1. What is Default replication factor and how will you change it at file level?

Answer: Replication factor dictates how many copies of a block should be kept in your cluster. The replication factor is 3 by default and hence any file you create in HDFS will have a replication factor of 3 and each block from the file will be copied to 3 different nodes in your cluster.

2. Why do we need replication factor > 1 in production Hadoop cluster?

Replication ensures the availability of the data. ... You can configure the Replication factor in you hdfs-site. xml file. Here, we have set the replication Factor to one as we have only a single system to work with Hadoop i.e. a single laptop, as we don't have any Cluster with lots of the nodes

3. How will you combine the 4 part-r files of a mapreduce job?

Hadoop -getmerge command is used to merge multiple files in an HDFS(Hadoop Distributed File System) and then put it into one single output file in our local file system. We want to merge the 2 files present inside are HDFS i.e. file1. txt and file2. txt, into a single file output

4. What are the Compression techniques in HDFS and which is the best one and why?

LZO and Snappy, on the other hand, both optimize for speed and are around an order of magnitude faster than gzip, but compress less effectively. Snappy is also significantly faster than LZO for decompression. The winner by pure compression is 7z, which isn't surprising to us. We've seen 7z come on the top of file compression benchmarks time and time again

5. How will you view the compressed files via HDFS command?

There are 2 types : Step 1: Copy any compressed file to your hdfs dir:

[s0998dnz@hdpm1 ~]$ hadoop fs -put logs.tar.gz /tmp/

Step 2: Now you can use in-build hdfs text command to read this .gz file. This command-line will automatically find the right decompressor for any simple text file and print the uncompressed data to standard output:

6. What is Secondary Namenode and its Functionalities? why do we need it?

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. So we now understood all Secondary Namenode does puts a checkpoint in filesystem which will help Namenode to function better.

7. What is Backup node and how is it different from Secondary namenode?

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?

The NameNode uses a transaction log called the EditLog to persistently record every change that occurs to file system metadata. ... The FsImage is stored as a file in the NameNode's local file system too. 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?

The default size of a block in HDFS is 128 MB (Hadoop 2. x) and 64 MB (Hadoop 1. x) which is much larger as compared to the Linux system where the block size is 4KB. 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

No idea

11. what is Balancing in HDFS?

No idea

12. What is expungne in HDFS ?

No idea