**Name: Bhavesh Padharia- Batch 2**

**MCQ Answer**

**Big Data and Hadoop - 01**

1. What is the default replication factor of Hadoop cluster?

a. 3

b. 2

c. 4

d. 1

Ans: 1-A

2. Which component in Hadoop Cluster is responsible for serving read and write requests from the file system's clients?

a. Name Node

b. Data Node

c. Both a & b

d. None of the above

Ans: 2-B

3. Which component of Hadoop Cluster manages the file system namespace and regulates access to files by clients?

a. Name Node

b. Data Node

c. Both a & b

d. None of the above

Ans: 3-A

4. If a file size of size 100 MB is stored on HDFS, what would be the split size?

a. 64 MB & 64 MB

b. 64 MB & 36 MB

c. 100 MB

d. None of the above

Ans: 4-B

5. State true or false: MR2 support various MPP modes for data processing?

a. FALSE

b. True

Ans: 5-A

6. Which comand of HDFS helps copy files from HDFS to Local file system?

a. copyFromLocal

b. copyToLocal

c. put

d. mv

Ans: get command is the answer

7. Which Eco system component of Hadoop is good for non sql programmers?

a. Hive

b. Hbase

c. Flume

d. Pig

Ans: 7-B

8. Block size of a Hadoop cluster is configurable by Administrator?

a. TRUE

b. FALSE

Ans: 8-A

9. The functions performed by DataNodes in Hadoop Cluster is/are?

a. Data Block Creation

b. Data Block Deletion

c. Data Block Replication

d. All above

Ans: 9-D

10. Find error in below command:

hdfs dfs -put /home/user1/abc.txt

a. Target name missing

b. Source name should include hdfs://

c. No error

Ans: 10-A

11. Hadoop block size should be multiple of which unit?

a. 32 MB

b. 50 MB

c. 64 MB

d. 70 MB

Ans: 11-C

12. Which component of the hadoop cluster manages data on slave nodes?

a. Name node

b. Data Node

c. Task Tracker

d. Job Tracker

Ans: 12-C

13. MR1 and MR2 are two modes of processing in Hadoop?

a. TRUE

b. FALSE

Ans: 13-A

14. What is Hadoop?

a. Open source software for reliable, scalable, distributed computing.

b. A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

c. Both a & b

d. None of the above

Ans: 14-C

15. Hadoop provides

a. A reliable distributed storage and processing system

b. Only distributed storage

c. Only processing system

d. None of the above

Ans: 15-A

**Assignment Questions**

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

The default replication factor is 3.  default replication for all files placed in HDFS. You can also change the replication factor on a per-file basis using the Hadoop FS shell. Alternatively, you can change the replication factor of all the files under a directory.

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

HDFS stores one copy of block of data in one node of one Rack and other two copies in different nodes of different Rack. This ensures Fault tolerance and High availability. Replication Factor: It is basically the number of times Hadoop framework replicate each and every Data Block

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

N/A

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

Compression techniques in HDFS are :

1 Gzip

2 Bzip2

3 Snappy

4 LZO

According to me BZip2

Bzip2 compresses more effectively than gzip. Bzip2's decompression speed is faster than its compression speed, but it is still slower than the other formats. 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.

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

hadoop fs -cat /hdfs\_location/part-00000.gz | zcat | head -n 20

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. 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?

Backup node as the name states its main role is to act as the dynamic Backup for the Filesystem Namespace(Metadata) in the Primary Namenode of the Hadoop Ecosystem.

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?

EditLogs is a transaction log that recorde the changes in the HDFS file system or any action performed on the HDFS cluster such as addtion of a new block, replication, deletion etc., It records the changes since the last FsImage was created, it then merges the changes into the FsImage file to create a new FsImage file.

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

It depends where this file is located. If it is in Local, i will use:

hadoop fs -copyFromLocal <source> <destination> This destination is hdfs location. If this file is on other database server, i can use sqoop for importing this file.

11. what is Balancing in HDFS?

HDFS provides a balancer utility. This utility analyzes block placement and balances data across the DataNodes. It keeps on moving blocks until the cluster is deemed to be balanced, which means that the utilization of every DataNode is uniform

12. What is expunge in HDFS ?

This expunge is used to empty the trash available in an HDFS system.

**HDFS Task-1**

1. What is the Namenode's URI and which file is it configured in?

Namenode's URI is hdfs://localhost:8020, it’s configured with fs.default.name property that’s specified in $HADOOP\_CONF\_DIR/core-site.xml

2. Where on a local file system will Namenode store its image and which file is it configured in?

Namenode will store its image under /home/bhavesh/Hadoop/hadoop\_work/exercise1/name, it's configured with dfs.namenode.name.dir property thats specified in $HADOOP\_CONF\_DIR/hdfs-site.xml

3. Where on a local file system will Datanode store its blocks and which file is it configured in?

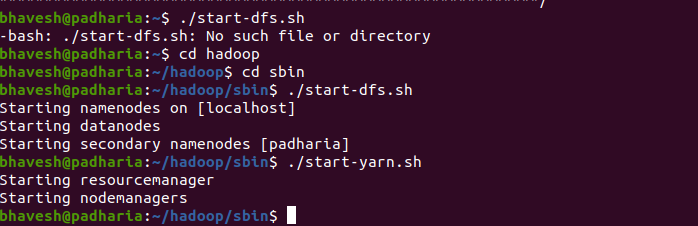
Datanode will store data blocks under /home/bhavesh/hadoop/dfsdata/data, it's configured with dfs.datanode.data.dir property that’s specified in $HADOOP\_CONF\_DIR/hdfs-site.xml

1. What is the block replication and which file is it configured in?

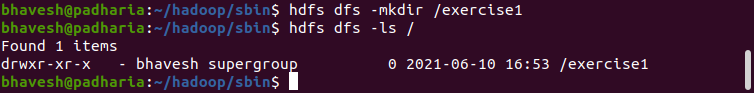
Replication is set to 1, it's configured with dfs.replication property that’s specified in $HADOOP\_CONF\_DIR/hdfs-site.xml

**Practical Task:**

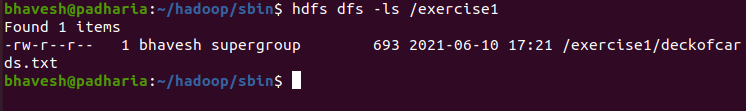
1. Start HDFS and verify that it's running



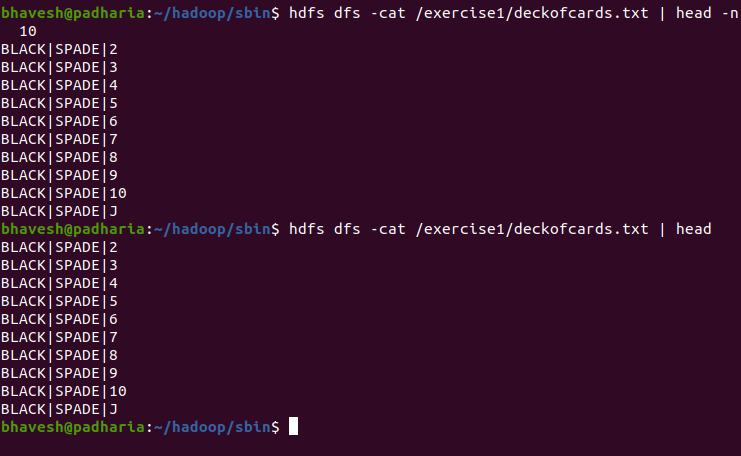
2. Create a new directory /exercise1 on HDFS?



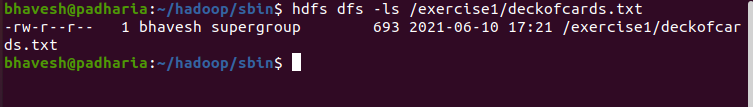
3. Upload GitHub repo sample\_data/deckofcards.txt to HDFS under /exercise1 directory



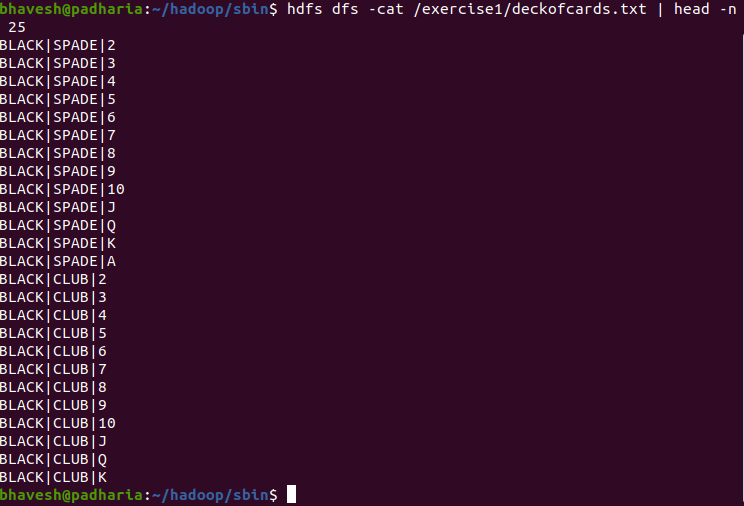
4. View the content of the /exercise1 directory



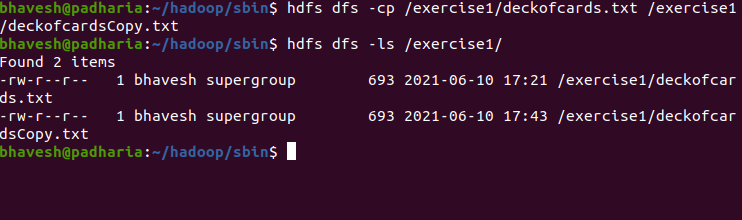
5. Determine the size of the hamlet.txt file in KB that resides on HDFS (not local directory)



6. Print the first 25 lines to the screen from deckofcards.txt on HDFS

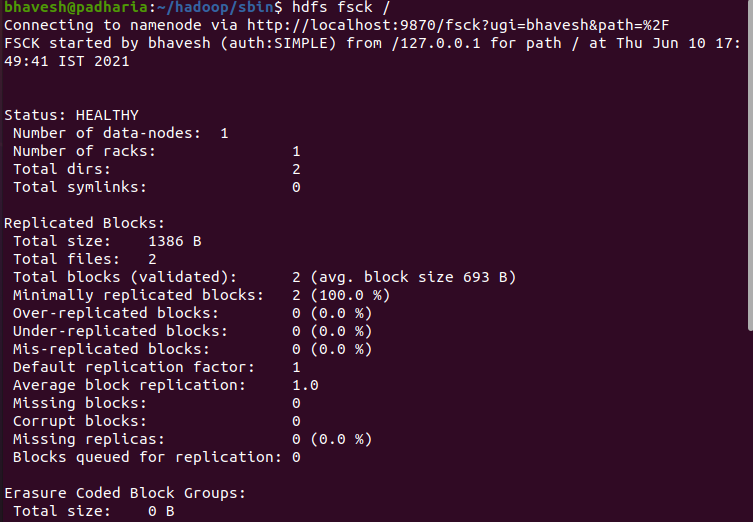


7. Copy deckofcards.txt to deckofcardsCopy.txt

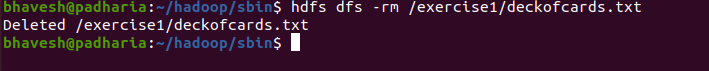


8. Copy deckofcards.txt back to local file system and name it deckofcards.copy.txt

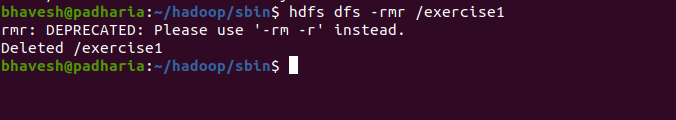
9. Check the entire filesystem for inconsistencies/problems



10. Delete deckofcards.txt from HDFS



11. Delete the /exercise1 directory from HDFS



12. Take a second to look at other available shell options.

