University of Pittsburgh
School of Computing and Information

INFSCI 2750: Cloud Computing

Spring 2020

Mini Project 1

Submitted by:

SL. No.	Name	Email address
1	Debdas Ghosh	deg107@pitt.edu
2	Piu Mallick	pim16@pitt.edu
3	Timothy M. Mizerak	tmizerak@pc.pitt.edu

Objective

The objective of this mini project is to get familiar with setting up the Hadoop system and to start programming in Hadoop.

Part 1: Setting up Hadoop

The first part of the project requires setting up Hadoop on the VMs. Each member of our team was assigned a VM, on which we installed and ran Hadoop. The goal of this part of the project is to build a Hadoop cluster. As the first step of our project, we made sure that Hadoop is set up properly on each of the VMs assigned to all the team members by running single node set up. Then, as the second step, we proceeded towards setting a 3-node cluster, with the assistance of the tutorial provided.

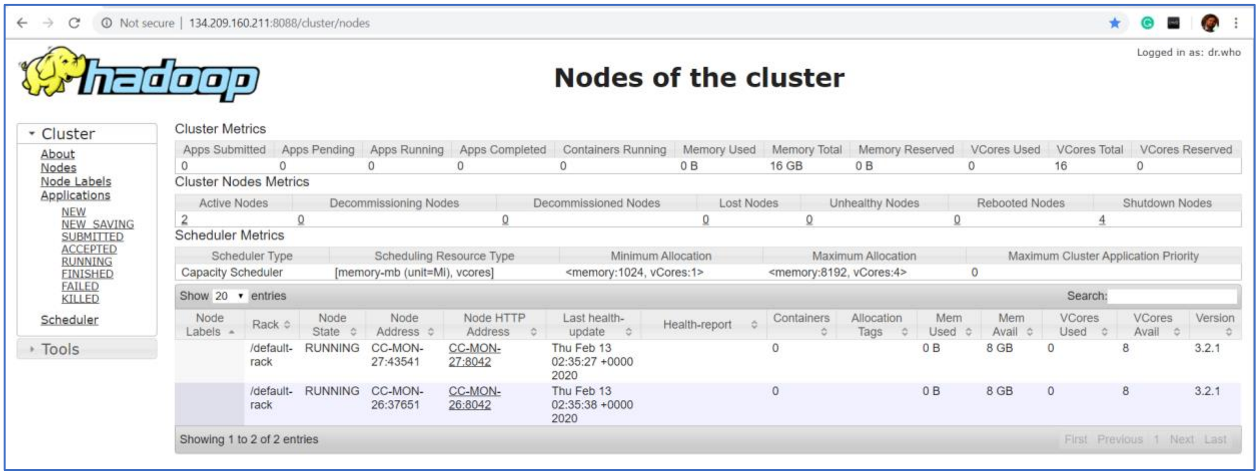
Segregating the 3 VMs as Master & Slaves:

Master → 134.209.160.211 (CC-MON-25)

Slave 1 → 64.225.17.118 (CC-MON-26)

Slave 2 → 138.197.96.66 (CC-MON-27)

Screenshots showing cluster set-up and active status of the nodes



Nodes of the cluster

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	16 GB	0 B	0	16	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	4

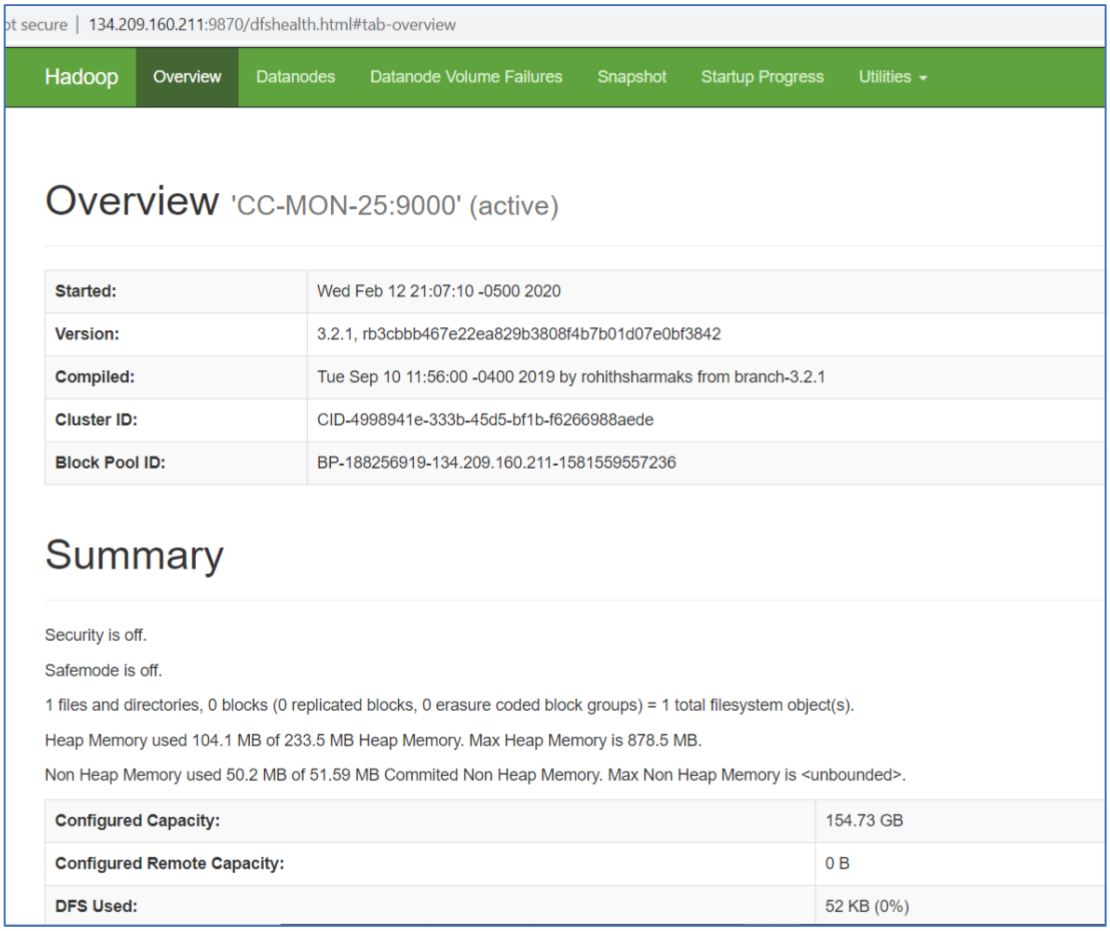
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	VCores Used	VCores Avail	Version
	/default-rack	RUNNING	CC-MON-27:43541	CC-MON-27:8042	Thu Feb 13 02:35:27 +0000 2020		0		0 B	8 GB	0	8	3.2.1
	/default-rack	RUNNING	CC-MON-26:37651	CC-MON-26:8042	Thu Feb 13 02:35:38 +0000 2020		0		0 B	8 GB	0	8	3.2.1

Showing 1 to 2 of 2 entries



Overview 'CC-MON-25:9000' (active)

Started:	Wed Feb 12 21:07:10 -0500 2020
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 11:56:00 -0400 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-4998941e-333b-45d5-bf1b-f6266988aede
Block Pool ID:	BP-188256919-134.209.160.211-1581559557236

Summary

Security is off.

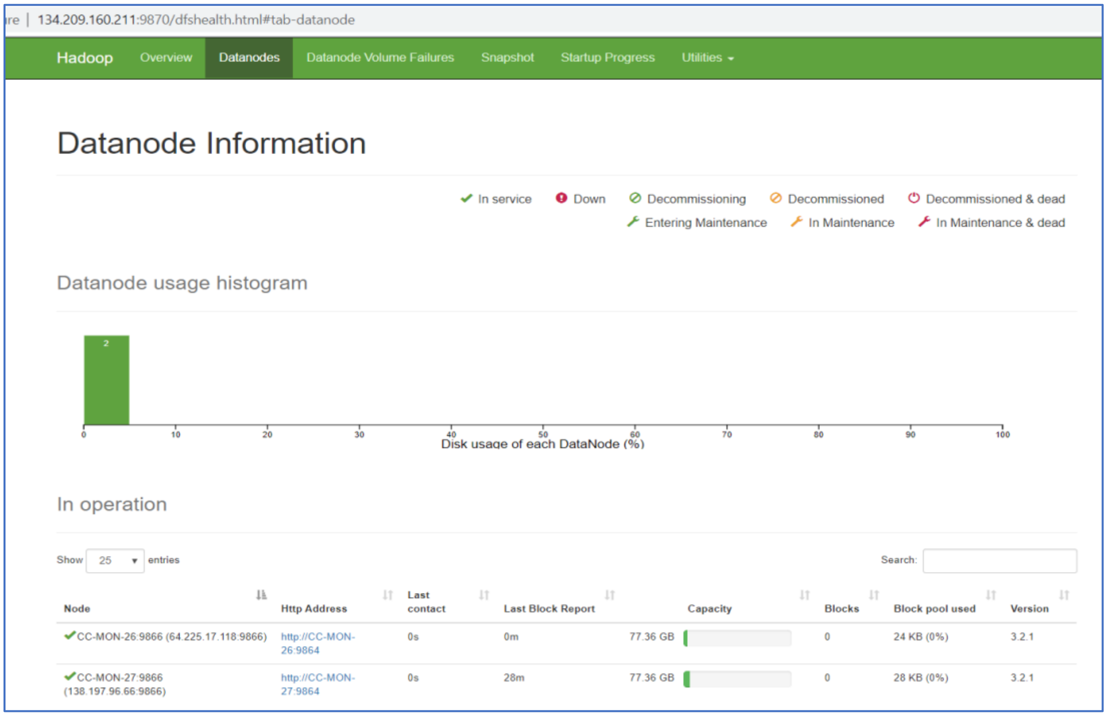
Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 104.1 MB of 233.5 MB Heap Memory. Max Heap Memory is 878.5 MB.

Non Heap Memory used 50.2 MB of 51.59 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	154.73 GB
Configured Remote Capacity:	0 B
DFS Used:	52 KB (0%)



Datanode Information

✓ In service ● Down ⊘ Decommissioning ⊘ Decommissioned ⏻ Decommissioned & dead
Entering Maintenance In Maintenance In Maintenance & dead

Datanode usage histogram

In operation

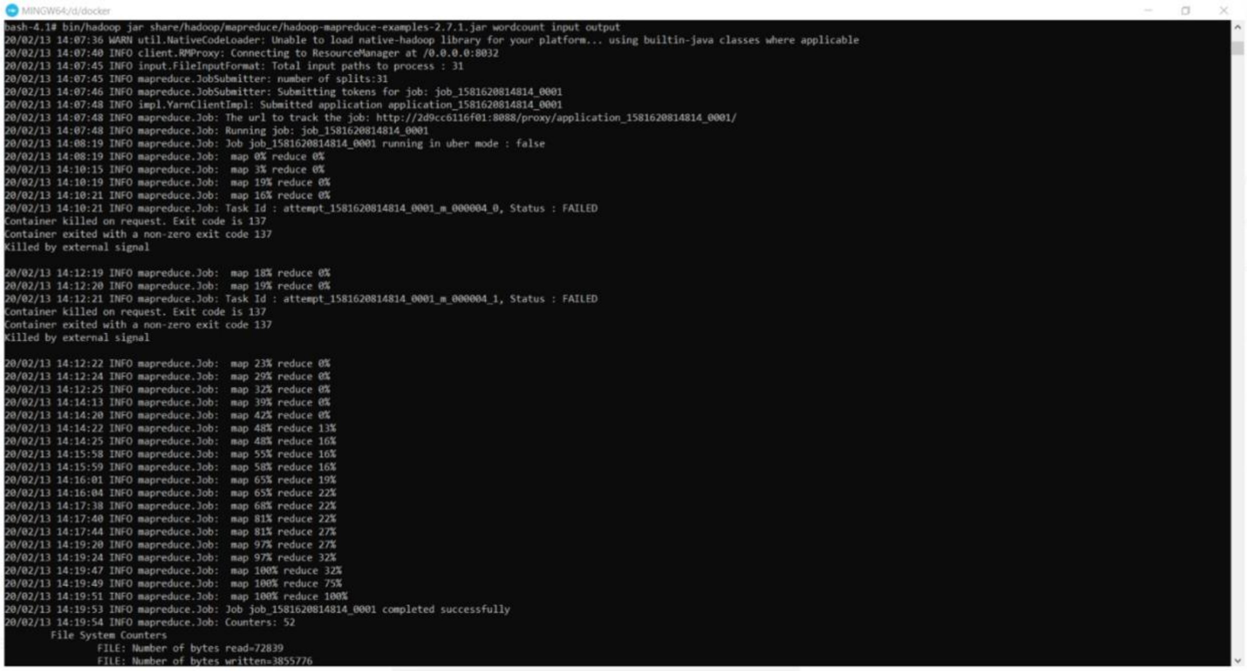
Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Version
✓CC-MON-26:9866 (64.225.17.118:9866)	http://CC-MON-26:9864	0s	0m	77.36 GB	0	24 KB (0%)	3.2.1
✓CC-MON-27:9866 (138.197.96.66:9866)	http://CC-MON-27:9864	0s	28m	77.36 GB	0	28 KB (0%)	3.2.1

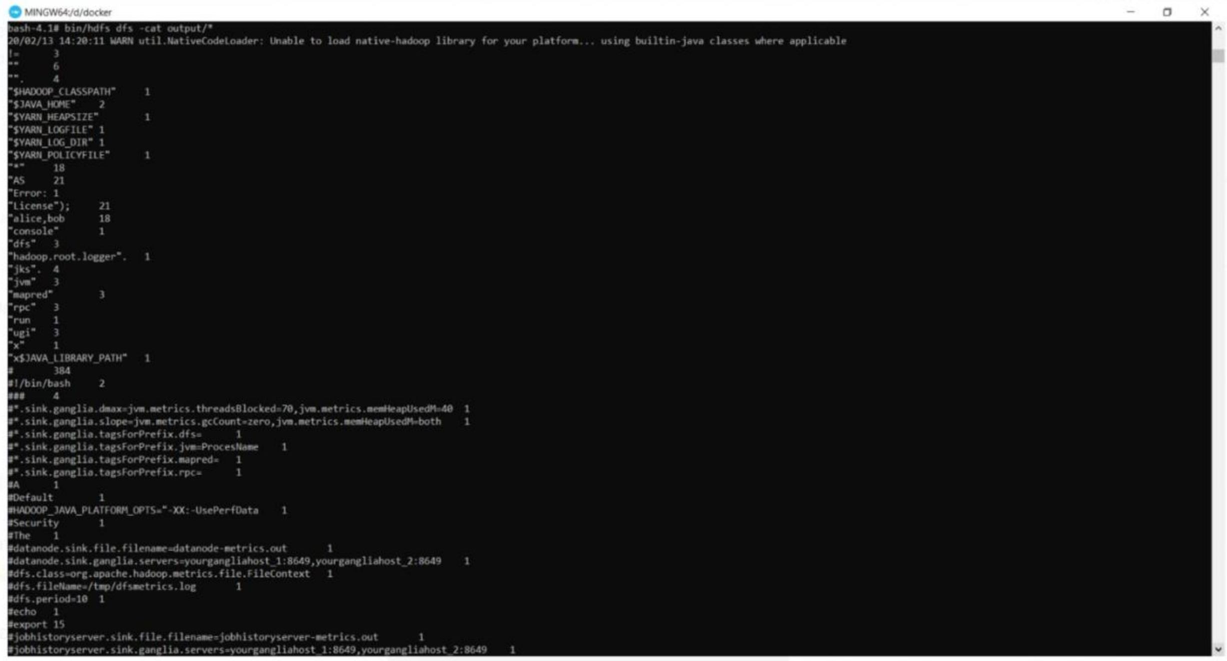
Part 2 → Requires us to build one Docker image to learn how to quick deploy a large-scale Hadoop Cluster in several minutes.

As a first step, we learnt what a '**Docker image**' is and the steps or the procedure to build it.

Running the default wordcount program that comes as part of the Hadoop package. Given an input file, the wordcount program prints the number of occurrences of each word in the file.

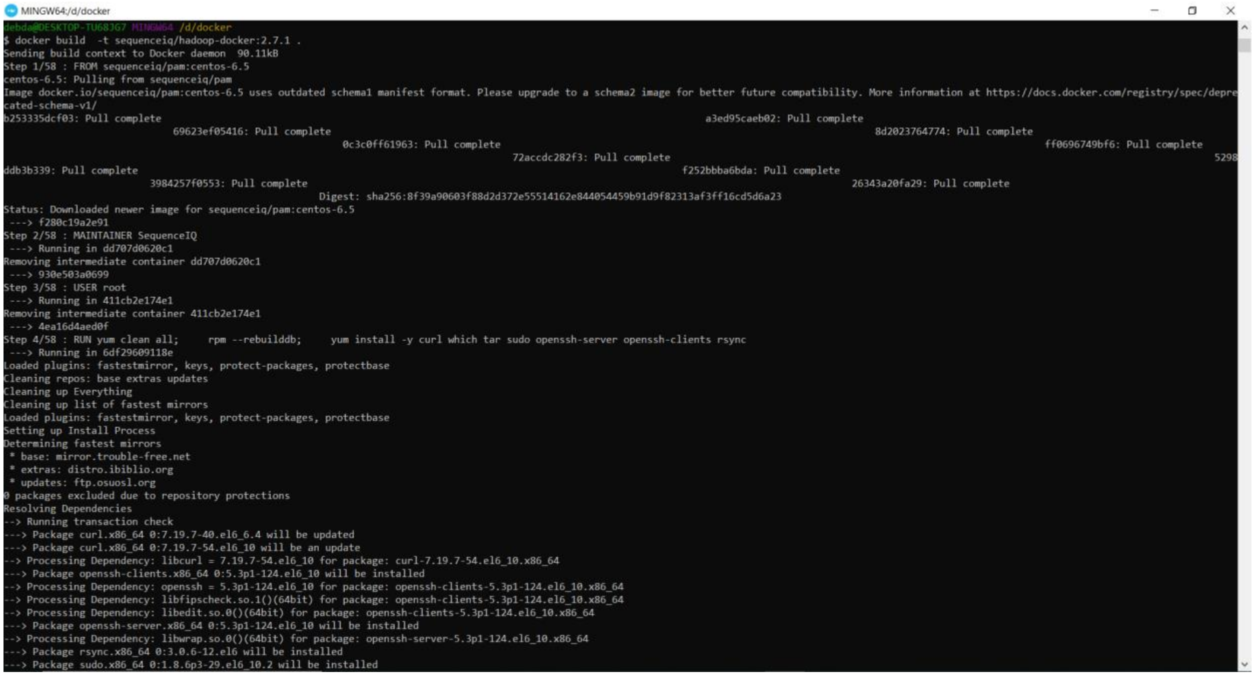


```
bash-4.1# bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar wordcount input output
20/02/13 14:07:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/02/13 14:07:40 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/02/13 14:07:45 INFO input.FileInputFormat: Total input paths to process : 31
20/02/13 14:07:45 INFO mapreduce.JobSubmitter: number of splits:31
20/02/13 14:07:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1581620814814_0001
20/02/13 14:07:48 INFO impl.YarnClientImpl: Submitted application application_1581620814814_0001
20/02/13 14:07:48 INFO mapreduce.Job: The url to track the job: http://2d9cc6116f01:8088/proxy/application_1581620814814_0001/
20/02/13 14:07:48 INFO mapreduce.Job: Running job: job_1581620814814_0001
20/02/13 14:08:19 INFO mapreduce.Job: Job job_1581620814814_0001 running in uber mode : false
20/02/13 14:08:19 INFO mapreduce.Job:  map 0% reduce 0%
20/02/13 14:10:15 INFO mapreduce.Job:  map 3% reduce 0%
20/02/13 14:10:19 INFO mapreduce.Job:  map 19% reduce 0%
20/02/13 14:10:21 INFO mapreduce.Job:  map 16% reduce 0%
20/02/13 14:10:21 INFO mapreduce.Job: Task Id : attempt_1581620814814_0001_m_000004_0, Status : FAILED
Container killed on request. Exit code is 137
Container exited with a non-zero exit code 137
Killed by external signal

20/02/13 14:12:19 INFO mapreduce.Job:  map 18% reduce 0%
20/02/13 14:12:20 INFO mapreduce.Job:  map 19% reduce 0%
20/02/13 14:12:21 INFO mapreduce.Job: Task Id : attempt_1581620814814_0001_m_000004_1, Status : FAILED
Container killed on request. Exit code is 137
Container exited with a non-zero exit code 137
Killed by external signal

20/02/13 14:12:22 INFO mapreduce.Job:  map 23% reduce 0%
20/02/13 14:12:24 INFO mapreduce.Job:  map 29% reduce 0%
20/02/13 14:12:25 INFO mapreduce.Job:  map 32% reduce 0%
20/02/13 14:14:13 INFO mapreduce.Job:  map 39% reduce 0%
20/02/13 14:14:20 INFO mapreduce.Job:  map 42% reduce 0%
20/02/13 14:14:22 INFO mapreduce.Job:  map 48% reduce 13%
20/02/13 14:14:25 INFO mapreduce.Job:  map 48% reduce 16%
20/02/13 14:15:58 INFO mapreduce.Job:  map 55% reduce 16%
20/02/13 14:15:59 INFO mapreduce.Job:  map 58% reduce 16%
20/02/13 14:16:01 INFO mapreduce.Job:  map 65% reduce 19%
20/02/13 14:16:04 INFO mapreduce.Job:  map 65% reduce 22%
20/02/13 14:17:38 INFO mapreduce.Job:  map 68% reduce 22%
20/02/13 14:17:40 INFO mapreduce.Job:  map 81% reduce 22%
20/02/13 14:17:44 INFO mapreduce.Job:  map 81% reduce 27%
20/02/13 14:19:20 INFO mapreduce.Job:  map 97% reduce 27%
20/02/13 14:19:24 INFO mapreduce.Job:  map 97% reduce 32%
20/02/13 14:19:47 INFO mapreduce.Job:  map 100% reduce 32%
20/02/13 14:19:49 INFO mapreduce.Job:  map 100% reduce 75%
20/02/13 14:19:51 INFO mapreduce.Job:  map 100% reduce 100%
20/02/13 14:19:53 INFO mapreduce.Job: Job job_1581620814814_0001 completed successfully
20/02/13 14:19:54 INFO mapreduce.Job: Counters: 52
        File System Counters
                FILE: Number of bytes read=72839
                FILE: Number of bytes written=3855776
```

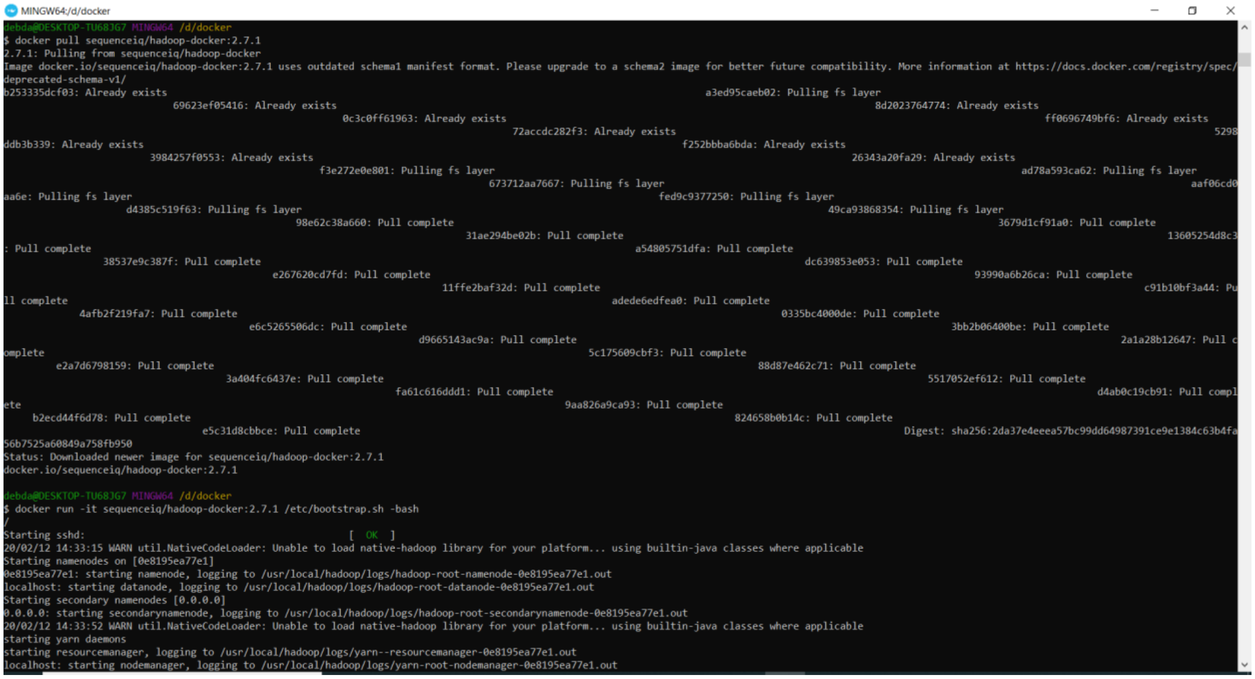


```
bash-4.1# bin/hdfs dfs -cat output/*
20/02/13 14:20:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
!=      3
""      6
"".     4
"$HADOOP_CLASSPATH"     1
"$JAVA_HOME"    2
"$YARN_HEAPSIZE"        1
"$YARN_LOGFILE" 1
"$YARN_LOG_DIR" 1
"$YARN_POLICYFILE"      1
"*"     18
"AS     21
"Error: 1
"License");     21
"alice,bob      18
"console"       1
"dfs"   3
"hadoop.root.logger".   1
"jks".  4
"jvm"   3
"mapred"        3
"rpc"   3
"run    1
"ugi"   3
"x"     1
"x$JAVA_LIBRARY_PATH"   1
#       384
#!/bin/bash     2
###     4
#*.sink.ganglia.dmax=jvm.metrics.threadsBlocked=70,jvm.metrics.memHeapUsedM=40  1
#*.sink.ganglia.slope=jvm.metrics.gcCount=zero,jvm.metrics.memHeapUsedM=both    1
#*.sink.ganglia.tagsForPrefix.dfs=      1
#*.sink.ganglia.tagsForPrefix.jvm=ProcesName    1
#*.sink.ganglia.tagsForPrefix.mapred=   1
#*.sink.ganglia.tagsForPrefix.rpc=      1
#A      1
#Default        1
#HADOOP_JAVA_PLATFORM_OPTS="-XX:-UsePerfData    1
#Security       1
#The    1
#datanode.sink.file.filename=datanode-metrics.out       1
#datanode.sink.ganglia.servers=yourgangliahost_1:8649,yourgangliahost_2:8649    1
#dfs.class=org.apache.hadoop.metrics.file.FileContext   1
#dfs.fileName=/tmp/dfsmetrics.log       1
#dfs.period=10  1
#echo   1
#export 15
#jobhistoryserver.sink.file.filename=jobhistoryserver-metrics.out       1
#jobhistoryserver.sink.ganglia.servers=yourgangliahost_1:8649,yourgangliahost_2:8649    1
```

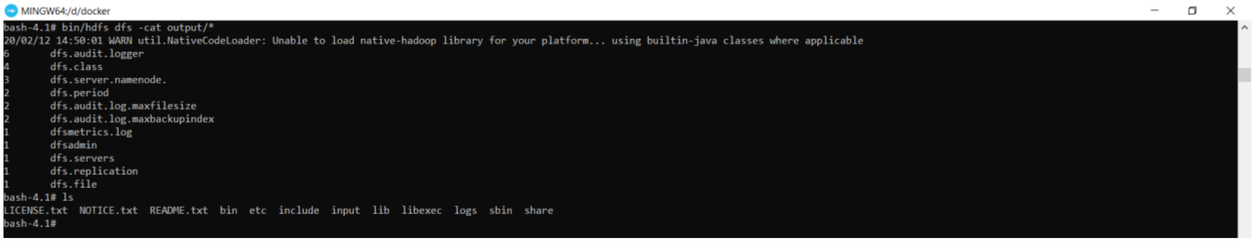
Building Docker Image



```
debda@DESKTOP-TU68JG7 MINGW64 /d/docker
$ docker build  -t sequenceiq/hadoop-docker:2.7.1 .
Sending build context to Docker daemon  90.11kB
Step 1/58 : FROM sequenceiq/pam:centos-6.5
centos-6.5: Pulling from sequenceiq/pam
Image docker.io/sequenceiq/pam:centos-6.5 uses outdated schema1 manifest format. Please upgrade to a schema2 image for better future compatibility. More information at https://docs.docker.com/registry/spec/depre
cated-schema-v1/
b253335dcf03: Pull complete                                                                                             a3ed95caeb02: Pull complete
                             69623ef05416: Pull complete                                                                                             8d2023764774: Pull complete
                                                          0c3c0ff61963: Pull complete                                                                                             ff0696749bf6: Pull complete
                                                                                       72accdc282f3: Pull complete                                                                                             5298
ddb3b339: Pull complete                                                                                             f252bbba6bda: Pull complete
                         3984257f0553: Pull complete                                                                                             26343a20fa29: Pull complete
                                                      Digest: sha256:8f39a90603f88d2d372e55514162e844054459b91d9f82313af3ff16cd5d6a23
Status: Downloaded newer image for sequenceiq/pam:centos-6.5
 ---> f280c19a2e91
Step 2/58 : MAINTAINER SequenceIQ
 ---> Running in dd707d0620c1
Removing intermediate container dd707d0620c1
 ---> 930e503a0699
Step 3/58 : USER root
 ---> Running in 411cb2e174e1
Removing intermediate container 411cb2e174e1
 ---> 4ea16d4aed0f
Step 4/58 : RUN yum clean all;     rpm --rebuilddb;     yum install -y curl which tar sudo openssh-server openssh-clients rsync
 ---> Running in 6df29609118e
Loaded plugins: fastestmirror, keys, protect-packages, protectbase
Cleaning repos: base extras updates
Cleaning up Everything
Cleaning up list of fastest mirrors
Loaded plugins: fastestmirror, keys, protect-packages, protectbase
Setting up Install Process
Determining fastest mirrors
 * base: mirror.trouble-free.net
 * extras: distro.ibiblio.org
 * updates: ftp.osuosl.org
0 packages excluded due to repository protections
Resolving Dependencies
--> Running transaction check
---> Package curl.x86_64 0:7.19.7-40.el6_6.4 will be updated
---> Package curl.x86_64 0:7.19.7-54.el6_10 will be an update
--> Processing Dependency: libcurl = 7.19.7-54.el6_10 for package: curl-7.19.7-54.el6_10.x86_64
---> Package openssh-clients.x86_64 0:5.3p1-124.el6_10 will be installed
--> Processing Dependency: openssh = 5.3p1-124.el6_10 for package: openssh-clients-5.3p1-124.el6_10.x86_64
--> Processing Dependency: libfipscheck.so.1()(64bit) for package: openssh-clients-5.3p1-124.el6_10.x86_64
--> Processing Dependency: libedit.so.0()(64bit) for package: openssh-clients-5.3p1-124.el6_10.x86_64
---> Package openssh-server.x86_64 0:5.3p1-124.el6_10 will be installed
--> Processing Dependency: libwrap.so.0()(64bit) for package: openssh-server-5.3p1-124.el6_10.x86_64
---> Package rsync.x86_64 0:3.0.6-12.el6 will be installed
---> Package sudo.x86_64 0:1.8.6p3-29.el6_10.2 will be installed
```



```
debda@DESKTOP-TU68JG7 MINGW64 /d/docker
$ docker pull sequenceiq/hadoop-docker:2.7.1
2.7.1: Pulling from sequenceiq/hadoop-docker
Image docker.io/sequenceiq/hadoop-docker:2.7.1 uses outdated schema1 manifest format. Please upgrade to a schema2 image for better future compatibility. More information at https://docs.docker.com/registry/spec/
deprecated-schema-v1/
b253335dcf03: Already exists                                                                                            a3ed95caeb02: Pulling fs layer
                             69623ef05416: Already exists                                                                                            8d2023764774: Already exists
                                                          0c3c0ff61963: Already exists                                                                                            ff0696749bf6: Already exists
                                                                                       72accdc282f3: Already exists                                                                                            5298
ddb3b339: Already exists                                                                                            f252bbba6bda: Already exists
                         3984257f0553: Already exists                                                                                            26343a20fa29: Already exists
                                                      f3e272e0e801: Pulling fs layer                                                                                          ad78a593ca62: Pulling fs layer
                                                                                   673712aa7667: Pulling fs layer                                                                                          aaf06cd0
aa6e: Pulling fs layer                                                                                          fed9c9377250: Pulling fs layer
                     d4385c519f63: Pulling fs layer                                                                                          49ca93868354: Pulling fs layer
                                                  98e62c38a660: Pull complete                                                                                             3679d1cf91a0: Pull complete
                                                                               31ae294be02b: Pull complete                                                                                             13605254d8c3
: Pull complete                                                                                             a54805751dfa: Pull complete
                 38537e9c387f: Pull complete                                                                                             dc639853e053: Pull complete
                                              e267620cd7fd: Pull complete                                                                                             93990a6b26ca: Pull complete
                                                                           11ffe2baf32d: Pull complete                                                                                             c91b10bf3a44: Pu
ll complete                                                                                             adede6edfea0: Pull complete
             4afb2f219fa7: Pull complete                                                                                             0335bc4000de: Pull complete
                                          e6c5265506dc: Pull complete                                                                                             3bb2b06400be: Pull complete
                                                                       d9665143ac9a: Pull complete                                                                                             2a1a28b12647: Pull c
omplete                                                                                             5c175609cbf3: Pull complete
         e2a7d6798159: Pull complete                                                                                             88d87e462c71: Pull complete
                                      3a404fc6437e: Pull complete                                                                                             5517052ef612: Pull complete
                                                                   fa61c616ddd1: Pull complete                                                                                             d4ab0c19cb91: Pull compl
ete                                                                                             9aa826a9ca93: Pull complete
     b2ecd44f6d78: Pull complete                                                                                             824658b0b14c: Pull complete
                                  e5c31d8cbbce: Pull complete                                                                                             Digest: sha256:2da37e4eeea57bc99dd64987391ce9e1384c63b4fa
56b7525a60849a758fb950
Status: Downloaded newer image for sequenceiq/hadoop-docker:2.7.1
docker.io/sequenceiq/hadoop-docker:2.7.1

debda@DESKTOP-TU68JG7 MINGW64 /d/docker
$ docker run -it sequenceiq/hadoop-docker:2.7.1 /etc/bootstrap.sh -bash
/
Starting sshd:                                             [  OK  ]
20/02/12 14:33:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [0e8195ea77e1]
0e8195ea77e1: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-0e8195ea77e1.out
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-0e8195ea77e1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-0e8195ea77e1.out
20/02/12 14:33:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-0e8195ea77e1.out
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-0e8195ea77e1.out
```



```
bash-4.1# bin/hdfs dfs -cat output/*
20/02/12 14:50:01 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
6       dfs.audit.logger
4       dfs.class
3       dfs.server.namenode.
2       dfs.period
2       dfs.audit.log.maxfilesize
2       dfs.audit.log.maxbackupindex
1       dfsmetrics.log
1       dfsadmin
1       dfs.servers
1       dfs.replication
1       dfs.file
bash-4.1# ls
LICENSE.txt  NOTICE.txt  README.txt  bin  etc  include  input  lib  libexec  logs  sbin  share
bash-4.1#
```

Part 3: Developing a Hadoop program

We are required to develop a Hadoop program that produces the **ngram-frequencies** of the text in a given input file. **n-gram is a contiguous sequence of n characters from a given sequence of text**.

An **n-gram of size 1** is referred to as a "**unigram**"; size **2** is a "**bigram**" (or, less commonly, a "**digram**"); size **3** is a "**trigram**". For example, the **2-gram frequency** in the text, “**Helloworld**” is as follows:

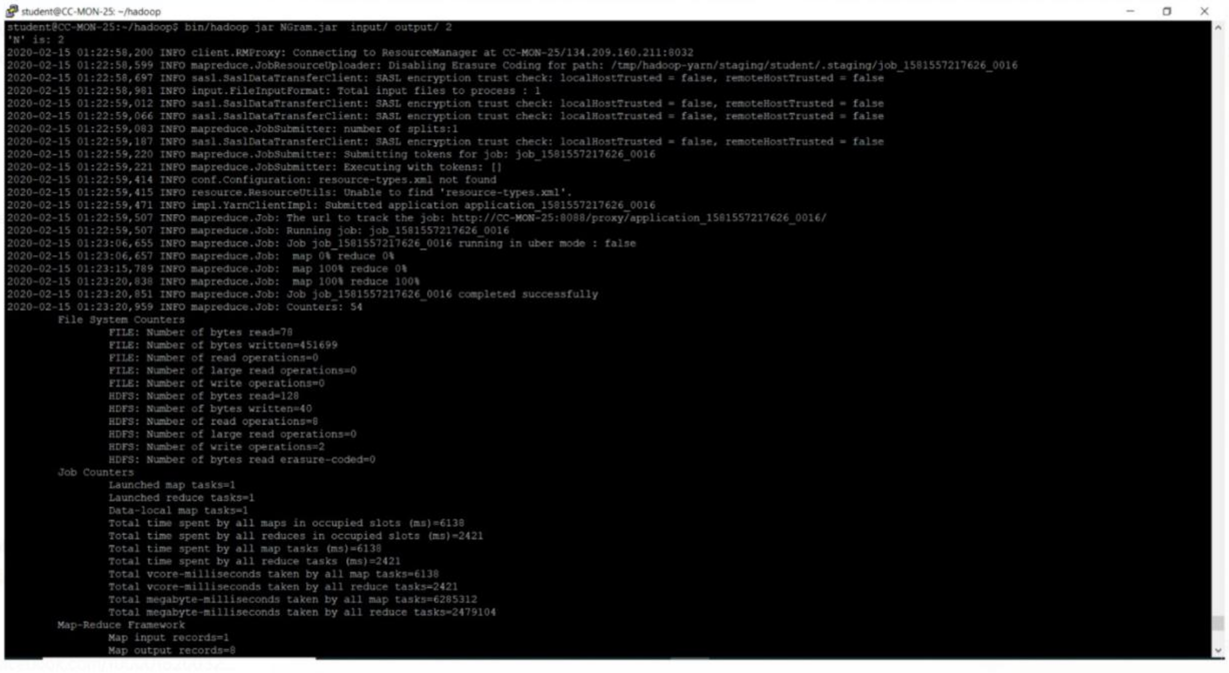
He-1, el-1, ll-1, lo-1, ow-1, wo-1, or-1, rl-1, ld-1

The program should accept **n** as an **input parameter** and produce the **n-gram frequencies** in the **text** as an **output file**.

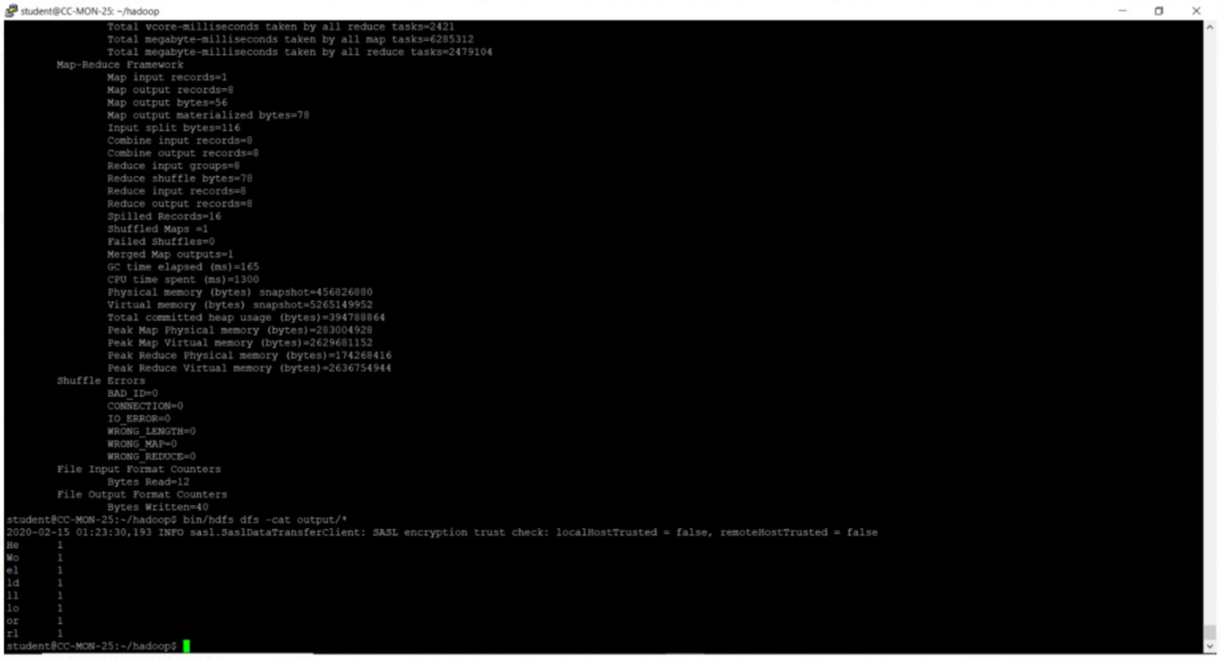
Answer:

Source code and **readme file** (explaining the procedure of the program) is provided separately in the zip file.

Output of the program:



```
student@CC-MON-25:~/hadoop$ bin/hadoop jar NGram.jar  input/ output/ 2
'N' is: 2
2020-02-15 01:22:58,200 INFO client.RMProxy: Connecting to ResourceManager at CC-MON-25/134.209.160.211:8032
2020-02-15 01:22:58,599 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/.staging/job_1581557217626_0016
2020-02-15 01:22:58,697 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 01:22:58,981 INFO input.FileInputFormat: Total input files to process : 1
2020-02-15 01:22:59,012 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 01:22:59,066 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 01:22:59,083 INFO mapreduce.JobSubmitter: number of splits:1
2020-02-15 01:22:59,187 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 01:22:59,220 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1581557217626_0016
2020-02-15 01:22:59,221 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-02-15 01:22:59,414 INFO conf.Configuration: resource-types.xml not found
2020-02-15 01:22:59,415 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-02-15 01:22:59,471 INFO impl.YarnClientImpl: Submitted application application_1581557217626_0016
2020-02-15 01:22:59,507 INFO mapreduce.Job: The url to track the job: http://CC-MON-25:8088/proxy/application_1581557217626_0016/
2020-02-15 01:22:59,507 INFO mapreduce.Job: Running job: job_1581557217626_0016
2020-02-15 01:23:06,655 INFO mapreduce.Job: Job job_1581557217626_0016 running in uber mode : false
2020-02-15 01:23:06,657 INFO mapreduce.Job:  map 0% reduce 0%
2020-02-15 01:23:15,789 INFO mapreduce.Job:  map 100% reduce 0%
2020-02-15 01:23:20,838 INFO mapreduce.Job:  map 100% reduce 100%
2020-02-15 01:23:20,851 INFO mapreduce.Job: Job job_1581557217626_0016 completed successfully
2020-02-15 01:23:20,959 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=78
                FILE: Number of bytes written=451699
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=128
                HDFS: Number of bytes written=40
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6138
                Total time spent by all reduces in occupied slots (ms)=2421
                Total time spent by all map tasks (ms)=6138
                Total time spent by all reduce tasks (ms)=2421
                Total vcore-milliseconds taken by all map tasks=6138
                Total vcore-milliseconds taken by all reduce tasks=2421
                Total megabyte-milliseconds taken by all map tasks=6285312
                Total megabyte-milliseconds taken by all reduce tasks=2479104
        Map-Reduce Framework
                Map input records=1
                Map output records=8
```



student@CC-MON-25: ~/hadoop

```
                Total vcore-milliseconds taken by all reduce tasks=2421
                Total megabyte-milliseconds taken by all map tasks=6285312
                Total megabyte-milliseconds taken by all reduce tasks=2479104
        Map-Reduce Framework
                Map input records=1
                Map output records=8
                Map output bytes=56
                Map output materialized bytes=78
                Input split bytes=116
                Combine input records=8
                Combine output records=8
                Reduce input groups=8
                Reduce shuffle bytes=78
                Reduce input records=8
                Reduce output records=8
                Spilled Records=16
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=165
                CPU time spent (ms)=1300
                Physical memory (bytes) snapshot=456826880
                Virtual memory (bytes) snapshot=5265149952
                Total committed heap usage (bytes)=394788864
                Peak Map Physical memory (bytes)=283004928
                Peak Map Virtual memory (bytes)=2629681152
                Peak Reduce Physical memory (bytes)=174268416
                Peak Reduce Virtual memory (bytes)=2636754944
        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=12
        File Output Format Counters
                Bytes Written=40
student@CC-MON-25:~/hadoop$ bin/hdfs dfs -cat output/*
2020-02-15 01:23:30,193 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
He      1
Wo      1
el      1
ld      1
ll      1
lo      1
or      1
rl      1
student@CC-MON-25:~/hadoop$
```

The steps to execute the jar file are mentioned in the readme file (inside the folder provided).

Part 4: Developing a Hadoop program to analyze real logs

In this part, we developed several MapReduce programs to analyze a real anonymous log to answer several questions based on the log. The log is in ‘**access_log.zip**’, provided in the **Course Web**.

The log file is in Common Log Format:
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lowpro.js HTTP/1.1"
200 10469
%h %l %u %t \"%r\" %>s %b

Where:

- %h is the IP address of the client
- %l is identity of the client, or "-" if it's unavailable
- %u is username of the client, or "-" if it's unavailable
- %t is the time that the server finished processing the request. The format is [day/month/year:hour:minute:second zone]
- %r is the request line from the client is given (in double quotes). It contains the method, path, query-string, and protocol or the request.
- %>s is the status code that the server sends back to the client. You will see see mostly status codes 200 (OK - The request has succeeded), 304 (Not Modified) and 404 (Not Found). See more information on status codes in W3C.org
- %b is the size of the object returned to the client, in bytes. It will be "-" in case of status code 304.

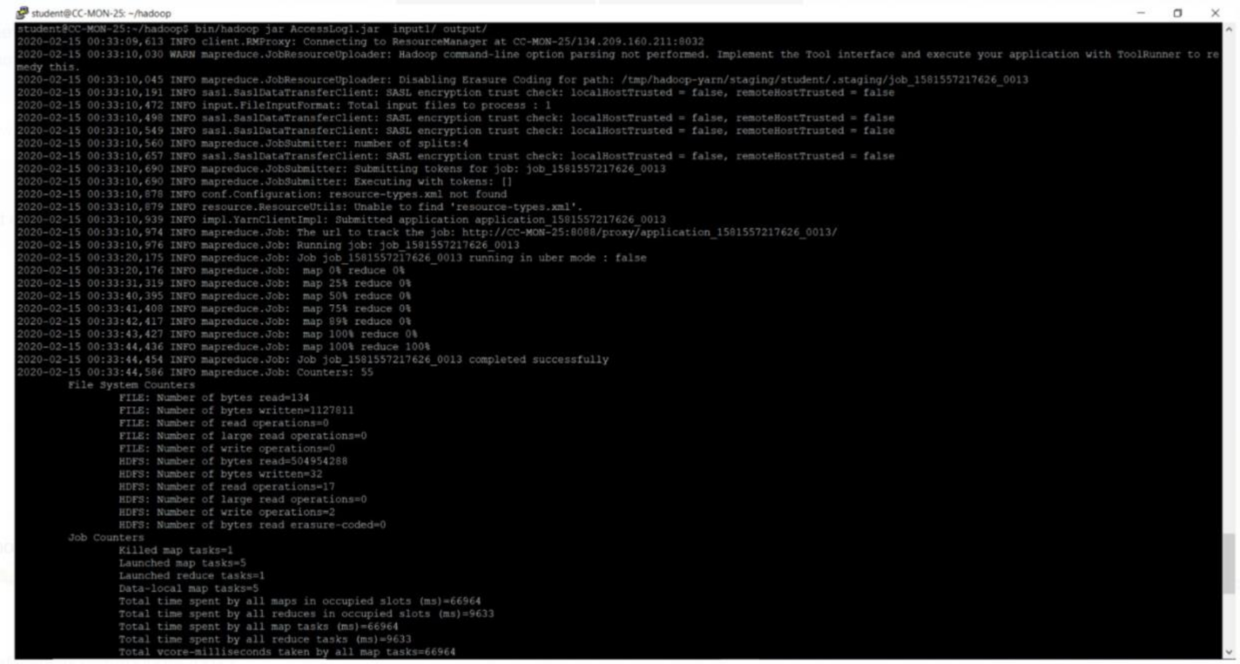
Answer:

Source code and **readme file** (explaining the procedure of the program) is provided separately in the zip file.

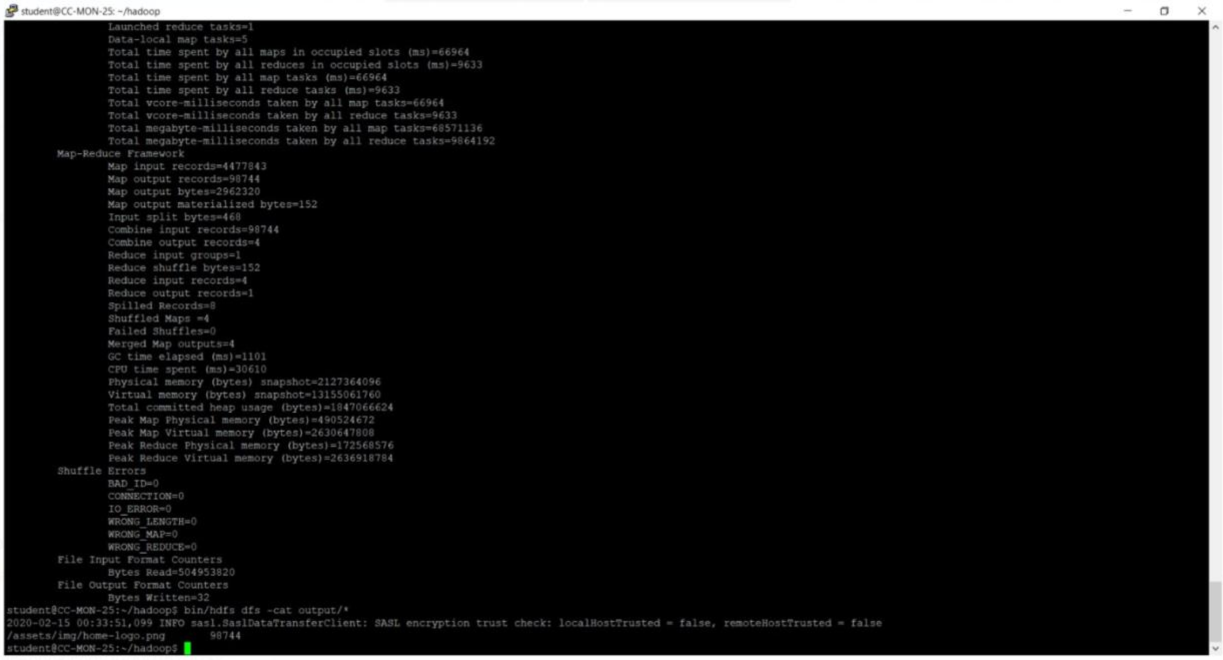
Problems:

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

   Answer: 98744



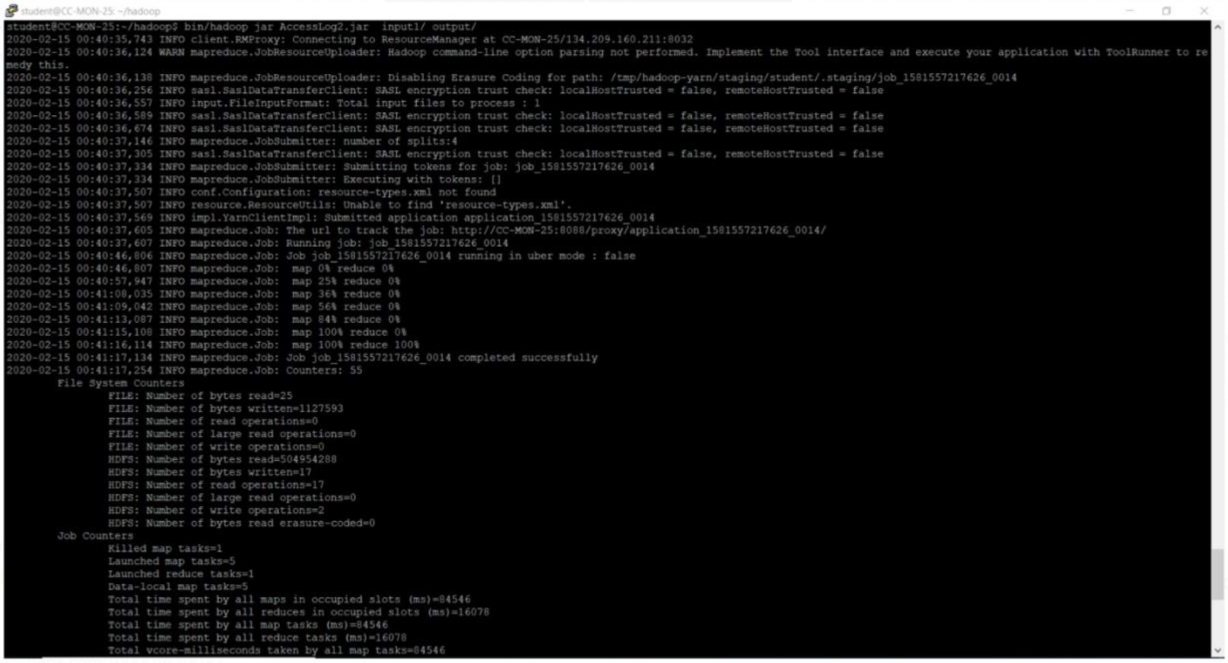
```
student@CC-MON-25:~/hadoop$ bin/hadoop jar AccessLog1.jar  input1/ output/
2020-02-15 00:33:09,613 INFO client.RMProxy: Connecting to ResourceManager at CC-MON-25/134.209.160.211:8032
2020-02-15 00:33:10,030 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to re
medy this.
2020-02-15 00:33:10,045 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/.staging/job_1581557217626_0013
2020-02-15 00:33:10,191 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:33:10,472 INFO input.FileInputFormat: Total input files to process : 1
2020-02-15 00:33:10,498 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:33:10,549 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:33:10,560 INFO mapreduce.JobSubmitter: number of splits:4
2020-02-15 00:33:10,657 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:33:10,690 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1581557217626_0013
2020-02-15 00:33:10,690 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-02-15 00:33:10,878 INFO conf.Configuration: resource-types.xml not found
2020-02-15 00:33:10,879 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-02-15 00:33:10,939 INFO impl.YarnClientImpl: Submitted application application_1581557217626_0013
2020-02-15 00:33:10,974 INFO mapreduce.Job: The url to track the job: http://CC-MON-25:8088/proxy/application_1581557217626_0013/
2020-02-15 00:33:10,976 INFO mapreduce.Job: Running job: job_1581557217626_0013
2020-02-15 00:33:20,175 INFO mapreduce.Job: Job job_1581557217626_0013 running in uber mode : false
2020-02-15 00:33:20,176 INFO mapreduce.Job:  map 0% reduce 0%
2020-02-15 00:33:31,319 INFO mapreduce.Job:  map 25% reduce 0%
2020-02-15 00:33:40,395 INFO mapreduce.Job:  map 50% reduce 0%
2020-02-15 00:33:41,408 INFO mapreduce.Job:  map 75% reduce 0%
2020-02-15 00:33:42,417 INFO mapreduce.Job:  map 89% reduce 0%
2020-02-15 00:33:43,427 INFO mapreduce.Job:  map 100% reduce 0%
2020-02-15 00:33:44,436 INFO mapreduce.Job:  map 100% reduce 100%
2020-02-15 00:33:44,454 INFO mapreduce.Job: Job job_1581557217626_0013 completed successfully
2020-02-15 00:33:44,586 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=134
                FILE: Number of bytes written=1127811
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=504954288
                HDFS: Number of bytes written=32
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=66964
                Total time spent by all reduces in occupied slots (ms)=9633
                Total time spent by all map tasks (ms)=66964
                Total time spent by all reduce tasks (ms)=9633
                Total vcore-milliseconds taken by all map tasks=66964
```



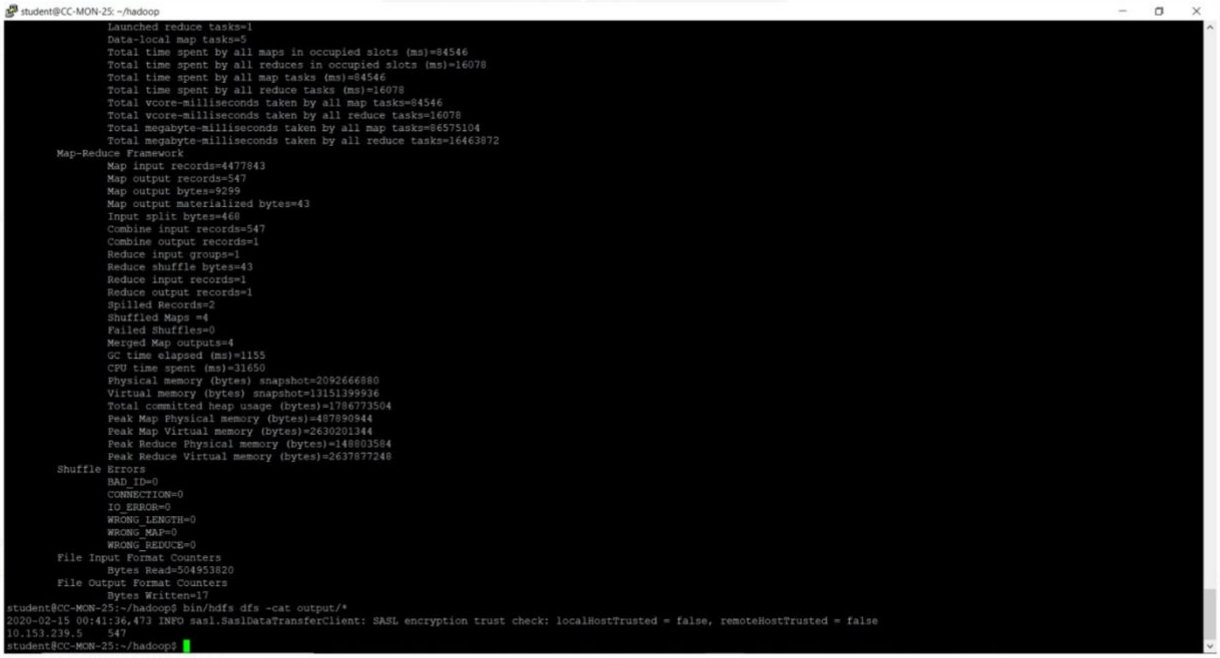
```
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=66964
                Total time spent by all reduces in occupied slots (ms)=9633
                Total time spent by all map tasks (ms)=66964
                Total time spent by all reduce tasks (ms)=9633
                Total vcore-milliseconds taken by all map tasks=66964
                Total vcore-milliseconds taken by all reduce tasks=9633
                Total megabyte-milliseconds taken by all map tasks=68571136
                Total megabyte-milliseconds taken by all reduce tasks=9864192
        Map-Reduce Framework
                Map input records=4477843
                Map output records=98744
                Map output bytes=2962320
                Map output materialized bytes=152
                Input split bytes=468
                Combine input records=98744
                Combine output records=4
                Reduce input groups=1
                Reduce shuffle bytes=152
                Reduce input records=4
                Reduce output records=1
                Spilled Records=8
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1101
                CPU time spent (ms)=30610
                Physical memory (bytes) snapshot=2127364096
                Virtual memory (bytes) snapshot=13155061760
                Total committed heap usage (bytes)=1847066624
                Peak Map Physical memory (bytes)=490524672
                Peak Map Virtual memory (bytes)=2630647808
                Peak Reduce Physical memory (bytes)=172568576
                Peak Reduce Virtual memory (bytes)=2636918784
        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=504953820
        File Output Format Counters
                Bytes Written=32
student@CC-MON-25:~/hadoop$ bin/hdfs dfs -cat output/*
2020-02-15 00:33:51,099 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
/assets/img/home-logo.png       98744
student@CC-MON-25:~/hadoop$
```

2. **How many hits were made from the IP: 10.153.239.5?**

Answer: 547



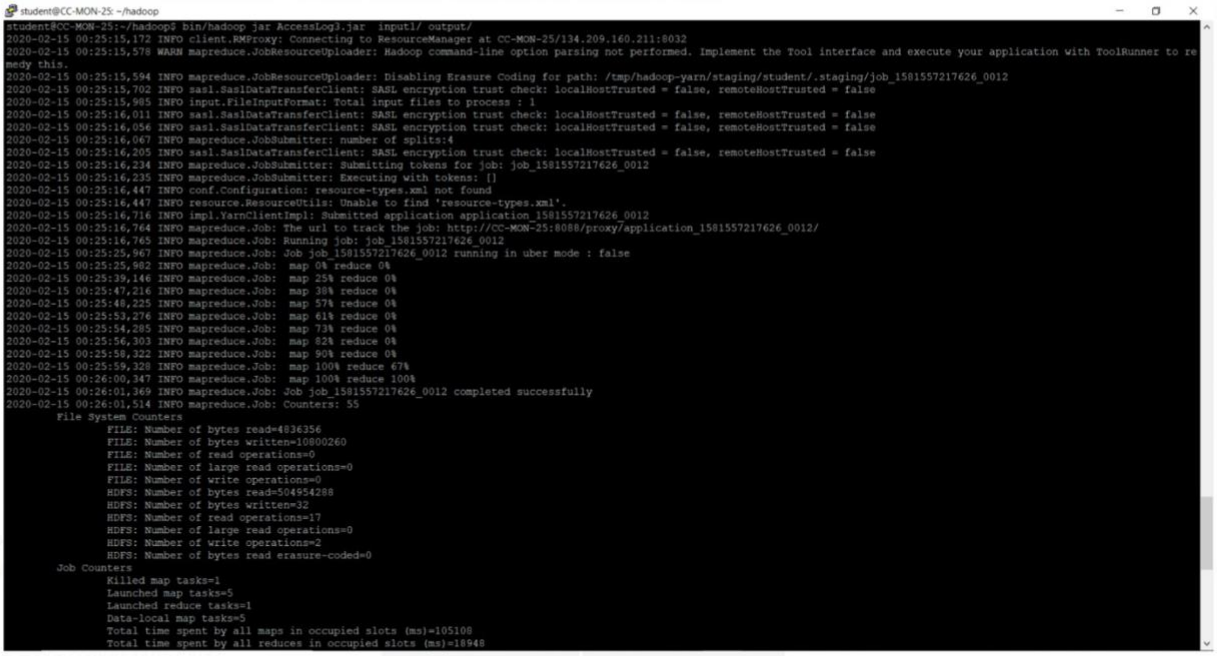
```
student@CC-MON-25:~/hadoop$ bin/hadoop jar AccessLog2.jar  input1/ output/
2020-02-15 00:40:35,743 INFO client.RMProxy: Connecting to ResourceManager at CC-MON-25/134.209.160.211:8032
2020-02-15 00:40:36,124 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to re
medy this.
2020-02-15 00:40:36,138 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/.staging/job_1581557217626_0014
2020-02-15 00:40:36,256 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:40:36,557 INFO input.FileInputFormat: Total input files to process : 1
2020-02-15 00:40:36,589 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:40:36,674 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:40:37,146 INFO mapreduce.JobSubmitter: number of splits:4
2020-02-15 00:40:37,305 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:40:37,334 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1581557217626_0014
2020-02-15 00:40:37,334 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-02-15 00:40:37,507 INFO conf.Configuration: resource-types.xml not found
2020-02-15 00:40:37,507 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-02-15 00:40:37,569 INFO impl.YarnClientImpl: Submitted application application_1581557217626_0014
2020-02-15 00:40:37,605 INFO mapreduce.Job: The url to track the job: http://CC-MON-25:8088/proxy/application_1581557217626_0014/
2020-02-15 00:40:37,607 INFO mapreduce.Job: Running job: job_1581557217626_0014
2020-02-15 00:40:46,806 INFO mapreduce.Job: Job job_1581557217626_0014 running in uber mode : false
2020-02-15 00:40:46,807 INFO mapreduce.Job:  map 0% reduce 0%
2020-02-15 00:40:57,947 INFO mapreduce.Job:  map 25% reduce 0%
2020-02-15 00:41:08,035 INFO mapreduce.Job:  map 36% reduce 0%
2020-02-15 00:41:09,042 INFO mapreduce.Job:  map 56% reduce 0%
2020-02-15 00:41:13,087 INFO mapreduce.Job:  map 84% reduce 0%
2020-02-15 00:41:15,108 INFO mapreduce.Job:  map 100% reduce 0%
2020-02-15 00:41:16,114 INFO mapreduce.Job:  map 100% reduce 100%
2020-02-15 00:41:17,134 INFO mapreduce.Job: Job job_1581557217626_0014 completed successfully
2020-02-15 00:41:17,254 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=25
                FILE: Number of bytes written=1127593
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=504954288
                HDFS: Number of bytes written=17
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=84546
                Total time spent by all reduces in occupied slots (ms)=16078
                Total time spent by all map tasks (ms)=84546
                Total time spent by all reduce tasks (ms)=16078
                Total vcore-milliseconds taken by all map tasks=84546
```



```
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=84546
                Total time spent by all reduces in occupied slots (ms)=16078
                Total time spent by all map tasks (ms)=84546
                Total time spent by all reduce tasks (ms)=16078
                Total vcore-milliseconds taken by all map tasks=84546
                Total vcore-milliseconds taken by all reduce tasks=16078
                Total megabyte-milliseconds taken by all map tasks=86575104
                Total megabyte-milliseconds taken by all reduce tasks=16463872
        Map-Reduce Framework
                Map input records=4477843
                Map output records=547
                Map output bytes=9299
                Map output materialized bytes=43
                Input split bytes=468
                Combine input records=547
                Combine output records=1
                Reduce input groups=1
                Reduce shuffle bytes=43
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1155
                CPU time spent (ms)=31650
                Physical memory (bytes) snapshot=2092666880
                Virtual memory (bytes) snapshot=13151399936
                Total committed heap usage (bytes)=1786773504
                Peak Map Physical memory (bytes)=487890944
                Peak Map Virtual memory (bytes)=2630201344
                Peak Reduce Physical memory (bytes)=148803584
                Peak Reduce Virtual memory (bytes)=2637877248
        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=504953820
        File Output Format Counters
                Bytes Written=17
student@CC-MON-25:~/hadoop$ bin/hdfs dfs -cat output/*
2020-02-15 00:41:36,473 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
10.153.239.5    547
student@CC-MON-25:~/hadoop$
```

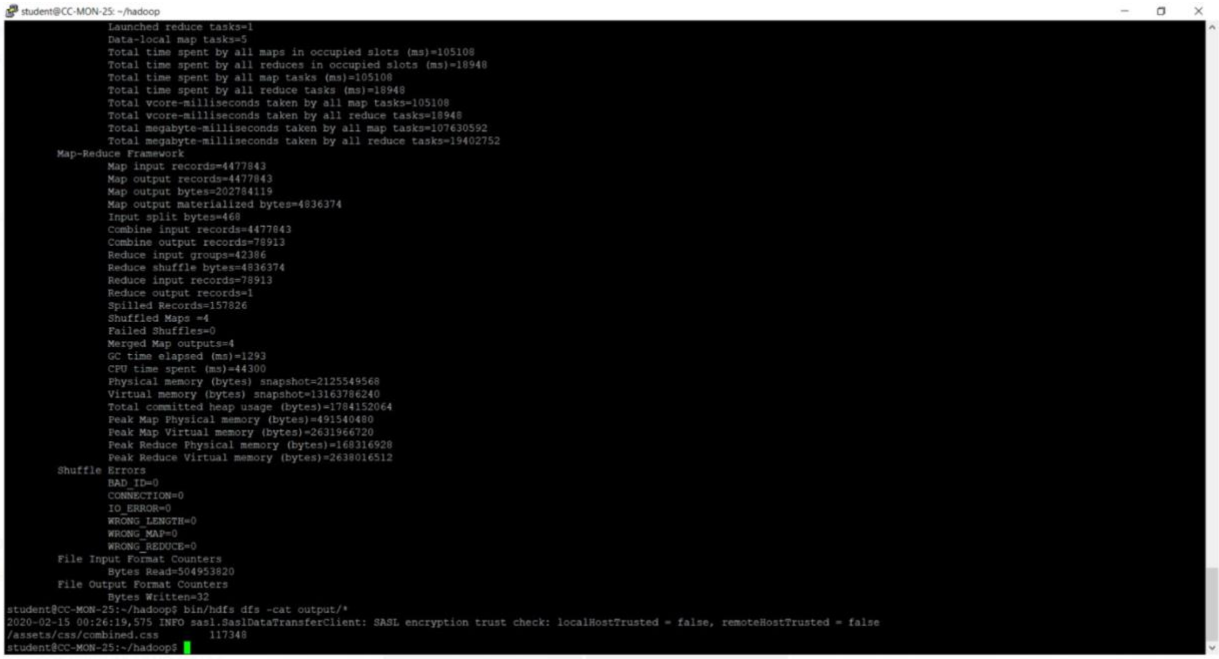
3. Which path in the website has been hit most? How many hits were made to the path?

Answer: /assets/css/combined.css   117348



student@CC-MON-25: ~/hadoop

```
student@CC-MON-25:~/hadoop$ bin/hadoop jar AccessLog3.jar  input1/ output/
2020-02-15 00:25:15,172 INFO client.RMProxy: Connecting to ResourceManager at CC-MON-25/134.209.160.211:8032
2020-02-15 00:25:15,578 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to re
medy this.
2020-02-15 00:25:15,594 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/.staging/job_1581557217626_0012
2020-02-15 00:25:15,702 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:25:15,985 INFO input.FileInputFormat: Total input files to process : 1
2020-02-15 00:25:16,011 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:25:16,056 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:25:16,067 INFO mapreduce.JobSubmitter: number of splits:4
2020-02-15 00:25:16,205 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:25:16,234 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1581557217626_0012
2020-02-15 00:25:16,235 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-02-15 00:25:16,447 INFO conf.Configuration: resource-types.xml not found
2020-02-15 00:25:16,447 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-02-15 00:25:16,716 INFO impl.YarnClientImpl: Submitted application application_1581557217626_0012
2020-02-15 00:25:16,764 INFO mapreduce.Job: The url to track the job: http://CC-MON-25:8088/proxy/application_1581557217626_0012/
2020-02-15 00:25:16,765 INFO mapreduce.Job: Running job: job_1581557217626_0012
2020-02-15 00:25:25,967 INFO mapreduce.Job: Job job_1581557217626_0012 running in uber mode : false
2020-02-15 00:25:25,982 INFO mapreduce.Job:  map 0% reduce 0%
2020-02-15 00:25:39,146 INFO mapreduce.Job:  map 25% reduce 0%
2020-02-15 00:25:47,216 INFO mapreduce.Job:  map 38% reduce 0%
2020-02-15 00:25:48,225 INFO mapreduce.Job:  map 57% reduce 0%
2020-02-15 00:25:53,276 INFO mapreduce.Job:  map 61% reduce 0%
2020-02-15 00:25:54,285 INFO mapreduce.Job:  map 73% reduce 0%
2020-02-15 00:25:56,303 INFO mapreduce.Job:  map 82% reduce 0%
2020-02-15 00:25:58,322 INFO mapreduce.Job:  map 90% reduce 0%
2020-02-15 00:25:59,328 INFO mapreduce.Job:  map 100% reduce 67%
2020-02-15 00:26:00,347 INFO mapreduce.Job:  map 100% reduce 100%
2020-02-15 00:26:01,369 INFO mapreduce.Job: Job job_1581557217626_0012 completed successfully
2020-02-15 00:26:01,514 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=4836356
                FILE: Number of bytes written=10800260
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=504954288
                HDFS: Number of bytes written=32
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=105108
                Total time spent by all reduces in occupied slots (ms)=18948
```

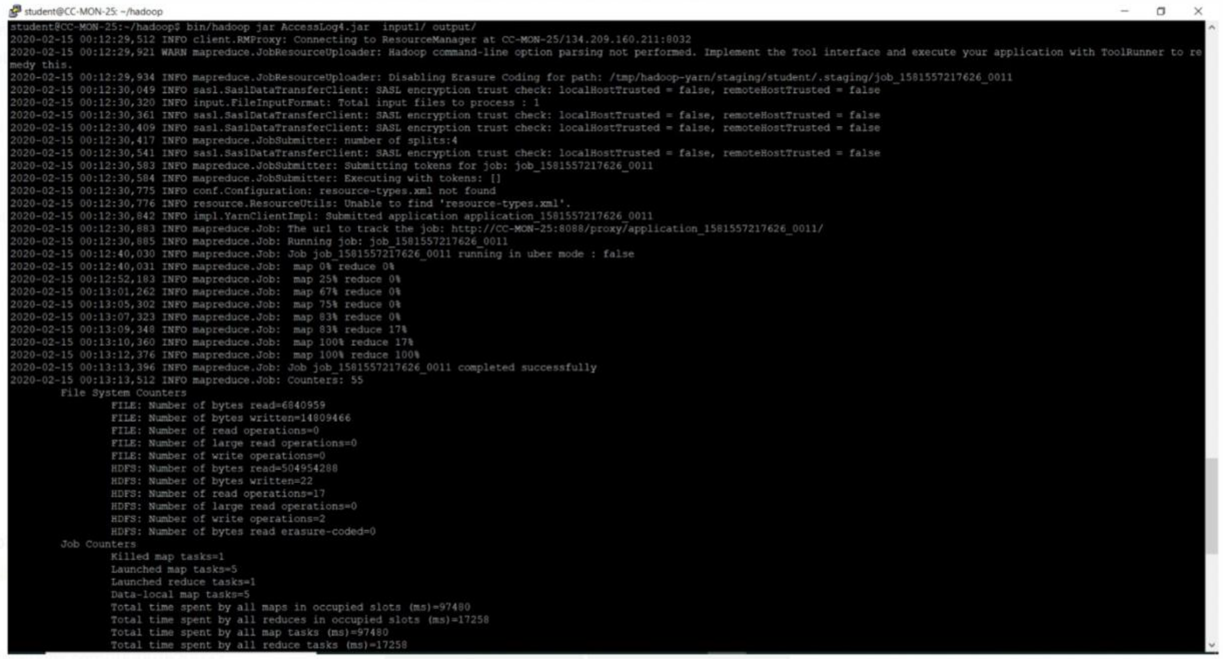


student@CC-MON-25: ~/hadoop

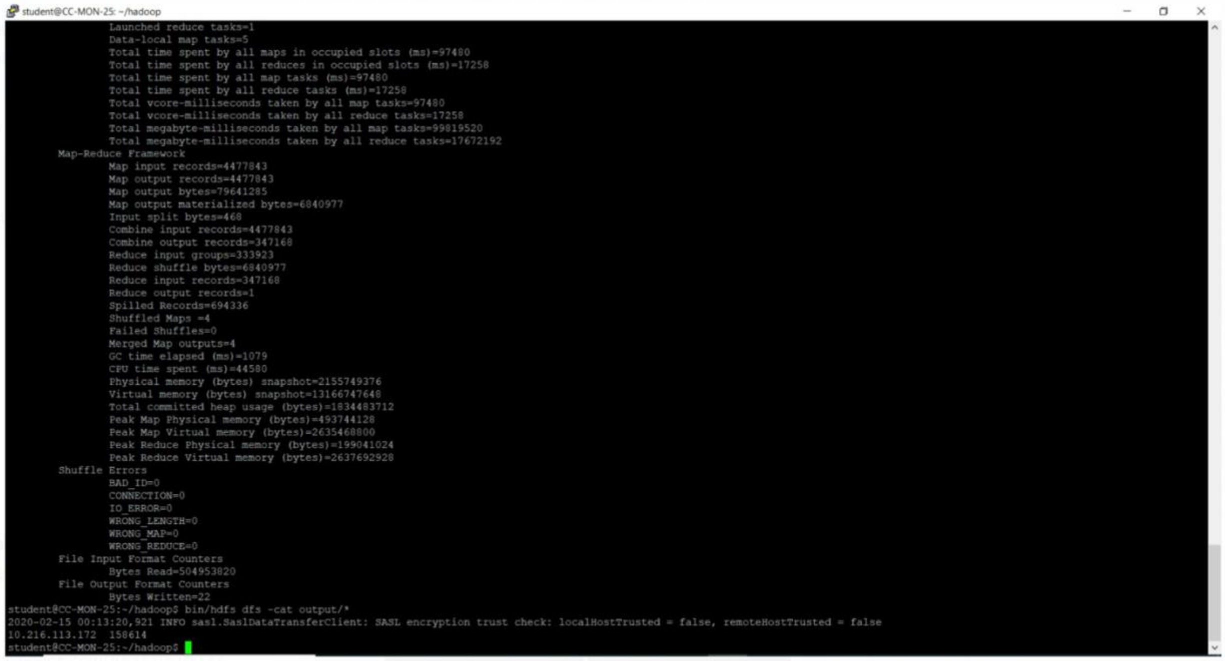
```
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=105108
                Total time spent by all reduces in occupied slots (ms)=18948
                Total time spent by all map tasks (ms)=105108
                Total time spent by all reduce tasks (ms)=18948
                Total vcore-milliseconds taken by all map tasks=105108
                Total vcore-milliseconds taken by all reduce tasks=18948
                Total megabyte-milliseconds taken by all map tasks=107630592
                Total megabyte-milliseconds taken by all reduce tasks=19402752
        Map-Reduce Framework
                Map input records=4477843
                Map output records=4477843
                Map output bytes=202784119
                Map output materialized bytes=4836374
                Input split bytes=468
                Combine input records=4477843
                Combine output records=78913
                Reduce input groups=42386
                Reduce shuffle bytes=4836374
                Reduce input records=78913
                Reduce output records=1
                Spilled Records=157826
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1293
                CPU time spent (ms)=44300
                Physical memory (bytes) snapshot=2125549568
                Virtual memory (bytes) snapshot=13163786240
                Total committed heap usage (bytes)=1784152064
                Peak Map Physical memory (bytes)=491540480
                Peak Map Virtual memory (bytes)=2631966720
                Peak Reduce Physical memory (bytes)=168316928
                Peak Reduce Virtual memory (bytes)=2638016512
        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=504953820
        File Output Format Counters
                Bytes Written=32
student@CC-MON-25:~/hadoop$ bin/hdfs dfs -cat output/*
2020-02-15 00:26:19,575 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
/assets/css/combined.css        117348
student@CC-MON-25:~/hadoop$ 
```

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

Answer: 10.216.113.172 158614



```
student@CC-MON-25:~/hadoop$ bin/hadoop jar AccessLog4.jar  input1/ output/
2020-02-15 00:12:29,512 INFO client.RMProxy: Connecting to ResourceManager at CC-MON-25/134.209.160.211:8032
2020-02-15 00:12:29,921 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to re
medy this.
2020-02-15 00:12:29,934 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/student/.staging/job_1581557217626_0011
2020-02-15 00:12:30,049 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:12:30,320 INFO input.FileInputFormat: Total input files to process : 1
2020-02-15 00:12:30,361 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:12:30,409 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:12:30,417 INFO mapreduce.JobSubmitter: number of splits:4
2020-02-15 00:12:30,541 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-02-15 00:12:30,583 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1581557217626_0011
2020-02-15 00:12:30,584 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-02-15 00:12:30,775 INFO conf.Configuration: resource-types.xml not found
2020-02-15 00:12:30,776 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-02-15 00:12:30,842 INFO impl.YarnClientImpl: Submitted application application_1581557217626_0011
2020-02-15 00:12:30,883 INFO mapreduce.Job: The url to track the job: http://CC-MON-25:8088/proxy/application_1581557217626_0011/
2020-02-15 00:12:30,885 INFO mapreduce.Job: Running job: job_1581557217626_0011
2020-02-15 00:12:40,030 INFO mapreduce.Job: Job job_1581557217626_0011 running in uber mode : false
2020-02-15 00:12:40,031 INFO mapreduce.Job:  map 0% reduce 0%
2020-02-15 00:12:52,183 INFO mapreduce.Job:  map 25% reduce 0%
2020-02-15 00:13:01,262 INFO mapreduce.Job:  map 67% reduce 0%
2020-02-15 00:13:05,302 INFO mapreduce.Job:  map 75% reduce 0%
2020-02-15 00:13:07,323 INFO mapreduce.Job:  map 83% reduce 0%
2020-02-15 00:13:09,348 INFO mapreduce.Job:  map 83% reduce 17%
2020-02-15 00:13:10,360 INFO mapreduce.Job:  map 100% reduce 17%
2020-02-15 00:13:12,376 INFO mapreduce.Job:  map 100% reduce 100%
2020-02-15 00:13:13,396 INFO mapreduce.Job: Job job_1581557217626_0011 completed successfully
2020-02-15 00:13:13,512 INFO mapreduce.Job: Counters: 55
        File System Counters
                FILE: Number of bytes read=6840959
                FILE: Number of bytes written=14809466
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=504954288
                HDFS: Number of bytes written=22
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=5
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=97480
                Total time spent by all reduces in occupied slots (ms)=17258
                Total time spent by all map tasks (ms)=97480
                Total time spent by all reduce tasks (ms)=17258
```



```
                Launched reduce tasks=1
                Data-local map tasks=5
                Total time spent by all maps in occupied slots (ms)=97480
                Total time spent by all reduces in occupied slots (ms)=17258
                Total time spent by all map tasks (ms)=97480
                Total time spent by all reduce tasks (ms)=17258
                Total vcore-milliseconds taken by all map tasks=97480
                Total vcore-milliseconds taken by all reduce tasks=17258
                Total megabyte-milliseconds taken by all map tasks=99819520
                Total megabyte-milliseconds taken by all reduce tasks=17672192
        Map-Reduce Framework
                Map input records=4477843
                Map output records=4477843
                Map output bytes=79641285
                Map output materialized bytes=6840977
                Input split bytes=468
                Combine input records=4477843
                Combine output records=347168
                Reduce input groups=333923
                Reduce shuffle bytes=6840977
                Reduce input records=347168
                Reduce output records=1
                Spilled Records=694336
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1079
                CPU time spent (ms)=44580
                Physical memory (bytes) snapshot=2155749376
                Virtual memory (bytes) snapshot=13166747648
                Total committed heap usage (bytes)=1834483712
                Peak Map Physical memory (bytes)=493744128
                Peak Map Virtual memory (bytes)=2635468800
                Peak Reduce Physical memory (bytes)=199041024
                Peak Reduce Virtual memory (bytes)=2637692928
        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=504953820
        File Output Format Counters
                Bytes Written=22
student@CC-MON-25:~/hadoop$ bin/hdfs dfs -cat output/*
2020-02-15 00:13:20,921 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
10.216.113.172  158614
student@CC-MON-25:~/hadoop$
```

The steps to execute the jar files are mentioned in the readme file (inside the folder provided).