W205 Exercise 2
Architecture

Rama Thamman

Introduction

This document describes the architecture of Twitter Application created for Exercise 2. As part of the functionality, the application facilitates capturing and analyzing live twitter data around certain interest areas to gain deeper understanding of the current social events and also gain immediate and real-time insights.

Real-World use case – this application can be used for capturing consumer trends or what people or taking about while a conference or show is under way. For example, capturing which consumer electronic gadget is trending at Consumer Electronics Show (CES).

The application captures live Twitter feed, parses the tweets into words, and stores the content in the database. Outline in this document includes overview of the architecture, directory and file structure, file dependencies, and steps for running the application.

Application Architecture

The application architecture uses a number of "Big Data" software to efficiently process high velocity live twitter feeds. Diagram below is a pictorial of the architecture.

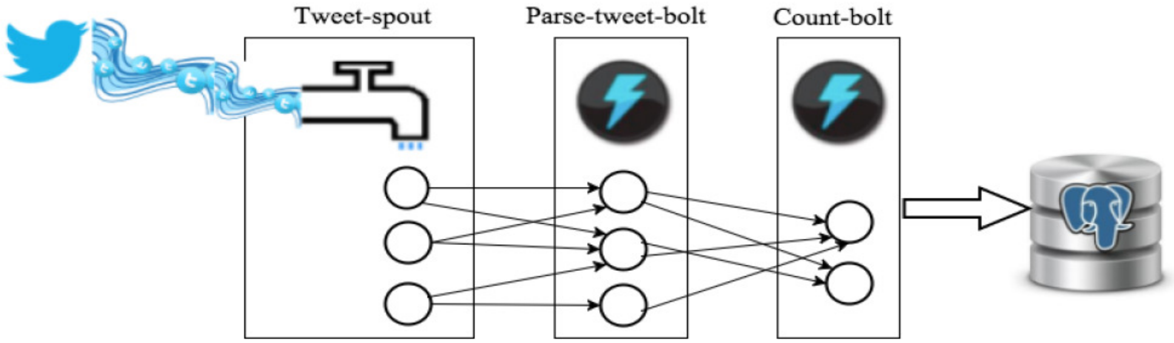

Figure 1: Application Topology

Tweet-spout captures live tweet feeds using Tweepy API and feeds the content to Parse-tweet-bolt. Twee-spout has three threads to processes the requests in parallel. This spout has appropriate timeout and error handling.

Parse-tweet-bolt takes the tweets fed by the spout and parses the tweets into words. Logic to exclude certain words or other parsing needs can be embedded in this bolt. There are three threads in this bolt that runs in parallel.

Count-bolt takes the words that are fed by Parse-tweet-bolt and updates the database with counts. Postgres database is used as the persistence layer. There are two threads in this bolt that runs in parallel.

Following components are used as part of the architecture:

- Apache Strom
  Apache Strom is a distributed real-time computation system. It is an open source software that make it easy to reliably process unbounded streams of data in real-time processing.
- Amazon EC2
  Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Service Cloud.
- Python
  Python programming language is used writing programs that facilitate the application.
- Twitter API
  An API called Tweepy is used for interfacing with Twitter to receive the feeds.
- Streamparse
  Streamparse facilitates execution of python code in real-time. It integrates python seamlessly with Apache Strom.
- Postgre SQL

  Postgre SQL is an open source relational database. The application uses Postgre for storing the words and its occurrence counts.

- Psycopg

  Psycopg is a popular PostgreSQL adapter for the Python programming language.

- Twitter Feed

  Live twitter feed is captured using a Twitter API called Tweepy. Tweepy facilitates communication with Twitter feeds to stream the tweets.

Folder Structure

- ex2
  - db_scripts
    - db_scripts.sql
  - serving_scripts
    - finalresults.py
    - histogram.py
  - tweetwordcount
    - src
      - bolts
        - parse.py
        - wordcount.py
      - spouts
        - tweets.py
    - topologies
      - tweetwordcount.clj
  - readme.txt
  - Architecture.pdf

File Dependencies

The application requires following software installed prior to execution.

- Apache Stream Strom
  - Please see https://drive.google.com/file/d/0B6706xGNaPPycWpIVU9YWUtKelU/view?usp=sharing
- Postgres
  - Version: 8.4.20

    - Install command for Psycopg; Install command - pip install psycopg2
- Python 2.7
    - Install command - $sudo yum install python27-devel –y
- Tweepy API
    - Install command: $pip install tweepy
    - Twitter application credentials. Visit https://apps.twitter.com and click on "Create New App".

Execution

Please see readme.txt for more information.