

#Question 1: What are HDFS and YARN?

HDFS and YARN:

HDFS is the distributed file system in Hadoop for storing big data.

MapReduce is the processing framework for processing vast data in the Hadoop cluster in a distributed manner.

YARN is responsible for managing the resources amongst applications in the cluster.

#Question 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.

#Question 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.

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

Multiple clients can't write into HDFS file at the similar time. When a client is granted a

permission to write data on data node block,
the block gets locked till the completion of a write operation.

If some another client request to write on the same block of the same file then it is not permitted

to do so.

#Question 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.

This is how Namenode handles the failure of data node.

#Question 6: What will you do when NameNode is down?

When the NameNode goes down, the file system goes offline. There is an optional SecondaryNameNode that can be hosted on a separate machine.

It only creates checkpoints of the namespace by merging the edits file into the fsimage file and

does not provide any real redundancy.

#Question 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

#Question 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.

#Question 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?

- a) 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.
- b) The default size of a block in HDFS is 128 MB (Hadoop 2.x) and 64 MB (Hadoop 1.x)
- c) Yes we can change the Hadoop block size.