Architecture – Exercise 2
Data Stream Processing and Visualization

Rohitha Bhushan

Application Details: This is a streaming application that makes use of Twitter API to parse and count tweets. The tweets are written to a Postgres DB. The words stored are further analyzed and graphically displayed.

Architecture Details: The streaming application uses streamparse which enables it to use Apache Storm with python to collect and analyze tweets. Storm uses components such as bolts and spouts to process these tweets. This application has a spout to collect the tweets which then flows to the bolts. There are 2 bolts – one for parsing the tweets and the other for counting the tweets. There is one Storm worker deployed for the spout and 2 each for the bolts. The Postgres database that the bolt writes into is names "tcount", and consists of one table "tweetwordcount". The table's schema has 2 columns : word & count. In order to analyze the words stored in Postgres DB, we have two other python programs : finalresults.py and histogram.py. The former gives a count of the number of times a word has been tweeted. It takes 0 or 1 argument(s). The latter returns a sorted list of words that occur between the 2 input arguments provided. (first<second).

Github : https://github.com/RohithaS/W205/tree/exercise_2
Repo: exercise_2

<u>File and folder structure in github</u>

- exercise_2 : the top level directory
  - README.txt : setup instructions
  - Architecture.pdf : this document
  - Plot.png : bar plot with top 20 words
  - Screesnshots : folder consisting of images of running application
  - TwitterCredentials.py : Included credentials from https://apps.twitter.com
  - extweetwordcount
      - finalresults.py
      - histogram.py
      - top20plot.py
    - topologies
      - extweetwordcount.clj
    - src
      - spouts
        - tweets.py
      - bolts
        - parse.py
        - wordcount.py

Dependencies:

Python 2.7, Streamparse, Lein, Numpy, Matplotlib, Psycopg2, VirtualEnv

Setup Instructions:

1. Create an EC2 instance with 30GB Root and AMI - UCB W205 Spring Ex2 Image
2. Open Ports – 4040, 50070, 8080, 22, 10000, 8020, 8088, 5432
3. Install Python 2.7 - follow instructions in Appendix/Lab6
4. Install Lein and Streamparse - follow instructions in Appendix
5. Run the command : pip install psycopg2
6. Create a project called extweetwordcount in Streamparse. (omitting the number '2' from the name avoids errors for some reason)
7. Clone my git repository: git clone https://github.com/RohithaS/W205.git
8. Install postgres : wget https://s3.amazonaws.com/ucbdatasciencew205/setup_ucb_complete_plus_postgres.sh
9. Run the command : setup_ucb_complete_plus_postgres.sh /dev/xvdf
10. Create Postgres database "tcount"

    ```
    psql -U postgres
    CREATE DATABASE tcount;
    \q
    ```

11. cd into exercise_2 repo
12. cd into extweetwordcount folder

    ```
    $ cd extweetwordcount
    $ sparse run
    ```

13. Run: python finalresults.py scene
14. Run: python histogram.py 8, 12
15. Run: python top20plot.py