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:

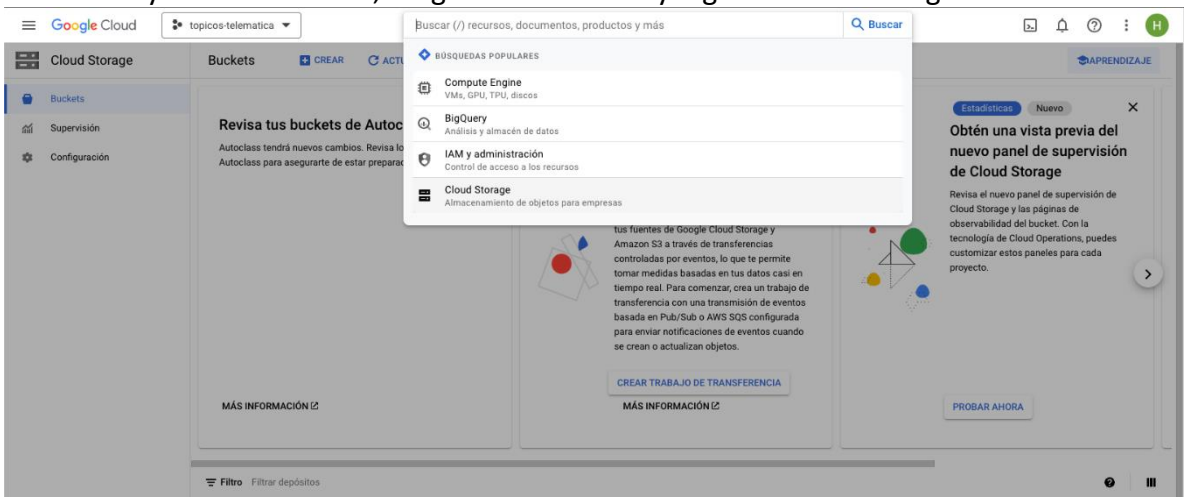
- Establecimiento de un Bucket en GCP Cloud Storage: Se logró la configuración de un bucket dentro del almacenamiento en la nube de GCP.
- Carga de Datos en el Bucket de Cloud Storage: Se completó con éxito la tarea de subir o montar los datos en el bucket previamente creado en Cloud Storage.
- Creación 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

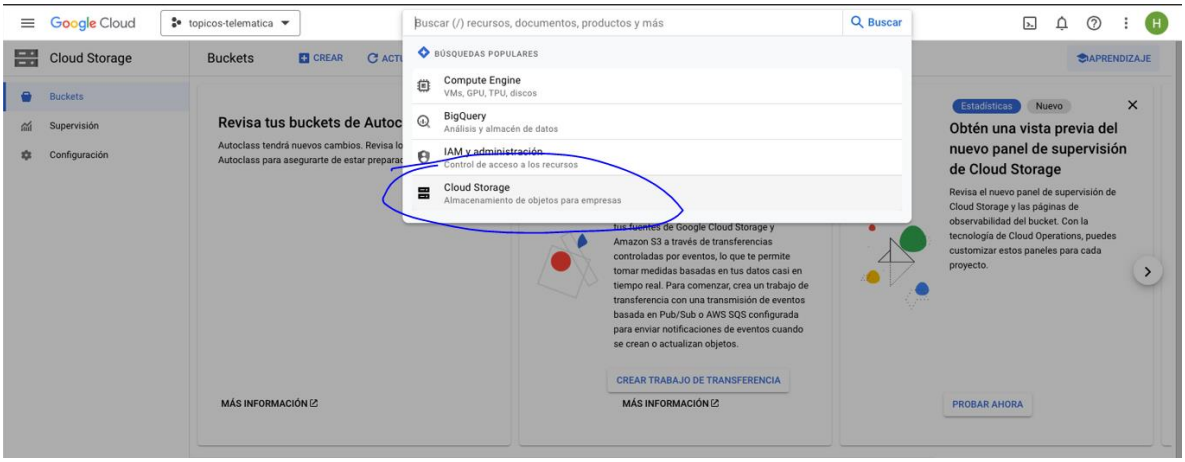
Guía

Primer Paso: Creación de un Bucket en GCP Cloud Storage

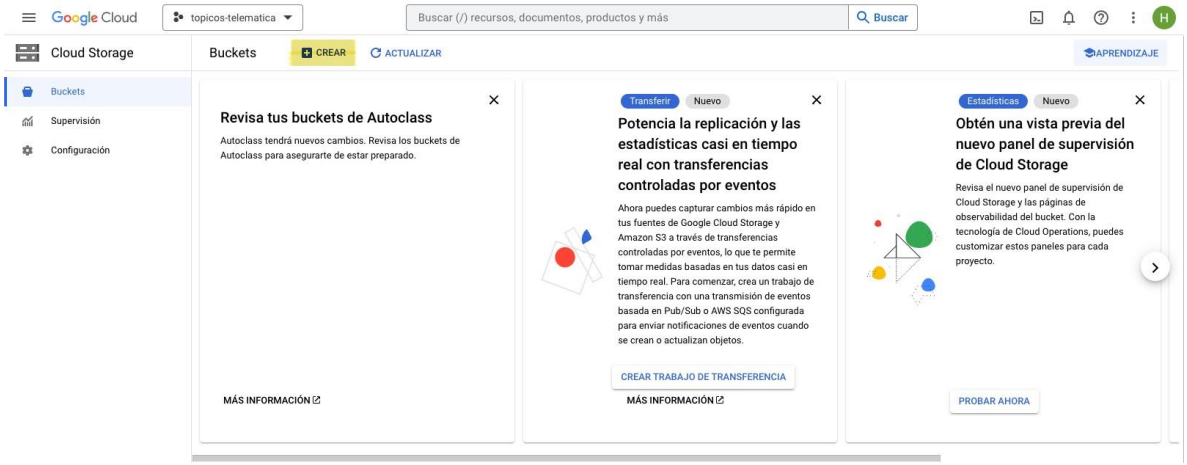
1. Iniciar sesión en GCP académico con su correo institucional y usar el código promocional que el profesor consiguió para usted.
2. Una vez haya iniciado sesión, dirigirse al buscador y digitar 'Cloud Storage'.



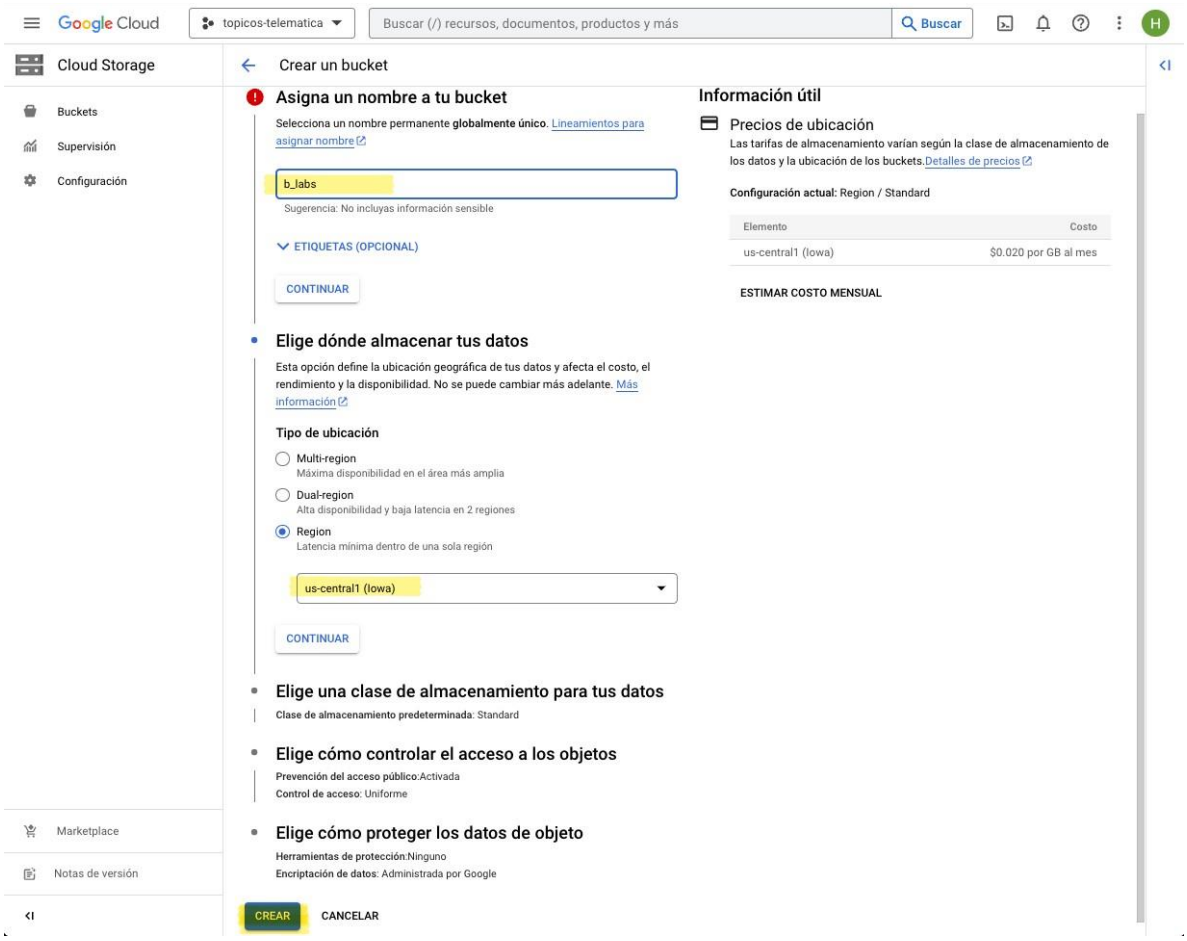
3. Luego de hacer click en 'Cloud Storage'.



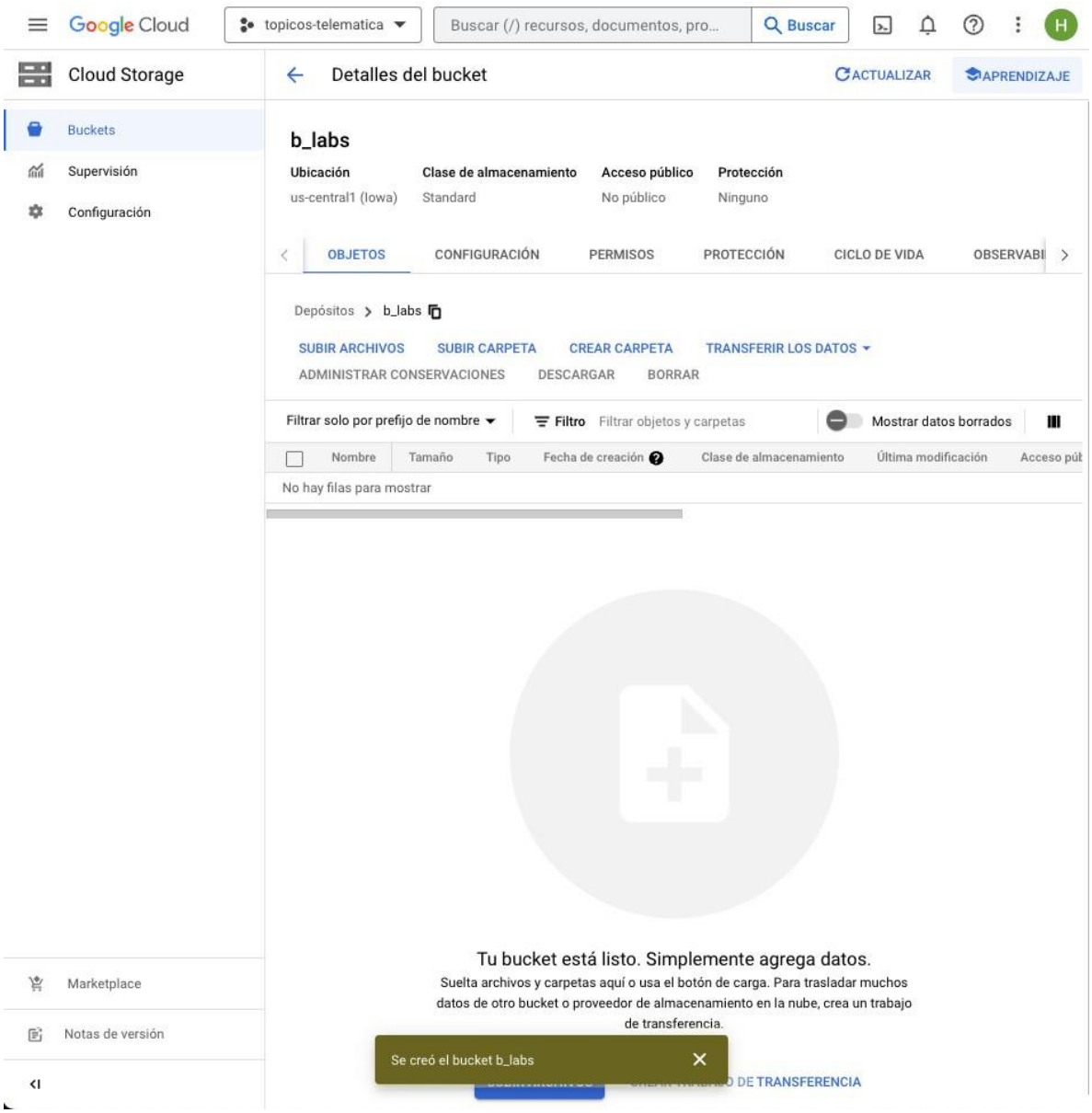
4. Luego verá la pestaña 'Buckets' en el panel izquierdo. Dar click ahí y luego click en 'CREAR'.



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.

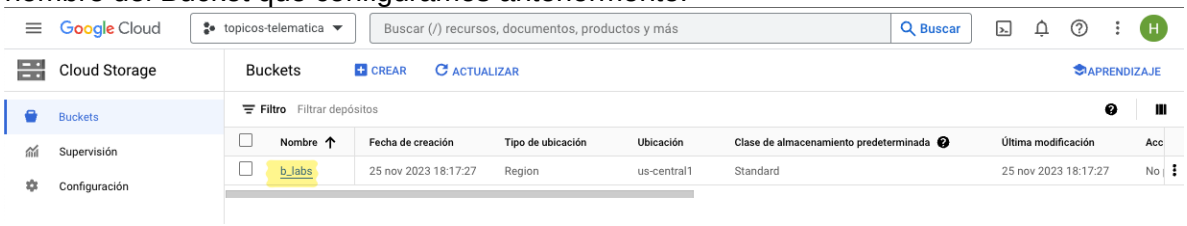


Así tendremos nuestro bucket creado y podemos buscarlo en la lista de buckets

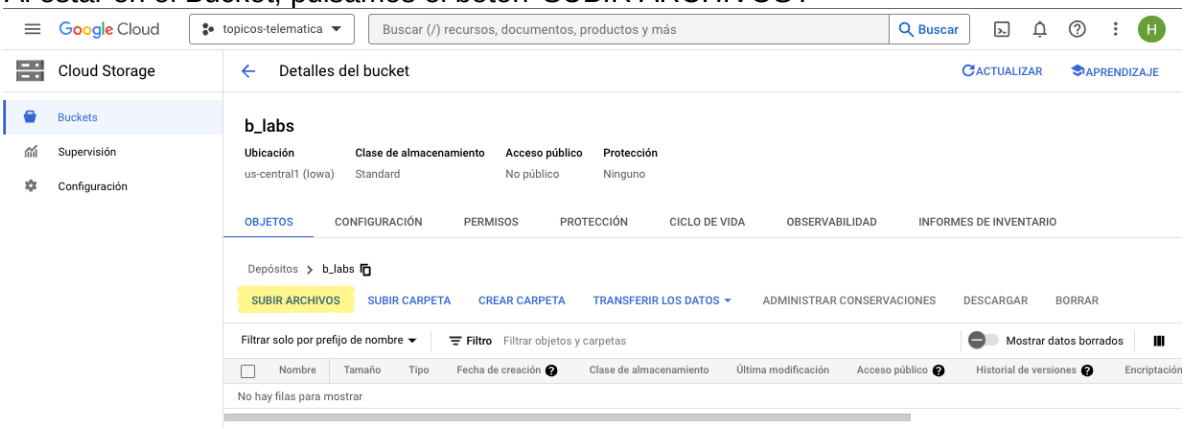

Google Cloud
topicos-telematica
Buscar (/) recursos, documentos, pro...
Buscar
Cloud Storage
Detalles del bucket
ACTUALIZAR
APRENDIZAJE
Buckets
Supervisión
Configuración
b_labs
Ubicación
us-central1 (Iowa)
Clase de almacenamiento
Standard
Acceso público
No público
Protección
Ninguno
OBJETOS
CONFIGURACIÓN
PERMISOS
PROTECCIÓN
CICLO DE VIDA
OBSERVABI
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 púb
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
TRANSFERENCIA
Marketplace
Notas de versión

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.

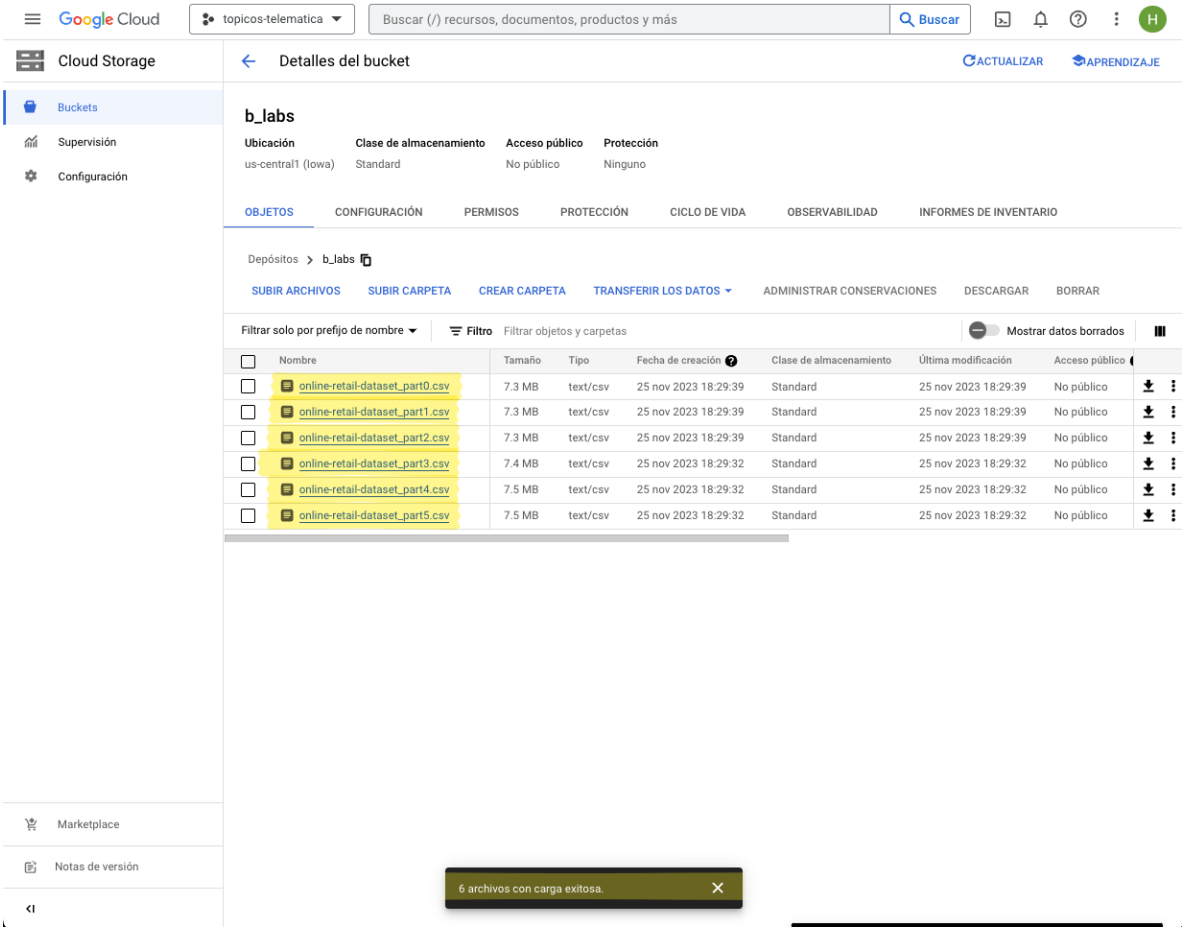


2. Al estar en el Bucket, pulsamos el botón 'SUBIR ARCHIVOS'.



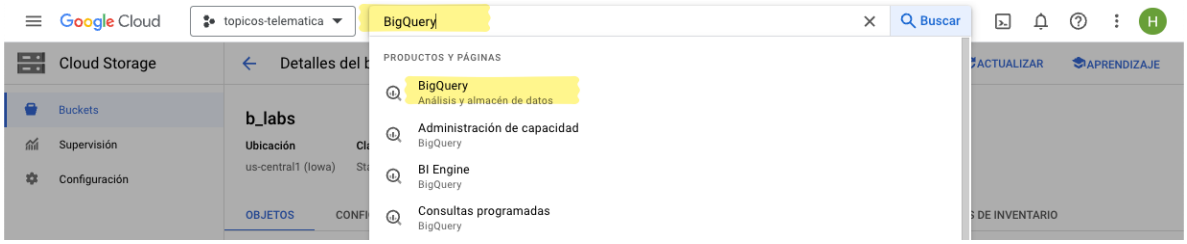
3. Tras seleccionar 'SUBIR ARCHIVOS', elegimos los archivos del dataset previamente descargados y hacemos clic en 'Abrir'.

4. Una vez creado se verá así:

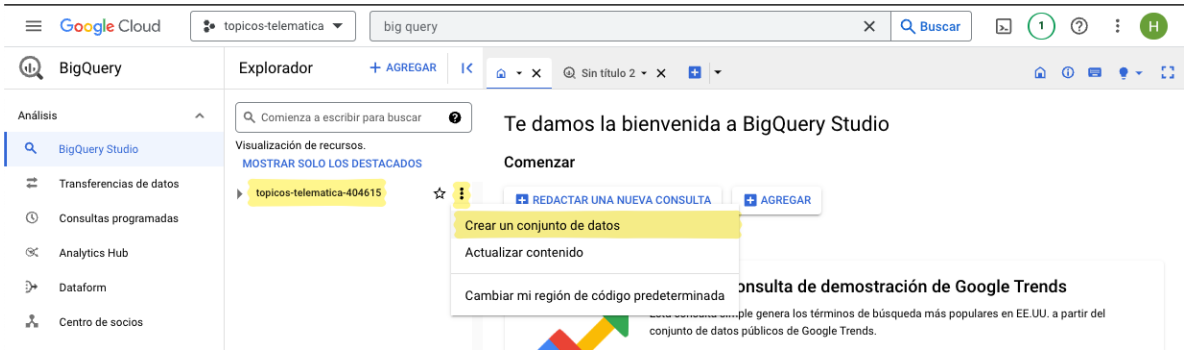


Tercer Paso: Configurar un Conjunto de Datos y una Tabla en BigQuery

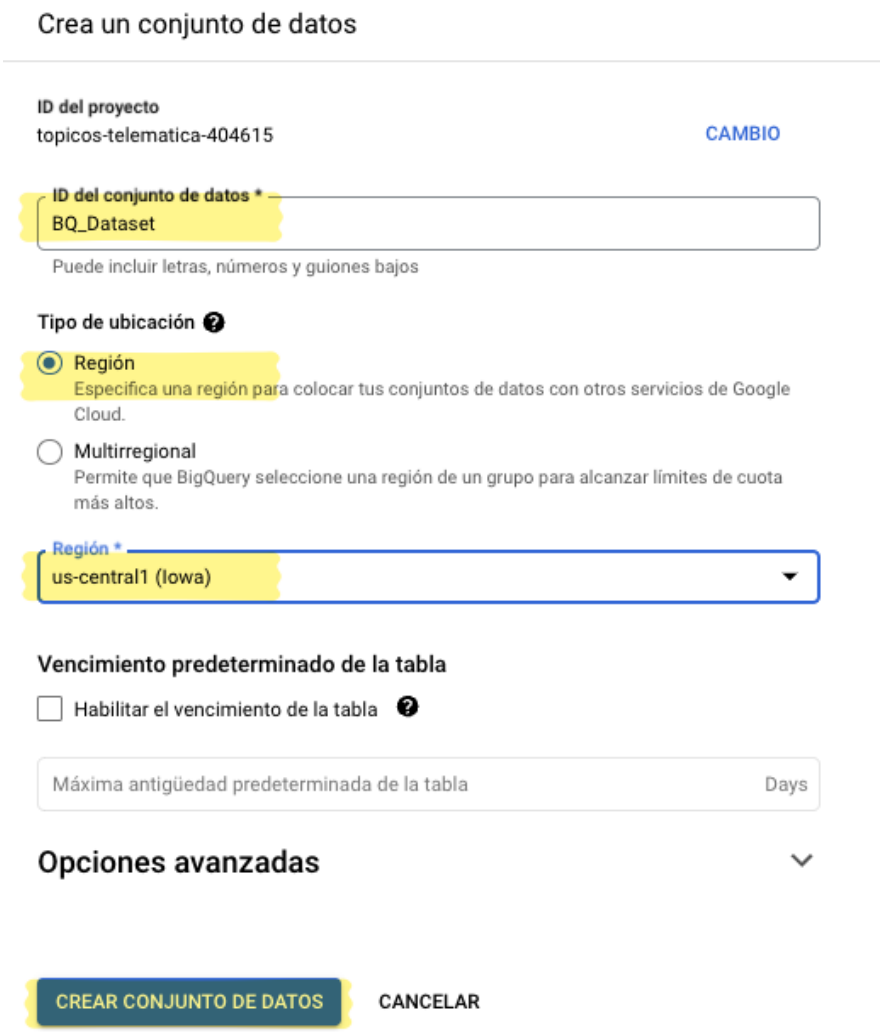
1. Ingresamos en el buscador de GCP, escribimos 'Big Query' y elegimos la primera opción que aparece con el nombre de 'Big Query'.



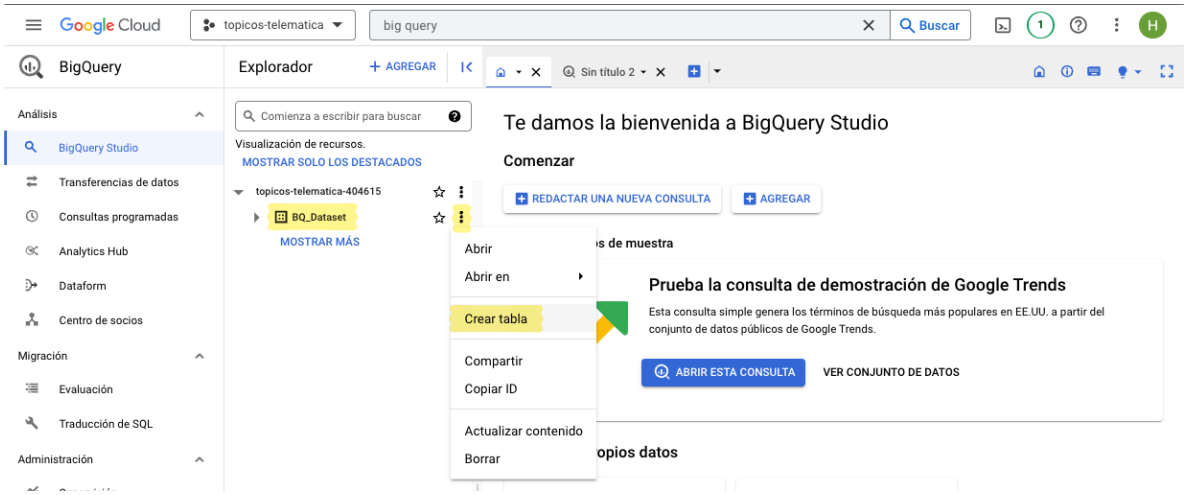
2. Una vez dentro, nos aseguramos de estar en la sección tópicos-telematica, que corresponde al nombre de nuestro proyecto.
3. Luego haremos click en los 3 puntos y seleccionamos la opción 'Crear un conjunto de datos'



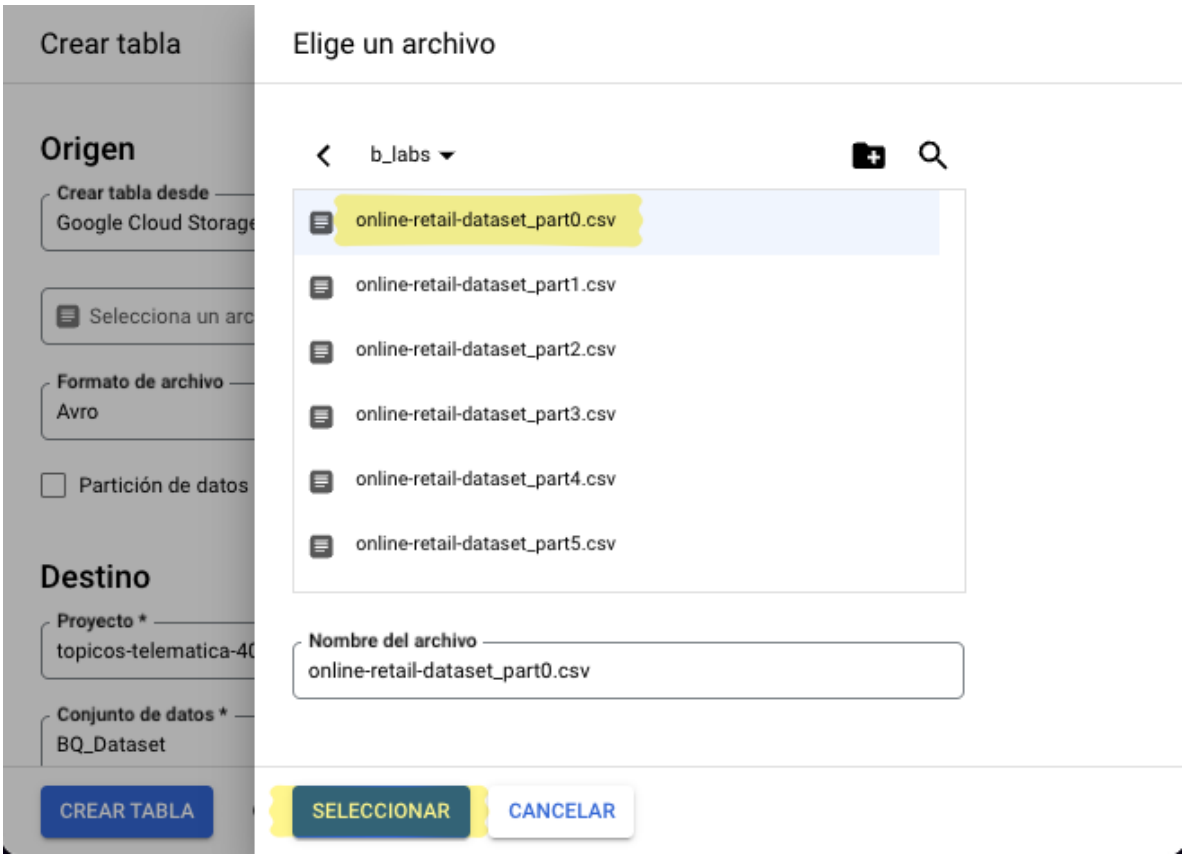
4. Luego:
   - Daremos nombre a nuestro conjunto de datos.
   - Seleccionaremos la región.
   - Escogemos la región: us-central (Iowa).
   - Por último, daremos click en 'CREAR CONJUNTO DE DATOS'.



5. 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'.

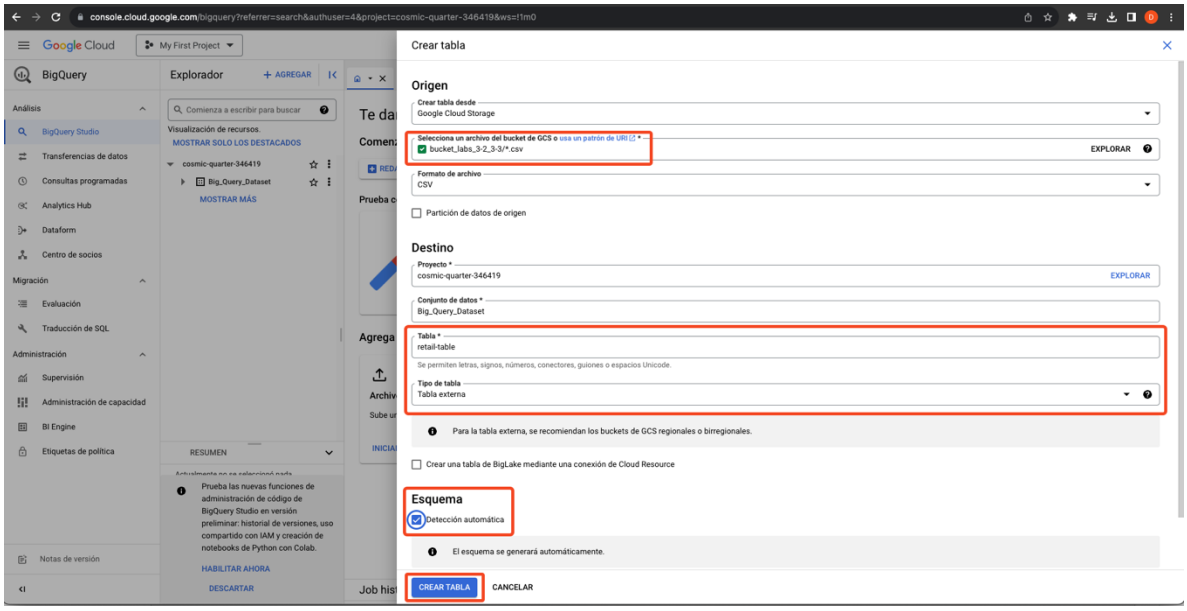


6. En esta sección escogeremos la opcion 'Goolge Cloud Storage' y luego haremos click en botón 'EXPLORAR'.
7. Seleccionamos el archivo y luego hacemos click en 'SELECCIONAR'.



8. 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'

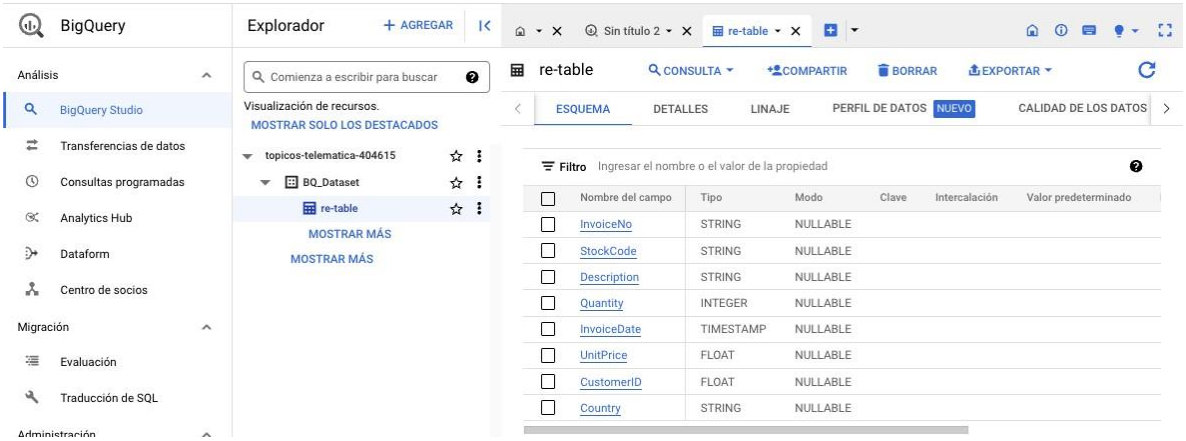
   Por último, hacemos click en 'CREAR TABLA'.



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

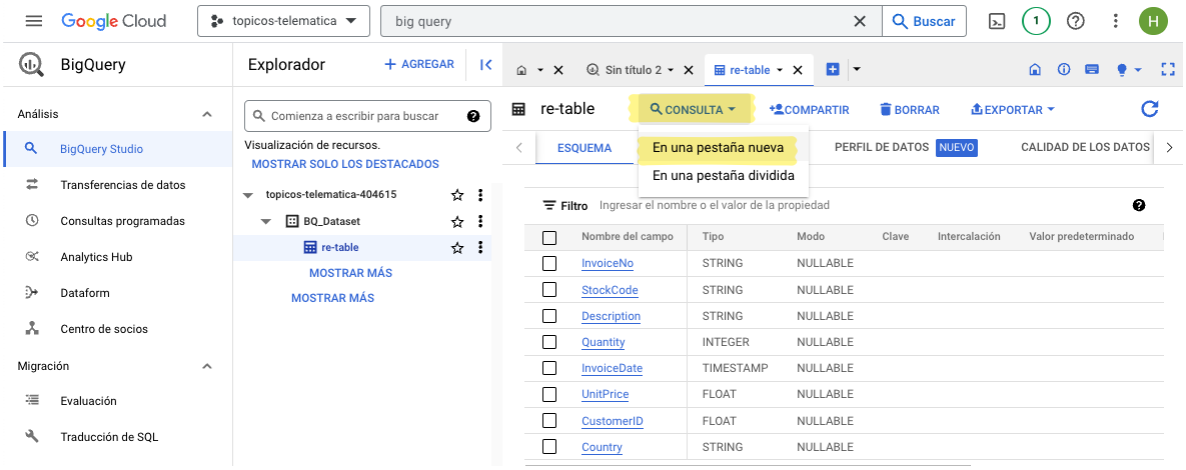
En estos pasos, primero procederemos con la auto-detección del esquema de los datos guardados en Cloud Storage. Luego, utilizaremos BigQuery para ejecutar consultas SQL sobre los datos almacenados en Cloud Storage.Nos dirigimos al buscador de GCP y digitamos 'Big Query' y seleccionamos la primera opcion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

1. 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.

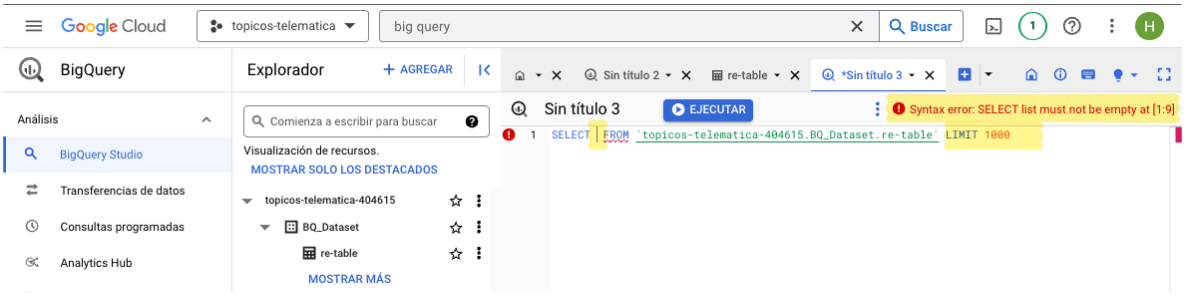


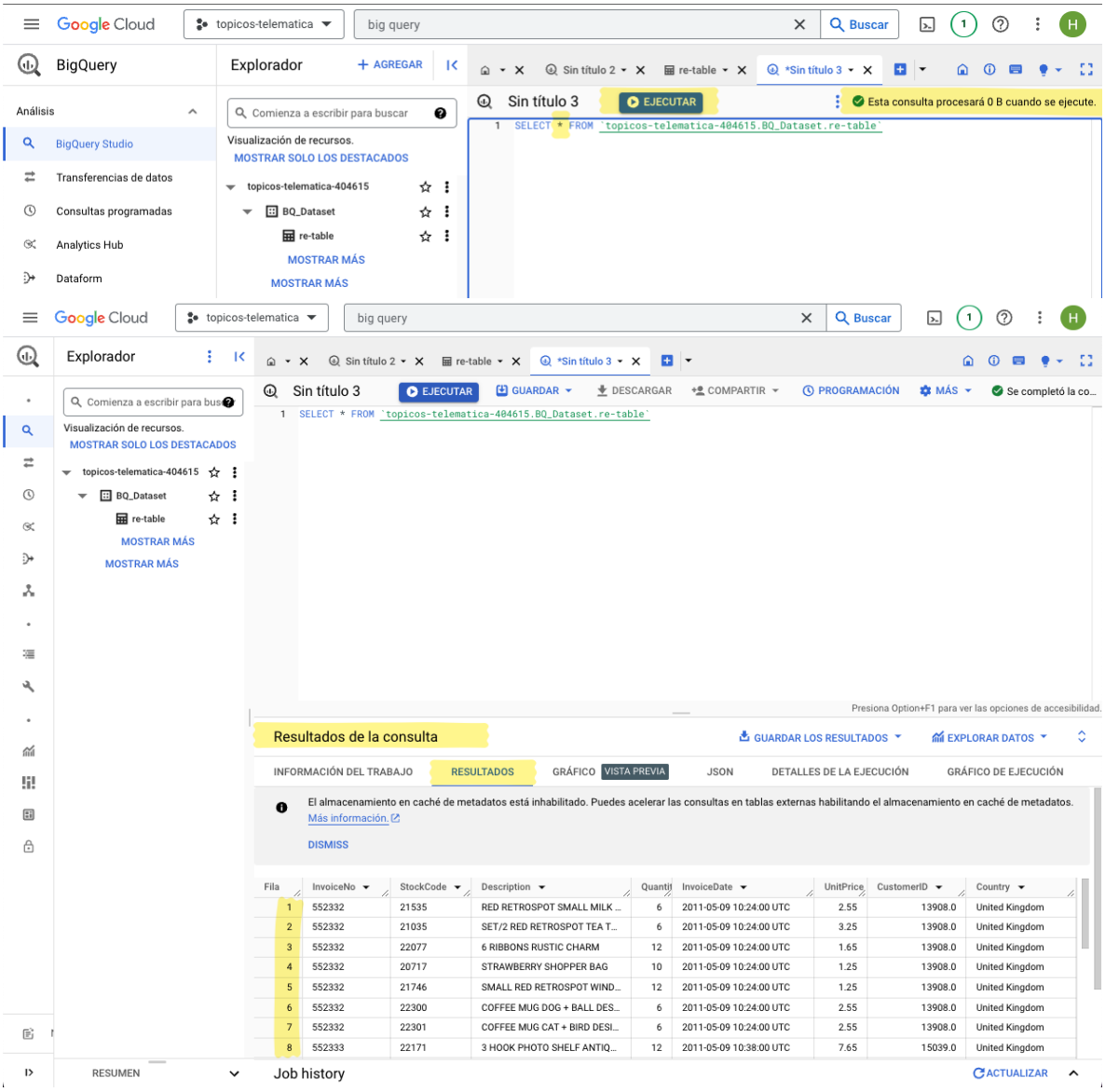
.

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



3. 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.





Con lo anterior completamos los labs.