**SESSION 5: MAPREDUCE INTRODUCTION**

**ASSIGNMENT 2**

**NAMENODE:**

* Namenode is the center piece of the HDFS. It acts as the master node.
* It stores the metadata of the HDFS – the directory tree of all files in the file system, and tracks the file across the cluster.
* Namenode knows the locations of the blocks of any particular file. For each block, the namenode returns the addresses of the datanodes that have a copy of that block.
* Namenode also keeps track of the corrupted data blocks.
* Namenode is a single point of failure in Hadoop. If the namenode is down, the whole cluster goes down and no job execution can take place.
* Namenode is the point of contact for the client applications.

**DATANODE:**

* Datanodes are responsible for storing the actual data in HDFS.
* Datanodes are in constant communication with the Namenode. They send heartbeat signals to the namenode, depicting they are alive and also, to share the block data that reside on them. This helps in keeping the namenode up to date with the most recent status of the cluster.
* Datanodes also communicate with each other, which is what they do during data replication.
* Datanodes are the slave nodes, and the actual processing of the computations occur on them.
* At cluster startup, datanodes report their block mappings to the namenode during a read-only mode called the safe mode.
* Collectively the datanodes make up the majority of the Hadoop Cluster providing fault tolerant data storage.

**RESOURCE MANAGER:**

* Resource Manager acts as the master node in YARN.
* Resource manager is the ultimate authority that arbitrates resources among all the applications in the system.
* Resource manager also takes care of the job scheduling.
* Job/Application is submitted by the YARN client to the Resource Manager.
* Scheduler and Application Manager – are two components of the Resource Manager.
* Resource Manager finds a suitable node manager to launch the application master in a container, which is often known as the First Container of an Application.

**NODE MANAGER:**

* Node Manager acts as the slave node to the Resource Manager, and is YARN’s per node agent.
* It takes care of the individual compute nodes in a Hadoop Cluster.
* Node Managers are responsible for containers, monitoring their resource usage (cpu, memory, disk, network) and reporting the same to the Resource Manager.
* Node Managers receive requests from Resource Managers about resource allocation to jobs.
* An important role of the Node Manager is to start the Application Master service within one of its containers.