              **CIS4560 Term Project Tutorial          https://lh5.googleusercontent.com/-zaX3ioEa_HraLDvx8Fd8LytWoQ30I6Wjp22nEvOttzGyjryuOD5OclyO8Pj9WRy-ZNidKGeQvbJXGeVhA8zSYcgHpk0qphXWOsy_Moi3n28g_MyguthYmJELvKTYqU4ed7p5ED7**

**Authors: Steven Hoang, Thanh Tu, Dat Le, Rafael Reyes, James Wu**

**Instructor:** [**Jongwook Woo**](https://www.linkedin.com/in/jongwook-woo-7081a85)

**Date: 12/10/2017**

**Lab Tutorial**

Steven Hoang ([shoang5@calstatela.edu](mailto:shoang5@calstatela.edu)), Thanh Tu ([ttu6@calstatela.edu](mailto:ttu6@calstatela.edu)), Dat Le ([datle9000@gmail.com](mailto:datle9000@gmail.com)), Rafael Reyes ([rafavreyes@gmail.com](mailto:rafavreyes@gmail.com)), James Wu ([wujames87@gmail.com](mailto:wujames87@gmail.com))

12/10/2017

**Citi Bike Data Analysis using Hive and Pig**

**Objectives**

**List what your objectives are.** In this hands-on lab, you will learn how to:

* Analyzing Data using Hive
* Visualize the Data in Excel
* Using Pig to load data and visualize
* Geospatial visualization

**Platform Spec**

·         IBM Bluemix BigInsights

·         CPU Speed: IOP 4.2

·         # of CPU cores: 4

·         # of nodes: 1 Management node, 1 Data node

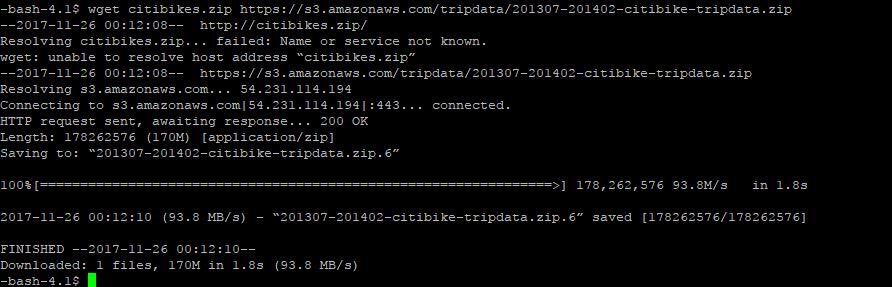
·         Total Memory Size: 1 terabyte

Analyzing Data using Hive

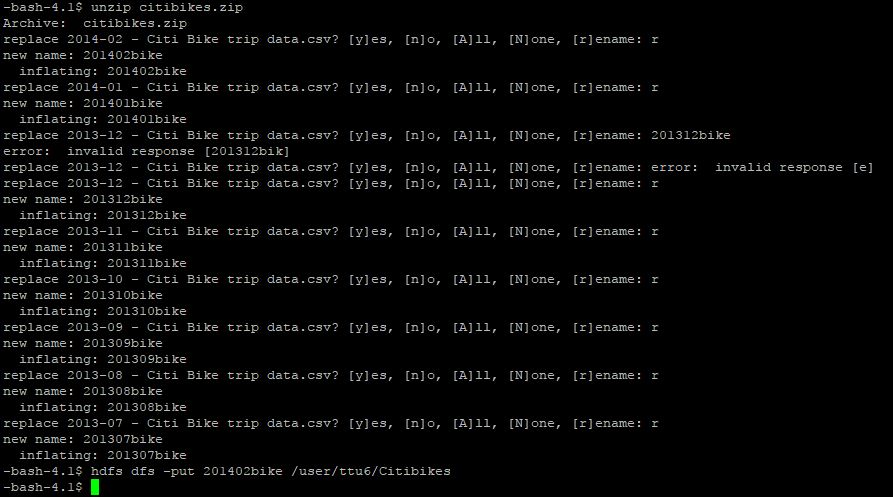
Step 1: Downloading and Unzipping the Data

Once Putty is launched, you must first download the data set and unzip it into HDFS using the following code.

1. wget citibikes.zip <https://s3.amazonaws.com/tripdata/201307-201402-citibike-tripdata.zip>



1. unzip citibikes.zip

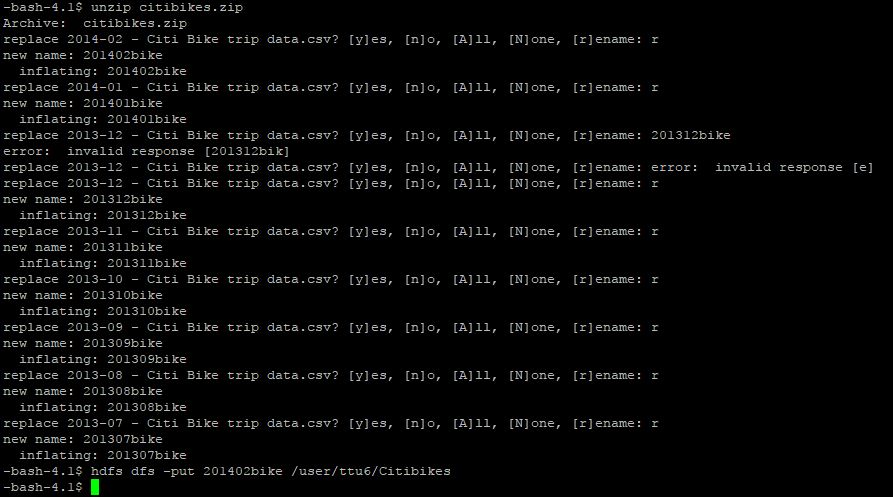


Before we can we can proceed to the rest of the steps, we must rename each file. In order to rename the files, unzip it twice.

Step 2: Creating a Folder and Insertion of Files

We must create a folder to house the files that we unzipped. We will be putting these files into our tables. Use the following code

1. hdfs dfs -mkdir /user/ttu6/Citibikes/
2. hdfs dfs -put 201307bike /user/ttu6/Citibikes



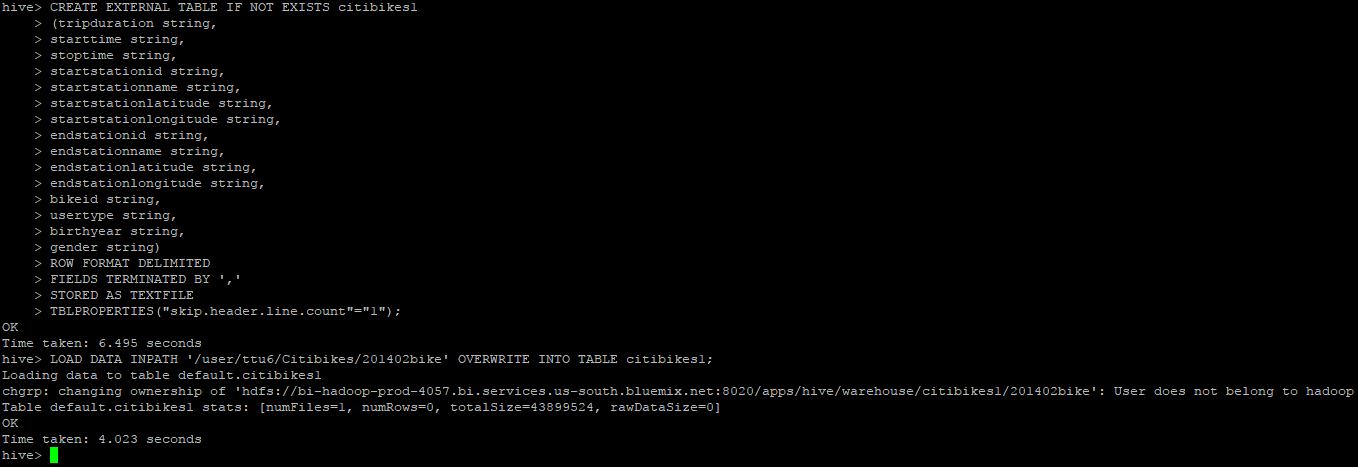
For this example, we will use 201307bike to demonstrate the table creation process.

Step 3: Table Creation and Data Loading

Now we can create the table. Doing this will allow us to query data from the data set. Use the following code to create it. First, launch hive with:

$ hive

1. CREATE EXTERNAL TABLE IF NOT EXISTS citibikes1  
   (tripduration string,  
   starttime string,  
   stoptime string,  
   startstationid string,  
   startstationname string,  
   startstationlatitude string,  
   startstationlongitude string,  
   endstationid string,  
   endstationname string,  
   endstationlatitude string,  
   endstationlongitude string,  
   bikeid string,  
   usertype string,  
   birthyear string,  
   gender string)  
   ROW FORMAT DELIMITED  
   FIELDS TERMINATED BY ','  
   STORED AS TEXTFILE  
   TBLPROPERTIES("skip.header.line.count"="1");
2. LOAD DATA INPATH '/user/ttu6/Citibikes/201307bike' OVERWRITE INTO TABLE citibikes1;



With the table created and the data loaded into the table, we can test out some queries.

Step 4: Query Testing

Try out these queries and analyze the data. The results will be displayed along with the following code.

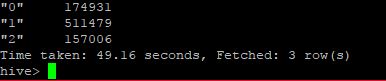
1. SELECT startstationname, COUNT(startstationname) AS pop\_station1 FROM citibikes1 GROUP BY startstationname ORDER BY popstation1 DESC limit 1;

https://lh4.googleusercontent.com/V2XOupKOBH5RQItAWYCgj1T4wXZUdKJIBG27QmEhRv2g9_JRknaZXQu4nUHnt_IBRN7FA7_TpZk6rmGPVdPDVoClFb9e2806VWaN1K3okcN09hph81kmfbhWn9FiKJM1Cp133kBw

1. SELECT endstationname, COUNT(endstationname) AS pop\_station1 FROM citibikes1 GROUP BY endstationname ORDER BY pop\_station1 DESC limit 1;

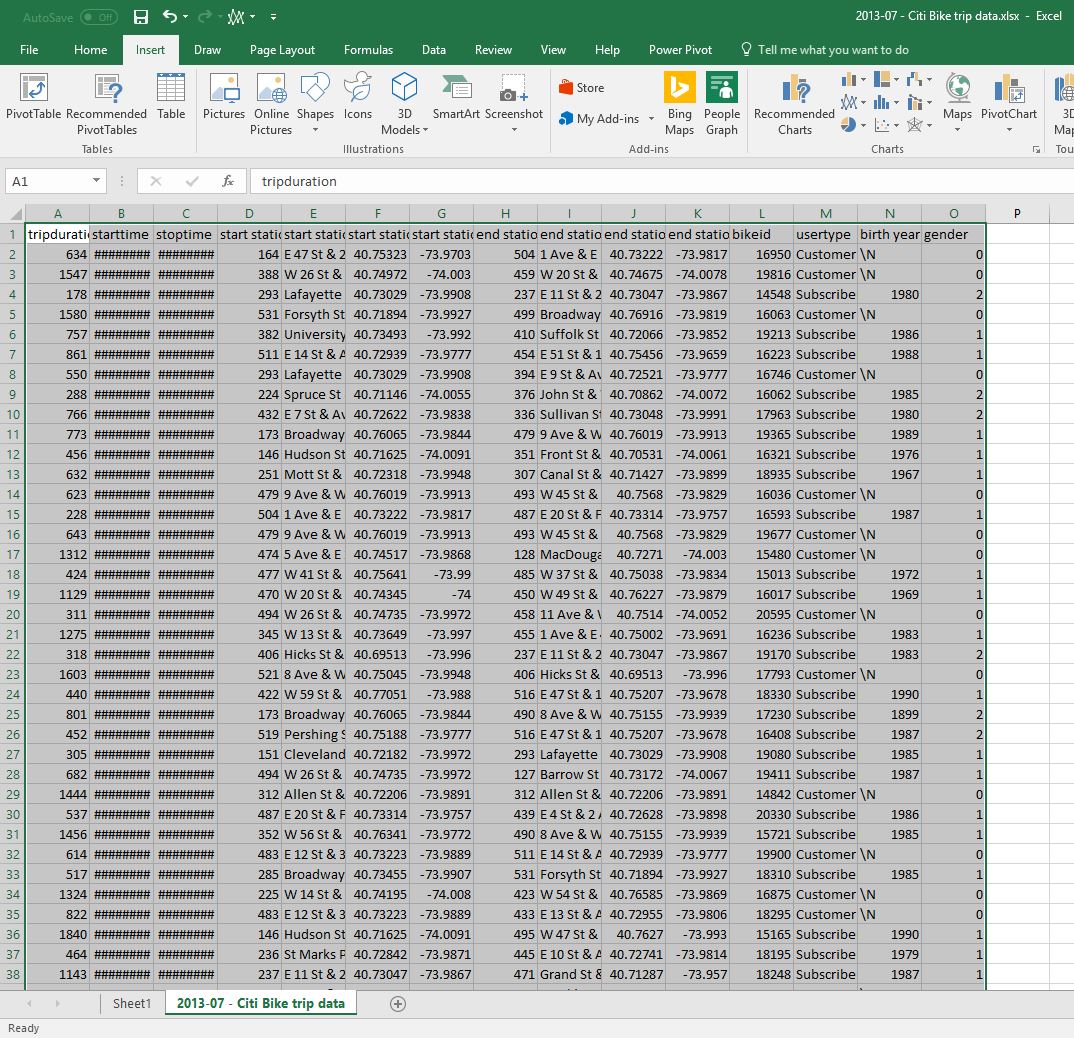
https://lh5.googleusercontent.com/W2weDIZExwUZzoHe5eIiAVmsUEwvquUuKoQPgOHnLTjxyGMYsKFgEgPhsgJtGH2iGsIFxJx4BsRbzfbRnSpGorLEsznCfPGdMK9T4gQgGrt30p1K2HGgPUUFtuSDBXreUmQBjDhO

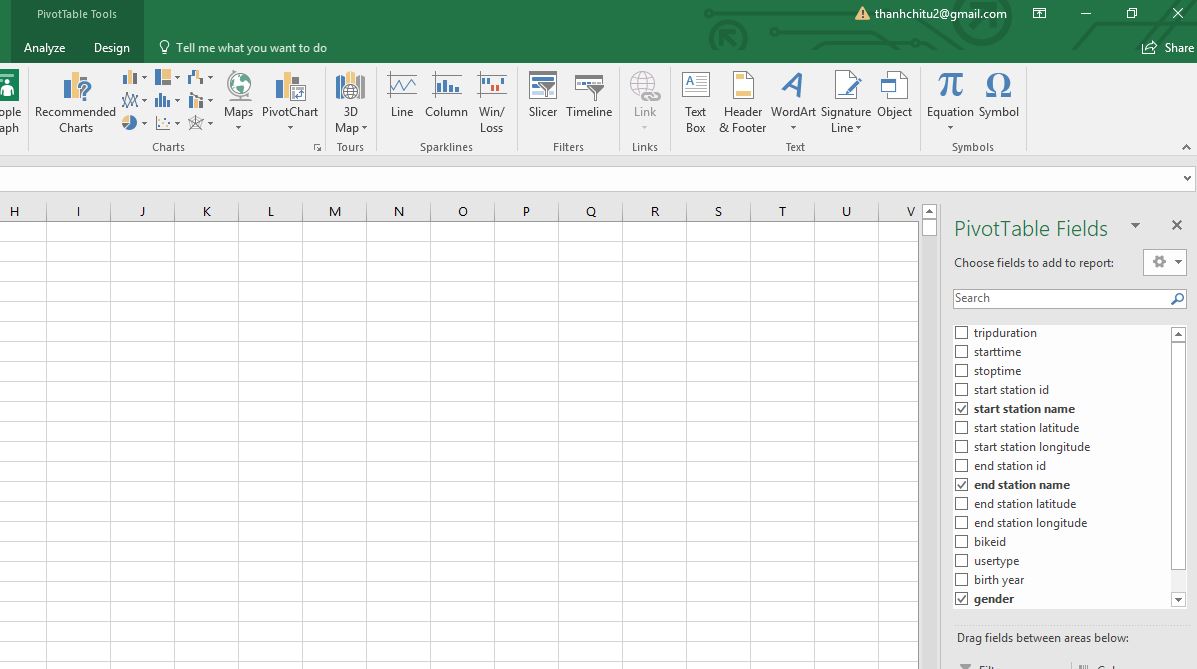
1. SELECT gender, COUNT(gender) from citibikes1 GROUP BY gender;

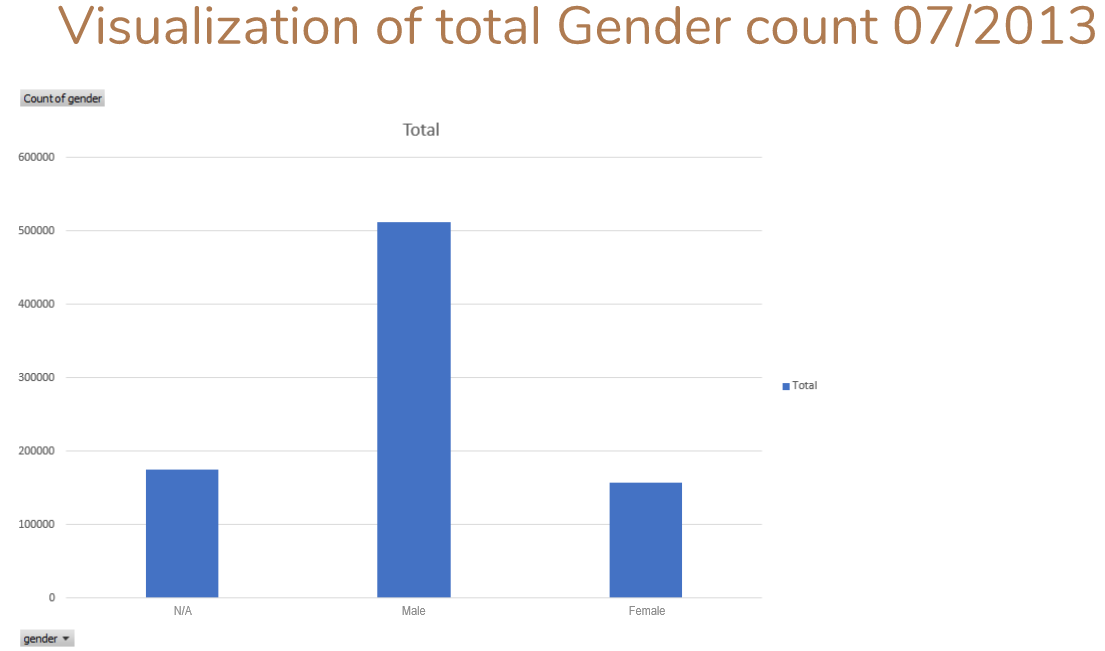


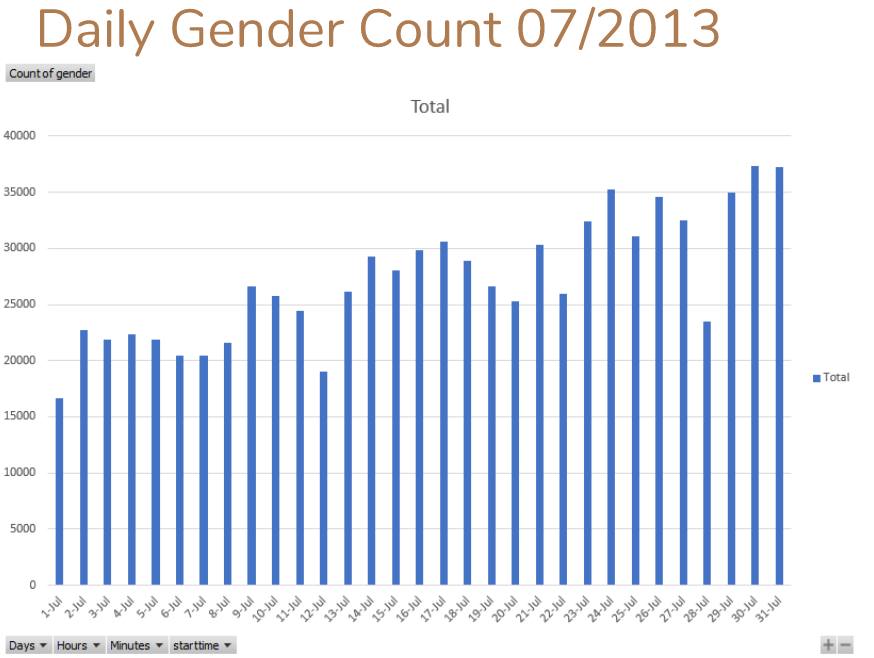
Step 5: Visualizing the Data

1. Remember to convert the csv files into excel files. Select all the data create a pivot table out of it.
2. On the right of the newly created pivot table, you will see the fields of the data. You can play around with whichever data you want to filter and see.
3. Once you have selected which fields you wish to see, you can create a chart/graph with that data. The following screenshots will show the process and some examples with gender and stations.



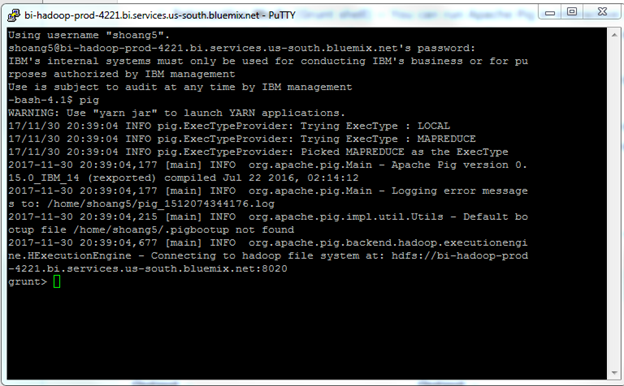






Step 1: Start the Grunt Shell Pig

**This step is to load into the Grunt shell using ‘Pig’**



Step 2: Loading and Storing the Data

**This step is to load the data files from '/user/shoang5/citi/' into Pig by executing the Pig Latin statement.**

grunt> data = LOAD '/user/shoang5/citi/' USING PigStorage(',') AS

(tripduration:chararray, starttime:chararray, stoptime:chararray,

start\_station\_id:int, start\_station\_name:chararray, start\_station\_latitude:double,

start\_station\_longitude:double, end\_station\_id:int, end\_station\_name:chararray, end\_station\_latitude:double, end\_station\_longitude:double, bikeid:int, usertype:int, birth\_year:int, gender:int);

1. To show the first 25 records of 2013-14.

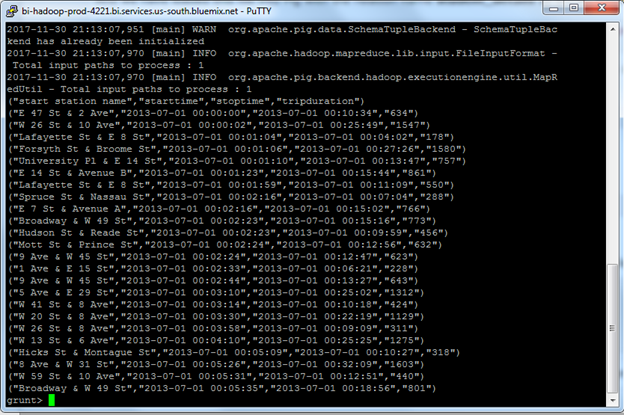
data\_subset = LIMIT data 25;

2.  Select specific columns from a relation.

specific\_columns = FOREACH data\_subset GENERATE start\_station\_name, starttime, stoptime, tripduration;

3. To read the 25 records using ‘DUMP’ specific\_columns.

            DUMP specific\_columns;



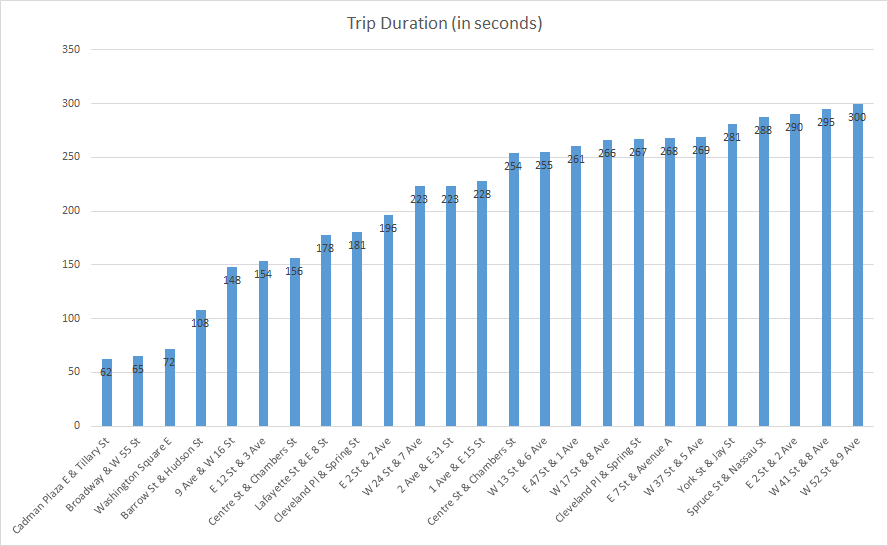
4. Store relationship data into a HDFS File.

           STORE specific\_columns INTO ‘user/shoang5/duration’ USING PigStorage(',');

Step 3: Visualization of Trip Duration (in seconds) for the first 25 stations

**Download the file part-r-00000 from ‘user/shoang5/duration’ and open it with Excel.**

1. Select A1 in excel and Insert a column chart to display the image below.



Step 1: Geospatial Visualization: Loading Data

**This step is to load the data files from '/user/shoang5/citi/' into Pig by executing the Pig Latin statement.**

grunt> data = LOAD '/user/shoang5/citi/' USING PigStorage(',') AS

(tripduration:chararray, starttime:chararray, stoptime:chararray,

start\_station\_id:int, start\_station\_name:chararray, start\_station\_latitude:chararray,

start\_station\_longitude:chararray, end\_station\_id:int, end\_station\_name:chararray,

end\_station\_latitude:double, end\_station\_longitude:double, bikeid:int, usertype:int, birth\_year:int,

gender:int);

Step 2: Remove Redundant Station Names and Storing the Data

**This step is to use the DISTINCT operator to remove redundant station names.**

1. Using DISTINCT operator to remove redundant information.

distinct\_start = DISTINCT data;

1. Select specific columns from a relation.

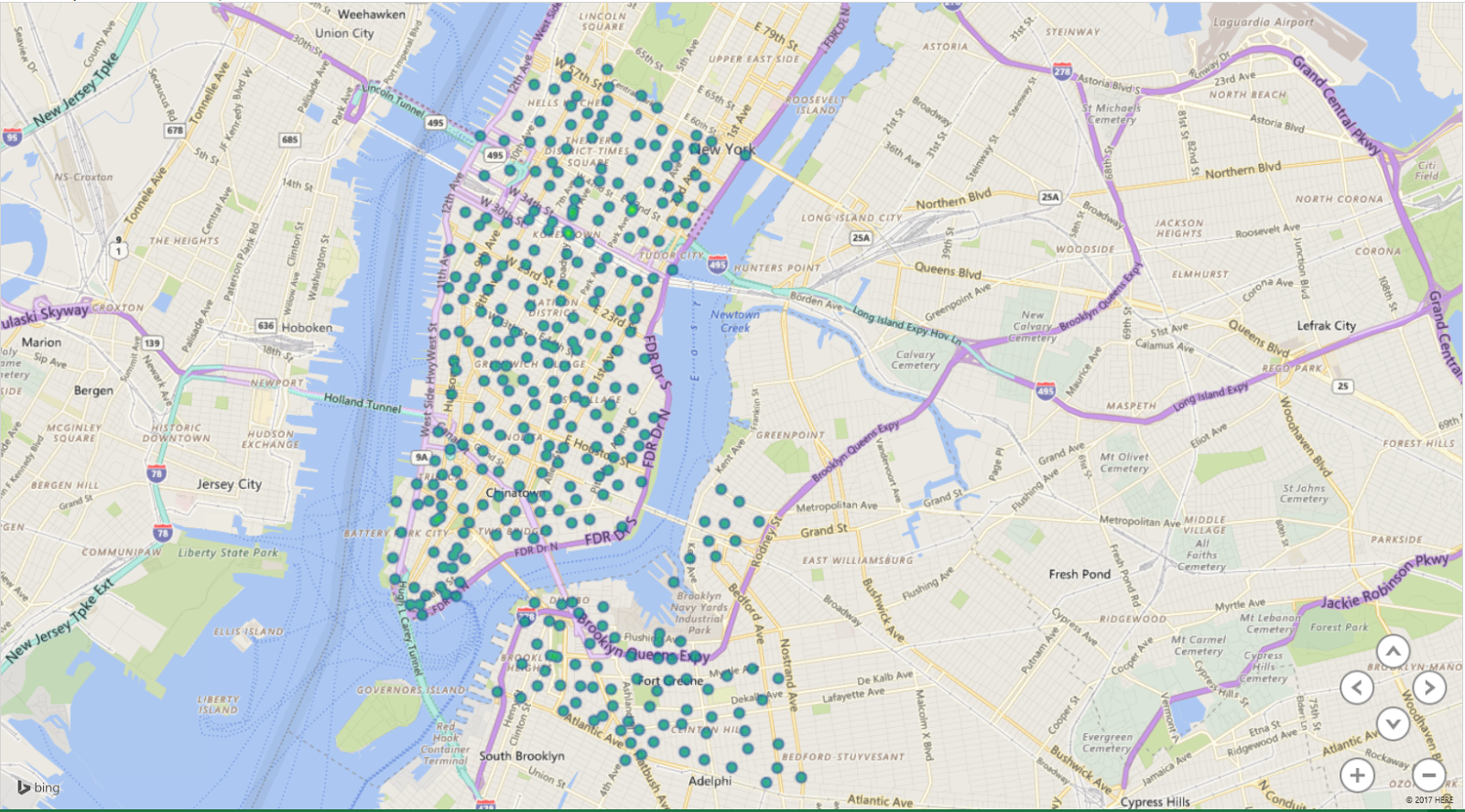
specific\_columns = FOREACH data GENERATE start\_station\_name, start\_station\_latitude, start\_station\_longitude;

1. Store relationship data into a HDFS File.

STORE specific\_columns INTO '/user/shoang5/geoStart' USING PigStorage(',');

Step 3: Geospatial Visualization

**This step is to use 3D map from Excel to visualize all the station names from 2013-14**

****

References

1.      <https://s3.amazonaws.com/tripdata/index.html>

2.      <https://github.com/dle31/Big-Data-Project>

3.      <http://www.calstatela.edu/centers/hipic/related-site>