**Fire Department Call Log Analysis in SF**

**Using Hadoop**

Shanmathi Arul Murugan , Ashwin.P.Karthik, Kaushik Sridharan

Department of Information Systems, California State University Los Angeles

[sarulmu@caltstatela.edu](mailto:sarulmu@caltstatela.edu) [akarthi@calstatela.edu](mailto:umalpan@calstatela.edu) [ksridha@calstatela.edu](mailto:ksridha@calstatela.edu)

**Abstract -** The paper analyses over four years of call log analysis data from San Francisco. The data is analyzed using Hadoop cluster and Hive QL. 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.

.

# Introduction

**San Francisco Fire Department (SFFD)**

The mission of the Fire Department is to protect the lives and property of the people from fires, natural disasters, and hazardous materials incidents; and to provide a work environment that values health, wellness and cultural diversity and is free of harassment and discrimination.

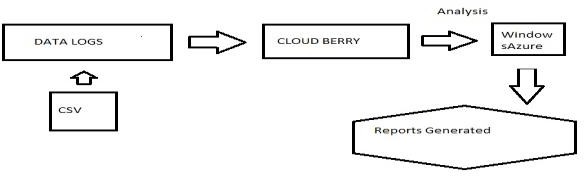
The San Francisco Fire Department was established in 1866 and serves an estimated population of 1.4 Billion people residing in San Francisco.

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

# Work Flow

Initially a data set comprising the details of the fire department call log analysis was downloaded from a trusted source. This data set contains of the name of the company their funding details, the acquirement details along with the function market. 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.These data logs in CSV format were uploaded in Azure using cloudberry, a tool that helps users to visualize and manage files in Microsoft azure efficiently. This tool serves as a bridge between the local system and the cloud data storage. The “COPY” function from cloudberry is used in this case to put the data logs into the cloud storage. 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.



## Data Storage

The Azure cluster comprises of the following architecture, a master and slave. A master node also called as the Head node in Azure manages the queries in the cluster and the data node which is the also called the worker node executes the commands. Hive QL is used as a querying language to extract the data from the data logs in the cluster.

## Data Analysis and Representation

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

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

The Following Hive Query was executed to obtain the total number of calls over 4years period:

**CREATE TABLE qn1**

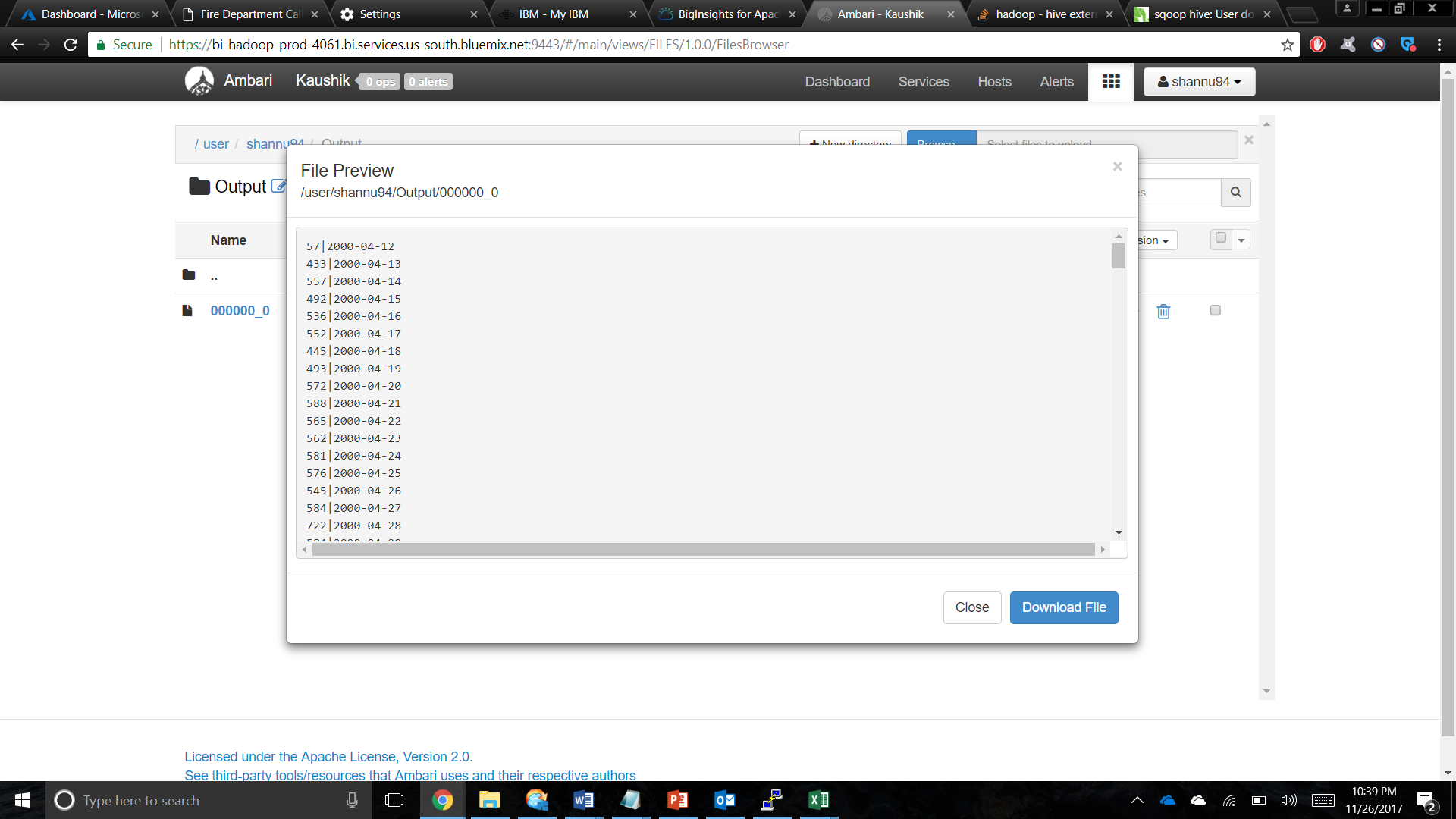
**ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'**

**STORED AS TEXTFILE LOCATION './Output/'**

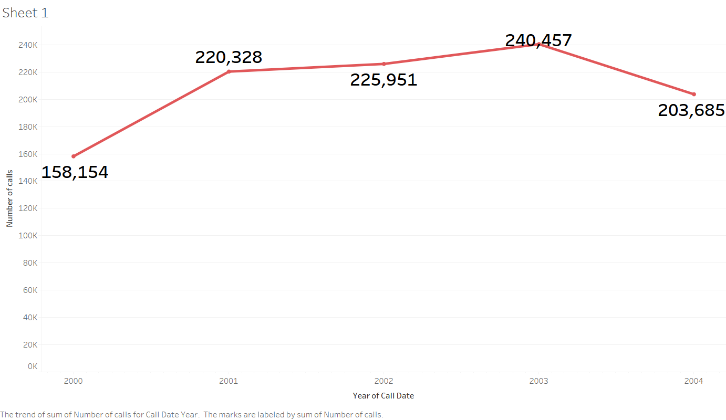
**AS**

**select count(Call\_Number), Call\_Date from firedept where Call\_Date IS NOT NULL group by Call\_Date;**

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



### Number of calls received over the four-year period



From the Line Chart, we see that the Fire Department of San Francisco has received about 240,457 calls in the year 2003 and 220,328 calls in the year 2001

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

A close up of a map

Description generated with very high confidence

This analysis shows that zip codes 94102,94103 leads with the most number of incidents that happened over the years in San Francisco.

### The trend of incidents that happened over the 4 years

A close up of a logo

Description generated with high confidence

The Packed Bubbles chart clearly states that from the year 2000 to 2004 the fire department received calls majorly for three issues, Medical Incidents leading the list.

* 1. **The number of high priority and low priority calls received per day** A screenshot of a cell phone

     Description generated with very high confidence

The Fire Department categorized each incident into two different priorities – High and Low. We analyzed that three consecutive years 2001, 2002 and 2003 has received the highest priority calls.

* 1. A screenshot of a cell phone

     Description generated with very high confidence**Intensity of each incident over the geospatial locations in San Francisco**

Among the various areas/cities in San Francisco, with the help of Heat Map the 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.

* 1. A screenshot of a video game

     Description generated with high confidence**Number of unit types dispatched for each incident during the year 2000 to 2004**

The analysis is on the Unit Types that are sent by the Fire department for each Incident (filtering the least 5). It can be inferred that the Administrative call type requires the least dispatch unit compared to others.

* 1. **A picture containing screenshot, map

     Description generated with high confidenceThe trend of incidents that happened over the years**

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)

# Conclusion

From the above analysis, we can conclude the following

1. Most calls were received in the YEAR 2003
2. Most incident prone areas are of zip codes 94102 and 94103
3. Leading Incident happened over the 4 years is Medical Incident
4. High priority calls were around the years 2001,2002,2003 and low priority calls were around 2000 and 2004
5. Medical Incident is the highest intensity over the geo spatial location
6. Administrative incident had the least requirement for dispatch units

# Limitations

From the available data, holding few parameters, we were able to provide the above-mentioned analysis. Had the data been more detailed i.e. holding information regarding the available data, holding few parameters, we were able to analyze only the type of incident over a span of 4 years. Had the data been more detailed i.e. holding information regarding the 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.

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

**Data Set Link –** <https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3>

# References

### Hadoop Tutorials

1. <https://hadoop.apache.org/>

### Apache Hive TM

1. <https://hive.apache.org/>

### How To Process Data with Apache Hive

1. [http://hortonworks.com/hadoop-tutorial/how-to-process- data-with-apache-hive/](http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/)

### Intro to Hive

1. [http://blog.cloudera.com/wp-content/uploads/2010/01/6- IntroToHive.pdf](http://blog.cloudera.com/wp-content/uploads/2010/01/6-IntroToHive.pdf)

### Demo Analyzing data with hue and hive

1. [http://blog.cloudera.com/blog/2013/04/demo-analyzing- data-with-hue-and-hive/](http://blog.cloudera.com/blog/2013/04/demo-analyzing-data-with-hue-and-hive/)

### HD Insights: Get Started Hadoop Tutorial

1. [https://azure.microsoft.com/en- us/documentation/articles/hdinsight-hadoop-tutorial-get- started-windows/](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-tutorial-get-started-windows/)

### Connecting excel with Hive

1. [https://azure.microsoft.com/en- us/documentation/articles/hdinsight-connect-excel-hive- odbc-driver/](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver/)

### HD Insights connect to excel power query

1. [https://azure.microsoft.com/en- us/documentation/articles/hdinsight-connect-excel-power- query/](https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-power-query/)

### Mapred Tutorial

1. [https://hadoop.apache.org/docs/r1.2.1/mapred\_tutorial. html](https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html)

### 10 Ways to query hadoop

1. [http://www.infoworld.com/article/2683729/hadoop/10- ways-to-query-hadoop-with-sql.html](http://www.infoworld.com/article/2683729/hadoop/10-ways-to-query-hadoop-with-sql.html)