**ASSIGNMENT 2.2**

1. **HDFS:**

* HDFS is a Java-based file system that provides scalable and reliable data storage, and it was designed to span large clusters of commodity servers
* HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a wide variety of concurrent data access applications.
* These files are stored in redundant fashion to rescue the system from possible data losses in case of failure.
* HDFS also makes applications available to parallel processing.
* HDFS follows the master-slave architecture and it has two elements namely namenode and data node.
* Namenode(MASTER): It is a software that can be run on commodity hardware. Manages the file system namespace. It also executes file system operations such as renaming, closing, and opening files and directories
* Datanode(SLAVE):These nodes manage the data storage of their system. Datanodes perform read-write operations on the file systems, as per client request.

1. **Hadoop cluster**

* Hadoop clusters are known for boosting the speed of data analysis applications.
* They also are highly scalable: If a cluster's processing power is overwhelmed by growing volumes of DATA, additional cluster nodes can be added to increase throughput.
* They are highly resistant to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails.
* Typically one machine in the cluster is designated as the NameNode and another machine the as JobTracker; these are the masters. The rest of the machines in the cluster act as both DataNode and TaskTracker; these are the slave.
* The only thing that is shared between nodes is the network that connects them.

1. **HDFS Blocks**

* Hadoop distributed file system also stores the data in terms of blocks.
* However the block size in HDFS is very large.
* The default size of HDFS block is 128MB.
* The files are split into 128MB blocks and then stored into the hadoop filesystem.
* The hadoop application is responsible for distributing the data blocks across multiple nodes.
* One file can contain many blocks
* In HDFS, it gets blocks of local file system contiguously to minimise head seek time.