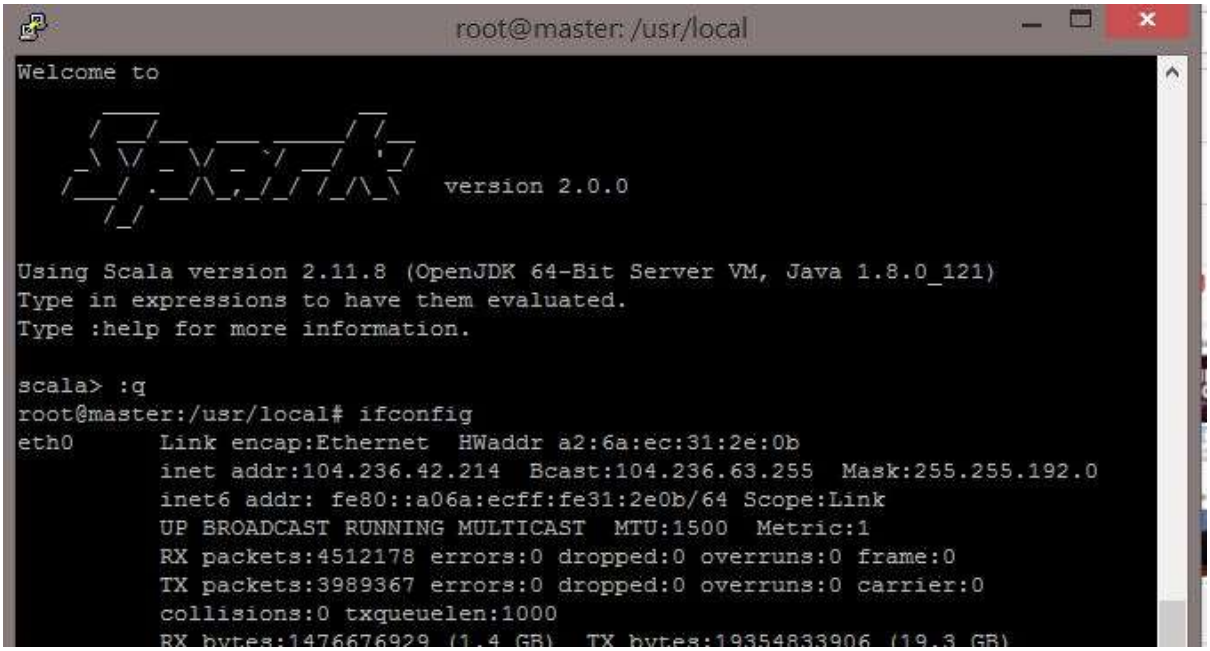
Sai charan Talipineni
sat122@pitt.edu

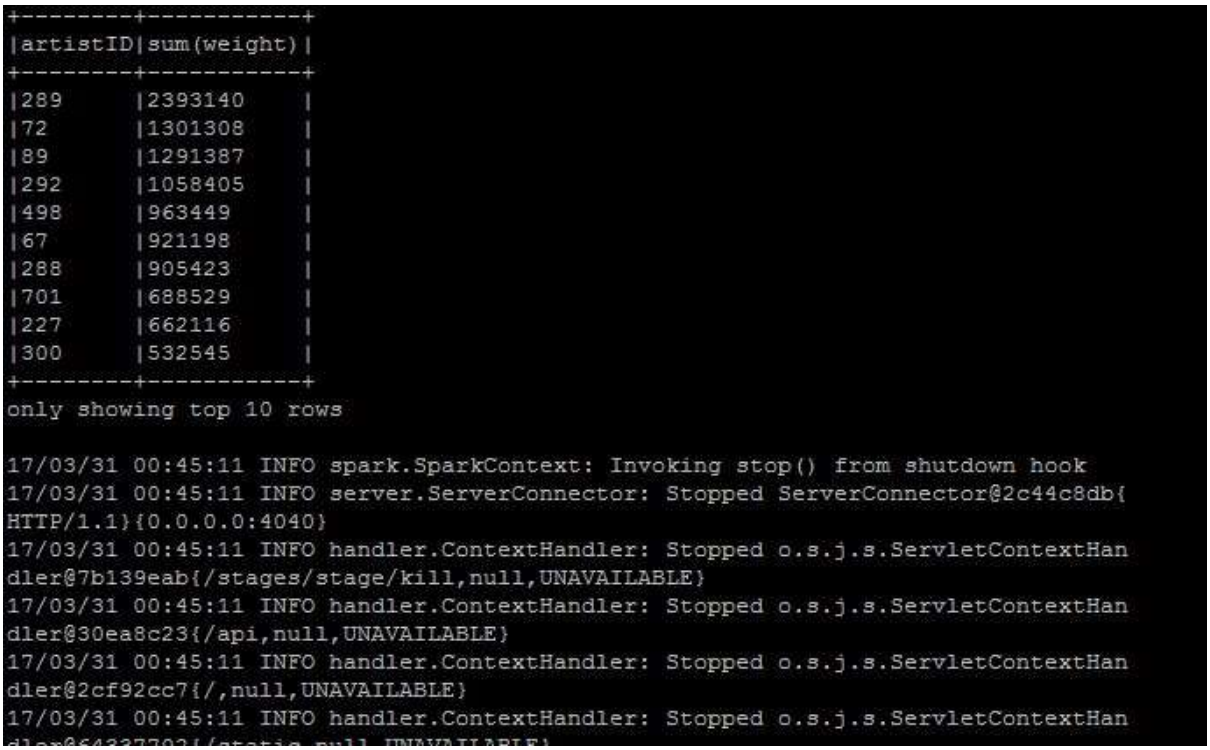
Sai Rakesh Ghanta
sag163@pitt.edu

Mini Project 2

1- Spark Shell



2- Task 2



Sai charan Talipineni
sat122@pitt.edu

Sai Rakesh Ghanta
sag163@pitt.edu

```
|288     |905423     |
|701     |688529     |
|227     |662116     |
|300     |532545     |
+--------+-----------+
only showing top 10 rows

17/03/31 00:45:11 INFO spark.SparkContext: Invoking stop() from shutdown hook
17/03/31 00:45:11 INFO server.ServerConnector: Stopped ServerConnector@2c44c8db{
HTTP/1.1}{0.0.0.0:4040}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@7b139eab{/stages/stage/kill,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@30ea8c23{/api,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@2cf92cc7{/,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@64337702{/static,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@76c7beb3{/executors/threadDump/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@751d3241{/executors/threadDump,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@3f2ef586{/executors/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@756cf158{/executors,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@2f058b8a{/environment/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@6175619b{/environment,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@165b8a71{/storage/rdd/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@7fd4acee{/storage/rdd,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@5b970f7{/storage/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@1a1da881{/storage,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@119f1f2a{/stages/pool/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@1a1da881{/storage,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@119f1f2a{/stages/pool/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@3d08f3f5{/stages/pool,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@55795845{/stages/stage/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@4fe89c24{/stages/stage,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@5dbe30be{/stages/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@89c10b7{/stages,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@72bd06ca{/jobs/job/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@625e134e{/jobs/job,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@5cc126dc{/jobs/json,null,UNAVAILABLE}
17/03/31 00:45:11 INFO handler.ContextHandler: Stopped o.s.j.s.ServletContextHan
dler@7e809b79{/jobs,null,UNAVAILABLE}
17/03/31 00:45:11 INFO ui.SparkUI: Stopped Spark web UI at http://104.236.42.214
:4040
17/03/31 00:45:11 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor
thread
```

Sai charan Talipineni
sat122@pitt.edu

Sai Rakesh Ghanta
sag163@pitt.edu

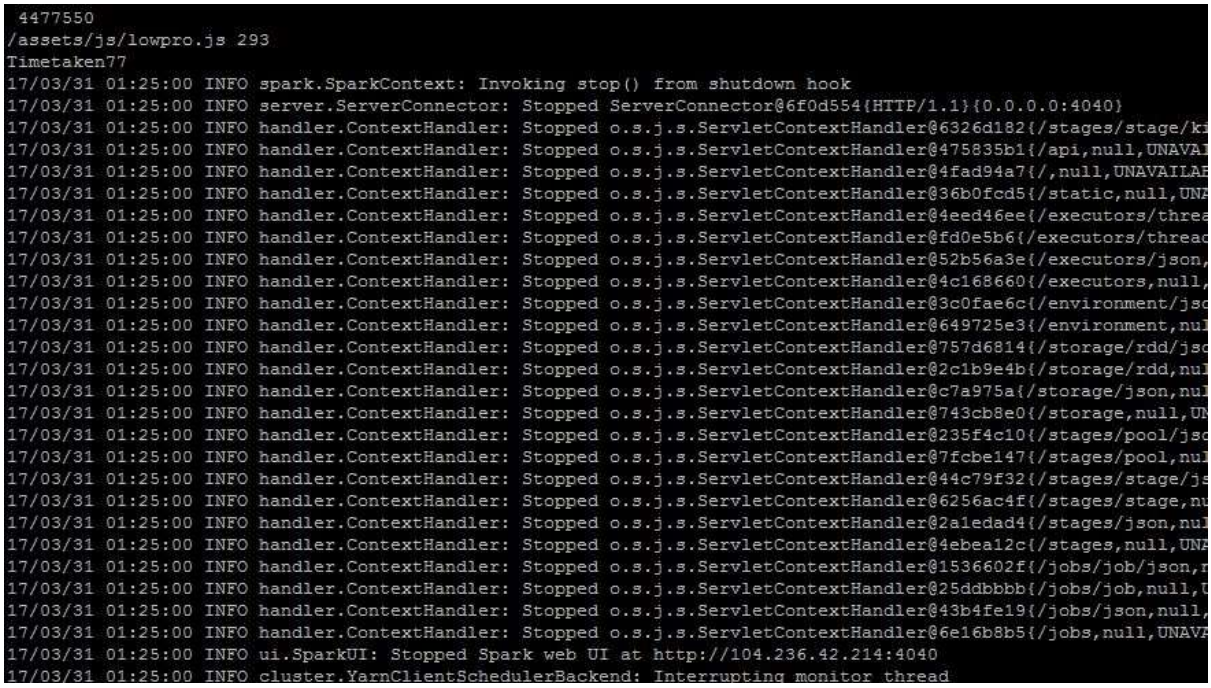
3- Task 3:

Note: Time is in Seconds

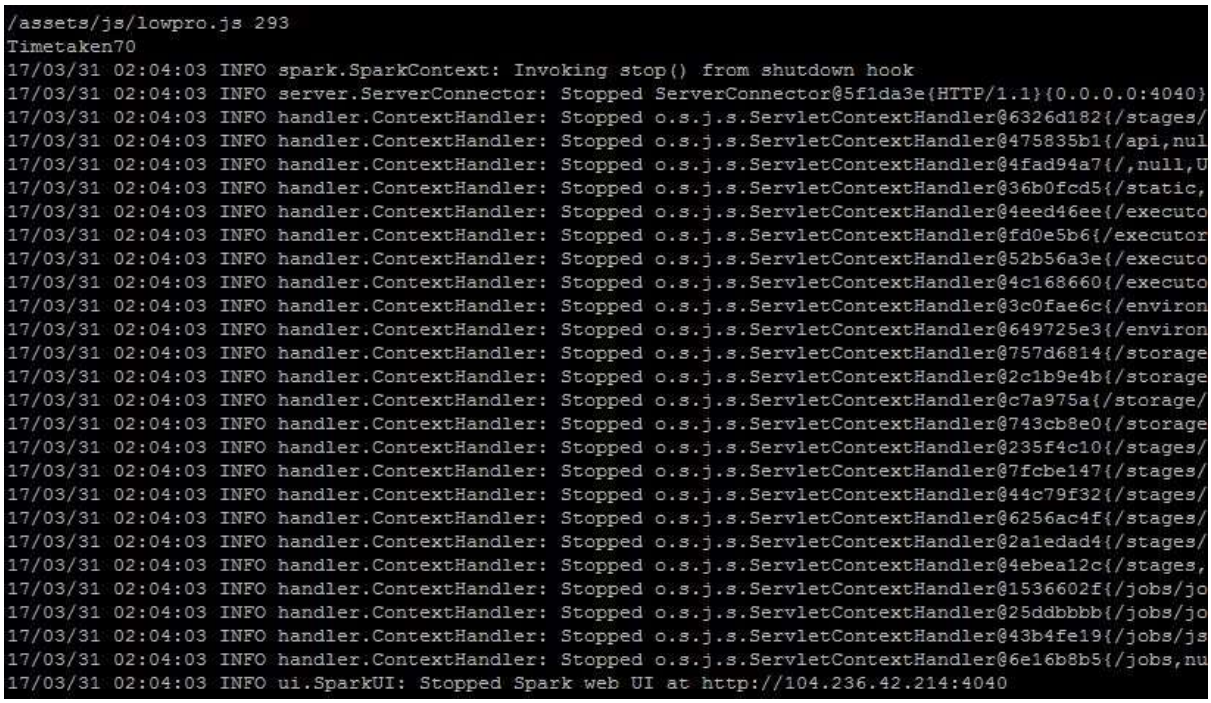
Without using cached RDD

a. How many hits were made to the website item “/assets/js/lowpro.js”?

293



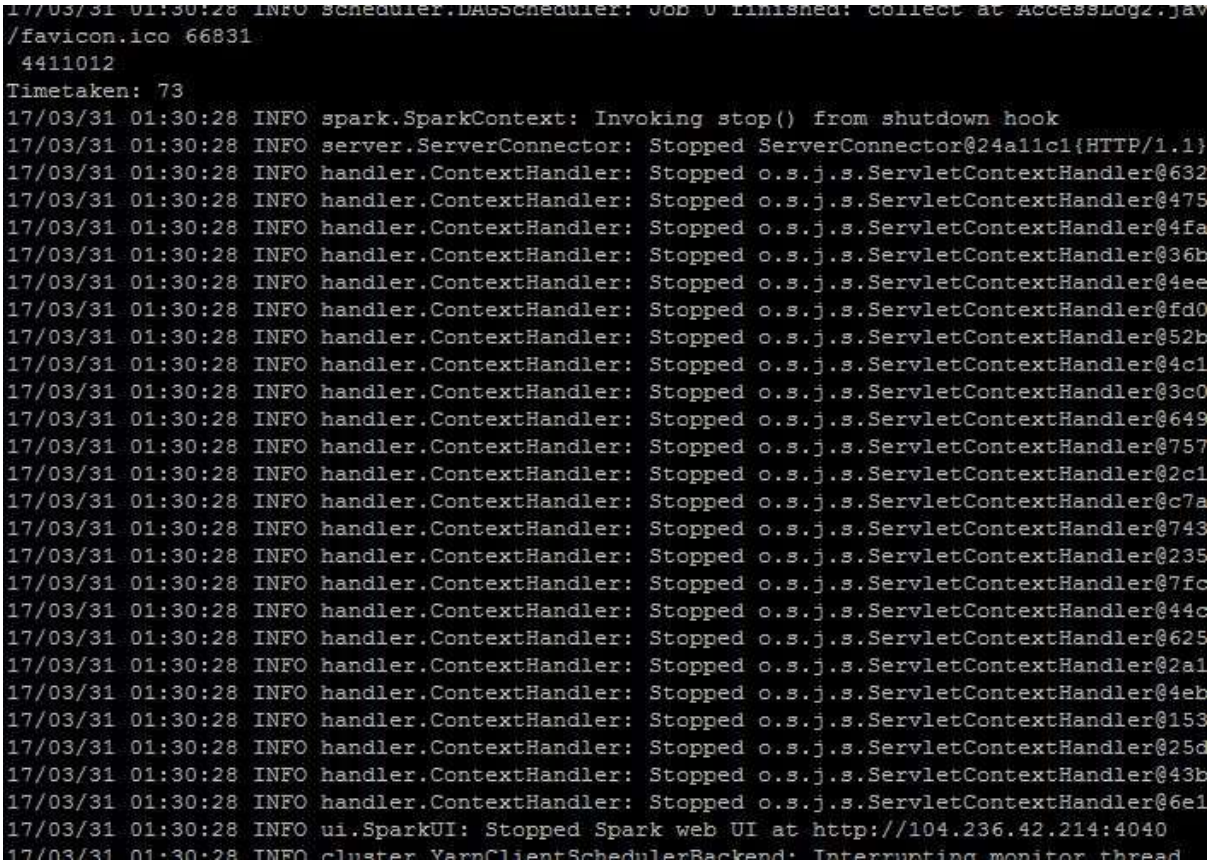
With cache:



Sai charan Talipineni
sat122@pitt.edu

Sai Rakesh Ghanta
sag163@pitt.edu

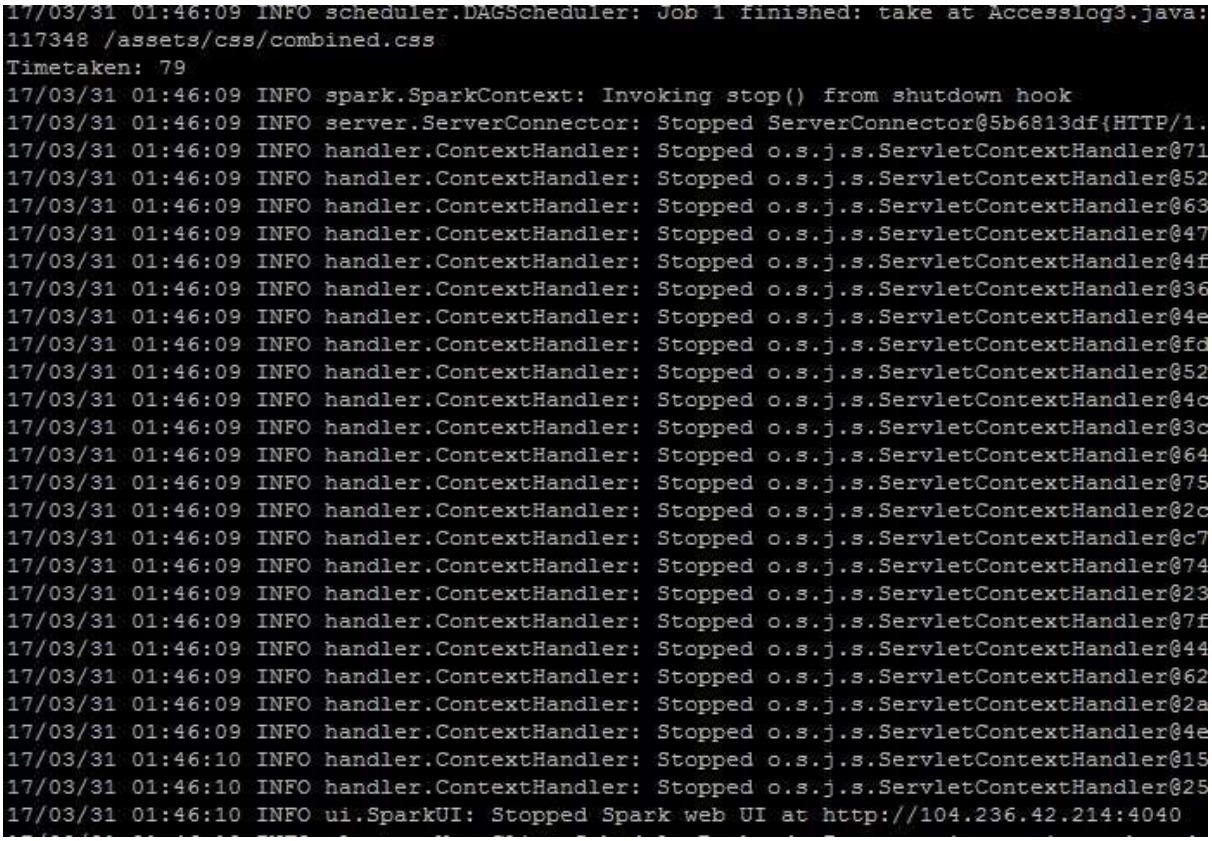
b. How many hits were made to the website item “/favicon.ico”?
66831



Sai charan Talipineni
sat122@pitt.edu

Sai Rakesh Ghanta
sag163@pitt.edu

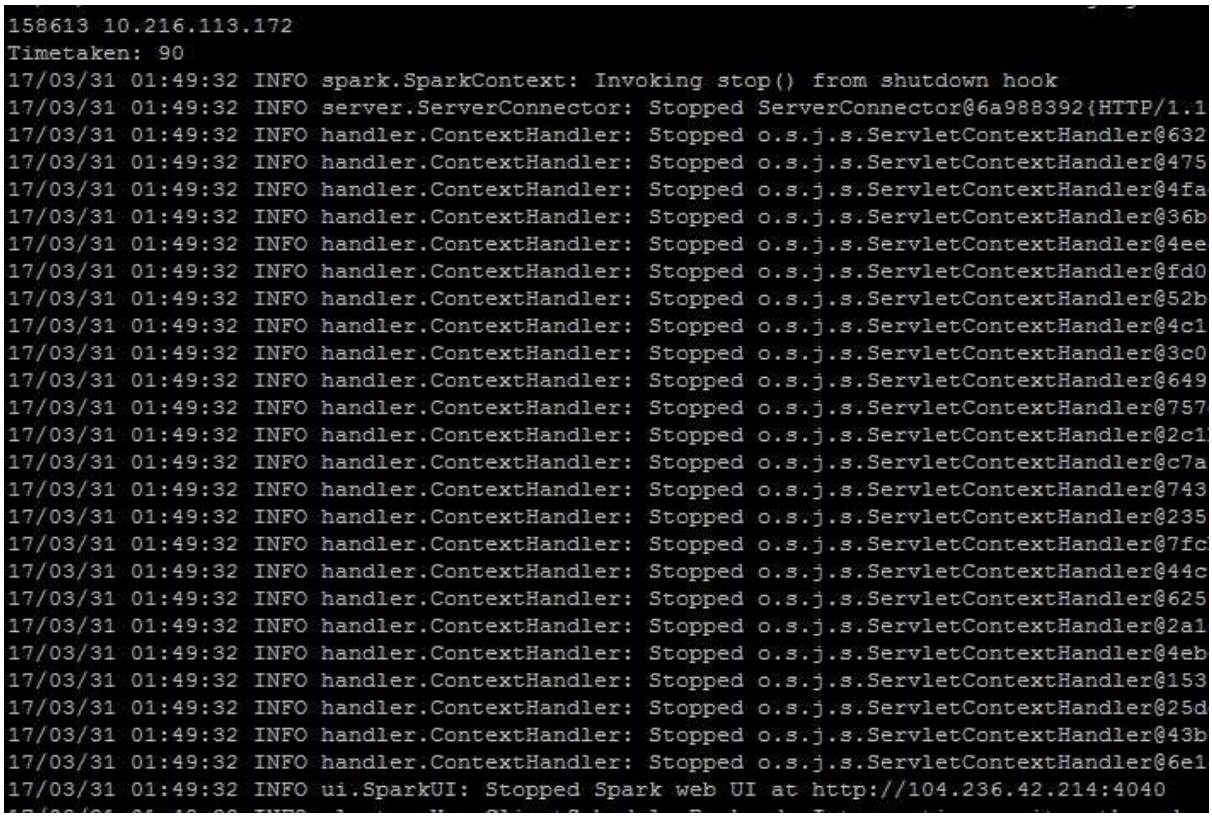
c. Which path in the website has been hit most? How many hits were made to the path?



Sai charan Talipineni
sat122@pitt.edu

Sai Rakesh Ghanta
sag163@pitt.edu

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

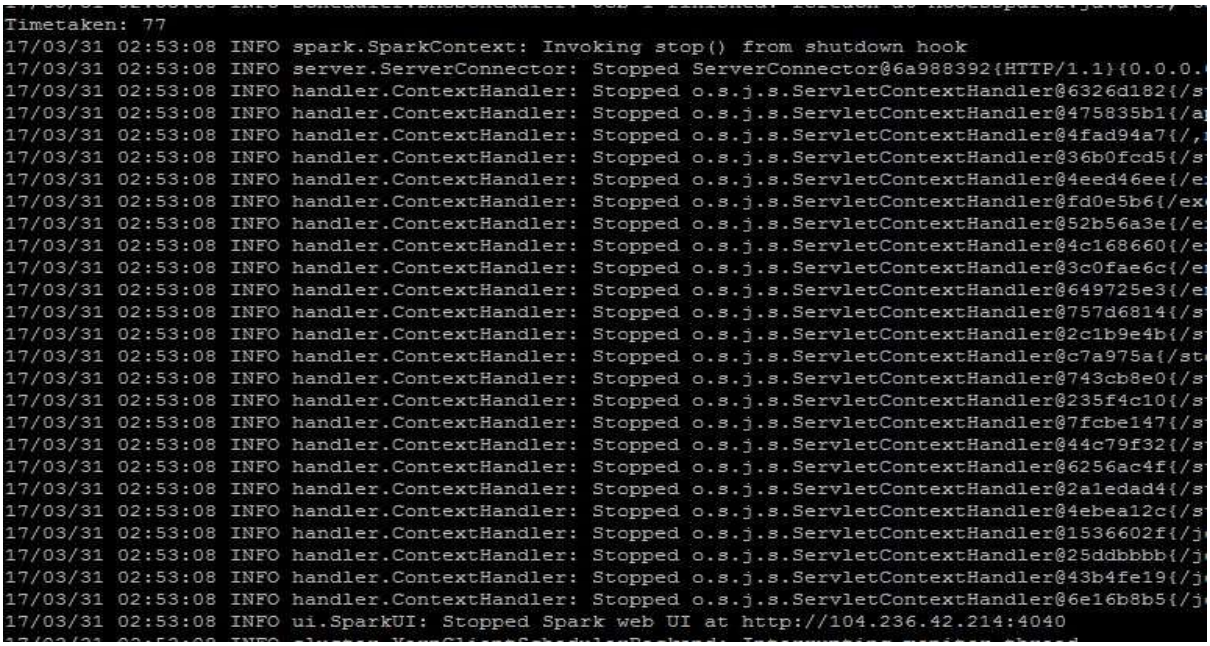


Task 3: part 2:

Loading the file every time to calculate counts

It took more time. I found the time difference of 15 seconds. I forgot to take the screen shot for the first time I ran it. When I tried it second time it is giving memory issue. So, I am unable to put screen shot for this one only.

Loading once and filtering in the middle



Sai charan Talipineni
sat122@pitt.edu

Sai Rakesh Ghanta
sag163@pitt.edu

Note: Readme file is in zip folder.

References:

http://spark.apache.org/docs/latest/programming-guide.html

http://spark.apache.org/examples.html

https://spark.apache.org/docs/0.9.1/java-programming-guide.html

http://www.javaworld.com/article/2972863/big-data/open-source-java-projects-apache-spark.html

https://www.simple-talk.com/cloud/cloud-data/start-big-data-apache-spark/