A Big Data Framework for Quality Assurance and Validation

S. Nachiyappan, Justus S

Abstract*: *Big data is a new technology, which is defined by large amount of data, so it is possible to extract value from the capturing and analysis process. Large data faced many challenges due to various features such as volume, speed, variation, value, complexity and performance. Many organizations face challenges while facing test strategies for structured and unstructured data validation, establishing a proper testing environment, working with non relational databases and maintaining functional testing. These challenges have low quality data in production, delay in execution and increase in cost. Reduce the map for data intensive business and scientific applications Provides parallel and scalable programming model. To get the performance of big data applications, defined as response time, maximum online user data capacity size, and a certain maximum processing capacity. In proposed, to test the health care big data . In health care data contains text file, image file, audio file and video file. To test the big data document, by using two concepts such as big data preprocessing testing and post processing testing. To classify the data from unstructured format to structured format using SVM algorithm. In preprocessing testing test all the data, for the purpose data accuracy. In preprocessing testing such as file size testing, file extension testing and de-duplication testing. In Post Processing to implement the map reduce concept for the use of easily to fetch the data.

Index Terms*: *Preprocessing, Map reduce in Post Processing, Structured data using SVM.

I. INTRODUCTION

Big data is new forms of information processing that promotes large volume, high Speed with communication assets, improved awareness, cost effective, decision making and process automation. Data represented large quantities is nothing but Big Data. True, there is no specific size parameter that defines this technology size. This is the safe way to measure the standard route of terabytes even pet bytes. The data travels from various directions, and the speed and volume will be terrible. Data will be replaced at a faster pace and therefore require more processing, especially for social media feeds. But it is not the only medium to get information. It comes from different sources and shapes. If you go through the data you can find text files, audio files, images, video files, presentations, sensor datas, data bases and log files. It depends purely on format. It can be in any structured or unstructured format or it can be also a corrupted file. The data which are collected from the various sources like social media and digital media will be constructive and structured.It is tough to analyze the types of data. There are many types of data like we categorize under structure and unstructured. It is very difficult to analyze all types of dataThere are some flexible solutions for DBMS and RDBMS such as Oracle. The RDBMS is used for structured query language or SQL to manage, define, query, and update data. However, suppose data size is irresistible, it seems that RDBMS can handle hard, and if done, the process becomes more expensive. It proves that relational databases are not capable of managing large data and some new technologies are needed for processing the data. Customary databases are accurate for structured data and not for unstructured data. Big data contains the three characteristics such as volume/variety and velocity always called as 3V's.Volume refers to an algorithm ability to deal with a large amount of data. The scale of the data set is the quantity for the clustering algorithms related to volume property, the higher the size, the handling outlines. The data set is a collection of data set properties. Classification of features, nominal, ordinal, interval and ratio. Many clustering algorithms support numerical and classification data. In large quantities, the size of the data set increases to maintain large data, and the dimensions do not even increase. It's a curse of size. In many clustering algorithms are capable of performing setbacks. Noise data can be grouped with data points. Variety indicates the ability of a clustering algorithm to perform various sets of data sets, such as numerical, classification, nominal and ordinal. A criterion for clustering algorithms is a set of data and cluster shape type. The size of the data set is smaller or larger, but clustering algorithms support larger data sets for large data mining. In cluster shape, the set of data cluster is based on size and type shape. Velocity refers to the calculation algorithm's calculations based on the complexity of the time period of the clustering algorithm. If the algorithm's calculations are too low, nothing algorithm has less run time. The algorithms run based on the Big O Option. The Artificial Neural Network algorithm is based on a cognitive approach, namely, a neural network without the hidden layer. Although this approach could lead to poor quality in classification, it was easily selected for construction. As with the SVM model we created a perception classification for each binary combination. A node has an input layer of a node for classification. Perception has an output layer that represents a number of two categories that belong to an example given either 0 or a 1.

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author

Nachiyappan S*, Assistant Prof (Sr.), SCSE, VIT University, Chennai.
Justus S, Associate Professor, SCSE, VIT University, Chennai.

Using the full feature set rules for input layer increases the computation, but stabilizes the feature set for comparison with the SVM algorithm.

II. RELATED WORK

Big Data does not mean that it is a very large volume of data it depends upon its features and it is differentiated by the "Very large data" and "huge data". There are many definitions for big data in literature and there are some definitions which plays a very important role. Big Data is Defined by IDC in 2011 : "Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis."[1]. This explains the four characters or four V's of Big data. Volume, Variety, Velocity and Veracity of data.

Big Data is defined as datasets whose size is very huge and it cannot be adopted in a traditional database tools to do all the data processing. This is a specific definition which defines big data in terms of its context not the metric. This was discussed in Mckinsey's report 2011 NIST has defined big data in some other way like " big data is where the data acquisition data volume and velocity or variety of data limits the ability to perform the analysis on data. There are certain limitations that which are needs to be addressed before processing it". There is also some other definitions which states that"software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units" [1].

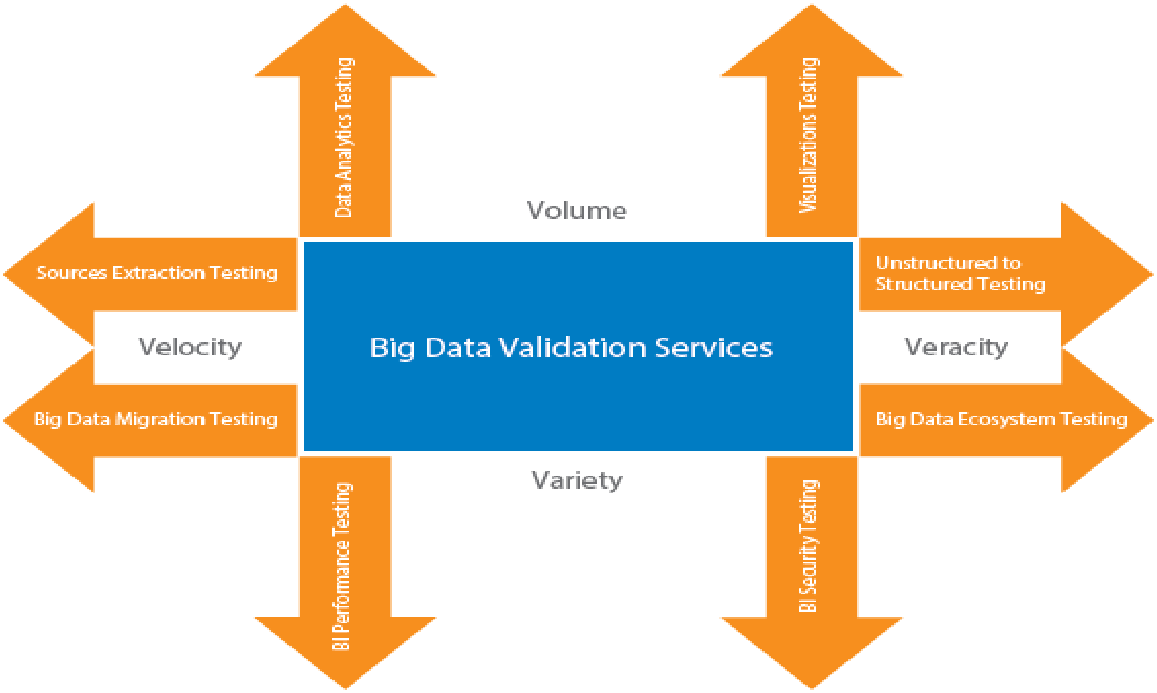
Data Analytics Testing
Visualizations Testing
Volume
Sources Extraction Testing
Unstructured to Structured Testing
Velocity
Big Data Validation Services
Veracity
Big Data Migration Testing
Big Data Ecosystem Testing
Variety
BI Performance Testing
BI Security Testing

Fig1. Big Data Validation Service

There ia s work which is carried out by an industry regarding big data testing, They have used the Big Data services for each and every V's. Here four types of testing's are done first is to test the velocity, when the data comes inside the system or storage the rate of speed which it is extracting and loading into target system. Second one is the volume testing which tests the amount of data in which the map reduce algorithms are used in specific to their business needs. Third one is the variety of data where the type of data is important to differentiate like structured or unstructured. If its unstructured data then the data has to be processed and it has to be converted into a structured format to process it. Fourth one is veracity of data where the truthiness of data is going to be the very important part as the validation and verification is concern. Fig1. Shows the big data validation services and how it is going to be processed.

III. METHODDOLOGY

A. *File Categorization using SVM Algorithm*

The file classification is a function that automatically separates the set of file extension from the classification from the predefined set. The concept of file classification is a standardized number of predefined categories or fractions. File classification can be defined as a function of automatically classifying electronic documents for their commenting classes based on their file extension. Each document is not exactly one, multiple or category. Using machine learning, learning classifications of targets, and automating those classifications automatically. This is a learning problem overseeing. Due to the overlapping of categories, each category is considered a separate binary classification problem.

Retrieval Number: A1912058119/19©BEIESP
DOI: 10.35940/ijrte.B1912.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

Classification helps to identify the correct category of domain in use, in this section I decided to divide the cloud file into four categories related to a particular file, which is split into an image file, video file, text file, and document file. For extraction. Then get the extension and classify the file extension and store it on the server. In this process we must use the SVM algorithm. SVM Algorithm Main concept classification

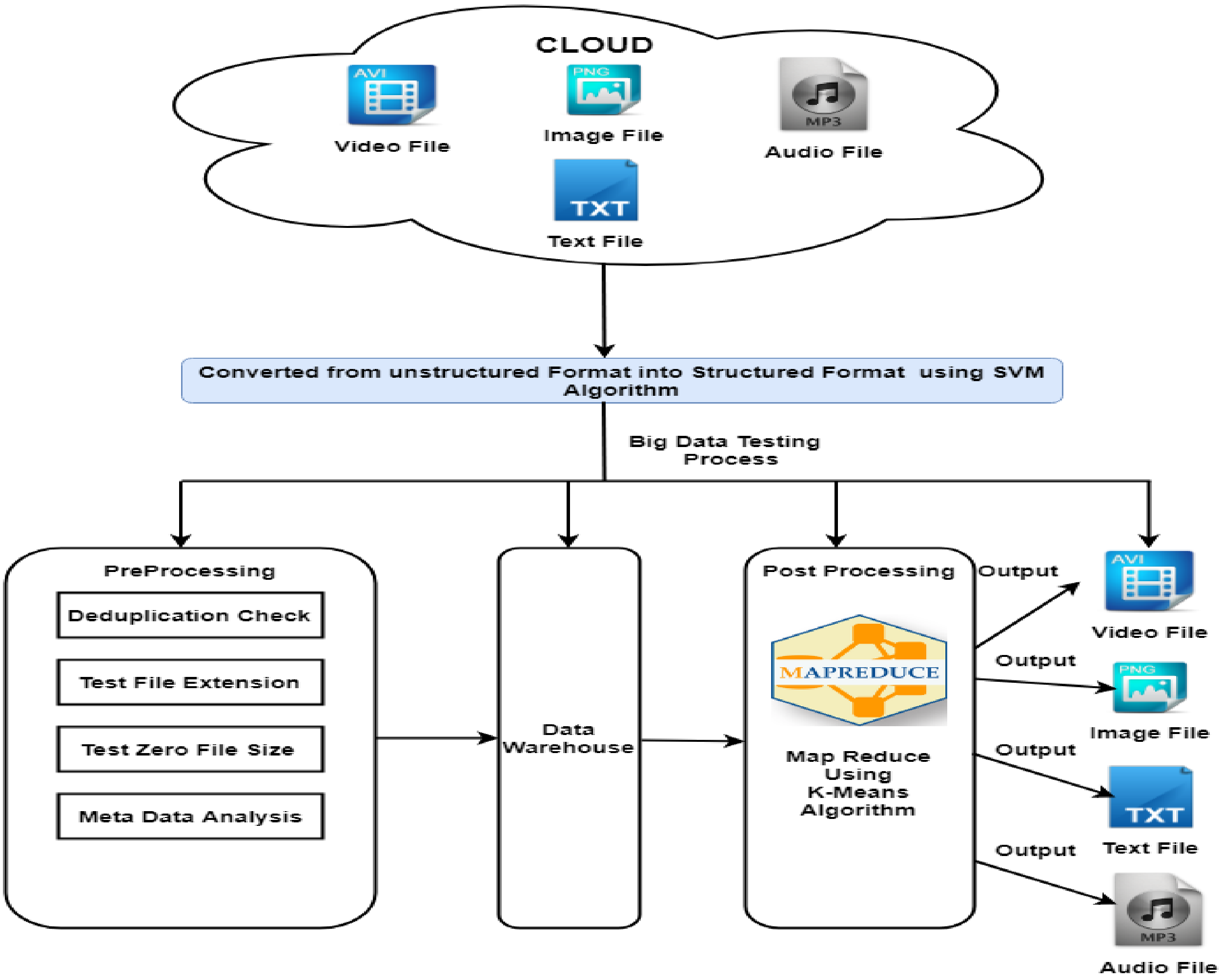

Fig 2: Overview of Big data testing

A. *De-duplication in Preprocessing Testing*

In big data preprocessing technique, we've got to check the de-duplication, zero file size, then the file extension. In de-duplication testing ,To transfer file the user and also the CSP perform each de-duplications. The de-duplication operation is a twin of that within the baseline approach. additional exactly, the user sends the file tag to the CSP for the file duplicate check. If a file duplicate is found, the user can run the POW protocol POWF with the CSP to prove the file possession. If no duplicate exists, CSP stores the cipher rtext with key and returns the corresponding pointers back to user for native storage. In de-duplication on the opposite hand of keeping the multiple information copies with an equivalent file content, de-duplication eliminates recurrent information by keeping solely single copy and referring alternative redundant information thereto single copy. The de-duplication to eliminates duplicate copies of an equivalent file. De-duplication also can be used at the block level, that eliminates duplicate blocks of information that occur in non identical files.

File size and File extension Testing

File size and file extension is the one of the pre process testing. Data has been collected from varied sources and when collection information the info the information set and uploading the data into the big information system and before process it, to validate the file is empty or not. If the file size is zero the file is not uploaded into the cloud server. Then the File extension validation helps us in many ways to confine the extension of file. In the file extension validation, to test the file size limit. For example, the image file contains some limit, if the size is exceeds it is not uploaded into the cloud

B. *Map Reduce in Post Processing*

Map reduce is that this programming paradigm that enables for large scalability across a whole lot or thousands of servers in a very big data cluster. The Map reduce is straightforward to grasp for those that area unit acquainted with clustered scale-out data processing solutions.

Map-Reduce Validation represent the checking of key-value pairs generation and validate the map-reduce by applying numerous business rules. The term Map reduce truly refers to 2 separate and distinct tasks that big data programs perform. the primary is that the map job, that takes a group of knowledge and converts it into another set of knowledge, wherever individual components area unit countermined into rows (key/value pairs). The scale back job takes the output from a map as input and combines those information rows into a smaller set of rows. In map scale back, the scale back job is often performed once the map job. The Health Care big data area unit hold on within the server. Within the user will fetch information quickly we've to use the map scale back.

Table 1. Quality Attributes of Big Data

S.No	Quality Variable	Explanation
1	Data correctness	The correctness of the data is validated with respect to format and data types.
2	Data consistency	This validated the data consistency in various angles it also refers to data gathering from various locations.
3	Data accuracy	This refers to closeness between the actual result and the expected result. Data from various sources are gathered and measured for its accuracy.
4	Data security	Security is one if the important concern which need to be addressed and validated for the applications security and its integrity in various perspectives

III. TEST PROCEDURE

When the Quality challenges for Big data is being discussed the data quality of applications are also considered. The Quality variables of enormous information applications were secret nowadays. Traditional quality factors following robustness, performance and security can be valid in big data. Now coming to big data validations and the quality challenges this work discuss about the quality and validation process of big data. On comparing to customary software testing with the big data application testing process is entirely different and they are discussed in this paper in a brief manner.

The test procedure for big data is as follows.

1) *Functional testing of big data, which includes rich test environments and domain-specific functions;*
2) *Non-function testing, includes performance, reliability, portability, Security, system consistency and Quality of Service*
3) *Big data Timing testing, checks timeliness of the system;*
4) *Big Data feature testing, targets user related system evolution and visualization*

These four steps are followed in testing the big data applications and feature testing which includes testing continuously with real time testing.

In addition the quality factors which are discussed in this paper are as follows:

Reliability:

This assures the reliability of the big data applications under some specific conditions how the system is going to perform. When a specific load is given to the system how it behaves.

Performance: How the big data applications performs in specific conditions and its also indicates about the performance of big data apps, such as availability and response time.

Correctness:

This speaks about the rightness of the big data applications.

Scalability:

Scalability is the factor which speaks about the applications flexibility to scale. In some situations it should support to scale some huge data and huge repositories and storages from period to period. In the same way that the applications scalability should be tested for its purpose.

Security:

The validation of security regarding the big data application is done here at different stages.

IV. RESULT

A. Data Accuracy

Data Quality is one of the important factor which needs to be considered when we go for any testing the first one we need to discus is data accuracy. Data accuracy is the important factor of data quality. It is the data stored in that field is correct or not. In this implementation the medical data set of sample 100000 records are taken as the test data set.

In data accuracy is higher when compare to preprocessing. After the pretesting the each cluster provides the correct accurate result. Before preprocessing the data is stored in unstructured format after preprocessing the data is formed in to structured data and its formed into different clusters. Cluster type such as image, video, document and text.

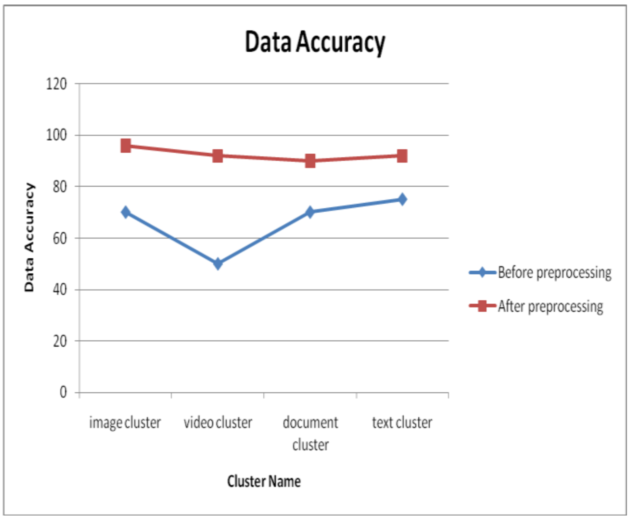

Fig 3: Data Accuracy

B. Data volume

In data volume, each cluster takes more storage space before pretesting. After that implementation of the pre testing the size of the data has been reduced. By means of de-duplication testing the duplicate data has been removed and the storage space has been reduced far better than before preprocessing. Because of the remove duplicate data, null value data and file categorization the storage space becomes low in each cluster.

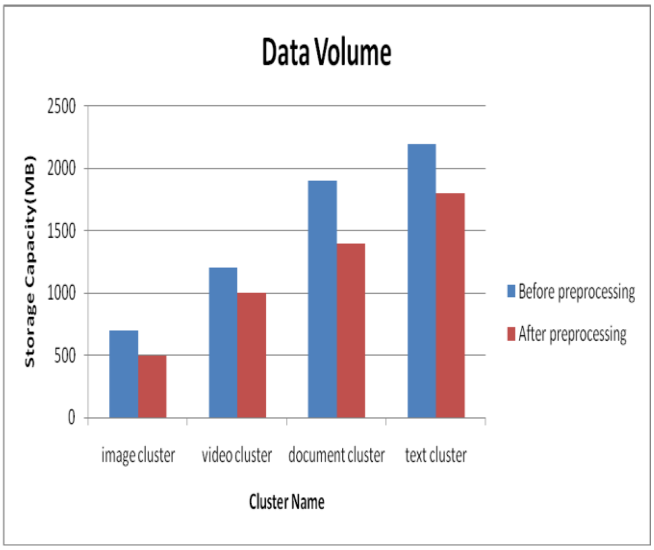

Fig. 4: Data Volume

V. CONCLUSION

Big data information is as yet advancing and analyzers and testers have a huge duty to recognize new thoughts for performing tests in the field of Big Data. A standout amongst the most testing things for an testers is to keep the pace with industry's evolving elements. In many aspects of the test, technical details behind the tester scene are unknown, but testing of Big Data Technology is quite different. There is no need to be strong in a Tester Fundamentals test, but in order to analyze many performance barriers and other problems, you need to know the minute details in the design of database designs. Big data testers should first learn parts of the big data Eco System. In this paper 10000 sample data is used entered big data in the same cluster mode. We turn out with two preprocess and post process testing results. The future work in this is to test information with numerous group frameworks.

We have to give the more accurate result by using different algorithms.

REFERENCES

1. Avita Katal, Mohammad Wazid, R H Goudar, “Big Data: Issues, Challenges, Tools and Good Practices”, IEEE, 2013.
2. Xiaoming Gao, Judy Qiu, “Supporting Queries and Analyses of Large-Scale Social Media Data with Customizable and Scalable Indexing Techniques over NoSQL Databases”, 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2014.
3. Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, “Big Data Privacy Issues in Public Social Media”, IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.
4. Vapnik (1995), The Nature of Statistical Learning Theory. Springer, Berlin
5. Burges, C.J.C. (1996). Simplified Support Vector Decision Rules. 13th International Conference on Machine Learning.
6. Pengcheng Zhang1, Xuewu Zhou1, Wenrui Li2, Jerry Gao3,4 (2017) A survey on quality assurance techniques for big data applications.
7. Quality Assurance for Big Data Applications – Issues, Challenges and Needs – Chuanqi Taq, Jearry Gao. 2016.
8. A Survey on Quality assurance techniques for big data applications, Pengcheng zhang, Xuewu Zhou, Jerry Gao, Chuanqi Tao. 2017.
9. Big Data - Testing Approach to Overcome Quality Challenges – Infosys White paper – Vol 11 no 1- 2013.
10. Big Data Testing Services, Infosys white paper – 2015

AUTHORS PROFILE

Prof. S. Nachiyappan is working in VIT University Chennai campus, Completed his PG in Anna university in 2004 and his area of research is software engineering and Big Data. He is having 5 years of Industry Experience and 10 + Years of teaching experience. He is a member of ACM professional Chapter.

Dr. S. Justus Worked in various industries as project manager and researcher, he has an over all 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.