



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 [\[https://catalog.data.gov/dataset/crime-data-from-2010-to-2019\]](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 [\[https://catalog.data.gov/dataset/crime-data-from-2020-to-present\]](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.

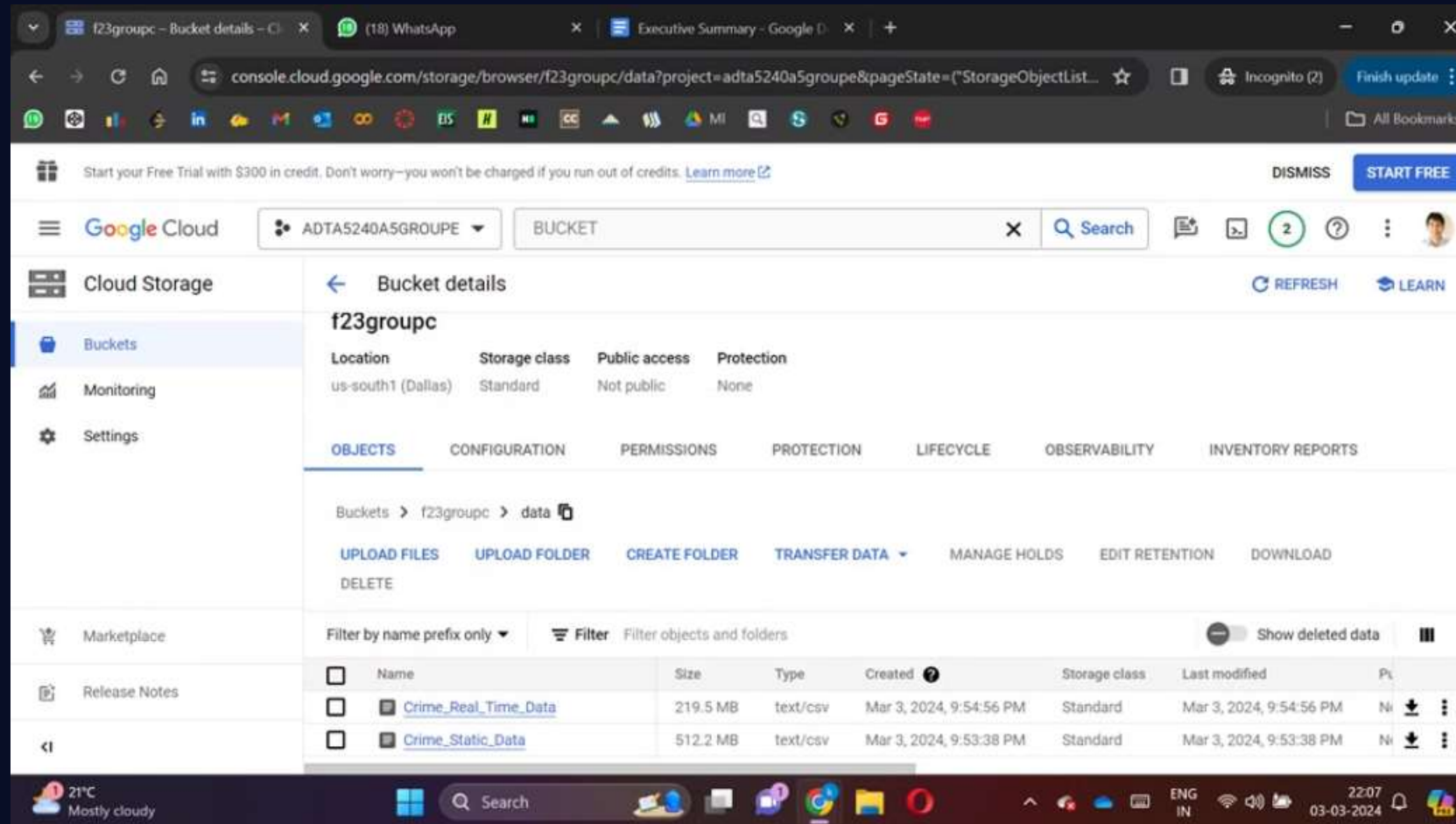
GCP Storage Bucket

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

The screenshot shows the Google Cloud console interface for the Dataproc Clusters page. The browser address bar displays `console.cloud.google.com/dataproc/clusters?project=adta5240a5groupe`. The left sidebar contains a navigation menu with categories like 'Jobs on Clusters', 'Serverless', 'Metastore Services', and 'Utilities'. The 'Clusters' link is selected. The main content area shows a table of clusters with the following columns: Name, Status, Region, Zone, Total worker nodes, Flexible VMs?, Scheduled deletion, Cloud Storage staging bucket, and Created. One cluster is listed with the name `dp-hadoop-spark-2-cluster-groupc`, status 'Running', region 'us-central1', zone 'us-central1-a', 2 worker nodes, and was created on Mar 3, 2024, at 8:36:32 PM. The cluster's Cloud Storage staging bucket is `fz3groupc`. Above the table, there are buttons for 'CREATE CLUSTER', 'REFRESH', 'START', 'STOP', and 'DELETE', along with a 'REGIONS' dropdown and a '+ 3 RECOMMENDED ALERTS' link. A search bar at the top of the table area prompts the user to 'Search clusters, press Enter'.

Name	Status	Region	Zone	Total worker nodes	Flexible VMs?	Scheduled deletion	Cloud Storage staging bucket	Created
dp-hadoop-spark-2-cluster-groupc	Running	us-central1	us-central1-a	2	No	Off	fz3groupc	Mar 3, 2024, 8:36:32 PM

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.

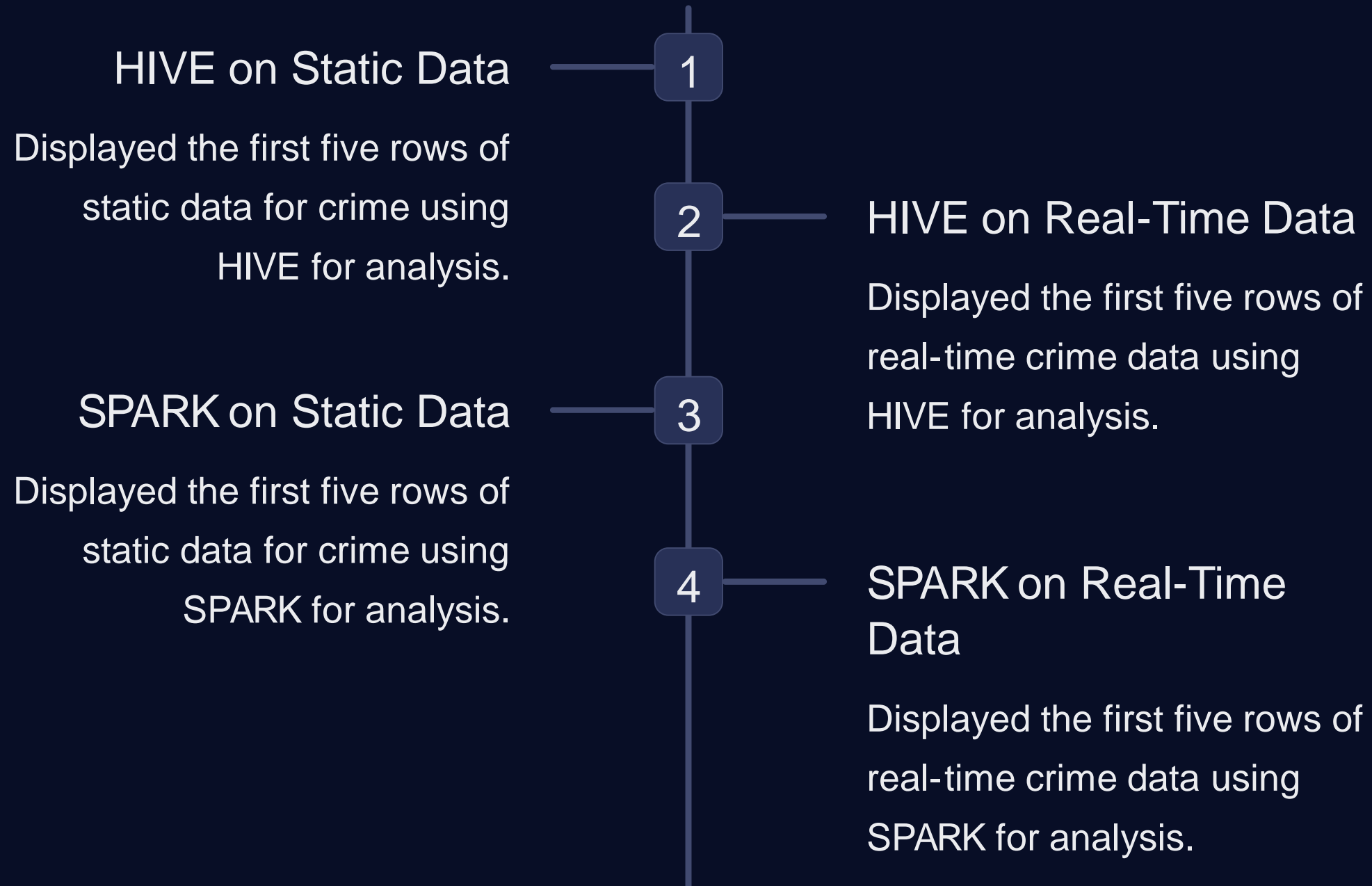
HDFS Echosystem

```
SSH-in-browser
[Icons: Upload File, Download File, Chat, Help, Settings]

-rw-r--r-- 1 dpksah199 dpksah199 230213943 Mar  4 03:16 Crime_Data_from_2020_to_Present.csv
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://A5GROUPE/data/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
BadRequestException: 400 Invalid bucket name: 'A5GROUPE'
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
CommandException: No URLs matched: gs://f23groupc/data/Crime_Real_Time_Data.csv
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/DATA/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
CommandException: No URLs matched: gs://f23groupc/DATA/Crime_Real_Time_Data.csv
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://A5GROUPE/DATA/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
BadRequestException: 400 Invalid bucket name: 'A5GROUPE'
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
CommandException: No URLs matched: gs://f23groupc/data/Crime_Real_Time_Data.csv
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Real_Time_Data Crime_Real_Time_Data
Copying gs://f23groupc/data/Crime_Real_Time_Data...
/ [1 files][219.6 MiB/219.6 MiB]
Operation completed over 1 objects/219.6 MiB.
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Static_Data Crime_Static_Data
Copying gs://f23groupc/data/Crime_Static_Data...
| [1 files][512.2 MiB/512.2 MiB]
Operation completed over 1 objects/512.2 MiB.
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ ls -l
total 1498712
-rw-r--r-- 1 dpksah199 dpksah199 537117667 Mar  4 03:15 Crime_Data_from_2010_to_2019.csv
-rw-r--r-- 1 dpksah199 dpksah199 230213943 Mar  4 03:16 Crime_Data_from_2020_to_Present.csv
-rw-r--r-- 1 dpksah199 dpksah199 230213943 Mar  4 04:31 Crime_Real_Time_Data
-rw-r--r-- 1 dpksah199 dpksah199 537117667 Mar  4 04:32 Crime_Static_Data
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -put Crime_Real_Time_Data /user/dpksah199/data/Crime_Real_Time_Data
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -put Crime_Static_Data /user/dpksah199/data/Crime_Static_Data
dpksah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -ls /user/dpksah199/data/Crime_Static_Data
Found 1 items
-rw-r--r--  2 dpksah199 hadoop  537117667 2024-03-04 04:35 /user/dpksah199/data/Crime_Static_Data/Crime_Static_Data
```

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



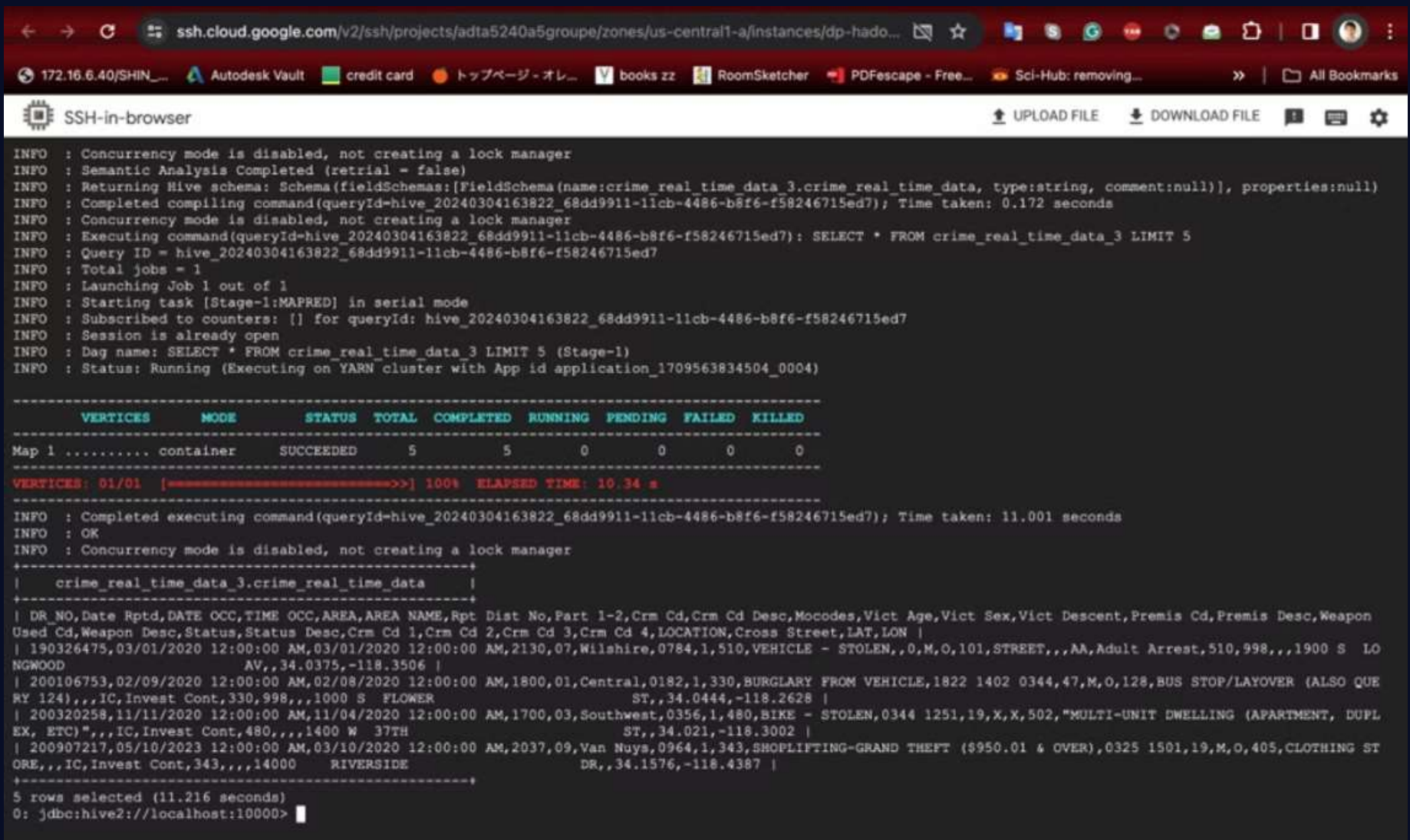
Hive on Static Data

```
ssh.cloud.google.com/v2/ssh/projects/adta5240a5groupe/zones/us-central1-a/instances/dp-hado...
172.16.6.40/SHIN... Autodesk Vault credit card トップページ - オレ... books zz RoomSketcher PDFescape - Free... Sci-Hub: removing... All Bookmarks
SSH-in-browser UPLOAD FILE DOWNLOAD FILE
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:crime_static_data_3.crime_static_data, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0); Time taken: 0.304 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0): SELECT * FROM Crime_Static_Data_3 LIMIT 5
INFO : Query ID = hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0
INFO : Session is already open
INFO : Dag name: SELECT * FROM Crime_Static_Data_3 LIMIT 5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1709563834504_0001)

-----
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 11 11 0 0 0 0
-----
VERTICES: 01/01 [=====] 100% ELAPSED TIME: 13.22 s
-----
INFO : Completed executing command(queryId=hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0); Time taken: 13.877 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| crime_static_data_3.crime_static_data |
+-----+
| DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA ,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon |
| Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON |
| 001307355,02/20/2010 12:00:00 AM,02/20/2010 12:00:00 AM,1350,13,Newton,1385,2,900,VIOLATION OF COURT ORDER,0913 1814 2000,48,M,H,501,SINGLE FAMILY DWELLING,, |
| ,AA,Adult Arrest,900,,,,,300 E GAGE AV,,33.9825,-118.2695 |
| 011401303,09/13/2010 12:00:00 AM,09/12/2010 12:00:00 AM,0045,14,Pacific,1485,2,740,"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)",0329,0,M,W,101,S |
| TREET,,,IC,Invest Cont,740,,,SEPULVEDA BL,MANCHESTER AV,33.9599,-118.3962 |
| 070309629,08/09/2010 12:00:00 AM,08/09/2010 12:00:00 AM,1515,13,Newton,1324,2,946,OTHER MISCELLANEOUS CRIME,0344,0,M,H,103,ALLEY,,,IC,Invest Cont,946,,,,,1300 |
| E 21ST ST,,34.0224,-118.2524 |
| 090631215,01/05/2010 12:00:00 AM,01/05/2010 12:00:00 AM,0150,06,Hollywood,0646,2,900,VIOLATION OF COURT ORDER,1100 0400 1402,47,F,W,101,STREET,102,HAND GUN,I |
| C,Invest Cont,900,998,,,CAHUENGA BL,HOLLYWOOD BL,34.1016,-118.3295 |
+-----+
5 rows selected (14.245 seconds)
0: jdbc:hive2://localhost:10000>
```

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



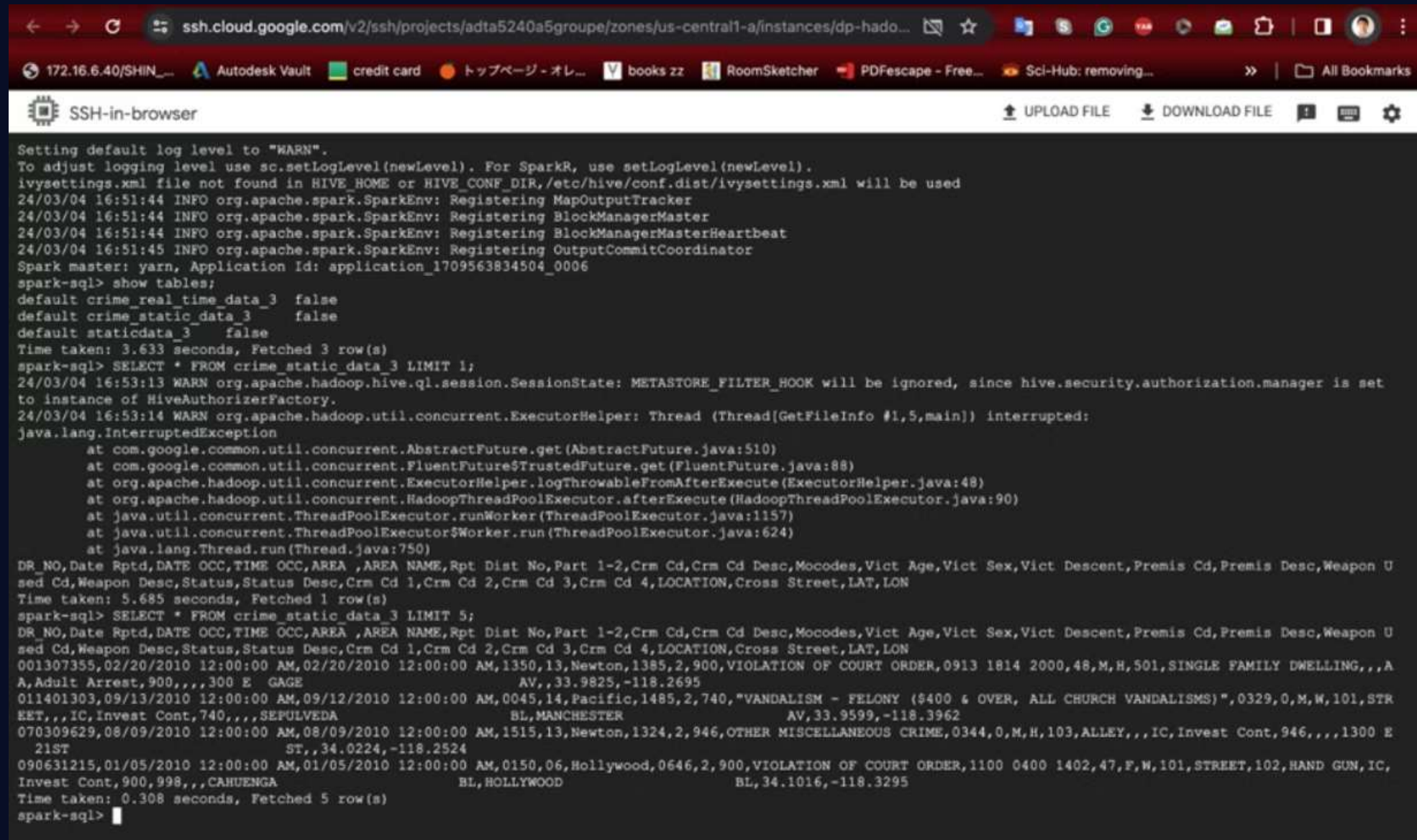
The screenshot shows an SSH-in-browser window with a terminal interface. The terminal displays Hive execution logs and the results of a query. The logs indicate that the query is running on a YARN cluster and has completed successfully. The query results are displayed as a table with 10 columns: VERTICES, MODE, STATUS, TOTAL, COMPLETED, RUNNING, PENDING, FAILED, and KILLED. The results show that the query completed successfully with 5 rows selected.

```
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:crime_real_time_data_3.crime_real_time_data, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7); Time taken: 0.172 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7): SELECT * FROM crime_real_time_data_3 LIMIT 5
INFO : Query ID = hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7
INFO : Session is already open
INFO : Dag name: SELECT * FROM crime_real_time_data_3 LIMIT 5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1709563834504_0004)

-----
VERTICES    MODE      STATUS    TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container    SUCCEEDED      5         5         0         0         0         0
-----
VERTICES: 01/01 [=====] 100% ELAPSED TIME: 10.34 s
-----
INFO : Completed executing command(queryId=hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7); Time taken: 11.001 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| crime_real_time_data_3.crime_real_time_data |
+-----+
| DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON |
| 190326475,03/01/2020 12:00:00 AM,03/01/2020 12:00:00 AM,2130,07,Wilshire,0784,1,510,VEHICLE - STOLEN,,0,M,O,101,STREET,,,AA,Adult Arrest,510,998,,,1900 S LO NGWOOD AV,,34.0375,-118.3506 |
| 200106753,02/09/2020 12:00:00 AM,02/08/2020 12:00:00 AM,1800,01,Central,0182,1,330,BURGLARY FROM VEHICLE,1822 1402 0344,47,M,O,128,BUS STOP/LAYOVER (ALSO QUE RY 124),,,IC,Invest Cont,330,998,,,1000 S FLOWER ST,,34.0444,-118.2628 |
| 200320258,11/11/2020 12:00:00 AM,11/04/2020 12:00:00 AM,1700,03,Southwest,0356,1,480,BIKE - STOLEN,0344 1251,19,X,X,502,"MULTI-UNIT DWELLING (APARTMENT, DUPL EX, ETC)",,,IC,Invest Cont,480,,,1400 W 37TH ST,,34.021,-118.3002 |
| 200907217,05/10/2023 12:00:00 AM,03/10/2020 12:00:00 AM,2037,09,Van Nuys,0964,1,343,SHOPLIFTING-GRAND THEFT ($950.01 & OVER),0325 1501,19,M,O,405,CLOTHING ST ORE,,,IC,Invest Cont,343,,,14000 RIVERSIDE DR,,34.1576,-118.4387 |
+-----+
5 rows selected (11.216 seconds)
0: jdbc:hive2://localhost:10000>
```

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

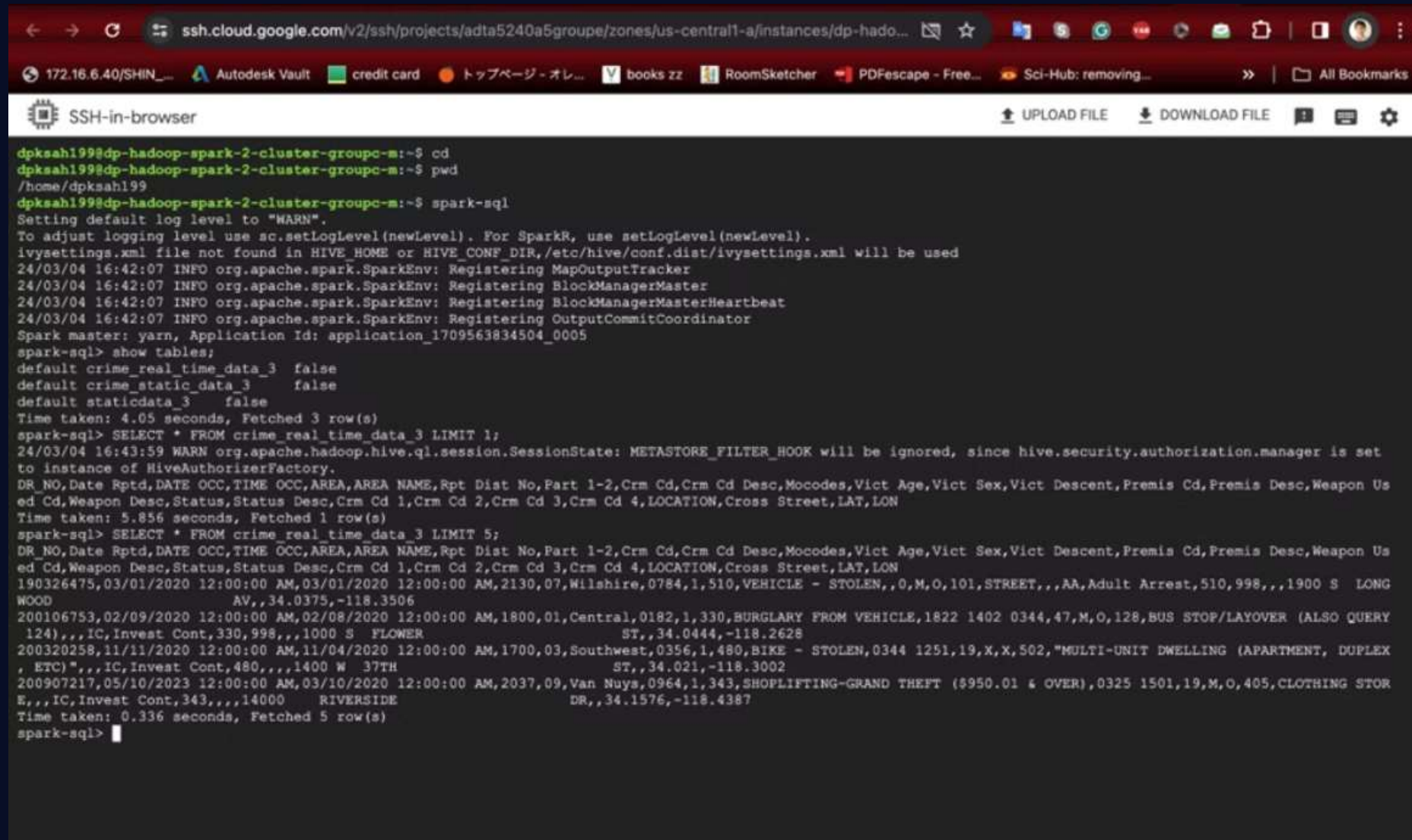


The screenshot shows a web browser window with an SSH-in-browser interface. The address bar shows a Google Cloud SSH connection to a Hadoop instance. The browser's bookmark bar contains various links like 'Autodesk Vault', 'credit card', and 'books zz'. The main content area displays the output of an SSH session. The session starts with setting the default log level to 'WARN'. It then shows Spark SQL commands and their outputs. The first command is 'show tables;', which returns three tables: 'crime_real_time_data_3', 'crime_static_data_3', and 'staticdata_3'. The second command is 'SELECT * FROM crime_static_data_3 LIMIT 1;', which returns a single row of data. The third command is 'SELECT * FROM crime_static_data_3 LIMIT 5;', which returns five rows of data. The output for the third command is truncated in the image.

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
24/03/04 16:51:44 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/03/04 16:51:44 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/03/04 16:51:44 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/03/04 16:51:45 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1709563834504_0006
spark-sql> show tables;
default crime_real_time_data_3 false
default crime_static_data_3 false
default staticdata_3 false
Time taken: 3.633 seconds, Fetched 3 row(s)
spark-sql> SELECT * FROM crime_static_data_3 LIMIT 1;
24/03/04 16:53:13 WARN org.apache.hadoop.hive.q1.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set
to instance of HiveAuthorizerFactory.
24/03/04 16:53:14 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA ,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon U
sed Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
Time taken: 5.685 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM crime_static_data_3 LIMIT 5;
DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA ,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon U
sed Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
001307355,02/20/2010 12:00:00 AM,02/20/2010 12:00:00 AM,1350,13,Newton,1385,2,900,VIOLETION OF COURT ORDER,0913 1814 2000,48,M,H,501,SINGLE FAMILY DWELLING,,,A
A,Adult Arrest,900,,,,300 E GAGE AV,,33.9825,-118.2695
011401303,09/13/2010 12:00:00 AM,09/12/2010 12:00:00 AM,0045,14,Pacific,1485,2,740,"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)",0329,0,M,W,101,STR
EET,,,IC,Invest Cont,740,,,,SEPULVEDA BL,MANCHESTER AV,33.9599,-118.3962
070309629,08/09/2010 12:00:00 AM,08/09/2010 12:00:00 AM,1515,13,Newton,1324,2,946,OTHER MISCELLANEOUS CRIME,0344,0,M,H,103,ALLEY,,,IC,Invest Cont,946,,,,1300 E
21ST ST,,34.0224,-118.2524
090631215,01/05/2010 12:00:00 AM,01/05/2010 12:00:00 AM,0150,06,Hollywood,0646,2,900,VIOLETION OF COURT ORDER,1100 0400 1402,47,F,W,101,STREET,102,HAND GUN,IC,
Invest Cont,900,998,,,CAHUENGA BL,HOLLYWOOD BL,34.1016,-118.3295
Time taken: 0.308 seconds, Fetched 5 row(s)
spark-sql>
```

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

Spark on Real time Data



```
ssh.cloud.google.com/v2/ssh/projects/adta5240a5groupe/zones/us-central1-a/instances/dp-hado...
172.16.6.40/SHIN_... Autodesk Vault credit card トップページ - オレ... books zz RoomSketcher PDFescape - Free... Sci-Hub: removing... All Bookmarks
SSH-in-browser UPLOAD FILE DOWNLOAD FILE
dpksahl99@dp-hadoop-spark-2-cluster-groupc-m:~$ cd
dpksahl99@dp-hadoop-spark-2-cluster-groupc-m:~$ pwd
/home/dpksahl99
dpksahl99@dp-hadoop-spark-2-cluster-groupc-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/ivysettings.xml will be used
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1709563834504_0005
spark-sql> show tables;
default crime_real_time_data_3 false
default crime_static_data_3 false
default staticdata_3 false
Time taken: 4.05 seconds, Fetched 3 row(s)
spark-sql> SELECT * FROM crime_real_time_data_3 LIMIT 1;
24/03/04 16:43:59 WARN org.apache.hadoop.hive.q1.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set
to instance of HiveAuthorizerFactory.
DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Us
ed Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
Time taken: 5.856 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM crime_real_time_data_3 LIMIT 5;
DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Us
ed Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
190326475,03/01/2020 12:00:00 AM,03/01/2020 12:00:00 AM,2130,07,Wilshire,0784,1,510,VEHICLE - STOLEN,,0,M,O,101,STREET,,,AA,Adult Arrest,510,998,,,1900 S LONG
WOOD AV,,34.0375,-118.3506
200106753,02/09/2020 12:00:00 AM,02/08/2020 12:00:00 AM,1800,01,Central,0182,1,330,BURGLARY FROM VEHICLE,1822 1402 0344,47,M,O,128,BUS STOP/LAYOVER (ALSO QUERY
124),,,IC,Invest Cont,330,998,,,1000 S FLOWER ST,,34.0444,-118.2628
200320258,11/11/2020 12:00:00 AM,11/04/2020 12:00:00 AM,1700,03,Southwest,0356,1,480,BIKE - STOLEN,0344 1251,19,X,X,502,"MULTI-UNIT DWELLING (APARTMENT, DUPLEX
, ETC)",,,IC,Invest Cont,480,,,1400 W 37TH ST,,34.021,-118.3002
200907217,05/10/2023 12:00:00 AM,03/10/2020 12:00:00 AM,2037,09,Van Nuys,0964,1,343,SHOPLIFTING-GRAND THEFT ($950.01 & OVER),0325 1501,19,M,O,405,CLOTHING STOR
E,,,IC,Invest Cont,343,,,14000 RIVERSIDE DR,,34.1576,-118.4387
Time taken: 0.336 seconds, Fetched 5 row(s)
spark-sql>
```

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.



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_Static_...

SCHEMA

DETAILS

LINEAGE

DATA

Filter Enter property name or value

	Field name	Type	Mod
<input type="checkbox"/>	DR_NO	INTEGER	NUL
<input type="checkbox"/>	Date_Rptd	TIMESTAMP	NUL
<input type="checkbox"/>	DATE_OCC	TIMESTAMP	NUL
<input type="checkbox"/>	TIME_OCC	INTEGER	NUL
<input type="checkbox"/>	AREA_	INTEGER	NUL
<input type="checkbox"/>	AREA_NAME	STRING	NUL
<input type="checkbox"/>	Rpt_Dist_No	INTEGER	NUL
<input type="checkbox"/>	Part_1_2	INTEGER	NUL
<input type="checkbox"/>	Crm_Cd	INTEGER	NUL
<input type="checkbox"/>	Crm_Cd_Desc	STRING	NUL
<input type="checkbox"/>	Mocodes	STRING	NUL
<input type="checkbox"/>	Vict_Age	INTEGER	NUL
<input type="checkbox"/>	Vict_Sex	STRING	NUL

EDIT SCHEMA

Job history

*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
```

Query results

SAVE RESULTS

EXPLORE DATA

RESULTS

CHART

JSON

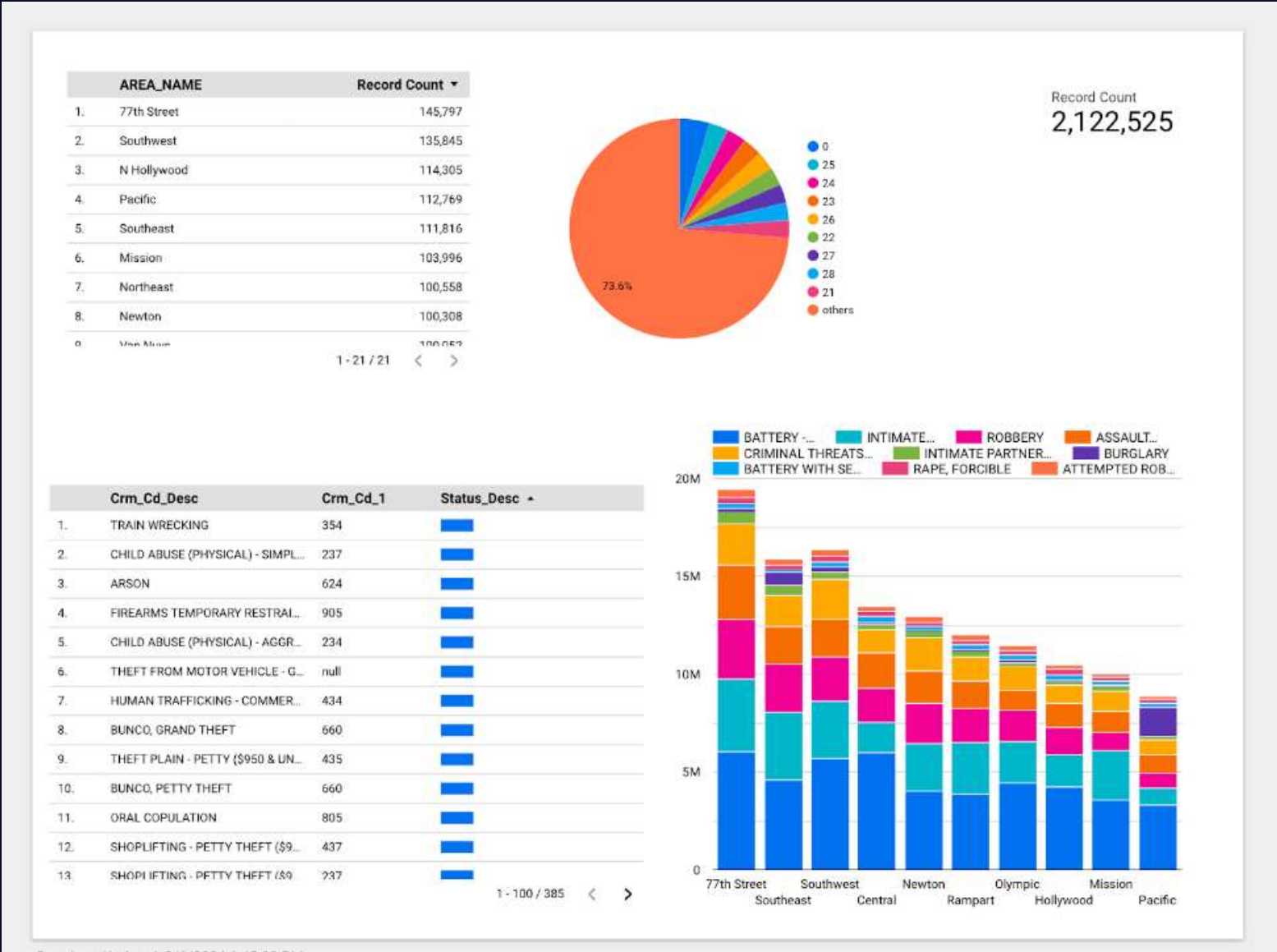
EXECUTION DETAILS

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

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.