International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

Automated Testing for Big Data Environment using Multi Model Structure Validation

S. Nachiyappan, S. Justus

Abstract*: Today, the usage of data volume has increased manifold. The domain of Big Data utilizes an attribute referred to as 'Predictability'. This extraordinary feature basically draws prediction and based on user's need, exhibits data to the user. The database systems prevailing today are quiet complex in nature. The software verification involves an essential phase of testing. The quality of database system depends on what is the quality of testing. Because of the high level of database complexity, testing too has become complicated which in turn has led to intensive workload. The evolution of Big Data has led many enterprises to make a shift towards big data technique such as 'Hadoop'. The proposed work recommends the technique of big data for carrying out processing of heterogeneous data items. On such technique is the MapReduce which being a well-known programming model build to achieve the proposed issues. The research utilizes machine learning techniques and recommends the approach of big data validation and testing via MMSV (Multi Model Structure Validation) in context to data handling. The MMSV model performs data analysis and comprehends it for making future forecast. Researchers have built methodologies keeping the automated testing components centralized for enhancing overall quality and efficiency of testing methods. Testing involves checking of massive data volume gathered from various sources into primary and central data warehouse. It's revealed that the technique of MMSV and big data yields improvised output compared to the prevailing system. Moreover the recommended system consumes low execution time. These techniques assure to attend all the issues thereby providing possible remedies and tactics.

***Index Terms*: Big Data, Automated testing, Hadoop, MapReduce, Multi Model Structure Validation techniques (MMSV), Validation.**

I. INTRODUCTION

Lately, there has been hype in the concept of big data which has extensively spread to big enterprise and government sector and a motivational tool for information advancement [1]. Though there are 4V attribute that distinguish big data in contrast to rest of the data, these being: volume, value, variety and velocity [2]. The attributes of velocity and volume denotes unexpected huge data volume and speed. Since big data is complicated and vast in variety, techniques need to be imbibed for extracting values [3]. For enhancing big data research, innovative algorithms, processing infra that is highly scalable and yielding great performance along with analytics strategy has been built for sustaining big data research [4].In last few years industries and academic circles introduced various big data analytics system namely Apache Spark, Cloudera, Impala, Apache Hive, IBM Big SQL, Transwarp Inceptor and much more, for attending concerning issues.

Manuscript published on 30 June 2019.
* Correspondence Author (s)
Prof. S. Nachiyappan, VIT University, Chennai (Tamil Nadu), India.
S. Justus, VIT University, Chennai (Tamil Nadu), India.

Large enterprises make use of big data analytics approach for constructing business application and achieving decision support based on the data. This being the reason that validation and testing of big data analytics system stands of utmost significant research in context to big data domain. The proposed research proposes big data technique for processing text data items. The Hadoop technique assures to be yielding high performance and scalability framework for executing big data. Moreover for storing and fetching of big data, Hadoop databases are made use of. MapReduce is yet another model for parallel data processing utilized for processing massive information for data intensive applications like big data.The work proposes the MMSV (Multi Model Structure Validation) model which is being highly implemented for big-data analysis. For achieving the availability and reliability of big data; big data applications and framework need to undergo testing and validation. But there stands few issues related to the testing and validation because of the four attributes of big data. For instance, to achieve high performance and efficiency of big data analysis, testing and validation process are extremely important but the fact is that huge data volume and its multiple variety portrays a big issue of concern for performing the same. The data validation method proposed in this work involves following stages: HDFS (Hadoop Distributed file System) by making use of MapReduce, Data Pre-processing, MMSV (Multi Model Structure Validation techniques) by utilizing validation. For carrying out the processing of big data, MMSV technique is implemented, here too there is difficulty in validation because of the massive data and anonymous results. Remarkable work has been carried out on big data's quality assurance as well as testing and validation of ML algorithms, but limited work is carried out on systematic validation and verification of overall big data. The proposed work emphasizes on the validation and testing related to big data.
Following is the journal classification: section 2 describes work of previous author. In section 3 the MMSV technique is presented by utilizing validation along with aspects of various stages. Section 4 exhibits experimental output. Lastly section 5 presents the conclusion and recommends research work for future.

II. RELATED WORK

Pengcheng Zhang et.al emphasizes on the review of quality assurance methods related to big data applications thereby presenting big data attributes and quality characteristics. For assuring the quality of big data applications, prime methodologies are being put forward, these are: testing, MDA (model-driven architecture), verification, fault tolerance, monitoring and prediction techniques [5]. Ikbal taleb et.al presents across-the-board quality management infrastructure illustrating the prime quality evaluation practices to be performed via multiple Big Data levels.

Byte means of the above framework, the quality management can be escalated thereby offering a blueprint for researchers to effectively comprehend quality practices and emphasize the significance of monitoring the quality [6].Paulo Vinicius Cardoso et.al presents the checkpoint and implementation of Recovery technique with the help of Apache Hadoop which being a framework supporting distributed processing of huge datasets around assembly of computers. By the means of checkpoint technique, fault tolerance is being provided on HDFS (Hadoop Distributed File System). Once CR attributes are stated statically by Hadoop, a suitable checkpoint interval must be selected, which is a challenge in itself. Thereafter a dynamic measure is being recommended for checkpoint attribute configuration on Hadoop Distributed File System, attempting to make it adjustable to system usage context [7]. Mingang Chen et.al put forwards two sort of cases of benchmark testing concerning big data analytics system. Case 1 illustrates a brief automated system testing strategy for Tran warp Inceptor by the means of TPC-DS. The test involves SQL compatibility along with system's performance, reliability and functionality. Case 2 presents testing and comparing the performance of Hive and Spark SQL by the means of TPCx-Big Bench which being an application centric end to end standard. It's clearly revealed from the test output that the overall performance of Spark SQL is much improvised compared to Hive in regard to the workload of pure HQL and query using MapReduce [8]. Gao et.al presents sound and useful discussion concerning big data validation and quality assurance, taking into consideration important focuses, concepts and the validation task. Besides there is a discussion presented on comparing big data validation tools and various major players in proposed in the industry and the various challenges and requirements associated with big data quality assurance. Catering to these discussions can aid in improvisation of huge data quality assurance in near future [9]. Junhua Ding et.al proposes a model for severe validation of large size image data and verification of software system as well as ML algorithms. For automating the selection and validation of large scale image data in CMA, various machine learning algorithms are combined with image processing strategies. In order to perform testing of scientific software, a metamorphic testing strategy is being implemented. Since the scientific software are non-testable, machine learning technique is imbibed for generating test oracles for ensuring the competency of test coverage [10].Alexandra L Heureux et.al presents the issues confronted by emphasizing the cause effect association by shaping issues with respect to Big data vs. dimensions responsible for triggering the issue, that is the 4V namely Volume, Veracity, Velocity and Variety. Besides ML algorithms and techniques are presented to illustrate their potential in managing multiple issues with the aim of aiding the practitioner to choose suitable remedies for the concerned cases. Eventually, a matrix is illustrated correlating the strategies and issues confronted. Zhiyi Zhang et.al presents metamorphic testing that was recommended to lighten the oracle issue concerning the domain of software engineering. It turns out to be a significant technique for carrying our software verification and validation. Lately, the metamorphic testing (MT) has revealed effective application in various fields, whether it is deep learning or bio informatics [12]. Raya Rizk et.al emphasizes on the necessity of workflows and elaborates the incorporation of validation tool referred to as 'Diftong'. It's used to make a comparison among 2 databases of different workflow version thus aiding in detection and prevention of un-required modifications. For enumerating the output of database comparison, row and column based stats have been utilized. It's revealed that Diftong yields in precise output in the testing environment, delivering advantage to enterprises requiring validating the workflow results. With the automation of the above process, one can minimize and even discard the threat of human error [13]. Chunli Xie et.al presents an overview of the case concerning the big data quality that deals with the study of actual big data quality, its dimension, process of data validation and tools. Data is being generated with huge variety, quick velocity. The big data is of supreme quality. According to the study, because of the poor quality, there can be mistaken data costs on the big data analysis output. The process of data validation aids in recognizing and improvising the data quality [14]. Feras A. Batarseh et.al has constructed a protective federal data analytical system, requiring a robust and tested model. The industry has shortage of government oriented models. FedDMV offers resolutions along with steps to enable growth of data analytics system amidst government restrictions. FedDMV handles unstructured data coming from various sources, manual steps that are automated, performs data validation and increasing the security. The experimental work output is jotted down and framed in manuscript [15]. Rachana Jannapureddy et.al inspects a structure of auto scaling across cloud environment focusing to reduce the resource cost by setting the virtual node automatically on the basis of data load that is being real time. For the AWS (Amazon Web Services) cloud scenario, a CEAS- cost effective auto scaling model is recommended. The CEAS model is responsible to increase the computing resource related to Hadoop cluster for minimizing the usage of resources during high or low work load, thereby enhancing the data processing and analysis in sufficient time. For analyzing and inspecting the frameworks effectiveness, real time sentiment analysis is carried over on the reviews and tweets of the universities individual that are being posted on social network [16]. Andrian Yang et.al presents a discussion on the implementation of contemporary cloud computing and big data programming models like Spark and MapReduce for successfully imbibing divide and conquers principle within distributed computing domain. In the field of big data informatics, software validation is of prime concern which is being overlooked. The process of software validation basically judges whether the program is capable enough to successfully complete the assigned task. Because of the huge input space and complicated algorithms, measuring the preciseness of computational output of big data bio informatics is arduous. A discussion is presented on the sophisticated strategy that relies upon multiple executions like metamorphic testing which can be utilized for imbibing the technique of bioinformatics quality assurance. [17]. Junhua Ding et.al presents a model for promising the quality over big data services. It incorporates the technique of an iterative metamorphic testing for carrying over the testing on non-testable sciebtific software along with a test strategy for validating ML (machine learning algorithms) with stratified 10-fold cross validation. In order to guarantee the efficiency of the model, the algorithms and software are being verified and validated in CMA [18].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Mr. Kunal Sharma et.al presents the model of generalized data validation amidst Hadoop and RDBMS. Big data enterprises are switching over Hadoop, Hadoop Ecosystem Projects namely Hive and HBase for storage of data. By the means of data migration process, originations are transferring data from RDBMS to Hadoop. This data is being further utilized for performing analysis. This migration of data may at times result in discrepancy because of various parameters, leading to imprecise data analysis [19]. Teemu Kanstren et.al presents the comprehensive system architecture, collection of various types of data, initial data analysis and comprehension till date. It's revealed from the test that there is a ceaseless evolution in the network with integration and testing of new services and application of new big data. The 5GTN (5G test network) is constructed for enhancing analysis and growth of 5G techniques and latest services residing on them. 5GTN is an holistic approach for monitoring and generating tests that aids in creation and gathering of multiple test and datasets and result analysis which can be utilized as foundation for building new techniques, services, models and optimizations. Combination of network, data volume and service parameters space results in various big data issues [20]. Jorge Veiga attends the above mentioned issue by carrying out a comparative assessment of Hadoop, Spark and Flink by utilizing big data workloads, taking into consideration parameters such as scalability and performance. Few workload factors are being altered namely input data size, HDFS block size, interconnect network or thread configuration for characterizing the framework's behavior [21].

III. PROPOSED WORK

A. *Overview*

Testing of big data analytics system confronts various issues because of the 4V attributes and complications involved in big data system. A framework has been built for precise validation of large text data and testing of ML algorithms and analytics system. The method recommended for validation and testing yields in difficulty of big data analytics system. Distributed architectures namely testing and validation are being imbibed. Test datasets must comply with 4V attributes as well as resemble typical business environment. The proposed research work aims on validation and big data testing in an automated manner. The technique of MMSV (Multi Model Structure Validation) is utilized for big data processing which being problematic for validating big data and results that are not known. Output yield by MMSV reveals and exhibits corrupted and non-corrupted files and format. Albeit, there exist noteworthy work on quality assurance concerning big data as well as testing and validation of ML algorithms

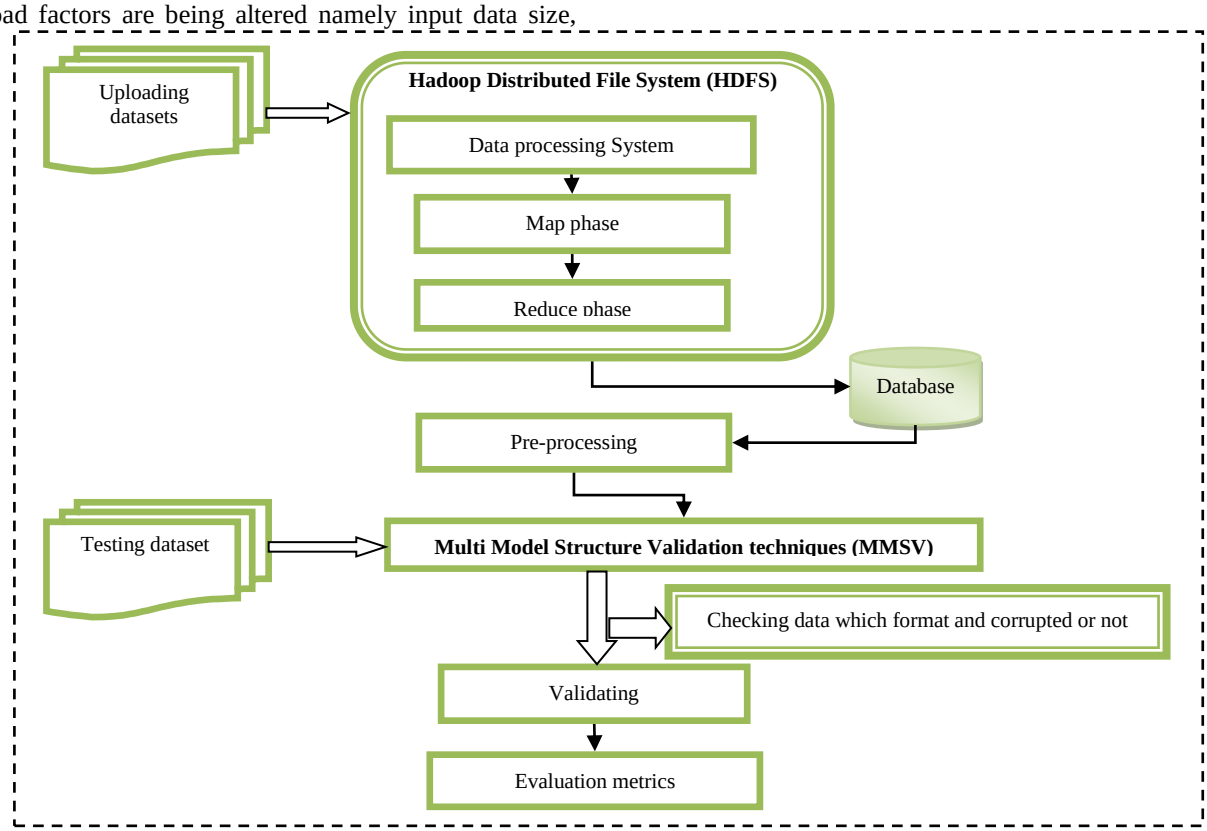

Fig 1: Proposed Architecture

B. *Upload Dataset*

Data uploading involves loading of different sorts of real time data that is unstructured, semi structured or structured. To achieve this HDFS is applied on Hadoop system in order to store and process massive text data volume that is unstructured. Like text document, file, email etc…) besides it has the support for RDBMS and machine-readable data like XML and CSV. Data upload status can be monitored by HDFS with the help of relational database within the dashboard along with unstructured data such as web logs or emails.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

C. *Hadoop Distributed File System (HDFS)*

The Hadoop technique offers a model with distributed file system to carry out the inspection and transformation of huge data sets via MapReduce model. Using HDFS, huge files can be stored over the machines in a huge cluster. Files are being stored as sequence of blocks; all blocks are of same size excluding the last block. To achieve fault tolerance, file blocks are replicated. Both the replication factor as well as block size can be configured for each file. No: of file replicas can be mentioned using an application. Similarly the replication factor can be mentioned during the creation of the file which can be modified later. HDFS files can be written once with only one writer for a given time. The programming paradigm of MapReduce aids in processing datasets that is of large scale within computer clusters. It involves two functions namely map () and reduce (). Users are granted to imbibe their processing logic by mentioning the functions of customized map () and reduce () function.

Map Phase – the map () function accepts an input value and on basis of that generates a list of intermediate key values.

(in_key,in_value)>list(out_key,intermediate_value)

Reduce Phase – the MapReduce() function assembles all intermediate pairs relying upon the intermediate keys, thereafter passing them over to reduce() function for generating the output. database is utilized for storing the MapReduce data.

(out_key,list(intermediate_value))>list(out_value)

D. *Data Preprocessing*

The method of data pre-processing is considered as the final step in quality management of big data. The various stages it involves are: cleansing, integration, normalizing, transforming, data integration for enhancing the quality prior to be utilized during the processing step. Data cleansing focuses on completeness and consistency of data. Various quality settings such as quality report, baseline model and requirements are necessary for achieving data quality enhancement. It helps in determining the expected data quality, scores and dimensions.

E. *Multi Model Structure Validation Techniques*

The technique of MMSV performs testing and validation of big data sequence of testing dataset is generated by fetching the identity number from the database. The file generated can be either corrupted or uncorrupted which will be validated by the means of MMSV. Response time, data size, data type and reliability are being validated. A brief algorithm is presented below.

Input: *Sequence of data m* $[(x_1, y_1),.....,(x_m, y_m)]$ *with identity number* $Y=(1...K)$;
Number of iterations I;

Initialization:
Testing dataset $D_1(s) = 1/m$

Procedure:
for (i=1, i<=I, i++)
{
Select training data subset S_t *, drawn from the distribution* D_t
Call the training data with subset S_t
Find the file format
If (m=text)
{
Generate the Validation $V_t = X \rightarrow Y$
Calculate Validation V_t
else
{
Data not readable ←Display
}
}
Output: *Show validated data*

F. *Data Validation*

This includes defining validity, completeness and consistency of data and assuring whether validated data is precise, credible and meaningful. According to the report, the major time consumed in big data projects is utilized towards data cleaning and preparation purpose. Framing customized expression for verifying data quality like for instance, checking data type, phone number, formats, positive and negative values, more or less than range etc…

G. *Evaluation Metrics*

For the evaluation of the trust score, there are many characteristics that are responsible. For the selection of data provider for processing of big data, the cloud's potential to process bug data must be taken into account, in regard to the 4 attributes of big data namely velocity, variety, veracity and volume. Therefore in order to select the appropriate data provider, the big data attributes play a vital role.

Reliability:

$$SR=\frac{TR-(IC+DoS\ incidents)}{TR}$$

Where,
SR= the task success ratio
TR= the total number of task requests
IC= the number of illegal connections
DoS incident = the number of denial of service incidents

Response Time: the real execution time 'D' represents the time utilized in milliseconds amidst sending a request and receiving the last byte of the response.

Availability: it is the ratio of the no: of received responses against the no: of sent requests.

Throughput:

$$RH=\frac{TR}{(End\ time-Start\ time)}\times 1000$$

Where,
RH=Request Handled per sec
End time = last request response time
Start time = first request start time.

IV. RESULT AND DISCUSSION

In the domain of big data, assuring validation is one of the vital issues of concern. Proposed techniques and strategies derived from the study and an industry has revealed that validation has not attained a satisfying maturity level. Evaluating the validation of big data in contrast to the value produced by it for its user's is of high significance. Moreover for attaining high quality evaluation output, thoroughly reviewed data testing and validation plan along with appropriate assessment scheme, suitable validation strategies and right tools and platforms is important for carrying out various validation evaluation tasks.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Table 1 depicts the overall performance of the recommended technique which is compared with respect to Spark SQL, RDBMS and Hive. The MMSV technique yields in improvised reliability, time and data-size and data type. It depicts high performance in contrast to rest of the techniques.

Table 1: Comparison of validation Techniques

S. No	Techniques	Data size (Volume)	Time (Velocity)	Reliability (Veracity)	Data Type (Variety)
1	Hive	50GB	456 Sec	No	Structured, Semi structured data and Un structured data
2	Spark SQL	50GB	510 Sec	No	Structured, Semi structured data and Un structured data
3	Relational Data Base Management System (RDBMS)	40GB	164.68 Sec	Yes	Not Mentioned
4	Multi Model Structure Validation (MMSV)	50GB	120 Sec	Yes	Structured data

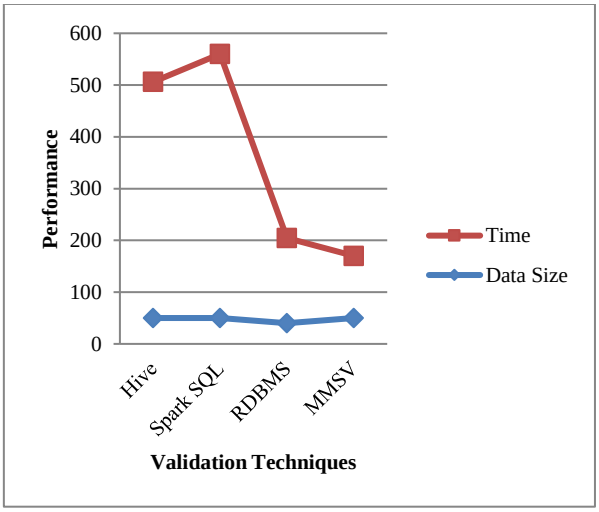

Fig 2: Performance Comparison of Validation Techniques

Fig 2 displayed above depicts the overall performance of the recommended technique which is compared with respect to Spark SQL, RDBMS and Hive. The MMSV technique yields in improvised reliability, time and data-size and data type. It depict high performance in contrast to rest of the techniques.

Table 2 depicts the accuracy performance of the recommended technique which is compared with respect to SVM, Navies Bayes and Google Prediction API. The MMSV technique yields in improvised accuracy. It depicts high accuracy in contrast to rest of the techniques.

Table 2: Comparison of overall Accuracy

S.No	Techniques	Accuracy (%)
1	Support Vector Machine (SVM)	90
2	Navies Bayes	90.21
3	Google Prediction API	89.98
4	Multi Model Structure Validation Technique (MMSV)	94

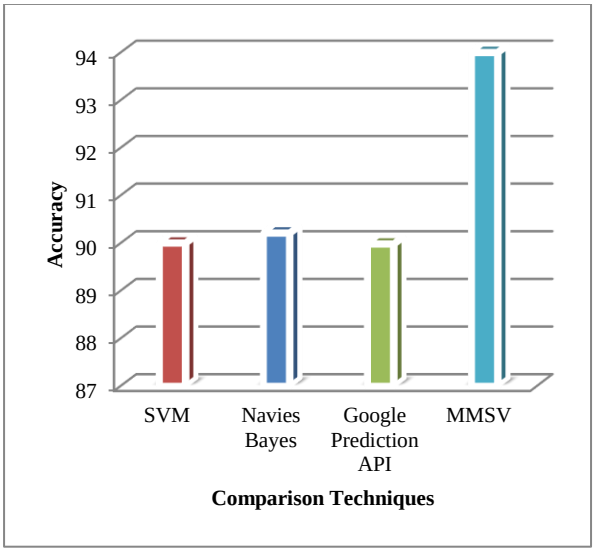

Fig 3: Accuracy Comparison of Accuracy Techniques

Fig. 3 displayed above depicts the accuracy performance of the recommended technique which is compared with respect to SVM, Navies Bayes and Google Prediction API. The MMSV technique yields in improvised accuracy.

It depicts high accuracy in contrast to rest of the techniques.

V. CONCLUSION

The vast growth and development of big data applications and techniques, there is more emphasis on big data analytics system by the industries and researches. There is comparison carried out on the performance of big data analytics system which grants you to adjust system parameters thereby optimizing system performance. The proposed research evaluates the issue involved in testing and validation of big data analytics system and provides a comprehensive view of methodologies and test techniques.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

REFERENCES

1. Alexandre Langeois, Eduardo Cunha de Almeida, Anthony Ventresque "Poster: BDTest, a System to Test Big Data Frameworks", © IEEE, International Conference on Software Testing, Verification and Validation Workshops, 2017, p.p.395-397.
2. Teemu Kanstren "Experiences in Testing and Analyzing Data Intensive Systems", © IEEE International Conference on Software Quality, Reliability and Security, 2017, p.p. 589-590.
3. Jing Wang, Dayong Ren "Research on Software Testing Technology under the Background of Big Data", © IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2018, p.p. 2679-2682.
4. Adiba Abidin, Divya Lal, Naveen Garg and Vikas Deep "Comparative Analysis on Techniques for Big Data Testing", © IEEE, Comparative Analysis on Techniques for Big Data Testing, 2016, p.p.219-223.
5. Pengcheng Zhang, Xuewu Zhou, Wenrui Li, Jerry Gao "A survey on quality assurance techniques for big data applications", © IEEE, International Conference on Big Data Computing Service and Applications, 2017, p.p. 313-319.
6. Ikbal taleb, Mohamed Adel Serhani and Rachida Dssouli "Big Data Quality: A Survey", © Research gate, Big Data Congress, 2018.
7. Paulo Vinicius Cardoso and Patr´ıcia Pitthan Barcelos "Validation of a Dynamic Checkpoint Mechanism for Apache Hadoop with Failure Scenarios", © IEEE, 2018.
8. Mingang Chen and Wenjie Chen, Lizhi Cai "Testing of big data analytics systems by benchmark", © IEEE, International Conference on Software Testing, Verification and Validation Workshops, 2018, p.p.231-238.
9. Jerry Gao, Chunli Xie and Chuanqi Tao "Big Data Validation and Quality Assurance – Issuses, Challenges, and Needs", © IEEE, Symposium on Service-Oriented System Engineering, 2017, p.p.433-441.
10. Junhua Ding, Xin-Hua Hu, and Venkat Gudivada, "A Machine Learning Based Framework for Verification and Validation of Massive Scale Image Data", © IEEE, TRANSACTION ON BIG DATA, 2017, p.p. 1-18.
11. Alexandra L Heureux, Katarina Grolinger, Hany F. ElYamany, Miriam A. M. Capretz "Machine Learning with Big Data: Challenges and Approaches", © IEEE Access, 2017, p.p. 1-22.
12. Zhiyi Zhang, Xiaoyuan Xie "Towards testing big data analytics software: the essential role of metamorphic testing", © Springer, Biophysical Reviews, 2019, p.p. 123- 125.
13. Raya Rizk, Steve McKeever, Johan Petrini and Erik Zeitler "Diftong: a tool for validating big data workflows", © Springer, journal of big data, 2019, p.p. 1-27.
14. Chunli Xie, Jerry Gao and Chuanqi Tao "Big Data Validation Case Study", IEEE Third International Conference on Big Data Computing Service and Applications, 2017, p.p. 281-286.
15. Feras A. Batarseh, Ruixin Yang and Lin Deng "A comprehensive model for management and validation of federal big data analytical systems", © Springer, Big Data Analytics, 2017, p.p. 1-22.
16. Rachana Jannapureddy, Quoc-Tuan Vien, Purav Shah and Ramona Trestian "An Auto-Scaling Framework for Analyzing Big Data in the Cloud Environment", © MDPI, Applied Sciences, 2019, p.p. 1-16.
17. Andrian Yang, Michael Troup, Joshua W.K. Ho "Scalability and Validation of Big Data Bioinformatics Software", © Elsevier, Computational and Structural Biotechnology, 2017, p.p. 379–386.
18. Junhua Ding, Dongmei Zhang and Xin-Hua Hu "A Framework for Ensuring the Quality of a Big Data Service", © IEEE International Conference on Services Computing, 2016, p.p. 82-89.
19. Mr. Kunal Sharma, Dr. Vahida Attar "Generalized Big Data Test Framework for ETL Migration", © IEEE, International Conference on Computing, Analytics and Security Trends (CAST), 2016, p.p.528-532.
20. Teemu Kanstren, Jussi Liikka, Jukka Makela, Markus Luoto, Jarmo Prokkola "Preliminary Big Data in a 5G Test Network", © IEEE International Conference on Big Data, 2016, p.p. 2722-2727.
21. Jorge Veiga, Roberto R. Exp´osito, Xo´an C. Pardo, Guillermo L. Taboada, Juan Touri˜no "Performance Evaluation of Big Data Frameworks for Large-Scale Data Analytics", ©IEEE International Conference on Big Data, 2016, p.p. 424-431.

AUTHORS PROFILE

Prof. S. Nachiyappan done his PG in Anna University in 2004 and he worked in industry for 5 years his research areas are software engineering , big data, knowledge engineering and Software Testing. He is working as an academician for more than 10 years. Overall he has experience of 15 years in both IT and Engineering Colleges. He is a member of ACM Professional Chapter. At present he is working with VIT University Chennai campus.

Dr. S. Justus has rich experience in managing projects he worked as a project manager in Infosys and he turned into academician and guided various projects and students, he has a very good skill set in working with software engineering, knowledge engineering software metrics and various IT software products, he has an overall experience of 17+ years in both IT and Academic. He has guided more than 15 PG students for the project and has published various papers in national and international journals. He is a member of ISTE, IEEE, IAENG.