Fire Department Call Log Analysis in San Francisco

Aaditya Alekar Albert Dabney Ilavarasan Anbundurai
Nirmal SubramaninnnYangyang Jia
Department of Information Systems, California
State University Los Angeles

Abstract

This paper analysis the fire department call log in San Francisco from 2000 to 2004. We use Hadoop cluster and Hive QL to analyze the data set. 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. By going through all the analyzation process, we are able to find out 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.

Key words: Fire department San Francisco Analyzation

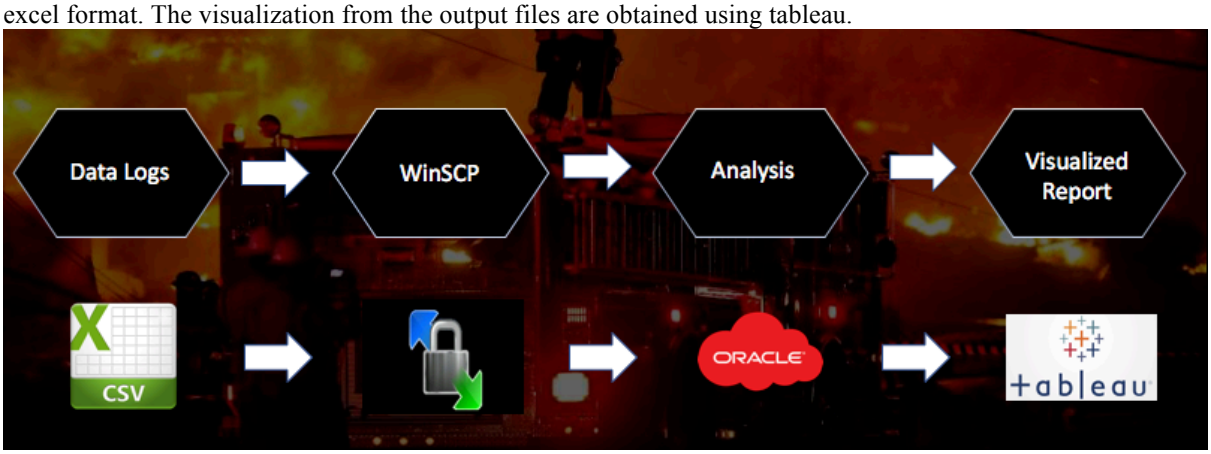
1. Introduction

San Francisco Fire Department (SFFD)

The mission of the Fire Department is to protect the lives and property of the people of San Francisco from fires, natural disasters, and hazardous materials incidents; to save lives by providing emergency medical services; to prevent fires through prevention and education programs.
The San Francisco Fire Department was established in 1866 and serves an estimated population of 1.4 Billion people residing in San Francisco.
In this project, we focus on analyzing the several incidents that occurred between the year 2000 to 2004, its criticality, priority given to each incident, the units that were involved in resolving each case, how the incidents have varied over the years in different parts of San Francisco

2. Work Flow

First, we downloaded this data set about fire department call log analysis from a trusted source. The data set has a rich data on call log analysis from year 2000 – 2004 about 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. By going through all the analyzation process, we are able to find out 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. Then, we using cloudberry to uploaded these data logs in Azure. The data is analyzed using the hive queries and the output files are stored in cloud which is then downloaded and opened in excel format. The visualization from the output files are obtained using tableau.



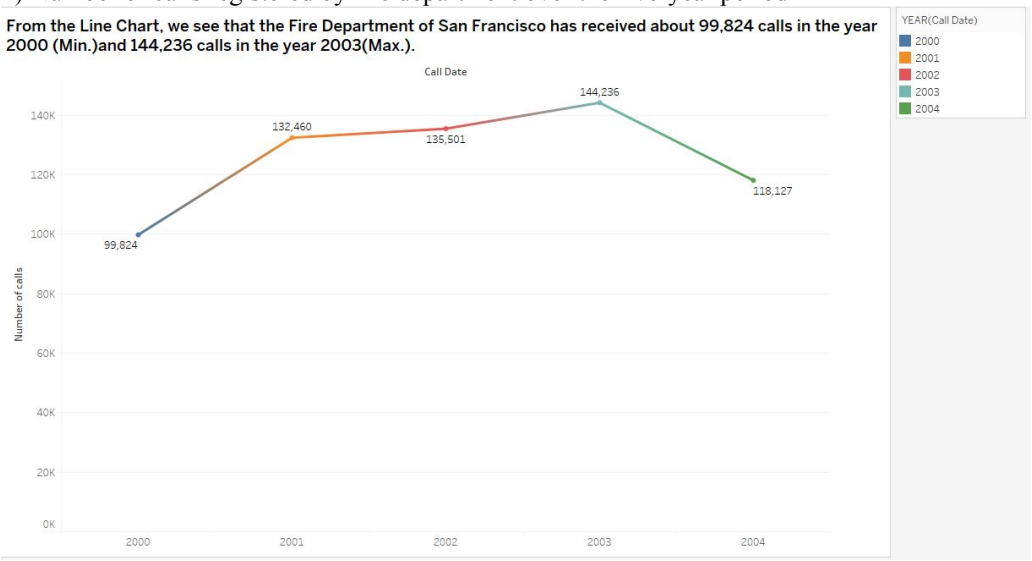
2. 1 Data Analysis and Representation

In this Project, we have considered six main parameters to analyze the data.

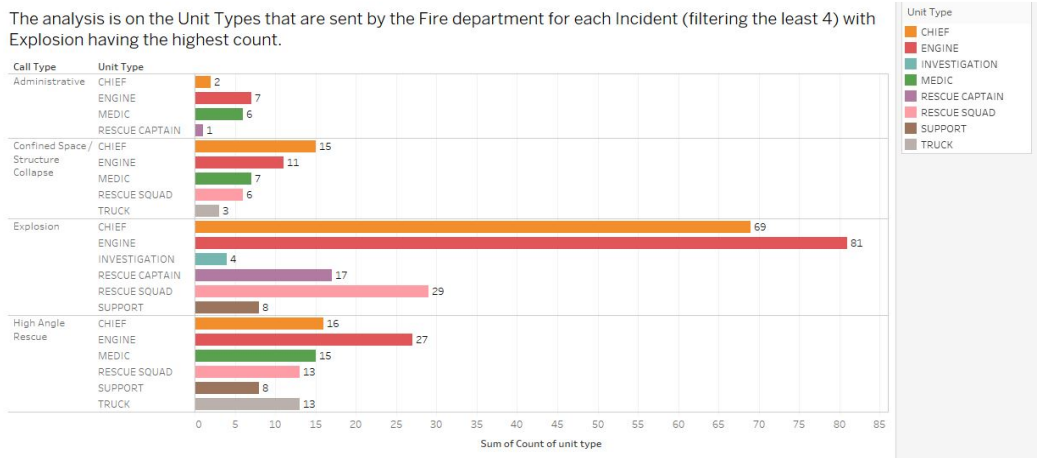
1. Which year received the most number of calls
2. Which incident had the least requirement of dispatch units
3. Which zip code is the most incident prone zone
4. Which is the leading Incident happened over the 5 years
5. What are the years that received the high priority and low priority?
6. Which incident has the highest intensity over the geo spatial location

Below is the screenshot for the output of the above query –

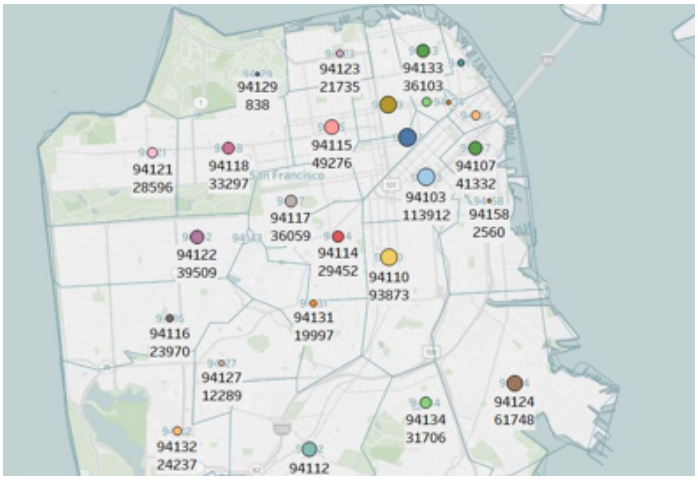
1) Number of calls registered by fire department over the five-year period



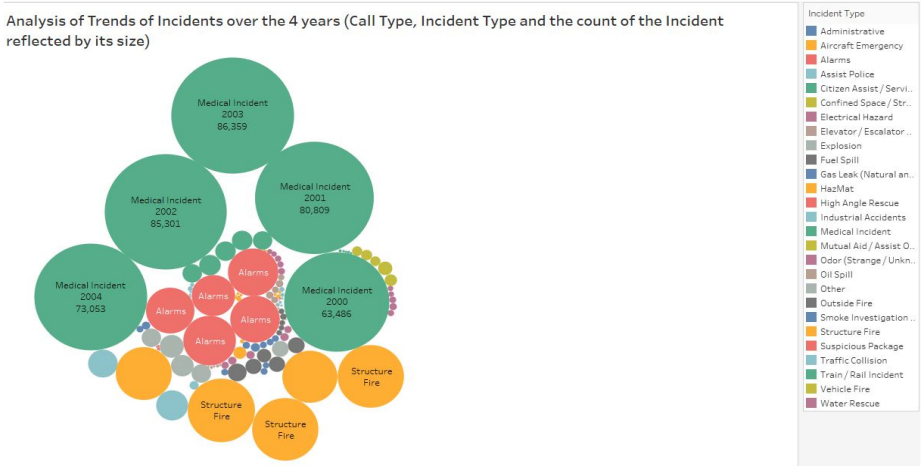
2) Number of units dispatched according to the different types of calls registered



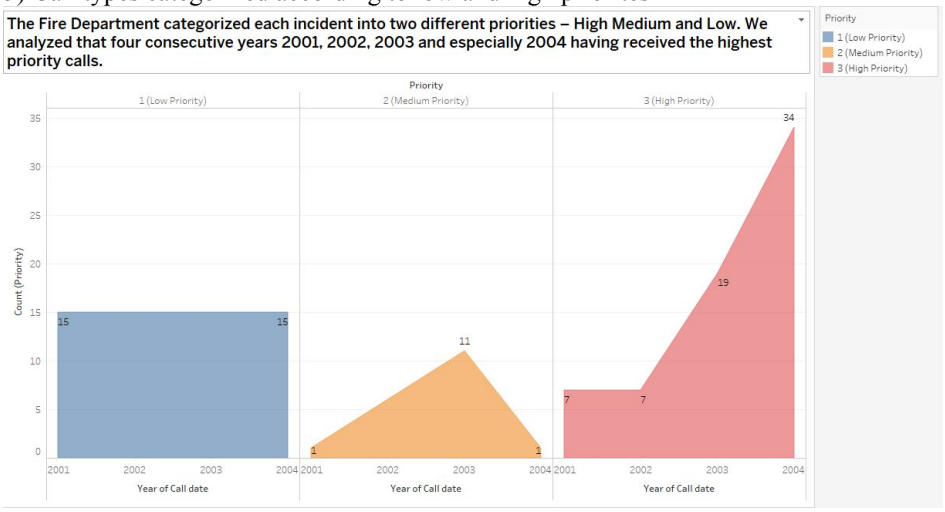
3) The number of incidents that happened in a particular zip code:



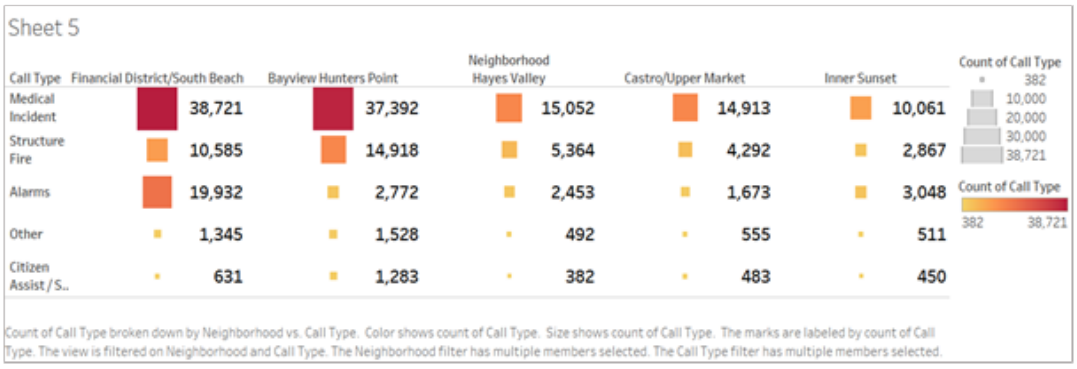
4) Analysis of incidents over the period of five years.



5) Call types categorized according to low and high priorites



6) Count of call types according to neighborhood



3. Summary

From the above analysis, we can conclude the following

- Most calls were received in the YEAR 2003
- Administrative incident had the least requirement for dispatch units from San Francisco Fire Department
- Most incident prone areas are of zip codes 94102, 94103 and 94109
- Medical Incident was observed to see highest no of count
- Low priority calls have remained constant, Medium priority calls had highs and lows, High priority calls varies over the years.
- Five major areas to have reported most incidents are Financial District/South Beach, Bayview Hunters Point, Neighborhood Hayes Valley, Castro/Upper Market and Inner Sunset.

4. Limitations

- The Analysis of the Dataset revolved around different Incident types due to Limited parameters.
- The dataset should have been more specific i.e. it should have more information pertaining to the actual cause of the incident, it would have been possible to analyze and infer the major reasons for the incident and provided a solution
- Certain incidents could be done as crimes too, had the data about incident caused by crimes be given, we could have analyzed the crimes that cause the incidents
- Had the data about damage been specified, we would have been able to analyze each incident in depth
- The data set should have included the correct count of People injured in each incident in the area or zip code

GitHub Link: https://github.com/ianbudu/CIS5200/tree/master

References

Hadoop Tutorials
[1] https://hadoop.apache.org/

Apache Hive TM
[2] https://hive.apache.org/

How To Process Data with Apache Hive
[3] http://hortonworks.com/hadoop-tutorial/how-to-process- data-with-apache-hive/

Intro to Hive
[4] http://blog.cloudera.com/wp-content/uploads/2010/01/6- IntroToHive.pdf

Demo Analyzing data with hue and hive
[5] http://blog.cloudera.com/blog/2013/04/demo-analyzing- data-with-hue-and-hive/

HD Insights: Get Started Hadoop Tutorial
[6] https://azure.microsoft.com/en- us/documentation/articles/hdinsight-hadoop-tutorial-get- started-windows/