Name : Deep Bhatt

Assignment 1

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

Ans : 3 is the replication Factor.

To change the factor at the file level , you need to just use one Command that is : “hdfs dfs –setrep –w 3 /user/hdfs/file.txt ”

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

Ans :

We need >1 because we need to ensure that the data is stored at different locations in different blocks(capacity would be 128). Also, Importantly , to ensure the security of the data.

This is known as FAULT TOLERENCE.

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

Ans :

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

Ans :

Here are some of the methods Listed :

1. GZIP.
2. BZIP2.
3. Snappy.
4. LZO.

- According to me and in my opinion , LZO would be the most Benefited method to be used.

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

Ans :

Step a : Copy the compressed file to the HDFS dir by $hadoop fs-put filename

Step b : Now use the fs text command to view the file.

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

Ans :

Secondary Namenode is the Helper Node for the Hadoop ecosystem.

As we know , Namenode is used to store the metainformation that is meta Data.

If the editlogs of the Name node which grows significantly and stores all types of details of the data and then the namenode tries to apply to the fsimage at the restart time which makes the process more lengthier and it is that moment that the secondary name node comes into picture.

Secondary namenode takes the part of noting the checkpoint on the namenode.

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

Ans :

Backup node is nothing which is also used to create checkpoints in Hadoop. It is an node that is used to do extended checkpointing in the Hadoop.

The main advantage of using backup node is that it can used at the time of online streaming of data.

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

Ans :

The namenode stores information in a file in 2 types that is Fsimage and the EditLogs.

Fsimage will Image snaps which will have the storage information of the file.

Edit logs will have detailed activities being performed on the HDFS.

They are just a assistant to each other where fsimage will help to store the storage information and edit logs will have detailed activities on the HDFS.

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

Ans :

Default block size is 128MB.

It is large due to 2 major Reasons :

1. Cost of Seek.
2. Reduce the Meta Data Information.

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

Ans :

It will be done automatically as it is Hadoop responsibility.

**11. what is Balancing in HDFS?**

Ans : Still Concept Is Unknown to me.

**12. What is expunge in HDFS ?**

Ans : Still Concept Is Unknown to me.

MCQ Answersheet :

|  |  |
| --- | --- |
| Question Number | Answersheet |
| 1 | A |
| 2 | D |
| 3 | A |
| 4 | D |
| 5 | B |
| 6 | B |
| 7 | D |
| 8 | B |
| 9 | D |
| 10 | C |
| 11 | C |
| 12 | D |
| 13 | A |
| 14 | C |
| 15 | A |

Practical Assignment:













