1.If 8TB is the available disk space per node (10 disks with 1 TB, 2 disk for operating system etc. were excluded.). Assuming initial data size is 600 TB. How will you estimate the number of data nodes (n)?

**Data Size  = 600TB**

**RF=3**

**Intermediate =1**

**Total Requirement= (3+1)\*600=2400TB**

**Avaliable Disk size =8TB**

**Total no nodes required =2400/8⇒300Nodes**

2. You have a directory ProjectPro that has the following files – HadoopTraining.txt, \_SparkTraining.txt, #DataScienceTraining.txt, .SalesforceTraining.txt. If you pass the ProjectPro directory to the [Hadoop MapReduce jobs](https://www.projectpro.io/hadoop%20mapreduce%20wordcount%20example-tutorial/hadoop-mapreduce-wordcount-tutorial), how many files are likely to be processed?

**Only 2 files will be processed because .SalesforceTraining.txt,\_SparkTraining.txt are hidden files.**

**Files starting with \_ and . are hidden files.**

**In hadoop hidden files are not directly processed.**

3) Imagine that you are uploading a file of 500MB into HDFS.100MB of data is successfully uploaded into HDFS and another client wants to read the uploaded data while the upload is still in progress. What will happen in such a scenario, will the 100 MB of data that is uploaded will it be displayed?

**Answer is No because we want to read 100 Mb data. Our block size is 128Mb**

**So to read at least 1 block and its replicas should be created on datanode.**

4) When decommissioning the nodes in a Hadoop Cluster, why should you stop all the task trackers?

7)What are the steps followed by the application while running a YARN job when calling a SubmitApplication method?

**Query→Client M/c→ RM→NameNode→Application Master→ RM(AM Request for Container)---->Conatiner Created—>Once job is finished Container and Resources get terminated and then AM also terminated.**

8)Suppose you want to get an HDFS file into a local directory; how would you go about it?

**Hadoop fs -copyToLocal <HDFS Source> <Local Destination>**

**Hadoop fs -get <HDFS Source> <Local Destination>**

9)Suppose you have one table in HBase. It is required to create a Hive table on top of it, where there should not be any manual movement of data. Changes made to the HBase table should be replicated in the Hive table without explicitly making any changes to it. How can you achieve this?

10) What command will you use to copy data from one node in Hadoop to another?

**hdfs dfs -distcp hdfs://namenodeA/apache\_hadoop hdfs://namenodeB/hadoop**

**For Block Location command: hadoop fsck <file path> -files -blocks -locations**

11) How can you kill an application running on YARN?

**sudo yarn application -kill <application-ID>**

12) In MapReduce tasks, each reduce task writes its output to a file named part-r-nnnnn. Here nnnnn is the partition ID associated with the reduce task. Is it possible to ultimately merge these files? Explain your answer.

**Yes We can merge this file by using**

**hadoop fs -getmerge <file 1 file 2> <mergedsinglefile>**

13) There is a YARN cluster in which the total amount of memory available is 40GB. There are two application queues, ApplicationA and ApplicationB. The queue of ApplicationA has 20 GB allocated, while that of ApplicationB has 8GB allocated. Each map task requires an allocation of 32GB. How will the fair scheduler assign the available memory resources under the DRF (Dominant Resource Finder) Scheduler?

14) How does a NameNode know that one of the DataNodes in a cluster is not functioning?

**After every three second datanode sends heartbeat signals to namenode. By that signal namenode knows that whether datanode is active or not**

15) How can you determine the number of map tasks and reduce tasks based on requirements?

**We determine no of map and reducer tasks on the basis logical split**

**Logical split=input split=No of mappers**

17)What are the differences between -copyFromLocal and -put command

**-copyFromLocal ===> here we have to give complete path of source file and and destination file**

**-put =====> Here we don't provide the complete path**