ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

BIG DATA VALIDATION NEEDS AND CHALLENGES

S.Nachiyappan[1], S.Justus[2]
[1,2]SCSE, VIT University, Chennai, India.
Email: nachiyappan.s@vit.ac.in

Received on: 18.10.2016 *Accepted on: 11.11.2016*

Abstract:

Big Data is a big topic in software development today. When it comes to practice, software testers may not yet fully understand what Big Data is exactly. What testers do know is that you need a plan for testing it. The problem here is the lack of a clear understanding about what to test and how deep inside a tester should go. There are some key questions that must be answered before going down this path. Since most Big Data lacks a traditional structure, what does Big Data quality look like? And what are the most appropriate software testing tools? Many of us improperly believe that Big Data is just a large amount of information. This is a completely incorrect approach. For example, a 2 petabyte Oracle database alone doesn't constitute a Big Data situation – just a high load one. To be very precise, Big Data is a series of approaches, tools and methods for processing of high volumes of structured and (most importantly) of unstructured data.

Introduction

The Internet is filled with a lot of information on what big data is, the tools that are used to capture, manage and process big data sets and its characteristics such as Volume, Variety, Velocity and Veracity. However, there is limited content available when it comes to devising a test strategy for big data applications or how big data needs to be approached from a testing point of view.The traditional tester wrote simple read/write queries against the database to store and retrieve data. Slowly, the size of data started increasing due to business needs and newer technologies like Data Warehousing mandated specialized skills which created a whole lot of designations within the tester community who were referred to as "Database Testers", "ETL Testers" and "Data Warehouse Testers". Now with the advent of big data, things are only getting from complex to worse for the tester. What should the tester expect from big data? What kind of challenges does it put forth to the tester? Does he need new skills?As mentioned in the beginning, the internet has so much of definition and explanation about what Big Data is but very limited information about testing it. In this article let's try and understand how traditional data processing is different from processing large data sets and look at how testers can approach them. Apparently, traditional data processing dealt with large data sets but no as huge as what we have now. As Wikipedia puts it, “The world’s technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s as of 2012, every day 2.5 exabytes of data were created. That kind of size is not easy for RDBMS to handle and that is where libraries like Apache Hadoop helps. Mike Olson, the CEO of Cloudera says, “The Hadoop platform was designed to solve problems where you have a lot of data – perhaps a mixture of complex and structured data – and it doesn’t fit nicely into tables. It’s for situations where you want to run analytics that are deep and computationally extensive, like clustering and targeting”. As we all know, reading from the disk is much slower than reading from RAM (Random Access Memory) and that is how traditional data processing works.

This is not suitable when processing huge data sets. Hadoop helps here with its HDFS (Hadoop Distributed File System), which lets you store large amount of data on a cloud of machines. On top of HDFS, Hadoop provides an API to process the stored data which is Map-Reduce. The idea is since the data is stored in a distributed manner across nodes, it can be processed well in that manner where each node can process the data stored on it instead of getting hit by performance degradation issues by moving it over the network. The last step in the process is to extract the data output from the second step and loading them into downstream systems which can be data warehouses of other systems that might use the data for further processing. Now, testers who have worked with Data Warehousing (DWH) applications can immediately relate to the ETL (Extraction, Transformation and Load) model when they read about the 3 stages of data processing. What they need to keep in mind is, while the processing sounds similar, the strategy deployed to test DWH applications (e.g. sampling when doing manual testing, using automation tools for testing) may not be very useful if tried as silos in big data implementations for the reasons given below. However, understanding and past experience working on DWH applications can definitely help the tester plan big data testing better.

- Huge size of Data sets
- Data Variety(Web logs, sensor networks, Social media networks, photo/video images)
- Unstructured vs Structured Data(DWH supports structured data while Big data supports both)

In the first stage which is the pre-Hadoop process validation, major testing activities include comparing input file and source systems data to ensure extraction has happened correctly and confirm that files are loaded correctly into the HDFS (Hadoop Distributed File System). There is a lot of unstructured or semi structured data at this stage.

The next stage in line is the map-reduce process which involves running the map-reduce programs to process the incoming data from different sources. The key areas of testing in this stage include business logic validation on every node and then validating them after running against multiple nodes, making sure that the map reduce program/process is working correctly and key value pairs are generated correctly and validating the data post the map reduce process. The last step in the map reduce process stage is to make sure that the output data files are generated correctly and are in the right format. The third or final stage is the output validation phase. The data output files are generated and ready to be moved to an EDW (Enterprise Data Warehouse) or any other system based on the requirement. Here, the tester needs to ensure that the transformation rules are applied correctly, check the data load in the target system including data integrity and confirm that there is no data corruption by comparing the target data with the HDFS file system data.

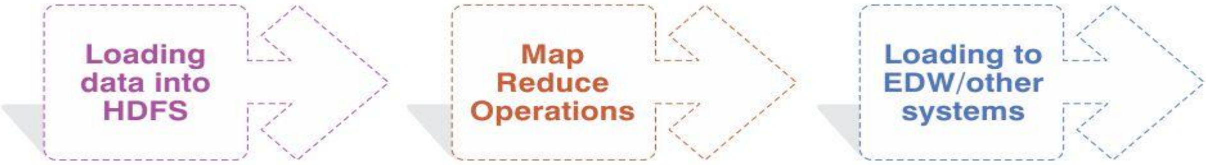

Fig.1. Stages of data processing

Testing Approach:

Big data is still emerging and a there is a lot of onus on testers to identify innovative ideas to test the implementation. Testers can create small utility tools using excel macros for data comparison which can help in deriving a dynamic structure from the various data sources during the pre Hadoop processing stage.

For instance, Aspire designed a test automation framework for one of our large retail customers' big data implementation in the BDT (Behavior Driven Testing) model using Cucumber and Ruby. The framework helped in performing count and data validation during the data processing stage by comparing the record count between the Hive and SQL tables and confirmed that the data is properly loaded without any truncation by verifying the data between Hive and SQL tables.

Similarly when it comes to validation on the map-reduce process stage, it definitely helps if the tester has good experience on programming languages. The reason is because unlike SQL where queries can be constructed to work through the data MapReduce framework transforms a list of key-value pairs into a list of values. A good unit testing framework like Junit or PyUnit can help validate the individual parts of the MapReduce job but they do not test them as a whole.Building a test automation framework using a programming language like Java can help here. The automation framework can focus on the bigger picture pertaining to MapReduce jobs while encompassing the unit tests as well. Setting up the automation framework to a continuous integration server like Jenkins can be even more helpful. However, building the right framework for big data applications relies on how the test environment is setup as the processing happens in a distributed manner here. There could be a cluster of machines on the QA server where testing of MapReduce jobs should happen. This paper is written todiscuss the urgent issues in big data quality assurance andquality services. Many practitioners in the real field arelooking for the answers to the following questions.

 What is the big data quality? Is the same as data quality?

 What are the key factors to affect big data quality?

 Why do we have poor big data quality?

 How to validate big data quality?

 Where is the big data quality assurance program?

 What are the challenges and needs to control big data quality?

Challenges faced in leveraging big data.

There are various challenges faced while leveraging Big Data they are as follows:

1. Lack of proper human resource: Proper human resource is required for leveraging Big Data. As in data analysis specialists are required who are good in business understanding and are capable of dealing with large quantity of data. There are some common mistakes made by companies in hiring people for working on Big Data they are – companies prioritize candidates having good industry experience over data analytics experience and may have insufficient skill set for the desired role. [1]

2. Too much stress on technical aspects of Big Data than on analytics: This has been the biggest challenge in building data driven culture as most of the companies approaching Big Data are focusing data collection and storage rather than communication tools and data analysis. And involvement of human power is required for conducting data analysis and bringing it into action. [1]

3. Misalignment between Big Data And Company's visions: This is the most common challenge faced by an organization while trying to exploit Big Data. Here there is a misalignment between company's vision and goals and the role of data in achieving those goals, required investments and who is responsible for this investment. This can cause scattering of Big Data initiatives.[1]

4. The extent of Operational transformation required: Transformation is a comprehensive change in a data centric company that requires adjusting of operational processes so as to realize benefits of data. But as these operational implications and its transformations are not properly understood this may lead in inefficient capturing of insights by the organization and may even result in cancellation of other viable Big Data initiatives.[1]

5. Improper estimation of value drivers: Sometimes companies try to start big which results in improper estimation of value drivers how they can be reached which may lead there whole initiative to fail.[1]

Need for Big data Testing

Big Data is a collection of large and complex data sets. So in dealing with this huge amount of data and executing it on multiple nodes there is a higher risk of having bad data and even data quality issues may exist at every stage of the processing [1]. Some of the needs are:

1. Increasing need for integration of huge amount of data available: With multiple sources of data available, it has become important to facilitate integration of all data. But this integration forces organization to have constantly clean and reliable data and this can be ensured through end to end testing of all the data sources available.

2. Instant Data Collection and Deployment issues: In today's world data collection and live deployment is very important to meet business needs of an organization. But challenges like proper data collection etc. can be overcome only by testing the application before live deployment.

3. Real-time scalability issues: Big Data Applications are built to match the level of scalability involved in a given scenario. Critical errors in the architectural elements governing the design of Big Data Applications can lead to worst situations. So hardcore testing involving data sampling techniques coupled with high end performance testing capabilities are essential to meet the scalability problems that Big Data Applications poses.

Data functional testing basically helps in identifying data related issues which may be caused either due to some node configuration errors or may be due to some code errors. Testing at all phases ensures that data being processed is completely error free. Functional testing provides validations at all phases of data processing. Apart from functional testing, non-functional testing like failover testing, performance testing and security testing are also needed to be performed on data to validate the quality of data and its performance as the quantity of data is quite huge.

Challenges-Big data Testing

1. Performance: One characteristic of Big Data is that the data is highly volatile, which is more often unstructured in format generated from various sources such as web logs, sensors embedded in devices, GPS systems etc .and therefore any conventional Performance Testing/Performance Engineering concepts are of no good. What hasn't changed through is the Business need for making best possible decisions quickly. To go through and access detailed information needed from this huge volume of data that too at a very high speed with increase in the degree of granularity makes this challenge worse.[3]

2. Scalability: Scalability is the ability of a system to handle larger workloads by enlarging the system in a straightforward manner. In practice, however, it is often impossible to plan for the scenarios that will most benefit from highly scalable systems. Workloads can drastically expand due to business growth, new application features, and usage patterns To work with this huge volume of data requires distributing parts of the problem to multiple machines/nodes to handle in parallel, data should be able to scale very rapidly and quickly across multiple data centers and cloud as well if needed. Whenever/wherever multiple machines are used in cooperation with one another, the probability of failures increases. In case of multiple-machine environment, failure is the main thing about which developers worry. If the machine has crashed, then there is no way for the program to recover. Also Synchronization between multiple machines remains the biggest challenge. [3]

3. Continuous Availability and Data Security: In today's world when organizations rely on data to generate revenues for business applications, data should never go down. It should be always available. In some NoSQL systems a certain amount of downtime is built in by default. Thus some approach is needed to overcome this challenge. Big Data contains massive amount of information, which may also contain personal ID information, account details, credit card data and some other sensitive information. Thus security of this sensitive information is needed. It possesses a major threat or challenge to secure all sensitive data because of this huge volume of data available. And most of the NoSQL Big Data solutions have very few mechanisms for securing Big Data. [3]

4. Meeting speed of data, understanding it and addressing data Quality: In today's competitive environment it is required that company should be capable of quickly analyzing and finding out relevant data needed by them for business purpose. Understanding Big Data is a big challenge as it is needed to get the data in right shape as required so that visualization can be used for data analysis. Addressing data quality is again a major challenge. As even if we can find and analyze data quickly it should be in proper format and context so that the targeted audience can easily consume it. If quality of data is not proper or accurate then it will affect decision-making capabilities of an organization. [3]

5. Node Failure: As in Big Data environment in case of Hadoop, Cassandra etc. architecture data is distributed across various nodes so there are chances of node failures like name node failure, data node failure and network failure. These node failures need to be handled carefully so as to prevent data from being unavailable even due to single node failure as it may result in big loss to Organizations depending on that data. Thus all Nodes of the architecture need to be tested against any kind of failure. [3]

Big data quality validation process

This section discusses the validation process for big data services. As shown in Figure 2, there are five types of big data services. Here we summarize the commonly used services: a) data collection, b) data cleaning, c) data transformation, d) data loading, and e) data analysis. For multisource, data need aggregation before loading. The detailed illustration for data service is as follows.

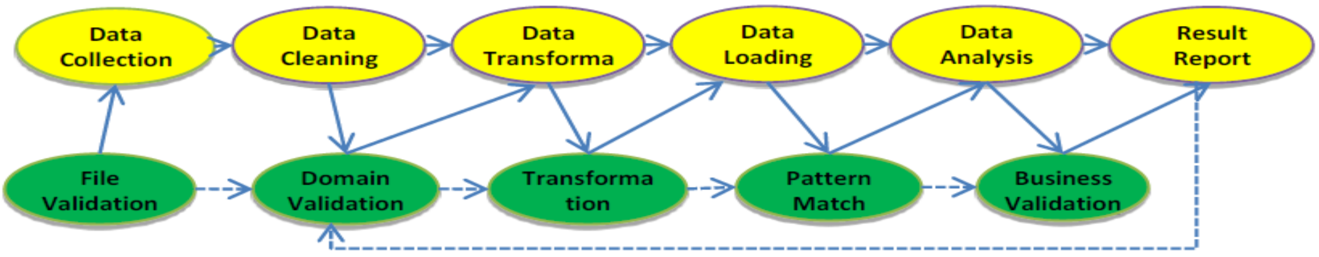

Fig. 2.Single Data Source Validation Process.

- **Data collection** is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.

- **Data cleaning** is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. The major purpose is to detect and identify incomplete, incorrect, inaccurate, irrelevant, data parts in data sets and they can be replaced, modified, or deleted [9].

- **Data transformation** converts a set of data values from the data format of a source data system into the data format of a destination data system [10].

- **Data Loading** refers to data loading activities in which data are loaded into a big data repository, for example, a Hadoop environment or a NoSQL Big database. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals—for example, hourly.

- Data Analysis- a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains [10].

- **Data Aggregation** refers to the compiling of information from databases with intent to prepare combined datasets for data processing [11].

Big Data Validation Tools

Recently, numerous data validation tools are developed by different vendors. According to [12], 42% organizations use some sophisticated data management tools, but don't have specified data-oriented roles or business-wide technological tools in place. Table 1 lists a selected popular commercial big data validation tools. These tools have been compared in terms their operational environments, supported data sources, data validation, and current successful applications.

S.No.	Name of the tool	Advantages
1	Datameer	Datameer offers the first data analytics solution built on Apache Hadoop that helps business users access, analyze and use massive amounts of data. Founded by Hadoop veterans in 2009. , Datameer Analytics Solution (DAS), provides unparalleled access to data with minimal IT resources. DAS can scale to 4,000 servers and petabytes of data, while also delivering low TCO. Datameer is based in San Mateo, California
2	Talend Open Studio	Talend software for big data enables data-driven businesses to immediately begin to integrate data from historical, live and emerging data sources. By simplifying the development skills needed to handle big data, Talend gives organizations the tools to instantly unlock the value in their data without needing to invest in new architecture or training for developers. Ultimately, instead of focusing on how to access and integrate data, Talend lets enterprises focus on everything they can do with the data to gain competitive advantage.
3	informatica	Informatica Big Data Integration enables you to easily and more quickly integrate more data from more data sources:

		• Ingest data at any speed • Process data quickly with flexibility and repeatability • Deliver data anywhere Informatica Big Data Security discovers and classifies data to drive a comprehensive 360-degree view of the proliferation, usage, provenance, and protection of sensitive data so you can: • Classify sensitive data with 360-degree visibility • De-identifies data so it can be safely used in development and production environments • Ensures compliance with corporate policies and industry regulations
4	IBM Query Surge	Ensure data quality with QuerySurge. QuerySurge is the collaborative data testing solution that finds bad data in Big Data and provides a holistic view of your data's health. It ensures that the data you extract from sources remains intact in the target by analyzing and quickly pinpointing any differences in your Big Data at every touchpoint. • Easily **automate** your manual testing effort for repeatability • Provide testing across **different platforms** – Hadoop, MongoDB, Oracle, Teradata, IBM, Microsoft, Cloudera, HortonWorks, Amazon, DataStax, MapR, all of the other Hadoop and NoSQL vendors, flat files, Excel, web services and XML • **Speed up testing** up to 1,000 x while providing up to 100% data coverage • Continuous Delivery - integrates an out-of-the-box **DevOps solution** for most Build, ETL & **QA management software** • Deliver shareable, automated email **reports** and data health dashboards • Provide a huge Return On Investment **(ROI)**, as much as 1,600%
5	Microsoft Azure HDInsight	• A managed Apache Hadoop, Spark, R, HBase, and Storm cloud service made easy • A Data Lake service • Scale to petabytes on demand • Crunch all data—structured, semi-structured, unstructured • Develop in Java, .NET, and more • Skip buying and maintaining hardware

		• Spin up Apache Hadoop, Spark, and R clusters in the cloud • Use Excel or your favourite BI tool to visualise Hadoop data • Connect on-premises Hadoop clusters with the cloud
6	SAP HANA	SAP HANA is a distributed computing solution for business. It leverages and extends the Apache Spark execution framework to provide enriched interactive analytics on enterprise and Hadoop data.

Conclusion

There are a number of major challenges and needs in big data quality validation and assurance. Here are typical ones. Lack of awareness and good understanding of big data quality validation and assurance. For building enterprises with useful big data, we must first establish quality big data team by providing them well-defined big data quality assurance program. Lack of well-defined enterprise-oriented big data quality assurance standards and programs in assisting data quality validation, control, and management. Define and develop well-defined big data quality. Validation and assurance standards and programs for enterprises Lack of available research results on big data quality models and quality evaluation metrics. Define and develop big data quality models and metrics for big data application services so these will be useful to assess big data quality and related usability Lack of well-established big data certification program and standards Establish big data quality certification programs and standards to ensure and evaluate big data quality as a third party for vendors, enterprises and the public.

References

1. White Paper - Testing the Giant : Testing Big Data - **Neha Maheshwari, Prateek Chaturvedi –** Infosys – Hyderabad.
2. White Paper - Big Data Testing : Best Practices – TechArchis
3. White Paper – A Primer on Big Data Testing - QA Consultants.
4. https://conference.eurostarsoftwaretesting.com/2013/testing-in-a-big-data-world-2/
5. http://blog.aspiresys.com/testing/approach-to-big-data-testing-part-2/
6. http://www.softwaretestingmagazine.com/knowledge/big-data-how-to-test-the-elephant/
7. https://www.quora.com/What-are-the-best-methods-for-testing-big-data-applications
8. http://www.querysurge.com/solutions/testing-big-data

9. W. W. Eckerson, Data quality and the bottom line: Achieving business success through a commitment to high quality data‖. Data, Warehousing Institute, 2002.

10. J. W. Osbourne, ―Notes on the Use of Data Transformation‖, Practical Assessment, Research & Evaluation, 2002, 8(6): n6.

11. I. Taleb, R. Dssouli R, and M. A. Serhani, ―Big Data Pre-processing: A Quality Framework‖, IEEE International Congress on Big Data, Pages: 191.

12. David Loshin (Author), The Practitioner's Guide to Data Quality Improvement (The Morgan Kaufmann Series on Business Intelligence), 1st Edition, Morgan Kaufmann; 1 edition, October 29, 2010.

13. https://www.talend.com/resource/big-data.html

14. https://www.informatica.com/in/products/big-data.html

15. http://www.querysurge.com/solutions/testing-big-data

16. https://azure.microsoft.com/en-in/services/hdinsight

17. Jerry Gao, Chunli Xie, Chuanqi Tao – March 2016 Big Data Validation and Quality Assurance – issues, Challenges and Needs.