Architectural Overview
W205 Sec 04. Summer 2017
John V. Tabbone

Overview

This system captures streaming twitter data, parses the data and stores them in a postgres database. The system is constructed on the Apache Storm framework, which abstracts 'sources' and 'sinks' of data. A 'Spout' is an emitter of data. It is part of the framework that publishes information. In this system the spout reads tweets from a twitter stream. The tweets are emited to 'bolts' which are consumers of data. In this system, the spout sends data to a bolt that parses the data. The bolts relays the parsed information to another bolt, which serializes the data in the postgres data.

Various python programs are employed to analyze the information from postgress.

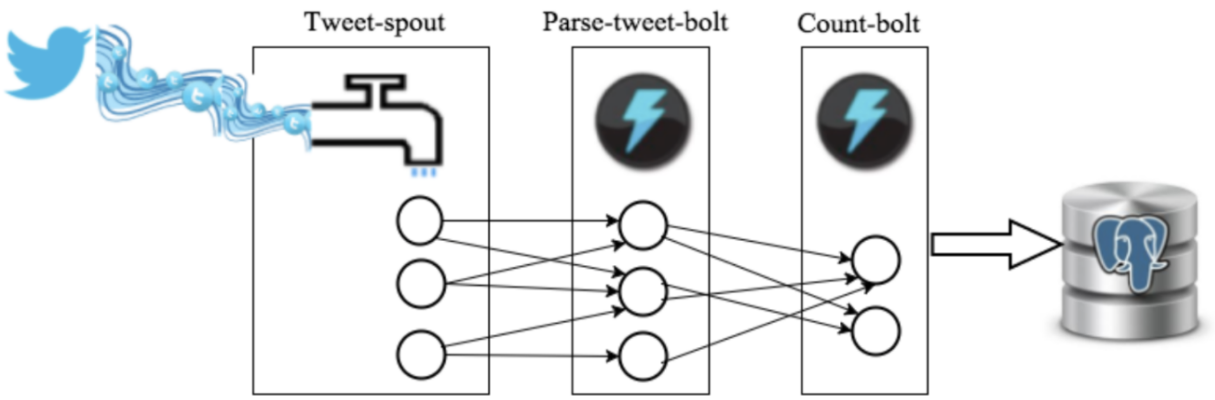

Figure 1: Application Topology

Architectural Overview
W205 Sec 04. Summer 2017
John V. Tabbone

System Components

Apache Storm: A middleware package that assists in message routing
PostgressSQL: Open Source database package
Psycopg: The python library used to connect to postgress
Tweepy: Python library used to read tweets
Streamparse: A software framework to ceate stream-enabled structures in Spark

Dependencies

The system must meet the HW and software levels described in the lab. The EC2 instance must be UCB MIDS W205 EX2-FULL

Directory Structure

(tested in /usr/local)

```
exercise_2/
        finalresults.py
        histogram.py
        Twittercredentials.py
        extweetwordcount/
                src/
                        bolts/
                                wordcount.py
                                parse.py
                        spouts/
                                tweets.py
                                words.py
                topologies/
                        extweetwordcount.clj
```