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)?
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, how many files are likely to be 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?
4. When decommissioning the nodes in a Hadoop Cluster, why should you stop all the task trackers?
5. When does a NameNode enter the safe mode?
6. Did you ever run a lopsided job that resulted in out of memory ever? If yes, then how did you handle it?
7. What are the steps followed by the application while running a YARN job when calling a SubmitApplication method?
8. Suppose you want to get an HDFS file into a local directory; how would you go about it?
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?
11. 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.
12. 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?
13. How does a NameNode know that one of the DataNodes in a cluster is not functioning?
14. How can you determine the number of map tasks and reduce tasks based on requirements?
15.