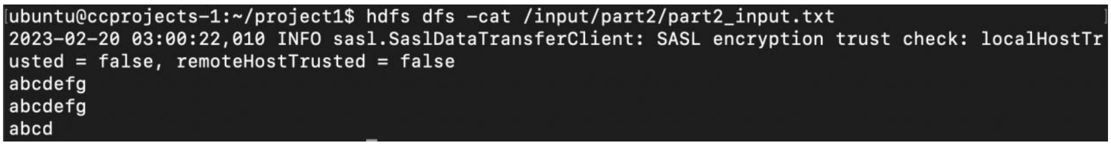
Report for Part2 and Part3

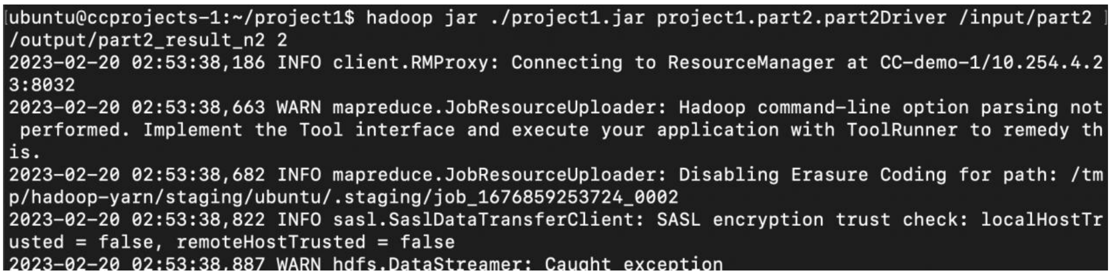
1. Part2

Input file:



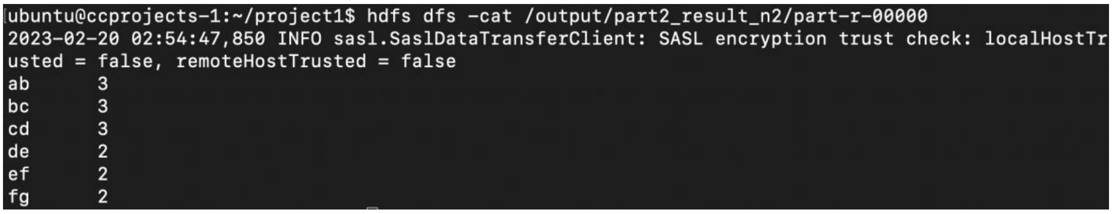
```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /input/part2/part2_input.txt
2023-02-20 03:00:22,010 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTr
usted = false, remoteHostTrusted = false
abcdefg
abcdefg
abcd
```

For n-gram = 2, run the following command:



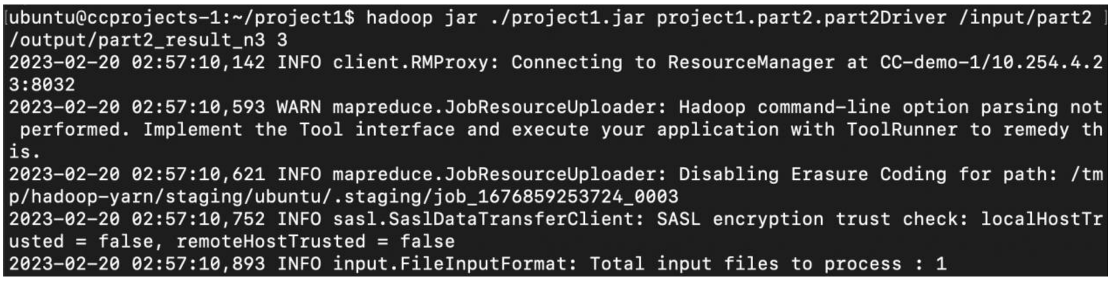
```
ubuntu@ccprojects-1:~/project1$ hadoop jar ./project1.jar project1.part2.part2Driver /input/part2
/output/part2_result_n2 2
2023-02-20 02:53:38,186 INFO client.RMProxy: Connecting to ResourceManager at CC-demo-1/10.254.4.2
3:8032
2023-02-20 02:53:38,663 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not
 performed. Implement the Tool interface and execute your application with ToolRunner to remedy th
is.
2023-02-20 02:53:38,682 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tm
p/hadoop-yarn/staging/ubuntu/.staging/job_1676859253724_0002
2023-02-20 02:53:38,822 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTr
usted = false, remoteHostTrusted = false
2023-02-20 02:53:38.887 WARN hdfs.DataStreamer: Caught exception
```

The result: output file for n-gram=2:



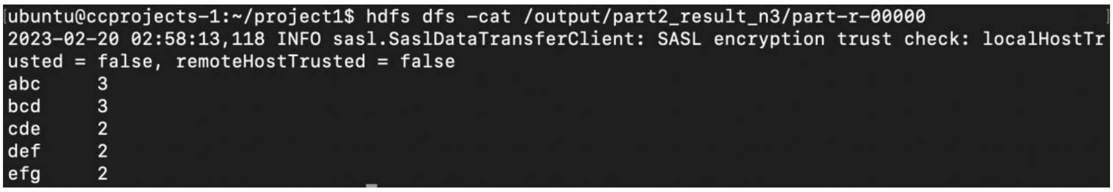
```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /output/part2_result_n2/part-r-00000
2023-02-20 02:54:47,850 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTr
usted = false, remoteHostTrusted = false
ab      3
bc      3
cd      3
de      2
ef      2
fg      2
```

For n-gram = 3, run the following command:



```
ubuntu@ccprojects-1:~/project1$ hadoop jar ./project1.jar project1.part2.part2Driver /input/part2
/output/part2_result_n3 3
2023-02-20 02:57:10,142 INFO client.RMProxy: Connecting to ResourceManager at CC-demo-1/10.254.4.2
3:8032
2023-02-20 02:57:10,593 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not
 performed. Implement the Tool interface and execute your application with ToolRunner to remedy th
is.
2023-02-20 02:57:10,621 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tm
p/hadoop-yarn/staging/ubuntu/.staging/job_1676859253724_0003
2023-02-20 02:57:10,752 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTr
usted = false, remoteHostTrusted = false
2023-02-20 02:57:10,893 INFO input.FileInputFormat: Total input files to process : 1
```

The result: output file for n-gram=3:



```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /output/part2_result_n3/part-r-00000
2023-02-20 02:58:13,118 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTr
usted = false, remoteHostTrusted = false
abc     3
bcd     3
cde     2
def     2
efg     2
```

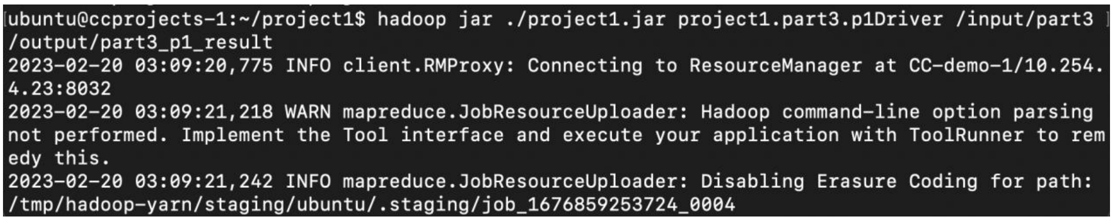
2. Part3

2.1 Problem1

Question:How many hits were made to the website item “/assets/img/home-logo.png”?

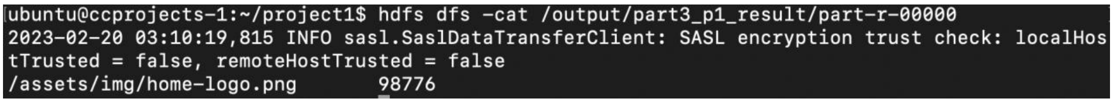
Answer: 98776

Command:



```
ubuntu@ccprojects-1:~/project1$ hadoop jar ./project1.jar project1.part3.p1Driver /input/part3
/output/part3_p1_result
2023-02-20 03:09:20,775 INFO client.RMProxy: Connecting to ResourceManager at CC-demo-1/10.254.
4.23:8032
2023-02-20 03:09:21,218 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing
not performed. Implement the Tool interface and execute your application with ToolRunner to rem
edy this.
2023-02-20 03:09:21,242 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/ubuntu/.staging/job_1676859253724_0004
```

Result:



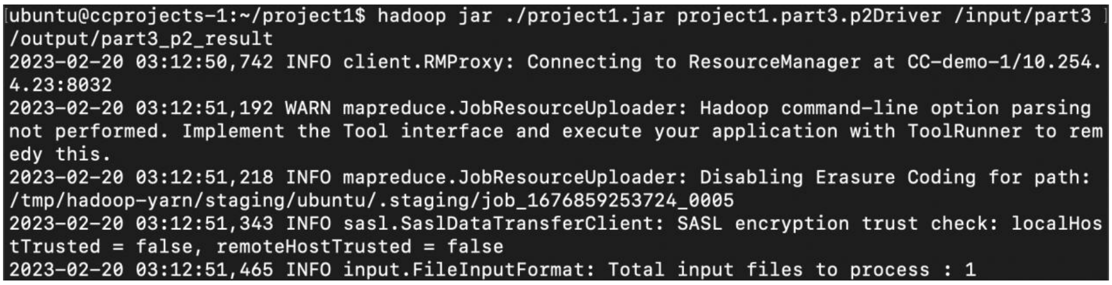
```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /output/part3_p1_result/part-r-00000
2023-02-20 03:10:19,815 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
/assets/img/home-logo.png       98776
```

2.2 Problem2

Question: How many hits were made from the IP: 10.153.239.5 ?

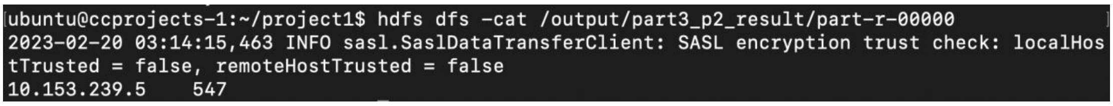
Answer: 547

Command:



```
ubuntu@ccprojects-1:~/project1$ hadoop jar ./project1.jar project1.part3.p2Driver /input/part3
/output/part3_p2_result
2023-02-20 03:12:50,742 INFO client.RMProxy: Connecting to ResourceManager at CC-demo-1/10.254.
4.23:8032
2023-02-20 03:12:51,192 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing
not performed. Implement the Tool interface and execute your application with ToolRunner to rem
edy this.
2023-02-20 03:12:51,218 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/ubuntu/.staging/job_1676859253724_0005
2023-02-20 03:12:51,343 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
2023-02-20 03:12:51,465 INFO input.FileInputFormat: Total input files to process : 1
```

Result:



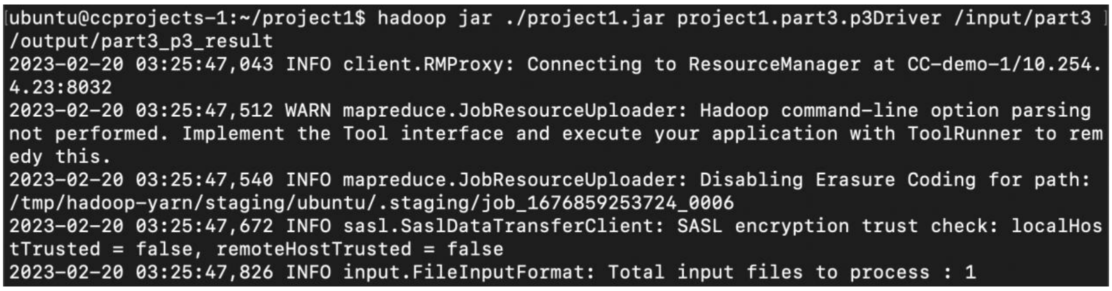
```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /output/part3_p2_result/part-r-00000
2023-02-20 03:14:15,463 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
10.153.239.5    547
```

2.3 Problem3

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

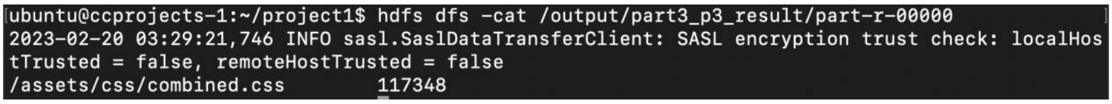
Answer: “/assets/css/combined.css” has been hit most. 117348

Command:



```
ubuntu@ccprojects-1:~/project1$ hadoop jar ./project1.jar project1.part3.p3Driver /input/part3
/output/part3_p3_result
2023-02-20 03:25:47,043 INFO client.RMProxy: Connecting to ResourceManager at CC-demo-1/10.254.
4.23:8032
2023-02-20 03:25:47,512 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing
not performed. Implement the Tool interface and execute your application with ToolRunner to rem
edy this.
2023-02-20 03:25:47,540 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/ubuntu/.staging/job_1676859253724_0006
2023-02-20 03:25:47,672 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
2023-02-20 03:25:47,826 INFO input.FileInputFormat: Total input files to process : 1
```

Result:



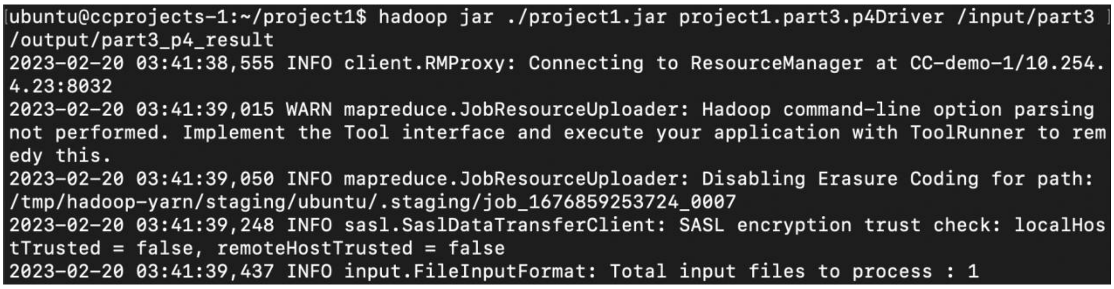
```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /output/part3_p3_result/part-r-00000
2023-02-20 03:29:21,746 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
/assets/css/combined.css        117348
```

2.4 Problem4

Question: Which IP accesses the website most? How many accesses were made by it?

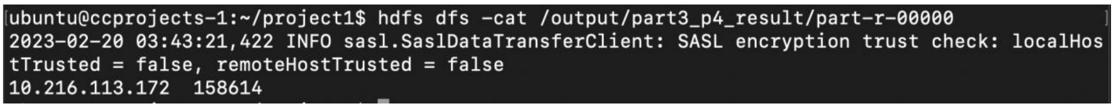
Answer: 10.216.113.172 accesses most. 158614

Command:



```
ubuntu@ccprojects-1:~/project1$ hadoop jar ./project1.jar project1.part3.p4Driver /input/part3
/output/part3_p4_result
2023-02-20 03:41:38,555 INFO client.RMProxy: Connecting to ResourceManager at CC-demo-1/10.254.
4.23:8032
2023-02-20 03:41:39,015 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing
not performed. Implement the Tool interface and execute your application with ToolRunner to rem
edy this.
2023-02-20 03:41:39,050 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path:
/tmp/hadoop-yarn/staging/ubuntu/.staging/job_1676859253724_0007
2023-02-20 03:41:39,248 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
2023-02-20 03:41:39,437 INFO input.FileInputFormat: Total input files to process : 1
```

Result:



```
ubuntu@ccprojects-1:~/project1$ hdfs dfs -cat /output/part3_p4_result/part-r-00000
2023-02-20 03:43:21,422 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHos
tTrusted = false, remoteHostTrusted = false
10.216.113.172  158614
```