Community: Residential Neighborhoods

Issue: Crime and Safety

Introduction

This presentation provides an overview of our project on Crime and safety in Residential Areas, detailing the data lifecycle steps such as Data Generation, Collection, Processing, Storage, Management, Analysis, Visualization, Interpretation; Tools used such as GCP, GCP Storage Bucket, Dataproc, Hadoop Architecture, HDFS, HIVE, and Spark were utilized at different phases; and next steps for the data science or analyst teams. The project focuses on addressing the rise in criminal activity in residential neighborhoods, with a specific emphasis on ensuring safety and security.

Data Generation

1

Static Data Set

We obtained a static data set from [https://catalog.data.gov/dataset/crime-data-from-2010-to-2019] to analyze historical crime data and trends.

2

Streaming Data Set

Additionally, we collected streaming data from [https://catalog.data.gov/dataset/crime-data-from-2020-to-present]to incorporate real-time crime data into our analysis.

Data Collection

- We have collected these Static dataset and Streaming Dataset which is in format as a Raw Data is mainly from the official website of United states Government DATA.GOV.
- These data sets mainly have the data of crime in the city of Los Angeles.
- Static Data set is the data which has incidents of crime in the city of Los Angeles from 2010 - 2019.
- The streaming Data Set is the data from 2020 - Present which is Up-to-Date.
- In the step of Data Cleaning the raw data is altered into the necessary data by removing the unwanted data.
- where we obtained some data by selecting necessary columns and obtaining null percentage for each column for both Static Data set and Streaming Data set.

Processing and Storage

Google Cloud Platform

We utilized various tools such as BigQuery, Hive and Spark provided by the Google Cloud Platform (GCP) to create an efficient solution for addressing the rise in criminal activity.

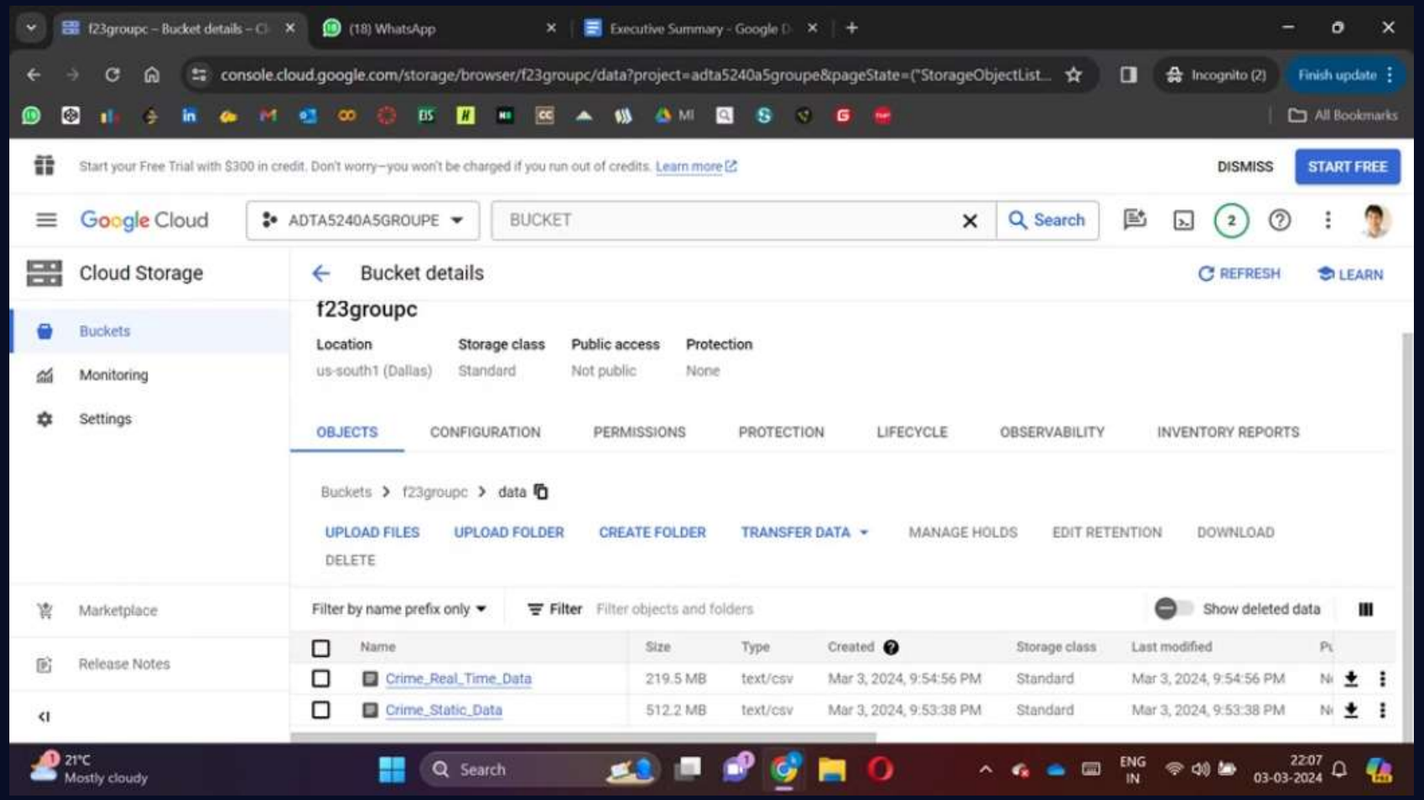
GCP Storage Bucket

In our GCP project, we established a new directory f23groupc and uploaded the sanitized dataset to a GCP storage bucket.

Dataproc

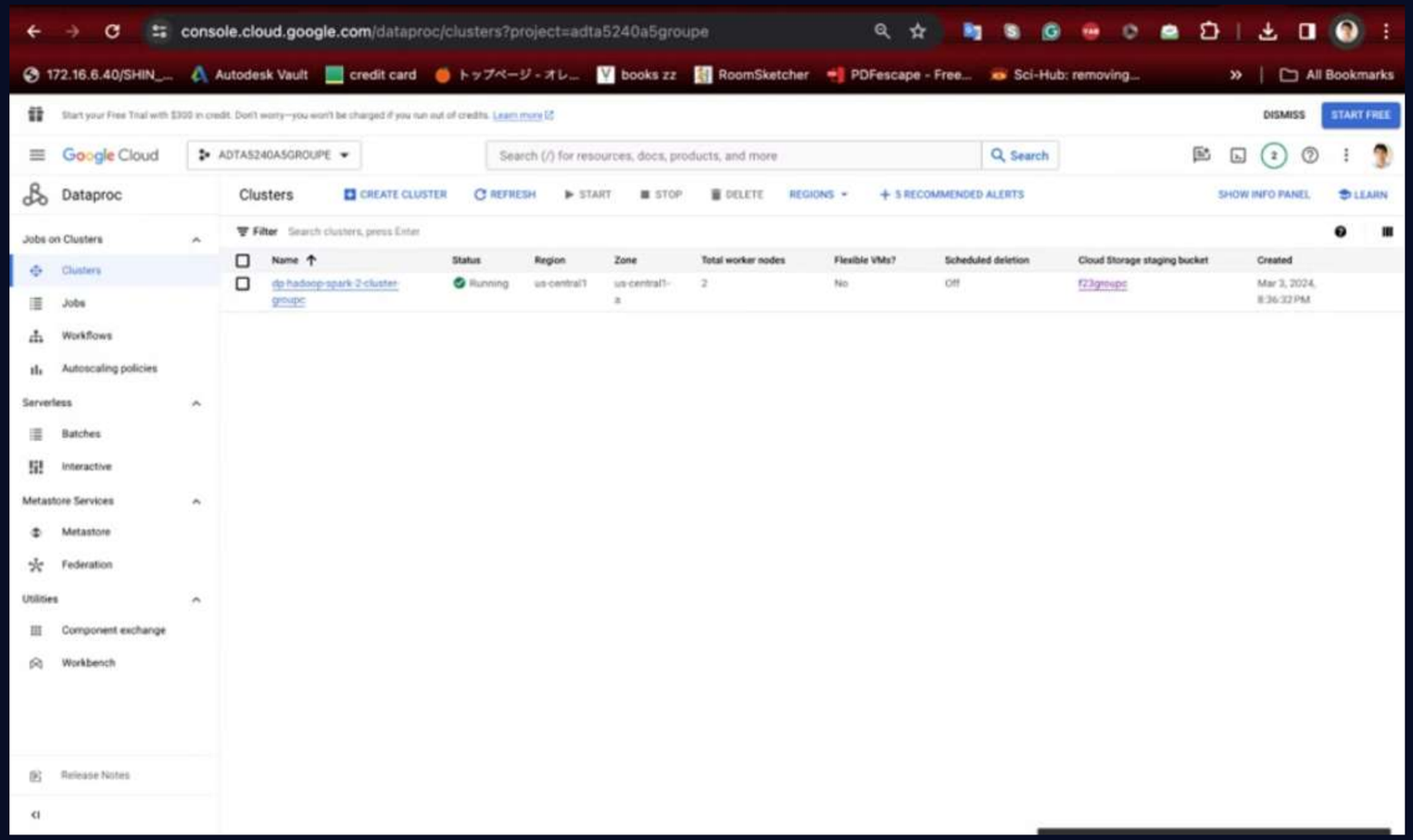
We created a Hadoop architecture using Google Dataproc to process the collected data effectively.

Creating a Bucket

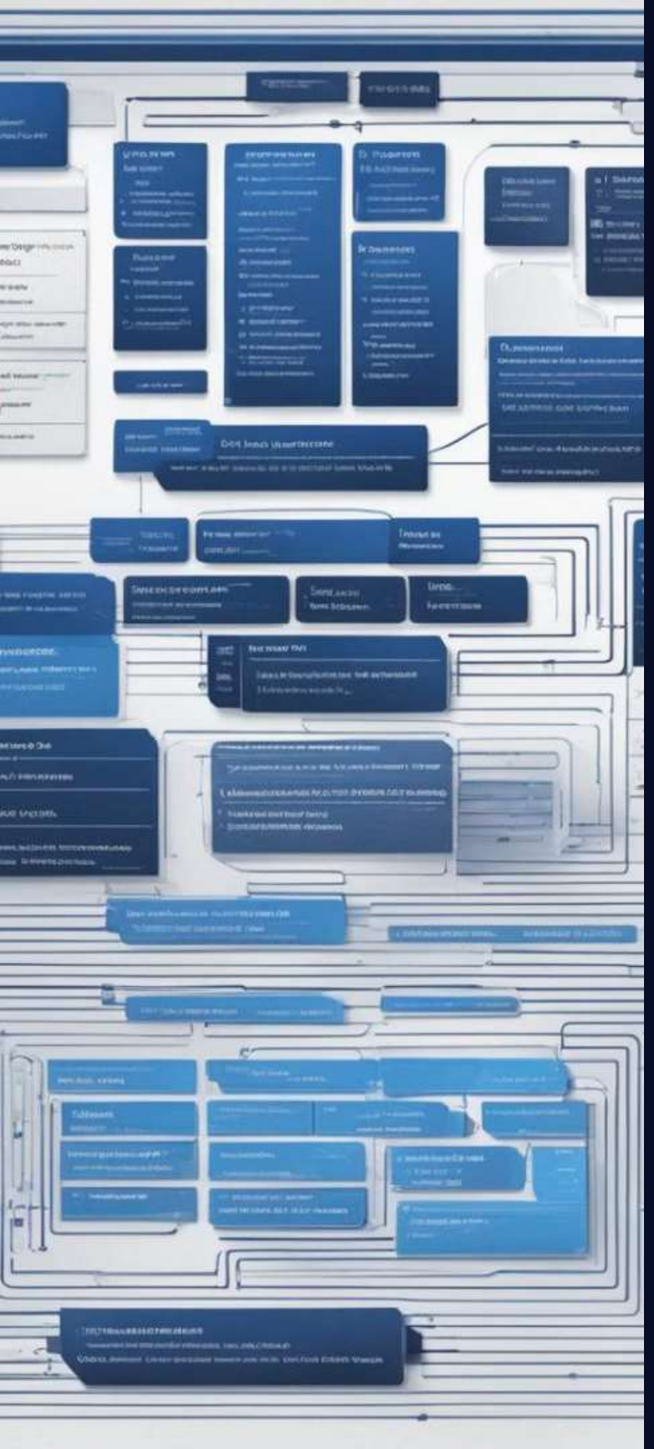


Here we created a bucket on Google Cloud Platform as f23groupc with some necessary requirements like location.

Dataproc



we have uploaded the data sets as Crime_Real_Time_Data and Crime_Static_Data into the folder data/. and have started the Cluster.



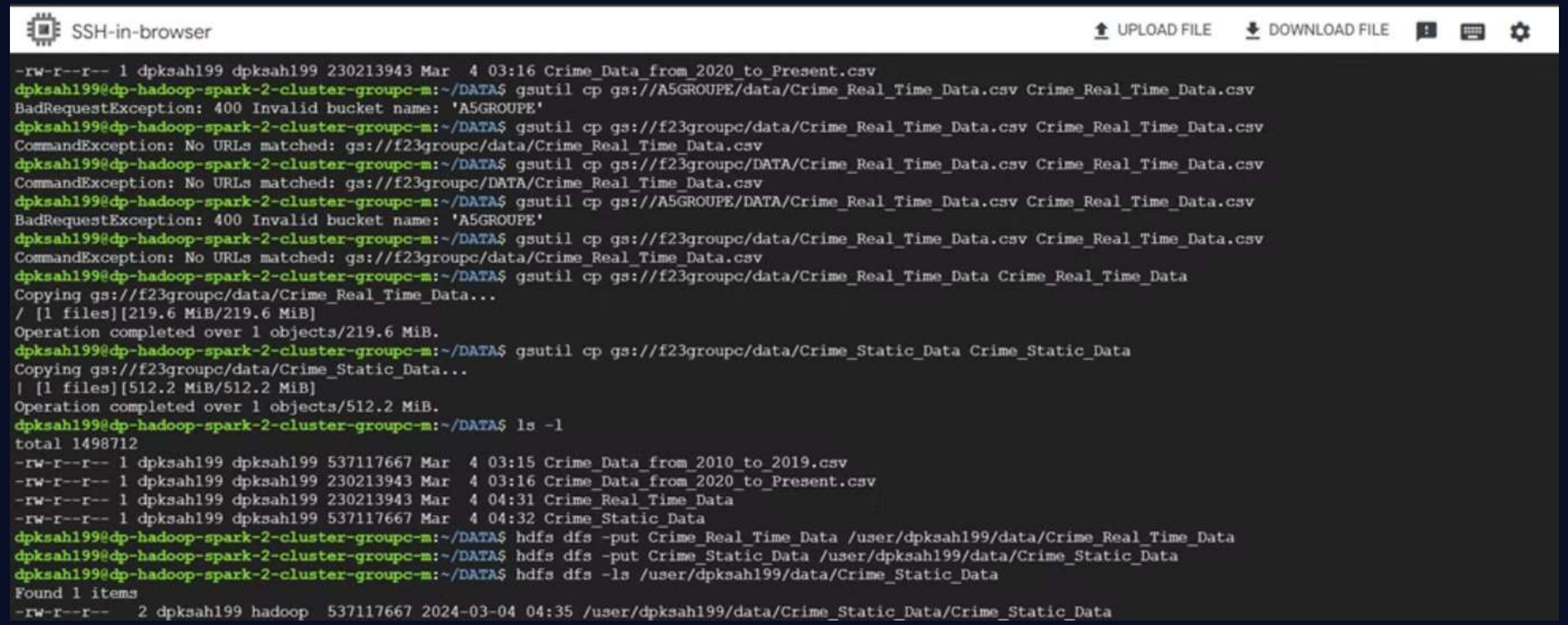
HDFS Integration

1 HDFS Data Loading

We loaded the datasets which we have uploaded into the Manager Node and integrated it into the Hadoop Distributed File System (HDFS) Ecosystem.

By the commands gsutil cp gs://f23groupc/data/Crime_static_Data Crime_Static_Data and Gsutilgs://f23groupc/data/Crime_Real_Time_DataCrime_Real_Time_Data we can copy the overall data into the integrated system by HDFS.

HDFS Echosystem



Here we used Hadoop to create HDFS Echosystem where we simply copy Data which we loaded into Manager Nodes to Intergrate the Data.

Data Management

HIVE on Static Data

Displayed the first five rows of static data for crime using HIVE for analysis.

SPARK on Static Data

Displayed the first five rows of static data for crime using SPARK for analysis.

1

2

3

4

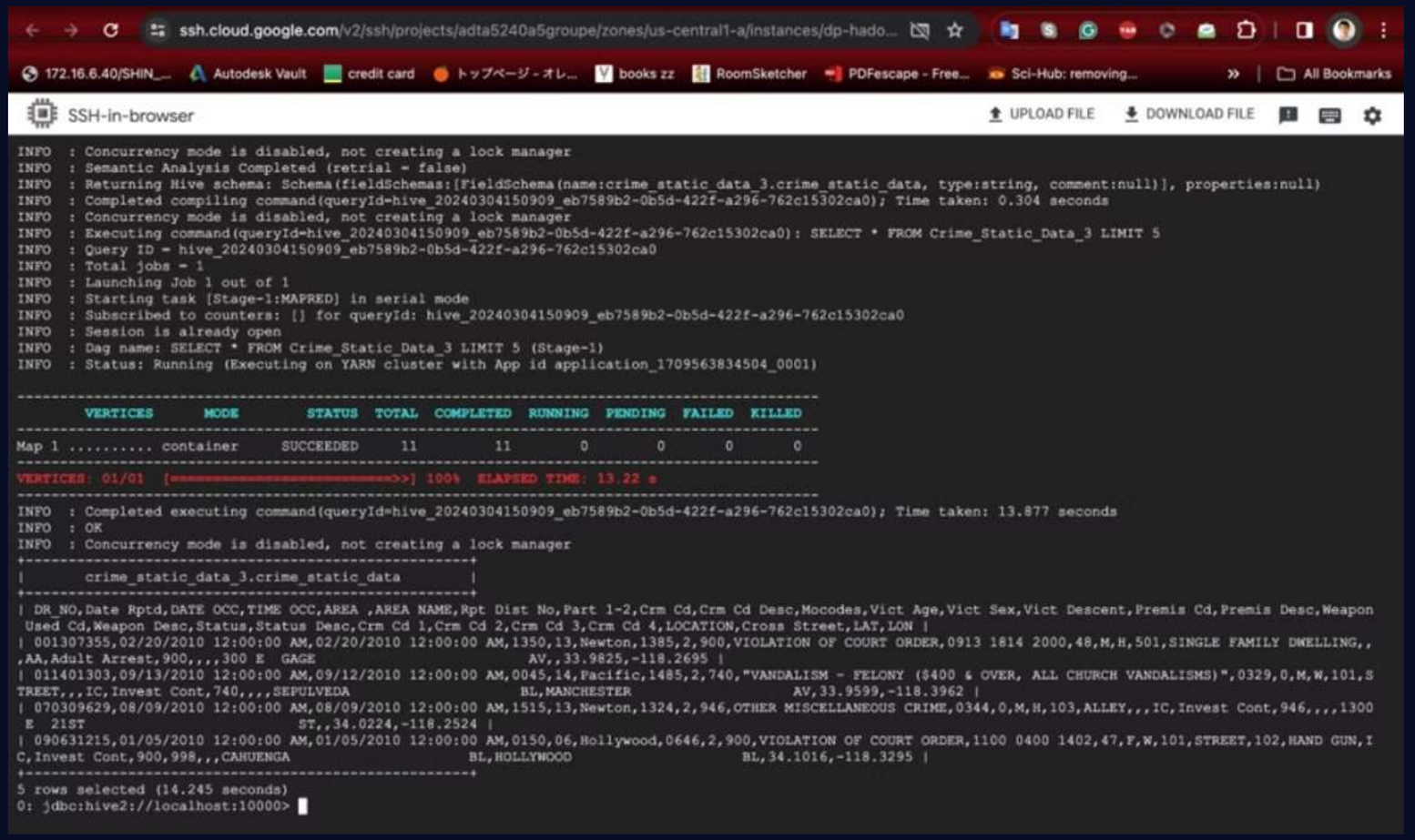
HIVE on Real-Time Data

Displayed the first five rows of real-time crime data using HIVE for analysis.

SPARK on Real-Time Data

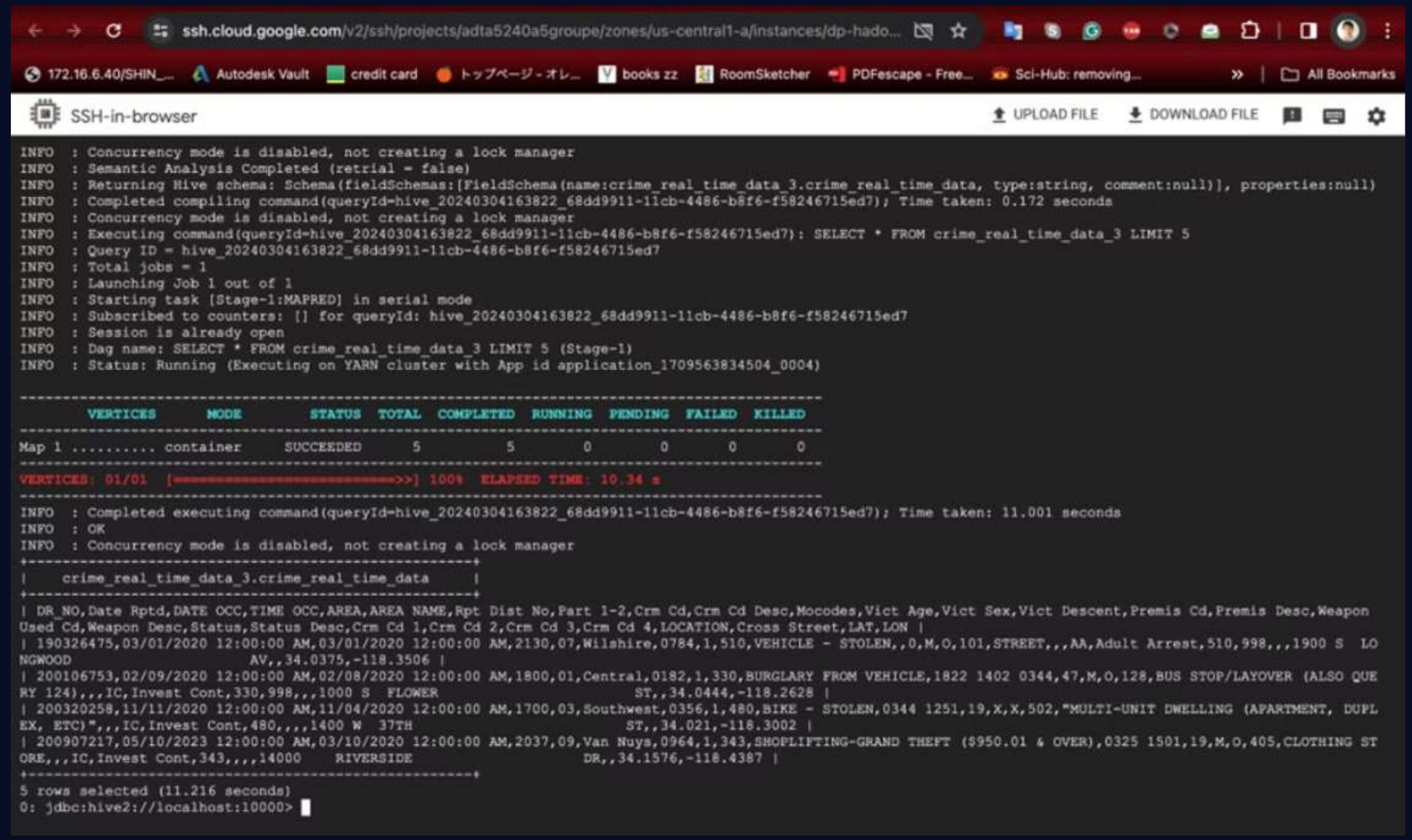
Displayed the first five rows of real-time crime data using SPARK for analysis.

Hive on Static Data



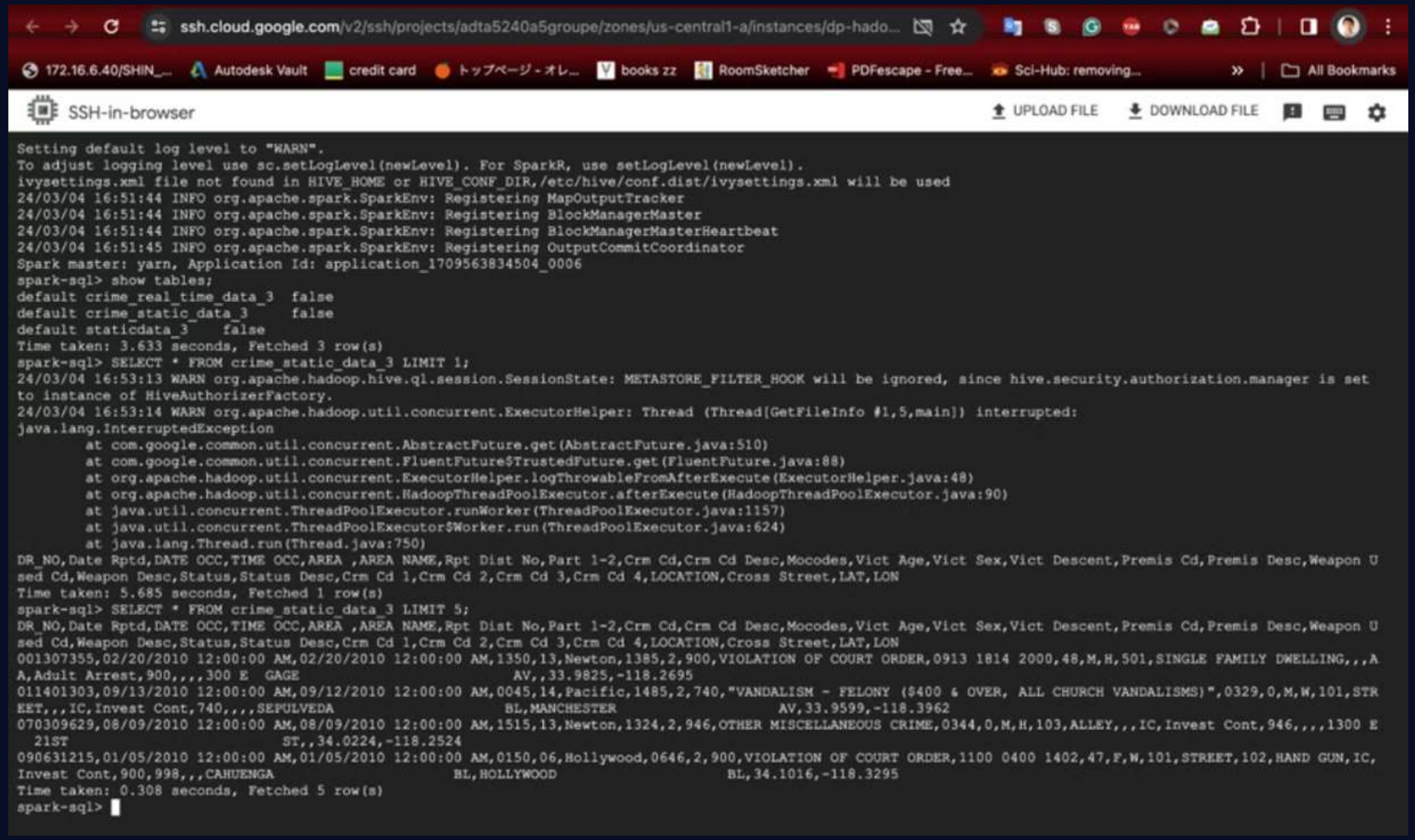
Here we use Hive on Static Data by creating a table crime_static_data_3 and it took 14.245 seconds.

HIVE on Real Time Data



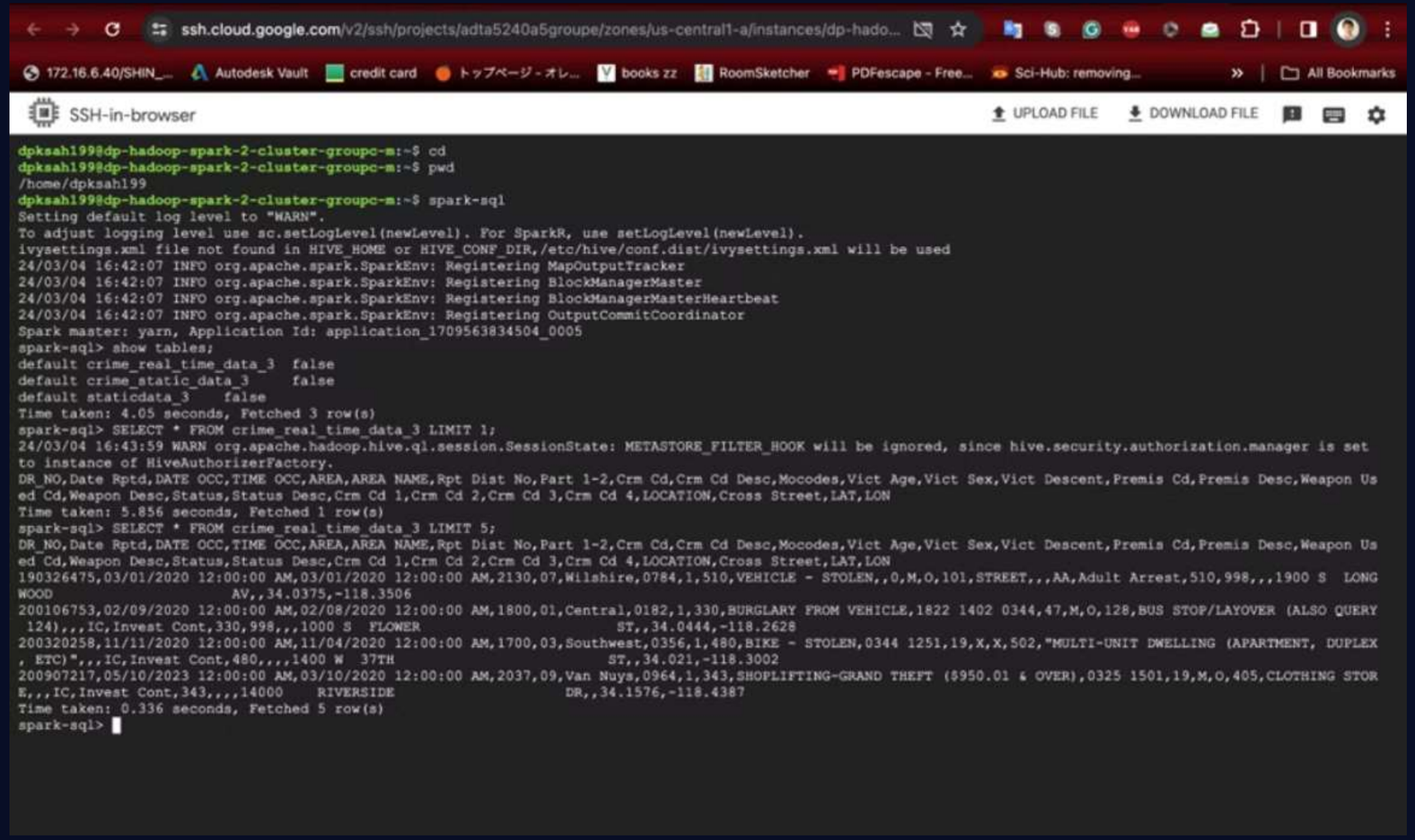
Here on Real time data it shows the various rows of data with the names in which it takes 11.216 seconds.

Spark on Static Data



We use Sql query to retrieve Data from the table and it takes 0.308 seconds.

Spark on Real time Data



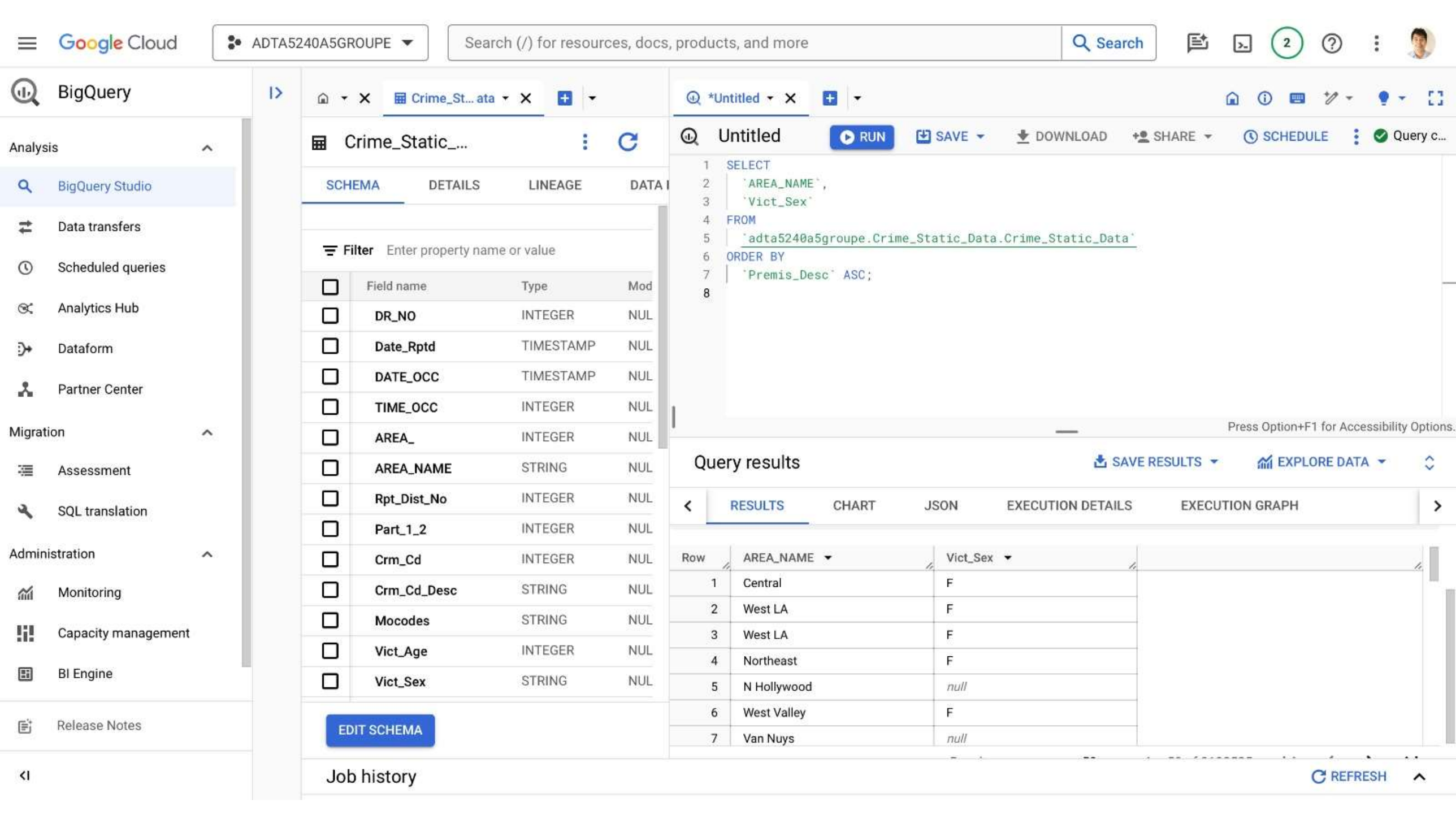
This is used to fetch the data in rows where it shows the limitations and its values; here it takes 0.336 seconds.

Analysis and BigQuery

1

BigQuery Performance

Compared the performance of BigQuery for data analysis tasks. Here We have given a query for retrieving data from the Datasets.


Google Cloud
ADTA5240A5GROUPE
Search (/) for resources, docs, products, and more
Search
BigQuery
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Migration
Assessment
SQL translation
Administration
Monitoring
Capacity management
BI Engine
Release Notes
Crime_St... ata
Crime_Static_...
SCHEMA
DETAILS
LINEAGE
Filter
Enter property name or value
Field name
Type
Mod
DR_NO INTEGER NUL
Date_Rptd TIMESTAMP NUL
DATE_OCC TIMESTAMP NUL
TIME_OCC INTEGER NUL
AREA_ INTEGER NUL
AREA_NAME STRING NUL
Rpt_Dist_No INTEGER NUL
Part_1_2 INTEGER NUL
Crm_Cd INTEGER NUL
Crm_Cd_Desc STRING NUL
Mocodes STRING NUL
Vict_Age INTEGER NUL
Vict_Sex STRING NUL
EDIT SCHEMA
*Untitled
Untitled
RUN
SAVE
DOWNLOAD
SHARE
SCHEDULE
Query c...
1 SELECT
2 `AREA_NAME`,
3 `Vict_Sex`
4 FROM
5 `adta5240a5groupe.Crime_Static_Data.Crime_Static_Data`
6 ORDER BY
7 `Premis_Desc` ASC;
8
Press Option+F1 for Accessibility Options.
Query results
SAVE RESULTS
EXPLORE DATA
RESULTS
CHART
JSON
EXECUTION DETAILS
EXECUTION GRAPH
Row AREA_NAME Vict_Sex
1 Central F
2 West LA F
3 West LA F
4 Northeast F
5 N Hollywood null
6 West Valley F
7 Van Nuys null
Job history
REFRESH

Data Visualization

Here we obtain the Visualization output as the above where it shows all the status by using different colors.

Next Steps for Data scientist

1 **Analysis Completion**

Concluding the analysis and preparing actionable insights for the data science and analyst teams.

2 **Report Generation**

Creating a comprehensive report based on the analysis and findings to guide future strategies and decisions.

Interpretation and Conclusion

- By studying databases from the past and even the present, data science helps us to have a broad perspective on a variety of topics and extract pertinent information for decision-making.
- In the particular instance of public security, it is feasible to pinpoint the locations that experience the highest rates of crime, tighten regulations on firearms, devise plans to reduce violence, and bolster neighborhood security.

- The utilization of data science fosters proactive rather than reactive measures in addressing public security concerns. Through continuous analysis and adaptation, data-driven strategies can lead to sustainable improvements in public safety outcomes.