View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UB ScholarWorks

Trends analysis of Vehicle Collisions using Hadoop and Spark

Akhilesh Ajit Borgaonkar; Advised by : Dr. Jeongkyu Lee
Department of Computer Science and Engineering
University of Bridgeport, Bridgeport, CT

Abstract

Vehicle collisions ranks amongst the main reasons for increased fatalities since few years. Vehicle collisions is the leading reason for fatalities and injuries in the age group below 24 age. The number of lives taken by vehicle collision in 20th century was equivalent to the number of fatalities in World War II. Therefore, necessary actions are required to be taken to bring down the number of collision incidents. The trends in vehicle collision incidents focuses on analyzing the changes in collision incident patterns over past few years. The main intention of this project is to find the pattern of incidents happening daily with respect to time, day, location and reason. If the trends are analyzed, the prediction of similar incidents can be confirmed and precautionary actions can be taken.

Introduction

It has been estimated that approximately 1.25 million fatalities and a million more injuries have been reported worldwide due to vehicle collisions in 2013. It was the leading cause of death majorly for the age group below 24. Subsequently, there has been a marked decrease in the number of fatalities and injuries in 2014 and later. But, still the number of incidents is not low. Vehicle collision is still amongst the major reason of fatalities today. The total number of fatalities due to vehicle collision in 20th century is approximately close to 60 million which is equivalent to the fatalities in world war II. I think this issue is something to be looked about. By analyzing the trends and predicting the incidents, we can bring down the number of fatalities and causalities. Deaths due to vehicle collision is amongst the leading causes of death since the invention of vehicles to be specific, it is 9th cause of every death worldwide. There have been few efforts to bring down the deaths caused by vehicles by various safety campaigns and acts to install mandatory safety equipment in every vehicle.

The trends in vehicle collision incidents gives a broad view about the changes in pattern of vehicle collisions. It will give you an idea about the maximum number of vehicle collision incidents happened due to the mistake of driver or pedestrian or other factors. The efforts are to analyze this trends per specific areas for better visualization. The trends that are to be analyzed include the growth/decline in the number of deaths and injuries caused by vehicle collisions, the trends in incidents for a reason in an area and finding the safest areas to travel at the time of the day according to the trends in collision.

Problem Statements

1. Finding the trend in decline or growth in number of fatalities or injuries caused by vehicle collisions over last few years.
2. Finding trends in collisions due to a reason in an area.
3. Aggregating all trend results to find the safest areas in city to travel at a specific hour of a day.

Methodologies

1. MapReduce Programming

   To analyze the dataset using MapReduce methodology, one mapper class, 5 reducer classes, 5 partitions are implemented. The key values omitted from mapper are sorted and shuffled to input reducers and later aggregated to get the count of similar keys.

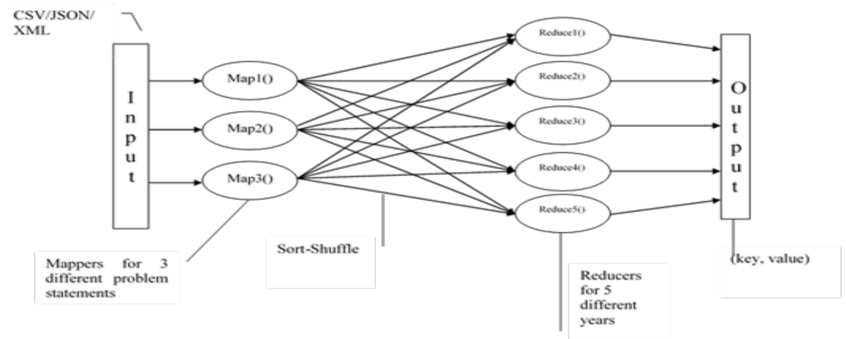

Figure 1. MapReduce Modelling

2. Hive

   To analyze the dataset, a simple hive query is designed using HiveQL. Hive provides SQL-like interface to manipulate the large datasets on HDFS. The designed query carries out MapReduce functions and results into number of incidents per location.

3. Spark

   Spark is another faster and simple implementation of MapReduce programming. I have implemented the spark query using mapper and reducer functions to get the results for safest time to travel in the city.

Results

The results after implementation on NYPD sample data are as follows:

1. Most number of fatalities due to vehicle collisions were recorded in 2015 in Brooklyn.
2. The most seldom reason for majority of vehicle collisions in the state of New York happened for the reason “Driver inattention/Distraction” in Manhattan.
3. Safest time to travel is before 8 AM

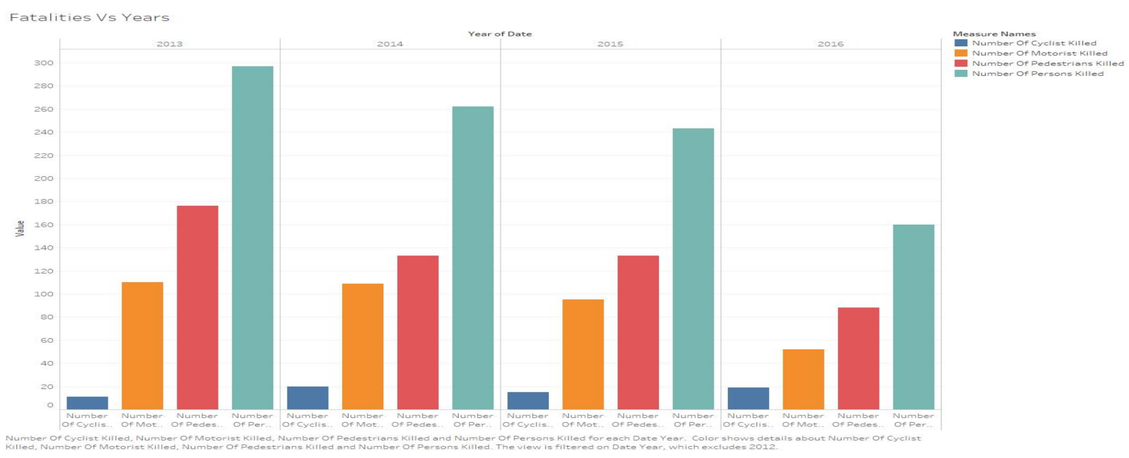

Figure 2-A. Fatalities vs Years

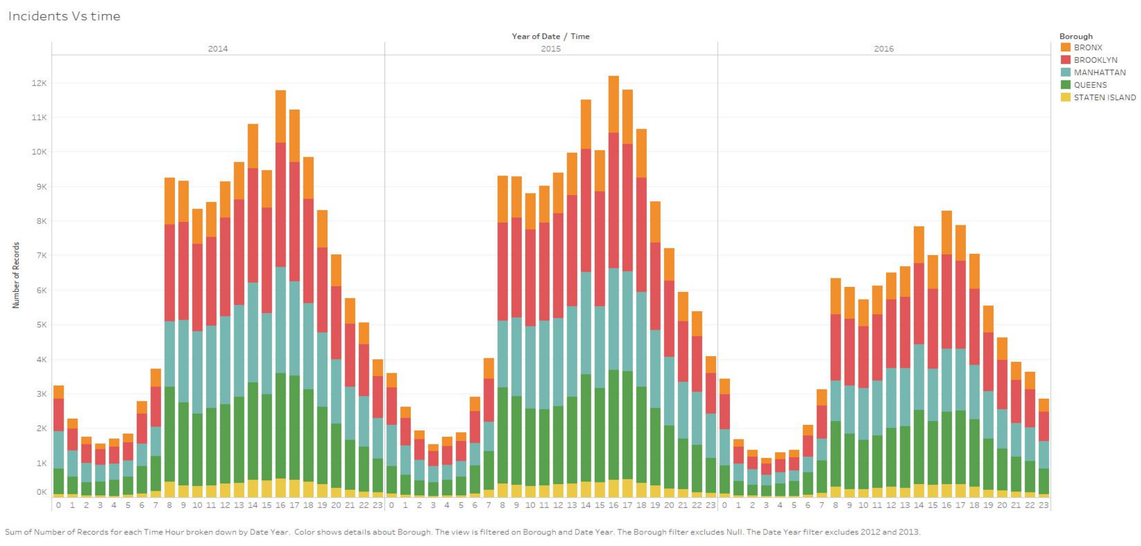

Figure 2-B. Incidents vs Time (Hours)

Conclusion

Hadoop is an emerging technology helping to analyze large amount of data at once. This project helped me learn how to upload datasets on Hadoop file system and different practices to analyze the data on Hadoop file system. This project taught me how to implement MapReduce Programming, Hive Query designing and Spark programming using python.

References

1. D. Jeffrey, S. Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM, vol. 51, pp. 107-113, Jan. 2008.
2. Tom White, “Hadoop: The definitive guide”, 3rd Edition, O’reilly publications, May 2012.