

Towards the Application of Test Driven Development in Big Data Engineering

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

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

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

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

Abstract—Big data analytics have claimed an important role in today's society. Consequently, ways of improving the design and development of the corresponding applications are highly sought after. One rather current proposition is the application of test driven development (TDD) in the big data domain. The idea behind it is to increase the quality and the flexibility of the developed solutions. However, the application of TDD is often seen as a rather challenging task. Therefore, to increase the accessibility and facilitate its use, it is necessary to provide information that give orientation for interested developers. Hence, the publication at hand focuses on the question which considerations, information, and resources need to be provided to facilitate the widespread utilization of TDD in the big data domain. In doing so, it can be used as the foundation for an inventory of the current state of the related literature but also as a call for action to fill remaining gaps by conducting corresponding research endeavours.

Keywords—big data, big data engineering, test driven development, TDD, software engineering, quality assurance, testing

I. INTRODUCTION

With today's society producing and capturing more and more data [1, 2], its purposeful utilization has become a driving force for many industries as well as the public sector [3]. This phenomenon is usually referred to by the term big data (BD) [4]. Consequently, the development of the corresponding applications along with the improvement of the underlying tools and techniques have gained tremendous importance and interest among practitioners and scientists alike [5, 6]. Consequently, the corresponding market's value is projected to grow from 241 billion U.S. dollars in 2021 to 655 billion U.S. dollars by 2029 [7], which not only pertains to big corporations but also to small and medium-sized enterprises [8]. Yet, to reap the potential benefits [9], besides the existence of issues that can be solved by or the chance for the discovery of opportunities that can be seized through the analysis of data [10], mainly three aspects have to be regarded [11]. (i) It is important to assure the quality of the utilized data [12]. (ii) Further, the willingness of the responsible decision makers to actually make use of the BD systems and to do so in a sincere manner instead of just trying to justify decisions that were already made beforehand is needed [13]. (iii) Finally, the BD analytics applications themselves have to work properly [14]. However, while the actual creation of the necessary means for BD analytics is widely discussed, their testing is often somewhat neglected [6].

Therefore, it seems reasonable to counteract this issue by directing the scientific discourse towards this important topic. This includes the discussion of ways how to improve currently applied approaches but also the exploration of new ones. In the latter category falls the rather recent proposition of applying the approach of test driven development (TDD) to the BD domain [15]. Yet, while it is supposed to bring several benefits, depending on the respective situation and requirements, it is also associated with considerable challenges. However, as of now, to our knowledge, despite a growing number of corresponding publications, there has been no publication that sheds light on the question, what is actually needed to facilitate the widespread implementation of the concept. Therefore, the publication at hand aims to answer the following research question (RQ).

RQ: Which considerations, information, and resources need to be provided to facilitate the widespread use of test driven development in the big data domain?

To provide an answer to the RQ, the publication is structured as follows. After this introduction, an overview of the most important concepts is provided as a foundation. This is succeeded by the exploration of what is needed to make TDD in the BD domain more accessible. Afterwards, the findings are discussed, which is followed by the conclusion.

II. BACKGROUND

In the following sub-sections, the most relevant concepts for the understanding of this paper are briefly outlined.

A. Big Data

Big data is one of the most influential trends of our time [16–18]. Consequently, the corresponding scientific discourse is also intense [6]. Nevertheless, there is still no generally adopted definition for the term itself [4].

However, the widely accepted definition of the National Institute of Standards and Technology (NIST) states that BD "consists of large datasets that primarily exhibit the characteristics of volume, velocity, variety, and/or variability and require a scalable architecture for efficient storage, processing, and analysis" [19].

The first property, volume, refers to the amount of data, in terms of the number and/or size of files, that must be processed by the corresponding applications [20]. Velocity, in turn, refers to two aspects, the speed at which data is received and the timeliness expected for application results [21].

Variety refers to the heterogeneity of the data, expressed in terms of, among other things, the different structuring (structured, semi-structured, unstructured), the use of diverse units of measurement and formats, and the various contexts from which they come [22]. Finally, variability expresses the fact that the characteristics mentioned above, as well as the questions that BD is used to answer and the content of the data, can change over time [23–25].

However, in addition to these four characteristics, there are many more that are referred to in some of the available publications, highlighting the complexity of the domain [18].

B. Test Driven Development

TDD is a development approach that is attributed with the ability to improve the quality (at the cost of reduced speed) of an implementation [26] by increasing the test coverage as well as the developed system's design through decomposition into single components. In doing so, this reduced the overall complexity of the systems while helping to avoid errors and increases maintainability [27, 28].

In the traditional software development approach, new features are first designed, then implemented, and afterwards tested. Yet this order is altered for TDD. The design remains the first step, however, the desired functionality is broken down into small parts that will be implemented over several iterations [29]. Subsequently, the corresponding tests for the part that shall be realized next are written. To ensure that they actually test new aspects, they are then executed and should fail, since the actual implementation is still lacking [30]. If they do not, based on the premise, they have to be revised. After the tests fail, the actual productive coding is performed, which shall yield the desired functionality. Here, the main focus is to create working code, capable to fulfil the intended task. Other aspects, such as its elegance, are insignificant as long as the previously written tests pass [27]. Once this is achieved, the code is refactored to improve aspects like compliance with standards, readability, best practices and code conventions, and overall quality [30]. At this stage, the previously written tests are used to ensure that no errors occur during this process. Besides impacting the test coverage and the length of the test cycles [31], the focus on well tested incremental changes and small tasks [32] also influences the developed solution's design. Usually, unit tests form the backbone of TDD. Yet, other types of tests, such as system, test, or integration tests can be used to complement them [33, 34].

C. TDD in BD

As previously highlighted, the use of TDD is a promising new proposition on how to change the common BD engineering [35], when an especially high quality of the application is required. To this end, the use of microservices has been proposed as a technical foundation [15]. Since an essential element of TDD is to break the desired application into small parts, and microservices enable precisely this, great synergy can be exploited by their utilization [36]. Their use allows each business functionality to be realized as a separate service, which also offers the possibility of independent scaling according to the respective workloads and necessities. Moreover, it also has an impact on the implementation process, as the development of the respective services can be distributed across different teams.

Because of the capsulated nature and the sole provisioning of dedicated interfaces, strict tool requirements for the development can be repealed. This includes, inter alia, the programming environment and used languages. Instead, the developers can use the technology they deem most appropriate for the task at hand. TDD also increases the flexibility in a second context. The tests that are created allow for simpler and less risky changes to the developed application, as these can be immediately validated by the existing tests. This provides faster feedback, helps to avoid newly introduced errors and, consequently, increases the users' confidence, which might, in turn, help the application's acceptance.

III. MAKING THE APPLICATION OF TDD FOR BD MORE ACCESSIBLE

Since the application of TDD in general is often seen as rather challenging, respectively that it necessitates familiarization [37], the same applies to its application in a BD context. Hence, to make it more accessible and facilitate its use, the provisioning of resources that give orientation for interested developers seems reasonable. To structure this, in the following, an overview of necessary or, at least, helpful types of information and insights is given. This, in turn, can be used as the foundation of an inventory of the current state of the related literature but also as a call for action to fill remaining gaps by conducting corresponding research endeavours.

The natural starting point, after already establishing the increase of the quality and flexibility of the developed BD applications as the overall objective, is to provide an overview of the challenges that await the developers of the corresponding applications, independently of the chosen mode of development. Thereby, a foundation for the ensuing considerations is provided, and also awareness is created, why comprehensive testing is required at all. This appears especially meaningful, since the testing of software is reported as less popular compared to its actual development [38]. Therefore, ways to increase the corresponding motivation and, thus, hopefully, also the likelihood of the task being properly conducted and adequately prioritized are highly important.

Once the general challenges that shall be (at least somewhat) overcome through the new approach are illustrated, it is naturally necessary to outline the general idea and concept behind the application of TDD in BD as well as the justification for doing it. This can be seen as the cornerstone of all the ensuing considerations and insights. Here, it shall be described on a rather high level how the approach is realized, which types of tests should be implemented, how potential application scenarios could look like, and which benefits TDD could bring in these.

However, as almost always, there are not only advantages but also challenges and potential issues that come with the use of TDD when building BD applications. This includes technical aspects but also the underlying processes and factors regarding the developers themselves. To provide a balanced display and also to prepare prospective developers and raise their awareness, these also need to be prominently highlighted and discussed. Further, when possible, ways how to deal with the challenges should be presented, acting as an additional aid and easing the endeavour.

In line with this, it is also important to consider under which conditions TDD should be practiced at all when

building a BD application. Because even though the approach brings several advantages, it is by no means a panacea and, depending on the setting, requirements, available resources, and objectives, the drawbacks might also outweigh them. Hence, especially for development teams that are inexperienced in TDD, but generally interested, some support concerning the decision, if it might be beneficial in their case could provide a lot of value. Moreover, besides the advantage for the prospective applicants, this might also prove valuable when it comes to the approach's overall popularization because first time users that do not have a successful experience, might be reluctant to give it another chance. Therefore, preventing them from using TDD in a scenario that might be not well suited for it can be seen as a measure that helps to avoid driving them away.

Once the decision to use TDD was made, the actual development needs to be structured in a comprehensive and easy to follow manner. This applies as well to the big picture as to a more detailed view. For this purpose, process models have proven themselves many times as a valuable tool across varying domains [39], with the (big) data science engineering process (BDSEP) [40] being an example from the BD domain. Therefore, they can also be assumed to be the means of choice for this case. By outlining which actions should be conducted in which situation, developers do not need to focus on this aspect anymore, can avoid potential pitfalls, and can instead concentrate on other aspects of their work, as, for instance, coming up with suitable test cases.

Besides that, a collection of best practices and application guidelines needs to be compiled to amend the structured process steps with insights in how to properly conduct them and which aspects should be especially being paid attention to. This can be with respect to technical details or ways of implementing something, but also regarding managerial decisions, the structuring of the teamwork, or all other facets that might have considerable impact on the final process and its results. By providing prospective developers this information, which is based on the experiences from actual applications, the likelihood of failure as well as the risk that costly changes have to be done at a later point of the implementation can be reduced. Further, the developers' time is at least partly freed up for conducting the actual work instead of being used to figure out a feasible course of action.

Moreover, since BD applications usually combine various capabilities (e.g., data pre-processing, data transfer, storing, analysis, visualization), it is also necessary to provide orientation how these can generally be tested. While each specific implementation might bring its own challenges, they will most likely at least somewhat resemble others from the same category so that a corresponding overview can be a valuable starting point for further considerations. Based on this, prospective developers can then get inspiration how to address the testing of their own creations and do the corresponding focused research.

However, besides those rather structured recommendations and guidelines, it can also be useful to have more comprehensive practical insights in the form of actual use cases and implementation studies such as, for instance, in [41] and [42]. While these reports might not always outline every step in detail, they provide another perspective and are proven examples of what is possible and practical while potentially also highlighting issues, challenges, and edge-cases that might have been overseen in purely theoretical

considerations. Thereby, they constitute a valuable complement to the other materials to further guide but also inspire prospective applicants of the TDD approach. Yet, the real benefits only emerge with a large number of diverse cases. Hence, over time, this category should most likely turn into the most extensive part of the material collection.

An overview of the discussed materials that help prospective developers in realizing their own TDD endeavours in the BD domain is given in Fig. 1. There, for each segment the leading question that it contains the answer for is given, providing quick and easy orientation.

However, while this division into distinct items appears to provide a good balance between comprehensiveness and focus on the specific aspects, other structures that differ from this proposition (e.g., more comprehensive items that comprise several points at once) might be feasible as well as long as the general content is covered in a clear way.

IV. CONCLUSION

With today's society being heavily reliant on data and their processing, BD application play an important role in many domains. Therefore, the exploration and advancement of their creation is very relevant and, consequently, also actively discussed by scientists and practitioners alike. However, their testing is oftentimes somewhat underrepresented in those considerations. Yet, it is still a crucial part of the development process.

One rather recent proposition was the application of TDD in the BD domain, which was explored more in-depth in this publication. Hereby, the focus was set on the question how it can be made more accessible to facilitate its use. For this purpose, a list of resources was proposed to support prospective applicants in their own corresponding endeavours. Further, the distinct items were explained and their value in the given context highlighted. Hereby, attention is directed towards the important but not always accordingly treated topic of testing in the course of BD engineering. This applies especially to the TDD approach, potentially sparking further interest and leading to more research in the future.

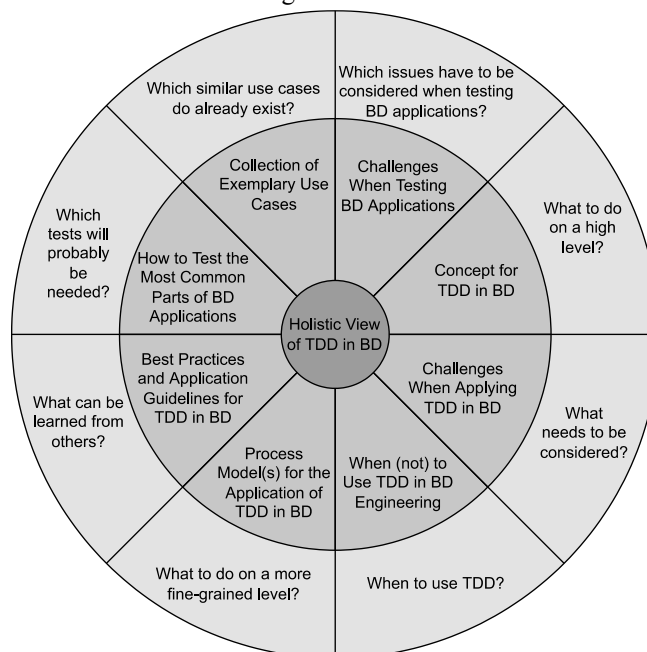


Fig. 1. Materials to facilitate the application of TDD in BD engineering

Beyond that, this paper's contribution is twofold. On one hand, it can be used as the foundation for an inventory of the current state of the related literature and on the other hand as a continuous call for action, to conduct corresponding research endeavours as long as there are still remaining gaps to be filled.

REFERENCES

- [1] C. Dobre and F. Xhafa, "Intelligent services for Big Data science," *Future Generation Computer Systems*, vol. 37, pp. 267–281, 2014, doi: 10.1016/j.future.2013.07.014.
- [2] S. Yin and O. Kaynak, "Big Data for Modern Industry: Challenges and Trends [Point of View]," *Proc. IEEE*, vol. 103, no. 2, pp. 143–146, 2015, doi: 10.1109/JPROC.2015.2388958.
- [3] M. Lee et al., "How to Respond to the Fourth Industrial Revolution, or the Second Information Technology Revolution? Dynamic New Combinations between Technology, Market, and Society through Open Innovation," *JOLtmC*, vol. 4, no. 3, p. 21, 2018, doi: 10.3390/joitmc4030021.
- [4] M. Volk, D. Staegemann, and K. Turowski, "Providing Clarity on Big Data: Discussing Its Definition and the Most Relevant Data Characteristics," in *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Valtta, Malta, 2022, pp. 141–148.
- [5] C. Acciarini, F. Cappa, P. Boccardelli, and R. Oriani, "How can organizations leverage big data to innovate their business models? A systematic literature review," *Technovation*, vol. 123, p. 102713, 2023, doi: 10.1016/j.technovation.2023.102713.
- [6] D. Staegemann, M. Volk, A. Nahhas, M. Abdallah, and K. Turowski, "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, 2019.
- [7] P. Taylor, Big data - statistics & facts. [Online]. Available: <https://www.statista.com/topics/1464/big-data/#topicOverview> (accessed: Jun. 5 2023).
- [8] P. Maroufkhani, M. Iranmanesh, and M. Ghobakhloo, "Determinants of big data analytics adoption in small and medium-sized enterprises (SMEs)," *IMDS*, vol. 123, no. 1, pp. 278–301, 2023, doi: 10.1108/IMDS-11-2021-0695.
- [9] 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, doi: 10.1080/07421222.2018.1451955.
- [10] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, "Big data monetization throughout Big Data Value Chain: a comprehensive review," *J Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-019-0281-5.
- [11] D. Staegemann, M. Volk, N. Jamous, and K. Turowski, "Understanding Issues in Big Data Applications - A Multidimensional Endeavor," in *Proceedings of the Twenty-fifth Americas Conference on Information Systems*, Cancun, Mexico, 2019.
- [12] M. Abdallah, A. Hammad, and W. AlZyadat, "Towards a Data Collection Quality Model for Big Data Applications," in *Lecture Notes in Business Information Processing, Business Information Systems Workshops*, W. Abramowicz, S. Auer, and M. Stróžyna, Eds., Cham: Springer International Publishing, 2022, pp. 103–108.
- [13] W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *The Journal of Strategic Information Systems*, vol. 26, no. 3, pp. 191–209, 2017, doi: 10.1016/j.jsis.2017.07.003.
- [14] D. Staegemann et al., "A Preliminary Overview of the Situation in Big Data Testing," in *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security*, Prague/Virtual Event, 2021, pp. 296–302.
- [15] D. Staegemann, M. Volk, N. Jamous, and K. Turowski, "Exploring the Applicability of Test Driven Development in the Big Data Domain," in *Proceedings of the 31st Australasian Conference on Information Systems (ACIS)*, Wellington, New Zealand, 2020.
- [16] G. Lampropoulos, "Artificial Intelligence, Big Data, and Machine Learning in Industry 4.0," in *Encyclopedia of Data Science and Machine Learning*, J. Wang, Ed.: IGI Global, 2022, pp. 2101–2109.
- [17] M. Ghasemaghahi and G. Calic, "Assessing the impact of big data on firm innovation performance: Big data is not always better data," *Journal of Business Research*, vol. 108, no. 2, pp. 147–162, 2020, doi: 10.1016/j.jbusres.2019.09.062.
- [18] M. Volk, D. Staegemann, and K. Turowski, "Big Data," in *Springer Reference Wirtschaft, Handbuch Digitale Wirtschaft*, T. Kollmann, Ed., Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 1–18.
- [19] W. L. Chang and N. Grady, "NIST Big Data Interoperability Framework: Volume 1, Definitions," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication (NIST SP), 2019. Accessed: Nov. 29 2021.
- [20] P. Russom, Big Data Analytics: TDWI Best Practices Report Fourth Quarter 2011.
- [21] 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, doi: 10.1016/j.ijinfomgt.2014.10.007.
- [22] A. Gani, A. Siddiqi, 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, doi: 10.1007/s10115-015-0830-y.
- [23] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and Good practices," in *Sixth International Conference on Contemporary Computing*, Noida, India, 2013, pp. 404–409.
- [24] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, 2014, doi: 10.1109/TKDE.2013.109.
- [25] D. Staegemann, M. Volk, C. Daase, and K. Turowski, "Discussing Relations Between Dynamic Business Environments and Big Data Analytics," *CSIMQ*, no. 23, pp. 58–82, 2020, doi: 10.7250/csimq.2020-23.05.
- [26] D. Staegemann, M. Volk, E. Lautenschlager, M. Pohl, M. Abdallah, and K. Turowski, "Applying Test Driven Development in the Big Data Domain – Lessons From the Literature," in *2021 International Conference on Information Technology (ICIT)*, Amman, Jordan, 2021, pp. 511–516.
- [27] L. Crispin, "Driving Software Quality: How Test-Driven Development Impacts Software Quality," *IEEE Softw.*, vol. 23, no. 6, pp. 70–71, 2006, doi: 10.1109/MS.2006.157.
- [28] F. Shull, G. Melnik, B. Turhan, L. Layman, M. Diep, and H. Erdogmus, "What Do We Know about Test-Driven Development?," *IEEE Softw.*, vol. 27, no. 6, pp. 16–19, 2010, doi: 10.1109/MS.2010.152.
- [29] D. Fucci, H. Erdogmus, B. Turhan, M. Oivo, and N. Juristo, "A Dissection of the Test-Driven Development Process: Does It Really Matter to Test-First or to Test-Last?," *IEEE Trans. Software Eng.*, vol. 43, no. 7, pp. 597–614, 2017, doi: 10.1109/tse.2016.2616877.
- [30] K. Beck, *Test-Driven Development: By Example*, 20th ed. Boston: Addison-Wesley, 2015.
- [31] D. Janzen and H. Saiedian, "Test-driven development concepts, taxonomy, and future direction," *Computer*, vol. 38, no. 9, pp. 43–50, 2005, doi: 10.1109/MC.2005.314.
- [32] L. Williams, E. M. Maximilien, and M. Vouk, "Test-driven development as a defect-reduction practice," in *Proceedings of the 14th ISSRE*, Denver, Colorado, USA, 2003, pp. 34–45.
- [33] C. Daase, D. Staegemann, M. Volk, and K. Turowski, "Creation of a Framework and a Corresponding Tool Enabling the Test-Driven Development of Microservices," *JSW*, vol. 18, no. 2, pp. 55–69, 2023, doi: 10.17706/jsw.18.2.55-69.
- [34] R. S. Sangwan and P. A. Laplante, "Test-Driven Development in Large Projects," *IT Prof.*, vol. 8, no. 5, pp. 25–29, 2006, doi: 10.1109/MITP.2006.122.
- [35] 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.
- [36] A. Shakir, D. Staegemann, M. Volk, N. Jamous, and K. Turowski, "Towards a Concept for Building a Big Data Architecture with Microservices," in *Proceedings of the 24th International Conference on Business Information Systems*, Hannover, Germany/virtual, 2021, pp. 83–94.
- [37] D. Staegemann et al., "A Literature Review on the Challenges of Applying Test-Driven Development in Software Engineering," *CSIMQ*, no. 31, pp. 18–28, 2022, doi: 10.7250/csimq.2022-31.02.

- [38] R. Florea and V. Stray, "The skills that employers look for in software testers," *Software Qual J*, vol. 27, no. 4, pp. 1449–1479, 2019, doi: 10.1007/s11219-019-09462-5.
- [39] D. C. Wynn and P. J. Clarkson, "Process models in design and development," *Res Eng Design*, vol. 29, no. 2, pp. 161–202, 2018, doi: 10.1007/s00163-017-0262-7.
- [40] M. Volk, D. Staegemann, S. Bosse, R. Häusler, and K. Turowski, "Approaching the (Big) Data Science Engineering Process," in *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security*, Prague, Czech Republic, 2020, pp. 428–435.
- [41] D. Staegemann, M. Volk, M. Perera, and K. Turowski, "Exploring the Test Driven Development of a Fraud Detection Application using the Google Cloud Platform," in *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Valletta, Malta, 2022, pp. 83–94.
- [42] D. Staegemann et al., "Implementing Test Driven Development in the Big Data Domain: A Movie Recommendation System as an Exemplary Case," in *Proceedings of the 7th International Conference on Internet of Things, Big Data and Security, Online Streaming/Prague*, 2022, pp. 239–248.