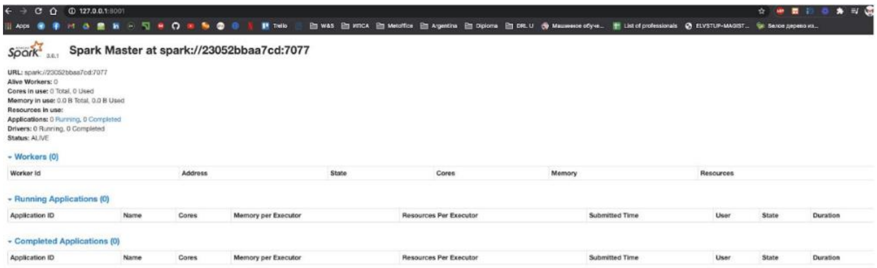
1) Install spark https://phoenixnap.com/kb/install-spark-on-ubuntu

2) start-master.sh --webui-port PORT - start spark app

3) stop-master.sh   - stop spark app



Spark Master at spark://23052bbaa7cd:7077

URL: spark://23052bbaa7cd:7077
Alive Workers: 0
Cores in use: 0 Total, 0 Used
Memory in use: 0.0 B Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

- Workers (0)

Worker Id	Address	State	Cores	Memory	Resources

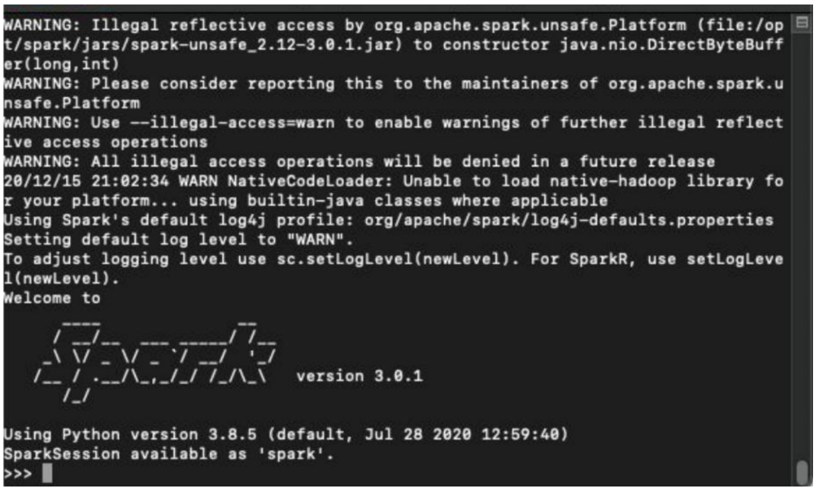
- Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

- Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration

4)

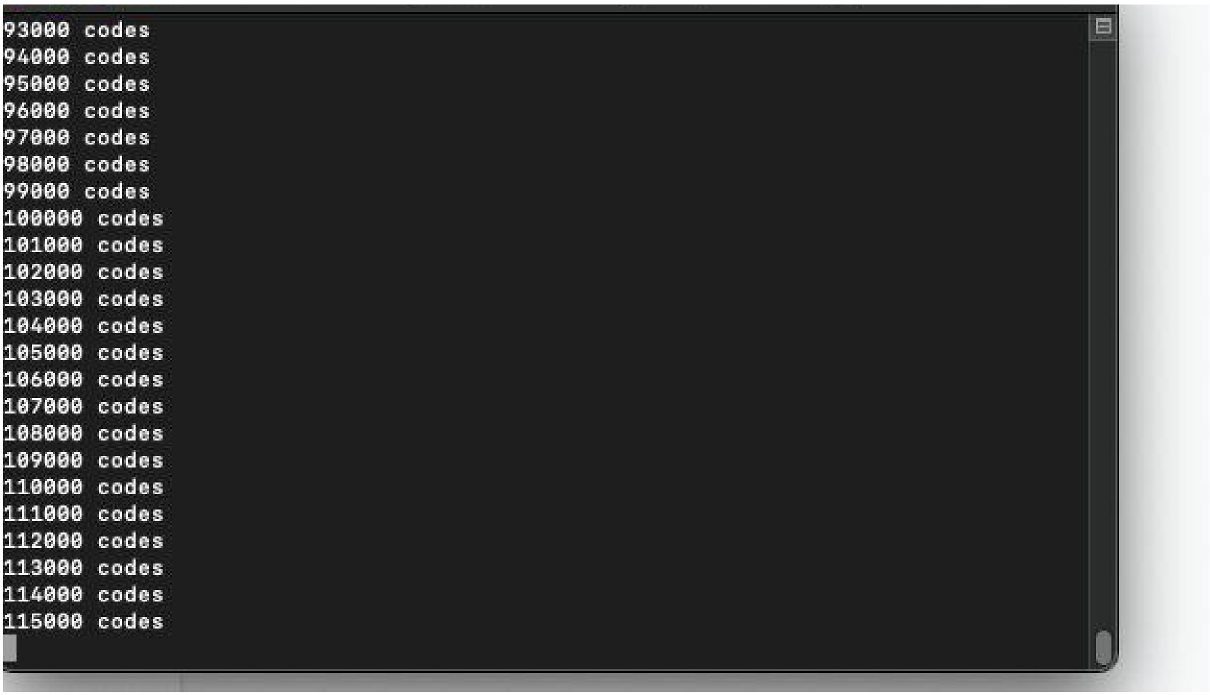


```
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/op
t/spark/jars/spark-unsafe_2.12-3.0.1.jar) to constructor java.nio.DirectByteBuff
er(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.u
nsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflect
ive access operations
WARNING: All illegal access operations will be denied in a future release
20/12/15 21:02:34 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.0.1
      /_/

Using Python version 3.8.5 (default, Jul 28 2020 12:59:40)
SparkSession available as 'spark'.
>>> 
```

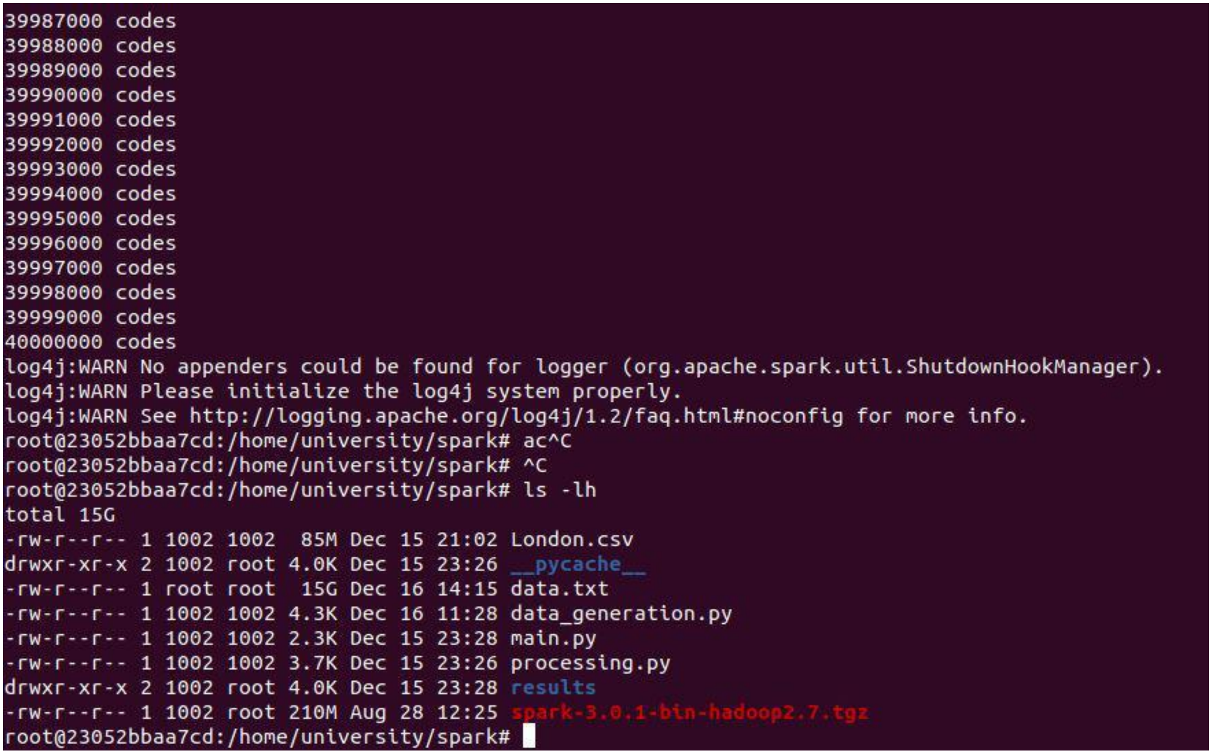
5) apt install python3-pip
6) pip3 install pyspark, geopy, faker, numpy
7) download London.csv
8) spark-submit data_generation.py

9)



```
93000 codes
94000 codes
95000 codes
96000 codes
97000 codes
98000 codes
99000 codes
100000 codes
101000 codes
102000 codes
103000 codes
104000 codes
105000 codes
106000 codes
107000 codes
108000 codes
109000 codes
110000 codes
111000 codes
112000 codes
113000 codes
114000 codes
115000 codes
```

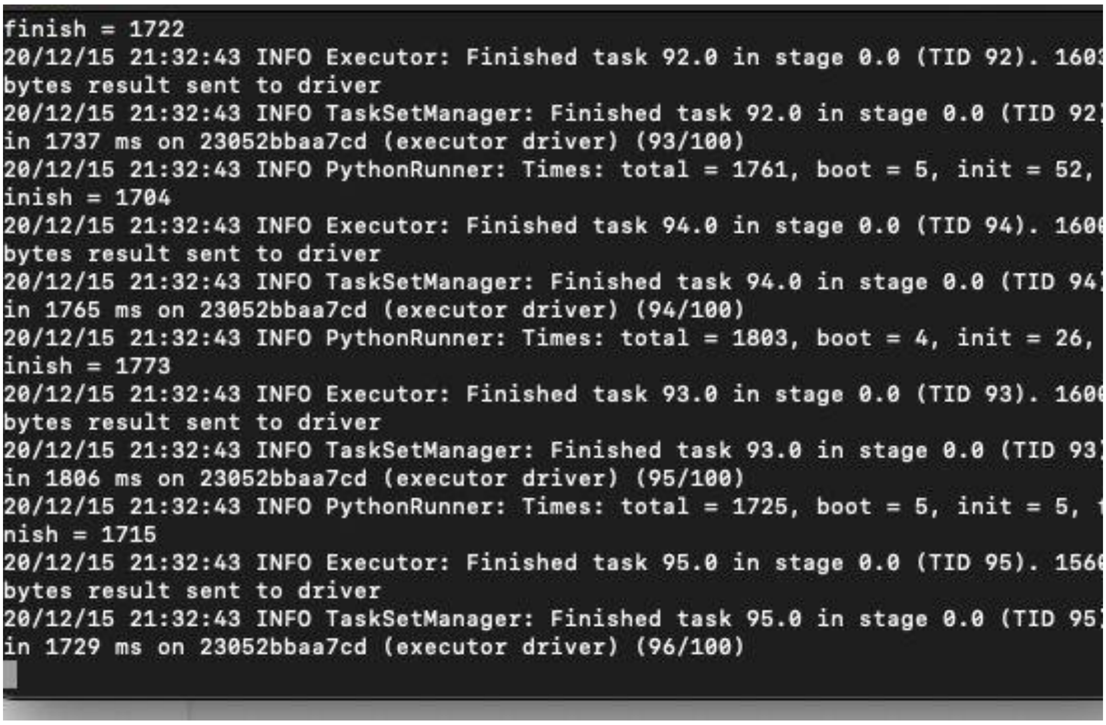
10)



```
39987000 codes
39988000 codes
39989000 codes
39990000 codes
39991000 codes
39992000 codes
39993000 codes
39994000 codes
39995000 codes
39996000 codes
39997000 codes
39998000 codes
39999000 codes
40000000 codes
log4j:WARN No appenders could be found for logger (org.apache.spark.util.ShutdownHookManager).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
root@23052bbaa7cd:/home/university/spark# ac^C
root@23052bbaa7cd:/home/university/spark# ^C
root@23052bbaa7cd:/home/university/spark# ls -lh
total 15G
-rw-r--r-- 1 1002 1002  85M Dec 15 21:02 London.csv
drwxr-xr-x 2 1002 root 4.0K Dec 15 23:26 __pycache__
-rw-r--r-- 1 root root  15G Dec 16 14:15 data.txt
-rw-r--r-- 1 1002 1002 4.3K Dec 16 11:28 data_generation.py
-rw-r--r-- 1 1002 1002 2.3K Dec 15 23:28 main.py
-rw-r--r-- 1 1002 1002 3.7K Dec 15 23:26 processing.py
drwxr-xr-x 2 1002 root 4.0K Dec 15 23:28 results
-rw-r--r-- 1 1002 root 210M Aug 28 12:25 spark-3.0.1-bin-hadoop2.7.tgz
root@23052bbaa7cd:/home/university/spark# 
```

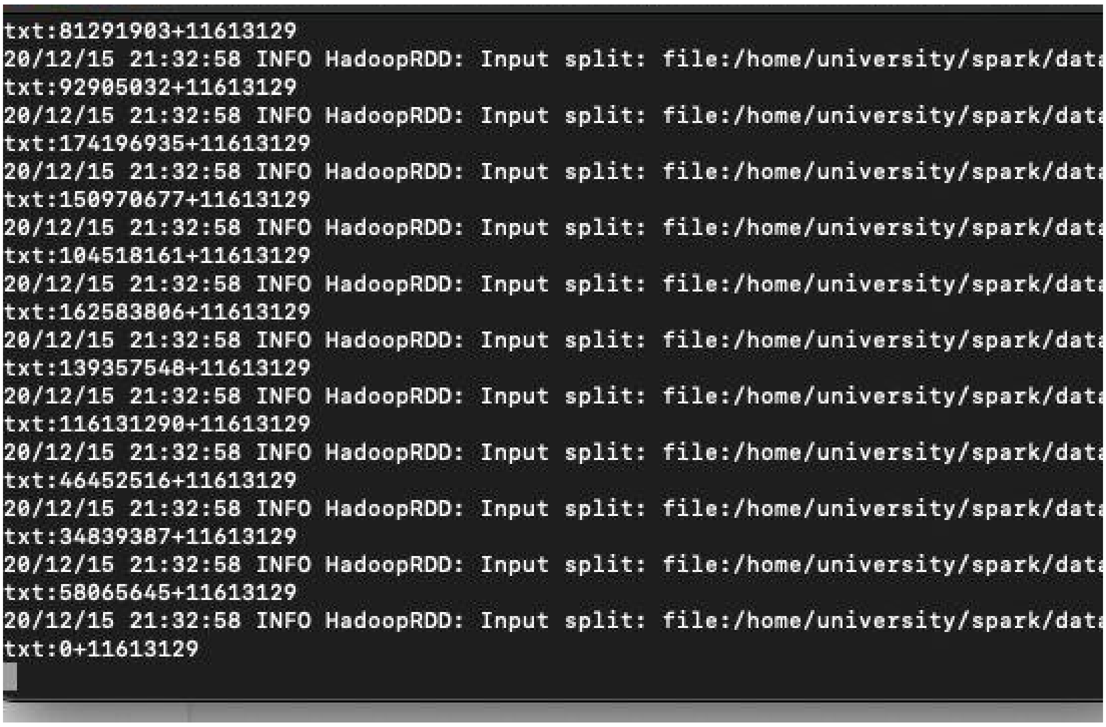
11) spark-submit main.py

12)



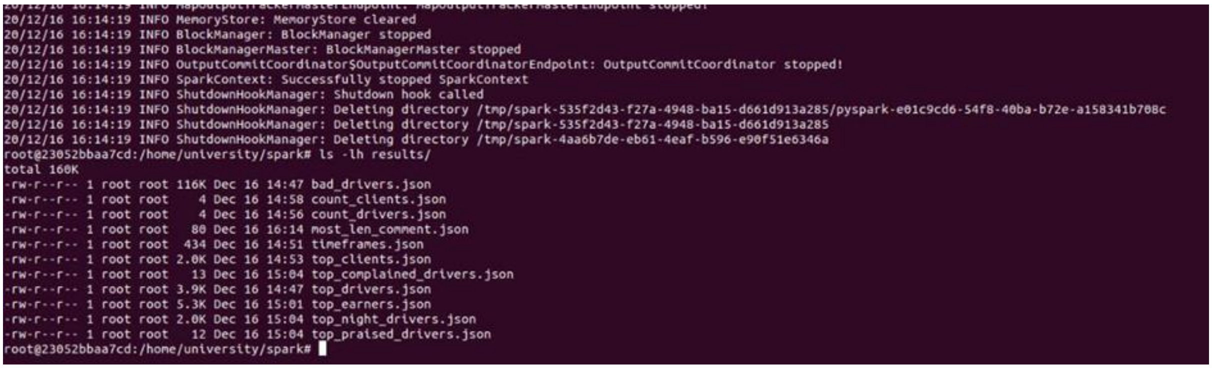
```
finish = 1722
20/12/15 21:32:43 INFO Executor: Finished task 92.0 in stage 0.0 (TID 92). 1603
bytes result sent to driver
20/12/15 21:32:43 INFO TaskSetManager: Finished task 92.0 in stage 0.0 (TID 92)
in 1737 ms on 23052bbaa7cd (executor driver) (93/100)
20/12/15 21:32:43 INFO PythonRunner: Times: total = 1761, boot = 5, init = 52,
inish = 1704
20/12/15 21:32:43 INFO Executor: Finished task 94.0 in stage 0.0 (TID 94). 1600
bytes result sent to driver
20/12/15 21:32:43 INFO TaskSetManager: Finished task 94.0 in stage 0.0 (TID 94)
in 1765 ms on 23052bbaa7cd (executor driver) (94/100)
20/12/15 21:32:43 INFO PythonRunner: Times: total = 1803, boot = 4, init = 26,
inish = 1773
20/12/15 21:32:43 INFO Executor: Finished task 93.0 in stage 0.0 (TID 93). 1600
bytes result sent to driver
20/12/15 21:32:43 INFO TaskSetManager: Finished task 93.0 in stage 0.0 (TID 93)
in 1806 ms on 23052bbaa7cd (executor driver) (95/100)
20/12/15 21:32:43 INFO PythonRunner: Times: total = 1725, boot = 5, init = 5, f
nish = 1715
20/12/15 21:32:43 INFO Executor: Finished task 95.0 in stage 0.0 (TID 95). 1560
bytes result sent to driver
20/12/15 21:32:43 INFO TaskSetManager: Finished task 95.0 in stage 0.0 (TID 95)
in 1729 ms on 23052bbaa7cd (executor driver) (96/100)
```

13)



```
txt:81291903+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:92905032+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:174196935+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:150970677+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:104518161+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:162583806+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:139357548+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:116131290+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:46452516+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:34839387+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:58065645+11613129
20/12/15 21:32:58 INFO HadoopRDD: Input split: file:/home/university/spark/data
txt:0+11613129
```

14)



```
20/12/16 16:14:19 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
20/12/16 16:14:19 INFO MemoryStore: MemoryStore cleared
20/12/16 16:14:19 INFO BlockManager: BlockManager stopped
20/12/16 16:14:19 INFO BlockManagerMaster: BlockManagerMaster stopped
20/12/16 16:14:19 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/12/16 16:14:19 INFO SparkContext: Successfully stopped SparkContext
20/12/16 16:14:19 INFO ShutdownHookManager: Shutdown hook called
20/12/16 16:14:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-535f2d43-f27a-4948-ba15-d661d913a285/pyspark-e01c9cd6-54f8-40ba-b72e-a158341b708c
20/12/16 16:14:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-535f2d43-f27a-4948-ba15-d661d913a285
20/12/16 16:14:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-4aa6b7de-eb61-4eaf-b596-e90f51e6346a
root@23052bbaa7cd:/home/university/spark# ls -lh results/
total 160K
-rw-r--r-- 1 root root 116K Dec 16 14:47 bad_drivers.json
-rw-r--r-- 1 root root    4 Dec 16 14:58 count_clients.json
-rw-r--r-- 1 root root    4 Dec 16 14:56 count_drivers.json
-rw-r--r-- 1 root root   80 Dec 16 16:14 most_len_comment.json
-rw-r--r-- 1 root root  434 Dec 16 14:51 timeframes.json
-rw-r--r-- 1 root root 2.0K Dec 16 14:53 top_clients.json
-rw-r--r-- 1 root root   13 Dec 16 15:04 top_complained_drivers.json
-rw-r--r-- 1 root root 3.9K Dec 16 14:47 top_drivers.json
-rw-r--r-- 1 root root 5.3K Dec 16 15:01 top_earners.json
-rw-r--r-- 1 root root 2.0K Dec 16 15:04 top_night_drivers.json
-rw-r--r-- 1 root root   12 Dec 16 15:04 top_praised_drivers.json
root@23052bbaa7cd:/home/university/spark# 
```

15) bad_drivers.json - 2 variant

16) top_drivers.json - 1 variant
17) top_night_drivers.json - 6 variant
18) top_clients.json - 4 variant
19) top_earners.json - 5 variant
20) top_complained_drivers.json - 8 variant
21) top_praised_drivers.json - 7 variant
22) timeframes.json - 3 variant
23) most_len_comment.json - 9 variant