Analysis Prediction and Visualization of Chicago Crime Rate by Area in Real-Time

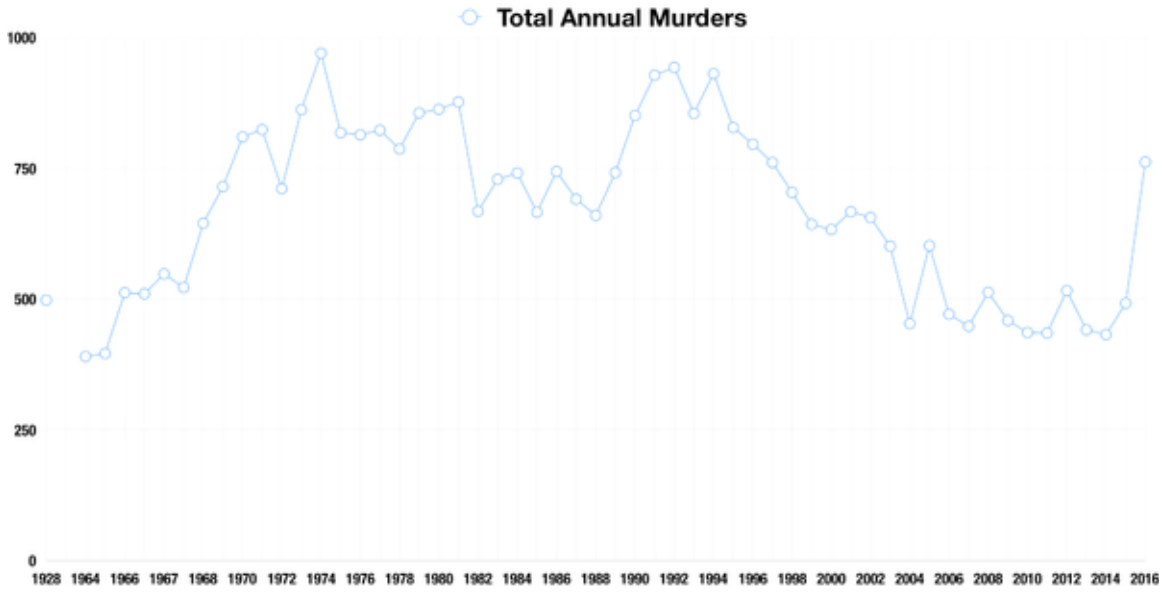
1. Problem Formulation

Chicago has the highest crime rate among all the cities in The United States of America, even greater than the major cities like Los Angeles and New York. Chicago was responsible for nearly half of 2016's increase in homicides in the US, though national crime rates stayed near historic lows. It is considered the most gang infested city in the United States, with a population of over 100,000 active members from nearly 60 different factions. Thus, there are a lot of casualties in their cross fire. Even the statistics say that murder, street crime and rapes cases have substantially increased over the last decade. Thus, resident safety becomes a major priority and should be kept in check. Hence, a system should be developed which ensures the safety of the residents.

The major goal of our project is to identify the areas in Chicago which have a high crime rate and in turn develop a system which would notify the resident the alert ratio of a certain area at a certain time. We ensure this by using the 'Chicago Crime Rate data set' which has important attributes such as the location, date, time, type of crime, etc. These will give us a clear idea of the type of crimes happening in certain area at a certain time which will then signify how dangerous that area is at that particular time. We achieve this objective by using the optimal regression method on the existing data set available from 2001-present and predict the same for the year 2017.

After this analysis, our second objective is to update the day-to-day crimes and append it to the current list in real-time. By doing this, we are enabled to analyze the change over time using a stream that is processed by Apache Storm. This method will pull the metadata from the crimes committed for that day, allowing us to see if we are correct in our predictions.

The final objective is to use this to analyze the crime-per-area in Chicago and notify the high alert areas to the users on the basis of the time. This allows citizens to avoid areas of high crime and stay safer in Chicago than they otherwise would be without this information.



2.Strategy To Solve The Problem

Using the crime data from 2001 to present obtained from the crime.cityofchicago.org website, we find which areas of the city are the most crime prone. The method for this will be to run a mapReduce on the raw data, using mappers to extract the area as the key and the crime rate per day as the value. When the reduction is finished, we will have the total crimes per region over the past 16 years. This will allow us to pinpoint the “most crime-prone” areas of chicago that we will use for the second phase of our strategy. We will archive this file and it’s results in case we need to re-run the mapReduce later.

For the second portion, we will use regression to predict future crime rates in these selected areas. To do this, we will take the past crime, type, area, date and time, per area. Each area will be a completely separate calculation using the same model to predict the future data per crime-prone region.

For the third portion, we will write a web application in Google Web Toolkit that compares our predictions with the actual data coming in for that day. The first thing in this program will be a timekeeping mechanism that records a timestamp to 2 files for redundancy. Upon starting up the program, the first thing it does is

check the files to see if it has been 24 hours since the last check. If it has, then it starts HDFS, pulls the data from the website and appends it to the end of the crime statistics file. The next phase is to pull up the archived preprocessed predictions file and show it next to the actual data that came in for that day. In this way, we can test the correctness of our predictions on a day-to-day basis.

3. Functions targeted by your software

What follows is a more detailed point-by-point breakdown of our proposed solution:

1: Set up HDFS and run a mapreduce on the raw data:

- Setting up Hadoop cluster
- Pulling the relevant data from the raw data to feed into our machine learning algorithm

2: Machine Learning

- Get the data from the original map reduce using a new mapper
- Map and reduce to apply regression technique to the data to predict future crime rate
- Final conditioned data will be crime rate per region for future days

3: User Interface

- Take the data from the original prediction and show it next to crime data for that day.
- Download new data if 24 hours or more has elapsed and append to the HDFS crime data.

4: Testing

- Unit testing for each module.
- Integrating the 3 different programs
- testing on corner cases, border cases and predictable points of program fault/failure
- test on a large range of other inputs, especially for the UI and Machine learning section.

4. The Dataset

The Fields for this dataset (a single Relation gained from the Chicago crime data website: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data) are as follows:

ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description,

Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y

Coordinate, Year, Updated On, Latitude, Longitude, Location

Of these, Yash will be using the District and Ward location information as well as Date and primary type to find the most crime-prone areas in the past. Everything else will be discarded by the preliminary map reduce as well as by the day-to-day updates.

5. Project timeline (weekly plan)

The following table shows bi-weekly plan for the project.

Sr. No.	Work Task	Team Member Roles (Development of Code)	Period For coding (weeks/days)	Period for Unit Testing (weeks/days)
1.	Setting up Hadoop and load crime data into the HDFS.	Yash Raikar, Yashad Samant John Scherer - Combined work	1 day	-
2.	Hadoop Pre-computing and analysis	Yash Raikar Will take the dataset and extract relevant data, the output of the hadoop will be the most crime prone areas by district and ward. From these we will choose to use districts or wards or both to implement the second and third phases of the assignment	1 week	3 days
3.	Machine Learning using Regression technique	Yashad Samant will take the output of the Map Reduce on the raw data and use the optimal regression technique to infer future crime rates from the past crime data. This will be used as a static (or dynamic day-by-day prediction) to be compared to actual data.	1 week	3 days
4.	Web application development	John Scherer Using GWT, Will create a User interface using HTML and CSS, with a Java logical layer. Does an HTTP request to the chicago database and appends	1 week	3 days

		to the previous day's data in HDFS. Will display predictions next to actual crime data for that day.		
5	Final Jar deployment on Hadoop cluster and Testing	Yash Raikar, yashad Samant John Scherer Testing, (unit and integration), Deployment.	2 weeks	-

6. Bibliography

- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data
- https://en.wikipedia.org/wiki/Crime_in_Chicago#Chicago_street_gangs