Pallavi Aggarwal

pallavia@bu.edu

Abstract

This is a Term End Project Report for MET CS 779

Bostom Crime Analysis

Using Apache Spark

Table of Contents

[Introduction: 2](#_Toc27340263)

[Big Data 2](#_Toc27340264)

[Apache Spark 4](#_Toc27340265)

[Benefits of Apache Spark over other alike platforms: 5](#_Toc27340266)

[Apache Spark Ecosystem 5](#_Toc27340267)

[Advantages of Apache Spark 6](#_Toc27340268)

[How spark works on a cluster 6](#_Toc27340269)

[Databricks 7](#_Toc27340270)

[Boston Dataset 8](#_Toc27340271)

[Loading the Data 9](#_Toc27340272)

[Data Processing 9](#_Toc27340273)

[Analysis 11](#_Toc27340274)

[Analyse Trends Over the years 12](#_Toc27340275)

[What is the most type of crime? 13](#_Toc27340276)

[Motor Vehicle Accident Response Analysis 14](#_Toc27340277)

[Where do crime happen? 15](#_Toc27340278)

[Roxbury Analysis 18](#_Toc27340279)

[When does crime happen? 20](#_Toc27340280)

[Analyse the shooting incidents 21](#_Toc27340281)

[Predictive Models (Python) 23](#_Toc27340282)

[Model 1 - Predicting Type of Crime 23](#_Toc27340283)

[Model 2 – Predicting the district where crime will happen 25](#_Toc27340284)

[Reference 26](#_Toc27340285)

# Figures

[Figure 1: Boston Crime Index 2](#_Toc27339739)

[Figure 2: V’s of Big Data: 3](#_Toc27339740)

[Figure 3: Apache Spark 4](#_Toc27339741)

[Figure 4: Benefits of Apache Spark 5](#_Toc27339742)

[Figure 5: Apache Spark Ecosystem 6](#_Toc27339743)

[Figure 6: MapReduce technique 7](#_Toc27339744)

[Figure 7: Spark running on a cluster 8](#_Toc27339745)

[Figure 8: Databricks 9](#_Toc27339746)

[Figure 9: Importing the Dataset 10](#_Toc27339747)

[Figure 10: Data Processing - 1 11](#_Toc27339748)

[Figure 11: Data Processing - 2 11](#_Toc27339749)

[Figure 12: Data Processing - 3 12](#_Toc27339750)

[Figure 13: Data Processing - 4 12](#_Toc27339751)

[Figure 14 - Data Processing - 5 12](#_Toc27339752)

[Figure 15: Dataset conversion from Pandas to Spark to SQL 13](#_Toc27339753)

[Figure 16: Crime analysis 13](#_Toc27339754)

[Figure 17: Time series graph 14](#_Toc27339755)

[Figure 18: Most Committed Crimes 14](#_Toc27339756)

[Figure 19: Crime type analysis 15](#_Toc27339757)

[Figure 20: Analyse Motor Vehicle Accident 15](#_Toc27339758)

[Figure 21: Analyse Motor Vehicle Accident for each hour 16](#_Toc27339759)

[Figure 22: Most unsafe districts 16](#_Toc27339760)

[Figure 23: Most unsafe streets 17](#_Toc27339761)

[Figure 24: Crime report plot 18](#_Toc27339762)

[Figure 25: Roxbury View 18](#_Toc27339763)

[Figure 26: Crime in Roxbury 19](#_Toc27339764)

[Figure 27: Unsafe streets in Roxbury 19](#_Toc27339765)

[Figure 28: Time analysis of Roxbury 20](#_Toc27339766)

[Figure 29: Monthly analysis of Crime of Boston 20](#_Toc27339767)

[Figure 30: Hourly analysis of crime in Boston 21](#_Toc27339768)

[Figure 31: Weekday analysis of Crime in Boston 21](#_Toc27339769)

[Figure 32: Shooting analysis 22](#_Toc27339770)

[Figure 33: Shooting across years 22](#_Toc27339771)

[Figure 34: Shooting in each district 23](#_Toc27339772)

[Figure 35: Machine Learning Packages 24](#_Toc27339773)

# Introduction:

According to FBI Crime Data, Boston is not a safe community. When compared to Massachusetts, Boston has a higher crime rate than 97% of the cities. The probability of being a victim to a criminal act is 1 in 34 in Boston which is considerable high.

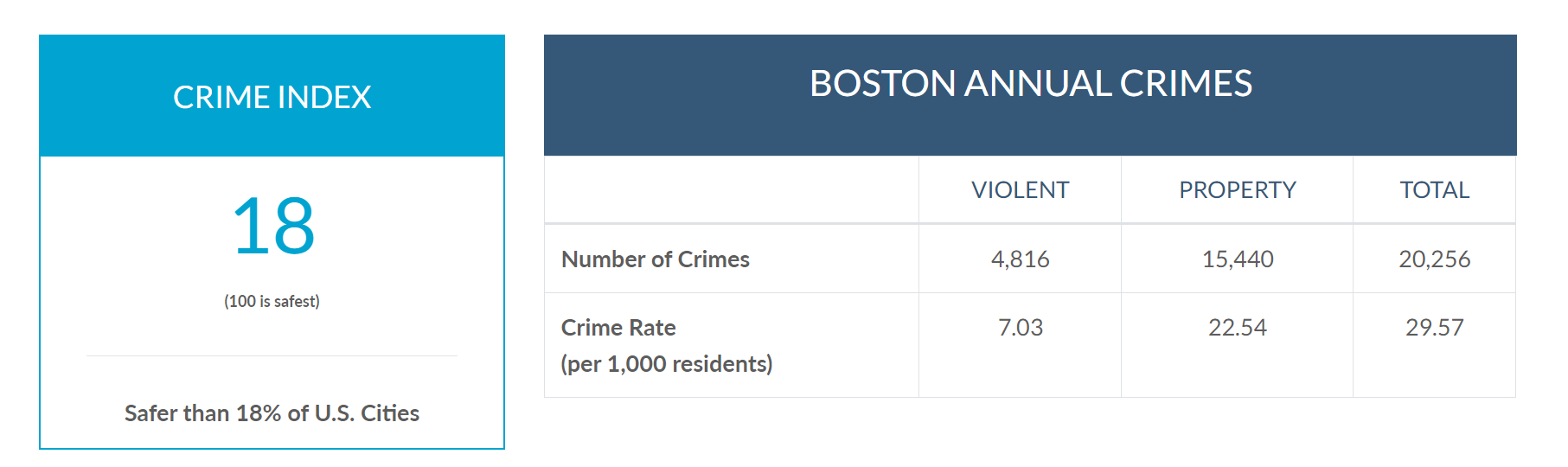


Figure : Boston Crime Index

(Source: <https://www.neighborhoodscout.com/ma/boston/crime>)

The objective of this research is to analyse crime in Boston city from 2015 to Present Date. The objective is to understand the crime with the help of data, where do crimes happen, what kind of crimes happen. Have the crime trends changed over the years? Is there any way to predict crimes before they happen?

To answer all these questions, we will analyse a dataset about Boston Crime using Big Data Analytics using Apache Spark

## Big Data

The best way to understand Big Data is by understanding Gartner’s definition from 2001: Big data is data that contains greater variety arriving in increasing volumes and with ever-higher velocity. This is known as the three Vs. Simply put, big data is a large, complex dataset collected from different sources. These datasets are so huge that the traditional data processing software fail. These huge datasets are usually unstructured data and can be in any form, text, image, video or any other forms.

Volume: Big data deals with huge amount of data. For some organizations it might be terabytes of data, for others it might tens of terabytes.

Velocity: Velocity is the rate at which the data is received. Real time data is received and operated on quickly. The data is directed to memory instead of the drive to make the process faster.

Variety: Variety refers to the many types of data that are available. Traditionally the datasets were structured and were relational in nature, with Big Data we deal with huge amount of unstructured data from different sources in different formats.

For example, Twitter gathers Terabytes of data every day in the form of text, comments, likes/dislikes, images, videos, audio etc. (Volume) All this data is unstructured or semi structured (Variety). This data is then processed in real time to make sense of it, support ads and support the metadata. (Velocity)

Another V has emerged over the past few years:  veracity.

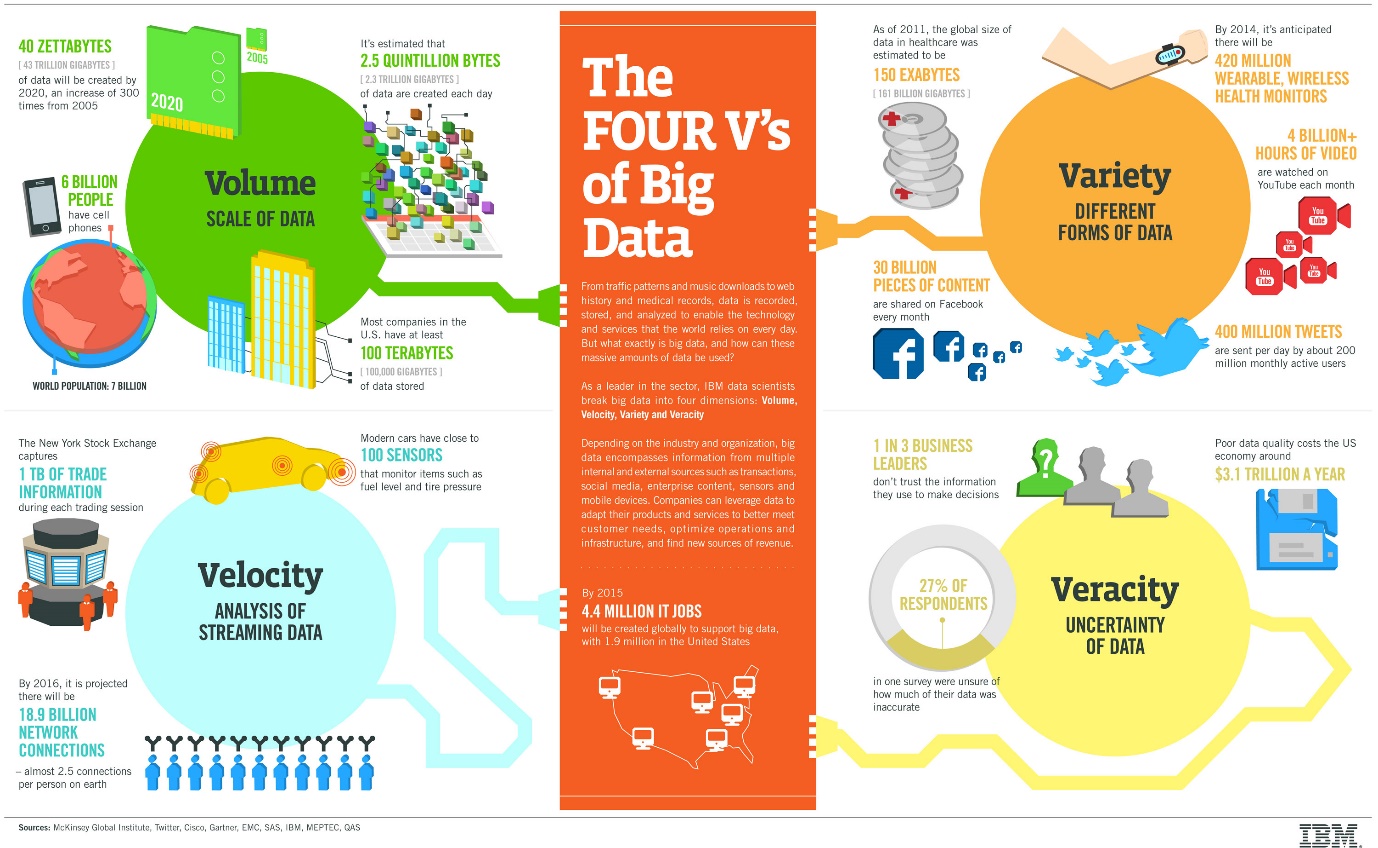


Figure : V’s of Big Data:

(Source: <https://www.ibmbigdatahub.com/infographic/four-vs-big-data>)

Today Data is the one of the largest assets of some of the world’s biggest technology companies. Large teams are working to build data driven products to bring more value to the organization. Now that the cost of data storage and computing has reduced to due to recent technological advancements, it is easier to store large amount of data and process it at a cheap cost.

The development of open-source framework such as Spark, Hadoop is essential as they help the growth of big data as they make it easier to work with big data at a cheaper cost. With the advancement of the Internet of Things (IoT), more objects and devices are connected to the internet, gathering data on customer usage patterns and product performance. The emergence of machine learning has produced still more data.

While big data has come far, its usefulness is only just beginning. Cloud computing has expanded big data possibilities even further. The cloud offers truly elastic scalability, where developers can simply spin up ad hoc clusters to test a subset of data.

To work with big data:

1. Integrate: In traditional databases ETL(extract, transform and load) technique was used to integrate databases but with Big data since the data is from multiple sources in different formats, ETL cannot be used. New techniques are used to integrate terabytes of data to analyse it
2. Manage: Large amount of data requires large storage. The storage solution is cloud. You can store your data on cloud, as it is cost effective, supports current computing requirements and spin up resources as needed.
3. Analyse: Analyse the transformed data

This is the final step where a processed data is analysed to answer the business questions.

Since traditional relational databases cannot handle big data, many organizations use NOSQL databases. The use frameworks like Hadoop, Yarn, MapReduce, Spark, Kafka etc to analyse data.

Big Data Use Cases-

* Product Development
* Predicting Sales
* Inventory Management
* Fraud Detection and other Risk Analysis
* Predict Failures and maintain them
* Customer Experience
* Operation Efficiency

## Apache Spark

Apache Spark is the largest 100% open source project in data processing. It is a lightning-fast unified analytics engine for big data and machine learning. It was developed by UC Berkeley’s AMPLab in 2009. Since its release, Apache Spark, has been adopted by a enterprises across different sectors. Technology companies such as Netflix, eBay have processed petabytes of data on clusters of over 8000 nodes.



Figure : Apache Spark

(Source: <https://databricks.com/spark/about>)

Spark is preferred over other cloud-based software as Spark combines different techniques and processes into a single software. Data can be selected, transformed and analysed all in the same framework which traditionally takes several different software and processes.

Spark jobs perform multiple operations consecutively, in memory, only spilling to disk when required by memory limitations. Spark simplifies the management of these disparate processes, offering an integrated whole – a data pipeline that is easier to configure, run, and maintain. In use cases such as ETL, these pipelines can become extremely rich and complex, combining large numbers of inputs and a wide range of processing steps into a unified whole that consistently delivers the desired result.

### Benefits of Apache Spark over other alike platforms:

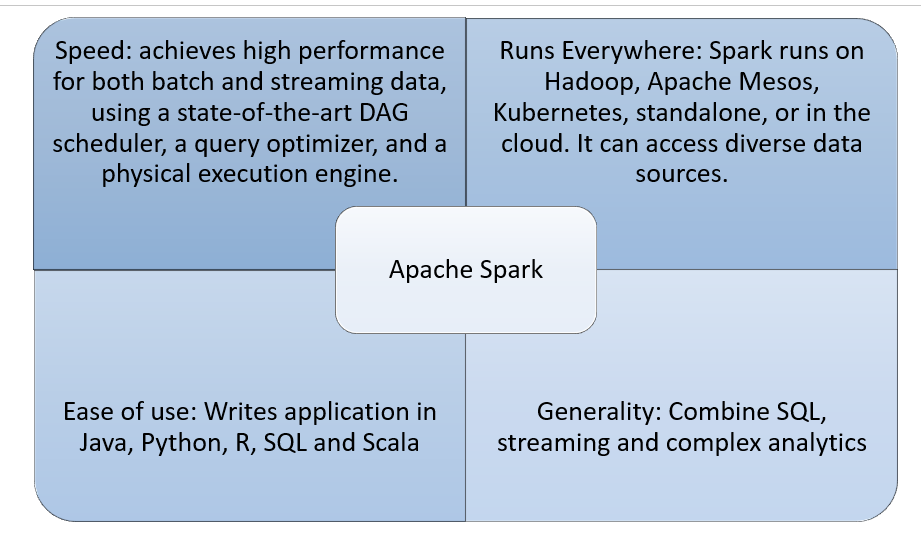


Figure : Benefits of Apache Spark

(Source: <http://spark.apache.org/>)

### Apache Spark Ecosystem

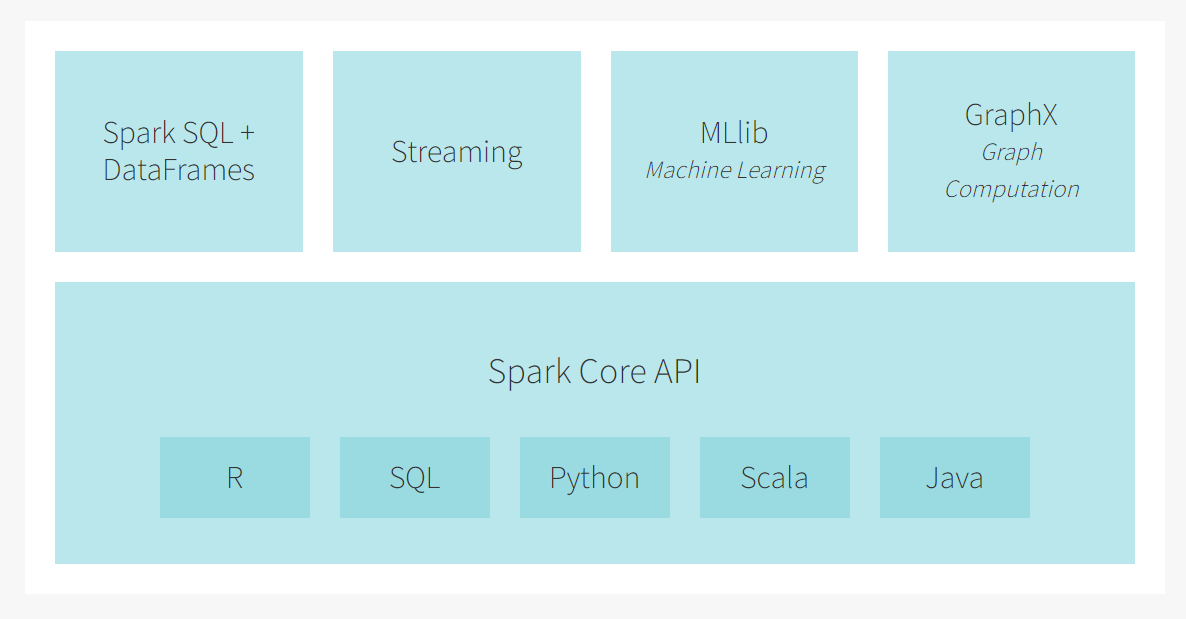


Figure : Apache Spark Ecosystem

(Source: <https://databricks.com/spark/about>)

### Advantages of Apache Spark

Before Spark, there was Hadoop which used MapReduce Technology. MapReduce is framework that enables Google to index web pages across clusters, MapReduce Framework (2004). In MapReduce multiple jobs work on clusters in parallel using an iterative algorithm, this requires a lot of reading and writing on the disk therefore the overall process is slow.

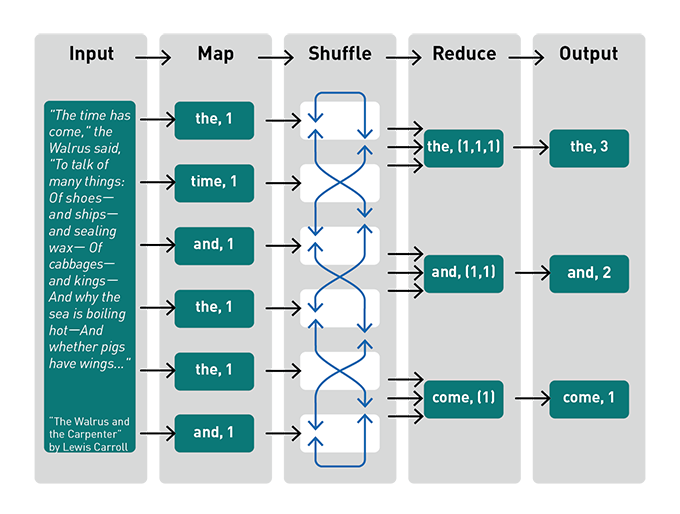


Figure : MapReduce technique

(Source: <https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/>)

The goal of the Spark project was to keep the benefits of MapReduce's scalable, distributed, fault-tolerant processing framework while making it more efficient and easier to use. Spark is designed for speed:

Spark runs multi-threaded lightweight tasks inside of JVM processes, providing fast job startup and parallel multi-core CPU utilization.

Spark caches data in memory across multiple parallel operations, making it especially fast for parallel processing of distributed data with iterative algorithms.

Spark provides a rich functional programming model and comes packaged with higher level libraries for SQL, machine learning, streaming, and graphs.

### How spark works on a cluster

* The Driver Program hosts Spark session to coordinate all the independent processes(nodes) in a cluster.
* The cluster manager assigns tasks for each partition
* Once the task is completed a new output partition is created and the operation is repeated on this new partition. (Iterative algorithm)
* Results are sent back to Driver Program and stored on Disk

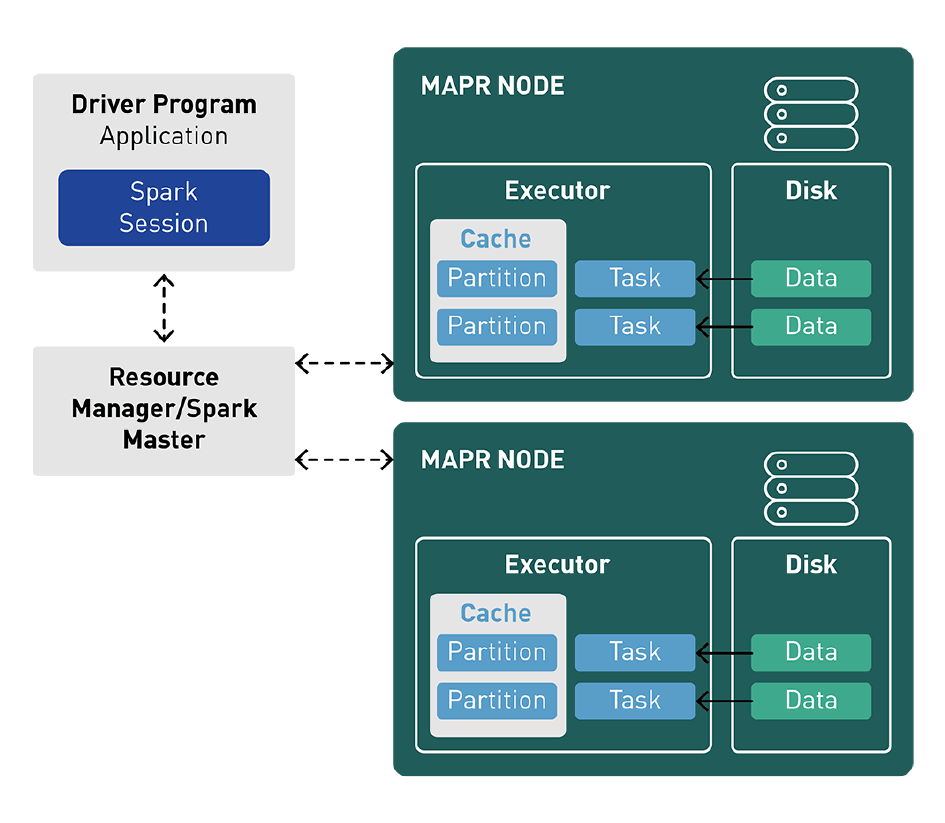


Figure : Spark running on a cluster

(Source: <https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/>)

## Databricks

Databricks Unified Cloud based Big Data Processing Platform. It enables collaborates data science and machine learning on the data. It was founded in 2013 by researchers at University of California, Berkeley to speed up processing jobs in Hadoop System.

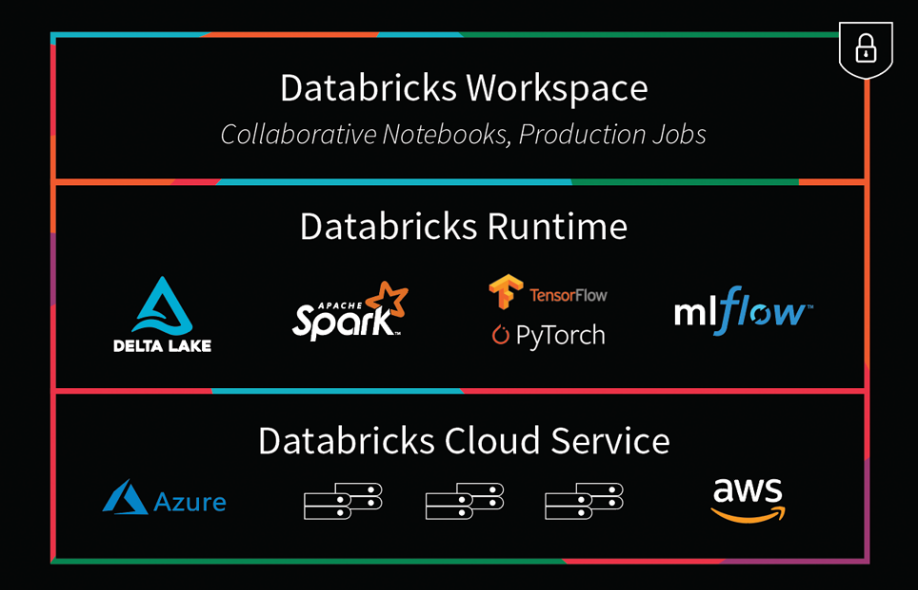


Figure : Databricks

(Source: <https://databricks.com/spark/comparing-databricks-to-apache-spark>

# Boston Dataset

The dataset on Boston Crime was taken from [Boston Gov](https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system). The crime reports in the dataset are provided by the Boston Police Department (BPD). The dataset has the crime reports recorded from June 2015 to the present date.

The dataset has ~ 440,000 incident reports(rows) and 17 columns.

The columns are:

* + INCIDENT\_NUMBER – A unique id of the report
  + OFFENSE\_CODE – A unique id for the offense group
  + OFFENSE\_CODE\_GROUP - All the crimes are segregated in different groups depending on the type of crime
  + OFFENSE\_DESCRIPTION – description of the crime
  + DISTRICT – A unique code for each district in Boston Area (by BPD)
  + REPORTING\_AREA – A unique code for each reporting area (by BPD)
  + SHOOTING - If there was shooting or not in that particular crime
  + OCCURRED\_ON\_DATE – date on which the crime happened - timestamp
  + YEAR
  + MONTH
  + DAY\_OF\_WEEK
  + HOUR
  + UCR\_PART - Boston is divided among 3 UCR parts
  + STREET – the name of the street where crime happened
  + Lat – Coordinates of the crime location - Latitude
  + Long – Coordinates of the crime location - Longitude
  + Location – Coordinates of the crime location

Now a .csv format file was downloaded and loaded in Spark as a Spark Dataframe.

All the processing is done in a Databricks Notebook. [Source Code](C://Users/palla/OneDrive/Desktop/CS%20779%20Project.html)

# Loading the DatA



Figure : Importing the Dataset

The file was converted to Pandas dataframe (Python) and to a SQL table as the analysis is done in both the languages interchangeably.

# Data Processing

Once the data is loaded and converted into a Pandas datafame, the data is processed for analysis. The data processing is done in Python.

The steps taken for transforming the data:

1. Find out all the null values in the dataset

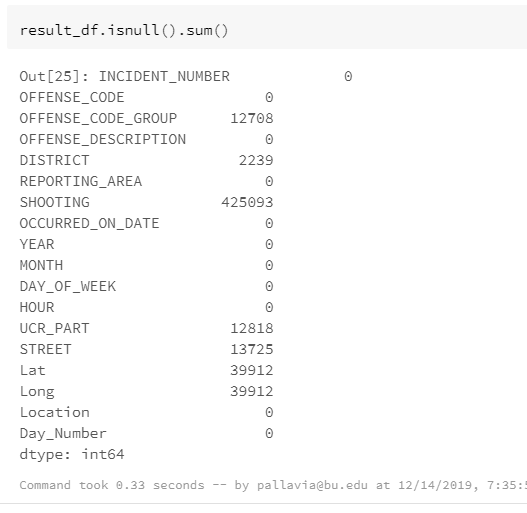


Figure : Data Processing - 1

1. Note in the above snippet, the attribute shooting shows huge number of nulls. On deeper analysis, it was understood, that all the null and 0 in the column meant that there was no shooting involved in the incident and all the ‘Y’ and 1s meant there was shooting. Therefore, instead of dropping the column, the shooting column was transformed into a binary form. 1(there was a shooting); 0 (there was no shooting).

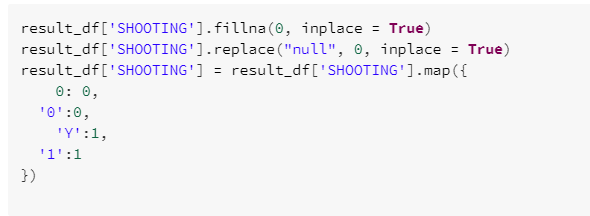


Figure : Data Processing - 2

1. As the remaining null rows were less than 10% of the overall dataset, therefore they were removed from the final dataframe. This reduced the size of the dataset to almost ~400,000 rows.
2. Columns incident number and offense description were dropped from the dataframe as they were not required for the analysis.

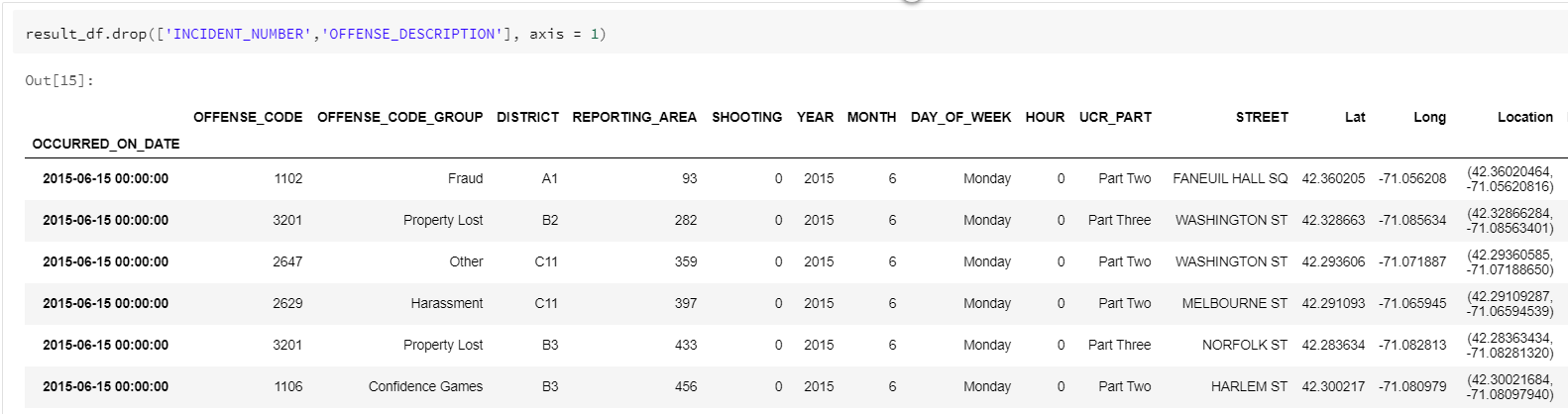


Figure : Data Processing - 3

1. New columns are added for better analysis

Day\_Number – The day of the month(Date) is extracted from the OCCURRED\_ON\_DATE column

District\_Name – Name of the district on the basis of the district\_code is added

(Source: <https://bpdnews.com/districts>)

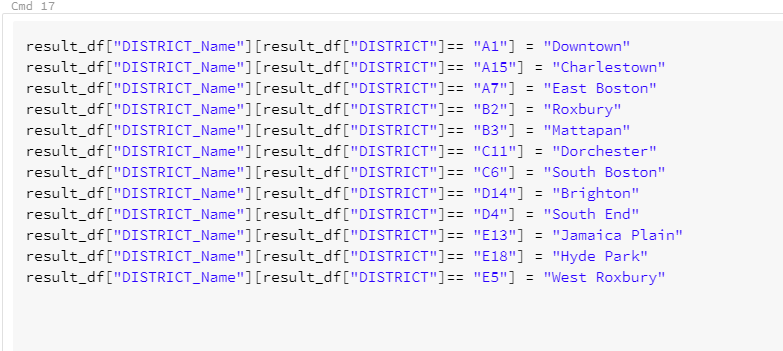


Figure : Data Processing - 4

1. Duplicate rows (714) are removed.



Figure - Data Processing - 5

1. Final shape of the dataframe after data processing: 396389 x 18

# Analysis

The transformed Pandas dataframe is converted to spark dataframe which is converted to SQL Table.

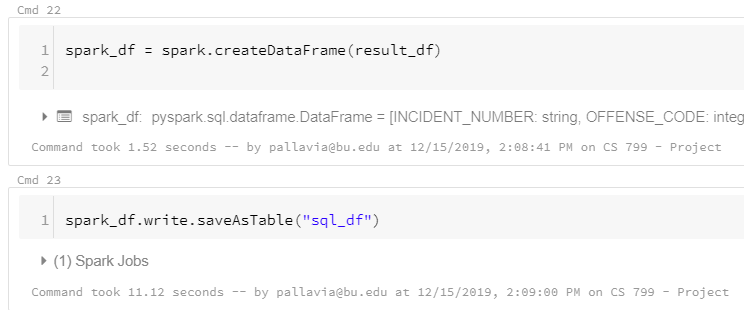


Figure : Dataset conversion from Pandas to Spark to SQL

## Analyse Trends Over the years

The crimes are analysed over the years, if they have changed in any way.

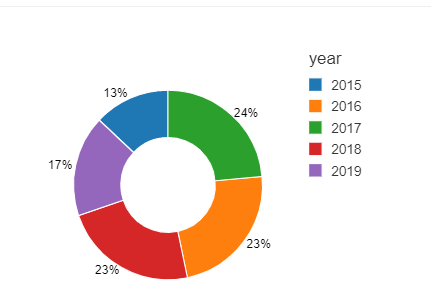
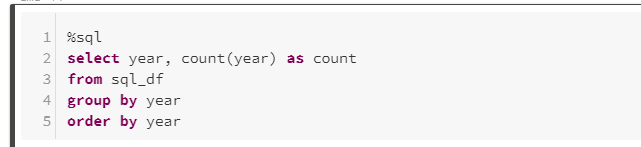


Figure : Crime analysis

Note: Please note that the dataset only has dataset from Dune-December for 2015.

Timeseries plot for the dataset for each month for each year

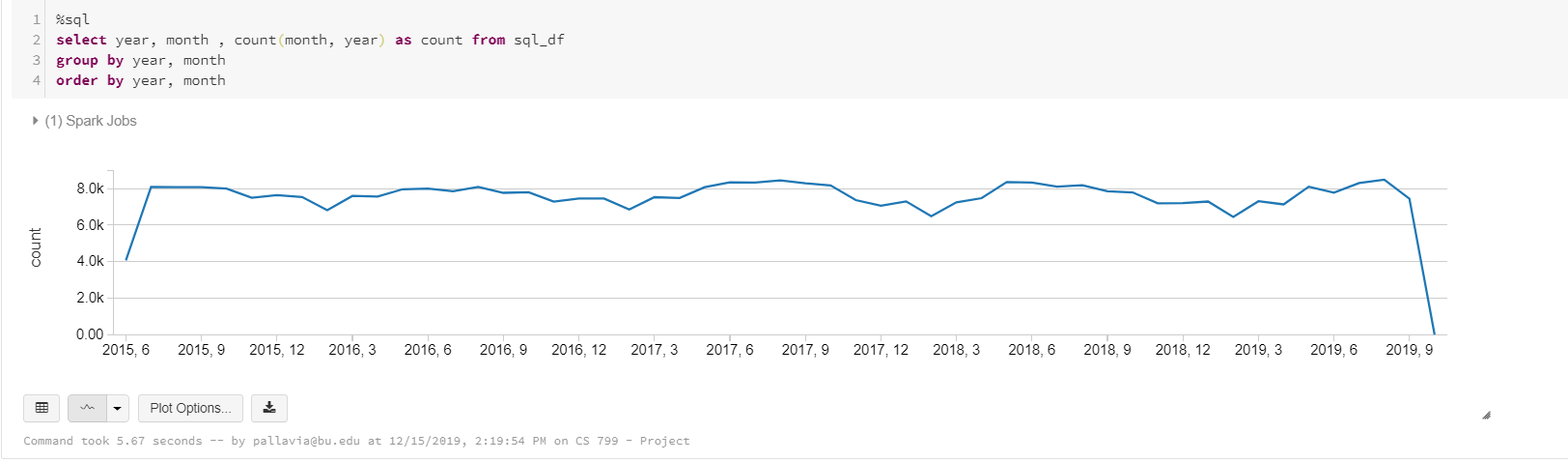
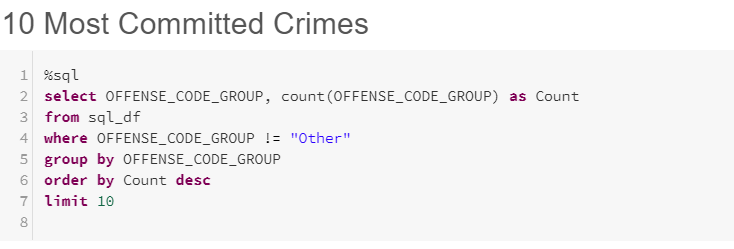


Figure : Time series graph

From the overall analysis across years, we notice there was an increase in crime 2017.

We can also see from the time series data that each year crime rates drops in January, February and March. This might be because the temperature in Boston drops really low therefore less cars will be on the road hence less accidents. Less reports of larceny are also reported as most of the people stay home making it difficult to break in.

## What is the most type of crime?



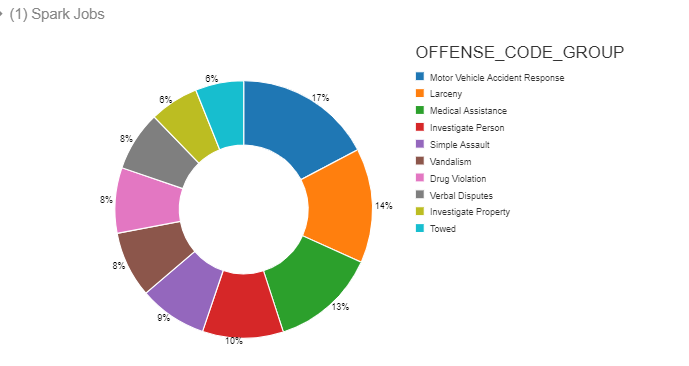


Figure : Most Committed Crimes

The above graph only displays the top 10 crimes reported in Boston. Vehicle accidents is the most reported crime in Boston Area followed by Larceny.

On a deeper analysis of the top 4 crimes, we can see each crime trends in almost same manner. A increase in 2017 is seen in all the crimes, Medical Assistance and Investigate Person shows a steeper increase.

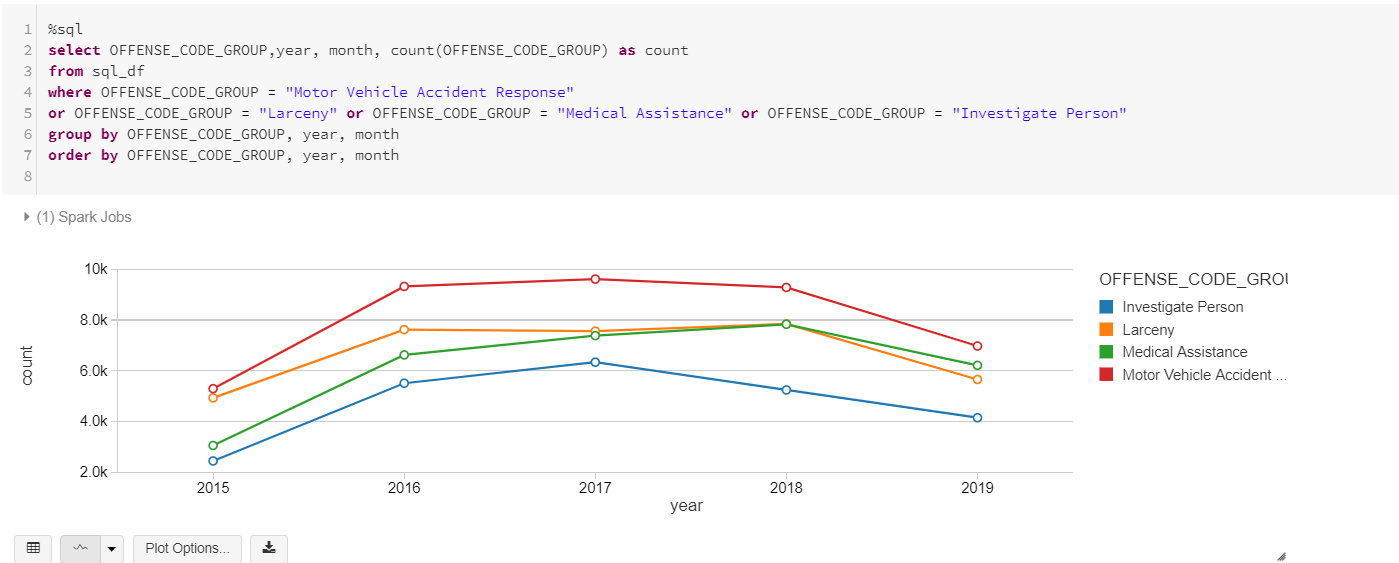


Figure : Crime type analysis

### Motor Vehicle Accident Response Analysis

A view is created to study Vehicle accident closely and increase the performance.

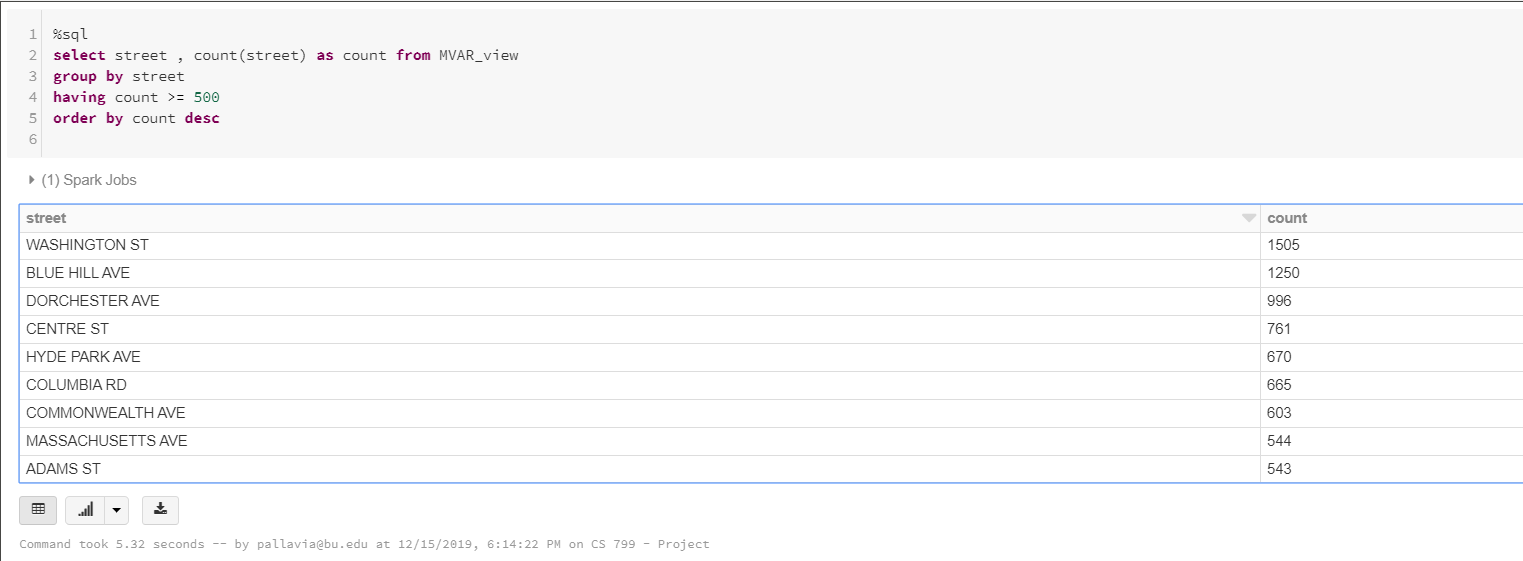


Figure : Analyse Motor Vehicle Accident

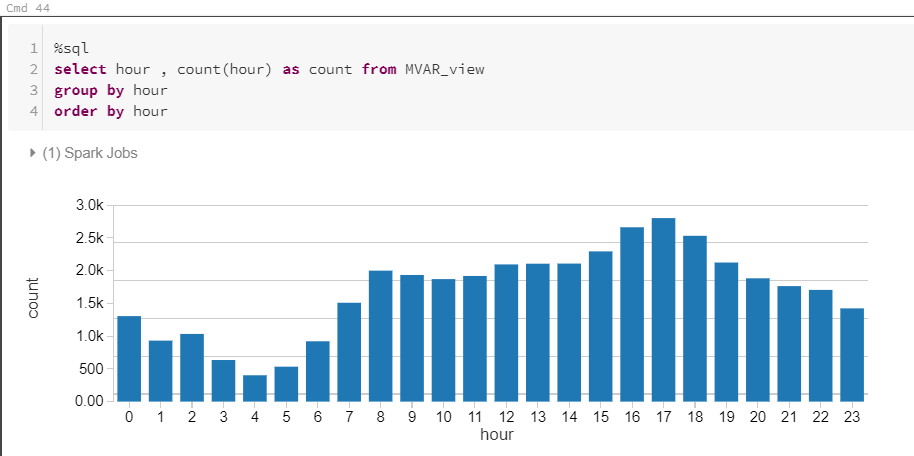


Figure : Analyse Motor Vehicle Accident for each hour

Most of the Vehicle accidents happen on Washington St in between 3 pm - 6 pm.

## Where do crime happen?

We want to study which areas are safe and which are unsafe based on the number crimes reported in these districts in the last 5 years. Boston is divided between 11 districts broadly.

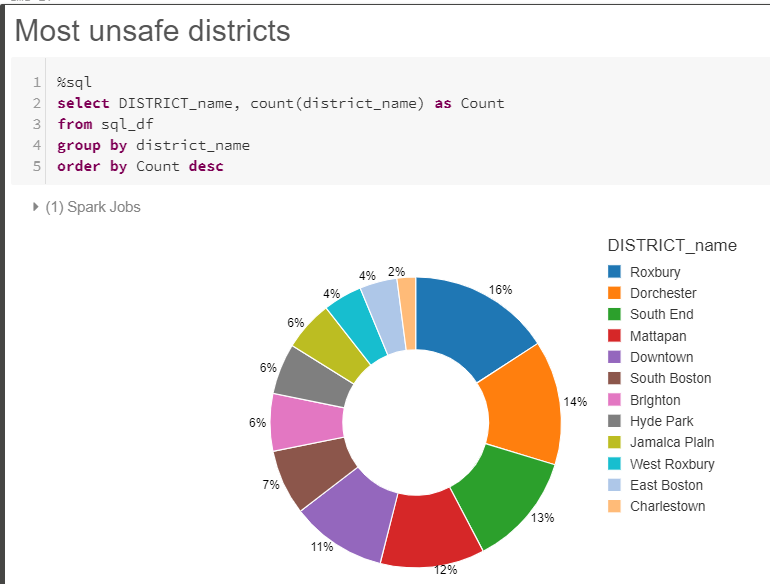


Figure : Most unsafe districts

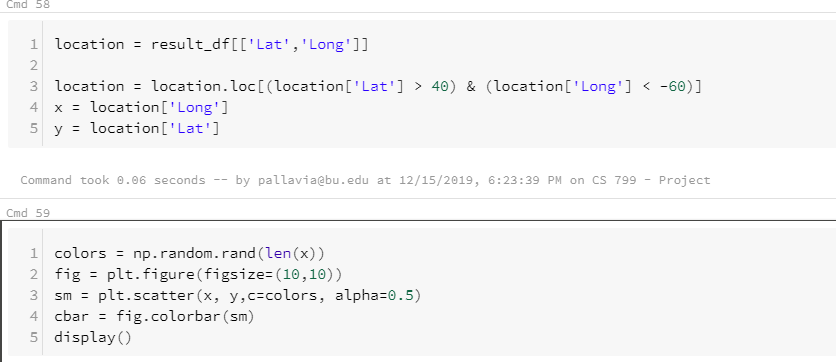
We can see Roxbury is reported to be the most unsafe place. 16% of all the crimes happen in Roxbury followed by Dorchester. Charlestown is the safest place to live with only 2% of overall crimes happening there. You know where to shift!!

Crimes were analysed for each street as well, Washington Street being the most unsafe. If you will notice here the top 10 streets are long streets which cover large amount of area, Washington St being the longest street of Boston. Therefore, I believe the plot below shows this. Longer the street, more people more crime.



Figure : Most unsafe streets

I have plotted all the crimes according to their location.



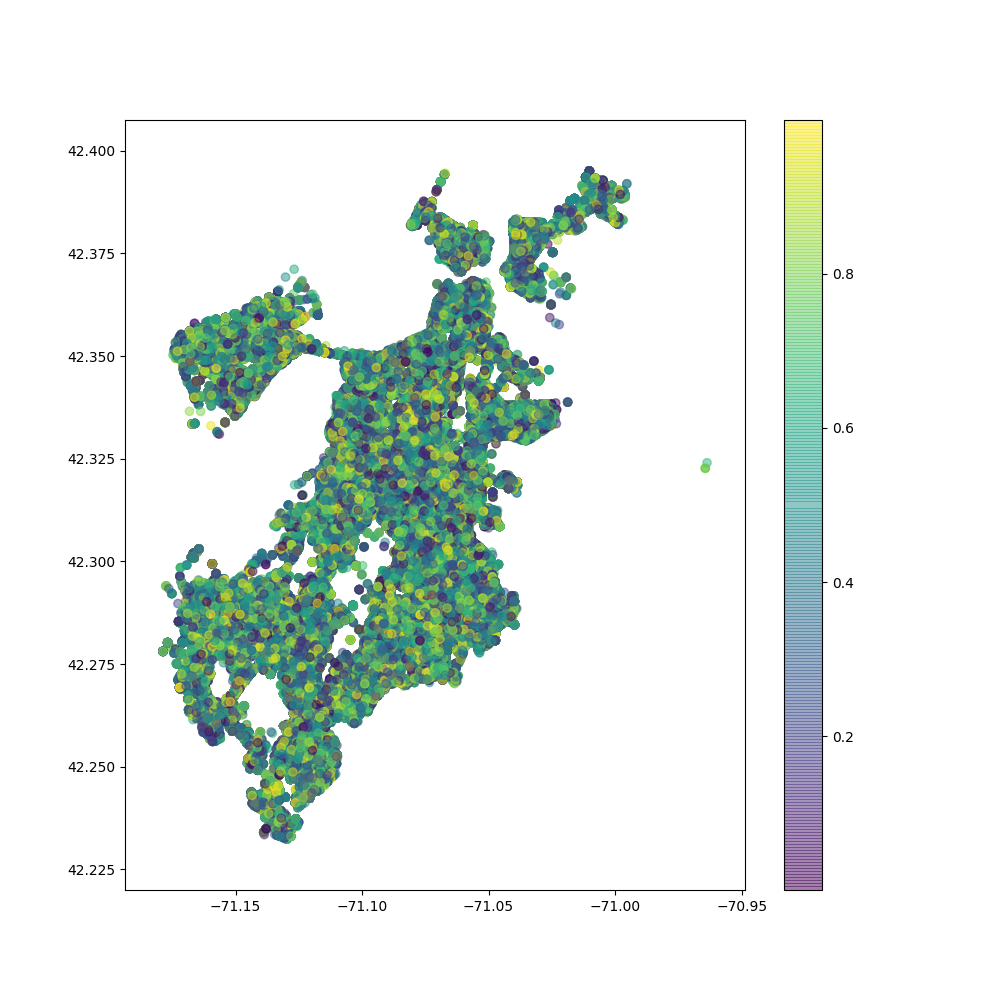


Figure : Crime report plot

### Roxbury Analysis

We will study Roxbury in depth to get a better understanding of the area. I have created a view to store all the data only about Roxbury. This will help me increase the efficiency.

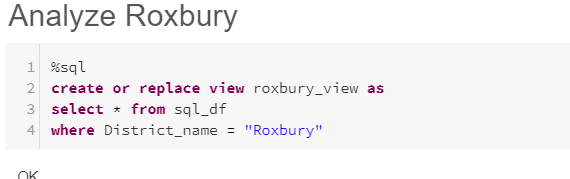


Figure : Roxbury View

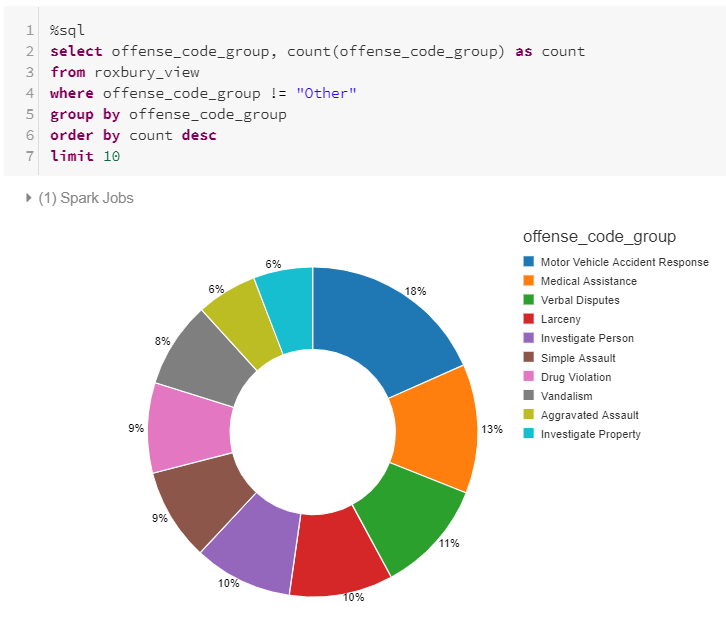


Figure : Crime in Roxbury

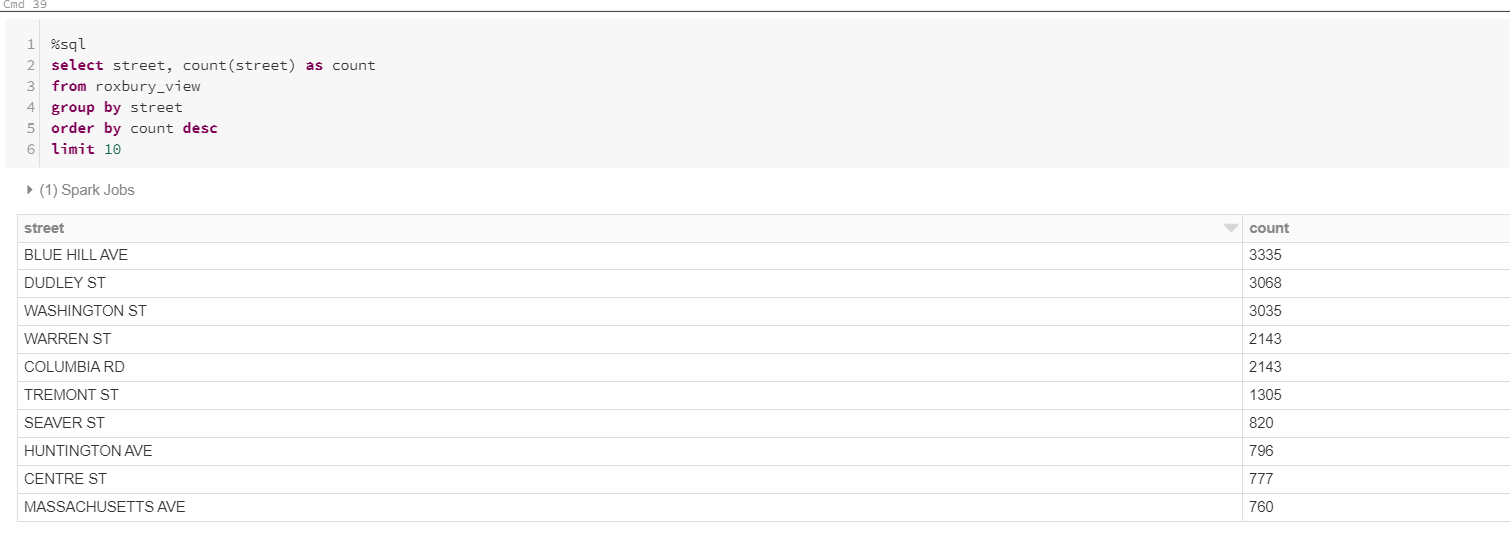


Figure : Unsafe streets in Roxbury

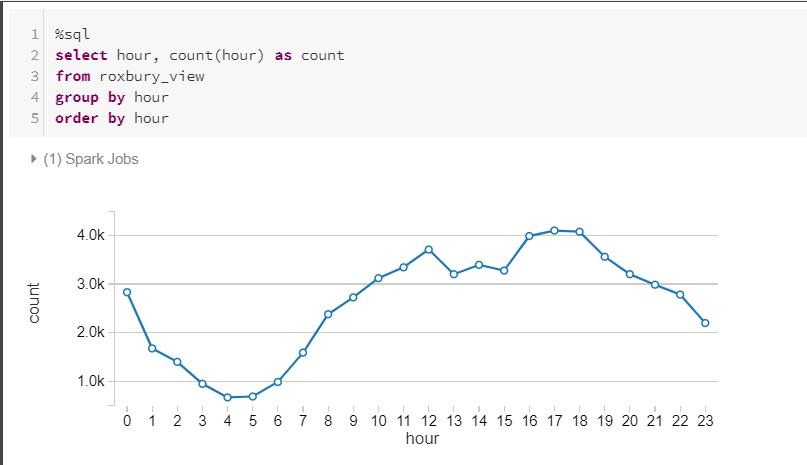


Figure : Time analysis of Roxbury

The most common crime is Motor Vehicle Accident for Roxbury. Most of the crimes happen on Blue Ave hill, Dudley St and Washington St in Roxbury. Most of the crimes happen in afternoon and evening.

## When does crime happen?

Once we know where crime happens, it is important to know when does most of the crime happen.

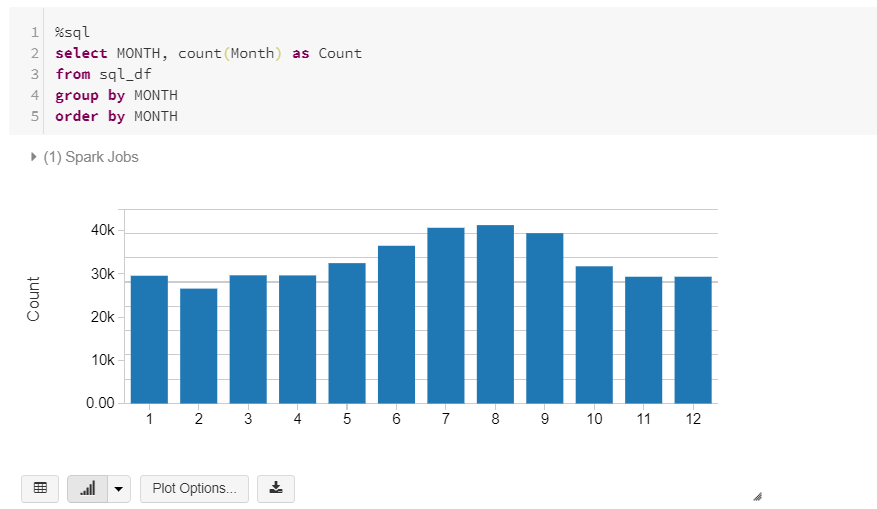


Figure : Monthly analysis of Crime of Boston

We can see crime rate starts to increase in May and continues to increase till September. During the summer months the crime rates are higher as more people on the road hence more accidents can take place. We can see in months October – December the crime rate is not low even though its Fall and comparatively colder outside. Due to festivities, a lot of people are on vacations and their homes are empty making them an easy target.

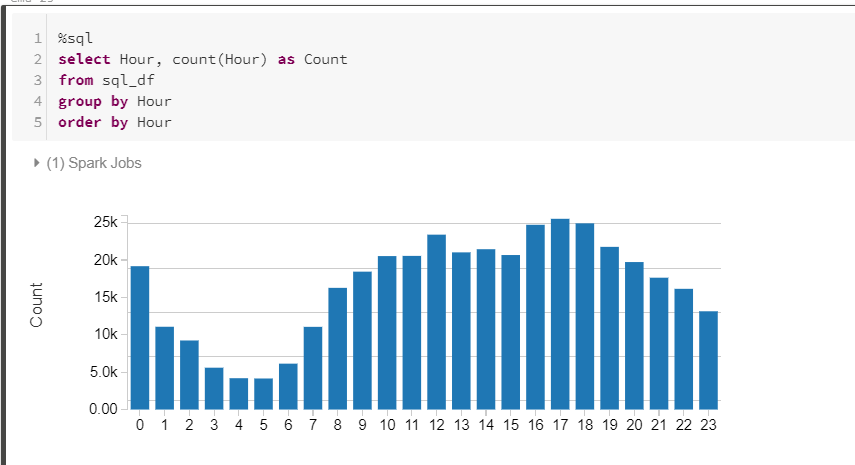


Figure : Hourly analysis of crime in Boston

Criminal activities usually take place between 4 pm – 7 pm. There is a sudden increase in reports at 12 pm.

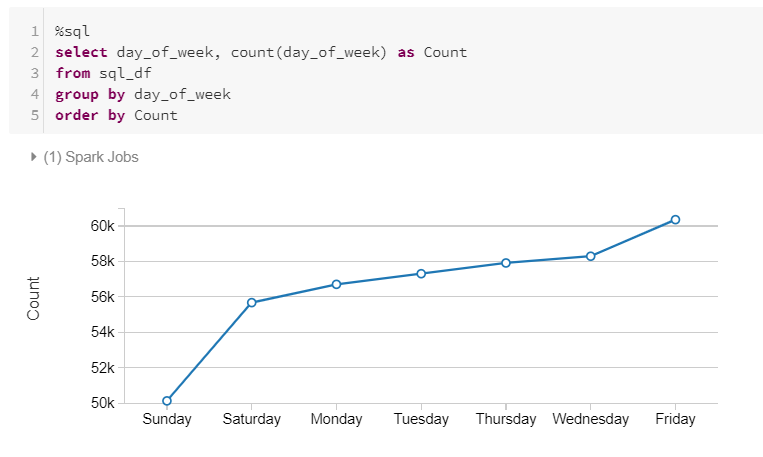


Figure : Weekday analysis of Crime in Boston

There is more crime on the weekdays as compared to the weekends.

## Analyse the shooting incidents

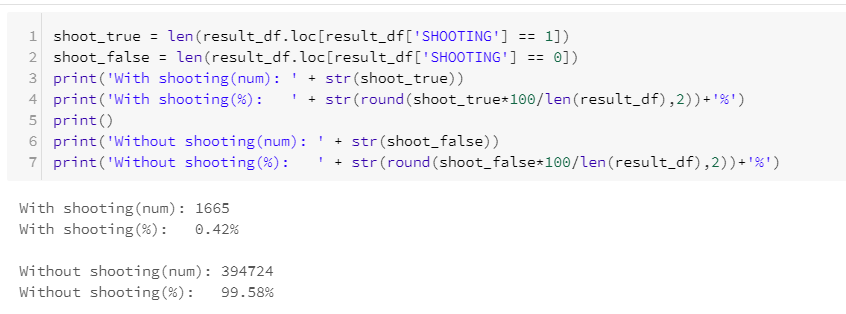


Figure : Shooting analysis

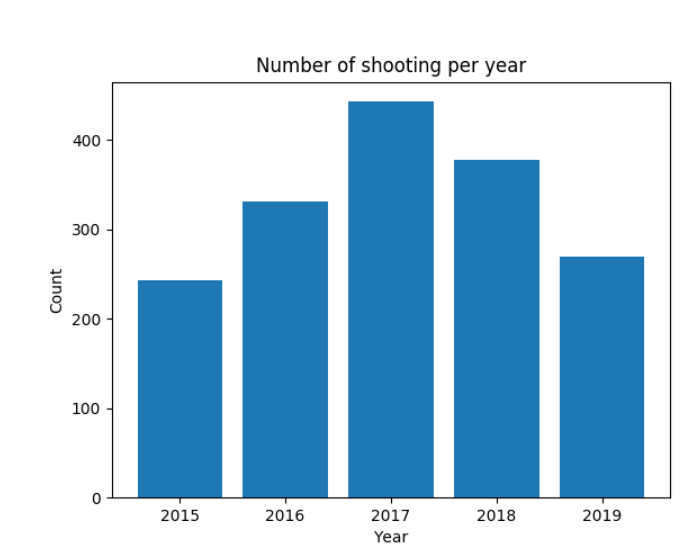
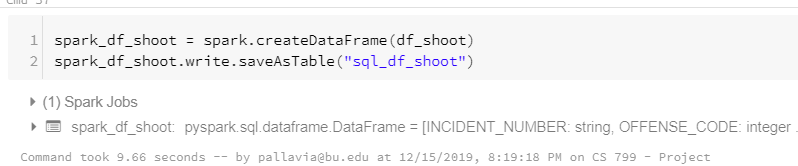


Figure : Shooting across years

There are 1665 reports of shooting incidents. There was a spike in shooting incidents in 2017

For further analysis the Pandas dataframe is converted to SQL table



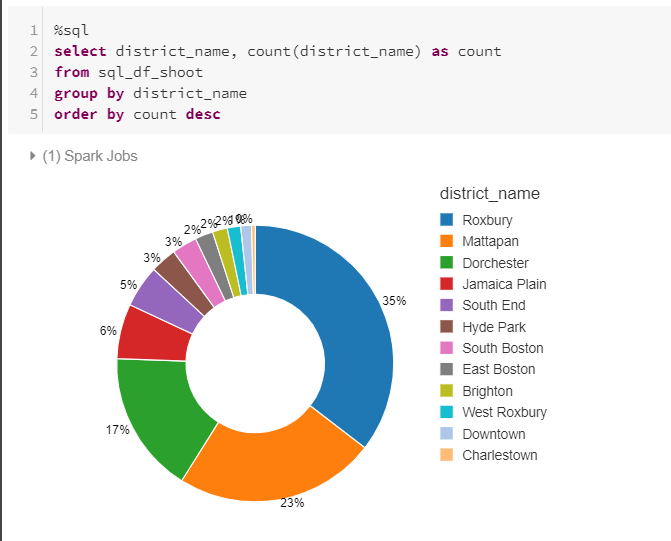


Figure : Shooting in each district

75% of the overall shooting happen in Roxbury, Mattapan and Dorchester.

Now that we have analysed the data in detail, we will try to predict the type of crime and the area in which the crime is to happen.

# Predictive Models (Python)

## Model 1 - Predicting Type of Crime

To reduce the size of dataset, data of top 5 crimes is used for this model.

This model will predict the type of crime,

* Motor Vehicle Accident Response
* Larceny
* Medical Assistance
* Investigate Person
* Drug Violation

The independent variable, or the variables that will predict the crime type will be:

* District
* Reporting Area
* Year
* Month
* Date
* Hour
* UCR Part
* Street
* Latitude
* Longitude
* Weekday

Classifiers used: Random Forest and Logistic Regression

1. Packages used for both the models

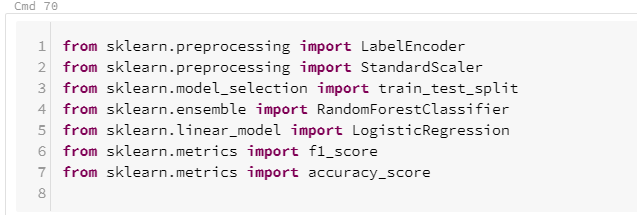
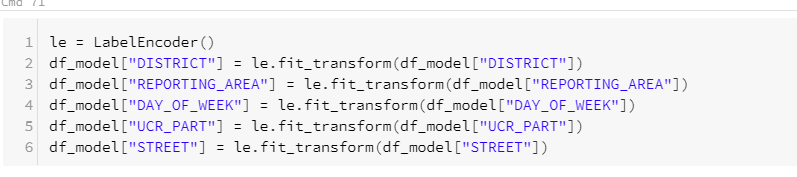
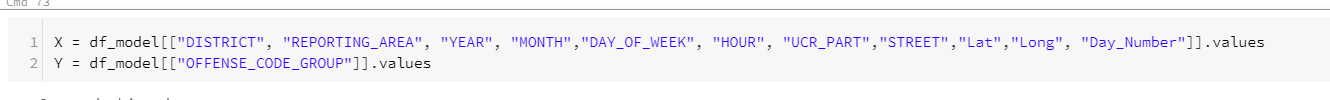


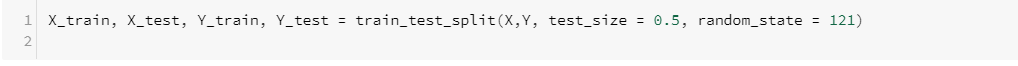
Figure : Machine Learning Packages

1. Transforming the data to make it suitable for classifier

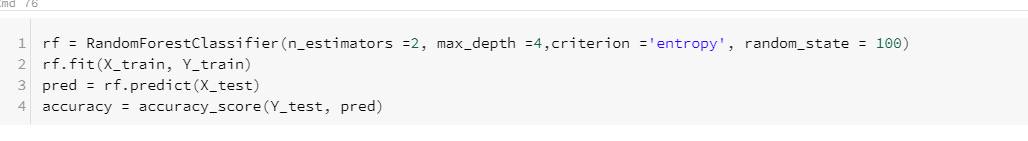




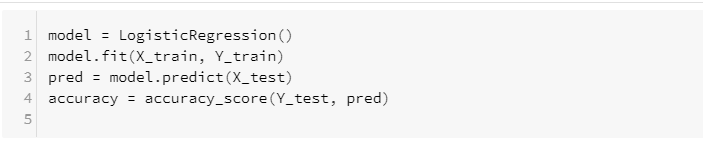
1. Splitting the data into test and train



1. Calculating accuracy using Random Forest Classifier



1. Calculating accuracy using Logistic Regression for comparison



|  |  |
| --- | --- |
| Classifier | Accuracy |
| Random Forest | 63.28% |
| Logistic Regression | 63.55% |

We can note that Logistic regression has a slightly higher accuracy that Random Forest. This means that given all the attributes, the model can predict the type of crime accurately 65% of the times.

### Model 2 – Predicting the district where crime will happen

To reduce the size of dataset, data of top 5 most unsafe districts are used for this model.

This model will predict the district,

* Roxbury
* Dorchester
* South End
* Mattapan
* Downtown

The independent variable, or the variables that will predict the crime type will be:

* Street
* Reporting Area
* Year
* Month
* Date
* Hour
* UCR Part
* Offense Code
* Latitude
* Longitude
* Weekday

Classifiers used: Random Forest and Logistic Regression

NOTE: Since the code is same, I’m not putting the screenshots.

|  |  |
| --- | --- |
| Classifier | Accuracy |
| Random Forest | 85.52% |
| Logistic Regression | 65.34% |

This means, 85% of the times the district is predicted accurately if all the other factors are fed in.

Conclusion

I hope this analysis will help you understand the criminal activities in Boston a little better. The objective of the analysis was to learn the varied functionalities of Apache Spark. The project included different functionalities and visualization in SQL and Python. I really enjoyed using Python and SQL fluidity of the software, I could use any coding language according to my comfort.

Although I’m using Community Edition df Databricks for this project and could only get my hands on a single cluster the processing of large dataset was quick.

# Reference

* Apache Spark Lightning-fast unified analytics engine. Retrieved on 2019, December 1 from <http://spark.apache.org/>
* Apache Spark. Retrieved on 2019, December 1 from <https://databricks.com/spark/about>
* Big Data Analytics. Retrieved on 2019, December 2 from <https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>
* Comparing Apache Spark and Databricks. Retrieved on 2019, December 1 from <https://databricks.com/spark/comparing-databricks-to-apache-spark>
* Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
* Real Estate and demographic data. Retrieved on 2019, December 1 from <https://www.neighborhoodscout.com/ma/boston/crime>
* Sicular S. (2013, May 27). Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. Retrieved from <https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#6998d96f42f6>
* Spark 101: What Is It, What It Does, and Why It Matter. Retrieved on 2019, December 1 from <https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/>
* The Four V’s of Big Data. Retrieved on 2019, December 1 from <https://www.ibmbigdatahub.com/infographic/four-vs-big-data>
* What is Big Data? Retrieved on 2019, December 2 from <https://www.oracle.com/big-data/guide/what-is-big-data.html>