Overcoming the Complexity of Quality Assurance for Big Data Systems: An Examination of Testing Methods

Christian Daase[a], Daniel Staegemann[b] and Klaus Turowski[c]
Institute of Technical and Business Information Systems, Otto-von-Guericke University, Magdeburg, Germany

Keywords: Big Data Analytics, Software Testing, Quality Assurance, Systematic Literature Review.

Abstract: As the complexity and diversity of big data systems reaches a new level, testing the solutions developed is becoming increasingly difficult. In this study, a systematic literature review is conducted on the role of testing and related quality assurance techniques in current big data systems in terms of applied strategies and design guidelines. After briefly introducing the necessary knowledge about big data in general, the methodology is explained in a detailed and reproducible manner, including the reasoned division of the main question into two concise research questions. The results show that methods such as individual experiments, standardized benchmarking, case studies and preparatory surveys are among the preferred approaches, but also have some drawbacks that need to be considered. In conclusion, testing alone may not guarantee a perfectly operating system, but can serve to minimize malfunctions to a limited number of special cases by revealing its principal weaknesses.

1 INTRODUCTION

The technological advancements of our increasingly digitized world do not only influence our daily lives, but also pose new challenges to those who contribute to this rapid development. In contrast to former times, when data were usually gathered in small quantities within an experimental context, today's data acquisition is characterized by high volumes, velocities, and varieties (Dremel et al. 2020). Reasons for this development are new possibilities to analyze documents, images, geospatial and three-dimensional data (Lonetti and Marchetti 2018; Saheb and Izadi 2019), both structured and unstructured (Abidin et al. 2016). These can be collected first-hand with modern technologies such as sensors and camera systems (Hamilton and Sodeman 2020), by taking advantage of data people voluntarily share via social media sites (Abidin et al. 2016), or through their buying behavior in e-commerce systems (Mithas et al. 2013). Regarding the characteristics, these masses of data are generally referred to as *big data,* or *big data analytics* (BDA) if the processing is meant. The arguments found in the literature for calling BDA a *"game changer"* (Wamba et al. 2017), *"driving force for innovation"* (Chen et al. 2018), or *"next big thing for innovation"* (Côrte-Real et al. 2017) are manifold and to a certain degree interconnected. Along with the goal of allowing for predictive analytics (Hamilton and Sodeman 2020), companies try to use BDA to uncover hidden patterns or interrelationships to support their decision making (Abidin et al. 2016) and the optimization of their organizational activities (Staegemann et al. 2020). However, the actual value of BDA systems may depend on the industry in which they operate. In a study by Müller et al. (2018), the authors found that although the productivity of a company can increase by an average of 4 percent, this only holds true for IT-intensive or highly competitive industries. BDA may also be employed to answer research questions in the fields of medicine and healthcare (Hamilton and Sodeman 2020; Mithas et al. 2013; Saheb and Izadi 2019), crime data analysis (Jain and Bhatnagar 2016), human resources management (Hamilton and Sodeman 2020) and supply chain or delivery route optimization (Côrte-Real et al. 2019). However, to benefit from BDA, it must be considered which requirements the

[a] https://orcid.org/0000-0003-4662-7055
[b] https://orcid.org/0000-0001-9957-1003
[c] https://orcid.org/0000-0002-4388-8914

Daase, C., Staegemann, D. and Turowski, K.
Overcoming the Complexity of Quality Assurance for Big Data Systems: An Examination of Testing Methods.
DOI: 10.5220/0012742100003705

In *Proceedings of the 9th International Conference on Internet of Things, Big Data and Security (IoTBDS 2024)*, pages 358-369
ISBN: 978-989-758-699-6; ISSN: 2184-4976

respective scenario places upon the system. Operating environments can vary in terms of data sources and structures, velocity of data acquisition, and speed of their obsolescence due to changing external influences (Staegemann et al. 2020). Software engineers must therefore consider necessary adjustments in the development and life cycle of their solutions. As usual in software engineering, a developed solution only delivers its maximum value if it works as intended. Since BDA is no exception to this statement, its utilization can also lead to severe failures and malicious effects if the system was not sufficiently tested (Abidin et al. 2016). While debugging is generally a huge unknown in terms of development time, as it adds about 100 to 200 percent of the amount of productive coding time to the total development time (Martin 2007), BDA additionally poses new challenges depending on the specific requirements, from processing capacities to adapted visualization techniques, and privacy issues due to legal restrictions (Jain and Bhatnagar 2016). The significance of these issues becomes even more apparent when considering that more than half of the BDA initiatives developed fail to achieve their strategic goal (Côrte-Real et al. 2019), either because of the underestimated complexity and diversity of the technologies involved (Lonetti and Marchetti 2018; Ordonez and Bellatreche 2020) or the unpredictability of the outcomes (Mithas et al. 2013). A major shortcoming in the testing of BDA solutions is that realistic, high-volume data sets are difficult to simulate with the same velocity as they would occur in a real-world scenario (Lonetti and Marchetti 2018).

To bridge the research gap on how BDA solutions can be tested to avoid the mentioned pitfalls, this paper conducts a systematic literature review (SLR), in which theoretical contributions as well as practical developments are examined. The focus can be consolidated into the following central question:

What is the role of testing in current BDA developments in terms of quality assurance (QA) of the results and sophistication of test strategies?

Necessary background knowledge on big data and software testing is given in the subsequent section. The research questions are formalized in section 3.1, in which the central question is divided into two questions. Furthermore, a detailed research justification is given. The results of the SLR are presented and discussed in the fourth section with separate subsections for the quantitative and qualitative results. Finally, in the fifth section a conclusion on the role of testing in BDA systems and a short outlook on possible future research are given.

2 BACKGROUND

In this section, necessary prior knowledge on the two areas of big data and software testing in general is outlined.

2.1 Big Data

In general, big data is a rapidly evolving topic whose definition changes frequently, and so do the requirements regarding the areas of application. While a few years ago mainly the plain volume of data was of importance for practitioners and academics (Staegemann et al. 2020), today's perception of big data is much broader in scope. Although its characteristics nowadays seem to be inconsistent, recurring terms can be observed in current literature, referred to as the five Vs, meaning *volume*, *velocity*, *variety*, *veracity* and *value* (Jain and Bhatnagar 2016; Wamba et al. 2017). Most of the found literature adopts a subset of these terms in order to describe the meaning of big data. While the volume is undeniably an integral component for the understanding of big data, it could be found that for example the veracity is sometimes neglected (Lonetti and Marchetti 2018). Other publications refrain from considering the value as an autonomous characteristic of big data (Abidin et al. 2016; Ordonez and Bellatreche 2020). With these considerations taken into account, the publication at hand incorporates the definition of big data with only the first four Vs (i.e., without value) in the following investigation.

In a broader sense, BDA can be considered as a "*socio-technical phenomenon*" (Dremel et al. 2020) not only consisting of the technological aspects, but also the intentions of the analysis themselves and the expectations and efforts towards further data-driven developments. Therefore, it is no surprise that researchers try to connect other recent topics of interest with the BDA domain, for example advanced technologies of artificial intelligence (Saheb and Izadi 2019), cloud computing (Daase et al. 2023; Ordonez and Bellatreche 2020) or the internet of things (Lonetti and Marchetti 2018). The synthesis of the understanding for the already high complexity of these domains with the assumption that BDA is only supported by an organization if its value can be clearly demonstrated (Mithas et al. 2013) reveals the necessity of finding ways to ease the testing and with it the presentation of the value of the BDA solution.

2.2 Software Testing

The fundamental role of testing in a software development cycle and its impact on the quality of the outcome of a project are well known among software engineers, but the majority of developers either refrain from writing an adequate amount of tests or overestimate their efforts in terms of reliability and code coverage (Palomba et al. 2016). Furthermore, apart from human error in testing, also the technological aspects pose problems regarding the test quality. There is a trend towards automating the process of generating and processing test data sets (Palomba et al. 2016), leading to new paradigms of how a development cycle might be designed, notably test-driven development (TDD). The expectation is that the permanent execution of tests that are usually written before the actual program code, can significantly reduce the effort for debugging, since each newly occurring error must have just been added (Bissi et al. 2016; Martin 2007). A study conducted by Bissi et al. (2016) found that this focus on testing through TDD improved the internal software quality in about 76 percent and the external software quality in about 88 percent of the publications examined. However, about 44 percent reported a decrease in productivity.

3 RESEARCH JUSTIFICATION

Both topics individually show an increasing interest in the literature, as a quick screening in the scientific database Scopus, which claims to be the largest abstract and citation database (Kitchenham and Charters 2007), indicates. A search carried out in February 2024 for the term *"big data"* in article titles yielded only 6 publications in 2010 with a steadily growing number rising to 6602 articles found for 2021 and a slight decrease afterwards. A search for *testing*, this time restricted to the subject of computer science, also resulted in 1819 articles for 2010 and a peak of 2706 for 2023. Since it could be argued that this is due to an increasing number of scientific publications in general, a more reliable indicator of the growing interest in both topics in combination could be the percentage of articles on big data solutions that take testing into account. Therefore, two other queries in Scopus were used for searching the abstracts of all publications:

(A): *"big data analytic*" OR "big data solution*" OR "big data system*"*

(B): (A) in brackets and in addition *AND test**

As further explained in the review protocol in section 3.2, this query also corresponds to the search phrase of the later review part carried out in Scopus. The asterisks were added to include the plural forms and slight variations. Figure 1 and Figure 2 depict the relationship between articles about BDA and articles with an additional mention of testing. How and to what extent testing is ultimately handled by these articles is the subject of the following main sections. On the left-hand side, the absolute numbers of publications on BDA and similar (i.e., query A) and with the additional requirement of mentioning test strategies (i.e., query B) are juxtaposed. In 2020, the events of the COVID-19 pandemic led to the cancellation or postponement of several conferences (Agarwal and Sunitha 2020), which is one reason for the sudden decrease in newly published articles. In Figure 2, the percentage share of B in A is graphically displayed, indicating that testing is an emerging issue within BDA, with a quota of more than 12 percent in 2023 at the time of this study. A search for *"big data" AND test* AND "literature review"* in the titles of all articles yielded no results, suggesting that no thorough systematic literature review on the role of testing in BDA has been conducted as of today. However, due to the rapid developments in information science, searches in digital libraries are difficult to replicate after a short period of time (Kitchenham and Charters 2007).

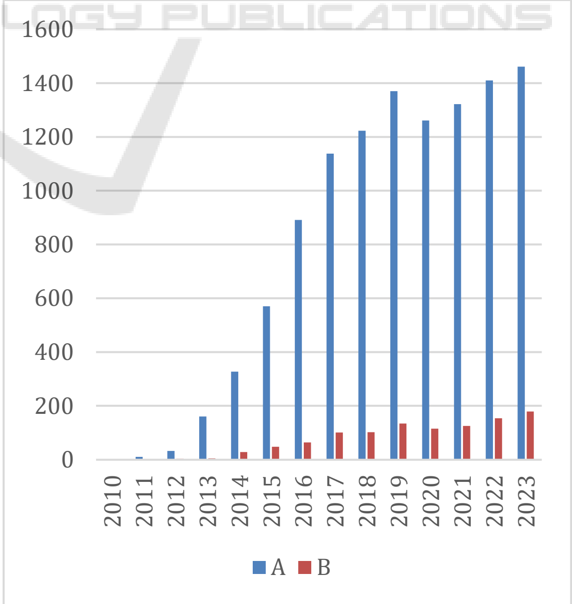

Figure 1: Publications on BDA with (A) and without (B) testing considered.

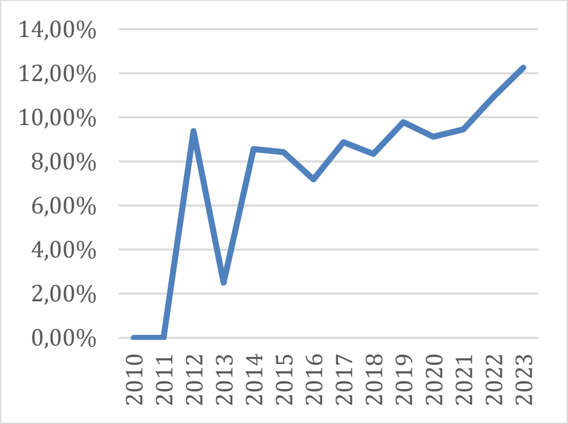

Figure 2: Percentage of B in A.

3.1 Research Questions

The specification of the RQs clarifies the perspective with which the publications were evaluated and the context in which the results are synthesized (Kitchenham and Charters 2007). In addition, this contributes to concluding the purpose of the literature review in concise sentences (Okoli 2015). Moreover, the RQs as a whole determine the composition of the query that is sent to the database to be searched. Kitchenham and Charters (2007) suggest that in essence the query should be composed of three elements: firstly, the *population*, which in this case refers to the application areas of big data, secondly, the *intervention*, which is testing among different variations of this term and thirdly, *outcomes*. The search term used in this paper does not contain a segment for outcomes because the role of testing itself is to be explored. Possible influences on productivity, reliability or software quality are not analyzed in isolation, rather the occurrence of such terms themselves is investigated. In order to keep the RQs summarized in Table 1 short and concise, all questions refer to the examined publications found with query B from the previous section.

Table 1: Research questions focused during the literature review.

ID	Question
RQ1	Which quality assurance methods are primarily used before and during the development time and as how reliable can their respective results be classified?
RQ2	How should tests be designed according to guidelines and what issues could arise?

3.2 Review Protocol

In order to provide a sufficient publication basis for this literature review, Scopus is chosen. In contrast to a full-text database, Scopus offers only abstracts and bibliographic data of the articles presented, but in return combines results from different actual full-text databases such as IEEE Xplore, ACM Digital Library and ScienceDirect (Staegemann et al. 2020).

The search query used is identical to the one used in section 3 for the derivation of the research interest between the topics of big data systems and testing. Since this query is the result of multiple test searches in order to provide an appropriate foundation of relevant publications, this paragraph serves the purpose to formalize the search phrase in accordance with the guidelines of Kitchenham and Charters (2007). As stated in the previous subsection, the search query consists of the *population* (i.e., different terms related to BDA) and the *intervention* (i.e., testing and related terms). The elements of each segment are combined with an *OR* while the segments themselves are combined with the *AND* operator. Table 2 lists the different terms used for this query. In the literature, *system*, *solution*, or *analytics* in general have been proved to be frequent synonyms for this topic. Moreover, with the asterisks phrases like *analytical* or *systematic* are included. *Test** on the other hand includes terms such as *testing, tested* or *test-driven*. The query is searched for in the abstracts and titles of the articles. The keywords are not considered because Scopus does not distinguish between author keywords and indexed keywords in the fields to be searched. Test searches have shown that these automatically added indexed keywords do not always correspond exactly to the topic of the paper. Furthermore, it is assumed that important keywords are also present in the abstract.

Table 2: Composition of the search query.

Population	Intervention
Big data analytic*	Test*
Big data system*	
Big data solution*	

After defining the query, the search must be refined by applying certain inclusion and exclusion criteria as an integral part of any systematic review process (Kitchenham and Charters 2007). Furthermore, these criteria are necessary to increase the density of relevant articles in the final publication base and to reduce the large number of articles to a practically manageable collection (Okoli 2015). Since the application of some criteria requires different levels of insight, Table 3 lists the SLR stages used in this paper.

Table 3: SLR stages and related RQs.

Stage ID	Description
0	Before the SLR
1	Reading title and abstract
2	Reading introduction and conclusion
3	Reading the full text in-depth

For the definition of the inclusion and exclusion criteria, the recommendations of Okoli (2015) are considered. Table 4 shows the criteria and the corresponding SLR stages in which they are applied. For example, the publication language, year, and source type are already considered using the built-in mechanics of Scopus. Contrary to section 3, where a marginal number of articles before 2012 could be found in Scopus for BDA related terms alone, the addition of the intervention segment caused that no articles could be found for this period. The lack of publications for the time before the early 2010s is consistent with the findings of other researchers that the general term *big data* did not become a buzzword before the year 2012 (Che et al. 2013; Ghandour 2015; Hong et al. 2020). The selection process in section 4 starts with the number of publications found after adapting the criteria indicated with (0), which can be automatically applied through the built-in mechanisms the database.

Table 4: Inclusion/exclusion criteria and corresponding SLR stages.

Inclusion	Exclusion
(0) Written in English	(1) Duplicates/proceedings introduction
(0) Published between 2012 and 2023	(1) Only review/no own contribution
(0) Source type is journal or conference proceeding	(1) Testing in non-BDA context/of no concrete solution
(0) Finalized publication	(2) No conclusion on influence of testing
(1) Big data related main subject	(3) No explanation of testing strategies
(2) Thoughtful considerations for testing	(3) No careful test execution (projects only)

4 RESULTS AND DISCUSSION

The search carried out in Scopus led to the search process illustrated in Figure 3. With an initial number of 870 articles, the publication base decreased significantly, as different topics used the term *test* in other contexts than a technical big data solution. In the health sector in particular, research focused on patient and drug testing drove up the number of articles that apparently use and evaluate BDA. Furthermore, about 10 percent of the articles found were introductions to proceedings or workshops, duplicates or had abstracts indicating that the article does not contain a sufficient contribution to the topic covered in this study.

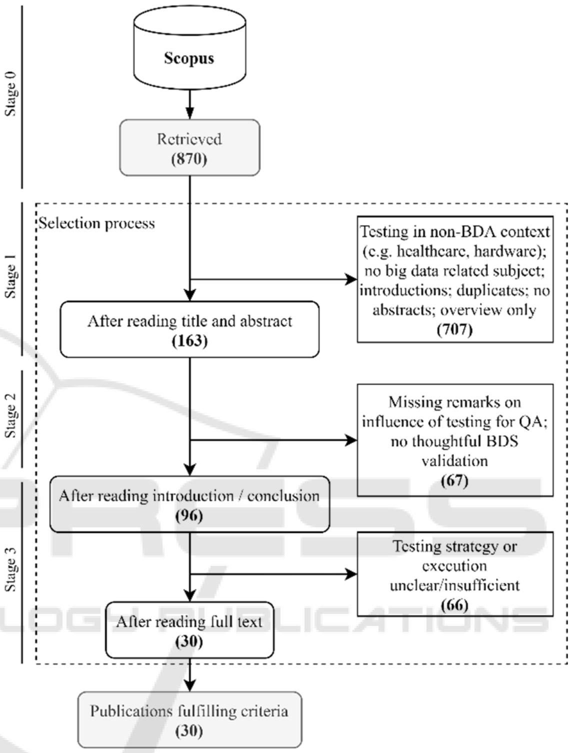
Stage 0
Scopus
Retrieved
(870)
Stage 1
Selection process
Testing in non-BDA context (e.g. healthcare, hardware); no big data related subject; introductions; duplicates; no abstracts; overview only
(707)
After reading title and abstract
(163)
Stage 2
Missing remarks on influence of testing for QA; no thoughtful BDS validation
(67)
After reading introduction / conclusion
(96)
Stage 3
Testing strategy or execution unclear/insufficient
(66)
After reading full text
(30)
Publications fulfilling criteria
(30)

Figure 3: Conducted search workflow.

In the following, the research questions established in section 3.1 are answered as far as the material fulfilling the corresponding review stages allows.

4.1 Quantitative Results

The found articles can be divided into seven categories. *Benchmarking* refers to techniques that compare a solution's performance against some predefined metric. *Experiments* are individual testing procedures in which a BDA solution's outcome is compared against an expected result. *Case studies* are comprehensive sets of activities in a realistic environment. *Surveys* mean that a BDA solution is instantiated and afterwards stakeholders are questioned about their experience regarding the results. *Discussions*, in turn, are partially opinion papers in which the topic of testing in BDA is considered from different angles. *Guidelines* mean sets of recommendations from experienced BDA developers based on profound prior knowledge. Finally, papers on challenges comprise lists of potential issues that need to be addressed when developing, deploying, and testing BDA systems. Table 5 summarizes the findings.

Table 5: Categorization of the found publications**.**

Category	Publications
Benchmarking	(Chen et al. 2018; Demirbaga et al. 2022; Ghazal et al. 2013; Hart et al. 2023; Skracic and Bodrusic 2017; Xia et al. 2019; X. Zhang et al. 2023; Zheng et al. 2017)
Experiment	(Covioli et al. 2023; Draughon et al. 2023; Fahrudin et al. 2022; Gulzar et al. 2018; Peng et al. 2020; Peng Zhang et al. 2018)
Case study	(Barba-González et al. 2019; Caputo et al. 2023; Eugeni et al. 2022; Prom-on et al. 2014; Sariyer et al. 2022)
Survey	(Côrte-Real et al. 2017; Shahbaz et al. 2019; C. Zhang et al. 2019)
Discussion	(Pengcheng Zhang et al. 2017)
Guidelines	(Shapira and Chen 2016; Staegemann et al. 2019; Tao and Gao 2016)
Challenges	(Alexandrov et al. 2013; Gulzar et al. 2019; Janković et al. 2018; Rabl et al. 2015)

4.2 Qualitative Results

The RQs on the primarily applied quality assurance methods and on how tests should be designed derive their answers from different subsets of the publications listed in Table 5. In the following, the methods of software testing (e.g., experiments, benchmarking, and case studies) and the more subjective conduct of surveys are explained based on the findings of this literature review. After that, the guidelines and extracted challenges are consulted to give an answer to the question how to organize test scenarios, and which issues could arise.

4.2.1 Surveys

The quality factors of big data applications are well known and partly similar to traditional software systems, including performance, reliability, correctness, security and the more application-specific scalability (Pengcheng Zhang et al. 2017). However, chronologically QA does not begin with the testing of a final system, but with pre-analyses on the issues of which purpose the developed solution should fulfill and how the acceptance by the intended user can be ensured. Although surveys in terms of asking participants for their opinion might lack objectivity (Bertrand and Mullainathan 2001), especially when determining the correctness of functional aspects, the subjective acceptance factors can still be investigated. Precautionary as well as during runtime, surveys can help in accumulating knowledge on user experience and satisfaction and are therefore considered as an adequate method of QA in the context of this paper. Shahbaz et al. (2019), for example, found that security and trust on information are important concerns of users by surveying 400 BDA end users in Pakistani hospitals, resulting in 224 valid responses. The authors concluded that the absence of sufficient trust leads to a higher resistance against a proposed BDA system which in turn negatively affects the overall usefulness. The survey of C. Zhang et al. (2019), also conducted in healthcare, observed similar issues that have to be considered before an actual implementation. Moreover, this research team suggests that the usefulness of a BDA system can depend on the knowledge of a user in the application field, meaning that knowing what and how data is stored and utilized within electronic health records (EHR) can have a significant influence. While the issue of storing data is linked to the software architecture and can therefore be viewed in connection with typical technical quality factors of BDA systems (Tao and Gao 2016), the knowledge about the EHRs is solely dependent on the training of the user. Although C. Zhang et al. (2019) pointed out that the lack of domain specific knowledge might have serious downturns on a systems performance, Shahbaz et al. (2020) found that the usefulness (i.e., the scope of the system) is more important than the perceived ease of use. Therefore, the potential for improvement of a BDA system can be higher in ensuring a better training for the employees than in the attempt of optimizing the user experience with the possible risk of removing parts of the functionality. Another dimension of suitability can be referred to as *technology-task fit* (TTF) (Shahbaz et al. 2019) which can be used as a term describing to what extend the features of the system match the intended task of a user. Organizations that heavily rely on economic success need to consider the aforementioned issues as well, as the study by Côrte-Real et al. (2017) shows. By surveying 500 European firms with 175 valid responses, the researchers found that organizational agility in particular is vital in competitive markets.

Beyond that, external knowledge was found to be more important than internal knowledge, which is why the call for an increase of the TTF has to be regarded in terms of adjustments on data sources and the corresponding analysis processes to make a BDA system more advantageous in a business context. Apart from these findings, the first part of an answer to RQ1 has to take the reliability of surveys into account. Although the complexity and the effort can be significantly lower than with other QA methods, results of surveys tend to be viewed as less reliable because the data mainly consists of the participants subjective opinions (Bertrand and Mullainathan 2001). Nonetheless, the easily obtained data can serve as a sufficient basis for important insights before an actual development or during runtime to identify necessary adjustments.

4.2.2 Experiments, Benchmarks, Case Studies

One technique for QA of big data applications is testing during development time (Pengcheng Zhang et al. 2017). The order of the remarks is based on the complexity and scope of the proposed methods, starting with individual experiments, through benchmarks to case studies. However, these methods might overlap regarding goals and work steps. Examining the literature reveals that studies conducting experiments and benchmarks often primarily concentrate on the performance of a system (e.g., transactions per minute, processing time) (Caputo et al. 2023; Draughon et al. 2023; Ghazal et al. 2013; Peng et al. 2020; Skracic and Bodrusic 2017; Xia et al. 2019; Peng Zhang et al. 2018). Zheng et al. (2017) criticize this focus, noting that also the reliability of a system has to be considered more intensive. Along with concerns about security, Peng et al. (2020) also recognize risks to overall performance if insufficient attention is paid to these two issues. In order to evaluate whether a tradeoff between security and performance is inevitable, the research team around Peng in this study conducted an experiment comparing two algorithms of access control in a BDA environment. The authors found that their newly developed algorithm enables a reliable security while affecting the performance less than the older solution. It can be concluded that these two dimensions of quality are not exclusive and can have a negligible tradeoff if carefully considered. Another issue for the reliability of tests in BDA can be associated with abstractions of data and workloads. A major difference between big data systems and traditional software resides in the considerably higher hardware requirements (Shapira and Chen 2016; Staegemann et al. 2019). Data scientists usually test their solutions with small samples of data on local workstations and hope that they will work equally well later when implemented in an expensive production cloud (Gulzar et al. 2018). While a cloud based testing environment might be thinkable depending on the resources of the respective organization, modern BDA systems also strive to take advantage of other current architectural ideas like edge computing and similar technologies (Xia et al. 2019). Those aspirations can even exceed the possibilities of a well-equipped company as such networks of devices can be of highly variable extend. Thus, certain risks when transferring a sufficiently locally tested solution into a real-world scenario can remain. Further studies detect dangers of abstraction due to a shortage of realistic test datasets and a resulting insufficient coverage of system behavior regarding the intended use case (Alexandrov et al. 2013; Rabl et al. 2015). In an attempt to address this issue, Gulzar et al. (2019) developed *BigTest*, a systematic input generation tool, and *BigSift*, an automated debugging toolkit (Gulzar et al. 2018), with a note in 2018 that the evolution of debugging in BDA is still in its early days. In the same year, Peng Zhang et al. (2018) followed a different approach in the application area of high-frequency trading by suggesting that transaction speed has already reached peak because of the current hardware capabilities. Therefore, the authors believe that more complex data analysis models are a key concept of further advancing traders performance. However, the study is limited to an approach to improve the performance of such complex models by adopting a parallel processing architecture which was then evaluated against competing approaches.

The incorporation of benchmarks into the testing strategy can constitute an advanced addition to single experiments with synthetic datasets. Alongside the usual functional performance measurements, benchmarks can enable more precise predictions regarding the price performance of a big data system as well, responding to the rising pressure to evaluate this quality factor (Ghazal et al. 2013). Beyond that, the scope of benchmarks can also include the verification of the correctness and the opportunity to compare different BDA systems (Chen et al. 2018). However, benchmarks must be standardized to gain meaningful comparisons of multiple solutions. In the late 1980s, this requirement led to the formation of benchmark consortia such as the Transaction Processing Performance Council (TPC) and the Standard Performance Corporation (SPEC) (Ghazal et al. 2013). While at that time the need for benchmarks for data warehouses was a crucial aspect, current software trends draw attention to benchmarks in the area of big data. Ghazal et al. (2013) responded to the call for more standardization by developing *BigBench* which was accepted by the TPC as a benchmark since its finalization. Apart from the notable interest in benchmarks in the literature, indications can be found that this technique may require a considerably higher effort than single experiments. Chen et al. (2018) present six phases of testing BDA systems by benchmarking, including a requirement analysis, preparing the test environment, preparing test datasets and workloads, loading the data, executing the tests and analyzing the results. While these phases may overlap with those of an experiment, they are not optional in the case of benchmarks and may exceed the requirements of small experiments, since they cover all system properties at once. The issue of generating realistic test data intensifies as the scenario has to be designed in a way that the special characteristics of big data (i.e., the Vs) are considered (Chen et al. 2018; Covioli et al. 2023; Pengcheng Zhang et al. 2017). Xia et al. (2019), for example, complement their benchmarking efforts with an evaluation in a real testbed. Thus, the problem of insufficient realistic test datasets can partly be tackled. Despite the popularity of experiments and benchmarks both can show a decisive pitfall when not adequately standardized. Organizations may tend to use self-defined and thus biased scenarios to make performance claims, a practice that Ghazal et al. (2013) refer to as *benchmarketing*. Unintentionally, such unrealistic claims can also occur when the benchmark is designed without sufficient consideration of real-life requirements (Shapira and Chen 2016). The reliability of both explained methods is therefore highly dependent on their degree of standardization.

The most time-consuming approach to testing a BDA system found in the literature are case studies in the real world, which is why only two publications of satisfactory depth were identified. Prom-on et al. (2014) conducted two different case studies, one to categorize social media posts into positive and negative opinions, and one for the prediction of traffic problems in Bangkok. Since the reliability heavily relies on the quality of the applied testing methodology, the time required already increases during the planning phase. Furthermore, in this case it is not possible to automate the test routine because only a properly functioning system could determine the correctness of the categorization, which in turn would be the system under test itself. The other publication utilizing case studies, written by Barba-González et al. (2019), introduces a framework to enhance analytical processes with semantic annotations. Their case studies were also conducted in the traffic sector and on a classification task, but in a comparative context. The reliability of the results is therefore less of a concern since the relative comparison is more comprehensible in terms of evaluating one solution over another than verifying the absolute results of a single system. The findings of the authors mirror the ones of Shahbaz et al. (2019) who found that more precise knowledge of the semantics of a system is beneficial for increasing the TTF and therefore the value of the software solution.

The previously explained findings from the literature are summarized in Table 6 listing the advantages and disadvantages of the QA techniques of conducting surveys, experiments, benchmarks and case studies. The table does not claim completeness as many more aspects of QA may have to be considered in any respective scenario.

Table 6: QA techniques and reliability in BDA.

Method	Advantages	Disadvantages
Survey	• Easy to perform • Provides prior knowledge for acceptance and user experience	• Questionable reliability and objectivity of data
Experiment	• Fast implementation on local workstations	• Uncertainty when transferred to real world scenario due to lack of realistic test datasets
Benchmark	• Comparison of different solutions • Standardized suitable for advertisement against competitors	• Possible laborious adjustments • Risk of benchmarketing
Case study	• Realistic conditions	• Time-consuming evaluation

Complexity / Reliability

4.2.3 Test Design and Possible Issues

As mentioned, Pengcheng Zhang et al. (2017) identified testing as an integral part of QA, but their study does not contain sufficient guidelines on how to design appropriate test scenarios. The most detailed and matured publication found on this subject is the industry experience paper by Shapira and Chen (2016). Besides conditions that have to be taken into account while testing, especially in terms of benchmarking, the paper also examines attributes of well-performed tests and reasons for their thoughtful execution. Among others, the authors demand the regard for realistic hardware and workload choices as well as the consideration of the systems properties such as the size of data and the number of nodes and tasks. A certain knowledge of the application domain is therefore a beneficial requirement for the testers or users, as other researchers have also noted (Shahbaz et al. 2019; C. Zhang et al. 2019). Especially emphasized by Shapira and Chen (2016) are the components of a proper documentation of the testing process. All necessary information to guarantee the reproducibility of the tests must be carefully recorded, including the configuration, hardware and workloads. The authors also make clear that the results must be unambiguous and that in tests that investigate the influence of a single parameter, it is essential to ensure that exclusively this parameter has been changed. Furthermore, according to the paper, outcomes are only reasonable if a model of the expected system behavior exists. Due to the explorative nature of BDA applications, this could be a difficulty since the desired output is not always known in a defined form (Staegemann et al. 2019). A possible solution for this test oracle problem can be metamorphic testing (MT). Tao and Gao (2016) explain, that a set of expected properties, known as metamorphic relation, specifies how the output would change following a particular change of the input. Another approach mentioned by the authors, assuming a sufficient number of known combinations of valid inputs and outputs, can be a trained classifier. Thus, the testing could be automated and newly produced outputs can be checked for their correctness. In all cases, the reusability and reproducibility of the tests should be considered, as this is generally viewed as an advantage (Shapira and Chen 2016; Staegemann et al. 2019). A possible dilemma for this demand are privacy issues if problematic system behavior occurs outside of internal tests and the users inputs, consisting of private data, would be mandatory for reproducibility (Alexandrov et al. 2013; Rabl et al. 2015; X. Zhang et al. 2023). Based on this, a further attribute of comprehensive testing is the coverage of every involved system component and status. Not only a shortage of realistic test data can lead to complications, but also the unpredictable behavior of external users. Although a randomly based input strategy could give a chance of covering neglected scenarios, it is unlikely to reveal every weak point, which is why a systematic testing strategy is generally more sensible (Gulzar et al. 2019). Nonetheless, the challenge of varying data origins and structures remains as it is a permanent requirement for BDA applications to integrate new sources (Janković et al. 2018). Hence, the security of the system might be endangered if those new sources are not under direct control of the user (Staegemann et al. 2019).

Summarizing the findings to answer RQ2, the design of a BDA test strategy might orientate on the following key concept. First, realistic datasets should be used whenever possible, otherwise a data model should first be created to generate synthetic datasets and workloads. Second, any procedure and configuration should be strictly documented to enable the reusability and reproducibility of the performed tests. Third, appropriate strategies which cover ideally every scenario and special case should be evaluated beforehand, especially if exact outcomes are unknown. Fourth, the security of the BDA system should always be considered for the case of integrating new data sources, structures and other risky adjustments.

5 CONCLUSION

This work presents a comprehensive literature review in the abstract and citation database Scopus to examine the role and adopted strategies of testing in current BDA approaches in the context of QA. Although Scopus provides the essential information to get hands on publications from different full-test databases like IEEE Xplore, ACM Digital Library and ScienceDirect (Staegemann et al. 2020) it does not cover the full range of every relevant database such as, for example, Springer Link. One improvement for this work might be the integration of the full-text databases themselves as considerations of the main topic could be hidden inside the text body and neglected in the title or abstract of an article. Moreover, a few terms that were found to occur frequently (e.g. *big data application*, *validation, evaluation*) could be a sensible addition for the search string.

Nonetheless, the main approaches on how to ensure the quality of a BDA project could be identified and thoroughly investigated. The most popular strategies here are standardized benchmarks, as provided by the TPC, because of their comparability, surveys in advance and at runtime because of their simplicity and case studies because of their elaborate results. However, every approach suffers possible downsides such as the lack of realism when benchmarking with synthetic data, the uncertainty of reliability of participants answers in surveys, and the necessary effort for real-world case studies. Aside discussions on those techniques, this work covers quality factors on *how* they should be applied according to various scholars. In this regard, the reproducibility and realism of test data turned out to be decisive attributes of well-conducted testing routines. Concluding it can be stated that despite all efforts and especially because of the existing lack of realistic test data, BDA applications will remain complex systems with a wide range of issues and long patch histories at runtime (Huang et al. 2015). In response to the initial question about the role of testing in current BDA developments, the systematic literature research carried out here supports the answer that testing is currently not meant to provide a perfectly running system, but rather to limit malfunctions to special cases that cannot be detected even by extensive testing or preliminary analysis.

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*. IEEE.

Agarwal, V., & Sunitha, B. (2020). COVID – 19: Current pandemic and its societal impact. In *International Journal of Advanced Science and Technology*, *29*, 432–439.

Alexandrov, A., Brücke, C., & Markl, V. (2013). Issues in big data testing and benchmarking. In *Proceedings of the Sixth International Workshop on Testing Database Systems - DBTest '13*. ACM Press.

Barba-González, C., García-Nieto, J., Roldán-García, M. d. M., Navas-Delgado, I., Nebro, A. J., & Aldana-Montes, J. F. (2019). BIGOWL: Knowledge centered Big Data analytics. In *Expert Systems with Applications*, *115*, 543–556.

Bertrand, M., & Mullainathan, S. (2001). Do People Mean What They Say? Implications for Subjective Survey Data. In *American Economic Review*, *91*(2), 67–72.

Bissi, W., Serra Seca Neto, A. G., & Emer, M. C. F. P. (2016). The effects of test driven development on internal quality, external quality and productivity: A systematic review. In *Information and Software Technology*, *74*, 45–54.

Caputo, F., Keller, B., Möhring, M., Carrubbo, L., & Schmidt, R. (2023). Advancing beyond technicism when managing big data in companies' decision-making. In *Journal of Knowledge Management*, *27*(10), 2797–2809.

Che, D., Safran, M., & Peng, Z. (2013). From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. In *Lecture Notes in Computer Science. Database Systems for Advanced Applications*. Springer Berlin Heidelberg.

Chen, M., Chen, W., & Cai, L. (2018). Testing of Big Data Analytics Systems by Benchmark. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE.

Côrte-Real, N., Oliveira, T., & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. In *Journal of Business Research*, *70*, 379–390.

Côrte-Real, N., Ruivo, P., Oliveira, T., & Popovič, A. (2019). Unlocking the drivers of big data analytics value in firms. In *Journal of Business Research*, *97*, 160–173.

Covioli, T., Dolci, T., Azzalini, F., Piantella, D., Barbierato, E., & Gribaudo, M. (2023). Workflow Characterization of a Big Data System Model for Healthcare Through Multiformalism. In *Lecture Notes in Computer Science. Computer Performance Engineering and Stochastic Modelling*. Springer Nature Switzerland.

Daase, C., Volk, M., Staegemann, D., & Turowski, K. (2023). The Future of Commerce: Linking Modern Retailing Characteristics with Cloud Computing Capabilities. In *Proceedings of the 25th International Conference on Enterprise Information Systems*. SCITEPRESS - Science and Technology Publications.

Demirbaga, U., Wen, Z., Noor, A., Mitra, K., Alwasel, K., Garg, S., Zomaya, A. Y., & Ranjan, R. (2022). AutoDiagn: An Automated Real-Time Diagnosis Framework for Big Data Systems. In *IEEE Transactions on Computers*, *71*(5), 1035–1048.

Draughon, G., Lynch, J., & Salvino, L. (2023). Integrated Vision-Body Sensing System for Tracking People in Intelligent Environments. In *Lecture Notes in Civil Engineering. European Workshop on Structural Health Monitoring*. Springer International Publishing.

Dremel, C., Herterich, M. M., Wulf, J., & vom Brocke, J. (2020). Actualizing big data analytics affordances: A revelatory case study. In *Information & Management*, *57*(1), 103121.

Eugeni, M., Quercia, T., Bernabei, M., Boschetto, A., Costantino, F., Lampani, L., Spaccamela, A. M., Lombardo, A., Mecella, M., Querzoni, L., Usinger, R., Aliprandi, M., Stancu, A., Ivagnes, M. M., Morabito, G., Simoni, A., Brandão, A., & Gaudenzi, P. (2022). An industry 4.0 approach to large scale production of satellite constellations. The case study of composite sandwich panel manufacturing. In *Acta Astronautica*, *192*, 276–290.

Fahrudin, T. M., Riyantoko, P. A., & Hindrayani, K. M. (2022). Implementation of Big Data Analytics for Machine Learning Model Using Hadoop and Spark Environment on Resizing Iris Dataset. In *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. IEEE.

Ghandour, A. (2015). Big Data Driven E-Commerce Architecture. In *International Journal of Economics, Commerce & Management (IJECM)*, *3*.

Ghazal, A., Rabl, T., Hu, M., Raab, F., Poess, M., Crolotte, A., & Jacobsen, H.-A. (2013). BigBench. In *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*. ACM Press.

Gulzar, M. A., Mardani, S., Musuvathi, M., & Kim, M. (2019). White-box testing of big data analytics with complex user-defined functions. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2019*. ACM Press.

Gulzar, M. A., Wang, S., & Kim, M. (2018). BigSift: automated debugging of big data analytics in data-intensive scalable computing. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018*. ACM Press.

Hamilton, R. H., & Sodeman, W. A. (2020). The questions we ask: Opportunities and challenges for using big data analytics to strategically manage human capital resources. In *Business Horizons*, *63*(1), 85–95.

Hart, M., Dave, R., & Richardson, E. (2023). Next-Generation Intrusion Detection and Prevention System Performance in Distributed Big Data Network Security Architectures. In *International Journal of Advanced Computer Science and Applications*, *14*(9).

Hong, Y., Li, Z [Zheng], & Wang, J. (2020). Business Value of Telecom Operators' Big Data. In *Journal of Physics: Conference Series*, *1437*.

Huang, J., Zhang, X [Xuechen], & Schwan, K. (2015). Understanding issue correlations. In *Proceedings of the Sixth ACM Symposium on Cloud Computing - SoCC '15*. ACM Press.

Jain, A., & Bhatnagar, V. (2016). Crime Data Analysis Using Pig with Hadoop. In *Procedia Computer Science*, *78*, 571–578.

Janković, S., Mladenović, S., Mladenović, D., Vesković, S., & Glavić, D. (2018). Schema on read modeling approach as a basis of big data analytics integration in EIS. In *Enterprise Information Systems*, *12*(8-9), 1180–1201.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering: Technical report, Version 2.3 EBSE Technical Report*.

Lonetti, F., & Marchetti, E. (2018). Emerging Software Testing Technologies. In *Advances in Computers*. Elsevier.

Martin, R. C. (2007). Professionalism and Test-Driven Development. In *IEEE Software*, *24*(3), 32–36.

Mithas, S., Lee, M. R., Earley, S., Murugesan, S., & Djavanshir, R. (2013). Leveraging Big Data and Business Analytics [Guest editors' introduction]. In *IT Professional*, *15*(6), 18–20.

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), 488–509.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. In *Communications of the Association for Information Systems*, *37*.

Ordonez, C., & Bellatreche, L. (2020). Guest Editorial—DaWaK 2018 Special Issue—Trends in Big Data Analytics. In *Data & Knowledge Engineering*, *126*.

Palomba, F., Di Nucci, D., Panichella, A., Oliveto, R., & Lucia, A. de (2016). On the diffusion of test smells in automatically generated test code. In *Proceedings of the 9th International Workshop on Search-Based Software Testing - SBST '16*. ACM Press.

Peng, J., Zhou, H., Meng, Q., & Yang, J. (2020). Big data security access control algorithm based on memory index acceleration in WSNs. In *EURASIP Journal on Wireless Communications and Networking*, *2020*(1).

Prom-on, S., Ranong, S. N., Jenviriyakul, P., Wongkaew, T., Saetiew, N., & Achalakul, T. (2014). DOM: A big data analytics framework for mining Thai public opinions. In *2014 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*. IEEE.

Rabl, T., Danisch, M., Frank, M., Schindler, S., & Jacobsen, H.-A. (2015). Just can't get enough. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*. ACM Press.

Saheb, T., & Izadi, L. (2019). Paradigm of IoT big data analytics in the healthcare industry: A review of scientific literature and mapping of research trends. In *Telematics and Informatics*, *41*, 70–85.

Sariyer, G., Ataman, M. G., Mangla, S. K., Kazancoglu, Y., & Dora, M. (2022). Big data analytics and the effects of government restrictions and prohibitions in the COVID-19 pandemic on emergency department sustainable operations. In *Annals of Operations Research*, 1–31.

Shahbaz, M., Gao, C., Zhai, L., Shahzad, F., & Arshad, M. R. (2020). Moderating Effects of Gender and Resistance to Change on the Adoption of Big Data Analytics in Healthcare. In *Complexity*, *2020*, 1–13.

Shahbaz, M., Gao, C., Zhai, L., Shahzad, F., & Hu, Y [Yanling] (2019). Investigating the adoption of big data analytics in healthcare: the moderating role of resistance to change. In *Journal of Big Data*, *6*(1).

Shapira, G., & Chen, Y. (2016). Common Pitfalls of Benchmarking Big Data Systems. In *IEEE Transactions on Services Computing*, *9*(1), 152–160.

Skracic, K., & Bodrusic, I. (2017). A Big Data solution for troubleshooting mobile network performance problems. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE.

Staegemann, D., Volk, M., Daase, C., & Turowski, K. (2020). Discussing Relations Between Dynamic Business Environments and Big Data Analytics. In *Complex Systems Informatics and Modeling Quarterly*(23), 58–82.

Staegemann, D., Volk, M., Nahhas, A., Abdallah, M., & Turowski, K. (2019). Exploring the Specificities and Challenges of Testing Big Data Systems. In *2019 15th*

International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE.

Tao, C., & Gao, J. (2016). Quality Assurance for Big Data Application – Issuses, Challenges, and Needs. In *International Conferences on Software Engineering and Knowledge Engineering. In Proceedings of the 28th International Conference on Software Engineering and Knowledge Engineering*. KSI Research Inc. and Knowledge Systems Institute Graduate School. https://doi.org/10.18293/SEKE2016-166

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. In *Journal of Business Research*, *70*, 356–365.

Xia, Q., Bai, L., Liang, W., Xu, Z., Yao, L., & Wang, L. (2019). QoS-Aware Proactive Data Replication for Big Data Analytics in Edge Clouds. In *Proceedings of the 48th International Conference on Parallel Processing: Workshops*. ACM.

Zhang, C., Ma, R., Sun, S., Li, Y., Wang, Y., & Yan, Z. (2019). Optimizing the Electronic Health Records Through Big Data Analytics: A Knowledge-Based View. In *IEEE Access*, *7*, 136223–136231.

Zhang, P [Peng], Gao, Y., & Shi, X. (2018). QuantCloud: A Software with Automated Parallel Python for Quantitative Finance Applications. In *2018 IEEE International Conference on Software Quality, Reliability and Security (QRS)*. IEEE.

Zhang, P [Pengcheng], 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)*. IEEE.

Zhang, X [Xiang], Lu, B., Zhang, L [Lyuzheng], Pan, Z., Liao, M., Shen, H., Zhang, L [Li], Liu, L., Li, Z [Zuxiang], Hu, Y [YiPao], & Gao, Z. (2023). An enhanced grey wolf optimizer boosted machine learning prediction model for patient-flow prediction. In *Computers in Biology and Medicine*, *163*, 107166.

Zheng, Y., Xu, L., Wang, W., Zhou, W., & Ding, Y. (2017). A Reliability Benchmark for Big Data Systems on JointCloud. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE.