Big Query

Agenda

1. Introducción a SQL y BigQuery
2. El Flujo de Datos en BigQuery
3. Sintaxis y Consultas en SQL
4. Funciones Avanzadas en SQL
5. Buenas Prácticas y Documentación

Con qué puedo ayudar ?

Un poco sobre mi.

Hesus García: DevOps Engineer

Analista de datos 🚀 con 5 años de experiencia en consultoría IT y análisis de datos. Ayudo a interpretar datos y elevar su calidad para mejorar la toma de decisiones. ✅

GitHub LinkedIn Mi Sitio Web

Introducción a SQL y BigQuery

¿Qué es SQL y BigQuery?

SQL

SQL (Structured Query Language) es el lenguaje estándar para gestionar bases de datos relacionales. Fue desarrollado inicialmente por IBM y formalizado en los años 70 por Edgar F. Codd. SQL permite realizar operaciones como consultar, insertar, actualizar y eliminar datos.

BigQuery

BigQuery es una herramienta enterprise que permite almacenar y analizar grandes volúmenes de datos sin la necesidad de realizar configuraciones complejas de infraestructura. Es un servicio completamente gestionado por Google Cloud, lo que significa que no es necesario configurar infraestructura compleja. Está diseñado para ser escalable y seguro, proporcionando una plataforma ideal para el análisis de datos a gran escala.

Arquitectura de BigQuery

BigQuery tiene una arquitectura robusta que facilita el manejo de grandes volúmenes de datos:

- Ingestión: Permite extraer datos de diversas fuentes como Google Sheets, Analytics, Drive y otras fuentes. BigQuery se encarga de la integración y gestión de estos datos de manera eficiente.
- Almacenamiento: Los datos se almacenan en una base de datos relacional altamente escalable. BigQuery utiliza un formato de almacenamiento columnar optimizado para consultas rápidas.
- Extracción y Carga: BigQuery facilita la construcción y modificación de datos mediante SQL, permitiendo a los usuarios crear productos finales listos para su uso o venta.

Instalación de BigQuery

BigQuery es un servicio completamente gestionado por Google Cloud, lo que significa que no es necesario configurar infraestructura compleja. Está diseñado para ser escalable y seguro, proporcionando una plataforma ideal para el análisis de datos a gran escala.

¡Sigan Adelante!

Ahora, pongamos a prueba lo que hemos aprendido con un par de preguntas:

¿Qué significa SQL?

A) Extraer, Transformar, Listar

B) Entrar, Tocar, Leer

C) Extraer, Transformar, Cargar

D) Structured Query Language

Consola Cloud de Google Cloud Platform

GCP (Google Cloud Platform) es el servicio de nube empresarial de Google. Ofrece una variedad de servicios que permiten realizar procesos ingenieriles, desde la creación de bases de datos hasta la interconexión de servicios y despliegue de aplicaciones web y modelos de Machine Learning.

Servicios GCP

Google Cloud Platform ofrece numerosos servicios para satisfacer diversas necesidades de computación en la nube:

- **Compute Engine**: Máquinas virtuales escalables y personalizables.
- **App Engine**: Plataforma como servicio (PaaS) para aplicaciones web.
- **Cloud Storage**: Almacenamiento de objetos en la nube.
- **BigQuery**: Almacenamiento y análisis de datos a gran escala.
- **Cloud Functions**: Ejecución de código sin servidor.
- **Pub/Sub**: Mensajería en tiempo real para la transmisión de datos.

El panel de BigQuery

Dentro de la consola de GCP, BigQuery tiene su propio panel que facilita la administración de datos. Aquí se pueden realizar tareas como:

- Query History: Historial de consultas ejecutadas.
- Saved Queries: Consultas guardadas para uso futuro.
- Job History: Historial de trabajos y tareas ejecutadas.
- Crear Proyecto: Facilita la organización y gestión de datos a través de proyectos específicos.

Cómo crear un proyecto

Para crear un proyecto en BigQuery:

1. Navega a la consola de GCP.
2. Selecciona "Crear Proyecto".
3. Asigna un nombre y un ID al proyecto.
4. Configura detalles adicionales según sea necesario.

Los proyectos son fundamentales para organizar y gestionar los datos de manera eficiente.

Creación de nuevo proyecto

Para crear un nuevo proyecto:

1. Nombre del proyecto: Define un nombre significativo.
2. ID del proyecto: Asegúrate de que sea único.
3. Detalles del proyecto: Revisa y ajusta según sea necesario.

Cada proyecto en GCP actúa como un contenedor independiente para tus recursos y datos.

Dataset

Un dataset en BigQuery es un contenedor que organiza tablas y vistas. Actúa como una unidad raíz donde se agrupan los datos. Para agregar datos:

1. Haz clic en "Agregar datos".
2. Selecciona la fuente de datos deseada.
3. Sigue los pasos para importar datos a tu dataset.

¿Qué es una tabla?

Una tabla en BigQuery es una estructura que organiza datos en filas y columnas. Por ejemplo, la tabla `ga_sessions` contiene datos sobre sesiones de usuarios en Google Analytics.

- Campo: Columna en la tabla que contiene un tipo específico de datos.
- Tipo de datos: Define el formato de los datos (string, integer, float, etc.).
- Nullable: Indica si el campo puede tener valores nulos.

Metadatos

Los metadatos proporcionan información sobre la estructura y el contenido de la tabla, como la fecha de creación, el esquema de la tabla y los permisos. Son cruciales para la gestión y el entendimiento de los datos.

Importante: Quiz

La tabla `ga_sessions` muestra datos duplicados. Según las reglas de normalización, no deben existir registros duplicados en una tabla. ¿Está la tabla `ga_sessions` normalizada?

Reflexiona sobre la necesidad de tener datos siempre normalizados y discute las diferencias entre datos estructurados y semiestructurados.

Editor de Query

El editor de Query en BigQuery permite escribir y ejecutar consultas SQL. La estructura básica de una consulta incluye:

- SELECT: Especifica las columnas a seleccionar.
- FROM: Indica la tabla de origen.
- WHERE: Filtra los resultados según condiciones específicas.

SQL se ejecuta línea por línea, facilitando la manipulación y análisis de datos.

Sintaxis SQL

La sintaxis es el conjunto de reglas que define cómo deben estructurarse las consultas SQL. Es crucial para asegurar que las consultas se ejecuten correctamente y produzcan los resultados esperados.