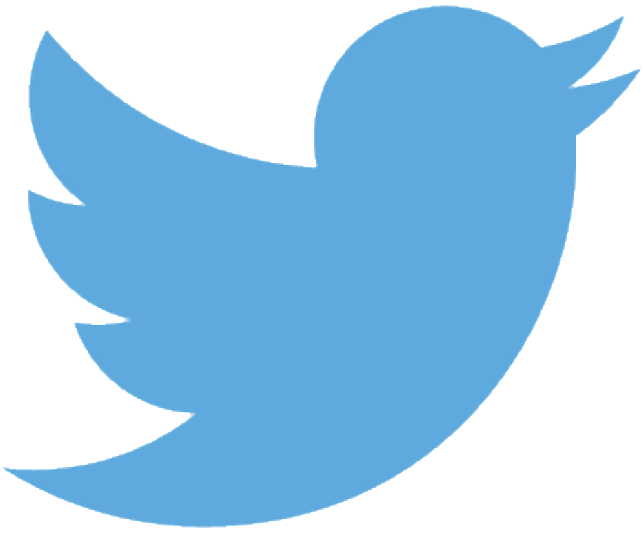
MIDS W205

Exercise 2 – Varadarajan Srinivasan

11/20/2016

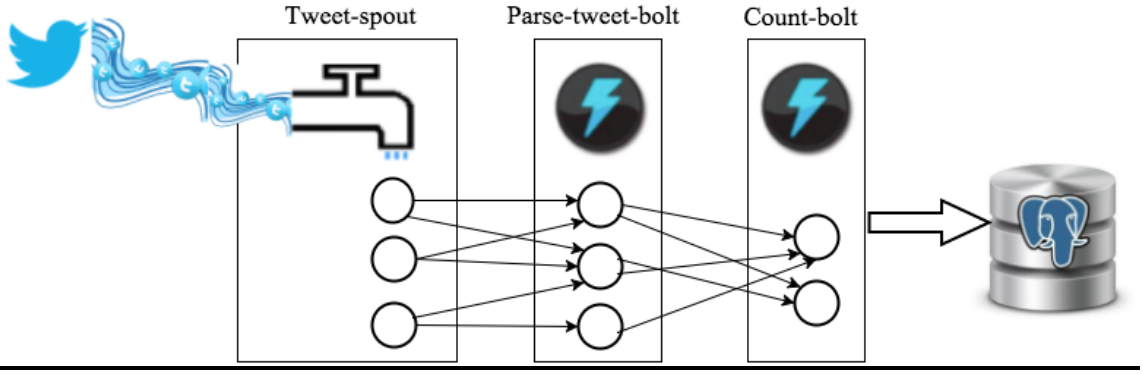
TWITTER STREAM PROCESSING USING APACHE STORM



Architecture:

In this application, live tweets are captured that show people's live interests. Tweets are processed in real-time to get insights and aggregated results are stored in a database. Figure 1 shows the overall architecture of the application. Figure 1 also shows the storm topology that has been developed as part of the application. Using Tweepy library, the application reads the live stream of tweets from twitter in the Tweet-spout component. The Parse-tweet-bolt parses the tweets, extracts the words from each parsed tweet and emits the words to the next bolt component (i.e Countbolt) in the topology. Count-bolt counts the number of each word in the received tuples and updates the counts associated with each words in the Tweetwordcount table inside the Tcount database. Tcount is a postgres database.

Figure 1



Description and Location of Files:

Name of Program	Location	Description
tweets.py	/root/EXTweetwordcount/src/spouts/	Tweet spout program. Reads from twitter live stream and spouts to bolts
parse.py	/root/EXTweetwordcount/src/bolts/	Bolt that Parses tweet, extracts words and passes it on to counter
wordcount.py	/root/EXTweetwordcount/src/bolts/	Bolt that counts the tweet and inserts into postgres database
tweetwordcount.clj	/root/EXTweetwordcount/topologies	Topology file for the architecture
finalresults.py	/root/EXTweetwordcount/	Reads and prints the data from postgres database. Two options to run the code: 1. Print count for a single word by passing word as parameter 2. Print all words with counts without passing any words as

		parameters
histogram.py	/root/EXTweetwordcount/	Takes 2 numbers as parameters (separated by comma) and prints all words that have count values within this range of numbers
top20words.py	/root/EXTweetwordcount/	Identifies top 20 words based on counts and plots a histogram
create_database_table.sql	/root/EXTweetwordcount/	Sql file that creates the tweetwordcount table
shell_wrapper.sh	/root/EXTweetwordcount/	Wrapper script that runs the whole script from start to end

Dependencies/Packages Used:

Before running the script, make sure following packages are installed:

- Apache Storm
- streamparse
- Postgresql
- Python 2.7
    - Tweepy
    - Pandas
    - Bokeh
    - Psycopg2
    - re

Steps to run the file:

Step 1: Clone the code base EXTweetwordcount from the github location

Step 2: Ensure postgres is running. You can check whether postgres is running by running the following command:

(i) ps aux | grep post

(ii) If postgres is not running. Start postgres: /data/start_postgres.sh

Step 3: Change directory into /root/EXTweetwordcount

Step 4: Run the shell wrapper script as bash shell_wrapper.sh

Step 5: Code should produce a print out all the outputs on the screen and create top20.html bar chart file