Exercise 2 Application Architecture

Nicholas Collins – W205 Spring 2016

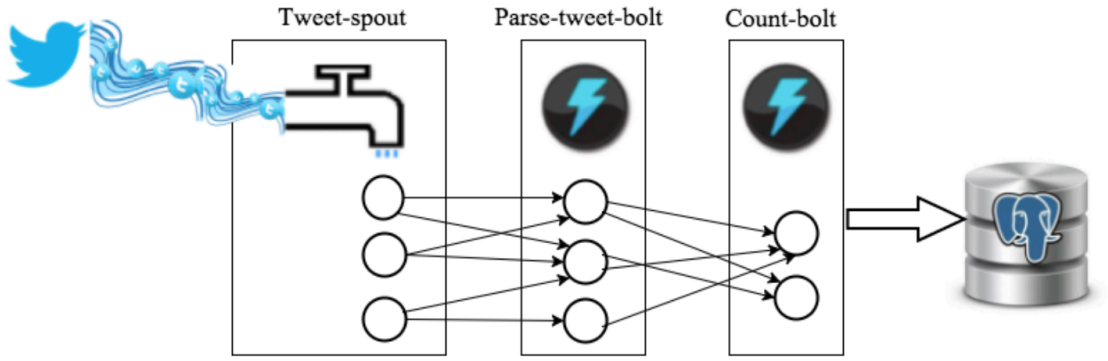
Application Idea

This application takes in a series of tweets, parses them into individual words, and then tallies the number of individual words seen across the tweets.

Description of Architecture

The user must create a twitter application that is used to get tweets. The tweets are then consumed by Apache Storm, and sent to a PostgreSQL database that collects the individual words seen in the tweets, and a count of each.

The Storm components consist of two bolts and one spout. The spout produces the tweets from the twitter application, and then sends them to a parse bolt. The parse bolt then parses the tweets into the individual words, and sends each of the words to a wordcount bolt. The wordcount bolt takes the words and adds them to the PostgresSQL database as new words are seen. If a word has already been seen before, the wordcount bolt will update the overall count for that word. The application is case-sensitive, so case differences will result in words being seen as distinct.



Streamparse is used as a framework for building and executing the Storm components.

The words and final counts for each word will appear in a table named `Tweetwordcount` in the `tcount` database.

As the application runs, words and counts will continue to accumulate in the table. If the application stops, the table and counts will remain. Whenever the application restarts, the table state will remain from where it was left and continue accumulating. To "reset" the state of the application, the user will have to truncate the `Tweetwordcount` table.

When the table contains data, the data can be queried using the `finalresults.py` and `histogram.py` python scripts as described in the "Run the Application" section below.

The application runs on a Linux server.

Directory and File Structure

ex2/	Main application directory that contains the application components including the python query scripts.
ex2/Tweetwordcount/	Location of the Storm (Streamparse) Tweetwordcount application code.
ex2/Tweetwordcount/src/	Location of the spout and bolt code for the application.
ex2/Tweetwordcount/topologies/	Location of the topology for the Storm application.

Dependencies

To run the application, you will need to install Storm (Streamparse), Python (2.7.3), and PostgreSQL (8.4.20). Other versions of these products may still work, though they have not been tested.

Running the Application

Place the credentials for your twitter application in the `tweets.py` file in the `/src/spouts` directory.

Place the credentials for connecting to the PostgreSQL database in the `wordcount.py` file in the `/src/bolts` directory.

You will need to create the `tcount` database in PostgreSQL, and run the create table code as shown in the `psycog-sample.py` file.

The python script `finalresults.py` takes a word as an argument and returns the number of occurrences of that word. If no argument is provided, it will return the number of occurrences of all words in alphabetical order.

The python script `histogram.py` takes the argument of two integers separated by a comma. It will return the words that have a number of occurrences between those two integers (inclusively).

To execute the app, you will need to call `sparse run` from the command line in the directory `ex2/Tweetwordcount/` as shown below.

```
[w205@ip-172-31-58-207 ex2]$ ls
Tweetwordcount          finalresults.py          psycopg-sample.py
Twittercredentials.py   hello-stream-twitter.py
Twittercredentials.pyc  histogram.py
[w205@ip-172-31-58-207 ex2]$ cd Tweetwordcount/
[w205@ip-172-31-58-207 Tweetwordcount]$ ls
README.md  _resources   fabfile.py  project.clj  tasks.py    virtualenvs
_build     config.json  logs        src          topologies
[w205@ip-172-31-58-207 Tweetwordcount]$ sparse run
Running Tweetwordcount topology...
Routing Python logging to /home/w205/ex2/Tweetwordcount/logs.
Running lein command to run local cluster:
lein run -m streamparse.commands.run/-main topologies/Tweetwordcount.clj -t 0 --op
tion 'topology.workers=2' --option 'topology.acker.executors=2' --option 'streampa
rse.log.path="/home/w205/ex2/Tweetwordcount/logs"' --option 'streamparse.log.level
="debug"'
```

As the application executes, you'll see the following type of information scroll along the screen…

```
360684 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt U.S: 1
360684 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt hell: 1
360686 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt Sled: 1
360689 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt 3rd: 1
360691 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt chill: 1
360692 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt or: 1
360694 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt with: 1
360694 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt ruins: 1
360697 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt for: 1
360697 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt beat: 1
360700 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt Canada: 1
360701 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt must: 1
360704 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt I'm: 1
360705 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt 4-1: 1
360706 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt be: 1
360708 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt with: 1
360709 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt puncturd: 1
360711 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt she: 1
360712 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt gonna: 1
360713 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt Naw: 1
360716 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt weird: 1
360718 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt fact: 1
360721 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt Yeah: 1
360723 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt Movie: 1
360724 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt stabby: 1
360726 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt is: 1
360728 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt bled: 1
360730 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt Netflix: 1
360730 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt The: 1
360733 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt shoot: 1
360735 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt has: 1
360737 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt my: 1
360738 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt clean: 1
360741 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt difficult: 1
360741 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt These: 1
360745 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt night: 1
360746 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt water: 1
360748 [Thread-43] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3486, name:count-bolt Right: 1
360749 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt were: 1
360754 [Thread-32] INFO  backtype.storm.task.ShellBolt - ShellLog pid:3440, name:count-bolt to: 1
```