Challenging Big Data Engineering: Positioning of Current and Future Development

Matthias Volk, Daniel Staegemann, Matthias Pohl and Klaus Turowski
Magdeburg Research and Competence Cluster Very Large Business Applications, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany

Keywords: Big Data, Engineering, Technology Selection, Decision Support System.

Abstract: This contribution examines the terms of big data and big data engineering, considering the specific characteristics and challenges. Deduced by those, it concludes the need for new ways to support the creation of corresponding systems to help big data in reaching its full potential. In the following, the state of the art is analysed and subdomains in the engineering of big data solutions are presented. In the end, a possible concept for filling the identified gap is proposed and future perspectives are highlighted.

1 INTRODUCTION

Big data and the accompanying technologies rank among the most popular and researched topics of the last years and have achieved significance influence on many aspects of daily life. Around the world scientists as well as practitioners are seeking to explore, utilize and extend its potentials (Mauro *et al.*, 2016; Jin *et al.*, 2015). For example the amount of data created by modern industry already exceeds a total volume of 1000 exabytes annually (Yin and Kaynak, 2015). Still, despite the increase of maturity accomplished by those efforts, there exists a plethora of challenges that needs to be solved in the future. One of those stems from the interest big data has attracted and the subsequently high number of companies being engaged and offering solutions, tools and services around the topic (Turck, 2018). Thus, to achieve the most optimal results, it is necessary to choose the best fitting tools and approaches for conducting those kinds of projects.

However, in doing so, lots of pitfalls may occur that may lead to a lower projects success or even a complete failure. Those can be related to the multifaceted nature of the data, their processing as well as the general management (Sivarajah *et al.*, 2017). For instance, even the pure testing of a constructed solution can be a sophisticated task (Staegemann *et al.*, 2019). Additionally, this situation is reinforced, by the huge shortage of people that are able to leverage big data (Debortoli *et al.*, 2014; Gardiner *et al.*, 2018; Sivarajah *et al.*, 2017).

1.1 Delimitation of Big Data

Apart from the described challenges, a second view emphasizes the opportunities that one could find in the analysis of big data, such as fraud detection in IT log analysis or information retrieval in social media (Zikopoulos, 2012, p. 12 ff.). However, dealing with the challenges and problems surrounding big data constitutes a condition for profitable insights. While, there is no sole explanation of the term itself, the definition of the National Institute of Standards and Technology (NIST) is widely accepted and will therefore also be used in the publication at hand (NIST, 2015). It introduces the four data characteristics named, volume, variety, velocity and variability.

Volume indicates the sheer amount of data that has to be handled in order to fulfill a given task. In case of big data, that volume excels the capabilities of commonly used systems and technologies. Therefore, it is required to apply approaches that are adjusted to meet those needs. For example transferring all data to a single high-end server for processing is impractical because of the bad ratio of time consumed by transfer to productive work. Although, there is no clearly defined number where *big* starts, today some companies are already working with data in the petabyte-region (Assunção *et al.*, 2015; Gandomi and Haider, 2015).

Variety refers to the heterogeneity of data and its origins. While a plethora of sources compared to a single one often allows for more comprehensive

Volk, M., Staegemann, D., Pohl, M. and Turowski, K.
Challenging Big Data Engineering: Positioning of Current and Future Development.
DOI: 10.5220/0007748803510358
In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security (IoTBDS 2019)*, pages 351-358
ISBN: 978-989-758-369-8

insights into a topic, the retrieval and handling of data also gets more demanding for the underlying technology the higher the number gets. Apart from that, variety also refers to the inner structure of the data (structured/semi-structured/unstructured) and properties like formatting, datatypes or used units. Those often have to be unified, when combining different sets of data. Nevertheless this process might be an additional source of flaws (Gani *et al.*, 2016).

Velocity denominates two things. On the one hand, it can refer to the speed at which those data are incoming, regardless of requirements regarding their processing. On the other hand, it can identify the speed at which the received data have to be processed. With the spread of smartphones, smart home technologies and the ubiquity of sensors, plenty of data are generated. Assessing those in a timely manner is not possible with the usage of only traditional database technology. However huge potential gains may arise, for instance when used to create personalized recommendations or optimize decisions regarding the staffing of a store (Gandomi and Haider, 2015; Sagiroglu and Sinanc, 2013).

Variability corresponds to the change of the other characteristics. Since the course of the real world is not always constant, the same applies for the generation of data. Therefore, systems handling those data have to accommodate this fact. This can for example be caused by special events that trigger resonance on social media. As a result, the produced volume might shortly increase, accompanied by a shift of the composition of received data towards the format of the affected platform (Katal *et al.*, 2013).

Table 1: Characteristics of big data.

Characteristic	Description
Volume	Volume indicates the amount of data that has to be handled.
Variety	Variety refers to the heterogeneity of data and its sources.
Velocity	Velocity denominates the speed in which data are incoming and the speed at which received data have to be processed.
Variability	Variability corresponds to the change of the other characteristics.

Those characteristics, depicted in Table 1, lead to certain requirements concerning the building of corresponding systems. For example the high volume makes it important to ensure a high degree of scalability to keep up with likely growing workloads. Since the potential of vertical growth is extremely limited, instead horizontal expansion is pursued. This approach offers far more possibilities and is usually handled by distributing the necessary calculations over a multitude of servers running on commodity hardware (McAfee and Brynjolfsson, 2012). It allows to handle huge amounts of data, while reducing costs for customization or high-tech premiums, therefore allowing for a better cost-performance ratio. Furthermore this approach constitutes a solution for the demands that go along with the aforementioned velocity. Meanwhile the variety results in a need to harmonize and convert related data for the sake of analyzing it. It is also common for data to be incomplete or incorrect, therefore lacking in data quality (Taleb *et al.*, 2018). This in turn necessitates measures for detection and correction of flaws concerning the data quality.

The activities of creating those systems fall into the category of engineering, which is concerned with the purposeful and planful design and construction of means to transform the reality towards a desired state (Rogers, 1983). Though, the generalistic nature of this explanation dimishes the applicability on today's information systems. Therefore, a more precise expression for the given challenges would be systems engineering. This term describes the combination of several interacting elements to solve a problem (Wasson, 2016).

However, this definition still lacks on the emphasis of software. The process of engineering software is defined as the "application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software" (IEEE, 2017, p. 8). Considering the interplay of different software and hardware solutions to fullfill the identified needs in big data system engineering, the aspects of both disciplines have to be factored in. Additionally, as a consequence of all highlighted specifics, depicted in Table 1, one can note that the engineering process in this particular area brings up even more deliberations.

In (ISO, 2014, p. 6) big data engineering is explained as "the storage and data manipulation technologies that leverage a collection of horizontally coupled resources to achieve a nearly linear scalability in performance." Yet, this statement ignores the specifics of the big data characteristics, possibly being too vague for practical application. Along with this, it doesn't accommodate the increased difficulty of testing due to the high complexity of the resulting systems (Tao and Gao, 2016).

Therefore we define big data engineering as "a systematic approach of designing, implementing, testing, running and maintaining scalable systems, combining software and hardware, that are able to

gather, store, process and analyze huge volumes of varying data, even at high velocities".

1.2 Challenges of Big Data Engineering

As pointed out, the challenges of big data engineering are strongly related to the specific requirements accompanying the task. Handling huge volumes of data in different formats at high speeds, while maintaining resiliency and data security, can be very challenging. Thus it is uncommon for typical enterprises, that are interested in the usage of big data, to develop their own system from scratch. Instead, existing solutions and tools are combined to fullfill the specific needs. Though, due to the sheer amount of those offerings, the proper selection of the solutions and the employees designated to administer and operate them has developed into a cumbersome and challenging task (Szyperski *et al.*, 2016). Since experts in the area of big data are rare, many enterprises lack the appropriate expertise for this purpose, potentially stalling or even impeding big data projects (Sagiroglu and Sinanc, 2013).

Therefore we argue, that big data can only unfold its full potential, if its application is made more accessible by supporting the engineering of corresponding systems, allowing more enterprises to participate.

2 STATE OF THE ART

Due to its popularity and the subsequent efforts, big data reached a high maturity in recent years, especially in terms of its definition, characteristics and applications. However, the general engineering activities are less extensively researched, compared to its foundation. In most of the cases, this is restricted to adjacent areas. As a consequence, only few contributions have been identified targeting multiple subdomains of big data engineering.

In (Fekete and Vossen, 2015) the authors highlight the complexity of the term big data in context of already existing solutions, such as data warehouse architectures. Through the continuous introduction of novel ideas and technologies, the difficulty of a big data-related project realization is steadily increasing. This led, for instance, to a widespread uncertainty about technology choices, combinations and especially their implementation, even though a goal and requirements exist. According to the authors, a layer-based reference architecture appears to be a suitable solution in context of this. As an artifact, the Goal-Oriented Business Intelligence Architectures (GOBIA) method is introduced, consisting out of a reference architecture and a five-stepped development process. Originated on a similar baseline, the contribution by (Volk *et al.*, 2018) provides a classification approach for big data technologies. Driven by the current situation, that lots of confusion in respect to big data technologies, their application and description exists, the *BDTonto* ontology was developed. The structure is predominantly oriented on the crucial steps of a big data project and the operations performed in each phase. Through the extraction and alignment of the needed technology information, a classification is utilized. Compared to the initially proposed GOBIA approach, requirements and reference architectures are neglected in here. However, the authors deal in other contributions with this subdomain, such as in (Volk *et al.*, 2017) with focus on requirements engineering and the project realization.

A description of similar activities, with regard to various technologies is also provided in (Oussous *et al.*, 2018). By highlighting challenges, located on the individual levels of big data, numerous descriptions and a comparison of a multitude of technologies are given. Another comprehensive approach is presented by (Lehmann *et al.*, 2016), in addition to a layered reference framework, also a method for the selection of big data technologies is provided. While the first attempts to distinguish and classify technologies, without explicitly naming specific manifestations, the second supports the selection procedure. Within the description of each layer, distinct properties are highlighted and exemplary technologies presented. The method itself, called Strategy Time Analytics Data Technology (S.T.A.D.T) Selection Framework (SSF), thereby represents a multi-stepped procedure. Starting with the strategy, a tactical plan, decomposing a use case into storage, processing and analytics, is formulated. Afterwards, each element is aligned to the layers and further process steps of SSF, which in combination describe the main building blocks. In the further steps, the remaining requirements of the use-case are investigated and integrated. By the identification of the processing types, used methods and the handled data, a technology selection is refined.

Despite the fact, that no clear definition or relation to big data engineering is given, all contributions are concerned with the realization of big data projects and the respective systems. The same applies for contributions, which are not further described within this section. Consequently the subdomains requirements engineering, project realization, technology selection and reference architectures were

deduced as major areas of interest. In the following, those are further examined, to highlight current challenges and future developments.

3 RELATED SUBDOMAINS

Within the following sub-section, each of the identified subdomains will be described in context of the area big data. Additionally to that, potentials as well as challenges will be identified and summarized in Table 2.

3.1 Requirements Engineering (RE)

When it comes to the realization of projects, one initial activity is always represented by the requirements engineering procedure. In context of system development, requirements describe the desired behaviour of the system, relevant properties as well as attributes (Sommerville and Sawyer, 1997). However big data projects reveal distinct properties compared to regular IT projects as it was previously highlighted, for instance, by the data characteristics. According to Volk et al., big data projects "can be described as an objective-oriented temporary endeavor with a precisely defined timeframe, whose implementation requires a combined use of specific big data technologies" (Volk *et al.*, 2017, p. 3). As a consequence, the requirements engineering procedure in big data projects differs from conventional IT projects. By combining compound big data requirements with a classification framework and typical project realization approaches, in (Volk *et al.*, 2017) a method for the sense-making of those projects in early stages is presented. Hence, practitioners are forced to start with an in-depth planning.

The previously referred compound requirements are also part of the investigations conducted by (Arruda and Madhavji, 2018). In their literature review the current situation of RE in big data is analyzed. In doing so, the authors highlight the importance of a thorough and comprehensive observation of big data relevant facets. Additionally to the specific big data technologies, also their selection, combination, integration and rapid changes are mentioned. Due to the reason that some of the most prevailing challenges are related to the sole formulation, testability and recognition of big data characteristics, plenty of future research is required. Despite the fact that the authors were able to identify initial approaches, they conclude that only a little amount research was carried out so far (Arruda and Madhavji, 2018).

Further contributions, such as (Altarturi *et al.*, 2017), propose a big data requirements model, which can be used for the instantiation of the project. It basically consists out of multiple steps performed by the requirements engineer as well as the data scientist to obtain a comprehensive view from various perspectives. Yet, this selection of contributions represents only an excerpt of current investigations.

3.2 Technology Selection (TS)

An important aspect of big data engineering arises with the application of those kinds of projects. Currently a multitude of technologies, tools and services exist (Turck, 2018), that lead with their sheer amount to lots of confusion. Although an ambitioned community exists beyond almost every single solution, the selection of application-specific tools and technologies requires plenty of effort (Philip Chen and Zhang, 2014). Reinforced by the continuous appearance of innovations and alternatives, several potentials as well as challenges can be identified. At its first glance, an increase of new solutions appears to be daunting. Nevertheless, by the composition of very specific solutions, incorporating new introductions or consolidations, a tailored system architecture can be facilitated. Hence, practitioners pursuing always an up-to-date solution are able to adapt rapidly to recent competitive needs.

However, it should not be neglected, that each of those solutions has its own requirements and properties. The composition of an architecture requires lots of expertise and knowledge (Zicari *et al.*, 2016). This is not restricted to the initial requirements that are formulating functional aspects, data characteristics and other constraints. Further aspects, such as the compatibility to other manifestations or the used licences have to be considered as well. Thus it is not surprising that the demand of experts in this area is constantly rising (Debortoli *et al.*, 2014; Gardiner *et al.*, 2018). At the current time, only a fair amount of approaches have been found, that attempt to provide clarity in this particular field. Either by classifying existing tools and technologies (Oussous *et al.*, 2018; Volk *et al.*, 2018) or selecting them (Lehmann *et al.*, 2016). In most of the cases those are also linked to architectural recommendations, such as in the case of (Pääkkönen and Pakkala, 2015) that proposed a classification of various tools and technologies while introducing a compound reference architecture.

3.3 Project Realization (PR)

Regarding the expertise that is needed for data science in the context of big data engineering, a big data project has to be considered from a business, a statistics, a machine learning, a domain and an engineering perspective (NIST, 2015). On the one hand, the domain view depends on the use-case as well as data semantics and cannot be generalized. On the other hand, an analysis of costs, benefits and risks as a description of a business case covers the business-related aspects and leads to a decision if a project should be conducted or not (van Putten *et al.*, 2013; Taschner, 2017). Cost-benefit analysis and business case development can be highly complex and depend distinctively on the stakeholders of a project.

From a statistics/machine learning perspective, a concept like knowledge discovery in databases (KDD) (Fayyad *et al.*, 1996) or the further matured framework CRISP-DM (Shearer, 2000) are widely used within the scope of data-related projects. Based on a business and data understanding, the selected data can be prepared and modelled. Evaluated results are further deployable to the target (e.g. database, application) regarding the purpose of the project. These steps stake out the broad lines, even though, the detailed elaboration has to be conducted by data scientists for instance. The most important part is attributed to the engineering point of view.

Referencing to the already described definition provided by the NIST, the data life cycle, including data collection, preparation, analysis, visualization and accessing (NIST, 2015) guides the technical project development. Further, an information-related approach, involving data creation, information consolidation, information utilization, information preservation as well as information archiving (Thome and Sollbach, 2007, p. 22) could also be considered.

However, engineering is constantly challenging in regards to big data and an overall concept that meshes on the data level is desirable. The complexity from technical and project organization perspective is immense in respect to business objectives and needed key capabilities (Zicari *et al.*, 2016). An interdisciplinary project approach has to combine technical, domain-specific as well as business aspects and reveals project related skillsets.

3.4 Reference Architectures (RA)

In order to overcome the complexity of these related subdomains, especially the TS and PR, lots of effort is put into the development of architectural solutions. Those system architectures focus on fundamental concepts or properties of a system, to encompass all elements and their relationships (IEEE, 2011). Due to the nature of big data and its characteristics, described in section 1.1, the development demands the consideration of various requirements. As a consequence, reference architectures gained huge popularity in the area of big data, since the initial introduction of the Lambda architecture by Nathan Marz (2011).

According to Vogel et al. (2011), reference architectures combine the "general architecture knowledge and general experience with specific requirements for a coherent architectural solution for a specific problem domain. They document the structures of the system, the main system building blocks, their responsibilities, and their interactions" (Vogel *et al.*, 2011, p. 232). Further approaches are for instance the Kappa (Kreps, 2014), Bolster (Nadal *et al.*, 2017) and Solid architecture (Martínez-Prieto *et al.*, 2015). All of them constitute extensions or alterations of the Lambda architecture. The Kappa architecture, for instance, devotes the reduction of the maintenance by scaling the initially two proposed layers into only a single one (Kreps, 2014).

Apart from the application of very specific approaches, also some general reference architectures exist, serving more as a kind of a *best-practice*. While the first mostly offers specific implementation and technology selection details, the latter only provide a rough structure. Those are predominantly based on the previously described project realization workflows, consisting out of various lifecycle steps. However, the use of a specific approach does not necessarily imply that concrete technical details are described. Depending on the scope, this can also be limited on functional aspects only (Vogel *et al.*, 2011). One highly regarded approach was proposed by Pääkkönen and Pakkala (Pääkkönen and Pakkala, 2015). In here, the main components were identified and mapped to the crucial steps of a project realization, through the comparison of real-world use cases.

However, while considering one of those solutions, the problem remains at which point in time a specific reference architecture should be applied. Depending on a multitude of attributes, the selection of the most suitable approach is a comprehensive task. The final decision eventually has high influence on the project and, therefore, will decide on its success or failure. Thus, this particular subdomain offers, as all other subdomains, lots of potentials, such as best-practices or decision support for technology selection. Nevertheless, also huge challenges can be

identified, like the careful choice of a reference architecture. A summary, highlighting all identified potentials and challenges of all subdomains, is given in Table 2.

Table 2: Derived potentials and challenges.

Domain	Potentials	Challenges
RE	detailed planning; consideration of various big data relevant aspects	requires an depth investigation; formulation of the requirements; frequent changes;
TS	distinct features; tailored solutions; interoperability; innovations and alternatives;	multifaceted data; compatibilities; system requirements and license; existing systems; outdated solutions;
PR	bus. objectives; domain-specific understanding; interdisciplinary project approach	skill matching; collaboration strategies; complexity of projects
RA	best-practices; imp. details; decision-support;	identifying required input information; choosing the RA; implementation

4 POTENTIAL SOLUTIONS

According to the aforementioned subdomains, one can observe, that the area of big data engineering, and thus the realization of these specific projects, is getting more and more complex. Thus, we argue, that in future development the focus will shift on the facilitation of the contiguous big data engineering tasks. Besides the sole application of best-practices and reference architectures, described in beforehand, a realization of a comprehensive supporting solution appears to be promising. By covering all identified subdomains, this could be realized, for instance, by a comprehensive knowledge-driven decision support system (DSS). Considering the previously made observation, this could be structured and used as exemplarily illustrated in Figure 1. As an input, the decision maker needs to deliver the needed information. This can be managed as described by Volk et al. (Volk *et al.*, 2017) through an initial RE procedure, during which a combination of requirements and characteristics could be developed. In doing so, also important operations should be introduced to the system, which are often related to the various phases of the data life-cycle. This includes for instance data generation and transformation methods. The benefits of this initial procedure are two-folded. First, the project itself is getting planned in a thorough manner. Second, the input information, required for the DSS, deliver as many details as possible. The input data itself could be realized by directly inserting the developed requirements. Afterwards, these could be decomposed and analysed by applying additional natural language processing techniques. Alternatively, preconfigured input options, mapping the basic structure of the requirements, could be provided. In any case, a cooperative strategy appears to be desirable, at which the respective user interacts with the systems. Whenever adjustments are needed, modifications of the initial inputs should be allowed. This can be realised by the provided user interface that represents a crucial part of a DSS (Nižetić *et al.*, 2007).

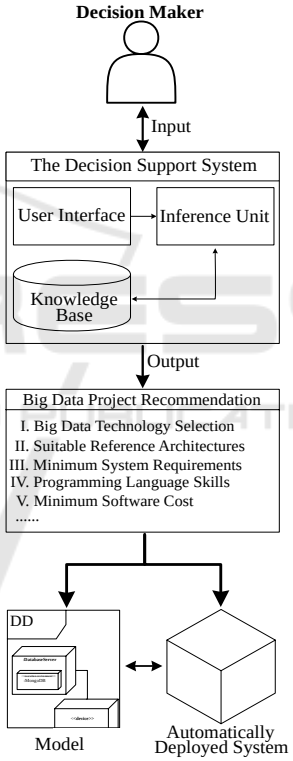

Figure 1: The proposed solution.

Afterwards the DSS analyses the data through the application of various models. In this particular case, a rule-based inference unit could be applied, due to the diversity of the data (Nižetić *et al.*, 2007). Through the use of an easy to adapt and extend solution, such as the *BDTonto* ontology or the *SSF* (cp. Section 2), a comprehensive knowledge base can be provided. Compared to other approaches, this would assure long-term usability, even in such a fast evolving field. The generated output of the DSS can

be manifold. Besides the sole recommendation about technological considerations, also suitable reference architectures and specifications required for implementation purposes could be determined. This might also include information, needed for the relevant stakeholders of the projects and, thus, persons directly involved. For example, a list of skills, that are mandatory for the implementation and utilization of the planned system, could be forwarded to the human resources department. This would facilitate the recruitment of appropriate experts, as it was tried by other authors through the investigation of current job descriptions (Gardiner *et al.*, 2018). Furthermore, the generated output could be redirected to automatically create a model for a better overview, a deployment or both. Hence, the system might be also capable to pass through the results to another system realizing an automated provisioning of the given recommendation. In consequence of an enhanced use of such a solution, multiple benefits are expectable, for both researchers as well as practitioners. First and foremost, the general relation between the various subdomains of big data can be uncovered. By providing such kind of a clarification, in the *jungle* of big data, the needs of the required knowledge of a related project can be specified in more detail. As a consequence, this might have a beneficial influence on the general demand of big data experts, data scientist and other related position titles. Because of the initial decision support, specific needs in terms of skills and knowledge can be identified, especially for current job descriptions. By reducing the general confusion, the acceptancy and willingness of an application of big data may rise. Thus, enterprises could profit on a large scale, even though such a system would be only utilized as a first quick-check whether it is reasonable to apply big data technologies or not.

5 CONCLUSIONS

In this work, the term of big data engineering was investigated and also defined. Along with this, the current challenges as well as potentials were highlighted for each subdomain. As it has been observed, currently a lot of uncertainty and confusion exists. A first step facilitating this situation was presented through the idea of a structured utilization of a DSS. In the future, those systems could be highly beneficial, freeing the time needed for the planning for the actual realization.

REFERENCES

Altarturi, H. H., Ng, K.-Y., Ninggal, M. I. H., Nazri, A. S. A. and Ghani, A. A. A. (2017), "A requirement engineering model for big data software", In *2017 IEEE Conference on Big Data and Analytics, Kuching, IEEE, pp. 111–117.*

Arruda, D. and Madhavji, N. H. (2018), "State of Requirements Engineering Research in the Context of Big Data Applications", In *Kamsties, E. (Ed.) Requirements Engineering: Foundation for Software Quality: 24th International Working Conference, Vol. 10753, Springer, pp. 307–323.*

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S. and Buyya, R. (2015), "Big Data computing and clouds: Trends and future directions", *Journal of Parallel and Distributed Computing, 79-80, pp. 3–15.*

Debortoli, S., Müller, O. and Vom Brocke, J. (2014), "Comparing Business Intelligence and Big Data Skills", *Business & Information Systems Engineering, Vol. 6 No. 5, pp. 289–300.*

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "Knowledge Discovery and Data Mining: Towards a Unifying Framework", In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 82–88.*

Fekete, D. and Vossen, G. (2015), "The GOBIA Method: Towards Goal-Oriented Business Intelligence Architectures".

Gandomi, A. and Haider, M. (2015), "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management, Vol. 35 No. 2, pp. 137–144.*

Gani, A., Siddiqa, A., Shamshirband, S. and Hanum, F. (2016), "A survey on indexing techniques for big data: taxonomy and performance evaluation", *Knowledge and Information Systems, Vol. 46 No. 2, pp. 241–284.*

Gardiner, A., Aasheim, C., Rutner, P. and Williams, S. (2018), "Skill Requirements in Big Data: A Content Analysis of Job Advertisements", *Journal of Computer Information Systems, Vol. 58 No. 4, pp. 374–384.*

IEEE (2011), ISO/IEC/IEEE Systems and software engineering -- Architecture description, *IEEE.*

IEEE (2017), ISO/IEC/IEEE International Standard - Systems and software engineering - Software life cycle processes, *1st ed., IEEE.*

ISO (2014), Big data: Preliminary Report, available at: *www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-jtc1.pdf* (accessed 17 January 2019).

Jin, X., Wah, B. W., Cheng, X. and Wang, Y. (2015), "Significance and Challenges of Big Data Research", *Big Data Research, Vol. 2 No. 2, pp. 59–64.*

Katal, A., Wazid, M. and Goudar, R. H. (2013), "Big data: Issues, challenges, tools and Good practices", In *Parashar (Hg.) – 2013 sixth International Conference, pp. 404–409.*

Kreps, J. (2014), "Questioning the Lambda Architecture. The Lambda Architecture has its merits, but alternatives are worth exploring", available at:

https://www.oreilly.com/ideas/questioning-the-lambda-architecture (accessed 5 December 2018).

Lehmann, D., Fekete, D. and Vossen, G. (2016), Technology selection for big data and analytical applications, Working Papers, ERCIS - European Research Center for Information Systems, available at: *http://hdl.handle.net/10419/156084*.

Martínez-Prieto, M. A., Cuesta, C. E., Arias, M. and Fernández, J. D. (2015), "The Solid architecture for real-time management of big semantic data", *Future Generation Computer Systems, Vol. 47, pp. 62–79*.

Marz, N. (2011), "How to beat the CAP theorem", available at: *http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html* (accessed 5 December 2018).

Mauro, A. de, Greco, M. and Grimaldi, M. (2016), "A formal definition of Big Data based on its essential features", *Library Review, Vol. 65 No. 3, pp. 122–135*.

McAfee, A. and Brynjolfsson, E. (2012), "Big Data: The Management Revolution", *Harvard Business Review, Vol. 91 No. 5, pp. 1–9.*

Nadal, S., Herrero, V., Romero, O., Abelló, A., Franch, X., Vansummeren, S. and Valerio, D. (2017), "A software reference architecture for semantic-aware Big Data systems", *Information and Software Technology, Vol. 90, pp. 75–92.*

NIST (2015), "NIST Big Data Interoperability Framework: Volume 1, Definitions" (accessed 15 January 2019).

Nižetić, I., Fertalj, K. and Milašinović, B. (2007), "An Overview of Decision Support System Concepts".

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A. and Belfkih, S. (2018), "Big Data technologies: A survey", *Journal of King Saud University - Computer and Information Sciences, Vol. 30 No. 4, pp. 431–448.*

Pääkkönen, P. and Pakkala, D. (2015), "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems", *Big Data Research, Vol. 2 No. 4, pp. 166–186.*

Philip Chen, C. L. and Zhang, C.-Y. (2014), "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", *Information Sciences, Vol. 275, pp. 314–347.*

Rogers, G. F. C. (1983), The Nature of Engineering. A Philosophy of technology, The Macmillan Press LTD.

Sagiroglu, S. and Sinanc, D. (2013), "Big data: A review", In 2013 International Conference on Collaboration Technologies and Systems (CTS), *IEEE, San Diego, pp. 42–47.*

Shearer, C. (2000), "The CRISP-DM Model: The New Blueprint for Data Mining", *Journal of Data Warehousing, Vol. 5 No. 4.*

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017), "Critical analysis of Big Data challenges and analytical methods", *Journal of Business Research, Vol. 70, pp. 263–286.*

Sommerville, I. and Sawyer, P. (1997), Requirements engineering: A good practice guide, Wiley.

Staegemann, D., Hintsch, J. and Turowski, K. (2019), "Testing in Big Data: An Architecture Pattern for a Development Environment for Innovative, Integrated and Robust Applications", In *Proceedings of the WI2019, pp. 279–284.*

Szyperski, C., Petitclerc, M. and Barga, R. (2016), "Three Experts on Big Data Engineering", *IEEE Software, Vol. 33 No. 2, pp. 68–72.*

Taleb, I., Serhani, M. A. and Dssouli, R. (2018), "Big Data Quality: A Survey", In *2018 IEEE International Congress on Big Data, IEEE, Los Alamitos, pp. 166–173.*

Tao, C. and Gao, J. (2016), "Quality Assurance for Big Data Application – Issuses, Challenges, and Needs", In *The 28th International Conference on Software Engineering and Knowledge Engineering, KSI Research Inc., Redwood City, pp. 375–381.*

Taschner, A. (2017), "Definitionen", In *Business Cases: Ein anwendungsorientierter Leitfaden, Springer Fachmedien Wiesbaden, pp. 5–10.*

Thome, G. and Sollbach, W. (2007), Grundlagen und Modelle des Information Lifecycle Management, Xpert.press, *Springer-Verlag Berlin Heidelberg*.

Turck, M. (2018), "The Big Data Landscape", available at: *http://dfkoz.com/big-data-landscape/* (accessed 11 January 2019).

van Putten, B.-J., Brecht, F. and Günther, O. (2013), "Challenges in Business Case Development and Requirements for Business Case Frameworks", Invan Putten, B.-J. (Ed.) Supporting Reuse in Business Case Development, *Springer, pp. 8–22.*

Vogel, O., Arnold, I., Chughtai, A. and Kehrer, T. (2011), Software Architecture, Springer Berlin Heidelberg.

Volk, M., Jamous, N. and Turowski, K. (2017), "Ask the Right Questions - Requirements Engineering for the Execution of Big Data Projects", In *23rd Americas Conference on Information Systems, AIS, Boston*.

Volk, M., Pohl, M. and Turowski, K. (2018), "Classifying Big Data Technologies - An Ontology-based Approach", In *24rd Americas Conference on Information 2018, AIS, New Orleans*.

Wasson, C. S. (2016), System analysis, design, and development: Concepts, principles, and practices, Wiley series in *systems engineering and management, 2nd edition, John Wiley & Sons*.

Yin, S. and Kaynak, O. (2015), "Big Data for Modern Industry: Challenges and Trends [Point of View]", *Proceedings of the IEEE, Vol. 103 No. 2, pp. 143–146.*

Zicari, R. V., Rosselli, M., Ivanov, T., Korfiatis, N., Tolle, K., Niemann, R. and Reichenbach, C. (2016), "Setting Up a Big Data Project: Challenges, Opportunities, Technologies and Optimization", In *Emrouznejad, A. (Ed.) Big Data Optimization: Recent Developments and Challenges, Springer, pp. 17–47.*

Zikopoulos, P. (2012), Understanding big data, McGraw-Hill.