

Big Data (HDFS) Assignment

Problem Statement: Using Command line of HDFS, perform following tasks.

1. Create a directory /hadoop/hdfs/ in HDFS

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir -p /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$
```

Browse Directory

/							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	31/8/2024, 4:52:54 pm	0	0 B	flights
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 11:05:07 am	0	0 B	hadoop
drwxr-xr-x	hadoop	supergroup	0 B	21/11/2022, 3:25:40 pm	0	0 B	hbase
drwxrwxrwx	hadoop	supergroup	0 B	21/11/2022, 3:12:23 pm	0	0 B	tmp
drwxr-xr-x	hadoop	supergroup	0 B	21/11/2022, 3:11:50 pm	0	0 B	user
drwxr-xr-x	hadoop	supergroup	0 B	31/8/2024, 2:30:16 pm	0	0 B	wordcount

Hadoop, 2018.

Browse Directory

/hadoop							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 11:05:07 am	0	0 B	hdfs

Hadoop, 2018.

2. Create a temp directory in Hadoop. Run HDFS command to delete "temp" directory.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir /hadoop/temp
hadoop@hadoop-VirtualBox:~$
```

Browse Directory

/hadoop							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 11:05:07 am	0	0 B	hdfs
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 11:09:30 am	0	0 B	temp

Hadoop, 2018.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -rmdir /hadoop/temp
hadoop@hadoop-VirtualBox:~$
```

Browse Directory

/hadoop							
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 11:05:07 am	0	0 B	hdfs

Hadoop, 2018.

3. List all the files/directories for the given hdfs destination path.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop
Found 1 items
drwxr-xr-x - hadoop supergroup          0 2024-09-03 11:05 /hadoop/hdfs
```

4. Command that will list the directories in /hadoop folder.
hdfs dfs -ls -d /hadoop/*

```
hadoop@hadoop-VirtualBox:~/assignment6$ hdfs dfs -ls /hadoop/hdfs
Found 5 items
-rw-r--r-- 1 hadoop supergroup 7413 2024-09-06 06:31 /hadoop/hdfs/Hadoop_Installation_
-rw-r--r-- 1 hadoop supergroup 5951 2024-09-06 06:31 /hadoop/hdfs/Hbase_Installation.t
drwxr-xr-x - hadoop supergroup 0 2024-09-03 16:13 /hadoop/hdfs/adir
drwxr-xr-x - hadoop supergroup 0 2024-09-03 16:12 /hadoop/hdfs/dira
-rwxrwx--x 3 hadoop supergroup 17 2024-09-06 23:52 /hadoop/hdfs/hdfstest.txt
hadoop@hadoop-VirtualBox:~/assignment6$ hdfs dfs -ls /hadoop/hdfs | grep ^d
drwxr-xr-x - hadoop supergroup 0 2024-09-03 16:13 /hadoop/hdfs/adir
drwxr-xr-x - hadoop supergroup 0 2024-09-03 16:12 /hadoop/hdfs/dira
hadoop@hadoop-VirtualBox:~/assignment6$
```

5. Command to list recursively all files in hadoop directory and all subdirectories in hadoop directory

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -R /hadoop
drwxr-xr-x - hadoop supergroup          0 2024-09-03 11:05 /hadoop/
hadoop@hadoop-VirtualBox:~$
```

6. List all the directory inside /hadoop/hdfs/ directory which starts with 'dir'.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir /hadoop/hdfs/dir1
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir /hadoop/hdfs/dira
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir /hadoop/hdfs/adir
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -d /hadoop/hdfs/dir*/
drwxr-xr-x - hadoop supergroup          0 2024-09-03 16:12 /hadoop/hdfs/d
drwxr-xr-x - hadoop supergroup          0 2024-09-03 16:12 /hadoop/hdfs/d
hadoop@hadoop-VirtualBox:~$
```

7. Create a temp.txt file. Copies this file from local file system to HDFS

```
hadoop@hadoop-VirtualBox:~$ sudo nano /home/hadoop/Downloads/temp.txt
[sudo] password for hadoop:
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/temp.txt /hadoop/
```

8. Copies the file from HDFS to local file system.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyToLocal /hadoop/hdfs/temp.txt /home/hadoop/Downlo
hadoop@hadoop-VirtualBox:~$ ls /home/hadoop/Downloads/
airports_1.csv      'Harry Potter and the Deathly Hallows.txt'  output
airports.csv       HarryPotter_Sorcerers_Stone.txt            People.json
bank_edited.json   Hbase_Installation.txt                    raw_flight_data
Employee_Advance.csv  hivexmlserde-1.0.5.3.jar                 raw_flight_data
flights            HR_Employee.csv                          temp.txt
Hadoop_Installation_Doc.txt  json-serde-1.3.8-jar-with-dependencies.jar
```

9. Command to copy from local directory with the source being restricted to a local file reference.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/temp.txt /hadoop
hadoop@hadoop-VirtualBox:~$
```

10. Command to copies to local directory with the source being restricted to a local file reference.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyToLocal /hadoop/hdfs/temp.txt /home/hadoop/Downlo
hadoop@hadoop-VirtualBox:~$ ls /home/hadoop/Downloads/
airports_1.csv      'Harry Potter and the Deathly Hallows.txt'  output
airports.csv       HarryPotter_Sorcerers_Stone.txt            People.json
bank_edited.json   Hbase_Installation.txt                    raw_flight_data
Employee_Advance.csv  hivexmlserde-1.0.5.3.jar                 raw_flight_data
flights            HR_Employee.csv                          temp.txt
Hadoop_Installation_Doc.txt  json-serde-1.3.8-jar-with-dependencies.jar
```

11. Command to move from local directory source to Hadoop directory.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/temp.txt /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$ sudo rm /home/hadoop/Downloads/temp.t
```

12. Deletes the directory and any content under it recursively.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/dir1
Found 1 items
-rw-r--r--  1 hadoop supergroup      32 2024-09-06 06:06 /hadoop/hdfs/dir1/temp
hadoop@hadoop-VirtualBox:~$ hdfs dfs -rm -r /hadoop/hdfs/dir1
Deleted /hadoop/hdfs/dir1
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs
Found 2 items
drwxr-xr-x  - hadoop supergroup      0 2024-09-03 16:13 /hadoop/hdfs/adir
drwxr-xr-x  - hadoop supergroup      0 2024-09-03 16:12 /hadoop/hdfs/dira
```

13. List the files and show Format file sizes in a human-readable fashion.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/Hadoop_Installation_Doc.txt /hadoop/
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/Hbase_Installation.txt /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du -h /hadoop/hdfs/
7.2 K /hadoop/hdfs/Hadoop_Installation_Doc.txt
5.8 K /hadoop/hdfs/Hbase_Installation.txt
0 /hadoop/hdfs/adir
0 /hadoop/hdfs/dira
```


14. Take a source file and outputs the file in text format on the terminal.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /hadoop/hdfs/Hadoop_Installation_Doc.txt
Prerequisite of HADOOP
-----
sudo addgroup hadoop
sudo adduser --ingroup hadoop hduser

Give Root Previlage:
-----
sudo visudo
myuser ALL=(ALL:ALL)ALL

su - hduser

su - hduser
Download Java 1.8

Download Hadoop

wget http://redrockdigimark.com/apachemirror/hadoop/common/hadoop-2.7.5/hadoop-2.7.5.tar.g
wget link jdk 1.8

Move Java Tar file to /usr/local
sudo mv jdk1.8 /usr/local/
Untar it to location
Move Java1.8 to java
Move Hadoop version to Hadoop
-----
cd /usr/local
sudo tar xvfz filename
sudo mv jdk_1.8 bin java
sudo mv hadoop_2.7.5_bin hadoop
-----
Once it is completed move to home dir hduser
cd

java environmental setup
-----
sudo update-alternatives --install "/usr/bin/java" "java" "/usr/local/java/bin/java" 1
sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/local/java/bin/javac" 1
sudo update-alternatives --install "/usr/bin/javaws" "javaws" "/usr/local/java/bin/javaws" 1
sudo update-alternatives --set java /usr/local/java/bin/java
sudo update-alternatives --set javac /usr/local/java/bin/javac
sudo update-alternatives --set javaws /usr/local/java/bin/javaws

verify it by command
-----
java -version
```

15. Display the content of the HDFS file test on your /user/hadoop2 directory.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /user/hadoop2/test2.txt
and this is content of file: test2.txt
hadoop@hadoop-VirtualBox:~$
```

16. Append the content of a local file test1 to a hdfs file test2.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -appendToFile /home/hadoop/Downloads/test1.txt /user/hadoop2/test2.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /user/hadoop2/test2.txt
and this is content of file: test2.txt
file1.txt contains some content
```

17. Show the capacity, free and used space of the filesystem

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -df /
Filesystem                Size         Used    Available    Us
hdfs://localhost:9000    42005135360  215244800    28012314624
```

18. Shows the capacity, free and used space of the filesystem. Add parameter Formats the sizes of files in a human-readable fashion.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -df -h /
Filesystem                Size         Used    Available    Use%
hdfs://localhost:9000    39.1 G    205.3 M    26.1 G        1%
```

19. Show the amount of space, in bytes, used by the files that match the specified file pattern.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du /*/*.txt
1164038 /wordcount/Harry Potter and the Deathly Hallows.t
439742 /wordcount/HarryPotter_Sorcerers_Stone.txt
```

20. Show the amount of space, in bytes, used by the files that match the specified file pattern. Formats the sizes of files in a human-readable fashion.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du -h /*/*.txt
1.1 M /wordcount/Harry Potter and the Deathly Hallows.txt
429.4 K /wordcount/HarryPotter_Sorcerers_Stone.txt
```

21. Check the health of the Hadoop file system.

← → ↻ local:localhost:50070/explorer.html#/hadoop/hdfs 120% ☆

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

/hadoop/hdfs Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	7.24 KB	6/9/2024, 6:31:25 am	1	128 MB	Hadoop_Installation_Doc.txt
-rw-r--r--	hadoop	supergroup	5.81 KB	6/9/2024, 6:31:40 am	1	128 MB	Hbase_Installation.txt
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 4:13:03 pm	0	0 B	adir
drwxr-xr-x	hadoop	supergroup	0 B	3/9/2024, 4:12:56 pm	0	0 B	dira
-rw-r--r--	hadoop	supergroup	0 B	6/9/2024, 11:24:54 pm	3	128 MB	hdfstest.txt

Hadoop, 2018.

25. Write command to display number of replicas for hdfstest.txt file.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -stat %r /hadoop/hdfs/hdfstest.txt
3
```

26. Write command to Display the status of file "hdfstest.txt" like block size, filesize in bytes.

```
hadoop@hadoop-VirtualBox:~$ hdfs fsck /hadoop/hdfs/hdfstest.txt
Connecting to namenode via http://localhost:50070/fsck?ugi=hadoop&path=%2Fhadoop%2Fhdfs%2Fhdfstest.txt
FSCK started by hadoop (auth:SIMPLE) from /127.0.0.1 for path /hadoop/hdfs/hdfstest.txt at Sat Sep 07 08:49:37 IST 2024

/hadoop/hdfs/hdfstest.txt: Under replicated BP-1998725987-127.0.1.1-1668676850545:blk_1073743145_2322. Target Replicas is 3 but found 1 rep
Status: HEALTHY
Total size: 17 B
Total dirs: 0
Total files: 1
Total symlinks: 0
Total blocks (validated): 1 (avg. block size 17 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 1 (100.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 2 (66.666664 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Sat Sep 07 08:49:37 IST 2024 in 2 milliseconds

The filesystem under path '/hadoop/hdfs/hdfstest.txt' is HEALTHY
```

27. Write HDFS command to change file permission from rw - r - r to rwx-rw-x for hdfstest.txt.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -chmod 761 /hadoop/hdfs/hdfstest.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/hdfstest.txt
-rwxrw--x 3 hadoop supergroup 17 2024-09-06 23:52 /hadoop/hdfs/hdfstest.txt
hadoop@hadoop-VirtualBox:~$
```