Assignment 5 Big Data (HDFS)

Sanjose N , Data Engineering Batch

a) Create a directory /hadoop/hdfs/ in HDFS

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

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir -p /hadoop/temp
hadoop@hadoop-VirtualBox:~$ hdfs dfs -rmr /hadoop/temp
rmr: DEPRECATED: Please use 'rm -r' instead.
Deleted /hadoop/temp
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /
Found 11 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-04 17:36 /FileStream
drwxr-xr-x   - hadoop supergroup          0 2024-09-04 17:57 /FileStreamed
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 09:16 /FileStreamedd
drwxr-xr-x   - hadoop supergroup          0 2024-09-04 17:53 /FileStreams
drwxr-xr-x   - hadoop supergroup          0 2024-09-04 01:16 /airlines
drwxr-xr-x   - hadoop supergroup          0 2024-08-31 16:51 /flights
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 11:03 /hadoop
drwxr-xr-x   - hadoop supergroup          0 2022-11-21 15:25 /hbase
drwxrwxrwx   - hadoop supergroup          0 2022-11-21 15:12 /tmp
drwxr-xr-x   - hadoop supergroup          0 2024-09-03 17:05 /user
drwxr-xr-x   - hadoop supergroup          0 2024-08-31 12:46 /wordcount
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls
Found 5 items
-rw-r--r--   1 hadoop supergroup     439742 2024-08-31 12:29 HarryPotter_Sorcerers_Stone.txt
-rw-r--r--   1 hadoop supergroup    1164038 2024-08-31 12:40 Harry_Potter_and_the_Deathly_Hallows.txt
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 10:15 company
drwxrwxrwx   - hadoop supergroup          0 2022-11-21 15:11 warehouse
drwxr-xr-x   - hadoop supergroup          0 2024-08-31 14:30 wordcount
hadoop@hadoop-VirtualBox:~$
```

d) Command that will list the directories in /hadoop folder.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/
Found 1 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 10:54 /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$
```

e) 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 2024-09-05 23:14 /hadoop/hdfs
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:13 /hadoop/hdfs/subdir
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:14 /hadoop/hdfs/subdir2
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/ | grep ^d | grep '/dir'
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:31 /hadoop/hdfs/directory1
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:31 /hadoop/hdfs/directory2
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ sudo nano temp.txt
[sudo] password for hadoop:
hadoop@hadoop-VirtualBox:~$ hdfs dfs -put temp.txt /hadoop/hdfs/
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/
Found 5 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:31 /hadoop/hdfs/directory1
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:31 /hadoop/hdfs/directory2
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:13 /hadoop/hdfs/subdir
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:14 /hadoop/hdfs/subdir2
-rw-r--r--   1 hadoop supergroup         38 2024-09-05 23:42 /hadoop/hdfs/temp.txt
hadoop@hadoop-VirtualBox:~$
```

h) Copies the file from HDFS to local file system.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -get /hadoop/hdfs/sample1.txt /home/hadoop/Downloads/
hadoop@hadoop-VirtualBox:~$ ls /home/hadoop/Downloads
airports_1.csv                  Harry_Potter_and_the_Deathly_Hallows.txt     People.json
airports.csv                    HarryPotter_Sorcerers_Stone.txt              raw_flight_data.csv
bank_edited.json                Hbase_Installation.txt                       sample1.txt
Employee_Advance.csv            hdfstream                                    sample.txt
flight_data.csv                 hivexmlserde-1.0.5.3.jar                    'Spark Streaming on HDFS.ipynb'
flights                         HR_Employee.csv                             'SQL Query.txt'
FMCG_data.csv                   json-serde-1.3.8-jar-with-dependencies.jar   Telco_Customer_Churn.csv
Hadoop_Installation_Doc.txt     output                                       transactions.csv
hadoop@hadoop-VirtualBox:~$
```

i) 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/users.txt /hadoop/hdfs/directory2
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyToLocal /hadoop/hdfs/directory2/users.txt /home/hadoop/Downloads/testfolder
hadoop@hadoop-VirtualBox:~$
```

k) Command to move from local directory source to Hadoop directory.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -put /home/hadoop/Downloads/rrvv.csv /hadoop/hdfs/
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/
Found 9 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-06 21:12 /hadoop/hdfs/directory2
-rw-r--r--   3 hadoop supergroup          0 2024-09-06 22:38 /hadoop/hdfs/hdfstest.txt
-rw-r--r--   1 hadoop supergroup   72491854 2024-09-06 23:08 /hadoop/hdfs/rrvv.csv
-rw-r--r--   1 hadoop supergroup         21 2024-09-06 00:07 /hadoop/hdfs/sample.txt
-rw-r--r--   1 hadoop supergroup         22 2024-09-06 00:11 /hadoop/hdfs/sample1.txt
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:13 /hadoop/hdfs/subdir
drwxr-xr-x   - hadoop supergroup          0 2024-09-05 23:14 /hadoop/hdfs/subdir2
-rw-r--r--   1 hadoop supergroup         38 2024-09-05 23:42 /hadoop/hdfs/temp.txt
-rw-r--r--   1 hadoop supergroup         35 2024-09-06 21:32 /hadoop/hdfs/test2.txt
hadoop@hadoop-VirtualBox:~$
```

l) Deletes the directory and any content under it recursively.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/directory1/
Found 1 items
-rw-r--r--   1 hadoop supergroup         22 2024-09-06 08:07 /hadoop/hdfs/directory1/sample1.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -rmr /hadoop/hdfs/directory1
rmr: DEPRECATED: Please use 'rm -r' instead.
Deleted /hadoop/hdfs/directory1
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du -h
429.4 K  HarryPotter_Sorcerers_Stone.txt
1.1 M    Harry_Potter_and_the_Deathly_Hallows.txt
0        company
0        warehouse
28.1 K   wordcount
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du -h /
1.5 M    /FileStream
1.5 M    /FileStreamed
429.4 K  /FileStreamedd
1.5 M    /FileStreams
20.8 M   /airlines
69.1 M   /flights
81       /hadoop
13.8 K   /hbase
0        /tmp
278.7 M  /user
1.1 M    /wordcount
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /hadoop/hdfs/sample.txt
This is sample text.
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /user/hadoop2/test.txt
This is the test file for hadoop2 directory.
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -appendToFile /home/hadoop/test1.txt /hadoop/hdfs/test2.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /hadoop/hdfs/test2.txt
This is test1 document for Hadoop.
hadoop@hadoop-VirtualBox:~$
```

q) Show the capacity, free and used space of the filesystem

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -df /
Filesystem                  Size       Used    Available  Use%
hdfs://localhost:9000  41954803712  403922944  27626147840    1%
hadoop@hadoop-VirtualBox:~$
```

r) 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  385.2 M     25.7 G    1%
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -R / | grep '\.txt$' |  awk '{print $8}' |xargs hdfs dfs -du
439742  /FileStream/HarryPotter_Sorcerers_Stone.txt
1164038  /FileStream/Harry_Potter_and_the_Deathly_Hallows.txt
439742  /FileStreamed/HarryPotter_Sorcerers_Stone.txt
1164038  /FileStreamed/Harry_Potter_and_the_Deathly_Hallows.txt
439742  /FileStreamedd/HarryPotter_Sorcerers_Stone.txt
439742  /FileStreams/HarryPotter_Sorcerers_Stone.txt
1164038  /FileStreams/Harry_Potter_and_the_Deathly_Hallows.txt
206  /hadoop/hdfs/directory2/users.txt
21  /hadoop/hdfs/sample.txt
22  /hadoop/hdfs/sample1.txt
38  /hadoop/hdfs/temp.txt
35  /hadoop/hdfs/test2.txt
439742  /user/hadoop/HarryPotter_Sorcerers_Stone.txt
1164038  /user/hadoop/Harry_Potter_and_the_Deathly_Hallows.txt
45  /user/hadoop2/test.txt
1164038  /wordcount/Harry_Potter_and_the_Deathly_Hallows.txt
hadoop@hadoop-VirtualBox:~$
```

t) 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 -ls -R / | grep '\.txt$' |  awk '{print $8}' |xargs hdfs dfs -du -h
429.4 K  /FileStream/HarryPotter_Sorcerers_Stone.txt
1.1 M  /FileStream/Harry_Potter_and_the_Deathly_Hallows.txt
429.4 K  /FileStreamed/HarryPotter_Sorcerers_Stone.txt
1.1 M  /FileStreamed/Harry_Potter_and_the_Deathly_Hallows.txt
429.4 K  /FileStreamedd/HarryPotter_Sorcerers_Stone.txt
429.4 K  /FileStreams/HarryPotter_Sorcerers_Stone.txt
1.1 M  /FileStreams/Harry_Potter_and_the_Deathly_Hallows.txt
206  /hadoop/hdfs/directory2/users.txt
21  /hadoop/hdfs/sample.txt
22  /hadoop/hdfs/sample1.txt
38  /hadoop/hdfs/temp.txt
35  /hadoop/hdfs/test2.txt
429.4 K  /user/hadoop/HarryPotter_Sorcerers_Stone.txt
1.1 M  /user/hadoop/Harry_Potter_and_the_Deathly_Hallows.txt
45  /user/hadoop2/test.txt
1.1 M  /wordcount/Harry_Potter_and_the_Deathly_Hallows.txt
hadoop@hadoop-VirtualBox:~$
```

u) Check the health of the Hadoop file system.

```
hadoop@hadoop-VirtualBox:~$
hadoop@hadoop-VirtualBox:~$ hdfs fsck /
Connecting to namenode via http://localhost:50070/fsck?ugi=hadoop&path=%2F
FSCK started by hadoop (auth:SIMPLE) from /127.0.0.1 for path / at Fri Sep 06 22:22:31 IST 2024
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....Status: HEALTHY
 Total size:    392976523 B
 Total dirs:    91
 Total files:   1329
 Total symlinks:                0
 Total blocks (validated):      1321 (avg. block size 297484 B)
 Minimally replicated blocks:   1321 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Fri Sep 06 22:22:31 IST 2024 in 260 milliseconds


The filesystem under path '/' is HEALTHY
hadoop@hadoop-VirtualBox:~$
```

v) Command to turn off the safemode of Name Node.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfsadmin -safemode leave
Safe mode is OFF
hadoop@hadoop-VirtualBox:~$
```

w) HDFS command to format NameNode.

```
hdfs namenode -format
```

x) Create a file named hdfstest.txt and change it number of replications to 3.

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -touchz /hadoop/hdfs/hdfstest.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -setrep 3 /hadoop/hdfs/hdfstest.txt
Replication 3 set: /hadoop/hdfs/hdfstest.txt
hadoop@hadoop-VirtualBox:~$
```

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

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

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du -s /hadoop/hdfs/hdfstest.txt
0  /hadoop/hdfs/hdfstest.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -stat %b /hadoop/hdfs/hdfstest.txt
0
hadoop@hadoop-VirtualBox:~$
```

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/directory2/
Found 1 items
-rwxrw-r--   1 hadoop supergroup        206 2024-09-06 21:12 /hadoop/hdfs/directory2/users.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -chmod 761 /hadoop/hdfs/directory2/users.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/directory2/
Found 1 items
-rwxrw---x   1 hadoop supergroup        206 2024-09-06 21:12 /hadoop/hdfs/directory2/users.txt
hadoop@hadoop-VirtualBox:~$
```