# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic) CIS5200 Term Project Tutorial

#### Authors: Shwetha Seetharam , Vimal Roy Kancharla , Rania George;

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

#### Date: 12/14/2018

**Lab Tutorial**

Shwetha Seetharam ([slnu2@calstatela.edu](mailto:slnu2@calstatela.edu))

Vimal Roy Kancharla ([vkancha@calstatela.edu](mailto:vkancha@calstatela.edu))

Rania Elyas Eshak (relyase[@calstatela.edu](mailto:ykalais@calstatela.edu))

**Data Analysis on Hazardous Air Pollutants**

**Objectives**

Air Pollution is the major issue affecting the global environment. Hazardous air pollutants are toxic for human health which can cause lung cancer, respiratory infections, and heart disease. This project aims at performing data analysis and providing insights on Air Pollution Dataset using HIVE and Presenting visualization on Power BI.

In this hands-on lab, you will learn how to:

* Download the Dataset and Clean it using Excel
* Upload CSV file to Hadoop Cluster
* Create a Hive table
* SQL commands to perform the analysis.
* Visualization using Power BI.

**Platform Spec**

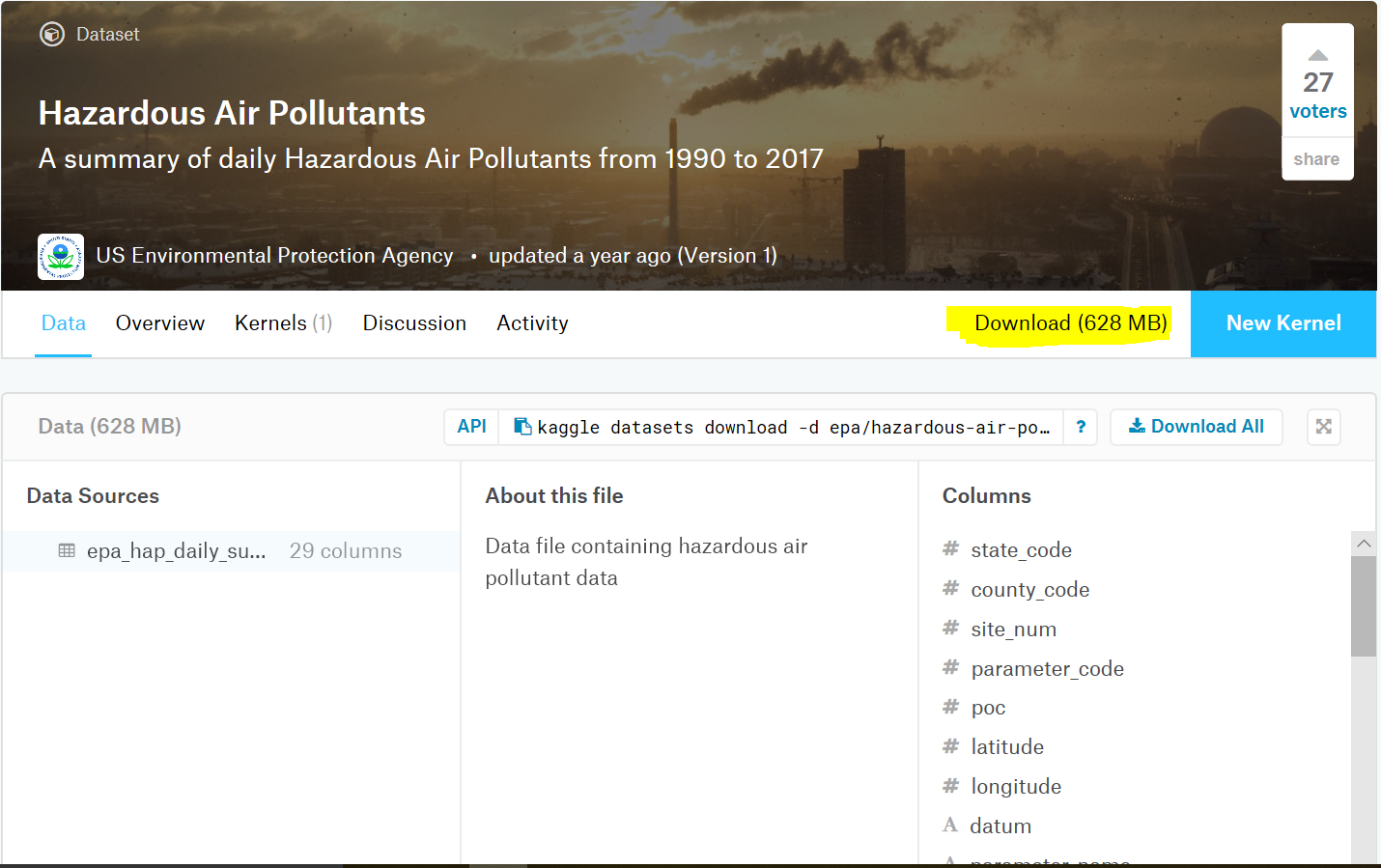
* Oracle Big Data Compute Edition: 5 nodes
* OCPUs: 10
* Memory: 150 GB
* Storage: 678 GB
* HDFS Capacity: 147 GB

Step 1: Download the Dataset

**Download the Dataset on your PC using the link given below**

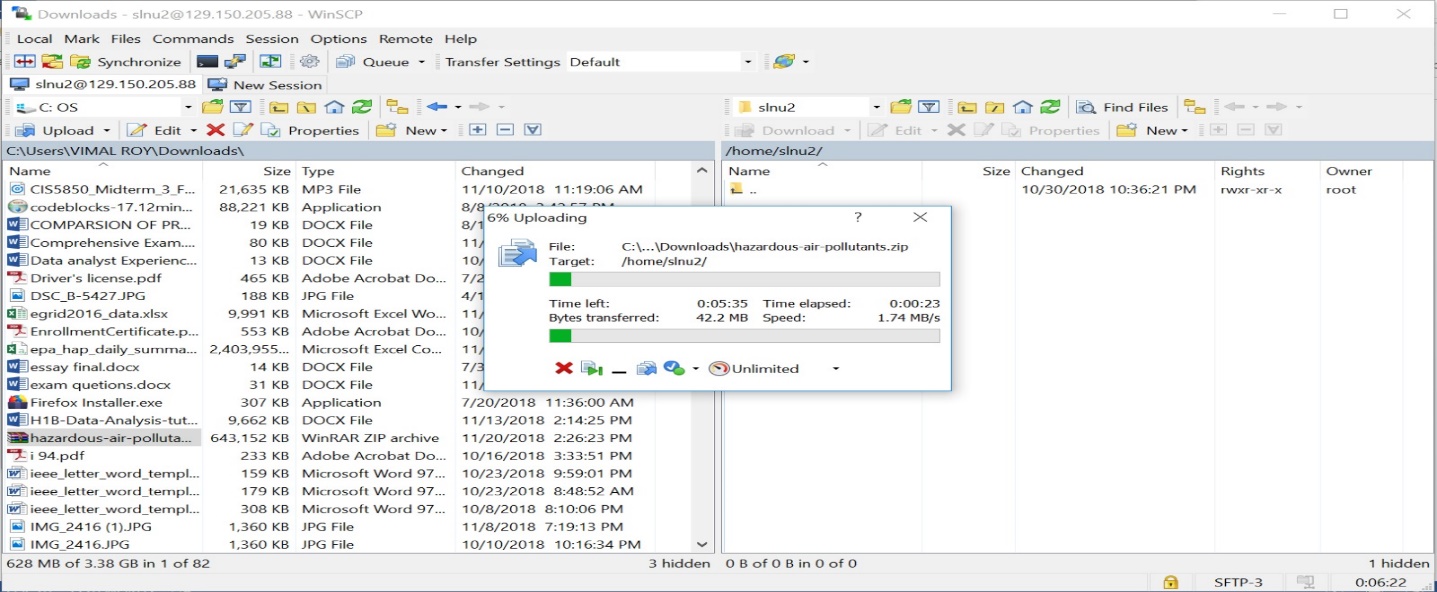
<https://www.kaggle.com/epa/hazardous-air-pollutants/data>

Keep it as a zip file.



Step 2: Upload CSV file to Hadoop Directory

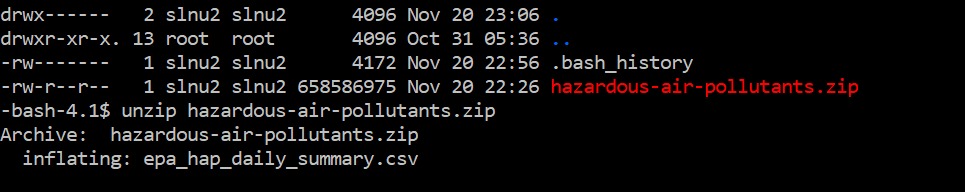
Before uploading the CSV file to Hadoop directory. We need to first transfer it to local directory using WINSCP Software.



Enter the below command in shell to unzip and view the uploaded **hazardous-air-pollutants.zip**

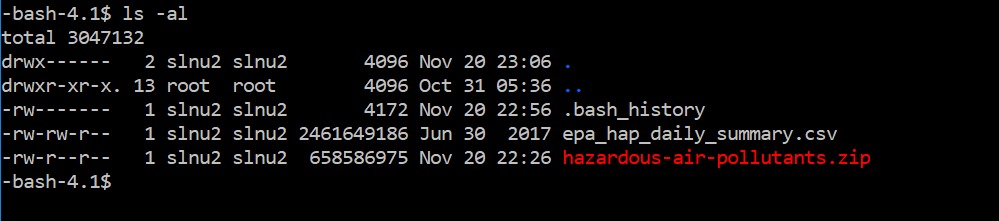
$ unzip hazardous-air-pollutants.zip

$ ls -al



Enter the below command in shell to view the uploaded **epa\_hap\_daily\_summary.csv**

$ ls -al

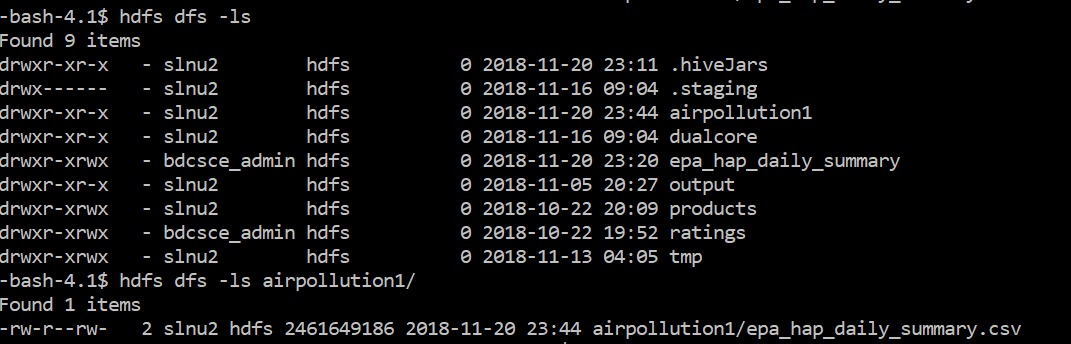


Now, you have to upload this CSV file to hdfs folder. Run the following HDFS commands to create and list the **airpollution1** directory in HDFS.

$ hdfs dfs -mkdir airpollution1

$ hdfs dfs -put epa\_hap\_daily\_summary.csv airpollution1/

$ hdfs dfs -ls airpollution1/



Step 3: Create Hive table to query Air Pollution data:

Open hive shell environment as follows:

$ beeline

Beeline link

!connect jdbc:hive2://cis5200s3-bdcsce-4.compute-608214094.oraclecloud.internal:2181,cis5200s3-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200s3-bdcsce-3.compute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin

Use this command to choose your database name.

0:jdbc:hive2://cis5200s3-bdcsce-4.compute-60> use slnu2;

(**Note:** Don’t forget to replace the red part with your account name.)

The following Hive statement creates an external table for Air Pollution data. External tables preserve the data in the original file format, while allowing Hive to perform queries against the data within the file.

In the hive shell CLI, you need to copy and paste the following HiveQL code to create an external table **airpollution.**

CREATE TABLE IF NOT EXISTS airpollution

(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/slnu2/airpollution1/epa\_hap\_daily\_summary'

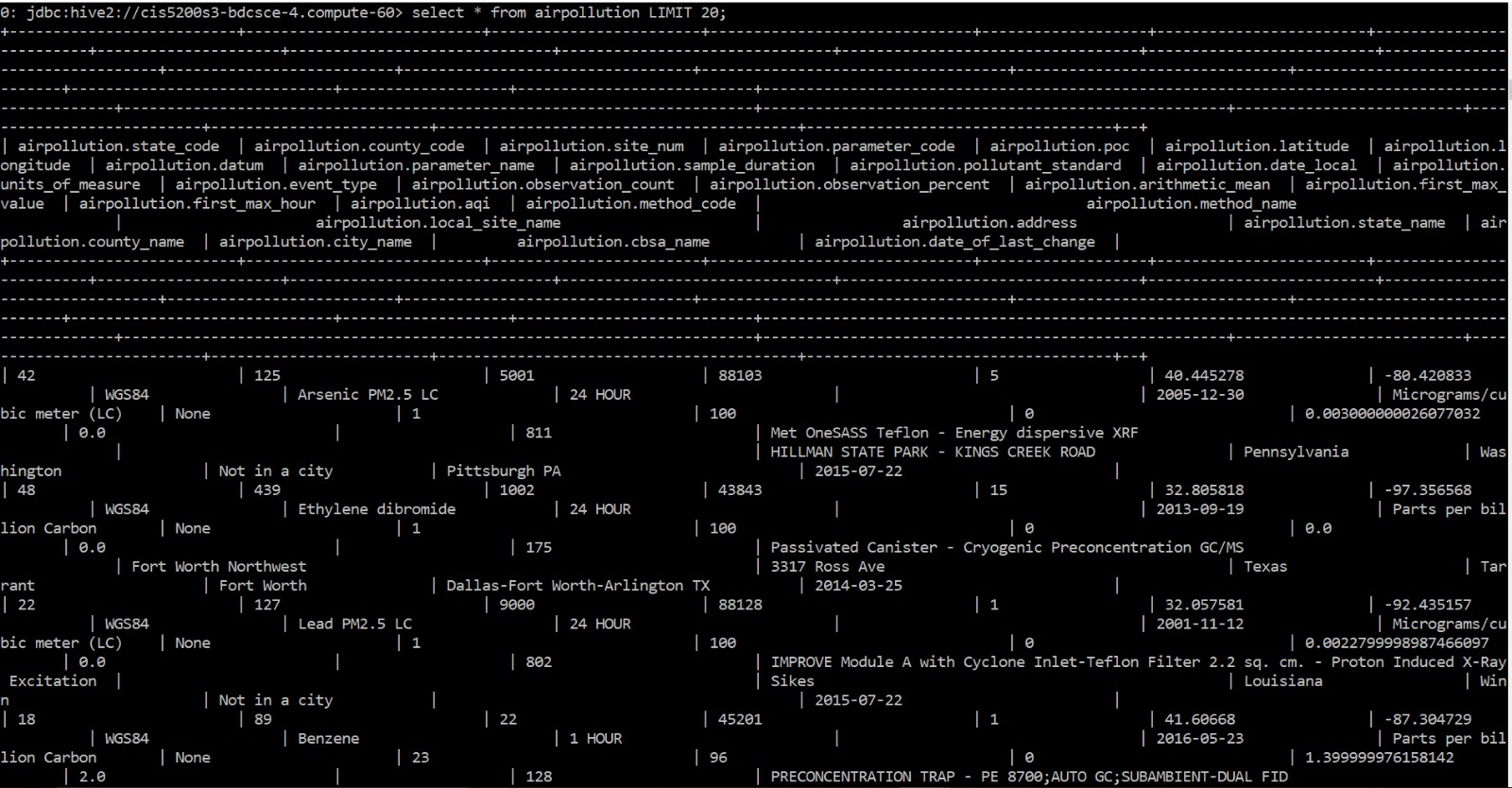
TBLPROPERTIES ('skip.header.line.count'='1');

Then in the Hive shell, you need to check if the table “**airpollution**” is shown.

0:jdbc:hive2://cis5200s3-bdcsce-4.compute-60> show tables;

To query the contents of **airpollution** table

0:jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select \* from airpollution LIMIT 20;



$ hdfs dfs -mkdir ./q1

$ hdfs dfs -mkdir ./q2

$ hdfs dfs -mkdir ./q3

$ hdfs dfs -mkdir ./q4

$ hdfs dfs -mkdir ./q5

$ hdfs dfs -mkdir ./q6

$ hdfs dfs -ls .

Create different directories in the hdfs to store the output of the queries

Step 4: Querying on Dataset

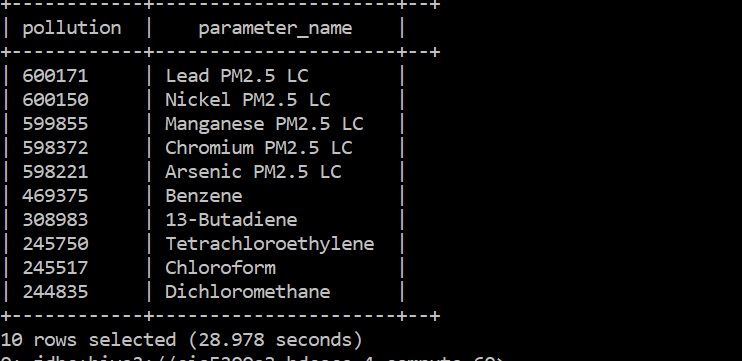
**Query 1: Top 10 Pollutants**.

For “Air pollution” details:

select count(\*) as pollution,parameter\_name from airpollution group by parameter\_name

order by pollution DESC limit 10;

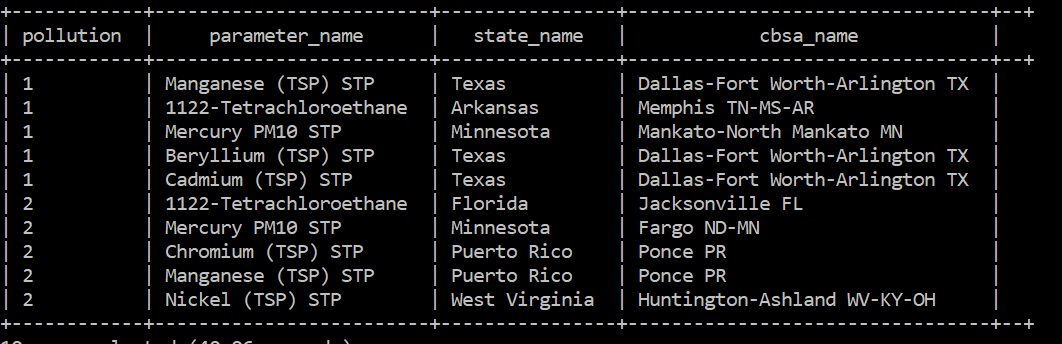
**OUTPUT:**



**Query 2: Top 10 pollutants by state**

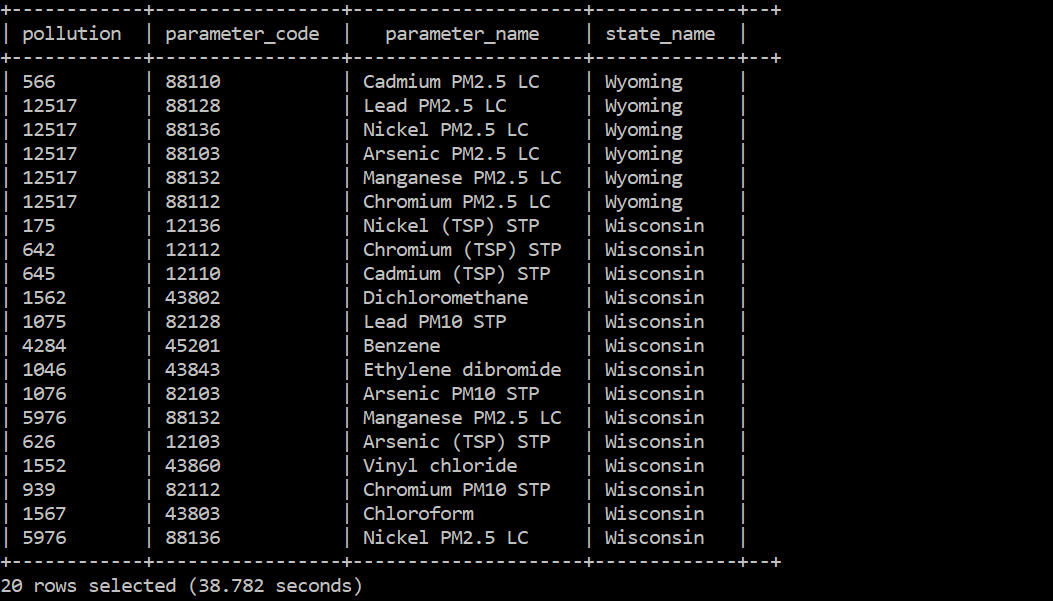
select count(\*) as pollution,parameter\_name,state\_name,cbsa\_name from airpollution where cbsa\_name !=”” group by parameter\_name,state\_name,cbsa\_name order by pollution ASC LIMIT 10;

**OUTPUT:**



**Query 3: bottom 20 pollutants name ordered state-wise**.

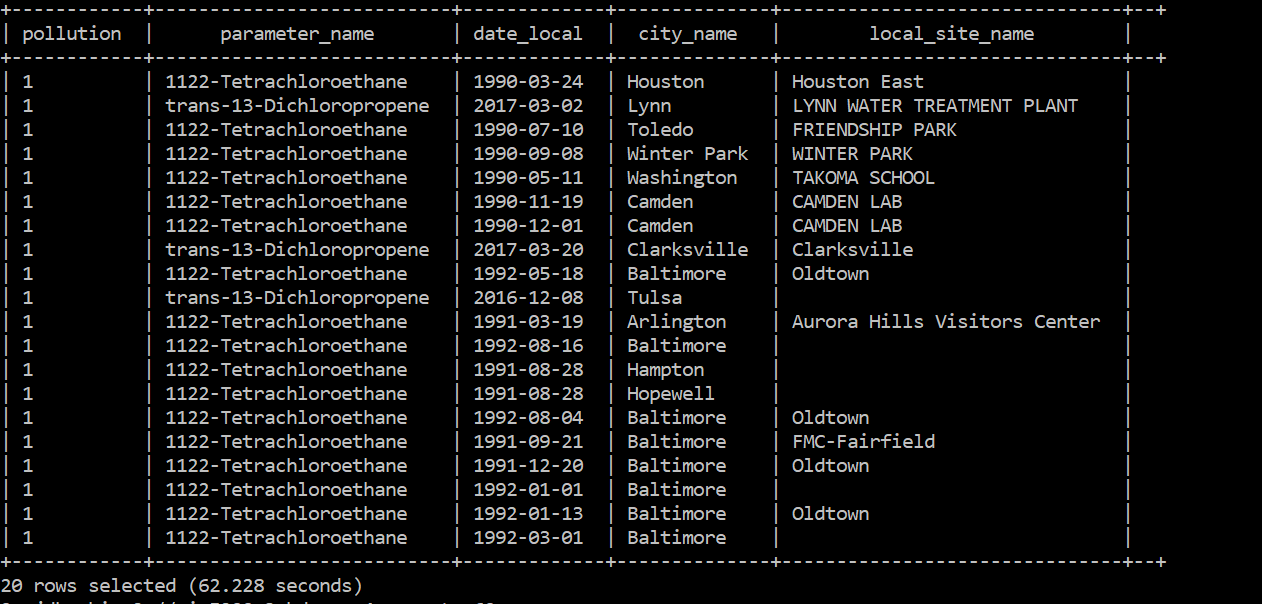
select count(\*) as pollution,parameter\_code,parameter\_name,state\_name from airpollution where state\_name!="" group by parameter\_name,state\_name,parameter\_code order by state\_name DESC LIMIT 20;



**Query 4: incident occurred at a site and the pollutants ordered date wise.**

select count(\*) as pollution,parameter\_name, date\_local,city\_name,local\_site\_name from airpollution group by parameter\_name,date\_local,city\_name,local\_site\_name order by pollution ASC LIMIT 20;

**OUTPUT:**

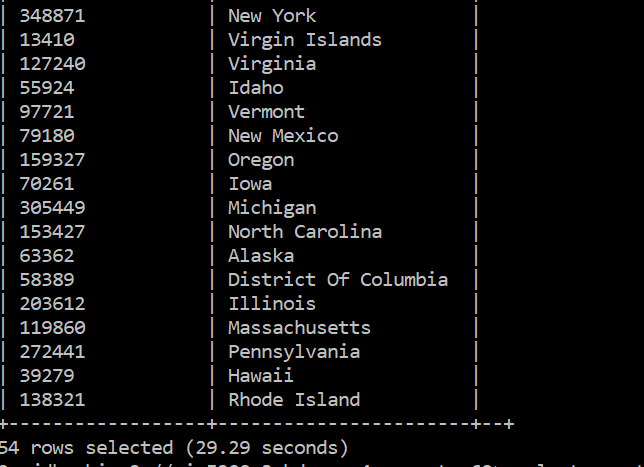


**Query 5: Total Number of pollutants in each state.**

select count(\*) as pollution\_count, state\_name from airpollution group by state\_name limit 60;

**OUTPUT:**





**Query 6: Top 10 pollutants ordered observation percent wise excluding “California”.**

select count(\*) as pollution,parameter\_code,parameter\_name,state\_name,observation\_count,observation\_percent from airpollution where state\_name!="California" group by parameter\_name, parameter\_code, observation\_count, observation\_percent, state\_name order by observation\_percent ASC LIMIT 10;

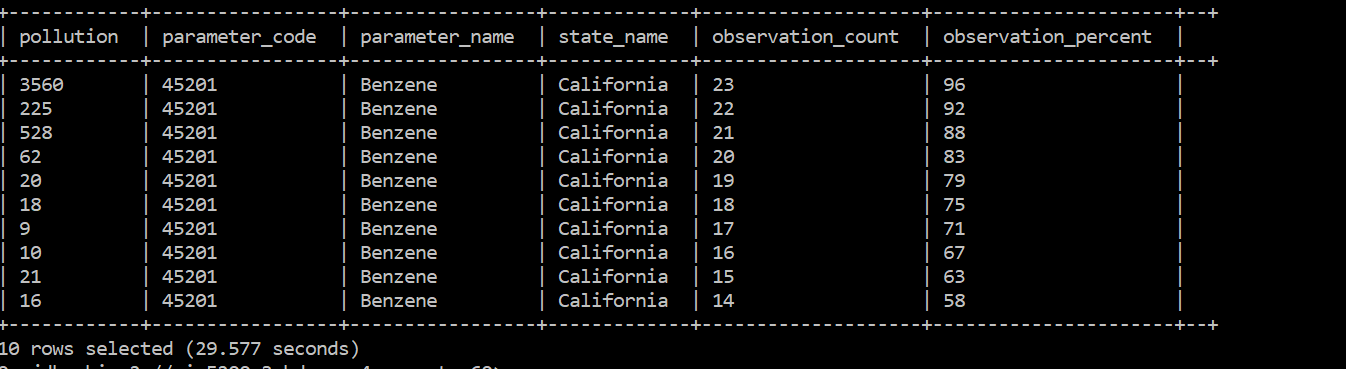
**OUTPUT:**



**Query 7: Pollutants ordered observation count wise in “California”.**

Select count(\*) as pollution,parameter\_code,parameter\_name,state\_name,observation\_count,observation\_percent from airpollution where state\_name ="California" group by parameter\_name, parameter\_code, observation\_count, observation\_percent, state\_name order by observation\_count ASC LIMIT 10;

**OUTPUT:**



STEP 5: Creating Hive tables and Downloading the data into your PC

Since the dataset is too big, you need to create a table for each Hive Queries:

CREATE TABLE top10pollutants\_name

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

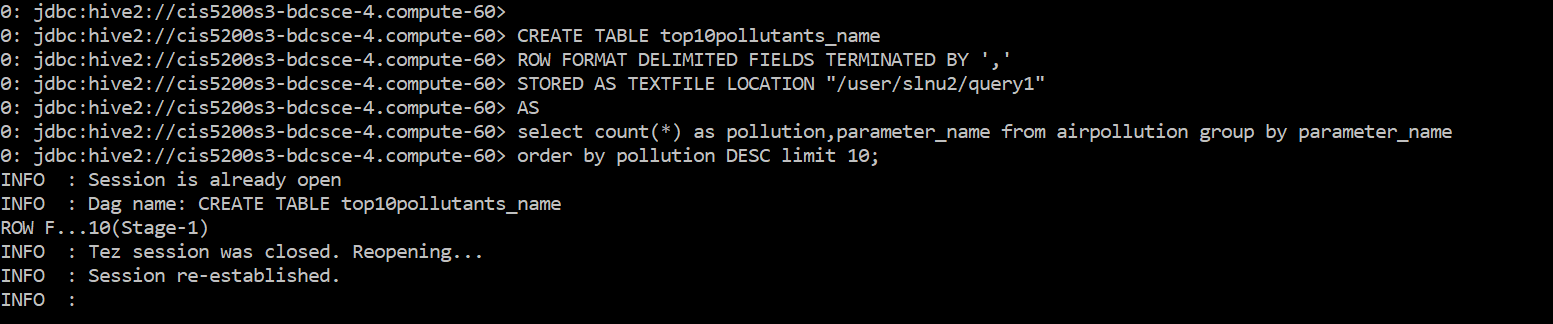
STORED AS TEXTFILE LOCATION "/user/slnu2/query1"

AS

select count(\*) as pollution,parameter\_name from airpollution group by parameter\_name

order by pollution DESC limit 10;

Replace RED TEXT with each query that you need to create a table. Replace YELLOW TEXT with the corresponding table name that you want to name it and Replace GREEN TEXT with the text file storage location.

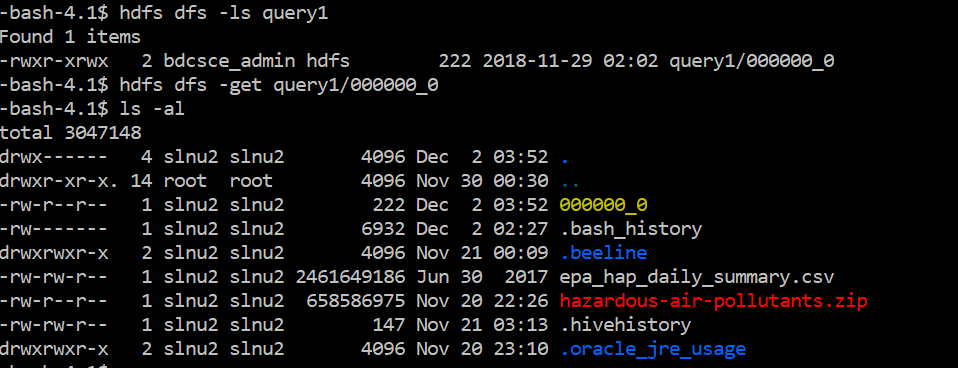


Download the output file 000000\_0 at the HDFS path ‘user/slnu2/query1’

Hdfs dfs –get query1/000000\_0

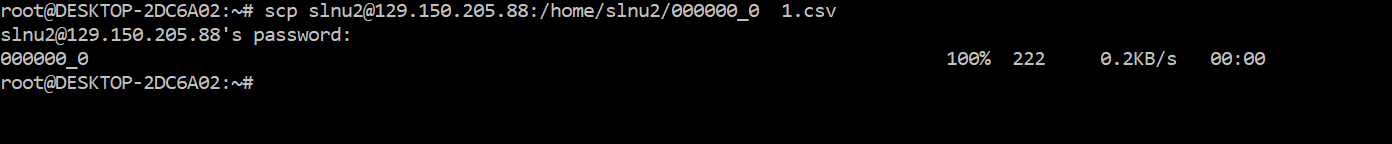
ls -al

Hdfs dfs –ls query1



After the above query is done for each table, open a new terminal

scp [slnu2@129.150.205.88](mailto:slnu2@129.150.205.88) :/home/slnu2/000000\_0 1.csv



The RED TEXT shows the downloaded data file.

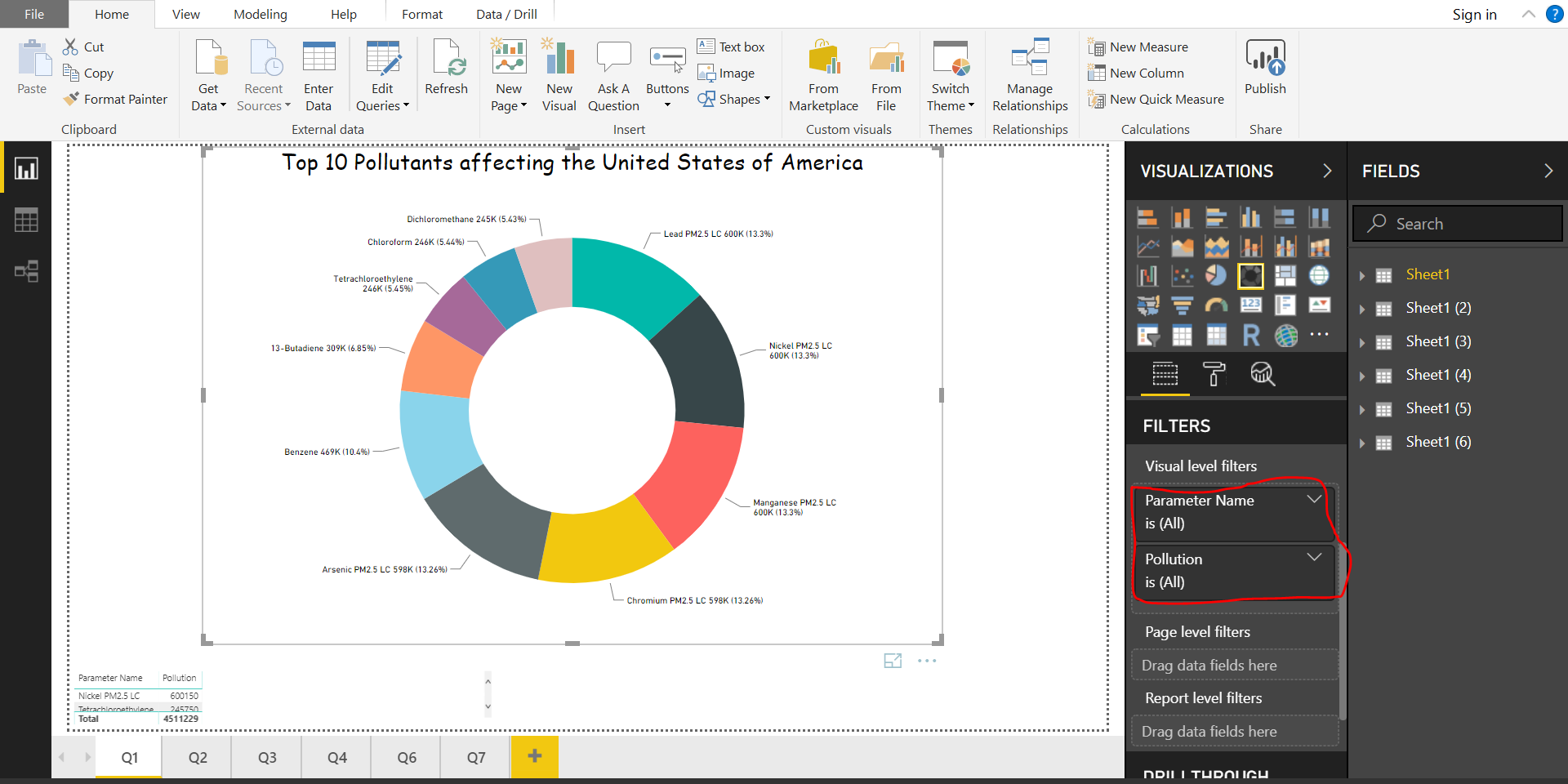
STEP 6: Visualization of Data

**Query 1: The top 10 Pollutants which are affecting the United States of America.**

Open the CSV file “1.csv” in Power BI. After loading the data in Power BI, we will click of sheet 1 as shown in the picture below to create our visualization.



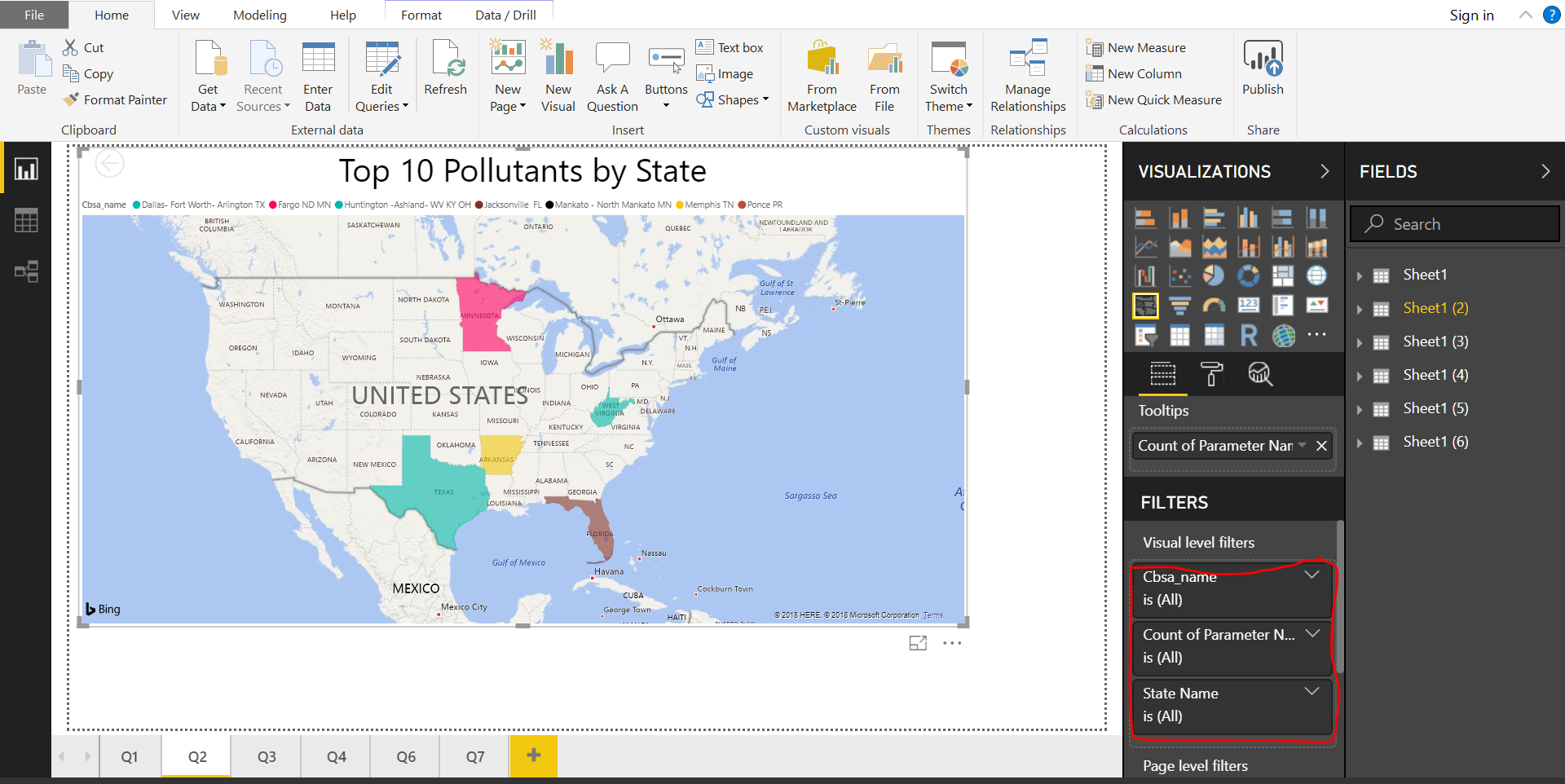
After clicking on sheet 1 worksheet, we drag the parameter names and pollution from fields to create the Donut Chart.



**Query 2: Top 10 pollutants by state in USA.**

Open the CSV file “2.csv” in Power BI. After loading the data in Power BI, we will click of sheet 2 as shown in the picture below to create our visualization.

After clicking on sheet 2 worksheet, we drag the change to Count parameter names, State Name and cbsa\_name from fields to create the Filled Maps.



**Query 3: Average Pollution rate of bottom 20 pollutants**

Open the CSV file “3.csv” in Power BI. After loading the data in Power BI, we will click of sheet 3 as shown in the picture below to create our visualization.



Go to Sheet 3, Drag and drop **Parameter name, State Name & Pollution change it to average**.

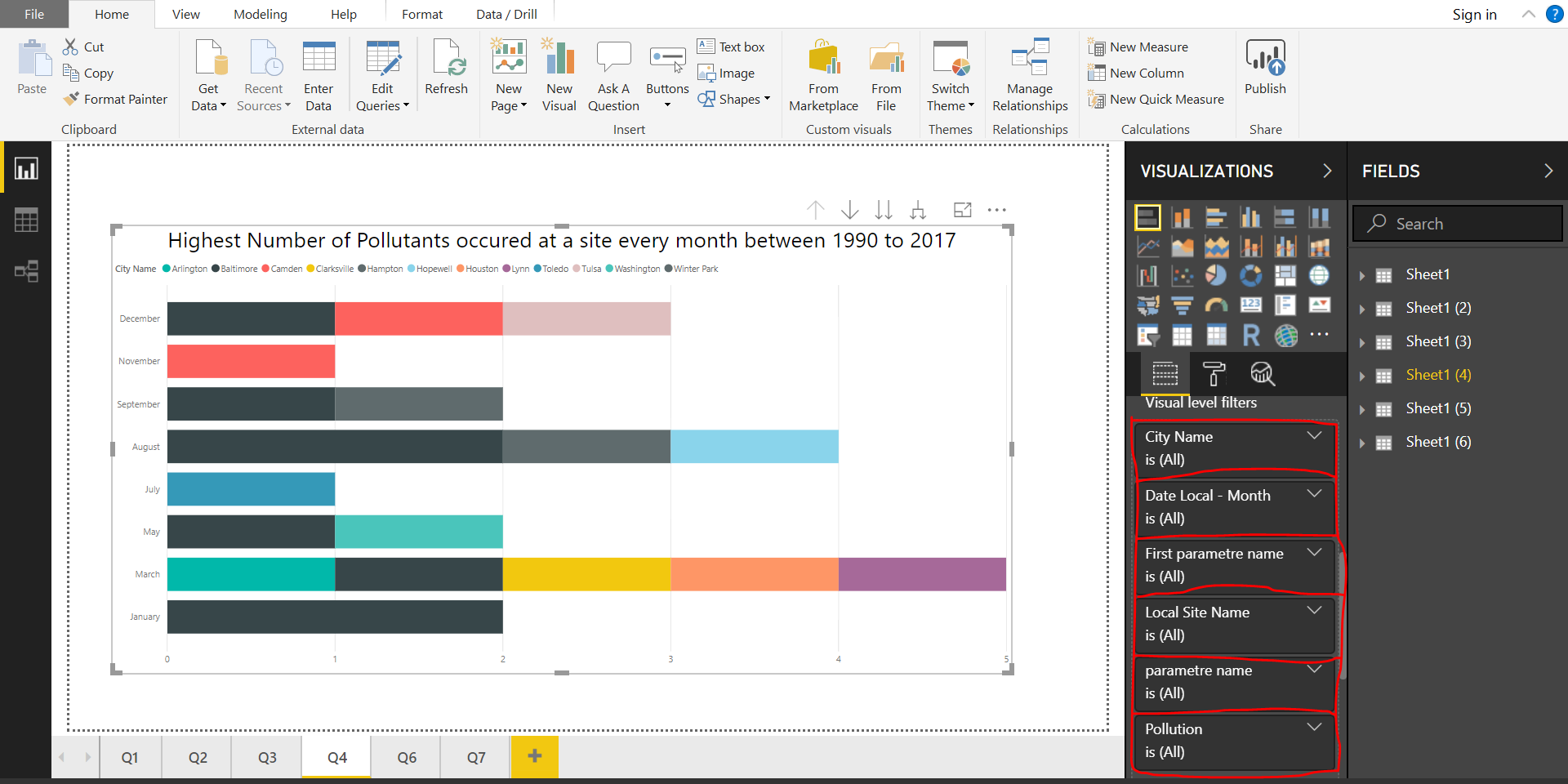
Choose the graph type as Statcked Bar Plot.

Based on your requirements, you can select pollution parameter name in the visual level filters and we have enabled the drill through option which enhances the visualization.

**Query 4: Highest number of pollutants occurred at a site every month between 1990 to 2017.**

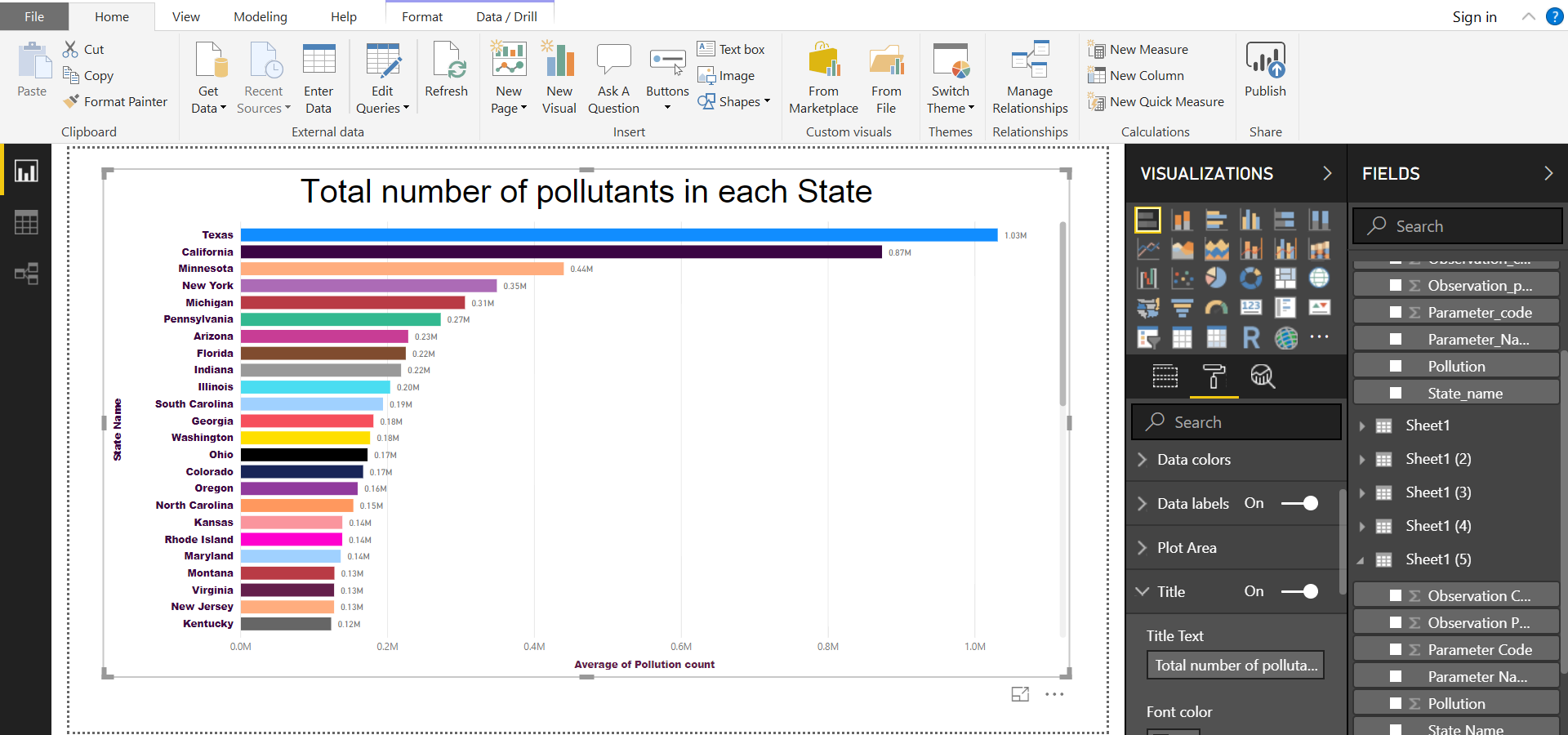
Open the CSV file “4.csv” folder in excel.

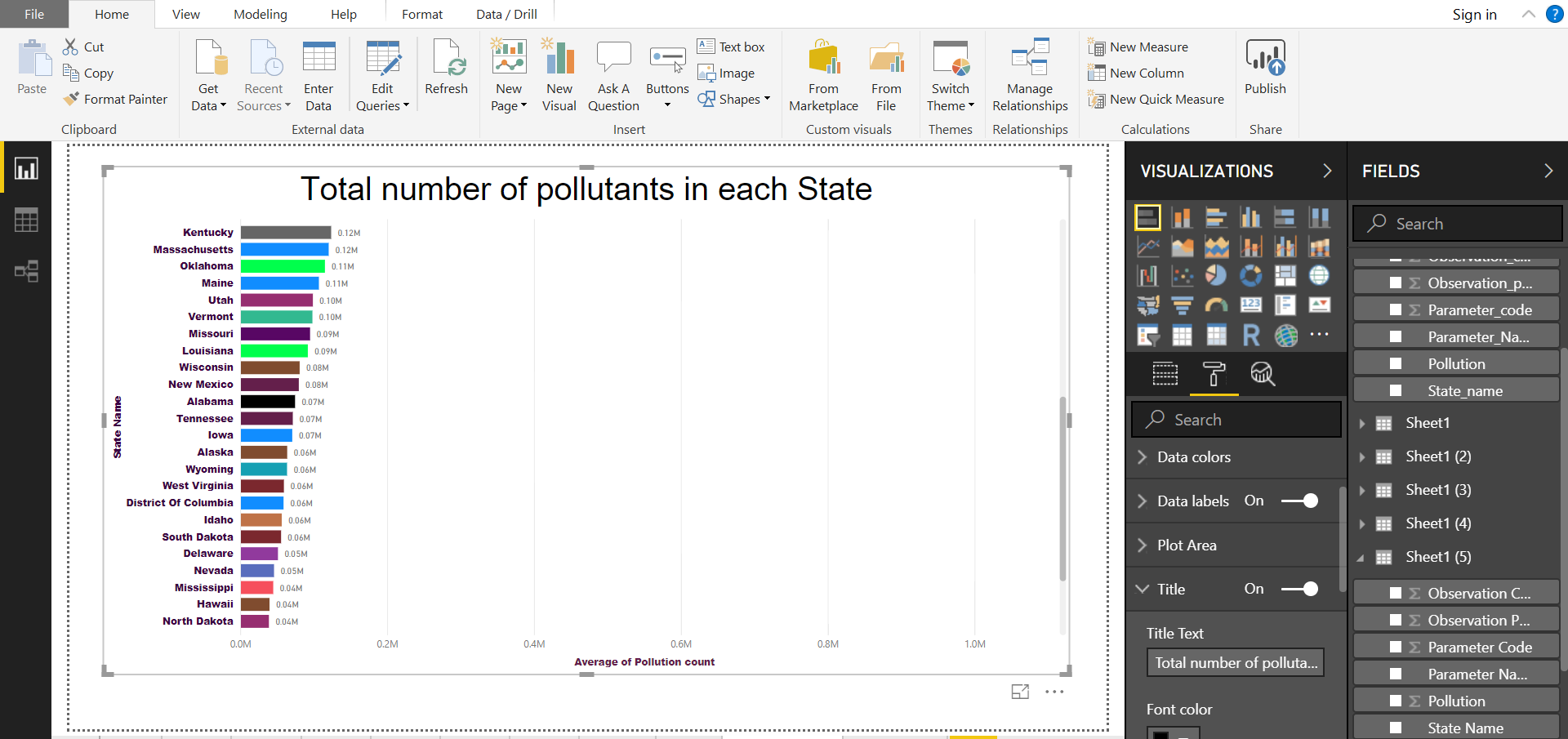
Drag and drop City name, parameter namey Name. Date Local and filter it to Month only local Site Name, Parameter name & pollution. Choose Clustered bar chart. And add data colors for differentiation purposes.



**Query 5: Total number of pollutants in each state.**

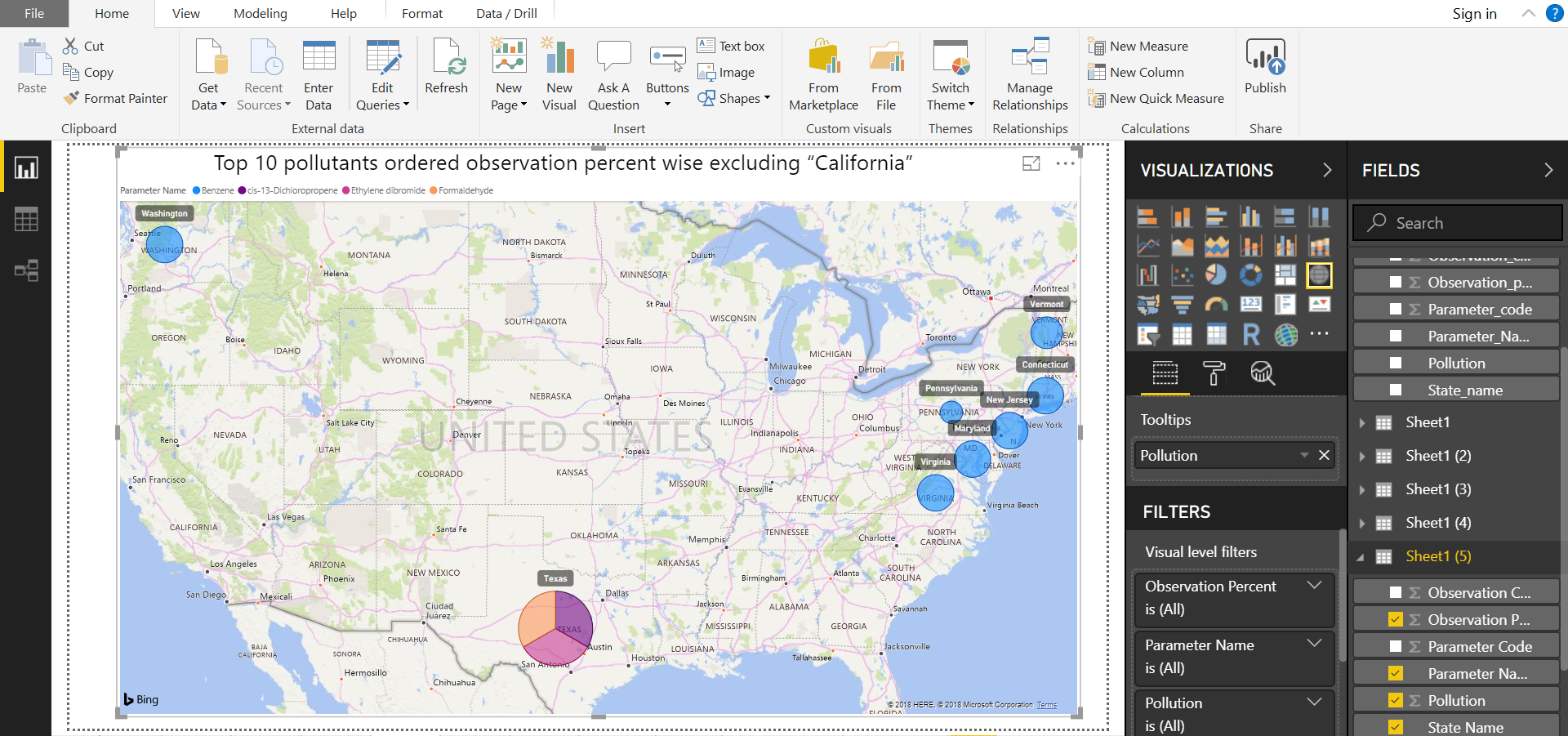
Open the CSV file “5.csv” in Power Bi. After loading the data in Power BI, we will click of sheet 5 as shown in the picture below to create our visualization. Drag State name & pollution (Convert it to Average).





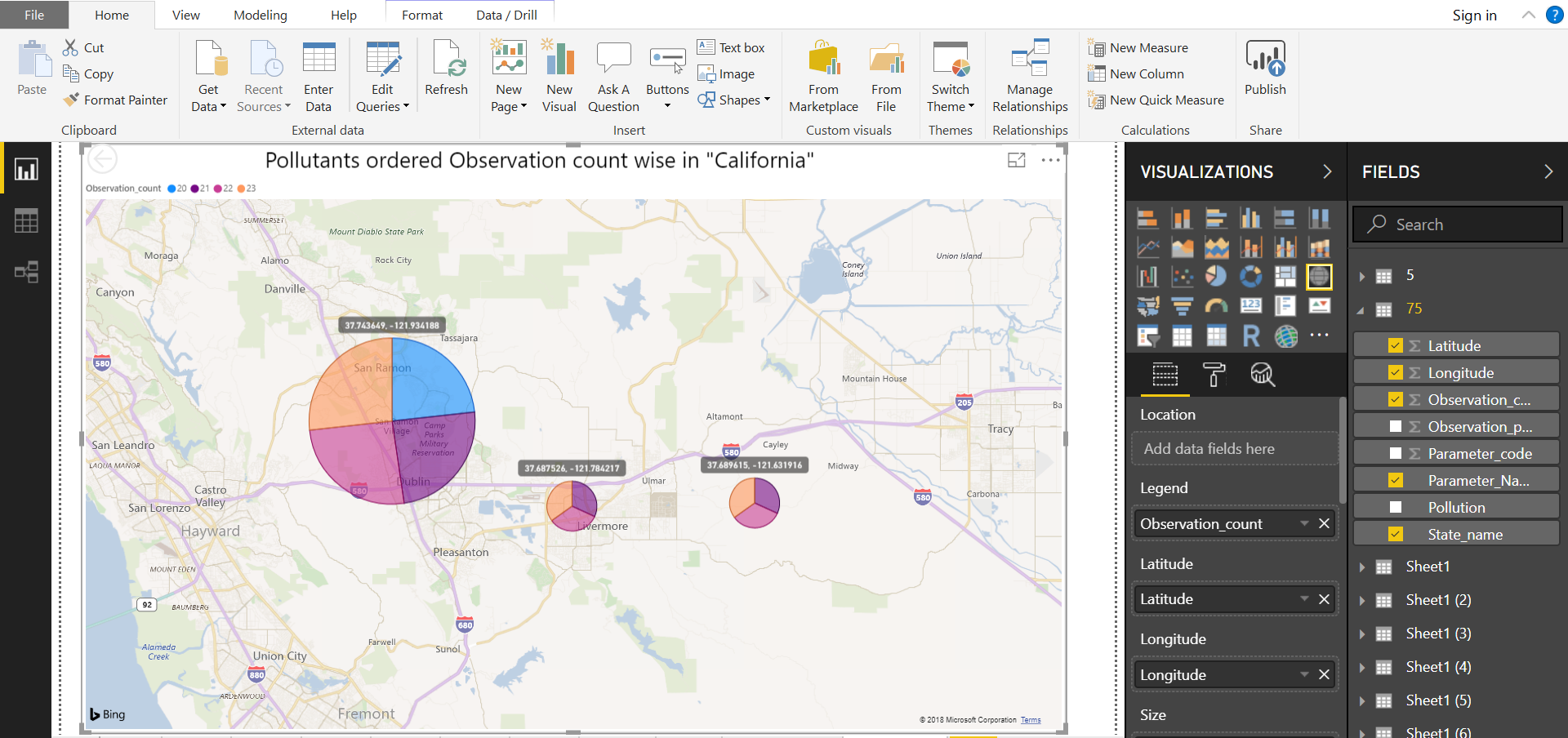
**Query 6: Top 10 Pollutants ordered observation percent wise excluding “California”.**

Open the CSV file from “6.csv” in Power BI. The Map lists the top 10 highly pollutant state.



**Query 7: Pollutants ordered Observation count wise in “California”.**

Open the CSV file from “7.csv” file in Powerbi and choose filledMaps.



Summary

In this tutorial, we learned how Oracle cloud can be used to make analysis on a data using Apache Hive. We went through a flow to understand how the raw data is first uploaded to HDFS, and then loaded to Hive tables for performing queries. And, we learned how to import the results of Hive queries into Power BI and make visualizations using the same.

References

[1] http://hive.apache.org/

[2] https://app.powerbi.com

[3] https://www.kaggle.com/epa/hazardous-air-pollutants/data

[4]Lab Tutorial from CIS 5200

[5]Classmate: Monica Mishra