Hazardous Air Pollutants in USA from 1990 to 2017 Analysis in Hive using IBM BigInsights

Krishna Ghorpade
Graduate Student
College of Business & Economics
Department of Information Systems
California State University
Los Angeles, USA
Tel. +1 (323) 405-5066
e-mail: kghorpa@calstatela.edu

Songyun Qian
Graduate Student
College of Business & Economics
Department of Information Systems
California State University
Los Angeles, USA
Tel. +1 (626) 429-1450
e-mail: sqian2@calstatela.edu

Himani Patel
Graduate Student
College of Business & Economics
Department of Information Systems
California State University
Los Angeles, USA
Tel. +1 (650) 534-4487
e-mail: hpatel60@calstatela.edu

Shahnawaz Khan
Graduate Student
College of Business & Economics
Department of Information Systems
California State University
Los Angeles, USA
Tel. +91 9960188188
e-mail: skhan30@calstatela.edu

Jongwook Woo
Graduate Professor
Department of Information Systems,
California State University
Los Angeles
Tel. 323-343-2916,
Fax. 323-343--5209
e-mail : jwoo5@caltstatela.edu

Abstract: This paper aims at performing data analysis in Hive using IBM BigInsights for the Hazardous Air Pollutants in USA from 1990 to 2017 along with transfering data with WINSCP and visualization with Tableau. From the data which was available we were able to categorize the air pollutants by different parameters like which pollutant was measured the most followed by city, state, date etc. The data set is a daily summary file, containing data for every monitor in the EPA database.

Keywords:- data analysis, transferring data, visualization, categorize, pollutant, summary file, EPA database etc.

1. Introduction

Air pollution is a serious problem in the world right now. Hazardous air pollutants, also known as toxic air pollutants or air toxics are poisonous for human body. Those hazardous air can cause cancer or other serious health problems, such as reproductive problems to abnormality by birth time. This data set is from the Environmental Protection Agency (EPA) tracking 187 air pollutants from 1990 to 2017 [1] . The data set is a daily summary file, containing data for every monitor in the EPA database. You will learn how to:

- Analyze data to determine which air pollutants are measured the most
- Analyze data to determine which cities have the highest and lowest air pollutants measured
- Analyze data to determine which states have the highest and lowest air pollutants measured
- Analyze data to determine which dates have the highest and lowest air pollutants measured
- Transfer data using **WINSCP**
- Visualize data in **Tableau**

2. General Instructions

In this project, we have analyzed and visualized Air Pollution Data. We have downloaded Air Pollution Data from Kaggle.com, upload to Google Drive, then download to the local system in Bluemix BigInsights. Then you will learn how to upload it to HDFS. You will figure out how to manipulate and analyze air pollution data in HDFS using HiveQL [2]. You will also practice how to visualize the result in Tableau.

2.1 The Project Flow

2.1.1 Air Pollutants data downloaded from Kaggle.com

The first stage of this project was downloading the dataset from www.kaggle.com. To download it, we needed to create an account on their website. Without creating an account we won't be able to download the dataset. After downloading it we need to save the dataset in the zip file.

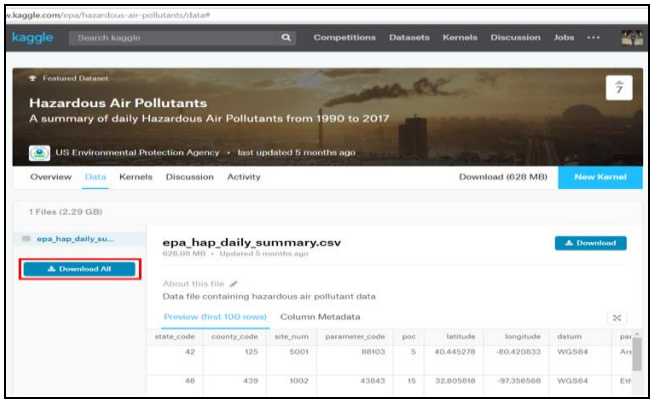

Figure 1: Downloading the dataset

The Figure 1 describes about the dataset which we have downloaded. This dataset is in .csv format and has total twenty nine columns. This data set is from the Environmental Protection Agency (EPA) tracking 187 air pollutants from 1990 to 2017 [3]. The data set is a daily summary file, containing data for every monitor in the EPA database.

2.1.2 Air Pollutants data loaded into Google Drive

The next stage was loading the dataset into the google drive. It was necessary to upload into Google Drive so that we can genereate the link and can be used easily to upload the dataset into the table. The dataset has to be in zip file before loading. This can be explained below.

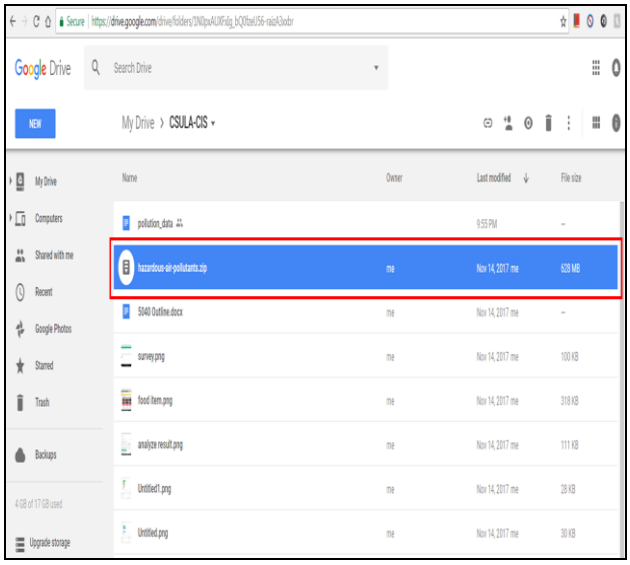

Figure 2: Data uploaded into Google Drive

2.1.3 Air Pollutants data loaded into BigInsights

In this stage we have to right click the zip file after uploading the dataset it onto Goggle Drive, and a sharable link is seen. We have to copy the link so that we can use it in the putty terminal. The link is given as https://drive.google.com/open?id=19pPy4pSVarMM6YAAnJjThox2lcThmFxJ We need to remotely access your BigInsights that you executed in your Bluemix account using ssh. You can download the data zip file Hazardous Air Pollutants from Google Drive. (Note: Don't forget to replace the red part of the link with the link you have generated). The commands in the terminal are as follows.

$ curl -c /tmp/cookies "https://drive.google.com/uc ?export=download&id=19pPy4pSVarMM6YAAnJjTho x2lcThmFxJ"> /tmp/intermezzo.html

$ curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/intermezzo.html | grep -Po 'uc-download-link" [^>]* href="\K[^"]*' | sed 's/\&/\&/g')" >

After running the above two commands we are able to get the results which is visible in the following figure.

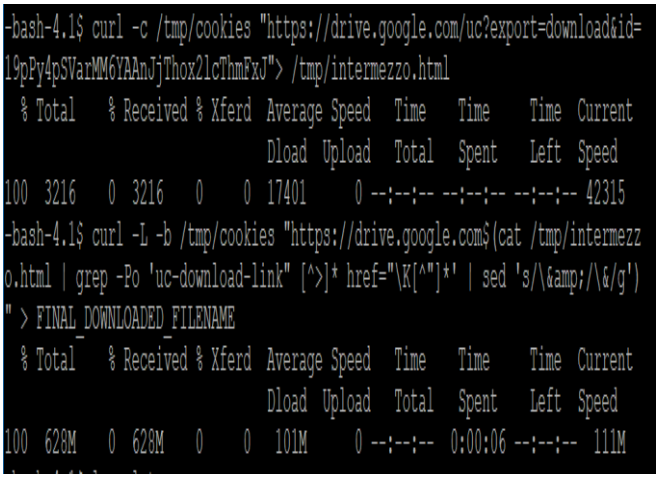

Figure 3: Data loaded into table

$ ls-alrt (check if this file FINAL_DOWNLOADED_FILENAME is available on path)

$ mv FINAL_DOWNLOADED_FILENAME air_pollutant.zip

$ unzip air_pollutant.zip

The above queries will check if this file FINAL_DOWNLOADED_FILENAME is available on the path and unzip the file.
If the data is present, we can see it in the Figure 4.

```
-bash-4.1$ ls -alrt
total 653032
-rw-------. 1 sqian2 biusers  10096122 Nov 21  2016 tweetsbi.csv
drwxr-xr-x. 19 root   root        4096 Nov 15 03:30 ..
drwx------. 3 sqian2 biusers      4096 Nov 15 17:49 .pki
-rw-------. 1 sqian2 biusers      1145 Nov 15 21:28 .bash_history
drwxr-xr-x. 3 sqian2 biusers      4096 Nov 15 21:32 .
-rw-------. 1 sqian2 biusers 658586975 Nov 15 21:32 FINAL_DOWNLOADED_FILENAME
-bash-4.1$ mv FINAL_DOWNLOADED_FILENAME air_pollutant.zip
-bash-4.1$ unzip air_pollutant.zip
Archive:  air_pollutant.zip
  inflating: epa_hap_daily_summary.csv
```

Figure 4: Data is present

$ ls -alrt (check csv is available or not)

```
-bash-4.1$ ls -alrt
total 3056992
-rw-------. 1 sqian2 biusers   10096122 Nov 21  2016 tweetsbi.csv
-rw-------. 1 sqian2 biusers 2461649186 Jun 30 18:54 epa_hap_daily_summary.csv
drwxr-xr-x. 19 root   root         4096 Nov 15 03:30 ..
drwx------. 3 sqian2 biusers       4096 Nov 15 17:49 .pki
-rw-------. 1 sqian2 biusers       1145 Nov 15 21:28 .bash_history
-rw-------. 1 sqian2 biusers  658586975 Nov 15 21:32 air_pollutant.zip
drwxr-xr-x. 3 sqian2 biusers       4096 Nov 15 21:33 .
```

Figure 5: CSV Availability

2.1.4 Create Hive table to Query Air Pollutants data

The following Hive statement creates an external table that allows Hive query stored in HDFS. External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.The Hive statement below create a new table named, air_pollution, by describing the fields within the files, the delimiter (comma) between fields [4]. This will allow you to create Hive queries over your data.

Open Hive shell environment as follow
$ hive

In the Hive shell, you need to copy and paste the following Hive QL code to create an external table "air_pollution". (**Note:** Don't forget to replace the red part with your account name)

hive> CREATE TABLE IF NOT EXISTS air_pollution (state_code DECIMAL,county_code DECIMAL, site_num DECIMAL,parameter_code DECIMAL, poc DECIMAL,latitude DECIMAL (10,6), longitude DECIMAL (10,6), datum string, parameter_name string,sample_duration string, pollutant_standard string,date_local date, units_of_measure string,event_type string, observation_count DECIMAL, observation_percent DECIMAL, arithmetic_mean DECIMAL, first_max_value float, first_max_hour float, aqi string, method_code DECIMAL, method_name string, local_site_name string, address string, state_name string, county_name string, city_name string, cbsa_name string, date_of_last_change date)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/sqian2/epa_hap_daily_summary'
TBLPROPERTIES ('skip.header.line.count'='1');

The next step is you have to load the data into the table. (**Note:** Don't forget to replace the red part with your account name)
hive>load data local inpath '/home/sqian2/epa_hap_daily_summary.csv' into table air_pollution;

```
hive> CREATE TABLE IF NOT EXISTS air_pollution
    > (state_code DECIMAL,
    > county_code DECIMAL,
    > site_num DECIMAL,
    > parameter_code DECIMAL,
    > poc DECIMAL,
    > latitude DECIMAL (10,6),
    > longitude DECIMAL (10,6),
    > datum string,
    > parameter_name string,
    > sample_duration string,
    > pollutant_standard string,
    > date_local date ,
    > units_of_measure string,
    > event_type string,
    > observation_count DECIMAL,
    > observation_percent DECIMAL,
    > arithmetic_mean DECIMAL,
    > first_max_value float ,
    > first_max_hour float,
    > aqi string,
    > method_code DECIMAL,
    > method_name string,
    > local_site_name string,
    > address string,
    > state_name string,
    > county_name string,
    > city_name string,
    > cbsa_name string,
    > date_of_last_change date)
    >
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE LOCATION '/user/sqian2/epa_hap_daily_summary'
    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.092 seconds
hive> load data local inpath '/home/sqian2/epa_hap_daily_summary.csv' into table air_pollution;
Loading data to table default.air_pollution
Table default.air_pollution stats: [numFiles=1, totalSize=2461649186]
OK
Time taken: 46.352 seconds
```

Figure 6: Table Created

Then in the Hive shell, you need to check if the table "air_pollution" is shown:
hive>show tables;

Now you can query the content of the air_pollution table:
hive>select * from air_pollution limit 10;

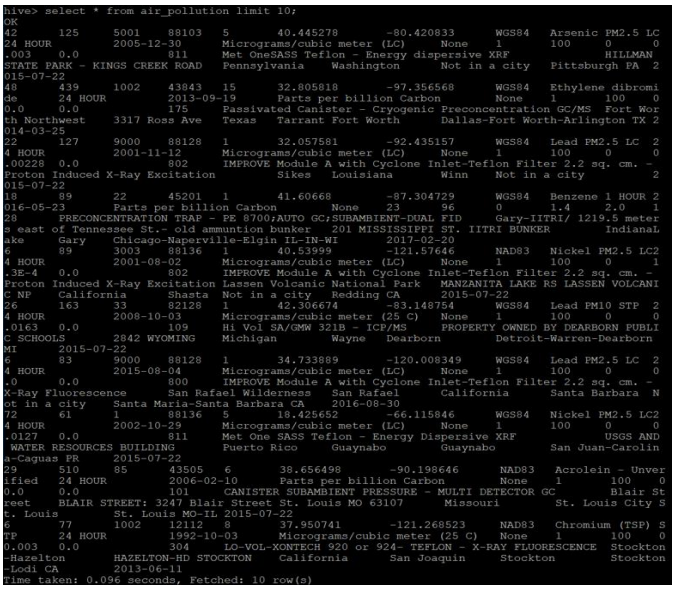

Figure 7: Displaying the first 10 values

2.1.5 Creating Hive Queries to Analyze Data.

The following Hive Queries will show you the top 10 pollutants:

Hive>Select count(*) as pollution,parameter_name from air_pollution group by parameter_name order by pollution DESC limit 10;

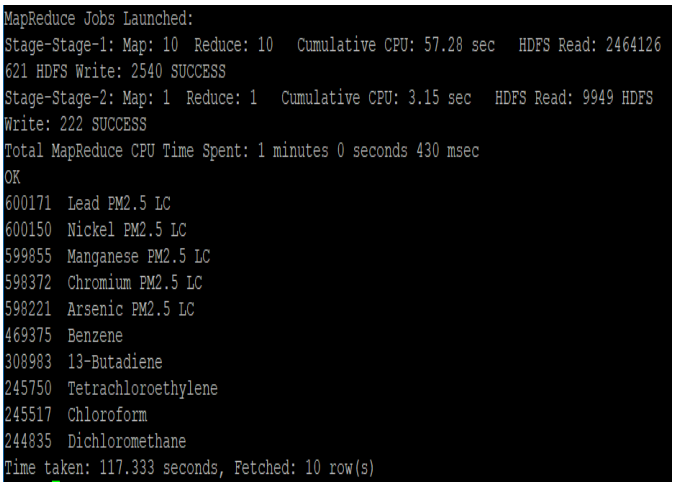

Figure 8: Displaying top 10 Pollutants

The following Hive Queries will show you the top 10 pollutants by city:

Hive> Select count(*) as pollution, parameter_name, cbsa_name from air_pollution where cbsa_name != "" group by parameter_name,cbsa_name order by pollution DESC limit 10;

Figure 9: Top 10 Pollutants by city

The following Hive Queries will show you the last 10 pollutants by city:

Hive> Select count(*) as pollution, parameter_name, cbsa_name from air_pollution group by parameter_name,cbsa_name order by pollution ASC limit 10;

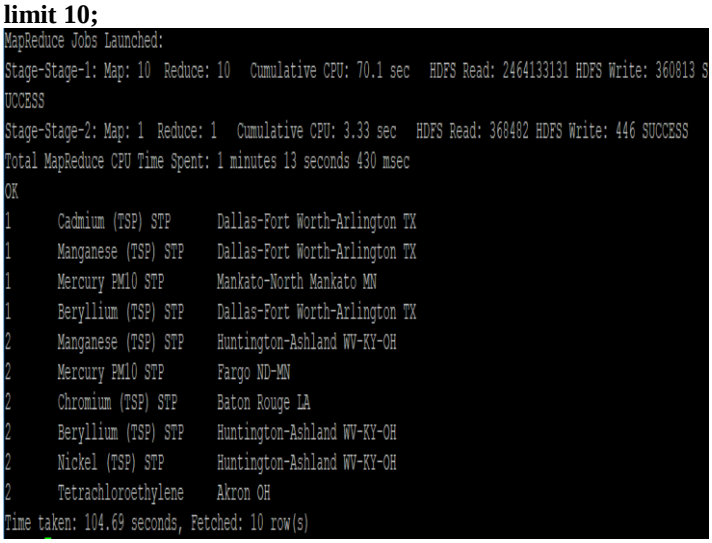

Figure 10: Last 10 Pollutants by City

The following Hive Queries will show you the top 10 pollutants by state:

Hive> Select count(*) as pollution, parameter_name, state_name from air_pollution group by parameter_name,state_name order by pollution DESC limit 10;

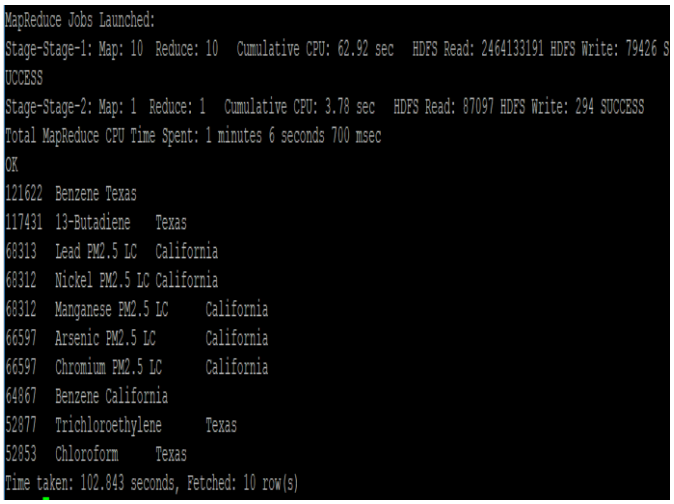

Figure 11: Top 10 Pollutants by State

The following Hive Queries will show you the last 10 pollutants by state:

Hive> Select count(*) as pollution, parameter_name, cbsa_name from air_pollution group by parameter_name,cbsa_name order by pollution ASC limit 10;

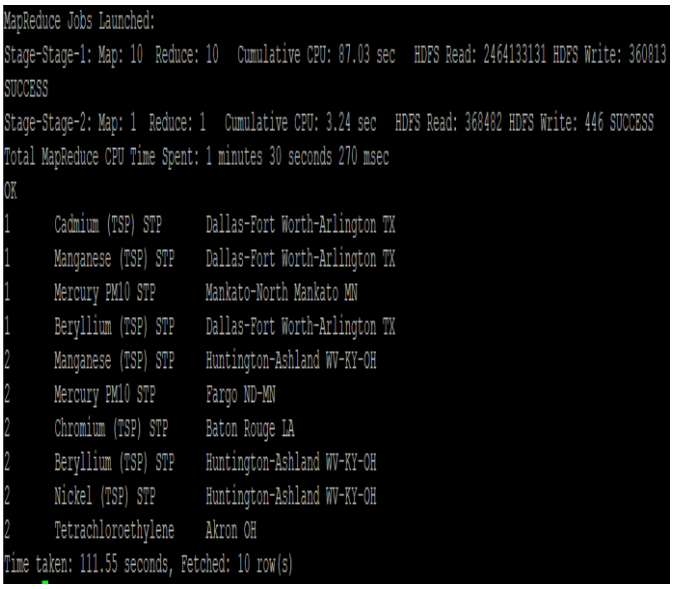

Figure 12: Last 10 Pollutants by State

The following Hive Queries will show you the top 20 pollutants by date:

Hive> Select count(*) as pollution, parameter_name, date_local from air_pollution group by parameter_name, date_local order by pollution DESC limit 20;

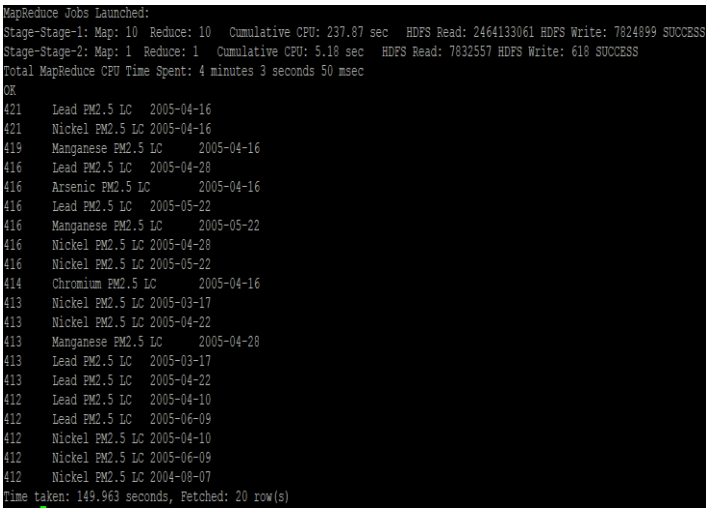

Figure 13: Top 20 Pollutants by Date

The following Hive Queries will show you the last 20 pollutants by date:

Hive> Select count(*) as pollution, parameter_name, date_local from air_pollution group by parameter_name, date_local order by pollution ASC limit 20;

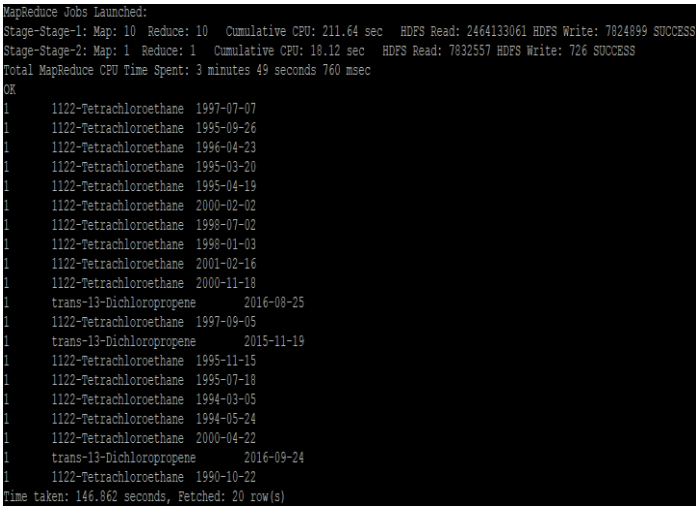

Figure 14: Last 20 pollutants by Date

2.2 Create Tables for Tableau

In these sections it will described in detail about different tables which are created in Tableau [5]. Since the dataset is too big, you need to create a table for each Hive Queries:

hive -e "use default ; Select count(*) as pollution,parameter_name,cbsa_name from air_pollution group by parameter_name,cbsa_name order by pollution ASC limit 10;"| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > last10_city.csv

Replace RED TEXT with each query that you need to create a table. Replace BLUE TEXT with the corresponding file name that you want to name it. After the

above query is done for each table, open WINSCP, and log in WINSCP as you would in putty using your Host name, user name, and password.

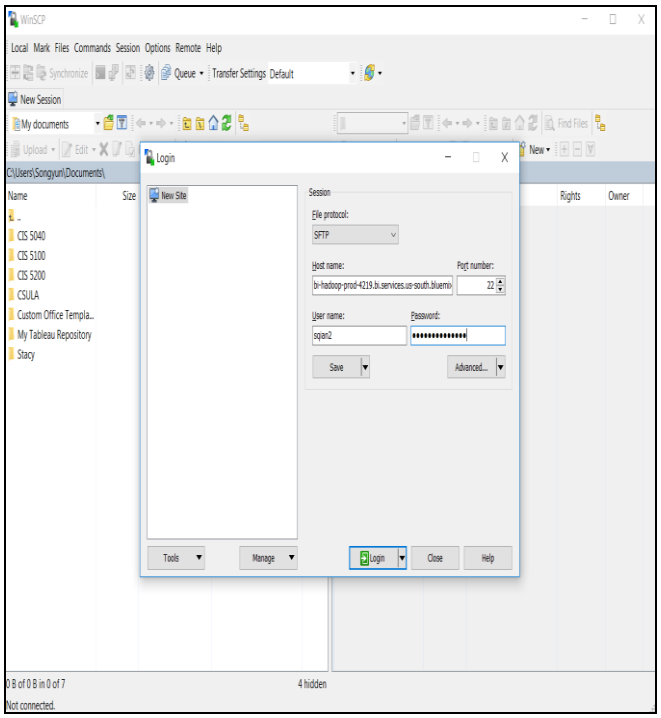

Figure 15: Create Tables for Tableau

Drag the csv files you have created to a local location on your computer on the left side of the window. You should have 7 CSV files downloaded.

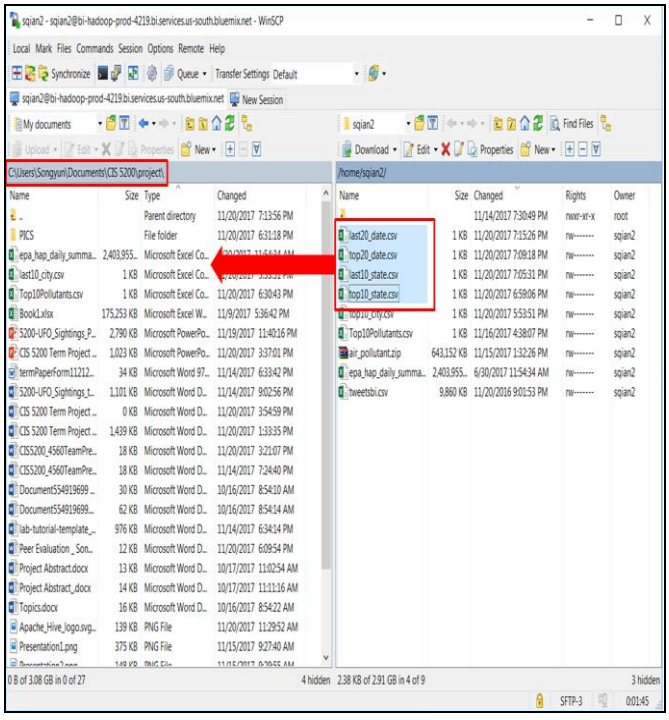

Figure 16: Download the CSV Files.

Once the csv files are downloaded, you need to open Tableau on your local computer [6]. Tableau to open data file directly from Tableau and Visualization.

Open Tableau, and open the file according to the following order.

1. Top 10 pollutants

First Rename F1 to Count, F2 to Pollutants. Select Sheet 1 next to Data Source, and drag Count to Rows and Pollutants to Columns. Drag Pollutants to Color, and you will get this chart.

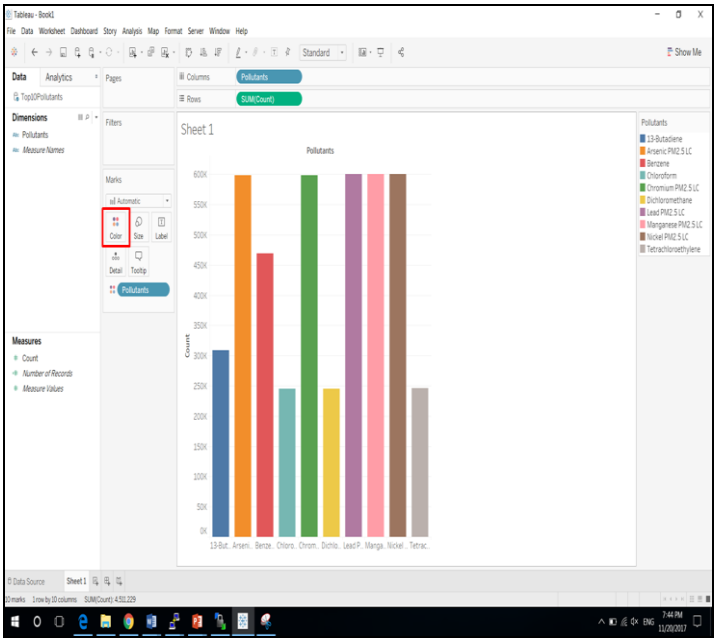

Figure 17: Top 10 Pollutants by Chart

2.Top 10 Pollutants by City

Rename F1 to Count, F2 to Pollutants, F3 to CBSA_NAME. Select Sheet 1 next to Data Source, and change CBSA's geographic role to CBSA MSA (USA). Drag Longitude(generated) to Columns, Latitude(generated) to Rows, CBSA to color, Count to Size, Pollutants to Detail. Select Show me, and select Geo Map:

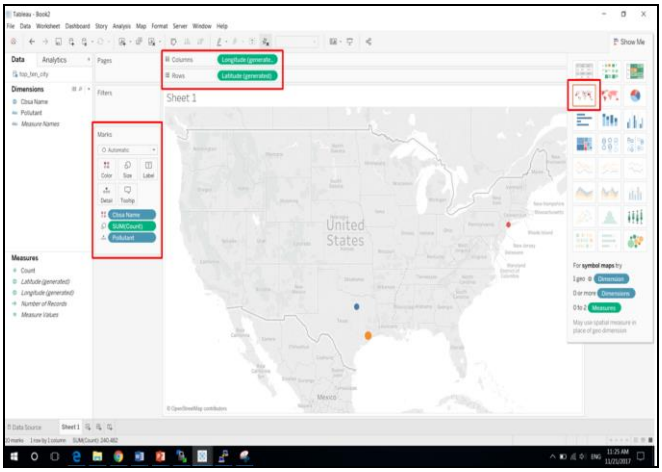

Figure 18: Top 10 Pollutants by City (Map view)

Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and CSBA to Columns and Count to Rows. Drag Pollutants to Color, you will get this:

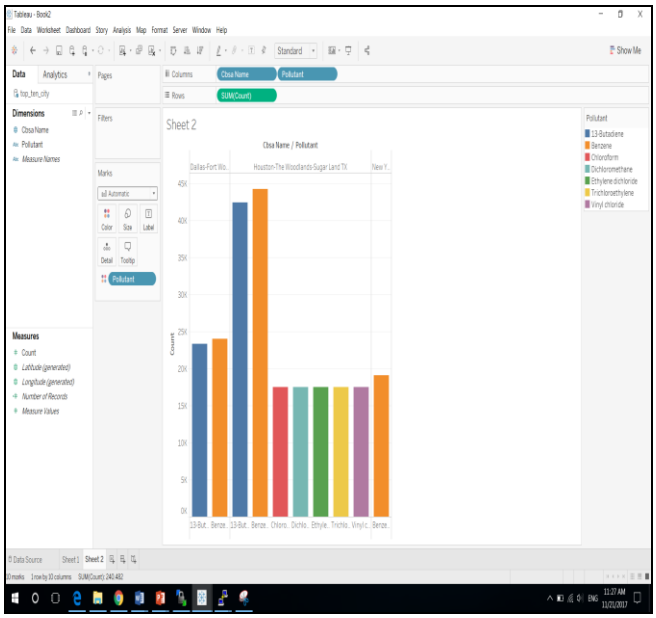

Figure 19: Top 10 Pollutants by City (chart)

3.Last 10 Pollutants by City

Rename F1 to Count, F2 to Pollutants, F3 to CBSA_NAME. Select Sheet 1 next to Data Source, and change CBSA's geographic role to CBSA MSA (USA). Drag Longitude(generated) to Columns, Latitude(generated) to Rows, CBSA to color, Count to Size, Pollutants to Detail. Select Show me, and select Geo Map.

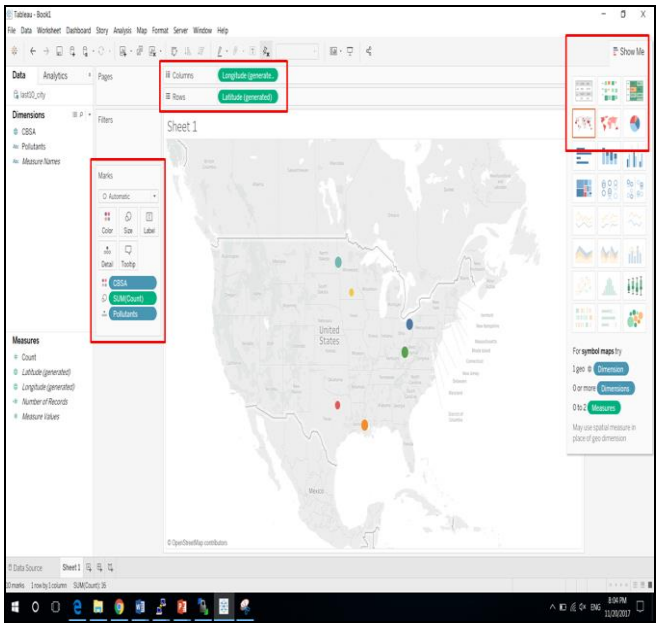

Figure 20: Last 10 Pollutants by City (Map View)

Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and CSBA to Columns and Count to Rows. Drag Pollutants to Color, you will get this:

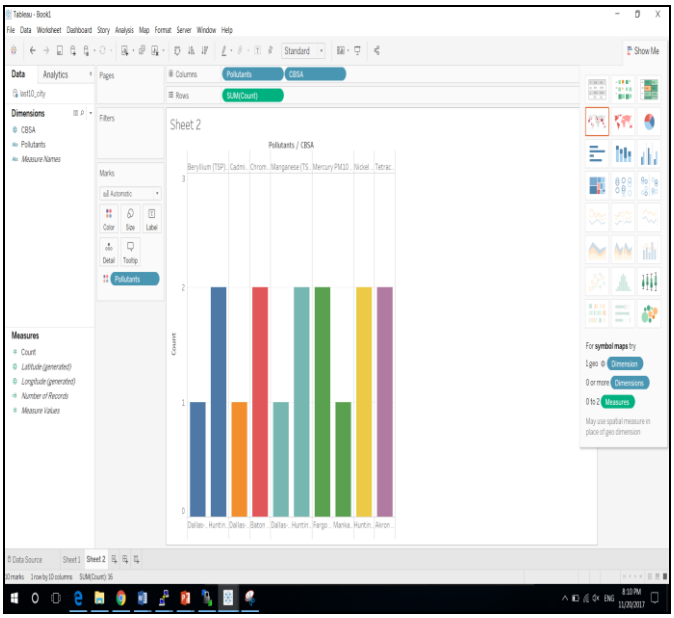

Figure 21: Last 10 Pollutants by City (chart)

4. Top 10 Pollutants by state

Change F1 to count, F2 to Pollutants, F3 to State. Select Sheet 1 next to Data Source, change State's geographical role to State/Province. Drag Longitude to Columns, Latitude to Rows, Pollutants to Color, Count to Size, State to Details. And select Geo Map.

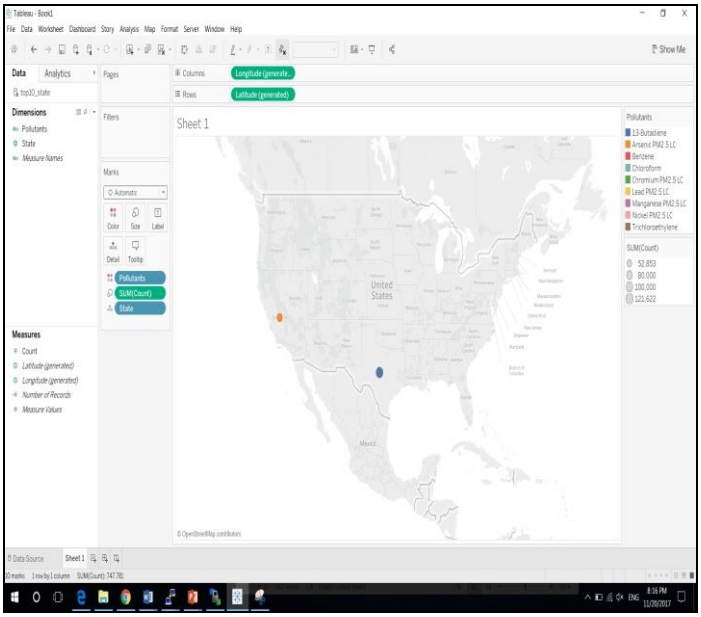

Figure 22: Top 10 Pollutants by state (Map View)

Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and State to Columns and Count to Rows. Drag State to Color, the bar chart will only generate

two colors due to the top 10 are only in Texas and California**.**

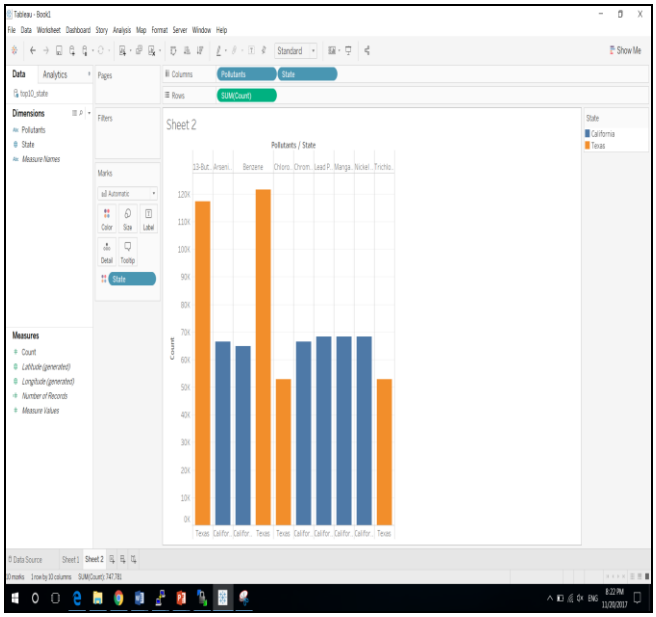

Figure 23: Top 10 Pollutants by state (chart)

5. Last 10 Pollutants by State

Change F1 to count, F2 to Pollutants, F3 to State. Select Sheet 1 next to Data Source, change State's geographical role to State/Province. Drag Longitude to Columns, Latitude to Rows, Pollutants to Color, Count to Size, State to Details and select Geo Map.

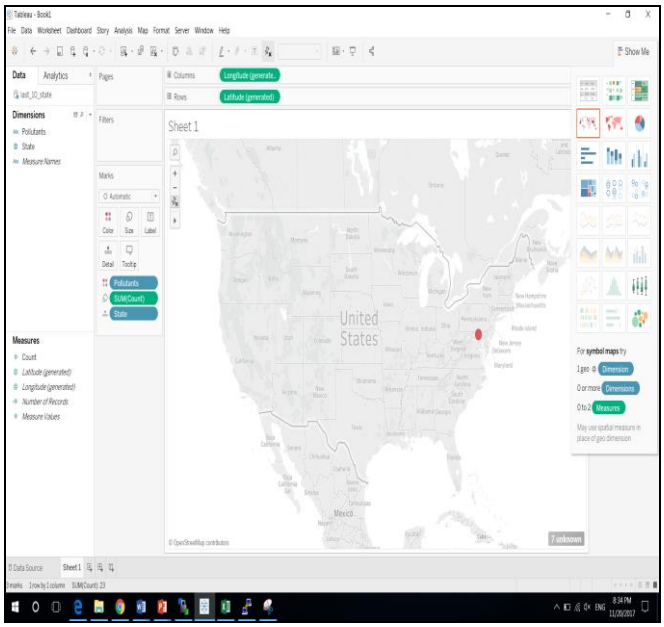

Figure 24: Last 10 Pollutants by State (Map View)

Create a new Worksheet by selecting the icon next to the Sheet 1. Drag Pollutants and State to Columns and Count to Rows. Drag State to Color, the bar chart will only generate two colors due to the top 10 are only in Country of Mexico and Maryland.

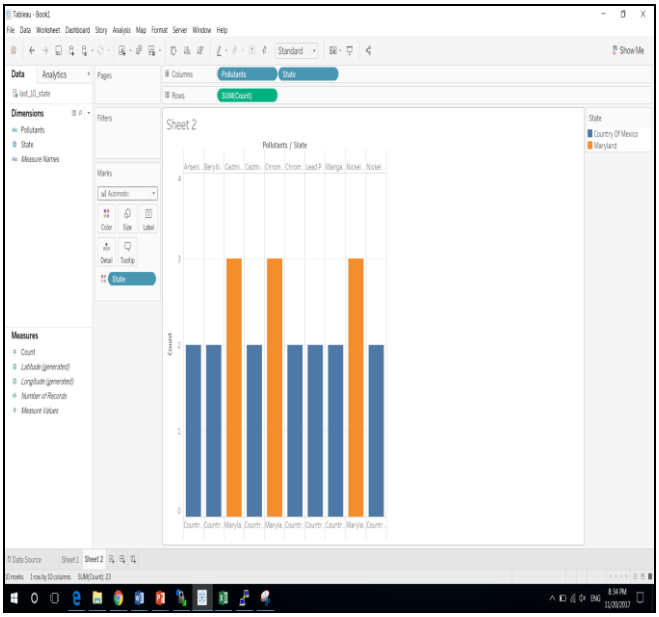

Figure 25: Last 10 Pollutants by State (chart)

6. Top 20 Pollutants by Date

Change F1 to count, F2 to Pollutants, F3 to Date. Select Sheet 1 next to Data Source, drag Date to Columns, Count to Rows, Pollutants to Color. Choose Day (May 8, 2015 format) for date, Sort Day of Date descending by Count, Choose stack bar**.**

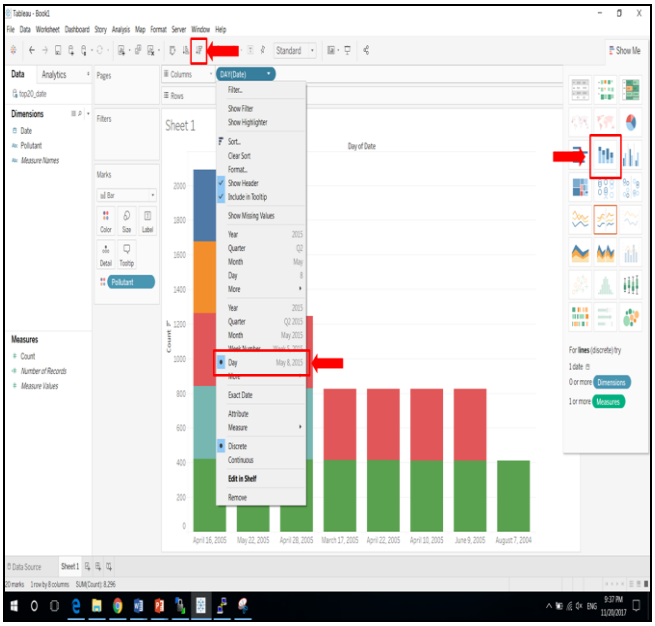

Figure 26: Top 20 Pollutants by Date (chart)

7. Last 20 Pollutants by Date

Change F1 to count, F2 to Pollutants, F3 to Date Select Sheet 1 next to Data Source, drag Date to Columns, Count

to Rows, Pollutants to Color. Choose Day (May 8, 2015 format) for date, Sort Day of Date descending by Count, Choose stack bar.

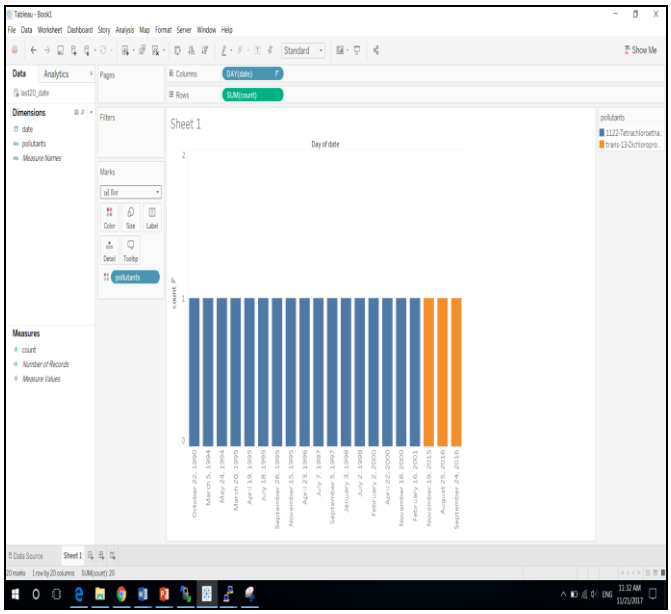

Figure 27: Last 20 Pollutants by Date

3. Conclusion

From the above analysis, we can see that interesting sets of trends and patterns exists in large data sets which helps us to get a better understanding of the data.

Recent advancement in cloud technologies helps us to harness the power of parallel processing of a cluster of computers with little investment and almost no maintenance of the underlying computer hardware.

The Air pollution is a serious problem in the world right now. Hazardous air pollutants, also known as toxic air pollutants or air toxics are poisonous for human body.

We have Analyzed data to determine which air pollutants are measured the most. We have found that pollutant Lead PM2.5 has maximum pollutant factor.

The pollutant Benzene has affected the city of The Woodlands-Sugar Land in Texas the most and the pollutant Cadmium (TSP) STP has affected the city of Dallas-Fort Worth-Arlington Texas the least.

The pollutant Benzene has affected the state of Texas the most and the pollutant Cadmium (TSP) STP the state of Texas the least.

The pollutant Lead PM2.5 has affected the United States the most and the pollutant 1122-Tetrachloroethane has affected the least since 1997 till 2017.

References

[1] Air pollution dataset of United States from 1990 to 2017. https://www.kaggle.com/epa/hazardous-air-pollutants/data

[2] Apache Hive http://hive.apache.org/

[3] Apache Hive Query Language Manual https://cwiki.apache.org/confluence/display/Hive/LanguageManual

[4] CIS 5200 Lab – Hive Twitter Sentiment Data Analysis using BigInsights of Bluemix https://app.box.com/file/96513790564

[5]CIS 5200 Lab – Analyzing social media and customer sentiment with IBM analytics engine and Tableau https://app.box.com/file/247447839736

[6]Hive SQL Syntax Checker https://sql.treasuredata.com/