Primeros pasos en BigQuery

KEEPCODING
Tech School

BigQuery

- Organizaciones
- Proyectos
- Conjuntos de datos
- Tablas:
  - Nativas de BQ.
    - Particionadas
    - Wildcard
    - Clusterizadas
  - Externas.
    - Ficheros alojados en Cloud Storage
    - Tabas Bigtable
    - Cloud SQL
    - Cloud Spanner
    - Google Drive
- Vistas: es una tabla virtual definida por una consulta SQL, con ella podemos restringir los campos a los que acceden los usuarios en una tabla.

- Funciones

BigQuery

Proyecto

En Google Cloud Platform, un "proyecto" es un espacio de trabajo lógico que te permite organizar y gestionar todos los recursos y servicios que utilizas en la nube. Es como una carpeta virtual donde alojas tus aplicaciones, datos y configuraciones.
Imagina que estás construyendo una aplicación en la nube usando servicios de Google, como almacenamiento, bases de datos, aprendizaje automático, etc. En lugar de tener todos estos servicios mezclados en un solo espacio, los agrupas en un proyecto. Este proyecto te proporciona un entorno aislado para tus recursos, lo que facilita la administración y la colaboración.
En resumen, un proyecto en Google Cloud Platform es una manera organizada y segura de trabajar con los servicios en la nube de Google, ofreciéndote un espacio centralizado para administrar y desarrollar tus aplicaciones y servicios.

BigQuery

Conjunto de datos (dataset)

En BigQuery, un "dataset" o "conjunto de datos" es básicamente un contenedor lógico que te permite organizar y estructurar tus tablas. Puedes pensar en un dataset como una carpeta que almacena conjuntos de tablas relacionadas entre sí.

BigQuery

Tablas particionadas

En BigQuery, las "tablas particionadas" son un tipo especial de tablas que están diseñadas para mejorar el rendimiento y reducir costos al trabajar con conjuntos de datos grandes. La partición se refiere a la forma en que se organizan físicamente los datos en la tabla.

La razón principal para tener tablas particionadas en BigQuery es facilitar la gestión y el análisis eficiente de grandes volúmenes de datos, especialmente cuando se trabaja con información temporal, como series temporales. Al particionar una tabla, se divide lógicamente en segmentos basados en una columna específica, llamada "columna de partición". Esta columna suele representar un valor de fecha o timestamp.

Algunas razones para utilizar tablas particionadas en BigQuery incluyen:

- Mejora del rendimiento: Las consultas que filtran los datos en función de la columna de partición pueden ser más eficientes, ya que el sistema solo necesita escanear las particiones relevantes en lugar de la tabla completa.
- Reducción de costos: Al limitar la cantidad de datos escaneados durante una consulta, se pueden reducir los costos de procesamiento, ya que solo se pagan los recursos utilizados en las particiones relevantes.
- Gestión temporal: Si tus datos tienen una dimensión temporal, como registros de eventos con marcas de tiempo, la partición por fecha puede ayudarte a realizar análisis específicos por períodos.

BigQuery

Tablas comodín o wildcard

En BigQuery, las tablas con "wildcard" (comodín) son una funcionalidad que te permite realizar consultas o realizar operaciones en múltiples tablas que comparten un patrón de nombres. Esto es útil cuando tienes conjuntos de datos que están distribuidos en varias tablas con nombres similares.
Cuando usas un wildcard en BigQuery, puedes especificar un patrón de nombres de tablas que deseas incluir en tu consulta. El comodín (*) se utiliza para representar cualquier cadena de caracteres en el nombre de la tabla. Esto facilita la ejecución de consultas en conjuntos de datos extensos o la realización de operaciones en tablas que siguen un patrón específico.

BigQuery

Tablas clusterizadas o agrupadas

En BigQuery, las tablas clusterizadas son tablas que se organizan físicamente en base a los valores de una o más columnas específicas. La clusterización de tablas puede mejorar el rendimiento de las consultas al agrupar datos similares físicamente juntos en el almacenamiento subyacente.
Cuando creas una tabla clusterizada en BigQuery, especificas una o más columnas como "columnas de clusterización". Las filas con valores similares en estas columnas se agrupan juntas en el almacenamiento, lo que puede acelerar las consultas que filtran o agrupan en función de esas columnas.

BigQuery

Índices agrupados o no agrupados

La diferencia principal entre índices agrupados y no agrupados se encuentra en cómo organizan físicamente los datos en una tabla en una base de datos relacional.

- Índice Agrupado: Las filas de datos están organizadas en el mismo orden que el índice (generalmente basado en la clave primaria).
- Índice No Agrupado: Las filas de datos y las entradas del índice se almacenan por separado, y las filas no están necesariamente organizadas según el índice.

BigQuery

Tablas Externas

En BigQuery se pueden crear tablas externas, es decir, tablas cuyos datos están alojados en otros productos del ecosistema Google. Por ejemplo:

- Ficheros (CSV, JSON, Avro, etc.) alojados en Google Cloud Storage (GCS)
- Tablas de BigTable
- Tablas de CloudSQL
- Tablas de Cloud Spanner
- Spreadsheet de Google Drive

BigQuery

Vistas

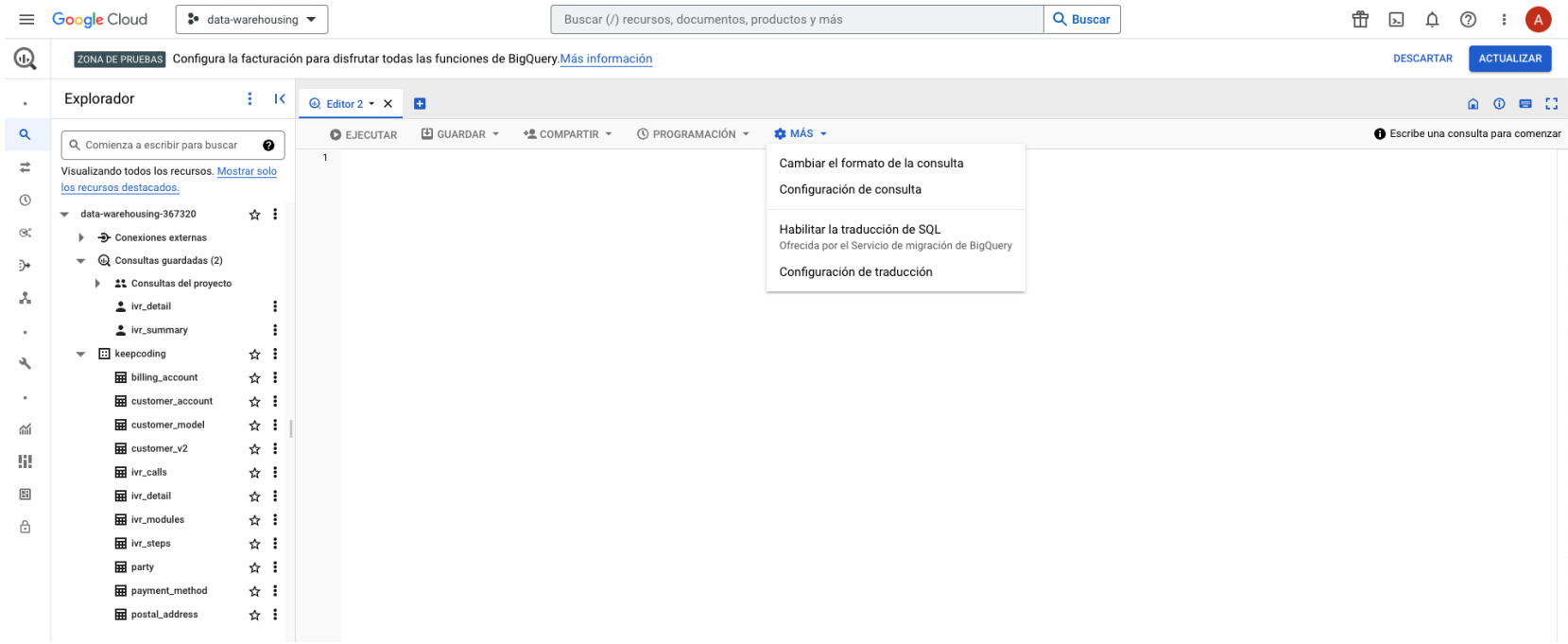
Una vista no almacena físicamente los datos; en cambio, proporciona una consulta predefinida que se ejecuta dinámicamente cada vez que se accede a la vista. Esto permite crear vistas personalizadas de los datos sin duplicar la información y facilita la gestión y el acceso a los datos de manera más eficiente.
Puedes utilizar vistas para controlar el acceso a los datos. Por ejemplo, puedes crear una vista que solo muestre ciertas columnas o filas a usuarios específicos, ocultando detalles innecesarios.

BigQuery

Funciones (UDF)

En BigQuery, las UDF (User-Defined Functions o Funciones Definidas por el Usuario) son funciones que puedes definir y utilizar dentro de tus consultas SQL. Estas funciones permiten extender la funcionalidad de BigQuery al permitirte escribir tu propia lógica personalizada y reutilizable. Se puede usar lenguaje SQL o Javascript.

UI de BigQuery



KEEPCODING

Tech School

Madrid | Barcelona | Bogotá

Datos de contacto