Exploring the Specificities and Challenges of Testing Big Data Systems

Daniel Staegemann
MRCC VLBA
Otto-von-Guericke University
Magdeburg, Germany
daniel.staegemann@ovgu.de

Matthias Volk
MRCC VLBA
Otto-von-Guericke University
Magdeburg, Germany
matthias.volk@ovgu.de

Abdulrahman Nahhas
MRCC VLBA
Otto-von-Guericke University
Magdeburg, Germany
abdulrahman.nahhas@ovgu.de

Mohammad Abdallah
Department of Software Engineering
Al-Zaytoonah University of Jordan
Amman, Jordan
m.abdallah@zuj.edu.jo

Klaus Turowski
MRCC VLBA
Otto-von-Guericke University
Magdeburg, Germany
klaus.turowski@ovgu.de

***Abstract*— Today, the amount and complexity of data that is globally produced increases continuously, surpassing the abilities of traditional approaches. Therefore, to capture and analyze those data, new concepts and techniques are utilized to engineer powerful big data systems. However, despite the existence of sophisticated approaches for the engineering of those systems, the testing is not sufficiently researched. Hence, in this contribution, a comparison of traditional software testing, as a common procedure, and the requirements of big data testing is drawn. The determined specificities in the big data domain are mapped to their implications on the implementation and the consequent challenges. Furthermore, those findings are transferred into six guidelines for the testing of big data systems. In the end, limitations and future prospects are highlighted.**

Keywords*— *Big Data, System, Engineering, Verification, Validation, Testing, Benchmarking, Technologies, Guidelines

I. INTRODUCTION

Big data and its accompanying technologies dramatically changed many aspects of today's world by allowing the purposeful analysis of information that would have been forfeited just a few years ago [1]. The derived opportunities of those sources of knowledge affect a plethora of application areas like healthcare [2–4], civil protection [5, 6], business [7–10], opinion mining [11] and transportation [12]. Apart from the wide applicability, the ever increasing rate of data generation exemplifies the importance of technologies to create value from this fairly new resource that has been described as *digital oil* [13]. While the amount of data created by modern industry in the year 2015 is stated with about 1000 exabytes, it is predicted to increase 20-fold by 2025 [14].

Due to its ubiquitousness, versatility, and impact, this topic unites researchers and practitioners in their desire to push the boundaries and facilitate more and more advanced solutions in numerous areas [15–17]. This interest shows in a multitude of different facts and metric. From a scientific perspective, the number of publications, whose title incorporates "big data", in the scientific literature meta-database Scopus has grown immensely. While 3.056 of those publications are published between 2011 to 2014, the number between 2015 and 2018 adds up to 15.128 entries. Additionally, as Fig. 1Fig. *1* depicts, there is a continuing and significant annual growth in numbers in this timespan.

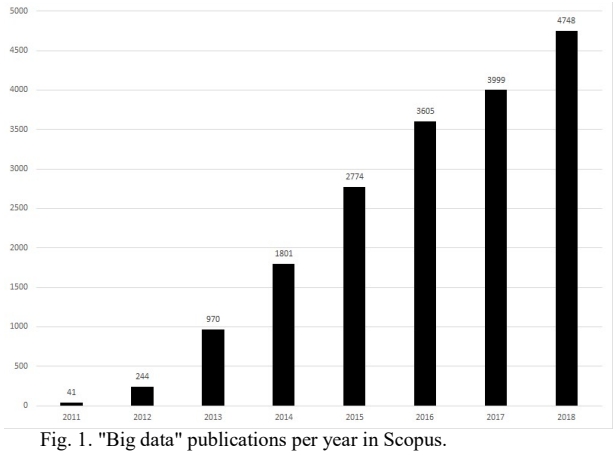

Fig. 1. "Big data" publications per year in Scopus.

However, not only the scientific figures thrive, but also the economical point of view suggests an outstanding significance of the topic. The market intelligence provider International Data Cooperation (IDC), for example, predicts that the worldwide market for big data and business analytics solution will reach $189.1 billion in 2019 and increase to $274.3 billion by 2022 [18]. Accompanying this growth in financial volume, the number of big data companies has reached a four-digit number [19]. Furthermore, [20] have shown the potential for increasing a company's productivity through the usage of big data analytics, therefore proving the factual economical value.

While the opportunities and attention are immense, the same applies to the challenges accompanying them [21]. One often overlooked challenge is the testing of the created big data systems landscape [22]. Even though the single components might have been tested extensively on their own, as it is often the case with popular applications, it is still required to ensure the correctness of their interaction. Thus, to assure the validity of the obtained findings. In this regard, the resulting task resembles traditional software testing. While the process of testing in the domain of conventional software has already been extensively researched [23], the same does not hold true for big data systems [24]. Therefore, to support the big data engineering procedure [25], the following research question will be answered in the course of this work:

978-1-7281-5686-6/19/$31.00 ©2019 IEEE
DOI 10.1109/SITIS.2019.00055

RQ1: What are the specificities of testing big data applications compared to common software testing?

RQ2: How can the identified differences be taken into account in the creation of test scenarios?

To answer the research questions, the publication is structured as follows. After introducing and motivating the topic in the first section, the second section discusses the fundamentals of big data and software testing. Subsequently, big data systems and software systems are compared, the challenges and requirements of big data testing are explored and guidelines for the testing are presented and discussed. The work ends with a conclusion that also includes limitations and future prospects.

II. FUNDAMENTALS

In the following, the domains of big data and software testing, constituting the fundamentals of the publication at hand, are introduced and the most important concepts explained.

A. Big Data

To answer the research questions, at first it is necessary to clarify the meaning of big data. While the term itself has no universally applied definition, there are several explanations, which are all describing the same phenomenon, but often slightly differ in detail. One of the most popular definitions is provided by the National Institute of Standards and Technology (NIST) and states, 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" [26].

Volume indicates the amount of data that has to be processed by the system to fulfill a given task. This can, on the one hand, refer to the number of records that have to be managed and, on the other hand, to the size of the handled data [27]. Either way, a tremendous increase of the volume is noticeable. Sometimes the relevant metrics are changed from gigabytes/terabytes to petabytes or even zettabytes [28]. In addition, a growing number of data points is generated and registered for the various subjects of interest [29]. Velocity also refers to two, not necessarily identical challenges. It can denominate the pace of incoming data that have to be handled by the system, but also the required speed when fulfilling a processing request [30]. Variety describes the multitude of different sources, data types, structures (structured/ semi-structured/ unstructured) and notational conventions that can be present in a single data analytics application [31]. While gathering that diverse information can yield significant benefits in terms of the gained insights, their integration can pose a major challenge. Variability represents the changes regarding the other dimensions. Since the real world is in a constant state of change, the data that are deemed relevant, as well as their amount are also continuously evolving. This dynamic must be taken into account during the design stage of big data systems to ensure a higher level of flexibility and scalability. Furthermore, distinct events can cause a short-term alteration of the composition of the received data, therefore posing additional challenges in terms of the system's flexibility [32]. Another important characteristic, despite not being mentioned in the initial definition, is the veracity. It "refers to the accuracy of the data" [26] and describes the trustworthiness of different data sources. This in turn affects the effort that has to be put into the preprocessing of the data before the actual analysis can take place. Furthermore, the validity signifies the temporal component. While data might have a high veracity, it is possible, that they are too old for the contained information to be still useful for the purpose of certain analysis, possibly even leading to wrong results [26, 28].

Although the given definitions are providing some orientation, there is, despite attempts for clarification [33, 34], no universal definition at which point the characteristics, depicted in Table 1 apply.

TABLE 1. BIG DATA CHARACTERISTICS

CHARACTERISTIC	DESCRIPTION
Volume	Volume represents the (high) amount of data the system is confronted with.
Velocity	Velocity refers to the speed at which the data have to be handled.
Variety	Variety corresponds to the heterogeneity of the data and its sources.
Variability	Variability denominates the variation relating to the other characteristics.
Veracity	Veracity stands for the accuracy and therefore the trustworthiness of the data.
Validity	Validity signifies the task related assessment of the data's actuality.

Because the manifestation of those characteristics can immensely vary, depending on the wanted results and the prevailing conditions, the required solutions and therefore the developed systems are highly individual [34, 35].

B. Software Testing

Since the research questions are targeted on the potential distinctions between common software testing and the testing of big data applications, it is self-evident, that an understanding of both is required. Hence, to emphasize the practice of software testing, the activity itself, the motivations, and approaches are described in more detail. Software testing can be defined as "a process, or a series of processes, designed to make sure computer code does what it was designed to do and, conversely, that it does not do anything unintended" [36]. Although the testing itself is no productive endeavor, it constitutes a crucial, auxiliary activity to ensure the quality of the created software. The oversight of existing issues or the deliberate decision to ignore them can cause severe consequences [37]. As a result, companies worldwide spend about one-quarter of their total IT budget on quality assurance and testing, exemplifying its importance [38].

Generally speaking, software testing comprises the design, the execution and analysis of test cases, and is possibly followed by the reporting of the results [39]. There are several different styles (black box, white box, gray box) that reflect the tester's amount of insights into the system [40]. Comprehensive testing usually occurs on different levels of growing scale (unit, integration, system) [41]. Furthermore, the testing can be conducted with varying intentions and reasons. Examples of those are the initial testing of new contents, ensuring, that the software runs at all (smoke test), making sure, that changes did not negatively affect already functioning components (regression test) and assuring the principal's satisfaction with the finished product (acceptance test) [42]. An overview of those described concepts, that in conjunction allow to categorize a testing endeavor, is given in Fig. 2.

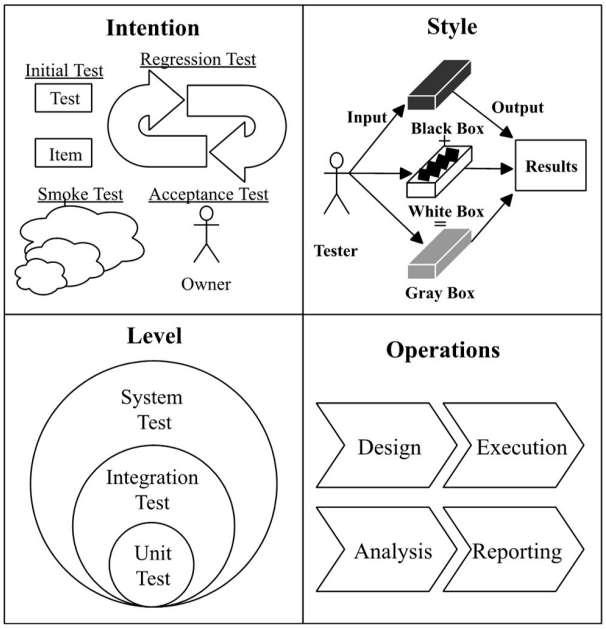

Fig. 2. Concepts in software testing.

III. COMPARISON OF BIG DATA AND SOFTWARE SYSTEMS

Even though big data applications heavily rely on software to fulfill their designated tasks, it would not be sufficient to observe them as common software, including the accompanying testing regimes. Instead, big data testing is a category of its own, which is concerned with hardware as well as software aspects and additional challenges [22, 43].

A. Testing of the Systems

Since the characteristics, depicted in Table 1, cannot be handled by traditional approaches, they require the creation of suitably adapted systems. A certain magnitude of volume, for instance, exceeds the possibilities of a system's vertical growth. This results in the necessity to expand horizontally, leading to a widespread network of servers to collectively handle the given tasks [26]. Additionally, it is prevalent to use commodity hardware for the sake of cost reduction, which in turn increases the system's heterogeneity [44]. This results in additional risks and potential incompatibilities, which have to be covered in the testing process. The distributed nature also introduces reliance on communication between the components for the sake of coordination and task fulfillment. Accordingly, an outage of components is an omnipresent threat that is amplified by the reduced reliability of the used commodity hardware in comparison to highly specialized business solutions [45]. Hence, in the domain of big data, it is necessary to incorporate the examination of the resiliency of a system. While those decentralized structures also allow handling the characteristic of velocity by distributing workloads and tasks across numerous heterogeneous servers, the specifications and therefore performance might vastly differ. Since the usefulness of big data applications often heavily relies on non-functional properties, those have to be tested under consideration of the aforementioned constraints [46]. Another consequence of a huge supply of data lies in the challenge of determining which are to be used and how to do so. While typical software commonly has a well-defined and verifiable behavior, big data applications are commonly intended to generate new information out of existing data. For this reason, there are additional aspects of the data sourcing and the reasoning of the underlying computations that have to be assessed. Hence, while in traditional software, the inputs and desired outputs or reactions are usually known and the congruency is verified, this approach does often not apply for big data applications. In consequence of their nature and purpose, those systems often constitute black boxes without a corresponding test oracle, which exacerbates the detection of logical flaws and therefore poses additional challenges [47]. Besides that task, the consequential variety of data requires pre-processing to harmonize the different formats, structures, and conventions as well as measures to incorporate inputs from a multitude of different sources, which might use completely different interfaces. Furthermore, it is common for data to be incorrect or incomplete [48], necessitating procedures for improving data quality. Thus, to ensure the quality of the obtained results. Those steps of integrating sources and preparing their data have to be checked as well, which results in adding another layer of complexity to the testing process. This leads to new challenges, since there might be a big number of rules and dependencies, whose adherence has to be verified, because errors in this stage can have a huge impact on the quality of the later analysis [49].

Another major difference in comparison to traditional software testing can be caused by the variability in regards to the other characteristics. While the possible valid inputs of software are usually known and modifications are pre-planned, the input-characteristics of big data applications can change over time or caused by events, therefore necessitating a modification of the application itself [26, 32]. Those modifications, in turn, add uncertainty, which has to be reflected by the testing. An additional factor of uncertainty occurs, when (a high number of) external sources are used. Since there is no option to directly control, how the data are provided, manipulations by the source or through malicious attacks on the source are possible. Those encounters might either compromise the data quality (cp. veracity) or even try to invoke damage to the system itself.

In general, while software testing is limited to testing software, testing big data systems has to deal with socio-technical systems, which adds additional aspects and challenges that have to be factored in [22]. Therefore, even though traditional software testing is already a demanding task, testing in big data is even more challenging, because its characteristics and stipulations cause increased complexity and uncertainty. This circumstance is depicted in Fig. 3, representing a summary of the made investigations and therefore an answer on the previously formulated *RQ1*. This results in additional challenges that have to be overcome to realize an effective quality assurance. While most of those challenges originally stem from the aforementioned characteristics, the area of application also often plays a crucial role in complicating the task by adding a certain degree of complexity to the issue. These dimensions, influencing the baseline situation of big data endeavors, lead to specific conditions for the implementation of the aspired solutions. As a result, the realization of the according systems is accompanied by additional challenges, which are not as prevalent in traditional software engineering, but have to be considered and therefore also covered in big data engineering and the according testing.

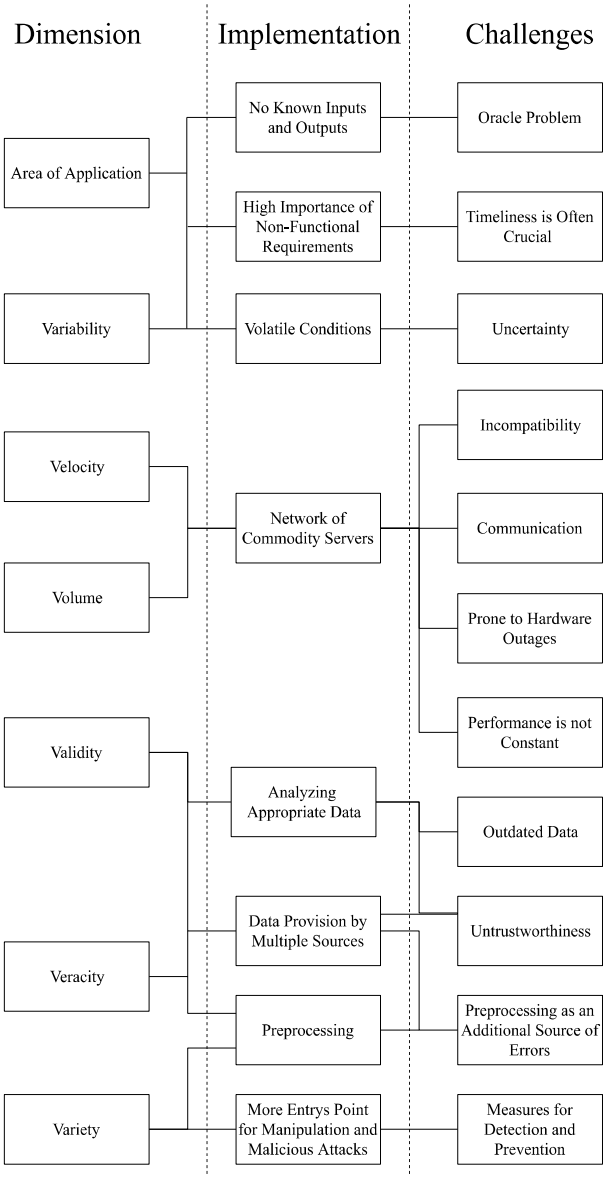

Fig. 3. Mapping of the data characteristics, implementation details, and occurring challenges.

B. Challenges and Requirements of Big Data Testing

Due to the fact that big data applications highly rely on software, many aspects and procedures of traditional software testing also persist in the big data domain. Therefore, the approaches depicted in Fig. 2 are still relevant and can in principle also be extended to the hardware components and the composition of software and hardware. However, derived from the differences that have been highlighted, additional tasks emerge, that have to be performed, to provide comprehensive quality assurance of big data systems. With regards to the content, it might, depending on the use case, be necessary to constantly review the undertaken calculations and the conducted usage of data sources for this purpose. This necessity arises due to the agility of the examined subjects and circumstances, degradation of data sources, changes in the explored questions, as well as the constantly emerging new technologies and insights in the field [26, 50]. Regarding technical aspects, the distributed and heterogeneous nature of the big data landscapes requires extensive tests of the used components, since they are a part of the system under test and might possibly induce errors due to faultiness [37]. Furthermore, the communication between the components, including possible data transformations, has to be tested and the system's reaction to the outage of nodes has to be assessed. Another consequence of the multitude of nodes and especially sources, which might not be under direct control, is the plethora of potential weak points for possible attackers. This might concern attacks on the system itself, but also attempts to manipulate the data and therefore the results of the analysis. For this reason, the security of the system and, depending on the use case, also the ability to detect manipulations have to be tested regarding those circumstances. In general, in the area of big data, it is an ambitious task to create or find ways to validate data. This also applies to the determination or creation of a test oracle. Because of the explorative nature of the applications, the desired outcome is often not known, impeding the according testing.

Since non-functional aspects, like response times, often play a major role, it is also necessary to extensively benchmark the application to assure conformity with those requirements. Especially the timely detection of the need to scale and the ability to do so have to be taken into account, including the needed time and possible capacity limits. When an applications sole purpose is a timely evaluation of data, for example in High-Frequency Trading, a delay of several seconds is not only a nuisance but might effectively render the whole application useless, stressing the importance of this aspect [51].

Another demand, derived from the dynamic and constant change in the application areas is the convertibility and extensibility of the tests to allow the adjustment to changes of the system under test. Since changes in the used algorithms, technologies and data sources are to be expected, the testing solutions should provide an according amount of flexibility. A possible solution for this task might be a highly modular structure, which allows swapping elements according to the prevailing needs [47].

Therefore, when facing big data applications, testers should follow the six guidelines depicted in Fig. 4. Those constitute the additionally needed steps on top of the common software testing procedures to take account of the specificities of those systems and represent an answer on the formulated *RQ2*.

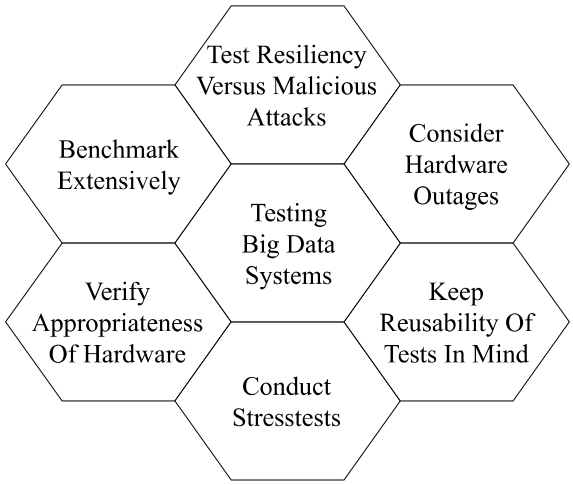

Fig. 4. Developed guidelines for testing big data applications.

It should be noted, however, that the respective relevancy of each of those guidelines depends on the concrete use case. Depending on the circumstances, some of those might not, or only to a lesser extent, be applicable to a testing scenario. For example, an application that has no direct or indirect external connections and only receives internal sensor data by production facilities is not prone to external attacks or manipulation. Therefore, testing the resiliency versus malicious attacks is unessential. On the other hand, when using cloud services, the appropriateness of the hardware can rarely by verified, due to factors like a lack of direct access and knowledge of the used components. Hence, while the guidelines help in understanding the specificities of big data testing, their concrete application and implementation still remains in the big data engineers or testers responsibility. Furthermore, finding an appropriate test oracle in the big data domain remains a challenging and highly individual task that is facilitated neither by common software testing practices nor by the application of the guidelines.

C. Discussion

Because big data systems are heavily relying on software, the concepts of software testing that are depicted in Fig. 2 are also applicable to the testing of big data systems. Nevertheless, while the major intentions for a test scenario are the same, the three possible styles persist and the general operations remain unchanged, the levels slightly differ. A unit, for instance, might comprise a combination of software and hardware, instead of just a piece of software. An example could be a server, which can be considered a unit in big data testing, even though it is not only relying on software or hardware, but on the conjunction of both. Though, since the application areas and characteristics of big data create additional challenges, it is necessary, to consider them in the creation of test scenarios. By applying the guidelines presented in Fig. 4, most of them can be tackled, increasing the value of quality assurance.

However, even considering all of those stipulations, depicted in Fig. 3, and finding an applicable test oracle, does not assure a highly productive and effective big data application. Since it is also necessary to incorporate further aspects, besides the system itself, implementing a holistic quality assurance process, accommodating the socio-technical nature and the complexity of the endeavour is required [22]. Consequently, future research should focus on examining the identified challenges one by one in detail, allowing for deeper insights and possibly new solution approaches. Considering both of the mentioned domains, this can be approached from a software as well as a systems engineering perspective. Hence, in the future the testing of big data systems would not only be observed in terms of the deployed software and the correct way of functioning but also the underlying architecture and its connections to the environment.

IV. CONCLUSION

Due to the characteristics that define big data, properly handling them and effectively generating value is a demanding task. The paper at hand showed that the systems that are used for this purpose exceed the scope of common software since they combine sophisticated software with complex hardware formations. For this reason, traditional software testing is not sufficient and further measures are required. Those are determined based on the additional challenges depicted in Fig. 3 and constitute the six guidelines for taking account of the specificities when testing big data applications in comparison to common software. The conjunction of the specificities with the guidelines also constitutes the answer to the second research question.

The increased clarity concerning the domain and the guidelines will support scientists as well as practitioners in their endeavors to understand, apply and advance the subject of big data, therefore contributing to the formation of tomorrow's system creation and management.

The following step could be the accumulation of best practices and use case independent techniques for each of those guidelines to further support the dissemination of big data analytics. Furthermore, reliable techniques for the definition of test oracles are still to be developed, leaving this as one of the most significant future challenges in the regarded domain.

REFERENCES

[1] N. Khan *et al.,* "Big data: survey, technologies, opportunities, and challenges," *The Scientific World Journal*, vol. 2014, pp. 1–18, 2014.

[2] Y. Wang, L. Kung, W. Y. C. Wang, and C. Cegielski, "Developing a Big Data-Enabled Transformation Model in Healthcare: A Practice Based View," in *Proceedings of Thirty Fifth International Conference on Information Systems*, 2014.

[3] J. Kallinikos and N. Tempini, "Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation," *Information Systems Research*, vol. 25, no. 4, pp. 817–833, 2014.

[4] A. Farseev and T.-S. Chua, "Tweet Can Be Fit," *ACM Transactions on Information Systems*, vol. 35, no. 4, pp. 1–34, 2017.

[5] K. Domdouzis, B. Akhgar, S. Andrews, H. Gibson, and L. Hirsch, "A social media and crowdsourcing data mining system for crime prevention during and post-crisis situations," *Journal of Systems and Information Technology*, vol. 18, no. 4, pp. 364–382, 2016.

[6] D. Wu and Y. Cui, "Disaster early warning and damage assessment analysis using social media data and geo-location information," *Decision Support Systems*, vol. 111, pp. 48–59, 2018.

[7] D. Staegemann, M. Volk, and K. Turowski, "Mobile Procurement Management," in *Springer Reference Wirtschaft, Handbuch Digitale Wirtschaft*, T. Kollmann, Ed., Wiesbaden: Springer Fachmedien Wiesbaden, 2019, pp. 1–15.

[8] T. Nguyen, L. Zhou, V. Spiegler, P. Ieromonachou, and Y. Lin, "Big data analytics in supply chain management: A state-of-the-art literature review," *Computers & Operations Research*, vol. 98, pp. 254–264, 2018.

[9] K. Nagorny, P. Lima-Monteiro, J. Barata, and A. W. Colombo, "Big Data Analysis in Smart Manufacturing: A Review," *International Journal of Communications, Network and System Sciences*, vol. 10, no. 03, pp. 31–58, 2017.

[10] Y. Tang, J. J. Xiong, Y. Luo, and Y.-C. Zhang, "How Do the Global Stock Markets Influence One Another? Evidence from Finance Big Data and Granger Causality Directed Network," *International Journal of Electronic Commerce*, vol. 23, no. 1, pp. 85–109, 2019.

[11] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.
[12] H. Lee, N. Aydin, Y. Choi, S. Lekhavat, and Z. Irani, "A decision support system for vessel speed decision in maritime logistics using weather archive big data," *Computers & Operations Research*, vol. 98, pp. 330–342, 2018.
[13] X. Yi, F. Liu, J. Liu, and H. Jin, "Building a network highway for big data: architecture and challenges," *IEEE Network*, vol. 28, no. 4, pp. 5–13, 2014.
[14] S. Yin and O. Kaynak, "Big Data for Modern Industry: Challenges and Trends [Point of View]," *Proceedings of the IEEE*, vol. 103, no. 2, pp. 143–146, 2015.
[15] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: a systematic literature review and research agenda," *Inf Syst E-Bus Manage*, vol. 16, no. 3, pp. 547–578, 2018.
[16] Z. A. Al-Sai, R. Abdullah, and M. h. husin, "Big Data Impacts and Challenges: A Review," in *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan, Apr. 2019 - Apr. 2019, pp. 150–155.
[17] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.
[18] International Data Cooperation, *IDC Forecasts Revenues for Big Data and Business Analytics Solutions Will Reach $189.1 Billion This Year with Double-Digit Annual Growth Through 2022.* [Online] Available: https://www.idc.com/getdoc.jsp?containerId=prUS44998419. Accessed on: May 21 2019.
[19] M. Turck and D. Obayomi, *The Big Data Landscape.* [Online] Available: http://dfkoz.com/big-data-landscape/. Accessed on: May 21 2019.
[20] O. Müller, M. Fay, and J. Vom Brocke, "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, 2018.
[21] O. Hummel, H. Eichelberger, A. Giloj, D. Werle, and K. Schmid, "A Collection of Software Engineering Challenges for Big Data System Development," in *44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Prague, 2018, pp. 362–369.
[22] D. Staegemann, M. Volk, N. Jamous, and K. Turowski, "Understanding Issues in Big Data Applications - A Multidimensional Endeavor," in *Twenty-fifth Americas Conference on Information Systems*, Cancun, 2019.
[23] V. Garousi and M. V. Mäntylä, "A systematic literature review of literature reviews in software testing," *Information and Software Technology*, vol. 80, pp. 195–216, 2016.
[24] C. Tao and J. Gao, "Quality Assurance for Big Data Application – Issuses, Challenges, and Needs," in *The 28th International Conference on Software Engineering and Knowledge Engineering*, 2016, pp. 375–381.
[25] M. Volk, D. Staegemann, M. Pohl, and K. Turowski, "Challenging Big Data Engineering: Positioning of Current and Future Development," in *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security*, Heraklion, Crete, Greece, 2019, pp. 351–358.
[26] NIST, *NIST Big Data Interoperability Framework: volume 1, definitions, version 2.* [Online] Available: https://bigdatawg.nist.gov/_uploadfiles/NIST.SP.1500-1r1.pdf. Accessed on: Jan. 31 2019.
[27] P. Russom, *Big Data Analytics: TDWI Best Practices Report Fourth Quarter 2011.* [Online] Available: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf. Accessed on: May 22 2019.
[28] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *CODATA*, vol. 14, no. 2, pp. 1–10, 2015.
[29] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 42–47.
[30] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.
[31] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and Information Systems*, vol. 46, no. 2, pp. 241–284, 2016.
[32] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and Good practices," in *Sixth International Conference on Contemporary Computing*, M. Parashar et al., Eds., 2013, pp. 404–409.
[33] D. Laney, "Information Economics, Big Data and the Art of the Possible with Analytics," 2012.
[34] M. Volk, S. W. Hart, S. Bosse, and K. Turowski, "How much is Big Data? A Classification Framework for IT Projects and Technologies," in *Twenty-second Americas Conference on Information Systems*, San Diego, 2016.
[35] A. Amado, P. Cortez, P. Rita, and S. Moro, "Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis," *European Research on Management and Business Economics*, vol. 24, no. 1, pp. 1–7, 2018.
[36] G. J. Myers, T. Badgett, and C. Sandler, *The art of software testing,* 3rd ed. Hoboken, N.J: J. Wiley & Sons, 2011.
[37] R. Patton, *Software testing*. Indianapolis: SAMS, 2001.
[38] Capgemini, Sogeti, HPE, and Micro Focus, *Proportion of budget allocated to quality assurance and testing as a percentage of IT spend from 2012 to 2018.* [Online] Available: https://www.statista.com/statistics/500641/worldwide-qa-budget-allocation-as-percent-it-spend/. Accessed on: May 23 2019.
[39] P. Ammann and J. Offutt, *Introduction to software testing*. Cambridge: Cambridge University Press, 2008.
[40] M. Kaur and R. Singh, "A Review of Software Testing Techniques," *International Journal of Electronic and Electrical Engineering*, vol. 7, no. 5, pp. 463–474, 2014.
[41] L. Copeland, *A Practicioner's Guide to Software Test Desing,* 11th ed. Boston: Artech House Publihsers, 2010.

[42] R. Binder, *Testing object-oriented systems: Models, patterns and tools,* 7th ed. Boston, Madrid: Addison-Wesley, 2006.
[43] M. Abdallah, "Big Data Quality Challenges," in *2019 International Conference on Big Data and Computational Intelligence (ICBDCI)*, Pointe aux Piments, Mauritius, 2019, pp. 1–3.
[44] A. McAfee and E. Brynjolfsson, "Big Data: The Management Revolution: Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision-making culture.," *Harvard Business Review*, vol. 91, no. 5, pp. 1–9, 2012.
[45] G. Wang, L. Zhang, and W. Xu, "What Can We Learn from Four Years of Data Center Hardware Failures?," in *47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks: 26-29 June 2017, Denver, Colorado : proceedings*, Denver, CO, USA, 2017, pp. 25–36.
[46] M. Gudipati, S. Rao, N. Mohan, and N. K. Gajja, "Big Data : Testing Approach to Overcome Quality Challenges," vol. 11, pp. 65–73, 2013.
[47] D. Staegemann, J. Hintsch, and K. Turowski, "Testing in Big Data: An Architecture Pattern for a Development Environment for Innovative, Integrated and Robust Applications," in *Proceedings of the WI2019*, 2019, pp. 279–284.
[48] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality: A Survey," in *2018 IEEE International Congress on Big Data*, 2018, pp. 166–173.
[49] W. Verbeke, C. Bravo, and B. Baesens, *Profit driven business analytics: A practitioner's guide to transforming big data into added value*. Hoboken, New Jersey: John Wiley & Sons, Inc, 2017.
[50] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "Big data. The parable of Google Flu: traps in big data analysis," (eng), *Science (New York, N.Y.)*, vol. 343, no. 6176, pp. 1203–1205, 2014.
[51] B. Fang and P. Zhang, "Big Data in Finance," in *Big Data Concepts, Theories, and Applications*, S. Yu and S. Guo, Eds., Cham: Springer International Publishing, 2016, pp. 391–412.