testing BigData using hadoop EcoSystem
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 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 MB 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.

Table of contents

03 What is Big Data?
03 Hadoop and Big Data
03 Hadoop Explained
04 Hadoop Architecture
05 Hadoop Eco System Testing
08 Testing Hadoop in Cloud Environment
09 How Atos Is Using Hadoop
12 Testing Accomplishments
15 Conclusion
15 Appendices

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.
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**

<table>
<thead>
<tr>
<th>Elements</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed Filesystem</td>
<td>Apache HDFS, CEPH File system</td>
</tr>
<tr>
<td>Distributed Programming</td>
<td>MapReduce, Pig, Spark</td>
</tr>
<tr>
<td>NoSQL Databases</td>
<td>Cassandra, Apache HBASE, MongoDB</td>
</tr>
<tr>
<td>SQL-On-Hadoop</td>
<td>Apache Hive, Cloudera Impala</td>
</tr>
<tr>
<td>Data Ingestion</td>
<td>Apache Flume, Apache Sqoop</td>
</tr>
<tr>
<td>Service Programming</td>
<td>Apache Zookeeper</td>
</tr>
<tr>
<td>Scheduling</td>
<td>Apache Oozie</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Mahout, Mahout</td>
</tr>
<tr>
<td>Benchmarking</td>
<td>Apache Hadoop Benchmarking</td>
</tr>
<tr>
<td>Security</td>
<td>Apache Ranger, Apache Knox</td>
</tr>
<tr>
<td>System Deployment</td>
<td>Apache Ambari, Cloudera Hue</td>
</tr>
<tr>
<td>Applications</td>
<td>PivotalR, Apache Nutch</td>
</tr>
<tr>
<td>Development Frameworks</td>
<td>Jumpdbane</td>
</tr>
<tr>
<td>BI Tools</td>
<td>BRT</td>
</tr>
<tr>
<td>ETL</td>
<td>Talend</td>
</tr>
</tbody>
</table>
Hadoop testers have to learn the components of the Hadoop ecosystem from 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 helps validate the code. Thereby the tester individually tests the MapReduce job but does 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.

Testing Types
Testing in Hadoop Ecosystem can be categorized as below:

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

MapReduce
Programming at the MapReduce level means working with the Java APIs and manually loading data files into HDFS. Testing MapReduce requires writing a script to decide if the workflow is working as expected. 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 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 RDDs. Spark performance is up to 100 times faster than hadoop mapreduce for some applications. From a map Reduce 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. Performance 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.

Testing includes:
- Checking reports that are exported from Jasper Report Server.
- Reports are exported in specified format correctly.
- report 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 actual 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 in sync with data ingestion tools such as Flume and Sqoop, which can be used to move data into and out of a 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. 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 involves data loading is happening as per business rules, counts match, it appropriately rejects, replaces with default values and reports invalid data. Validating is an 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

Using Hadoop Eco System
- Testing - BigData Using Hadoop Eco System

Challenges and Best Practices
In traditional approach, there are several challenges in terms of validation of data traversal and load testing. Hadoop testers distributed programming paradigm with the combination of Talend (open source Big data tool), can we explore list of big data tasks work flow. Following this, you develop a framework to validate and verify the workflow, tasks and tasks complete. You can also identify the testing tool and its integration for automation. 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 definitely needs for performance and load testing in big data implementations.
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-tenant or single-tenant

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

Broader Impact
The main key features that leverage Big data 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

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 are the 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

<table>
<thead>
<tr>
<th>Machine-generated</th>
<th>Structured data</th>
<th>Semi-structured data</th>
<th>Unstructured data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input data, click-stream data, gaming-related data, sensor data</td>
<td>Electronic data interchange (EDI), SWIFT, XML, RSS feeds, sensor data</td>
<td>Satellite images, scientific data, photographs and video, radar or sonar data</td>
<td>Internal company test data, social media, mobile data, website content</td>
</tr>
<tr>
<td>Web log data (e.g. weblog), point of sale data (e.g. when something is bought), financial data</td>
<td>Emails, spreadsheets, incident tickets, CRM records</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Structured data
- Machine-generated: Input data, click-stream data, gaming-related data, sensor data
- Human-generated: Web log data (e.g. weblog), point of sale data (e.g. 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 test 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:

1. Enterprise Analytics: Better decision-making, enabled by customized business intelligence solutions
2. Consumer Analytics: Driven by real-time data flows, delivering immediate access to actionable intelligence

Atos has developed three connected circles to show the main areas of focus and activity for Atos:

- Digital transformation
- Operational & decision support
- Continuous optimization

The impact of analytics is felt in every sector:
- Retail
- Financial
- Government
- Healthcare
- Transport
- Manufacturing
- Telecom
- Energy

Driving the need to analyze and harvest
- Customer Data
- Citizen Data
- Employee Data
- Machine Data
- Enterprise Data
- Vertical Data
- Open data

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 the 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.

MIC/CH 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 Test where using Jmeter, System was being fed with Streaming Data on a continual basis.
Conclusion

Big data is still emerging and 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 ecosystem 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