***MIGRACIÓN EN AMÉRICA: DESIGUALDAD ECONÓMICA Y DESPLAZAMIENTO FORZADO***

**DOCUMENTACIÓN FASE 2 DEL PROYECTO**

ESQUEMA ENTIDAD-RELACIÓN, DICCIONARIO DE DATOS Y FLUJO DE TRABAJO EN GOOGLE CLOUD PLATFORM

**EQUIPO DE TRABAJO CONSULTORA DATOS ESTRATÉGICOS:**

ADA PARHUANA - DATA SCIENTIST

ALEXANDER RIOS - DATA ENGINEER

CRISTIAN ANDRES - DATA ENGINEER

RICARDO SORIA - DATA SCIENTIST

SANTIAGO ZAPATA - DATA ANALYST

**DIRECCIÓN DE PROYECTO:**

MAICO BERNAL - SCRUM MASTER

**MAYO 2023**

**Introducción**

En el presente documento se presentan los detalles técnicos del diseño de la base de datos relacional a implementar en nuestro Data Warehouse y los detalles de la estructuración del proyecto en Google Cloud Platform. Encontrarán algunos enlaces de interés, el diccionario de datos, el esquema entidad-relación, el diagrama del flujo de trabajo en GCP y el detalle del paso a paso realizado durante todo el proceso en la nube.

**Enlaces de interés**

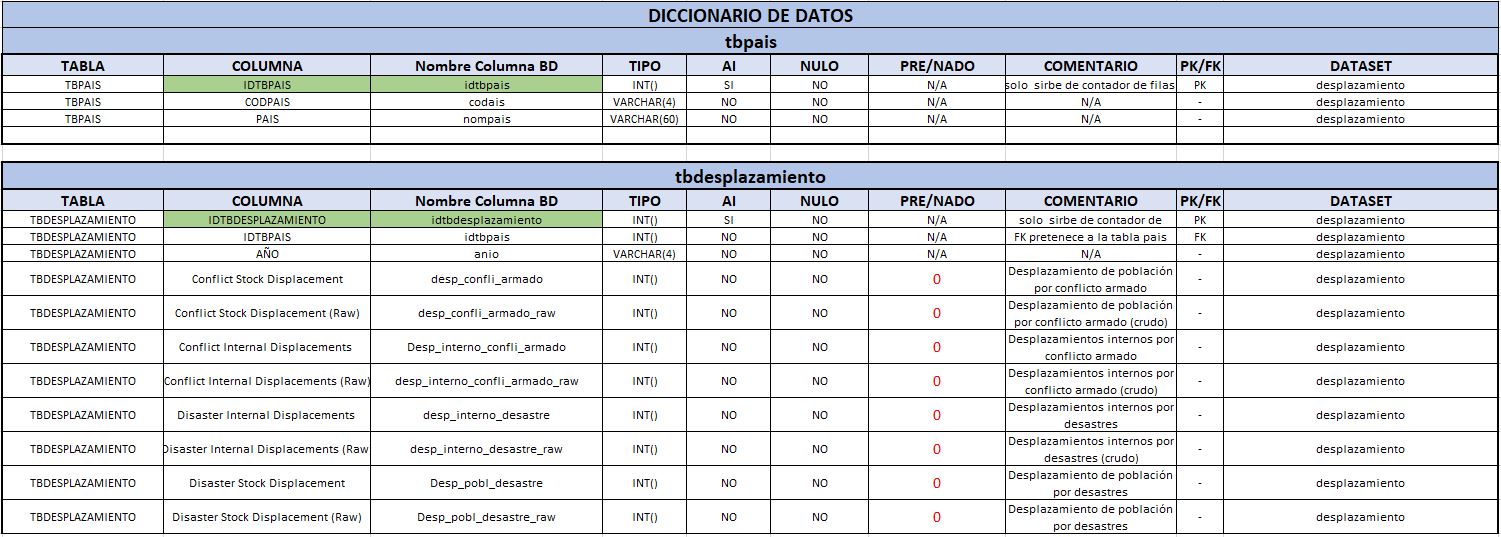
Presentación interactiva: <https://lookerstudio.google.com/reporting/eebc7c1c-eb2d-4415-939f-b1b1186d423e>

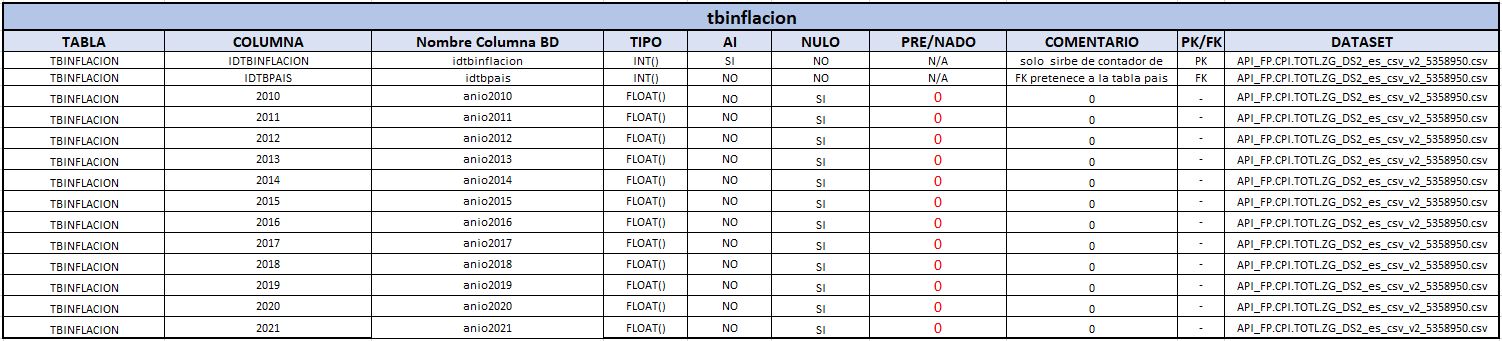
Repositorio: <https://github.com/Adapa22/PF-MigracionEnAmerica>

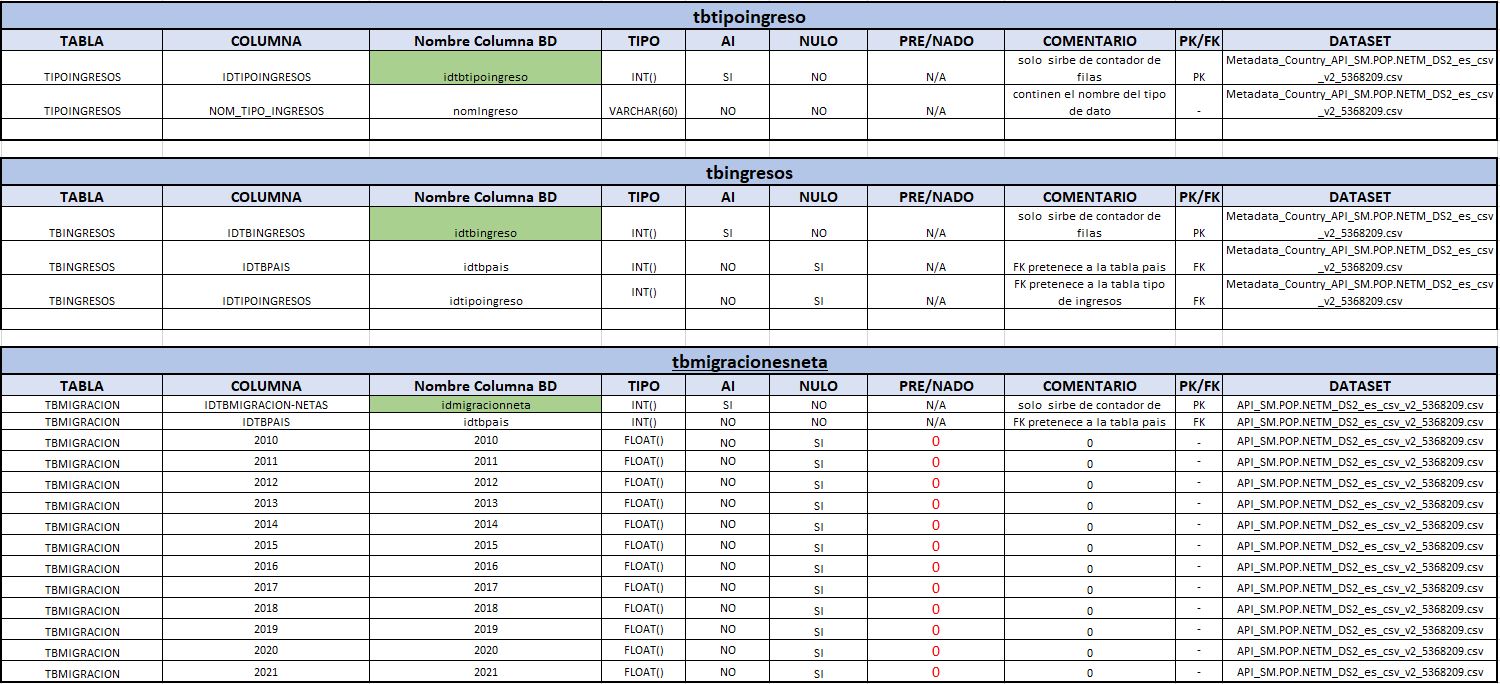
Video de prueba pipeline: <https://www.youtube.com/watch?v=FzxQB865p-Y>

Diccionario de datos completo: <https://docs.google.com/spreadsheets/d/1YiBxH3nfGt5mb2SwQruVHwFhLUIzOSaV/edit?usp=sharing&ouid=103160295414729886986&rtpof=true&sd=true>

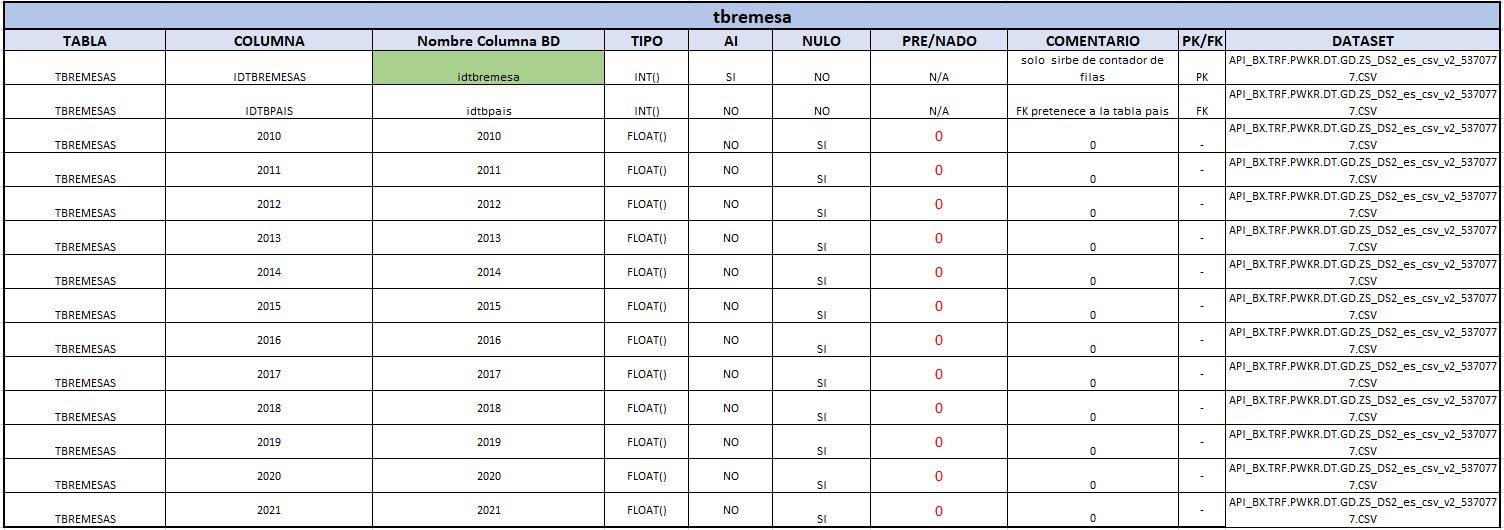
**Diccionario de datos:** se establece una descripción detallada de todas las tablas con base en los conjuntos de datos empleados. En el diccionario se puede observar los nombres de las tablas y de las columnas, el tipo de dato de cada columna, la descripción de cada variable, el uso de llaves primarias y foráneas, etc. En las siguientes imágenes se muestra la totalidad del diccionario de datos. También disponible aquí: <https://docs.google.com/spreadsheets/d/1YiBxH3nfGt5mb2SwQruVHwFhLUIzOSaV/edit?usp=sharing&ouid=103160295414729886986&rtpof=true&sd=true>



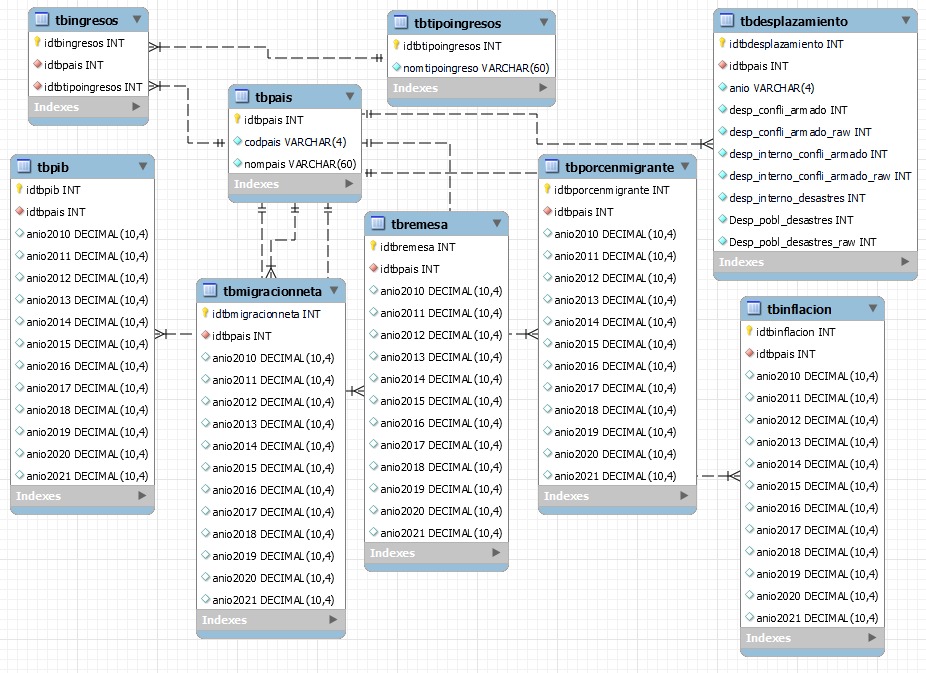








**Esquema ER (Entidad-Relación):** este es un modelo de datos que resulta útil para el diseño de la base de datos, ya que permite visualizar cómo se relacionan los datos antes de construir la estructura de la base de datos.



**Justificación del uso de Google Cloud Platform**

Decidimos realizar el proceso de ETL y Data Warehouse en la plataforma de Google Cloud debido a la flexibilidad y amplia gama de tecnologías que ofrece, muchas de ellas orientadas hacia los procesos de Big Data, Machine Learning, Data Analytics, entre otros. Luego también permite una fácil conexión entre diferentes herramientas incluyendo Looker Studio para la visualización.

Aparte de las ventajas mencionadas que ofrece Google Cloud Platform, realizar nuestro proceso de ETL nube también presenta muchas otras ventajas como:

***Accesibilidad remota:*** es accesible desde cualquier ubicación tanto para el trabajo colaborativo del equipo de desarrollo como para el uso de la solución propuesta por parta de la fundación Migrantes Sin Fronteras.

***Costos:*** permite un costo menor para el proyecto al no requerir infraestructuras costosas de hardware y software.

***Seguridad:*** GCP brinda seguridad y protección a los datos, así como copias de seguridad automáticas y control de acceso, lo que mejora la seguridad del proyecto y reduce el riesgo de pérdida de datos.

Las herramientas de Google Cloud Platform que vamos a emplear son las siguientes:

**Cloud Storage:** es un servicio de almacenamiento en la nube integrado a las demás herramientas.

**Dataflow:** facilita la creación de pipelines (canalizaciones) de datos que pueden integrarse con otros servicios de Google Cloud y de terceros.

**BigQuery:** es un servicio de análisis de datos que permite almacenar y consultar grandes conjuntos de datos con una gran velocidad y facilidad.

**Cloud Functions:** permite ejecutar código en respuesta a eventos específicos (trigger), como la carga de archivos en Cloud Storage.

**Looker Studio:** permite crear informes y dashboards a partir de los datos alojados en BigQuery.

**Vertex AI:** es un servicio de aprendizaje automático que permite crear, entrenar e implementar modelos de Machine Learning.

**Flujo de trabajo/arquitectura de los datos en Google Cloud Platform (resumen)**

Inicialmente se creará un bucket (herramienta de almacenamiento en la nube que funciona de forma similar a una carpeta) en el cual se alojarán el código de los pipelines y el data lake a donde llegarán los conjuntos de datos en crudo después de realizar la extracción.

Cuando estén subidos y listos los pipelines, se ejecutan mediante la terminal de Google Shell para que estos se encarguen de alojar los conjuntos de datos en el data lake y realizar unas transformaciones iniciales.

Luego mediante la herramienta BigQuery se tomarán estos datos desde el data lake para transformarlos y añadirlos a la base de datos (Data Warehouse).

Una vez se tengan los datos almacenados en el Data Warehouse, estos estarán disponibles para ser tomados por la herramienta de visualización Looker Studio para la posterior creación de un dashboard y otras herramientas de aprendizaje automático como es el caso de Vertex AI.

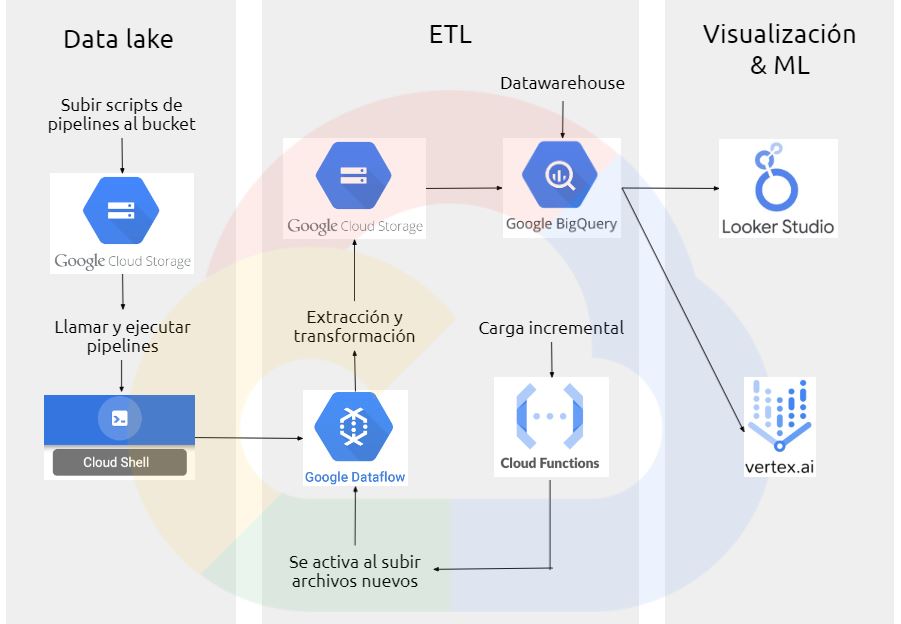
Adicionalmente, se emplearán funciones de Cloud Functions que permitan realizar la carga incremental de datos. A estas funciones se les define un activador o *trigger* para que se ejecuten de forma automática. En este caso el activador que se tomará es el de subir un archivo a un bucket.

El proceso de carga incremental de datos con Apache Beam permite la adición de nuevos datos a una tabla existente sin perder los registros antiguos que ya se encuentran allí. Esto se logra mediante el uso de una clave de identificación única para cada registro en la tabla y comparando esta clave con los nuevos datos que se van a cargar. Si se encuentra una coincidencia, se actualiza el registro existente con los nuevos valores. De lo contrario, se agrega un nuevo registro a la tabla. Este enfoque garantiza que la tabla siempre contenga los datos más recientes y precisos. Además, al utilizar Apache Beam para realizar este proceso, se obtiene una solución escalable y fácilmente configurable para la carga incremental de grandes cantidades de datos en tiempo real.

En este diagrama, podemos ver que el proceso comienza con la lectura de los datos nuevos y antiguos. A continuación, se lleva a cabo una transformación que utiliza una clave de identificación única para comparar los nuevos datos con los existentes. Si se encuentra una coincidencia, se actualiza el registro existente; de lo contrario, se agrega un nuevo registro.

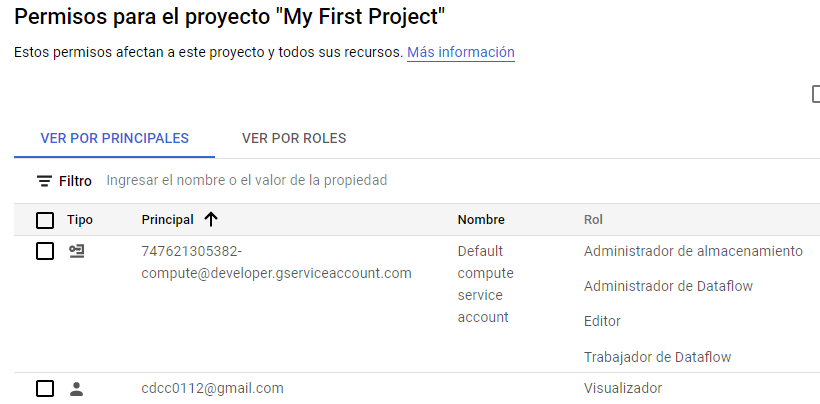
Después de la transformación, los datos actualizados se cargan en la tabla existente, manteniendo los datos antiguos que ya se encuentran allí. Este proceso se puede repetir de forma regular para mantener actualizada la tabla con los nuevos datos.

**Diagrama flujo de trabajo en la nube**

****

**Paso a paso del proceso a realizar en Google Cloud Platform**

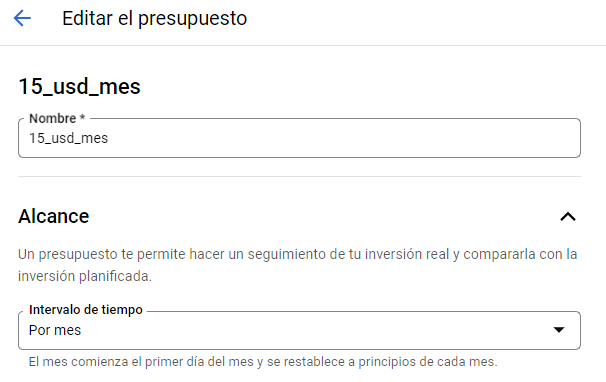
Se crean roles y permisos para las aplicaciones y para el equipo de trabajo:



Se habilitan las API necesarias para el correcto funcionamiento de las diversas herramientas:



Se configura el presupuesto para el proyecto:



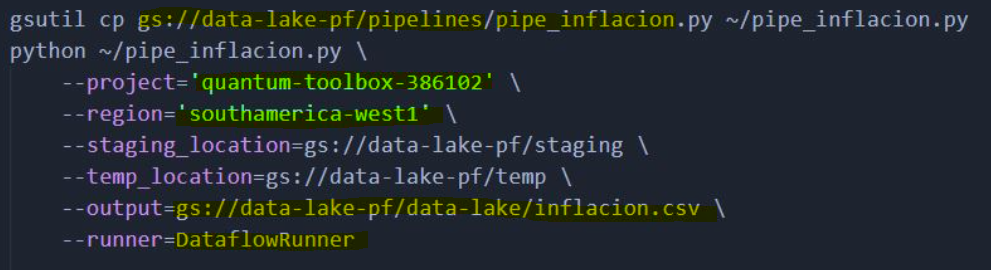
Se crea y configura un bucket en Cloud Storage:



Se almacenan en el bucket los pipelines que se van a ejecutar luego:



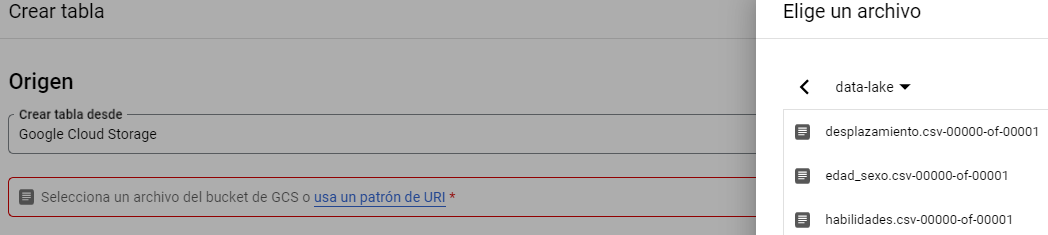
Se ejecutan los pipelines mediante la terminal Cloud Shell definiendo parámetros como ubicación, id del proyecto, carpetas temporales, ubicación de salida y “runner” en este caso dataflow, es decir, la herramienta que va a realizar el flujo de datos (demo funcionamiento: <https://www.youtube.com/watch?v=FzxQB865p-Y> )



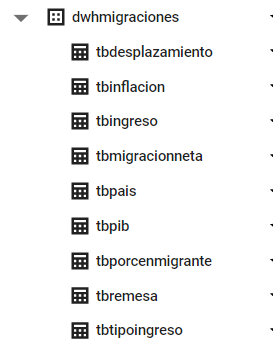
Después de ejecutar los pipelines, automáticamente se suben los archivos procesados al data lake:



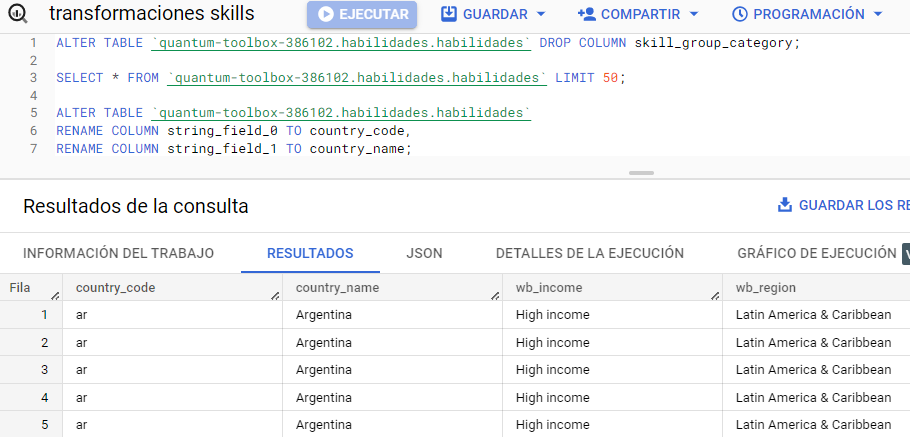
Con BigQuery se toman los datos desde el data lake para crear las tablas:



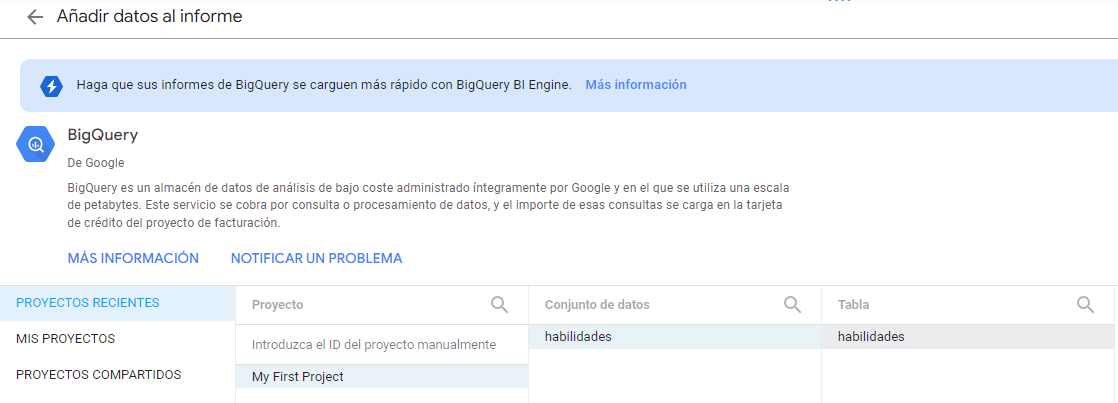
Se crea el Data Warehouse en BigQuery con las sentencias respectivas las cuales son similares a MySQL pero es necesario cambiar la sintaxis:



Con BigQuery se realizan transformaciones y se pueden almacenar las sentencias SQL para reutilizarlas en el futuro:



Los datos del Data Warehouse se pueden tomar desde Looker Studio y otras herramientas:



Para realizar la carga incremental de los datos nuevos se pueden emplear Cloud Functions que se activan automáticamente al cargar un archivo en el bucket seleccionado:

