BDAD Project Proposal

Crime in NYC(How Safe NYC really is?)

Team members

Yashasvi Choukikar	yc4953
Lakshmi Sai Anmol Ramayanam	lr2947
Sanjeet Vijaya Shetty	ss14431

Project Synopsis:

In this project we intend to focus on the crimes committed in New York City. By analyzing data from complaints registered with the NYPD, arrests made by NYPD, distressed 911 calls described with various subcategories, we aim to find valuable insights of crime statistics across 5 boroughs of New York City. Our findings we believe shall help diagnosing and prescribing solutions such as delays in 911 dispatch help, prevention of certain types of crime in key areas using preemptive measures as well as demographics vis-a-vis crime-type to identify better programmes to alleviate the issues. By discovering the locations of most dangerous crimes , the timing stats can help in establishing better patrolling and prevention policies.

Keywords:Big Data, Apache Spark, Scala, Tableau, Spark SQL, MS Excel

Data Sources:

All of the below datasets are publicly available , cited from official NYC data sites. It is historical data, collected over a period of time.

1. **NYC 911 calls**: (worked on by : Yashasvi Choukikar)
   https://data.cityofnewyork.us/Public-Safety/NYPD-Calls-for-Service-Historic-/d6zx-ckhd
   The NYPD is the owner of the dataset below. To communicate with callers and the NYPD, phone takers and dispatchers use this information. A system entry is represented by each record. Both entries created by the general public and ones started by NYPD employees are included in the data. The information can be used for problems that the NYPD is addressing. It was released to the public on 5/3/2021 and contains historical data.
   File size: 6.8 GB

2. **NYPD Complaints**:(worked on by :Sanjeet Vijaya Shetty)
   The NYPD also owns the dataset that follows. This dataset comprises every legitimate felony, misdemeanor, and violation crime that the New York City Police Department (NYPD) received reports of from 2006 through the end of the previous year. It was made public on November 16, 2016, and it is updated yearly.
   https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i
   File Size: 2.51 GB

3. **NYPD Arrests**:(worked on by : Lakshmi Sai Anmol Ramayanam)
   A list of each arrest made in New York City from the beginning of 2006 till the conclusion of the preceding year. It includes a breakdown of each arrest that the NYPD made in NYC from 2006 through the end of the previous calendar year. Every three months, the Office of Management Analysis and Planning manually extracts this data, reviews it, and then posts it on the NYPD website.
   https://catalog.data.gov/dataset/nypd-arrests-data-historic
   File size : 1.13 GB

4. **NYC hate crimes**:(worked on by : Yashasvi Choukikar)
   The NYPD owns the dataset that follows. Dataset of confirmed hate crimes in New York City from the years 2019-2022. This dataset comprises where the hate crime is committed and segmented into various categories of crimes like felony, aggravated assault and so on. The data is updated manually every quarter and the statistics and information is made publicly available.

https://data.cityofnewyork.us/Public-Safety/NYPD-Hate-Crimes/bqiq-cu78
File size : 313KB

5. **NYC 2010-2020 Census Data**:(worked on by : Sanjeet Vijaya Shetty)
   The NYC government owns the dataset that follows. Dataset does a comparison in the census statistics between the years 2010 and 2020. This dataset comprises different information about the basic housing and demographic information for different NYC neighborhoods. The updation of census data happens every 10 years.
   https://www.nyc.gov/site/planning/https://catalog.data.gov/dataset/new-york-city-population-by-borough-1950-2040planning-level/nyc-population/2020-census.page
   File size: 2 MB

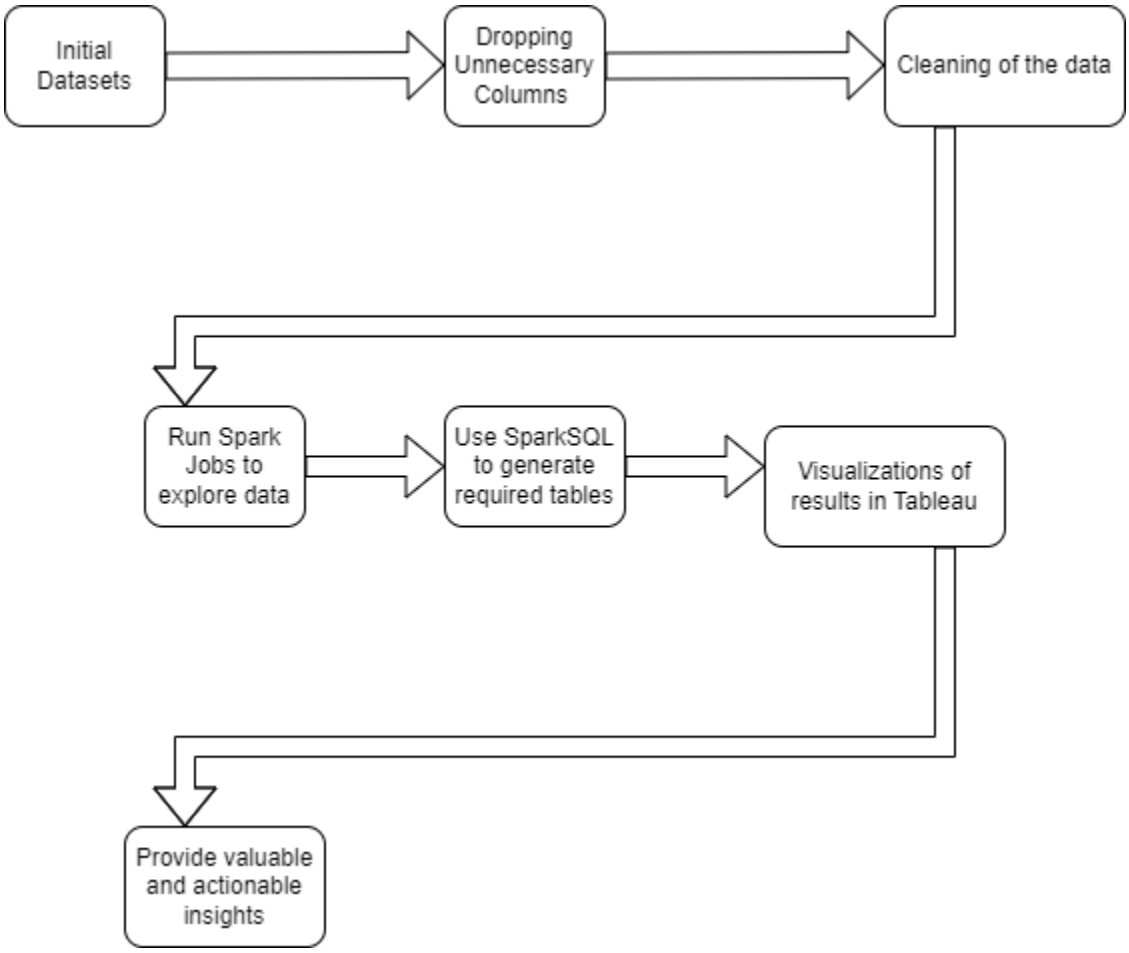
6. **NYC population by Borough**:(worked on by :Lakshmi Sai Anmol Ramayanam)
   Summary table of New York City population numbers and percentage shares by Borough, including school-age (5 to 17), 65 and Over, and total population for over the years 1950-2040. The data is updated and maintained by the NYC government and is made publically available.
   https://catalog.data.gov/dataset/new-york-city-population-by-borough-1950-2040
   File size: 2KB

Technology:

By integrating Scala with Hadoop, we can store data for longer periods of time and use APIs to search against it. The default memory allocation and maxResultSize setting need to be increased when working with a large dataset. It is possible that jobs occasionally fail simply because the default timeout limits are too low, making it impossible to save a huge data frame to the database or some bucket. Several workloads, including interactive queries, real-time analytics, machine learning, and graph processing, can be handled on Apache Spark. Multiple workloads can be combined smoothly by a single application.

We would use Scala APIs to process the datasets in order to keep the information that is important while removing the columns that don't. We would also check the acquired data's integrity. We intend to use Tableau for data visualization once we have completed this and obtained our datasets. Finally, based on the results, we will be able to perform our prescriptive analytics.

Design Diagram



As we are still in the planning stage, based on the datasets profiling and the insights the flow could vary.

References:

1.https://medium.com/@subpath/managing-huge-datasets-with-scala-spark-9840ad760424

2.https://www.analyticsvidhya.com/blog/2022/02/comparing-r-and-tableau-for-data-visualisation/