1. What is Default replication factor and how will you change it at file level?
Ans: The replication factor dictates how many copies of a block should be kept in your cluster. By default, its value is 3 and it can be changed based on the requirement.

2. Why do we need replication factor > 1 in production Hadoop cluster?
Ans: As we have only a single system to work with Hadoop.

3. How will you combine the 4 part-r files of a mapreduce job?
Ans: The Mapreduce Join operation is used to combine two datasets. This process involves in writing lots of code to perform the actual join operation. Joining two datasets started by comparing the size of each dataset.

4. What are the Compression techniques in HDFS and which is the best one and why?
Ans: The LZO compression format is composed of many smaller (~256K) blocks of compressed data, allowing jobs to be split along block boundaries.

5. How will you view the compressed files via HDFS command?
Ans: Step 1: Copy any compressed file to your hdfs dir:
Step 2: Now you can use in-build hdfs text command to read this .gz file.

6. What is Secondary Namenode and its Functionalities? why do we need it?
Ans: It periodically merges the namespace image with the edit log to prevent the edit log from becoming too large.
- Requires similar hardware as Namenode machine.

It maintains the filesystem tree and the metadata for all the files and directories in the tree.
- Secondary Namenode whole purpose is to have a checkpoint in HDFS. Its just a helper node for namenode.That's why it also known as checkpoint node inside the community.

7. What is Backup node and how is it different from Secondary namenode?
Ans: But unlike Secondary NameNode or Checkpoint Node, the Backup node does not need to download fsimage and edits files from the active NameNode to create a checkpoint, as it already has an up-to-date state of the namespace in it's own main memory.

8. What is FSimage and editlogs and how they are related?
Ans: The FsImage is stored as a file in the NameNode's local file system. The NameNode keeps an image of the entire file system namespace and file Blockmap in memory.

9. what is default block size in HDFS? and why is it so large?
Ans: The default size of a block in HDFS is 128 MB. The reason of having this huge block size is to minimize the cost of seek and reduce the meta data information generated per block.

10. How will you copy a large file of 50GB into HDFS in parllel

Ans: HDFS is meant to handle huge files. Given the capabilities of HDFS, 80 GB file may not even quilify for huge file if you have average sized cluster.
When you copy file from local to HDFS, it stored in the blocks of 64MB(default) on multiple data nodes. Also each block is replicated (default replication factor is 3) to other data nodes to make data always available in case any node fails. So while writing itself, it creates pallalel chunks of your file.

11. what is Balancing in HDFS?
Ans : HDFS provides a balancer utility. This utility analyzes block placement and balances data across the DataNodes. It keeps on moving blocks until the cluster is deemed to be balanced, which means that the utilization of every DataNode is uniform.

12. What is expunge in HDFS ?
Ans: expunge: This command is used to empty the trash available in an HDFS system.
Syntax: $ hadoop fs –expunge.

----

1. What is the default replication factor of Hadoop cluster? (A)
a. 3 b. 2 c. 4 d. 1

2. Which component in Hadoop Cluster is responsible for serving read and write requests from the file system's clients? (B)
a. Name Node b. Data Node c. Both a & b d. None of the above

3. Which component of Hadoop Cluster manages the file system namespace and regulates access to files by clients? (C)
a. Name Node b. Data Node c. Both a & b d. None of the above

4. If a file size of size 100 MB is stored on HDFS, what would be the split size? (D)
a. 64 MB & 64 MB b. 64 MB & 36 MB c. 100 MB d. None of the above

5. State true or false: MR2 support various MPP modes for data processing? (B)
a. FALSE b. True

6. Which comand of HDFS helps copy files from HDFS to Local file system? (C)
a. copyFromLocal b. copyToLocal c. put d. mv

7. Which Eco system component of Hadoop is good for non sql programmers? (A)
a. Hive b. Hbase c. Flume d. Pig

8. Block size of a Hadoop cluster is configurable by Administrator? (A)
a. TRUE b. FALSE

9. The functions performed by DataNodes in Hadoop Cluster is/are?(A)
a. Data Block Creation b. Data Block Deletion c. Data Block Replication d. All above

10. Find error in below command: hdfs dfs -put /home/user1/abc.txt (C)
a. Target name missing b. Source name should include hdfs:// c. No error

11. Hadoop block size should be multiple of which unit? (C)
a. 32 MB b. 50 MB c. 64 MB d. 70 MB

12. Which component of the hadoop cluster manages data on slave nodes? (B)
a. Name node b. Data Node c. Task Tracker d. Job Tracker

13. MR1 and MR2 are two modes of processing in Hadoop? (A)
a. TRUE b. FALSE

14. What is Hadoop? (C)
a. Open source software for reliable, scalable, distributed computing. b. A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. c. Both a & b d. None of the above

15. Hadoop provides (A)
a. A reliable distributed storage and processing system b. Only distributed storage c. Only processing system d. None of the above

---
By Manikanta.