Exercise 2

W-205

Author: Saurabh Jugalkishor Jaju

Date: 09 April 2017

Table of Contents

1. Introduction

1.1. Purpose

The purpose of this document is to present the detailed scope, architecture and dependencies of the Apache Storm based Tweet word count System. It explains the design of the system. It also explains the dependencies and constraints of the system.

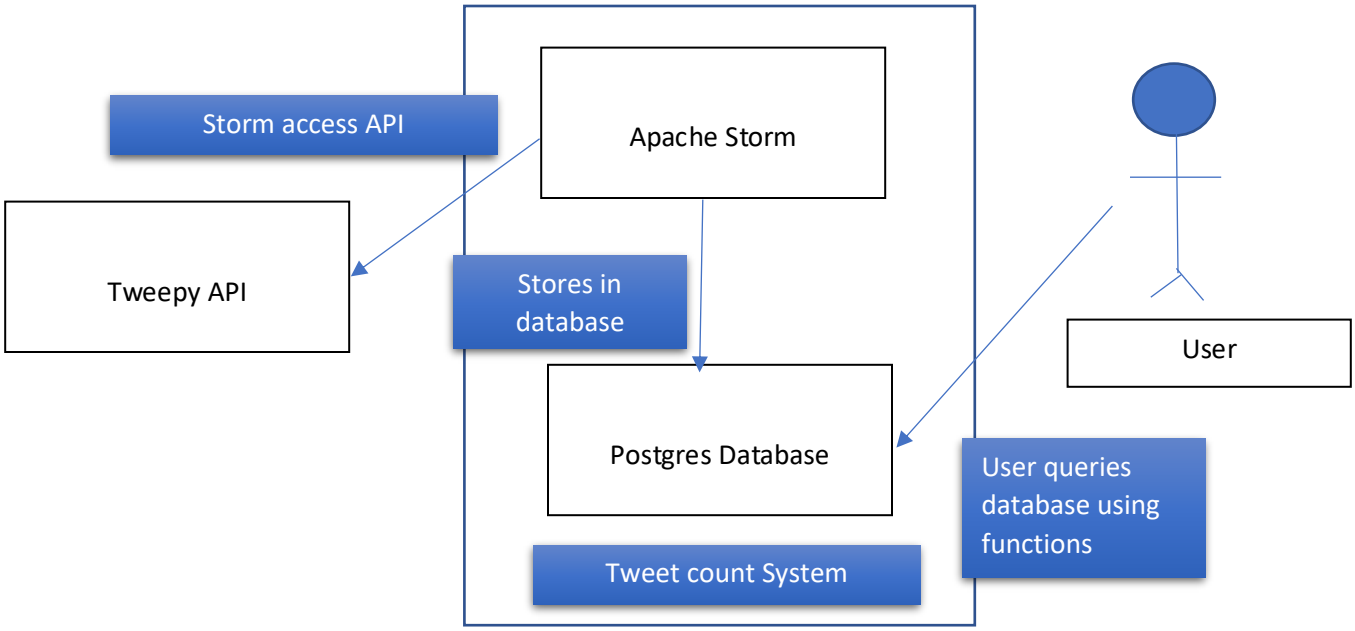
1.2. Scope

The goal of the system is to capture the tweets in real-time and count the number of occurrences of words in all the tweets and store the word count in a postgres table. This system can be used as a base for multiple applications like twitter chatbots. But such advanced applications are out of the scope of the system.

2. Overall Description

2.1. System Environment

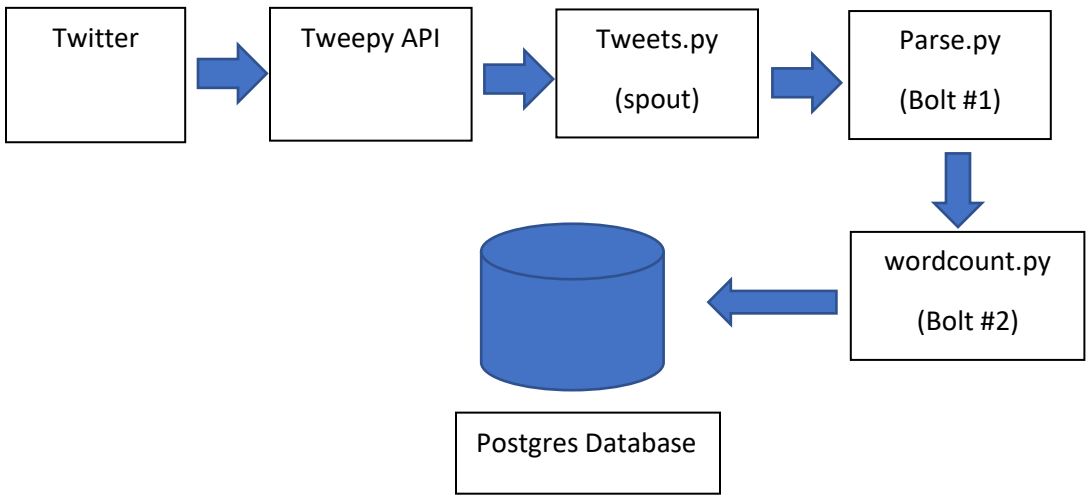
The system is setup on AWS. It utilizes the Apache storm and Postgres framework to generate wordcount of tweets in real-time.



2.2. Architecture

2.2.1. Architecture Diagram

Following is the architecture of the Tweet count system internally. This is the path on which the data flows through the modules. The following diagram describes this flow.



Flow of data within the system

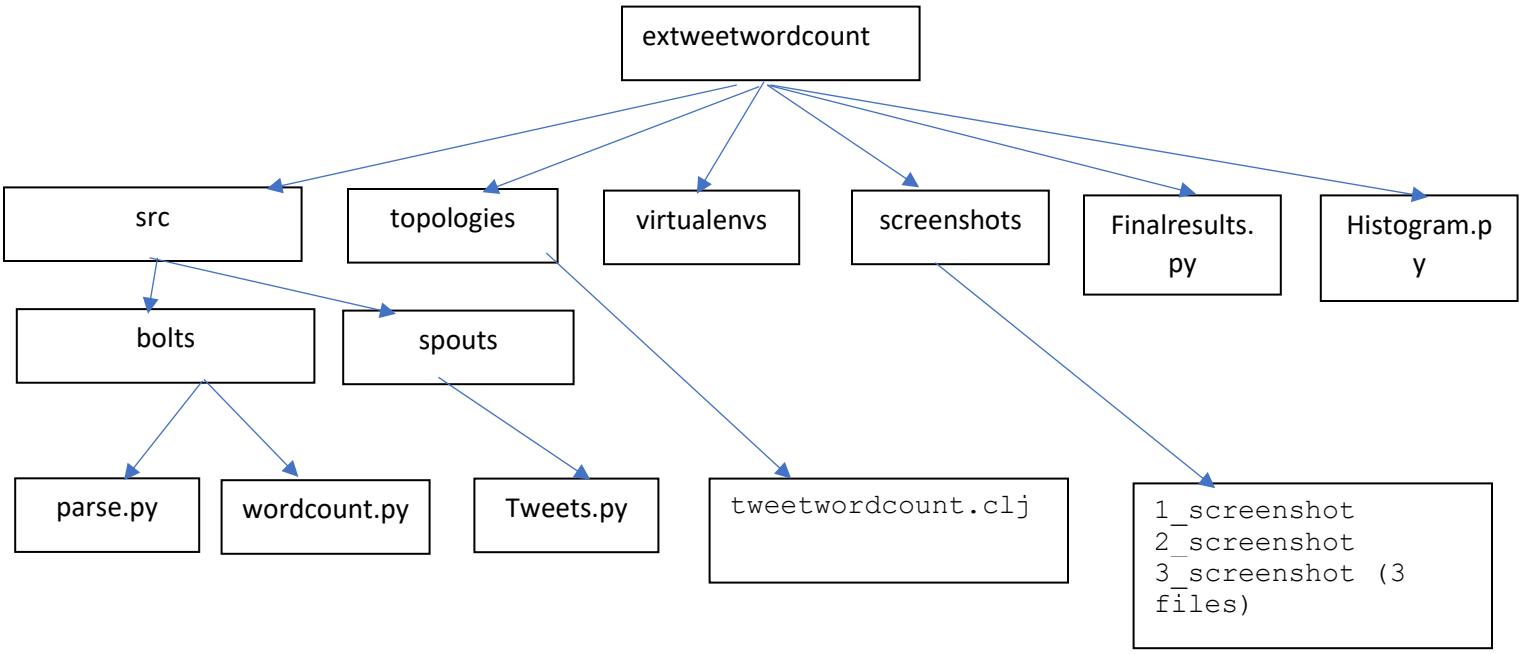
2.2.2 Architecture Description

The directory and file structure of the application:

```
extweetwordcount
    ├── Architecture.pdf
    ├── barplot.R
    ├── config.json
    ├── fabfile.py
    ├── finalresults.py
    ├── hello-stream-twitter.py
    ├── histogram.py
    ├── project.clj
    ├── psycopg-sample.py
    ├── README.md
    ├── screenshots
    │   ├── 1_screenshot_Spout_and_Bolt_launch
    │   ├── 2_screenshot_twitterStream.png
    │   └── 3_Screenshot_resultScript_output.png
    ├── src
    │   ├── bolts
    │   │   ├── __init__.py
    │   │   ├── parse.py
    │   │   └── wordcount.py
    │   └── spouts
    │       ├── __init__.py
    │       ├── tweets.py
    │       └── words.py
    ├── tasks.py
    ├── topologies
    │   ├── tweetwordcount.clj
    │   └── tweetwordcount.clj.save
    ├── Twittercredentials.py
    └── virtualenvs
        └── wordcount.txt
```

The Directory is mainly divided in four directories:

The following are the important files and directories to note:



File Structure