Trustworthiness of Big Data

Akhil Mittal
Technical Test Lead
Infosys Limited

ABSTRACT
Big data refers to large datasets that are challenging to store, search, share, visualize, and analyze and so the **Testing**. Information is emerging at volatile rate, coming into organization from diverse areas and in numerous formats. Traditional DW testing approach is inadequate due to Technology Changes, Infrastructure (DB/ETL on Cloud) and Big Data. Big Data validation is not only around validation of just what is different; it's also about validation of new integrated components to what you already have.

There is unique testing prospects exists as poor data quality is still a major and exponentially growing problem. It's a digital world, which is causing massive increases in the volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources) of data. As a result, concern for realistic data sets, data accuracy, consistency and data quality is now a critical issue. The paper tries to explore testing challenges in Big Data adoption and outline a testing strategy to validate high volume, velocity and variety of information.

General Terms
Big Data, Data warehousing, Testing.

Keywords
ETL, Hadoop, big data validation, big data testing

1. INTRODUCTION
A joint report by NASSCOM and CRISIL Global Research & Analytics suggests that by 2015, Big Data is expected to become a USD 25 billion industry, growing at a CAGR of 45 per cent. Managing data growth is the number two priority for IT organizations over the next 12-18 months. In order to sustain growth, enterprises will adopt next generation data integration platforms and techniques fueling the demand for Quality Assurance mechanisms around the new data perspectives.

To keep up with pace, organizations need to consider on their big data validation approach and laid it in a strategic manner.

What exactly is Big Data? It refers to data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. Big Data has vast characteristic and few of them are showed below

Figure 1: Characteristics of Big Data

There are four fundamental characteristics on which organizations need to define their test strategy. These are:

- Data Volume
- Data Variety
- Data Velocity and
- Data Veracity

All these characteristics support to understand data flow and defining data quality

Characteristics	Description	Attribute	Driver
Volume	Sheer amount of data generated/ingested/analyzed	World's "digital universe" is in the process of generating 1.8 Zettabytes of information - with continuing exponential growth – projecting to 35 Zettabytes in 2020	• Increase in data sources • Higher resolution sensors
Velocity	How fast data is being produced and changed and the speed with which data must be received, understood, and processed	**Accessibility**: Information when, where, and how the user wants it, at the point of impact **Applicable**: Relevant, valuable information **Time value**: real-time analysis yields improved data-driven decisions	• Increase in data sources • Improved thru-put connectivity • Enhanced computing power of data generating devices
Variety	Information coming from new sources both inside and outside the walls of the enterprise	**Structured** **Unstructured** –unstructured or human generated information **Semi-structured** **Complexity** – where data sources are moving and residing	• Mobile • Social Media • Videos • Chat • Genomics • Sensors
Veracity	The quality and origin of received data	The quality of Big Data may be good, bad, or undefined due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	• Data-based decisions require traceability and justification

Figure 2: Big Data Characteristics, Attributes and Driver

Every second, data is coming in large volume from a wide variety of applications/interfaces in different formats and rapidly changing. The goal is Veracity, measure of trustworthiness and accuracy of data.

Organization data contains valuable information and this valuable information is not limited to internal data. For effective analytical business intelligence, organizations need to understand. Analyze to meet their sales, productivity etc. goals. Having said that, analyzing this valuable information is not easy as it seems. There are tools available to extract and process this valuable information from disparate sources but the real issue is to know if process data is trustworthy, accurate and meaningful?

Many organizations have already adopted big data technology and reaping exponential benefits but some organizations are still not sure of big data truth.

2. DATA QUALITY

Data Quality is a relative term and holistically handling a data quality issues is a big challenge for any organization and it is even more challenging in a sense that people have different notation about data quality while defining and validating data. There are essentially three types of data quality

- Data Validity: Does the data do what it is supposed to do?
  - ✓ Orange is a name of company as well as fruit
- Data Completeness/Accuracy: Is the data good enough to do the business?
  - ✓ Is all processed data needed to make a decision
  - ✓ Is the data upon which calculations performed trustworthy?
- Data Consistency: Does the data do the same thing all the time?
  - ✓ Is the same set of data captured and stored in different branches for different customers?

Once organizations are clear of data flow, what is required and expected from different systems and have clear understanding of data, they would be able to define effective test strategy. When it comes to Big Data, organizations are struggling to build required infrastructure, expertise and skill set to leverage big data to positively impact their business. Unstructured data, complex business rules, complex implementation algorithms, regulatory compliance and lack of standard big data validation approach puts a tremendous pressure on independent testing team to effectively prepare for a big test effort. There is clearly a need to define data validity, data completeness and data consistency for big data validation framework to accomplish the goals and validate data is trustworthy, accurate and meaningful.

3. CHALLENGES AND OPPORTUNITIES

Organizations have been making business decisions based on transactional data stored in relational databases. To derive business value from non-traditional, unstructured and social media data, organizations need to adopt right technology and infrastructure to analyze the data to get new insights and business intelligence analysis. Having said that, to make sure the data used for this purpose should be complete, trustworthy, consistent and accurate.

Validating Big Data is one of the biggest challenges faced by organizations because of lack of knowledge on what to test and how much data to test as incoming data volumes are exploding in complexity, variety, speed and volume. Organizations had the data but, did not have the information. The following are some of the needs and challenges that make it imperative for Big Data applications to be tested thoroughly:

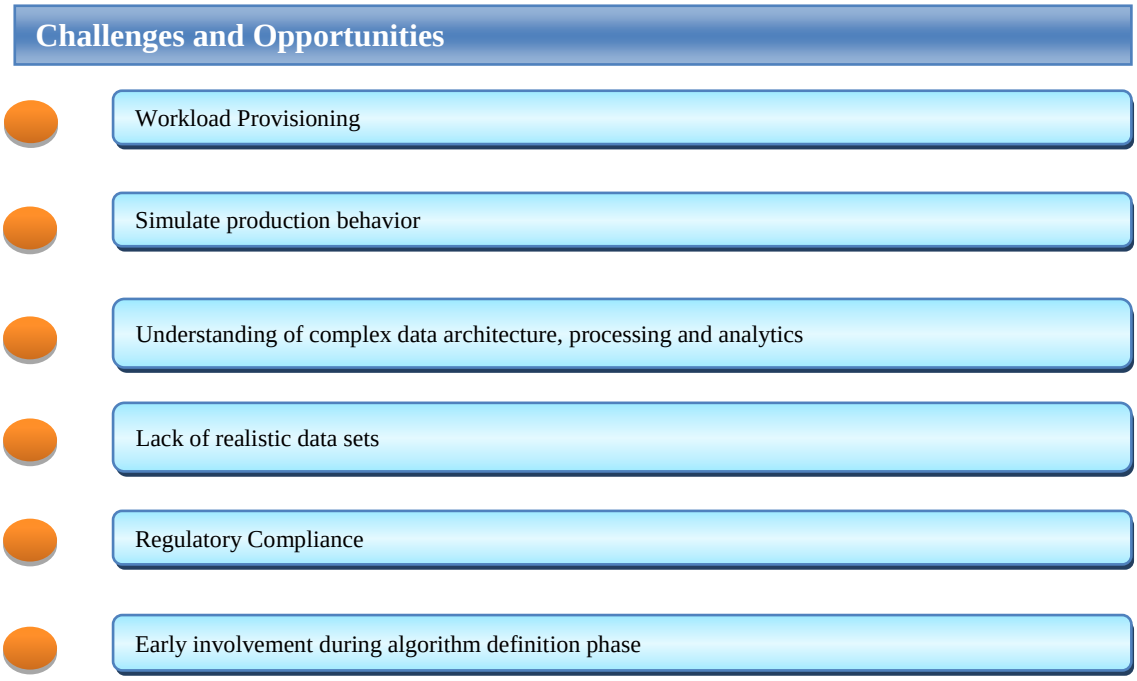

Fig 3: Challenges and Opportunities

3.1 Know and Understand your Data

Understanding data and relationship is one of the important tasks before going for any test strategy or scenarios. It is not easy to decide how much to test and most important what to test. Data volume is huge, coming on very fast pace and validating this data volume and velocity is quite a big challenge. If validation team able to understand the data and value of data from business perspective, it will be a win win situation.

- Architecture, development and independent testing team need to explore the source systems and its data, identify patterns and relationships that could exploit in testing.
- Understand data relationship and business rules
- Derive the structure dynamically from data sources
- Understand statistical correlation between different sets of data and their relevance for business users

3.2 Workload Provisioning

Traditional benchmarks focus on simple workloads but Big Data architecture differs from standard relational database systems with respect to data and workloads. Traditional benchmarks used by the database community are inadequate but for evaluating and testing big data systems, more complex workloads that involve machine learning algorithms, statistical correlation and information extraction are required. A large problem for benchmarking and testing of big data systems is the lack of realistic data sets.

3.3 Sentiments v/s Test

Unstructured data sources (such as text documents, tweets, social media posts, sensor data etc.) provide a rich supplement to the data feed into big data system. The biggest thing with understanding unstructured text is that it's not just necessarily about the text, but the context (sentiments). E.g. Consumer are discussing about new phone launch and sharing their feedback. We need to capture and process those sentiments and convert them into meaningful data sets for effective business analysis and decision making. The real question is, is the final decision correct? Is the data on which analysis has done trustworthy? Organizations need to define an effective test strategy to validate these sentiments

3.4 Compliance and Security

Poor data affects the ability to comply with government regulations such as Sarbanes-Oxley, Basel II and HIPAA. Lack of compliance can lead to unnecessary fines and levies. The compliance should be defined during requirements and implemented as part of big data technology.

Due to increasing volume of unstructured data, organizations has a unique challenge to address the rules, compliance and protecting the data at different levels. Are organizations keeping tabs on the regulated information in all that mix? It gives a unique challenge and opportunity to independent testing team to validate compliance issues. Independent testing team needs to think to automate compliance and security validation process else validating this huge data against all compliance is very tedious and complex.

- Who all in the organizations have access to personal identification information?
- Does regulatory compliance say HIPAA applicable to the entire data or subset of data?
- Is data coming from social media sites secured?

3.5 User Behavior

We don't know what final end users are expecting unless we collaborate with them and include them into reviewing out test scenarios and test approach. In most cases the independent testing team will not have access to end users. This is a core challenge as the tester herein simulates user behavior without much interaction with them. Look for data sampling wherever possible, including any call logs, user feedback and usage patterns, and collaborative filters from previous releases, keeping in mind that ongoing brainstorming with the product team is important in arriving at test scenarios and an optimized test matrix

3.6 Effective Collaboration

It's a new technology and everyone is on learning curve to understand the algorithm behind big data processing. As part of independent validation team, collaborate with developers and architects during early development lifecycle is very important to learn big data technology and gain a leading edge. While traditional RDBMS testing is not going away, when you get to big data testing you are talking about file systems, dynamic and unstructured data, and newer technology jargon, concepts, and implementation strategies such as HADOOP, NoSql, and Map Reduce.

Testing team has tremendous learning as well as validation opportunity during algorithm definition phase. During algorithm definition phase, testing team needs to learn how data will be extracted from heterogeneous sources, what is the pre and post processing algorithm and is there any filtering/transformation logic before data goes to data warehouse or analytical tool.

4. PROPOSED TEST STRATEGY AND SOLUTIONS

4.1 Introduction

Irrespective of Data Warehouse Testing (DWT) or a BIG Data Testing, the elementary factor that's of interest the independent validation team, is the '**Data**'. At the fundamental level, the data validation in both these cases involves validation of data against the source systems, for the defined business rules. It's easy to think that, if we know how to test a data warehouse, we know how to test the BIG Data storage system. But, unfortunately, that is not the case!!

Data management and data quality principles for big data are the same as they have been in the past for traditional data. But priorities may change, and certain data management and data quality processes, such as metadata, data integration, data standardization and data quality must be given increased emphasis.

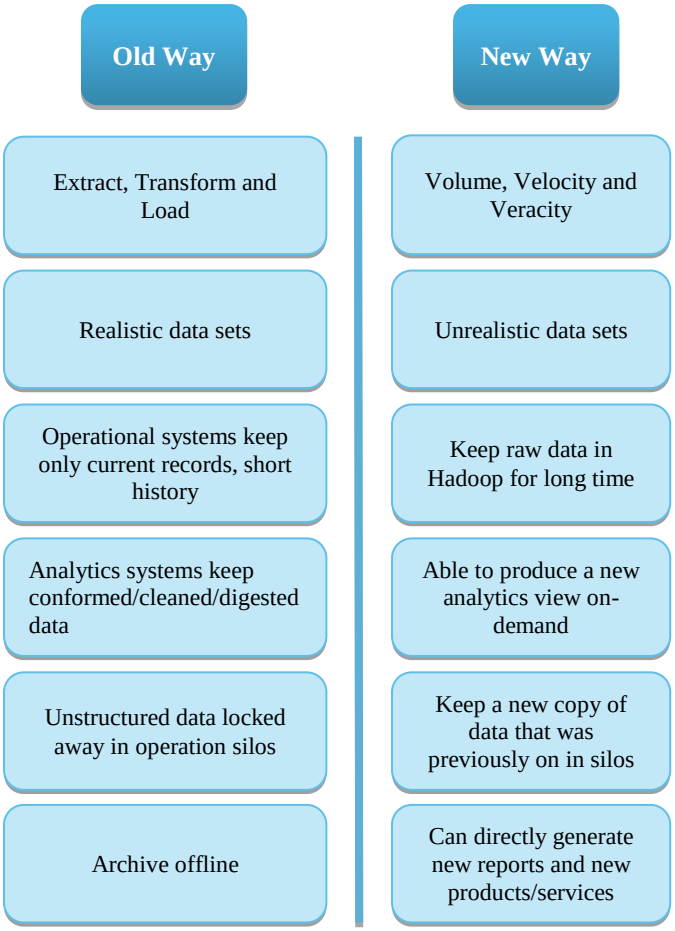

Fig 4: Data warehouse Testing v/s Big Data Testing

In the world of big data, where data may be used in ways not originally intended, data elements need to be defined, organized, created and tested in a way that maximizes long term potential use and does not hinder future usefulness.

Before defining low level test strategy, attention should be given to

- ❖ Understand data
- ❖ Understanding the Current Business Process Flows
- ❖ Break down the analyzed information into data elements and
- ❖ Draw your validation strategy

Understand Data:

Data understanding is very essential before laid down test data validation strategy. It includes understanding of

- Where the data came from
- Types of data
- How the data will be used
- What are the workload types
- Who will use the data and
- What decisions will be made with the data
- Which data comes under compliance and security

Understanding the Current Business Process Flows:

Once independent testing team has understanding of data and data flow, Begin understands the current-state operations environment with an analysis of the information required to support the processes. Asking questions such as

- What information is an input to the process?
- What information is changed or created during the process?
- What happens to the information once the process is complete? and
- How is the output distributed along down-stream processes?

Break down the analyzed information into data elements:

Once the process information is analyzed broke down into required data elements. The resultant answers provide a baseline set of data on which to begin to examine data quality. Unstructured data, for instance, may be broken down and condensed in a structured format.

Draw your validation strategy:

- Define data quality benchmarks to ensure the data will be fit for its intended use
- Define the key data qualities attributes to be measured
- Capture the data lineage and traceability (for example, data flows from the underlying business processes) aspects
- Define new principles and processes for big data scenarios and see what all traditional methods will be relevant
- List all tools and techniques that can help prevent big data quality issues
- Plan for environment configuration
- Plan for integration and end to end testing

Before all this, make a specific mindset, skillset and deep understanding of the technologies and practical approaches to big data world.

5. TEST APPROACH AND STRATEGY

5.1 Requirement Elicitation

Now days, organizations recognize the value of big data but problem is with business requirements. Big data requirements are very different than transactional systems requirements.

Requirement elicitation is an important task while defining the big data test strategy. Independent validation team should be able to understand end user requirements and how end users are going to use data in decision making. During this stage, validation team should also understand and develop big data reporting strategy.

- Collaborate with business users and get business concerns from business users/end users
- Create a solution roadmap on how validation approach will address their concern
- Map requirements against various data sources and final reports, identify data rules and patterns
- Define how interfaces, workflows, transformations
- Outline a Risk Based Approach to define Data Risk, Business Risk and Legal Risk against each requirement and get them approved from business users
- Capture and list down how non-functional requirements, negative scenarios and exception scenarios
- Last but not least, illustrate how requirements are defined to make sure data going out from big data system is compliant to various compliance and security.

Before moving into requirement elicitation phase, understand how big data is integrated or going to be integrated into current business and analytics processes.

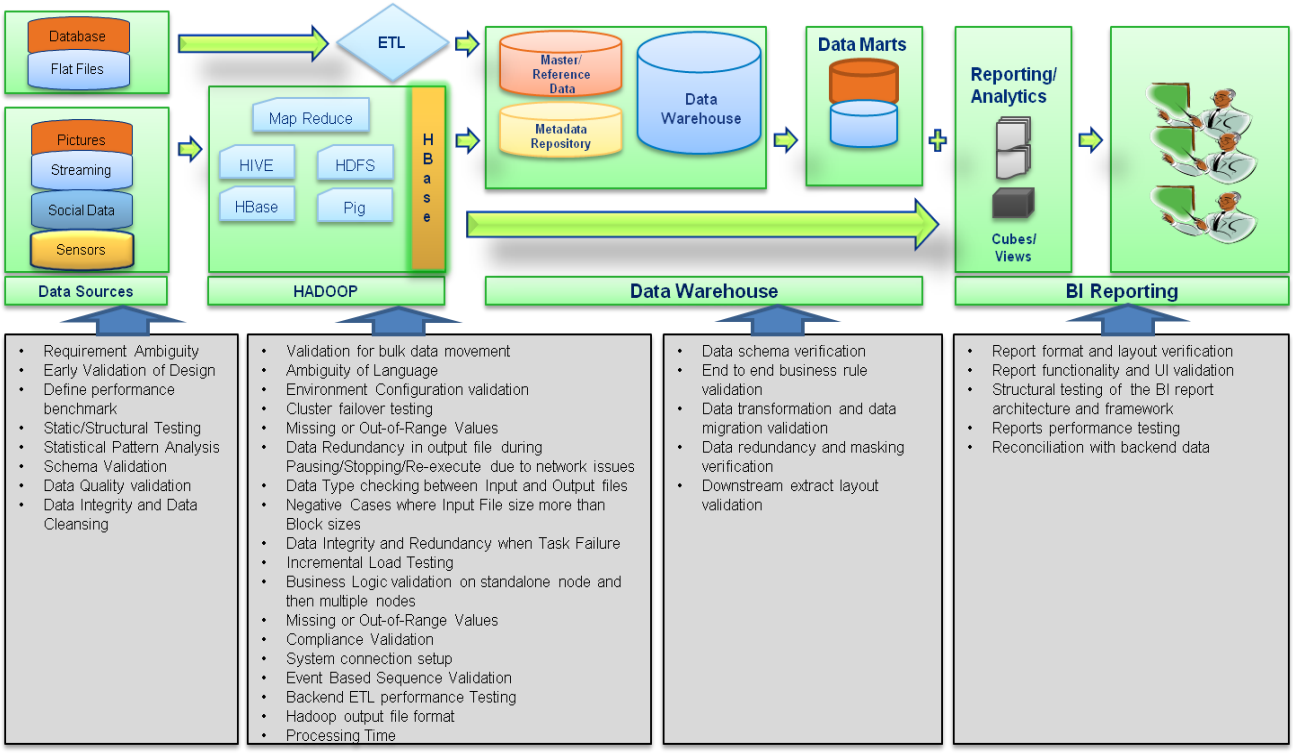

Fig 5: Big Data flow and Test Strategy

5.2 Data Mapping

In transactional world, there is a need to analyze data mappings due to mergers and acquisitions, database migrations and various business intelligence initiatives. Now data in question is unstructured and from heterogeneous sources which are new and keep on adding. Thanks to big data technology which gave an opportunity to independent validation team to validate inherent slips associated with invalid data, duplication and contradictory definitions.

- Define and ensure right external data source is mapped to the right internal data source.
- Check for the data duplicates between datawarehouse and HDFS and whether there is any real business benefit in duplicating the data.
- Apply a table top validation approach wherein independent validation team will work with BSA's to validate expected outcome matches is correct, complete and accurate as per business rules and transformation logic. Table top is a type of static validation approach to validate the requirements on paper than to execute workflow.
- Validate data is extracted correctly and as per defined business rules

5.3 Artifacts

- Artifacts relating to each data element [incoming and outgoing] including business rules and mappings, must be documented and verified
- Create data quality metadata that includes data quality attributes, measures, business rules, mappings, cleansing routines, data element profiles and controls

5.4 Test Environment

Setting up an optimal test environment for big data testing is the major challenge due to large volume and processing of data. Big data systems are based on file systems while datawarehouse systems are based on RDBMS.

- Record and validate all profile file, log file, error file paths
- Evaluate in advance number of nodes required in each test environment
- In case of cloud setup, make sure data privacy requirements are not breached
- Validate all objects and software required as part of setup and batch launcher id's have access to execute Hadoop workflows
- Change some of the parameters in Hadoop profile file and validate behavior of the system
- Make sure test environment is configured as its indented to use in production

5.5 Test data

Like other functional testing, test data needs to be setup in big data based on test scenarios. The objective of big data testing is to validate map reduce process, data validation [structured and unstructured], data process validation to ensure data correctness and accuracy.

- Create a realistic data sets to test Hadoop processing
- Change the input data, replicate the data files and validate if map reduce has taken care of all possible variations of input data
- Make sure data masking strategy is in place in case of sensitive data
- There are number of public data sources available freely to conduct test runs which can support is discovering issues with code and underlying algorithms.

5.6 Failover validation

Apart from functional testing, failover testing plays a critical role in ensuring scalability of the process. Start small by testing your nodes.

- Plan for testing with additional (at least one cluster) clusters added, removed dynamically.
- Validate all system access is intact after node/cluster is added, removed
- Validate alternate data paths once one or two node removed from the system and note down change in total execution time
- Validate Hadoop eco system is working as expected after node failure or network failure

5.7 Performance

If the Hadoop system is not able to process structured and non-structured data, the purpose of setting up Hadoop is lost.

- Validate load balancing among different nodes at different loads
- Measure completion time, throughput, memory utilization etc. for various load/time parameters
- How much data can the hadoop system actually support? At what threshold does it start to degrade?
- Make sure to test hadoop system with maximum number of heterogeneous sources with real time information coming at fast pace at high volume.
- Keep the number of nodes same and gradually increase data volume
- Same as above, but drop few nodes in between and note down performance issue.
- Validate the data set with maximum aggregation and transformation
- Run map reduce jobs in parallel with data loading. System may show locking issues

5.8 Tools

- Validation tools for datawarehouse testing are based on SQL (Structured Query language). For the comparison purpose, the datawarehouse testers use either the excel based macros or full-fledged UI based automation tools. For BIG Data, there are no defined tools.

- Learn tool like PIGLatin which is statement based and does not require complex coding. But, since both HIVE and PIGLatin are evolving, the need for writing MapReduce programs, to make the testing comprehensive cannot be ruled out. This poses tremendous challenge to the testers. Either they work towards adding the scripting skills or wait for industry to come out with powerful automation tools which provide easy interfaces to query and compare the data in HDFS with the external sources.

5.9 Generic

- As part of big data independent testing team, you may have to work with 'Unstructured or Semi Structured' data most of the time. The tester needs to seek the additional inputs on 'how to derive the structure dynamically from the given data sources' from the business/development teams.
- In traditional Data warehouse testing, tester can validate complete data with required tools but in Big data testing, considering the huge data sets for validation, even 'Sampling' strategy is a challenging. Here there is an opportunity to leverage industry best practices, tools and scripting languages to automate the data validation process.
- Validate data showing in report is matching with data in Hadoop
- Validate input files to Hadoop are correctly split, moved and replicated in different nodes
- Validate output file generated is of correct format, with correct schema and constraints, complete data
- Validate data against transformation rules, if any

6. FUTURE DIRECTION/LONG TERM FOCUS

There exists a tremendous opportunity as well as challenges for testers. Big data testers have to learn the components of the big data eco system from the scratch. Till the time, the market evolves and fully automated testing tools are available for big Data validation, the tester does not have any other option but to acquire the same skill set as the big data developer in the context of leveraging the big data technologies like Hadoop, NoSql etc. This requires a tremendous mindset shift for both the testers as well as the testing units within the organization

7. CONCLUSION

In order for businesses to take full advantage of the benefits that latest technologies has allowed, foundations for effective validation strategy need to be laid out in a strategic manner. The success and ultimate bottom line of business rely directly on how effectively independent testing team validates structured and unstructured data of large volume. Obtaining timely and meaningful information is difficult, if not possible, often hampering decision making and in many cases the cost of obtaining information is prohibitive. Organizations faced with this situation must fully realize the benefits of a big data initiative, create an effective test strategy and appropriate test environment as part of their overall validation strategy.

8. REFERENCES

[1] Big data Overview, Wikipedia.org at http://en.wikipedia.org/wiki/Big_data

[2] Big data: Testing Approach to Overcome Quality Challenges, Infosys.com at http://www.infosys.com/infosys-labs/publications/Documents/testing-approach.pdf

[3] What are best methods for testing big data applications?, quora.com at http://www.quora.com/What-are-best-methods-for-testing-big-data-applications

[4] Testing BIG Data Implementations - How is this different from Testing DWH Implementations?, infosysblogs.com at http://www.infosysblogs.com/testing-services/2012/07/testing_big_data_implementation.html