Hadoop Task -1

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

Default Replication factor is 3

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

replication factor > 1 so that it can work as single system.

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

Merge the output files into single file

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

GZIP

BZIP

LZO

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

By using -ls command

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

It stores the data and maintain the data.The secondary NameNode merges the fsimage and the editslogs files periodically and keeps edits log size within a limit.

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

Backup node: It update the data.

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

FSimage:files stored on OS and contains directory structure.

Editlogs: it is transaction log.it holds the information about metadata.

9. what is default block size in HDFS? and why is it so large?

Default block size in HDFS: 128 MB

It is large because to reduce data information and also cost.

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

When we copy file from hdfs it stored in the blocks of default size on multiple data nodes.so while writing itself it create parllel. It use command of distcp

Hadoop distcp file1 file2

11. what is Balancing in HDFS?

Utilization of every DataNode is uniform.It keeps on moving blocks until the cluster

12. What is expunge in HDFS ?

It is used to clear the data.