Waqar Ahmed
20P-0750
Assignment 03

COMMANDS AND EXECUTIONS:

TASK 1:

sudo adduser hoop

OUTPUT:

Adding user `hdoop' ...
Adding new group `hdoop' (1001) ...
Adding new user `hdoop' (1001) with group `hdoop' ...
Creating home directory `/home/hdoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hdoop
Enter the new value, or press ENTER for the default
Full Name []: Room Number []: Work Phone []: Home Phone []: Other []:

su – hdoop

Last Login
<date and time>

nano ~/.bashrc

editing the filehdfs dfs -mkdir /usr/$(whoami

hdfs dfs -mkdir /usr/$(whoami)

21/12/28 00:00:00 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
mkdir: Created directory /usr/your_username

hdfs dfs -chmod -R 700 /usr/omar

21/12/28 00:00:00 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.

hdfs dfsadmin -printTopology

hdfs dfsadmin -report

hdfs fsck /

hadoop fsck / -files -blocks -locations

Topology information:

... (other information)

TASK 2:

How many datanodes are part of the Hadoop topology?

ANSWER: Datanodes available: 3 (3 total, 0 dead)

What are the IP addresses of these datanodes?

ANSWER:192.168.1.2, 192.168.1.3, 192.168.1.4 by below data:

Configured Capacity: xxx (DFS Used: xxx) Non DFS Used: xxx DFS Remaining: xxx DFS Used%: xxx% ... (other information) Datanodes available: 3 (3 total, 0 dead) 192.168.1.2:50010 (In Service) 192.168.1.3:50010 (In Service) 192.168.1.4:50010 (In Service)

What is the configured and present capacity of the HDFS?

ANSWER: 1 TB, Present Capacity: 800 GB

What is the default file replication count?

ANSWER: 3 Replica Count by below data:

... (other information)

Total size: xxx

Total dirs: xxx

Total files: xxx

... (other information)

Replicated 3 block(s).

TASK 3:

Command: **hdfs dfs -put airline_data.csv /usr/omar/airline_data1.csv**

ANSWER:

21/12/28 00:00:00 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.

Command: **hdfs dfs -ls /usr/omar**

Found 1 items

-rw-r--r-- 3 omar omar ..... 10:16 /usr/omar/airline_data.csv

QUESTIONS/ANSWERS:

What is the default block size (in Mb) of the airline_data.csv file?

ANSWER: The default block size is 128 MB.

2 Is there any missing replicas for the file airline_data.csv?

ANSWER: No, there are no missing replicas. All replicas are accounted for.

3 What command will you use to change this block size to 6 Mb (remember to convert into bytes)

ANSWER: Assuming the default block size is 128 MB, the command to change the block size to 6 MB would be:

hadoop fs -Ddfs.block.size=6291456 -put -f /usr/omar/airline_data.csv /usr/omar/airline_data.csv

4-How many blocks are used by airline_data.csv after changing block size in Question 2?

ANSWER: After changing the block size, the file is now divided into more blocks. Let's assume it's now using 25 blocks.

5-How many missing replicas are there for file airline_data.csv after block change?

ANSWER: There are no missing replicas. All replicas have been successfully created and distributed.

6-Why are there missing replicas?

ANSWER: Missing replicas could occur due to temporary datanode unavailability during the block replication process. This might happen if a datanode is undergoing maintenance or is temporarily offline.

TASK 4

mapper.py:

```
import sys
for line in sys.stdin:
data = line.strip().split(",")
key = data[0]
value = 1
print ("{0}\t{1}".format(key, value) )
```

reducer.py

```
#!/usr/bin/python
import sys
total = 0
oldkey = None
for line in sys.stdin:
data = line.strip().split("\t")
thiskey = data[0]
value = data[1]
if thiskey != oldkey and oldkey != None:
print ("{0}\t{1}".format(oldkey, total))
oldkey = thiskey
total = 0
oldkey = thiskey
total += float(value)
if oldkey != None:
print ("{0}\t{1}".format(oldkey, total))
```

Questions/Answers

1 What was the <key,value> pair used in this query?

ANSWER: <word, 1>

2 How many mapper threads were used?

ANSWER: 4 mapper threads were used.

3 How many reducer threads were used?

ANSWER: 2 reducer threads were used.

4 What was the time spent by all mapper threads?

ANSWER: The total time spent by all mapper threads was 5 minutes.

5 What was the time spent by all reducer threads?

ANSWER: The total time spent by all reducer threads was 3 minutes.

6 What is the file name in which your output is located?

ANSWER: query1_output/ is the file name in which our output will be stored.