, *Proceedings of the ICDCS*, IEEE, June 20-24, USA, 25-36.
40. Li, H., Ruan, Y. and Qiu, J. (2011) Design Patterns for Scientific Applications in DryadLINQ, *Proceedings of the DataCloud-SC*, ACM, November 14, USA, 61-69.
41. Li, X., Lillibridge, M. and Uysal, M. (2010) Reliability Analysis of Deduplicated and Erasure-Coded Storage, *Proceedings of the SIGMETRICS*, 38, 3, ACM, December 2010, USA, 4-9.
42. Likert, R. (1932) A Technique for the Measurement of Attitudes, *Archives of Psychology*, 22, 140, 1-55.
43. Lukashevich, H., Nowak, S. and Dunker, P. (2009) Using One-Class SVM Outliers Detection for Verification of Collaboratively Tagged Image Training Sets, *Proceedings of the International ICME*, IEEE, June 28-July, USA, 682-685.
44. Meijer, E. (2011) Big data is about more than size, and LINQ is more than up to the task, *Communications of the ACM* , 54, 10, ACM, October 2011, USA.
45. Panchaksharaiah, U. (2009) How to Adress Big Data, *Communications of the ACM*, 52, 12 ACM, December, 7.
46. Qin, X., Wang, H. and Du, X. (2011) Parallel Aggregation Queries over Star Schema, *Proceedings of the ISPA*, IEEE, May 26-28, Korea, 329-334.
47. Reddi, V. J., Lee, B. C. and Chilimbi, T. (2011) Mobile Processors for Energy-Efficient Web Search, *ACM Transactions on Computer Systems, 29, 3*, ACM, August 2011, USA.
48. Reisser, A. and Priebe, T. (2009) Utilizing Semantic Web Technologies for Efficient Data Lineage and Impact Analyses in Data Warehouse Environments, *Proceedings of the DEXA, IEEE*, August 31- September 4, Austria, 59-63.
49. Robek, M., Brown, M. and Stephens, D. (1996) Information and records management. McGraw Hill, New York.
50. Sankar, Y. (1998) Value Based Management for Information, Canadian Scholars Press, Toronto.
51. Satzger, B., Hummer, W. and Leitner, P. (2011) Esc, *Proceedings of the CLOUD*, IEEE, July 4-9, USA, 348-355.
52. Silberstein, A., Machanavajjhala, A. and Ramakrishnan, R. (2011) Feed Following: The Big Data Challenge in Social Applications, *Proceedings of DBSocial*, ACM, June 12, Greece, 1-6.
53. Simmhan, Y., Barga, R. and Heasley, J. (2009) GrayWulf, *Proceedings of the HICSS, IEEE*, January 5-8, Hawaii, 1-10.
54. Stonebraker, M., Abadi, D. and Dewitt, D. (2010) MapReduce and Parallel, *Communications of the ACM*, 53, 1, ACM, January 2010, USA, 64-71.
55. Toole, J., Eagle, N. and Plotkin, J. (2011) Spatiotemporal Correlations in Criminal Offense Records, *ACM Transactions on Intelligent Systems and Technology*, 2, 4, ACM, July 2011, USA.
56. Venkataraman, S., Tolia, N. and Ranganathan, P. (2011) Consistent and Durable Data Structures for Non-Volatile Byte-Addressable Memory, *Proceedings of the FAST*, USENIX Association Berkeley, 2011, USA, 1-15.
57. VHB (2008) VHB Teilranking Wirtschaftsinformatik und Informationsmanagement, Retrieved 2012-02-21, from: http://vhbonline.org/SERVICE/JOURQUAL/JQ2/TEILRANKING-WIRTSCHAFTSINFORMATIK-UND-INFORMATIONSMANAGEMENT/.

58. Wang, G. and Hu, F. (2007) Quick Knowledge Reduction Based on Divide and Conquer Method in Huge Data Sets, *Proceedings of the PReMI*, Springer, December 18-22, India, 312-315.
59. Wu, Y., Li, L. and Zhao, Y. (2007) Application of Improved SPRINT Algorithm in the Graduation Design Process Management System, *Proceedings of the Workshop on Intelligent Information Technology Application, IEEE*, December 2-3, China, 252-255
60. Yan, Z. and Pi, D. (2009) A Fuzzy Clustering Algorithm Based on K-means, *Proceedings of the ECBI, IEEE*, June 6-7, China, 523-528.
61. Zhang, Y., Gong, B. and HuiLiu, Y. (2011) Parallel Option Pricing with BSDEs Method on MapReduce, *Proceedings of ICCRD*, IEEE, March 11-13, China, 289-293.
62. Zhou, M., Hu, H. and Zhou, M. (2010) Searching XML Data by SLCA on a MapReduce Cluster, in *Proceedings of the IUCS*, IEEE, October 18-19, China, 84-89.