**Hadoop 2.x has the following three Major Components**

1. HDFS

2. MapReduce

3. YARN

1. **HDFS** :- 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.

NameNode: NameNode does not store actual data or dataset. NameNode stores Metadata i.e. number of blocks, their location, on which Rack, which Datanode the data is stored and other details. It consists of files and directories.

DataNode: Datanode is responsible for storing actual data in HDFS. It performs read write operation as per requests from the clients.

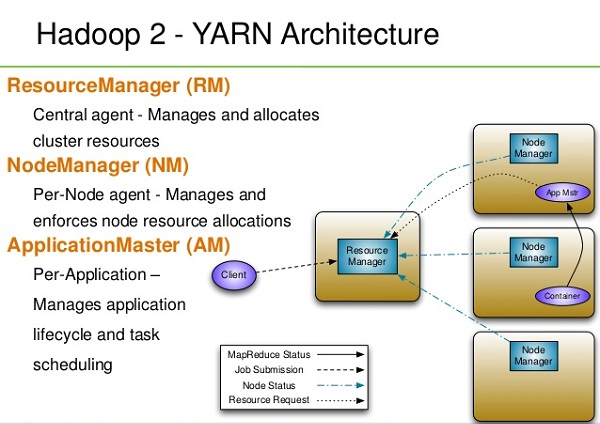
2. **MapReduce**: It is a Java-based system created by Google where the actual data from the HDFS store gets processed efficiently. MapReduce breaks down a big data processing job into smaller tasks. MapReduce is responsible for the analyzing large datasets in parallel before reducing it to find the results.

MapReduce works by breaking the processing into two phases:

i. Map phase : Map function takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs)

ii. Reduce phase : Reduce function takes the output from the Map as an input and combines those data tuples based on the key and accordingly modifies the value of the key

3. **YARN:** YARN stands for Yet Another Resource Negotiator. It is called as the operating system of Hadoop as it is responsible for managing and monitoring workloads. It allows multiple data processing engines such as real-time streaming and batch processing to handle data stored on a single platform.



**Data Access component of Hadoop Ecossytem:**

1. PIG: It is a convenient tools used for analyzing 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.

2. Hive: It 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.

**Data Integration Components of Hadoop Ecosystem:**

1. **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.

2. **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.

**Data Storage Component of Hadoop Ecosystem :**

1.  **HBase**: It 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.

**Monitoring, Management and Orchestration Components of Hadoop Ecosystem :**

1. **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.

2. **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.

3. **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.

4. **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.

5. **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.