Evaluación 2 - Trabajando en Big Data: herramientas y procesos esenciales

U.S. Homicide Reports, 1980-2014

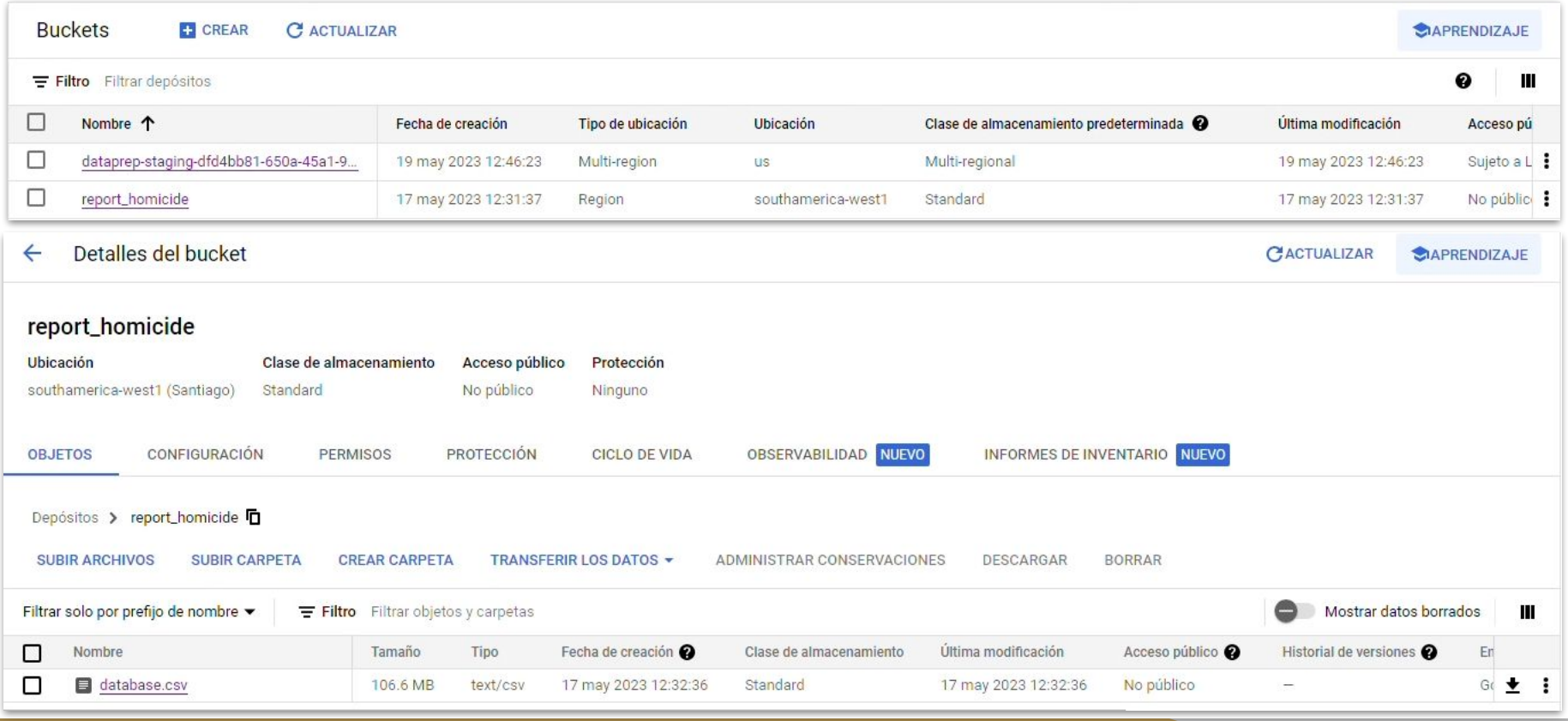
Desarrollo:

Para este proyecto, hemos seleccionado el conjunto de datos "U.S. Homicide Reports, 1980-2014". Este conjunto de datos contiene información detallada sobre los informes de homicidio en Estados Unidos durante el período de 1980 a 2014. Incluye datos como la ubicación del incidente, el año y mes en que ocurrió, características demográficas de las víctimas y perpetradores, tipos de crímenes, armas utilizadas y mucho más.

Kaggle (Data)

Carga de datos en Cloud Storage:

Para comenzar, cargamos el conjunto de datos en un Google Cloud Bucket. Utilizamos el servicio de almacenamiento en la nube de Google para almacenar y gestionar los archivos del conjunto de datos. Esto nos permite acceder y procesar los datos de manera eficiente utilizando herramientas como BigQuery y Dataflow.

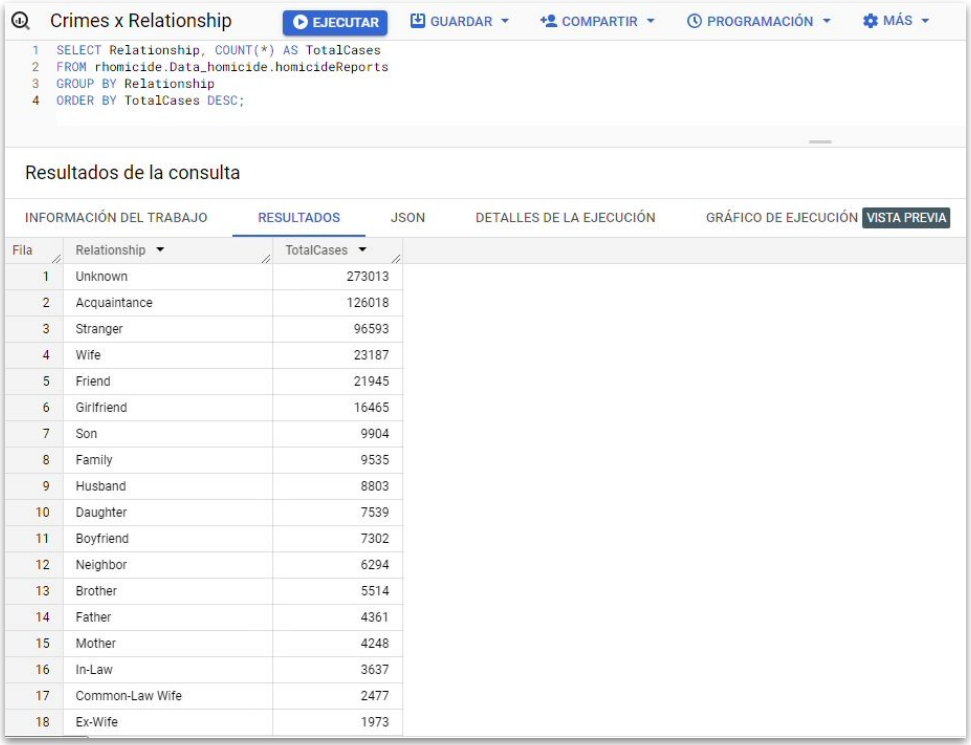


Análisis de datos en BigQuery:

En esta etapa, utilizamos BigQuery, una plataforma de análisis de datos de Google Cloud, para realizar varias consultas en el conjunto de datos. Algunas de las consultas que realizamos incluyen:

- Promedio de edad de las víctimas: Calculamos el promedio de edad de las víctimas en los informes de homicidio.
- Tasa de resolución de crímenes por ciudad: Determinamos la tasa de resolución de crímenes por ciudad.
- Crímenes por mes y año: Analizamos la cantidad de crímenes registrados por mes y año.
- Crímenes por mes: Exploramos la distribución de crímenes a lo largo de los meses.
- Crímenes por relación: Investigamos la relación entre víctimas y perpetradores en los crímenes.
- Crímenes por arma: Analizamos los tipos de armas utilizadas en los crímenes.
- Las 5 ciudades con más crímenes: Identificamos las ciudades con la mayor cantidad de crímenes.
- Total de crímenes por estado: Calculamos el número total de crímenes por estado.
- Agrupación por sexo de la víctima y tipo de crimen: Realizamos un análisis comparativo entre el sexo de la víctima y el tipo de crimen.
- Total de límites: Calculamos el número total de límites (registros) en el conjunto de datos.
- Total por raza de la víctima: Analizamos la cantidad de crímenes según la raza de la víctima.

Análisis de datos en BigQuery:

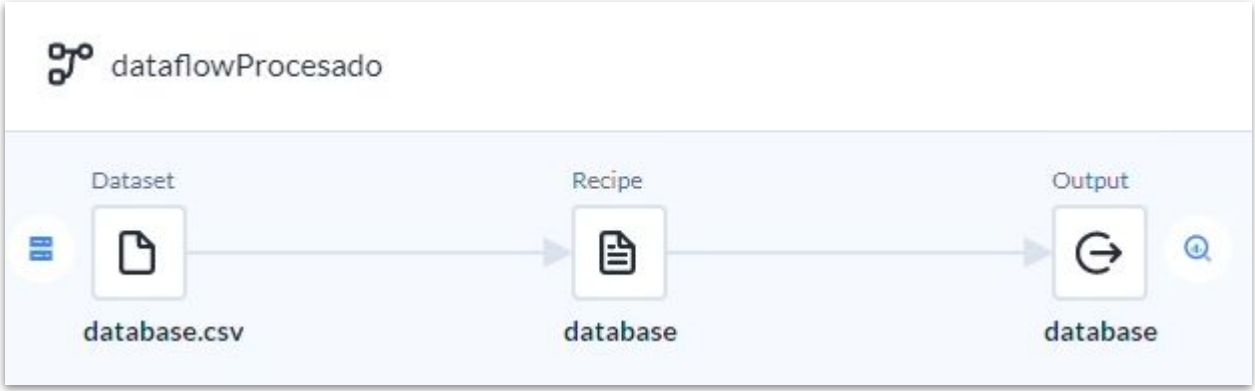


Fila	Relationship	TotalCases
1	Unknown	273013
2	Acquaintance	126018
3	Stranger	96593
4	Wife	23187
5	Friend	21945
6	Girlfriend	16465
7	Son	9904
8	Family	9535
9	Husband	8803
10	Daughter	7539
11	Boyfriend	7302
12	Neighbor	6294
13	Brother	5514
14	Father	4361
15	Mother	4248
16	In-Law	3637
17	Common-Law Wife	2477
18	Ex-Wife	1973

Fila	Weapon	TotalCrimes
1	Knife	94962
2	Blunt Object	67337
3	Handgun	317484
4	Shotgun	30722
5	Rifle	23347
6	Unknown	33192
7	Firearm	46980
8	Drowning	1204
9	Fire	6173
10	Suffocation	3968
11	Strangulation	8110
12	Gun	2206
13	Fall	190
14	Poison	454
15	Drugs	1588
16	Explosives	537

Limpieza de datos en Dataprep:

Después de realizar las consultas, utilizamos Dataflow, el servicio de procesamiento de datos de Google Cloud, para llevar a cabo la limpieza de datos. Implementamos pipelines de Dataflow para filtrar y transformar los datos según nuestras necesidades. Esto incluye la eliminación de datos incompletos o irrelevantes, la normalización de formatos y cualquier otra transformación necesaria para obtener datos limpios y coherentes.



Limpieza de datos en Dataprep:

database

Edit recipe | Branch recipe

Recipe | Data

Steps Preview

1. **Break into rows** using '\n' as a delimiter
2. **Split** column1 into 24 columns on ','
3. **Set** all columns to IF(STARTSWITH(TRIM($col), '\"', false) && ENDSWITH(TRIM($col), '\"', false), SUBSTITUTE(TRIM($col), `{start}"|"{end}`, '', false), $col)
4. **Replace** '\"\"' in all columns with '\"'
5. Convert row 1 to **header**
6. **Replace mismatched** values in State with 'Rhode Island'
7. **Lock** Victim Sex **type** to String
8. **Lock** Perpetrator Sex **type** to String
9. **Rename** 24 columns by replacing ' ' with '_'
10. **Rename** 24 columns to lowercase
11. **Delete** agency_code

Steps	11
Updated	Last Friday at 4:03 PM
Created	Last Friday at 3:04 PM

agency_type	city	state
7 Categories	1,385 Categories	51 Categories
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Anchorage	Alaska
Municipal Police	Juneau	Alaska
Municipal Police	Nome	Alaska
Municipal Police	Bethel	Alaska
County Police	North Slope	Alaska
County Police	North Slope	Alaska
County Police	North Slope	Alaska
County Police	North Slope	Alaska
County Police	North Slope	Alaska
Municipal Police	Kenai Peninsula	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska
State Police	Juneau	Alaska

database

Run

Jobs (3) | Manual settings | Scheduled settings

Latest job

Job 19468307 • Completed
DANIEL . SANTIBANEZ MONDACA • Finished Last Friday at 4:06 PM

record_id	agency_name	agency_type	city
445282	Ada	Municipal Police	Pontotoc
538764	Ada	Sheriff	Ada
538763	Ada	Sheriff	Ada
620200	Ada	Municipal Police	Pontotoc
511799	Ada	Municipal Police	Pontotoc
125812	Bay	Sheriff	Bay
375462	Bay	Sheriff	Bay
145749	Bay	Sheriff	Bay
125813	Bay	Sheriff	Bay
088772	Bay	Sheriff	Bay

View on BigQuery | View details

The preview above shows the current data in the job destination. It might not reflect the output from this particular job run.

Previous jobs

Job 19468067 • Completed
DANIEL . SANTIBANEZ MONDACA • Finished Last Friday at 3:45 PM

Job 19467995 • Canceled
DANIEL . SANTIBANEZ MONDACA • Canceled Last Friday at 3:28 PM

Análisis de datos limpios en BigQuery:

group by x victim sex and c...ype

EJECUTAR GUARDAR

```
SELECT Victim_Sex, Crime_Type, COUNT(*) AS Total
FROM `rhomicide.Data_homicide.homicideReports`
GROUP BY Victim_Sex, Crime_Type;
```

Resultados de la consulta

INFORMACIÓN DEL TRABAJO | RESULTADOS | JSON | DETALLES DE LA EJEC

Fila	Victim_Sex	Crime_Type	Total
1	Female	Murder or Manslaughter	141097
2	Male	Murder or Manslaughter	487268
3	Male	Manslaughter by Negligence	6857
4	Female	Manslaughter by Negligence	2248
5	Unknown	Murder or Manslaughter	973
6	Unknown	Manslaughter by Negligence	11

total x victim race

EJECUTAR GUARDAR

```
SELECT
  Victim_Race,
  COUNT(Victim_Race) AS Total
FROM
  `rhomicide.Data_homicide.homicideReports`
GROUP BY
  Victim_Race;
```

Resultados de la consulta

INFORMACIÓN DEL TRABAJO | RESULTADOS | JSON

Fila	Victim_Race	Total
1	Black	299899
2	White	317422
3	Native American/Alaska Native	4567
4	Unknown	6676
5	Asian/Pacific Islander	9890

Top 5 cities with most crimes

EJECUTAR

```
SELECT City, COUNT(*) AS TotalCrimes
FROM rhomicide.Data_homicide.homicideReports
GROUP BY City
ORDER BY TotalCrimes DESC
LIMIT 5;
```

Resultados de la consulta

INFORMACIÓN DEL TRABAJO | RESULTADOS | JSON

Fila	City	TotalCrimes
1	Los Angeles	44511
2	New York	38431
3	Cook	22383
4	Wayne	19904
5	Harris	16331

Visualización de datos en Looker Studio:

Después de realizar el análisis de datos en BigQuery, utilizamos Looker Studio para visualizar los resultados. Looker es una plataforma de visualización de datos que nos permite crear paneles interactivos, gráficos y tablas para comunicar de manera efectiva los insights obtenidos. Creamos visualizaciones atractivas y personalizadas para representar los resultados de nuestras consultas, lo que nos permite explorar los datos de forma intuitiva y compartir los hallazgos con el equipo y los interesados.

Looker U.S. Homicide Reports, 1980-2014