What is Big Data

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 6 years.

VVV:

Sources of Big Data

These data come from many sources like

* **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
* **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
* **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
* **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
* **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.
* **Health Insurance**

3V's of Big Data

1. **Velocity:** The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.
2. **Variety:** Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.
3. **Volume:** The amount of data which we deal with is of very large size of Peta bytes.

Use case

An e-commerce site XYZ (having 100 million users) wants to offer a gift voucher of 100$ to its top 10 customers who have spent the most in the previous year. Moreover, they want to find the buying trend of these customers so that company can suggest more items related to them.

Issues

Huge amount of unstructured data which needs to be stored, processed and analyzed.

Solution

**Storage:** This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion. It works on Write once, read many times principle.

**Processing:** Map Reduce paradigm is applied to data distributed over network to find the required output.

**Analyze:** Pig, Hive can be used to analyze the data.

**Cost:** Hadoop is open source so the cost is no more an issue.

What is Hadoop

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing.It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

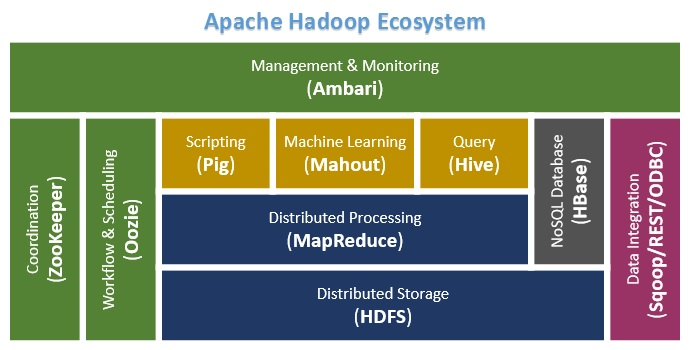
Modules of Hadoop

1. **HDFS:** Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
2. **Yarn:** Yet another Resource Negotiator is used for job scheduling and manage the cluster.
3. **Map Reduce:** This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
4. **Hadoop Common:** These Java libraries are used to start Hadoop and are used by other Hadoop modules.

Advantages of Hadoop

* **Fast:** In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
* **Scalable:** Hadoop cluster can be extended by just adding nodes in the cluster.
* **Cost Effective:** Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
* **Resilient to failure:** HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.

**BigData Hadoop Architecture**



VM or Cloud Environment (AWS/Azure/GoogleCloud)

Linux Cloud

HDFS

Python

NOSQL : mongodb

SparkSQL

**Defining Architecture Components of the Big Data Ecosystem**

**Core Hadoop Components**

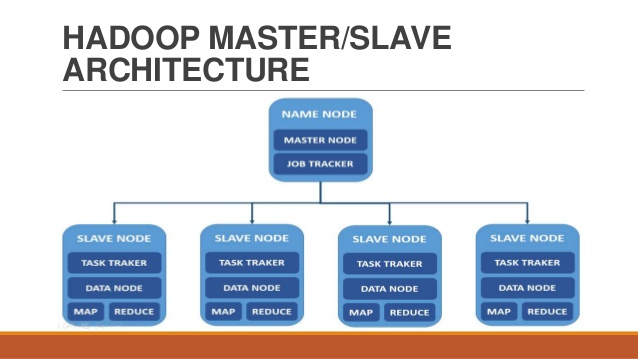
The Hadoop Ecosystem comprises of 4 core components –

**1) Hadoop Common-**

Apache Foundation has pre-defined set of utilities and libraries that can be used by other modules within the Hadoop ecosystem. For example, if HBase and Hive want to access HDFS they need to make of Java archives (JAR files) that are stored in Hadoop Common.

**2) Hadoop Distributed File System (HDFS) -**

The default big data storage layer for Apache Hadoop is HDFS. HDFS is the “Secret Sauce” of Apache Hadoop components as users can dump huge datasets into HDFS and the data will sit there nicely until the user wants to leverage it for analysis. HDFS component creates several replicas of the data block to be distributed across different clusters for reliable and quick data access. HDFS comprises of 3 important components-NameNode, DataNode and Secondary NameNode. HDFS operates on a Master-Slave architecture model where the NameNode acts as the master node for keeping a track of the storage cluster and the DataNode acts as a slave node summing up to the various systems within a Hadoop cluster.



**HDFS Use Case-**

Nokia deals with more than 500 terabytes of unstructured data and close to 100 terabytes of structured data. Nokia uses HDFS for storing all the structured and unstructured data sets as it allows processing of the stored data at a petabyte scale.

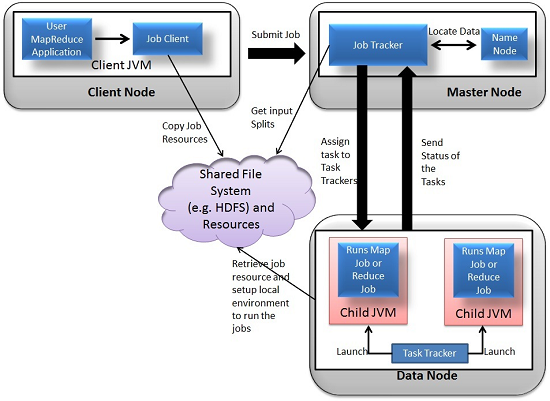
**3) MapReduce- Distributed Data Processing Framework of Apache Hadoop**

MapReduce is a Java-based system created by Google where the actual data from the HDFS store gets processed efficiently. MapReduce breaks down a big data processing job into smaller tasks. MapReduce is responsible for the analysing large datasets in parallel before reducing it to find the results. In the Hadoop ecosystem, Hadoop MapReduce is a framework based on YARN architecture. YARN based Hadoop architecture, supports parallel processing of huge data sets and MapReduce provides the framework for easily writing applications on thousands of nodes, considering fault and failure management.

The basic principle of operation behind MapReduce is that the “Map” job sends a query for processing to various nodes in a Hadoop cluster and the “Reduce” job collects all the results to output into a single value. Map Task in the Hadoop ecosystem takes input data and splits into independent chunks and output of this task will be the input for Reduce Task. In The same Hadoop ecosystem Reduce task combines Mapped data tuples into smaller set of tuples. Meanwhile, both input and output of tasks are stored in a file system. MapReduce takes care of scheduling jobs, monitoring jobs and re-executes the failed task.

MapReduce framework forms the compute node while the HDFS file system forms the data node. Typically in the Hadoop ecosystem architecture both data node and compute node are considered to be the same.

The delegation tasks of the MapReduce component are tackled by two daemons- Job Tracker and Task Tracker as shown in the image below –

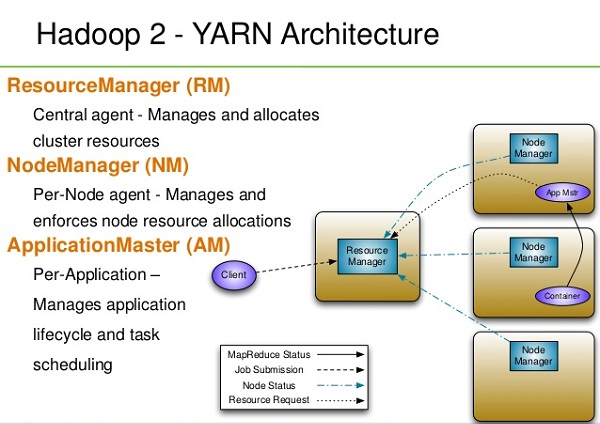


**MapReduce Use Case:**

Skybox has developed an economical image satellite system for capturing videos and images from any location on earth. Skybox uses Hadoop to analyse the large volumes of image data downloaded from the satellites. The image processing algorithms of Skybox are written in C++. Busboy, a proprietary framework of Skybox makes use of built-in code from java based MapReduce framework.

**4)YARN**

YARN forms an integral part of Hadoop 2.0.YARN is great enabler for dynamic resource utilization on Hadoop framework as users can run various Hadoop applications without having to bother about increasing workloads.



**Key Benefits of Hadoop 2.0 YARN Component-**

* It offers improved cluster utilization
* Highly scalable
* Beyond Java
* Novel programming models and services
* Agility

**YARN Use Case:**

Yahoo has close to 40,000 nodes running Apache Hadoop with 500,000 MapReduce jobs per day taking 230 compute years extra for processing every day. YARN at Yahoo helped them increase the load on the most heavily used Hadoop cluster to 125,000 jobs a day when compared to 80,000 jobs a day which is close to 50% increase.

The above listed core components of Apache Hadoop form the basic distributed Hadoop framework. There are several other Hadoop components that form an integral part of the Hadoop ecosystem with the intent of enhancing the power of Apache Hadoop in some way or the other like- providing better integration with databases, making Hadoop faster or developing novel features and functionalities. Here are some of the eminent Hadoop components used by enterprises extensively -

**Data Access Components of Hadoop Ecosystem-  Pig and Hive**

* **Pig-**

**​**Apache Pig is a convenient tools developed by Yahoo for analysing huge data sets efficiently and easily. It provides a high level data flow language Pig Latin that is optimized, extensible and easy to use. The most outstanding feature of Pig programs is that their structure is open to considerable parallelization making it easy for handling large data sets.

**Pig Use Case-**

The personal healthcare data of an individual is confidential and should not be exposed to others. This information should be masked to maintain confidentiality but the healthcare data is so huge that identifying and removing personal healthcare data is crucial. Apache Pig can be used under such circumstances to de-identify health information.

* **Hive-**

**​** Hive developed by Facebook is a data warehouse built on top of Hadoop and provides a simple language known as HiveQL similar to SQL for querying, data summarization and analysis. Hive makes querying faster through indexing.

**Hive Use Case-**

Hive simplifies Hadoop at Facebook with the execution of 7500+ Hive jobs daily for Ad-hoc analysis, reporting and machine learning.

**Data Integration Components of Hadoop Ecosystem- Sqoop and Flume**

* **Sqoop**

**​​**Sqoop component is used for importing data from external sources into related Hadoop components like HDFS, HBase or Hive. It can also be used for exporting data from Hadoop o other external structured data stores. Sqoop parallelized data transfer, mitigates excessive loads, allows data imports, efficient data analysis and copies data quickly.

**Sqoop Use Case-**

Online Marketer Coupons.com uses Sqoop component of the Hadoop ecosystem to enable transmission of data between Hadoop and the IBM Netezza data warehouse and pipes backs the results into Hadoop using Sqoop.

* **Flume-**

**​**Flume component is used to gather and aggregate large amounts of data. Apache Flume is used for collecting data from its origin and sending it back to the resting location (HDFS).Flume accomplishes this by outlining data flows that consist of 3 primary structures channels, sources and sinks. The processes that run the dataflow with flume are known as agents and the bits of data that flow via flume are known as events.

**Flume Use Case –**

Twitter source connects through the streaming API and continuously downloads the tweets (called as events). These tweets are converted into JSON format and sent to the downstream Flume sinks for further analysis of tweets and retweets to engage users on Twitter.

**Data Storage Component of Hadoop Ecosystem –HBase**

**HBase –**

HBase is a column-oriented database that uses HDFS for underlying storage of data. HBase supports random reads and also batch computations using MapReduce. With HBase NoSQL database enterprise can create large tables with millions of rows and columns on hardware machine. The best practice to use HBase is when there is a requirement for random ‘read or write’ access to big datasets.

**HBase Use Case-**

Facebook is one the largest users of HBase with its messaging platform built on top of HBase in 2010.HBase is also used by Facebook for streaming data analysis, internal monitoring system, Nearby Friends Feature, Search Indexing and scraping data for their internal data warehouses.

**Monitoring, Management and Orchestration Components of Hadoop Ecosystem- Oozie and Zookeeper**

* **Oozie-**

**​**Oozie is a workflow scheduler where the workflows are expressed as Directed Acyclic Graphs. Oozie runs in a Java servlet container Tomcat and makes use of a database to store all the running workflow instances, their states ad variables along with the workflow definitions to manage Hadoop jobs (MapReduce, Sqoop, Pig and Hive).The workflows in Oozie are executed based on data and time dependencies.

**Oozie Use Case:**

The American video game publisher Riot Games uses Hadoop and the open source tool Oozie to understand  the player experience.

* **Zookeeper-**

**​**Zookeeper is the king of coordination and provides simple, fast, reliable and ordered operational services for a Hadoop cluster. Zookeeper is responsible for synchronization service, distributed configuration service and for providing a naming registry for distributed systems.

**Zookeeper Use Case-**

Found by Elastic uses Zookeeper comprehensively for resource allocation, leader election, high priority notifications and discovery. The entire service of Found built up of various systems that read and write to   Zookeeper.

Several other common Hadoop ecosystem components include: Avro, Cassandra, Chukwa, Mahout, HCatalog, Ambari and Hama. By implementing Hadoop using one or more of the Hadoop ecosystem components, users can personalize their big data experience to meet the changing business requirements. The demand for big data analytics will make the elephant stay in the big data room for quite some time.

**AMBARI**

A Hadoop component, Ambari is a RESTful API which provides easy to use web user interface for Hadoop management. Ambari provides step-by-step wizard for installing Hadoop ecosystem services. It is equipped with central management to start, stop and re-configure Hadoop services and it facilitates the metrics collection, alert framework, which can monitor the health status of the Hadoop cluster. Recent release of Ambari has added the service check for Apache spark Services and supports Spark 1.6.

**MAHOUT**

Mahout is an important Hadoop component for machine learning, this provides implementation of various machine learning algorithms. This Hadoop component helps with considering user behavior in providing suggestions, categorizing the items to its respective group, classifying items based on the categorization and supporting in implementation group mining or itemset mining, to determine items which appear in group.

**Apache Kafka**

A distributed public-subscribe message  developed by LinkedIn that is fast, durable and scalable.Just like other Public-Subscribe messaging systems ,feeds of messages are maintained in topics.

**Apache Kafka Use Cases**

* Spotify uses Kafka as a part of their log collection pipeline.
* Airbnb uses Kafka in its event pipeline and exception tracking.
* At FourSquare ,Kafka powers online-online and online-offline messaging.
* Kafka power's MailChimp's data pipeline.