UE18CS322 Lecture Notes : Topic – Hadoop Distributed File System

- What is Hadoop HDFS cluster and in how many different modes can you configure a Hadoop HDFS cluster?
    - Ans: Hadoop cluster is a set of machines configured in a master slave model. The master node of the Hadoop cluster runs the HDFS namenode and the slave nodes run the HDFS datanode. There are also secondary namenodes and standby namenodes in the cluster
- The standby namenode is used for performance. True/False - Justify
    - False: The **secondary** namenode is used for improving the performance of the primary active namenode. The standby namenode is used for fault tolerance.
- What is the usage of HDFS for the read operation?
    - The hdfs client connects to the namenode to retrieve the datanodes on which the blocks are resident. The blocks are then retrieved by connecting to the first data node on the list of datanodes returned by the namenode. In case the first datanode is not able to return the data, it connects to the next datanode on the list.
- Describe the use of Namenode? Why is it the most important process in HDFS
    - The namenode contains the metadata that maps the user file block to the datanodes that contain a copy of the block. It is the most important service as in its absence, the cluster will fail to function. As the namenode is running on the master node, any failure to the master node, renders the entire cluster unusable.
- Can namenode and datanode be on commodity hardware?
    - Yes. Both can be on commodity hardware. Neither namenode or datanode is hardware, but are services running on the hardware.

UE18CS322 Lecture Notes : Topic – Map Reduce

- Q3 short answer from chapter4, T2
    - Input Split – the input file in HDFS that is provided as input to the MR job is split into multiple chunks and a mapper is started for each chunk. Each of these chunks is called an input split. It is best if the input split is the same as the HDFS block size to maximize performance.
- Q4 short answer from chapter4, T2
    - When one of the tasks out of a 100 tasks fails, the map reduce framework will attempt to restart the task probably on a different node.
- Q1 long answer from chapter 4, T2
    - Shuffling is the process of collecting all the values associated with a key produced by the mapper and then deciding which reducer should process a particular key; this is based on the partition function.
    - Sorting is the process of sorting all the keys prior to sending the data from the mapper to the reducer.