

Lab 3.2: Lab Sspark y Lab 3.3 HIVE y SparkSQL, GESTIÓN DE DATOS VIA SQL

Objetivo

Objetivo: Desarrollar un Data Warehouse sencillo y eficaz empleando BigQuery como la herramienta principal para el análisis de datos dentro del entorno de Google Cloud Platform (GCP).

Aspectos Resueltos:

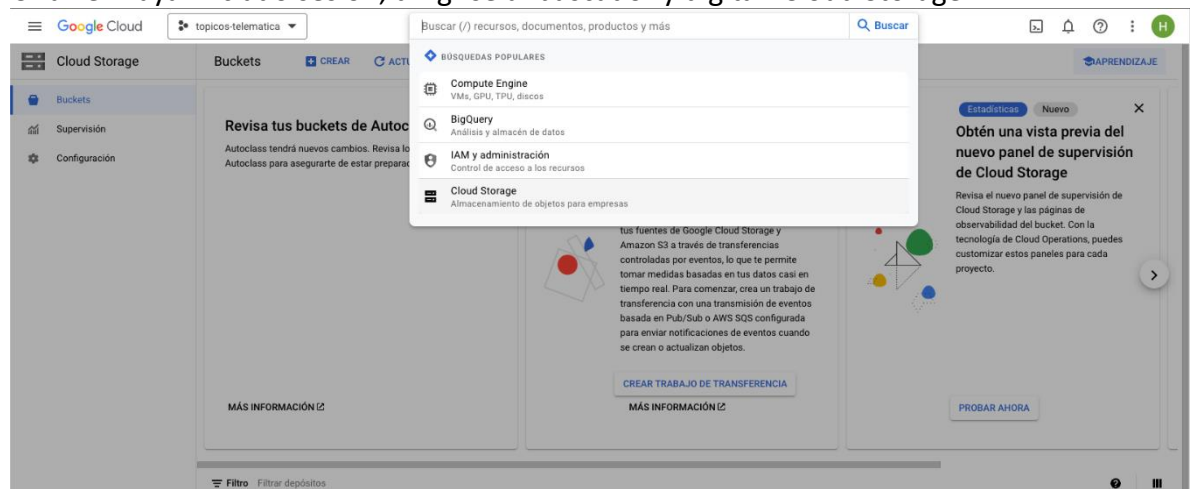
- Establecimiento de un Bucket en GCP Cloud Storage: Se logró la configuración de un bucket dentro del almacenamiento en la nube de GCP.
- Carga de Datos en el Bucket de Cloud Storage: Se completó con éxito la tarea de subir o montar los datos en el bucket previamente creado en Cloud Storage.
- Creación de un Conjunto de Datos y Tabla en BigQuery: Se llevó a cabo la formación de un conjunto de datos y una tabla en la plataforma de BigQuery.
- Determinación Automática del Esquema de Datos en Cloud Storage: Se realizó la auto-inferencia del esquema de los datos alojados en Cloud Storage.
- Ejecución de Consultas SQL en BigQuery sobre Datos de Cloud Storage: Se efectuaron consultas SQL mediante la plataforma BigQuery, enfocadas en los datos almacenados en Cloud Storage.

Ejecución

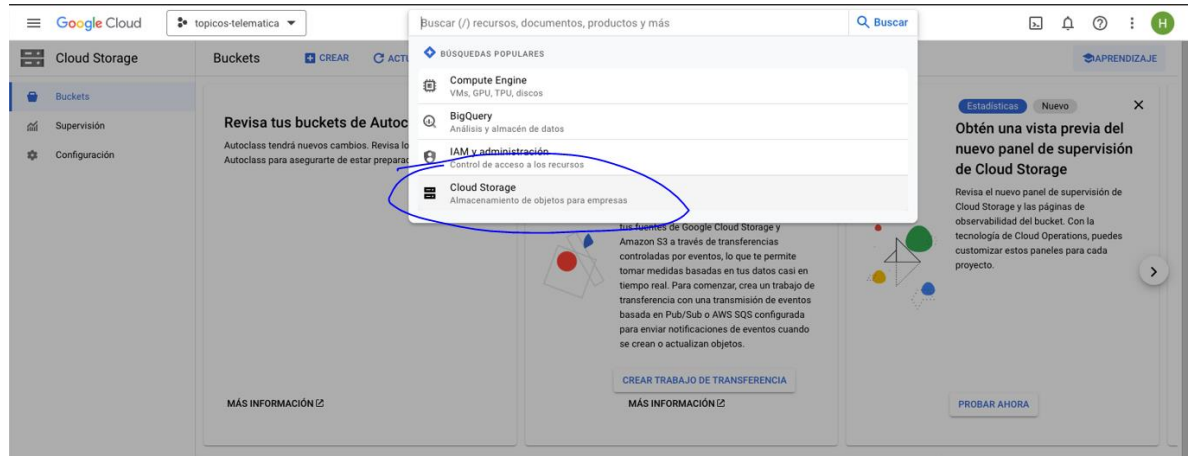
Guía

Primer Paso: Creación de un Bucket en GCP Cloud Storage

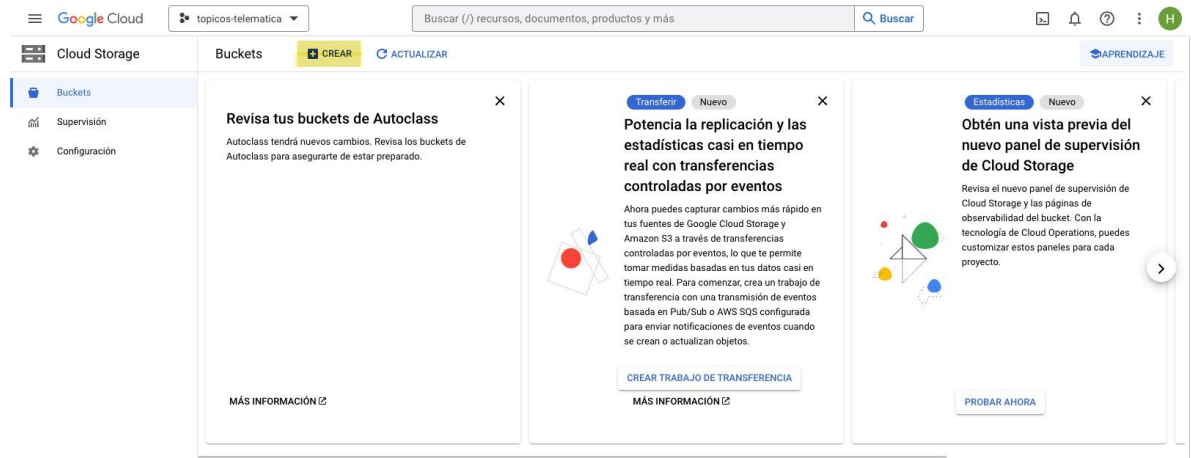
1. Iniciar sesión en GCP académico con su correo institucional y usar el código promocional que el profesor consiguió para usted.
2. Una vez haya iniciado sesión, dirigirse al buscador y digitar 'Cloud Storage'.



3. Luego de hacer click en 'Cloud Storage'.



4. Luego verá la pestaña 'Buckets' en el panel izquierdo. Dar click ahí y luego click en 'CREAR'.



5. A continuación, pondremos un nombre a nuestro Bucket que sea de nuestro agrado. En la parte donde nos pide seleccionar la ubicación de los datos, elegiremos la opción 'Región' y después nos decidiremos por la región 'us-central1 (Iowa)'. Para terminar, simplemente hacemos clic en el botón CREAR que se encuentra abajo.

Google Cloud

topicos-telematica

Buscar (/) recursos, documentos, productos y más

Buscar

H

Cloud Storage

Buckets

Supervisión

Configuración

Marketplace

Notas de versión

Crear un bucket

Asigna un nombre a tu bucket

Selecciona un nombre permanente globalmente único. [Lineamientos para asignar nombre](#)

b_labs

Sugerencia: No incluyas información sensible

ETIQUETAS (OPCIONAL)

CONTINUAR

Elige dónde almacenar tus datos

Esta opción define la ubicación geográfica de tus datos y afecta el costo, el rendimiento y la disponibilidad. No se puede cambiar más adelante. [Más información](#)

Tipo de ubicación

☐ Multi-region

Máxima disponibilidad en el área más amplia

☐ Dual-region

Alta disponibilidad y baja latencia en 2 regiones

☒ Region

Latencia mínima dentro de una sola región

us-central1 (Iowa)

CONTINUAR

Elige una clase de almacenamiento para tus datos

Clase de almacenamiento predeterminada: Standard

Elige cómo controlar el acceso a los objetos

Prevención del acceso público: Activada

Control de acceso: Uniforme

Elige cómo proteger los datos de objeto

Herramientas de protección: Ninguno

Encriptación de datos: Administrada por Google

CREAR

CANCELAR

Información útil

Precios de ubicación

Las tarifas de almacenamiento varían según la clase de almacenamiento de los datos y la ubicación de los buckets. [Detalles de precios](#)

Configuración actual: Region / Standard

| Elemento | Costo |
|--------------------|-----------------------|
| us-central1 (Iowa) | \$0.020 por GB al mes |

ESTIMAR COSTO MENSUAL

Así tendremos nuestro bucket creado y podemos buscarlo en la lista de buckets

Google Cloud

topicos-telematica

Buscar (/) recursos, documentos, pro...

Buscar

Cloud Storage

Buckets

Supervisión

Configuración

Marketplace

Notas de versión

← Detalles del bucket

ACTUALIZAR

APRENDIZAJE

b_labs

Ubicación

us-central1 (Iowa)

Clase de almacenamiento

Standard

Acceso público

No público

Protección

Ninguno

OBJETOS

CONFIGURACIÓNPERMISOSPROTECCIÓN

CICLO DE VIDAOBSERVABLE

Depósitos > b_labs

SUBIR ARCHIVOS

SUBIR CARPETA

CREAR CARPETA

TRANSFERIR LOS DATOS

ADMINISTRAR CONSERVACIONES

DESCARGAR

BORRAR

Filtrar solo por prefijo de nombre

Filtro

Filtrar objetos y carpetas

Mostrar datos borrados

Nombre

Tamaño

Tipo

Fecha de creación

Clase de almacenamiento

Última modificación

Acceso público

No hay filas para mostrar

Tu bucket está listo. Simplemente agrega datos.

Suelta archivos y carpetas aquí o usa el botón de carga. Para trasladar muchos datos de otro bucket o proveedor de almacenamiento en la nube, crea un trabajo de transferencia.

Se creó el bucket b_labs

×

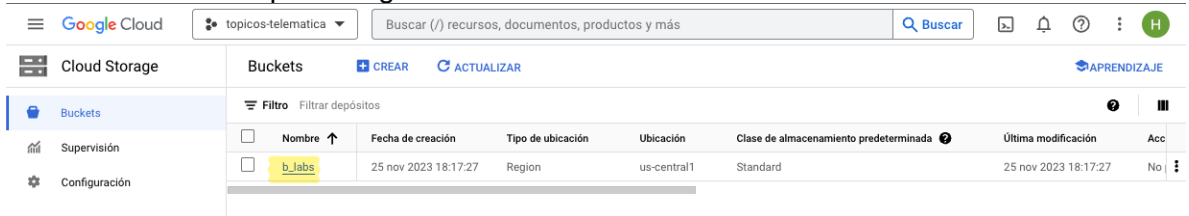
TRANSFERIR DATOS

DE TRANSFERENCIA

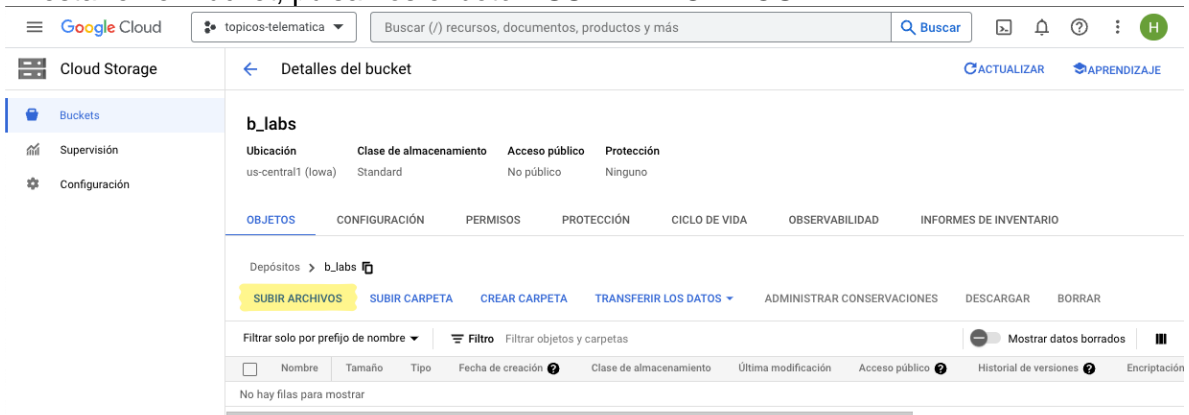
Segundo Paso: Subir/Montar datos al Bucket creado de Cloud Storage

Lo primero que necesitamos es descargar el dataset para ello podemos clonar el repo del profesor donde se encuentra el dataset <https://github.com/st0263eafit/st0263-232>

1. Abrimos el buscador de GCP, tecleamos 'Cloud Storage' y hacemos clic en el nombre del Bucket que configuramos anteriormente.



2. Al estar en el Bucket, pulsamos el botón 'SUBIR ARCHIVOS'.



3. Tras seleccionar 'SUBIR ARCHIVOS', elegimos los archivos del dataset previamente descargados y hacemos clic en 'Abrir'.

4. Una vez creado se verá así:

The screenshot shows the Google Cloud Storage interface for a bucket named 'b_labs'. The bucket is located in 'us-central1 (Iowa)' with 'Standard' storage class, 'No public' access, and 'None' protection. The 'OBJETOS' (Objects) tab is selected, displaying a list of six CSV files: 'online-retail-dataset_part0.csv' through 'online-retail-dataset_part5.csv'. Each file is 7.3 MB or 7.5 MB, created on Nov 25, 2023, and has 'Standard' storage class. A notification at the bottom states '6 archivos con carga exitosa.' (6 files uploaded successfully).

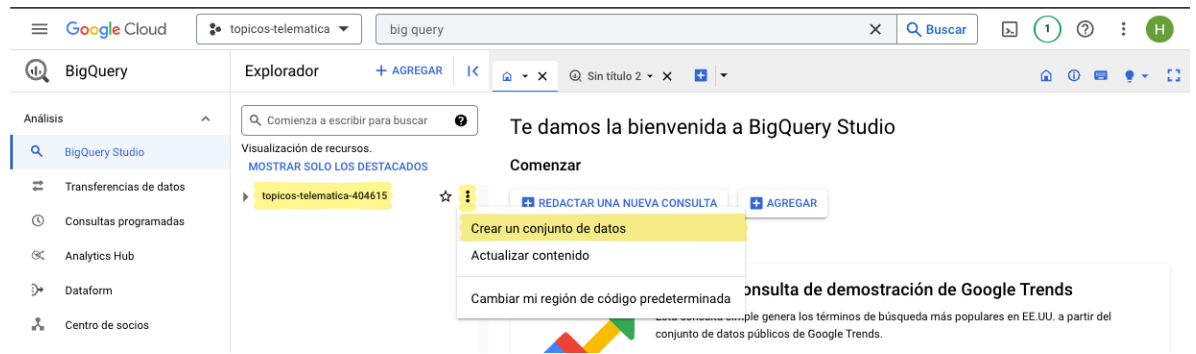
| Nombre | Tamaño | Tipo | Fecha de creación | Clase de almacenamiento | Última modificación | Acceso público |
|---------------------------------|--------|----------|----------------------|-------------------------|----------------------|----------------|
| online-retail-dataset_part0.csv | 7.3 MB | text/csv | 25 nov 2023 18:29:39 | Standard | 25 nov 2023 18:29:39 | No público |
| online-retail-dataset_part1.csv | 7.3 MB | text/csv | 25 nov 2023 18:29:39 | Standard | 25 nov 2023 18:29:39 | No público |
| online-retail-dataset_part2.csv | 7.3 MB | text/csv | 25 nov 2023 18:29:39 | Standard | 25 nov 2023 18:29:39 | No público |
| online-retail-dataset_part3.csv | 7.4 MB | text/csv | 25 nov 2023 18:29:32 | Standard | 25 nov 2023 18:29:32 | No público |
| online-retail-dataset_part4.csv | 7.5 MB | text/csv | 25 nov 2023 18:29:32 | Standard | 25 nov 2023 18:29:32 | No público |
| online-retail-dataset_part5.csv | 7.5 MB | text/csv | 25 nov 2023 18:29:32 | Standard | 25 nov 2023 18:29:32 | No público |

Tercer Paso: Configurar un Conjunto de Datos y una Tabla en BigQuery

1. Ingresamos en el buscador de GCP, escribimos 'Big Query' y elegimos la primera opción que aparece con el nombre de 'Big Query'.

The screenshot shows the Google Cloud search interface with 'BigQuery' entered in the search bar. The search results list several options: 'BigQuery' (Análisis y almacén de datos), 'Administración de capacidad BigQuery', 'BI Engine BigQuery', and 'Consultas programadas BigQuery'. The first option, 'BigQuery', is highlighted.

2. Una vez dentro, nos aseguramos de estar en la sección tópicos-telematica, que corresponde al nombre de nuestro proyecto.
3. Luego haremos click en los 3 puntos y seleccionamos la opción 'Crear un conjunto de datos'



4. Luego:

- Daremos nombre a nuestro conjunto de datos.
- Seleccionaremos la región.
- Escogemos la región: us-central (Iowa).
- Por último, daremos click en 'CREAR CONJUNTO DE DATOS'.

Crea un conjunto de datos

ID del proyecto

topicos-telematica-404615

[CAMBIO](#)

ID del conjunto de datos *

BQ_Dataset

Puede incluir letras, números y guiones bajos

Tipo de ubicación ?

☒ Región

Especifica una región para colocar tus conjuntos de datos con otros servicios de Google Cloud.

☐ Multirregional

Permite que BigQuery seleccione una región de un grupo para alcanzar límites de cuota más altos.

Región *

us-central1 (Iowa)

Vencimiento predeterminado de la tabla

☐ Habilitar el vencimiento de la tabla ?

Máxima antigüedad predeterminada de la tabla

Days

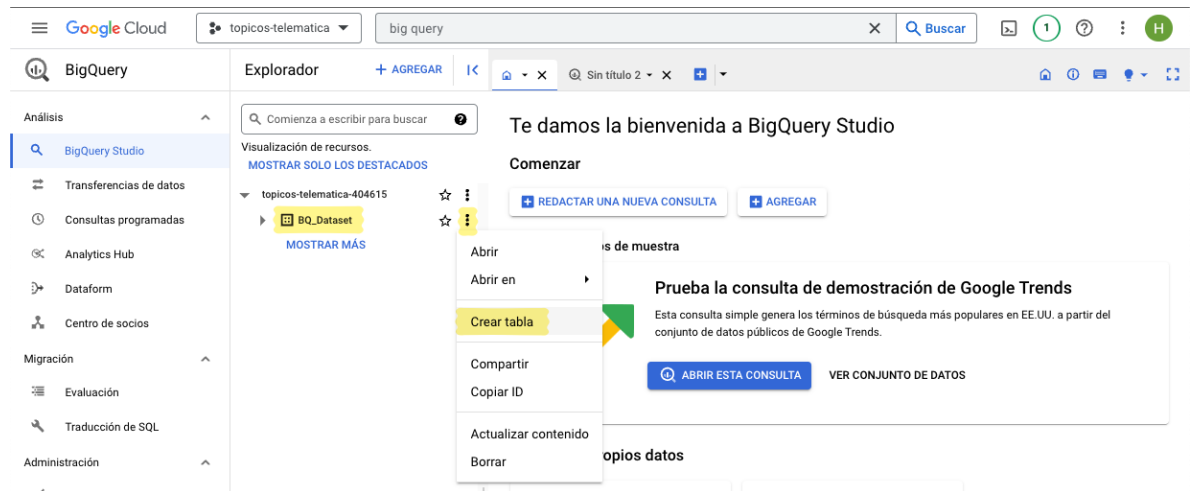
Opciones avanzadas



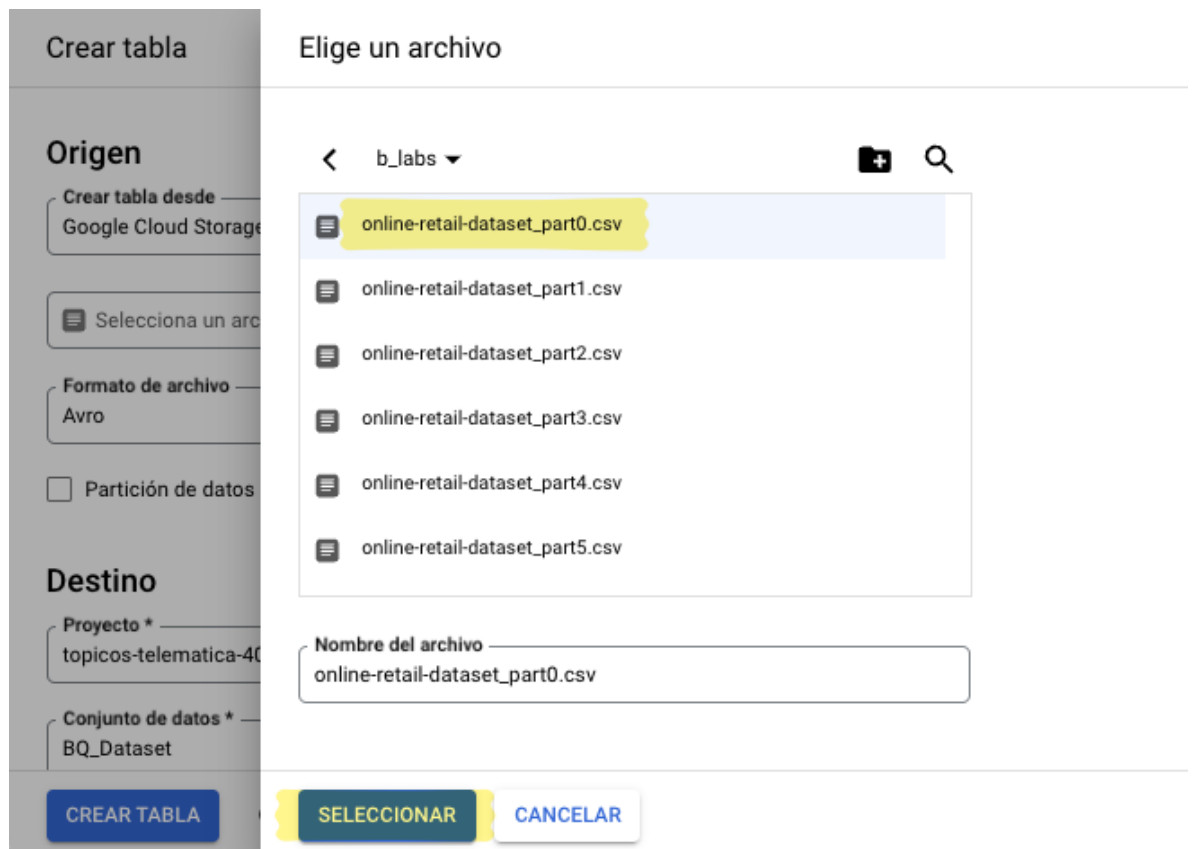
CREAR CONJUNTO DE DATOS

CANCELAR

- Posteriormente, en el panel izquierdo de GCP, observaremos que nuestro conjunto de datos ha sido creado. El siguiente paso es la creación de una tabla. Para ello, debemos hacer clic en los tres puntos y seleccionar la opción 'Crear tabla'.

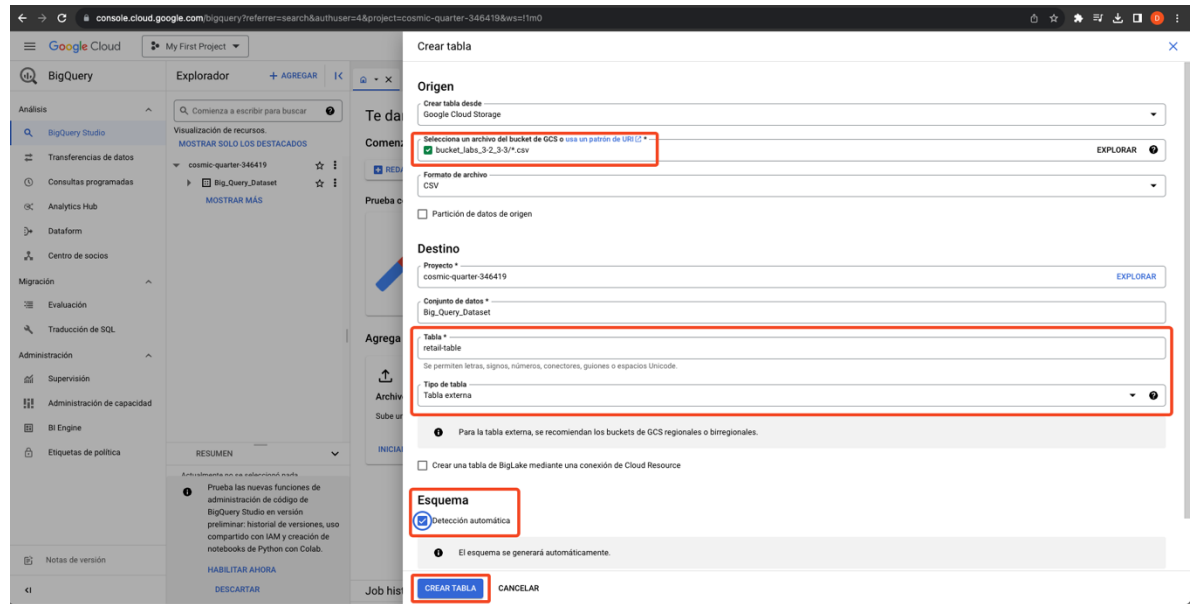


- En esta sección escogeremos la opción 'Google Cloud Storage' y luego haremos click en botón 'EXPLORAR'.
- Seleccionamos el archivo y luego hacemos click en 'SELECCIONAR'.



- En este paso debemos corregir el apartado 'Selecciona un archivo del bucket de GCS o usa un patrón de URI' borrando el nombre del archivo y reemplazándolo por `nombre_bucket/*.csv`
Luego seleccionamos la opción 'Detección automática'

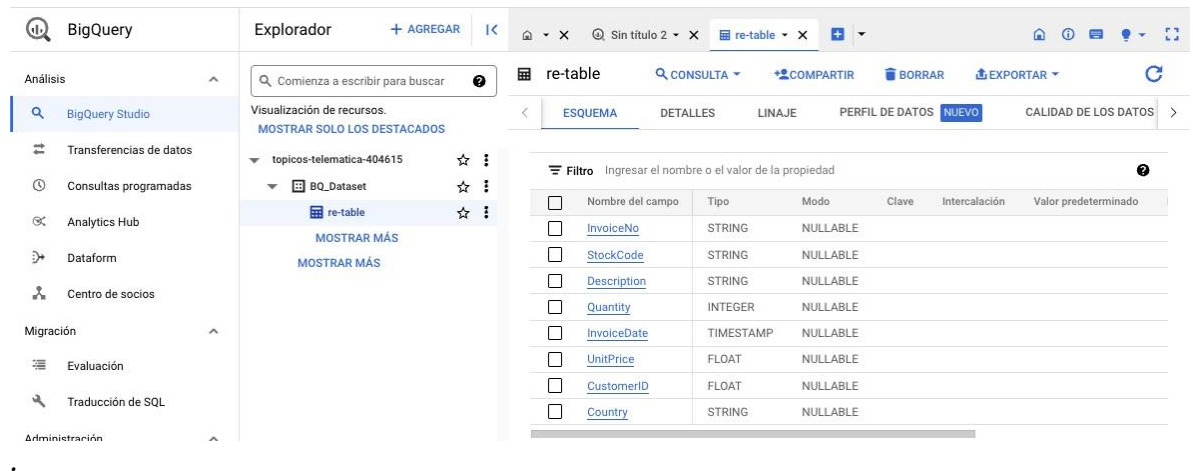
Por último, hacemos click en 'CREAR TABLA'.



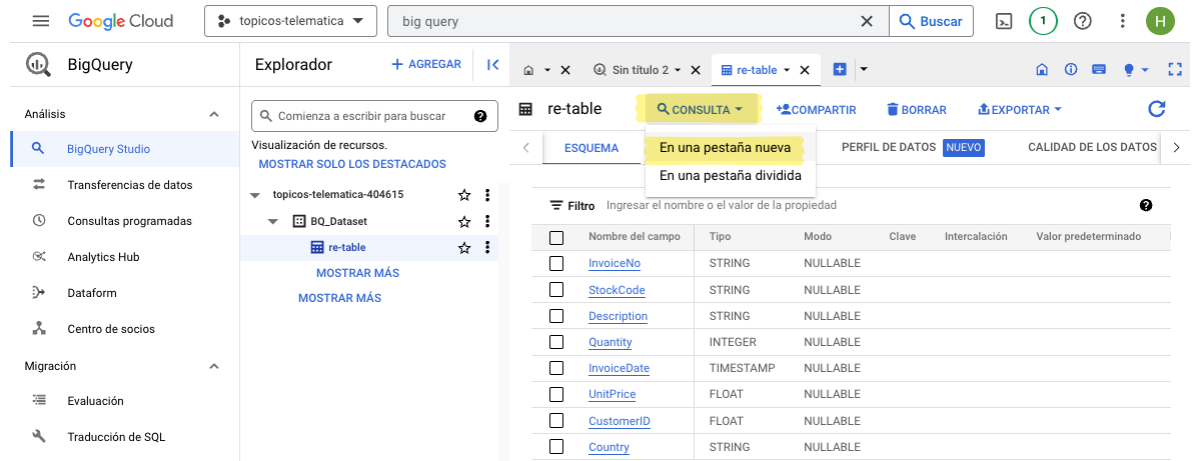
Cuarto y Quinto Paso: Determinación Automática del Esquema de los Datos en Cloud Storage y Realización de Consultas SQL en BigQuery

En estos pasos, primero procederemos con la auto-detección del esquema de los datos guardados en Cloud Storage. Luego, utilizaremos BigQuery para ejecutar consultas SQL sobre los datos almacenados en Cloud Storage. Nos dirigimos al buscador de GCP y digitamos 'Big Query' y seleccionamos la primera opción que es 'Big Query'.

- En la sección donde se encuentra el nombre del proyecto elegimos el conjunto de datos que creamos anteriormente y seguido de este el nombre de la tabla.



- Una vez allí, hacemos click en 'CONSULTA' y este desplegará 2 opciones, de las cuales escogeremos la que dice 'En una pestaña nueva'.



- Al abrir la nueva pestaña, veremos una interfaz similar a la mostrada en la imagen adjunta, donde procederemos a ejecutar la línea 1, destinada a realizar una consulta SQL a la tabla. Hay que tener en cuenta los siguientes pasos:
 - Primero, es necesario añadir el símbolo '*' al principio de la línea de código, ya que por defecto no se incluye. La estructura general de la consulta es: **SELECT * FROM 'nombre_proyecto.nombre_conjunto_de_datos.nombre_tabla'**
 - En mi caso, la consulta se formuló así: **SELECT * FROM 'cosmic-quarter-346419.Big_Query_Dataset.retail-table'**
 - Segundo, una vez realizada la corrección, hacemos clic en 'EJECUTAR'.
 - Tercero, tras ejecutar, los datos se visualizarán en la tabla.

Google Cloud | topics-telematica | big query

BigQuery | Explorador | + AGREGAR

Análisis | BigQuery Studio | Transferencias de datos | Consultas programadas | Analytics Hub

Sin título 3 | EJECTUAR | Syntax error: SELECT list must not be empty at [1:9]

```
1 SELECT FROM `topics-telematica-404615.BQ_Dataset.re-table` LIMIT 1000
```

Sin título 3 | EJECTUAR | Esta consulta procesará 0 B cuando se ejecute.

```
1 SELECT FROM `topics-telematica-404615.BQ_Dataset.re-table`
```

Sin título 3 | EJECTUAR | GUARDAR | DESCARGAR | COMPARTIR | PROGRAMACIÓN | MÁS | Se completó la co...

Resultados de la consulta | GUARDAR LOS RESULTADOS | EXPLORAR DATOS

INFORMACIÓN DEL TRABAJO | RESULTADOS | GRÁFICO | VISTA PREVIA | JSON | DETALLES DE LA EJECUCIÓN | GRÁFICO DE EJECUCIÓN

El almacenamiento en caché de metadatos está inhabilitado. Puedes acelerar las consultas en tablas externas habilitando el almacenamiento en caché de metadatos. [Más información](#)

DISMISS

| Fila | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
|------|-----------|-----------|------------------------------|----------|-------------------------|-----------|------------|----------------|
| 1 | 552332 | 21535 | RED RETROSPOT SMALL MILK ... | 6 | 2011-05-09 10:24:00 UTC | 2.55 | 13908.0 | United Kingdom |
| 2 | 552332 | 21035 | SET/2 RED RETROSPOT TEA T... | 6 | 2011-05-09 10:24:00 UTC | 3.25 | 13908.0 | United Kingdom |
| 3 | 552332 | 22077 | 6 RIBBONS RUSTIC CHARM | 12 | 2011-05-09 10:24:00 UTC | 1.65 | 13908.0 | United Kingdom |
| 4 | 552332 | 20717 | STRAWBERRY SHOPPER BAG | 10 | 2011-05-09 10:24:00 UTC | 1.25 | 13908.0 | United Kingdom |
| 5 | 552332 | 21746 | SMALL RED RETROSPOT WIND... | 12 | 2011-05-09 10:24:00 UTC | 1.25 | 13908.0 | United Kingdom |
| 6 | 552332 | 22300 | COFFEE MUG DOG + BALL DES... | 6 | 2011-05-09 10:24:00 UTC | 2.55 | 13908.0 | United Kingdom |
| 7 | 552332 | 22301 | COFFEE MUG CAT + BIRD DES... | 6 | 2011-05-09 10:24:00 UTC | 2.55 | 13908.0 | United Kingdom |
| 8 | 552333 | 22171 | 3 HOOK PHOTO SHELF ANTIQ... | 12 | 2011-05-09 10:38:00 UTC | 7.65 | 15039.0 | United Kingdom |

RESUMEN | Job history | ACTUALIZAR

Con lo anterior completamos los labs.