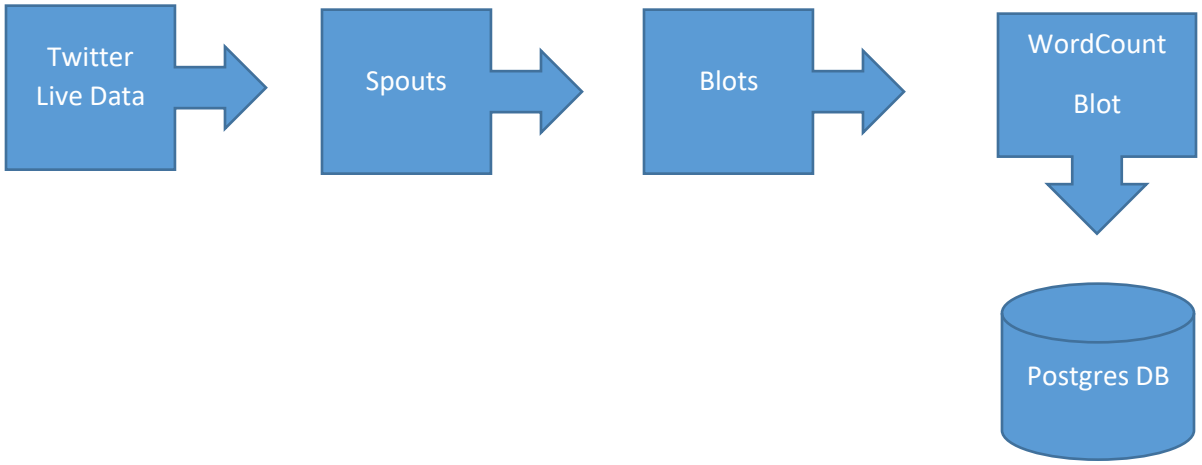
MIDS W205 Storing and Retrieving Data

Exercise 2 Lab

Exercise Purpose:

Explore streaming application with Twitter data by deploying existing codebase, in order to understand the architecture and process flow.

Application Flow:



How it works:

The application will start with streaming live Twitter data by using spouts to retrieve any tweets that contain basic English letters. Subsequently, these data are being passed through bolts for further splitting and cleaning up.

After the second step, we use the wordcount bolt to count the same English word. The counting results are being stored in a Postgres database named tweetwordcount.

Directory of the Github folder w205_fall_2017_exercise2:

- **Extweetwordcount** : folder that contains streamparse project
  subfolder: w205_fall_2017_exercise2/extweetwordcount/src/ contains bolts and spouts code
- **finalresults.py**: script for checking user input's word's count
- **histogram.py:** script to show all words within a range of count

- **Screenshot:** folder that contains screenshots as required for this lab

File dependencies:

There are some Python package dependencies:

Packages needed: tweepy, sys, psycopg2, time, simplejson

Steps to run the application:

1. Create a specific AWS EC2 instance
2. Install Python, psycopg2, and Tweepy
3. Install Postgres
4. Create Twitter API credentials
5. Run sparse run under extweetwordcount
6. While the sparse run is executing, one can simply call finalresults.py script or histogram.py script to retrieve information