CIS 4560 Term Paper: Los Angeles and Chicago Crime Data Analysis Using Hive

Blanco, Andrea, Kawashima, Keith,

Najera, Jose, Macias, Jennifer, Nguon, Chiv Haghnazarian, Osheen

Department of Information Systems, California State University

Los Angeles

**Abstract:** Los Angeles and Chicago are two of the most popular cities in America and we decided to examine their crime rates by extracting datasets provided by Gov.data. The datasets cover different crimes committed in both cities of Los Angeles and Chicago. The crime data specifies the type of crime committed, date & time committed, location (address & coordinates), sex, age, ethnicity, and if any weapons used. We will analyze the data using IBM’s Biginsights for Apache using Hadoop clusters to store our data sets and analyze if crime rate has increased or decreased each year between both cities and compare which city has the most robbery and criminal homicide crime by year and within the age group of 10 - 18 years of age. We will visualize the data on graphs comparing by year the amount of crime and also try to visualize the data using the 3D maps to show where the crime was committed. This is vital information for the police department and anyone looking to move into Los Angeles or Chicago.

**1. Introduction**

First, what is Big Data? Many are left questioned to what is Big Data. Big Data is the analysis of large and complex data sets that no ordinary data processing application can handle. Big Data analysis consists of extracting or capturing large scales of data by storing, analyzing, accessing, sharing, transferring, visualizing, updating the data sets. We decided to extract two large data sets that measure up to 1.7 gigabits. The datasets we extracted are crime data sets from the cities of Los Angeles and Chicago. These two data sets consist of columns of data that specify the type of crime committed, date & time committed, location (address & coordinates), sex, age, ethnicity, and if any weapon used. We will be utilizing IBM’s BigInsight for Apache Hadoop to create Hadoop clusters to analyze both Los Angeles and Chicago crime data sets to get insight on how crime differs between the two different cities and compare the increase of crime year by year. With this analysis, we can see which city has the most crimes between the two and this can be vital information for the Los Angeles and Chicago police department or anyone considering moving into either of these cities.

In this analysis, you'll see step by step on how we executed the crime data sets using IBM’s BigInsight Clusters to create:

• Download and upload multiple csv files

• Create Hive tables to query those crime data

• Create Hive queries to analyze the crime data

• Use Microsoft Excel to connect to BigInsight (directly or using an ODBC connection) to retrieve the analyzed data

• Use Excel 3D Map for 3D visualization

• Use Excel Charts for visualization

**2. Prerequisites**

A detailed review of all necessary prerequisites will be explained, which are required to complete the lab exercise in its entirety. Of the prerequisites required is basic knowledge of the Apache Hadoop and Hive, IBM Bluemix infrastructure and the application of the software to large data structures.

**2.1 Apache Hadoop**

Apache Hadoop is mainly used in storing, processing and analyzing big data sets. There are four core components that make this possible in a robust software. the first is that work is distributed file system related amongst many computers. Open source software is the foundation on which Apache Hadoop has thrived in the professional world. Since the software is open source at its core, many companies contribute to its success. Including but not limited to Cloudera, Yahoo, Facebook, Apple[[1]](#footnote-1).

**2.2 Variations of Hadoop**

There are also many variations of Apache Hadoop existent in the large and growing Ecosystem. These include Pig, Hive, Mahout, Oozie, Impala, Hue. We will be mainly focusing on using Hive and Hadoop. As defined by the official Apache Hadoop website, “Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models[[2]](#footnote-2) ”. In using a group of clustered computers, we are able to definitively analyze raw data and output correct information. These computers are mainly run by Hadoop Distributed File Systems.

**2.3 MapReduce**

System used to process data in a Hadoop cluster that we are required to create to complete our data analysis thoroughly. There are two phases, MapReduce. The map task operates on a discrete portion of the data set directed by the user. Once mapping has commenced, MapReduce will distribute the intermediate data to node which perform the Reduce phase.

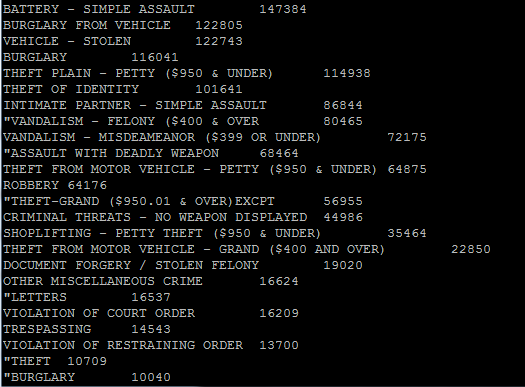
**2.4 Hive**

The use of Hive is a requirement for the data analysis in HDFS systems. Hive uses SQL like language called HiveQl and this is what is used to analyses data and generate MapReduce jobs that run on Hadoop clusters. Writing queries in HiveQl enables a professional to submit jobs to the cluster[[3]](#footnote-3). The Hive interpreter and execution engine parse, plan, generate, submit and monitor progress of MapReduce jobs.

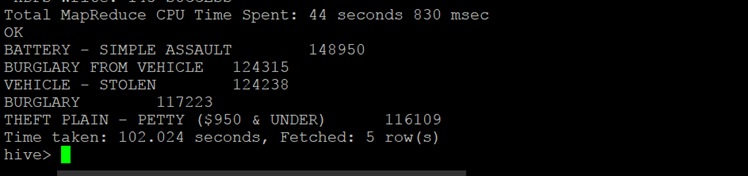
**3. Data Analysis**

When we extract both Los Angeles and Chicago’s data sets, initially we have to download both of the CSV files into the Hadoop Cluster using SSH command scripts. Then we created separate file directory for both Los Angeles and Chicago and then uploaded the cities files in their own separate directory. Implementing this step insures that the two data set files are uploaded inside their separate directory and ready to manipulate the data. Then we created Hive tables to query specific data that is stored inside the clusters. This step allows us to create a separate external table to preserve the dataset its initial file format, utilizing Hive to perform queries against the data within the file. This also helps with analyzing and manipulating the data we extracted because now it is stored in an external table, ready to be queried easily. Then we created a separate external table, but this table was created for the purpose to narrow and only extract the useful columns we want to analyze instead of the whole. This makes it easier and more efficient to analyze the data sets because we have the specific data columns we want to work with.

Then we created a range of different queries for both Los Angeles and Chicago data so that we can analyze and access easily. We created seven different queries for the Los Angeles data set and the first one is “Most common crimes from the Los Angeles Dataset from highest to lowest” and the results consisted of battery - simple assault being the highest crime committed in Los Angeles to Burglary being the lowest as shown on Figure 3.1.

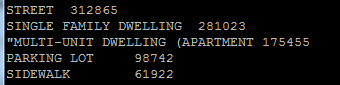


**Figure 3.1** Most common crimes from the Los Angeles Dataset from highest to lowest

The second query we created was the “Top 5 most common crimes in Los Angeles Dataset” and the top 5 consisted of Battery or Assault being the number one, then burglary from vehicle, stolen vehicle, burglary, and theft plain – petty as shown on figure 3.2. 

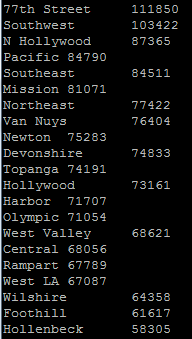
**Figure 3.2** Top 5 most common crimes in Los Angeles Dataset

These results can be vital to the city of Los Angeles because this can bring awareness to individuals or police officers to prevent these crimes from happening. The next query we created was the “Top 5 Premises where crimes were most committed in Los Angeles” and the results demonstrate that street was the highest premise where crime was committed and then it was single family dwelling, multi-unit dwelling, parking lot, and sidewalk as show on figure 3.3.

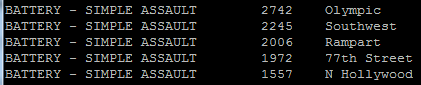


**Figure 3.3** Top 5 Premises where crimes were most committed in Los Angeles

The fifth query we created was “Area with most crimes committed” and only extracted 21 area names. The number one being, 77th Street in Los Angeles, the second area is Southwest, and the third area being North Hollywood as shown on Figure 3.4.

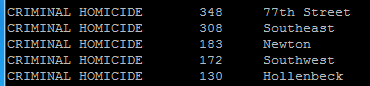
  
**Figure 3.4** Area with most crimes committed

This information can be beneficial for the Los Angeles Police Department because they can utilize our analysis by placing more police officers in those hot spot areas where crimes are committed the most and prevent future crimes from happening. The next query we created was “How frequent the most common crime is in the top 5 locations”. The results consist of two properties, it demonstrates the pattern of the most common crimes committed in the top 5 locations of the city of Los Angeles as shown on figure 3.5. The top 5 locations is Olympic, Southwest, Rampart, 77th Street, and North Hollywood with the crime battery – simple assault. This is vital data because the LAPD can use our analysis to pinpoint the actual pattern of common crimes committed in the most frequent location of Los Angeles and possibly use it to increase public safety.



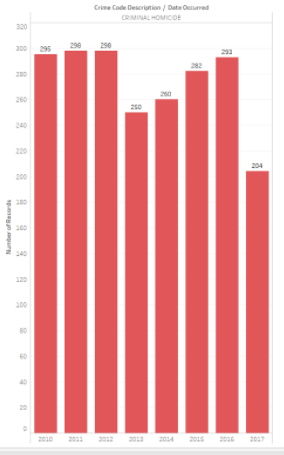
**Figure 3.5** how frequent the most common crime is in the top 5 locations

The next query we created was “how frequent “Criminal Homicide” is in the dataset” The results consist of how common Criminal Homicide is in our dataset and according to the result it is 2,180. Finally, the last query we created was “how frequent “Criminal Homicide” is in the top 5 locations”. The results demonstrate how common is criminal homicide in the top 5 locations of Los Angeles. The top 5 locations with criminal homicide are 77th Street, Southeast, Newton, Southwest, and Hollenbeck as shown on figure 3.6. This data can be useful for the LAPD because this can definitely increase public safety and decrease the number of deaths in Los Angeles by placing police officers in those five areas.

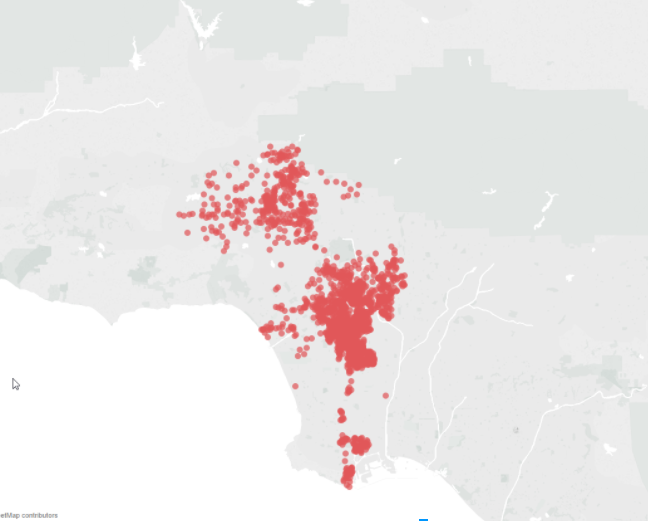


**Figure 3.6** Top 5 locations with criminal homicide

Using the tools available in Hive, we were able to filter results from our queries and store them directly to an output file. The final output file was then downloaded using Ambari as a text file. The file was then uploaded to a geospatial and temporal graphing software called Tableau as depicted in figure 3.7, which is a temporal bar graph depicting the year and the amount of homicides committed in Los Angeles. This process enabled us to create an accurate geospatial map using Tableau as presented in figure 3.8, which displays Los Angeles Homicides by geographic location.

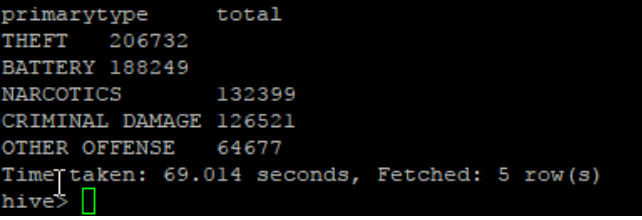


**Figure 3.7** Temporal graph showing the year and the amount of Homicides committed in Los Angeles.

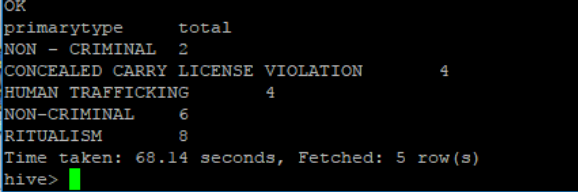


**Figure 3.8** Los Angeles Homicides by geographic location are present using Tableau

We have explored dataset from gov.data and applied principle to select more relevant type and locations of crimes in Chicago for estimation the difference in type of crime and locations. The first query is a hive statement which creates an external table named, CHICrimedata that allows Hive to query data stored in HDFS. The next hive queries create and select highest to lowest crime from CHICrimdata data. Our result consisted within 33 different crimes within the city. Therefore, our next two queries narrow down to the 5 top common crimes and top 5 least common crime in Chicago.

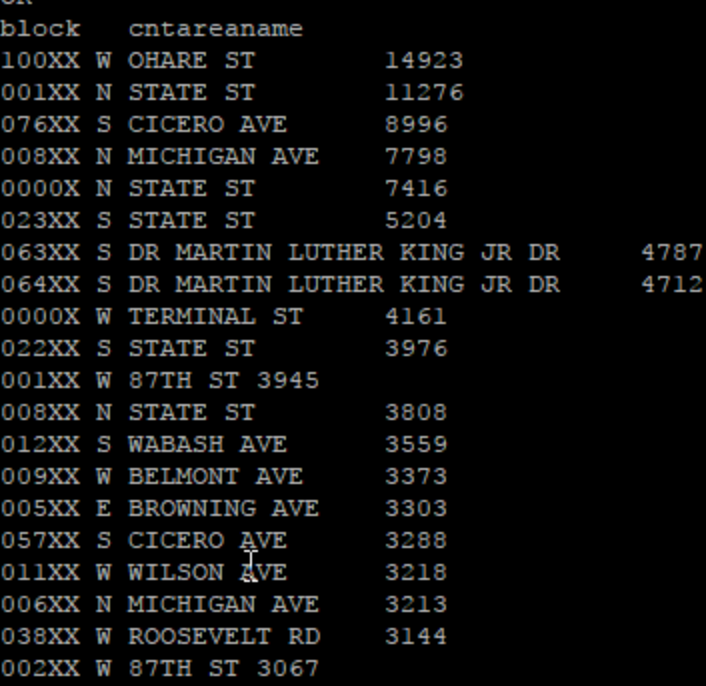
****

**Figure 3.8** The 5 most common crime in Chicago



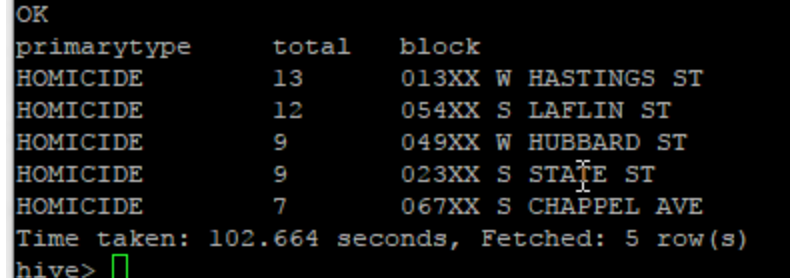
**Figure 3.9** Top 5 least common crime in Chicago

For our next query, we created the top 3 location descriptions of where crimes were committed in Chicago and the result for the main locations are street, residence, apartment and sidewalk with the total of crime committed for each. Street being the most common description bring awareness in the community. Chicago is divided within district. Therefore, next query we count how many crimes are committed in each district and group by X and Y coordinate. We also created a query for the total top 20 area with most crime committed which is group by block.

****

**Figure 3.10** Top 20 blocks with most crime committed

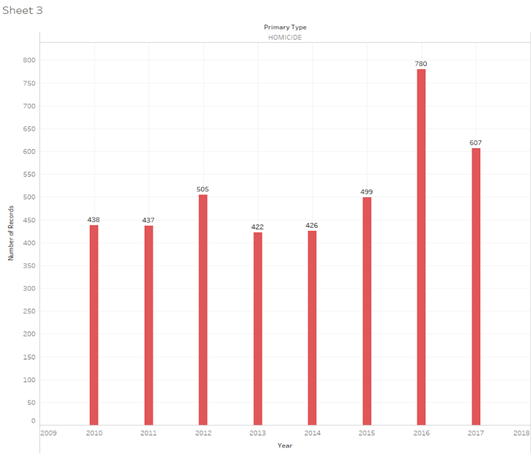
The sixth query show how frequent the most common crime is in the top 5 locations. The result shows “Theft” is the most common crime and is group by the top 5 blocks within the city from our query before. For our last three queries, we created queries based on criminal homicide. First, we look the total number of homicide committed in the city, the result show 8,787 homicide committed within 2001 to present. Next query, show how frequent homicide is in the top 5 locations.

****

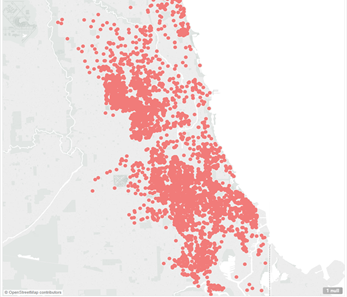
**Figure 3.11**. Result for homicide top 5 location

For our final query, we have created a query listing all homicide type crime with latitude and longitude coordination. This help the community to understand where most homicide crime is committed.

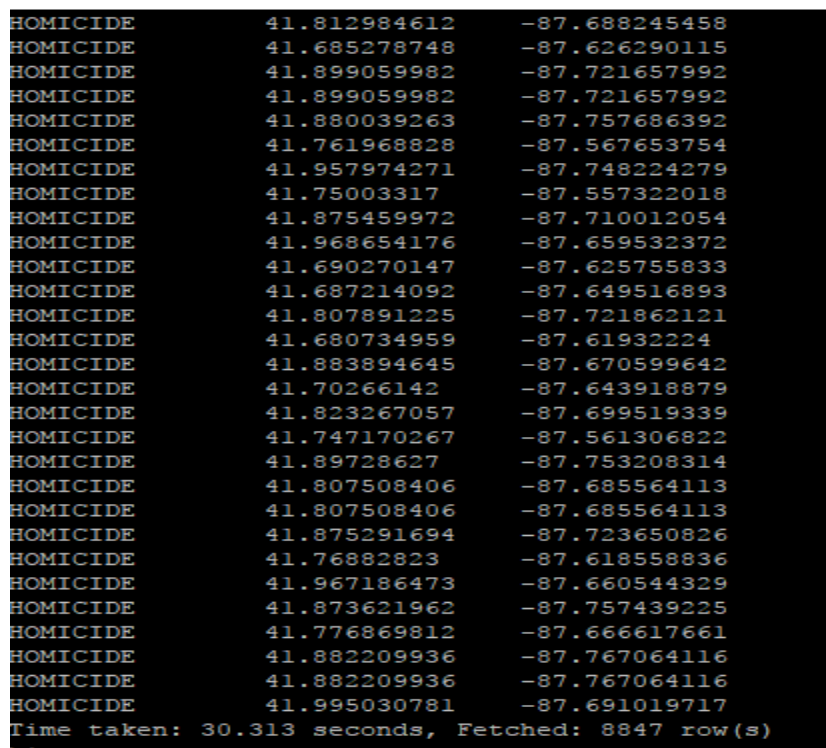
Using the tools available in Hive, we were able to filter results from our queries and store them directly to an output file. The final output file was then downloaded using Ambari as a text file. The file was then uploaded to a geospatial and temporal graphing software called Tableau as depicted in figure 3.12,which is a temporal bar graph depicting the year and the amount of homicides committed in Chicago**.**  This process enabled us to create an accurate geospatial map using Tableau as presented in figure 3.13, which displays Chicago Homicides by geographic location.



**Figure 3.12** Temporal graph showing the year and the amount of Homicides committed in Chicago.



**Figure 3.13** Chicago Homicides by geographic location are present using Tableau

****

**Figure 3.14** Result of homicide crime with latitude and longitude coordination

### References

[1][https://s3-us-west-1.amazonaws.com/dreab/Crime\_Data\_from\_2010\_to\_Pre](https://s3-us-west-1.amazonaws.com/dreab/Crime_Data_from_2010_to_Present.csv)

[2]<https://s3-us-west-1.amazonaws.com/dreab/Crimes_-_2001_to_present.csv>

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. [↑](#footnote-ref-3)