[[1]](#footnote-1)

Analysis of Safety Measures and Crime Prevention in LA City

Srivatsan Ananthakrishnan, Department of Computer Science, University of Bridgeport, CT

*Abstract*—: Los Angeles (LA) Police Department provides open datasets showing their progress of city’s performance and services in terms of crime rate and safety. This dataset was created by LA Police Department and it is updated on a regular basis. The crime data was obtained from the US government website as csv format [5].

In this project, we present an analysis of crime data which comprises over 2 million records and we analyze solutions to various problems in terms of multiple factors. The solutions for those problems has been analyzed using some of the BigData solution techniques like MapReduce(using python), Hive, Spark(pyspark) and Big query in Google Cloud Platform(GCP).

**Keyword— Spark; BigQuery; MapReduce; hive**

# Introduction

The crime datasets are considered to be one of the largest dataset which contains millions of records and cover millions of people. This dataset is updated by Los Angeles Police Department who feed the updated crime to this dataset on a daily basis. The purpose of this analysis mainly focus on analyzing the safety area to live and rank them based on their crime rate. The other idea deals with predicting the approximate crime occurrence in a specific area/demise in the future through the analysis [5].

Thus the above mentioned goals are expected with the proper analysis of the Crime dataset of LA City. The history of LA city crime rate is way ahead compared to other important cities in the United States. With the information we have on the crime data of LA city, we can analyze the crime rate from 2010 to present and see whether the rate has been reduced or not.

This analysis helps us in finding how the crime rates affect the Asian Young women’s, who are the major victims for the past 5 years. Thus, with the data we have, an analysis is performed on the city, place, demise, area etc., and all possible fields collected from the dataset and to come with a solution in which place, area, demise ranks most number of Asian young women victim and which is the most ranked crime type (shooting, vehicle stolen etc., )

In particular, this dataset consists of almost 250,000 incidents that involves only Asian with Female gender from 2010 to present. Thus, analyzing on this particular feeds will definitely route through in coming up with some serious analysis report that helps in minimizing the Asian victims in Los Angeles city in the near future.

# CRIME DATA ANALYSIS

The chosen crime data has been analyzed in various steps by deriving a set of problem statements. The analysis has been done in order to explore various big data modules the Hadoop file system (HDFS) for storage and other Hadoop MapReduce, HIVE, Big Query and Spark mainly for analytical purpose.

*Hadoop Distributed File System (HDFS)*

The HDFS is a storage space where we can store huge amounts of data and can easily access the stored data. The HDFS is highly fault tolerant and stores the data across multiple machines which gives a way for parallel processing. The files are stored in redundant fashion to rescue that helps to prevent the loss of data during any kind of system failure. In this study, the chosen big data (Crime data) has been stored in HDFS [1].

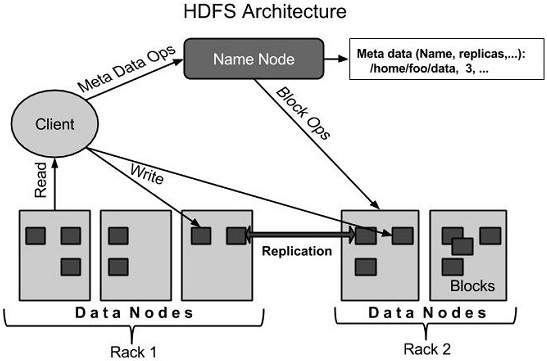


Figure 1. *HDFS architecture*

*MapReduce*

Map Reduce is a programming model mainly used to access and analyze the big data stored in the HDFS. It mainly comprises of two different tasks namely the mapper and reducer. The two tasks can be written using any of the programming languages like Java, Python, C++ etc. The two tasks can be written based on some algorithms.  Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model [2].

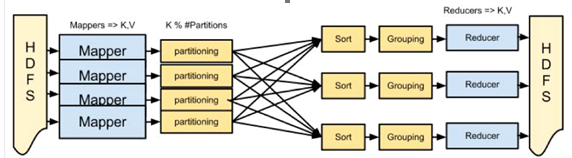


Figure 2. *MapReduce framework*

*Hive*

Hive is a programming tool that is mainly used to process the structured data in Hadoop. It is placed on the top of Hadoop mainly to summarize the big data by making the querying/analyzing easy.

Features of Hive

* It stores schema in a database and processed data into HDFS.
* It is designed for OLAP.
* It provides SQL type language for querying called HiveQL or HQL.
* It is familiar, fast, scalable, and extensible. [3]



Figure 3. *Hive Architecture*

*BIG QUERY*

Big Query is an enterprise data warehouse, a product of google that makes storing and querying massive datasets easier. Big query uses super-fast SQL queries by using the processing power of Google’s infrastructure [4].

*SPARK*

Apache Spark is practically a fast and general engine for big data processing. In memory data processing of Spark improved the performance up to 10 times faster than on-disk data processing. In addition, Spark offers over 80 high-level applications that can interactive with different programming languages such as Java, Scala, R and Python (which is used in this paper). The Spark with combined version of Python is pyspark. This feature shows the Spark ease of use compared to Hadoop platform. Spark can easily handle data streaming and real-time data processing. Also, it can easily interact with SQL databases and data frames. In this paper, the implementation for various problems has been solved using Spark by interacting with SQL databases where the dataset is converted to a data frame which made to use SQL functions to extract the desired result [6].

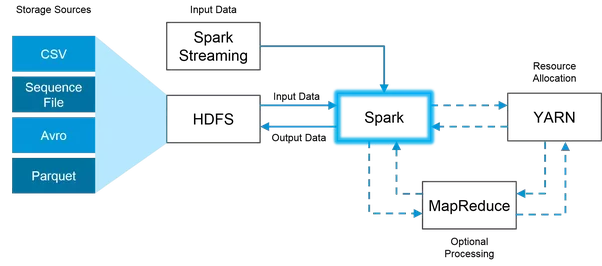


Figure 4. *Spark framework*

# Problem Statement and Implementation

The Crime Data has been analyzed through a set of problem statements as follows,

1. Analyzing how safe the city is for Asian Women to live. This Problem statement is solved using Hadoop MapReduce.
2. Analyzing number of crimes in each and every crime category of all the areas in the city. This Problem statement is solved by Hive QL.
3. Analysis of total number of crimes in each type from 2010-present using Spark (pyspark).
4. Analyzing total number of crimes area wise and which crime has occurred the most in each area using Spark
5. Top 5 number of crimes occurred in each type from 2010- present by using Google BigQuery
6. Top 10 type of crimes happened to women's in the year 2017 is analyzed using Google BigQuery.
7. Number of crimes that has involved theft of $950.01 and over every year is analyzed and solved using Google BigQuery.
8. Analysis of number of crimes occurred in a particular area in the years 2017 and 2018 using Google BigQuery.

# RESULTS AND EVAULATION

The Analysis of all the problem solving techniques which is discussed in previous section is being evaluated based on the time taken to process the BigData that produces results for those problem statements.

Apache Spark has more advanced cluster compute engine than MapReduce and MapReduce limits to batch processing, thus Spark runs bit faster than MR.

1. MapReduce: 4minutes 43 seconds

2. HiveQL: 196.882 seconds

3. Spark: 2 minutes 15 seconds

4. BigQuery: 35-50 seconds

# conclusion

In this study, we obtained a statistical report on the safety measures and crime preventions in LA city which is rated high amongst other US major cities. I expect to gain knowledge in setting up HDFS, uploading the data to the HDFS, solving the computations over the data using the MapReduce and spark programming, Hive and Google BigQuery. Other than the technical skills, this will improve the analytical skills too.

References

## [1] <https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm>

## [2] <https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm>

## [3] https://www.tutorialspoint.com/hive/hive\_introduction.htm

## [4] <https://cloud.google.com/bigquery/what-is-bigquery>

*[5] Sarpreet Kaur, Dr. Williamjeet Singh, Department of Computer Engineering, Punjabi University, Patiala, India “Systematic Review of Crime Data Mining”, ISSN No. 0976-5697.*

*[6] Lenin Mookiah, William Eberle and Ambareen Siraj Tennessee Technological University Cookeville, TN 38505. “Survey of Crime Data Analysis and Prediction”, ISSN No. 0976-5697.*

# APPENDIX

# Supporting Materials for Implementing Crime Data Analysis

Link to download the Crime Dataset used in the project: (.csv recommended)

<https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq/data>

Source Code

*MapReduce Programming*



*Hive*



*Spark Programming*



*BigQuery in GCP*



*Results of BigQuery (.csv)*

**

1. [↑](#footnote-ref-1)