Visualization of Current Twitter trends using Hive

Project Readme

Presented to
Professor Su Chen

Department of Computer Science
San Jose State University

for CS185C

By

Aparna Kale [013534079]
Gayatri Hungund [013738426]
Parth Patel [013705718]

December 2018

Presentation Slides:

https://prezi.com/view/YSCfI82KVwC9Zt8QCVWp/

GitHub:

Following is the link to the source code of the project:
https://github.com/ppat94/CS185C_Counting_Hashtags.git

Video Tutorial:

Following is the link to the video tutorial to run the project code. You can go through it while performing below steps:
https://drive.google.com/open?id=1vuQF1cHBG_lbDL2EIVE_2CAls88OkaA3

Contact Details:

In case of any issues please contact-

a. Aparna Kale (669) 231-9027 [aparna.kale@sjsu.edu]
b. Gayatri Hungund (669) 268-4040 [gayatri.hungund@sjsu.edu]
c. Parth Patel (551) 229-5476 [parth.b.patel@sjsu.edu]

Installation Specifications:

Mapr Hadoop 2.7.0-mapr-1710
Mapr Hive mapr-hive-2.1
Npm 6.4.1
Node 8.14.0
Java 1.8.0_191
Python 3.6.5

There are 2 ways to execute the project:

1. **Using .ova image [recommended]**
2. **Complete installation and deployment**

1. Using .ova image [recommended]:

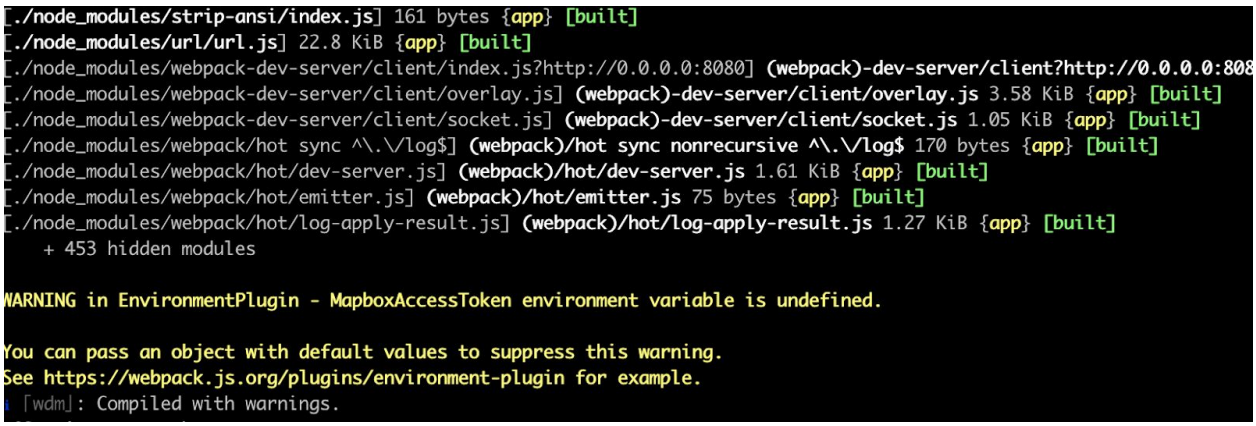
***Note**: All the following commands should be executed using the **root** user with password "**mapr**".*

1. Download the .ova image from the following link: https://drive.google.com/open?id=1zqmiNRkLkquSRdOB7QbsMUjNu5WucwGi
2. Import the .ova image in VirtualBox and start the virtual machine
3. Start the virtual machine from the virtual box and wait for it be up completely.
4. Now, on your host computer, ssh into machine using mapr user and the ip: 192.168.56.101 using the following command:
   ssh mapr@192.168.56.101
   Note: On being prompted for password, enter password as "mapr"
5. If there is trouble in logging into the virtual machine and error message "`WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!`" is displayed, then, execute the following command:
   ssh-keygen -R 192.168.56.101
6. Open a new tab in terminal and ssh into VM using the command mentioned in step 4.
7. Navigate to the CS185C_Counting_Hashtags/src using the following command:
   cd Desktop/CS185C_Counting_Hashtags/src
8. Grant permissions and execute the start.sh script by using the following command. Make sure to log in as root and enter the password as "mapr"
   >su root
   >chmod +x run.sh
   >./start.sh
   >You will be prompted to "replace hash.txt?" for which, enter "y" as a response.
   >Further you will also be prompted to "replace __MACOSX/._hash.txt?" for which, enter "y" as the response.
   This script starts the cron job. You will get the message as -

```
You are using pip version 9.0.1, however version 18.1 is available.
You should consider upgrading via the 'pip install --upgrade pip'
command.
```

```
Requirement already satisfied: numpy in
/usr/lib64/python3.6/site-packages
You are using pip version 9.0.1, however version 18.1 is available.
You should consider upgrading via the 'pip install --upgrade pip'
command.
---------------------------
Started Cron Job Execution
Setting up watches.
Watches established.
```

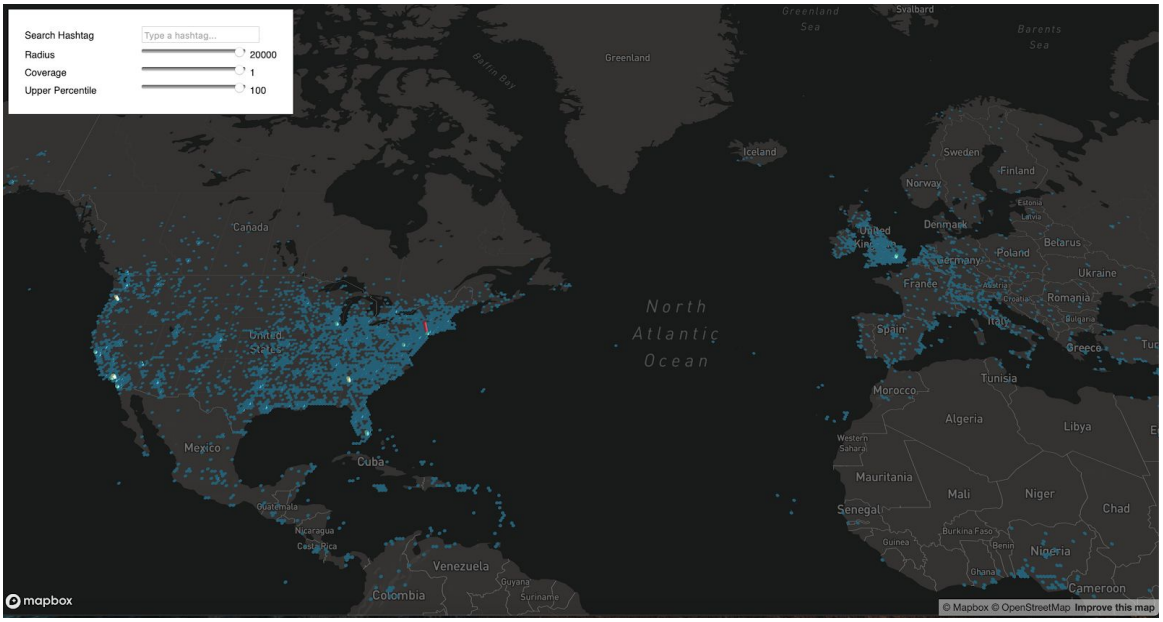
9. Now let this process run. Run the run.sh in another tab which we opened earlier and make sure to connect as root.
   >cd Desktop/CS185C_Counting_Hashtags/src
   a)Grant permissions and execute the run.sh script by using the following command:
   >chmod +x run.sh
   >./run.sh
   The above script will perform twitter streaming and terminate after 30 seconds.
   b)In the cron job terminal, you will see the JDBC code, node module getting built. The web server starts after that.
   Wait for the following message to appear on the screen before opening the browser:



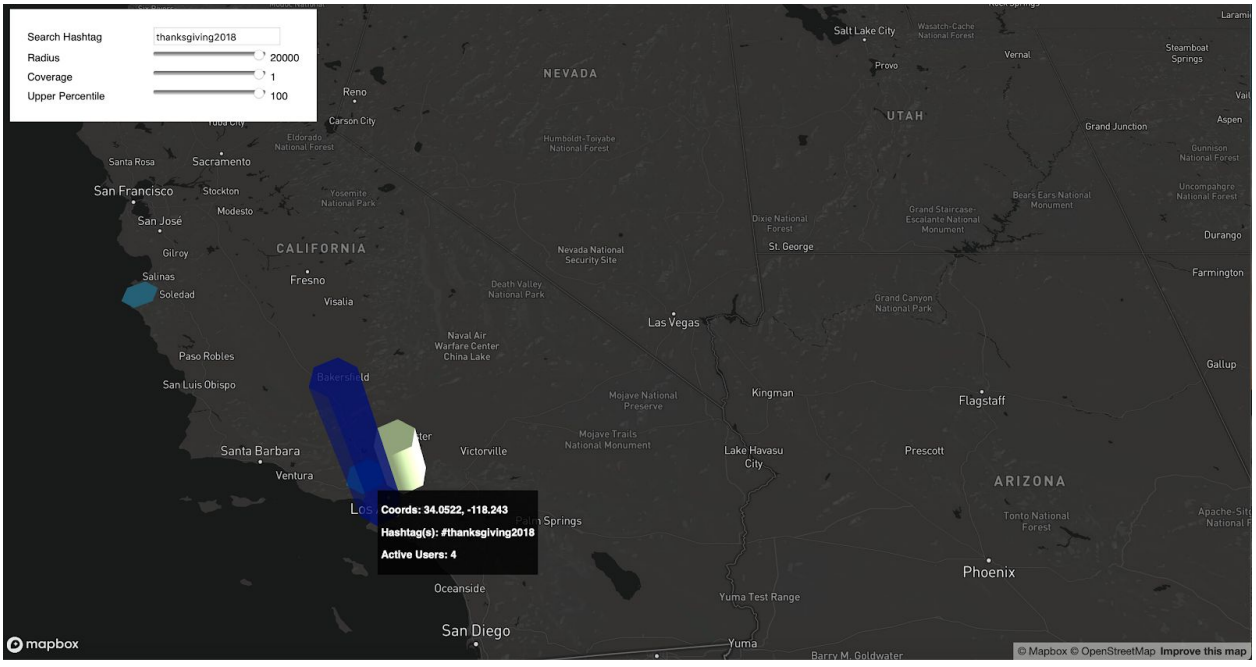
10. To find the trending tweets open the following URL in the **Safari** web browser:

    http://0.0.0.0:8080

11. Now select the search-box and hit the Enter **key 2 times** to show all the current trends in the world. (Data visualization can sometimes take time to populate the data on the webpage due to large volume of data).



12. Type a hash-tag, for example, “thanksgiving2018” and hit Enter key twice to see the change in the data and view only trends related to the hashtag specified in the searchbox.



Note: Browser compatibility: Due to limited sessionStorage on google chrome and firefox, the project currently is tested to be working on Safari.

Troubleshooting:

Unable to ssh:

Error message: Host is down
Solution:
Login to super user using “su” command and password “mapr”
Execute the following command: service network start
Check the IP address for interface enps08 to be 192.168.56.101
If no IP address is assigned, reboot the machine.

2. Complete installation and deployment

Data Generation and preprocessing:

With Twitter's API tweepy, streamed more than 10GB of data. This data was then parsed and collected only those tweets which had hashtags and geolocation coordinates. We finally collected around 3.4L tweets. Since collecting the data took days, our script will download tweets for total 30 seconds in the chunk of 10 seconds. After the data is generated, this will be given as input to DataProcessor.py which will extract the required tweets and create hash.csv under processed/ directory.

Hadoop and Hive Setup:

Pre-requisites:

Before installation, ensure that the java version is 1.8 and JAVA_PATH is correctly set in the ~/.bashrc file

Set the JAVA_HOME if needed and source the bashrc file to apply changes.

Installation:

1. Install mapr-hadoop by following instructions mentioned in the Lab Assignment 1.

2. Install mapr-hive using following link to official documentation: https://mapr.com/docs/61/AdvancedInstallation/InstallingHive.html

3. In Step 2 of the Mapr official documentation for Hive installation, click on "Configuring Database for Hive Metastore"

4. Follow the instructions to "Use MySQL as metastore for Hive"

5. Run the configure.sh as mentioned in Step 3 of official installation documentation and start the hive shell by using "hive" command.
6. Verify correct installation of the Hadoop and hive before moving further to building project environment.

Building project environment:

Pre-Requisites:

1. Install Node 8.14.0 by using the following commands:

    a. curl -sL https://rpm.nodesource.com/setup_10.x | sudo bash -

    b. yum install nodejs

    c. Verify the successful installation of node using the following command:
    node —version

2. Install git using the following command:
    a. yum install git

3. Navigate to Desktop directory and use the following command to clone the project from GitHub:
    git clone https://github.com/ppat94/CS185C_Counting_Hashtags.git

4. Now navigate to the project directory by using the command:
    cd CS185C_Counting_Hashtags/src/

Note: Before triggering the script make sure that all commands are executed from root user and all hadoop services are started.

5. To check the status of all active Hadoop services use the following command:

jps

6. If some services are not enlisted in the process list, execute the following command:

service mapr-warden start

7. Now, fire the script:

>chmod +x start.sh

>./start.sh

start.sh script file installs the dependencies and starts the cron job which triggers the execution of java code on changes appearing in the data file.

8. Open another tab in the terminal to run the following script after executing start.sh in the CS185C_Counting_Hashtags/src/ directory:

>chmod + run.sh

>./run.sh

run.sh script generates the data using TwitterLiveStreamer.py and preprocessed using DataProcessor.py file. TwitterLiveStreamer will run for 10seconds three times to generate the data which will be appended to the data we generated. As soon as the file is appended, cron job will copy the hash.txt file to the Hadoop filesystem and it will trigger the Hadoop database queries which create the database schema. This code will generate JSON file which will be fed to the node module. Node dependencies are then installed and web-pack server is started.

9. The user interface can now be accessed by the user by using the following in safari browser:

http://0.0.0.0:8080

10. Now select the search-box and hit the Enter key 2 times to show all the current trends in the world. (Data visualization can sometimes take time to populate the data on the webpage due to large volume of data)

11. Type a hash-tag, for example, “thanksgiving2018” and hit Enter key twice to see the change in the data and view only trends related to the hashtag specified in the searchbox.