The Emerging Big Data System - Testing Perspective

Table of Contents

Abstract	3
Typical Data Warehouse Architecture	3
The Emergence of Big Data	4
Hadoop Architecture & Components	4
The Emerging Data Warehouse	5
Testing Hadoop System	5
Hadoop System - Data Flow	6
Conclusion	6

Abstract

The purpose of this paper is to delve into the evolution of the data paradigm in this information age, where data has truly exploded to Terabytes and Petabytes. From an enterprise perspective, this is both a challenge & an opportunity to tap into the new data sources. These data sources are referred to as Big Data. The existing data warehouse technologies cannot handle this data glut and will require adaptation. Therefore, data warehousing and big data technologies complement each other to derive maximum use from the hitherto untapped area of unstructured and semi-structured data getting generated endlessly in the Internet, sensor technologies etc.. We will see the basic Big Data architecture around Hadoop and also the testing approaches in a Hadoop system.

Typical Data Warehouse Architecture

In a typical DW architecture, we will have data from diverse & disparate sources arriving at a central location, for processing into a Data warehouse. This data is cleansed and scrubbed before transformations are applied to be finally pushed into the data warehouse. This data warehouse now represents the single-version of truth for the enterprise, including the wealth of historical data accumulated over several years. This data can now be analyzed/ sliced/diced across multiple-dimensions for various decision supports, to help senior management take informed decisions.

Major ETL Tools in Market : Informatica, Abi initio, DataStage, Talend Major BI Tools in Market : Hyperion, Cognos, OBIEE, Business Objects

Though the data are from different sources, these are mostly structured data coming out of organized systems / platforms. The ETL tool's major function is to convert the data into a consistent format. For example, the date in different sources may be in different formats, so that in the final analysis, reports are also consistent and accurate at the data warehouse. The major function of the analytical tool is to summarize and aggregate data across various dimensions.

A pharmaceutical company will be interested in knowing how its various products have performed across different locations as compared to competition, in a specific time period. The senior management may want to make certain decisions on R&D. For these, the BI reports can be very helpful. With these reports the management can drill into the details for more clarity on certain aspects of the business.

The data warehouse discussed, contains majorly the transactional data in a structured format. The historical data after a point becomes a drag and hits the data warehouse performance. So it needs to be maintained for better results.

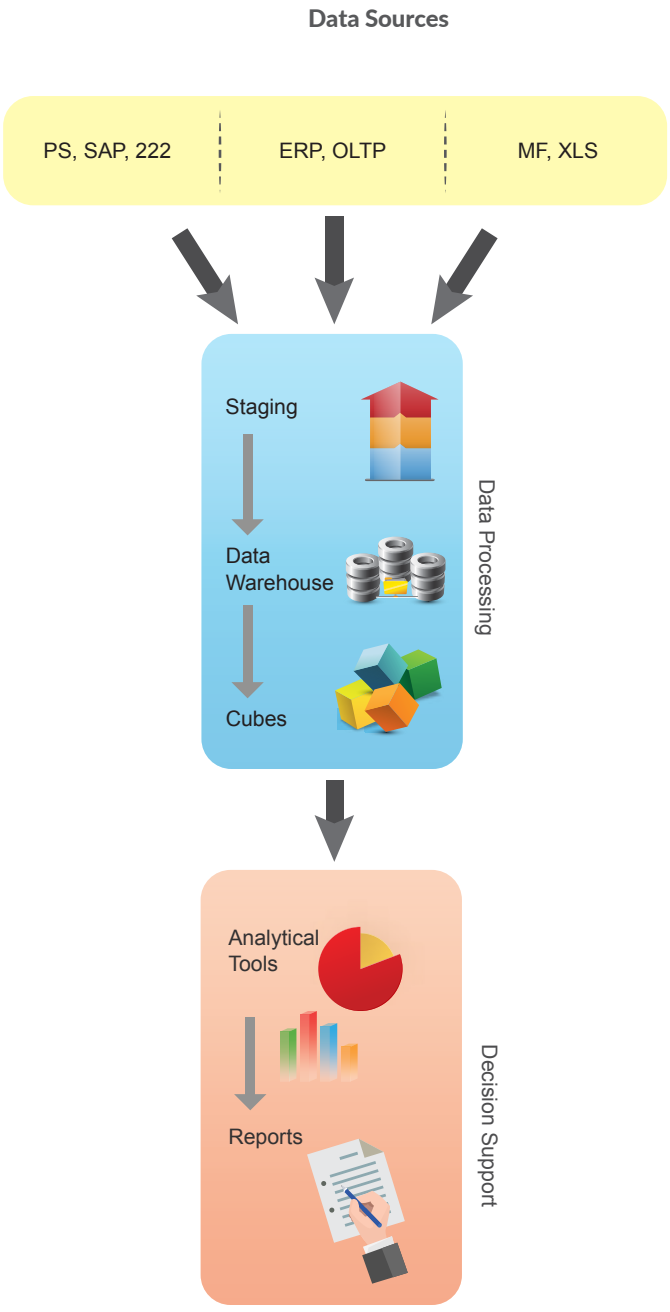

Fig 1: Basic Data warehouse architecture

The Emergence of Big Data

With the explosion of Internet, followed by the crash in prices of storage devices, it has become imperative for enterprises to tap into additional sources of data to make much better and focused decisions. But the challenges are multi-fold, in terms of volume, variety and velocity of data, referred as the 3Vs. Welcome to the world of Big Data!

Now-a-days, IBM can get quick feedback on specific products from its vast product portfolio from customers through Facebook, Twitter. The advancements in sensor technology have triggered a new wave of data that can help industries such as retail and logistics. Web sites and system generated log files from various applications in an enterprise can be analyzed for error patterns, user preferences, traffic analysis and conversion ratios (how many people finally bought something from the web site, how many just browsed). Credit card & insurance companies can utilize big data to identify frauds. For example, if a credit card had been swiped in 2 places that are far apart, in a short time, the system can now flag it as suspect and send alerts to the various stakeholders for action. An enterprise can analyze email patterns for possible abuses, also for arriving at the efficiency of its sales personnel in closing deals.

But these data formats are unstructured and huge in terms of storage, running to Terabytes (TB). So the complexity is in the data format, its size as well as the speed with which they arrive. Making sense of this data presents a huge opportunity as well, in the wealth of information they provide. This calls for newer technologies for data processing as the traditional systems cannot accommodate these formats.

Some unstructured / semi-structured Data Sources:

Big Data Technologies :

Hadoop Architecture & Components

Apache Hadoop is a scalable, fault-tolerant system for data storage and processing. It is economical and reliable, which makes it perfect to run data-intensive applications on commodity hardware.

It contains the following components:

HDFS: The Hadoop Distributed File System (HDFS) is the storage system responsible for replicating and distributing the data throughout the computing cluster

MapReduce: It is the software framework which developers use to write the applications which process the data stored throughout HDFS

Hue: Sits on top of Hadoop, helps interface with the various Hadoop components

ZooKeeper: It is responsible for coordinating the configuration data and process sync Database / Data warehouse for Hadoop: HBase, Hive is the DB, DW used with Hadoop

The Emerging Data Warehouse

With unstructured as well as semi-structured data joining the data sources, the current DW technologies will have to accommodate the Big Data technologies as well. In fact, these two technologies complement each other so that the transformed/ processed unstructured, semi-structured data find their way into the existing data warehouse. With this, the data warehouse continues to remain as the single truth of data, from which BI reports are generated. The semi-structured, unstructured data is pushed into the HDFS (Hadoop Distributed File System) using specific APIs. The HDFS holds the semi-structured, unstructured data sources. The MR framework in Hadoop uses the HDFS as input, does the processing on commodity hardware and finally writes-back into HDFS.

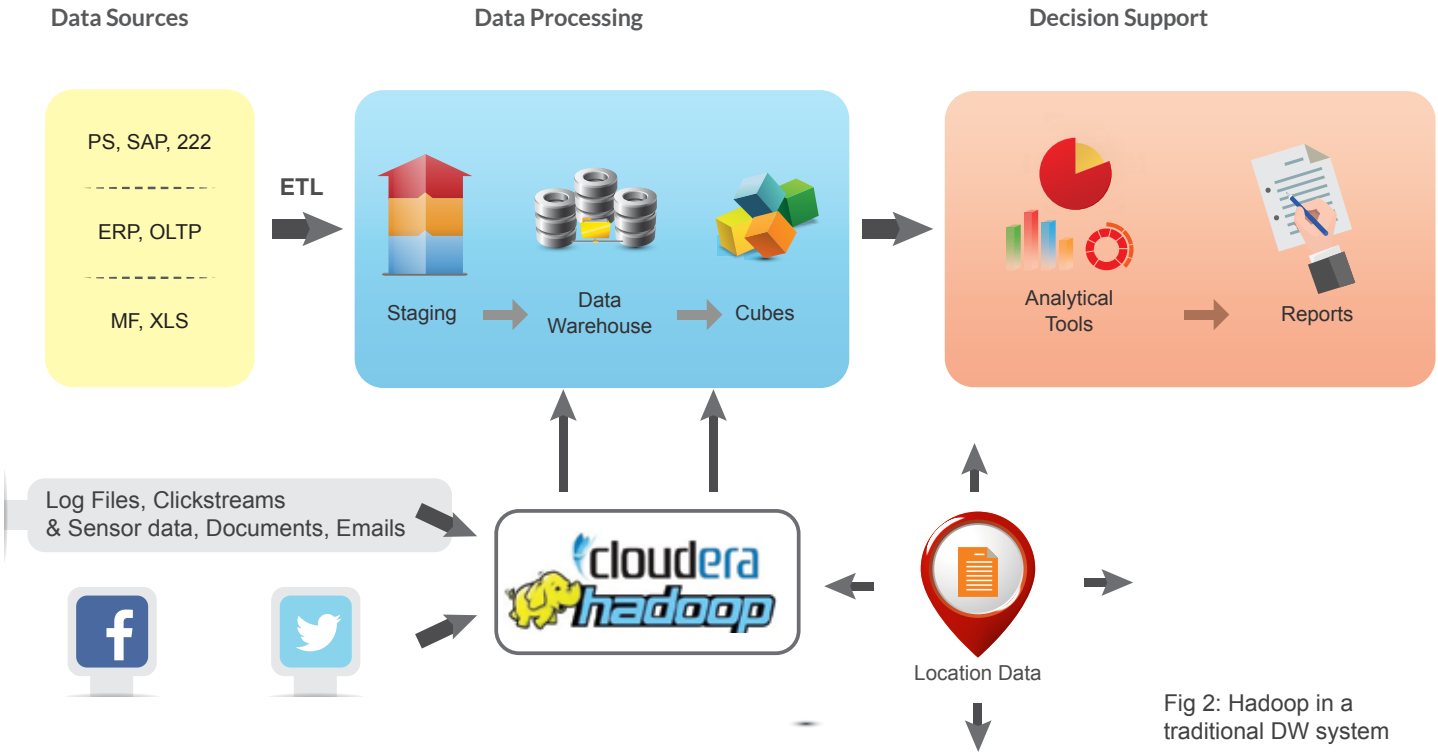

Fig 2: Hadoop in a traditional DW system

Testing Hadoop System

The testing steps broadly involve the following:

1. **Testing the Input system (HDFS)**
   - To validate if the HDFS has data in the correct format
   - To validate if all required source data have been moved into HDFS

2. **Testing the Output of the MR process**
   - To validate if the MR process generates the correct output in terms of Key-value pairs
   - To validate if the output is in sync with the HDFS source in data and format
   - The summarized and aggregated data in the output (HDFS/Hive), tally with that of the input (HDFS)
   - To validate if specific data transformations are as per the requirement
   - To make sure that the output remains the same if Hadoop is run on a single-node or on multiple nodes
   - When the HDFS output is finally moved into the data warehouse, it is required that HDFS data aggregation/summarization tally with the data moved into the data warehouse.HDFS can be queried using Pig, while the data warehouse can be queried using SQL

3. **Testing the BI Reports**
   - Here we validate the reports for layout/ format and data

4. **4. ETL Validations**
   - The ETL Validation strategy remains the same as that of a typical data warehouse

Hadoop System - Data Flow

With unstructured as well as semi-structured data joining the data sources, the current DW technologies will have to accommodate the Big Data technologies as well. In fact, these two technologies complement each other so that the transformed/ processed unstructured, semi-structured data find their way into the existing data warehouse. With this, the data warehouse continues to remain as the single truth of data, from which BI reports are generated. The semi-structured, unstructured data is pushed into the HDFS (Hadoop Distributed File System) using specific APIs. The HDFS holds the semi-structured, unstructured data sources. The MR framework in Hadoop uses the HDFS as input, does the processing on commodity hardware and finally writes-back into HDFS.

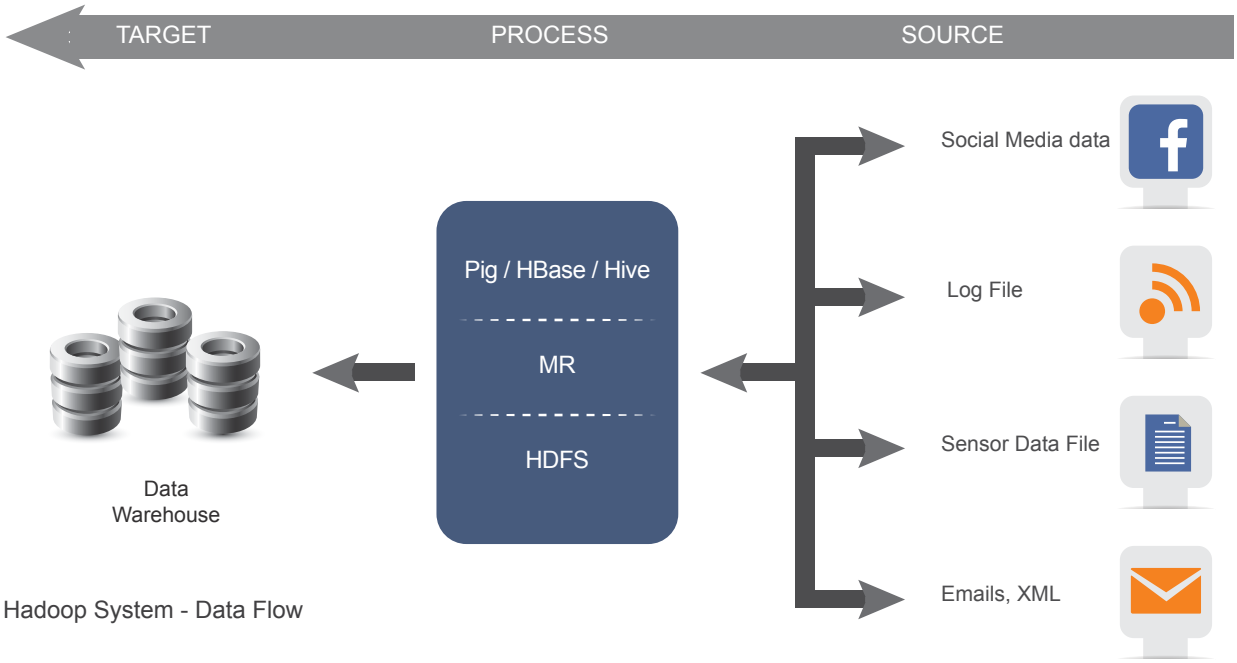

Fig 3 : Hadoop System - Data Flow

Conclusion:

Success in a data testing project calls for a clear test strategy. This cannot change in a big data environment, where the complexity has increased considerably due to the arrival of semi-structured/ unstructured data. Automation of the testing process can give a decisive edge in this complex data environment.

About Author:

Nagarajan K R, manages the BI Testing Practice in Hexaware technologies. He has over 15 years of IT experience in web, BI technologies. His current interests includes testing Big Data applications, coming up with utility tools for testing them and as a result the changing Dataware house architecture.

About Hexaware

Hexaware is the fastest growing next-generation provider of IT, BPO and consulting services. Our focus lies on taking a leadership position in helping our clients attain customer intimacy as their competitive advantage. Our digital offerings have helped our clients achieve operational excellence and customer delight by 'Powering Man Machine Collaboration.' We are now on a journey of metamorphosing the experiences of our customer's customers by leveraging our industry-leading delivery and execution model, built around the strategy— 'Automate Everything, Cloudify Everything, Transform Customer Experiences.'

We serve customers in Banking, Financial Services, Capital Markets, Healthcare, Insurance, Manufacturing, Retail, Education, Telecom, Professional Services (Tax, Audit, Accounting and Legal), Travel, Transportation and Logistics. We deliver highly evolved services in Rapid Application prototyping, development and deployment; Build, Migrate and Run cloud solutions; Automation-based Application support; Enterprise Solutions for digitizing the back-office; Customer Experience Transformation; Business Intelligence & Analytics; Digital Assurance (Testing); Infrastructure Management Services; and Business Process Services.

Hexaware services customers in over two dozen languages, from every major time zone and every major regulatory zone. Our goal is to be the first IT services company in the world to have a 50% digital workforce.

NA Headquarters
Metro 101, Suite 600,101 Wood Avenue South, Iselin,
New Jersey - 08830
Tel: +001-609-409-6950
Fax: +001-609-409-6910

India Headquarters
152, Sector – 3
Millennium Business Park
'A' Block, TTC Industrial Area
Mahape, Navi Mumbai – 400 710
Tel : +91-22-67919595
Fax : +91-22-67919500

EU Headquarters
Level 19, 40 Bank Street,
Canary Wharf,
London - E14 5NR
Tel: +44-020-77154100
Fax: +44-020-77154101

APAC Headquarters
180 Cecil Street,
#11-02, Bangkok Bank Building,
Singapore 069546
Tel : +65-63253020
Fax : +65-6222728