A survey on quality assurance techniques for big data applications

Pengcheng Zhang[1], Xuewu Zhou[1], Wenrui Li[2], Jerry Gao[3,4]
[1]College of Computer and Information, Hohai University, Nanjing, China
[2]School of Mathematics & Information Technology, Nanjing Xiaozhuang University, Nanjing, P.R. China
[3]San Jose State University, San Jose, CA, [4]Taiyuan University of Technology, China
Email Address: {pchzhang@hhu.edu.cn; jerry.gao@sjsu.edu}

***Abstract*—** With the rapid advance of big data and cloud computing, building high quality big data systems in different application fields has gradually became a popular research topic in academia and industry as well as government agencies. However, more quality problems lead to application errors. Although the current research work has discussed how to ensure the quality of big data applications from several aspects, there is no systematic discussion on how to ensure the quality of large data applications. Therefore, a systematic study on big data application quality assurance is very necessary and critical. This paper focuses on the survey of quality assurance techniques of big data applications, and it introduces big data properties and quality attributes. It mainly discusses the key approaches to ensure the quality of big data applications and they are testing, model-driven architecture (MDA), monitoring, fault tolerance, verification and also prediction techniques. In addition, this paper also discusses the impact of big data characteristics on big data applications.

Index Terms*— *Quality Assurance, Big data, Big data application, MDA, Testing, Verification, Fault tolerance, Monitoring, Prediction

I. INTRODUCTION

According to IDC report, the Big Data technology market will grow at "a 27% compound annual growth rate (CAGR) to $32.4 billion through 2017" [1]. It shows that large-scale data computing and big data application services become more and more popular and have more influences on people's daily lives. Big data applications are now widely used in many aspects, such as *monitoring systems*, *forecasting*, and *statistical reporting* applications. However, big data applications pose new challenges for Quality Assurance (QA) engineers due to the large big data characteristics (e.g., velocity of arriving data, volume of data) [2], [3]. For examples, because of the volume and timeliness of the data, verification the accuracy of big data prediction systems is a difficult task, and it is a hard job to validate the correctness of a big data prediction system due to the large scale data size and the feature of timeliness. Therefore, quality assurance techniques for big data applications become a key concern and research topic. Although there are many published papers addressing data quality assurance in the past, a few of them focused on the systematic study on the quality assurance techniques for big data applications. Towards this research direction, the main purpose of this paper is to investigate literature relevant for the quality assurance techniques for big data applications so that it can provide a comprehensive reference to the challenges of quality assurance approaches for big data applications.

Unlike existing work, this paper provides the contributions in the following aspects:

- It discusses quality assurance approaches for big data applications, mainly from the six aspects: testing, model-driven architecture (MDA), monitoring, fault tolerance, verification and prediction for big data applications.
- It also combines quality assurance techniques with big data characteristics while it considers the quality assurance of big data applications, and it explores the big data 4V properties of existing quality assurance techniques for big data application.

The rest of the paper is organized as follows. Section II reviews related work. Section III introduces the different types of big data applications, and the quality assurance approaches. Section IV provides an overview and comparison of the existing approaches for quality assurance of big data applications, specifically in testing, model-driven architecture (MDA), monitoring, fault tolerance, verification and prediction. Section V discussed big data 4V properties and the quality assurance of big data applications. Section VI concludes the paper.

II. RELATED SURVEY

Many scholars have investigated the analysis of big data quality assurance. Let us consider the most interesting approaches from our point of view results obtained by them. Because of the widespread use of big data applications, big data quality assurance research has been tried by scholars. However, due to the huge volume of generated data, the fast velocity of arriving data, and the large variety of heterogeneous data, the quality of data is far from perfect [4] . Therefore, big data quality assurance in big data service applications and academic research has become an important and critical issue due to 4V in big data applications. In general, big data quality assurance refers to the study and application of various assurance processes, methods, standards, criteria, and systems to ensure the quality of big data in terms of a set of quality parameters.

Gao et al. [2] provide informative discussions for big data validation and quality assurance, including the essential concepts, focuses, and validation process. Moreover, they present a comparison among big data validation tools and several major players in industry are discussed. Also, they discuss the big data quality assurance issues, challenges and needs. Furthermore, these discussions may bring great benefits to the future of large data quality assurance. We have collected some data quality parameters from the published papers, and we have presented in Table I. It includes quality parameter and the corresponding attribute meaning.

978-1-5090-6318-5/17 $31.00 © 2017 IEEE
DOI 10.1109/BigDataService.2017.42

Table I. Quality Parameters for Big Data

Quality Parameters	Attribute Meaning
Data accuracy	It refers to the degree of closeness between the observed result and the true value or value that is accepted as true. Therefore, we can know this quality parameter is typically used to measure the collected sensor data by comparing the multiple sources.
Data correctness	This data quality parameter is much helpful to evaluate the correctness of big data sets in term of data types, formats, and so on.
Data consistency	Data consistency refers to data collection methods, schedules, and locations. It is much helpful to evaluate the consistency of the big data sets in abundant and different angles.
Data security	This quality parameter could be helpful to evaluate the security of the given big data sets in different perspectives.

They also discussed big data quality verification tools and players. They compare with tools in terms of operating environment, supported data sources, data validation, and current successful applications.

Now, when big data quality assurance is discussed, the quality of big data applications is also concerned. Of course, the quality factors of big data applications have gradually opened the mystery. Conventional quality factor such as *performance*, *robustness*, *security*, etc., can be applicable onto big data applications. From the published papers in [5], Tao et al. focus on big data system validation and quality assurance, and the paper includes informative discussions about essential *quality parameters*, *primary focuses*, and *validation process*. Compared with traditional software testing, they discussed the big data application specific test process. The test procedure comprises the following steps [6] .

Step 1: System function testing, including rich oracles, intelligent algorithms, learning capability, as well as domain-specific functions;

Step 2: System non-function testing, including system consistency, security, robustness, and QoS (Quality of Service);

Step 3: System feature testing, checks usability, system evolution, visualization, and so on;

Step 4: System timeliness testing, targets time related feature testing, including continuous testing, real-time testing, life-time testing, and others.

In addition, they also discuss the quality factors of different systems, including prediction systems, recommendation systems and so on. Based on those, we can draw out the quality factors of big data applications, and presented below:

- **Performance:** This factor indicates the performance of the big data applications, such as availability, response time, etc.
- **Reliability:** This factor helps to evaluate the durability of the big data applications when the required function is performed within a specified time period under specified conditions.
- **Correctness:** This is a quality factor used to assess the correctness of big data applications.
- **Scalability:** This quality factor means that big data application should be able to support large data sets now and in the future, and all components of big data application can be extended to address the growing complexity of complex data sets.
- **Security:** This factor helps to evaluate security of the big data application in various perspectives at the different levels.

Our brief survey of the literature has demonstrated that although big data quality assurance has been studied, quality assurance techniques for big data applications has also been studied. However, there has been little scientific research aimed at understanding, defining, classifying and communicating quality assurance techniques of big data applications. Consequently, there is no clear way to deal with quality assurance of big data applications. Therefore, discussing quality assurance techniques of big data applications is very necessary.

III. The SURVEY FRAMEWORK

In this section, we briefly summarize the articles which we researched, and we describe the articles in several sections. We have studied new research results in the last five years, discussed the application domain of big data applications, and show whether the quality assurance technique is applied at design-time or run-time. In addition, we also discussed the big data applications functional properties or non-functional properties (e.g., performance, reliability, availability, etc.), which are very important.

We all know that big data has its own properties, such as *Volume*, *Velocity*, *Variety* and *Veracity*. *Volume* means the sheer size of the databases. *Variety* means the different types of data which can be stored within a single data container, and everything from discrete numeric and string values to texts and images and to video films and audio recordings. All of this can be stored and retrieved in various sequences or combinations [7] . *Velocity* means the speed with which the objects can be retrieved and put together. The search algorithms are constructed in such a way that many multiple search paths are executed parallel to one another. In the end the results of the different searches are joined together to form a consistent whole. We discussed the quality assurance of big data application, so big data itself unique 4V properties (i.e., volume, velocity, variety, and veracity) are also focused. For quality assurance, quality assurance techniques are particularly important. Therefore, we are mainly from these aspects to analyze the article. In the Table II, we have a simple induction for the articles which we have researched.

Through the analysis of Table II, and related articles, large data applications are widely applicable to many areas, especially in recent years. Quality assurance technology of big data applications are rapid developed. Consequently, we can conclude that there are six main ways, including testing, model-driven architecture (MDA), monitoring, fault tolerance, verification and prediction to typically ensure the quality of big data applications. In the next part,

we will conduct a detailed description and analysis of the approaches of these six aspects.

Table II. Comparison of Quality Assurance Approaches of Big Data Application

Year	Reference	Application Domain	Technique	Design-time or Run-time	Functional or Non-functional Properties	Properties
2014	[8]	Application Testing	Testing	Design-time	Performance	4V characteristics of big data
2015	[7]	Big data bases	Testing	Design-time	Validity Consistency	Volume, Variety
2015	[9]	Big Data and Cloud Computing to process large data.	Testing	Design-time	NULL	Volume, Variety, Velocity
2015	[10]	Data-intensive software systems	MDA	Design-time	Reliability Safety Efficiency	Volume, Velocity
2015	[11]	Not Mentioned Specific Application Domain	MDA	Design-time	Performance	Not Mentioned
2012	[12]	Enterprise Application Performance Management	Monitoring	Not Mentioned	Performance	Velocity
2016	[22]	Not Mentioned Specific Application Domain	Monitoring	Design-time	Reliability	Volume, Velocity
2015	[13]	Distributed storage systems	Fault tolerance	Design-time	Performance	Velocity
2012	[14]	Modern cloud computing systems and so on	Prediction	Run-time	Reliability	Volume
2015	[15]	MapReduce	Verification	Design-time	Integrity, Performance	Not Mentioned

IV. THE SURVEY APPROACHES

Research shows that quality assurance techniques of big data application are mainly these aspects – MDA, Testing, Verification, Fault tolerance, Monitoring, and Prediction.

A. Model-Driven Architecture (MDA)

MDA derives from the well-known idea of separating the specification of system operations from the system. MDA provides a way (through related tools) to standardize a platform-independent application, selects a specific implementation platform for the application, and then transforms application specifications to a specific implementation platform. The three main goals of MDA are: to achieve portability, interoperability, and reusability through architectural separation [16], [17] .

The model driven approach is a well-known one and has been widely exploited in many areas of software engineering. The goal of the MDA is to design applications in a model-driven approach which is more abstract than the implementation of the techniques. For example, Alodib et al. [11] propose an extension to automate the integration of the Hadoop platform. This is intended to break up each problem into multiple sub-tasks using a simple programming model (MapReduce). After the analysis is calculated, the results are submitted to the Score table linked to the protocol service. The approach harnesses the capability of Model-Driven Architecture (MDA) to automate the creation, and integration of the architecture.

Largely, due to existing models and QA techniques ignore properties of data such as volumes, velocity and so on. Casale et al. [10] present the research agenda of DICE. It is a quality-aware MDE technology for big data cloud applications. And its goal is to developing a quality engineering tool chain offering simulation, verification, and architectural optimization for Big Data applications. They also present the main challenge in this approach. These challenges are due to the fact that data operations and data characteristics cannot be fully described.

Etani [18] describes database application model and its service for drug discovery introducing their proposed software development process in MDA into their research process. The issue of veracity can be solved when pinpoint data are selected from drug properties in big data analytics with domain model. Our approach of software development process in MDA will be useful for developing a big data application and a new service by “veracity” of big data.

All in all, MDA provides a complete solution for integration of big data applications at different lifecycle stages. It advocates the use of formalized system models as the core of application integration. Consequently, we can know MDA is an important method for quality assurance of big data application.

B. Testing

Application testing is a test of the entire product to verify whether the application meets the requirements specification definition, and to identify inconsistent with the requirements specification or contradictory places, so as to propose a more complete solution.

The volume and variety of big data presents a particular challenge to the testing of the big data application. Therefore, Sneed et al. [7] consider that there is no other way but to automate the test process to test the applications. Due to the volume and variety of big data, they think it is impossible to test big data application manually, and testing big data need new processes and higher degree of automation. People need automated tools to scan through the big data and check the validity and consistency of the content.

The performance of big data applications is particularly important. Performance testing is a test method, which belongs to a typically non-functional testing. During performance testing, the system tests by simulating various normal and abnormal peak load conditions to reduce operational, upgrade, or patch deployment risk through performance testing (such as information systems) to achieve a user response time load. But the existing performance testing techniques are not suitable for the big data application. Liu [8] proposes test technique for performance testing. The technique provided testing goal analysis, testing design, load design for big data applications. The characters for different big data applications could be supported to consider specific multiple test data design method under this framework. This performance technique is used to test some applications and demonstrated its effectiveness.

Jesús Morán et al. [9] propose a testing technique named MRFlow, which is based on data flow test criteria and oriented to transformations analysis between the input and the output, and it can test defects in MapReduce programs. MapReduce is a programming model for parallel computing of large-scale data sets. MapReduce achieves reliability by distributing the large-scale operations on the data set to each node on the network. Moreover, they tested the technology, and the testing results are better.

In summary, the testing for big data quality assurance is very important, especially because of the big data properties, testing is a very good quality assurance method. And we can learn that the testing will work in many cases which we meet.

C. *Verification*

Applications based on big data are now widely used, such as recommendation, prediction and decision systems. Research shows that current research rarely explores how to effectively verify big data applications to ensure the quality of big data applications. Big data properties have taken many challenges for big data applications. For example, because of the volume of data and the timeliness of data, it is a very difficult task to verify the correctness of big data applications.

Gao et al. [5] have discussed the validation methods for big data application. They discussed and reviewed existing research results in software testing methods that have been used to validate various types of big data applications, including data mining programs, bioinformatics programs, and learning-based applications. And those methods include program-based software testing, classification-based testing, metamorphic testing (MT), learning-based testing, crowd-sourced testing, data model-based testing, rule-based software testing and so on.

Result integrity is one of the most important security issues in cloud-based big data computing scenarios. Wang et al. [15] present MtMR, a Merkle tree-based verification method to ensure the high integrity of the MapReduce tasks. MtMR covers MapReduce in a hybrid cloud environment and performs two rounds of Merkle tree-based verification for the pre-reduction and restoration phases. In each round of verification, MtMR samples a small portion of the reduced task input/output records on the private cloud, and then performs Merkle tree-based verification of all task input/output records. After analysis, they believe that MTMR can significantly improve the comprehensive, while can produce moderate performance overhead.

Traditional software verification models and standards have been unable to meet the quality requirements of big data applications (because of the existence of big data properties)[19] . Although many scholars have studied the quality verification problem of big data applications, but not enough, the quality verification and assurance of big data application challenges remain.

D. *Fault tolerance*

The so-called fault tolerance refers to the existence of the fault in the case of the system does not fail, still is able to work properly. Fault tolerance is rather a fault, not an error. The use of fault tolerance to ensure the quality of big data applications can usually measure in terms of application reliability, availability, and testability.

Due to the trends towards Big Data, people want to provide large storage systems, and those are accessible by many servers. The shared storage has been the performance bottleneck and a single-point of failure. Lundberg et al. [13] suggest that we introduce a cache in the distributed storage system. The cache system must be fault tolerant so that no data is lost when the hardware failure happened. According to the study, we know that the cache system is a way to improve the performance of most systems.

As we all known, NoSQL databases are critical for supporting big data applications, because they can handle a large number (i.e., volume) of highly variable (i.e., variety) user-generated content while guaranteeing fault tolerance, availability, and scalability. However, all NoSQLs are somewhat different from each other, even if they are considered to belong to the same database family. Scavuzzo et al. [20] pose an efficient and fault tolerance data migration method. In general, data migration should be able to tolerate faults or interruptions by recovering to the last correct state, since NoSQL typically stores large amounts of data, which means long-running migration tasks, but on the contrary, higher risk of faults will happen. However, their approach tolerates a sudden fault of any component involved in the data migration process without any data loss. Experiments show that the method used to perform the data migration is efficient, fault tolerance, and really can improve the NoSQL technology interoperability.

Likewise, there is an increasing interest in the reliability and availability of big data cloud applications. And fault tolerance is a very effective means to solve the problem of reliability and usability. Jhawar et al. [21] focus on describing repetitive faults in typical cloud computing applications, analyzing the impact of faults on user applications, and investigating fault tolerance solutions corresponding to each type of failure. And they also talk

about providing fault tolerance as a service to user applications as an effective means of addressing reliability and availability issues.

From those researches, we can know that the fault tolerance is helpful to quality assurance of big data applications.

E. Monitoring

In recent years, a large number of structured, semi-structured and unstructured data is generated. These data are huge, complex, and rapidly changing. If the data cannot be filtered, the real-time monitoring of information cannot be achieved. Therefore, one of the biggest challenges with big data applications is how to analyze and process huge amounts of data in real time. And real-time monitoring is an effective way to ensure the quality of large data applications. Therefore, improving the real-time performance of large data monitoring is very necessary.

In order to improve the real-time performance of big data monitoring, Shi et al. [22] dish a dual cloud architecture to take full advantage of cloud resources and network bandwidth. They also propose a real-time monitoring algorithm based on user evaluation in Hadoop platform, which uses a combination of computing nodes. The monitoring algorithm can eliminate nonsense data such as spam, malice evaluation, brush score, brush reputation and brush list by establishing user evaluation system. As a result, it can significantly reduce the amount of data, but also can greatly improve the operational efficiency. Thus, it can ensure real-time monitoring information, reliability and accuracy.

Distributed systems are typically big data applications. State monitoring has been widely used to detect critical events and anomalies in distributed systems. Unfortunately, existing distributed state monitoring methods are usually designed based on the condition that we assume always-online distributed monitoring nodes and reliable inter-node communicate. Therefore, based on these methods, it often produces misleading results, which leads to various problems being introduced to rely on state monitoring results to perform automatic management tasks of the user. Meng et al. [23] introduced a new state monitoring approach, and this method exposed and handled communication dynamics such as message delay and loss in Cloud monitoring environments. Firstly, by quantitatively estimating the accuracy of monitoring results, it can capture uncertainties which are introduced by messaging dynamics. This characteristic is useful to distinguish trustworthy monitoring results from one heavily deviated from the truth. Secondly, they can configure the monitoring algorithm, which minimizes monitoring errors. And there are also other methods related to monitoring, which we can find in paper [24], [25] .

Therefore, we can know that big data brings some trouble to big data applications, and using special monitoring approaches can improve the quality assurance of big data applications and improve reliability, performance and other non-functional properties.

F. Prediction

Big data applications will have a variety of failures. If we can predict the upcoming failure; it will greatly improve the quality of large data applications. Therefore, the prediction technique for big data quality assurance is an effective way.

Yang et al. [26] design a general framework named Hdoctor for hard drive failure prediction. Hdoctor demonstrated a number of innovations, and building time-dependent features to characterize Self-monitoring, Analysis and Reporting Technology (SMART) value transitions during disk failures is the important one. Meanwhile, Hdoctor automatically collects/labels samples and updates model, and works well for all kinds of disk failure prediction in their intelligent data center.

Existing production applications are short of real-time performance status of production process active perception, resulting in the production abnormal conditions processed lag, leading to the frequency problems of deviations in production tasks execution and planning. To address this problem, Zhang et al. [27] advance they should extend an advanced identification technology to the manufacturing field to acquire the real-time performance data. Based on the sensed real-time manufacturing data, they present a prediction method which applies the Dynamic Bayesian Networks (DBN) theory and methods. Achieving the prediction of the performance status of production system and potential anomalies is the goal of the method, and it can provide the important and abundant prediction information. All in all, Dynamic Bayesian Networks theory and method is used to make the mathematical modeling of performance prediction for production system based on manufacturing big data.

In modern cloud computing systems, thousands of cloud servers are interconnected through multiple layers of networks. Faults are common in such large and complex systems. In order to predict the failure, we should monitor the system implementation process, and collect health-related runtime performance data. Guan et al. [14] present an unsupervised failure detection method based on an ensemble of Bayesian models. It characterizes the normal system execution state and detects anomalous behavior. The tagged data is available after the system administrator verifies the exception. Then, supervised learning based on decision tree classifier is used to predict future failures. There are other predictive methods in paper [28], [29] and other papers which we do not know.

Dealing with faults which have been happened may be very difficult, and fault prediction is particularly important. Therefore, it is necessary to discuss the fault prediction method of big data applications. Therefore, I think prediction will play an important role in quality assurance of big data applications.

V Discussion

By reading a lot of literature, we can summarize a number of approaches to ensure the quality of big data applications, including MDA, Testing, Verification, Fault tolerance, Monitoring, and Prediction. In TABLE III, we further summarize functional or non-functional properties involved in these six aspects, as well as the big data properties.

As we can see from the TABLE III, in the process of considering big data application quality assurance,

performance of this non-functional property is basically the main consideration. However, the big data properties have a great impact on quality assurance of big data application. As shown in TABLE III, we know one of big data properties which are common for mostly approaches is variety. However, the variety often is solved by NOSQL which can handle structured data, semi-structured data, and unstructured data of big data. NoSQL databases are key to supporting Big Data applications, since they enable handling large quantities (i.e., volume) of highly-variable (i.e., variety), user-generated contents while guaranteeing fault tolerance, availability (i.e., velocity) and scalability [20] .

TABLE III. Approaches, Functional or Non-functional Properties, 4V properties of big data application

Approaches	**Functional or Non-functional Properties**	**4V properties**
Model-Driven Architecture (MDA)	Performance, Scalability	Veracity, Volume, Variety
Testing	Availability, Performance	Variety, Velocity
Verification	Performance, Reliability	Volume, Variety
Fault tolerance	Performance, Scalability	Variety, Volume
Monitoring	Performance, real-time	Variety, Velocity
Prediction	Performance, Dependability	Variety, Veracity

Not only that, according to the research, we can get big data properties of the challenges, and how to use the novel technique to solve the problems. Consequently, we can summarize the big data properties, challenges, as well as the techniques for those challenges in following form.

TABLE IV. Properties, Challenges and Techniques

Properties	**Challenge**	**Novel Technique**
Volume	Storage/Scale	Distributed File Systems
Velocity	Fast Processing	Parallel Programming
Variety	Heterogeneity	NOSQL Databases

When we consider the big data properties and quality requirements, it is anticipated to aid requirement analysts in the specification of quality requirements while keeping big data properties in mind.

There are some major issues and challenges in big data application quality assurance. Here are typical ones.

Issue #1 - Lack of awareness and good understanding of quality assurance techniques for big data applications.

With the fast development of big data technologies and analytics approaches, more big data applications and service systems are developed to be used in many areas of our daily life. Consequently the increasing deployment of big data applications and services dishes quality assurance concerns. Then, most people will find ways to solve a specific problem until the big data application problems happened. Hence, according to real world practitioners, there is a clear demand on understanding the quality assurance of big data application. This brings the first demand of big data application quality assurance.

Need #1- Full understanding the quality assurance techniques to solve the special functions and needs of big data applications and services.

Issue #2 - Lack of approaches to solve quality assurance issues in different big data applications.

For specific big data applications, there are specific ways to solve the quality assurance problem. However, there is currently no strictly defined approach to solve the problem. Therefore, it brings the second demand of big data application quality assurance. For example, testing oracle may be a big issue for big data applications due to the 4V properties.

Need #2- Define and develop well-defined big data application quality assurance standards, and define some approaches to solve quality assurance issues. Those approaches can be extracted from the six aspects of this paper.

Issue #3 – Lack of solutions to coordinate big data properties with quality assurance techniques.

Today, big data applications, such as social media, generate more data in a short period of time than was previously available requiring new techniques for quality assurance. Existing techniques have no adequate scalability and facing challenges because of big data properties such as Volume, Velocity, Variety and Veracity. Therefore, it brings the last demand of big data application quality assurance.

Need #3 - Consider functional or non-functional properties with big data properties together to ensure the quality of big data applications.

In addition, we have general approaches to deal with big data properties:

- Distributed File Systems for Volume;
- Parallel Programming for Velocity;
- NOSQL Databases for Variety (structured, semi-structured and unstructured data).

VI. Conclusion and Future Work

This paper focuses on the quality assurance of big data application. It mainly discusses the state-of-art approaches to ensure the quality of big data applications. The surveyed approaches are mainly testing, model-driven architecture (MDA), monitoring, fault tolerance, verification and prediction. In addition, this paper discusses the impact of big data characteristics on big data applications.

Although researchers have proposed some quality assurance techniques for big data applications, the challenge of big data applications still exists. Consequently, how to effectively ensure the quality of big data applications is still a hot research issue. In the follow-up study, we should conduct more research based on big data 4V properties. We can try to deal with big data 4V problems and we can also consider functional or non-functional properties of big data applications with big data properties together to ensure the quality of big data applications.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61572171) and the Fundamental Research Funds for the Central Universities (No. B15020191).

REFERENCES

[1] Big Data Technology and Services at $32.4 Billion in 2017 - IDC[J]. San/lan, 2013.

[2] Gao J, Xie C, Tao C. Big Data Validation and Quality Assurance -- Issuses, Challenges, and Needs[C]// IEEE, IEEE International Symposium on Service-Oriented System Engineering. IEEE, 2016: 433-441.

[3] Garg N, Singla S, Jangra S. Challenges and Techniques for Testing of Big Data[J]. Procedia Computer Science, 2016, 85: 940-948.

[4] Yesudas M, Menon S G, Nair S K. High-Volume Performance Test Framework using Big Data[C]// International Workshop on Large-Scale Testing. ACM, 2015: 13-16.

[5] Tao C, Gao J. Quality Assurance for Big Data Applications– Issues, Challenges, and Needs[C]// The Twenty-Eighth International Conference on Software Engineering and Knowledge Engineering. 2016.

[6] Guerriero M, Tajfar S, Tamburri D A, et al. Towards a model-driven design tool for big data architectures[C]// The, International Workshop. 2016: 37-43.

[7] Sneed H M, Erdoes K. Testing big data (Assuring the quality of large databases) [C]// IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops. IEEE, 2015: 1-6.

[8] Liu Z. Research of performance test technology for big data applications[C]// IEEE International Conference on Information and Automation. IEEE, 2014: 53-58.

[9] Jesús Morán, Riva C D L, Tuya J. Testing data transformations in MapReduce programs[C]// The, International Workshop. 2015: 20-25.

[10] Casale G, Ardagna D, Artac M, et al. DICE: Quality-Driven Development of Data-Intensive Cloud Applications[C]// IEEE/ACM, International Workshop on Modeling in Software Engineering. ACM, 2015: 78-83.

[11] Alodib M, Malik Z. A Big Data approach to enhance the integration of Access Control Policies for Web services[C]// IEEE/ACIS, International Conference on Computer and Information Science. IEEE, 2015: 41-46.

[12] Rabl T, Mez-Villamor S, Sadoghi M, et al. Solving big data challenges for enterprise application performance management[J]. Proceedings of the Vldb Endowment, 2012, 5(12): 1724-1735.

[13] Lundberg L, Grahn H, Ilie D, et al. Cache Support in a High Performance Fault-Tolerant Distributed Storage System for Cloud and Big Data[C]// Parallel and Distributed Processing Symposium Workshop. IEEE, 2015: 537-546.

[14] Guan Q, Zhang Z, Fu S. Ensemble of Bayesian Predictors and Decision Trees for Proactive Failure Management in Cloud Computing Systems[J]. Journal of Communications, 2012, 7(1): 52-61.

[15] Wang Y, Shen Y, Wang H, et al. MtMR: Ensuring MapReduce Computation Integrity with Merkle Tree-based Verifications[J]. 2016: 1-1.

[16] Xuan P, Zheng Y, Sarupria S, et al. SciFlow: A Dataflow-Driven Model Architecture for Scientific Computing using Hadoop[C]// IEEE Big Data 2013 Workshops: Big Data and Science - Infrastructure and Services. IEEE, 2013: 36-44.

[17] Klein J, Buglak R, Blockow D, et al. A reference architecture for big data systems in the national security domain[C]// International Workshop on Big Data Software Engineering. 2016: 51-57.

[18] Etani N. Database application model and its service for drug discovery in Model-driven architecture[J]. Journal of Big Data, 2015, 2(1): 1-17.

[19] Hussain M, Almourad M B, Mathew S S. Collect, Scope, and Verify Big Data -- A Framework for Institution Accreditation[C]// International Conference on Advanced Information NETWORKING and Applications Workshops. IEEE, 2016: 187-192.

[20] Scavuzzo M, Tamburri D A, Nitto E D. Providing big data applications with fault-tolerant data migration across heterogeneous NoSQL databases[C]// International Workshop on Big Data Software Engineering. 2016: 26-32.

[21] Jhawar R, Piuri V. Chapter 7 - Fault Tolerance and Resilience in Cloud Computing Environments[M]// Computer and Information Security Handbook. Elsevier Inc. 2013: 125-141.

[22] Shi G, Wang H. Research on Big Data Real-Time Public Opinion Monitoring under the Double Cloud Architecture[C]// IEEE Second International Conference on Multimedia Big Data. IEEE Computer Society, 2016: 416-419.

[23] Meng S, Iyengar A K, Rouvellou I M, et al. Reliable State Monitoring in Cloud Datacenters[C]// IEEE, International Conference on Cloud Computing. IEEE, 2012: 951-958.

[24] Iuhasz G, Dragan I. An Overview of Monitoring Tools for Big Data and Cloud Applications[C]// International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. 2015: 363-366.

[25] Zareian S, Fokaefs M, Khazaei H, et al. A big data framework for cloud monitoring[C]// The, International Workshop. 2016: 58-64.

[26] Yang W, Hu D, Liu Y, et al. Hard Drive Failure Prediction Using Big Data[C]// Reliable Distributed Systems Workshop. IEEE, 2015: 13-18.

[27] Zhang Y, Liu S, Si S, et al. Production system performance prediction model based on manufacturing big data[C]// IEEE, International Conference on Networking, Sensing and Control. IEEE, 2015.

[28] Xu J, Li H. The Failure Prediction of Cluster Systems Based on System Logs[M]// Knowledge Science, Engineering and Management. Springer Berlin Heidelberg, 2013:526-537.

[29] Dai D, Chen Y, Kimpe D, et al. Provenance-based object storage prediction scheme for scientific big data applications[C]// IEEE International Conference on Big Data. IEEE, 2014: 271-280.