# Big Data DBMS Assessment: A Systematic Mapping Study

Maria Isabel Ortega(✉), Marcela Genero, and Mario Piattini

Institute of Technology and Information Systems,
University of Castilla-La Mancha, Ciudad Real, Spain
misabel.ortega2@alu.uclm.es,
{marcela.genero,mario.piattini}@uclm.es

**Abstract.** The tremendous prosperity of big data systems that has occurred in recent years has made its understanding crucial for both research and industrial communities. Big Data is expected to generate an economy of 15 billion euros over the next few years and to have repercussions that will more or less directly change the way in which we live. It is, therefore, important for organizations to have quality Database Management Systems (DBMSs) that will allow them to manage large volumes of data in real time and according to their needs. The last decade has witnessed an explosion of new Database Management Systems (DBMSs) which deal not only with relational Data Bases but also with non-relational Data Bases. Companies need to assess DBMS quality in order, for example, to select which DBMS is most appropriate for their needs. The main research question formulated in this research is, therefore, *"What is the state of the art of Big Data DBMS assessment?"*, which we attempt to answer by following a well-known methodology called "Systematic Mapping Studies" (SMS). This paper describes an SMS of papers published until May 2016. Five digital libraries were searched, and 19 papers were identified and classified into five dimensions: quality characteristics of Big Data DBMSs, techniques and measures used to assess the quality characteristics, DBMSs whose quality has been measured, evolution over time and research methods utilized. The results indicate that there are several benchmarks, which are principally focused on the performance of MongoDB and Cassandra, and that the interest in Big Data DBMS quality is growing. Nonetheless, more research is needed in order to define and validate a quality model that will bring together all the relevant characteristics of DBMSs for Big Data and their respective measures. This quality model will then be employed as a basis on which to build benchmarks for DBMSs, covering not only the diversity of DBMSs and application scenarios and types of applications, but also diverse and representative real-world data sets.

**Keywords:** Big data · DBMS · Quality · Benchmark

## 1 Introduction

The technological advances we have been experiencing in recent years, such as cloud computing, the Internet of Things and social networks, have led to a continuous increase in data, which are accumulating at an unprecedented rate. The term Big Data


Y. Ouhammou et al. (Eds.): MEDI 2017, LNCS 10563, pp. 96–110, 2017.
DOI: 10.1007/978-3-319-66854-3_8

was coined to represent the large amount and many types of digital data, including documents, images, videos, audio and websites. All of the aforementioned technologies were the forerunners to the arrival of what has been called the Big Data era.

Big Data is expected to generate an economy of 15 billion euros over the next few years, and it will have very many repercussions that will change the way in which we live to a greater or lesser extent. This future, which is so impressive as regards numbers and seems so promising, signifies that the appearance of Big Data has attracted the attention of industry, academia and governments.

In fact, the McKinsey Global Institute [1] estimated that data volume was growing by 40% per year, and would grow to 44 times its initial size between 2009 and 2020. However, the volume of data is not the only important characteristic. Most of the tech industry follows Gartner's '3Vs' (Volume, Velocity and Variety) model to define Big Data [3], and Dijcks [2] recently added one more characteristic to this model: Value. Many other authors also propose that Veracity should be considered.

It is, therefore, important for organizations to have quality Database Management Systems (DBMSs) that will allow them to manage large volumes of data in real time and according to their needs.

For all of the above reasons, the last decade has witnessed an explosion of new Database Management Systems (DBMSs) which deal not only with relational data bases but also with non-relational data bases (NoSQL databases). And companies need to assess quality characteristics for the current and emerging DBMSs in order, for example, to compare which is more appropriate according to the actual needs.

The main research question formulated in this research is, therefore, *"What is the state of the art of Big Data DBMS assessment?"* which we attempt to answer by following a well-known methodology called "Systematic Mapping Studies" (SMS). A systematic mapping study provides an objective procedure with which to identify the nature and extent of the research that is available to answer a particular research question. These kinds of studies also help to identify gaps in current research in order to suggest areas for further investigation. They therefore also provide a framework and background in which to appropriately develop future research activities [1].

The remainder of the paper is organized as follows. Section 2 presents a brief discussion of related work. This is followed by an outline of the SMS and a description of the activities of the SMS process in Sect. 3. Section 4 presents the complete procedure followed to develop the SMS, whilst the main results obtained are presented in Sect. 5. The paper concludes with a discussion of the results and outlines future work.

## 2 Related Work

To the best of our knowledge, the relevant literature contains no systematic literature reviews (SLR) or SMS that tackle Big Data DBMS quality. It is, however, true that there are some works whose aim is to provide the state of art regarding different issues related to Big Data:

- Mathisen et al. [2] presents a systematic mapping review that provides an overview of empirical papers dealing with Big Data and categorizes them according to the

3 V's. These authors conclude that no systematic review of empirical work has been carried out to date in the field of Big Data.

- Ruixan [3] presents a bibliometrical analysis of the Big Data research in China. They conclude that research based on Big Data now has an outline, although most papers that present the theoretical step of the research lack sufficient practical sustenance, and they consequently recommend intensifying efforts based on both theory and practice.
- Jeong and Ghani [4] carried out a review of semantic technologies for Big Data, concluding that their analysis shows that there is a need to put more effort into suggesting new approaches. They also note that tools need to be created with the purpose of encouraging researchers and practitioners to realize the true power of semantic computing and support them as regards solving the crucial issues of Big Data.
- Wang and Krishnan [5] present a review whose aim is to provide an overview of the characteristics of clinical Big Data. They describe some commonly employed computational algorithms, statistical methods, and software tool kits for data manipulation and analysis, and discuss the challenges and limitations in this field.
- Polato et al. [6] conducted a systematic literature review to assess research contributions to Apache Hadoop. The objective was to detect possible gaps, providing motivation for fresh research, and outline collaborations with Apache Hadoop and its environment, categorizing and quantifying the central topics dealt with in literature.
- Hashem et al. [7] assessed the rise of Big Data in cloud computing, studying research challenges focused on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Finally, they provided an overview of open research topics that require substantial research efforts.

The literature review presented in this paper is different from those mentioned above in that it tackles Big Data DBMS quality, which has not been researched to date. Moreover, this literature review has been carried out in a systematic and rigorous manner, following the guidelines provided in [8, 13].

## 3 SMS Outline

A systematic mapping study consists of three activities: planning, execution, and reporting [8]. Each of these activities is divided into several steps. The first step when developing an SMS is the definition of the review protocol, which establishes a controlled procedure with which to conduct the review. The execution activity includes data retrieval, study selection, data extraction, and data synthesis. Finally, the reporting activity presents and interprets the results.

### 3.1 Planning the Review

The aim of this SMS is to gather all existing proposals regarding the assessment of the quality characteristics of DBMSs for Big Data. To this end, the following research question was formulated:

*"What is the state of the art of the Big Data DBMS assessment?"*

As this question is too broad to answer, we have split it into five research questions, which are shown in Table 1.

**Table 1.** Research questions

| Research questions | Main motivation |
|---|---|
| RQ1. Which quality characteristics of Big Data DBMSs have been investigated by researchers? | To identify the quality characteristics of DBMSs with which to manage Big Data that have been addressed by researchers, and map them onto the quality characteristics proposed in ISO/IEC 25010 [9] |
| RQ2. Which techniques and quality measures are used to assess the quality characteristics? | To identify which quality assurance techniques for Big Data DBMSs have been used and which measures have been proposed to assess the quality characteristics of Big Data DBMSs |
| RQ3. Which DBMSs have been evaluated by researchers and how is the data represented in them? | To identify which DBMSs have been evaluated and what kind of data representation is used |
| RQ4. How has the research into the quality of Big Data DBMSs evolved over time? | To discover the importance that has been placed on empirical studies on the topic of Big Data DBMS quality over time |
| RQ5. What research methods have been used to investigate the quality of Big Data DBMSs? | To determine whether or not the research has been validated. Also, to discover which research method was used to validate it |

### 3.2 Search Strategy

The research question was decomposed into individual elements related to the technology (technology acceptance model), the study type (evaluation) and the response measure (correlation with actual effort) used, in order to obtain the main search terms. Secondly, key words obtained from known primary studies were assessed in order to obtain other main terms. Synonyms for the main terms were then identified. Finally, the search string was constructed using the Boolean "AND" to join the main terms and the Boolean "OR" to include synonyms. This process enabled the main search terms and alternative terms (spellings, synonyms and terms related to the major terms) to be defined, as is shown in Table 2.

The final search string was: "("Database Management System" OR DBMS OR Warehouse OR "Data system") AND (evaluat* OR measur* OR assess* OR test* OR

**Table 2.** Search string terms

| Main terms | Alternative terms |
|---|---|
| Database management system | ("Database management system" OR DBMS OR warehouse OR "data system") |
| Evaluate | (evaluat* OR measur* OR assess* OR test* OR analys* OR select* OR compar* OR adquisi* OR implement* OR benchmark) |
| Big data | ("Big Data" OR "NewSQL" OR "No SQL" OR NoSQL) |

analys* OR select* OR compar* OR adquisi* OR implement* OR benchmark) AND ("Big Data" OR "New SQL" OR "No SQL" OR NoSQL))".

The search was performed in digital libraries that contain a wide variety of computer science journals. The search was specifically performed in Scopus database, Science@Direct, IEEE Digital Library, Springer database and ACM Digital Library. As we wished to guarantee the reliability of the elements that would be studied, we analyzed only journal papers, workshop papers and conference papers. Table 3 summarizes the search strategy defined.

**Table 3.** Search strategy

| | |
|---|---|
| Databases | Scopus<br>Science@Direct (subject computer science)<br>IEEE digital library<br>ACM digital library<br>Springer database |
| Target items | Journal papers<br>Workshop papers<br>Conference papers |
| Search applied to | Title<br>Abstract<br>Keywords |
| Language | Papers written in English |
| Publication period | Until May 2016 (inclusive) |

### 3.3 Selection Criteria and Procedure

The intention of this SMS was to discover all papers that present any research related to Big Data DBMS Quality, that are written in English and have been published until May 2016. The start of the publication period was not established because we wished to discover since when Big Data DBMS quality proposals have existed. Papers were excluded according to the selection criteria shown in Table 4.

The study selection procedure was executed with the final string defined above, and was conducted in two stages. In the first stage, the selection of the studies was executed by reviewing the title, the abstract and the keywords of the studies; only those papers that dealt with Big Data DBMS quality were selected. The set of papers selected in the first stage was used as the basis for the second stage, which consisted of reading the full texts of these papers and applying the inclusion and exclusion criteria.

**Table 4.** Inclusion and exclusion criteria

| | |
|---|---|
| Inclusion criteria | Journals, conferences and workshop papers<br>Papers written in English<br>Papers published until May 2016 (inclusive) |
| Exclusion criteria | Papers not focusing on DBMS quality<br>Papers focusing on data quality<br>Papers available only in the form of abstracts or PowerPoint presentations<br>Duplicate papers (the same paper in different databases)<br>Papers in which Big Data DBMS quality is mentioned only as a general introductory term, or in which there are no proposals related to quality among the paper′s contributions |

### 3.4 Data Extraction and Synthesis Procedure

A set of five dimensions was used to classify the research, based on the research questions described above. This classification scheme was developed prior to the first round of data extraction and was subsequently refined after the pilot data had been extracted and analyzed. The possible categories are based on the results found during the review. A summary of the classification scheme is presented in Table 5. The detailed classification scheme is available at http://alarcos.esi.uclm.es/DBMS-BigData-Quality.

**Table 5.** Summary of the classification scheme

| Dimensions | Categories |
|---|---|
| Quality characteristic | Product quality in use model: efficiency<br>Product quality model: performance efficiency, adaptability, availability and usability |
| Techniques and measures | YCSB, YCSB ++, LUBM, TPCX-HS, BigDataBench and Others |
| DBMS | MongoDB, Cassandra, Riak, HBase, Neo4j, Hadoop, Redis, CouchDB, MySQL, Phoenix, Spark, Hive, Pig, Oracle and DB2 |
| Time evolution | The year of the publication |
| Research method | Proposal, evaluation, validation, philosophical, opinion or personal experience [10] |

## 4 Conducting the Review

The SMS was carried out by following all the steps of the protocol defined previously. Nonetheless, as the definition of the protocol is iterative, we have made some modifications to it during the execution. The version of the protocol presented in the previous section is the final one.

The SMS was completed in 9 months, and this period included the time needed for planning, conducting and reporting. 957 papers were initially founded. We found 430 studies in Scopus, 382 studies in ACM, 3 studies in Science Direct and 142 studies in IEEE. No studies were found in Springer.

After applying the inclusion and exclusion criteria and reviewing the title and abstract of each paper, the number of papers selected was reduced to 86. As will be observed, we selected 58 studies form Scopus, 19 studies form ACM, 2 studies form Science Direct and 7 studies form IEEE.

17 papers were also subsequently excluded because they were duplicated (the same paper in a different database). As is shown in Fig. 1, we removed 15 studies from Scopus and 2 studies from ACM.

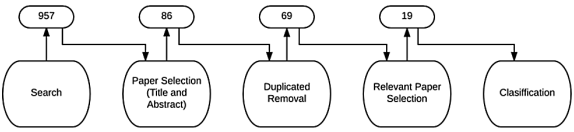


**Fig. 1.** Selection process

Inclusion and exclusion criteria were applied to the full text and 40 more papers were discarded. The final 19 papers were analyzed and their results were synthesized and interpreted. Figure 1 shows the selection process employed. The list of the primary studies selected is available at http://alarcos.esi.uclm.es/DBMS-BigData-Quality.

Table 6 summarizes the chronology of activities rigorously performed to carry out the SMS. The identification and selection of studies took place between February 2016 and November 2016. This period included the protocol refinement.

**Table 6.** Review outline

| Chronology | Step | Activities | Outcome |
|---|---|---|---|
| March 2016 | Planning | Protocol development | Reviewed protocol |
| May 2016 | Conducting | Data retrieval | Metadata information of 957 papers |
| | | Paper selection (title and abstract) | Metadata information of 86 papers selected |
| | | Removal of duplicates | Metadata information of 69 papers selected |
| | | Extraction of files of the papers | Repository of papers (69 papers) |
| July 2016 | Planning | Protocol improvement<br>Pilot data extraction | Data extraction form (classification scheme refined), 69 papers reviewed |
| August 2016 | Conducting | Paper selection, classification (full text) | Data extraction form complete, 19 papers classified |
| | | Data synthesis | |
| November 2016 | Reporting | Report on the stages and activities undertaken during the development of the SMS | Final report of the SMS |

## 5 Reporting Results and Data Synthesis

In this section, the answers to each of the questions formulated in Sect. 3 are presented and interpreted, in addition to which the dimensions covered by the questions are combined.

### 5.1 RQ1. Which Quality Characteristics of Big Data DBMSs Have Been Investigated by Researchers?

The process used to match the characteristics in the ISO/IEC 25010 standard [9] with the characteristics investigated in the paper is described as follows. The full text of the paper was read in order to search for quality characteristics, and we then looked at the standard for the characteristics that best matched the characteristics found in the paper. In the review of the full text of the selected papers, it was found that in the majority the authors used several terms to refer to the quality characteristics being researched. These terms were analyzed until the characteristics that best fitted them was found in the standard.

The results obtained for RQ1 revealed that most of the papers selected addressed only one quality characteristic or sub-characteristic. We found that the quality model most frequently investigated is the product quality model. The characteristics of the quality product model most frequently researched were performance efficiency (89.47%), distantly followed by usability (10.53%) the adaptability sub-characteristic. The reliability was most frequently researched through the use of the availability sub-characteristic, with the same amount of appearances as those of the usability characteristic (5.26%). Table 7 shows which paper evaluates each characteristic.

**Table 7.** Distribution of papers per characteristics of the ISO 25010 product quality model

| Characteristic | Reference |
|---|---|
| Performance efficiency | [P01] [P02] [P03] [P04] [P05] [P06] [P08] [P09] [P10] [P11] [P13] [P14] [P15] [P16] [P17] [P18] [P19] |
| Usability | [P12] |
| Adaptability | [P06] [P07] |
| Availability | [P09] |

We also found that only one article researched quality in use characteristics (efficiency (5.26%)). Table 8 shows which characteristic(s) or sub-characteristic(s) are evaluating each paper.

**Table 8.** Distribution of papers per characteristics of the ISO 25010 quality in use model

| Characteristic | Reference |
|---|---|
| Efficiency | [P09] |

These results could be explained by the fact that researchers are principally concerned with the rapid treatment of large volumes of data with the purpose of obtaining the value of the data, which is why they might research performance efficiency. Surprisingly, the security characteristic, which is usually crucial when selecting a Big Data DBMS in order to assess the privacy and integrity of the information, has not been addressed as regards the security of Big Data DBMSs.

### 5.2 RQ2. Which Techniques and Quality Measures Are Used in Order to Assess the Quality Characteristics?

Benchmarking provides us with the possibility of evaluating quality characteristics by comparing them with a standard. Various standards have been imposed in order to measure the quality of Big Data DBMSs, among others, and particularly to measure the performance of DBMSs. In conceptual terms, a big data benchmark aims to generate application-specific workloads and tests capable of processing big data with the 5 V properties (volume, velocity, variety, value and veracity) [11] in order to produce meaningful evaluation results [12].

The SMS revealed that most of the benchmarks that have been carried out are proposal of benchmarks (53.63%). These are followed by the Yahoo! Cloud Serving Benchmark (YCSB) at 31.58%, which is very distantly followed and with the same result of utilization by YCSB++ (5.26%), LUBM benchmark (5.26%), TPCx-HS (5.26%) and BigDataBench (5.26%). The results show the lack of consensus as regards the use of a benchmark when the intention is to ensure Big Data DBMS quality. At this point, the importance of achieving standardization is tangible, in order to ensure that all systems are measured with the same established criteria that will facilitate their comparison (Table 9).

**Table 9.** Metrics and techniques per primary studies

| Reference | Technique | Metrics |
|---|---|---|
| [P01] | YCSB | *Latency*: relating time spend with the number of operations per second |
| [P02] | YCSB | *Latency*: relating time spend with the number of operations per second |
| [P03] | Proposal of benchmark | *Resource Utilization*: memory, CPU utilization, Garbage Collection (GC) statistics, heap memory usage, IO wait, disk read and write throughput, disk usage, OS load, etc.<br>*Datastore*: Read and write throughput, pending read and write requests count, read and write latency, compactions completed, pending compactions, etc. |
| [P04] | LUMB benchmark | *vertical joins*: Cost (q, sdb) = \|T\|, where \|T\| is the number of pages in the table T. If an index is defined in the triple table, cost (q, sdb) = P (index) + sel(t) ∗ \|T\|, |

(*continued*)

**Table 9.** (*continued*)

| Reference | Technique | Metrics |
|---|---|---|
| | | where P(index) is the cost of index scanning and sel(t) is the selectivity of the triple pattern t as defined in [13]<br>*binary joins*: the selection is made in the property tables. Cost q, sdb) = \|T p\| where Tp is the property table of the property of the query triple pattern. With an index on the selection predicate, cost(q,sdb) = P(index) + sel∗\|Tp\|, where sel is the selectivity of the index<br>*horizontal joins*: the selection targets the tables of the class domain of the property of the query triple pattern. Cost(q,sdb) = Tcp∈dom(p)(\|Tcp\|), where Tcp are the tables corresponding to the classes domain of the property of the query triple pattern. If there is an index defined in the selection predicate, cost(q, sdb) = T cp∈dom(p)(P (index) + sel ∗ \|T cp\|) where sel is the index selectivity |
| [P05] | Proposal of benchmark benchmark | *general statistics (STATS)*: the algorithm counts the numbers of vertices and edges in the graph and computes the mean local clustering coefficient<br>*breadth-first search (BFS)*: the algorithm traverses the graph starting from a seed vertex, and first visits all the neighbors of a vertex before moving to the neighbors of the neighbors<br>*connected components (CONN)*: for each vertex, the algorithm determines the connected component it belongs to<br>*community detection (CD)*: the algorithm detects groups of nodes that are more strongly connected to each other than they are connected to the rest of the graph<br>*graph evolution (EVO)*: the algorithm predicts the evolution of the graph according to the "forest fire" model |
| [P06] | YCSB and YCSB++ | Not specified |
| [P07] | Proposal of benchmark | *Load balancing* |
| [P08] | Proposal of benchmark | Not specified |
| [P09] | TPCx-HS | *Performance (HSph@SF):* the effective sort throughput of the benchmarked configuration:<br>•HSph@SF = SF /(T /3600)<br>Where:<br>•SF is the Scale Factor<br>•T is the total elapsed time for the run-in seconds<br>*Price-performance metric:*<br>•$/HSph@SF = P HSph @ SF |

(*continued*)

**Table 9.** (*continued*)

| Reference | Technique | Metrics |
|---|---|---|
| | | Where:<br>•P is the total cost of ownership of the system being tested<br>*System Availability Date:* when the benchmarked systems are generally available to any customer<br>*TPCx-HS Energy Metrics:* expected to be accurate representations of system performance and energy consumption. The approach and methodology are explicitly detailed in this specification and the TPC Benchmark Standards, as defined in TPC- Energy |
| [P10] | YCSB | *Speed limit on a single node:* workload operations which consist of update heavy, read heavy, read only, read latest, short ranges and read-modify-write<br>*Latency*: relating time spent with the number of operations per second<br>*Workloads*: workload operations which consist of 95% of read and 5% of update sent by each client on non-master nodes |
| [P11] | YCSB | *Latency*: relating time spent with the number of operations per second |
| [P12] | YCSB | *Latency*: relating time spent with the number of operations per second |
| [P13] | Proposal of benchmark | *Latency*: relating time spent with the number of operations per second |
| [P14] | Proposal of benchmark | *RPS in short*: the number of processed requests per second<br>Latency: relating time spent with the number of operations per second<br>*OPS in short*: number of operations per second<br>*DPS in short*: data processed per second<br>*MIPS*: million instructions per second<br>*MPKI*: MIS-Predictions per 1000 Instructions (branch prediction) |
| [P15] | BigDataBench | Not specified |
| [P16] | Proposal of benchmark | *Latency*: relating time spent with the number of operations per second |
| [P17] | Proposal of benchmark | Not specified |
| [P18] | Proposal of benchmark | Not specified |
| [P19] | Proposal of benchmark | *Query response time, tuning overhead, data arrival to query time, storage size* and *monetary cost* |

### 5.3 RQ3. Which DBMSs Have Been Evaluated by Researchers and How Is the Data Represented in Them?

The results show that Cassandra (36%) and MongoDB (31%) stand out as the dominant DBMSs. They are followed by Hadoop (26%) and HBase (21%). There are other DMBSs that make a medium number of appearances in the papers, such as MySQL (15%), Riak (15%), Hive (10%), Redis (10%) and Neo4j (10%). The DBMSs which appear the least are Pig (5%), Spark (5%), Phoenix (5%), CouchDB (5%), Google (5%), Graph (5%), Giraph (5%), DB2 (5%) and Oracle (5%) (Fig 2).

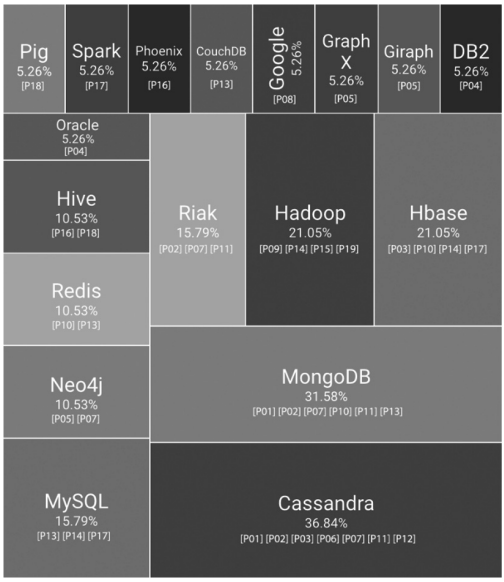


**Fig. 2.** Percentage of DBMSs evaluated in the primary studies

At this point, we should highlight the lack of maturity of the systems, thus making the use of larger and more complex systems such as Apache Hadoop, unnecessary. It is also noteworthy that more conventional DBMSs, such as Oracle or DB2, appear to be falling behind and giving way to new systems as a first alternative.

### 5.4 RQ4. How Has the Research into the Quality of Big Data DBMSs Research Evolved Over Time?

The question shows the apparent evolution of the quality of Big Data DBMS over time. It can be observed that it has been rising. This may be owing to the ever-growing weight of Big Data systems in our society and therefore to the importance of their efficiency and reliability. In the year 2016, only 1 item has been found concerning the quality of Big Data systems. This is probably because the review was finalized in May of that year and many of the referenced articles had not yet been published.

### 5.5 RQ5. What Research Methods Have Been Used to Investigate the Quality of Big Data DBMSs?

This question was answered by using the classification of research approaches proposed by Wieringa et al. [10], as recommended in Petersen et al. [14]. The scheme also presents the classification of non-empirical research, which contains the categories of proposal papers, evaluation papers, validation papers, philosophical papers, opinion papers and personal experience papers. The results showed that proposal (42%) stood out as the dominant research method. The second most common research method used was evaluation (32%); in third place was validation (16%), and finally in last place was opinion (1.0%) (Table 10).

**Table 10.** Distribution of papers per characteristics by research method

| Characteristic | Reference |
|---|---|
| Proposal | [P05] [P06] [P07] [P08] [P10] [P12] [P13] [P18] |
| Evaluation | [P01] [P02] [P03] [P04] [P11] [P15] |
| Validation | [P14] [P16] [P19] |
| Opinion | [P09] [P17] |

The results of this classification show that almost half of the primary studies are proposals or evaluations in laboratory contexts, and it is therefore evident that more validation is needed in industrial settings.

### 5.6 Combining Several RQs and Additional Information Extracted

Figure 3 shows the combination of the quality characteristics evaluated in the SMS, the quality characteristic, the DBMS, the research method and the techniques. The aim of this section is to show the evidence regarding SG quality found in this SLM, combining some research questions with additional information extracted from primary studies. This figure shows that, of the 19 studies analyzed:

- 9 of the primary studies focused on evaluating the performance efficiency and none of them used any of the existing standard benchmarks.
- The DBMSs which have been most frequently used to evaluate any of the quality characteristics are: MongoDB, Cassandra and Hadoop. Moreover, these DBMSs are principally used to ensure the performance efficiency of the DBMSs.
- The most frequently evaluated and modified benchmark is the YCBS benchmark. However, most of the primary studies are proposals and additionally proposed new techniques or benchmarks with which to assess the quality of Big Data DBMSs.

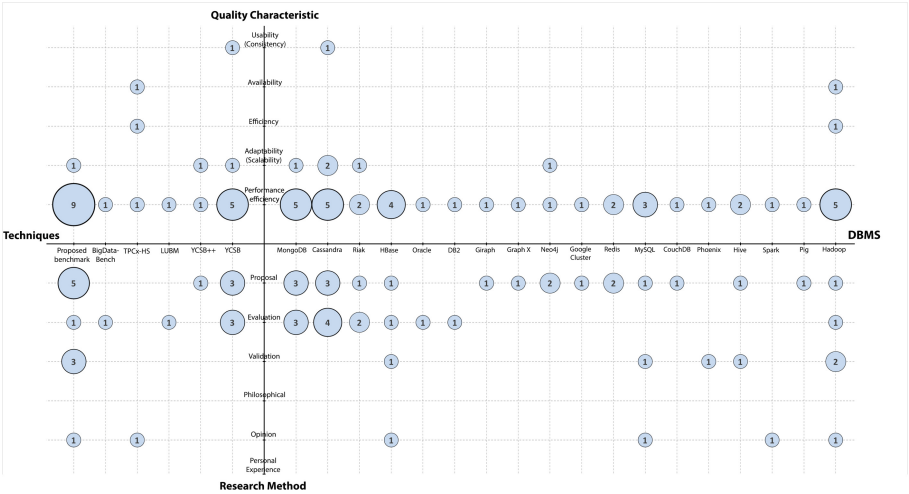


**Fig. 3.** Combination of quality characteristic, DBMSs, research method and techniques

## 6 Conclusions

Several efforts have been made in recent years to assess Big Data DBMS quality, but further work is needed. In this work we have, therefore, identified the different proposals regarding Big Data DBMS quality, in an attempt to answer the questions raised based on five facets: the quality characteristic investigated (Q1), the techniques and metrics used to assess the quality characteristics (Q2), the DMBSs used (Q3), the evolution of quality research over time (Q4) and the research method (Q5).

The results of the systematic mapping study presented in the previous sections allow us to state that there is an increasing interest in Big Data DBMS quality assessment. However, much still needs to be done. Thus, as future work we shall continue to advance in this line of work. We shall define and validate a quality model for Big Data DBMSs, integrating different exiting proposals, while in the long term we intend to build benchmarks based on the quality model that will cover not only the diversity of DBMSs and application scenarios and types of applications, but also diverse and representative real-world data sets.

**Acknowledgements.** This work has been funded by the SEQUOIA project (Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional FEDER, TIN2015-63502-C3-1-R).

## References

1. Budgen, D., Turner, M., Brereton, P., Kitchenham, B.: Using mapping studies in software engineering. In: Proceedings of PPIG, pp. 195–204 (2008)

2. Mathisen, B.M., Wienhofen, L.W.M., Roman, D.: Empirical big data research: a systematic literature mapping. Journal of ArXiv preprint arXiv:1509.03045 (2015)
3. Yang, R.: Bibliometrical analysis on the big data research in China. J. Digit. Inf. Manag. **11**, 383–390 (2013)
4. Jeong, S.R., Ghani, I.: Semantic computing for big data: approaches, tools, and emerging directions (2011–2014). ACM TIIS **8**, 2022–2042 (2014)
5. Wang, W., Krishnan, E.: Big data and clinicians: a review on the state of the science. Proc. JMIR Med. Inform. **2**, e1 (2014)
6. Polato, I., Ré, R., Goldman, A., Kon, F.: A comprehensive view of Hadoop research—a systematic literature review. J. Netw. Comput. Appl. **46**, 1–25 (2014)
7. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of "big data" on cloud computing: review and open research issues. J. Inf. Syst. **47**, 98–115 (2015)
8. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Technical report, EBSE Technical report EBSE-2007-012007
9. ISO/IEC: ISO/IEC 25010 - Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models (2010)
10. Wieringa, R., Maiden, N., Mead, N., Rolland, C.: Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. J. Requirements Eng. **11**, 102–107 (2006)
11. IBM: IBM big data platform (2016). http://www-01.ibm.com/software/data/bigdata/
12. Tay, Y.: Data generation for application-specific benchmarking. J. VLDB Challenges Vis. (2011)
13. Stocker, M., Seaborne, A., Bernstein, A., Kiefer, C., Reynolds, D.: SPARQL basic graph pattern optimization using selectivity estimation. In: Proceedings of the 17th International Conference on World Wide Web, pp. 595–604 (2008)
14. Petersen, K., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: an update. J. Inf. Softw. Technol. **64**, 1–18 (2015)