Dataset Used : 2002 Crime Data from Chicago Data Portal
Link of Reference :
https://data.cityofchicago.org/Public-Safety/Crimes-2002/g9qy-h66j/about_data

Original Dataset Schema :

root
|-- CaseNumber: string (nullable = true)
|-- Date: string (nullable = true)
|-- Block: string (nullable = true)
|-- IUCR: string (nullable = true)
|-- PrimaryType: string (nullable = true)
|-- Description: string (nullable = true)
|-- LocationDescription: string (nullable = true)
|-- Arrest: string (nullable = true)
|-- Domestic: string (nullable = true)
|-- Beat: string (nullable = true)
|-- Ward: string (nullable = true)
|-- FBICode: string (nullable = true)
|-- X_Coorddinate: string (nullable = true)
|-- Y_Coorddinate: string (nullable = true)
|-- Year: string (nullable = true)
|-- Latitude: string (nullable = true)
|-- Longitude: string (nullable = true)
|-- Location: string (nullable = true)
|-- _rescued_data: string (nullable = true)

Data Ingestion :

Data has been uploaded on a personally hosted amazon S3 bucket, which has been accessed using the security credentials from amazon aws console and are being read in the notebook through AutoLoader using the following code :

Storage in the Bronze Layer :

Created a bronze database to which the data frame is being written as a Delta Table with the name bronze_table. This table contains the raw data which is being ingested from the s3 bucket.

Transformations for Silver Layer :

Created a silver database to which a transformed dataframe with the following transformations :

1. Check Primary Crime type for null values
2. Check Date and Location for Null Values
3. Convert the Longitude and Latitude into double from string.
4. Convert the X_Coordinates and Y_Coordinates into double from String
5. Remove the rows where Date and Location are Null.

After the transformations , the data has been written using a streaming table into the silver database for permanent storage and about 11000 rows have been removed due to null values.

Gold Layer :

The Following transformations have been performed to make sure that the data fits the machine learning and dashboard use case :

1. A Timestamp column has been added converting the Date from String to Timestamp Object
2. Watermark has been Added to each and every transformation dataframe for writing in storage.
3. Crime Type dataframe has been formed to contain each crime and the number of times they have occurred throughout the year.
4. MonthlyCrime Count data frame has been aggregated to depict the number of crimes in each month which can help in detecting what month the crime rates are highest.
5. ArrestRate Data Frame has been formed to contain the Each Crime Type, their total cases and the number of arrests made which can be used to depict the ratio of crime and it’s arrest which can help is knowing the reason behind low arrest rates.