Bigdata Assignment 1.5

Hadoop 2.x has the following three Major Components:

- HDFS
- YARN
- MapReduce

HDFS(Hadoop distributed file System):

It is the primary storage system of Hadoop. Bsically , it is a java based distributed file system that provides scalable, fault tolerance, reliable and cost efficient data storage for Big data that runs on commodity hardware. It is configured with default configuration for many installations. Through shell like commands it can be interacted. It has 3 components name node , secondary name node and data node .

Name Node – It is the master component of HDFS. It stores the metadata which is the directories structures and track of all the files in the folders.

It does not store actual data. It contains fsimage(file system image) which has the directory structure of HDFS, replication factor , file and access permissions ,block size and no of blocks constituting the file. Also it has edits which stores the all modifications done which later on updated with fsimage.

Secondary NameNode – It has the responsibility of checkpointing that is merging of edits with fsimage . As it is time consuming as well as resource consuming it is taken care by other machine i.e Secondary NameNode. It gets the editlogs from the namenodes at regular intervals and creates the checkpoinit.Then copies the fsimage back to namenode.

DataNode – It stores actual data blocks of file in HDFS on its own local disk.It is the slave node.It sends signals to NameNode periodically (called as Heartbeat) to verify it is active or not . It performs all the block operation including periodic checksum and also whete the blocks is tobe stored and what is the replication factor. This is the main workhorse of Hadoop HDFS.

YARN(Yet Another Resource Negotiator) -

It comprises of 2 components-:

- Resource Manager - It is the master daemon of Yarn which manages the available resources (CPU and memory) among all the applications in an disrtribted environment.It has scheduler which is responsible for allocating the resources to the running application. Also it has application manager which monitors the different nodes in the cluster .
- Node Manager – It slave daemon of Yarn. It is responsible for containers which monitors the resource usage and reporting the same to the ResourceManager so that it can distribute tasks or jobs t the datanodes.

MapReduce – It is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks. MapReduce works by breaking the processing into two phases:

- the map phase
- the reduce phase

Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer.

The programmer also specifies two functions that is the map function

and the reduce function. There is a shuffle and sort phase in between.

Map phase takes input in Key-Value pairs and produces output in the form of Key-Value pair which are grouped together on the basis of Key. Then the key and its associated set of values are sent to the Reduce phase where Reduce method applied on the key and associated list of values. Then the output of Reduce is written to HDFS.