

UNIVERSIDAD
COMPLUTENSE
DE MADRID



Google Cloud BigQuery

Autor: Pedro Pablo Malagón Amor

Actualizado Enero 2022

Datawarehouse empresarial en la nube

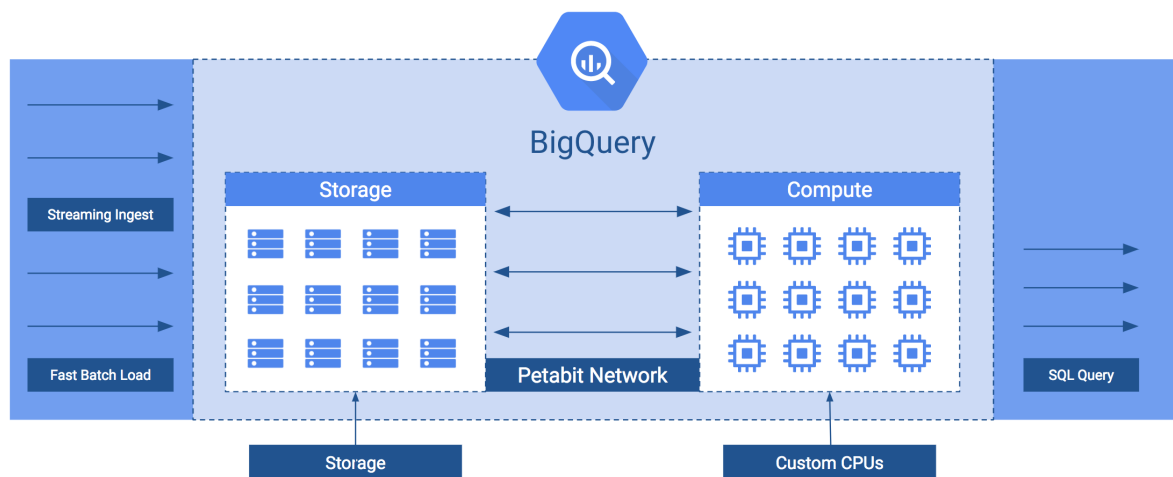
BigQuery es el servicio datawarehouse empresarial de Google de bajo coste, totalmente administrado y apto para analizar petabytes de datos. BigQuery no requiere servidor y como no hay que administrar ninguna infraestructura ni se necesita un administrador de bases de datos, cada uno puede centrarse en analizar los datos para obtener información importante mediante el conocido lenguaje SQL.

BigQuery permite consultas SQL súper rápidas utilizando la potencia de procesamiento de la infraestructura de Google

Cada uno simplemente tiene que mover sus datos a BigQuery o enlazarlo como tablas externas a BigQuery para poder lanzar consultas por ejemplo sobre cloud storage de nuestro dataprocc

El almacenamiento y consulta de grandes conjuntos de datos puede ser largo y costoso sin el hardware y la infraestructura adecuada. Google BigQuery es un datawarehouse que resuelve este problema permitiendo consultas SQL súper rápidas utilizando el poder de procesamiento de la infraestructura de Google. Simplemente enlaza tus datos a BigQuery y dejar a la nube de Google hacer el trabajo duro.

BigQuery se cimienta en la tecnología que utiliza Google



Cada uno podéis controlar el acceso tanto al proyecto como a sus datos en función de las necesidades de tu empresa, como por ejemplo, permitirle a otros ver o consultar sus datos.

Se puede acceder a BigQuery utilizando una interfaz de usuario web, una herramienta de línea de comandos, o haciendo llamadas a la API de REST de BigQuery utilizando una variedad de bibliotecas de clientes como Java, .NET o Python.

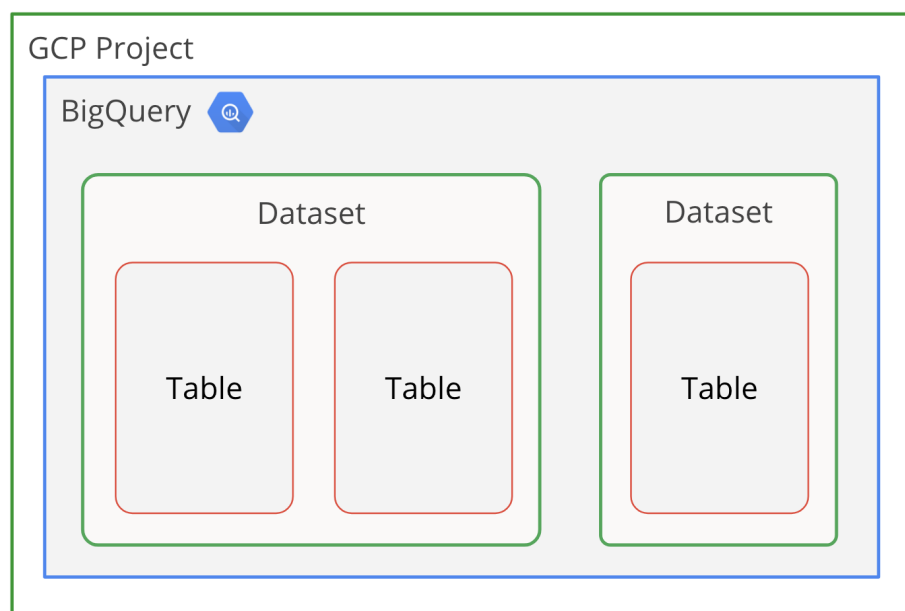
También hay una variedad de herramientas de terceros que puede usar para interactuar con BigQuery, así como visualizar los datos o cargar los datos.

BigQuery está completamente administrado, esto significa que no necesita desplegar ningún recurso, como discos duros y máquinas virtuales.

Proyectos

Los proyectos de Bigquery son una agrupación lógica como contenedor de alto nivel en Google Cloud Platform. Almacenan información sobre la facturación y los usuarios autorizados y que tienen datos en BigQuery. Cada proyecto tiene un nombre y un identificador único.

BigQuery Proyectos



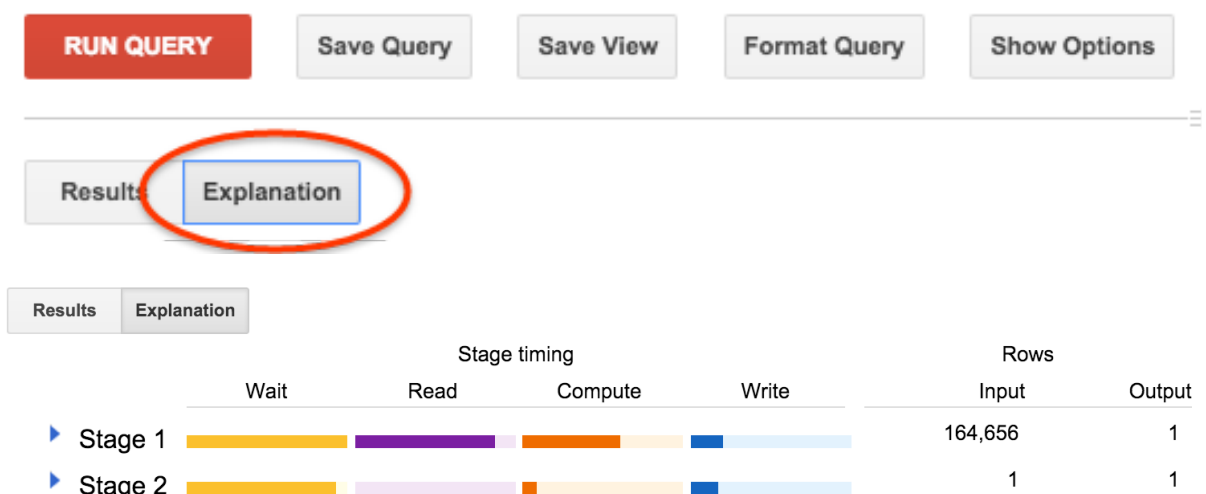
Velocidad y escala

BigQuery escanea terabytes en cuestión de segundos y petabytes en cuestión de minutos.

Se puede cargar los datos desde Google Cloud Storage, Google Cloud Datastore o transmitirlos a BigQuery para permitir analizarlos en tiempo real.

Por lo tanto, con BigQuery se puede escalar fácilmente tu base de datos de gigabytes a petabytes.

BigQuery maneja los aspectos técnicos del almacenamiento de los datos, incluida la compresión, el cifrado, la replicación, el ajuste del rendimiento y la escala. BigQuery almacena los datos en el formato de datos columnar Capacitor y ofrece los conceptos de base de datos estándar de tablas, particiones, columnas y filas.



Carga de Datos en BigQuery

Se puede cargar datos en el almacenamiento de BigQuery a través de:

- Cargas por lotes (batch)
- Transmisión (streaming)

Y puede realizar las siguientes operaciones de datos:


- Copiar tablas

- Consulta de tablas usando SQL
- Modificar datos usando SQL DML
- Exportar datos

Cargar Datos en BigQuery

En Bigquery se puede cargar datos de las siguientes fuentes:

- De Google Cloud Storage
- Desde una fuente de datos legible (como tu máquina local)
- Al insertar registros individuales usando inserciones de transmisión
- Usar declaraciones DML para realizar inserciones masivas
- Usar un pipeline de Google Cloud Dataflow para escribir datos en BigQuery


Load
gs://soydata/yob2015.txt to soydata-158417:SoyDataSet.nombres

Job ID	soydata-158417:bqijob_161ab0c4_15bc39498a0
Creation Time	May 1, 2017, 12:35:50 PM
Start Time	May 1, 2017, 12:35:51 PM
End Time	May 1, 2017, 12:35:54 PM
Destination Table	soydata-158417:SoyDataSet.nombres
Write Preference	Write if empty
Source Format	CSV
Source URI	gs://soydata/yob2015.txt (Open in GCS)
Autodetect Schema	true

Se puede compartir los datos almacenados con otros usuarios, mediante roles y permisos de Gestión de identidades y accesos (IAM) y controles de acceso al conjunto de datos.

BigQuery también admite la consulta de datos que no están en el almacenamiento de BigQuery. Una fuente de datos externa (también conocida como fuente de datos federada) es una fuente de datos que puede consultar directamente aunque los datos no estén almacenados en BigQuery. En lugar de cargar o transmitir los datos, crea una tabla que hace referencia a la fuente de datos externa.

BigQuery ofrece soporte para consultar datos directamente desde:

- Google Cloud Bigtable
- Google Cloud Storage

- Google Drive

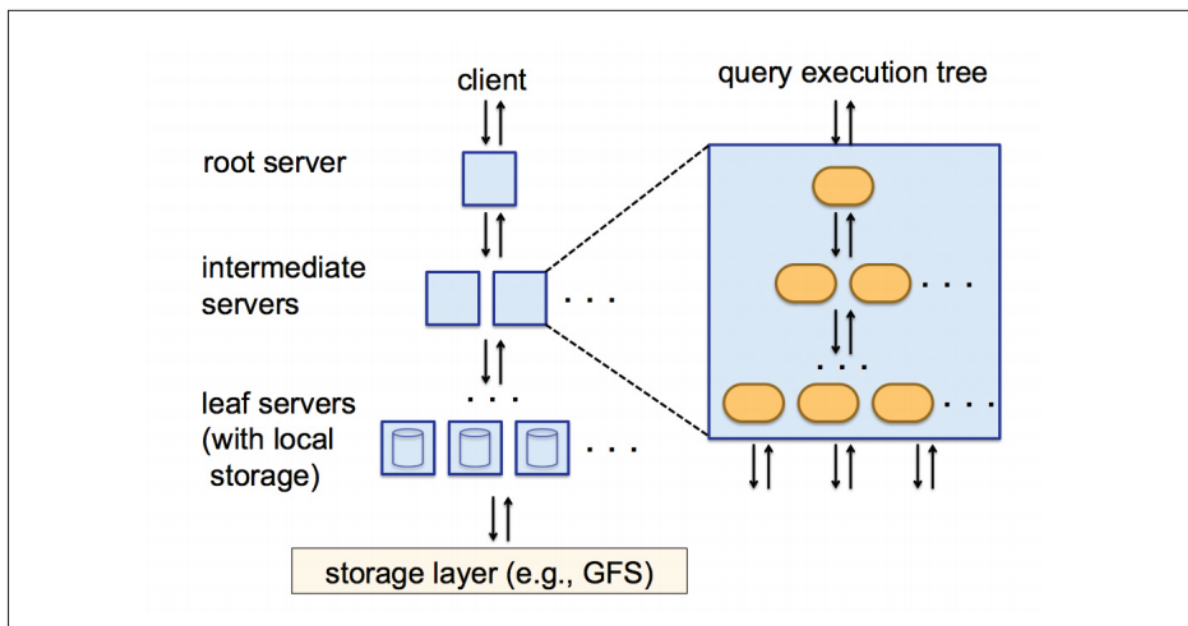
BigQuery Slots

Un Slot de BigQuery es una unidad de capacidad computacional necesaria para ejecutar consultas SQL. BigQuery calcula automáticamente cuántos Slots requiere cada consulta, dependiendo del tamaño de la consulta y la complejidad.

La mayoría de los clientes encuentran que la capacidad de Slots predeterminada es más que suficiente. Sin embargo, un grupo más grande de slots podría mejorar el rendimiento de consultas muy grandes o muy complejas, así como el rendimiento de cargas de trabajo altamente concurrentes.

Se puede decidir, establecer y verificar cuántos Slots quiere utilizar en su proyecto siempre que lo desee y Supervisar BigQuery usando Stackdriver.

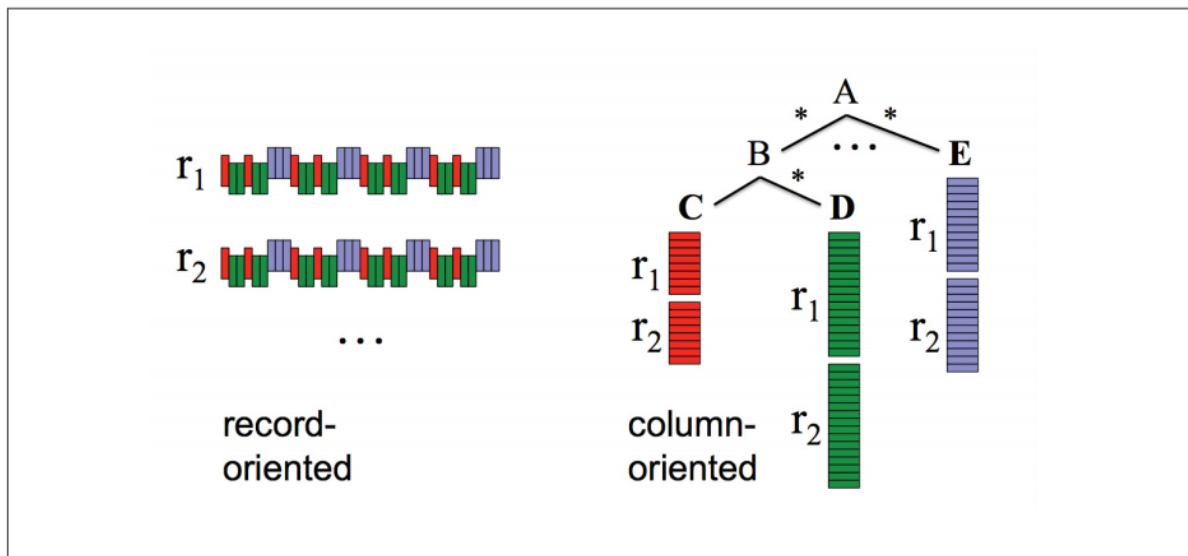
BigQuery gestiona automáticamente su cuota de slots basada en la historia de uso.



Tree architecture of Dremel

Tablas particionadas

BigQuery soporta el particionamiento de Tablas, Una tabla particionada es una tabla especial que se divide en segmentos, llamados particiones, que hacen que sea más fácil administrar y consultar sus datos. Al dividir una tabla grande en particiones más pequeñas, puede mejorar el rendimiento de la consulta y reducir el número de bytes que se facturan al restringir la cantidad de datos que se escanean. BigQuery ofrece tablas con fechas divididas, lo que significa que la tabla se divide en una partición separada para cada fecha.



Columnar storage of Dremel

Límites de tablas con particiones

- Cada tabla particionada puede tener hasta 2.500 particiones.
- Límite diario: 2,000 actualizaciones de partición por tabla, por día.
- Límite de velocidad: 50 actualizaciones de partición cada 10 segundos.

```
#legacySQL
/* Often performs better */
SELECT
  field1
FROM
  mydataset.table1
WHERE
  _PARTITIONTIME > DATE_ADD(TIMESTAMP('2016-04-15'), -5, "DAY")
```


Precios de almacenamiento a largo plazo

Cada partición de una tabla particionada se considera por separado para el precio de almacenamiento a largo plazo. Si una partición no se ha modificado en los últimos 90 días, los datos en esa partición se consideran almacenamiento a largo plazo y se cobran al precio con descuento.

Formatos de datos compatibles en BigQuery

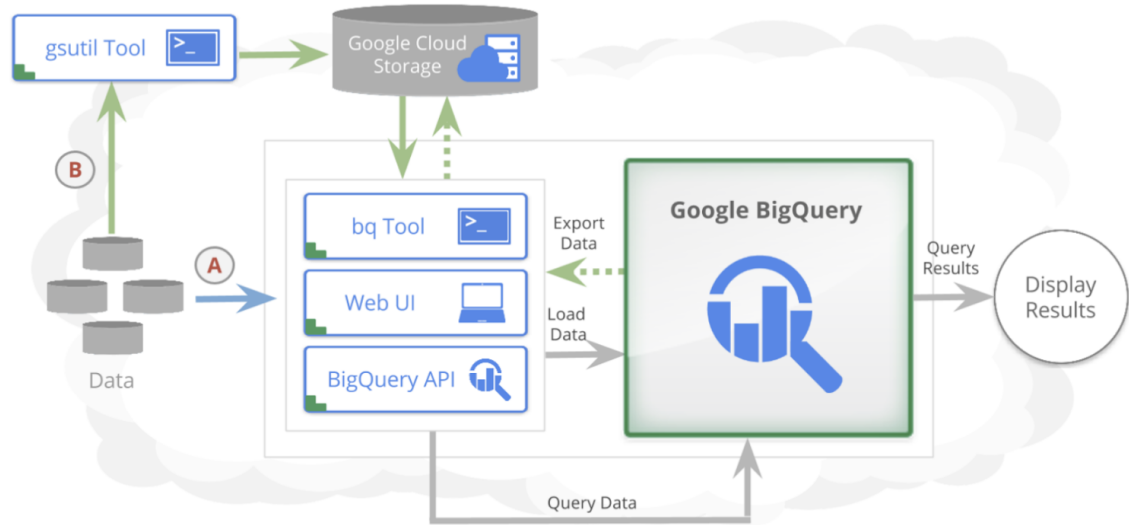
BigQuery admite cargar datos de Cloud Storage y fuentes de datos legibles en los siguientes formatos:

Google Cloud Storage:

- CSV
- JSON
- Avro
- Copias de seguridad de Google Cloud Datastore

Otras Fuentes de datos legible (como tu máquina local):

- CSV
- JSON
- Avro



Cargando datos codificados

BigQuery admite la codificación UTF-8 para datos anidados, repetidos y planos. BigQuery admite la codificación ISO-8859-1 para datos planos en archivos CSV.

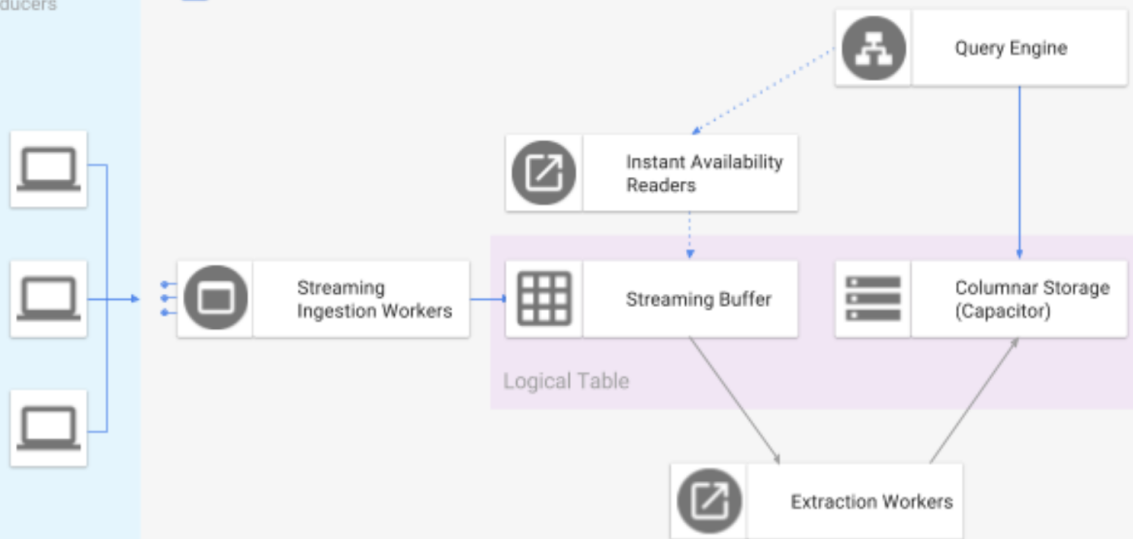
Ingestión flexible de datos

Se permite la carga de los datos desde Google Cloud Storage o Google Cloud Datastore o bien transmitirlos a BigQuery a 100.000 filas por segundo para analizarlos en tiempo real.

BigQuery Streaming Ingestion - Internals

Streaming Data
Producers

 **BigQuery Service**



Totalmente integrado

Además de las consultas SQL, se puede leer y escribir datos fácilmente en BigQuery a través de Cloud Dataflow, Spark y Hadoop.

Create Table

Source Data ☒ Create from source ☐ Create empty table

Repeat job	Select Previous Job	?
Location	Google Cloud Storage	gs://soydata/yob2015.txt
File format	CSV	View Files

Destination Table

Table name	SoyDataSet	nombres
Table type	Native table	?

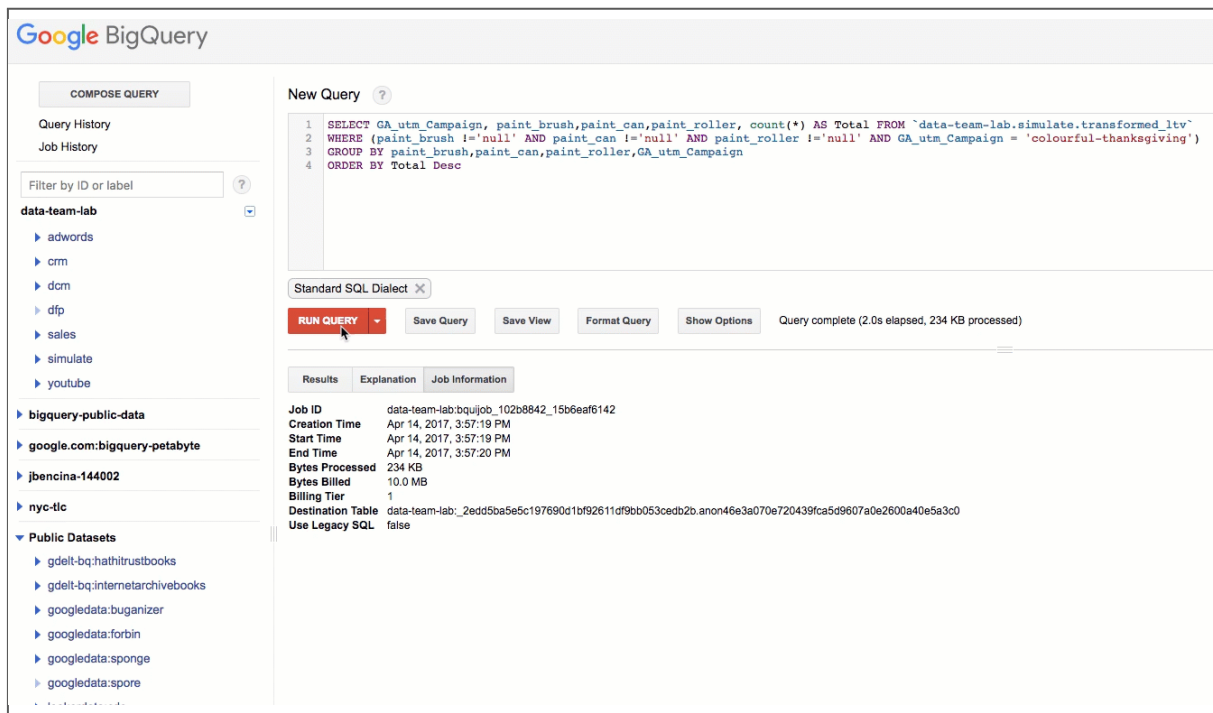
Schema ☒ Automatically detect ?

Schema will be automatically generated.

Options

Write preference Write if empty ?

Create Table



Seguridad y fiabilidad

BigQuery cifra los datos y crea una réplica automáticamente para garantizar la máxima seguridad, disponibilidad y durabilidad. Además también puede proteger aún más sus datos con potentes LCA basadas en funciones que se configuran y controlan mediante nuestro sistema Google Cloud Identity & Access Management.

Conexión con los productos de Google

Se puede exportar automáticamente los datos de Google Analytics Premium a BigQuery, visualizarlos con Google Data Studio y analiza conjuntos de datos almacenados en Google Cloud Storage.

Colaboraciones e integraciones

Varios partners y desarrolladores externos de Google Cloud Platform han creado distintas integraciones en BigQuery para cargar y procesar los datos, así como para generar visualizaciones interactivas de estos. Entre nuestros partners se incluyen Looker, Tableau, Qlik, Talend, Google Analytics, SnapLogic, Microstrategy y muchos más.

[BigQuery Data Transfer Service](#) automatiza el movimiento de los datos desde las aplicaciones de SaaS de partners a Google BigQuery.

