



Facultad de Ingeniería
Gestión de datos.

Taller 1 – Primer Semestre 2023

Taller Final.

Preparación de datos, visualización y modelo de recomendación

Joshep Andersson Blanco Reres ^{a,c} , Laura Estrada Saldarriaga ^{a,c}

Fabián Camilo Peña^{b,c}

a Estudiante de Maestría en Inteligencia Artificial.

b Profesor, Departamento de Ingeniería de Sistemas

c Pontificia Universidad Javeriana, Bogotá, Colombia



Facultad de Ingeniería

Gestión de datos.

Taller 1 – Primer Semestre 2023

Contenido

SEGUNDA ENTREGA. CONSTRUCCIÓN DEL ETL	2
1. Arquitectura ETL	3
2. CONSTRUCCIÓN DASHBOARD	6
ANEXO 1. Detalles del pipeline de automatización ejecutado en VertexAI	9
ANEXO 2. Detalles de la configuración del bucket y de la base de datos.	11
ANEXO 3. Conexión de la base de datos en GCP a Google DataStudio (Visualizador.)	15

ENTREGA FINAL

El repositorio se encuentra en el siguiente link.

https://github.com/Joshep1229/spotify_project

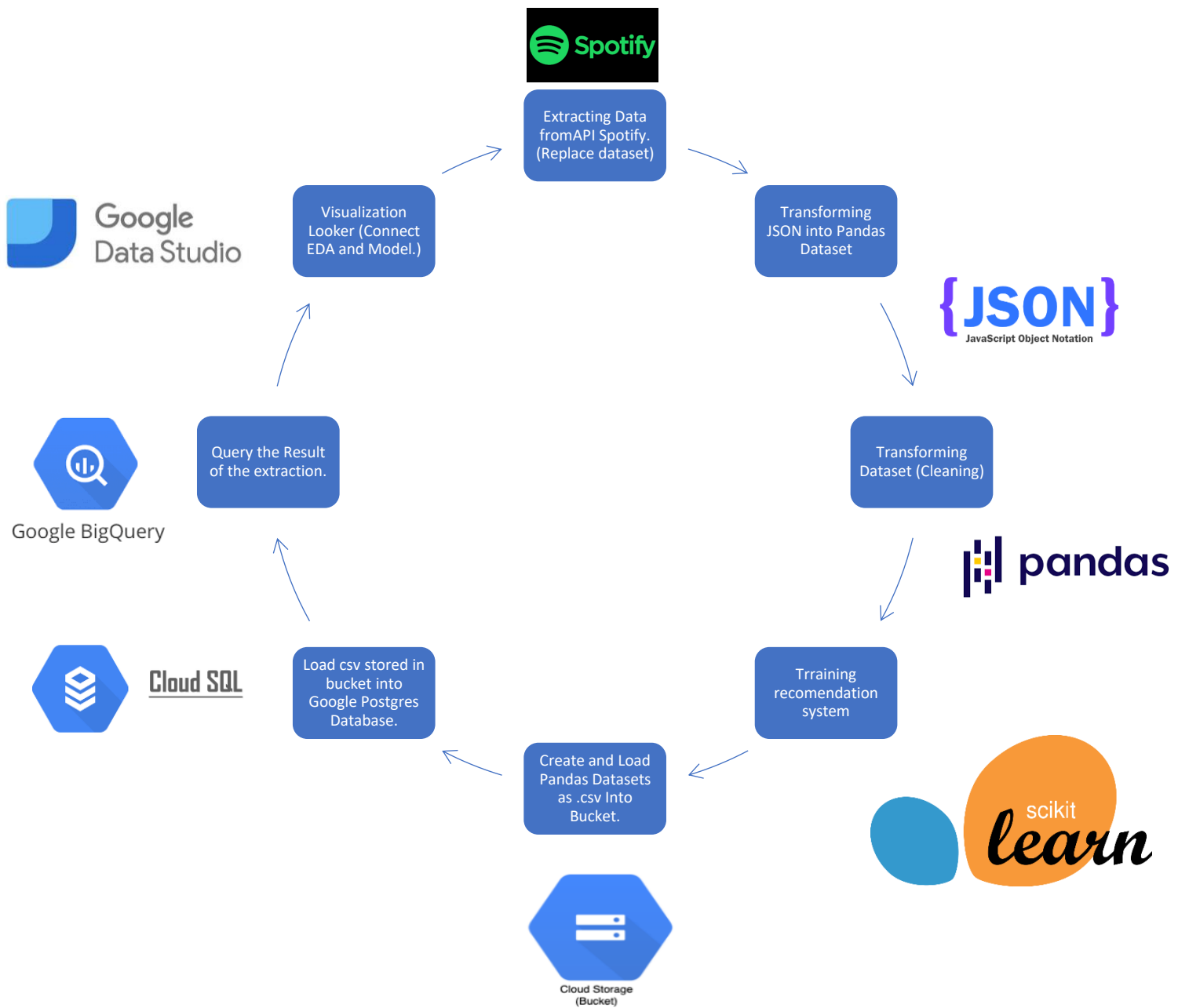
1. Arquitectura ETL

Orchestrator:



Vertex AI

Pipeline of Informa...



Este proyecto contiene una solución end to end **en la nube** para ejecutar desde la extracción hasta la visualización y el modelo de recomendación de forma automática usando VertexAI.

Las playlists usadas fueron el top 50 global, el top 50 Brazil y el top 50 Corea del sur. Los datos generados y el modelo, se actualizan **semanalmente**.

Esto permite que el sistema **siempre generará recomendaciones nuevas y personalizadas de los mejores 50 canciones de los 3 países.** Tomando en cuenta las canciones que ingresen o salgan del top.

Flujo en la nube en detalle:

1. Cargar Información desde la ****API de spotify****.
2. Preprocesarla y generar datasets (Artistas, canciones, género, datos de modelo y gráfico de similitud.)
3. Procesar ****modelo de recomendaciones**** y guardar los resultados como un csv (Recomendaciones)
3. Guardar los dataset como formato .csv en un ****bucket en GCP****. (Previa conexión configurada en el notebook.)
4. Con SQL cargar el archivo de csv guardado en el bucket en la base de datos de ****Google Bigquery de GCP**** (que está en postgres).
5. Conectar la base de datos de Google Bigquery a ****Google Datastudio (Visualizador)****
6. Visualizar los resultados en Google datastudio para visualizar la información.
 - 6.1 La primera hoja contiene el dashboard.
 - 6.2 La segunda hoja contiene el modelo de recomendaciones generadas por el modelo.

2. SISTEMA DE RECOMENDACIÓN.

Para el sistema de recomendación, se ejecutó el modelo en el flujo preestablecido de los datos

Se ejecuta el modelo con los parámetros de 'danceability', 'energy', 'loudness' y 'valence'.

Se intentó hacer un modelo más complejo agrupando variables como la duración, el género e incluso si la playlist era coreana, brasilera o global, sin embargo sesgaba mucho el resultado y se perdía la variedad en las canciones.

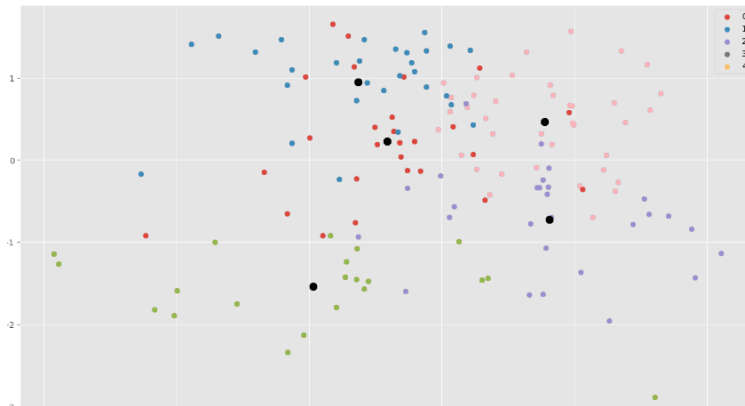
En éste caso quisimos tener un equilibrio entre recomendar música similar, pero también mostrarle elementos distintos al usuario, esa es la razón por la que escogimos 3 mercados muy dispares en la comparación. (Mercado coreano, brasilero y global.)

Por ello, también agrupamos cuatro elementos a nuestro modelo, la energía, la bailabilidad, el ruido y el valence, o el qué tan positiva sonaba la canción. Dejamos por fuera elementos como el género o el país donde se generó la playlist, para darle variedad.

```
[327] from sklearn.cluster import KMeans
import numpy as np

X=np.array(Datos_Modelo[['danceability', 'energy', 'loudness','valence']])
X_norm = (X - X.mean(axis=0)) / X.std(axis=0)

kmeans = KMeans(n_clusters=5).fit(X_norm)
label = kmeans.fit_predict(X_norm)
centroids = kmeans.cluster_centers_
print(centroids)
```



El modelo tiene 5 centroides a los que se les calcula la distancia al centroide y posteriormente la distancia entre los puntos usando la distancia euclidiana.

Esta matriz cuadrada de 143 canciones (50 por cada top, menos las repetidas), se convierte a un formato largo, que contiene:

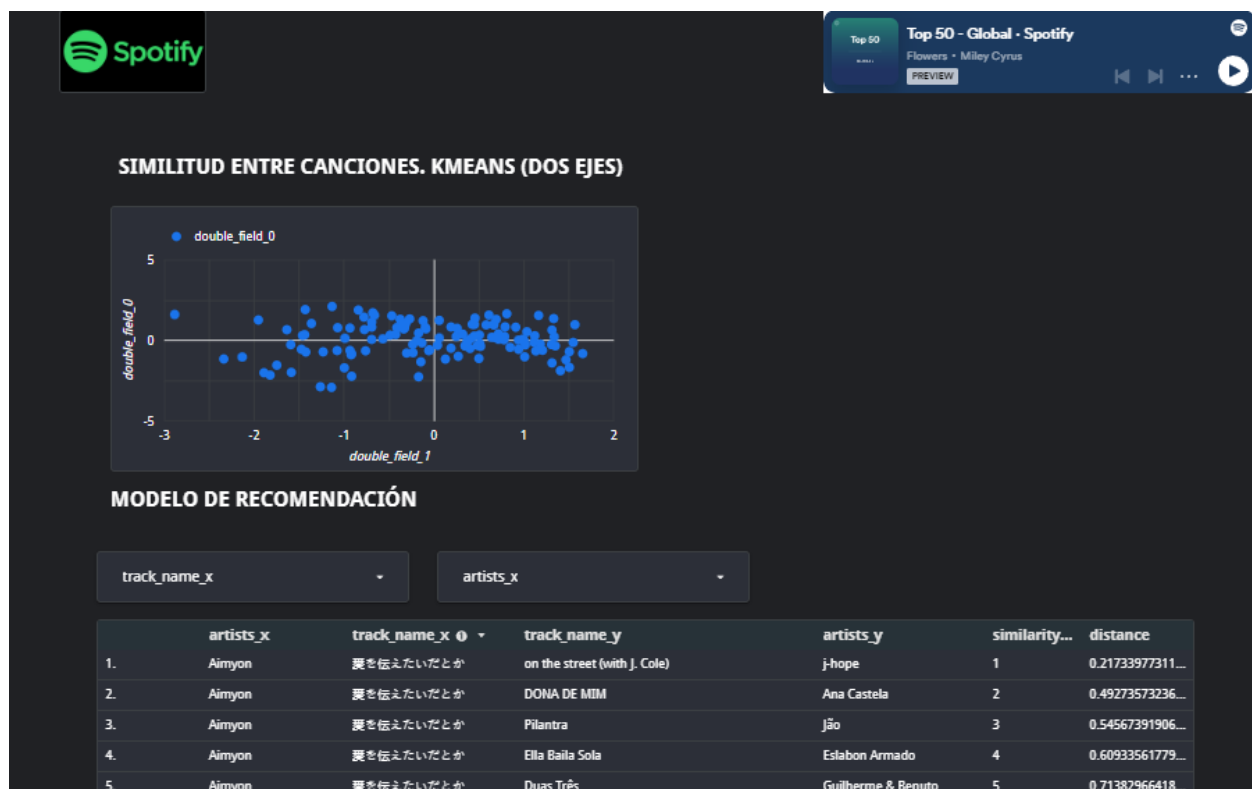
Canción A, Artista	Canción A, Arista	Canción B, Arista	Distancia	Ranking
Bad Bunny	Ojitos lindos	TQG Karol G	0,42	1
Bad Bunny	Ojitos lindos	Mor Ferxxo	0.67	2
Bad Bunny	Ojitos lindos	Para Elisa Mozart	4.52	149

Así sucesivamente hasta tener todos los pares posibles para todas las canciones que contienen el dataset.

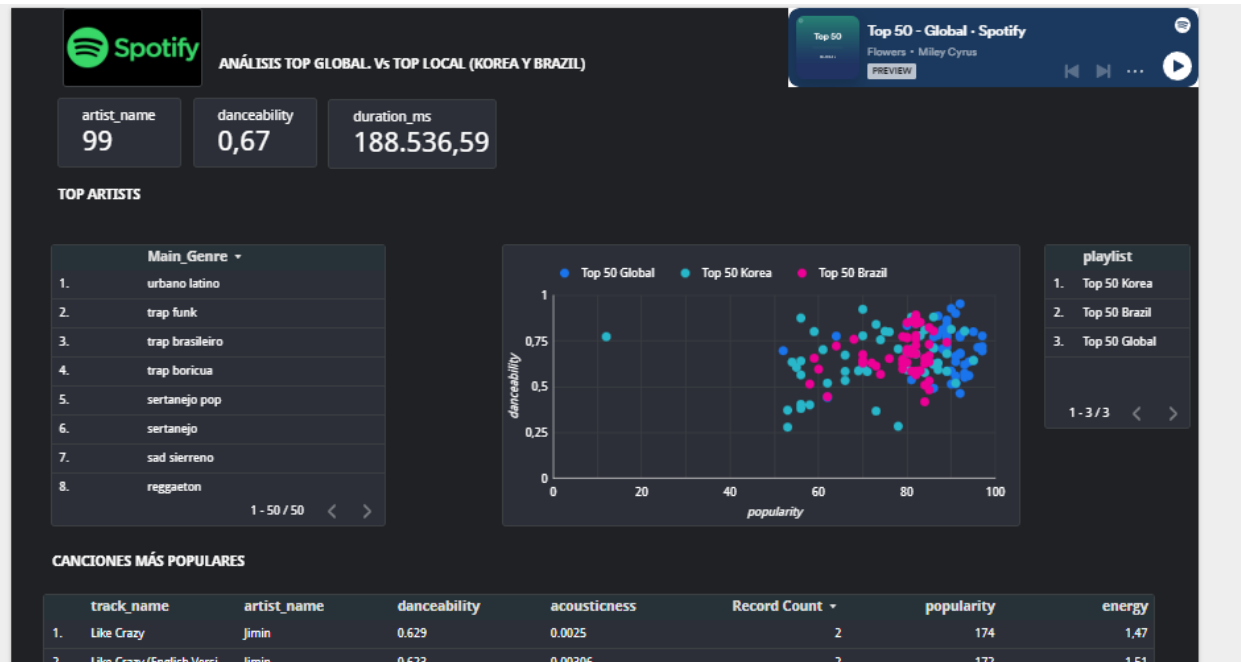
Esta base se carga como una base de datos en bigquery, no se entrena un modelo y se realiza la exportación a un joblib, por la naturaleza de la base de datos, debido a que si entran y salen canciones del top de los tres países, es mejor reentrenar el modelo con los nuevos elementos de datos que se generen.

Se crea una base de datos y no un API rest, porque desde el visualizador, podemos conectar el sistema de recomendación como un conjunto de datos en big query y visualizar las recomendaciones de todos los artistas (además de mantener en la nube la solución.).

Sistema de recomendación integrado a Datastudio.



3. CONSTRUCCIÓN DASHBOARD



El dashboard se genera usando datastudio en conexión con BigQuery, y contiene información importante sobre el total de artistas analizados, el tipo de canciones examinadas, y métricas. Clave, como el promedio de duración de las canciones.

Nos da una visual interesante de cómo se diferencian las canciones dependiendo del país, su nivel de popularidad y de su bailabilidad, podemos observar cómo Corea del Sur tiene canciones más planas en comparación al top de Brasil o al top global.

Son variables importantes para el negocio, porque la compañía puede indagar sobre las métricas que la podrían hacer triunfar en cada mercado. Y de cómo podría acercarse a cada país. También permite visibilizar artistas que podrían ser muy populares fuera de sus fronteras al encontrarle similitudes con los artistas de otros mercados, como se verá en el sistema de recomendación.

También se inserta la app de Spotify para que el usuario pueda disfrutar de la música mientras realiza la exploración del mismo.

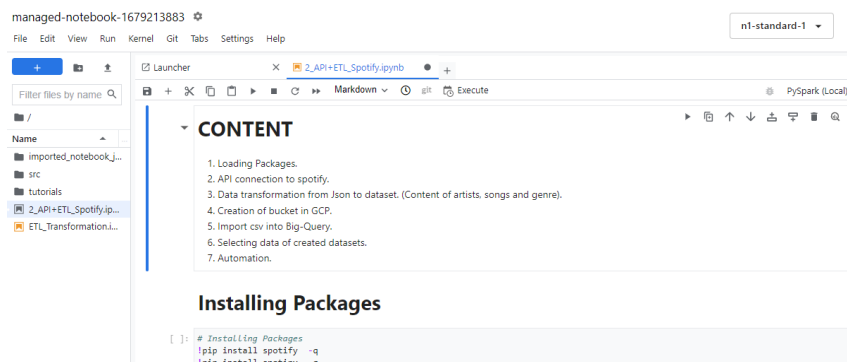
ANEXO 1. Detalles del pipeline de automatización ejecutado en VertexAI

AUTOMATIZACIÓN USANDO VERTEX AI.

Vertex AI es un sistema que permite ejecutar, compilar y escalar modelos de datos (MLOps), para el caso nuestro, ha sido usado para automatizar el proceso de ETL, pero también se podría ejecutar el sistema de recomendación.

Éste sistema crea toda la lógica y la ejecución automáticamente en Google Composer. Para tal finalidad, se debe crear una máquina virtual que ejecute el Jupyter notebook dentro de la aplicación, hemos escogido la más básica de las opciones, una CPU de 1 GB sin tarjeta gráfica.

Luego de la creación del entorno, subimos el notebook y se verá se la siguiente manera.



Le damos en el ícono del reloj y mostrará las siguientes opciones de programación de la tarea.

Submit notebooks to Executor

Accelerator type
None

Environment
Custom Container

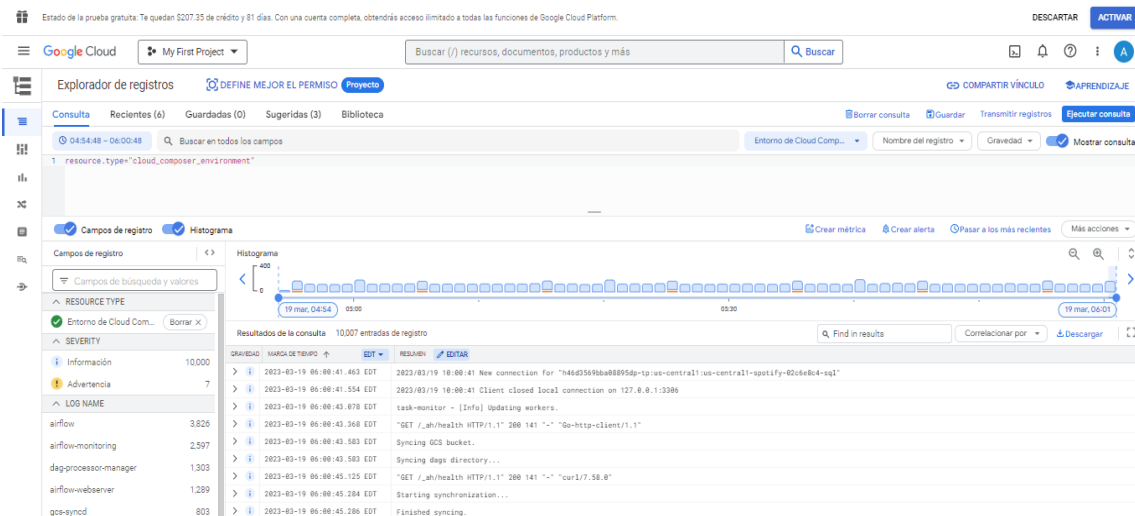
Docker container image
us-docker.pkg.dev/deeplearning-platform/gcr.io/release.base-cu113:m1
Note: Custom container must be a [derivative container](#).

Type
Schedule-based recurring executions

Repeat every
week

Repeat at
09:00

Al ejecutarse, genera un log de ejecución que detalla tanto los errores, como las ejecuciones dadas por el programa.



Estado de la prueba gratuita: Te quedan \$207.35 de crédito y 81 días. Con una cuenta completa, obtendrás acceso limitado a todas las funciones de Google Cloud Platform.

DESCARTAR ACTUAR

Google Cloud My First Project

Buscar (/) recursos, documentos, productos y más

Explorador de registros

Consulta Recientes (6) Guardadas (0) Sugeridas (3) Biblioteca

04:54:48 - 06:00:48

Buscar en todos los campos

Entorno de Cloud Com... Nombre del registro Gravedad Mostrar consulta

1 resource.type="cloud_composer_environment"

Campos de registro Histograma

Campos de búsqueda y valores

RESOURCE TYPE

Entorno de Cloud Com... (Borrar X)

SEVERITY

Información 10,000

Advertencia 7

LOG NAME

airflow 3,826

airflow-monitoring 2,597

dag-processor-manager 1,303

airflow-webserver 1,289

gcp-synod 803

Resultados de la consulta: 10,007 entradas de registro

SEVERIDAD	FECHA DE TIEMPO	RESUMEN
Información	2023-03-19 06:00:41.463 EDT	2023/03/19 16:00:41 New connection for "h46d569ba88895dp-tp-us-central1-us-central1-spotify-62c6b64-sql"
Advertencia	2023-03-19 06:00:41.554 EDT	2023/03/19 16:00:41 Client closed local connection on 127.0.0.1:3386
Información	2023-03-19 06:00:43.678 EDT	task-monitor - [Info] Updating workers.
Información	2023-03-19 06:00:43.368 EDT	"GET /_ah/health HTTP/1.1" 200 141 "-" "Go-http-client/1.1"
Información	2023-03-19 06:00:43.583 EDT	Syncing GCS bucket.
Información	2023-03-19 06:00:43.583 EDT	Syncing dags directory...
Información	2023-03-19 06:00:45.125 EDT	"GET /_ah/health HTTP/1.1" 200 141 "-" "curl/7.58.0"
Información	2023-03-19 06:00:45.284 EDT	Starting synchronization...
Información	2023-03-19 06:00:45.286 EDT	Finished syncing.

También podemos ver los errores de ejecución del notebook, puntualmente.

Bucket Creado

← Detalles del bucket ACTUALIZAR APRENDIZAJE

spotify11

Ubicación: us-east1 (Carolina del Sur) | Clase de almacenamiento: Standard | Acceso público: No público | Protección: Ninguna

OBJETOS CONFIGURACIÓN PERMISOS PROTECCIÓN CICLO DE VIDA OBSERVABILIDAD **NUEVO**

Depósitos > spotify11

SUBIR ARCHIVOS SUBIR CARPETA CREAR CARPETA TRANSFERIR LOS DATOS ADMINISTRAR CONSERVACIONES DESCARGAR BORRAR

Filtrar solo por prefijo de nombre Filtro: Filtrar objetos y carpetas Mostrar datos borrados

<input type="checkbox"/>	Nombre	Tamaño	Tipo	Fecha de creación	Clase de almacenamiento	Última modificación	Acceso público	Historial de versiones	Encriptación	Fecha de vencimiento de la retención	Conservaciones
No hay filas para mostrar											

Estado de la prueba gratuita: Te quedan \$400.00 de crédito y 91 días. Con una cuenta completa, obtendrás acceso ilimitado a todas las funciones de Google Cloud Platform.

Google Cloud My First Project cuentas de ser Buscar 4 ? J

IAM y administración Cuentas de servicio + CREAR CUENTA DE SERVICIO BORRAR ADMINISTRAR ACCESO ACTUALIZAR APRENDIZAJE

Cuentas de servicio del proyecto "My First Project"

Una cuenta de servicio representa una identidad de servicio de Google Cloud, como el código en ejecución en las VM de Compute Engine, las apps de App Engine o los sistemas que se ejecutan fuera de Google. [Obtén más información sobre las cuentas de servicio.](#)

Las políticas de la organización se pueden usar para asegurar las cuentas de servicio y bloquear sus características riesgosas, como el otorgamiento automático de IAM, la creación y carga de claves, o la creación misma de cuentas de servicio. [Obtén más información sobre las políticas de la organización para cuentas de servicio.](#)

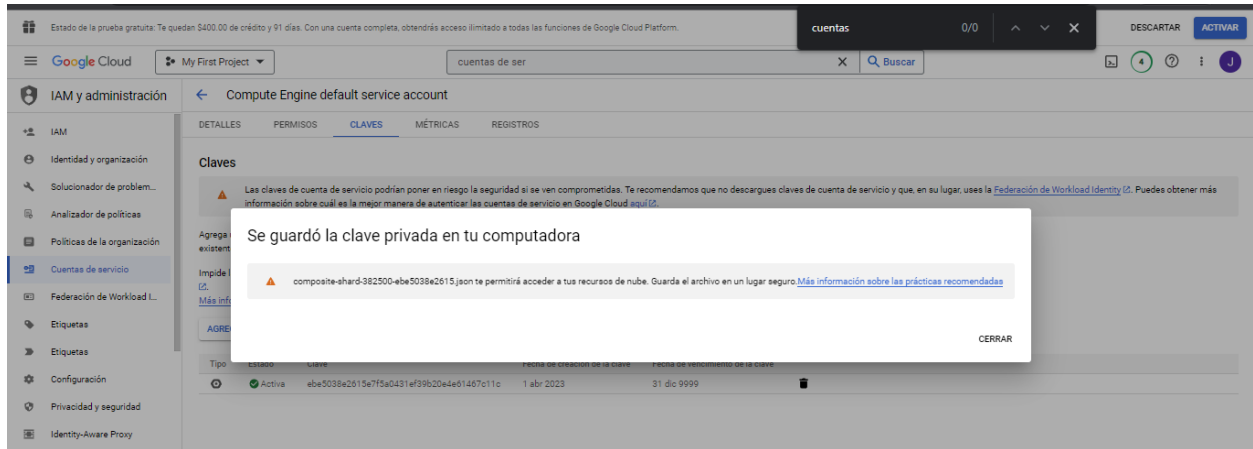
Filtro: Ingresar el nombre o el valor de la propiedad

<input type="checkbox"/>	Correo electrónico	Estado	Nombre	Descripción	ID de clave	Fecha de creación de la clave	ID de cliente de OAuth2	Acciones
<input type="checkbox"/>	270888618758-compute@developer.gserviceaccount.com	✓	Compute Engine default service account		No hay claves		11672137609083116234	Acciones Administrar detalles Administrar permisos Administrar claves Ver métricas Ver registros Inhabilitar Borrar

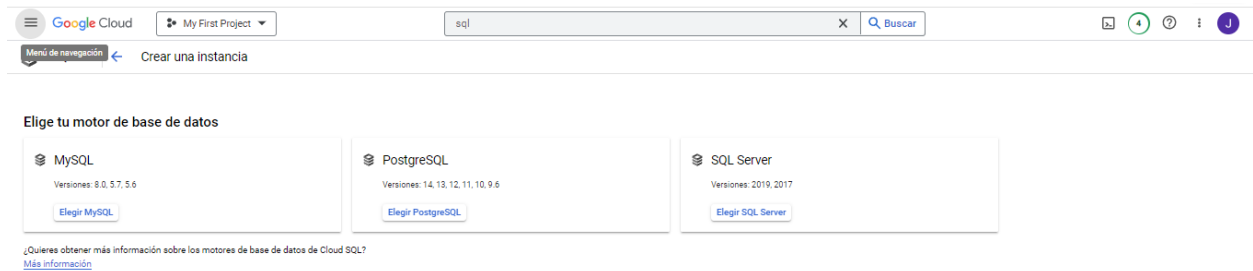
Se copió el URI de gutil al portapapeles

Operaciones de My First Project cargadas COMPETENCIAS CLAVES DE UN Completada

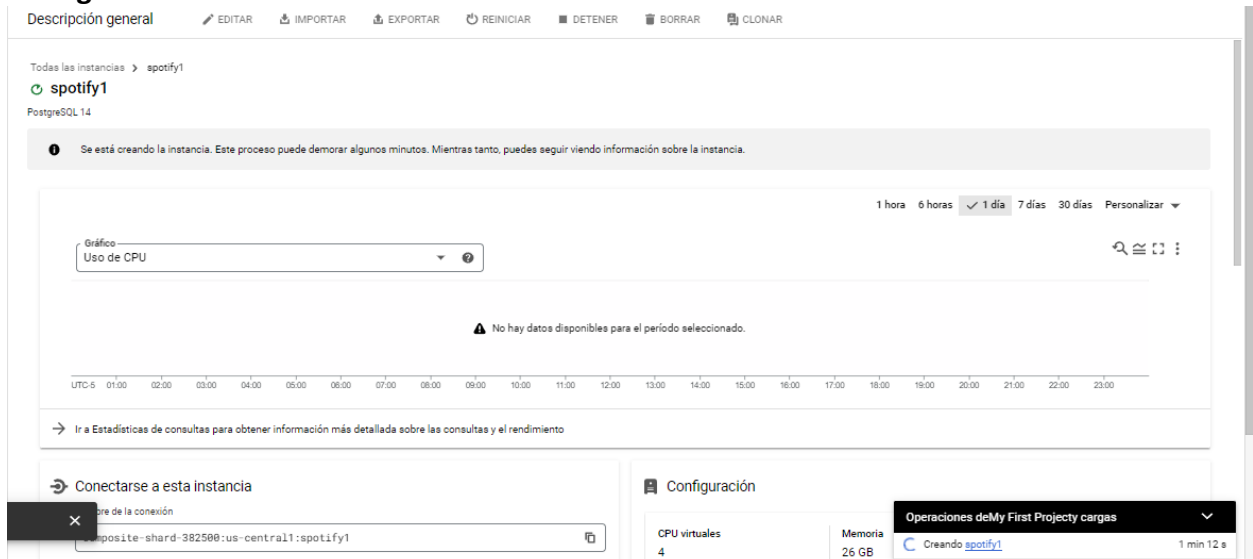
Creación clave de bucket



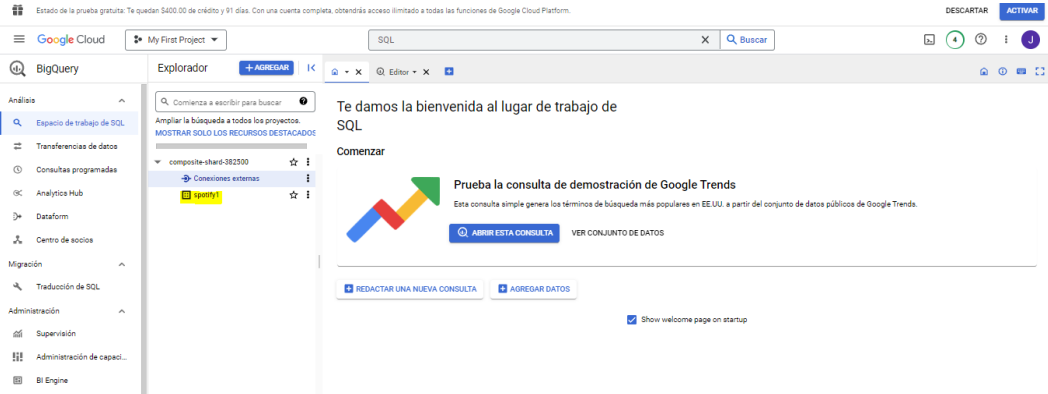
Creación de SQL Cloud



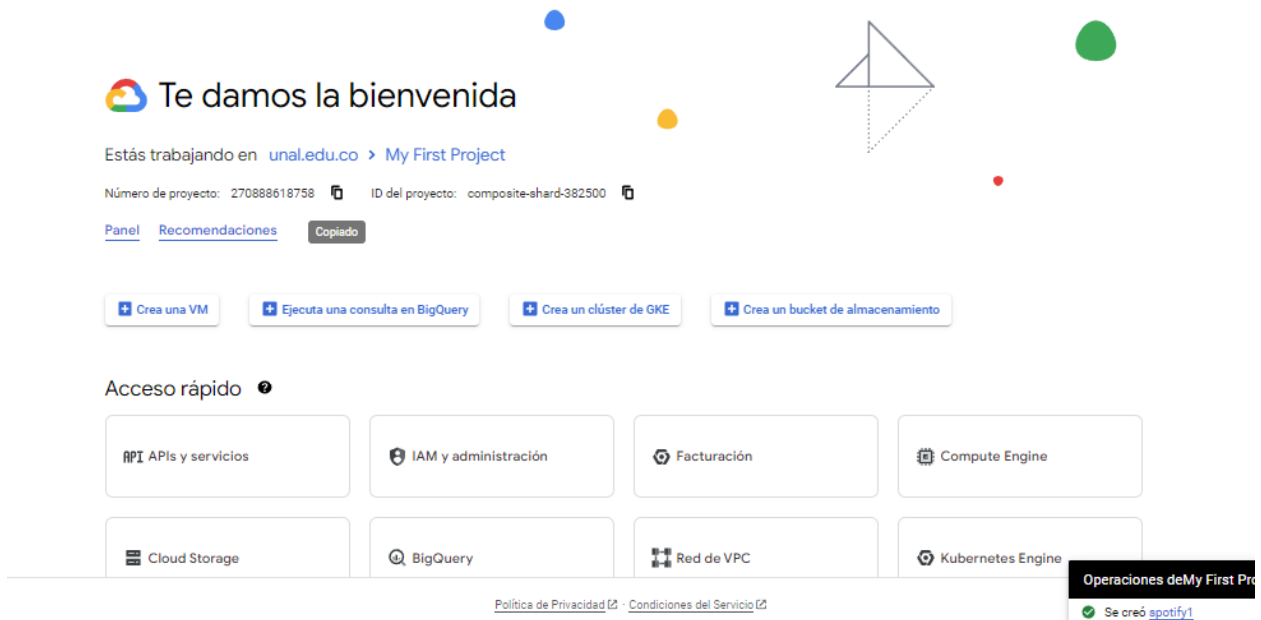
Configuración de la instancia



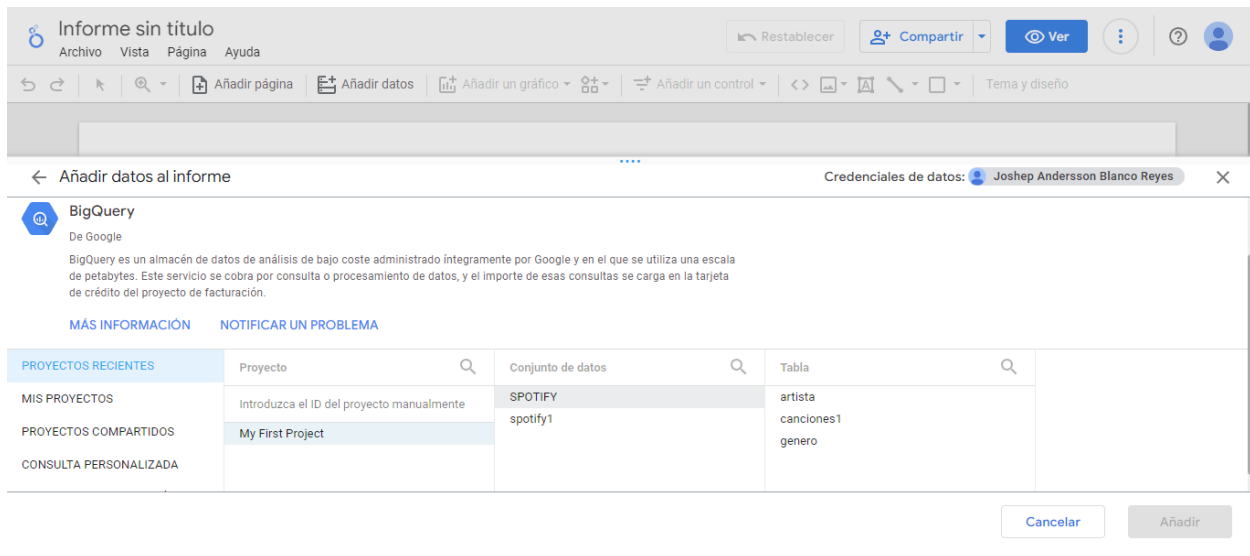
Creación BBDD Big Query



Obtener Id del proyecto:



ANEXO 3. Conexión de la base de datos en GCP a Google DataStudio (Visualizador.)



The screenshot shows the Google Data Studio interface. At the top, there's a header bar with the title 'Informe sin título' and navigation links like 'Archivo', 'Vista', 'Página', and 'Ayuda'. Below this is a toolbar with various icons for adding elements like pages, data, charts, and controls. The main content area is titled 'Añadir datos al informe' (Add data to report). It features a 'BigQuery' section with a description of the service. Below this, there's a table with columns for 'Proyecto' (Project), 'Conjunto de datos' (Dataset), and 'Tabla' (Table). The 'Proyecto' column has a search bar and a list of projects, including 'My First Project'. The 'Conjunto de datos' column has a search bar and a list of datasets, including 'SPOTIFY' and 'spotify1'. The 'Tabla' column has a search bar and a list of tables, including 'artista', 'canciones1', and 'genero'. At the bottom right, there are 'Cancelar' (Cancel) and 'Añadir' (Add) buttons.

PROYECTOS RECIENTES	Proyecto	Conjunto de datos	Tabla
MIS PROYECTOS	Introduzca el ID del proyecto manualmente	SPOTIFY	artista
PROYECTOS COMPARTIDOS	My First Project	spotify1	canciones1
CONSULTA PERSONALIZADA			genero