**Big Data Assignment 3.1**

**Q.** List the Components of Hadoop 2.x and explain each component in detail.

**Ans.** Hadoop 2.x has some common Hadoop API which can easily be integrated with any third party applications to work with Hadoop. It has some new Java APIs and features in HDFS and MapReduce which are known as HDFS2 and MR2 respectively. New architecture has added the architectural features like HDFS High Availability and HDFS Federation. Hadoop 2.x not using Job Tracker and Task Tracker daemons for resource management now on-wards, it is using YARN (Yet Another Resource Negotiator) for Resource Management. There are 3 major components of Hadoop 2.x are HDFS, MapReduce & YARN.

**Explaination :-**

**1) 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.

**2) Map Reduce :-** [MapReduce](https://www.dezyre.com/hadoop-course/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](https://www.dezyre.com/article/hadoop-mapreduce-vs-apache-spark-who-wins-the-battle/83" \o "Hadoop vs Spark" \t "_blank) 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.

**3) YARN :-** YARN stands for Yet Another Resource Negotiator. It is new Component in Hadoop 2.x Architecture. It is also know as “MR V2”. [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. Under YARN 3 major components will work and that are –

**i) ResourceManager** – It is a central agents which manage all allocates cluster resources.  
**ii) NodeManager –** It is a Pre node agents which manages and enforces node resource allocation.  
**iii) ApplicationMaster** – It is a Pre Application which manage application life cycle and task scheduling.