**FIRE DEPARTMENT CALL LOG ANALYSIS USING AZURE HD INSIGHTS AND TABLEAU**

By

Shanmathi Arul Murugan ([sarulmu@calstatela.edu](mailto:sarulmu@calstatela.edu)), Ashwin.P.Karthik([akarthi@calstatela.edu](mailto:akarthi@calstatela.edu)) and Kaushik Sridharan ([ksridha@calstatela.edu](mailto:ksridha@calstatela.edu))

12/10/2016

**Objective:**

This tutorial is about the Fire Department Call Log analysis using Hadoop cluster in Azure HD insights and visualizations using tableau. You will gain working knowledge about creating a Hadoop cluster in Azure HD insights, using Cloudberry explorer to import data into HDFS and visualizing the analyzed data in tableau.

The data set has a rich data on call log analysis from year 2000 – 2004. It is categorized by the Call Number, Incident Number, Call Type, Call Date, Day, Month, Year, Address, Zip code, Box, Priority, ALS Unit, Number of Alarms, Unit Type, Neighborhood, Locations. The analyzation results will give us an idea about the year that received the most number of calls, most incident prone area, leading incident that happened over the four years, the incident that required least dispatch unit etc.

**Technologies Used:**

* Microsoft Azure HD Insight
* Cloudberry Explorer for Azure Blob Storage
* Tableau

**Pre-requisites:**

**Setup Tableau with the License:**

If you have the Tableau software with active license you are good to go. Else you can get the Tableau for students from the below link,

<http://www.tableau.com/academic/students>

You can use your institution ID to register and get free tableau desktop license for 1-year.

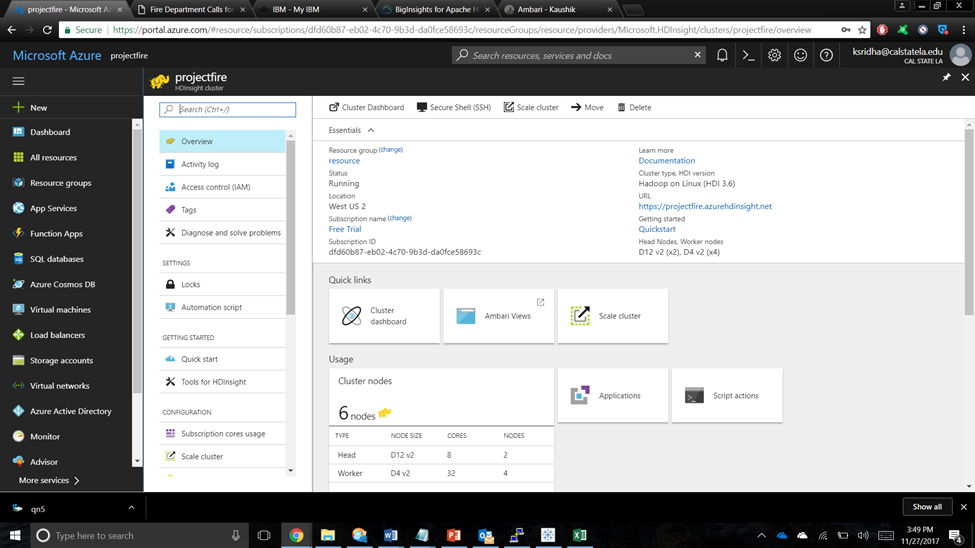
**Creating a Hadoop Cluster on Microsoft Azure HD insights:**

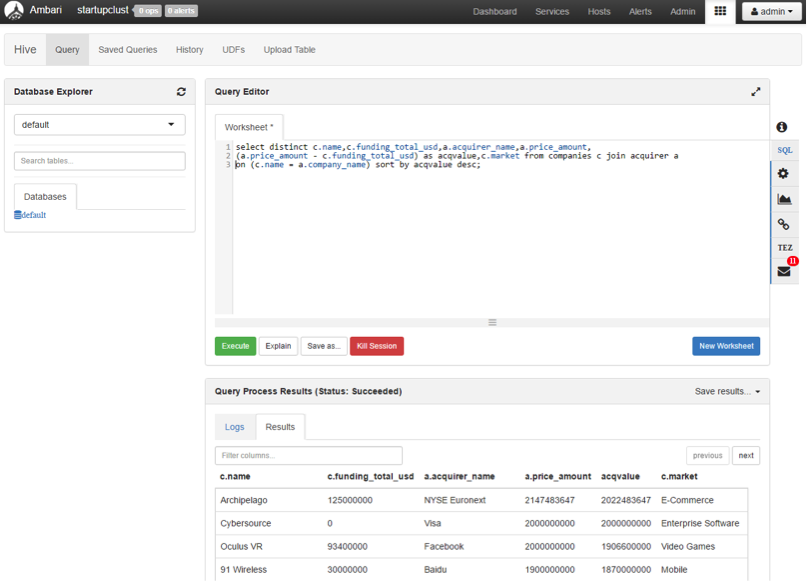
You have to sign up Microsoft Azure cloud computing account. You may go to https://azure.microsoft.com/en-us/free/ in order create a 30 days free trial with 200$ credit limit. (Make sure you sign up with free trail and **not with Pay-as-you-go subscription** which will charge you from your card, once the free credit is over)

* In Azure portal, Click New 🡪 Data + Analytics 🡪 HDInsight
* Choose the cluster type as Hadoop and give the authentication details.
* Choose the hardware configuration as below,

|  |
| --- |
| Number Of Head Nodes = 2 | CPU = 8 Cores |
| Number Of Worker Nodes = 2 |CPU = 8 Cores |
| Processor - 2.4 GHz Intel Xeon® E5-2673 v3 |
| Ram – 24 GB |
| Disks – 16 |
| Local SSD – 400 GB |

* The cluster creation may take a while.
* Once the cluster is created you can see the cluster dashboard as below,

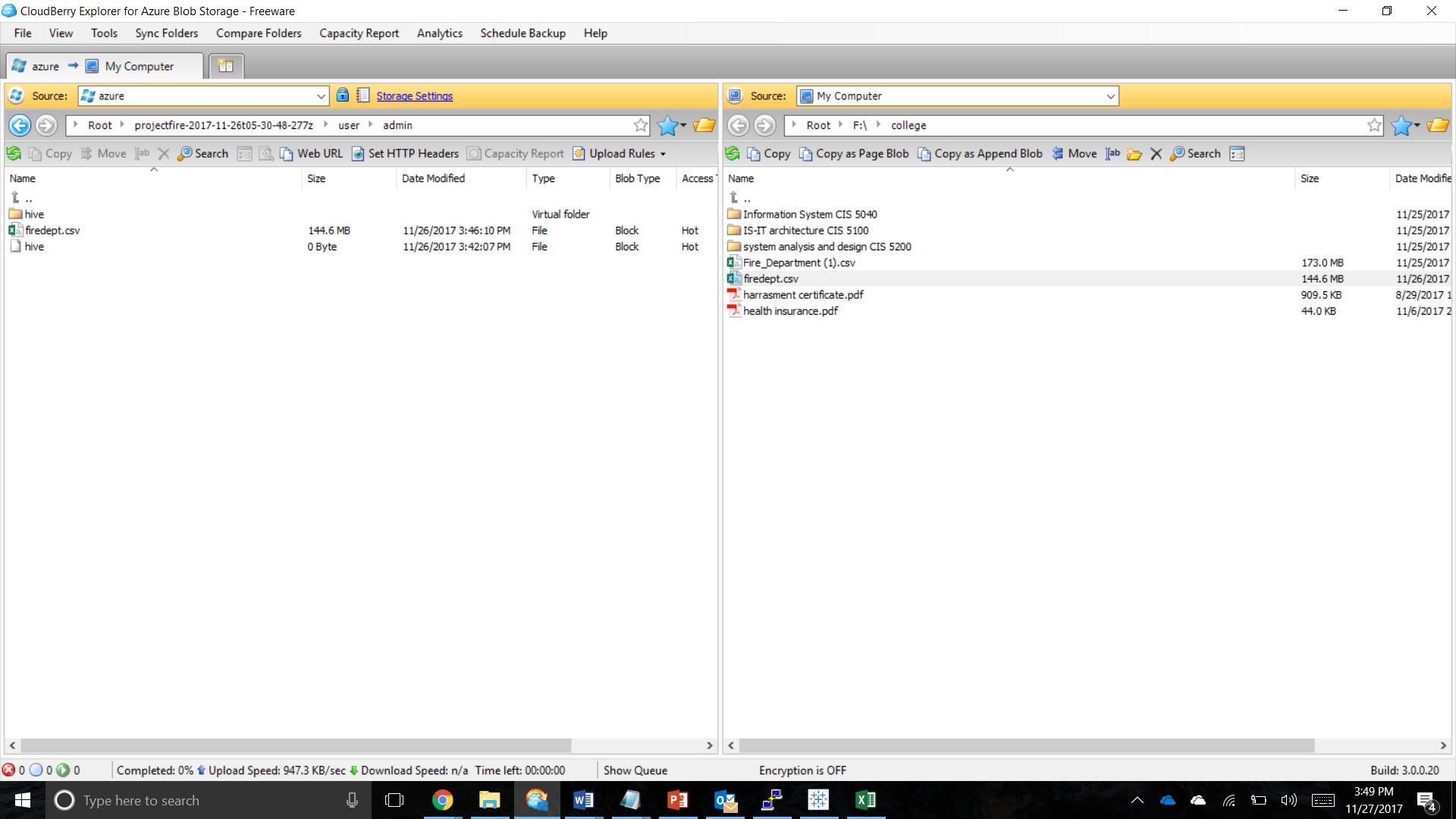


* In the cluster dashboard, you will see a Hive query editor, where you will execute all the Hive commands. The hive editor looks like below,

**Importing Data into HDFS using Cloudberry:**

* The download link for the dataset used for our project is given below,

<https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3>

* Total Size : 1.5GB
* Once the data is downloaded into local disk, use the below link to download the Cloudberry Explorer for Azure Blob Storage which is a freeware,
* <http://www.cloudberrylab.com/explorer/microsoft-azure.aspx>
* Once the cloudberry is downloaded, connect to the Azure storage blob with the credentials of the storage account you used while creating a HD insight Hadoop cluster.
* Once the Azure storage blob is connected, you can see local disks on one screen and your HDFS on the other side as below,
* The data now can be copied to the HDFS just by dragging and dropping the files from our local system or by Right click 🡪 Move option on the file.
* Once the data is moved to the HDFS, you are good to create the external tables for these CSV files.

**Analyze Startup data in Hive on Azure:**

In the Hive editor, execute the following queries to create an external table for Fire Dept. This will create a schema for the table and load the data from the CSV files we moved to HDFS using cloudberry.

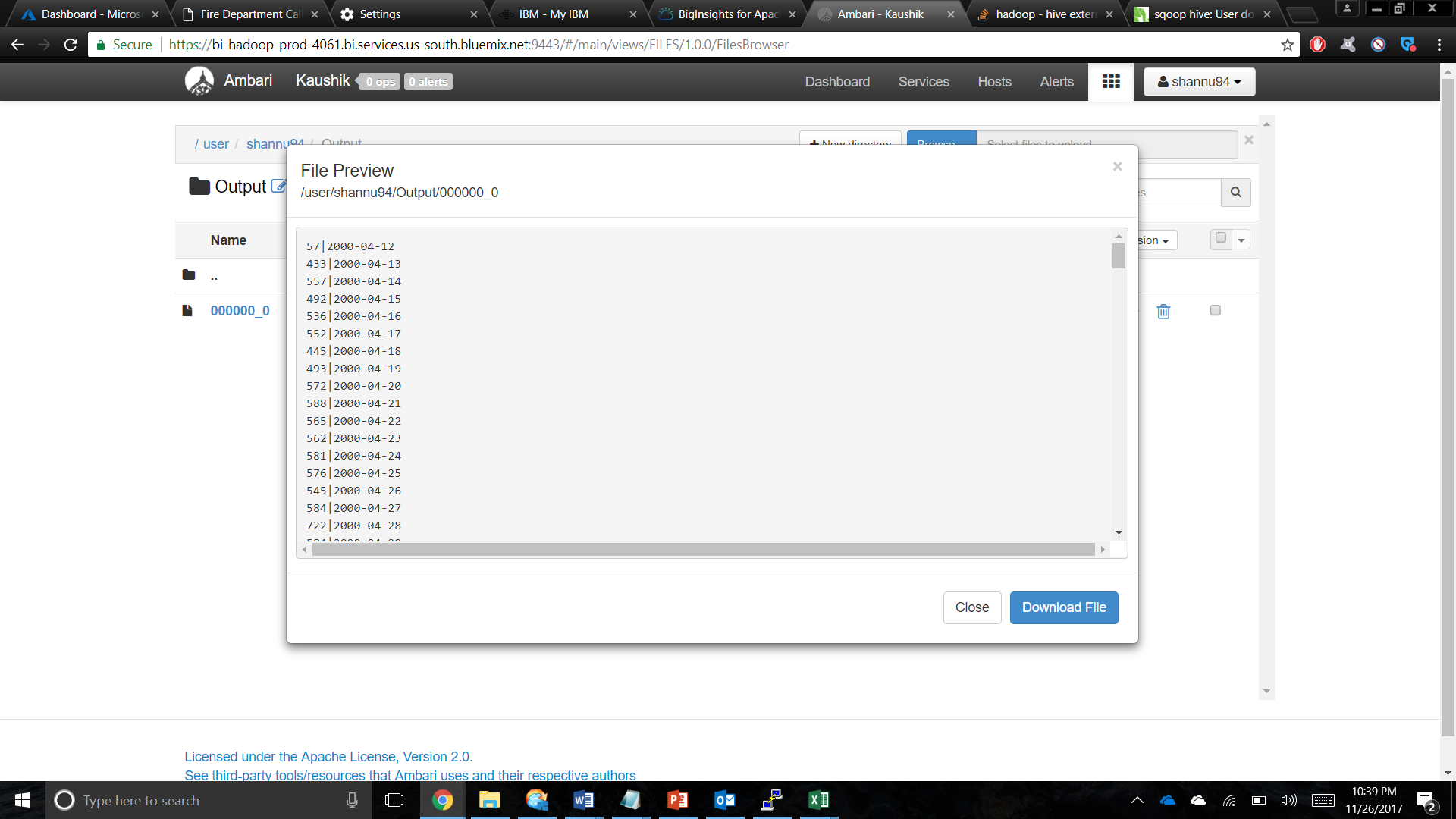
|  |
| --- |
| DROP TABLE firedept;  CREATE EXTERNAL TABLE IF NOT EXISTS  firedept(Call\_Number BIGINT,Incident\_Number INT,Call\_Type STRING,Call\_Date DATE,Day INT,Month INT,Year INT,Call\_Final\_Disposition STRING,Address STRING,Zipcode BIGINT,Box INT,Priority INT,ALS\_Unit STRING,Number\_of\_Alarms INT,Unit\_Type STRING,Neighborhood STRING,Lattitude DOUBLE,Longitude DOUBLE)  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  tblproperties ("skip.header.line.count"="1");  LOAD DATA INPATH '/user/shannu94/firedept/firedept.csv'  OVERWRITE INTO TABLE firedept; |

Once the external table creation is successfully completed, execute the following HiveQL to analyze various factors from our dataset.

1. **The number of calls received by the fire department over the 4 years period**

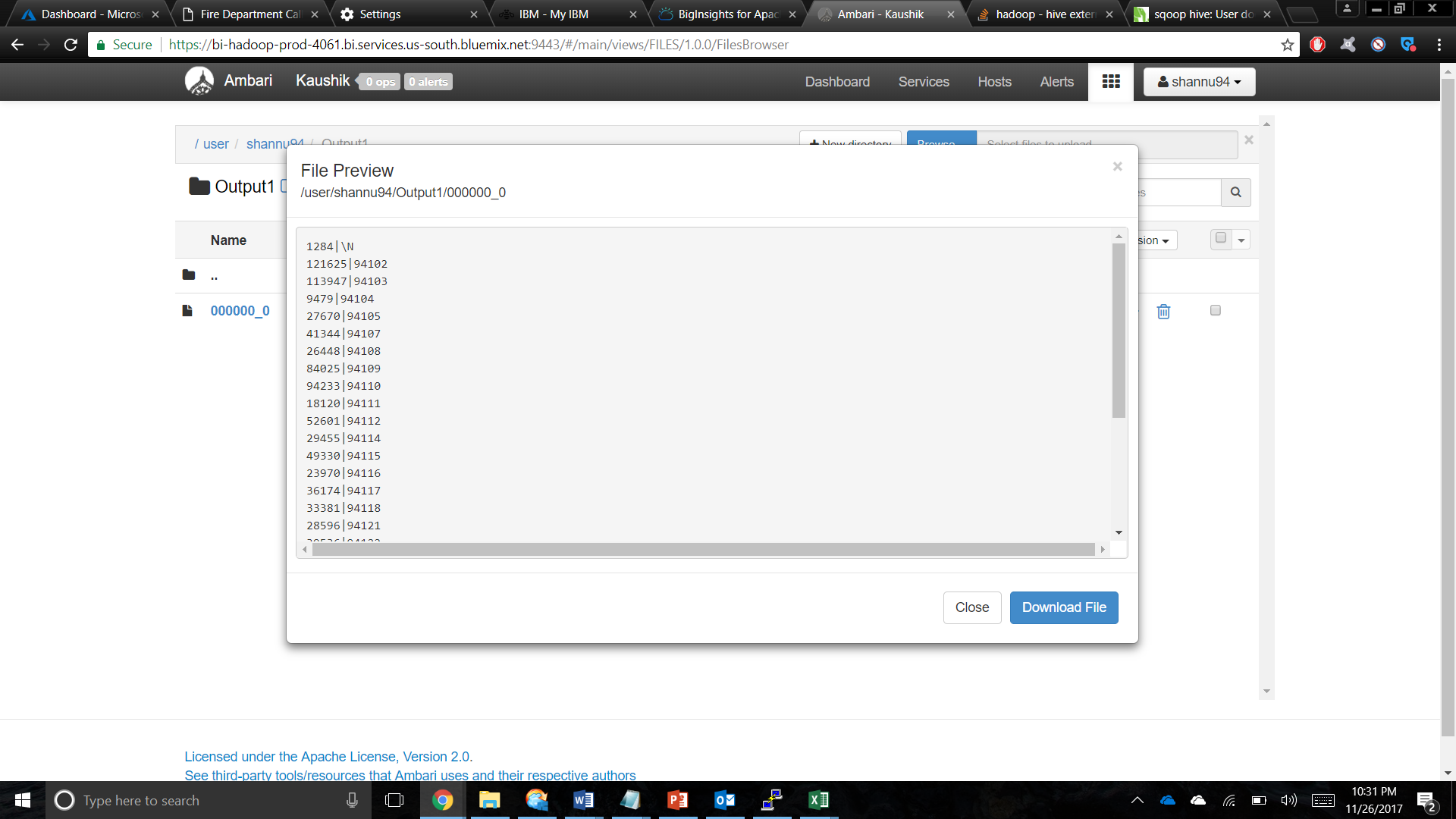
|  |
| --- |
| CREATE TABLE qn1  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output/'  AS  select count(Call\_Number), Call\_Date from firedept where Call\_Date IS NOT NULL group by Call\_Date; |

**Output:**



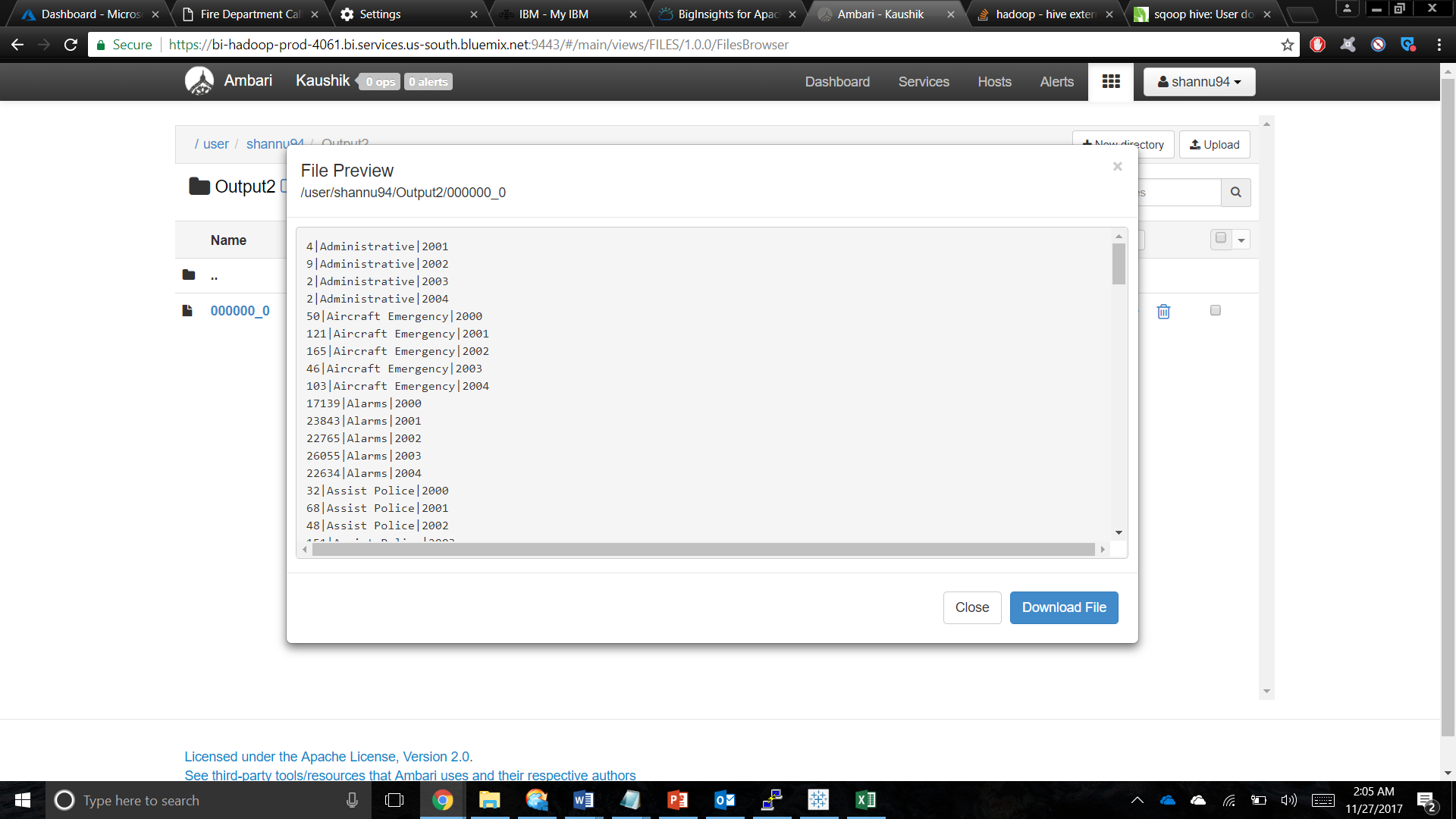
1. **The number of incidents that happened in a particular zip code:**

|  |
| --- |
| CREATE TABLE qn2  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output1/'  AS  select count(Incident\_Number),Zipcode from firedept WHERE Incident\_Number IS NOT NULL group by Zipcode; |

**Output:**

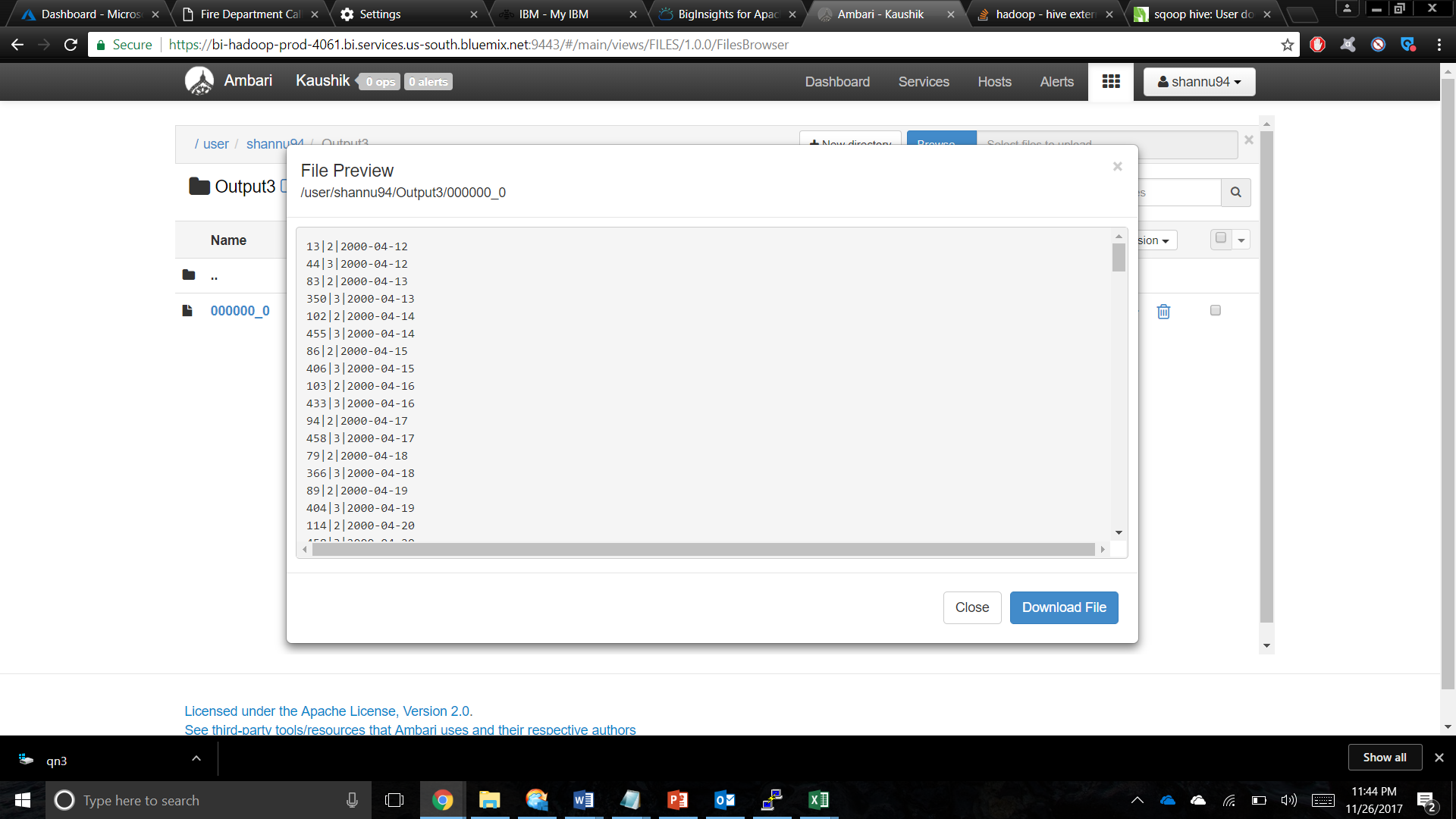
1. **What is the trend of each incident over the years?**

|  |
| --- |
| CREATE TABLE qn3  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output2/'  AS  select count(call\_type), call\_type, year from firedept group by call\_type, year; |

**Output:**

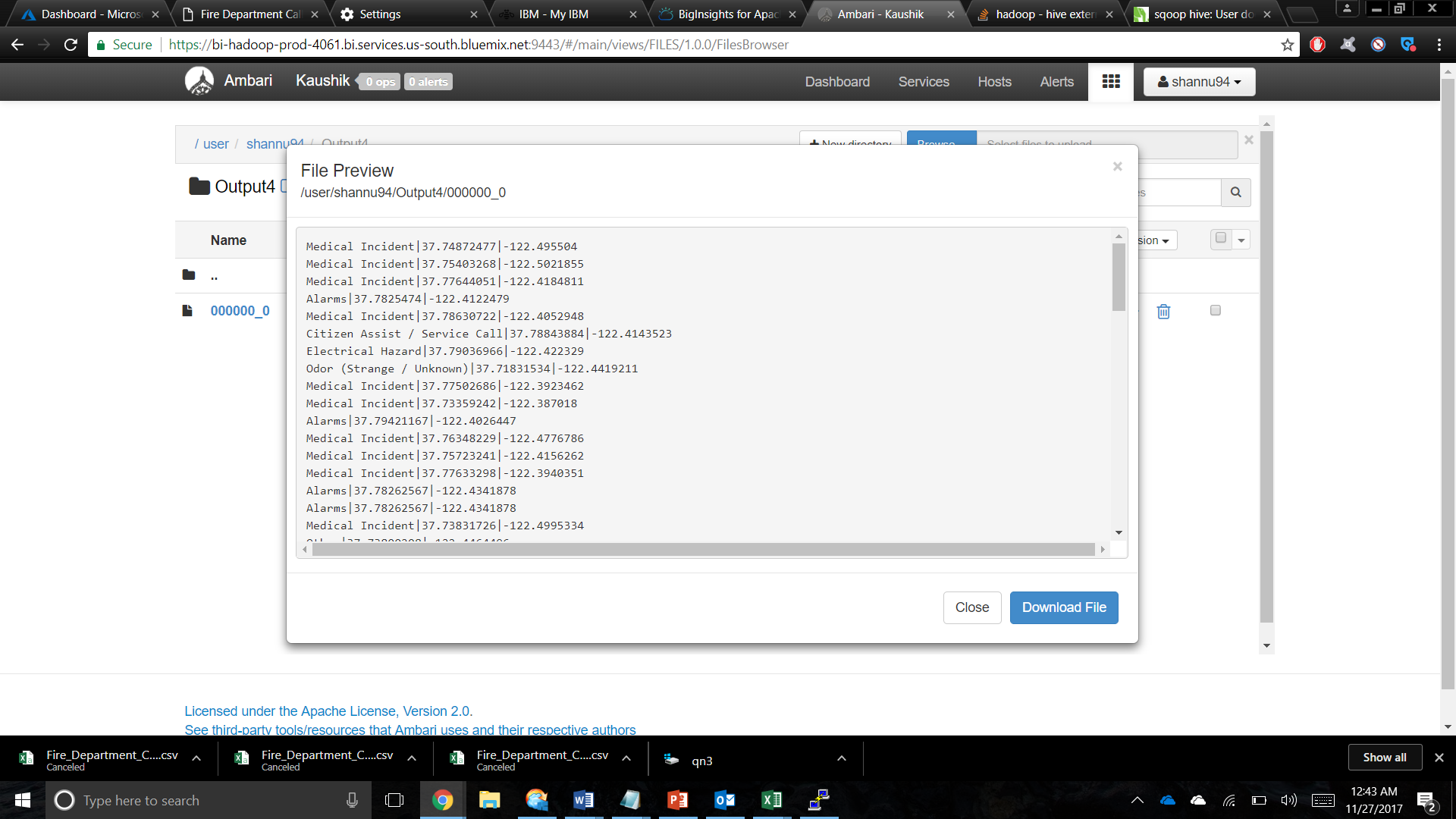
1. **The number of high priority and low priority calls received per day**

|  |
| --- |
| CREATE TABLE qn4  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output3/'  AS  Select count(priority), priority, call\_date from firedept group by call\_date, priority; |

**Output:**

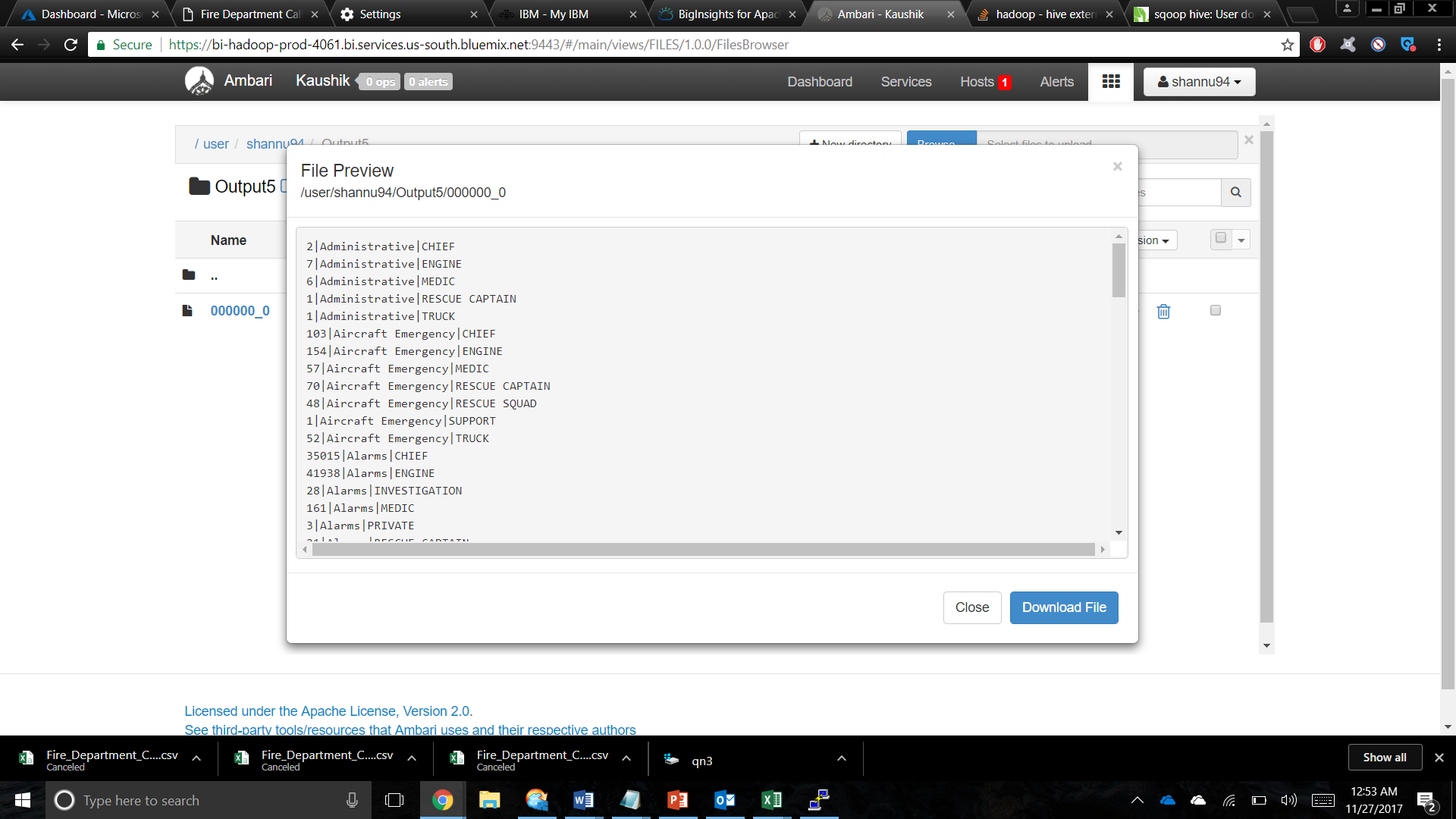
1. **Intensity of each incident over the geospatial locations in San Francisco**

|  |
| --- |
| CREATE TABLE qn5  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output4/'  AS  Select call\_type, Lattitude, Longitude, Neighborhood from firedept; |

**Output:**

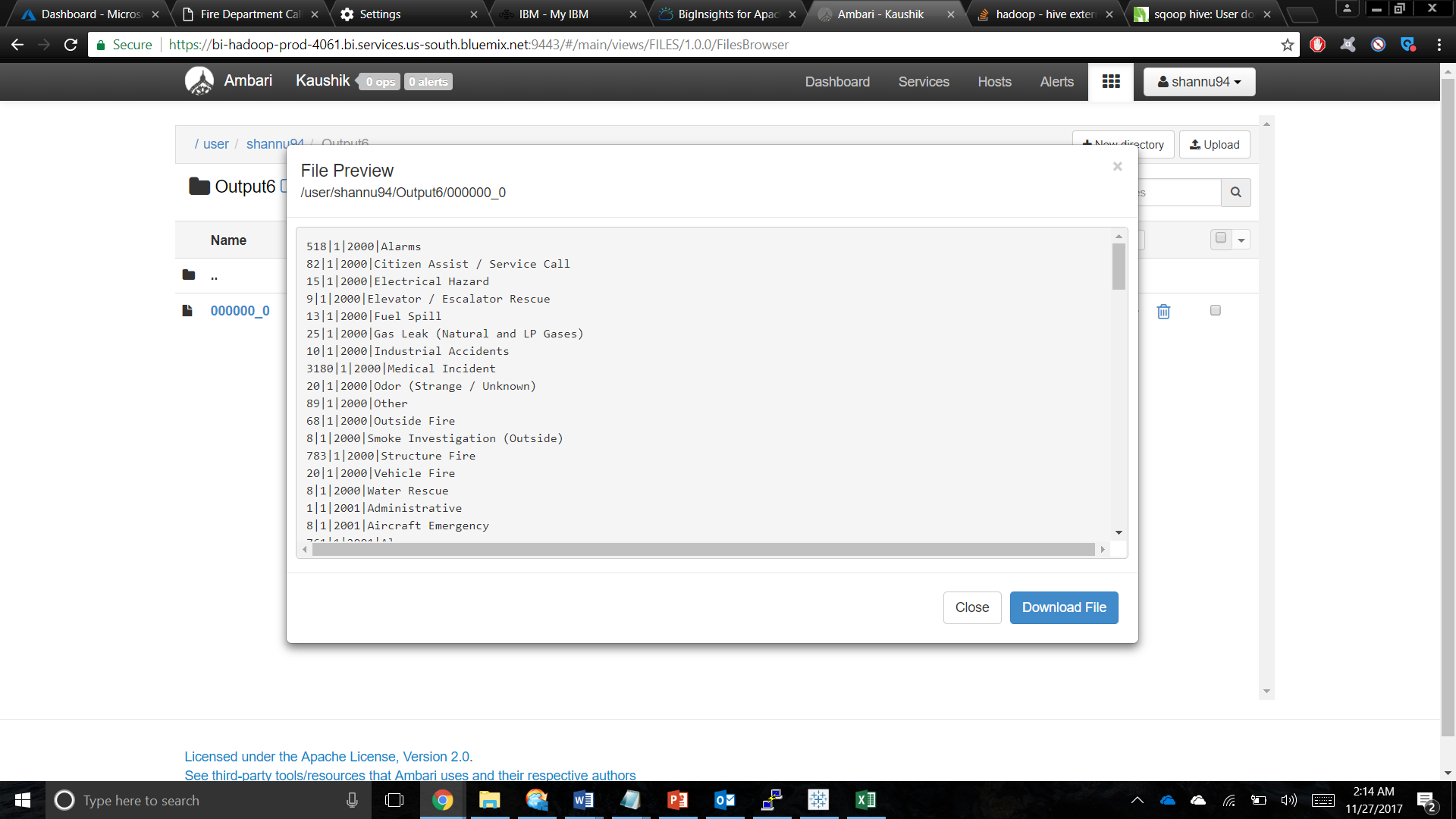
1. **Number of unit types dispatched for each incident during the year 2000 to 2004**

|  |
| --- |
| CREATE TABLE qn6  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output5/'  AS  Select count(unit\_type), call\_type, unit\_type from firedept group by call\_type, unit\_type; |

**Output:**

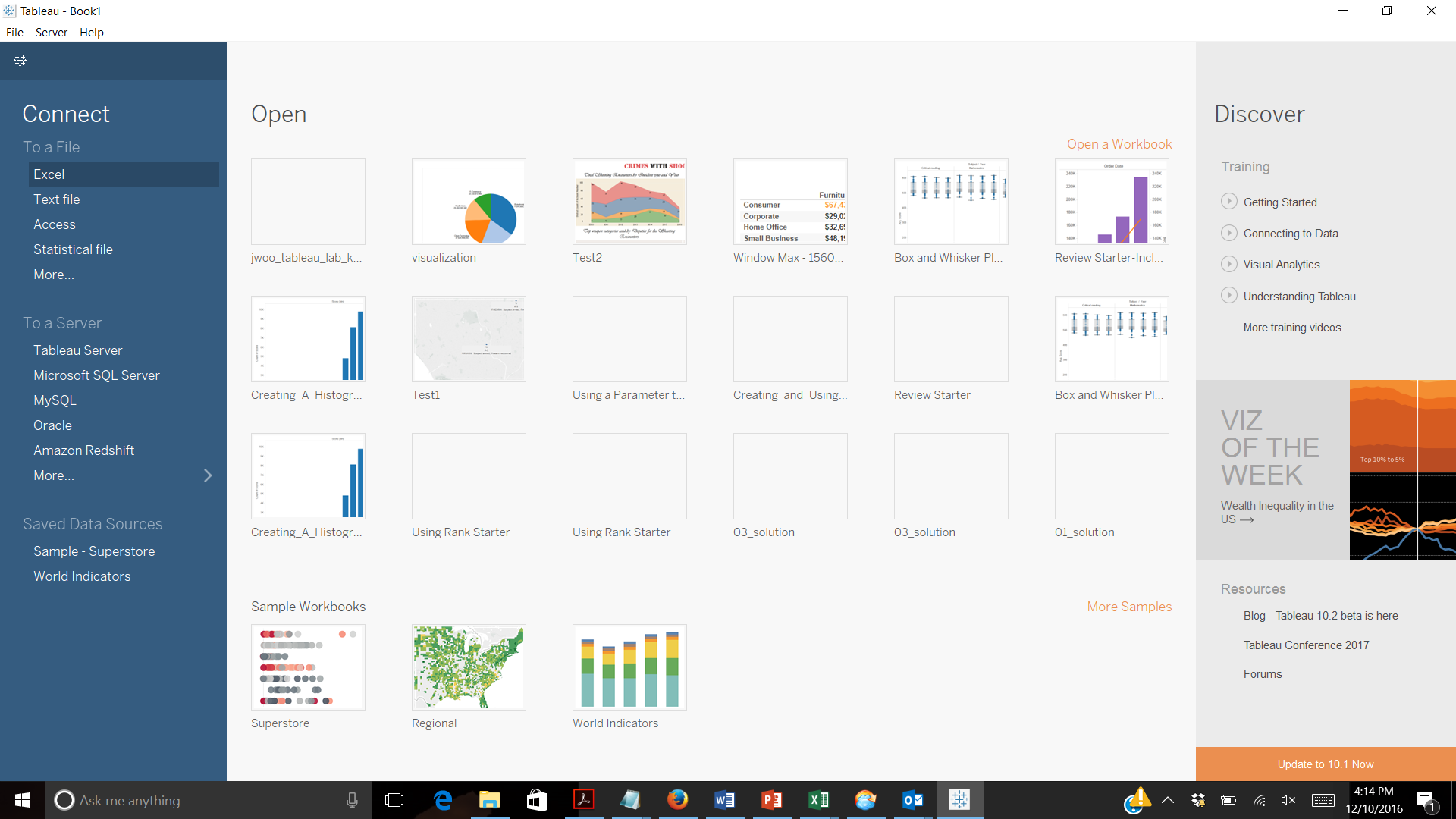
1. **The trend of incidents that happened over the years**

|  |
| --- |
| CREATE TABLE qn7  ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'  STORED AS TEXTFILE LOCATION './Output6/'  AS  Select count(call\_type), month, year, call\_type from firedept group by month,year,call\_type; |

**Output:**

**Visualizations using Tableau:**

* Download the analyzed files from HDFS to local system using Cloudberry.
* Save the files with .XLS (excel format)
* Open Tableau desktop and import the excel file as below,



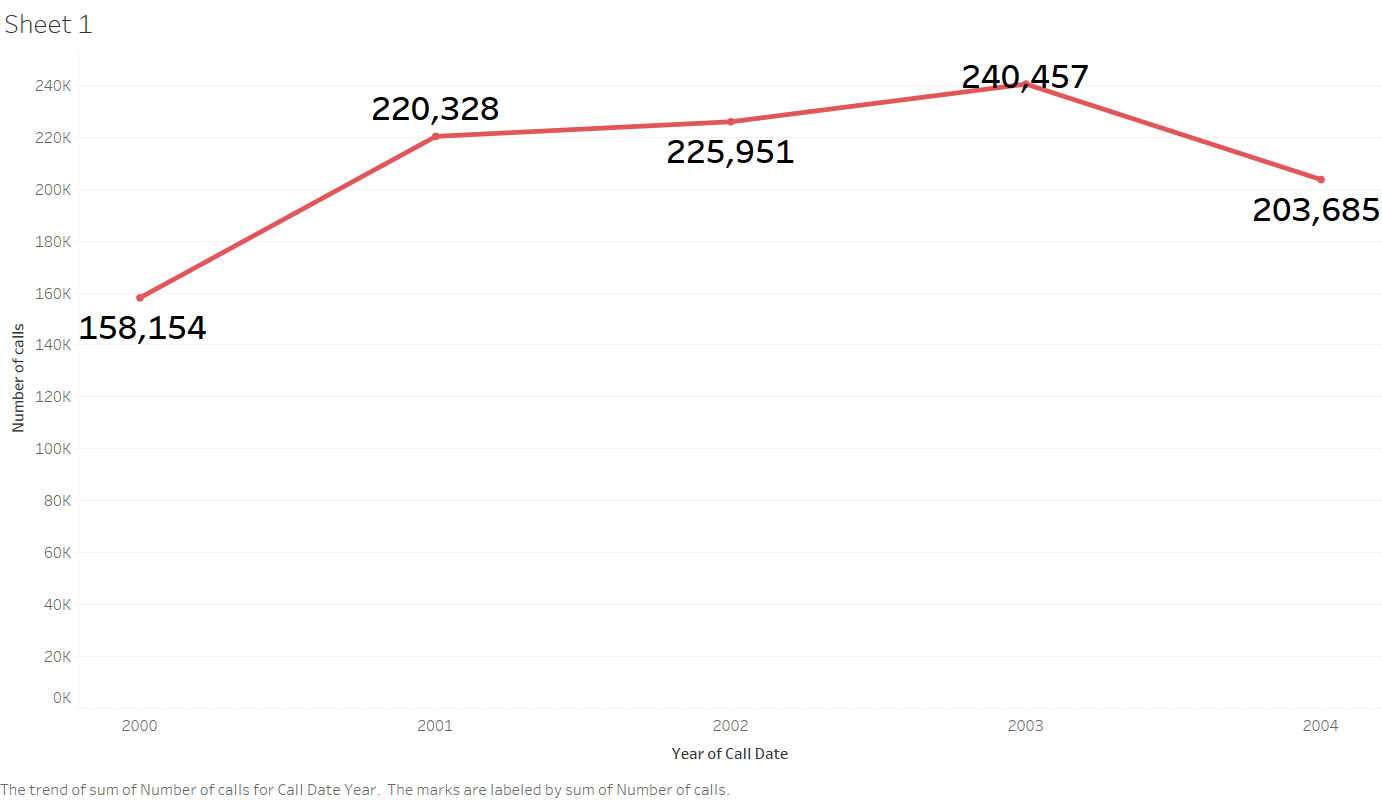
The visualization can be made in tableau by dropping appropriate measures and dimensions in rows, columns, filters and in marks. The following are the visualizations for the various factors that are analyzed,

1. **The number of calls received by the fire department over the 4-year period**

**Rows:** Number of calls

**Column:** Year of call date

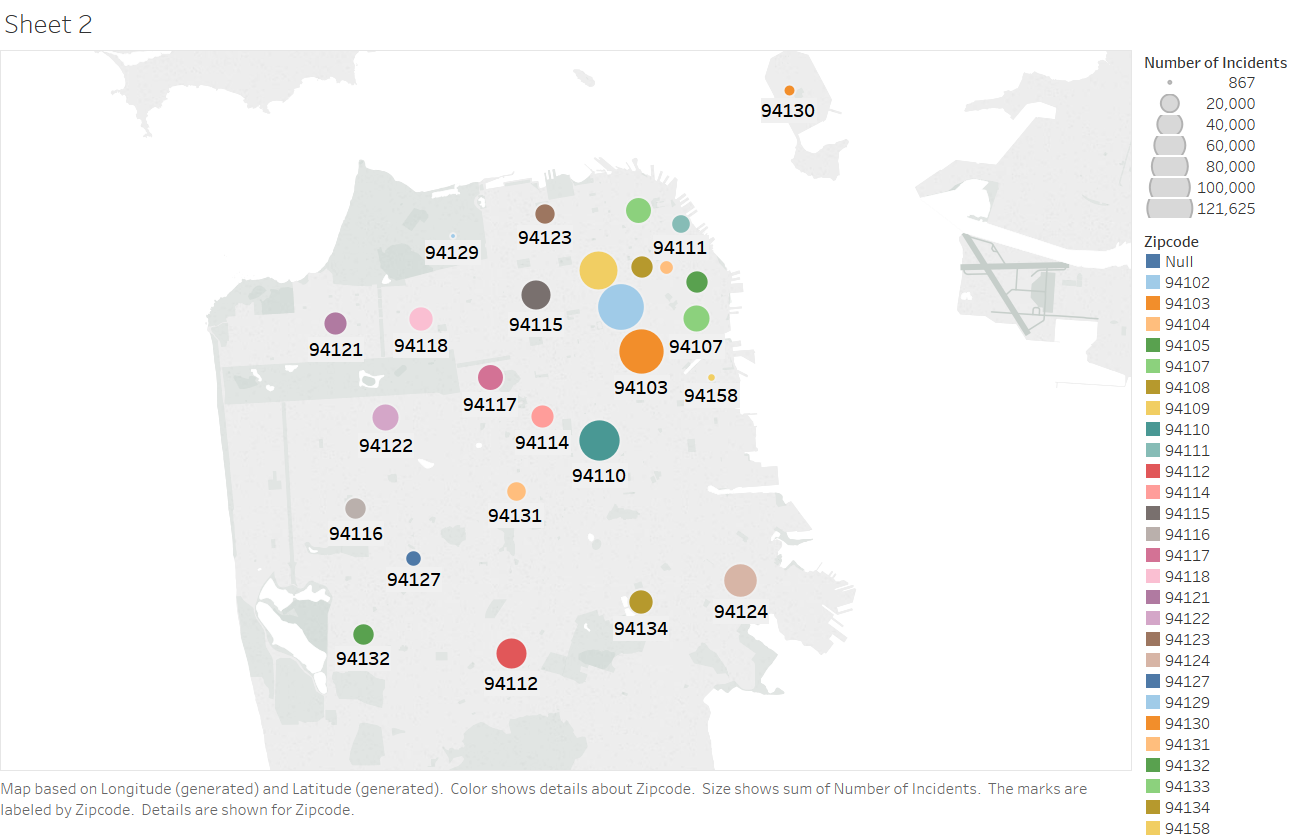
Select the Line chart from the type of charts



1. **The number of incidents that happened in a particular zip code**

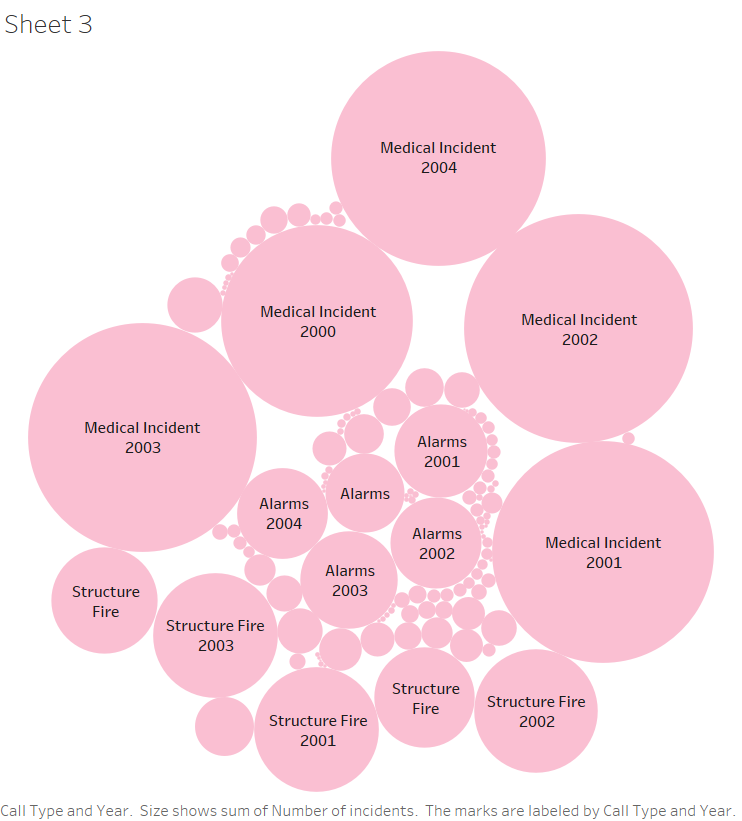
Select Maps from Show Me on the right top

Add zipcode and number of incidents to the details card in the marksheet

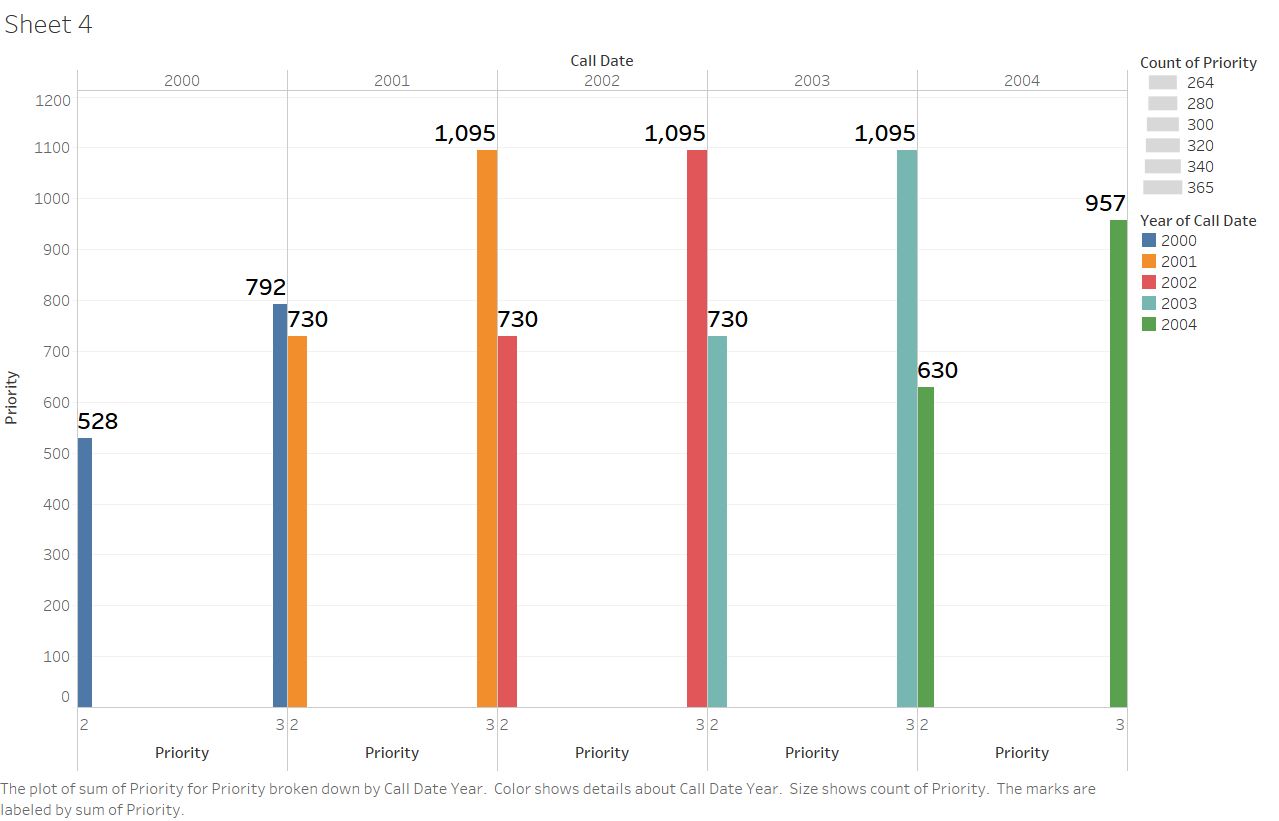


1. **What is the trend of each incident over the years?**

select the bubble chart from the show me tab



1. **The number of high priority and low priority calls received per day**



**Column**: Priority

**Row:** Count of Priority

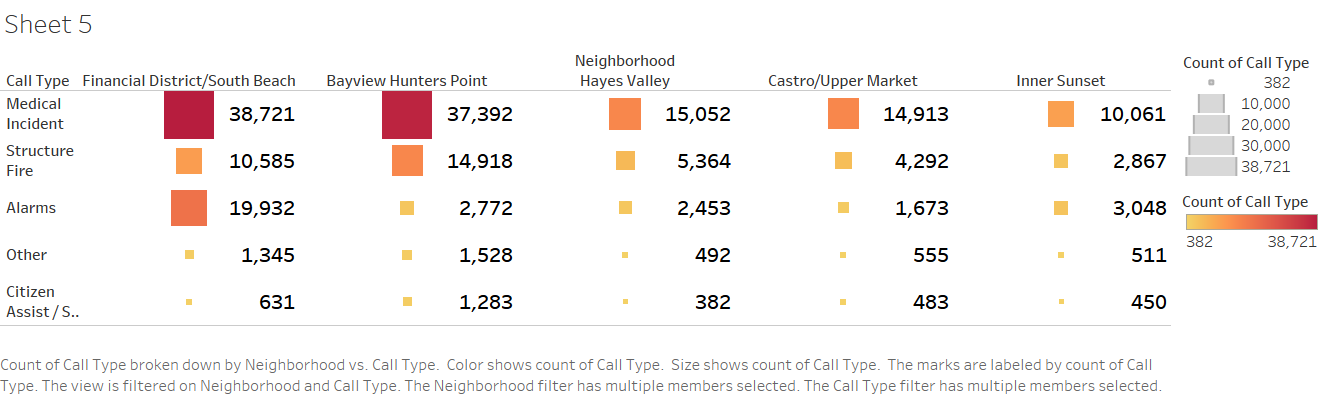
Select the Bar chart from the show me tab

1. **Intensity of each incident over the geospatial locations in San Francisco**

Select the heat map from show me tab

**Rows : Call Type**

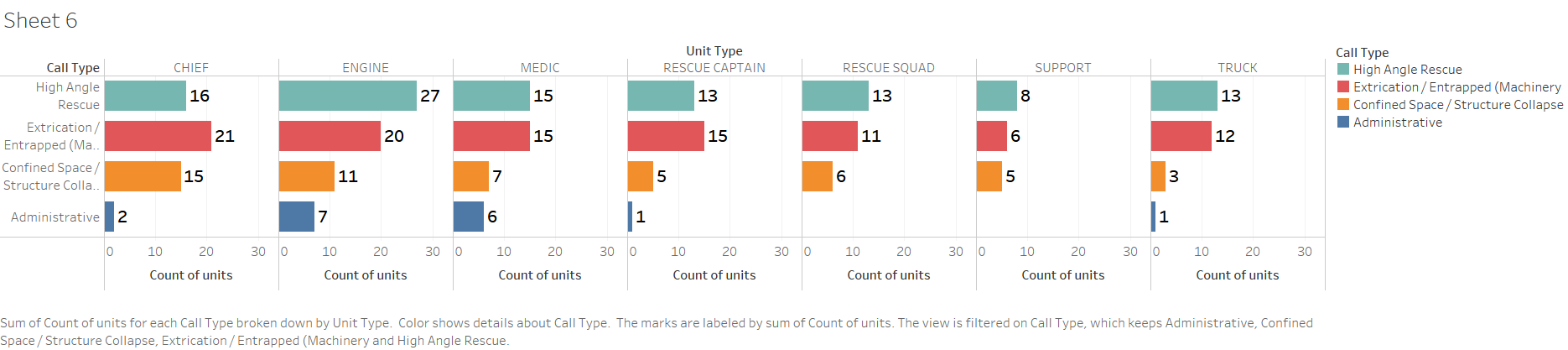
**Columns: neighborhood**



1. **Number of unit types dispatched for each incident during the year 2000 to 2004**

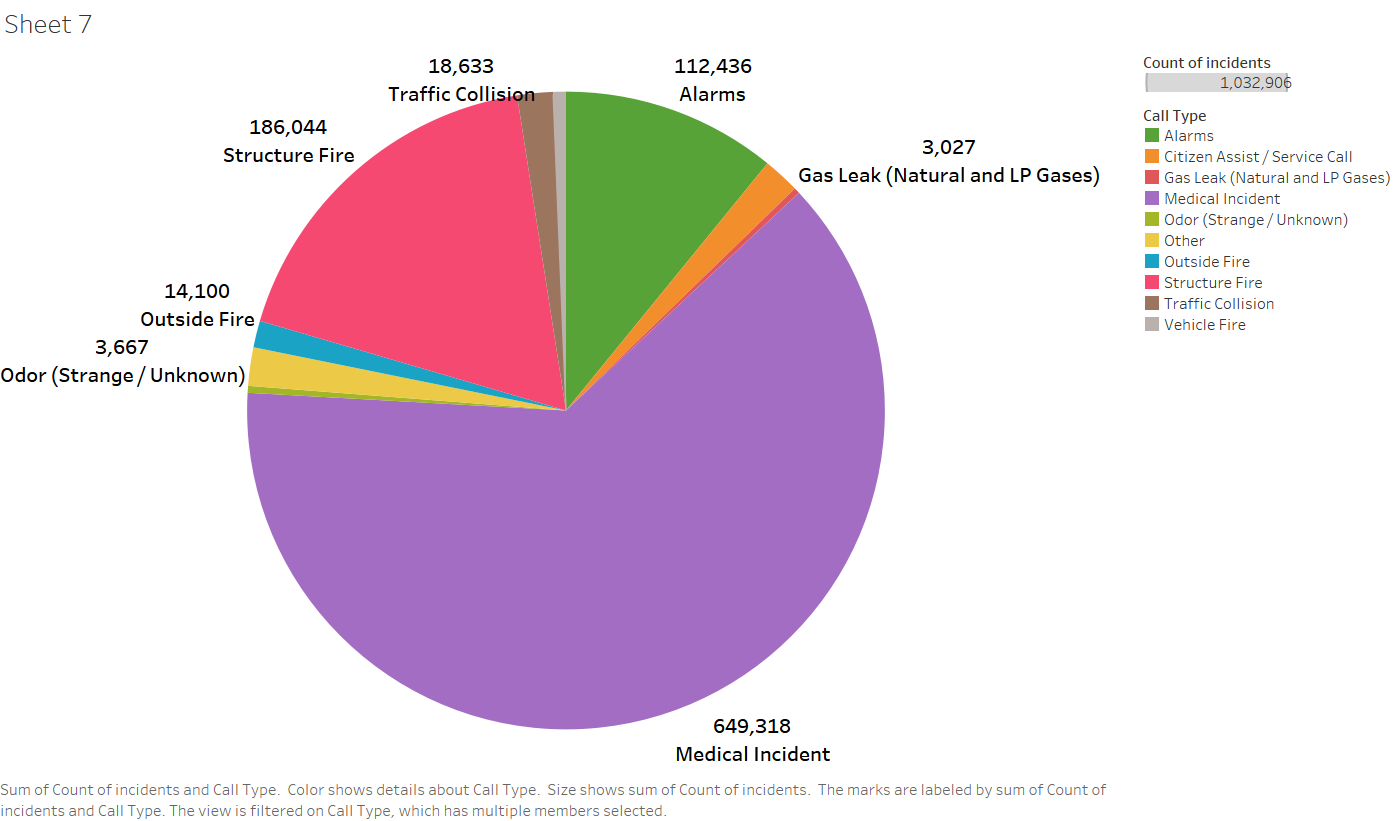
**Column:** Call Type

**Rows:** Unit Type, Count of units



1. **The trend of incidents that happened over the years**

Select the pie chart from show me tab



1. Medical Incident
2. Structure Fire
3. Alarms
4. Other
5. Citizen Assist / Service Call
6. Traffic Collision
7. Outside Fire
8. Vehicle Fire
9. Odor (Strange/Unknown)
10. Gas Leak (Natural and LP Gases)

**GITHUB:**

**Source code:** <https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3>

### Github Link: <https://github.com/ShanmathiA/5200Project>

**REFERENCES**

1. Create HDInsight at Azure: https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-tutorial-get-started-windows

2. http://hadooptutorial.info/tableau-integration-with-hadoop/

3. http://www.cloudberrylab.com/explorer/microsoft-azure.aspx