211720104014

BIG DATA HOME WORK 2

Question No:1

1) What are HDFS and YARN?

HDFS and YARN

- HDFS is Hadoop's distributed file system for storing large amounts of data.
- The Hadoop cluster's MapReduce processing paradigm enables distributed processing of enormous amounts of data.
- YARN is in charge of allocating resources among the cluster's applications.

Question No:2

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

Various Hadoop daemons

I. Name Node
II. Data Node
III. Secondary Name Node
IV. Resource Manager
V. Node Manager

Role of the Hadoop daemons

- The Node Manager controls the memory resources within the Node and Memory Disk through the Slaves System.
- A separate Node Manager Daemon is active on each Slave Node in a Hadoop cluster.
- Additionally, the Resource Manager receives this monitoring data from the system.

Question No:3

3) Why does one remove or add nodes in a Hadoop cluster frequently?

- Basically, a Manager node in a Hadoop cluster will be installed on dependable hardware with high settings, while the Slave nodes will be installed on generic hardware.

- Thus, the likelihood of a data node crash is higher.
- As a result, administrators will be adding and removing data nodes from clusters more regularly.

Question No:4

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

- No more than one client may write simultaneously to an HDFS file.
- The block is locked until the write operation has been completed when a client is given authorization to write data on a data node block.
- It is not permissible to do so if another client requests to write to the same block of the same file.

Question No:5

5) How does Name Node tackle Data Node failures?

- Based on the defined replication factor in the HDFS-site.
- xml file, data blocks on the failing Data node are duplicated on other Data nodes.
- The Name node will once more control the replication factor after the failing data nodes have recovered.
- This is how Name node responds to a data node failure.

Question No:6

6) What will you do when Name Node is down?

- The file system goes offline when the Name Node goes down.
- A Secondary NameNode that can be hosted on a different machine is an optional component.
- The only true redundancy it offers is the creation of namespace checkpoints through the merging of the changes file and the fsimage file.

211720104014

Question No:7

7) How is HDFS fault tolerant?

- Due to the great fault tolerance of the HDFS, in the event that one system fails, the other machine holding a duplicate of the data immediately takes over.
- Distributed data storage is one of the key components of HDFS that gives Hadoop its tremendous capability.
- Data is separated into many chunks and stored in nodes in this instance.

Question No:8

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

    a. As opposed to small data chunks saved in numerous files, HDFS is more effective at maintaining a large number of data sets in a single file.

Question No:9

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?

I. The smallest continuous area on your hard disk where data is stored is called a block.

II. Each file is stored as a block by HDFS, which distributes it throughout the Hadoop cluster.

III. A block in HDFS has a default size of 128 MB (for Hadoop 2.x) and 64 MB (Hadoop 1.x)

IV. Yes, it is possible to alter the Hadoop block size.