Report

Group Members:

- Bhavana Devulapally
- Shusrita Venugopal
- Neha Navarkar

Our submission (attached zip file) contains separate folders for each Part. Following is the list of files submitted:
Part 1:
1. Dockerfile
2. bootstrap.sh
3. core-site.sh
4. hdfs-site.xml
5. mapred-site.xml
6. yarn-site.xml

Part 2:
1. ngrammaper.py
2. ngramreducer.py
3. input.txt

Part 3:
Contains mapper and reducer for each problem.

Part 1: Setting up Hadoop in Docker

In this project, we explored Docker images and learned how to build them. Docker is a powerful virtualization tool similar to the previous tool, Sandbox, but with additional features for system developers. It combines the functionality of a version control system (VCS) with that of a programming language, such as Java, for virtual machines.

1.To begin, we set up a single node in local Docker by building the Docker image based on the Dockerfile in the "local_demo.zip" file, which contains the necessary code and files for our project.
2. We then setup SSH keys using ssh-keyget command to generate public-private RSA key pair.
3. We configured start-dfs.sh and start-yarn.sh, to initialize the core components of a Hadoop cluster, enabling data storage and resource management for distributed computing tasks. These scripts are typically executed on the master node (or nodes) of the Hadoop cluster to start the necessary services.

4. We configured stop-dfs.sh and stop-yarn.sh, to gracefully stop the core components of a Hadoop cluster, ensuring that data is safely persisted, and resources are released. These scripts are typically executed on the master node (or nodes) of the Hadoop cluster to stop the necessary services.
5. We configured hdfs-site.xml to configure properties related to HDFS, such as replication factor, block size, and data node settings. It contains parameters that govern the behavior of the HDFS components, including the NameNode, DataNode, and Secondary NameNode. Some common properties defined in hdfs-site.xml include:

- dfs.replication: Specifies the default replication factor for file blocks in HDFS.
- dfs.blocksize: Defines the default block size for files in HDFS.
- dfs.namenode.name.dir: Specifies the directory where the NameNode stores its metadata.
- dfs.datanode.data.dir: Defines the directory where DataNode stores HDFS blocks.

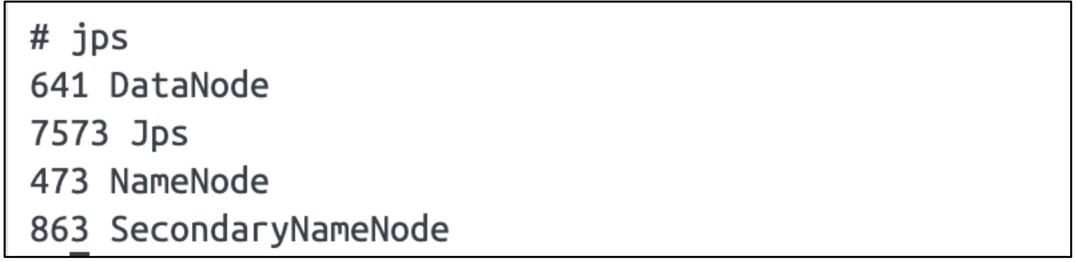
6. We configured mapred.xml to configure properties related to the MapReduce framework, including JobTracker and TaskTracker settings. It defines parameters that control job execution, task scheduling, and resource allocation within the Hadoop cluster. Some common properties defined in mapred.xml include:

- mapred.job.tracker: Specifies the hostname and port of the JobTracker.
- mapred.map.tasks: Defines the number of map tasks to run per job.
- mapred.reduce.tasks: Specifies the number of reduce tasks to run per job.
- mapred.tasktracker.map.tasks.maximum: Defines the maximum number of map tasks that can be executed concurrently on a TaskTracker node.
- mapred.tasktracker.reduce.tasks.maximum: Defines the maximum number of reduce tasks that can be executed concurrently on a TaskTracker node.

7. We then executed The bin/hadoop namenode -format command initializes and prepares the Hadoop Distributed File System (HDFS) by creating necessary metadata structures and clearing any existing data. It ensures a clean and consistent starting state for the HDFS, essential for setting up new clusters or reinstalling HDFS.
8. Here we encountered error for java_home on MAC devices. We set JAVA_HOME and PATH in /opt/hadoop/etc/hadoop/hadoop-enc.sh file using the following commands:

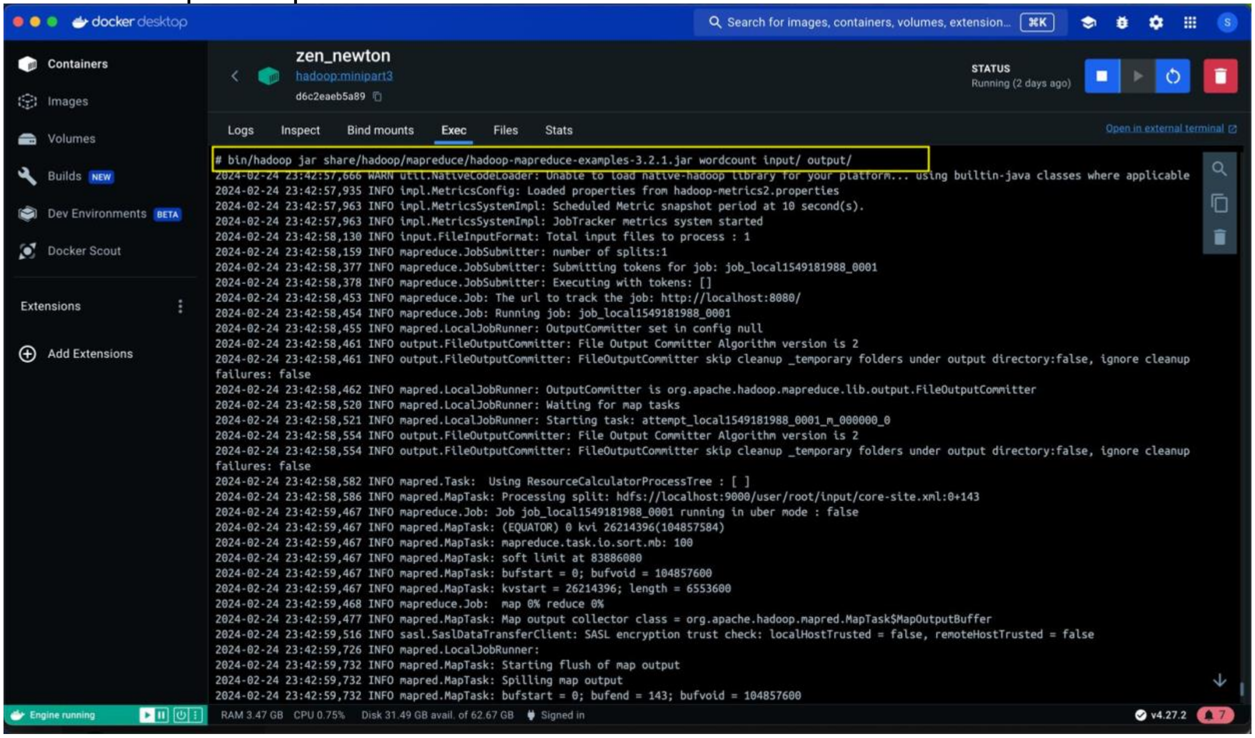
- cd /opt/hadoop/etc/hadoop
- nano hadoop-env.sh
- In the last line of Hadoop-env.sh file add export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

9. Further, we executed ./start-dfs.sh command to start the HDFS daemons i.e. NameNode and DataNode daemons.
10. With JPS command we verified that Hadoop daemons (like NameNode, DataNode, ResourceManager) are up and running. As we know, the NameNode specifically listens on port 9000.



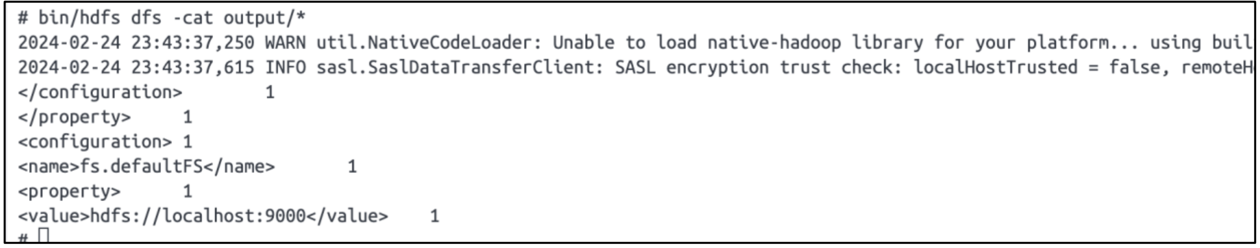
```
# jps
641 DataNode
7573 Jps
473 NameNode
863 SecondaryNameNode
```

Wordcount Example:
bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar wordcount input/ output/



```
# bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar wordcount input/ output/
2024-02-24 23:42:57,666 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 23:42:57,935 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-02-24 23:42:57,963 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-02-24 23:42:57,963 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-02-24 23:42:58,130 INFO input.FileInputFormat: Total input files to process : 1
2024-02-24 23:42:58,159 INFO mapreduce.JobSubmitter: number of splits:1
2024-02-24 23:42:58,377 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1549181988_0001
2024-02-24 23:42:58,378 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-02-24 23:42:58,453 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-02-24 23:42:58,454 INFO mapreduce.Job: Running job: job_local1549181988_0001
2024-02-24 23:42:58,455 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-02-24 23:42:58,461 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-02-24 23:42:58,461 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup
failures: false
2024-02-24 23:42:58,462 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-02-24 23:42:58,520 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-02-24 23:42:58,521 INFO mapred.LocalJobRunner: Starting task: attempt_local1549181988_0001_m_000000_0
2024-02-24 23:42:58,554 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-02-24 23:42:58,554 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup
failures: false
2024-02-24 23:42:58,582 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-02-24 23:42:58,586 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/root/input/core-site.xml:0+143
2024-02-24 23:42:59,467 INFO mapreduce.Job: Job job_local1549181988_0001 running in uber mode : false
2024-02-24 23:42:59,467 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-02-24 23:42:59,467 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-02-24 23:42:59,467 INFO mapred.MapTask: soft limit at 83886080
2024-02-24 23:42:59,467 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-02-24 23:42:59,467 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-02-24 23:42:59,468 INFO mapreduce.Job:  map 0% reduce 0%
2024-02-24 23:42:59,477 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-02-24 23:42:59,516 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-02-24 23:42:59,726 INFO mapred.LocalJobRunner:
2024-02-24 23:42:59,732 INFO mapred.MapTask: Starting flush of map output
2024-02-24 23:42:59,732 INFO mapred.MapTask: Spilling map output
2024-02-24 23:42:59,732 INFO mapred.MapTask: bufstart = 0; bufend = 143; bufvoid = 104857600
```

Output:

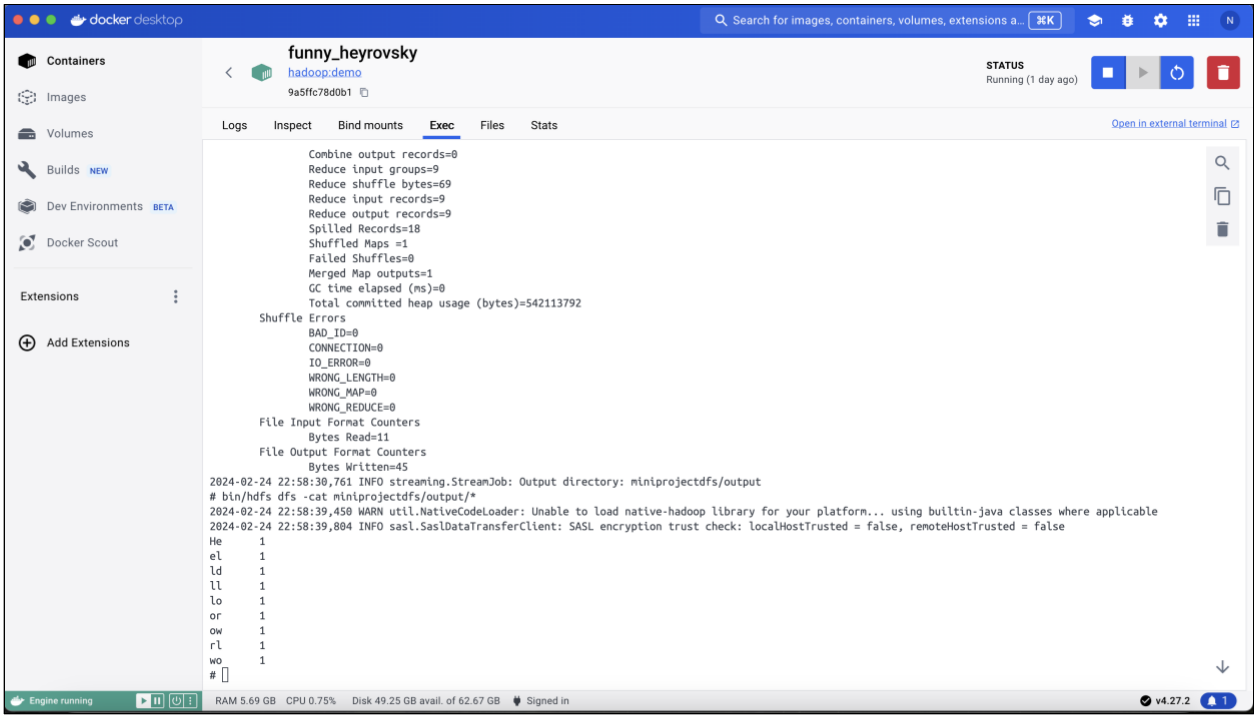


```
# bin/hdfs dfs -cat output/*
2024-02-24 23:43:37,250 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using buil
2024-02-24 23:43:37,615 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteH
</configuration>        1
</property>     1
<configuration> 1
<name>fs.defaultFS</name>       1
<property>      1
<value>hdfs://localhost:9000</value>    1
# 
```

Part 2: Hadoop program that produces the n-gram frequencies of the text “Helloworld”.

We executed the below command for ngram frequency (n) ”2”.

bin/hadoop jar ./share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar -files /miniproject1/ngrammapper.py,/miniproject1/ngramreducer.py -input miniprojectdfs/ -output miniprojectdfs/output -mapper "python3 ngrammapper.py 2" -reducer "python3 ngramreducer.py"



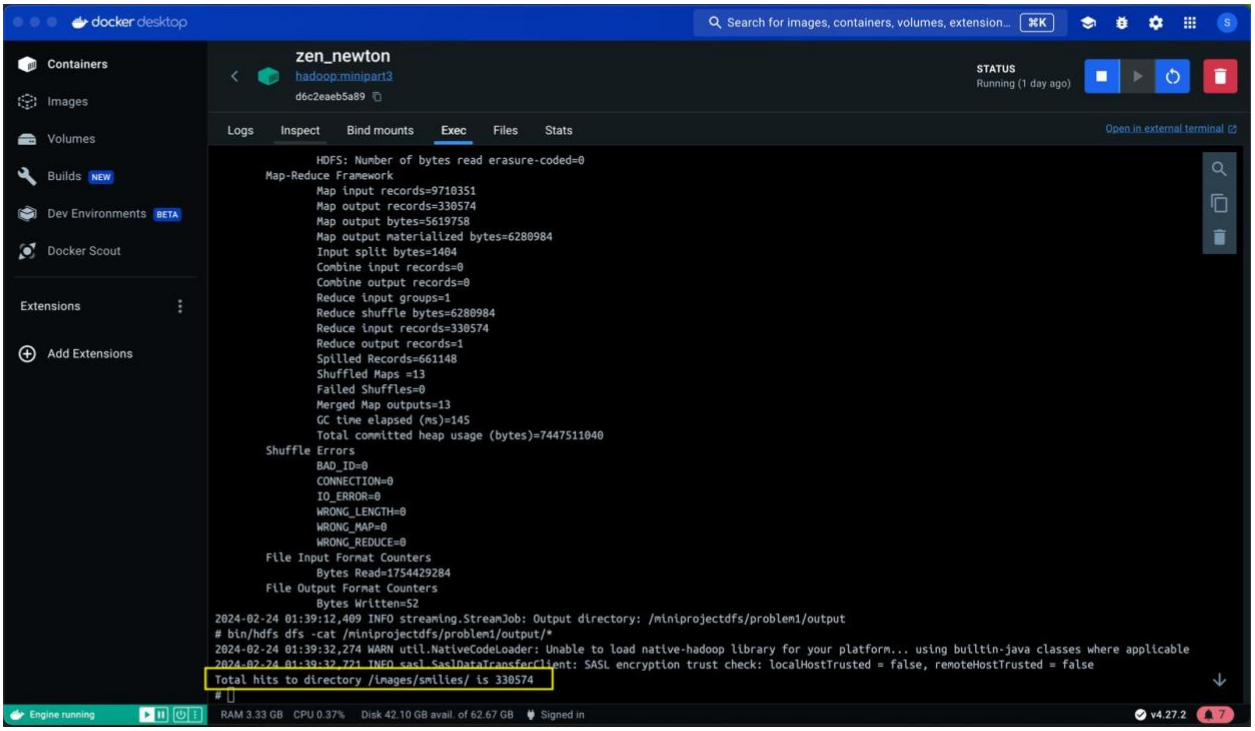
```
                Combine output records=0
                Reduce input groups=9
                Reduce shuffle bytes=69
                Reduce input records=9
                Reduce output records=9
                Spilled Records=18
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=542113792
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=11
        File Output Format Counters
                Bytes Written=45
2024-02-24 22:58:30,761 INFO streaming.StreamJob: Output directory: miniprojectdfs/output
# bin/hdfs dfs -cat miniprojectdfs/output/*
2024-02-24 22:58:39,450 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 22:58:39,804 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
He      1
el      1
ld      1
ll      1
lo      1
or      1
ow      1
rl      1
wo      1
# 
```

Part 3: Hadoop program to analyze real logs.

In this part, we developed MapReduce programs to analyze a real anonymous log file and answer several questions based on the log data. The log file, named access_log, follows the Common Log Format with additional fields that we ignored for the purpose of these problems.

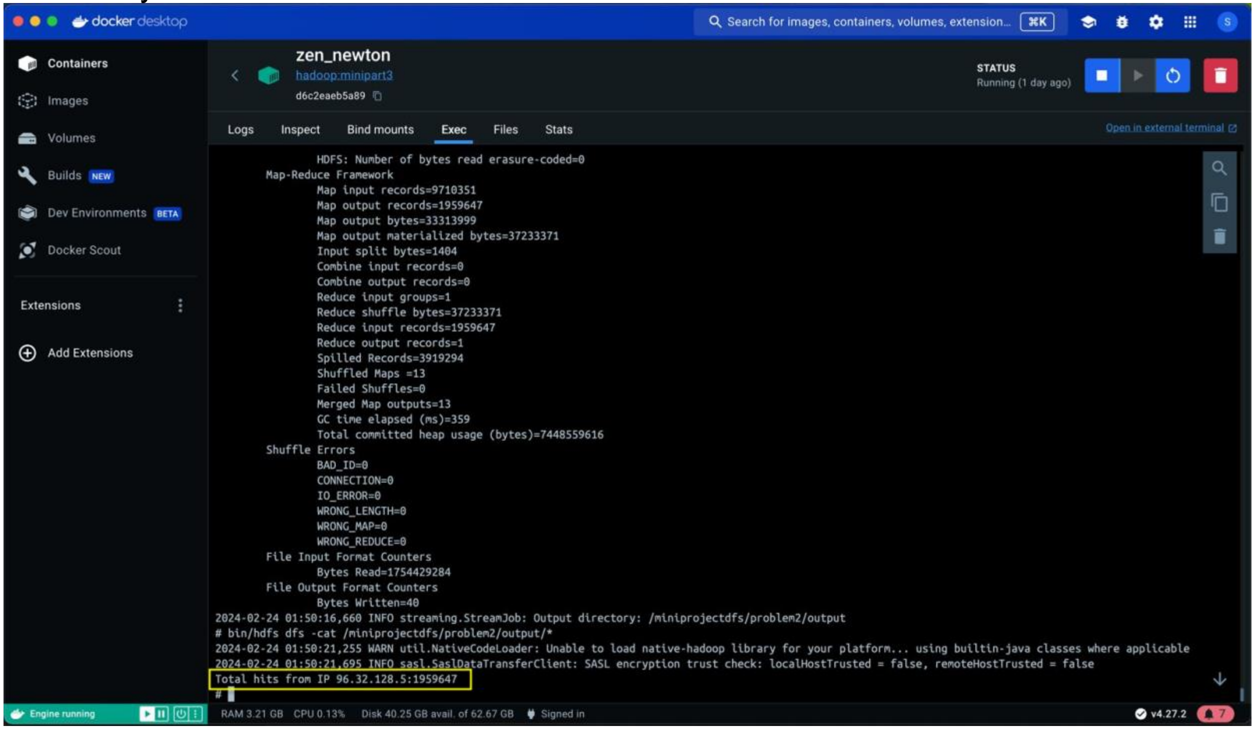
Here are the problems we addressed along with the solutions:

Problem 1 –
How many hits were made to the website directory “/images/smilies/”(including subdirectories and files)?



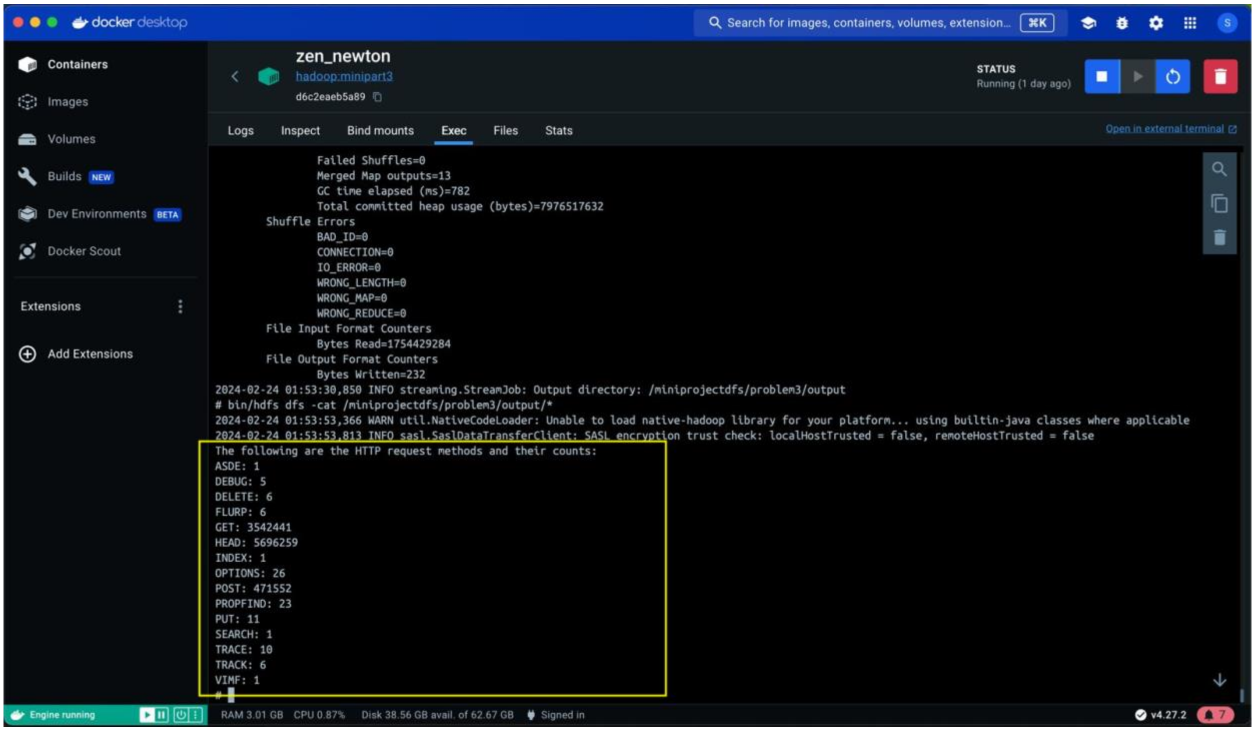
```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=9710351
                Map output records=330574
                Map output bytes=5619758
                Map output materialized bytes=6280984
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=6280984
                Reduce input records=330574
                Reduce output records=1
                Spilled Records=661148
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=145
                Total committed heap usage (bytes)=7447511040
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=52
2024-02-24 01:39:12,409 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem1/output
# bin/hdfs dfs -cat /miniprojectdfs/problem1/output/*
2024-02-24 01:39:32,274 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 01:39:32,721 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total hits to directory /images/smilies/ is 330574
# 
```

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



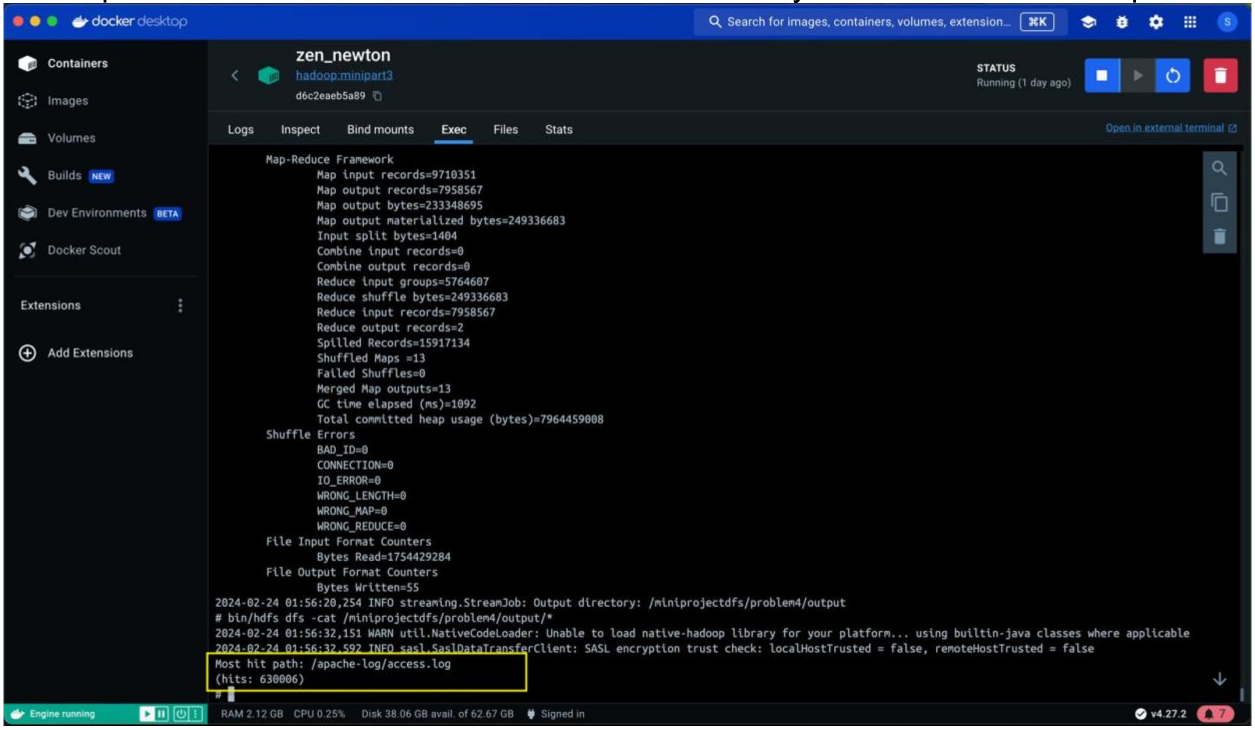
```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=9710351
                Map output records=1959647
                Map output bytes=33313999
                Map output materialized bytes=37233371
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=37233371
                Reduce input records=1959647
                Reduce output records=1
                Spilled Records=3919294
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=359
                Total committed heap usage (bytes)=7448559616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=40
2024-02-24 01:50:16,660 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem2/output
# bin/hdfs dfs -cat /miniprojectdfs/problem2/output/*
2024-02-24 01:50:21,255 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 01:50:21,695 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total hits from IP 96.32.128.5:1959647
# 
```

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



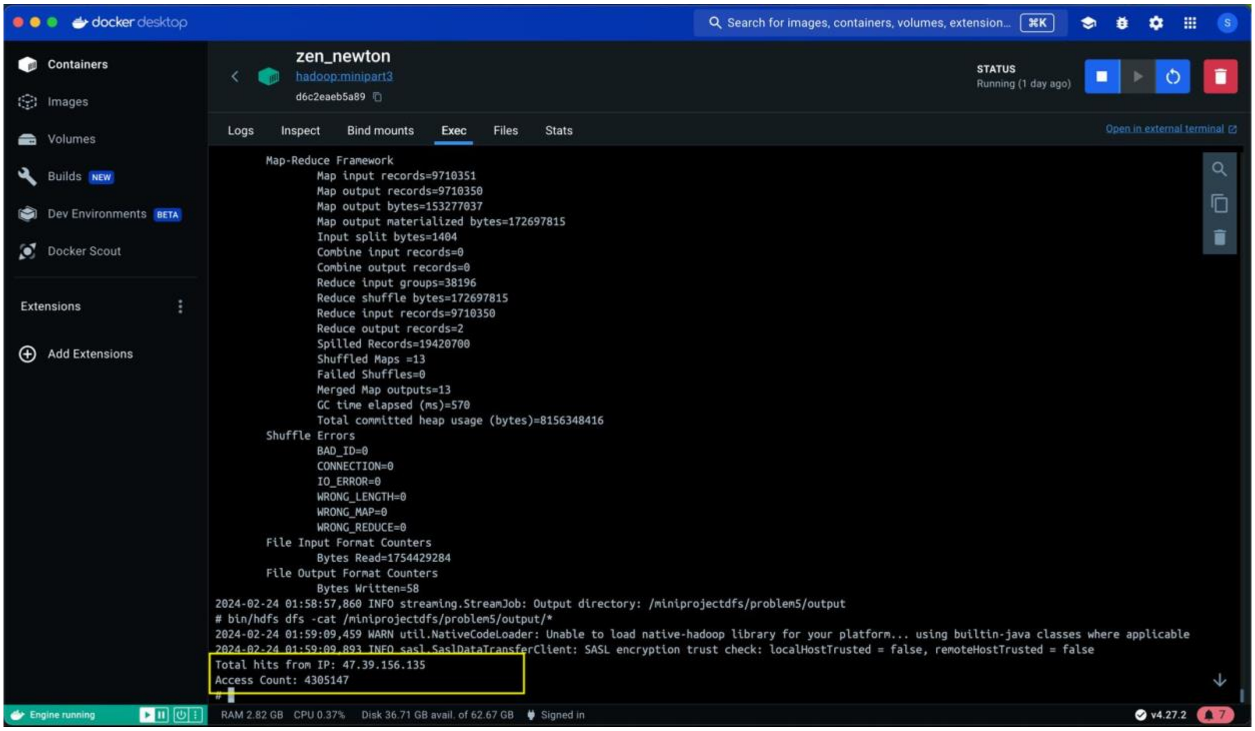
```
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=782
                Total committed heap usage (bytes)=7976517632
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=232
2024-02-24 01:53:30,850 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem3/output
# bin/hdfs dfs -cat /miniprojectdfs/problem3/output/*
2024-02-24 01:53:53,366 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 01:53:53,813 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
The following are the HTTP request methods and their counts:
ASDE: 1
DEBUG: 5
DELETE: 6
FLURP: 6
GET: 3542441
HEAD: 5696259
INDEX: 1
OPTIONS: 26
POST: 471552
PROPFIND: 23
PUT: 11
SEARCH: 1
TRACE: 10
TRACK: 6
VIMF: 1
#
```

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



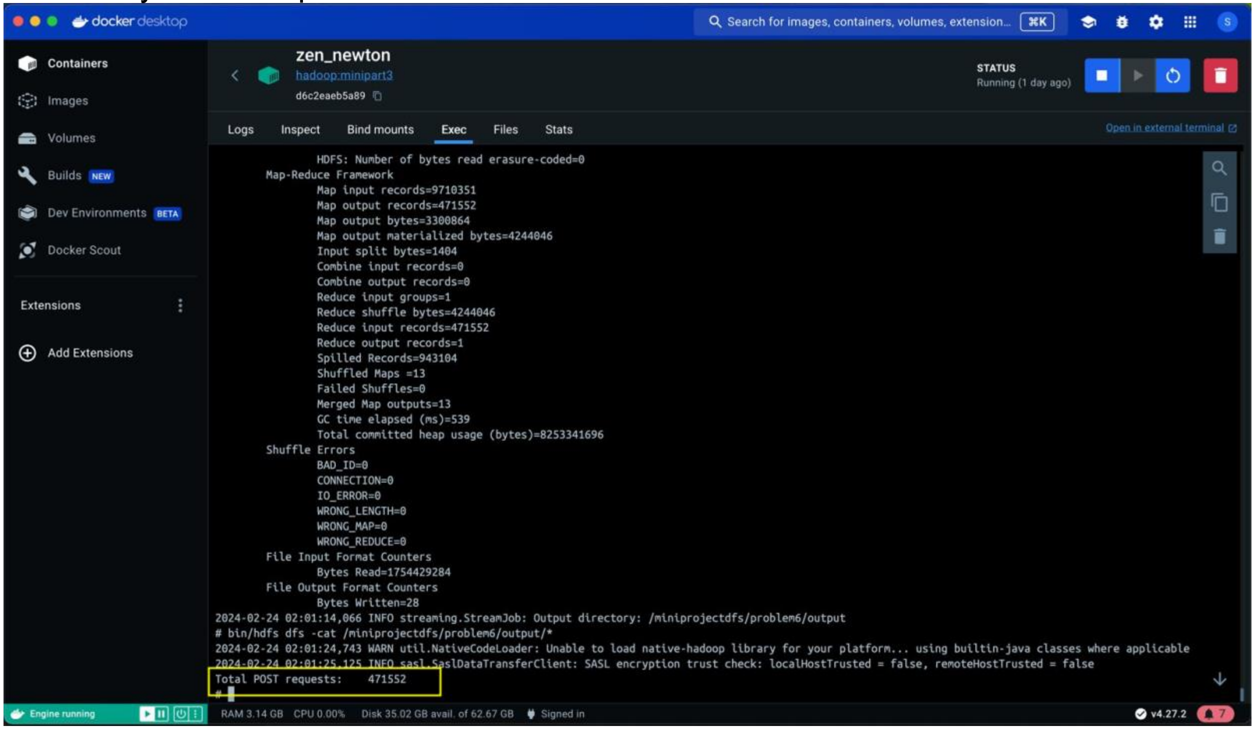
```
        Map-Reduce Framework
                Map input records=9710351
                Map output records=7958567
                Map output bytes=233348695
                Map output materialized bytes=249336683
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=5764607
                Reduce shuffle bytes=249336683
                Reduce input records=7958567
                Reduce output records=2
                Spilled Records=15917134
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=1092
                Total committed heap usage (bytes)=7964459008
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=55
2024-02-24 01:56:20,254 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem4/output
# bin/hdfs dfs -cat /miniprojectdfs/problem4/output/*
2024-02-24 01:56:32,151 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 01:56:32,592 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Most hit path: /apache-log/access.log
(hits: 630006)
#
```

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



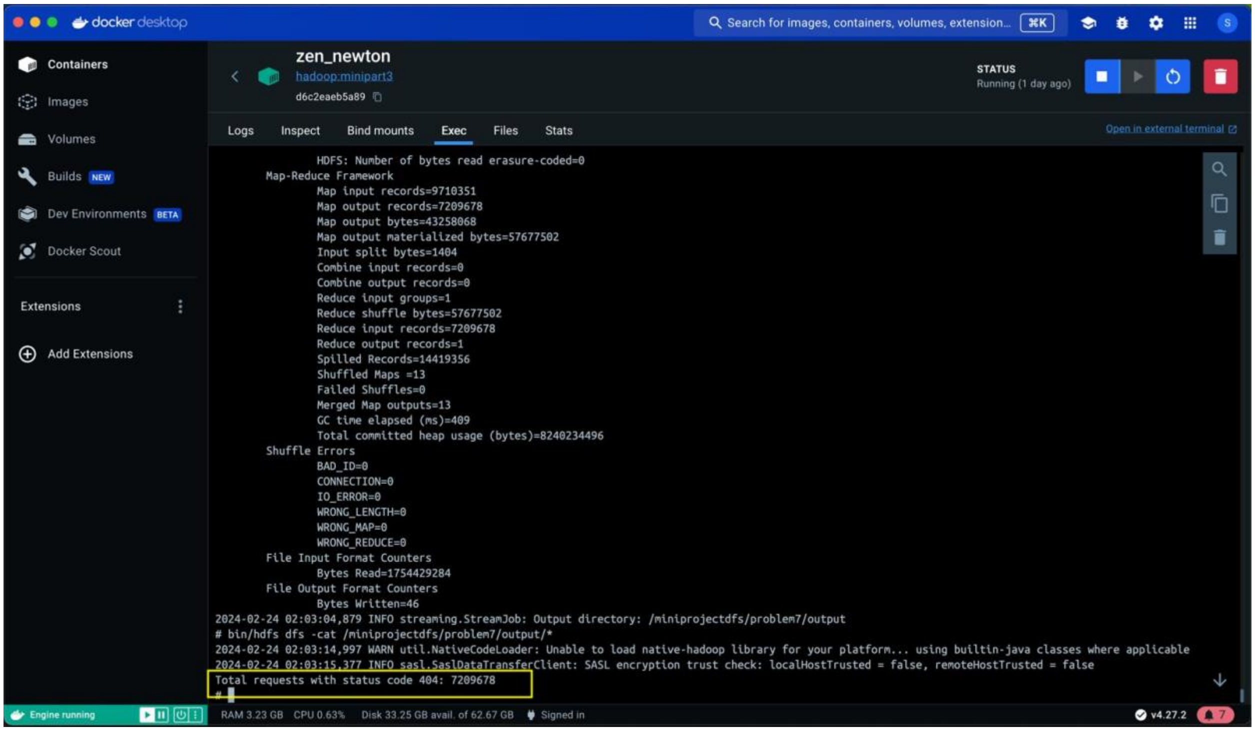
```
        Map-Reduce Framework
                Map input records=9710351
                Map output records=9710350
                Map output bytes=153277037
                Map output materialized bytes=172697815
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=38196
                Reduce shuffle bytes=172697815
                Reduce input records=9710350
                Reduce output records=2
                Spilled Records=19420700
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=570
                Total committed heap usage (bytes)=8156348416
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=58
2024-02-24 01:58:57,860 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem5/output
# bin/hdfs dfs -cat /miniprojectdfs/problem5/output/*
2024-02-24 01:59:09,459 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 01:59:09,893 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total hits from IP: 47.39.156.135
Access Count: 4305147
#
```

Problem 6 –
How many POST request were made?



```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=9710351
                Map output records=471552
                Map output bytes=3300864
                Map output materialized bytes=4244046
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=4244046
                Reduce input records=471552
                Reduce output records=1
                Spilled Records=943104
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=539
                Total committed heap usage (bytes)=8253341696
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=28
2024-02-24 02:01:14,066 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem6/output
# bin/hdfs dfs -cat /miniprojectdfs/problem6/output/*
2024-02-24 02:01:24,743 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 02:01:25,125 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total POST requests:    471552
#
```

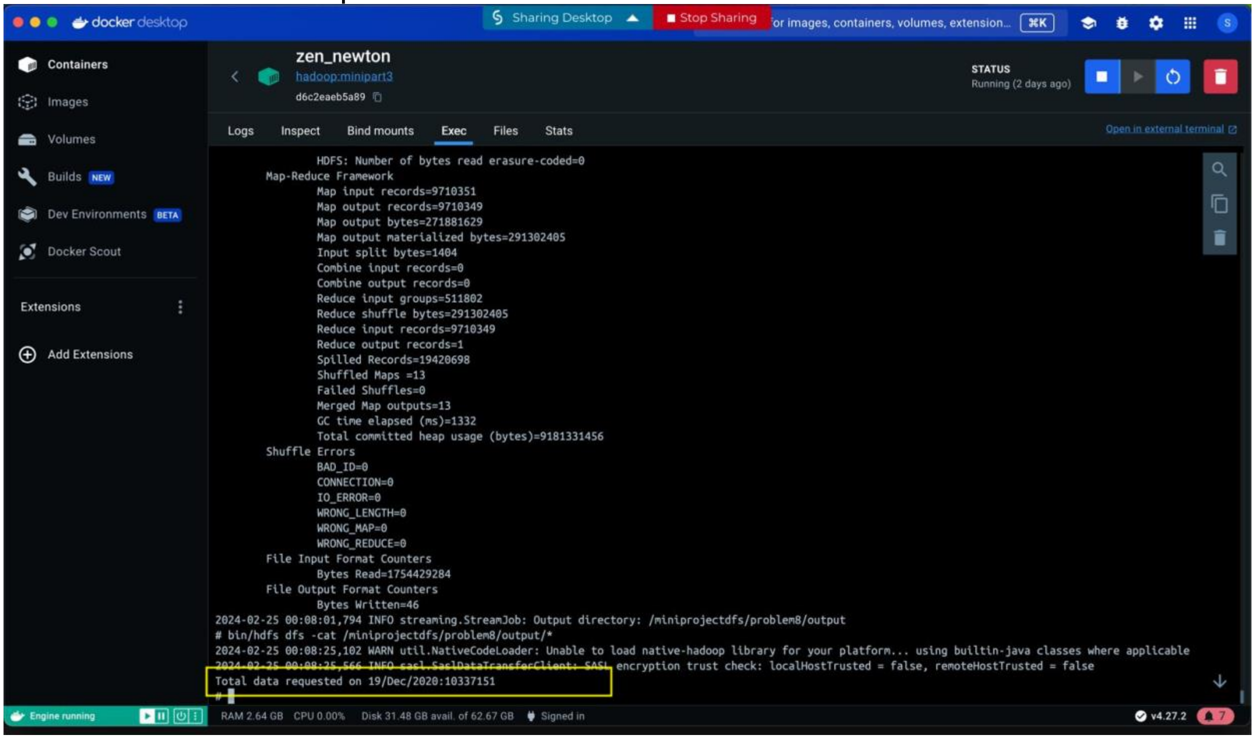
Problem 7 –
How many requests received a 404-status code?



```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=9710351
                Map output records=7209678
                Map output bytes=43258068
                Map output materialized bytes=57677502
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=57677502
                Reduce input records=7209678
                Reduce output records=1
                Spilled Records=14419356
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=409
                Total committed heap usage (bytes)=8240234496
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=46
2024-02-24 02:03:04,879 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem7/output
# bin/hdfs dfs -cat /miniprojectdfs/problem7/output/*
2024-02-24 02:03:14,997 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 02:03:15,377 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total requests with status code 404: 7209678
#
```

Problem 8 –

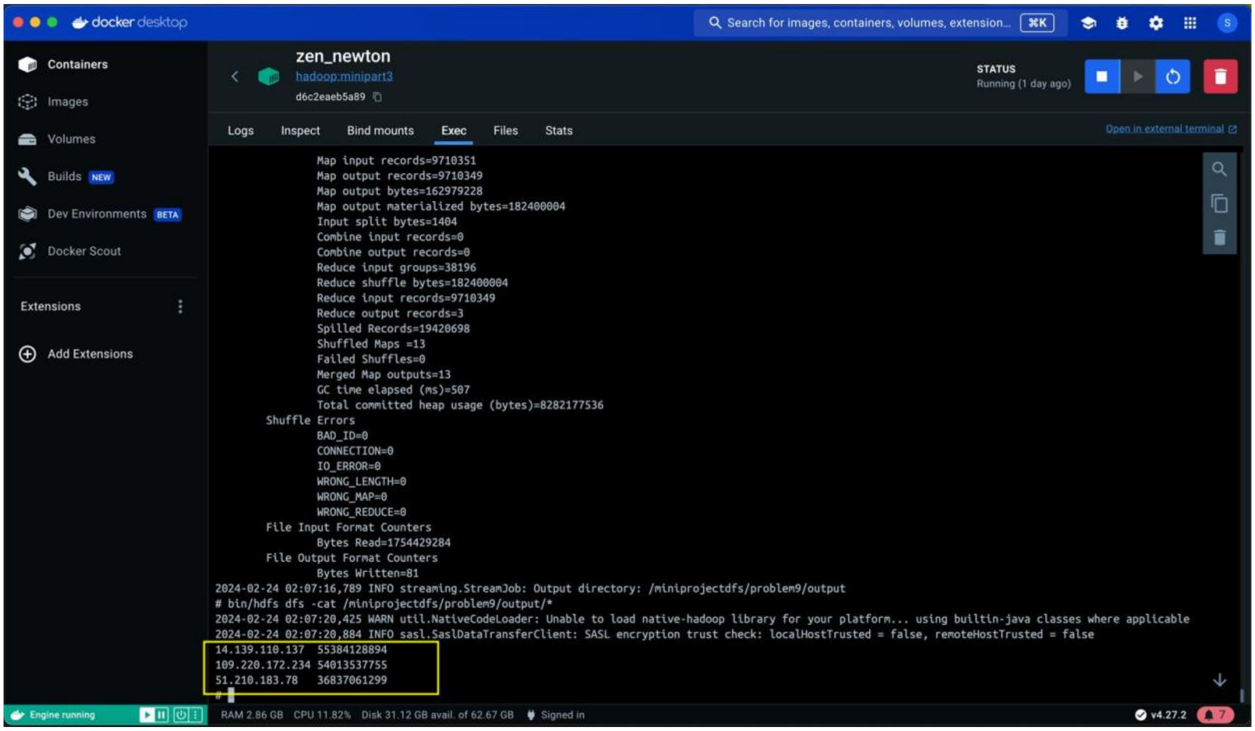
How much data was requested on 19/Dec/2020?



```
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=9710351
                Map output records=9710349
                Map output bytes=271881629
                Map output materialized bytes=291302405
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=511802
                Reduce shuffle bytes=291302405
                Reduce input records=9710349
                Reduce output records=1
                Spilled Records=19420698
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=1332
                Total committed heap usage (bytes)=9181331456
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=46
2024-02-25 00:08:01,794 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem8/output
# bin/hdfs dfs -cat /miniprojectdfs/problem8/output/*
2024-02-25 00:08:25,102 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-25 00:08:25,566 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Total data requested on 19/Dec/2020:10337151
#
```

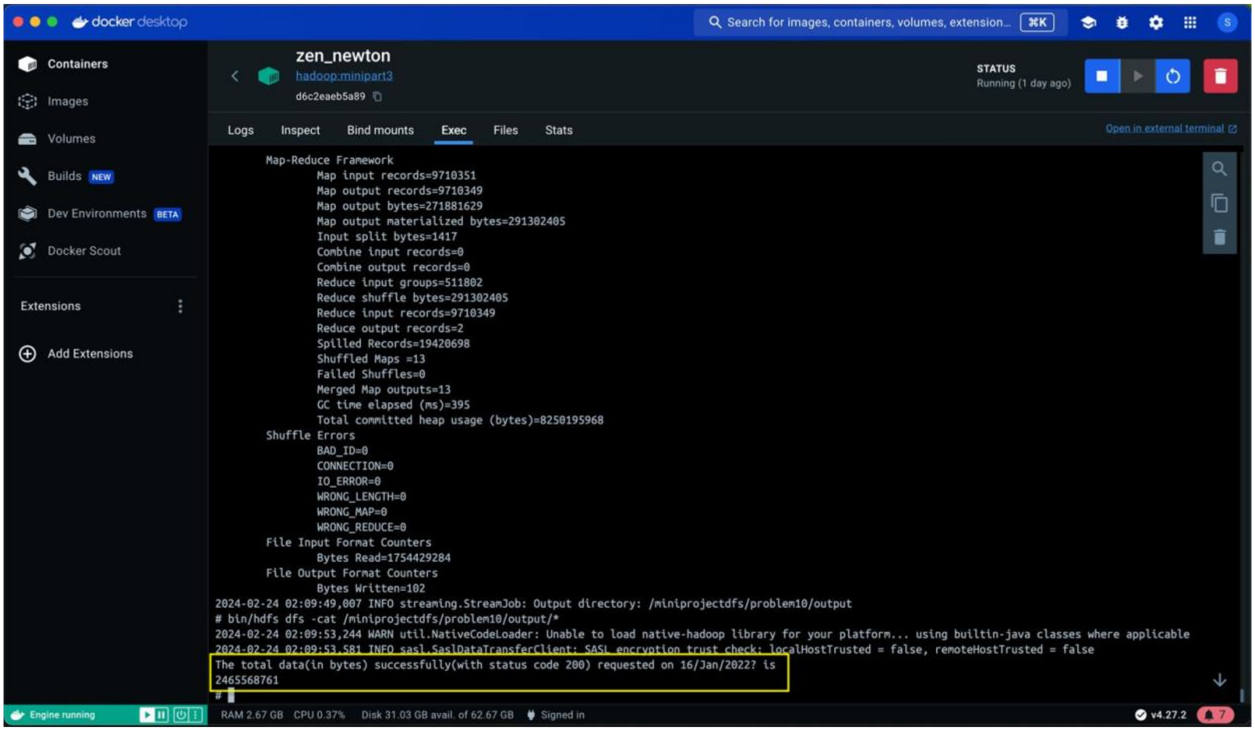
Problem 9 –

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



```
                Map input records=9710351
                Map output records=9710349
                Map output bytes=162979228
                Map output materialized bytes=182400004
                Input split bytes=1404
                Combine input records=0
                Combine output records=0
                Reduce input groups=38196
                Reduce shuffle bytes=182400004
                Reduce input records=9710349
                Reduce output records=3
                Spilled Records=19420698
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=507
                Total committed heap usage (bytes)=8282177536
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=81
2024-02-24 02:07:16,789 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem9/output
# bin/hdfs dfs -cat /miniprojectdfs/problem9/output/*
2024-02-24 02:07:20,425 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 02:07:20,884 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
14.139.110.137  55384128894
109.220.172.234 54013537755
51.210.183.78   36837061299
#
```

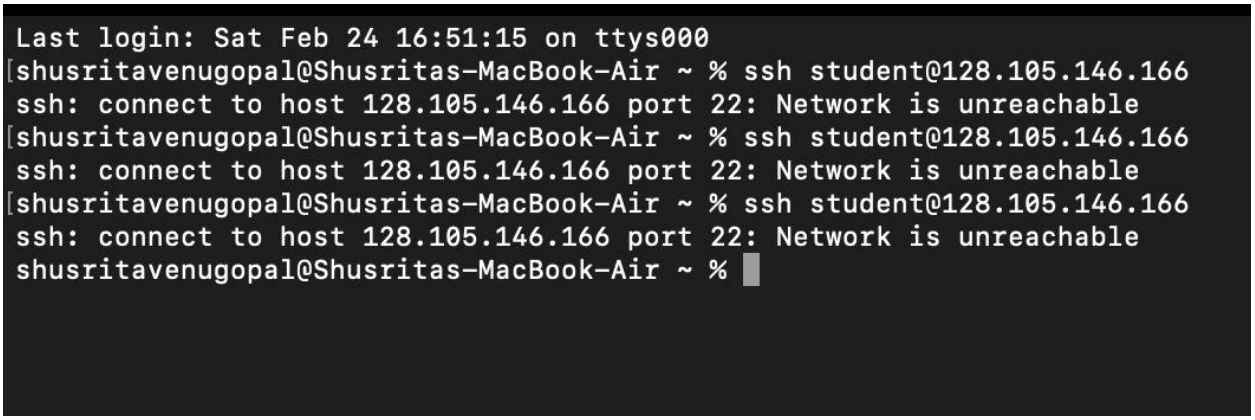
Problem 10 –
How much data (in bytes) was successfully (with status code 200) requested on 16/Jan/2022?



```
        Map-Reduce Framework
                Map input records=9710351
                Map output records=9710349
                Map output bytes=271881629
                Map output materialized bytes=291302405
                Input split bytes=1417
                Combine input records=0
                Combine output records=0
                Reduce input groups=511802
                Reduce shuffle bytes=291302405
                Reduce input records=9710349
                Reduce output records=2
                Spilled Records=19420698
                Shuffled Maps =13
                Failed Shuffles=0
                Merged Map outputs=13
                GC time elapsed (ms)=395
                Total committed heap usage (bytes)=8250195968
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1754429284
        File Output Format Counters
                Bytes Written=102
2024-02-24 02:09:49,007 INFO streaming.StreamJob: Output directory: /miniprojectdfs/problem10/output
# bin/hdfs dfs -cat /miniprojectdfs/problem10/output/*
2024-02-24 02:09:53,244 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-02-24 02:09:53,581 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
The total data(in bytes) successfully(with status code 200) requested on 16/Jan/2022? is
2465568761
#
```

Distributed mode Hadoop (VM)

We tried to step the project on VM mode but unfortunately, we were unable to connect to the IPs provided. Below is the error we ran into:



```
Last login: Sat Feb 24 16:51:15 on ttys000
[shusritavenugopal@Shusritas-MacBook-Air ~ % ssh student@128.105.146.166
ssh: connect to host 128.105.146.166 port 22: Network is unreachable
[shusritavenugopal@Shusritas-MacBook-Air ~ % ssh student@128.105.146.166
ssh: connect to host 128.105.146.166 port 22: Network is unreachable
[shusritavenugopal@Shusritas-MacBook-Air ~ % ssh student@128.105.146.166
ssh: connect to host 128.105.146.166 port 22: Network is unreachable
shusritavenugopal@Shusritas-MacBook-Air ~ % 
```

References

Patel, Mehul. “10 Essential Things to Know About Docker - Mehul Patel - Medium.” *Medium*, 26 July 2023, medium.com/@nomadicmehul/10-essential-things-to-know-about-docker-19b827235bfa.

Mehul - Codedamn. “Docker Tutorial 6: Running Ubuntu in Container.” *YouTube*, 29 Jan. 2020, www.youtube.com/watch?v=pBTD9bowokA.

Mehul - Codedamn. “Docker Tutorial 11: Dockerfile Introduction.” *YouTube*, 8 Mar. 2020, www.youtube.com/watch?v=wGY1WnSPUOk.

Running Hadoop on Ubuntu Linux (Single-Node Cluster). www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster.

How to Fix the Error JAVAHOME Is Not Set and Could Not Be Found After Hadoop Installation | Saturn Cloud Blog. 26 Oct. 2023, https://saturncloud.io/blog/how-to-fix-the-error-javahome-is-not-set-and-could-not-be-found-after-hadoop-installation/