



DBi DATA
BUSINESS
INTELLIGENCE

BIG DATA EN SANTANDER R O

Con el fin de contar con la mayor cantidad de información posible, se creó una vista específica en Google Analytics 360 llamada “SR Big Data” que contiene información de **todos los activos digitales de Santander Rio**.

Mediante esta vista, se ha configurado hace unos meses atrás la integración **Google Analytics360-BigQuery** que será explicada a continuación.

Como producto de esta integración, se sincronizan diariamente la información completa de todo lo que ocurre **dentro del ecosistema digital** de la compañía. Para conocer más en detalle cual sería esta información, les enviaremos un documento que describe las distintas estrategias actuales de medición.



INTEGRACIÓN GOOGLE ANALYTICS 360 - BIGQUERY

Una vez que se integra Google Analytics con BigQuery, comienza a **sincronizarse diariamente** los datos procesados por la plataforma. Esta sincronización sucede todos los días cada ocho horas aproximadamente y agrega información a una tabla intradiaria presente en el Dataset de Google Analytics.

Una vez al día, por la madrugada se **sincronizan automáticamente los datos de la tabla intradiaria a una tabla definitiva** conteniendo el total de información del día anterior a la fecha sincronizada.

Toda la información que se sincroniza en BigQuery **surge a partir de una vista (View)** en Google Analytics 360 por lo que toda la información que se podrá consultar en el Dataset vendrá actualizada siguiendo las reglas definidas en la configuración de dicha vista.

Los datos de un día se consideran definitivos cuando la importación diaria haya finalizado. Es posible que observe diferencias entre los datos intradiarios y los datos diarios en función de las sesiones de usuario activas que sobrepasen el límite de tiempo de la última importación de datos intradiarios.

En BigQuery, cada tabla definitiva que contiene información de un día específico se nombra a partir del siguiente formato: "ga_sessions_AAAAMMDD" mientras que los datos intradiarios se incluyen en tablas nombradas de la siguiente manera: "ga_sessions_intraday_AAAAMMDD".

Table Details:

ga_sessions_20170508 (2017-05-08) ⬆

Table Details:

ga_sessions_intraday_20170509 (2017-05-09) ⬆

dbi-santanderrio-2324477 ⬇

▼ 137275638 + ⬇

■ ga_sessions_ (125)

■ ga_sessions_intraday_ (15)



MODELO DE DATOS DE GOOGLE ANALYTICS 360

El esquema de datos de una tabla del Dataset se compone de filas y columnas **ordenadas por usuario (Client ID de GA)**.

Cada **fila** en las tablas del Dataset de Google Analytics 360, representa una sola **sesión (Visita de un usuario)** al ecosistema digital de Santander Rio.

Cada **columna** representa una **dimensión o métrica de Google Analytics 360** y contiene muchos campos, algunos de los cuales se pueden repetir y anidar según lo defina el esquema.

fullVisitorId	visitId	customDimensions		hits						totals		
		index	value	hitNumber	type	customDimensions		page		bounces	pageviews	transactions
		index	value			index	value	pagePath	pageTitle			
				hitNumber	type	customDimensions		page				
						index	value	pagePath	pageTitle			
						index	value					

fullVisitorId	visitId	customDimensions		hits						totals		
		index	value	hitNumber	type	customDimensions		page		bounces	pageviews	transactions
		index	value			index	value	pagePath	pageTitle			
				hitNumber	type	customDimensions		page				
						index	value	pagePath	pageTitle			
						index	value					

El esquema de datos de Google Analytics 360 está confirmado por **distintos grupos de datos** anidados según la finalidad y naturaleza que tiene cada dato dentro de la plataforma.

Así mismo, el esquema completo está conformado por los siguientes grupos de información:

- **Datos identificatorios del usuario y de la sesión:**
Permiten distinguir al usuario y cada sesión en particular realizada por el mismo.
- **Métricas de performance de la sesión:**
Contienen las métricas que permiten entender las características de una visita (duración, profundidad, rebotesn, cantidad de interacciones, cantidad de pantallas vistas, etc).
- **Información sobre las Fuentes del tráfico:**
Contiene toda la información acerca del medio y la fuente desde la cual se generó la visita.

- **Tecnología y Geolocalización:**

Este grupo guarda información detallada sobre el dispositivo usado (resolución de pantalla, sistema operativo, browser, etc) en una sesión y toda la información demográfica que se recolecta automáticamente (país, ciudad, región, etc).

- **Métricas de performance de contenidos:**

Contiene la información precisa de todas las pantallas visitadas por el usuario y de todas las interacciones realizadas en la sesión.

- **Métricas transaccionales:**

Es un conjunto de métricas y dimensiones orientadas a capturar información relacionada con los procesos de E-Commerce presentes un sitio web. Aquí también figura la información de performance de las publicidades internas.

- **Métricas de performance de Aplicaciones móviles:**

Contiene todos los datos de las aplicaciones móviles nativas (Android & iOS) que se estén midiendo.

ESQUEMA DE DATOS DE GOOGLE ANALYTICS 360

DBi

visitorId
visitNumber
visitId
visitStartTime
date

fullVisitorId
userId
channelGrouping
socialEngagementType

totals
totals.visits
totals.hits
totals.pageviews
totals.timeOnSite
totals.bounces
totals.transactions
totals.transactionRevenue
totals.newVisits
totals.screenviews
totals.uniqueScreenviews
totals.timeOnScreen
totals.totalTransactionRevenue

device
device.browser
device.browserVersion
device.browserSize
device.operatingSystem
device.operatingSystemVersion
device.isMobile
device.mobileDeviceBranding
device.mobileDeviceModel
device.mobileInputSelector
device.mobileDeviceInfo
device.mobileDeviceMarketingName

hits
hits.hitNumber
hits.time
hits.hour
hits.minute
hits.isSecure
hits.isInteraction
hits.isEntrance
hits.isExit
hits.referrer
hits.page
hits.page.pagePath
hits.page.hostname
hits.page.pageTitle
hits.page.searchKeyword
hits.page.searchCategory
hits.page.pagePathLevel1
hits.page.pagePathLevel2

hits.transaction
hits.transaction.transactionId
hits.transaction.transactionRevenue
hits.transaction.transactionTax
hits.transaction.transactionShipping
hits.transaction.affiliation
hits.transaction.currencyCode
hits.transaction.localTransactionRevenue
hits.transaction.localTransactionTax
hits.transaction.localTransactionShipping
hits.transaction.transactionCoupon
hits.item
hits.item.transactionId
hits.item.productName
hits.item.productCategory

Existen algunas diferencias entre los datos en la Interfaz de Google Analytics 360 y BigQuery: algunas de estas diferencias son menores, como el hecho de que un Pageview, se llama PAGE en BigQuery, mientras que otros son más importantes y deberán ser tenidos en cuenta para el procesamiento de información.

Además, existen algunas otras diferencias con respecto al esquema tradicional de bases de datos relacionales como se menciona a continuación:

- **No es recomendado en BigQuery seleccionar todos los campos a la vez:**
Las tablas están divididas en columnas para una mejor compresión de los datos pero son almacenados en distintos elementos por lo que ejecutar una selección en un número menor de columnas utiliza menos recursos y se ejecutará más rápido. (ej. `Select * from`).
- **Los datos se dividen en tablas por fecha:**
El conjunto de datos utiliza la ID de la vista de Google Analytics 360 como su nombre. Para consultar varias tablas y combinar el resultado como si fuera una tabla, BigQuery proporciona la función `table_date_range`.

- **En el esquema cada registro representa una sesión:**
El fullVisitorId y visitId, juntos, forman una clave única para cada sesión.
- **En BigQuery, un tipo de datos de campo puede ser un RECORD:**
Es decir, puede tener un registro completo, con campos y valores, dentro de un campo. En el esquema de Google Analytics 360, un buen ejemplo de esto es trafficSource. Dentro de este campo se almacena un registro que contiene todo tipo de detalles de trafficSource. Utilizando la notación de puntos, puede recuperar cualquiera de estos campos interno.

fullVisitorID	visitID	trafficSource														
		<table><tr><td>source</td><td>medium</td><td>campaign</td><td>keyword</td><td>referralPath</td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr></table>					source	medium	campaign	keyword	referralPath					
source	medium	campaign	keyword	referralPath												
		<table><tr><td>source</td><td>medium</td><td>campaign</td><td>keyword</td><td>referralPath</td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr></table>					source	medium	campaign	keyword	referralPath					
source	medium	campaign	keyword	referralPath												
		<table><tr><td>source</td><td>medium</td><td>campaign</td><td>keyword</td><td>referralPath</td></tr><tr><td></td><td></td><td></td><td></td><td></td></tr></table>					source	medium	campaign	keyword	referralPath					
source	medium	campaign	keyword	referralPath												

- **En BigQuery, un campo puede ser REPETIDO:**
Además de ser NULLABLE y REQUIRED como en bases de datos tradicionales. Un campo repetido, puede tener una colección de valores, en lugar de un solo valor, almacenados dentro de un campo. En el esquema de Google Analytics en BigQuery, así se almacenan las dimensiones personalizadas, como REGISTROS REPETIDOS dentro del registro principal que representa una sesión:

fullVisitorID	visitID	customDimensions												
		<table><tr><th>index</th><th>value</th></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr></table>	index	value										
index	value													
		<table><tr><th>index</th><th>value</th></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr></table>	index	value										
index	value													
		<table><tr><th>index</th><th>value</th></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr><tr><td></td><td></td></tr></table>	index	value										
index	value													

Adicionalmente a este documento, se proveerá un Excel confeccionado en base al modelo de datos de Google Analytics 360 indicando la estructura de campos y definición de formatos de cada uno de ellos.

Como el esquema de Google Analytics 360 siempre está evolucionando, con nuevos campos que se agregan y otros antiguos que quedan obsoletos, es conveniente visitar frecuentemente el siguiente link oficial de Google donde aparece actualizado el esquema completo de datos de Google Analytics 360:

<https://support.google.com/analytics/answer/3437719>

DBi DATA
BUSINESS
INTELLIGENCE

BIGQUERY

BigQuery se puede acceder programáticamente utilizando una consola de Google Cloud Platform o bien de forma online a través de una interfaz amigable que se accede desde la siguiente url: <http://bigquery.cloud.google.com>

Google BigQuery

COMPOSE QUERY

Query History
Job History

Filter by ID or label ?

dbi-santanderrio-2324477

- ▼ 137275638
 - ga_sessions_ (213)
 - ga_sessions_intraday_ (16)
- ▼ Public Datasets
 - bigquery-public-data:hacker_news
 - bigquery-public-data:noaa_gsod
 - bigquery-public-data:samples
 - bigquery-public-data:usa_names
 - gdelt-bq:hathitrustbooks
 - gdelt-bq:internetarchivebooks
 - lookerdata:cdc
 - nyc-tlc:green

Table Details: ga_sessions_20170828 (2017-08-28) ⬆

Refresh Query Table Copy Table Export Table

Schema Details Preview

visitorId	INTEGER	NULLABLE	Describe this field...
visitNumber	INTEGER	NULLABLE	Describe this field...
visitId	INTEGER	NULLABLE	Describe this field...
visitStartTime	INTEGER	NULLABLE	Describe this field...
date	STRING	NULLABLE	Describe this field...
totals	RECORD	NULLABLE	Describe this field...
totals.visits	INTEGER	NULLABLE	Describe this field...
totals.hits	INTEGER	NULLABLE	Describe this field...
totals.pageviews	INTEGER	NULLABLE	Describe this field...
totals.timeOnSite	INTEGER	NULLABLE	Describe this field...
totals.bounces	INTEGER	NULLABLE	Describe this field...
totals.transactions	INTEGER	NULLABLE	Describe this field...

La interfaz de BigQuery permite realizar consultas en tiempo real sobre todos los datasets que se encuentren habilitados en un proyecto.

Para ello, provee de una consola de Queries donde en lenguaje SQL adaptado (propio de bigquery) se podrán generar cualquier tipo de consulta en una o varias tablas.

Al momento de edición de una Query, Google provee estadísticas de performance y costos asociados a la ejecución de la misma por lo que es posible conocer el volumen, tiempo de ejecución y costos estimados.

The screenshot displays the Google BigQuery web interface. On the left, the 'Compose Query' sidebar includes links for 'Query History' and 'Job History', and a list of datasets under 'Analytics/BigQuery' such as 'AnalyticsImport', 'GoogleStore', and 'publicdata:samples'. The main area, titled 'New Query', contains a SQL query that selects and groups visit data by hour and quarter-hour. Below the query editor, a red 'RUN QUERY' button is visible, followed by a status message: 'Query complete (3.8s elapsed, 2.54 GB processed)'. The 'Query Results' section, dated '11:48am, 14 May 2013', shows a table with 6 rows of data. Navigation links for 'Download as CSV' and 'Save as Table' are present. At the bottom, pagination controls show 'First < Prev Rows 1-6 of 96 Next > Last'.

```
1 SELECT
2   CONCAT(LPAD(STRING(HOUR(SEC_TO_TIMESTAMP(visitStartTime))),
3     2,
4     '0'),
5     '- ',
6     STRING(INTEGER(FLOOR(MINUTE(SEC_TO_TIMESTAMP(visitStartTime)) / 15)))) AS hqh,
7   HOUR(SEC_TO_TIMESTAMP(visitStartTime)) AS hour,
8   FLOOR(MINUTE(SEC_TO_TIMESTAMP(visitStartTime)) / 15) AS quarter_hour,
9   COUNT(visitId) AS numberVisits
10 FROM
11   [GoogleStore.sessions_20130415]
12 GROUP BY
13   hour,
14   quarter_hour,
```

Row	hqh	hour	quarter_hour	numberVisits
1	00-0	0	0.0	85
2	00-1	0	1.0	105
3	00-2	0	2.0	89
4	00-3	0	3.0	80
5	01-0	1	0.0	89
6	01-1	1	1.0	99

Para citar algunos ejemplos de consultas que se pueden hacer en los datasets de BigQuery, utilizaremos el ejemplo "London Cycle Helmet" proporcionado por Google.

Este modelo se puede agregar a BigQuery, incorporándolo como un dataset de Datos públicos. Este dataset se encuentra en el proyecto `google.com:analytics-bigquery` y en el conjunto de datos `LondonCycleHelmet`.

Objetivo: Contar el número de sesiones generadas por búsquedas orgánicas:

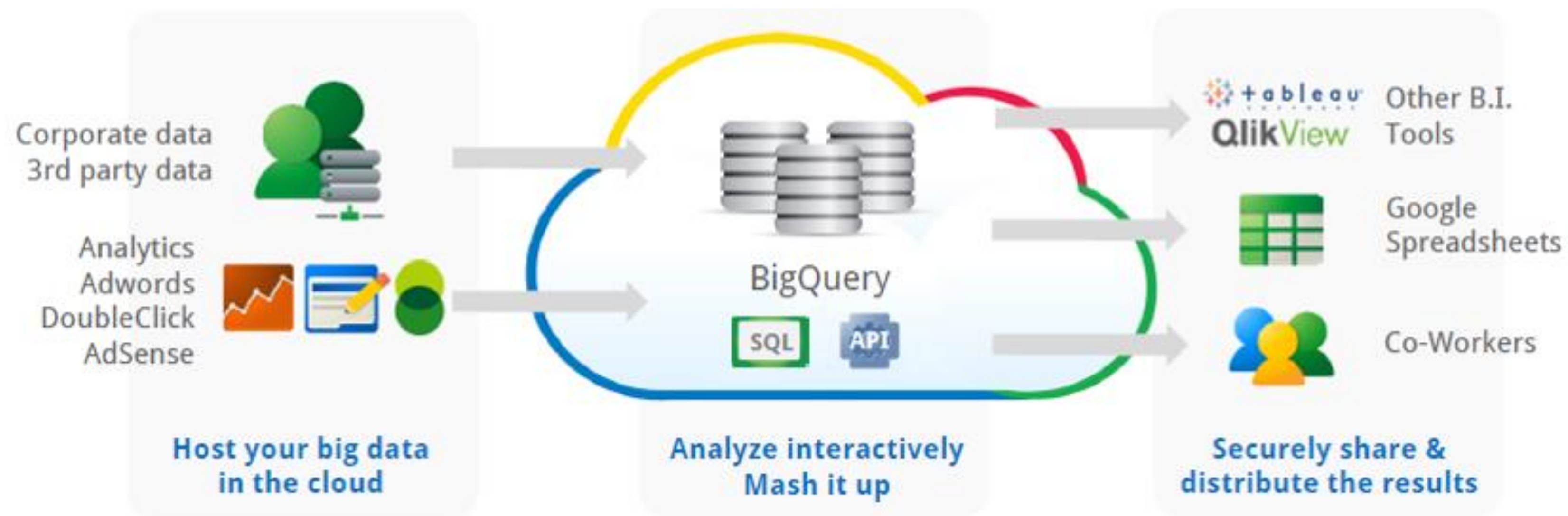
```
SELECT COUNT(totals.visits) AS visitCount
FROM [google.com:analytics-bigquery:LondonCycleHelmet.ga_sessions_20130910]
WHERE trafficSource.medium = "organic"
```

Objetivo: Obtener las tasas de rebote por fuente de tráfico (Medio)

```
SELECT trafficSource.medium AS medium,
       HOUR(SEC_TO_TIMESTAMP(visitStartTime)) AS sessionHour,
       COUNT(totals.bounces)/COUNT(totals.visits) AS bounceRate
FROM [google.com:analytics-bigquery:LondonCycleHelmet.ga_sessions_20130910]
GROUP BY medium, sessionHour
ORDER BY sessionHour, medium
```


Después de cargar los datos en BigQuery, se pueden exportar los datos en varios formatos. BigQuery puede exportar hasta **1 GB de datos por archivo** y admite la exportación a varios archivos individuales.

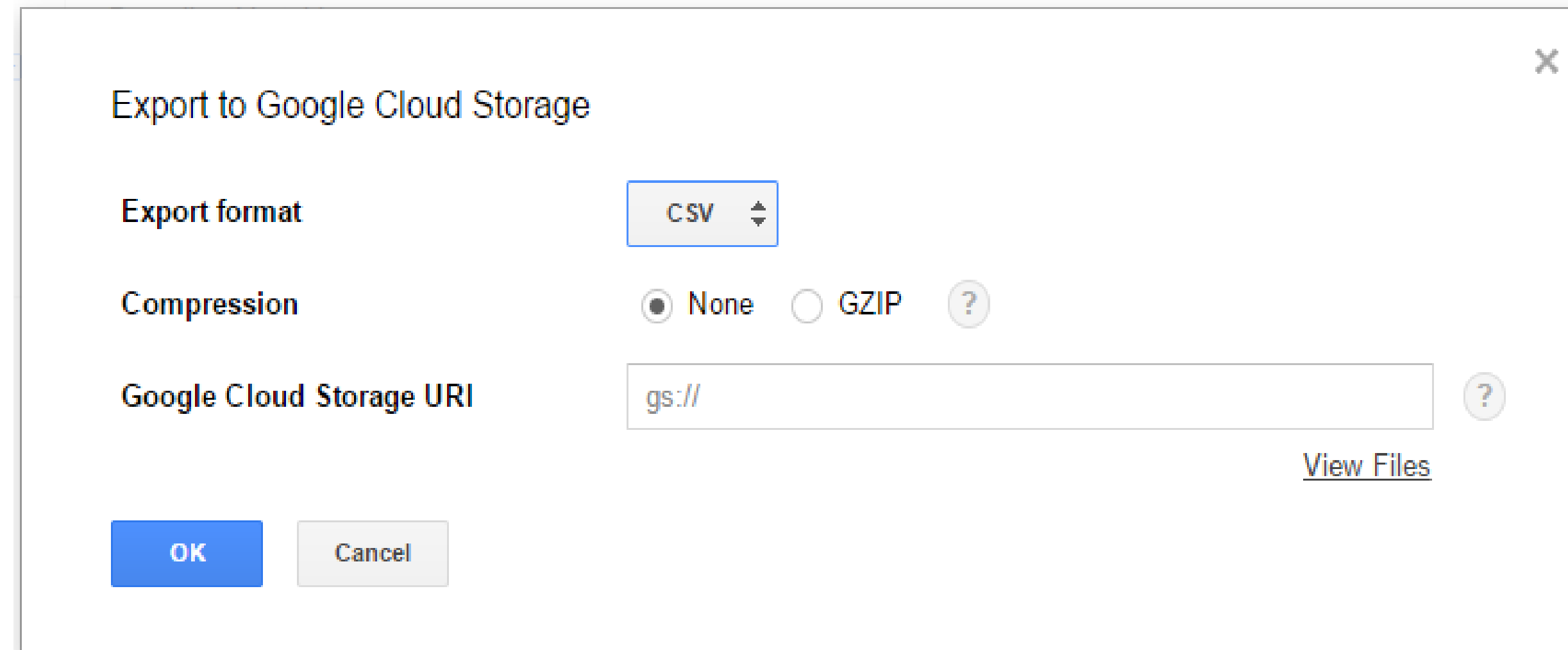
Para poder exportar datos de BigQuery existen dos grande métodos. Ninguno de ellos **genera costos asociados** pero sí existen límites en el volumen de información requerida:



Exportación manual desde la interfaz:

Se puede exportar datos bajando en formato CSV, JSON o AVRO cualquier tabla que se encuentre en la plataforma. Esta acción manual, genera un archivo en Google Cloud Storage que luego se podrá descargar fácilmente.

Cabe destacar que el proceso en BigQuery no genera costos asociados pero existen costos de almacenamiento en Google Cloud Storage que deberán tenerse presentes si se piensa guardar los archivos generados.



The image shows a dialog box titled "Export to Google Cloud Storage" with a close button (X) in the top right corner. It contains the following fields and controls:

- Export format:** A dropdown menu currently showing "CSV".
- Compression:** Two radio buttons, "None" (selected) and "GZIP", followed by a help icon (?).
- Google Cloud Storage URI:** A text input field containing "gs://", followed by a help icon (?).
- Buttons:** "OK" (blue) and "Cancel" (grey) at the bottom left.
- Link:** A "View Files" link at the bottom right.

Exportación programática por API:

Para exportar una tabla de BigQuery utilizando la API de BigQuery, se debe realizar una llamada al método `Jobs.insert` con la configuración adecuada, indicando el método de exportación escogido. Ejemplo:

```
{
  'jobReference': {
    'projectId': projectId,
    'jobId': uniqueIdentifier
  },
  'configuration': {
    'extract': {
      'sourceTable': {
        'projectId': projectId,
        'datasetId': datasetId,
        'tableId': tableId
      },
      'destinationUris': [cloudStorageURI],
      'destinationFormat': 'CSV',
      'compression': 'NONE'
    }
  }
}
```

En el caso de querer exportar una tabla que excede el tamaño máximo de salida de 1 GB por archivo, se deberá aprovechar la opción llamada "**Wildcard URI**" añadiendo un asterisco * en algún lugar del nombre de archivo.

Por ejemplo, una URI de almacenamiento en la nube de gs: //bookstore/melville-*.json en la configuración se convertirá realmente en una serie iterada de nombres de archivo incrementales de la siguiente manera:

```
gs://bookstore/melville-000000000000.json  
gs://bookstore/melville-000000000001.json  
gs://bookstore/melville-000000000002.json
```

Para mayor detalle sobre todos los procesos de exportación de datos desde BigQuery, se podrá consultar la documentación oficial de Google en la siguiente url: <https://cloud.google.com/bigquery/docs/exporting-data>

A long-exposure photograph of a city street at night, showing vibrant light trails from cars in red, orange, and blue. In the background, a modern building with a curved facade is visible under a dark sky. A security camera on a tall pole stands on the left side of the road. A semi-transparent dark grey banner is positioned across the middle of the image, containing the DBi logo and the text 'MUCHAS GRACIAS'.

DBi DATA
BUSINESS
INTELLIGENCE

MUCHAS GRACIAS