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

**Ans:** The Default replication factor is 3. We can change by two methods

* gedit /home/Alexprasanth/hadoop/etc/hadoop/hdfs-site.xml (in Values)
* Hadoop fs -setrep -w 3 /user/hdfs-site.xml

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

**Ans:** Then only we can access from different machines

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

**Ans:** By using -getMerge

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

**Ans:** Deflate, Gzip, Bzip2, LZO, LZ4, Snappy. Snappy is the best because it compression/decompression Library. It aim for very high speeds and reasonable compression.

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

**Ans:** hfs dis -cat /path/filename.gz | gunzip

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

**Ans:** We can assume that is as a backup node but its not, it simply gets edit logs from name node periodically and copies to FSimage. It is a helper node to Namenode. The main purpose is to have checkpoint in HDFS. So it is called as Checkpoint Node.

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

**Ans:** Backup node is an extended checkpoint node that performs checkpointing and also support online streaming of the system edits. And its keeps a copy of the namespace in main memory similar to Namenode.

1. What is FSimage and editlogs and how they are related?

**Ans:** FSimage - It contains the complete state of the file system at a point in time.

EditLogs - An edit file is a log that lists each file system change (file creation, deletion or modification) that was made after the most recent FSimage.

1. What is default block size in HDFS? and why is it so large?

**Ans:** The default block size is 128MB, The reason of having this huge block size is to minimize the cost of seek and reduce the meta data information generated per block.

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

**Ans:** Hadoop comes with a useful program called distcp for copying data to and from hadoop filesystems in parallel.

1. What is Balancing in HDFS?

**Ans:** HDFS provides a balancer utility. This utility analyses block placement and balances data across the DataNodes. It keeps on moving blocks until the cluster is deemed to be balance. Which means that the utilisation of every DataNode is uniform.

1. What is expunge in HDFS ?

**Ans:** This command is used to empty the trash available in HDFS system.

Syntax = hadoop fs -expunge