Faisal Patwary
3/31/2025

Report

Information Architecture

Data

Hate_crimes.xlsx (cleaned data)
Dictionary.xlsx

*** As the FBI states in its 2022 Uniform Crime Report, hate crimes recorded in its system have sufficient evidence that the offender was motivated by bias during the incident. This reduces the uncertainty that comes from subjective motivation. ***

For this project, I selected one dataset from the CDE website. The dataset I have chosen contains over 200,000 records that cover hate crimes recorded by the FBI across all of the US from 1991 to 2023. According to the website, the dataset was recently updated in September 2024 to include 2023 data. If I were to update my database I expect to wait until September 2025. As for the data value types, there is a combination of date values, integers, and texts. Some text fields have many characters that I have to take into consideration when building my dimensional model. The website did not offer a dictionary for my selected dataset so after cleaning my data I created my own.

Business Requirements

Thousands of hate crime incidents occur across the tristate region each year. My goal is to collect and store data from 2020-2023 enabling year-over-year comparisons to show us

- Shifts in incident frequency
- Commonly targeted groups within the Tri-state region

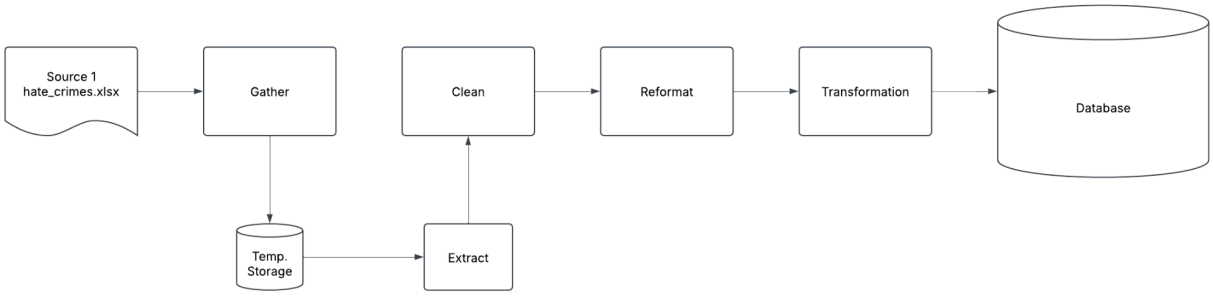
Functional Requirements

This system will allow users to do the following:

- Filter hate crime based on location, date, and types of offenses
- Create year-over-year comparison reports
- Display interactive charts that compare trends by year or location
- Download query results in different formats for external analysis

These functions not only will be useful for users who seek their own analytics but also supports our goals of understanding and visualizing shifts in trends year-over-year as well as answering what are the most commonly targetted groups within the tri-state region.

Model



Description

I have one source of data for this project. I am using hate_crimes.xlsx from the CDE website. Firstly, I downloaded the dataset of interest: hate crimes.

I would save this file and save a copy to Google Cloud's bucket storage in case something goes wrong with the original file.

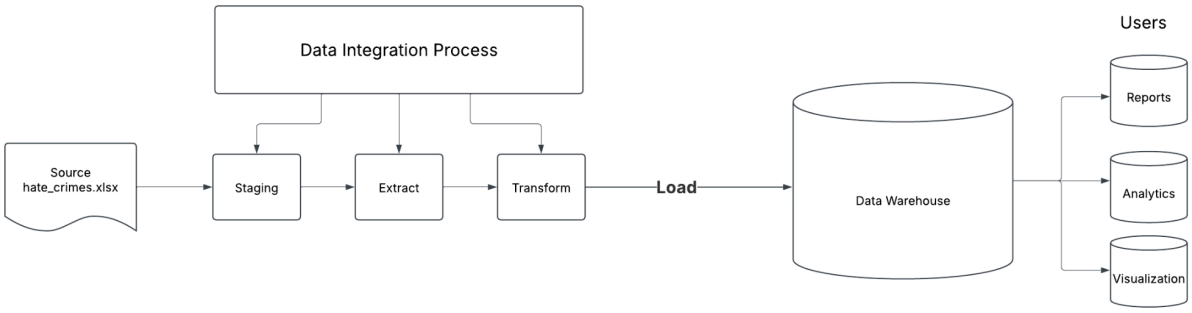
From there I will extract the data and begin by cleaning the data. I do so by reducing the 200,000+ records that contain data across all reporting states from the years 1991 to 2023. After getting the data I want I come down to 8,120 records that cover reportings across the tri-state region (New York, New Jersey, Connecticut) from 2020 to 2023. During this step of the process, I also deleted columns I deemed unnecessary for my purpose.

For my next step, there are values such as the date and counts that must be reformatted accordingly.

Having made the necessary changes I saved this new dataset and uploaded it to my Database.

Data Architecture

Model



Description

Having downloaded the data from the CDE website I was able to transform the data to fit my needs by doing the following: replacing null numeric values with 0, removing columns that did not suit my project, and filtering states and date to tailor to my business goal.

Due to the fact that I have one data source there is no need to merge anything. The data file does not require any ingestion as I am not consistently uploading new data frequently.

After going through the data integration process I load my data into Google Cloud's BigQuery. Users and businesses can pull data for analytical purposes. The data they retrieve can be in the form of aggregated data, summarized reports, and visual models. The users have access to all current (2020 - 2023) data of the tri-state regions.