Architecture of W205 Exercise 2

Real Time Data Processing Using Apache Storm

This exercise implements the Apache Storm topology illustrated in Figure 1.

Tweet-spout (tweet.py) uses tweepy library to read sampled live Twitter stream and emits tweets to Parse-tweet-bolt (parse.py). Parse-tweet-bolt cleans up tweets and emits words to Count-bolt (wordcount.py) Count-bolt counts each word and insert or update word count into tweetwordcount table in tcount Postgres database.

Two python scripts (finalresult.py and histogram.py) were created to query tweetwordcount table to show how streamed and processed data can be used by other applications.

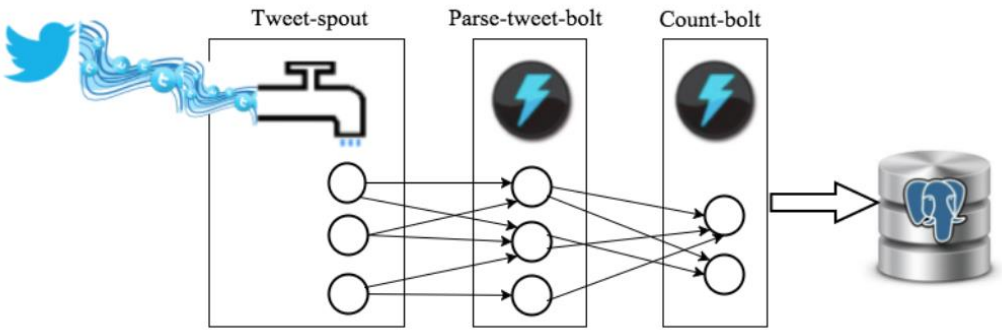

Figure 1: Application Topology

Table 1 shows the directories and files structure of the completed application. Please refer to the Readme.txt file on how to set up the environment and step-by-step instruction to run this application.

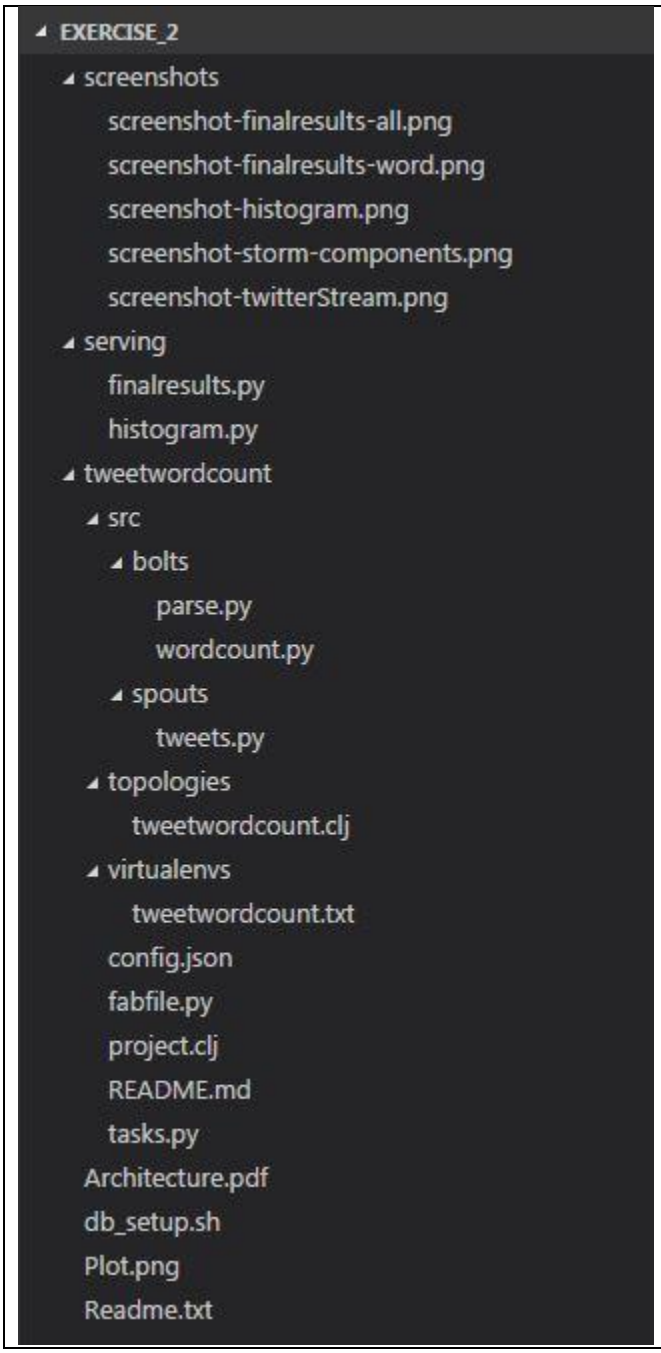	EXERCISE_2: This is the root folder. screenshots: Screenshots of serving scripts result and streaming in progress. serving: Scripts used to query data stored in database. tweetwordcount: This is the streaming application. Storm spouts and bolts components are under "src" and Storm topology is under "topologies". Architecture.pdf: This document. db_setup.sh: Bash script used to set up database and table in Postgres. Plot.png: Chart shows the top 20 most frequent words. Readme.txt: Step by step instruction to set up and run this application.

Table 1 Application Directories and File Structure