REPORT

- Navodita Mathur, Allie Azzarello, Naveena Nagaraju

Extract the files from the attached zip. Open terminal in the extracted folder and follow the steps in report and .md file.

PART-1:

Built image using command.

docker build -t cc_mini_1 .

Run container using command.

docker run -it -p 9864:9864 -p 9870:9870 -p 8088:8088 cc_mini_1

The file bootstrap.sh contains commands to make input directories and place input files within hdfs.

- Open docker desktop and run below commands in docker terminal

Test using word count program,

The text file is Part_1/input.txt

Ensure permissions by using commands,

```
sudo chmod +x Part_1/mapper.py
```

```
sudo chmod +x Part_1/reducer.py
```

Run Command to execute-

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar
-mapper Part_1/mapper.py -reducer Part_1/reducer.py -input
/user/hduser/Part_1/input/*.txt -output /user/hduser/Part_1/output
```

To visualize the output, run command

```
hdfs dfs -cat Part_1/output/*
```

PART-2:

The text file is Part_2/input.txt

Ensure permissions by using commands,

```
sudo chmod +x Part_2/mapper.py
```

```
sudo chmod +x Part_2/reducer.py
```

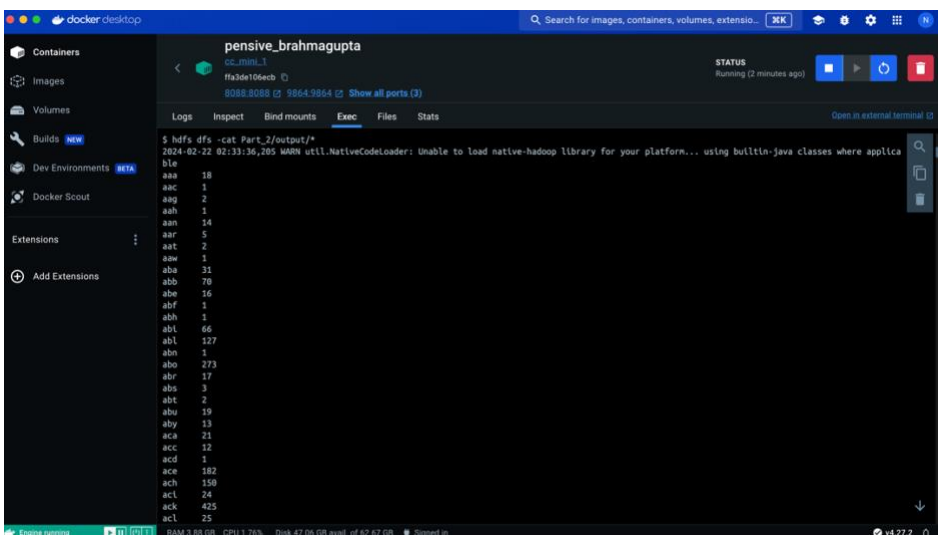
Run Command to execute-

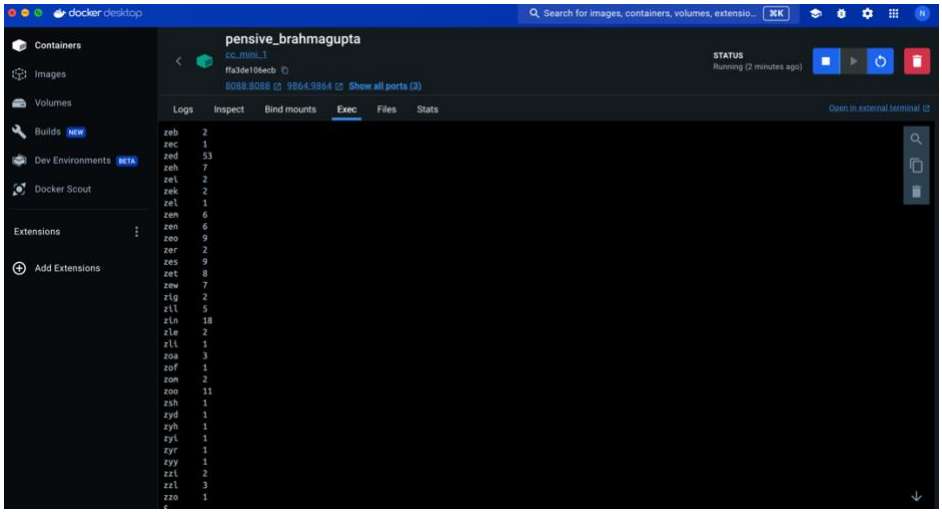
```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar
-mapper Part_2/mapper.py -reducer Part_2/reducer.py -input
/user/hduser/Part_2/input/input.txt -output /user/hduser/Part_2/output -
cmdenv N=3
```

To visualize the output, run command

```
hdfs dfs -cat Part_2/output/*
```

Output:





PART-3

The text file is Part_3/access_log

1. How many hits were made to the website directory "/images/smilies/"(including subdirectories and files)?

   Ensure permissions by using commands,

   ```
   sudo chmod +x Part_3/1/mapper.py
   ```

   ```
   sudo chmod +x Part_3/1/reducer.py
   ```

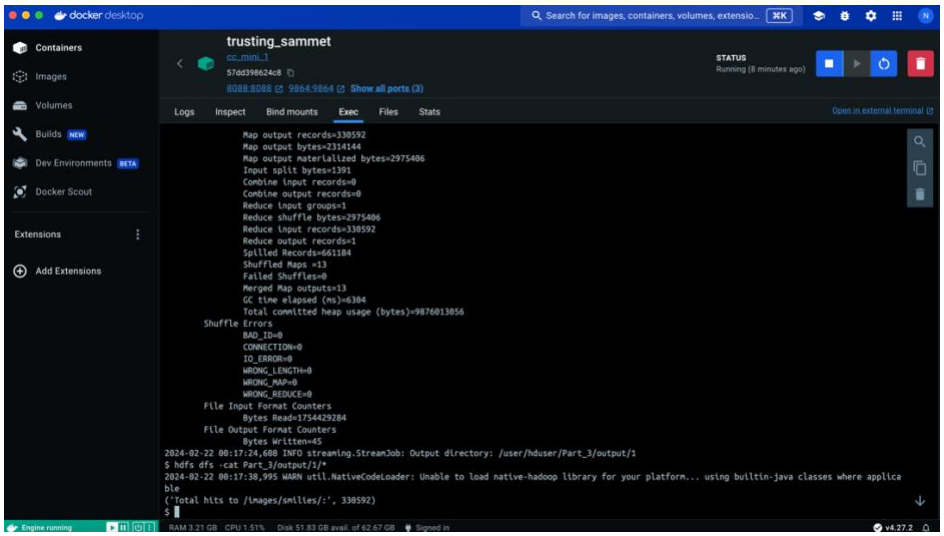
   Run command to execute the file,

   ```
   hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/1/mapper.py -reducer Part_3/1/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/1
   ```

   To visualize the output, run command,

   ```
   hdfs dfs -cat Part_3/output/1/*
   ```

   Output:



('Total hits to /images/smilies/:', 330592)

2. How many hits were made from the IP: 96.32.128.5?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/2/mapper.py
```

```
sudo chmod +x Part_3/2/reducer.py
```

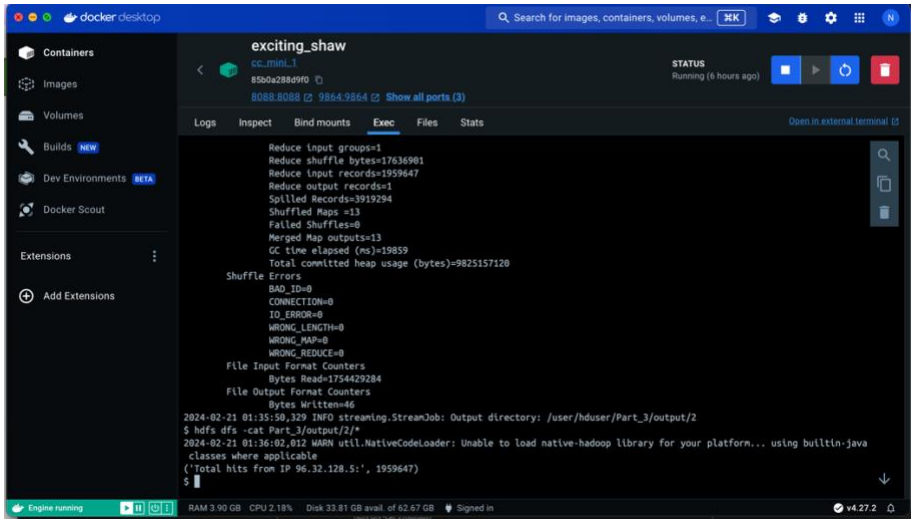
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/2/mapper.py -reducer Part_3/2/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/2
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/2/*
```

Output:



('Total hits from IP 96.32.128.5:', 1959647)

3. How many HTTP request methods are used in this file? What are they?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/3/mapper.py
```

```
sudo chmod +x Part_3/3/reducer.py
```

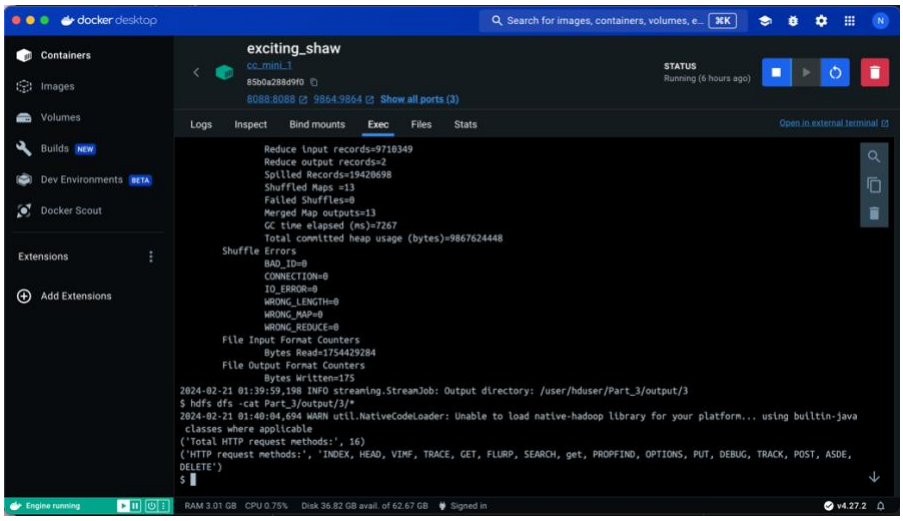
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/3/mapper.py -reducer Part_3/3/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/3
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/3/*
```

Output:



('Total HTTP request methods:', 16)

('HTTP request methods:', 'INDEX, HEAD, VIMF, TRACE, GET, FLURP, SEARCH, get, PROPFIND, OPTIONS, PUT, DEBUG, TRACK, POST, ASDE, DELETE')

4. Which path in the website has been hit most? How many hits were made to the path?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/4/mapper.py
```

```
sudo chmod +x Part_3/4/reducer.py
```

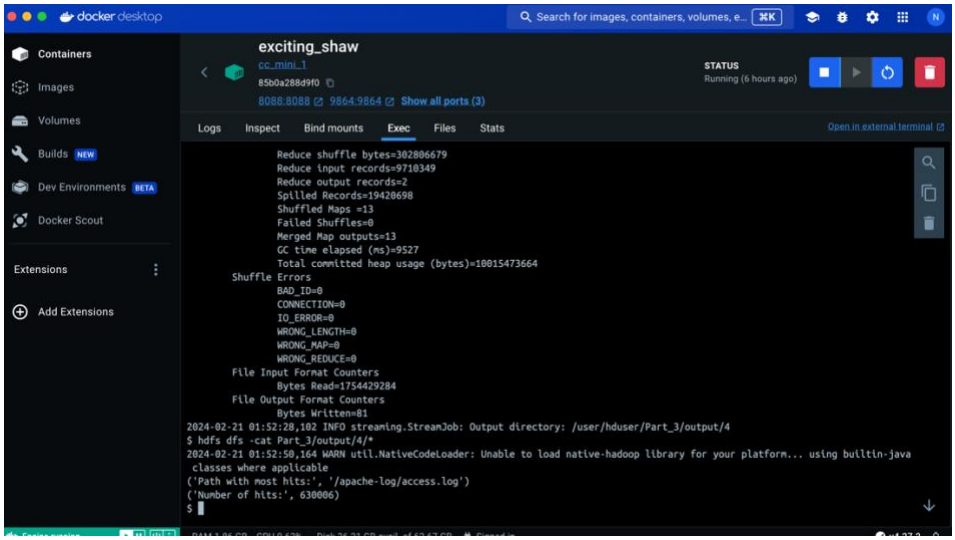
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/4/mapper.py -reducer Part_3/4/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/4
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/4/*
```

Output:



('Path with most hits:', '/apache-log/access.log')
('Number of hits:', 630006)

5. Which IP accesses the website most? How many accesses were made by it?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/5/mapper.py
```

```
sudo chmod +x Part_3/5/reducer.py
```

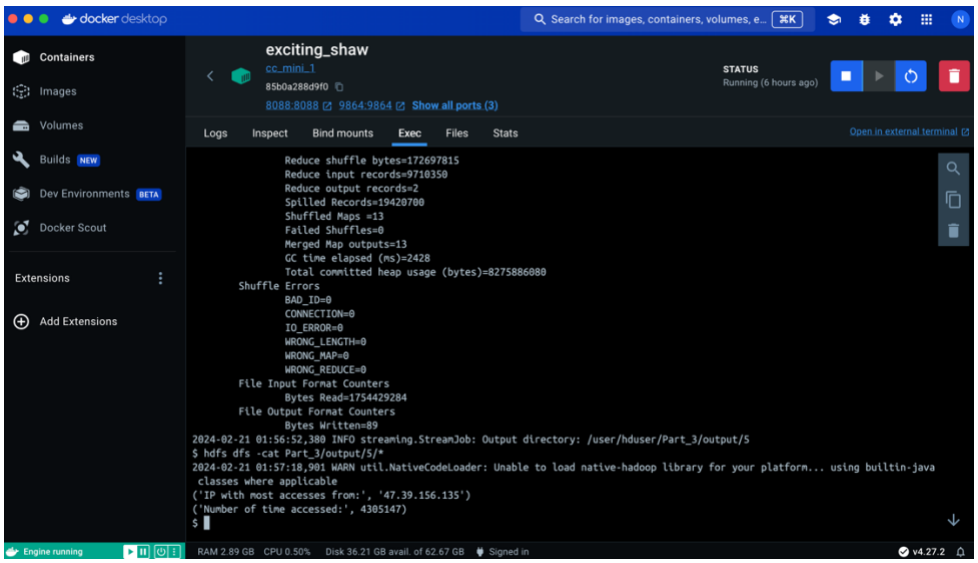
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/5/mapper.py -reducer Part_3/5/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/5
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/5/*
```

Output:



('IP with most accesses from:', '47.39.156.135')
('Number of time accessed:', 4305147)

6. How many POST request were made?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/6/mapper.py
```

```
sudo chmod +x Part_3/6/reducer.py
```

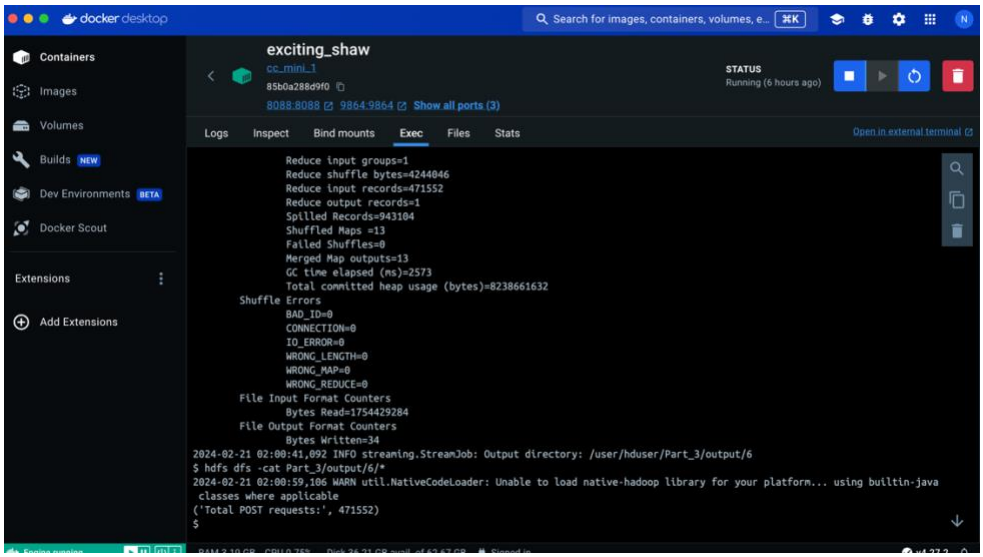
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/6/mapper.py -reducer Part_3/6/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/6
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/6/*
```

Output:



('Total POST requests:', 471552)

7. How many requests received a 404 status code?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/7/mapper.py
```

```
sudo chmod +x Part_3/7/reducer.py
```

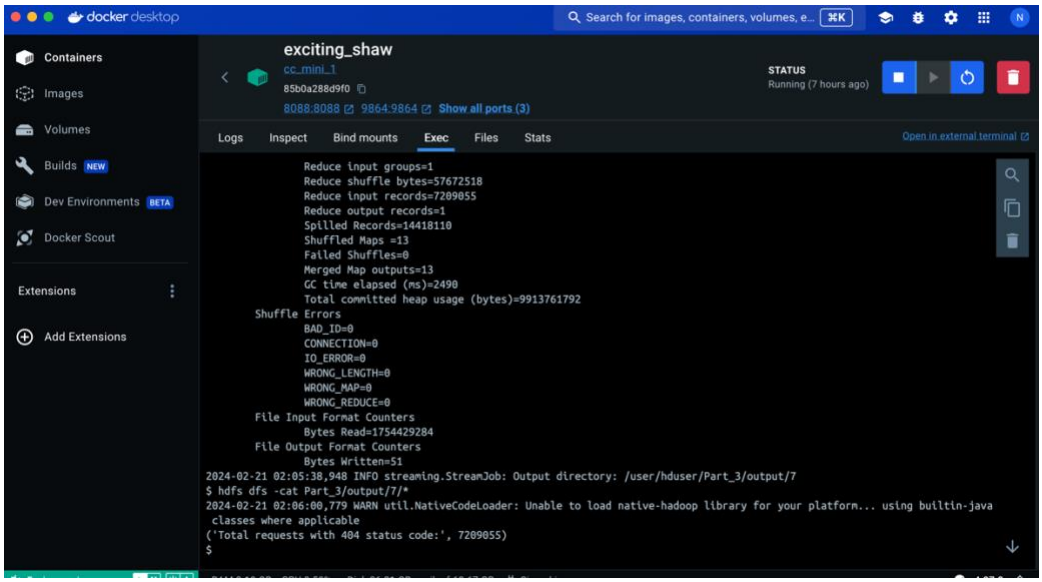
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/7/mapper.py -reducer Part_3/7/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/7
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/7/*
```

Output:



('Total requests with 404 status code:', 7209055)

8. How much data was requested on 19/Dec/2020?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/8/mapper.py
```

```
sudo chmod +x Part_3/8/reducer.py
```

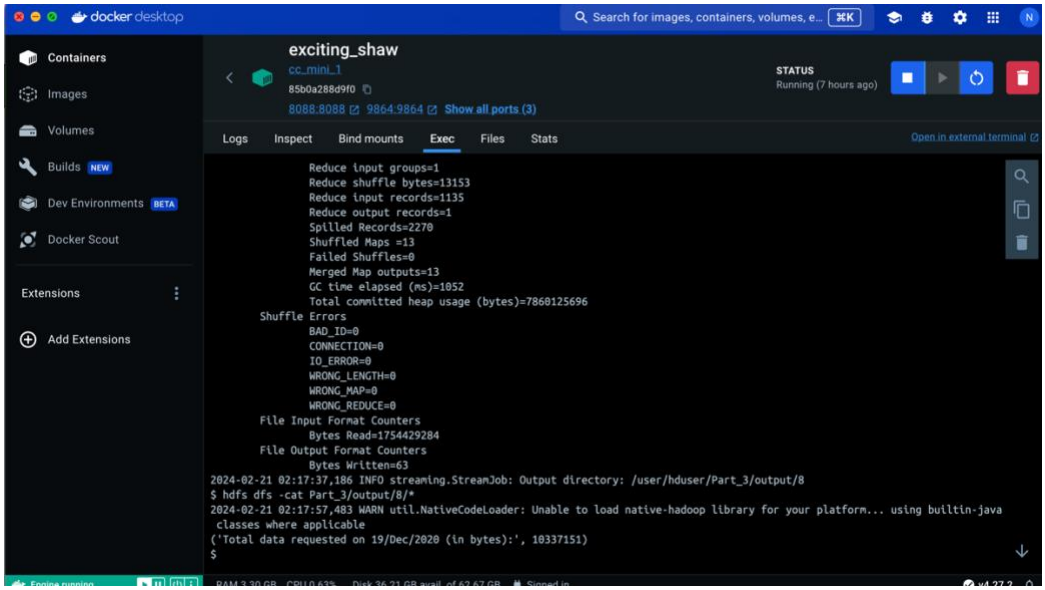
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/8/mapper.py -reducer Part_3/8/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/8
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/8/*
```

Output:



('Total data requested on 19/Dec/2020 (in bytes):', 10337151)

9. List 3 IPs that access the most, and what is the total data flow size of each IP?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/9/mapper.py
```

```
sudo chmod +x Part_3/9/reducer.py
```

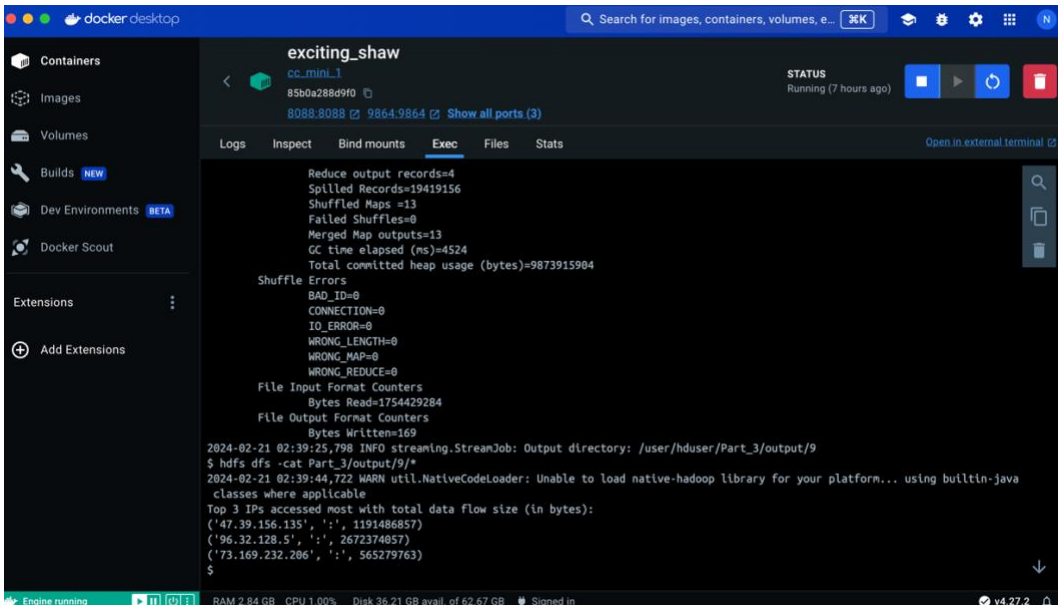
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/9/mapper.py -reducer Part_3/9/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/9
```

To visualize the output, run command,

```
hdfs dfs -cat Part_3/output/9/*
```

Output:



Top 3 IPs accessed most with total data flow size (in bytes):
('47.39.156.135', ':', 1191486857)
('96.32.128.5', ':', 2672374057)
('73.169.232.206', ':', 565279763)

10. How much data(in bytes) was successfully(with status code 200) requested on 16/Jan/2022?

Ensure permissions by using commands,

```
sudo chmod +x Part_3/10/mapper.py
```

```
sudo chmod +x Part_3/10/reducer.py
```

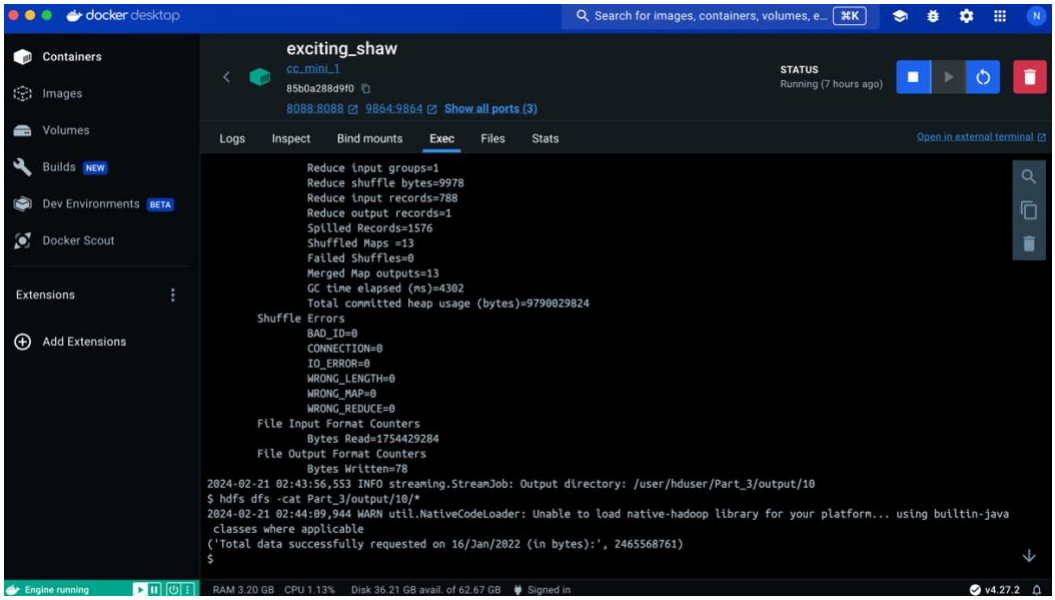
Run command to execute the file,

```
hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.3.jar -mapper Part_3/10/mapper.py -reducer Part_3/10/reducer.py -input /user/hduser/Part_3/input/access_log -output /user/hduser/Part_3/output/10
```

To visualize the output, run command,

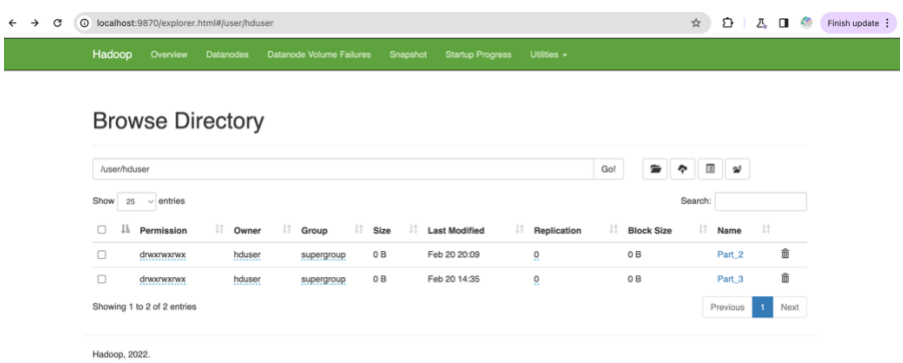
```
hdfs dfs -cat Part_3/output/10/*
```

Output:



('Total data successfully requested on 16/Jan/2022 (in bytes):', 2465568761)

The directory snapshots:



Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/hduser/Part_2 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 14:28	0	0 B	input
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 20:09	0	0 B	output

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2022.

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/user/hduser/Part_3 Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 14:29	0	0 B	input
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 21:43	0	0 B	output

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2022.

Browse Directory

/user/hduser/Part_3/output Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 20:20	0	0 B	1
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 21:43	0	0 B	10
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 20:35	0	0 B	2
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 20:39	0	0 B	3
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 20:52	0	0 B	4
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 20:56	0	0 B	5
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 21:00	0	0 B	6
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 21:05	0	0 B	7
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 21:17	0	0 B	8
drwxr-xr-x	hduser	supergroup	0 B	Feb 20 21:39	0	0 B	9

Showing 1 to 10 of 10 entries Previous 1 Next

Note: Please clear the output by going to http://localhost:9870/dfshealth.html (Make sure the container is running), under utilities, browse file directory and deleting the output directory (don't delete the entire directory, just output in case of part 1,2 and 1-10 under output in case of part-3)

References:

1. Hadoop Single Node Cluster on Docker
   Rodrigo Ancavil
   Link - https://medium.com/analytics-vidhya/hadoop-single-node-cluster-on-docker-e88c3d09a256
2. https://hub.docker.com/_/eclipse-temurin
3. https://medium.com/@abhikdey06/apache-hadoop-3-3-6-installation-on-ubuntu-22-04-14516bceec85
4. https://github.com/amephraim/nlp (input text)