Whitepaper

testing BigData using hadoop EcoSystem

Trusted partner for your **Digital Journey**

Atos

Table of contents

Author profile

Padma Samvaba Panda is a test manager of Atos testing Practice. He has 10 years of IT experience encompassing in software quality control, testing, requirement analysis and professional services. During his diversified career, He has delivered multifaceted software projects in a wide array of domains in specialized testing areas like (big data, crm, erp, data migration). He is responsible for the quality and testing processes and strategizing tools for big-data implementations. He is a CSM (Certified Scrum Master) and ISTQB-certified. He can be reached at padma.panda@atos.net

What is Big Data?

Big data is the term for a collection of large datasets that cannot be processed using traditional computing techniques. Enterprise Systems generate huge amount of data from Terabytes to and even Petabytes of information. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks. Specifically, Big Data relates to data creation, storage, retrieval and analysis that is remarkable in terms of volume, velocity, and variety.

Hadoop and Big Data

Hadoop is one of the tools designed to handle big data. Hadoop and other software products work to interpret or parse the results of big data searches through specific proprietary algorithms and methods.

Hadoop is an open-source program under the Apache license that is maintained by a global community of users. Apache Hadoop is 100% open source, and pioneered a fundamentally new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data.

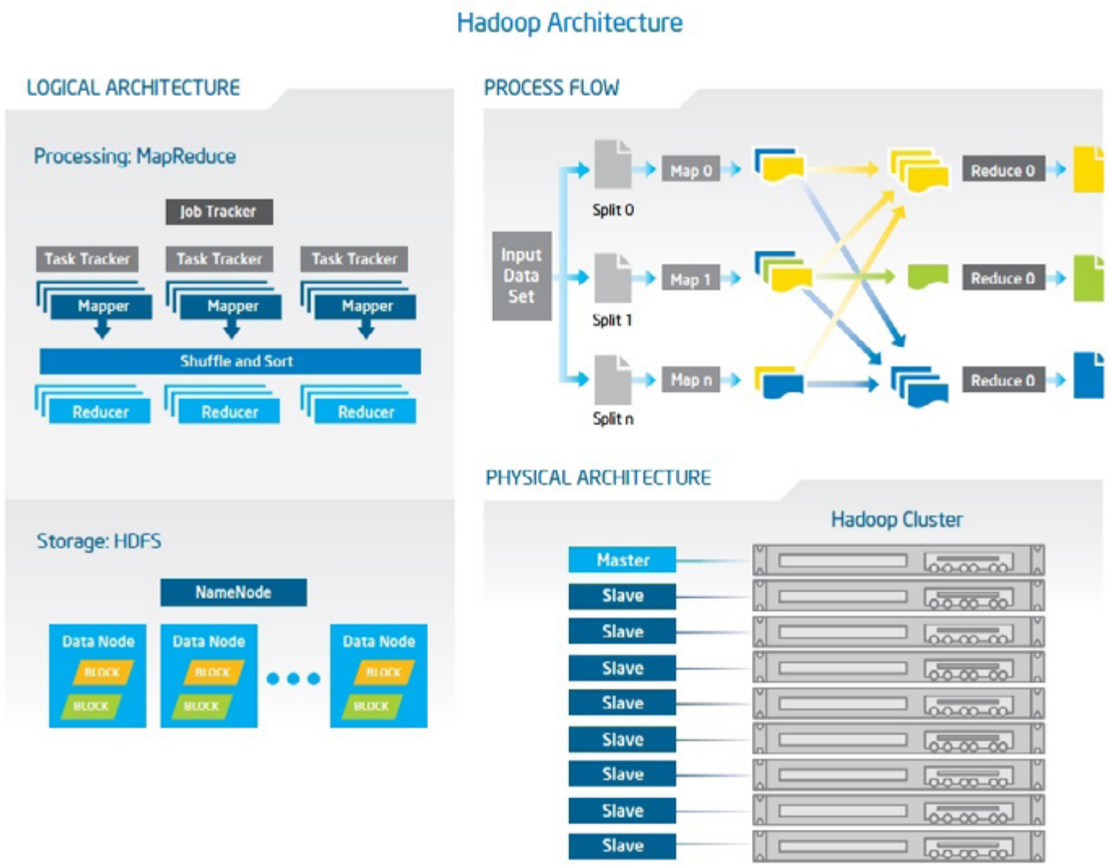
Hadoop Explained.....

Apache Hadoop runs on a cluster of industry-standard servers configured with direct-attached storage. Using Hadoop, you can store petabytes of data reliably on tens of thousands of servers while scaling performance cost-effectively by merely adding inexpensive nodes to the cluster.

The Apache Hadoop platform also includes the Hadoop Distributed File System (HDFS), which is designed for scalability and fault-tolerance. HDFS stores large files by dividing them into blocks (usually 64 or 128 MB) and replicating the blocks on three or more servers.

HDFS provides APIs for MapReduce applications to read and write data in parallel. Capacity and performance can be scaled by adding Data Nodes, and a single NameNode mechanism manages data placement and monitors server availability. HDFS clusters in production use today reliably hold petabytes of data on thousands of nodes.

Hadoop Architecture

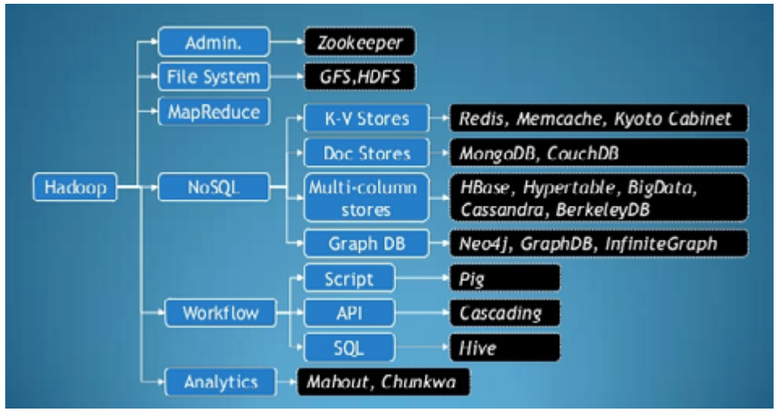


As Google, Facebook, Twitter and other companies extended their services to web-scale, the amount of data they collected routinely from user interactions online would have overwhelmed the capabilities of traditional IT architectures. So they built their own, they released code for many of the components into open source. Of these components, Apache Hadoop has rapidly emerged as the de facto standard for managing large volumes of unstructured data. Apache Hadoop is an open source distributed software platform for storing and processing data. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to the reduce tasks, which group them into final results. MapReduce uses JobTracker and TaskTracker mechanisms to schedule tasks, monitor them, and restart any that fail.

Hadoop Eco System Testing

Apache Hadoop is not actually a single product but instead a collection of several components, below screen provides the details of Hadoop Ecosystem.

Test Approach



Components

Elements	Components
Distributed Filesystem	Apache HDFS, CEPH File system
Distributed Programming	MapReduce,Pig,Spark
NoSQL Databases	Cassandra, Apache HBASE, MongoDB
SQL-On-Hadoop	Apache Hive, Cloudera Impala
Data Ingestion	Apache Flume, Apache Sqoop
Service Programming	Apache Zookeeper
Scheduling	Apache Oozie
Machine Learning	Mlib,Mahout
Benchmarking	Apache Hadoop Benchmarking
Security	Apache Ranger,Apache Knox
System Deployment	Apache Amabari ,Cloudera Hue
Applications	PivotalR,Apache Nutch
Development Frameworks	Jumpbune
BI Tools	BIRT
ETL	Talend

Hadoop testers have to learn the components of the Hadoop eco system from the scratch. Till the time, the market evolves and fully automated testing tools are available for Hadoop validation, the tester does not have any other option but to acquire the same skill set as the Hadoop developer in the context of leveraging the technologies.

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.

Challenges and Best Practices

In traditional approach, there are several challenges in terms of validation of data traversal and load testing. Hadoop involves distributed NoSQL databases instance. With the combination of Talend (open source Big data tool), we can explore list of big data tasks work flow. Following this, you can develop a framework to validate and verify the workflow, tasks and tasks complete. You can also identify the testing tool to be used for this operation. Test automation can be a good approach in testing big data implementations. Identifying the requirements and building a robust automation framework can help in doing comprehensive testing. However, a lot would depend on how the skills of the tester and how the big data environment is setup. In addition to functional testing of big data applications using approaches such as test automation, given the large size of data there are definitely needs for performance and load testing in big data implementations.

Testing Types

Testing in Hadoop Eco System can be categorized as below:

» Core components testing (HDFS, MapReduce)
» Data Ingestion testing (Sqoop,Flume)
» Essential components testing (Hive, Cassandra)

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. In functional testing with Hadoop, testers need to check data files are correctly processed and loaded in database, and after data processed the output report should generated properly also need to check the business logic on a standalone node and then on multiple nodes. Load in target system and also validating aggregation of data and data integrity.

Some of the Important Hadoop Component Level Test Approach

MapReduce

Programming Hadoop at the MapReduce level means working with the Java APIs and manually loading data files into HDFS. Testing MapReduce requires some skills in white-box testing. QA teams need to validate whether transformation and aggregation are handled correctly by the MapReduce code. Testers need to begin thinking as developers.

YARN

It is a cluster and resource management technology. YARN enables Hadoop clusters to run interactive querying and streaming data applications simultaneously with MapReduce batch jobs. Testing YARN involves validating whether MapReduce jobs are getting distributed across all the data nodes in the cluster.

Apache Hue

Hadoop has provided a web interface to make it easy to work with Hadoop data. It provides a centralized apoint of access for components like Hive, Oozie, HBase, and HDFS.From Testing point of view it involves checking whether a user is able to work with all the aforementioned components after logging in to Hue.

Apache Spark

Apache Spark is an open-source cluster computing framework originally developed in the AMPLab at UC Berkeley.
Spark is an in-memory data processing framework where data divided into smaller RDD.Spark performance is up to 100 times faster than hadoop mapreduce for some applications. From QA standpoint it involves validating whether spark worker nodes are working and processing the streaming data supplied by the spark job running in the namenode.Since it is integrated with other nodes (E.g. Cassandra) it should have appropriate failure handling capability. Performanace is also an important benchmark of a spark job as it is used as an enhancement over existing MapReduce Operation.

Jasper Report

It is integrated with Data Lake layer to fetch the required data (Hive). Report designed using Jaspersoft Studio are deployed on Jasper Report Server. Analytical and transactional data coming from Hive database is used by Jasper Report Designer to generate complex reports. The Testing comprises the following:
Jaspersoft Studio is installed properly.
Hive is properly integrated with Jasper Studio and Jasper Report Server via a JDBC connection. Reports are exported in specified format correctly. Auto Complete Login Form, Password Expiration Days and allow User Password Change criteria. User not having admin role cannot create new User and new Role.

Apache Cassandra

It is a non-relational, distributed, open-source and horizontally scalable database. A NoSQL database tester will need to acquire knowledge of CQL (Cassandra Query Language) in order to perform quality testing. It is independent of specific application or schema and can operate on a variety of platforms and operating systems. QA areas for Cassandra include data type checks, count checks, CRUD operations checks, timestamp and its format checks, checks related to cluster failure handling and data integrity and redundancy checks on task failure.

Flume and Sqoop

Big data is equipped with data ingestion tools such as Flume and Sqoop, which can be used to move data into and out of Hadoop. Instead of writing a stand-alone application to move data to HDFS, these tools can be considered for ingesting data, say for example from RDBMS since they offer most of the common functions.General QA checkpoints include successfully generating streaming data from Web sources using Flume, checks over data propagation from conventional data storages into Hive and HBase, and vice versa.

Talend

Talend is an ETL tool that simplifies the integration of big data without having to write or maintain complicated Apache Hadoop code. Enable existing developers to start working with Hadoop and NoSQL databases. Using Talend data can be transferred between Cassandra ,HDFS and Hive.Validating talend activities involve data loading is happening as per business rules,counts match, it appropriately rejects, replaces with default values and reports invalid data. Time taken and performance is also important while validating the above scenarios.

KNIME Testing

Konstanz Information Miner, is an open source data analytics, reporting and integration platform. We have integrated and tested KNIME with various components (Ex: R, Hive, Jasper Report Server).
Testing includes:
» Proper KNIME installation and configuration. (KNIME + R) integration with Hive.
» Testing KNIME analytical results using R scripts.
» Checking reports that are exported from KNIME analytical results in specified format from Hive database in Jasper Report.

Ambari

All administrative tasks (e.g.: configuration, start/stop service) are done from Ambari Web. Tester need to check that Ambari is integrated with all other applications. E.g.: Nagios, Sqoop, WebHDFS, Pig, Hive, YARN etc.

Apache Hive

Hive enables Hadoop to operate as a data warehouse. It superimposes structure on data in HDFS, and then permits queries over the data using a familiar SQL-like syntax. Since Hive is recommended for analysis of terabytes of data, the volume and velocity of big data are extensively covered in Hive testing from a functional standpoint, Hive testing requires tester to know HQL (Hive Query Language). It incorporates validation of successful setup of the Hive meta-store database; data integrity between HDFS vs. Hive and Hive vs. MySQL (meta-store); correctness of the query and data transformation logic, checks related to number of MapReduce jobs triggered for each business logic, export / import of data from / toHive, data integrity and redundancy checks when MapReduce jobs fail.

Apache Oozie

It's a really nice scalable and reliable solution in Hadoop ecosystem for job scheduling. Both Map Reduce and Hive, Pig scripts can be scheduled along with the job duration. QA activities involve validating an Oozie workflow. Validating the execution of a workflow job based on user defined timeline.

Nagios testing

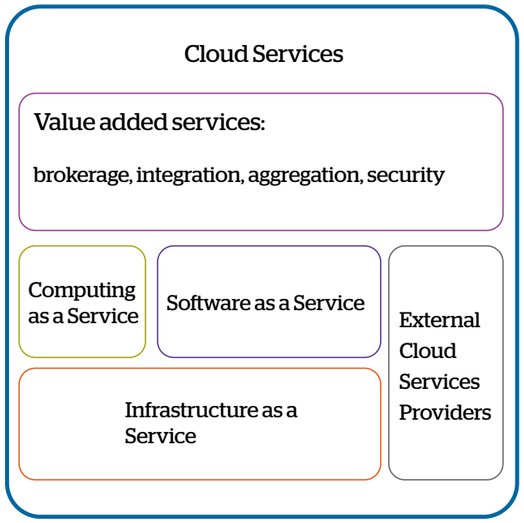
Nagios enables to perform health checks of Hadoop & other components and checks whether the platform is running according to normal behavior. We need to configure all the services in Nagios, so that we can check the health and performance from Nagios web portal.
- The health check includes:
  . If a process is running
  . If a service is running
  . If the service is accepting connections
  . Storage capacity on data nodes.

Testing Hadoop in Cloud Environment

Before Testing Hadoop in Cloud:

1. Document the high level cloud test infrastructure (Disk space, RAM required for each node, etc.)
2. Identify the cloud infrastructure service provider
3. Document the data security plan
4. Document high level test strategy, testing release cycles, testing types, volume of data processed by Hadoop, third party tools.

Technology principles
Big data in the cloud



Types

» Application, data, computing and storage
» Fully used or hybrid cloud
» Public or on-premise
» Multi-tennant or single -tennant

Characheristics

» Scalability
» Elasticity
» Resource pooling
» Self service
» Pay as you go

Broader Impact

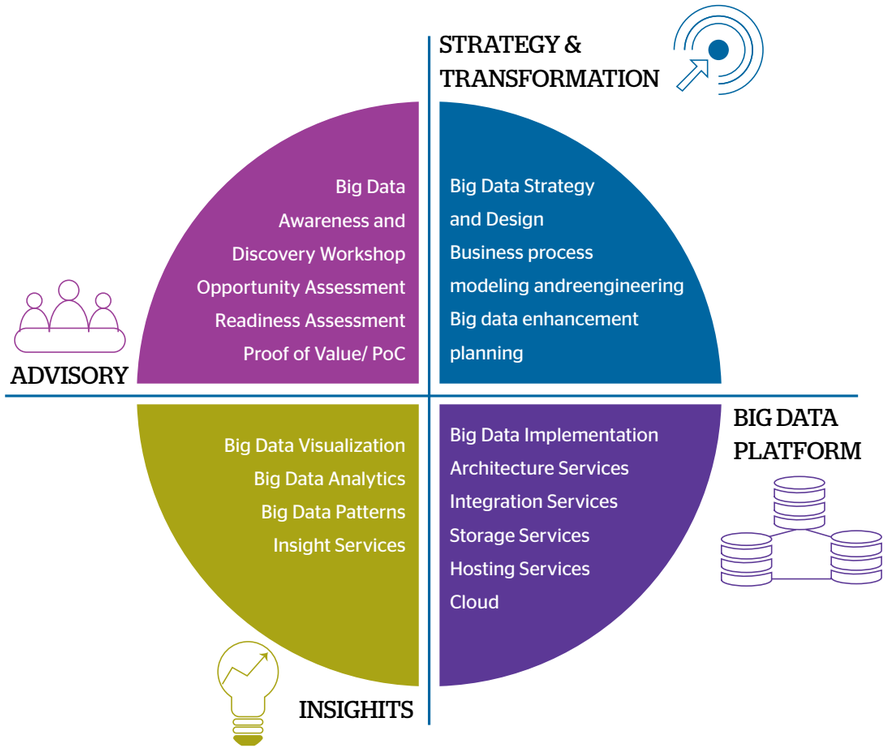
The main key features that leverage Bigdata test framework in cloud are:

» On demand Hadoop testbed to test Big data
» Virtualized application / service availability that need to be tested
» Virtualized testing tool suite- Talend and jmeter
» Managed test life cycle in Cloud
» Different types of Big Data test metrics in cloud
» Operations like import / export configurations and test artifacts in / out of the testbed.

How Atos Is Using Hadoop

Information Culture is changing... Leading to increased Volume, Variety & Velocity

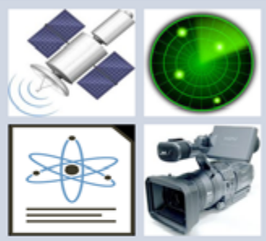
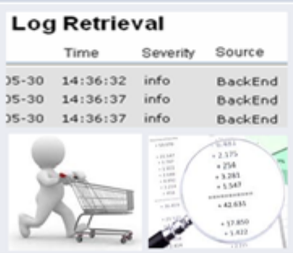
Atos and Big Data
Service Overview, from critical IT to Business support



We use a four-stage framework to deliver our Big Data Analytics solutions and services

Advisory	Strategy and transformation	Big Data platform	Insights
Workshops – we deliver successful workshops for clients across all markets Proof of Concept / Proof of Value – practical and ready PoC/PoV scenarios can be deployed.	Big Data strategy and design; how should your business approach big data Business process modeling and reengineering; what implications does big data have on your business.	Big Data implementation such as IDA Architecture, Integration, Storage and Hosting Services, with Canopy.	Big Data Visualization, Analytics, Patterns and Insight Services.

Data examples

	Structured data	Semi-structured data	Unstructured data
Machine-generated			
Human-generated			

Structured data

- **Machine-generated:** Input data, click-stream data, gaming-related data, sensor data
- **Human-generated:** Web log data [!= weblog], point-of-sale data (when something is bought), financial data

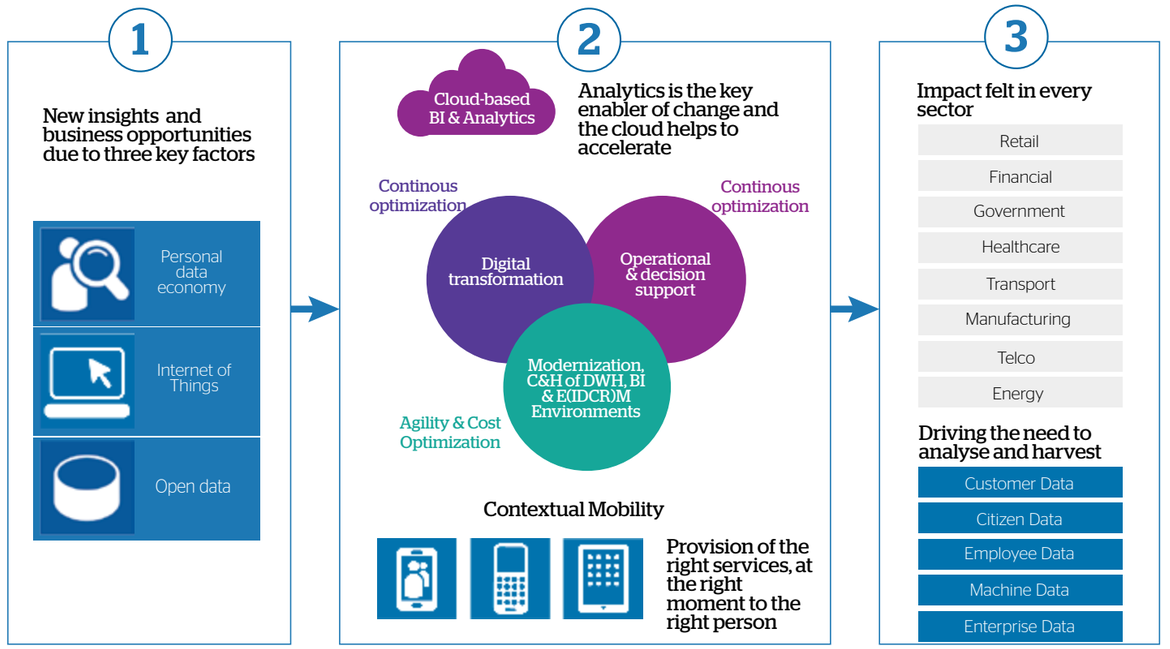
Semi-structured data

- **Machine-generated:** Electronic data interchange (EDI), SWIFT, XML, RSS feeds, sensor data
- **Human-generated:** Emails, spreadsheets, incident tickets, CRM records

Unstructured data

- **Machine-generated:** Satellite images, scientific data, photographs and video, radar or sonar data
- **Human-generated:** Internal company text data, social media, mobile data, website content

Economics of Big Data



Atos has a clear vision of the importance of Big Data as a Business Success factor. This section gives a single, overall view of how we see the analytics marketplace and how we operate within it.

Atos believes that analytics is the key to gaining true business insight and to achieving competitive advantage. Organizations must turn on analytics everywhere to realize this understanding and apply it effectively.

We have a three level approach on analytics:

Enterprise Analytics: Better decision-making, enabled by customized business intelligence solutions

Consumer Analytics: Driven by real-time data flows, delivering immediate access to actionable intelligence

Vertical & Industry-specific Analytics: Process optimization and improved operational efficiency through automated use of real-time data, intelligence, monitoring & analytics.

The top line of the graph shows the principal inputs, the flows of real-time or near real-time data that provide raw material for analytics. The three connected circles show the main areas of focus and activity for Atos:

Digital transformation is all about digitizing and optimizing business processes through proper application of workflow, data and information management, and analytics concepts.

Performance Management & Operational Intelligence is on one hand about financial data reporting and on the other hand about creating better decision support systems that support stakeholders in running their business or an organization.

M/C/H is about modernizing existing Data and Analytics environments to support challenges in performance, requirements, disruptive trends like mobility & cloud, and operational costs Finally, we see the ways in which different analytics-driven outputs lead to positive change in a wide range of different sectors, as you can see on the right hand side of the model.

Testing Accomplishments

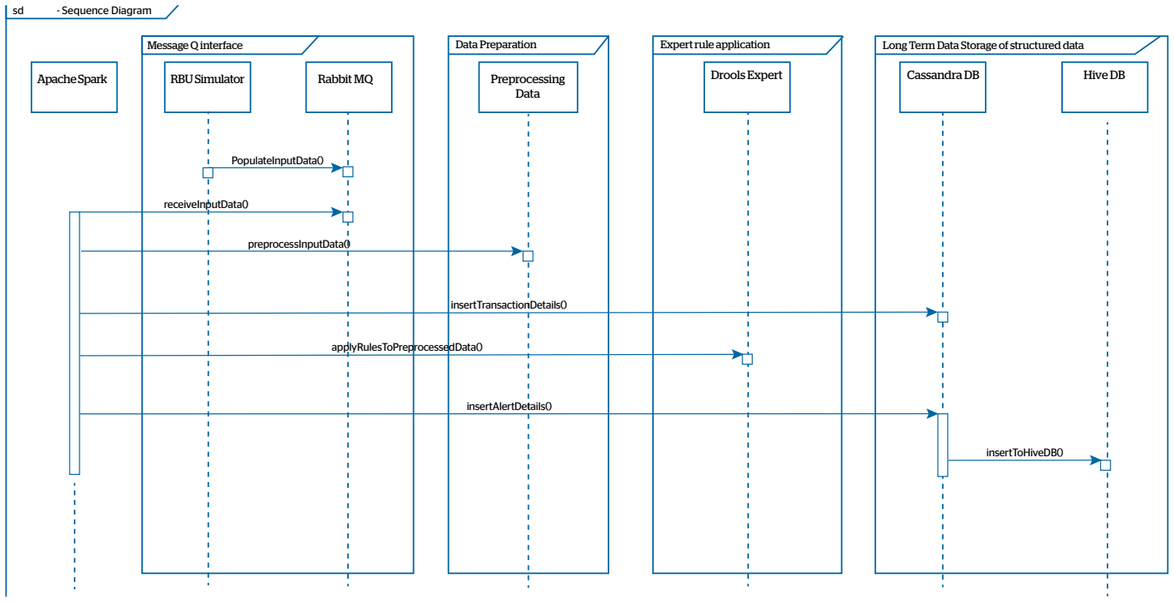
Here are the rules of Big Data testing:

1. Generate a lot of test data
2. Using continuous integration (CI) and automated builds
3. Create two or more test modules, with increasing load and execution time
4. Spin your clusters of Hadoop or HBase nodes as part of the test
5. Do performance testing early
6. Install proper monitoring.

Let's discuss some of the case studies...

Case Study 1 (Integration Testing)

Data from external sources should pass through various stages and processed data should be properly stored in Data Lake (Hive)



Test Data from a variety of sources to simulate real life scenarios was prepared as an external data source. Using the simulator data was fed to RabbitMQ where as a tester it was checked whether data was in proper format and then through a Spark job data was decoded using Spark-Cassandra connector it was inserted into Cassandra (NoSQL) where data was checked by tester for integrity and count using CQL languages. After through Talend jobs data from Cassandra to Hive via HDFS was transferred. Same was validated in Hive using Hive Query Language.

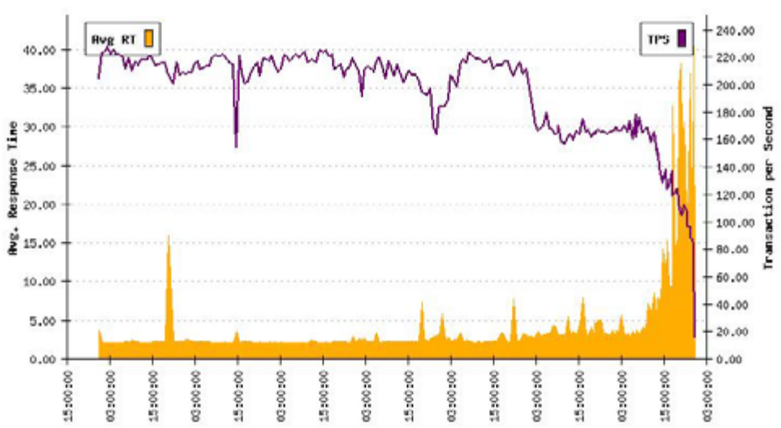
Case Study 2 (Performance Testing)

Streaming data ingestion continuously.

Volumes to be handled: 1.3 million transactions per day with a peak of 1,000 transactions/minute
Day Long Test (24 hr. test)
Long Run Test (7 day test).

To meet the above objectives Apache Jmeter was used as a performance testing tool. Plugin were used in Jmeter to enable it send messages to Hadoop via Messaging Queue. CSV data set config feature was used to feed the bulk data (file size 1.4 GB).Various listeners were configured(E.g. Summary, Response Time, Aggregate Report).Variety of Graphs were generated some of which are given below. Later on the above feature was also extended to Day Long test and Long Run Tes where using Jmeter, System was being fed with Streaming Data on a continual basis.

Test Timeline Charts
Load Generation Timeline
Average Response



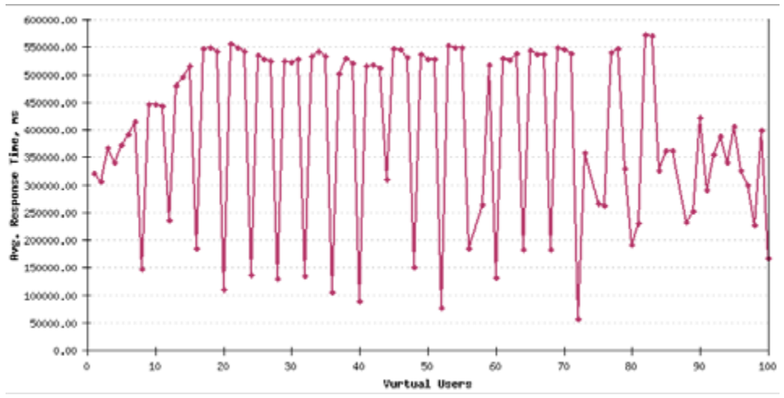
Case Study 3 (100 simultaneous users test)

Hadoop interfacing applications should support 100 simultaneous users without any performance degradation.

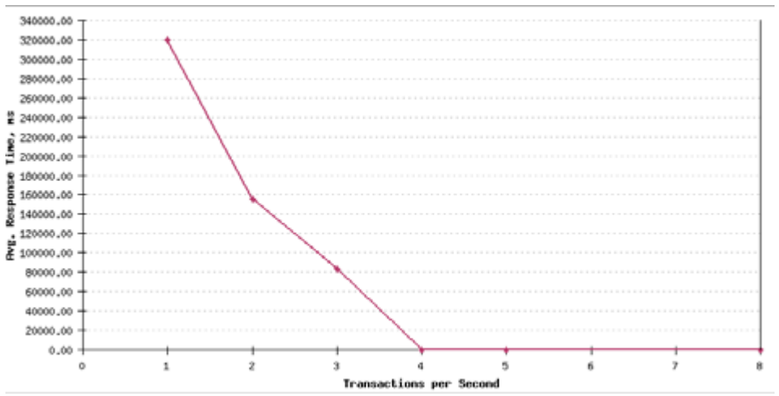
100 Simultaneous User Test:

Customer interfacing applications (E.g. Hue, Drools, RabbitMQ, Hive, Ambari, Jasper Report, and BIRT) were subjected to a 100 user test where those applications can support up to 100 simultaneous users without any degrade in performance. JMeter was also used here to generate scripts simulating the activity performed by the users.

Response Time vs Virtual Users. (Apache Hive)



Response Time vs TPS (Apache Hive)"



Conclusion

Big data is still emerging and a there is a lot of onus on testers to identify innovative ideas to test the implementation. One of the most challenging things for a tester is to keep pace with changing dynamics of the industry. While on most aspects of testing, the tester need not know the technical details behind the scene however this is where testing Big Data Technology is so different. A tester not only needs to be strong on testing fundamentals but also has to be equally aware of minute details in the architecture of the database designs to analyze several performance bottlenecks and other issues. Hadoop testers have to learn the components of the Hadoop 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. This requires a tremendous mindset shift for both the testers as well as the testing units within the organization. To be competitive, in the short term, the organizations should invest in the BIG Data specific training needs of the testing community and in the long term, should invest in developing the automation solutions for BIG Data validation.

Appendix

References

1. http://www.slideshare.net/pnicolas/overview-hadoop-ecosystem

About Atos

Atos SE (Societas Europaea) is a leader in digital services with pro forma annual revenue of circa € 12 billion and circa 100,000 employees in 72 countries. Serving a global client base, the Group provides Consulting & Systems Integration services, Managed Services & BPO, Cloud operations, Big Data & Cyber-security solutions, as well as transactional services through Worldline, the European leader in the payments and transactional services industry. With its deep technology expertise and industry knowledge, the Group works with clients across different business sectors: Defense, Financial Services, Health, Manufacturing, Media, Utilities, Public sector, Retail, Telecommunications, and Transportation.

Atos is focused on business technology that powers progress and helps organizations to create their firm of the future. The Group is the Worldwide Information Technology Partner for the Olympic & Paralympic Games and is listed on the Euronext Paris market. Atos operates under the brands Atos, Atos Consulting, Atos Worldgrid, Bull, Canopy, Unify and Worldline. Management