SATHYANATHAN H
211720104133 BIG DATA

HOMEWORK 2

Q 1 : What are HDFS and YARN?

HDFS stands for **Hadoop Distributed File System**. HDFS operates as a distributed file system designed to run on commodity hardware. HDFS is fault-tolerant and designed to be deployed on low-cost, commodity hardware.

YARN is **responsible for allocating system resources to the various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes**.

Q 2 : What are the various Hadoop daemons and their roles in a Hadoop cluster?

a)various Hadoop daemons
NameNode
DataNode
Secondary Name Node
Resource Manager
Node Manager

b)Role of the Hadoop daemons
The Node Manager works on the Slaves System that manages the memory resource within the Node and Memory Disk.
Each Slave Node in a Hadoop cluster has a single NodeManager Daemon running in it. It also sends this monitoring information to the Resource Manager.

Q 3: Why does one remove or add nodes in a Hadoop cluster frequently?
Basically, in a Hadoop cluster a Manager node will be deployed on a reliable hardware with high configurations, the Slave node's will be deployed on commodity hardware. So chance's of data node crashing is more .
So more frequently you will see admin's remove and add new data node's in a cluster.

Q 4: What happens when two clients try to access the same file in the HDFS?

HDFS works on write once read many. It means only one client can write a file at a time. **Multiple clients cannot write into an HDFS file at same time**. When one client is given permission by Name node to write data on data node block, the block gets locked till the write operations is completed.

Q 5: How does NameNode tackle DataNode failures?

Data blocks on the failed Datanode are replicated on other Datanodes based on the specified replication factor in hdfs-site.
xml file. Once the failed datanodes comes back the Name node will manage the replication factor again.

Q 6:What will you do when NameNode is down?

If NameNode fails, **the entire Hadoop cluster will fail**. Actually, there will be no data loss, only the cluster job will be shut down because NameNode is just the point of contact for all DataNodes and if the NameNode fails then all communication will stop

Q 7:How is HDFS fault tolerant?

The HDFS is highly fault-tolerant that if any machine fails, the other machine containing the copy of that data automatically become active.
Distributed data storage -
This is one of the most important features of HDFS that makes Hadoop very powerful. Here, data is divided into multiple blocks and stored into nodes

Q 8:Why do we use HDFS for applications having large data sets and not when there are a lot of small files?

HDFS is more efficient for a large number of data sets, maintained in a single file as compared to the small chunks of data stored in multiple files.

Q 9:How do you define “block” in HDFS? What is the default block size in Hadoop 1 and in Hadoop 2? Can it be changed?

Blocks are the smallest continuous location on your hard drive where data is stored. HDFS stores each file as blocks, and distribute it across the Hadoop cluster.