HDFS TASK-1

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

1. Replication factor by default is 3. It’s depend on the size of file, should copies many files and place in our cluser.

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

A) In HDFS need relication factor, because if one Datanode will fails, same these data will get another

Datanodes. For this we will get accureance of falut tolerant criteria.

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

A) In Mapreduce job there is one operation used to combine two large datasets. Here involves writing lots of code to get the “join operation”. And joining datasets checking the size of each dataset.

Like: hadoop fs -cat/user/… \* |hadoop fs –put…..

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

A) Here in HDFS compression is main important. Because it very expensive adding memory to the machine. So, we can simply tech,, is compressing.

There is 4 types of compressing:

Gzip,

Bzip2,

snappy,

Lzo

Here Bzip2 is the best , because it’s split easily, and Degree of comprasion is fast, and compression speed also fast.

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

A) By using this command we wll get compressed file.

Like : hdfs dfs –cat/path/file.gz |gunzip

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

A) Here secondary namenode is main purpose is checkpoint in HDFS. It act like **helper** node for namende. It will help namenode to function better.

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

A) The backup node doet not need any download fsimage and edits file from the namenode . It has own up to date of data and it’s own memory.

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

A) This Editlogs is apply any transaction that occur the record in HDFS file or any specific action will occur like (new block, delection,replicatin, copied etc . For everytime namenode restart , this editlog will apply to fsimage and get latest updated picture of the file system.

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

A) Mainly default size bloc kis 128 mb, or 64 mb for hadoop first version. Because large block size is to less the cost of finding and reduce of Metadata information.

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

A) If large file store from local to hdfs, by default 64mb, and it store multiple data nodes. By default 3 and if any node will store parallely, there it get chunks of your file. By using dist cp command get create new copy.

11. what is Balancing in HDFS?

A) The hdfs balancer is a tool for balance the data across the storage device of a cluster.

12. What is expunge in HDFS ?

A) This is also a command, it used to clean the trash available in hdfs system. And it’s wait for total trash wiil full, then it will clean(empty).