INFSCI 2750 – Cloud Computing

Mini Project 2 – Working with Apache Spark

Charu Sreedharan (chs263@pitt.edu)

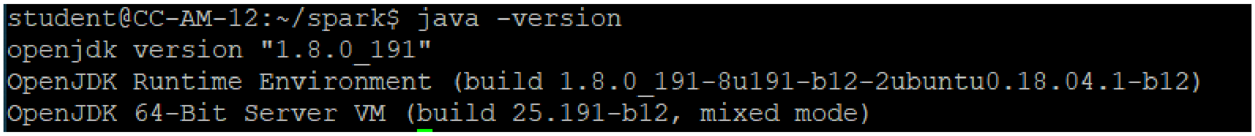
Lakshmi Ravichandran (lar146@pitt.edu)

Sanzil Madye (ssm59@pitt.edu)

Part 1 – Setting up Spark

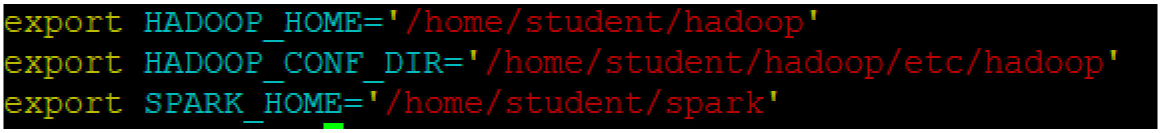
Configuring Spark on top of Hadoop cluster. Hadoop cluster setup was done in previous project. Below steps are followed to configure Spark on top of Hadoop.

1. Pre-requisites for running spark programs in YARN are done in previous project. Java JDK version 8 –



```
student@CC-AM-12:~/spark$ java -version
openjdk version "1.8.0_191"
OpenJDK Runtime Environment (build 1.8.0_191-8u191-b12-2ubuntu0.18.04.1-b12)
OpenJDK 64-Bit Server VM (build 25.191-b12, mixed mode)
```

Setting environment variables – HADDOP_HOME, HADOOP_CONF_DIR, SPARK_HOME in **~/.bashrc file** -



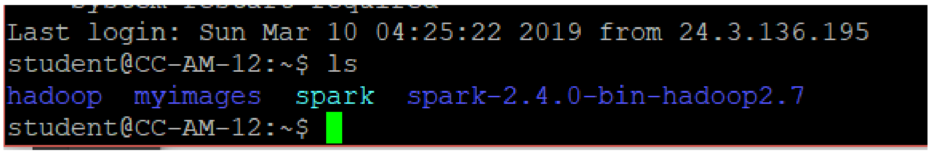
```
export HADOOP_HOME='/home/student/hadoop'
export HADOOP_CONF_DIR='/home/student/hadoop/etc/hadoop'
export SPARK_HOME='/home/student/spark'
```

Hadoop cluster started and running in all three machines -

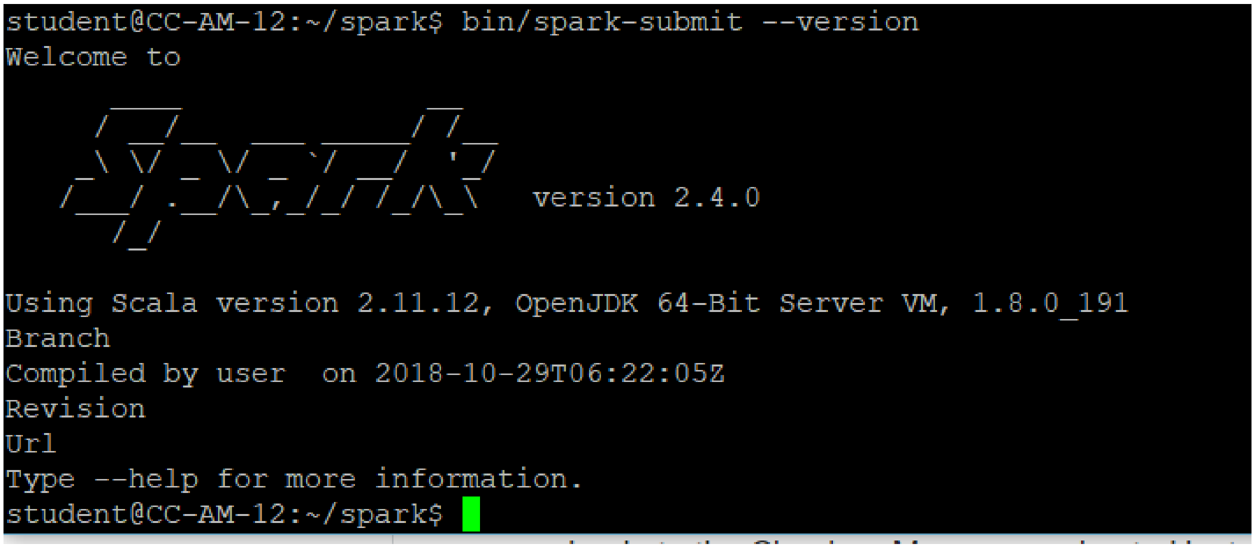
$HADOOP_HOME/sbin/start-all.sh
$HADOOP_HOME/bin/mapred --daemon start historyserver
$SPARK_HOME/sbin/start-history-server.sh

Spark version 11 – Download, unpack spark, install and rename

```
wget http://us.mirrors.quenda.co/apache/spark/spark-2.4.0/spark-2.4.0-bin-hadoop2.7.tgz
tar zxvf spark-2.4.0-bin-hadoop2.7.tgz
ln -s spark-2.4.0-bin-hadoop2.7 spark
```



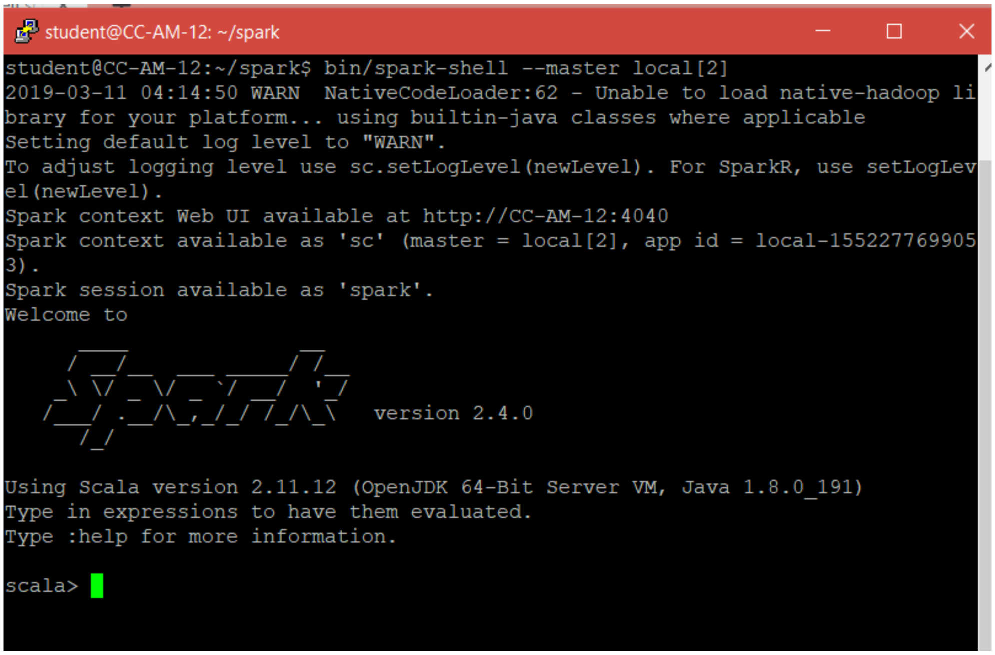
```
System restart required
Last login: Sun Mar 10 04:25:22 2019 from 24.3.136.195
student@CC-AM-12:~$ ls
hadoop  myimages  spark  spark-2.4.0-bin-hadoop2.7
student@CC-AM-12:~$
```



```
student@CC-AM-12:~/spark$ bin/spark-submit --version
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Scala version 2.11.12, OpenJDK 64-Bit Server VM, 1.8.0_191
Branch
Compiled by user  on 2018-10-29T06:22:05Z
Revision
Url
Type --help for more information.
student@CC-AM-12:~/spark$
```

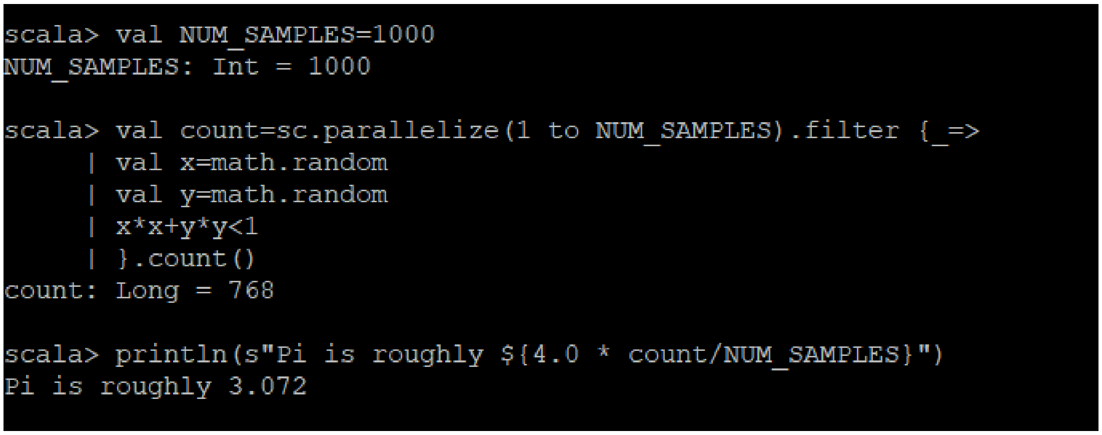
Install Scala and open spark shell in local mode

bin/spark-shell --master local[2]



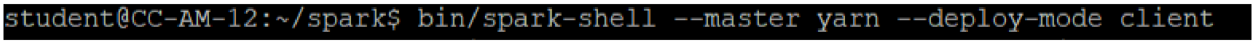
```
student@CC-AM-12:~/spark$ bin/spark-shell --master local[2]
2019-03-11 04:14:50 WARN  NativeCodeLoader:62 - Unable to load native-hadoop li
brary for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLev
el(newLevel).
Spark context Web UI available at http://CC-AM-12:4040
Spark context available as 'sc' (master = local[2], app id = local-155227769905
3).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_191)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Trying a scala program – To calculate pi value

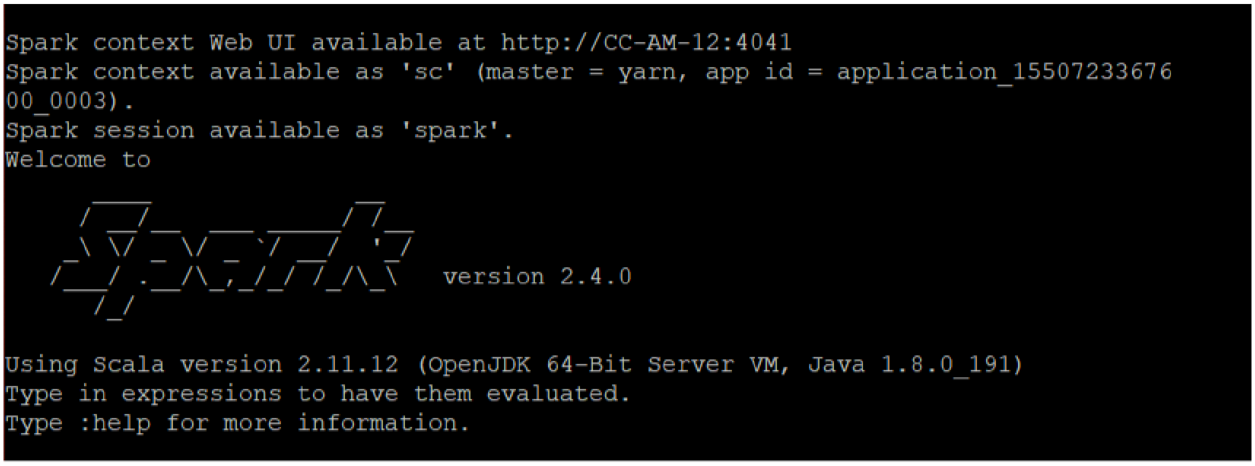


```
scala> val NUM_SAMPLES=1000
NUM_SAMPLES: Int = 1000

scala> val count=sc.parallelize(1 to NUM_SAMPLES).filter {_=>
     | val x=math.random
     | val y=math.random
     | x*x+y*y<1
     | }.count()
count: Long = 768

scala> println(s"Pi is roughly ${4.0 * count/NUM_SAMPLES}")
Pi is roughly 3.072
```

To run spark shell with YARN as the scheduler



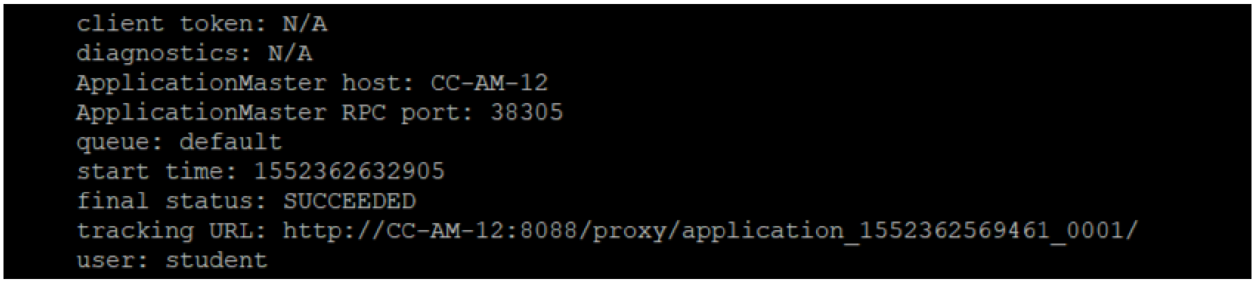
```
student@CC-AM-12:~/spark$ bin/spark-shell --master yarn --deploy-mode client
```



```
Spark context Web UI available at http://CC-AM-12:4041
Spark context available as 'sc' (master = yarn, app id = application_15507233676
00_0003).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_191)
Type in expressions to have them evaluated.
Type :help for more information.
```

Running pi program with YARN

bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn --deploy-mode cluster --driver-memory 512m --executor-memory 512m --executor-cores 1 --queue default examples/jars/spark-examples*.jar 10



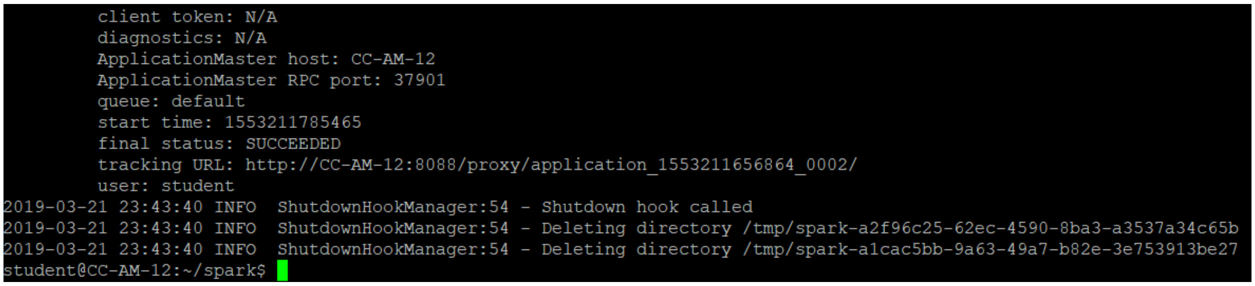
```
     client token: N/A
     diagnostics: N/A
     ApplicationMaster host: CC-AM-12
     ApplicationMaster RPC port: 38305
     queue: default
     start time: 1552362632905
     final status: SUCCEEDED
     tracking URL: http://CC-AM-12:8088/proxy/application_1552362569461_0001/
     user: student
```

Part 2 – Spark program for data analysis – To print listening count of each artist from the ‘*user_artists.dat’* data

ArtistListeningCounts.java file has the source code for counting total listening count of each artists.

To run the above code as a jar file in Spark with YARN as the scheduler

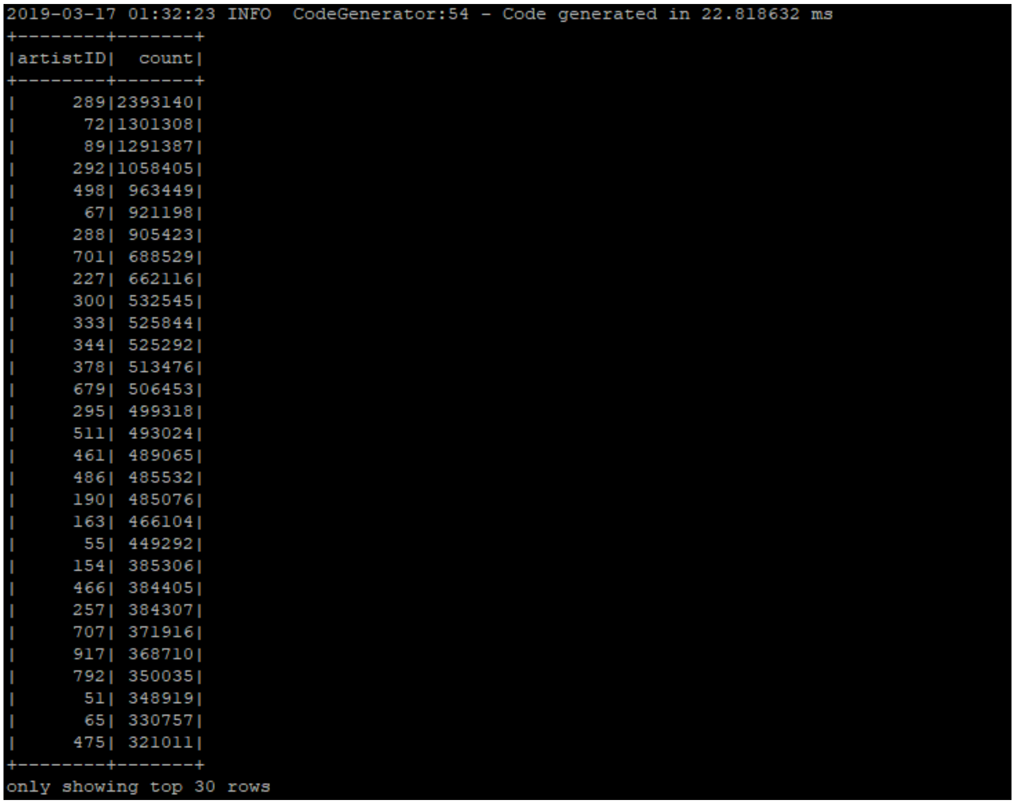
bin/spark-submit --class "com.pitt.cloudcomputing.ArtistListeningCounts" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Artist*.jar



```
        client token: N/A
        diagnostics: N/A
        ApplicationMaster host: CC-AM-12
        ApplicationMaster RPC port: 37901
        queue: default
        start time: 1553211785465
        final status: SUCCEEDED
        tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0002/
        user: student
2019-03-21 23:43:40 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-21 23:43:40 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-a2f96c25-62ec-4590-8ba3-a3537a34c65b
2019-03-21 23:43:40 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-a1cac5bb-9a63-49a7-b82e-3e753913be27
student@CC-AM-12:~/spark$
```

To view the output and the logs

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0002



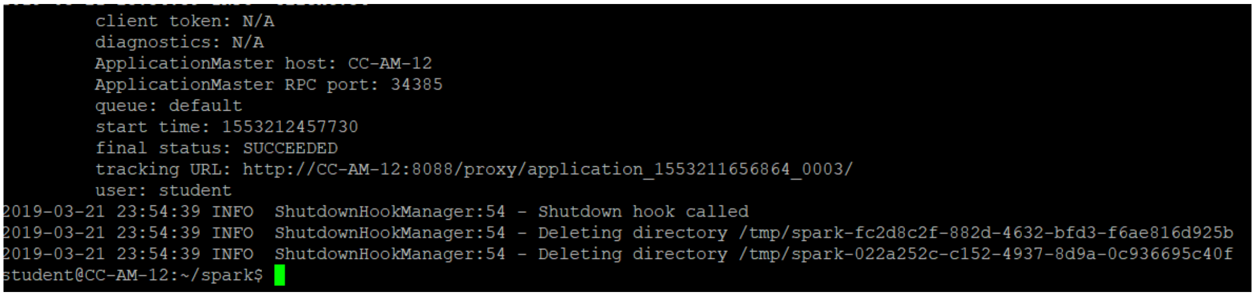
```
2019-03-17 01:32:23 INFO  CodeGenerator:54 - Code generated in 22.818632 ms
+--------+-------+
|artistID|  count|
+--------+-------+
|     289|2393140|
|      72|1301308|
|      89|1291387|
|     292|1058405|
|     498| 963449|
|      67| 921198|
|     288| 905423|
|     701| 688529|
|     227| 662116|
|     300| 532545|
|     333| 525844|
|     344| 525292|
|     378| 513476|
|     679| 506453|
|     295| 499318|
|     511| 493024|
|     461| 489065|
|     486| 485532|
|     190| 485076|
|     163| 466104|
|      55| 449292|
|     154| 385306|
|     466| 384405|
|     257| 384307|
|     707| 371916|
|     917| 368710|
|     792| 350035|
|      51| 348919|
|      65| 330757|
|     475| 321011|
+--------+-------+
only showing top 30 rows
```

Part 3 – Spark program to analyze web log – To answer questions from the log data set

(a) **Log-Analysis-1.java** file has the source code for counting and printing the number of hits for the website element /assets/img/loading.gif.

To run the above code as a jar file in Spark with YARN as the scheduler

bin/spark-submit --class "com.pitt.cloudcomputing.KeyCountEvaluator" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Log-Analysis-1*.jar

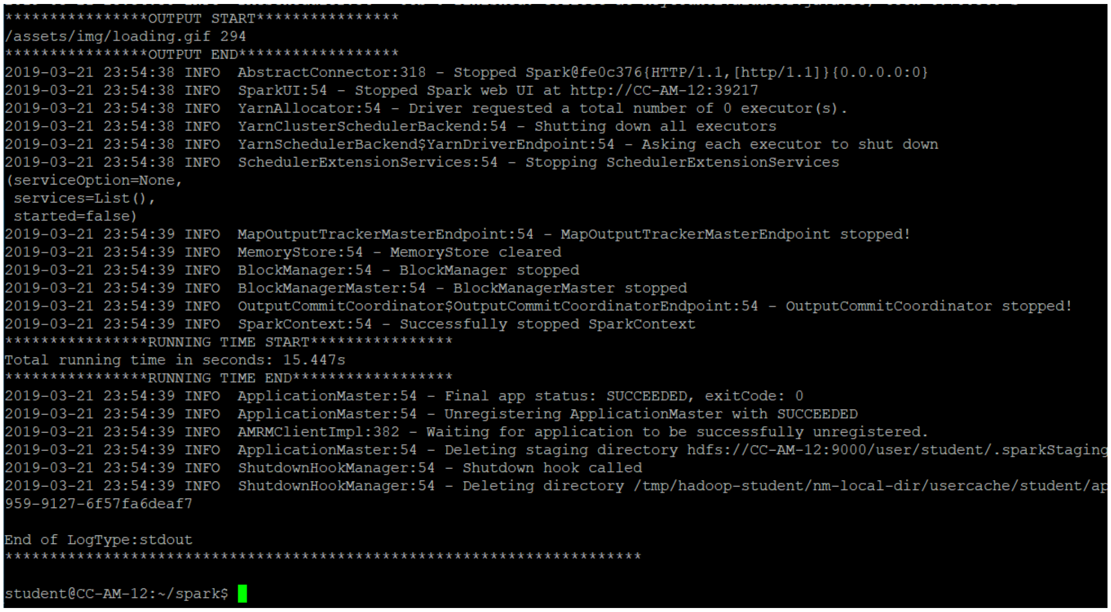


```
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-12
         ApplicationMaster RPC port: 34385
         queue: default
         start time: 1553212457730
         final status: SUCCEEDED
         tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0003/
         user: student
2019-03-21 23:54:39 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-21 23:54:39 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-fc2d8c2f-882d-4632-bfd3-f6ae816d925b
2019-03-21 23:54:39 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-022a252c-c152-4937-8d9a-0c936695c40f
student@CC-AM-12:~/spark$
```

The log file can be viewed here

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0003

The screenshot below gives the result of the program and the hits for the website element loading.gif



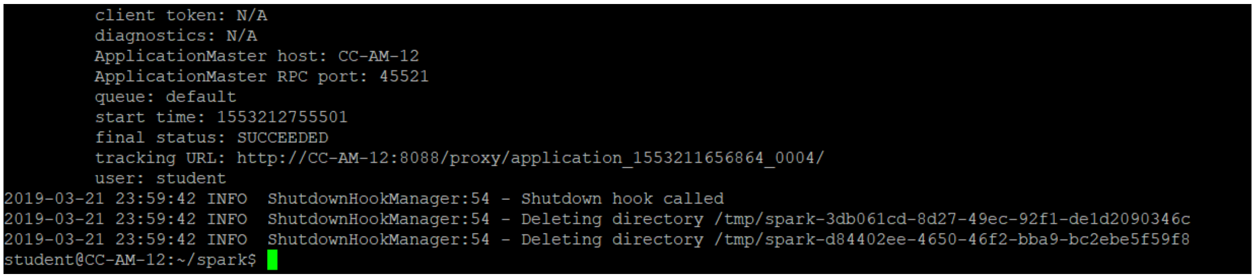
```
***************OUTPUT START****************
/assets/img/loading.gif 294
***************OUTPUT END*****************
2019-03-21 23:54:38 INFO  AbstractConnector:318 - Stopped Spark@fe0c376{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-21 23:54:38 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-12:39217
2019-03-21 23:54:38 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-21 23:54:38 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-21 23:54:38 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-21 23:54:38 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-21 23:54:39 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-21 23:54:39 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-21 23:54:39 INFO  BlockManager:54 - BlockManager stopped
2019-03-21 23:54:39 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-21 23:54:39 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-21 23:54:39 INFO  SparkContext:54 - Successfully stopped SparkContext
***************RUNNING TIME START****************
Total running time in seconds: 15.447s
***************RUNNING TIME END*****************
2019-03-21 23:54:39 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-21 23:54:39 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-21 23:54:39 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-21 23:54:39 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-12:9000/user/student/.sparkStaging
2019-03-21 23:54:39 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-21 23:54:39 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/ap
959-9127-6f57fa6deaf7

End of LogType:stdout
***********************************************************************

student@CC-AM-12:~/spark$
```

Answer : The element /assets/img/loading.gif has 294 hits

(b) Log-Analysis-2.java file has the source code for counting and printing the number of hits for the website element /assets/js/lightbox.js.

To run the above code as a jar file in Spark with YARN as the scheduler

bin/spark-submit --class "com.pitt.cloudcomputing.KeyCountEvaluator" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Log-Analysis-2*.jar

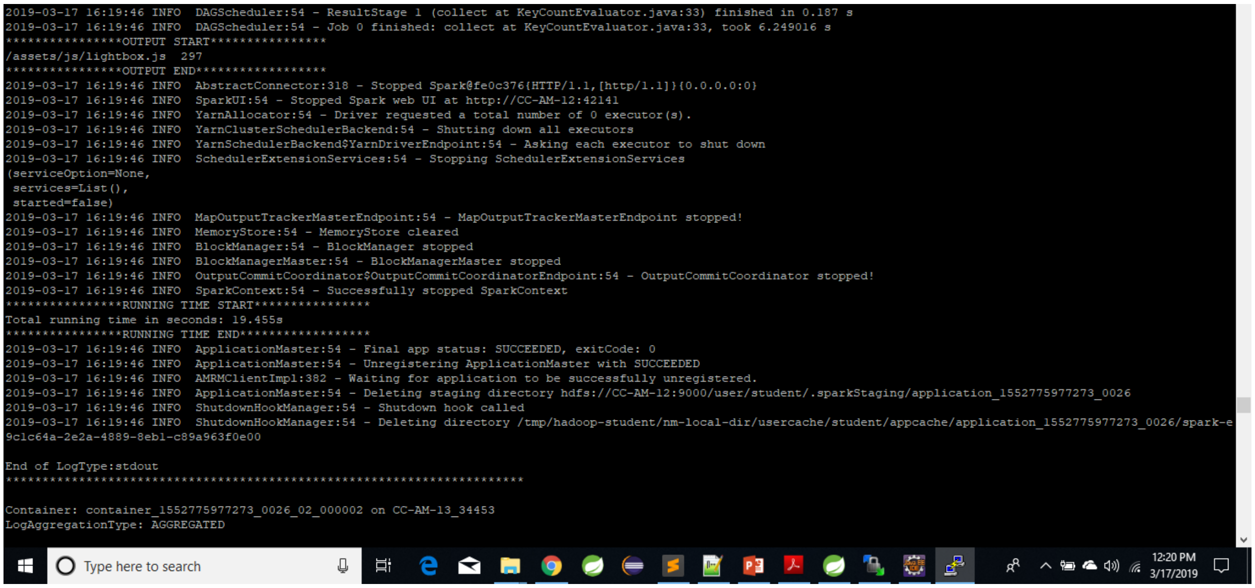


```
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-12
         ApplicationMaster RPC port: 45521
         queue: default
         start time: 1553212755501
         final status: SUCCEEDED
         tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0004/
         user: student
2019-03-21 23:59:42 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-21 23:59:42 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-3db061cd-8d27-49ec-92f1-de1d2090346c
2019-03-21 23:59:42 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-d84402ee-4650-46f2-bba9-bc2ebe5f59f8
student@CC-AM-12:~/spark$ 
```

The log file can be viewed here

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0005

The screenshot below gives the result of the program and the hits for the website element lightbox.js



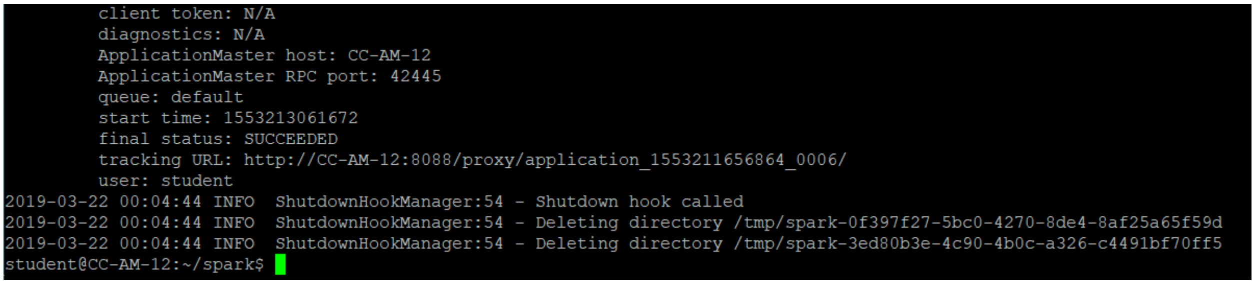
```
2019-03-17 16:19:46 INFO  DAGScheduler:54 - ResultStage 1 (collect at KeyCountEvaluator.java:33) finished in 0.187 s
2019-03-17 16:19:46 INFO  DAGScheduler:54 - Job 0 finished: collect at KeyCountEvaluator.java:33, took 6.249016 s
****************OUTPUT START****************
/assets/js/lightbox.js  297
****************OUTPUT END******************
2019-03-17 16:19:46 INFO  AbstractConnector:318 - Stopped Spark@fe0c376{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-17 16:19:46 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-12:42141
2019-03-17 16:19:46 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-17 16:19:46 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-17 16:19:46 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-17 16:19:46 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-17 16:19:46 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-17 16:19:46 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-17 16:19:46 INFO  BlockManager:54 - BlockManager stopped
2019-03-17 16:19:46 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-17 16:19:46 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-17 16:19:46 INFO  SparkContext:54 - Successfully stopped SparkContext
****************RUNNING TIME START****************
Total running time in seconds: 19.455s
****************RUNNING TIME END******************
2019-03-17 16:19:46 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-17 16:19:46 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-17 16:19:46 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-17 16:19:46 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-12:9000/user/student/.sparkStaging/application_1552775977273_0026
2019-03-17 16:19:46 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-17 16:19:46 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/appcache/application_1552775977273_0026/spark-e
9c1c64a-2e2a-4889-8eb1-c89a963f0e00

End of LogType:stdout
***********************************************************************

Container: container_1552775977273_0026_02_000002 on CC-AM-13_34453
LogAggregationType: AGGREGATED
```

Answer : The element /assets/js/lightbox.js has 297 hits

(c) Log-Analysis-3.java file has the source code for printing the website path that has the most hits (count the max URL) and the number of hits for that path.

To run the above code as a jar file in Spark with YARN as the scheduler

bin/spark-submit --class "com.pitt.cloudcomputing.MaxUrlCountEvaluator" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Log-Analysis-3*.jar

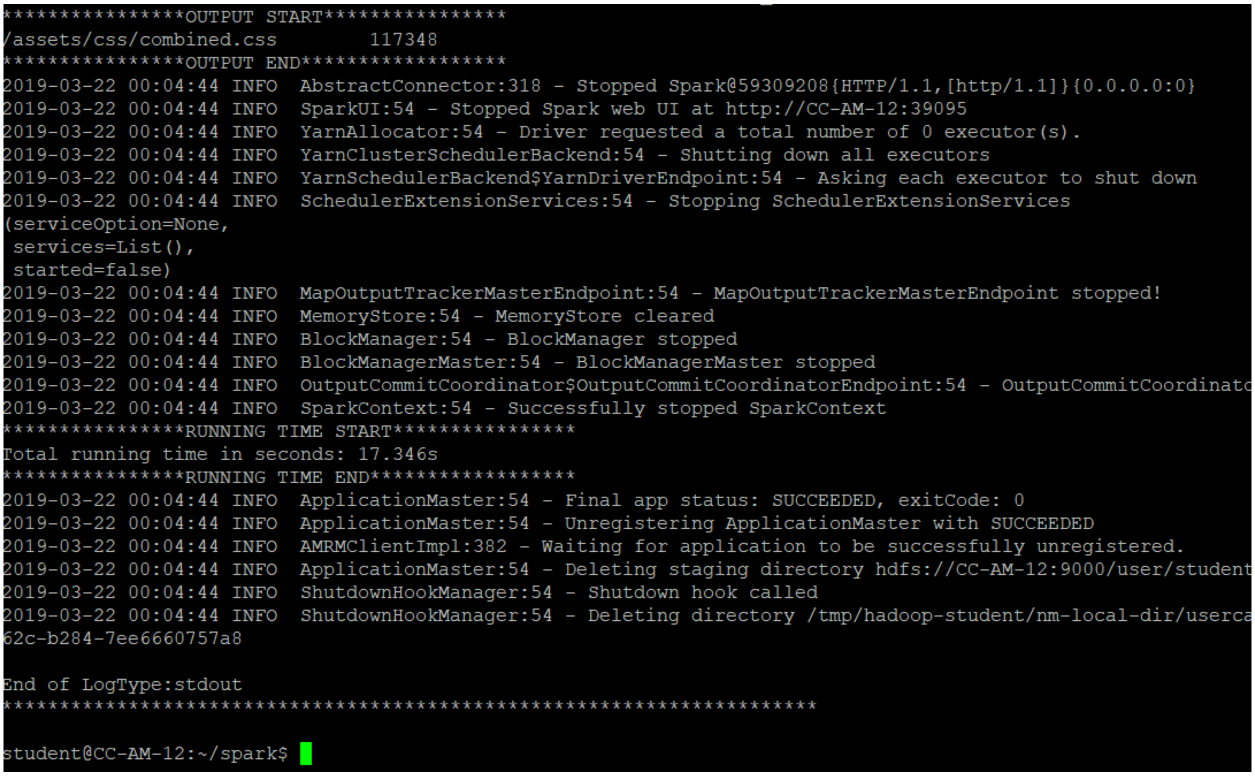


```
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-12
         ApplicationMaster RPC port: 42445
         queue: default
         start time: 1553213061672
         final status: SUCCEEDED
         tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0006/
         user: student
2019-03-22 00:04:44 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-22 00:04:44 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-0f397f27-5bc0-4270-8de4-8af25a65f59d
2019-03-22 00:04:44 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-3ed80b3e-4c90-4b0c-a326-c4491bf70ff5
student@CC-AM-12:~/spark$
```

The log file can be viewed here

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0006

The screenshot below gives the result of the program and the max hit URL and the count of hits.



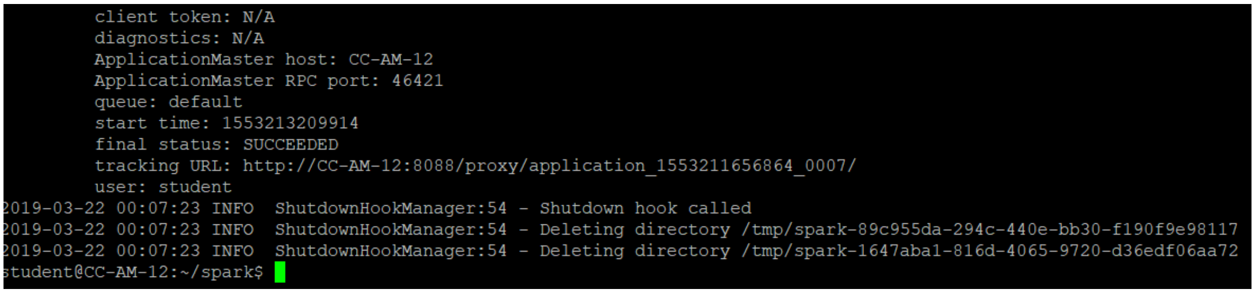
```
****************OUTPUT START*****************
/assets/css/combined.css        117348
****************OUTPUT END******************
2019-03-22 00:04:44 INFO  AbstractConnector:318 - Stopped Spark@59309208{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-22 00:04:44 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-12:39095
2019-03-22 00:04:44 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-22 00:04:44 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-22 00:04:44 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-22 00:04:44 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-22 00:04:44 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-22 00:04:44 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-22 00:04:44 INFO  BlockManager:54 - BlockManager stopped
2019-03-22 00:04:44 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-22 00:04:44 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinatc
2019-03-22 00:04:44 INFO  SparkContext:54 - Successfully stopped SparkContext
****************RUNNING TIME START****************
Total running time in seconds: 17.346s
****************RUNNING TIME END******************
2019-03-22 00:04:44 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-22 00:04:44 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-22 00:04:44 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-22 00:04:44 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-12:9000/user/student
2019-03-22 00:04:44 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-22 00:04:44 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/userca
62c-b284-7ee6660757a8

End of LogType:stdout
***********************************************************************

student@CC-AM-12:~/spark$
```

Answer : The website path with the max hits is /assets/css/combined.css and the count of the hits is 117348

(d) **Log-Analysis-4.java** file has the source code for finding the IP address with maximum accesses to the website and the number of accesses.

To run the above code as a jar file in Spark with YARN as the scheduler

bin/spark-submit --class "com.pitt.cloudcomputing.MaxIPCountEvaluator" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Log-Analysis-4*.jar

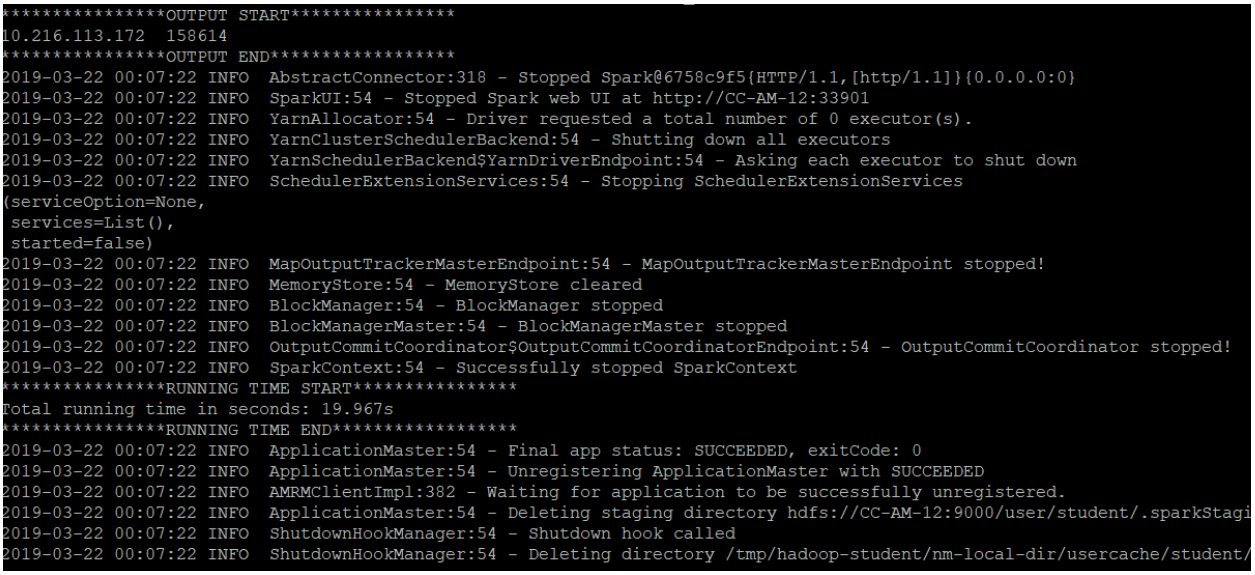


```
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-12
         ApplicationMaster RPC port: 46421
         queue: default
         start time: 1553213209914
         final status: SUCCEEDED
         tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0007/
         user: student
2019-03-22 00:07:23 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-22 00:07:23 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-89c955da-294c-440e-bb30-f190f9e98117
2019-03-22 00:07:23 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-1647aba1-816d-4065-9720-d36edf06aa72
student@CC-AM-12:~/spark$
```

The log file can be viewed here

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0007

The screenshot below gives the result of the program and the IP with maximum accesses and the count of the accesses to the website from that IP



```
****************OUTPUT START****************
10.216.113.172  158614
****************OUTPUT END******************
2019-03-22 00:07:22 INFO  AbstractConnector:318 - Stopped Spark@6758c9f5{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-22 00:07:22 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-12:33901
2019-03-22 00:07:22 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-22 00:07:22 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-22 00:07:22 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-22 00:07:22 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-22 00:07:22 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-22 00:07:22 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-22 00:07:22 INFO  BlockManager:54 - BlockManager stopped
2019-03-22 00:07:22 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-22 00:07:22 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-22 00:07:22 INFO  SparkContext:54 - Successfully stopped SparkContext
***************RUNNING TIME START****************
Total running time in seconds: 19.967s
***************RUNNING TIME END******************
2019-03-22 00:07:22 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-22 00:07:22 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-22 00:07:22 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-22 00:07:22 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-12:9000/user/student/.sparkStagi
2019-03-22 00:07:22 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-22 00:07:22 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/
```

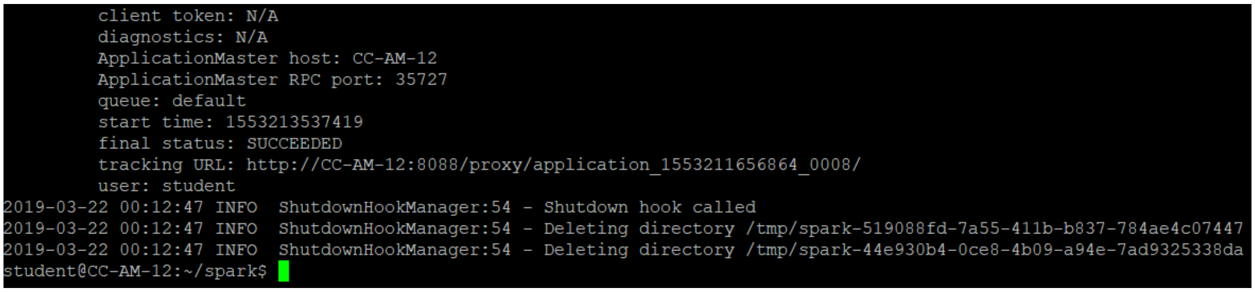
Answer : The IP with maximum accesses to the website is 10.216.113.172 and the number of accesses is 158614.

Part 3 – second task – to analyze and report the performance numbers of the program with and without cached RDD

With cached RDD : Log-Analysis-With-Cache.java file has the source code for printing the hit count for two keys (two website elements) with cached RDD. The file is loaded once and all the keys are counted and transferred to the RDD.

To run the above code as a jar file in Spark with YARN as the scheduler

bin/spark-submit --class "com.pitt.cloudcomputing.KeyCountEvaluatorWithCache" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Log-Analysis-With*.jar

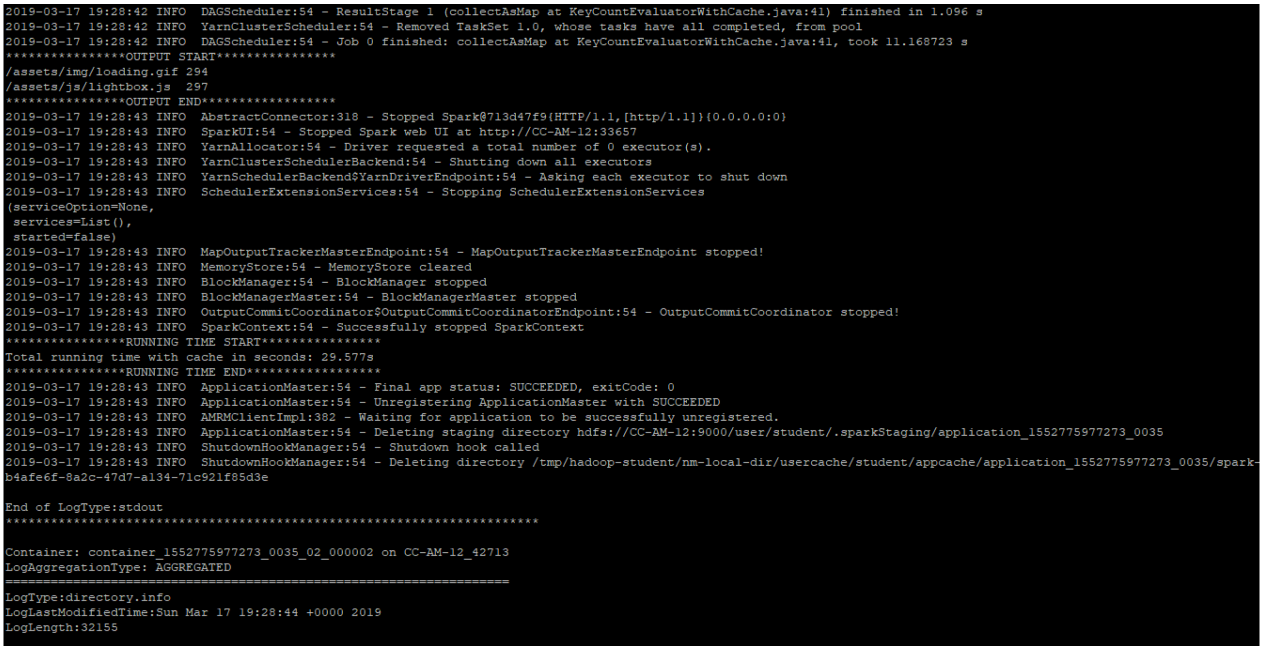


```
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-12
         ApplicationMaster RPC port: 35727
         queue: default
         start time: 1553213537419
         final status: SUCCEEDED
         tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0008/
         user: student
2019-03-22 00:12:47 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-22 00:12:47 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-519088fd-7a55-411b-b837-784ae4c07447
2019-03-22 00:12:47 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-44e930b4-0ce8-4b09-a94e-7ad9325338da
student@CC-AM-12:~/spark$
```

The log file can be viewed here

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0008

The screenshot below gives the result of the program and the total run time of the program to count the number of hits for both the keys.



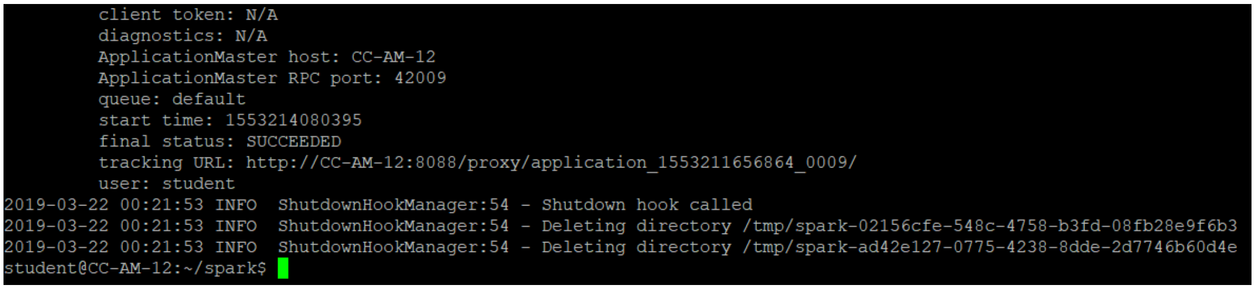
```
2019-03-17 19:28:42 INFO  DAGScheduler:54 - ResultStage 1 (collectAsMap at KeyCountEvaluatorWithCache.java:41) finished in 1.096 s
2019-03-17 19:28:42 INFO  YarnClusterScheduler:54 - Removed TaskSet 1.0, whose tasks have all completed, from pool
2019-03-17 19:28:42 INFO  DAGScheduler:54 - Job 0 finished: collectAsMap at KeyCountEvaluatorWithCache.java:41, took 11.168723 s
****************OUTPUT START****************
/assets/img/loading.gif 294
/assets/js/lightbox.js  297
****************OUTPUT END*****************
2019-03-17 19:28:43 INFO  AbstractConnector:318 - Stopped Spark@713d47f9{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-17 19:28:43 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-12:33657
2019-03-17 19:28:43 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-17 19:28:43 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-17 19:28:43 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-17 19:28:43 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-17 19:28:43 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-17 19:28:43 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-17 19:28:43 INFO  BlockManager:54 - BlockManager stopped
2019-03-17 19:28:43 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-17 19:28:43 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-17 19:28:43 INFO  SparkContext:54 - Successfully stopped SparkContext
****************RUNNING TIME START****************
Total running time with cache in seconds: 29.577s
****************RUNNING TIME END*****************
2019-03-17 19:28:43 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-17 19:28:43 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-17 19:28:43 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-17 19:28:43 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-12:9000/user/student/.sparkStaging/application_1552775977273_0035
2019-03-17 19:28:43 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-17 19:28:43 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/appcache/application_1552775977273_0035/spark-
b4afe6f-8a2c-47d7-a134-71c921f85d3e

End of LogType:stdout
***********************************************************************

Container: container_1552775977273_0035_02_000002 on CC-AM-12_42713
LogAggregationType: AGGREGATED
====================================================================
LogType:directory.info
LogLastModifiedTime:Sun Mar 17 19:28:44 +0000 2019
LogLength:32155
```

Answer : The total running time with cached RDD is 29.577s

Without cached RDD : Log-Analysis-No-Cache.java file has the source code for printing the hit count for two keys (two website elements) without cached RDD. For each key, the file is loaded once and the hits are counted and transferred to the RDD. In our example, the file is loaded twice.

To run the above code as a jar file in Spark with YARN as the scheduler

bin/spark-submit --class "com.pitt.cloudcomputing.KeyCountEvaluatorWithoutCache" --master yarn --deploy-mode cluster --driver-memory 1g --executor-memory 1g --executor-cores 1 --queue default jarfiles/Log-Analysis-No*.jar

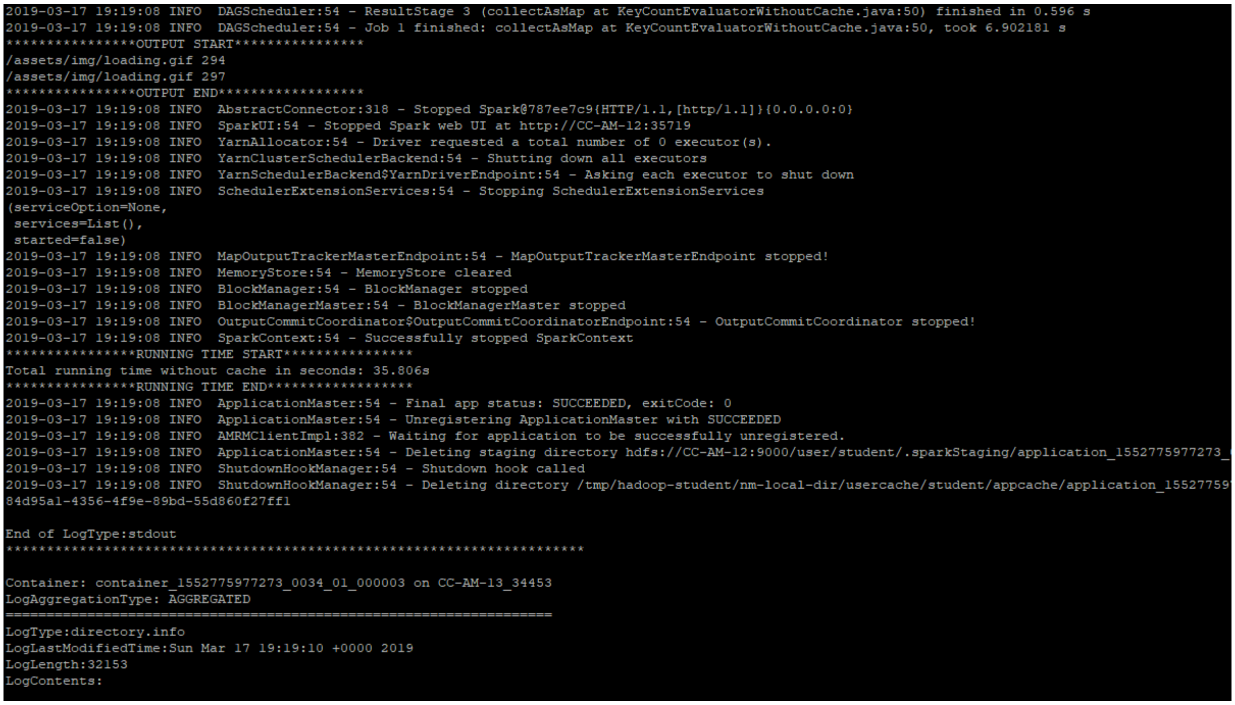


```
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: CC-AM-12
         ApplicationMaster RPC port: 42009
         queue: default
         start time: 1553214080395
         final status: SUCCEEDED
         tracking URL: http://CC-AM-12:8088/proxy/application_1553211656864_0009/
         user: student
2019-03-22 00:21:53 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-22 00:21:53 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-02156cfe-548c-4758-b3fd-08fb28e9f6b3
2019-03-22 00:21:53 INFO  ShutdownHookManager:54 - Deleting directory /tmp/spark-ad42e127-0775-4238-8dde-2d7746b60d4e
student@CC-AM-12:~/spark$
```

The log file can be viewed here

~/hadoop/bin/yarn logs -applicationId application_1553211656864_0009

The screenshot below gives the result of the program and the total run time of the program to count the number of hits for both the keys.

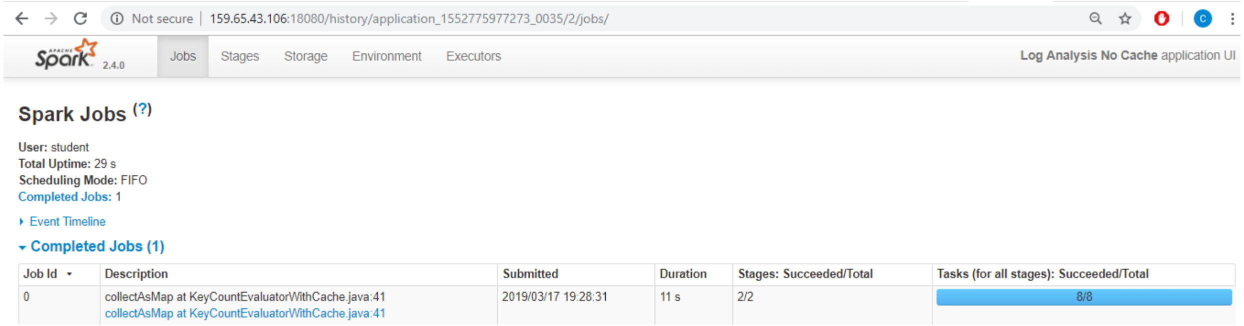


```
2019-03-17 19:19:08 INFO  DAGScheduler:54 - ResultStage 3 (collectAsMap at KeyCountEvaluatorWithoutCache.java:50) finished in 0.596 s
2019-03-17 19:19:08 INFO  DAGScheduler:54 - Job 1 finished: collectAsMap at KeyCountEvaluatorWithoutCache.java:50, took 6.902181 s
****************OUTPUT START****************
/assets/img/loading.gif 294
/assets/img/loading.gif 297
****************OUTPUT END******************
2019-03-17 19:19:08 INFO  AbstractConnector:318 - Stopped Spark@787ee7c9{HTTP/1.1,[http/1.1]}{0.0.0.0:0}
2019-03-17 19:19:08 INFO  SparkUI:54 - Stopped Spark web UI at http://CC-AM-12:35719
2019-03-17 19:19:08 INFO  YarnAllocator:54 - Driver requested a total number of 0 executor(s).
2019-03-17 19:19:08 INFO  YarnClusterSchedulerBackend:54 - Shutting down all executors
2019-03-17 19:19:08 INFO  YarnSchedulerBackend$YarnDriverEndpoint:54 - Asking each executor to shut down
2019-03-17 19:19:08 INFO  SchedulerExtensionServices:54 - Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2019-03-17 19:19:08 INFO  MapOutputTrackerMasterEndpoint:54 - MapOutputTrackerMasterEndpoint stopped!
2019-03-17 19:19:08 INFO  MemoryStore:54 - MemoryStore cleared
2019-03-17 19:19:08 INFO  BlockManager:54 - BlockManager stopped
2019-03-17 19:19:08 INFO  BlockManagerMaster:54 - BlockManagerMaster stopped
2019-03-17 19:19:08 INFO  OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:54 - OutputCommitCoordinator stopped!
2019-03-17 19:19:08 INFO  SparkContext:54 - Successfully stopped SparkContext
****************RUNNING TIME START****************
Total running time without cache in seconds: 35.806s
****************RUNNING TIME END******************
2019-03-17 19:19:08 INFO  ApplicationMaster:54 - Final app status: SUCCEEDED, exitCode: 0
2019-03-17 19:19:08 INFO  ApplicationMaster:54 - Unregistering ApplicationMaster with SUCCEEDED
2019-03-17 19:19:08 INFO  AMRMClientImpl:382 - Waiting for application to be successfully unregistered.
2019-03-17 19:19:08 INFO  ApplicationMaster:54 - Deleting staging directory hdfs://CC-AM-12:9000/user/student/.sparkStaging/application_1552775977273_
2019-03-17 19:19:08 INFO  ShutdownHookManager:54 - Shutdown hook called
2019-03-17 19:19:08 INFO  ShutdownHookManager:54 - Deleting directory /tmp/hadoop-student/nm-local-dir/usercache/student/appcache/application_15527759
84d95a1-4356-4f9e-89bd-55d860f27ff1

End of LogType:stdout
***********************************************************************

Container: container_1552775977273_0034_01_000003 on CC-AM-13_34453
LogAggregationType: AGGREGATED
====================================================================
LogType:directory.info
LogLastModifiedTime:Sun Mar 17 19:19:10 +0000 2019
LogLength:32153
LogContents:
```

Answer : The total running time without cached RDD is 35.806s

The program with cached RDD was able to run in 29.577s while the one without cached RDD takes 35.806s. If we consider only the total run time, with cached RDD we get a speedup of 1.2X for the whole program. The performance measurement was only done with one run and multiple runs will give us a more robust speedup number.

The screenshot below gives the output of the jobs summary web portal and the time for the sub-tasks of the program. With cached RDD, even though the total run time is 29s, the time for the text processing is 11s with cached file and output the result for the desired keys.


Not secure | 159.65.43.106:18080/history/application_1552775977273_0035/2/jobs/
Spark 2.4.0
Jobs
Stages
Storage
Environment
Executors
Log Analysis No Cache application UI

Spark Jobs (?)

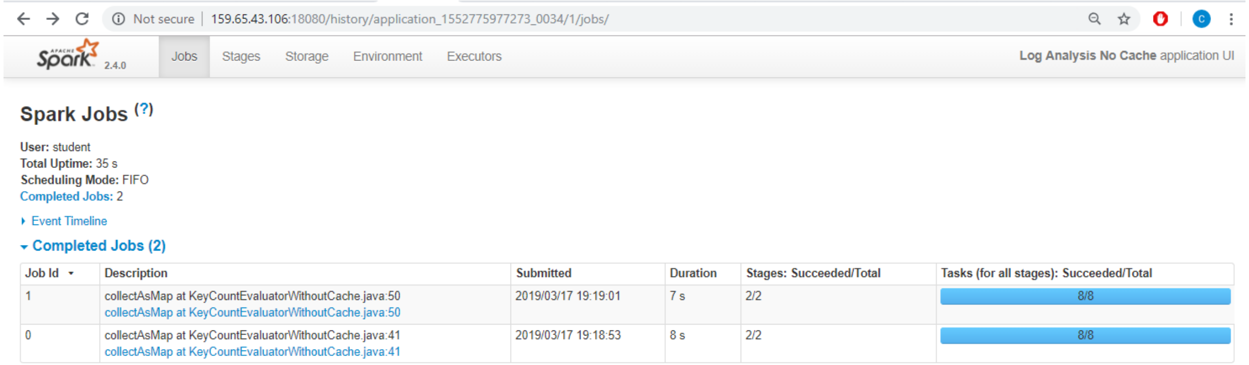
User: student
Total Uptime: 29 s
Scheduling Mode: FIFO
Completed Jobs: 1

Event Timeline

Completed Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	collectAsMap at KeyCountEvaluatorWithCache.java:41 collectAsMap at KeyCountEvaluatorWithCache.java:41	2019/03/17 19:28:31	11 s	2/2	8/8

The screenshot below gives the output of the jobs summary web portal and the time for the sub-tasks of the program. Without cached RDD, even though the total run time is 35s, the time for the text processing is 7s and 8s for the two separate sub tasks.


Not secure | 159.65.43.106:18080/history/application_1552775977273_0034/1/jobs/
Spark 2.4.0
Jobs
Stages
Storage
Environment
Executors
Log Analysis No Cache application UI

Spark Jobs (?)

User: student
Total Uptime: 35 s
Scheduling Mode: FIFO
Completed Jobs: 2

Event Timeline

Completed Jobs (2)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	collectAsMap at KeyCountEvaluatorWithoutCache.java:50 collectAsMap at KeyCountEvaluatorWithoutCache.java:50	2019/03/17 19:19:01	7 s	2/2	8/8
0	collectAsMap at KeyCountEvaluatorWithoutCache.java:41 collectAsMap at KeyCountEvaluatorWithoutCache.java:41	2019/03/17 19:18:53	8 s	2/2	8/8

The performance results show the importance of the use of cached RDD in caching the results of the intermediate tasks and use when for repetitive tasks, thereby speeding up the overall program.