Crime rate data analysis using Hadoop

Vinay Chennupati(305079520)

Department of Information Systems, California State University

Los Angeles

**Abstract**:

Safety is the major issue anywhere. There are a lot of crimes happening every day. It would be very insightful to analyze the crime rate data to identify frequency of crimes, types of crimes, areas with a higher number of crimes etc. These insights will then have the potential to aid in proactive preventive measures by police increasing the level of safety in certain are.

Our dataset is large and unstructured. Hence, storing data in Hadoop Distributed File System would be ideal for our project. Hadoop distributed file systems provides scalable and reliable data storage in addition to enabling analytical capabilities over large amounts of data. The Hadoop map reduce architecture model implementation will be instrumental in helping us process and analyze large data sets. We also intend to use HiveQL to query the data and project a structure onto the information. The utility of using Hadoop to consume large amounts of data and enable analytical insights is what we aim to demonstrate.

**Introduction**:

There are huge number of crimes reported around the world and the crime rate is significantly growing as indicated by statistics and evaluations published by different agencies across the world every day. This is a very unfortunate but indisputable fact which we have reluctantly accepted. Some types of crimes are very alarming and cause the agony. Crimes are performed because of various reasons like poverty, alcohol, drugs, circumstances etc. which are very complex and difficult to inter-relate.

     According to the Federal Bureau of Investigation there were a total of 9,475,816 crimes reported in the year of 2014 in the U.S i.e. an average 25,960 number of crimes were carried out per day and U.S had a crime index of 50.15 during this year. Of the annual number, a total of 243,750 crimes in Los Angeles city. Our analysis is primarily focused on LAPD crime data for 2014 for Los Angeles in order to provide insight into the types of crimes, areas with higher frequency of crimes, and time windows of high frequency crimes etc. This can potentially enable law enforcement divisions in crime apprehension, crime reduction, and crime evaluation to take proactive measures to control the crime rate and ensure safety.

**Background work:**

We have used the LAPD 2014 crime dataset for Los Angeles city to carry out detailed analysis on crime trends, geographical occurrence, frequency etc.

To add a different dimension to the analysis we considered California State University Los Angeles as our focal point and proceeded to project the data based on different parameters like time and distance. This would result in extracting key findings about crimes occurring around California State University Los Angeles and in Los Angeles.

**Data Storage:**

We intend to store our unstructured data in Hadoop Distributed File System (HDFS). HDFS provides reliable and scalable storage for storing and analyzing unstructured data. It runs on top of Linux operating system and this is designed to operate on commodity hardware. Hadoop stores data in blocks thereby providing strategic data processing and fault tolerance. Hadoop was designed to have a Master-Slave architecture in which slaves are data nodes and master is name node.

**Querying:**

Data stored in HDFS can be queried using different programming languages, for our analysis we used Hive Query Language to extract the data as it provides an SQL-like structure and operates on tables just like an RDBMS, even though it does not strictly follow the full SQL standard. Hive queries are translated into Map reduce jobs which run on top of Hadoop cluster.

**Services:**

We used Microsoft Azure cloud computing platform as it provides both PAAS (Platform as a service) and IAAS (Infrastructure as a service) in addition to supporting different programming languages. For using Hadoop on Microsoft Azure, Azure HDInsight must be created and deployed. HD Insight provisions Hadoop clusters designed to store and analyze large data sets. HD Insight uses HortonWorks Data Platform (HDP) Hadoop distribution. This provides an ecosystem of components of like Hive, Pig etc.

**Analysis of the dataset:**

Our datasets consists of total 2, 43,750 rows in comma separated values (csv) format, which provides information on type of crime, area, time, location and date and other information regarding the crime. In our analysis we have found that a total of 119 distinct number of crimes have been reported in the year of 2014 and an average of 20,312 crimes per month. It was observed that there was a significant growth in total number of crimes reported from the month of April and almost constant number of crimes reported till the month of October.

From all the kind of crimes reported, the most kind of crimes which were committed are theft which is 33% followed by

traffic 19% and assault of 17% the least reported crime is criminal activities which is of 2%.As we took our university as a focal point and calculated the distance to every crime reported from our university, we found that the most of number of crimes were committed within the radius of 5-10 miles, in which again theft is the major crime reported followed by traffic and assault. The same pattern is observed when we applied our analysis for radius greater than 10.

The other important parameter in our analysis is time, where we analyzed crimes committed for every two hours interval. We observed that the most number of crimes were reported during 10:00 AM to 10:00 PM. The least is during 4:00 AM to 6:00 AM and there is significant growth during 6:00 AM to 8:00 AM.

We extended our analysis to find the areas with most number of crimes within 5 miles from the university and also all the areas in Los Angeles city. In the areas under 5 miles from the university Hollenbeck which is 3 miles away has most number of crimes reported followed Central. Surprisingly even in this areas theft is primarily reported followed by traffic and assault. In Los Angeles city the most number of crimes were reported in 77th Street area followed by Hollywood.

**Data Visualization:**

Data visualization is projecting the relevant information in pictorial or graphical form. Hive query results can be visualized in Microsoft excel. We have used hive ODBC for data visualization in excel. Apart from Excel there are other powerful tools available for data visualization/dashboard creation like R and Tableau.

**Graphical representations of few data analyzing results:**

The below figures are samples of data visualization done using hive and excel.

Figure (i): Crime Frequency in 2 hour intervals

As per the above graph, the frequency of crimes were very high during the time period of 10:00 AM to 10:00 PM, this can be clearly visualized.

Figure (ii) Total number of crimes based on distance

The above graph shows the total number of crimes committed for different radius from the university. We segregated the data for every 5 miles. It is very clear that most number of crimes are committed within 5-10 miles.

Figure (iii) Crimes based on crime types

Above pie chart describes the total number of crimes reported based on types of crimes. As u can see, theft is the most occurred crime.

Fig (iv) Crime Analysis as per month.

The above bar graph depicts the count of crimes varying for each month. The crime rate is almost constant from May to October with subtle variations.

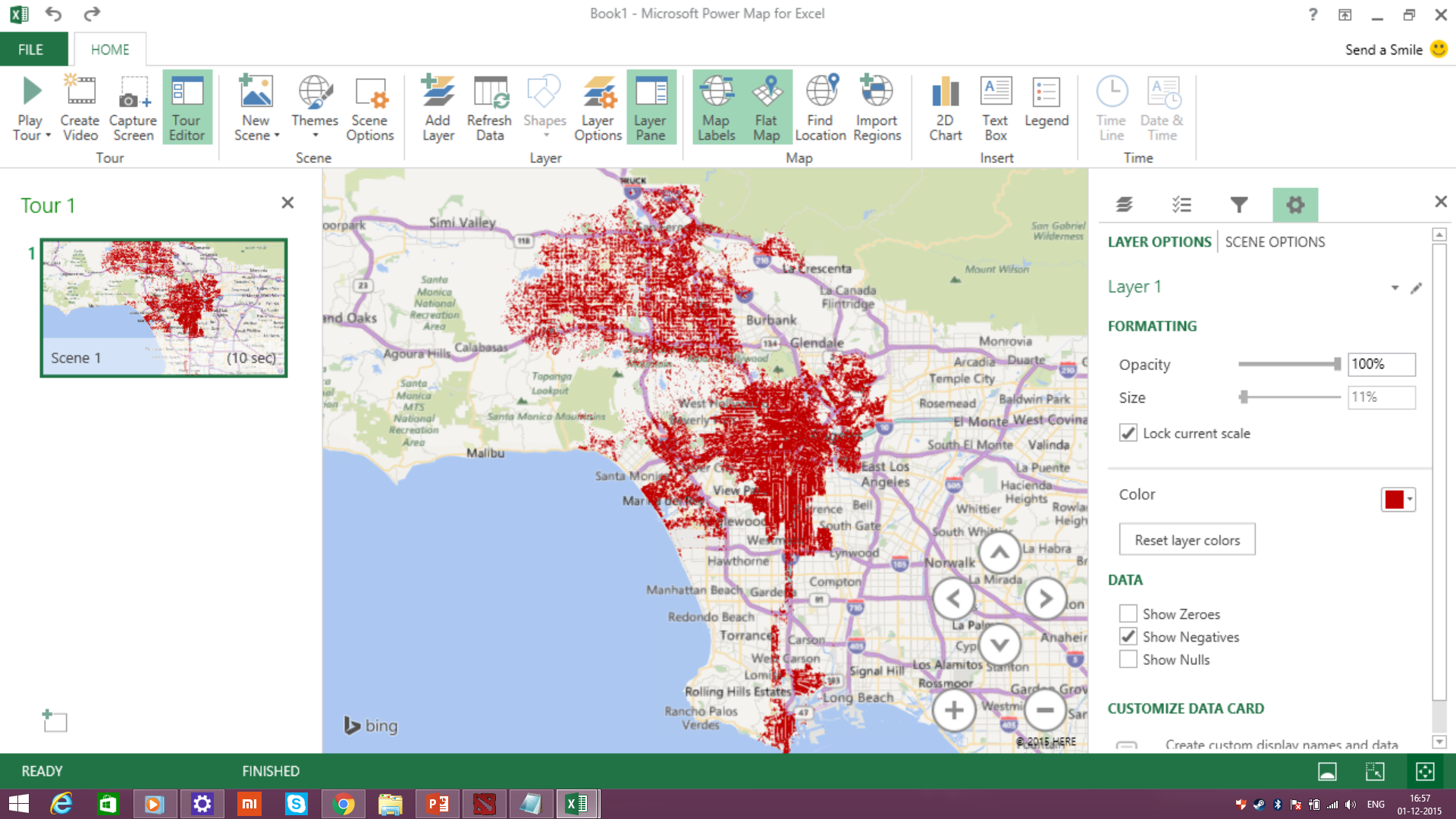


Fig (v) Geo graphical representation of all the crimes reported.

**Conclusion**:

In this project we have learnt the approach to leverage the Hadoop Distributed File System (HDFS) architecture. We have gained experience on querying the data stored in HDFS using the Hive Query Language. We also understand the mechanism used by Hive to convert queries to map reduce jobs and process the data. Also we understand the utility of using Hadoop in addition to its limitations. We also understand the difference between Hive and SQL.

In our detailed analysis we were able to project and visualize valuable data about crime occurrence in Los Angeles relative to California State University Los Angeles (CSU LA). Through our analysis we found that the major type of crime reported in Los Angeles is theft and most number of crimes were reported during 10:00 AM to 10:PM.This analysis can be instrumental in providing valuable insights regarding safety, crime frequency, time windows, and types of crimes to both law enforcement and students thereby being able to drive preventive and cautionary measures respectively.

**Dataset URL:**

<https://data.lacity.org/A-Safe-City/LAPD-Crime-and-Collision-Raw-Data-2014/eta5-h8qx>

**GitHub Code:**

[https://github.com/Ramdharan/Lapd\_Crime\_Data\_Analysis\_2014](https://data.lacity.org/A-Safe-City/LAPD-Crime-and-Collision-Raw-Data-2014/eta5-h8qx)

**References**:

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

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

<http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive/>

<http://blog.cloudera.com/wp-content/uploads/2010/01/6-IntroToHive.pdf>

<http://blog.cloudera.com/blog/2013/04/demo-analyzing-data-with-hue-and-hive/>

<https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-tutorial-get-started-windows/>

<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-power-query/>

<https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html>

<http://www.infoworld.com/article/2683729/hadoop/10-ways-to-query-hadoop-with-sql.html>

<https://docs.treasuredata.com/articles/hive-functions>

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF>

<http://www.w3schools.com/sql/default.asp>

<https://developer.yahoo.com/hadoop/tutorial/>