Exercise 2

The following is a description of the steps that I took to complete this exercise and the steps necessary to execute the code contained in my GitHub repository.

An alternative is to run the following shell script after you have selected your AMI

- GitHub: **w205_Exercise2_TimDavid/automation.sh**
    - If you do run this shell script, please follow the following directions:

```
$ cd /data
$ wget https://raw.githubusercontent.com/tddavid89/w205_Exercise2_TimDavid/master/automation.sh](https://raw.githubuse
$ chmod 777 /data/automation.sh
$ /data/automation.sh
```

    - Please note that as you are running this script as ***root***, you will get an error from *lein* and you will have to press ENTER to continue:

```
WARNING: You're currently running as root; probably by accident.
Press control-C to abort or Enter to continue as root.
Set LEIN_ROOT to disable this warning.
```

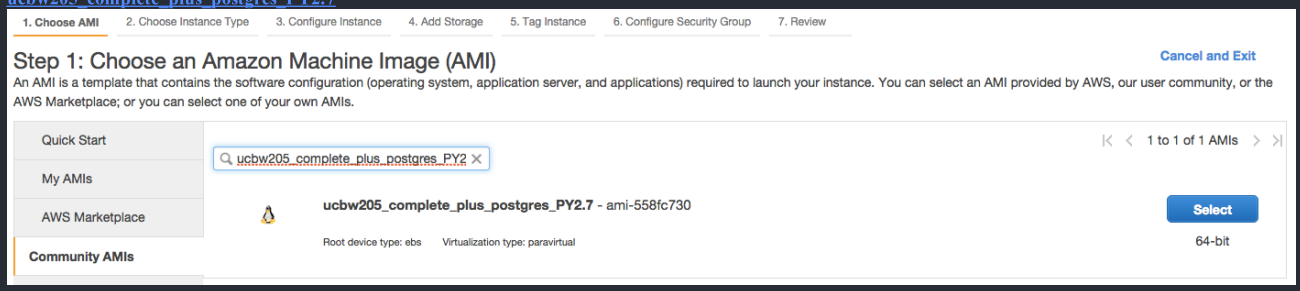
If you choose to run the code manually, here are the appropriate steps for setup:

INITIAL SETUP

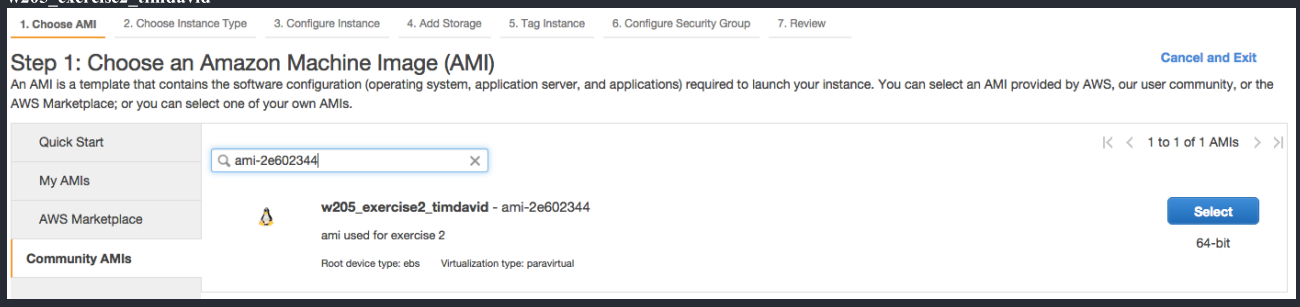
AMI

Please use one of the following AMI's in order to execute the code:

- ucbw205_complete_plus_postgres_PY2.7



- **w205_exercise2_timdavid**



User

All code was run as ***root***

Preamble

If you select **ucbw205_complete_plus_postgres_PY2.7** *as your AMI, please run the following code in order to set up the correct version of python:*

```
# Python correct version
$ yum install python27-devel -y
$ mv /usr/bin/python /usr/bin/python266
$ ln -s /usr/bin/python2.7 /usr/bin/python

#install ez_setup
$ curl -o ez_setup.py https://bootstrap.pypa.io/ez_setup.py
$ python ez_setup.py
$ /usr/bin/easy_install-2.7 pip
$ pip install virtualenv
$ wget --directory-prefix=/usr/bin https://raw.githubusercontent.com/technomancy/leiningen/stable/bin/lein
$ chmod a+x /usr/bin/lein
```

The remainder of this document is applicable to all AMI's :

Install Packages

In case they are not included, the following packages need to be installed:

```
$ pip install streamparse
$ pip install tweepy
$ pip install psycopg2
```

Create Folders

Need to create the folder where we will extract and store the files from the w205 GitHub repository

```
$ cd /data
$ mkdir ex2
$ cd ex2
```

Clone repository

Clone the repository, but only keep the exercise_2 folder:

```
$ git clone https://github.com/UC-Berkeley-I-School/w205-labs-exercises
$ cp ./w205-labs-exercises/exercise_2 ./exercise_2
$ rm -r w205-labs-exercises
```

Streamparse

Create a streamparse project called ***EX2tweetwordcount***:

```
$ sparse quickstart EX2tweetwordcount
```

```
[root@ip-172-31-58-210 exercise_2]# sparse quickstart EX2tweetwordcount

Creating your EX2tweetwordcount streamparse project...
    create    EX2tweetwordcount
    create    EX2tweetwordcount/.gitignore
    create    EX2tweetwordcount/config.json
    create    EX2tweetwordcount/fabfile.py
    create    EX2tweetwordcount/project.clj
    create    EX2tweetwordcount/README.md
    create    EX2tweetwordcount/src
    create    EX2tweetwordcount/src/bolts
    create    EX2tweetwordcount/src/bolts/__init__.py
    create    EX2tweetwordcount/src/bolts/wordcount.py
    create    EX2tweetwordcount/src/spouts
    create    EX2tweetwordcount/src/spouts/__init__.py
    create    EX2tweetwordcount/src/spouts/words.py
    create    EX2tweetwordcount/tasks.py
    create    EX2tweetwordcount/topologies
    create    EX2tweetwordcount/topologies/wordcount.clj
    create    EX2tweetwordcount/virtualenvs
    create    EX2tweetwordcount/virtualenvs/wordcount.txt
Done.

Try running your topology locally with:

        cd EX2tweetwordcount
        sparse run
```

Move/Copy Files

Copy Necessary Files From GitHub Repo to EX2tweetwordcount

```
$ cp /data/ex2/exercise_2/psycopg-sample.py /data/ex2/exercise_2/EX2tweetwordcount/psycopg-sample.py
$ cp /data/ex2/exercise_2/tweetwordcount/src/spouts/tweets.py /data/ex2/exercise_2/EX2tweetwordcount/src/spouts/tweets.py
$ cp /data/ex2/exercise_2/tweetwordcount/src/bolts/parse.py /data/ex2/exercise_2/EX2tweetwordcount/src/bolts/parse.py
$ cp /data/ex2/exercise_2/tweetwordcount/src/bolts/wordcount.py /data/ex2/exercise_2/EX2tweetwordcount/src/bolts/wordcount.py
$ cp /data/ex2/exercise_2/tweetwordcount/topologies/tweetwordcount.clj /data/ex2/exercise_2/EX2tweetwordcount/topologies/tweetwordc
```

Remove the files left from the original topology from streamparse folder

```
$ rm /data/ex2/exercise_2/EX2tweetwordcount/topologies/wordcount.clj
```

Clone my github and copy bolts/spouts to main project:

```
$ cd /data/ex2
$ git clone https://github.com/tddavid89/w205_Exercise2_TimDavid
$ cp /data/ex2/w205_Exercise2_TimDavid/scripts/wordcount.py /data/ex2/exercise_2/EX2tweetwordcount/src/bolts/wordcount.py
$ cp /data/ex2/w205_Exercise2_TimDavid/scripts/tweets.py /data/ex2/exercise_2/EX2tweetwordcount/src/spouts/tweets.py
```

File Structure

At this point, the file structure should look similar to this:

FILE STRUCTURE:

```
.
|-exercise_2
   |---EX2tweetwordcount
         |---_build
         |---config.json
         |---fabfile.py
         |---logs
         |---project.clj
         |---README.md
         |---_resources
```

```
        |---src
             |---bolts
                  |---__init__.py
                  |---parse.py
                  |---wordcount.py
             |---spouts
                  |---__init__.py
                  |---tweets.py
        |---tasks.py
        |---topologies
             |---tweetwordcount.clj
        |---virtualenvs
             |---wordcount.txt
   |---Exercise-2-Subject-205-Real Time Data Processing Using Apache Storm.pdf
   |---finalresults_limit20.py
   |---finalresults.py
   |---hello-stream-twitter.py
   |---histogram.py
   |---psycopg-sample.py
   |---README.md
   |---tweetwordcount
   |---Twittercredentials.py
   |---Twittercredentials.pyc
   |---wordcount

|-w205_Exercise2_TimDavid
   |---screenshots
        |---01_sparse_quickstart_EX2tweetwordcount.png
        |---02_sparse_run_t_300.png
        |---03_streamparse_mid_run.png
        |---04_finalresults_python_script_input_hello.png
        |---05_finalresults_python_script_part_1.png
        |---06_histogram_python_script.png
        |---AMI_selection_2.png
        |---AMI_selection.png
        |---architecture_diagram.png
   |---scripts
        |---create_table_tcount.py
        |---wordcount.py
        |---tweets.py
   |---twitterApplicationCodes
        |---finalresults.py
        |---histogram.py
   |---architecture.html
   |---architecture.md
   |---architecture.pdf
   |---automation.sh
   |---Plot.png
   |---Readme.html
   |---Readme.md
   |---Readme.pdf
   |---Readme.txt
```

Postgres

Make sure that Postgres is running:

```
$ /data/stop_postgres.sh
$ /data/start_postgres.sh
```

Log in to Postgres as user postgres and create database and table:

```
$ psql -U postgres
$ CREATE DATABASE tcount;
$ \c tcount
$ CREATE TABLE Tweetwordcount (word TEXT PRIMARY KEY  NOT NULL, count INT NOT NULL);
$ \q
```

Streamparse

Run Streamparse for 5 minutes to populate postgres table:

```
$ cd /data/ex2/exercise_2/EX2tweetwordcount
$ sparse run -t 300
```

```
[root@ip-172-31-58-210 EX2tweetwordcount]# sparse run -t 300
Running tweetwordcount topology...
Routing Python logging to /data/ex2/exercise_2/EX2tweetwordcount/logs.
Running lein command to run local cluster:
lein run -m streamparse.commands.run/-main topologies/tweetwordcount.clj -t 300 --option 'topology.workers=2' --option 'topology.acker.executors=2' --option 'streamparse.log.path="/data/ex2/exercise_2/EX2
tweetwordcount/logs"' --option 'streamparse.log.level="debug"'
WARNING: You're currently running as root; probably by accident.
Press control-C to abort or Enter to continue as root.
Set LEIN_ROOT to disable this warning.
```

```
36405 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt i: 21
36413 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt one: 13
36414 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt what: 6
36419 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt knew: 2
36420 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt it: 42
36421 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt was: 11
36424 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt feeling: 1
36434 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt coming: 2
36435 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt but: 18
36435 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt that: 22
36435 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt one: 13
36436 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt it: 43
36448 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt gained: 2
36451 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt actually: 1
36458 [Thread-23-tweet-spout] INFO  backtype.storm.spout.ShellSpout - ShellLog pid:12518, name:tweet-spout Empty queue excepti
36463 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt 8727: 1
36464 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt having: 2
36476 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt trouble: 1
36477 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt want: 14
36477 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt the: 127
36490 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt voting: 5
36492 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt too: 10
36494 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt im: 0
36495 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt to: 77
36504 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt signup: 0
36508 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt driving: 1
36517 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt please: 1
36519 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt stay: 3
36520 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt you: 69
36536 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt over: 6
36538 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt direction: 2
36550 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt sexy: 0
36550 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt some: 7
36553 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt to: 78
36554 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt of: 47
36563 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt kid: 1
36568 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt way: 5
36576 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt think: 6
36579 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt use: 5
36580 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt a: 83
36590 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt dimms: 1
36591 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt sigh: 0
36603 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt vibrator: 0
36605 [Thread-33] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12514, name:count-bolt i: 21
36608 [Thread-29] INFO  backtype.storm.task.ShellBolt - ShellLog pid:12504, name:count-bolt haha: 2
```

Code

Run twitter application codes:

```
$ python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/finalresults.py hello
[root@ip-172-31-58-210 EX2tweetwordcount]# python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/finalresults.py hello
hello: 13
```

```
$ python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/finalresults.py
[root@ip-172-31-58-210 EX2tweetwordcount]# python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/finalresults.py
: 302
0: 5
0001: 1
007: 1
01: 1
014658: 1
014659: 2
014702: 2
014703: 1
014705: 2
014706: 1
014707: 3
014708: 4
014709: 1
014710: 2
014711: 6
014712: 3
014713: 1
014716: 3
014717: 5
```

```
youu: 1
youuuu: 1
youve: 19
youx: 3
yr: 2
yrs: 2
yukon: 1
yummy: 1
yun: 1
yung: 1
yuri: 1
zaddyyyy: 1
zapiro: 1
zar23: 1
zaria: 1
zayn: 2
zeppeli: 1
zero: 1
zhu: 1
zimbabweans: 1
zone: 2
zoo: 1
zouis: 1
zrl: 1
zulu: 1
zuma: 5
zumas: 1
zyppah: 1
[root@ip-172-31-58-210 EX2tweetwordcount]#
```

```
$ python /data/ex2/w205 Exercise2 TimDavid/twitterApplicationCodes/histogram.py 5,10
[root@ip-172-31-58-210 EX2tweetwordcount]# python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/histogram.py 5,10
5: 34
6: 8
7: 13
8: 13
9: 4
10: 44
[root@ip-172-31-58-210 EX2tweetwordcount]# python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/histogram.py 10,5
The second number must be greater than the first number
[root@ip-172-31-58-210 EX2tweetwordcount]# python /data/ex2/w205_Exercise2_TimDavid/twitterApplicationCodes/histogram.py
you must include two numbers, separated by a comma
[root@ip-172-31-58-210 EX2tweetwordcount]#
```

Histogram

Here are the results of the top 20 most frequently tweeted words in the time that I ran streamparse: