Exercise 2 - Architecture

Olivier Zimmer - W205

Directory

- Exercise_2
    - README.txt
    - Screenshots
        - screenshot finalresults 1
        - screenshot finalresults 2
        - screenshot histogram
        - screenshot plot top 20 words
        - screenshot twitter stream
    - Tweecounts
        - Config.json
        - Fabfile.py
        - Finalresults.py
        - Histogram.py
        - Project.clj
        - README.md
        - Src
            - Bolts
                - parse.py
            - Spouts
                - tweetcounter.py
        - Tasks.py
            - tweets.py
        - Topologies
            - wordcount.clj
        - virtualenvs
    - plot

Application idea

The goal of the application is to extract streaming data from Twitter and analyze the word count. The data is being extracted in streaming through the Twitter API, parse it then each word is being counted based its occurrence in a tweet.

Description of the architecture

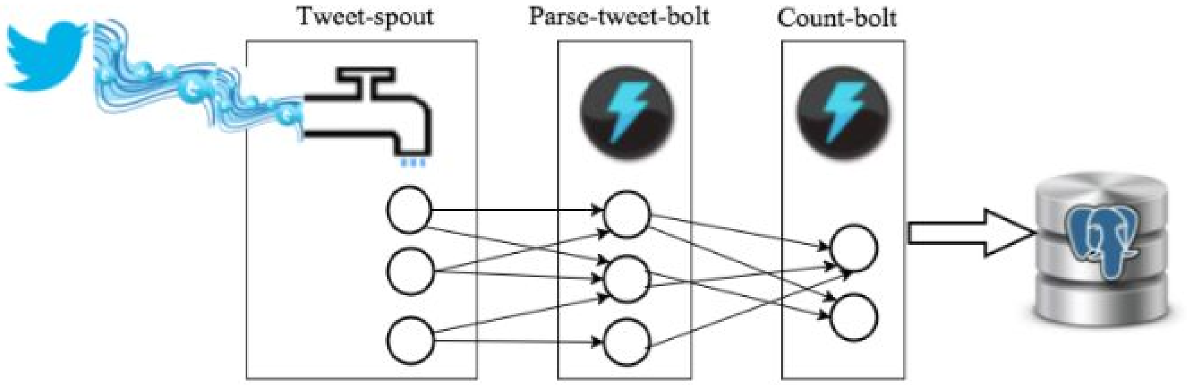

Figure 1: Application Topology

The infrastructure is leveraging a local Apache Storm cluster running on an Amazon EC2 set-up for that purpose. The setup of Storm is being through the streamparse package.

The application is extracting data from Twitter via the package Tweepy - by three spouts connected to the Twitter API. The spouts are returning the tweets. Our application topology is then plugging these tweets to three bolts in order to parse the tweets. The tweet parser is stripping away invalid words such as (@, RT…) and returns a list with all the valid words. Finally the output is sent to two bolts to count the words.

As the words are being counted - using psycopg to interface with postgres - the last bolts writes the count of words in a postgres database.

This database can be accessed independently in order to retrieve aggregates on the word count that has been extracted through serving scripts or through postgreSQL.

Necessary information to run the application

#See the Readme.txt
# How to run the application

# Pre-requisite:
#Storm running on amazon EC2
#Postgres running on the same instance

#Packages required:
#psycopg2, tweepy, streamparse, python 2.7, virtualenv

# 1. Need to create a database and a table in postgres
# From root linux shell

```
psql --username=postgres
CREATE USER w205 WITH PASSWORD 'postgres';
DROP DATABASE Tcount;
CREATE DATABASE Tcount;
ALTER DATABASE Tcount OWNER TO w205;
GRANT ALL ON DATABASE Tcount TO w205;
\q
```

```
psql --host=localhost --username=w205 --password --dbname=tcount
```
# Password for user w205: postgres

#Create a table
```
CREATE TABLE tweetwordcount (id BIGSERIAL PRIMARY KEY, word TEXT
PRIMARY KEY NOT NULL, count INT NOT NULL);
\q
```

#2. Run Tweetcount
#You will need to update your Twitter API credentials in src/spout/tweets.py
```
cd tweetcount
sparse run
```
#Control+C to kill the streaming

#3. Serving scripts
```
python finalresults.py hello
python finalresults.py hello
python histogram.py 5 10
```