**ASSIGNMENT-3.1**

**HADOOP 2.x AND COMPONENTS:**

Hadoop is a master/ slave architecture. The master being the namenode and slaves are datanodes. The namenode controls the access to the data by clients. The datanodes manage the storage of data on the nodes that are running on. Hadoop splits the file into one or more blocks and these blocks are stored in the datanodes. Each data block is replicated to 3 different datanodes to provide high availability of the hadoop system. The block replication factor is configurable.

**HADOOP COMPONENTS:**

The major components of hadoop are: 

* **Hadoop Distributed File System:** HDFS is designed to run on commodity machines which are of low cost hardware. The distributed data is stored in the HDFS file system. HDFS is highly fault tolerant and provides high throughput access to the applications that require big data.
* **Namenode:** Namenode is the heart of the hadoop system. The namenode manages the file system namespace. It stores the metadata information of the data blocks. This metadata is stored permanently on to local disk in the form of namespace image and edit log file. The namenode also knows the location of the data blocks on the data node. However the namenode does not store this information persistently. The namenode creates the block to datanode mapping when it is restarted. If the namenode crashes, then the entire hadoop system goes down. Read more about [Namemode](http://www.folkstalk.com/2013/10/namenode-secondary-safe-mode-hadoop.html)
* **Secondary Namenode:** The responsibility of secondary name node is to periodically copy and merge the namespace image and edit log. In case if the name node crashes, then the namespace image stored in secondary namenode can be used to restart the namenode.
* **DataNode:** It stores the blocks of data and retrieves them. The datanodes also reports the blocks information to the namenode periodically.
* **JobTracker:** JobTracker responsibility is to schedule the clients jobs. Job tracker creates map and reduce tasks and schedules them to run on the datanodes (tasktrackers). Job Tracker also checks for any failed tasks and reschedules the failed tasks on another datanode. Jobtracker can be run on the namenode or a separate node.
* **TaskTracker:** Tasktracker runs on the datanodes. Task trackers responsibility is to run the the map or reduce tasks assigned by the namenode and to report the status of the tasks to the namenode.

## CORE 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](https://www.dezyre.com/hadoop-course/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)YARN

[YARN](https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84) 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.