UGC JOURNAL NUMBER 44557

International Journal of Allied Practice, Research and Review

Website: www.ijaprr.com (ISSN 2350-1294)

Big Data Testing

Supriya and Suhani Jain
B.Tech, CSE, III year,
'College of Engineering and Technology,
Mody University of Science and Technology', Lakshmangarh, Sikar, Rajasthan, India

Abstract - Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. Big Data involves various tools, strategy, techniques, testing, challenges, and advantages. Big Data is a vast field. Day by day there is some new evaluation is taking place. There are many future aspects of Big Data, and there are many field in Big Data in which we need to work. Here we will discuss the Big Data strategy, some technique, and the Big Data testing – various types of testing, challenges, future aspects and the advantages.

Keywords - *Big Data, Variety of data sources*

I. Introduction

Big Data is a collection of data or the sets of the data which are grouped together. Big Data describes the volume of the data. It is classified into three types – unstructured, structured, and semi structured. The data which are not similar type are called unstructured, the data which are of same types are grouped as the structured and the data are both structured and unstructured is called the semi structured data. Testing of Big Data applications therefore is a very necessary process to be followed, so as to maintain as well as manage the important characteristics of Big Data like Volume, i.e. data size, Velocity, i.e. speed of change and Variety of data sources.

One of the best methods which can be used for testing Big Data apps is testing with Automation.

The Big Data has 3 vs- volume, variety and velocity.

Volumn - Here volume of the Data defines and tell us that the data has been collected from various sources like business, transport, media etc. And thus a large amount of data has been collected and by that the volume increases.

Velocity - The velocity is basically the speed by which the data are transferred from one place to another. At what speed the data is being share. Is it in a real time. The basic idea behind the velocity is that the maximum data should be transferred with the maximum speed.

Variety – There are different types of data like structured, unstructured- in that there are many others like video audio, text, messaging, and email. These all comes under the section variety.

II. Big Data Testing

2.1Big Data Testing is mainly divided into three steps:

2.1.1) Pre Hadoop Testing

This testing is performed before data enters the Hadoop System. It is tested that the data that is being ingested from various sources matches the defined schema or not. If the ingested data are invalid then they are not sent for further testing.When these data enter the HDFS then they are tallied with the source data. After that they are sent to their correct location.

2.1.2) Map Reduce Testing

This testing basically ensures that the Map reduce process works correctly which involves correct key-pair value generation. It also takes care of correct implementation of data segregation and aggregation rules**.** After this testing output files are compared with input files to make sure that data is processed correctly.

2.1.3) Output testing:

This is the final stage of testing. After this the output generated is sent to the data warehouse. This test guarantees that the generated output data file has no corruption**.**

- Validate data is aggregated and merge as post Map-Reduce Jobs.
- Verify that correct data is loaded into storage system & discard any intermediate data which is present.
- Verify that there is no data corruption by comparing output data with HDFS (or any storage system) data.

2.2 Architecture Testing

Architecture Testing is one of the crucial parts of Big Data testing. Lack of proper architectural testing can cause performance degradation

There are two important tests that must be done in Hadoop environment for successful architecture testing.

Performance testing involves testing of:

1) Job completion time

2) Memory utilization

3) Data Throughput

Failover testing involves testing of:

1) Data processing even in the case of data nodes failure

2.3Performance Testing

Performance Testing for Big Data includes two main actions

- **Data ingestion and Throughout**: In this stage it is checked that how fast the system can consume data from various sources. Testing involves different types of messages that the queue can process in a given time frame.
    - In addition to this it also checks at which rate data can be inserted into the data store.
- **Data Processing**: It involves the verifying of the speed with which the queries and the map reduce jobs are being executed. It also includes testing of the data processing in isolation when the underlying data store is populated within the data sets. For example running Map Reduce jobs on the underlying HDFS is the example of data processing.
- **Sub-Component Performance**: These systems are made up of multiple components, and it is essential to test each of these components in isolation. For example, how quickly message is indexed and consumed, map reduce jobs, query performance, search, etc.

The sequence in which the Performance testing can be performed

1. Setting and building of large cluster so that we can check for the testing.
2. Design and identify all the corresponding workloads
3. Prepare individual clients (Custom Scripts are created)

4. Execute the test and then analyzes the result (If objectives are not met then tune the component and re-execute)

5. Optimum Configuration must require.

2.4 List of few tools used in Big Data Testing

Data Ingestion - Kafka, Zookeeper, Sqoop, Flume, Storm, Amazon Kinesis. **Data Processing** - Hadoop (Map-Reduce), Cascading, Oozie, Hive, Pig. **Data Storage** - HDFS (Hadoop Distributed File System), Amazon S3,

2.5Test Environment Needs

Test Environment needs depends on the various type of application that is being tested. For Big data testing, test environment should encompass

- It should have large cluster with distributed nodes and data
- It must have minimum utilization of the CPU so that the performance remains same.
- It must have a large space so that any type of data can be accommodated and there is no shortage of the memory and space.

III. Big Data challenges

As the today's world, Big Data proves to be very helpful and positive for all of us. The life and the digital world become so simple and easy. But there is also some certain challenges which the process of Big Data faces. It's not just a Small process, but actually it has to take care of certain things. There are various challenges which are as follows-

1. Security of the dataset.
2. Proper maintaining and implementation of the data
3. Control over the data
4. Listing of the data in the proper arrangement
5. Dealing with the data growth
6. Validating the data
7. Checking of data after the regular interval of time
8. Deleting and modifying the data

9. Organize the data

10. Proper output

IV. Conclusion

Big Data testing is one of the main aspects. It is important thing which we need to know. The testing of Big Data is functional and non functional. Based on this the testing occurs. Moreover the testing is not so simple, there are many challenges and the different parameters which we need to take care. Without that the testing won't be possible. The testing includes many constraints also. There are many other fields on which there is some need of improvement. We can make the improvement as future aspects. More and more data should be structured and also automation theory can be used. Automation is one major approach in testing Big

Data testing, if the automation setup is built then the mechanism become more and more advanced.

V. References

1. **https://www.qubole.com/resources/big-data-challenges/**

2. **https://www.guru99.com/big-data-testing-functional-performance.html**

3.https://techbeacon.com/resources/continuous-testing-hewlett-packard enterprise?utm_source=tb&utm_medium=article&utm_campaign=inline-cta

4. **https://www.qualitestgroup.com/white-papers/iot-testing-the-big-challenge/**

5. **https://afourtech.com/iot-testing-services/**