Approaching the (Big) Data Science Engineering Process

Matthias Volk, Daniel Staegemann, Sascha Bosse, Robert Häusler 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, Data Science, Engineering, Process.

Abstract: For many years now, researchers as well as practitioners are harnessing well-known data mining processes, such as the CRISP-DM or KDD, to realize their data analytics projects. In times of big data and data science, at which not only the volume, variety and velocity of the data increases, but also the complexity to process, store and manage them, conventional solutions are often not sufficient and even more sophisticated systems are needed. To overcome this situation, in this positioning paper the (big) data science engineering process is introduced to provide a guideline for the realization of data-intensive systems. For this purpose, using the design science research methodology, existing theory and current literature from relevant subdomains are contextualized, discussed and adapted.

1 INTRODUCTION

In the last decade, big data was one of the major trends in the computer science domain. Every year, more and more researchers and practitioners are harnessing the power that comes with new technologies, techniques, and paradigms to conduct their data-intensive projects. Application areas include, but are not limited to, industry (Reis and Gins, 2017) or the government sector (Kim *et al.*, 2014). However, despite the increasing interest, the proven added value through its application (Müller *et al.*, 2018; Fosso Wamba *et al.*, 2015), and the growing maturity of the topic as well as the related technologies, most of these projects fail to meet the expectations. Although lots of general guidelines, best practices and workflows exist, such as presented in (Pääkkönen and Pakkala, 2015; Chen *et al.*, 2015; Volk *et al.*, 2017; Grady, 2016), these either provide isolated activities or do not cover relevant projects' steps from start-to-end. Additionally, some of the widely accepted approaches known from the data mining domain, like the knowledge discovery in databases (KDD) (Fayyad *et al.*, 1996), Cross-Industry-Standard-Process-for-Data-Mining (CRISP-DM) (Shearer, 2000) or Sample, Explore, Modify, Model and Assess (SEMMA) (Azevedo and Santos, 2008), are no longer easily applicable. This results especially from the missing consideration of the technical implementation as well as deployment and operational activities. Eventually, this complicates the applicability of useful processes and additionally confuses the users regarding the available options.

Due to the nature of big data, numerous elaborated decisions need to be made regarding the corresponding collection, preparation, processing and storing (Pääkkönen and Pakkala, 2015). This includes, for instance, the general identification of relevant components and specific technologies. Most of these tasks fall under the term of big data engineering that can be defined 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" (Volk *et al.*, 2019). In here, the authors already highlighted the importance of the discipline for the future construction of big data systems. At the same time, it became apparent that currently no systematic process exists, which covers the system engineering from a start-to-end perspective in the domain of big data. Especially, due to the reason that this topic is also heavily interwoven with other domains and builds a technical foundation for the realization of their projects, such as data science (Provost and Fawcett, 2013) or industry 4.0 (Xu and Duan, 2019), this may have a highly beneficial influence on all data-intensive systems. For the aforementioned reasons, we argue that one comprehensive approach that emerges from well-

Volk, M., Staegemann, D., Bosse, S., Häusler, R. and Turowski, K.
Approaching the (Big) Data Science Engineering Process.
DOI: 10.5220/0009569804280435
In *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security (IoTBDS 2020)*, pages 428-435
ISBN: 978-989-758-426-8

known theory could help to resolve the opacity of existing "guidelines". Hence, the following research question shall be answered in the course of this work:

"How can a system engineering process for data-intensive projects be modeled to provide coverage for each step from the planning to the development and operation?"

The outcome of this research is intended to provide a process model that shall serve as a guideline for researchers and practitioners to realize their data-intensive projects. In order to substantiate the argumentation and positioning of our statements on a scientific basis, the design science research methodology according to (Hevner *et al.*, 2004; Peffers *et al.*, 2007) is used. Resulting out of this, the contribution at hand is structured as follows. While the first section provided a brief motivation, information about the current situation and the main objective, the subsequent section discusses relevant theories and approaches. This shall mainly serve as a foundation for the design and development of the intended artifact, which takes place in the third section of this paper. Apart from the general description of the artifact itself, initially existing approaches are discussed and a critical discussion as well as an outlook on future research provided. The work ends with a conclusion.

2 BACKGROUND

To provide a better understanding of the topics we are positioning on, the theory and related work about each of the relevant domains are presented. This is not to deliver an overview but also to confirm the needed theory identification (Peffers *et al.*, 2007) for the actual design and development of the artifact. Hence, in the following section, relevant information, as well as related research articles, are presented. This comprises big data, data science, systems engineering and big data engineering processes.

2.1 Big Data

While big data is an intensely discussed topic that is dealt with by a multitude of publications, as shown in (Staegemann *et al.*, 2019b), there is still no single one, universally applied definition for the term itself. One of the most commonly used definitions is provided by the National Institute of Standards and Technology (NIST), stating that big data "consists of extensive datasets primarily in the characteristics of volume, velocity, variety, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis" (NIST, 2019). While volume indicates the sheer amount of data, regarding volume or size, that are to be processed (Russom, 2011), velocity stands for the speed with which those data are incoming as well as the timeliness in which results are expected by the users (Gandomi and Haider, 2015). Another challenge lies in the data's heterogeneity, including for example different structures (structured, semi-structured, unstructured), formats, units of measurement or contexts, subsumed under the term variety (Gani *et al.*, 2016). Moreover, variability signifies the ability of those aforementioned characteristics to change over time, with the same applying to the determined questions as well as the data's content (Katal *et al.*, 2013; Wu *et al.*, 2014). However, those characteristics do not even include the factor of the data quality, which is in turn highly influential on the quality of the obtained analysis results (Hazen *et al.*, 2014), or the task of verifying the validity of the developed application, adding even more challenges to the topic. Furthermore, besides big data's inherent complexity due to those factors and its multidimensional nature, combining technical, human and organizational aspects (Alharthi *et al.*, 2017), also the abundance of potentially available tools and techniques (Turck and Obayomi, 2019) increases the difficulty when attempting to engineer such a system.

2.2 Data Science

The volume of *big data* led not only to an increased demand of data-intensive systems, at the same time methodologies and theories to investigate those massive amounts were needed (Cao, 2017). As a consequence, the term data science evolved that can be described as "a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" (Provost and Fawcett, 2013). For the actual information extraction and knowledge creation, different types of analytics are frequently used today, such as descriptive-, predictive- or prescriptive analytics (Cao, 2017).Due to the origin from the data mining domain, for the actual realization of related projects, today often well-known approaches are used (Provost and Fawcett, 2013). This includes, for instance, the already mentioned processes KDD, SEMMA and most of all the CRISP-DM (Piatetsky, 2014). Although most of them differ in parts, all of them share a similar understanding when it comes to the exploration of the data in a structured way. Azevedo and Santos (2008), for instance, performed a thorough comparison of all three approaches. In their work, they highlighted that

the KDD can be more observed as an implementation of the other two approaches and that the CRISP-DM appears to be more complete than the others are. Apart from the sole preparation and analysis of the data, this approach explicitly incorporates the business understanding in an organizational context as well as the deployment of the targeted solution. This results in the six phases of *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation,* and *deployment* (Shearer, 2000; Azevedo and Santos, 2008).

2.3 Systems Engineering

The design and development of information systems remains one of the major disciplines for many years now. Due to this, it is not surprising that many researchers, as well as practitioners, are attempting to provide approaches and guidelines for the successful planning and engineering of the systems and software, such as (Hevner *et al.*, 2004; Peffers *et al.*, 2007; Mobus and Kalton, 2015; Nicholas and Steyn, 2012b; Sommerville, 2016). Commonly, each system, independently from its nature, follows a life cycle that runs from the initial identification of the need for the system over the system analysis, the design, the construction, and operation until the decommissioning (Mobus and Kalton, 2015). Nicholas and Steyn (2012b) refer to systems engineering (SE) as "a way to bring a whole system into being and to account for its whole life cycle". This is comparable to other definitions such as from the non-profit organization International Council on Systems Engineering (INCOSE). According to INCOSE, the term can be observed as a "transdisciplinary and integrative approach to enable the successful realization, use, and retirement of engineered systems, using systems principles and concepts, and scientific, technological, and management methods" (INCOSE, 2020). Despite those explanations, as well as, the needed integration of different concepts, technologies and (sub-) systems, sometimes from different domains, the SE can be seen as a meta-engineering discipline (Mobus and Kalton, 2015). Apart from the general description of the term and the life cycle of a system, the authors also developed one of the most widely accepted approaches to engineer those systems while covering the mentioned life cycle stages. The process is depicted in Figure 1. This approach synthesizes most of the existing approaches, such as from (Nicholas and Steyn, 2012b), but in more detail. The process covers seven main steps that range from the initial identification of the problem until the operation.

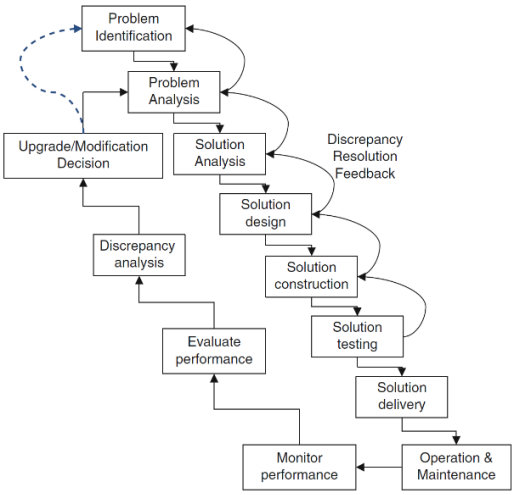

Figure 1: SE Life Cycle (Mobus and Kalton, 2015).

Within the *problem identification*, the same will be performed. Important here is that the actual problem will be discovered and not only (obvious) implications of it, providing a problem-centric instead of a solution-centric view. Afterward, the problem needs to be specified in more detail by developing requirements within the *problem analysis* stage. Inter alia, this can be realized through decomposition and separate observation of relevant sub-problems. Besides that, boundaries of the system functionalities can be determined. The identified problems are subsequently used as an input for the *solution analysis* that pursues to present a possible system specification. In doing so, smaller, logically independently acting parts of the system (sub-system) and their interconnections are identified. The specifications of those should conform to the needs ordinated from the problem analysis. After all relevant specifications were made, the *solution design* is taking place at which the physical aspects are determined. By the end, design documents are formulated, which serve as an input for the *solution construction*. Within this step, the actual systems, which is often referred to as the artifact, is developed. Eventually, the developed artifacts need to be evaluated, which will be performed in the *solution testing* phase. Although at this point a comprehensive verification will be performed, concurrent validations during each of the previous stages are recommended that are covered under the *discrepancy resolution feedback*. This in turn, may lower the need to perform changes in later stages. If everything was successfully developed, the solution is delivered and productively used. Continuing steps cover the monitoring, performance monitoring, and further analysis, which

may lead to modifications, upgrades or decommissioning (Mobus and Kalton, 2015).

2.4 Big Data Engineering

Due to the general characteristics of big data, the realization of related projects differs greatly from conventional IT projects. Most of all, the implementation is associated with much more data handling and interpretation, since the these differ quite strongly in terms of variety, velocity and volume (cf. 2.1). Hence, the development of suitable systems appears to be even more demanding. The engineering of systems in the area of big data is described by the term *big data engineering* (Volk *et al.*, 2019). Relevant domains and activities were already identified in numerous publications, but to our knowledge, no comprehensive start-to-end process exists that not only observes the general data analysis but also the technical implementation and operation. In most of the cases, only the project realization (Dutta and Bose, 2015; Mousannif *et al.*, 2016; Li *et al.*, 2016; Grady, 2016) or specific activities, needed for this, were thoroughly investigated. This includes, for instance, the general planning, requirements engineering steps (Volk *et al.*, 2017; Altarturi *et al.*), the identification of the suitable technologies (Lehmann *et al.*, 2016) and most of all relevant reference architectures (Martínez-Prieto *et al.*, 2015; Jay Kreps, 2014; Nadal *et al.*, 2017). Especially the latter can be highly beneficial when it comes to the limitation of available options for technologies to be applied and guidance during the construction of the system. However, the selection of these can be a very demanding task, mainly due to the same reason, the availability of numerous big data technologies, such as highlighted in (Turck and Obayomi, 2019). Hence the thorough planning and requirements engineering represents an initial step for the construction of the needed system (Volk *et al.*, 2019). Followed by activities that identifies relevant components and specify them in terms of their connection and technological implementation.

3 DESIGN AND DEVELOPMENT

In consideration of the previously described data science and engineering domain, it becomes apparent that the discipline of big data engineering unites both domains. Many researchers are aware of the importance of big data in concatenation with data science. Due to this, many attempts to provide guidelines for the realization of those *projects* exist. To provide an enhanced overview over the base of argumentation we are positioning on, subsequently, an excerpt from the current state of the art is presented. Afterward, the *(big) data science engineering process* (BDSEP) is presented.

3.1 State of the Art

By performing a three-stepped literature review according to the methodology of (Levy and J. Ellis, 2006) and the forward-backward procedure proposed by Webster and Watson (2002) relevant publications were identified. In the following, each of the found out papers is shortly described. Dutta and Bose (2015) introduced a holistic roadmap that attempts to guide organizations by the conceptualization, planning and implementation of big data projects. Although the previously mentioned data science processes are introduced within the contribution and noticeable similarities ascertained, no concrete details about their relationship to the workflow are described. An explicit connection between big data and the previously referred data science processes was made in (Grady, 2016). It presents a mixture of the KDD, the CRISP-DM and parts of the big data domain. In particular, a five-stepped procedure is developed, that covers the planning, collection, curating, analysis and acting. Another process model that interconnects the KDD with big data was presented by Li *et al.* (2016). In their contribution, a snail shell process model for knowledge discovery, the proposed eight-stepped procedure heavily relies on the key activities used in the KDD process and involves the lifecycle presentation of the CRISP-DM model. A similar approach was found in (Mousannif *et al.*, 2016). The authors propose a big data project workflow that describes the realization of big data projects step-by-step. Additionally to that, concrete technical implementation details, such as specific technologies, are addressed. These detailed system observations are even more concretized in (Chen *et al.*, 2015), which propose a new method called *Big Data system Design.* The procedure consists of ten essential steps, starting from the requirements analysis to the design and implementation. Here, reference architectures are considered as a suitable foundation. The same applies to the implicit application of system engineering-related activities, such as the decomposition of the solution for better understanding. Compared to previously described contributions, this work rather focuses on the technical implementation and thus the system engineering of big data related systems. However, the theoretic background is little described and data science-related activities not included. IBM

developed a *step-by-step* guide that extends the CRISP-DM. The Analytics Solutions Unified Method (ASUM) presents a hybrid approach that attempts to integrate agile as well as traditional principles in combination with big data relevant aspects (IBM, 2016). Yet, as in the case of the previous approaches, the process describes the needed steps without any concrete implementation details. Another approach that attempts to provide a general guideline for the realization of big data projects is presented in (Volk *et al.*, 2017). By including a specific requirements engineering strategy, the KDD and metrics to check the general sensibility of a big data project, a through project instantiation process are introduced.

3.2 The (Big) Data Science Engineering Process (BDSEP)

Again, in none of the approaches a completed process that enlightens the realization of data-intensive projects in a combination of the data science and systems engineering domain, was found. Instead, different combinations of all of the aforementioned domains were ascertained. Especially the CRISP-DM (Shearer, 2000) and SE methods, known from the recommended workflow by (Mobus and Kalton, 2015), were either implicitly or explicitly used. Due to this, we argue that the linkage of both approaches in addition to big data-related specifics appears sensible. Although both of the approaches attempt to achieve different goals, a closer comparison of each of the related steps reveal similarities. This applies not only for the general problem identification (business understanding) and for problem analysis (data analysis) but for the solution construction (modeling), solution testing (evaluation) and solution delivery (deployment) as well. Differences, in turn, are predominantly noticeable in terms of the main scope. While the CRISP-DM intends to rather focus on the analysis of the data, the SE pursues the engineering of the implementation. However, in both processes the supplemented steps are implicitly integrated. Due to the aforementioned reasons above, a mixture out of both of the approaches was chosen. In particular the system engineering process from (Mobus and Kalton, 2015) was used as a base and extended by the thorough data investigation. The concrete workflow of the process is depicted in Figure 2. Within this figure, the referred *foundation* comprises all steps of the SE procedure until the operation, in combination with the data understanding from the CRISP-DM. In contrast to the other steps, the data understanding was explicitly integrated, due to the importance of the data be processed. The BDSEP together with the most important information and the general focus are depicted in the second layer of the figure. In here, the workflow contains the steps as execution directives. It starts with the formulation of the vision or idea. Due to the closely related concept of IT project realizations, the starting point is not necessarily limited to a *problem*. Moreover, the general description of an overarching vision or idea for the project may appear also as a sufficient starting point. Furthermore, also contracts might be forming the base for the instantiation of the engineering. Independent from its origin, the detailed identification of the main scope is the result of this step. This serves as a transition to the in-depth analyzes of the *use case*. At this point, scenario descriptions and use case diagrams may serve as an additional help to formalize the problem and set its boundaries, such as highlighted in (Chen *et al.*, 2015; Sommerville, 2016). Apart from this, relevant stakeholders and the data to be used need to be determined in here. For instance, if the data is gathered multiple times from a multitude of data sources, sophisticated orchestration activities are later on required (Khalifa *et al.*, 2016). Furthermore, due to the strong relationship between the data and requirements in data-intensive environments, the

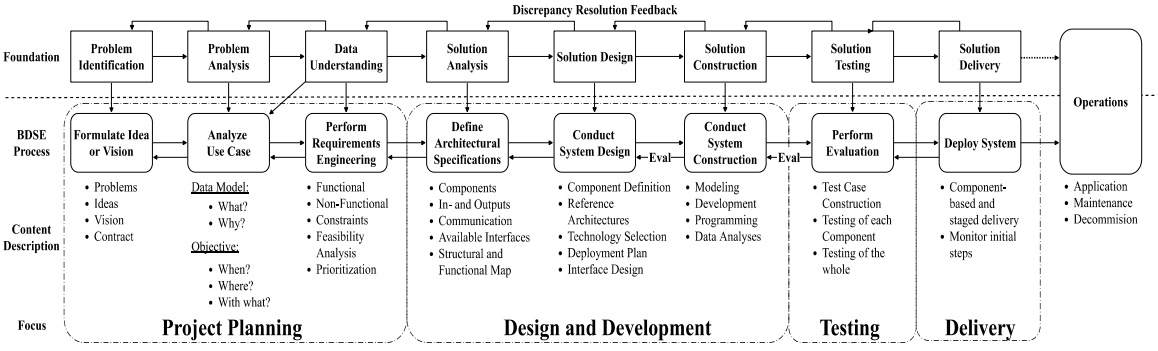

Figure 2: The (Big) Data Science Engineering Process.

specific characteristics should be uncovered, such as highlighted in (Volk *et al.*, 2017). Among other things, the template of Chen *et al.* (2015) could be taken into account, that comprises 14 essential data requirements. Further, the requirements engineering step finishes the general planning of the projects, by the development of the functional, non-functional requirements and constrains. While the functional requirements, define general functions of the system to be performed, the non-functional requirements are focusing on system properties (Sommerville, 2016). Prioritizations and feasibility analysis can be helpful at this process step (Nicholas and Steyn, 2012a). In any case, the requirements should be developed as thorough as possible, to avoid massive changes on the system architecture. After the project planning is finished, the design and development takes places. At the beginning, specifications are needed, at which basic components their relation, functionalities, performance and future tests are defined (Mobus and Kalton, 2015; Nicholas and Steyn, 2012b). Additional inputs and outputs as well as available interfaces are relevant in terms of this. For better depiction and understanding of the system, structural and functional maps could be used, known from the system decomposition (Mobus and Kalton, 2015). After each of the needed elements is determined, the system design is conducted. This includes most of all the definition of the component specifics. In the area of big data, multiple technologies exist that can be used for different purposes (Turck and Obayomi, 2019). Hence, the adequate selection of suitable solutions during the specification of the required architecture represents a sophisticated undertaking. At this point, best practices (Pääkkönen and Pakkala, 2015), reference architectures (Martínez-Prieto *et al.*, 2015; Jay Kreps, 2014; Nadal *et al.*, 2017) and decision support systems appear to be useful (Volk *et al.*, 2019). All of them intent to provide general guidelines for the construction of the system architecture, as well as, in parts, concrete implementation details and technology recommendation. In any case, the requirements originating from the previous phase need to be discussed in a thorough manner. However, not always are the made decisions final and in some cases further modification are required, for instance, in terms of the technologies or patterns to be used (Mobus and Kalton, 2015; Li *et al.*, 2016). After all of the required elements and their interconnections were identified, the actual combination and construction of the solution takes place. Apart from the development of the system itself, this includes the programming or modeling of the needed application running on the system. After the solution was constructed, it needs to be evaluated whether everything is working correctly or not. For that reason, a thorough testing procedure is needed, comprising significant test cases that cover the validation of the separate components as well as the system as a whole. However, the properties of the big data domain turn this into a highly sophisticated task (Staegemann *et al.*, 2019b). It is necessary to cover a variety of technologies, types and sources of data, connections and requirements. At the same time, the demand for future scalability and an often prevailing lack of knowledge regarding the correct outcome, which complicates a verification, pose additional challenges. Furthermore, even apparently small flaws like for example rounding errors can be built up during the processing, amounting to considerable derivations from the correct result (Yang *et al.*, 2018). At the same time, while being highly important and complex, the testing of big data applications is not sufficiently acknowledged in the literature (Staegemann *et al.*, 2019a). Subsequently to the successful evaluation, the developed solution can be deployed. In context of the described process, the step refers to the actual distribution of the solution in the targeted environment. In case of complex systems, Mobus and Kalton (2015) highlight that this should be realized in a staged process, to uncover unforeseen issues. Especially in the domain of big data, this should be recognized. Due to the high number of existing technologies and their versions, compatibility issues can easily emerge. This is not restricted to the dependencies between the used components, but also the targeted environment (Chen *et al.*, 2015). Hence, during the delivery, comprehensive monitoring activities are required. Eventually, the actual application of the developed solution and its further maintenance will be performed during the operation phase. As prescribed in most of the existing approaches, for each problem encountered in one of the steps, considerations and tasks of a previous step should be revised.

3.3 Discussion

The BDSEP covers relevant steps needed for the engineering of data-intensive systems. Instead of presenting a stepwise procedure that meticulously describes every single step in a detailed chronological way, we attempt to draw the attention on the big picture. This is especially due to the reason that each project differs quite strongly and the same applies to the engineering of the system. Sometimes sophisticated procedure may be required in larger projects, while smaller ones are rather lacking on a

general plan. Independent from its size, attention should be most importantly to the data to be stored, managed and processed. Compared to other approaches, we built our positioning upon existing theory, in particular from the data science, big data and systems engineering domain. Hence, for detailed information about the referred activities, potential users can make use of the referenced contributions. In the future, it is planned to evaluate this process in large scale. Consequently, possible shortcomings and best practices can be identified in more detail and contributed back to the BDSEP. While this approach serves as an initial starting point, providing an overview regarding the steps to be conducted, future observations and changes can reinforce the general applicability. This applies especially for the detailed investigation of particular steps and their relevant activities. Within the requirements engineering part, for instance, agile project management principles were not discussed in detail. For now, the requirements are considered as to be completed. In context of this, another direction could be realized through the test-driven development, at which the test of the relevant component or system is developed before the targeted element itself. Further principles that are worthy to be examined are related to the operations phase and their transition to it. Approaches, such as continuous delivery or DevOps in general, appear to be sensible, especially in context of fast changing fields of a *data-intensive* nature.

4 CONCLUSIONS

In the last decade, big data was one of the most regarded topics in the computer science domain. However, many issues are still existing today that are challenging the realization of corresponding projects and the needed systems. Although many processes, best practices and other relevant workflows for the realization of data-intensive projects arose, still a lot of insecurity about their applicability exists. By harnessing the design science research methodology, we uncovered intersection points of some of the most prominent approaches and adapted them to create a comprehensive (*big) data science engineering process*. This process unifies knowledge and best practices from the information systems engineering domain as well as data science processes to overcome the stated problem. Researchers and practitioners benefit from this artifact, especially when it comes to the structured planning and realization of data-intensive projects.

REFERENCES

Alharthi, A., Krotov, V. and Bowman, M. (2017), "Addressing Barriers To Big Data", *Business Horizons*, Vol. 60 No. 3, Pp. 285–292.

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

Azevedo, A. and Santos, M.F.d. (2008), "KDD, SEMMA and CRISP-DM: a parallel overview", *undefined*.

Cao, L. (2017), "Data Science", *ACM Computing Surveys*, Vol. 50 No. 3, pp. 1–42.

Chen, H.-M., Kazman, R., Haziyev, S. and Hrytsay, O. (2015), "Big Data System Development: An Embedded Case Study with a Global Outsourcing Firm", In *First International Workshop on Big Data Software Engineering - BIGDSE 2015*,IEEE,pp. 44–50.

Dutta, D. and Bose, I. (2015), "Managing a Big Data project: The case of Ramco Cements Limited", *International Journal of Production Economics*, Vol. 165, pp. 293–306.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, Vol. 17 No. 3, p. 37.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015), "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study", *International Journal of Production Economics*, Vol. 165, pp. 234–246.

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.

Grady, N.W. (2016), "KDD meets Big Data", InJoshi, J. (Ed.) *2016 IEEE International Conference on Big Data*, IEEE, Piscataway, NJ, pp. 1603–1608.

Hazen, B.T., Boone, C.A., Ezell, J.D. and Jones-Farmer, L.A. (2014), "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications", *International Journal of Production Economics*, Vol. 154, pp. 72–80.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004), "Design Science in Information Systems Research", *MIS Quaterly*, Vol. 28 No. 1, pp. 75–105.

IBM (2016), "Analytics Solutions Unified Method", available at:ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf (accessed 3 Feb. 2020).

INCOSE (2020), "Systems Engineering", available at: https://www.incose.org/systems-engineering (accessed 16 January 2020).

Jay Kreps (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 21 January 2020).

Katal, A., Wazid, M. and Goudar, R.H. (2013), "Big data: Issues, challenges, tools and Good practices", InParashar (Ed.) *Sixth International Conference on Contemporary Computing, Noida, India, 08.08.2013 - 10.08.2013*, IEEE, pp. 404–409.

Khalifa, S., Elshater, Y., Sundaravarathan, K., Bhat, A., Martin, P., Imam, F., Rope, D., Mcroberts, M. and Statchuk, C. (2016), "The Six Pillars for Building Big Data Analytics Ecosystems", *ACM Computing Surveys*, Vol. 49 No. 2, pp. 1–36.

Kim, G.-H., Trimi, S. and Chung, J.-H. (2014), "Big-data applications in the government sector", *Communications of the ACM*, Vol. 57 No. 3,pp. 78–85.

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.

Levy, Y. and J. Ellis, T. (2006), "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research", *Informing Science: The International Journal of an Emerging Transdiscipline*, Vol. 9, pp. 181–212.

Li, Y., Thomas, M.A. and Osei-Bryson, K.-M. (2016), "A snail shell process model for knowledge discovery via data analytics", *Decision Support Systems*, Vol. 91, pp. 1–12.

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.

Mobus, G.E. and Kalton, M.C. (2015), *Principles of Systems Science, Understanding Complex Systems*, Springer.

Mousannif, H., Sabah, H., Douiji, Y. and Oulad Sayad, Y. (2016), "Big data projects: just jump right in!", *International Journal of Pervasive Computing and Communications*, Vol. 12 No. 2, pp. 260–288.

Müller, O., Fay, M. and Vom Brocke, J. (2018), "The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics", *Journal of management information systems*, Vol. 35 No. 2, pp. 488–509.

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.

Nicholas, J.M. and Steyn, H. (Eds.) (2012a), *Project Management for Engineering, Business, and Technology,* Fourth Edition, Butterworth-Heinemann.

Nicholas, J.M. and Steyn, H. (2012b), "Systems Approach and Systems Engineering", InNicholas, J.M. and Steyn, H. (Eds.) *Project Management for Engineering, Business, and Technology*, pp. 46–82.

NIST (2019), *NIST Big Data Interoperability Framework: Volume 1, Definitions, Version 3*.

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.

Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007), "A design science research methodology for information systems research", *Journal of management information systems*, Vol. 24 No. 3, pp. 45–77.

Piatetsky, G. (2014), "CRISP-DM, still the top methodology for analytics, data mining, or data science projects - KDnuggets", available at: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (accessed 9 January 2020).

Provost, F. and Fawcett, T. (2013), "Data science and its relationship to big data and data-driven decision making", *Big data*, Vol. 1 No. 1, pp. 51–59.

Reis, M. and Gins, G. (2017), "Industrial Process Monitoring in the Big Data/Industry 4.0 Era: from Detection, to Diagnosis, to Prognosis", *Processes*, Vol. 5 No. 3, pp. 35–50.

Russom, P. (2011), "Big Data Analytics. TDWI Best Practices Report Fourth Quarter 2011".

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

Sommerville, I. (2016), *Software engineering,* 10. ed.

Staegemann, D., Volk, M., Jamous, N. and Turowski, K. (2019a), "Understanding Issues in Big Data Applications - A Multidimensional Endeavor".

Staegemann, D., Volk, M., Nahhas, A., Abdallah, M. and Turowski, K. (2019b), "Exploring the Specificities and Challenges of Testing Big Data Systems".

Turck, M. and Obayomi, D. (2019), "The Big Data Landscape", available at: http://dfkoz.com/big-data-landscape/ (accessed 13 January 2020).

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.

Volk, M., Staegemann, D., Pohl, M. and Turowski, K. (2019), "Challenging Big Data Engineering: Positioning of Current and Future Development", In *Proceedings of the IoTBDS 2019*, SCITEPRESS - Science and Technology Publications, pp. 351–358.

Webster, J. and Watson, R.T. (2002), "Analyzing the Past to Prepare for the Future: Writing a Literature Review", *MIS Quaterly*, Vol. 26 No. 2, pp. xiii–xxiii.

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014), "Data mining with big data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26 No. 1, pp. 97–107.

Xu, L.D. and Duan, L. (2019), "Big data for cyber physical systems in industry 4.0: a survey", *Enterprise Information Systems*, Vol. 13 No. 2, pp. 148–169.

Yang, M., Adomavicius, G., Burtch, G. and Ren, Y. (2018), "Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining", *Information Systems Research*, Vol. 29 No. 1, pp. 4–24.