Complex Systems Informatics and Modeling Quarterly (CSIMQ)
eISSN: 2255-9922
Published online by RTU Press, https://csimq-journals.rtu.lv
Article 159, Issue 28, September/October 2021, Pages 1–14
https://doi.org/10.7250/csimq.2021-28.01

Quality Assurance in Big Data Engineering - A Metareview

Daniel Staegemann*, Matthias Volk, and Klaus Turowski

Otto von Guericke University, Universitätsplatz 2, Magdeburg, 39106, Germany

daniel.staegemann@ovgu.de, matthias.volk@ovgu.de, klaus.turowski@ovgu.de

Abstract. With a continuously increasing amount and complexity of data being produced and captured, traditional ways of dealing with their storing, processing, analysis and presentation are no longer sufficient, which has led to the emergence of the concept of big data. However, not only the implementation of the corresponding applications is a challenging task, but also the proper quality assurance. To facilitate the latter, in this publication, a comprehensive structured literature metareview on the topic of big data quality assurance is presented. The results will provide interested researchers and practitioners with a solid foundation for their own quality assurance related endeavors and therefore help in advancing the cause of quality assurance in big data as well as the domain of big data in general. Furthermore, based on the findings of the review, worthwhile directions for future research were identified, providing prospective authors with some guidance in this complex environment.
Keywords: Big Data, Quality Assurance, Benchmark, Testing, Literature Review, Metareview.

1 Introduction

With the ongoing surge in data production [1], [2], the possibilities for gaining new insights and creating new knowledge based on those data are also tremendously increasing. Yet, with those theoretical opportunities also comes the challenge of actually realizing those potential gains. This is, however, a highly complex task, because traditional ways of storing, processing, analyzing or presenting data are oftentimes no longer sufficient to keep up with the new demands that accompany the development of the available inputs and desired outputs [3]. For this reason, new concepts, techniques, tools and tactics have emerged that are focused on dealing with huge amounts of diverse data, while adhering to time constraints [4]. In conjunction of the data itself, they are amalgamated under the term big data, with the corresponding data investigation for the purpose of gaining knowledge and insights being denoted as big data analytics (BDA). To provide value, those findings have to be converted into actions, which will then, for instance, help in winning customers, finding new product ideas, reducing maintenance costs, increasing the efficiency of an organization or predicting future developments [5], [6]. Correctly harnessing big

* Corresponding author

Reference: D. Staegemann, M. Volk, and K. Turowski, "Quality Assurance in Big Data Engineering - A Metareview," Complex Systems Informatics and Modeling Quarterly, CSIMQ, no. 28, pp. 1–14, 2021. Available: https://doi.org/10.7250/csimq.2021-28.01

Additional information. Author ORCID iD: D. Staegemann – orcid.org/0000-0001-9957-1003 and M. Volk – orcid.org/0000-0002-4835-919X. PII S225599222100159X. Received: 19 August 2021. Accepted: 25 October 2021. Available online: 31 October 2021.

data can therefore demonstrably increase an organizations performance [7]. Though, to warrant those benefits and also create trust in the designated users, the BDA applications have to be of high quality, to assure the validity of their outputs and facilitate their actual implementation [8], [9]. As with other products and services, this necessitates a rigorous quality assurance as part of the creation process, which is in this case are called big data engineering [10]. However, this aspect is often somewhat neglected [11], even though low quality BDA can severely impact the results in a negative manner [12].

If it comes to the scientific realm, there is a plethora of different approaches, tools, and suggestions on how to facilitate the quality assurance of big data application. Those include, inter alia, benchmarking suites [13], test data generators [14], ETL testing [15], ontology-based testing [16], metamorphic testing [17], the application of test driven development [18] or the simulation of databases [19]. Though, while this variety appears to be beneficial on the one hand, it can also cause disorientation and can turn into a distraction when attempting to get an overview of the larger picture. However, such an overview can be a highly valuable foundation for future research endeavors, which can build upon the existing knowledge instead of just involuntarily repeating it and which can also be motivated and guided by research gaps that have been previously identified [20]. To facilitate this process, literature reviews are usually the means of choice. Yet, those are usually limited on certain aspects of a topic, since the vastness of the available content would otherwise be overwhelming. When attempting to get a more comprehensive view of a domain, it is therefore sensible to conduct a metareview, which brings together the findings of those individual reviews. In doing so, not only a broader outline of the domain can be given, but also biases, which might be inherent to certain studies, can be somewhat compensated for.

To provide such a comprehensive view into the domain of quality assurance in big data engineering, this article seeks to answer the following research question (RQ) by conducting a metareview.

RQ: What is the current state of the art in the domain of quality assurance in big data engineering?

To compile the desired answer, the RQ is divided into several sub research questions (SRQ), which each deal with a certain aspect of the big picture.

SRQ 1: How many and which corresponding review papers can be found?
SRQ 2: Where are those reviews published and how high is their resonance?
SRQ 3: What are the respective findings of the existing review papers?
SRQ 4: Which aspects of the quality assurance in big data engineering can be considered rather mature?
SRQ 5: Which challenges have not been sufficiently solved and what are possible approaches?

The remainder of the article is structured as follows. Succeeding this introduction, in Section 2, the topic of big data is delineated more in-depth, to provide a clear understanding of the subject of discussion. Afterwards, in Section 3, the review protocol is outlined, followed by a presentation of the findings. Those are subsequently discussed in Section 4. Finally, in Section 5, a conclusion of the presented work is given, including a discussion of the study's limitations and avenues for future research.

2 Big Data

The domain of big data is highly complex and subject to numerous publications, focusing on a broad variety of aspects [21]. While it is neither possible, nor necessary, to comprehensively describe all of them to answer the RQ, a brief overview of the most relevant big data characteristics, its general application and the corresponding quality assurance is given to provide a solid foundation for the ensuing review process.

2.1 Characteristics

As opposed to what the term big data suggests, the concept is not only based on the amount of data, but multiple characteristics. Furthermore, despite its huge popularity, neither its origins are clarified beyond doubt [22], nor is there a single, universally applied definition in use. Instead, numerous researchers have proposed their own explanations, which are, however, usually rather similar concerning the general idea and mostly differ in detail [23], [24]. Many of those definitions are based on certain characteristics of the regarded data, which are especially pronounced and therefore necessitate special treatment. This approach can be traced back to the year 2001, where Doug Laney used volume, velocity and variety to describe the essence of big data [25]. One of the most popular explanations originating from the National Institute of Standards and Technology (NIST) has added also the aspect of variability. In NIST it is stated that big data "consists of extensive datasets primarily in the characteristics of volume, velocity, variety, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis" [3]. Besides those four characteristics, which are also often referred to as the *Vs* of big data, due to their common initial letter, there are numerous more being mentioned in the literature, as, for instance, in [26], where 51 of those Vs are proposed.

However, in the following, only the aforementioned core characteristics, including the variability, will be further regarded, since those are the most common and broadly applicable ones. *Volume* can signify two aspects, the huge number of data entries that have to be handled and an exceptionally big storage requirement of the data sets [3], [27]. *Velocity* also bears two meanings. While many experts in this field use this to address the speed of pure data processing, according to other voices this property also refers to the speed at which the data arrives [28]. *Variety* refers to the diversity of data in terms of its structure, resulting in the categories of structured, semi-structured and unstructured data. The first category describes data sets that have a fixed schema and can therefore usually be stored, managed and analyzed without major effort. Semi-structured data, on the other hand, contains partial information about the underlying structure and can be modified and extended, as is the case, for instance, when using the extensible markup language (XML) exchange format. In contrast, unstructured data can generally only be handled using special procedures. This comprises, for instance, pictures, sound files or videos. Furthermore, factors such as the content of the data, its origin, the underlying context, the language used, units used or formatting rules are also addressed [3], [23], [28]. *Variability*, as the fourth common characteristic, includes the changes in the data over time, which can, inter alia, occur due to trends or seasonal events [28].

2.2 Application

Since many organizations across various fields of activity can benefit from an improved decision making, the use of BDA has reached a variety of domains with highly diverse requirements and contexts. This list comprises, for instance, civil protection [29], healthcare [30], agriculture [31], manufacturing [32], sports [33] and the steering of urban infrastructure [34].

To turn the relevant occurrences in real life into a useful asset for guiding or even automatically triggering actions, there is a usually a complex chain of activities required. At first, the data have to be *acquired*. This could, for instance, happen through sensors, which monitor certain events and statuses, people inputting data, the repurposing of already existing and available data or the sourcing from external origins. Furthermore, those data have to be *stored*, even though, it is not always necessary to permanently keep them in storage, in many use cases historic data can be of value. Another common building block of BDA applications is the *pre-processing* of the data. Here, it is attempted to mitigate potential data quality issues, as well as to account for the data variety. The former means that, for instance, missing data, outliers, typos or invalid entries are accounted for, by either correcting, interpolating or excluding the respective entries, depending on the specified policy. The latter refers to the fusion of data whose amalgamation makes sense from

a content perspective but is hampered by different structures. An example for this would be the use of varying formats for the date specification, depending on the regarded country. While the data might be otherwise compatible, they can only be appropriately joined when unifying the data representations. Once the data are prepared, the actual analysis can ensue. Here, it can be differentiated between descriptive (how was the past), diagnostic (why is something the way it is), predictive (how will the future look like) and prescriptive (what should be done) knowledge [35]. Depending on the use case, a variety of approaches and algorithms can be applied, ranging from the creation of rather simple statistics to highly complex machine learning algorithms. Finally, the results of the analysis have to be either directly converted into *action*, for instance, by automatically reconfiguring a machine, or they have to be *visualized* to present the underlying information in an adequate way that allows the users to obtain the desired insights.

2.3 Quality Assurance

Because the benefits of harnessing big data can only be reaped when the quality in the whole process chain is adequate [8], it is obviously necessary to assure that this is the case [36]. However, BDA is reliant on the interplay of several aspects, rendering it a multidimensional endeavour [37]. Consequently, those facets are also relevant when it comes to the quality assurance. Regarding the BDA application itself, there are two aspects that have to be considered. On one hand, it is important to test that a system actually provides the desired functionality without the algorithms or the internal communication being erroneous, which might skew the results (e.g., if the data import from certain sources is erroneous, the data might be completely ignored or interpreted incorrectly) or even entirely prohibit their use and, on the other hand, it is also important to have a sufficient performance to keep up with the incoming workload as well as the corresponding requirements. For the latter, oftentimes benchmarks are used to warrant comparability of the results. They can, for instance, provide insights into the resource consumption, the processing speed or the maximum capacity of a solution. Thus, related issues can be identified and subsequently addressed.

Moreover, both approaches usually have to be repeated when the regarded system is changed, due to additions or modifications. Checking not only the new part of the system, but also the already existing parts, assures that they are not negatively impacted by the change. This practice goes by the name “regression testing”. It is especially important in large and highly complex systems, since these systems can easily become incomprehensible, which makes it hard to keep track of all dependencies and possible side effects. Besides the application’s actual implementation, the quality of the used data plays an important role [38], with the same applying to the aptitude of the personnel dealing with the systems, respectively being responsible for the corresponding decision making [12]. If the data that an analysis is based on are improper, wrong, insufficient or incomplete, the results will also be flawed. However, even if the analysis itself is generally well implemented, but its focus is not suited to facilitate solving the underlying issue of the concerned organization or if the results of the analysis are ignored, respectively skewed in order to support a certain position or strategy, the benefit is heavily reduced. However, since the RQ is focused on big data engineering, the data quality, the human component with respect to the solution’s use and also the monitoring of already running solutions are not relevant. Therefore, in the following review protocol, only the testing and the benchmarking of BDA applications are considered.

3 The Review

To answer the RQ, a structured literature review (SLR), oriented on the methodologies proposed by Webster and Watson [39], Levy and Ellis [40], and Okoli [41] is conducted. Since such reviews are not supposed to be an end in itself, but a valuable tool to the research community as a whole, it is necessary to assure rigor. Furthermore, a comprehensive description of the process that is as

detailed as possible should be given to enable others to retrace and evaluate the undertaken steps, so they can judge the review's value for their own work and possibly also build upon it [20].

Though, unlike in common SLRs, in a metareview, the focus is not on primary studies but on existing reviews concerning a topic, which are collated, analyzed and brought together to generate new insights and provide a broader overview of the underlying subject. However, apart from that, the general methodology remains the same.

3.1 Protocol

To start the whole search process, which is depicted in Figure 1, a set of promising databases to be considered was determined. Since the goal is to provide a comprehensive overview of the domain, this list should be rather extensive, to reduce the risk of missing out on relevant publications. For the presented review it was also not discriminated between meta databases and publisher bound databases and instead, both types were included, further increasing the likelihood of finding all the suitable contributions. Following this line of thought, *ACM Digital Library*, *AISeL*, *Emerald Insight*, *IEEE Xplore*, *Mary Ann Libert*, *Sage*, *Science Direct*, *SciTePress*, *Scopus*, *Springer Link*, *Taylor & Francis* and *Wiley* were included into the search process. While it was initially also planned to include the repository of IGI publishing, the search engine did not allow for a sufficiently elaborate filtering, therefore, it was dropped.

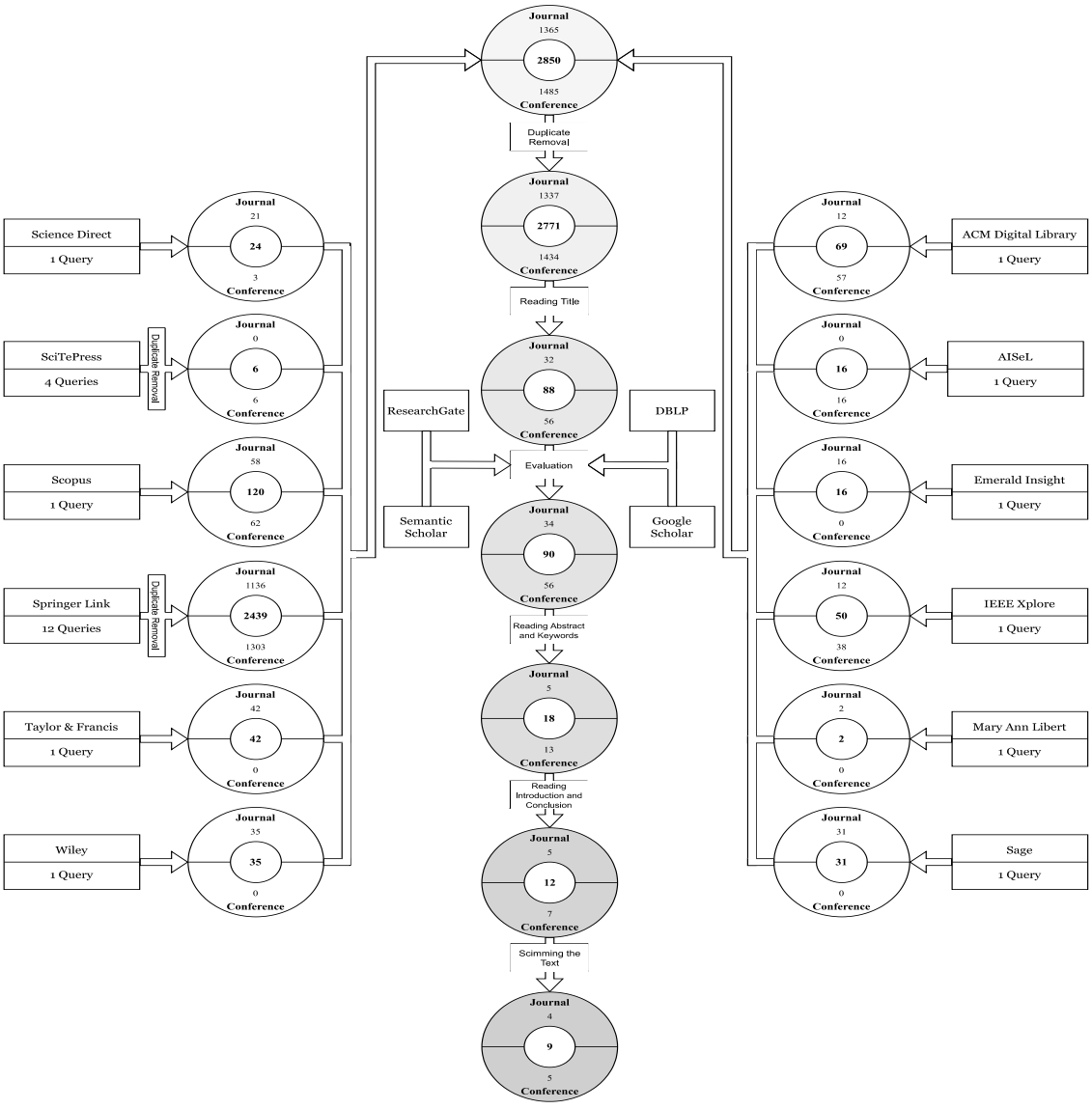

Figure 1. The Search Process

Based on the starting date of the literature review and to increase traceability by only considering completed years, the regarded timeframe was everything *before 2021* and, as depicted in Table 1, only *conference publications* and *journal articles* were considered, since those are peer-reviewed, assuring a certain degree of quality. Since the RQ focuses on getting an overview of quality assurance in big data engineering, this comes down to testing and benchmarking, which is reflected in the search terms that include *big data* as the regarded domain, relevant synonyms for *overview* and *test*, *benchmark* or *quality*. Besides the general process, Figure 1 also shows the specific number of findings for each search engine after accounting for the duplicates that resulted from the multiple queries that had to be conducted for SciTePress and Springer Link.

Table 1. Mapping of the Search Terms to the Used Engines

Search Engine	**Search Term**	**Type**
ACM Digital Library; AISeL; Emerald Insight; Mary Ann Libert; Sage; Taylor & Francis; Wiley	Title: "Big Data" AND ("Review" OR "Survey" OR "Overview" OR "State of the Art"); Anywhere: Quality OR Test* OR Benchmark*	Journal or Conference
IEEE Xplore	("Document Title": "Big Data") AND ("Document Title": "Review" OR "Document Title": "Survey" OR "Document Title": "Overview" OR "Document Title": "State of the Art") AND ("All Metadata": Test* OR "All Metadata": Quality OR "All Metadata": Benchmark*)	Journal or Conference
Science Direct	Title: "Big Data"; Title: "Review" OR "Survey" OR "Overview" OR "State of the Art"; Title/Abstract/Keywords: Quality OR Test OR "Benchmark	Journal or Conference
SciTePress 1	Title: "Big Data"; Title: "Review"	Journal or Conference
SciTePress 2	Title: "Big Data"; Title: "Survey"	Journal or Conference
SciTePress 3	Title: "Big Data"; Title: "Overview"	Journal or Conference
SciTePress 4	Title: "Big Data"; Title: "State of the Art"	Journal or Conference
Scopus	Title: "Big Data"; Title: "Review" OR "Survey" OR "Overview" OR "State of the Art"; Title/Abstract/Keywords: Quality OR Test* OR "Benchmark*	Journal or Conference
Springer Link 1	Title: "Big Data"; Anywhere: "Review" OR "Survey"; Anywhere: "Quality"	Journal
Springer Link 2	Title: "Big Data"; Anywhere: "Review" OR "Survey"; Anywhere: "Quality"	Conference
Springer Link 3	Title: "Big Data"; Anywhere: "Review" OR "Survey"; Anywhere: "Test*"	Journal
Springer Link 4	Title: "Big Data"; Anywhere: "Review" OR "Survey"; Anywhere: "Test*"	Conference
Springer Link 5	Title: "Big Data"; Anywhere: "Review" OR "Survey"; Anywhere: "Benchmark*"	Journal
Springer Link 6	Title: "Big Data"; Anywhere: "Review" OR "Survey"; Anywhere: "Benchmark*"	Conference
Springer Link 7	Title: "Big Data"; Anywhere: "Overview" OR "State of the Art"; Anywhere: "Quality"	Journal
Springer Link 8	Title: "Big Data"; Anywhere: "Overview" OR "State of the Art"; Anywhere: "Quality"	Conference
Springer Link 9	Title: "Big Data"; Anywhere: "Overview" OR "State of the Art"; Anywhere: "Test*"	Journal
Springer Link 10	Title: "Big Data"; Anywhere: "Overview" OR "State of the Art"; Anywhere: "Test*"	Conference
Springer Link 11	Title: "Big Data"; Anywhere: "Overview" OR "State of the Art"; Anywhere: "Benchmark*"	Journal
Springer Link 12	Title: "Big Data"; Anywhere: "Overview" OR "State of the Art"; Anywhere: "Benchmark*"	Conference

With the included search engine's filters not necessarily being identical, it was not possible to use the exact same search terms and settings for all of them. However, the conditions of the search were kept as similar as possible. The applied mapping of databases and search terms is shown in Table 1. As shown there, due to its specifics, the searches in Springer Link and SciTePress had to be split in several parts, which were later on merged again. Furthermore, while most search engines support wildcard symbols (*), Science Direct does not, which leads to the minimal deviation from the Scopus search term.

The obtained papers were subsequently merged in two sets, one containing conference publications and the other one journal articles. This resulted in 1485 entries for the former category and 1365 for the latter. However, those numbers already take the previous removal of duplicates within the same search engine (this applies to Springer Link and SciTePress) into account. Afterwards, those collections were cleansed from duplicates in their entirety, leaving 1434 unique conference papers and 1337 articles. Those papers were in the following filtered, using the inclusion and exclusion criteria presented in Table 2. While all the inclusion criteria had to be met for a paper to be accepted, the fulfilment of any exclusion criterion led to its dismissal.

Table 2. Inclusion and Exclusion Criteria

Inclusion Criteria	**Exclusion Criteria**
Published in Conference Paper or Journal Article	Not Written in English
Is Peer-Reviewed	Only Deals with Data Quality
Deals with Quality Assurance or Benchmarking of Big Data Applications	No Sufficient Referencing to Support Made Claims
Is a Review Paper	Insufficient Comprehensiveness
Was Written Before 2021	

In a first step, the papers' titles were read and those deemed unsuitable discarded. This already led to a reduction to 56 conference papers and 32 journal articles. To somewhat validate the comprehensiveness and also potentially add additional inputs, *DBLP*, *Google Scholar*, *ResearchGate* and *Semantic Scholar* were used, which were not considered for the initial search, since they tend to deliver numerous but oftentimes also irrelevant results. For those, the search parameters depicted in Table 3 were applied. While those might not provide full coverage, it should give at least a general idea, if the previous searches are majorly incomplete, or, in case of only limited additions, adequate.

Table 3. Details of the Additional Searches

Search Engine	**Search Term**	**Additional Info**
DBLP 1	"Big Data Review"	Regarded all entries.
DBLP 2	"Big Data Survey"	Regarded all entries.
DBLP 3	"Big Data Overview"	Regarded all entries.
DBLP 4	"Big Data State of the Art"	Regarded all entries.
Google Scholar 1; Semantic Scholar 1	(intitle:(Review OR Survey OR Overview OR "State of the Art") AND intitle: "Big Data") AND (intitle:(quality OR test* OR benchmark*)	Regarded the first 100 entries, since those are already sorted by relevance.
Google Scholar 2; Semantic Scholar 2	(intitle:(Review OR Survey OR Overview OR "State of the Art") AND intitle: "Big Data") AND (intext:(quality OR test* OR benchmark*)	Regarded the first 100 entries, since those are already sorted by relevance.
ResearchGate	"Big Data" AND ("Review" OR "Survey" OR "Overview" OR "State of the Art") AND (Quality OR Test* OR Benchmark*)	Regarded the first 100 conference papers, articles or book chapters. The latter because sometimes conference papers are considered book chapters by ResearchGate.

Those further searches resulted in the addition of two more journal articles to the regarded body of literature. While it shows that the initial search did not provide complete coverage, the relatively low number also indicates, that the search was already very comprehensive. In a next step, the gathered publications' abstracts and keywords were read, resulting in the remaining of 13 conference papers and five articles. After reading the introduction and conclusion, the set was reduced to seven contributions from conferences and five from journals. In a last filtering step, those were skimmed in their entirety, leaving five conference papers and four journal articles as the final set of reviews for answering the RQ, with no additional findings when conducting a backward search in their respective reference sections. Those numbers therefore constitute the first half of the answer to SRQ 1 and are also relevant concerning SRQ 2.

3.2 Findings

The set of relevant papers, resulting from the conducted review, is shown in Table 4, which is, therefore, answering SRQ 1. Furthermore, with the added information concerning the publication outlet and the current citation score, also SRQ 2 is answered. The depicted number of citations for each paper was queried in Google Scholar on February 18, 2021. In the following, each of the nine papers is briefly presented, thereby addressing SRQ 3.

Table 4. Overview of the Regarded Review Papers

Ref.	Title	Published In	Type	Focus	Year	Cit.
[42]	A Survey on Benchmarks for Big Data and Some More Considerations	Lecture Notes in Computer Science	Conference Paper	Benchmarking	2013	9
[43]	On Big Data Benchmarking	Lecture Notes in Computer Science	Conference Paper	Benchmarking	2014	38
[44]	Big Data Benchmark Compendium	Lecture Notes in Computer Science	Conference Paper	Benchmarking	2015	29
[45]	A Survey on Quality Assurance Techniques for Big Data Applications	Proceedings of the Big Data Service 2017	Conference Paper	General Quality Assurance Techniques	2017	9
[46]	Big Data DBMS Assessment: A Systematic Mapping Study	Lecture Notes in Computer Science	Conference Paper	Benchmarking	2017	2
[47]	Benchmarking Big Data Systems: A Review	IEEE Transactions on Services Computing	Journal Article	Benchmarking	2017	44
[48]	Benchmarking big data systems: A survey	Computer Communications	Journal Article	Benchmarking	2020	4
[49]	Big Data Systems: A Software Engineering Perspective	ACM Computing Surveys	Journal Article	Big Data Engineering (Relevant: Testing)	2020	0
[50]	Quality Assurance Technologies of Big Data Applications: A Systematic Literature Review	Applied Sciences	Journal Article	General Quality Assurance Techniques	2020	0

In [42], which is the earliest paper considered in this metareview, 18 benchmarks for measuring the performance of varying tools for (big) data management are presented and compared. The focus is on different types of databases and specific technologies like Hadoop or MapReduce. Furthermore, to account for big data's inherent variety that also heavily influences the corresponding systems, the authors state the necessity to not only benchmark just end-to-end or component based, but to combine both approaches to get a more comprehensive assessment of the evaluated systems. For the same reason, the selection of the data used for benchmarking to match real world workloads instead of just using SQL queries to gauge a system's performance is

highlighted as a necessity. Additionally, it is made a plea for considering scalability, energy efficiency, fault tolerance (by purposefully injecting failures into the target system) and also security aspects on top of the common performance metrics, when benchmarking big data systems.

One year later, [43] directed their attention to the requirements and challenges concerning the generation of data used for benchmarking big data applications. They point out the importance of a benchmark being application-specific and subsequently the need to identify typical workload behaviors as a foundation for the actual performance evaluation. Besides that, they state several requirements for a successful benchmarking test, namely the capability to adapt to different data formats, portability to a broad spectrum of representative software stacks, fairness (e.g. not comparing a system that is using the default configuration with a perfectly configured one), extensibility and usability. Furthermore, they gave an overview of existing benchmarks, which comprises ten instances and puts special emphasis on their data generation techniques. For the challenges, they especially highlight the controllability of the data velocity, the evaluation of the generated data's veracity and the enrichment of the data with meta information (e.g., arrival rate). Further, they call for the support of heterogeneous hardware platforms, a focus on the adaptability and reusability of existing components as well as an extension of the use cases of the benchmarks, since the respective applicability of the examined ones was rather narrow.

The most extensive overview of big data benchmarking tools was provided by [44]. In there, 20 benchmarks and two benchmarking platforms are described. Furthermore, it is highlighted, that the existing benchmarks are usually rather specific and that even though it is desirable to have an objective and universal benchmark, taking account of all the relevant aspects renders its creation a highly complex task. Besides that, as in the previous papers, the importance of workload diversity is stressed along with the possibility to integrate new workloads. It is also pointed out that workloads in a suite should be seen as complimentary and redundancy should be avoided. Additionally, an urgent need for new benchmark metrics is expressed, since the ones encountered in the course of the study were deemed insufficient.

An overview of quality assurance techniques was given in [45]. The paper lists and summarizes ten contributions that each deal with one of the following topics: testing, model-driven architecture (MDA), monitoring, fault tolerance, prediction and verification. Further, the functional or non-functional properties and big data Vs that are the most relevant with respect to those categories are listed. The authors also discussed three major issues, namely a lack of understanding of big data quality assurance techniques, a lack of quality assurance standards for big data and the challenge of keeping up with technological development. By including MDA and fault tolerance, it is highlighted that quality assurance can already be considered in the early design phases and in the choice of the development approach; and not just after something was implemented.

In the same year, a systematic mapping study of database management systems (DBMS) assessment was presented in [46]. There, it was analyzed, which DBMS were evaluated in the literature and which benchmarks were used for this purpose. During the study, the heterogeneity of the domain was highlighted, which showed in a plethora of encountered approaches. One additional finding was that a large proportion of the regarded primary studies was constituted by proposals and evaluations in a laboratory context, showcasing a lacking validation in industrial settings.

The first journal article to appear in this collection [47] also targets the domain of big data benchmarking. It proposes three categories, micro benchmarks (for individual components), end-to-end benchmarks (for the entire system) and benchmark suits, which combine the former two to provide comprehensive benchmarking capabilities. Furthermore, the systems are also divided into three types, Hadoop-related-systems, data stores and specialized systems, which are further partitioned into subsets. Subsequently, the derived categories are joined to form a matrix. It is then used to categorize 37 regarded benchmarks. Additionally, this overview is enriched by additional information and the underlying data generation techniques are also specifically discussed. Further, an overview of the encountered evaluation metrics is given. The papers main contribution is, therefore, to provide a comprehensive summary of the domains state of the art. The main

challenges for big data benchmarking that are pointed out in this article are the assurance of relevancy, portability and scalability as well as the generation of suitable test data and the assessment of their veracity.

In the latest of the collection's papers that are focused on benchmarking [48], six types of big data technologies are determined. Those are NoSQL databases, SQL Systems, batch processing, stream processing, graph processing and deep learning/machine learning. Example technologies for those types are stated, applicable benchmarks shown and corresponding scientific publications are introduced. This contribution once again exhibits that there is no universally applicable solution and benchmarking remains a rather individual and tool specific task. Nevertheless, the authors highlight the lack of collaborative efforts towards the creation of benchmarks and suggest the pursuit of community driven approaches.

While [49] is not a pure literature review on quality assurance in big data, but takes a broader perspective, it still extensively deals with those aspects. In the course of this work, several categories of quality assurance challenges are elaborated. Those are the verification of test results, the resource-intensity of the testing environments, which can make it really challenging to simulate real world scenarios, the generation of appropriate test data, error-tracing and the difficulties that come with distributed log files, verification, a weakened notion of data consistency and the assessment of data quality, which is, however, out of this review's scope. Specific to the software testing in the big data domain the paper states three dimensions. The first is the test objective with its three categories – data quality testing, functional testing and non-functional testing. As a second dimension, the granularity level states if algorithms, components, subsystems or the system as a whole are regarded. Finally, the test execution level states if a system is evaluated in a static fashion, with dynamic tests or if the running system is monitored. Furthermore, the paper highlights open research challenges such as the abandonment of costly test environments, the generally immature tool support for testing big data applications, scaling issues, the absence of a suitable testing approach that is geared towards the velocity as one of big data's characteristics and the oracle problem. To alleviate the latter for the quality assurance of BDA that are based on machine learning, the authors highlight metamorphic testing as a promising approach.

Finally, the latest contribution [50] can probably be seen as a continuation of [45]. This refers to the content, but also to the role of Pengcheng Zhang, who co-authored both papers. In comparison to traditional applications, for big data ones, it especially highlights the challenges of conducting statistical computation based on diverse data in large-scale, the utilization of machine learning techniques, decision-making under uncertain conditions and complex requirements for the visualization. Consequently, the corresponding solutions also have to be tested. In total, the paper covers 83 publications, which were found through a systematic literature review. It is therefore the most comprehensive one in the regarded list. The authors identify the most relevant big data quality attributes, illustrate their relations to the big data characteristics, give an overview of the big data quality assurance technologies that are the most frequently used and, as in [45], point out that quality assurance can already take place before the actual implementation. A main result of this work is that novel approaches for the testing in the big data domain are still needed, since there is a significant difference between the testing of traditional software and the testing of big data applications. Moreover, a prevailing necessity for crafting individualized quality assurance solutions is emphasized. Further highlighted challenges include a lack of awareness for and understanding of approaches for big data quality assurance and, in line with the findings of [49], a paucity of high quality tools for its execution.

4 Discussion

Looking at the found publications, a strong focus on benchmarking becomes apparent, whereas the testing of applications is rather underrepresented. However, there seems to be a growing interested in it, respectively a more comprehensive view of quality assurance. At least, when only regarding the timeframe starting in 2017, there would be a parity in the number of publications

found in this review. Even though this is only a small sample and, therefore, not representative, it indicates a growing interest in more comprehensive quality assurance aspects that go beyond pure benchmarking. Furthermore, it is very noticeable that there has been a shift from conference publications to journal articles, which can be seen as an indicator for a growing importance of the topic.

While the answers to SRQ 1, SRQ 2, and SRQ 3 were rather straightforward, SRQ 4 and SRQ 5 are more complex. One of the common themes of the analyzed publications is an emphasis on the individuality of the quality assurance of a given solution. While there are numerous benchmarks, their applicability is always limited to certain use cases. A universally usable benchmark does not exist and, given the challenges for its creation, it can also be assumed that this situation will prevail in the foreseeable future. For the testing, the situation is similar. Even though certain patterns have been identified, the actual implementation is still an individual endeavor. While a need for novel approaches for the testing in the big data domain was explicitly stated, since there is a significant difference between the testing of traditional software and the testing of big data applications, and the benchmarking generally seems to be understood better, it is also far from being able to be considered mature. This lack of maturity was also highlighted across the board as well as the existence of many big challenges that can be mostly attributed to the properties of big data and its particular characteristics. Therefore, the most promising directions for future research, with the most immediate positive effects, seems to be the development of frameworks and approaches for the testing and benchmarking of big data applications that provide a general structure and at the same time grant a huge amount of freedom on the individual level. However, since for now the benchmarking seems to be more intensely researched and there are, while not being universally applicable, at least numerous solutions for individual tools and techniques, the shift in research interest from benchmarking towards more general studies, but also the testing in specific, appears to be justified.

Another repeated theme in the regarded publications was the segmentation of the quality assurance into different levels. This is mentioned the first time in the earliest of the papers [42] and still prevails [49]. Therefore, the creation of tests or benchmarks that act on a solution's component and system level can be seen as a necessity to account for the regarded application's complexity, with at least the testing probably even calling for more granularity as indicated in [49] but also other works such as [18]. The latter, proposing the application of test driven development in big data engineering, also serves as an example for new approaches to the quality assurance in big data, which are being called for, as in [50], to help advance the domain as a whole.

5 Conclusion

With a continuously increasing amount and complexity of data being produced and captured, traditional ways of dealing with their storing, processing, analysis and presentation are no longer sufficient. This led to the emergence of the concept of big data as well as countless tools and techniques for the implementation of corresponding applications. Since its harnessing promises sizeable benefits, organizations all over the world are heavily invested in it. However, when striving for those benefits, it is not only necessary to handle this rather new source of insights, but also to assure that the quality of the developed solutions is high, since otherwise the obtained results might be flawed, which can, in turn, lead to their disregard or even detrimental effects. Yet, the domain of quality assurance in big data engineering is far from mature. To facilitate bridging this gap, this article examined *what is the current state of the art in the domain of quality assurance in big data engineering*. For this purpose, a comprehensive structured meta-literature review on the topic of big data quality assurance was conducted. The results will provide interested researchers and practitioners with a solid foundation for their own quality assurance related endeavors and, therefore, help in advancing the cause of quality assurance in big data as well as the domain of big data in general. Furthermore, based on the findings of the review, worthwhile directions for future research were identified, providing prospective authors with some guidance

in this rather complex environment. Those are mainly the development of high-level frameworks and approaches for the testing and benchmarking of big data applications, research concerning the segmentation of the quality assurance into different levels, the exploration of new approaches to the quality assurance in big data, and a general shift away from somewhat neglecting issues outside of benchmarking towards more general and testing related studies.

However, as for any review article and despite the authors' best efforts, a certain degree of subjectivity will always remain in the decisions concerning the inclusion or exclusion of the evaluated publications as well as in the analysis of those that were accepted into the final set. This might, in turn, influence the obtained results. Additionally, while the chosen approach of conducting a metareview allows for a broader overview compared to the use of primary sources, it, potentially, also reduces the level of detail that is captured and results in a high dependency on the preliminary of others. Even though the respective publications have all been rigorously evaluated through peer review, had to undergo additional screening in the course of this work and have, except for two of the most recent ones, also been recognized in the form of getting cited, this still constitutes a threat to validity that has to be taken into account. However, it also has to be recognized that the publication at hand just constitutes a snapshot in a rapidly developing environment where tasks, requirements, tools, techniques and methods constantly evolve. Therefore, the validness of some of its insights and propositions might be very limited in time. One possibility to track the domains development in the future as well as to further increase the findings' reliability and also enhance it could be to repeat the review in regular intervals and complement its insights with practitioner interviews.

References

[1] C. Dobre and F. Xhafa, "Intelligent services for Big Data science," *Future Generation Computer Systems*, vol. 37, pp. 267–281, 2014. Available: https://doi.org/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. Available: https://doi.org/10.1109/JPROC.2015.2388958

[3] W. L. Chang and N. Grady, *NIST Big Data Interoperability Framework: Volume 1, Definitions.* Available: https://doi.org/10.6028/NIST.SP.1500-1r2

[4] M. Volk, D. Staegemann, S. Bosse, R. Häusler, and K. Turowski, "Approaching the (Big) Data Science Engineering Process," *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security*, Prague, Czech Republic, 2020, pp. 428–435. Available: https://doi.org/10.5220/0009569804280435

[5] J. Vom Brocke, C. Sonnenberg, and A. Simons, "Value-oriented Information Systems Design: The Concept of Potentials Modeling and its Application to Service-oriented Architectures," *Bus. Inf. Syst. Eng.*, vol. 1, no. 3, pp. 223–233, 2009. Available: https://doi.org/10.1007/s12599-009-0046-3

[6] M. Ghasemaghaei and G. Calic, "Can big data improve firm decision quality? The role of data quality and data diagnosticity," *Decision Support Systems*, vol. 120, pp. 38–49, 2019. Available: https://doi.org/10.1016/j.dss.2019.03.008

[7] 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. Available: https://doi.org/10.1080/07421222.2018.1451955

[8] M. Janssen, H. van der Voort, and A. Wahyudi, "Factors influencing big data decision-making quality," *Journal of Business Research*, vol. 70, pp. 338–345, 2017. Available: https://doi.org/10.1016/j.jbusres.2016.08.007

[9] 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. Available: https://doi.org/10.7250/csimq.2020-23.05

[10] M. Volk, D. Staegemann, M. Pohl, and K. Turowski, "Challenging Big Data Engineering: Positioning of Current and Future Development," *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security*, Heraklion, Crete, Greece, 2019, pp. 351–358. Available: https://doi.org/10.5220/0007748803510358

[11] D. Staegemann, M. Volk, A. Nahhas, M. Abdallah, and K. Turowski, "Exploring the Specificities and Challenges of Testing Big Data Systems," *Proceedings of the 15th International Conference on Signal Image Technology & Internet based Systems*, Sorrento, 2019. Available: https://doi.org/10.1109/SITIS.2019.00055

[12] 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. Available: https://doi.org/10.1016/j.jsis.2017.07.003

[13] L. Wang *et al.*, “BigDataBench: A big data benchmark suite from internet services,” *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*, Orlando, FL, USA, 2014, pp. 488–499. Available: https://doi.org/10.1109/HPCA.2014.6835958

[14] A. Alexandrov, C. Brücke, and V. Markl, “Issues in big data testing and benchmarking,” *Sixth International Workshop on Testing Database Systems - DBTest '13*, New York, 2013, pp. 1–5. Available: https://doi.org/10.1145/2479440.2482677

[15] D. G. Tesfagiorgish and L. JunYi, “Big Data Transformation Testing Based on Data Reverse Engineering,” *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)*, Beijing, 2015, pp. 649–652. Available: https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.129

[16] T. Stepanova, A. Pechenkin, and D. Lavrova, “Ontology-based big data approach to automated penetration testing of large-scale heterogeneous systems,” *Proceedings of the 8th International Conference on Security of Information and Networks - SIN '15*, Sochi, Russia, 2015, pp. 142–149. Available: https://doi.org/10.1145/2799979.2799995

[17] Z. Zhang and X. Xie, “Towards testing big data analytics software: the essential role of metamorphic testing,” *Biophysical reviews*, vol. 11, no. 1, pp. 123–125, 2019. Available: https://doi.org/10.1007/s12551-018-0492-6

[18] D. Staegemann, M. Volk, N. Jamous, and K. Turowski, “Exploring the Applicability of Test Driven Development in the Big Data Domain,” *Proceedings of the ACIS 2020*, Wellington, New Zealand, 2020.

[19] S. Ashoke and J. R. Haritsa, “CODD: A dataless approach to big data testing,” *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 2008–2011, 2015. Available: https://doi.org/10.14778/2824032.2824123

[20] J. Vom Brocke, A. Simons, B. Niehaves, K. Reimer, R. Plattfaut, and A. Cleven, “Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process,” *Proceedings of the ECIS 2009*, Verona, Italy, 2009.

[21] A. Parlina, K. Ramli, and H. Murfi, “Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database,” *Information*, vol. 11, no. 2, pp. 69–94, 2020. Available: https://doi.org/10.3390/info11020069

[22] F. X. Diebold, “On the Origin(s) and Development of the Term ‘Big Data’,” PIER Working Paper NO. 12-037, 2012. Available: https://doi.org/10.2139/ssrn.2152421

[23] M. Volk, D. Staegemann, and K. Turowski, “Big Data,” *Springer Reference Wirtschaft, Handbuch Digitale Wirtschaft*, T. Kollmann, Ed., Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 1–18. Available: https://doi.org/10.1007/978-3-658-17345-6_71-1

[24] M. Al-Mekhlal and A. Ali Khwaja, “A Synthesis of Big Data Definition and Characteristics,” *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, New York, NY, USA, 2019, pp. 314–322. Available: https://doi.org/10.1109/CSE/EUC.2019.00067

[25] D. Laney, “3D Data Management: Controlling Data Volume, Velocity, and Variety,” Gartner, 2001.

[26] N. Khan, A. Naim, M. R. Hussain, Q. N. Naveed, N. Ahmad, and S. Qamar, “The 51 V’s of Big Data,” *Proceedings of the International Conference on Omni-Layer Intelligent Systems 2019*, Crete Greece, 2019, pp. 19–24. Available: https://doi.org/10.1145/3312614.3312623

[27] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, “Addressing big data issues in Scientific Data Infrastructure,” *International Conference on Collaboration Technologies and Systems*, San Diego, CA, USA, 2013, pp. 48–55. Available: https://doi.org/10.1109/CTS.2013.6567203

[28] 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. Available: https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[29] D. Wu and Y. Cui, “Disaster early warning and damage assessment analysis using social media data and geo-location information,” *Decision Support Systems*, vol. 111, pp. 48–59, 2018. Available: https://doi.org/10.1016/j.dss.2018.04.005

[30] S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares, “BIG DATA for Healthcare: A Survey,” *IEEE Access*, vol. 7, pp. 7397–7408, 2019. Available: https://doi.org/10.1109/ACCESS.2018.2889180

[31] K. Bronson and I. Knezevic, “Big Data in food and agriculture,” *Big Data & Society*, vol. 3, no. 1, 2016. Available: https://doi.org/10.1177/2053951716648174

[32] K. Nagorny, P. Lima-Monteiro, J. Barata, and A. W. Colombo, “Big Data Analysis in Smart Manufacturing: A Review,” *IJCNS*, vol. 10, no. 03, pp. 31–58, 2017. Available: https://doi.org/10.4236/ijcns.2017.103003

[33] F. R. Goes *et al.*, “Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review,” *European Journal of Sport Science*, pp. 1–16, 2020. Available: https://doi.org/10.1080/17461391.2020.1747552

[34] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, “Applications of big data to smart cities,” *J Internet Serv Appl*, vol. 6, no. 1, 2015. Available: https://doi.org/10.1186/s13174-015-0041-5

[35] K. Vassakis, E. Petrakis, and I. Kopanakis, “Big Data Analytics: Applications, Prospects and Challenges,” *Lecture Notes on Data Engineering and Communications Technologies, Mobile Big Data*, G. Skourletopoulos, G. Mastorakis, C. X. Mavromoustakis, C. Dobre, and E. Pallis, Eds., Cham: Springer International Publishing, 2018, pp. 3–20. Available: https://doi.org/10.1007/978-3-319-67925-9_1

[36] M. Abdallah, M. Muhairat, A. Althunibat, and A. Abdalla, “Big Data Quality: Factors, Frameworks, and Challenges,” *Compusoft: An International Journal of Advanced Computer Technology*, vol. 9, no. 8, pp. 3785–3790, 2020.

[37] D. Staegemann, M. Volk, N. Jamous, and K. Turowski, “Understanding Issues in Big Data Applications – A Multidimensional Endeavor,” *Proceedings of the Twenty-fifth Americas Conference on Information Systems*, Cancun, Mexico, 2019.

[38] L. Cai and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” *CODATA*, vol. 14, 2015. Available: https://doi.org/10.5334/dsj-2015-002

[39] J. Webster and R. T. Watson, “Analyzing the Past to Prepare for the Future: Writing a Literature Review,” *MISQ*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[40] Y. Levy and T. J. Ellis, “A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research,” *Informing Science: The International Journal of an Emerging Transdiscipline*, vol. 9, pp. 181–212, 2006. Available: https://doi.org/10.28945/479

[41] C. Okoli, “A Guide to Conducting a Standalone Systematic Literature Review,” *CAIS*, vol. 37, pp. 879–910, 2015. Available: https://doi.org/10.17705/1CAIS.03743

[42] X. Qin and X. Zhou, “A Survey on Benchmarks for Big Data and Some More Considerations,” *Lecture Notes in Computer Science, Intelligent Data Engineering and Automated Learning – IDEAL 2013*, D. Hutchison et al., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 619–627. Available: https://doi.org/10.1007/978-3-642-41278-3_75

[43] R. Han, X. Lu, and J. Xu, “On Big Data Benchmarking,” *Lecture Notes in Computer Science, Big Data Benchmarks, Performance Optimization, and Emerging Hardware*, J. Zhan, R. Han, and C. Weng, Eds., Cham: Springer International Publishing, 2014, pp. 3–18. Available: https://doi.org/10.1007/978-3-319-13021-7_1

[44] T. Ivanov *et al.*, “Big Data Benchmark Compendium,” *Lecture Notes in Computer Science, Performance Evaluation and Benchmarking: Traditional to Big Data to Internet of Things*, R. Nambiar and M. Poess, Eds., Cham: Springer International Publishing, 2016, pp. 135–155. Available: https://doi.org/10.1007/978-3-319-31409-9_9

[45] P. Zhang, X. Zhou, W. Li, and J. Gao, “A Survey on Quality Assurance Techniques for Big Data Applications,” *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, Redwood City, CA, USA, 2017, pp. 313–319. Available: https://doi.org/10.1109/BigDataService.2017.42

[46] M. I. Ortega, M. Genero, and M. Piattini, “Big Data DBMS Assessment: A Systematic Mapping Study,” *Lecture Notes in Computer Science, Model and Data Engineering*, Y. Ouhammou, M. Ivanovic, A. Abelló, and L. Bellatreche, Eds., Cham: Springer International Publishing, 2017, pp. 96–110. Available: https://doi.org/10.1007/978-3-319-66854-3_8

[47] R. Han, L. K. John, and J. Zhan, “Benchmarking Big Data Systems: A Review,” *IEEE Trans. Serv. Comput.*, vol. 11, no. 3, pp. 580–597, 2018. Available: https://doi.org/10.1109/TSC.2017.2730882

[48] F. Bajaber, S. Sakr, O. Batarfi, A. Altalhi, and A. Barnawi, “Benchmarking big data systems: A survey,” *Computer Communications*, vol. 149, pp. 241–251, 2020. Available: https://doi.org/10.1016/j.comcom.2019.10.002

[49] A. Davoudian and M. Liu, “Big Data Systems,” *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–39, 2020. Available: https://doi.org/10.1145/3408314

[50] S. Ji, Q. Li, W. Cao, P. Zhang, and H. Muccini, “Quality Assurance Technologies of Big Data Applications: A Systematic Literature Review,” *Applied Sciences*, vol. 10, no. 22, article 8052, 2020. Available: https://doi.org/10.3390/app10228052