**Assignment 3.3**

**Problem Statement:**

List the components of Hadoop 2.x and explain them in detail.

**Answer:**

**Components of Hadoop 2.x:**

* **HDFS**

**NameNode**:

(MASTER DAEMON -1 IN NUMBER)

• Contains the Hadoop FileSystem Tree and other metadata information about files and directories.

• Contains in-memory mapping of which blocks are stored in which datanode. It contains 1.fsimage (file system image) 2.Edits 1.fsimage (file system image) fsimage (file system image) contains:

• A directory structure of the HDFS

• Replication level of the files

• Modification and access times of files

• Access permissions of files and directories

• Block size of files

• Blocks constituting a file 2.Edits

• When any operation takes place in HDFS, the directory structure gets modified

• These modifications are stored in the memory as well as in the edits files further it has PASSIVE NAME NODE and DATANODES HEART BEAT AND BLOCK DETAILS to both active and passive node so that if active node fails this passive name node could be used so that single point failure could be avoided

**Secondary NameNode:**

• Performs house-keeping activities for Name Nodes, like the periodic merging of namespace and edits.

• This is not a backup for a NameNode. Connects to name node every hour. Data Node (SLAVE DAEMON –MANY IN NUMBER)

• Stores actual data blocks of files in HDFS on its own local disk.

• Sends signals to the NameNode periodically (called as Heartbeat) to verify whether it is active Sends block reporting to the NameNode on the cluster startup as well as periodically at every 10th Heartbeat.

• The Data Nodes are the workhorses of a system.

• They perform all the block operations including periodic checksum. They receive instructions from the name node of where to put the blocks and how to put them.

* **MAP REDUCE:**
* It is responsible for the operation to be performed on data YARN (Yet Another Resource Manager Tracker).
* It was responsible for how operation to be performed on data coordinating all task running on all cluster.
* It is responsible for maintaining resources and assigning the task to each datanode responsible for changing nodes in case of failure RESOURCE MANAGER(master daemon-1 IN NUMBER) 1.schedule task across node resource manager get the job and then assign it to each node Yarn schedules the job using 3 scheduler 1.FIFO(first in first out) 2.capacity scheduler 3.fair scheduler 1.FIFO
* The first job which comes to cluster get completed first but is rarely used as it may result in waiting of jobs 2.capacity scheduler
* Here the resource are allocated in que. Say there are 2 jobs namely searching and manipulation.
* Since searching takes less time and is processed quickly But manipulating takes higher time so if manipulating task come first it will take a long time and unnecessarily it makes the searching to wait as per FIFO.
* So in capacity scheduler it divides the resources such that 70% is used for manipulating and 30% is used for searching so both will take place 3.fair scheduler here the resourced among the resources proportionately so none of job waits. By default capacity scheduler is used.
* Node Manager (SLAVE DAEMON –MANY IN NUMBER) it contains application master which is responsible for running multiple processes on a single data node and it also asks for resources from the resource manager if the particular node is overloaded and resource manager then reallocates the particular job.