

Assignment 5

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

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

Output

```
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.

```
hdfs dfs -mkdir /temp
hdfs dfs -rmdir /temp
```

Output

```
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.6.jar -file /home/hadoop/MapReduceAssignment/mapper1.py -mapper mapper1.py -file /home/hadoop/MapReduceAssignment/reducer1.py -reducer reducer1.py -input /FMCG_data.csv -output /Assignment/output1/packageJobJar: [/home/hadoop/MapReduceAssignment/mapper1.py, /home/hadoop/MapReduceAssignment/reducer1.py] [] /tmp/streamjob3812317638367137624.jar tmpDir=null
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop fs -cat /Assignment/output1/part-00000
Zone      | Regional Zone | Total Supply | Demand | Status
-----|-----|-----|-----|-----
East      | Zone 1        | 872338       | 3492648 | Demand > Supply
East      | Zone 3        | 2526684     | 10741137 | Demand > Supply
East      | Zone 4        | 3366171     | 13399689 | Demand > Supply
East      | Zone 5        | 1768074     | 6473168  | Demand > Supply
East      | Zone 6        | 1274236     | 5757802  | Demand > Supply
North     | Zone 1        | 18466131    | 73722374 | Demand > Supply
North     | Zone 2        | 18966332    | 76564370 | Demand > Supply
North     | Zone 3        | 21335735    | 85152367 | Demand > Supply
North     | Zone 4        | 26254519    | 165327883 | Demand > Supply
North     | Zone 5        | 42893115    | 177698022 | Demand > Supply
North     | Zone 6        | 100249991   | 416423855 | Demand > Supply
South     | Zone 1        | 14682866    | 57645258 | Demand > Supply
South     | Zone 2        | 32407899    | 132741548 | Demand > Supply
South     | Zone 3        | 18810119    | 76623417 | Demand > Supply
South     | Zone 4        | 19238678    | 77721577 | Demand > Supply
South     | Zone 5        | 24113697    | 97755480 | Demand > Supply
South     | Zone 6        | 38235656    | 125142810 | Demand > Supply
West      | Zone 1        | 150385197   | 42388542 | Demand > Supply
West      | Zone 2        | 15146537    | 62796336 | Demand > Supply
West      | Zone 3        | 28617692    | 88085666 | Demand > Supply
West      | Zone 4        | 43804669    | 182935694 | Demand > Supply
West      | Zone 5        | 32242727    | 129837754 | Demand > Supply
West      | Zone 6        | 52661774    | 218129237 | Demand > Supply
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$
```

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

```
hdfs dfs -ls /
```

Output

```
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.6.jar -file /home/hadoop/MapReduceAssignment/mapper2.py -mapper mapper2.py -file /home/hadoop/MapReduceAssignment/reducer2.py -reducer reducer2.py -input /FMCG_data.csv -output /Assignment/output2/packageJobJar: [/home/hadoop/MapReduceAssignment/mapper2.py, /home/hadoop/MapReduceAssignment/reducer2.py] [] /tmp/streamjob325266464739746260.jar tmpDir=null
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop fs -cat /Assignment/output2/part-00000
Large 4.093814534369161
Mid 4.113473853892216
Small 4.028060694242361
Correlation between wh_capacity_size and num_refilled: 0.7349881101354251
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$
```

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

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

Output

```
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop fs -cat /Assignment/output2/part-00000
Large 4.093814534369161 224729805
Mid 4.113473053892216 222456958
Small 4.028060694242361 105348875
Correlation between wh_capacity_size and num_refilled: 0.7349881101354251
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$
```

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

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

Output

```
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.6.jar -file /home/hadoop/MapReduceAssignment/mapper3.py -mapper mapper3.py -file /home/hadoop/MapReduceAssignment/reducer3.py -reducer reducer3.py -input /FMCG_data.csv -output /Assignment/output3/packageJobJar: [/home/hadoop/MapReduceAssignment/mapper3.py, /home/hadoop/MapReduceAssignment/reducer3.py] [] /tmp/streamjob3830306654094982490.jar tmpDir=null
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop fs -cat /Assignment/output3/part-00000
Issues Total Count Avg_Weights Impact
-----
1 99133868 4644 21346.655469422913 High
2 41450553 2198 18858.304367606914 High
3 32129593 1818 17673.043454345436 Low
4 14896451 777 19171.75160875161 High
5 5788009 348 10632.209770114943 Low
-----
Correlation: -0.8128373527873256
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$
```

Activate Windows
Go to Settings to activate Windows.

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

```
hdfs dfs -ls /hadoop/hdfs | grep "^d" | grep "dir"
```

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

```
touch temp.txt
hdfs dfs -copyFromLocal ~/temp.txt /
```

```

hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop jar /usr/local/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.6.jar -file /home/hadoop/MapReduceAssign
ment/mapper4.py -mapper mapper4.py -file /home/hadoop/MapReduceAssignment/reducer4.py -reducer reducer4.py -input /FMCg_data.csv -output /Assignment/output4/
packageJobJar: [/home/hadoop/MapReduceAssignment/mapper4.py, /home/hadoop/MapReduceAssignment/reducer4.py] [] /tmp/streamjob1699339183848801444.jar tmpDir=null
hadoop@hadoop-VirtualBox:~/MapReduceAssignment$ hadoop fs -cat /Assignment/output4/part-00000
Issues Total      Count      Avg_Weights      Impact
-----
0      4930869      907      5436.459757442117      Low
4      5602095      1080      5187.125      Low
5      8645439      1350      6404.028888888889      Low
6      8158616      1055      7733.285308056872      Low
7      4393171      490      8965.655102040817      Low
8      4120684      405      10174.528395061729      Low
9      9165459      705      11660.809312977099      Low
10     8259859      636      12987.199685534591      Low
11     12270859      866      14169.583140877598      Low
12     11436927      738      15497.19105691057      Low
13     12163798      725      16777.652413793105      Low
14     14535116      820      17725.751219512196      Low
15     17281171      907      19053.11025358324      Low
16     19200310      937      20491.25933831377      Low
17     16410984      748      21947.83957219251      Low
18     24289887      1069      22722.06454630496      Low
19     24569176      1021      24063.83545453585      Low
20     27006058      1064      25381.63345864662      Low
21     18581712      686      27087.043731778427      Low
22     25472459      911      27960.98682766191      High
23     26797528      916      29254.943231441048      High
24     42904667      1423      30150.85523541813      High
25     39461458      1261      31293.781126090405      High
26     19953755      608      32826.89967105263      High
27     19849883      584      33989.525684931505      High
28     12281089      335      36659.96716417911      High
29     12068423      320      37713.821875      High
30     13109614      336      39016.708333333336      High
31     11698085      288      40618.350694444445      High
32     12244881      295      41508.07118644068      High
33     12648436      294      43008.3537414966      High

```

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

```
hdfs dfs -copyToLocal /temp.txt ~/
```

```

hadoop@hadoop-VirtualBox:~$ hdfs dfs -copyToLocal /temp.txt ~/
hadoop@hadoop-VirtualBox:~$ ls
derby.log  Downloads  MapReduceAssignment  Pictures  sample_text2.txt  sample_text.txt  Videos
Desktop    examples.desktop  metastore_db          Public   sample_text3.txt  Templates        Activate Wi
Documents  google-chrome-stable_current amd64.deb  Music    'Pyspark RDD.ipynb'  sample_text4.txt  temp.txt      Go to Settings t
hadoop@hadoop-VirtualBox:~$

```

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

```
hdfs dfs -put test1.txt /hadoop/hdfs/test1.txt
```

Output

```

hadoop@hadoop-VirtualBox:~$ hdfs dfs -mkdir -p /hadoop/hdfs/
hadoop@hadoop-VirtualBox:~$ hdfs dfs -put test1.txt /hadoop/hdfs/test1.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs
Found 1 items
-rw-r--r--  1 hadoop supergroup          35 2024-09-06 21:19 /hadoop/hdfs/test1.txt
hadoop@hadoop-VirtualBox:~$

```

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

```
hdfs dfs -get /hadoop/hdfs/test1.txt ~/test_to_local.txt
```

```

hadoop@hadoop-VirtualBox:~$ hdfs dfs -get /hadoop/hdfs/test1.txt ~/test_to_local.txt
hadoop@hadoop-VirtualBox:~$ ls

```

```
hadoop@hadoop-VirtualBox:~$ ls -l
total 108356
-rw-r--r-- 1 hadoop hadoop      627 Mar 15  2023 derby.log
drwxr-xr-x 2 hadoop hadoop    4096 Nov 17  2022 Desktop
drwxr-xr-x 2 hadoop hadoop    4096 Nov 17  2022 Documents
drwxr-xr-x 9 hadoop hadoop    4096 Sep  6 15:23 Downloads
-rw-r--r-- 1 hadoop hadoop    8980 Nov 17  2022 examples.desktop
-rw-r--r-- 1 hadoop hadoop 110831940 Aug 28 11:42 google-chrome-stable_current_amd64.deb
-rw-rw-r-- 1 hadoop hadoop      0 Sep  6 18:56 hdfstest.txt
drwxr-xr-x 2 hadoop hadoop    4096 Sep  6 09:21 MapReduceAssignment
drwxr-xr-x 5 hadoop hadoop    4096 Mar 15  2023 metastore_db
drwxr-xr-x 2 hadoop hadoop    4096 Nov 17  2022 Music
drwxr-xr-x 3 hadoop hadoop   12288 Sep  5 17:09 Pictures
drwxr-xr-x 2 hadoop hadoop    4096 Nov 17  2022 Public
-rw-r--r-- 1 hadoop hadoop   21590 Aug 30 17:49 'Pyspark RDD.ipynb'
-rw-r--r-- 1 root  root        31 Sep  4 16:59 sample_text2.txt
-rw-r--r-- 1 root  root        48 Sep  4 17:18 sample_text3.txt
-rw-r--r-- 1 hadoop hadoop      48 Sep  4 17:35 sample_text4.txt
-rw-r--r-- 1 root  root        55 Sep  4 16:40 sample_text.txt
drwxr-xr-x 2 hadoop hadoop    4096 Nov 17  2022 Templates
-rw-r--r-- 1 root  root        35 Sep  6 18:10 test1.txt
-rw-r--r-- 1 root  root        40 Sep  6 18:09 test2.txt
-rw-r--r-- 1 hadoop hadoop      35 Sep  6 21:23 test_to_local.txt
-rw-r--r-- 1 root  root       199 Sep  6 17:40 test.txt
drwxr-xr-x 2 hadoop hadoop    4096 Nov 17  2022 Videos
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -moveFromLocal ~/temp.txt /hadoop/hdfs
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -moveFromLocal ~/temp.txt /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls /hadoop/hdfs
Found 4 items
drwxr-xr-x - hadoop supergroup      0 2024-09-06 12:37 /hadoop/hdfs/new
-rw-r--r-- 1 hadoop supergroup      0 2024-09-06 12:37 /hadoop/hdfs/sample1.txt
-rw-r--r-- 1 hadoop supergroup      0 2024-09-06 12:37 /hadoop/hdfs/sample2.txt
-rw-r--r-- 1 hadoop supergroup      0 2024-09-06 15:30 /hadoop/hdfs/temp.txt
```

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

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -rm -r /hadoop/hdfs
Deleted /hadoop/hdfs
hadoop@hadoop-VirtualBox:~$
```

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

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



```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -h /
Found 13 items
drwxr-xr-x - hadoop supergroup 0 2024-09-06 09:49 /HR_Orc
drwxr-xr-x - hadoop supergroup 0 2024-09-06 09:56 /HR_partition
drwxr-xr-x - hadoop supergroup 0 2024-09-03 17:46 /airlines
drwxr-xr-x - hadoop supergroup 0 2024-09-02 09:18 /flights
drwxr-xr-x - hadoop supergroup 0 2024-09-03 17:44 /flights1
drwxr-xr-x - hadoop supergroup 0 2024-09-06 15:33 /hadoop
drwxr-xr-x - hadoop supergroup 0 2022-11-21 15:25 /hbase
-rw-r--r-- 1 hadoop supergroup 0 2024-09-06 12:46 /temp.txt
drwxr-xr-x - hadoop supergroup 0 2024-09-04 17:35 /test
-rw-r--r-- 1 hadoop supergroup 55 2024-09-04 16:54 /test_streaming
-rw-r--r-- 1 hadoop supergroup 55 2024-09-04 16:42 /textfile
drwxrwxrwx - hadoop supergroup 0 2022-11-21 15:12 /tmp
drwxr-xr-x - hadoop supergroup 0 2024-09-03 16:50 /user
```

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

```
cat test.txt
```

```
hadoop@hadoop-VirtualBox:~$ sudo nano test.txt
[sudo] password for hadoop:
hadoop@hadoop-VirtualBox:~$ cat test.txt
India, officially the Republic of India (ISO: Bhārat Gaṇarājya),
[21] is a country in South Asia. It is the seventh-largest country by area;
the most populous country with effect from June 2023
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -cat /user/hadoop2/test.txt
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /user/hadoop2/
India, officially the Republic of India (ISO: Bhārat Gaṇarājya),
[21] is a country in South Asia. It is the seventh-largest country by area;
the most populous country with effect from June 2023
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -appendToFile ~/test1.txt /test2.txt
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -appendToFile ~/test1.txt /test2.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -cat /test2.txt
This is the native text file from test2
This text is from test1 from local
hadoop@hadoop-VirtualBox:~$
```

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

```
hdfs dfs -df
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -df
Filesystem                Size      Used    Available  Use%
hdfs://localhost:9000  41954803712  214269952  27547213824    1%
hadoop@hadoop-VirtualBox:~$
```

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

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -df -h
Filesystem                Size      Used    Available  Use%
hdfs://localhost:9000  39.1 G  204.3 M    25.7 G    1%
```

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

```
hdfs dfs -du /*txt
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du /*txt
0  /hdfstest.txt
0  /temp.txt
199 /test.txt
75 /test2.txt
```

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 /*txt
```

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -du -h /*txt
0  /temp.txt
199 /test.txt
75 /test2.txt
```

u) Check the health of the Hadoop file system.

```
hdfs fsck /
```

```
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 18:38:20 IST 2024
.....
.....Status: HEALTHY
Total size:      204513266 B
Total dirs:      82
Total files:     1408
Total symlinks:   0
Total blocks (validated):    1399 (avg. block size 146185 B)
Minimally replicated blocks: 1399 (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 18:38:20 IST 2024 in 171 milliseconds
```

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

```
hdfs dfsadmin -safemode leave
```

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

w) HDFS command to format NameNode.

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

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

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

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

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

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

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

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -stat "%o %b" /hdfstest.txt
134217728 40
hadoop@hadoop-VirtualBox:~$
```

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

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

```
hadoop@hadoop-VirtualBox:~$ hdfs dfs -chmod 761 /hdfstest.txt
hadoop@hadoop-VirtualBox:~$ hdfs dfs -ls -l /
```

```
-rwxrw--x  3 hadoop supergroup      40 2024-09-07 01:05 /hdfstest.txt
```