Diane Margo, Michael Tang, Jerry Trieu, Ocean Lu
Professor Yang
CS 4650
12/10/2019

Capstone Project: ReadMe

# Introduction

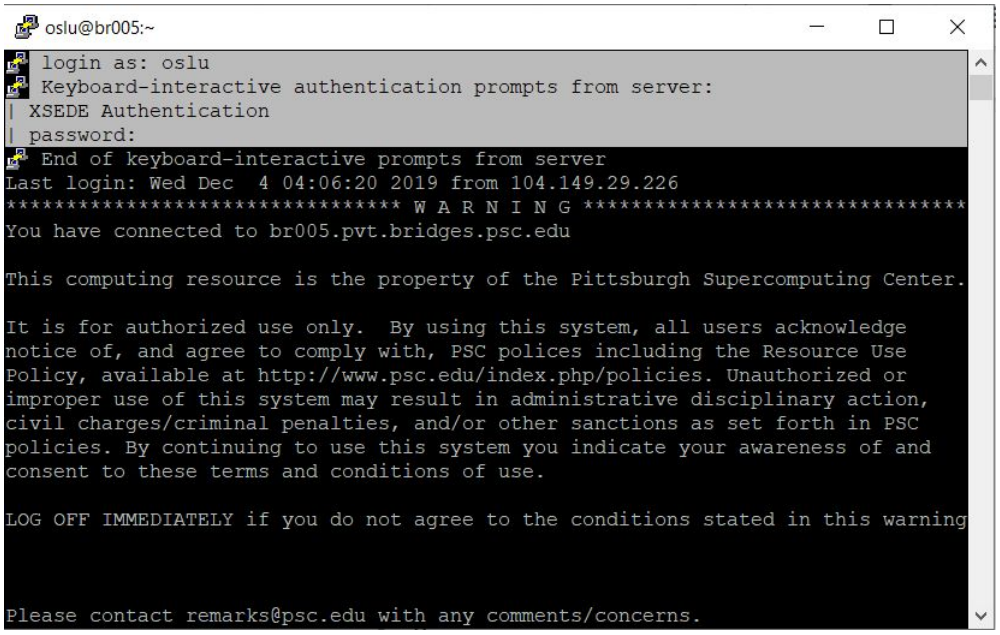
We are going to be utilizing the LA Parking Citation data by implementing Hadoop on XSEDE. Within the dataset are multiple columns filled with information such as frequent violation codes, locations, car body style/make, ticket frequency, the top time to be ticketed, etc. We are going to be finding correlations between these data columns and turning this data into useful information that can be utilized by the general public.

The programming model MapReduce is something that we will be using, as we will make a mapper program to obtain key-pair values from the dataset; after obtaining the key-pair values, we will be using the reducer program to get the output that we need. Essentially, the output we desire involves the frequency of the data in each of the desired columns, from greatest to least. We will analyze the data and label/mark important information that is relevant to the general public.

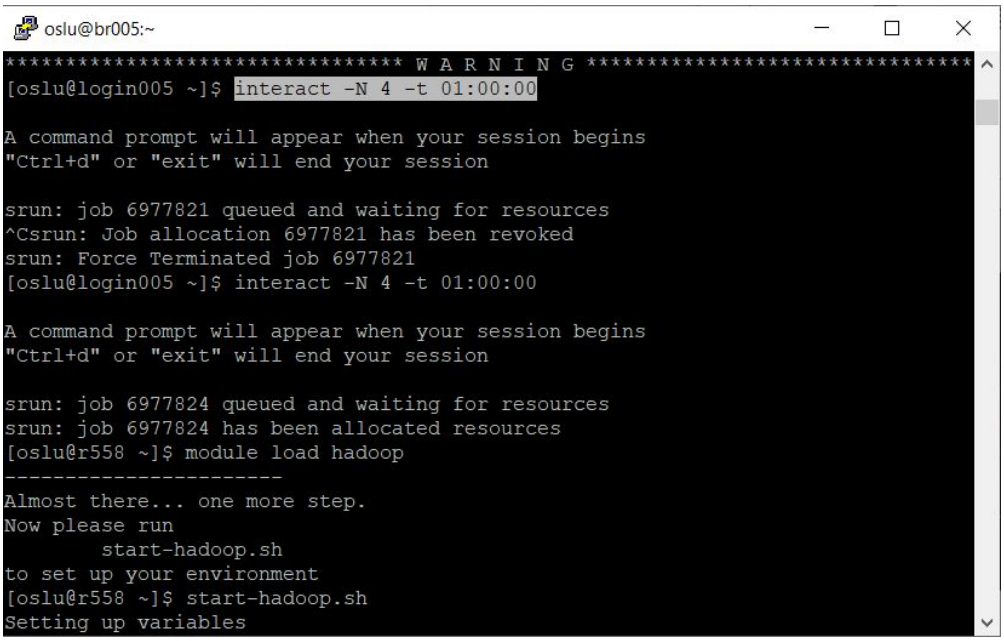
The procedure is very simple. First, prepare the data--we received our data from kaggle. Upload CSV file to HDFS. Write Mapper and Reducer in Python. Mapper and Reducer are specific to the type and kind of data being categorized and compiled. Change access permission on Mapper and Reducer and run the Map Reduce on Hadoop.

# Hadoop Specific Instructions with Screenshots

1. Ssh into bridges.psc.edu, with the port 2222




2. To start hadoop use the following commands:
   a. interact -N 4 -t 01:00:00



b. module load hadoop

```
oslu@br005:~
[oslu@r558 ~]$ module load hadoop
------------------------
Almost there... one more step.
Now please run
        start-hadoop.sh
to set up your environment
[oslu@r558 ~]$ start-hadoop.sh
Setting up variables
no command line arguments, using slurm vars
testing if we are a slave
slavelist is r566 r567 r649
i is r566
host is r558
i is r567
host is r558
i is r649
host is r558
HOSTDIR is /pylon5/ac5fq2p/oslu/6977824/r558
Loading modules
populating hadoop/spark directories
modifying config files
setting up daemons
path is /opt/packages/java/jdk1.8.0_211/bin:/opt/packages/java/jdk1.8.0_211/jre/
bin:/opt/packages/hive/hive/bin/:/opt/packages/hbase/hbase/bin/:/opt/packages/ha
```

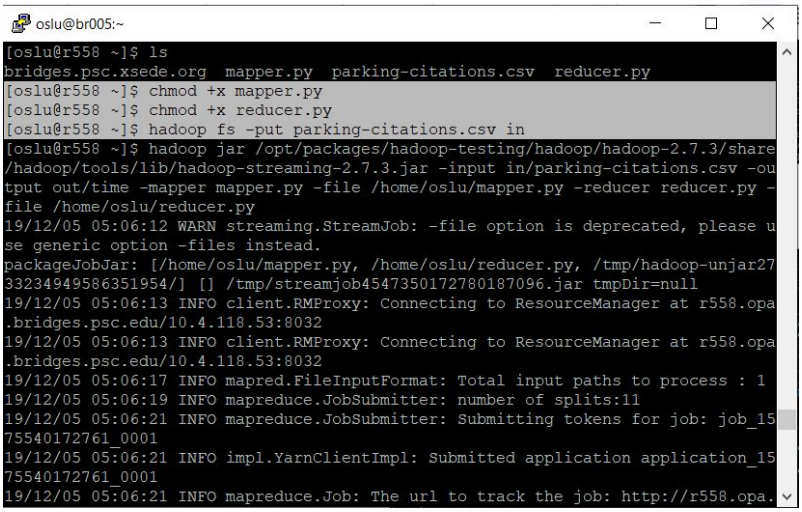
c. start-hadoop.sh

```
oslu@br005:~
[oslu@r558 ~]$ start-hadoop.sh
Setting up variables
no command line arguments, using slurm vars
testing if we are a slave
slavelist is r566 r567 r649
i is r566
host is r558
i is r567
host is r558
i is r649
host is r558
HOSTDIR is /pylon5/ac5fq2p/oslu/6977824/r558
Loading modules
populating hadoop/spark directories
modifying config files
setting up daemons
path is /opt/packages/java/jdk1.8.0_211/bin:/opt/packages/java/jdk1.8.0_211/jre/
bin:/opt/packages/hive/hive/bin/:/opt/packages/hbase/hbase/bin/:/opt/packages/ha
doop-testing/hadoop/hadoop/sbin/:/opt/packages/hadoop-testing/hadoop/hadoop/bin/
:/opt/packages/spark/default/bin/:/opt/packages/python/2_7_11_gcc/bin:/opt/packa
ges/xdusage/2.1-1/bin:/usr/lib64/qt-3.3/bin:/opt/intel/advisor_2019.5.0.602216/b
in64:/opt/intel/vtune_amplifier_2019.6.0.602217/bin64:/opt/intel/inspector_2019.
5.0.602103/bin64:/opt/intel/itac/2019.5.041/intel64/bin:/opt/intel/clck/2019.5/b
in/intel64:/opt/intel/compilers_and_libraries_2019.5.281/linux/bin/intel64:/opt/
```

3. Create a directory to store input files in the HDFS:
   a. hadoop fs -mkdir -p in

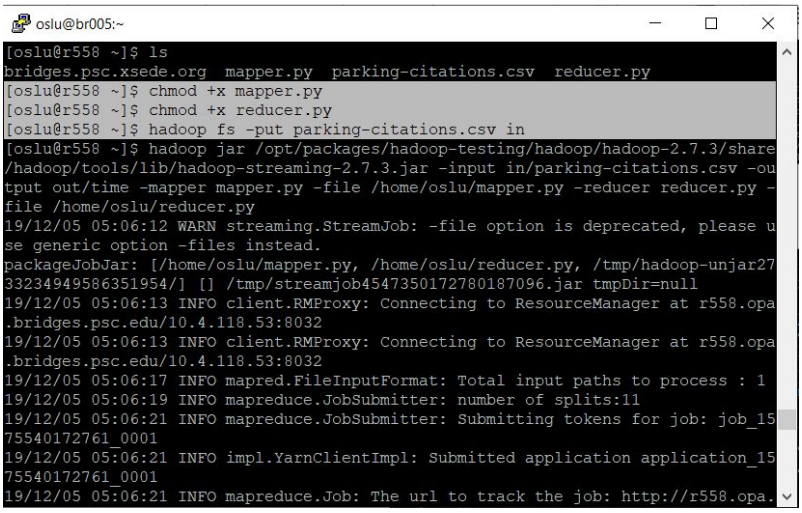
```
oslu@br005:~
r566: setting up daemons
r566: path is /opt/packages/python/2_7_11_gcc/bin:/opt/packages/java/jdk1.8.0_21
1/bin:/opt/packages/java/jdk1.8.0_211/jre/bin:/usr/lib64/qt-3.3/bin:/opt/package
s/xdusage/2.1-1/bin:/opt/intel/advisor_2019.5.0.602216/bin64:/opt/intel/vtune_am
plifier_2019.6.0.602217/bin64:/opt/intel/inspector_2019.5.0.602103/bin64:/opt/in
tel/itac/2019.5.041/intel64/bin:/opt/intel/clck/2019.5/bin/intel64:/opt/intel/co
mpilers_and_libraries_2019.5.281/linux/bin/intel64:/opt/intel/compilers_and_libr
aries_2019.5.281/linux/mpi/intel64/libfabric/bin:/opt/intel/compilers_and_librar
ies_2019.5.281/linux/mpi/intel64/bin:/opt/intel/debugger_2019/gdb/intel64/bin:/o
pt/packages/slurm/default/bin:/opt/packages/allocations:/opt/packages/interact/b
in:/usr/lib64/ccache:/usr/local/bin:/usr/bin:/opt/puppetlabs/puppet/bin:/opt/int
el/parallel_studio_xe_2019.5.075/bin:/sbin:/opt/puppetlabs/bin:/opt/package
s/slurm/default/bin:/opt/packages/hadoop-testing/hadoop/hadoop/bin:/opt/packages
/spark/default/bin
r566: Worker Node Detected, setting up
r566: starting nodemanager, logging to /pylon5/ac5fq2p/oslu/6977824/r566/logs/ya
rn-oslu-nodemanager-r566.pvt.bridges.psc.edu.out
r567: starting datanode, logging to /pylon5/ac5fq2p/oslu/6977824/r567/logs/hadoo
p-oslu-datanode-r567.pvt.bridges.psc.edu.out
r566: starting datanode, logging to /pylon5/ac5fq2p/oslu/6977824/r566/logs/hadoo
p-oslu-datanode-r566.pvt.bridges.psc.edu.out
[oslu@r558 ~]$ hadoop fs -mkdir -p in
[oslu@r558 ~]$ ls
bridges.psc.xsede.org  mapper.py  parking-citations.csv  part-00000  reducer.py
```

4. Load file to the HDFS storing it in the 'in' directory:
    a. hadoop fs -put parking-citations.csv in



```
oslu@br005:~
[oslu@r558 ~]$ ls
bridges.psc.xsede.org  mapper.py  parking-citations.csv  reducer.py
[oslu@r558 ~]$ chmod +x mapper.py
[oslu@r558 ~]$ chmod +x reducer.py
[oslu@r558 ~]$ hadoop fs -put parking-citations.csv in
[oslu@r558 ~]$ hadoop jar /opt/packages/hadoop-testing/hadoop/hadoop-2.7.3/share
/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input in/parking-citations.csv -ou
tput out/time -mapper mapper.py -file /home/oslu/mapper.py -reducer reducer.py -
file /home/oslu/reducer.py
19/12/05 05:06:12 WARN streaming.StreamJob: -file option is deprecated, please u
se generic option -files instead.
packageJobJar: [/home/oslu/mapper.py, /home/oslu/reducer.py, /tmp/hadoop-unjar27
33234949586351954/] [] /tmp/streamjob4547350172780187096.jar tmpDir=null
19/12/05 05:06:13 INFO client.RMProxy: Connecting to ResourceManager at r558.opa
.bridges.psc.edu/10.4.118.53:8032
19/12/05 05:06:13 INFO client.RMProxy: Connecting to ResourceManager at r558.opa
.bridges.psc.edu/10.4.118.53:8032
19/12/05 05:06:17 INFO mapred.FileInputFormat: Total input paths to process : 1
19/12/05 05:06:19 INFO mapreduce.JobSubmitter: number of splits:11
19/12/05 05:06:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
75540172761_0001
19/12/05 05:06:21 INFO impl.YarnClientImpl: Submitted application application_15
75540172761_0001
19/12/05 05:06:21 INFO mapreduce.Job: The url to track the job: http://r558.opa.
```

5. Give permissions
    a. chmod +x mapper.py
    b. chmod +x reducer.py



```
oslu@br005:~
[oslu@r558 ~]$ ls
bridges.psc.xsede.org  mapper.py  parking-citations.csv  reducer.py
[oslu@r558 ~]$ chmod +x mapper.py
[oslu@r558 ~]$ chmod +x reducer.py
[oslu@r558 ~]$ hadoop fs -put parking-citations.csv in
[oslu@r558 ~]$ hadoop jar /opt/packages/hadoop-testing/hadoop/hadoop-2.7.3/share
/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input in/parking-citations.csv -ou
tput out/time -mapper mapper.py -file /home/oslu/mapper.py -reducer reducer.py -
file /home/oslu/reducer.py
19/12/05 05:06:12 WARN streaming.StreamJob: -file option is deprecated, please u
se generic option -files instead.
packageJobJar: [/home/oslu/mapper.py, /home/oslu/reducer.py, /tmp/hadoop-unjar27
33234949586351954/] [] /tmp/streamjob4547350172780187096.jar tmpDir=null
19/12/05 05:06:13 INFO client.RMProxy: Connecting to ResourceManager at r558.opa
.bridges.psc.edu/10.4.118.53:8032
19/12/05 05:06:13 INFO client.RMProxy: Connecting to ResourceManager at r558.opa
.bridges.psc.edu/10.4.118.53:8032
19/12/05 05:06:17 INFO mapred.FileInputFormat: Total input paths to process : 1
19/12/05 05:06:19 INFO mapreduce.JobSubmitter: number of splits:11
19/12/05 05:06:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
75540172761_0001
19/12/05 05:06:21 INFO impl.YarnClientImpl: Submitted application application_15
75540172761_0001
19/12/05 05:06:21 INFO mapreduce.Job: The url to track the job: http://r558.opa.
```

6. Execute the program
    a. For example:
    hadoop jar
    /opt/packages/hadoop-testing/hadoop/hadoop-2.7.3/share/hadoop/tools/lib/hadoop
    -streaming-2.7.3.jar
    -input in/parking-citations.csv
    -output out/time -mapper mapper.py
    -file /home/oslu/mapper.py -reducer reducer.py

-file /home/oslu/reducer.py

```
oslu@br005:~
[oslu@r558 ~]$ hadoop jar /opt/packages/hadoop-testing/hadoop/hadoop-2.7.3/share
/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input in/parking-citations.csv -ou
tput out/time -mapper mapper.py -file /home/oslu/mapper.py -reducer reducer.py -
file /home/oslu/reducer.py
19/12/05 05:06:12 WARN streaming.StreamJob: -file option is deprecated, please u
se generic option -files instead.
packageJobJar: [/home/oslu/mapper.py, /home/oslu/reducer.py, /tmp/hadoop-unjar27
33234949586351954/] [] /tmp/streamjob4547350172780187096.jar tmpDir=null
19/12/05 05:06:13 INFO client.RMProxy: Connecting to ResourceManager at r558.opa
.bridges.psc.edu/10.4.118.53:8032
19/12/05 05:06:13 INFO client.RMProxy: Connecting to ResourceManager at r558.opa
.bridges.psc.edu/10.4.118.53:8032
19/12/05 05:06:17 INFO mapred.FileInputFormat: Total input paths to process : 1
19/12/05 05:06:19 INFO mapreduce.JobSubmitter: number of splits:11
19/12/05 05:06:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_15
75540172761_0001
19/12/05 05:06:21 INFO impl.YarnClientImpl: Submitted application application_15
75540172761_0001
19/12/05 05:06:21 INFO mapreduce.Job: The url to track the job: http://r558.opa.
bridges.psc.edu:8088/proxy/application_1575540172761_0001/
19/12/05 05:06:21 INFO mapreduce.Job: Running job: job_1575540172761_0001
19/12/05 05:07:04 INFO mapreduce.Job: Job job_1575540172761_0001 running in uber
 mode : false
19/12/05 05:07:04 INFO mapreduce.Job:  map 0% reduce 0%
```

7. Commands to see in/out directory:
   a. hdfs dfs -ls in/

```
[oslu@r558 ~]$ hdfs dfs -ls in/
Found 1 items
-rw-r--r--   2 oslu supergroup 1452001757 2019-12-05 05:06 in/parking-citations.
csv
[oslu@r558 ~]$
```

   b. hdfs dfs -ls out/

```
[oslu@r558 ~]$ hdfs dfs -ls out/time
Found 2 items
-rw-r--r--   2 oslu supergroup          0 2019-12-05 05:08 out/time/_SUCCESS
-rw-r--r--   2 oslu supergroup      16976 2019-12-05 05:08 out/time/part-00000
[oslu@r558 ~]$
```

8. Command to see the file outputted:
   a. hdfs dfs -cat out/time/part-00000
9. To retrieve the file and place into original hadoop directory

a. hadoop fs -get out/time/part-00000 /home/oslu

```
oslu@br005:~
                Reduce input records=9880902
                Reduce output records=2394
                Spilled Records=19761804
                Shuffled Maps =11
                Failed Shuffles=0
                Merged Map outputs=11
                GC time elapsed (ms)=604
                CPU time spent (ms)=59160
                Physical memory (bytes) snapshot=6002831360
                Virtual memory (bytes) snapshot=77213339648
                Total committed heap usage (bytes)=17901289472
        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=1452042717
        File Output Format Counters
                Bytes Written=16976
19/12/05 05:08:23 INFO streaming.StreamJob: Output directory: out/time
[oslu@r558 ~]$ hadoop fs -get out/time/part-00000 /home/oslu
```

10. Delete the out/ directory

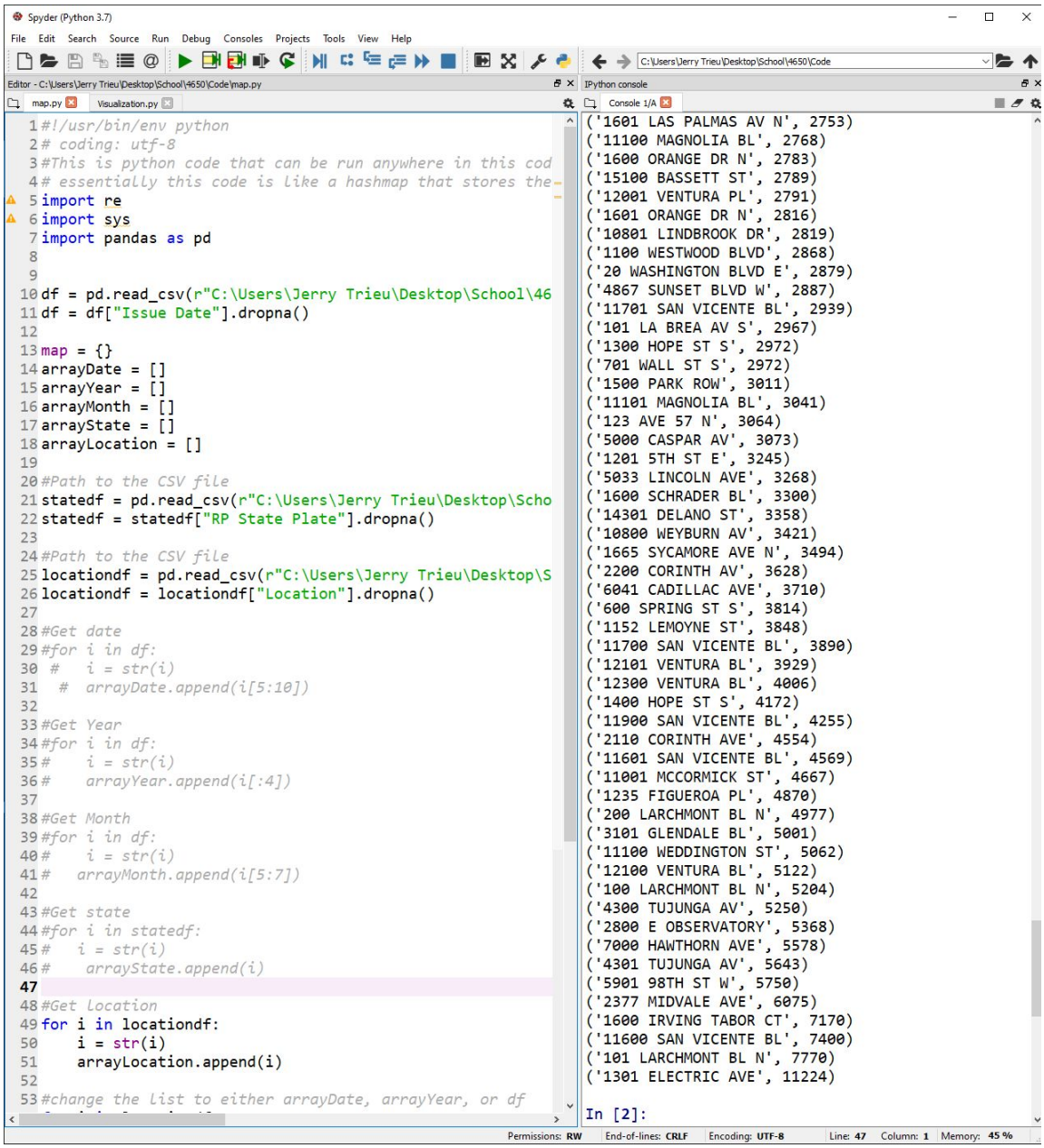
a. hdfs dfs -rm -r out/

```
[oslu@r558 ~]$ ^C
[oslu@r558 ~]$ hdfs dfs -rm -r out/
19/12/05 05:38:18 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted out
[oslu@r558 ~]$
```

How to Run map.py

In this python file it basically takes the total counts of a certain column and then sorts it

1) Open Spyder
2) Open up the map.py file
3) Run the file allow for it to run
4) Uncomment parts that you would want and comment out the parts you don't want and change the list to "arrayState", "arrayMonth", "arrayYear" etc.



How to run lacitiation.ipnb

1. Run cell by cell
2. Choose the file from locally

```
[ ] # most frequent body style, make, color, car stuff **diane wanted to add compare and contrast fo

    import numpy
    import pandas
```

```
[ ] from google.colab import files
    uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the curren

```
Saving parking-citations.csv to parking-citations (1).csv
```

```
[ ] import pandas as pd

    df = pandas.read_csv('parking-citations.csv')
    '''
```

3. Find out the general data via each cell.