Final Project
DAV-6100-Information Architecture

Group 3

Yeshiva University | KATZ SCHOOL

NYC Car Crash Data Warehouse Design

Overview

- Our project centers on analyzing and interpreting the Motor Vehicle Collisions data for New York City.
- The DW address critical questions surrounding incident patterns, location, collusion trends over time, and factors contributing to collisions.
- We aim to uncover insights that could bolster traffic management strategies, enhance public safety initiatives, and inform policy decisions.
- The project aims also at facilitating a better understanding of traffic dynamics in NYC, which could lead to improved road safety measures and a reduction in collision rates.
- The anticipated outcome is a robust data-driven framework that provides NYC authorities, policymakers, and the public with the tools necessary to comprehend and improve the multifaceted nature of urban traffic safety.

Roles & Responsibilities

Rekha & Jetendra - Data Engineer:

Integrates and manages data workflows into AWS cloud architecture. Also focuses on efficiency and relability of data collection and processing.

Priyank & Volkan - Data Analyst:

Interprets data patterns, providing insights and recommendations to support strategic decision-making.

Timeline of the Project

Week 4: Bus Matrix, Basic Use Cases

Week 5: Dimensional Model

Week 6-7: Conceptual, Logical and Physical Models

Week 8: Data Dictionary

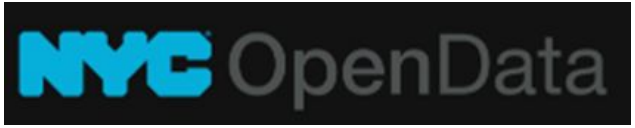
Week 9: Architecture Diagram

Week 10-12: ETL/ ELT

Week 13-14: Analytical Queries and Visualization

About Data

Motor Vehicle Collisions in NYC - records comprehensive details from all police-reported motor vehicle collisions in New York City. It requires the completion of police reports for incidents involving injuries, fatalities, or property damage exceeding $1000. It records various aspects such as the specifics of each vehicle involved, personal details of individuals affected (drivers, passengers, pedestrians, etc.), and the overall crash event characteristics. This data, which dates back to the adoption of the electronic reporting system in April 2016, serves as a pivotal resource for analyzing collision causes, assessing traffic safety policies, and developing preventative measures.



Data Source:
https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4

Summary of Datasets

Entities

Date: Represents the date and time aspects of a crash.

Contribution Factor: Captures the factors contributing to a crash.

Vehicle: Details about the vehicles involved in the crashes.

Person: Information about individuals involved in the crashes.

Crash Event: The central entity that records each crash incident.

Method of access

- Pull data from internet open-source
- MySQL connector and Python (Jupyter Notbook)
- Load into AWS S3 bucket

Data Quality

- The dataset includes fair amount of missing or incomplete information.
- We Identified one dataset is not been mostly aligned we can only 50% of data, so ignored it.
- Problems are detected by EDA and addressed cleaning and validation process

Data Profile

Source: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4

Number of Tables: 3

Number of Records: 10,990,900 rows in total, number of columns ranges from 25 to 29

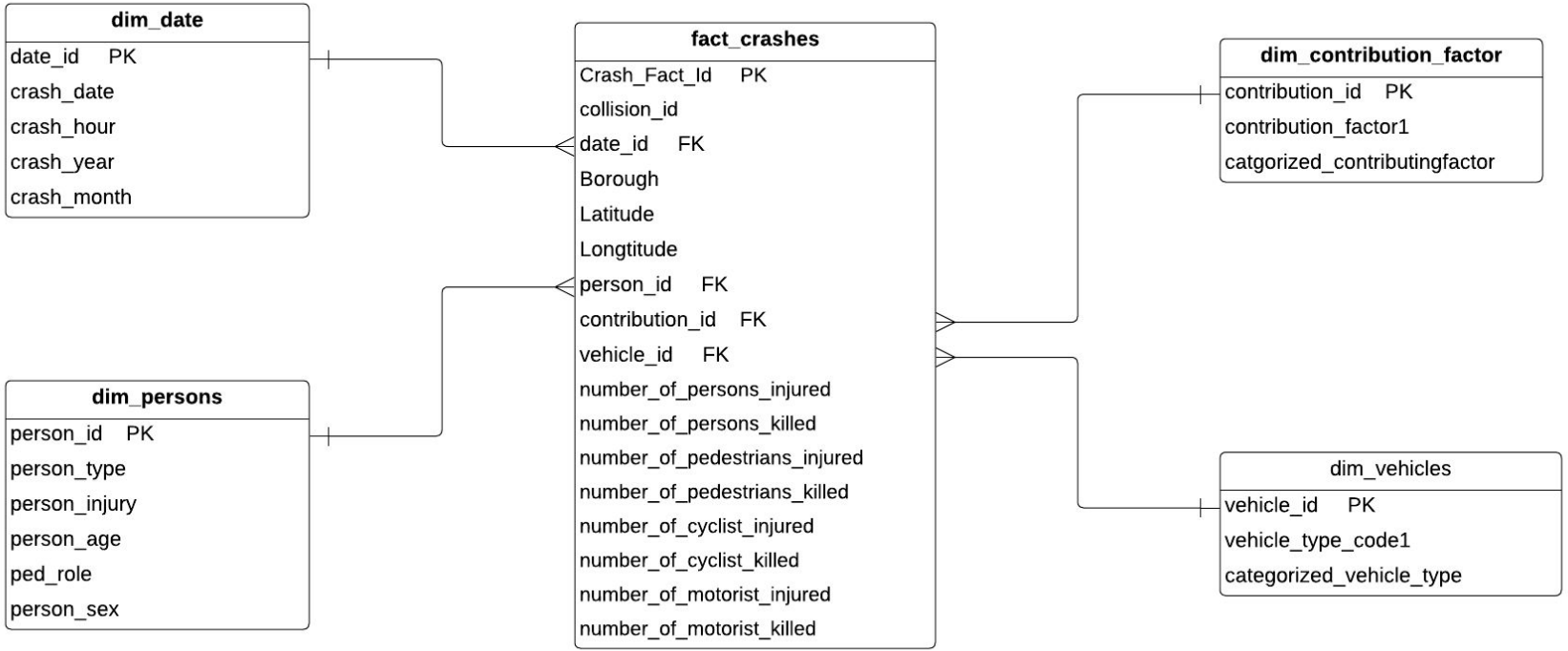
Data Type: Numeric, Text, Temporal, Geospatial

Data Acquisition Method: MySQL Connector through Python

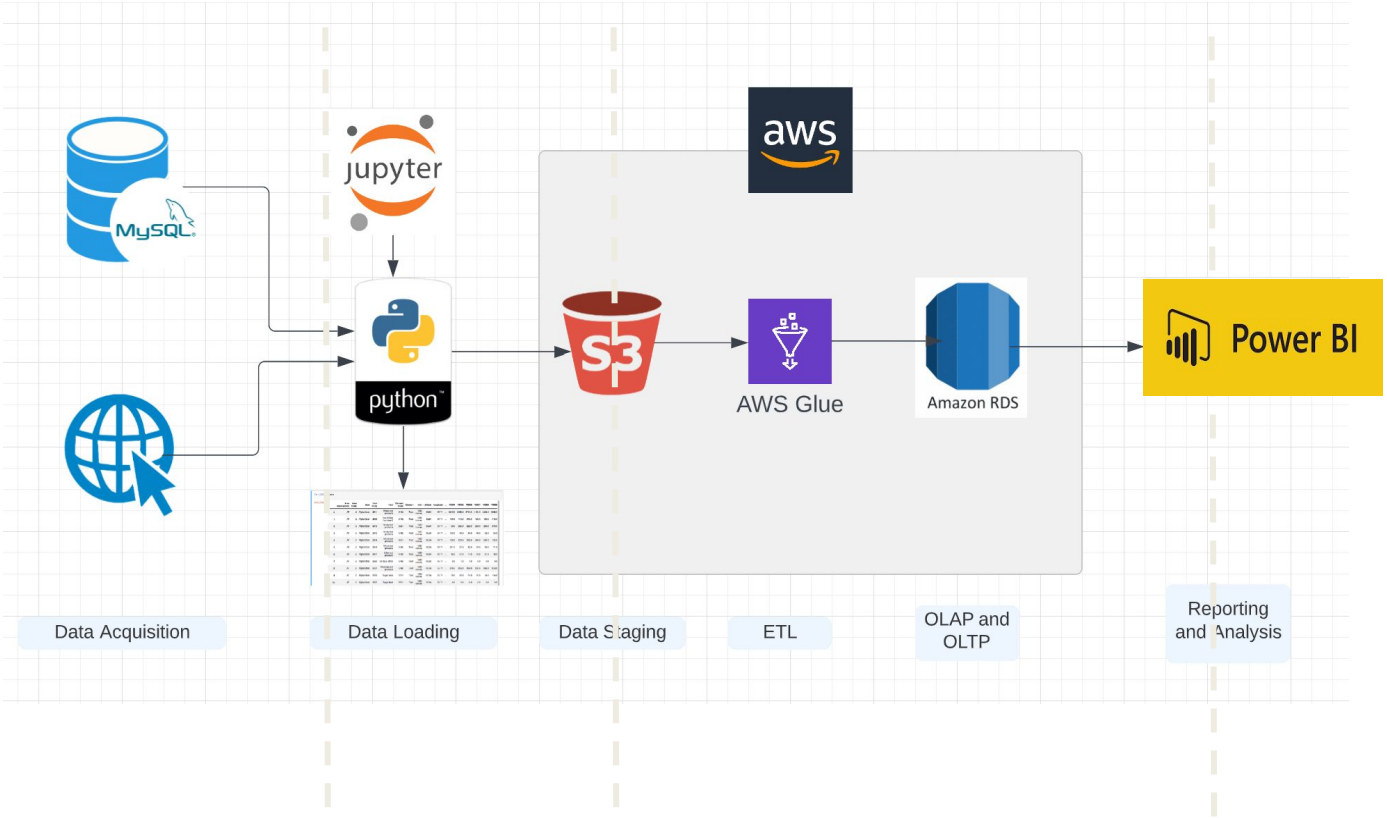
Bus Matrix

Facts	Date	Person	Contributing Factor	Vehicle	Location
Crash	X	X	X	X	X
Injury & Death	X	X		X	

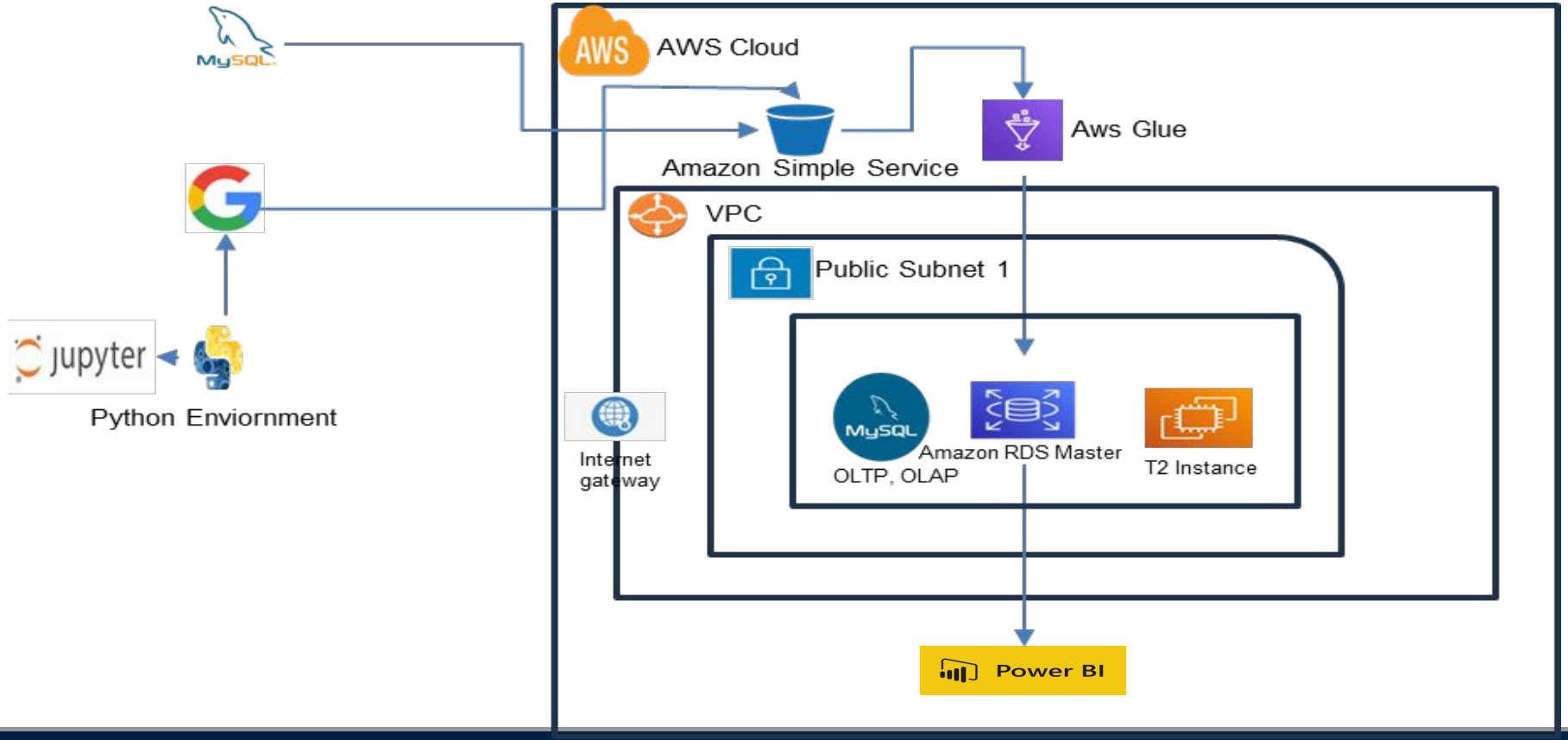
Dimensional Modal



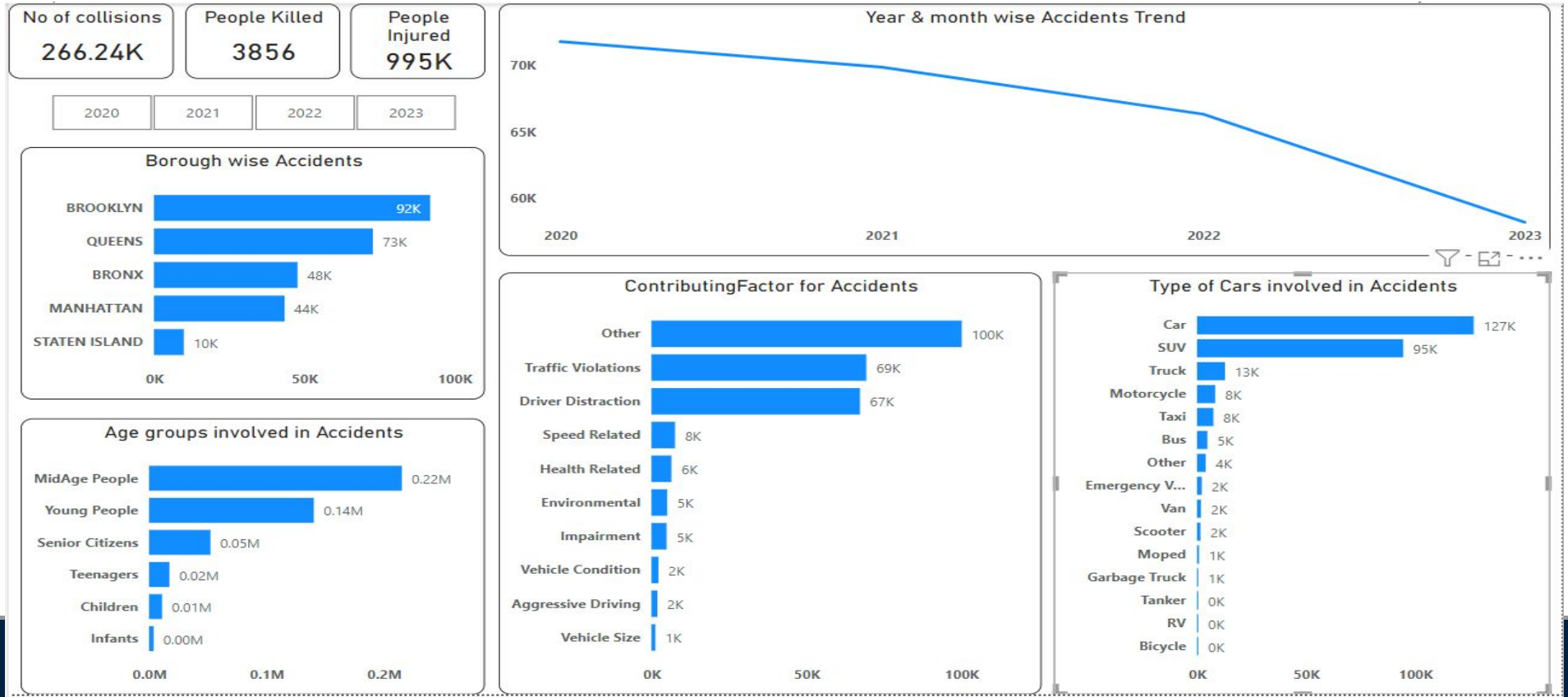
Work Flow


MySQL
jupyter
python
aws
S3
AWS Glue
Amazon RDS
Power BI
Data Acquisition
Data Loading
Data Staging
ETL
OLAP and OLTP
Reporting and Analysis

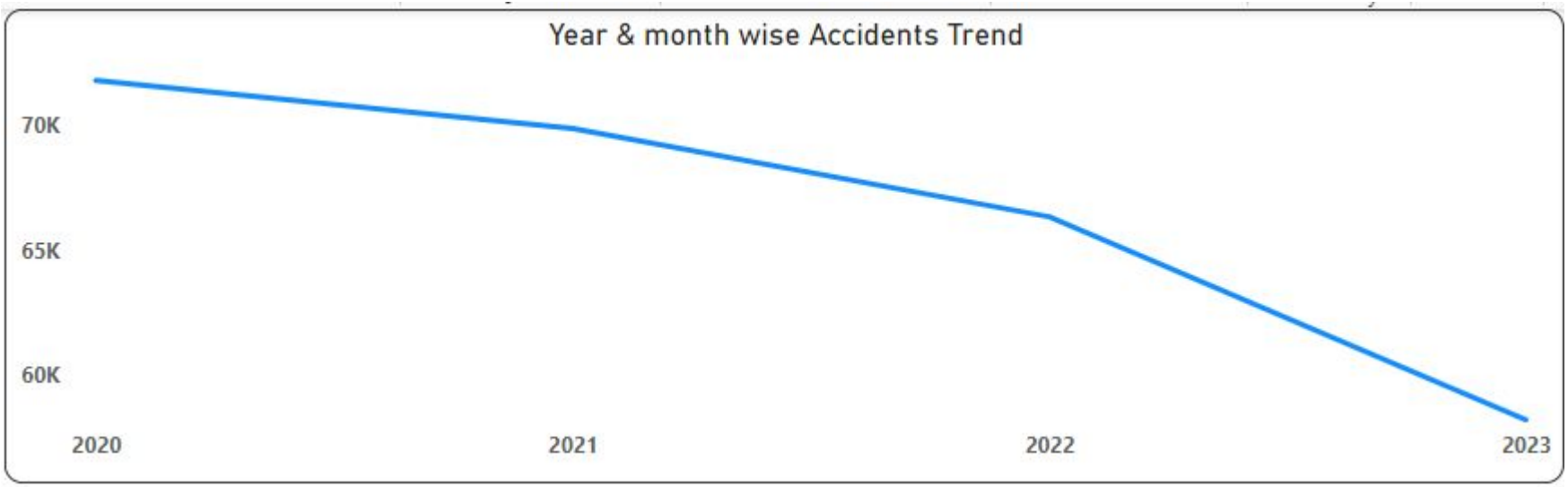
AWS Architecture



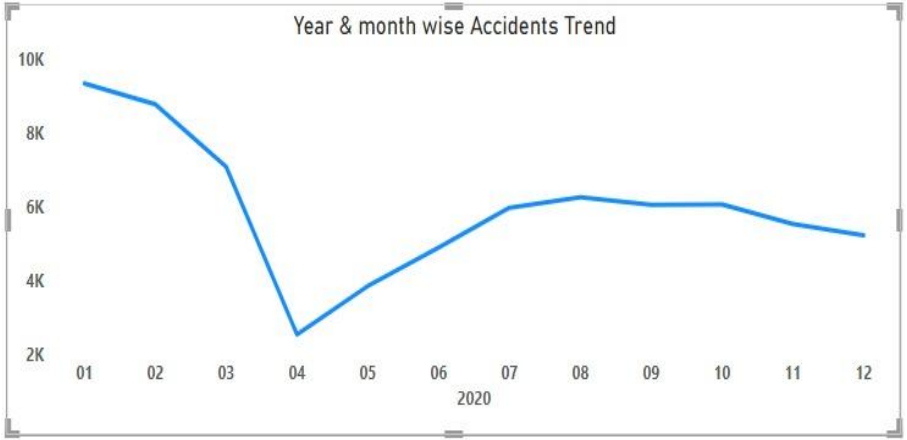
Dashboard

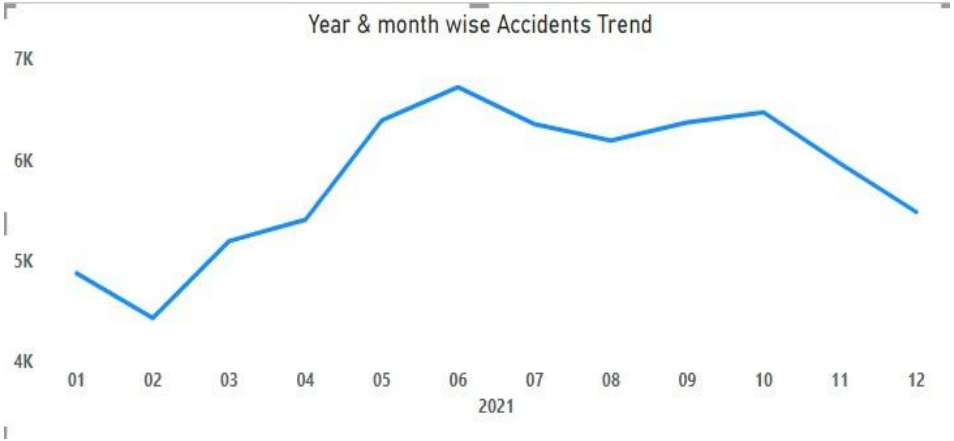

No of collisions
266.24K
People Killed
3856
People Injured
995K
2020
2021
2022
2023
Borough wise Accidents
BROOKLYN
92K
QUEENS
73K
BRONX
48K
MANHATTAN
44K
STATEN ISLAND
10K
0K
50K
100K
Age groups involved in Accidents
MidAge People
0.22M
Young People
0.14M
Senior Citizens
0.05M
Teenagers
0.02M
Children
0.01M
Infants
0.00M
0.0M
0.1M
0.2M
Year & month wise Accidents Trend
70K
65K
60K
2020
2021
2022
2023
ContributingFactor for Accidents
Other
100K
Traffic Violations
69K
Driver Distraction
67K
Speed Related
8K
Health Related
6K
Environmental
5K
Impairment
5K
Vehicle Condition
2K
Aggressive Driving
2K
Vehicle Size
1K
0K
50K
100K
Type of Cars involved in Accidents
Car
127K
SUV
95K
Truck
13K
Motorcycle
8K
Taxi
8K
Bus
5K
Other
4K
Emergency V...
2K
Van
2K
Scooter
2K
Moped
1K
Garbage Truck
1K
Tanker
0K
RV
0K
Bicycle
0K
0K
50K
100K

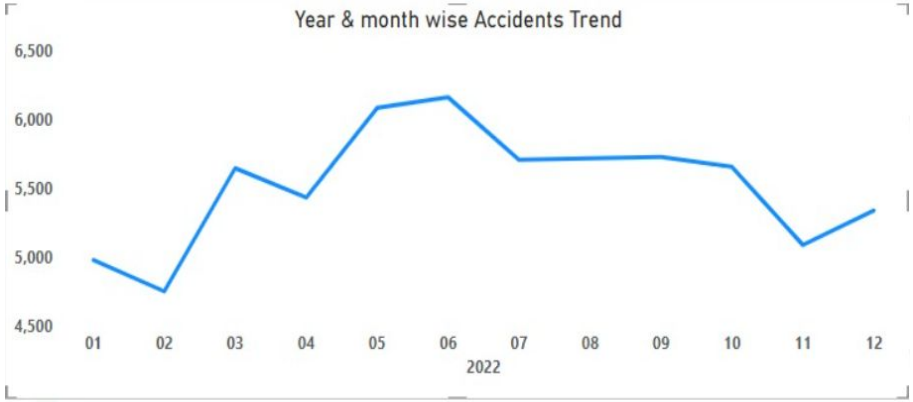
Visualization

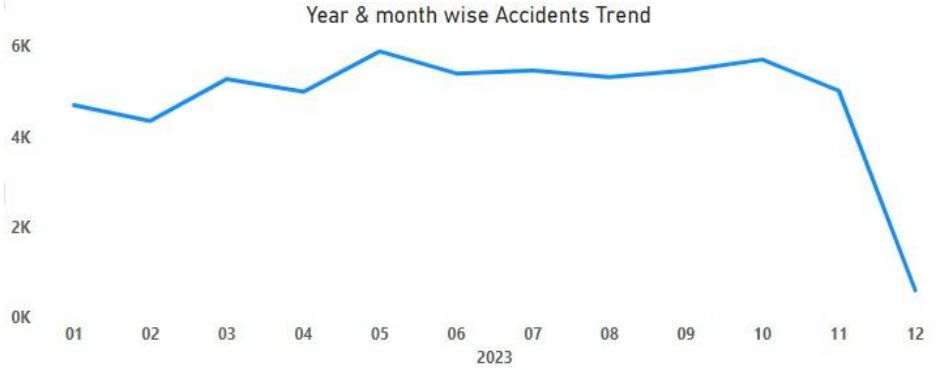


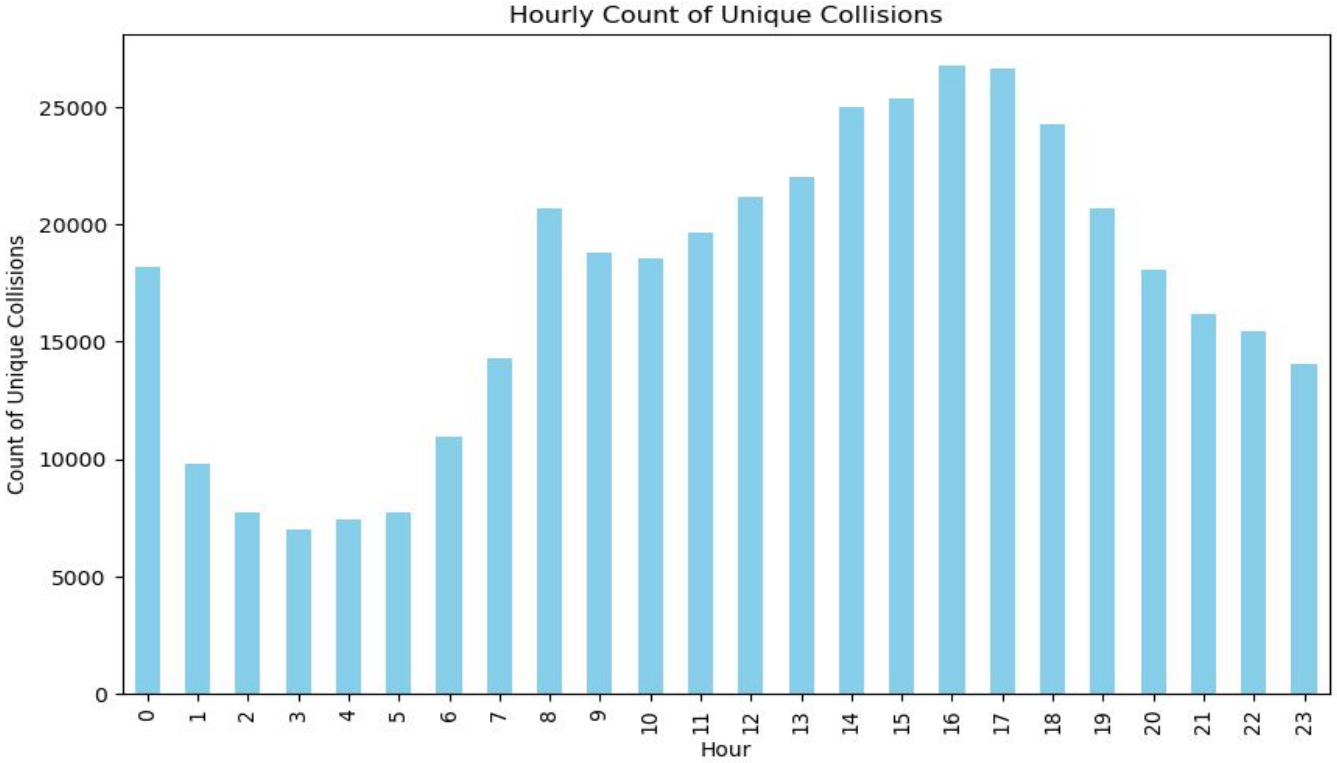
Visualization



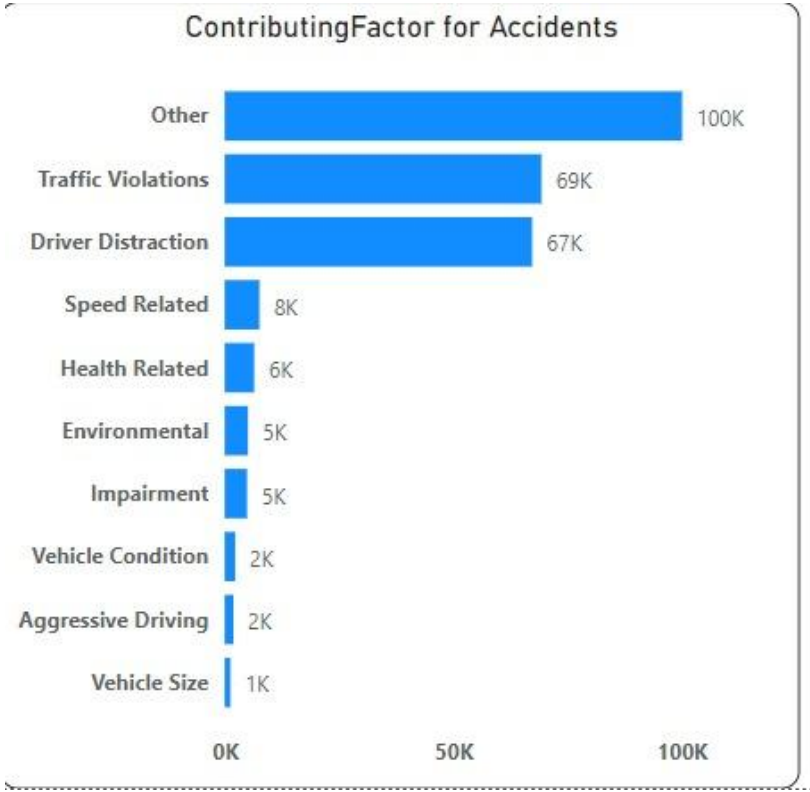


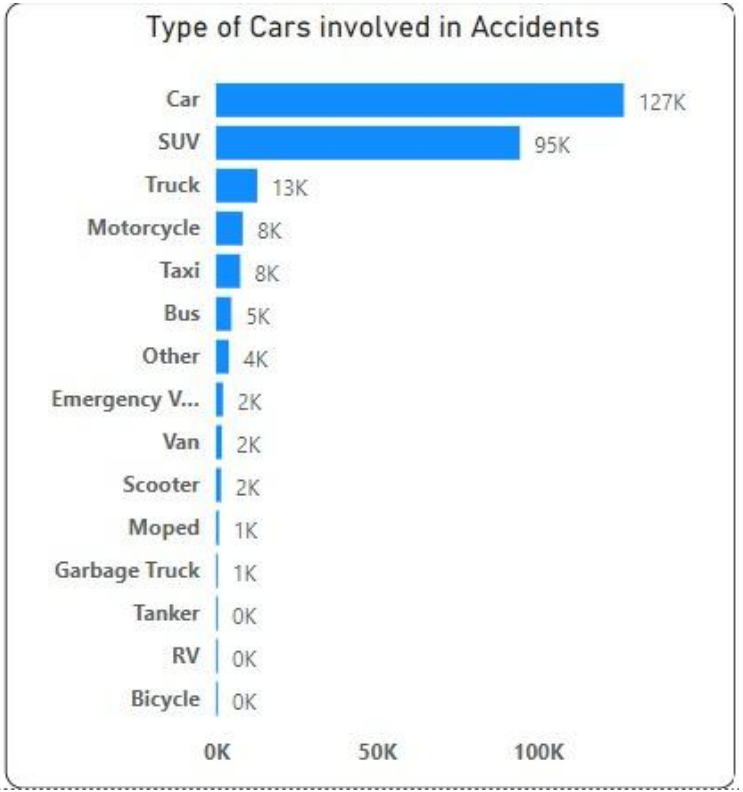




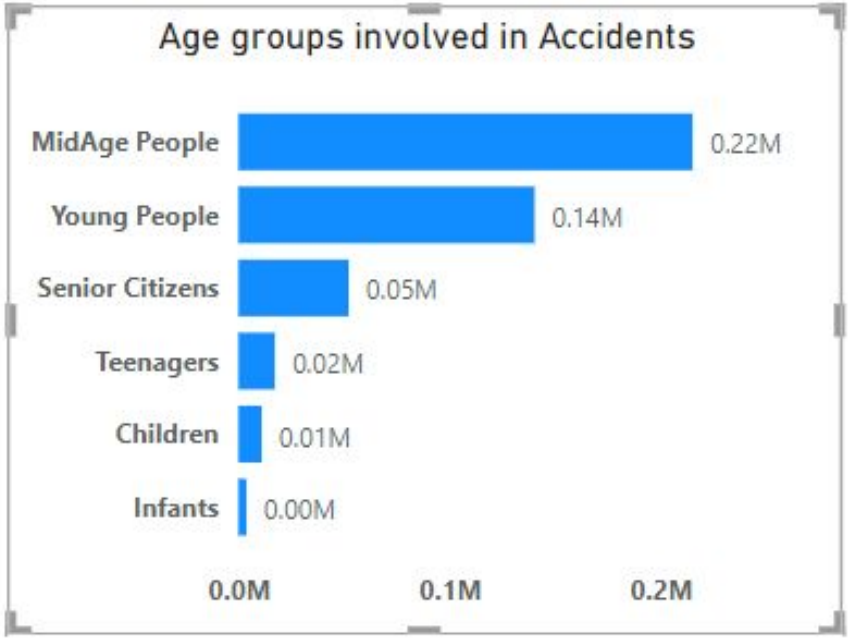

Hourly Count of Unique Collisions
Count of Unique Collisions
0
5000
10000
15000
20000
25000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

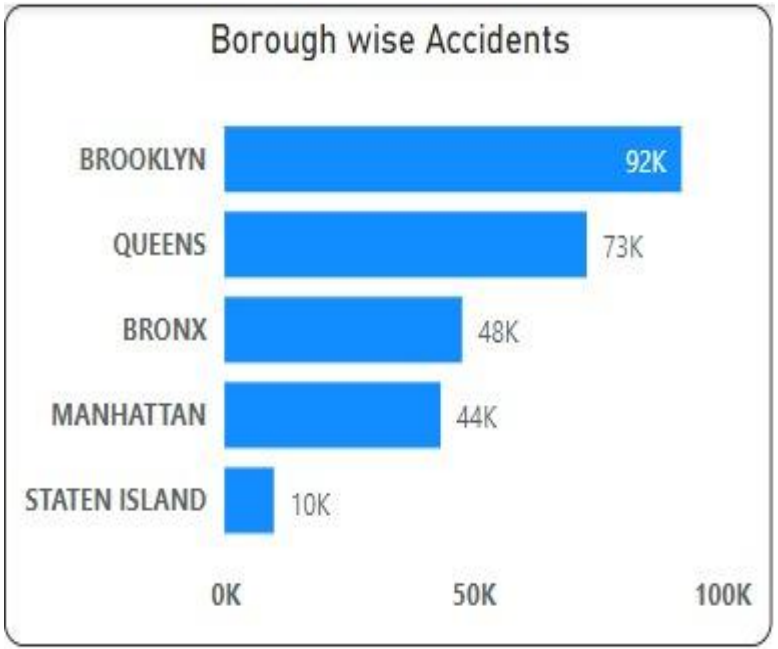
Visualization



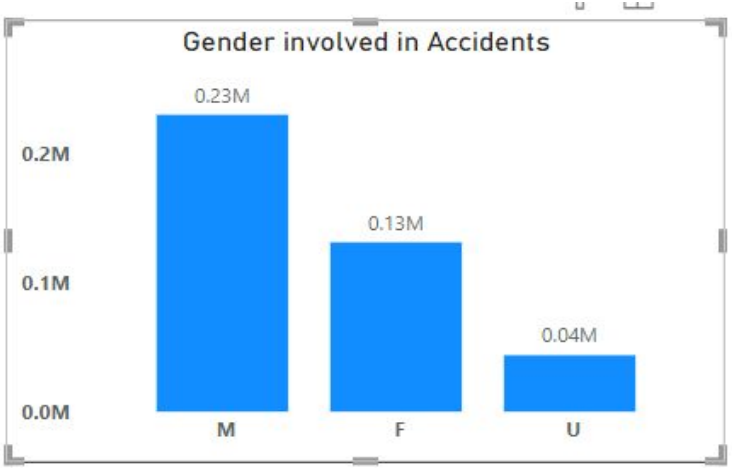


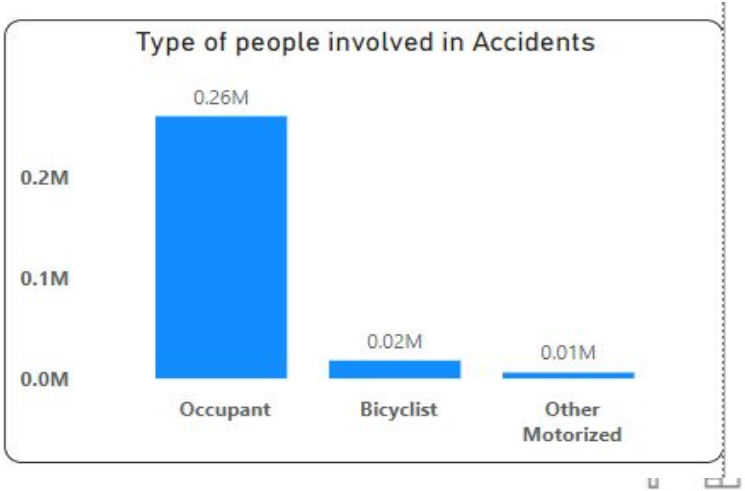
Visualization





Visualization



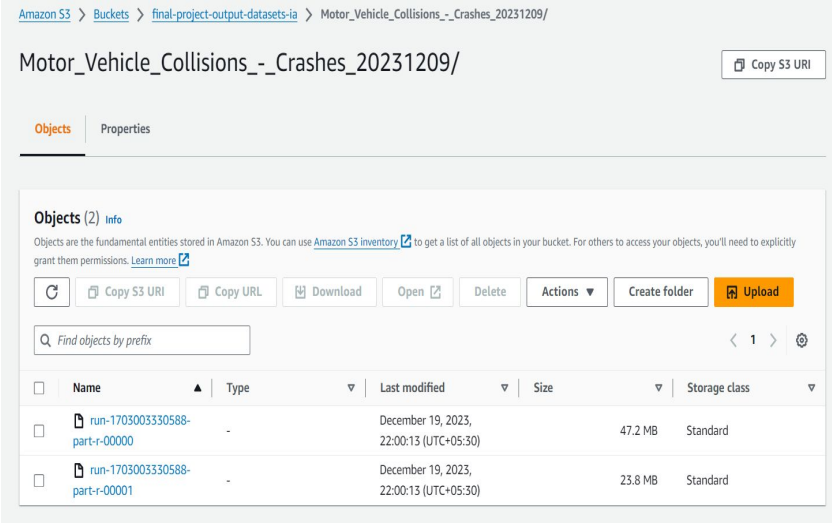


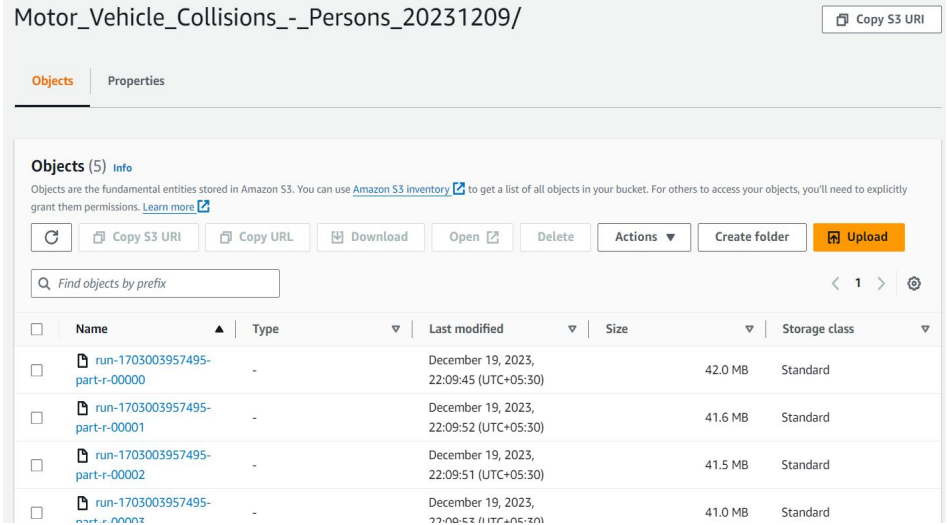
borough	Sum of Total Killed	Sum of Total Injured
BRONX	620	184988
BROOKLYN	1412	361836
MANHATTAN	589	128424
QUEENS	1019	278979
STATEN ISLAND	216	40680
Total	3856	994907

personsex	Sum of Total Killed	Sum of Total Injured
F	871	315635
M	2771	632828
U	214	46444
Total	3856	994907

Demo

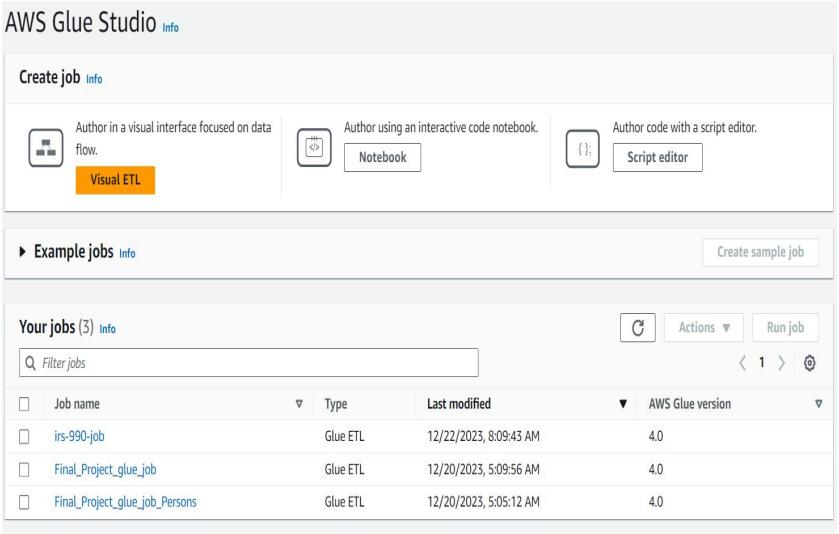
S3 Buckets



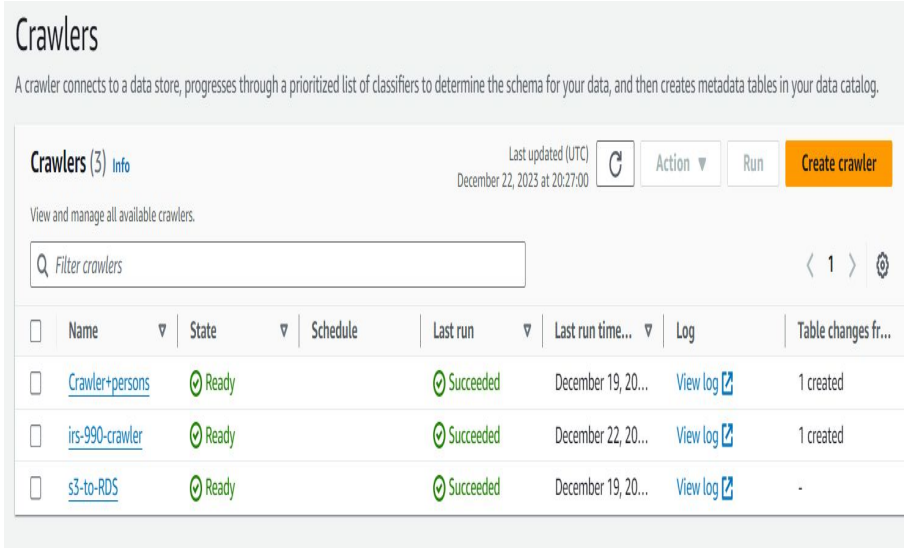


Demo

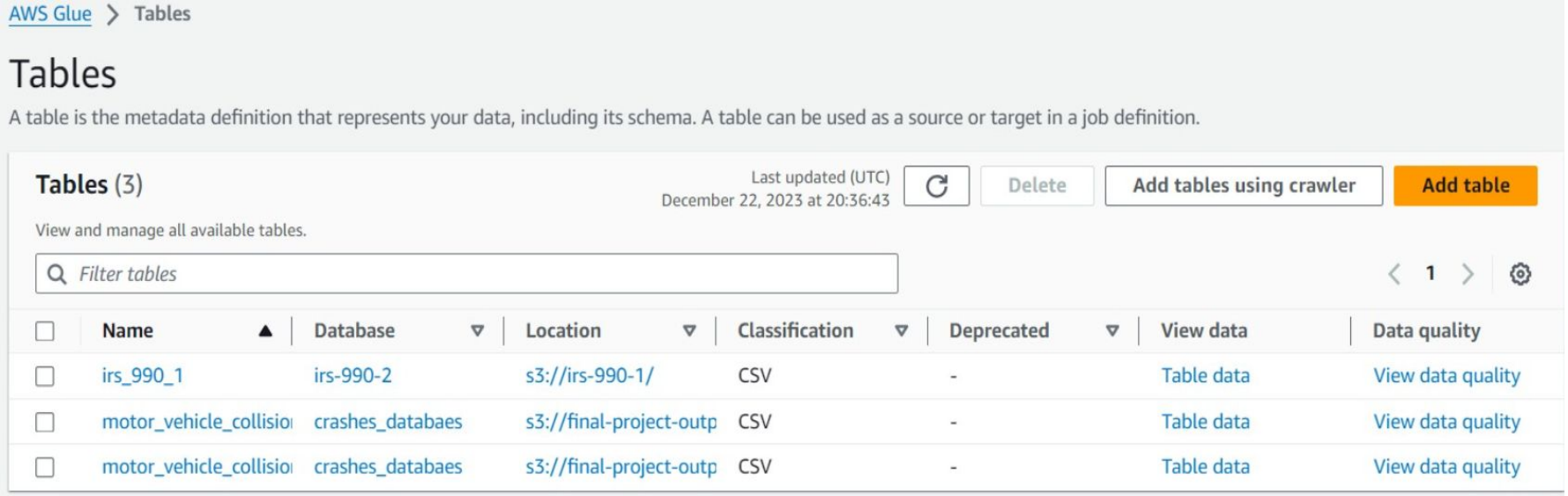
AWS Glue



Crawlers



Demo


AWS Glue > Tables
Tables
A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.
Tables (3)
Last updated (UTC)
December 22, 2023 at 20:36:43
Delete
Add tables using crawler
Add table
View and manage all available tables.
Filter tables
1
Name
Database
Location
Classification
Deprecated
View data
Data quality
irs_990_1
irs-990-2
s3://irs-990-1/
CSV
-
Table data
View data quality
motor_vehicle_collisio
crashes_databaes
s3://final-project-outp
CSV
-
Table data
View data quality
motor_vehicle_collisio
crashes_databaes
s3://final-project-outp
CSV
-
Table data
View data quality

Assumptions

Data Completeness and Accuracy: We assumed that the crash data captured in the fact table, along with the associated dimension tables (dim_date, dim_persons, dim_vehicles, dim_contribution_factor), is complete and accurate.

Stable and Consistent Data Sources: The project assumes that the data sources providing the information for the crash data are stable and consistent over time.

Policy Impact and Relevance Assumption: We assumed that the data and insights obtained from the fact table and related dimension tables will be directly relevant and actionable for policy makers and business stakeholders.

Challenges

i) Finding data

ii) Solving error from AWS connection

Thank You for Listening!