

Big Data (HDFS) Assignment

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

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

```
hdfs dfs -mkdir -p /hadoop/hdfs
```

output:

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/
Found 1 items
drwxr-xr-x  - hadoop supergroup          0 2024-09-02 16:28 /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$
```

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

```
# create dir
hdfs dfs -mkdir /hadoop/temp

# delete dir
hdfs dfs -rmdir /hadoop/temp
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir /hadoop/temp
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/
Found 2 items
drwxr-xr-x  - hadoop supergroup          0 2024-09-02 16:28 /hadoop/hdfs
drwxr-xr-x  - hadoop supergroup          0 2024-09-02 16:34 /hadoop/temp
hadoop@hadoop-VirtualBox:~$ hdfs dfs -rmdir /hadoop/temp
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop
Found 1 items
drwxr-xr-x  - hadoop supergroup          0 2024-09-02 16:28 /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -ls /
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /
Found 6 items
drwxr-xr-x   - hadoop supergroup          0 2024-08-31 16:51 /flights
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:34 /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 2022-11-21 15:11 /user
drwxr-xr-x   - hadoop supergroup          0 2024-08-31 14:28 /wordcount
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -ls /hadoop | grep '^d'
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop | grep '^d'
drwxr-xr-x   - hadoop supergroup          0 2024-09-07 07:55 /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$
```

e) Command to list recursively all files in hadoop directory and all subdirectories in hadoop directory

```
hdfs dfs -ls -R /hadoop
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -R /hadoop
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:28 /hadoop/hdfs
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:44 /hadoop/subdir1
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:44 /hadoop/subdir1/subsubdir1
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:44 /hadoop/subdir1/subsubdir2
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:45 /hadoop/subdir2
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:45 /hadoop/subdir2/subsubdir1
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:44 /hadoop/subdir3
hadoop@hadoop-VirtualBox:~$
```

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

- approach one:

```
hdfs dfs -ls -d /hadoop/hdfs/dir*
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -d /hadoop/hdfs/dir*
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:48 /hadoop/hdfs/dir1
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:48 /hadoop/hdfs/dir2
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:48 /hadoop/hdfs/dir_final
hadoop@hadoop-VirtualBox:~$
```

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

```
# create file
touch temp.txt

# copy this to hdfs
hdfs dfs -put ./temp.txt /hadoop/hdfs/
```

```
hadoop@hadoop-VirtualBox:~$ touch temp.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -put ./temp.txt /hadoop/hdfs/
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/*.txt
-rw-r--r--    1 hadoop supergroup          0 2024-09-02 16:56 /hadoop/hdfs/temp.txt
hadoop@hadoop-VirtualBox:~$
```

alternative:

```
hdfs dfs -copyFromDevice ./temp.txt /hadoop/hdfs/
```

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

```
hdfs dfs -get /hadoop/hdfs/temp.txt .
```

```
hadoop@hadoop-VirtualBox:~/test$ ls
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -get /hadoop/hdfs/temp.txt .
hadoop@hadoop-VirtualBox:~/test$ ls
temp.txt
hadoop@hadoop-VirtualBox:~/test$
```

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

```
hdfs dfs -copyFromLocal ./file.txt /hadoop/hdfs
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyFromLocal ./file.txt /hadoop/hdfs/
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs/
Found 1 items
-rw-r--r--    1 hadoop supergroup        15 2024-09-07 07:55 /hadoop/hdfs/file.txt
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -copToLocal /hadoop/hdfs/file.txt .
```

```
hadoop@hadoop-VirtualBox:~$ mkdir test
hadoop@hadoop-VirtualBox:~$ cd test
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -copyToLocal /hadoop/hdfs/file.txt .
hadoop@hadoop-VirtualBox:~/test$ ls
file.txt
hadoop@hadoop-VirtualBox:~/test$
```

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

```
hdfs dfs -moveFromLocal ./to_move /hadoop/hdfs/
```

```
hadoop@hadoop-VirtualBox:~/test$ touch to_move.txt
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -moveFromLocal ./to_move.txt /hadoop/hdfs/
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -ls /hadoop/hdfs/
Found 1 items
-rw-r--r--    1 hadoop supergroup          0 2024-09-04 08:42 /hadoop/hdfs/to_move.txt
hadoop@hadoop-VirtualBox:~/test$
```

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

```
hdfs dfs -rm -r /hadoop/hdfs/
```

```
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -ls -R /hadoop/hdfs
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 17:18 /hadoop/hdfs/dir1
-rw-r--r--   1 hadoop supergroup          0 2024-09-02 17:18 /hadoop/hdfs/dir1/test.txt
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:48 /hadoop/hdfs/dir2
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:48 /hadoop/hdfs/dir_final
drwxr-xr-x   - hadoop supergroup          0 2024-09-02 16:49 /hadoop/hdfs/notwithdir
-rw-r--r--   1 hadoop supergroup          0 2024-09-02 16:56 /hadoop/hdfs/temp.txt
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -rm -r /hadoop/hdfs
Deleted /hadoop/hdfs
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -rm -r /hadoop/hdfs
rm: `/hadoop/hdfs': No such file or directory
hadoop@hadoop-VirtualBox:~/test$
```

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

```
hdfs dfs -ls -h /hadoop/
```

```
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -ls -h /hadoop
Found 4 items
-rw-r--r--    1 hadoop supergroup      16 2024-09-03 08:53 /hadoop/sample.txt
drwxr-xr-x    - hadoop supergroup      0 2024-09-02 16:44 /hadoop/subdir1
drwxr-xr-x    - hadoop supergroup      0 2024-09-02 16:45 /hadoop/subdir2
drwxr-xr-x    - hadoop supergroup      0 2024-09-02 16:44 /hadoop/subdir3
hadoop@hadoop-VirtualBox:~/test$
```

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

```
# display source file in local
cat sample.txt
```

```
hadoop@hadoop-VirtualBox:~/test$ cat sample.txt
this is sample textfile
this is line two
hadoop@hadoop-VirtualBox:~/test$
```

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

```
# displaying content of file in hdfs
hdfs dfs -cat /hadoop2/sample.txt
```

```
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -put ./sample.txt /hadoop2/
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -cat /hadoop2/sample.txt
this is sample textfile
this is line two
hadoop@hadoop-VirtualBox:~/test$
```

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

```
hdfs dfs -appendToFile ./to_append.txt /assignment/file.txt
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /assignment/file.txt
this is line 1
hadoop@hadoop-VirtualBox:~$ cat to_append.txt
this is line 2
hadoop@hadoop-VirtualBox:~$ hdfs dfs -appendToFile ./to_append.txt /assignment/f
ile.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /assignment/file.txt
this is line 1
this is line 2
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -report
```

```
hadoop@hadoop-VirtualBox:~/assignment/q3$ hdfs dfsadmin -report
Configured Capacity: 20955348992 (19.52 GB)
Present Capacity: 10329325568 (9.62 GB)
DFS Remaining: 10325348352 (9.62 GB)
DFS Used: 3977216 (3.79 MB)
DFS Used%: 0.04%
Under replicated blocks: 1
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-----
Live datanodes (1):

Name: 127.0.0.1:50010 (localhost)
Hostname: hadoop-VirtualBox
Decommission Status : Normal
Configured Capacity: 20955348992 (19.52 GB)
DFS Used: 3977216 (3.79 MB)
Non DFS Used: 9535610880 (8.88 GB)
```

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

```
hdfs dfs -df -h
```

```
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -df -h
Filesystem              Size      Used    Available   Use%
hdfs://localhost:9000  39.1 G   71.4 M    26.8 G      0%
hadoop@hadoop-VirtualBox:~/test$
```

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

```
hdfs dfs -du -s /hadoop/hdfs/*.txt
```

```
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -du -s /hadoop/hdfs/*.txt
12 /hadoop/hdfs/file2.txt
0 /hadoop/hdfs/to_move.txt
hadoop@hadoop-VirtualBox:~/test$
```

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.

```
hdfs dfs -du -h /hadoop/hdfs/*.txt
```

```
hadoop@hadoop-VirtualBox:~/test$ hdfs dfs -du -h /hadoop/hdfs/*.txt
12 /hadoop/hdfs/file2.txt
34 /hadoop/hdfs/random.txt
0 /hadoop/hdfs/to_move.txt
hadoop@hadoop-VirtualBox:~/test$
```

u) Check the health of the Hadoop file system.

```
hdfs fsck /
```



```
hadoop@hadoop-VirtualBox:~/test$ 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 Wed Sep 04 09:00:06 IST 2024
.....Status: HEALTHY
Total size:      119353460 B
Total dirs:      70
Total files:     96
Total symlinks:   0
Total blocks (validated): 90 (avg. block size 1326149 B)
Minimally replicated blocks: 90 (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 Wed Sep 04 09:00:06 IST 2024 in 18 milliseconds

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

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

```
hdfs dfsadmin -safemode leave
```

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

w) HDFS command to format NameNode.

```
hdfs dfsadmin -format
```

```
hadoop@hadoop-VirtualBox:~$ hdfs namenode -format
24/09/04 21:44:44 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = hadoop-VirtualBox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.6
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/commons-beanutils-1.7.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jets3t-0.9.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-xc-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-json-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/asm-3.2.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-core-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-net-3.1.jar:/usr/local/hadoop/share/hadoop/common/lib/httpcore-4.2.5.jar:/usr/local/hadoop/share/hadoop/common/lib/activation-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-framework-2.7.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-6.1.26.jar:/usr/local/hadoop/share/hadoop/common/lib/xmlenc-0.52.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-server-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-io-2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-annotations-2.7.6.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-codec-1.4.jar:/usr/local/hadoop/share/hadoop/common/lib/httplib-core-3.1.0-
*****/
```

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

```
hdfs dfs -setrep 3 /hadoop/hdfstest.txt
```



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

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

```
hdfs dfs -stat %r /hadoop/hdfstest.txt
```

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

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

```
hdfs dfs -stat "%b %o" /hadoop/hdfstest.txt
```

```
hadoop@hadoop-VirtualBox:~/test/mr_assignments/task1$ hdfs dfs -stat "%b %o" /hadoop/hdfstest.txt
12 134217728
hadoop@hadoop-VirtualBox:~/test/mr_assignments/task1$
```

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

```
hdfs dfs -chmod 761 /hadoop/hdfstest.txt
```

```
hadoop@hadoop-VirtualBox:~/test/mr_assignments/task1$ hdfs dfs -ls /hadoop/hdfstest.txt
-rwxrw---x  3 hadoop supergroup      12 2024-09-05 06:41 /hadoop/hdfstest.txt
hadoop@hadoop-VirtualBox:~/test/mr_assignments/task1$
```