A Preliminary Overview of the Situation in Big Data Testing

Daniel Staegemann[1][a], Matthias Volk[1][b], Matthias Pohl[1][c], Robert Häusler[1], Abdulrahman Nahhas[1][d], Mohammad Abdallah[2][e] and Klaus Turowski[1]

[1]*Magdeburg Research and Competence Cluster Very Large Business Applications, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany*

[2]*Department of Software Engineering, Al-Zaytoonah University of Jordan, Amman, Jordan*

Keywords: Big Data, Testing, Quality Assurance, Benchmark, Literature Review, Taxonomy.

Abstract: Due to the constantly increasing amount and variety of data produced, big data and the corresponding technologies have become an integral part of daily life, influencing numerous domains and organizations. However, because of its diversity and complexity, the necessary testing of the corresponding applications is a highly challenging task that lacks maturity and is still being explored. While there are numerous publications dealing with this topic, there is no sufficiently comprehensive overview to conflate those isolated pieces of information to a coherent knowledgebase. The publication at hand highlights this grievance by means of an unstructured literature review, proposes a starting point for a corresponding taxonomy to bridge this gap and highlights future avenues for research.

1 INTRODUCTION

Today's society is influenced by an ongoing omnipresence of data that are created, captured, stored and analyzed (Jin et al., 2015). Not only is the amount of those data rapidly rising (Dobre and Xhafa, 2014; Yin and Kaynak, 2015), but also the demand for their fast processing increases (Kolajo et al., 2019). This lead to the establishment of the term big data (Gandomi and Haider, 2015; Diebold, 2012), respectively big data analytics, which address those new challenges and ways of coping with them. To gain new insights, organizations establish and develop their analytics capabilities, aiming to enhance their operational performance by improving their decision making, reducing costs, amending existing assets or services and establishing new ones (Becker, 2016; Vom Brocke et al., 2009b; Wamba et al., 2017). However, those analytic capabilities also have to be tested thoroughly to assure them working correctly (Staegemann et al., 2019b). Yet, to our knowledge, there is no universally followed procedure for doing so, leading to the existence and application of a multitude of methods, approaches and tools (Alexandrov et al., 2013; Ashoke and Haritsa, 2015; Han et al., 2018; Staegemann et al., 2019a), making this particular realm incomprehensible and hampering the achievement of significant advancements. The publication at hand aims at bridging this gap by highlighting the issue and proposing ways to cope with it. To achieve that, in the following sections, the concept of big data is explained in more detail, followed by a consideration of the corresponding approaches regarding the testing and closing with concluding remarks, which comprise an outlook on avenues for future research endeavors.

2 BIG DATA

Big data as a term has no universal and unified definition. Instead, it is rather a concept with a general

[a] https://orcid.org/0000-0001-9957-1003
[b] https://orcid.org/0000-0002-4835-919X
[c] https://orcid.org/0000-0002-6241-7675
[d] https://orcid.org/0000-0002-1019-3569
[e] https://orcid.org/0000-0002-3643-0104

Staegemann, D., Volk, M., Pohl, M., Häusler, R., Nahhas, A., Abdallah, M. and Turowski, K.
A Preliminary Overview of the Situation in Big Data Testing.
DOI: 10.5220/0010475002960302
In *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security (IoTBDS 2021)*, pages 296-302
ISBN: 978-989-758-504-3

idea but various interpretations regarding its delimitation and scope (Gandomi and Haider, 2015), (Volk et al., 2020d). Following a popular definition provided by the National Institute of Standards and Technology (NIST), 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).

This emphasizes on the one hand the distinctiveness based on the data's characteristics and on the other hand, the resulting requirements on the used architectures and technologies (Volk et al., 2020b). Despite those challenges, the incorporation of big data into an organization's operations has proven beneficial (Müller et al., 2018), (Raguseo and Vitari, 2018).

Furthermore, the potential application areas are manifold (Volk et al., 2020c), comprising the likes of healthcare (Safa et al., 2019), transportation management (Fiore et al., 2019), education (Häusler et al., 2020), agriculture (Bronson and Knezevic, 2016), manufacturing (Nagorny et al., 2017), civil protection (Wu and Cui, 2018) and many more.

However, besides the potential gains and the versatility, there is, due to the high complexity (Volk et al., 2020a), also a substantial susceptibility for errors, be it the occurrence of spurious correlations within the analyzed data (Calude and Longo, 2017), biased interpretations (Günther et al., 2017), the amplification of seemingly small errors during the processing (Yang et al., 2018) or the used analytics solutions becoming outdated (Staegemann et al., 2020a). Moreover, also unreliable data sources (Pu et al., 2018), faulty implementations (Staegemann et al., 2019b) and challenges regarding the data transfer within a solution itself (Staegemann et al., 2020d) can negatively impact the achieved results.

Thus, it is highly important to mitigate or preferably even completely prevent those issues, whose consequences could lead to a financial loss in a business scenario or even to deaths in the medical domain (Raghupathi and Raghupathi, 2014).

Table 1: Excerpt of the relevant body of literature.

Paper	Main content	Concepts
(Han et al., 2018)	Review regarding big data benchmarks	Benchmarking
(Alexandrov et al., 2013)	Data generation for testing/benchmarking	Testing/Benchmarking
(Ashoke and Haritsa, 2015)	Simulation of database environments	Testing/Benchmarking
(Staegemann et al., 2019a)	Proposal for modular testing approach	Testing
(Staegemann et al., 2020d)	Test approach for ETL process	ETL testing
(Staegemann et al., 2020c)	Proposal to adopt TDD in big data	TDD
(Casale et al., 2015)	Proposal to adopt MDE in big data	MDE
(Gulzar et al., 2019)	New white-box testing approach	Testing
(Wang et al., 2014)	Big data benchmark suite	Benchmarking
(Staegemann et al., 2019c)	Discussion of challenges in big data testing	General considerations
(Zhang et al., 2017)	Review regarding quality assurance techniques	General considerations
(Truică et al., 2020)	Generic document-oriented benchmark	Benchmarking
(Tao and Gao, 2016)	Overview regarding big data quality assurance	General considerations
(González-Aparicio et al., 2018)	Test transactional services in NoSQL DBs	Testing
(Nagdive et al., 2014)	Overview regarding benchmarking in big data	Benchmarking
(Abidin et al., 2016)	Overview of big data testing techniques	Testing
(Tesfagiorgish and JunYi, 2015)	ETL testing based on data reverse engineering	ETL testing
(Sharma and Attar, 2016)	Framework for validation of ETL process	ETL testing
(Xiong et al., 2013)	Overview regarding benchmarking in big data	Benchmarking
(Li et al., 2015)	Data generation for ETL process verification	ETL testing
(Gudipati et al., 2013)	Outline how to test big data applications	General considerations
(Madhavji et al., 2015)	Challenges of testing big data applications	General considerations
(Thangaraj and Anuradha, 2015)	Overview regarding big data testing	Testing
(Garg et al., 2016)	Challenges of testing and solution approaches	General considerations
(Stepanova et al., 2015)	Ontology-based testing approach	Testing
(Ivanov et al., 2016)	Overview of big data benchmarks	Benchmarking
(Zhang and Xie, 2019)	Proposal to adopt metamorphic testing	Testing
(Liu, 2014)	Process for big data benchmarking	Benchmarking
(Morán et al., 2015)	Testing for MapReduce programs	Testing
(Li et al., 2016)	Combinatorial data generation to test ETL	ETL testing

While they are multifarious and therefore require a plethora of diversified measures, at least some of them can be counteracted with a rigorous testing policy (Staegemann et al., 2019b).

3 TESTING IN BIG DATA

As indicated in the previous section, the general quality assurance of big data applications can pertain to different sectors. Namely, those are the data dimension, the human dimension and the technical dimension (Staegemann et al., 2019b). However, despite the undeniable importance of the first two, the following considerations will only be focused on the testing of the technical implementation and not on data quality or questions regarding the correct way of usage.

To obtain a starting point, without the complexity of other methods (Webster and Watson, 2002), (Levy and Ellis, 2006), an unstructured and exploratory look at the existing scientific literature of the target domain was taken, similar to the approach in (Laursen and Svejvig, 2016). The aim of this procedure was not to get a perfect analysis of the domain, but to create a preliminary overview of the different directions. While this does not allow to derive information regarding the importance of certain approaches or the corresponding development over time, it provides input on which later considerations can be based upon.

One of the first findings was that there are indeed numerous publications that are dealing with the issues of testing bit data regarding its functional or non-functional requirements, respectively considering its testability in general. While the publications depicted in Table 1 are by far not the entirety of the corresponding literature, they already convey an impression of the prevailing situation.

The concepts discussed in the analyzed literature roughly translate to the categories *testing*, *benchmarking, design approaches* and *general considerations*. However, considering the mapping of the contributions, the line between those categories and most of all testing and benchmarking was not always clear (Alexandrov et al., 2013), with some publications dealing with several aspects.

Within those categories, there were also varying approaches, leading to a further segmentation. Exemplarily, approaches for the verification of the extract transformation load (ETL) process, which is highly relevant in big data settings (Nwokeji et al., 2018), could be mentioned here. As additional concepts, the ideas of applying test-driven development (TDD) (Staegemann et al., 2020c) and model-driven engineering (MDE) (Casale et al., 2015) in the big data domain were suggested, constituting ways of preemptive quality assurance via specific design approaches.

Besides the differences regarding the explored aspects, the level of detail and depth of theobservations is also highly diverse. While some publications describe highly specific implementations (Gulzar et al., 2019), (Wang et al., 2014), others focus on more general and broadly applicable considerations (Han et al., 2018), (Staegemann et al., 2019c), (Zhang et al., 2017).

To provide a systematic overview, the creation of a big data application testing taxonomy appears sensible, especially since the literature is lacking in this matter as showcased in (Staegemann et al., 2020b).

This would allow to collate the existing body of literature in a structured way, making it more accessible to those interested in the topic and therefore facilitating the dissemination of knowledge (Raschen, 2005).

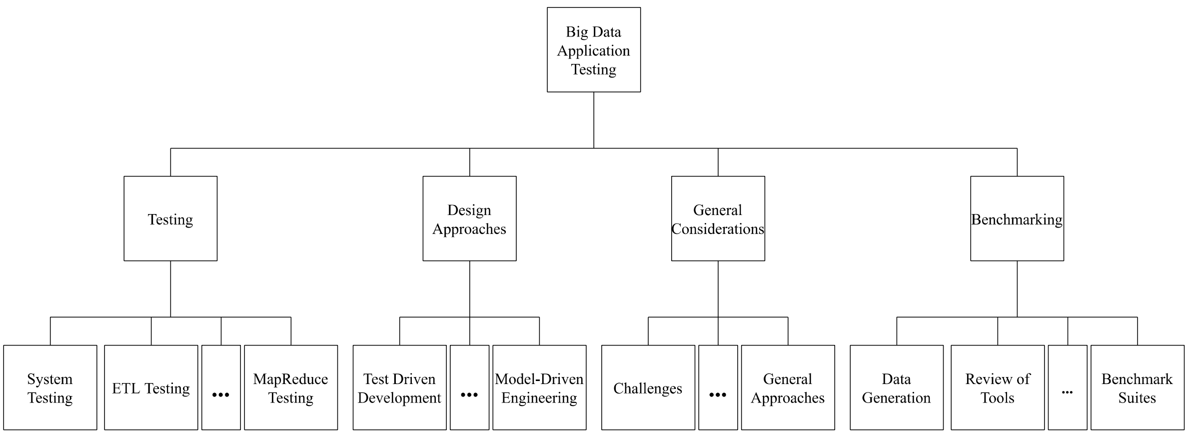

Figure 1: Starting point for a domain specific taxonomy.

An initial starting point for a taxonomy, based on the preliminary unstructured literature review is depicted in Figure 1. However, to actually reflect the complexity of the domain, it requires major revision and refinement based on a comprehensive and structured literature review. Furthermore, a conscientious review would not only allow to derive such a structure, but also to deeply explore the actual contents, facilitating an understanding of the undiscovered (potential) connections and interplays of the challenges, methods, approaches and tools (Vom Brocke et al., 2009a).

4 CONCLUDING REMARKS

As the unstructured literature review, despite its limited representativeness, showed, there is a vast body of knowledge regarding the topic of big data application testing. However, it is diverse and covers a variety of subtopics, making it hard for scientists and practitioners to effectively make use of it. To bridge this gap, a comprehensive overview, for example in the form of a taxonomy, would be useful.

For this purpose, a structured literature review appears to be the method of choice. Apart from establishing a structure, such a review would also allow to condense the findings, determine trends, identify gaps and challenges, but also to derive recommendations and best practices and therefore, to advance the development of the domain as a whole.

REFERENCES

Abidin, A., Lal, D., Garg, N., Deep, V., 2016. Comparative analysis on techniques for big data testing. In 2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds. 2016 International Conference on Information Technology (InCITe): Next-Generation IT Summit on the Theme "Internet of Things: Connect Your Worlds". Noida, 06.10.2016 - 07.10.2016: IEEE, pp. 219–223.

Alexandrov, A., Brücke, C., Markl, V., 2013. Issues in big data testing and benchmarking. In Vivek Narasayya, Neoklis Polyzotis (Eds.)Sixth International Workshop on Testing Database Systems - DBTest '13. The Sixth International Workshop on Testing Database Systems - DBTest '13. New York, 24.06.2013 - 24.06.2013. New York: ACM Press, pp. 1–5.

Ashoke, S., Haritsa, J. R., 2015. CODD: A dataless approach to big data testing. In *Proceedings of the VLDB Endowment* 8 (12), pp. 2008–2011.

Becker, T., 2016. Big Data Usage. In José María Cavanillas, Edward Curry, Wolfgang Wahlster (Eds.) New Horizons for a Data-Driven Economy, vol. 51. Cham: Springer International Publishing, pp. 143–165.

Bronson, K., Knezevic, I., 2016. Big Data in food and agriculture. In *Big Data & Society* 3 (1).

Calude, C. S., Longo, G., 2017. The Deluge of Spurious Correlations in Big Data. In *Found Sci* 22 (3), pp. 595–612.

Casale, G., Ardagna, D., Artac, M., Barbier, F., Di Nitto, E., Henry, A. et al., 2015. DICE: Quality-Driven Development of Data-Intensive Cloud Applications. In 2015 IEEE/ACM 7th International Workshop on Modeling in Software Engineering. 2015 IEEE/ACM 7th International Workshop on Modeling in Software Engineering (MiSE). Florence, Italy, 16.05.2015 - 17.05.2015: IEEE, pp. 78–83.

Diebold, F. X., 2012. On the Origin(s) and Development of the Term 'Big Data'. In *SSRN Journal*.

Dobre, C., Xhafa, F., 2014. Intelligent services for Big Data science. In *Future Generation Computer Systems* 37, pp. 267–281.

Fiore, S., Elia, D., Pires, C. E., Mestre, D. G., Cappiello, C., Vitali, M. et al., 2019. An Integrated Big and Fast Data Analytics Platform for Smart Urban Transportation Management. In *IEEE Access* 7, pp. 117652–117677.

Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. In *International Journal of Information Management* 35 (2), pp. 137–144.

Garg, N., Singla, S., Jangra, S., 2016. Challenges and Techniques for Testing of Big Data. In *Procedia Computer Science* 85, pp. 940–948.

González-Aparicio, M. T., Younas, M., Tuya, J., Casado, R., 2018. Testing of transactional services in NoSQL key-value databases. In *Future Generation Computer Systems* 80, pp. 384–399.

Gudipati, M., Rao, S., Mohan, N. D., Gajja, N. K., 2013. Big Data: Testing Approach to Overcome Quality Challenges. In *Infosys Labs Briefings* 11 (1), pp. 65–73.

Gulzar, M. A., Mardani, S., Musuvathi, M., Kim, M., 2019. White-box testing of big data analytics with complex user-defined functions. In Marlon Dumas, Dietmar Pfahl, Sven Apel, Alessandra Russo (Eds.)Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2019. The 2019 27th ACM Joint Meeting. Tallinn, Estonia, 26.08.2019 - 30.08.2019. New York, USA: ACM Press, pp. 290–301.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., Feldberg, F., 2017. Debating big data: A literature review on realizing value from big data. In *The Journal of Strategic Information Systems* 26 (3), pp. 191–209.

Han, R., John, L. K., Zhan, J., 2018. Benchmarking Big Data Systems: A Review. In *IEEE Trans. Serv. Comput.* 11 (3), pp. 580–597.

Häusler, R., Staegemann, D., Volk, M., Bosse, S., Bekel, C., Turowski, K., 2020. Generating Content-Compliant Training Data in Big Data Education. In Proceedings of the 12th CSEdu. 12th International Conference on Computer Supported Education. Prague, Czech Republic, 02.05.2020 - 04.05.2020: SCITEPRESS - Science and Technology Publications, pp. 104–110.

Ivanov, T., Rabl, T., Poess, M., Queralt, A., Poelman, J., Poggi, N., Buell, J., 2016. Big Data Benchmark Compendium. In Raghunath Nambiar, Meikel Poess (Eds.) Performance Evaluation and Benchmarking: Traditional to Big Data to Internet of Things, vol. 9508. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 135–155.

Jin, X., Wah, B. W., Cheng, X., Wang, Y., 2015. Significance and Challenges of Big Data Research. In *Big Data Research* 2 (2), pp. 59–64.

Kolajo, T., Daramola, O., Adebiyi, A., 2019. Big data stream analysis: a systematic literature review. In *J Big Data* 6.

Laursen, M., Svejvig, P., 2016. Taking stock of project value creation: A structured literature review with future directions for research and practice. In *International Journal of Project Management* 34 (4), pp. 736–747.

Levy, Y., Ellis, T. J., 2006. A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. In *InformingSciJ* 9, pp. 181–212.

Li, N., Escalona, A., Guo, Y., Offutt, J., 2015. A Scalable Big Data Test Framework. In 2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST). 2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST). Graz, Austria, 13.04.2015 - 17.04.2015: IEEE, pp. 1–2.

Li, N., Lei, Y., Khan, H. R., Liu, J., Guo, Y., 2016. Applying combinatorial test data generation to big data applications. In David Lo, Sven Apel, Sarfraz Khurshid (Eds.) Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering. The 31st IEEE/ACM International Conference. Singapore, Singapore, 03.09.2016 - 07.09.2016. New York, USA: ACM Press, pp. 637–647.

Liu, Z., 2014. Research of performance test technology for big data applications. In 2014 IEEE International Conference on Information and Automation (ICIA). 2014 IEEE International Conference on Information and Automation (ICIA). Hailar, Hulun Buir, China, 28.07.2014 - 30.07.2014: IEEE, pp. 53–58.

Madhavji, N. H., Miranskyy, A., Kontogiannis, K., 2015. Big Picture of Big Data Software Engineering: With Example Research Challenges. In 2015 IEEE/ACM 1st International Workshop on Big Data Software Engineering. 2015 IEEE/ACM 1st International Workshop on Big Data Software Engineering (BIGDSE). Florence, Italy, 23.05.2015 - 23.05.2015: IEEE, pp. 11–14.

Morán, J., La Riva, C. d., Tuya, J., 2015. Testing data transformations in MapReduce programs. In Tanja Vos, Sigrid Eldh, Wishnu Prasetya (Eds.)Proceedings of the 6th International Workshop on Automating Test Case Design, Selection and Evaluation - A-TEST 2015. the 6th International Workshop. Bergamo, Italy, 30.08.2015 - 31.08.2015. New York, USA: ACM Press, pp. 20–25.

Müller, O., Fay, M., Vom Brocke, J., 2018. The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. In *Journal of Management Information Systems* 35 (2), pp. 488–509.

Nagdive, A. S., Tugnayat, R. M., Tembhurkar, M. P., 2014. Overview on Performance Testing Approach in Big Data. In *International Journal of Advanced Research in Computer Science* 5 (8), pp. 165–169.

Nagorny, K., Lima-Monteiro, P., Barata, J., Colombo, A. W., 2017. Big Data Analysis in Smart Manufacturing: A Review. In *IJCNS* 10 (03), pp. 31–58.

NIST, 2019. NIST Big Data Interoperability Framework: Volume 1, Definitions, Version 3. Gaithersburg, MD.

Nwokeji, J., Aqlan, F., Anugu, A., Olagunju, A., 2018. Big Data ETL Implementation Approaches: A Systematic Literature Review (P). In Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering. The 30th International Conference on Software Engineering and Knowledge Engineering, 01.07.2018 - 03.07.2018: KSI Research Inc. and Knowledge Systems Institute Graduate School (International Conferences on Software Engineering and Knowledge Engineering), pp. 714–721.

Pu, W., Liu, Y.-F., Yan, J., Liu, H., Luo, Z.-Q., 2018. Optimal estimation of sensor biases for asynchronous multi-sensor data fusion. In *Math. Program.* 170 (1), pp. 357–386.

Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. In *Health information science and systems* 2.

Raguseo, E., Vitari, C., 2018. Investments in Big Data Analytics and Firm Performance: An Empirical Investigation of Direct and Mediating Effects. In *International Journal of Production Research* 56 (15), pp. 5206–5221.

Raschen, B., 2005. A resilient, evolving resource. In *Business Information Review* 22 (3), pp. 199–204.

Safa, B., Zoghlami, N., Abed, M., Tavares, J. M. R. S., 2019. BIG DATA for Healthcare: A Survey. In *IEEE Access* 7, pp. 7397–7408.

Sharma, K., Attar, V., 2016. Generalized Big Data Test Framework for ETL migration. In 2016 International Conference on Computing, Analytics and Security Trends (CAST). 2016 International Conference on Computing, Analytics and Security Trends (CAST). Pune, India, 19.12.2016 - 21.12.2016: IEEE, pp. 528–532.

Staegemann, D., Hintsch, J., Turowski, K., 2019a. 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.

Staegemann, D., Volk, M., Daase, C., Turowski, K., 2020a. Discussing Relations Between Dynamic Business Environments and Big Data Analytics. In *CSIMQ* (23), pp. 58–82.

Staegemann, D., Volk, M., Grube, A., Hintsch, J., Bosse, S., Häusler, R. et al., 2020b. Classifying Big Data Taxonomies: A Systematic Literature Review. In Proceedings of the 5th International Conference on Internet of Things, Big Data and Security. 5th International Conference on Internet of Things, Big Data and Security. Prague, Czech Republic, 07.05.2020 - 09.05.2020: SCITEPRESS - Science and Technology Publications, pp. 267–278.

Staegemann, D., Volk, M., Jamous, N., Turowski, K., 2019b. Understanding Issues in Big Data Applications - A Multidimensional Endeavor. In Proceedings of the Twenty-fifth Americas Conference on Information Systems. Cancun, Mexico, 15.08.2019 - 17.08.2019.

Staegemann, D., Volk, M., Jamous, N., Turowski, K., 2020c. Exploring the Applicability of Test Driven Development in the Big Data Domain. In Proceedings of the ACIS 2020. 31st Australasian Conference on Information Systems. Wellington, New Zealand, 01.12.2020 - 04.12.2020.

Staegemann, D., Volk, M., Jamous, N., Venkatesh, R., Hart, S. W., Bosse, Sascha, Turowski, Klaus, 2020d. Improving the Quality Validation of the ETL Process using Test Automation Automation. In Proceedings of the Twenty-sixth Americas Conference on Information Systems. Salt Lake City, United States, 10.08.2020 - 14.08.2020.

Staegemann, D., Volk, M., Nahhas, A., Abdallah, M., Turowski, K., 2019c. Exploring the Specificities and Challenges of Testing Big Data Systems. In Proceedings of the 15th International Conference on Signal Image Technology & Internet based Systems. Sorrento, 26.11.2019 - 29.11.2019.

Stepanova, T., Pechenkin, A., Lavrova, D., 2015. Ontology-based big data approach to automated penetration testing of large-scale heterogeneous systems. In Oleg Makarevich, Josef Pieprzyk, Ron Poet, Atilla Elçi, Manoj Singh Gaur, Mehmet Orgun et al. (Eds.) Proceedings of the 8th International Conference on Security of Information and Networks - SIN '15. the 8th International Conference. Sochi, Russia, 08.09.2015 - 10.09.2015. New York, New York, USA: ACM Press, pp. 142–149.

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

Tesfagiorgish, D. G., JunYi, L., 2015. Big Data Transformation Testing Based on Data Reverse Engineering. In 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom). 2015 IEEE 12th Intl. Conf. on Ubiquitous Intelligence and Computing, 2015 IEEE 12th Intl. Conf. on Autonomic and Trusted Computing and 2015 IEEE 15th Intl. Conf. on Scalable Computing and Communications and its Associated Workshops (UIC-ATC-ScalCom). Beijing, 10.08.2015 - 14.08.2015: IEEE, pp. 649–652.

Thangaraj, M., Anuradha, S., 2015. State of art in testing for big data. In 2015 IEEE International Conference on Computational Intelligence and Computing Research. 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). Madurai, India, 10.12.2015 - 12.12.2015: IEEE, pp. 1–7.

Truică, C.-O., Apostol, E.-S., Darmont, J., Assent, I., 2020. TextBenDS: a Generic Textual Data Benchmark for Distributed Systems. In *Inf Syst Front*.

Volk, M., Staegemann, D., Bosse, S., Häusler, R., Turowski, K., 2020a. Approaching the (Big) Data Science Engineering Process. In Proceedings of the 5th International Conference on Internet of Things, Big Data and Security. 5th International Conference on Internet of Things, Big Data and Security. Prague, Czech Republic, 07.05.2020 - 09.05.2020: SCITEPRESS - Science and Technology Publications, pp. 428–435.

Volk, M., Staegemann, D., Jamous, N., Pohl, M., Turowski, K., 2020b. Providing Clarity on Big Data Technologies. In *International Journal of Intelligent Information Technologies* 16 (2), pp. 49–73.

Volk, M., Staegemann, D., Trifonova, I., Bosse, S., Turowski, K., 2020c. Identifying Similarities of Big Data Projects–A Use Case Driven Approach. In *IEEE Access* 8, pp. 186599–186619.

Volk, M., Staegemann, D., Turowski, K., 2020d. Big Data. In Tobias Kollmann (Ed.)Handbuch Digitale Wirtschaft, vol. 58. Wiesbaden: Springer Fachmedien Wiesbaden (Springer Reference Wirtschaft), pp. 1–18.

Vom Brocke, J., Simons, A., Niehaves, B., Reimer, K., Plattfaut, R., Cleven, A., 2009a. Reconstructing the Giant. On the Importance of Rigour in Documenting the Literature Search Process. In Proceedings of the ECIS 2009. 17th European Conference on Information Systems. Verona, Italy, 08.06.2009 - 10.06.2009.

Vom Brocke, J., Sonnenberg, C., Simons, A., 2009b. Value-oriented Information Systems Design: The Concept of Potentials Modeling and its Application to Service-oriented Architectures. In *Bus. Inf. Syst. Eng.* 1 (3), pp. 223–233.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-f., Dubey, R., Childe, S. J., 2017. Big Data Analytics and Firm Performance: Effects of Dynamic Capabilities. In *Journal of Business Research* 70, pp. 356–365.

Wang, L., Zhan, J., Luo, C., Zhu, Y., Yang, Q., He, Y. et al., 2014. BigDataBench: A big data benchmark suite from internet services. In 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA). 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA). Orlando, FL, USA, 15.02.2014 - 19.02.2014: IEEE, pp. 488–499.

Webster, J., Watson, R. T., 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. In *MISQ* 26 (2), pp. xiii–xxiii.

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

Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y. et al., 2013. A characterization of big data benchmarks. In 2013 IEEE International Conference on Big Data. 2013 IEEE International Conference on Big Data. Silicon Valley, CA, USA, 06.10.2013 - 09.10.2013: IEEE, pp. 118–125.

Yang, M., Adomavicius, G., Burtch, G., Ren, Y., 2018. Mind the Gap: Accounting for Measurement Error and Misclassification in Variables Generated via Data Mining. In *Information Systems Research* 29 (1), pp. 4–24.

Yin, S., Kaynak, O., 2015. Big Data for Modern Industry: Challenges and Trends [Point of View]. In *Proc. IEEE* 103 (2), pp. 143–146.

Zhang, P., Zhou, X., Li, W., Gao, J., 2017. A Survey on Quality Assurance Techniques for Big Data Applications. In 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). Redwood City, CA, USA, 06.04.2017 - 09.04.2017: IEEE, pp. 313–319.

Zhang, Z., Xie, X., 2019. Towards testing big data analytics software: the essential role of metamorphic testing. In *Biophysical reviews* 11 (1), pp. 123–125.