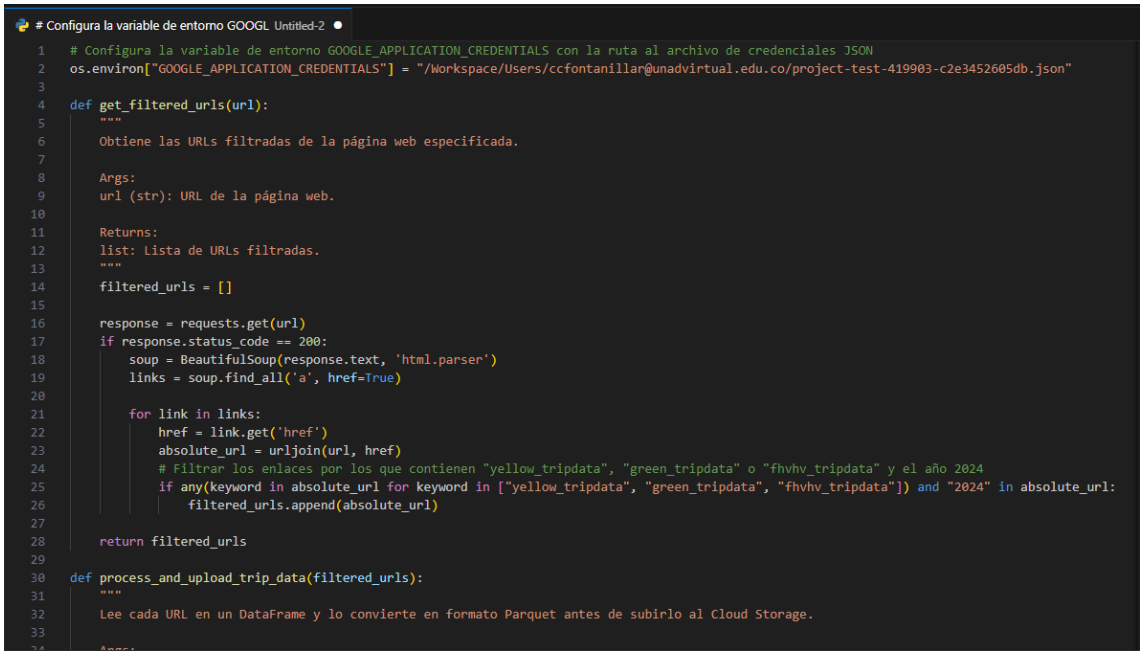
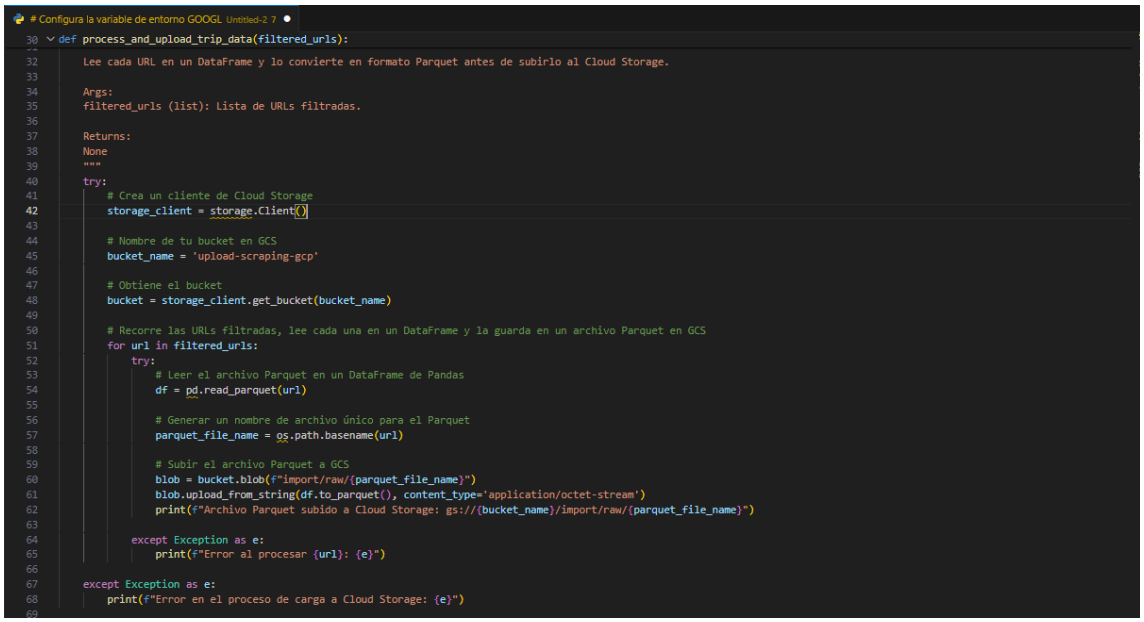
# Documentación Ingeniería de Datos

## Extracción y carga de datos

El objetivo era extraer los datos de la pagina NYC Taxi & Limousine Commission, del periodo de 2018 hasta 2023 para Yellow taxi, Green taxi y For-Hire Vehicle para Green y Yellow taxi 2018 a 2023 y para For-Hire Vehicle los años 2022 a 2023


Para lograr este objetivo decidimos usar un enfoque de web scraping utilizando la biblioteca de Python Beautiful Soup para extraer los nuevos enlaces disponibles en la pagina creamos el siguiente código:





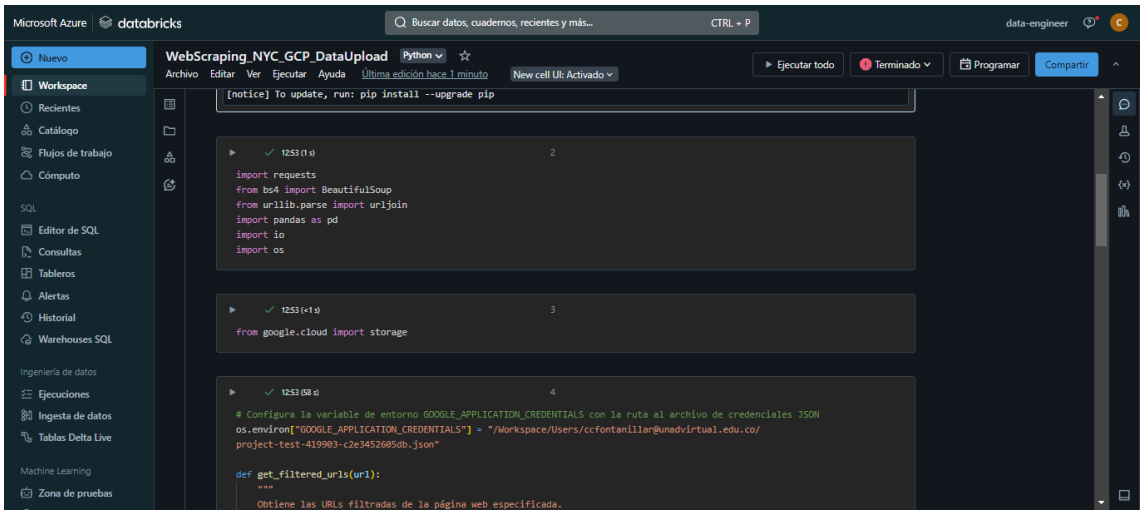

Este código consiste en dos funciones principales: `get_filtered_urls(url)` y `process_and_upload_trip_data(filtered_urls)`.

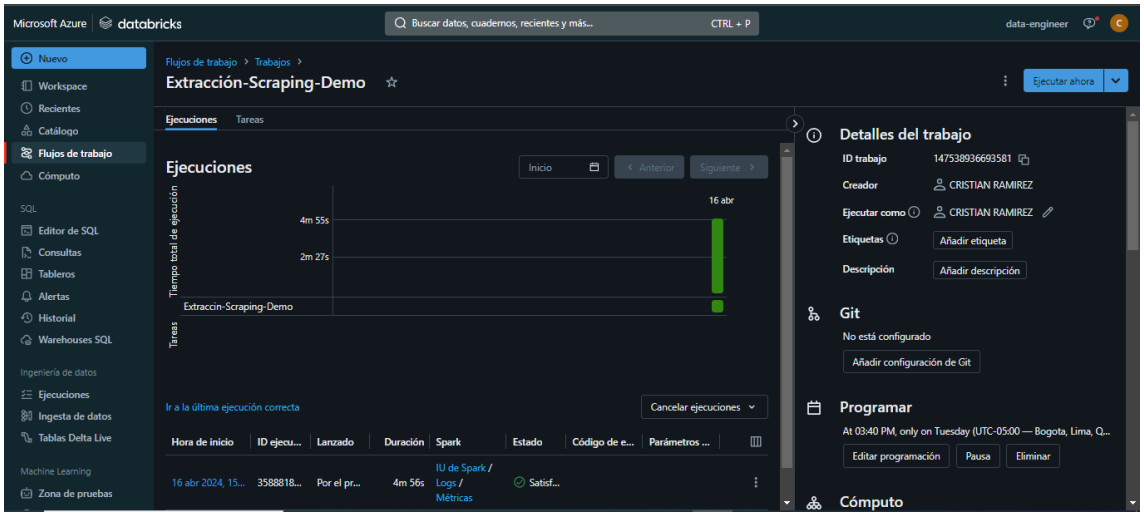
La función `get_filtered_urls(url)` toma una URL como entrada y utiliza bibliotecas como `requests` y `BeautifulSoup` para obtener el contenido HTML de la página web y analizarlo en busca de enlaces. Luego, filtra estos enlaces según ciertos criterios, como la presencia de palabras clave específicas y el año 2024 en la URL, y devuelve una lista de URLs que cumplen con estos criterios.

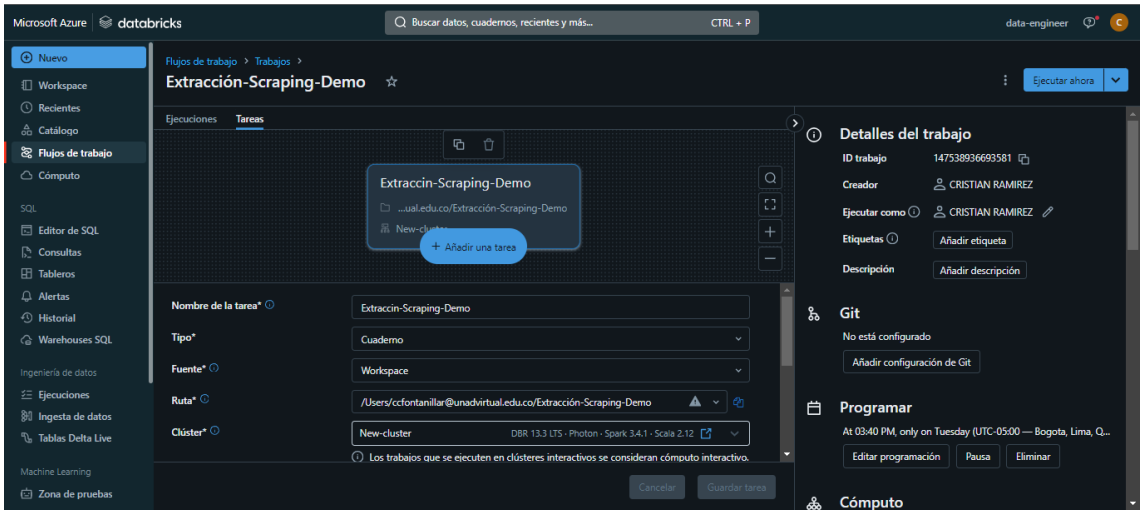
La función `process_and_upload_trip_data(filtered_urls)` toma la lista de URLs filtradas como entrada. Itera sobre cada URL, lee su contenido como un DataFrame de Pandas (asumiendo que son archivos Parquet), y luego los sube al Google Cloud Storage en una ubicación específica dentro de un bucket especificado. Esta función maneja excepciones para capturar cualquier error que pueda ocurrir durante el proceso y muestra mensajes de error apropiados.

Además de estas funciones, el código incluye un bloque principal que define la URL del sitio web del cual se desean extraer los datos. Luego, utiliza las funciones mencionadas anteriormente para realizar el proceso completo de obtención, procesamiento y carga de datos.

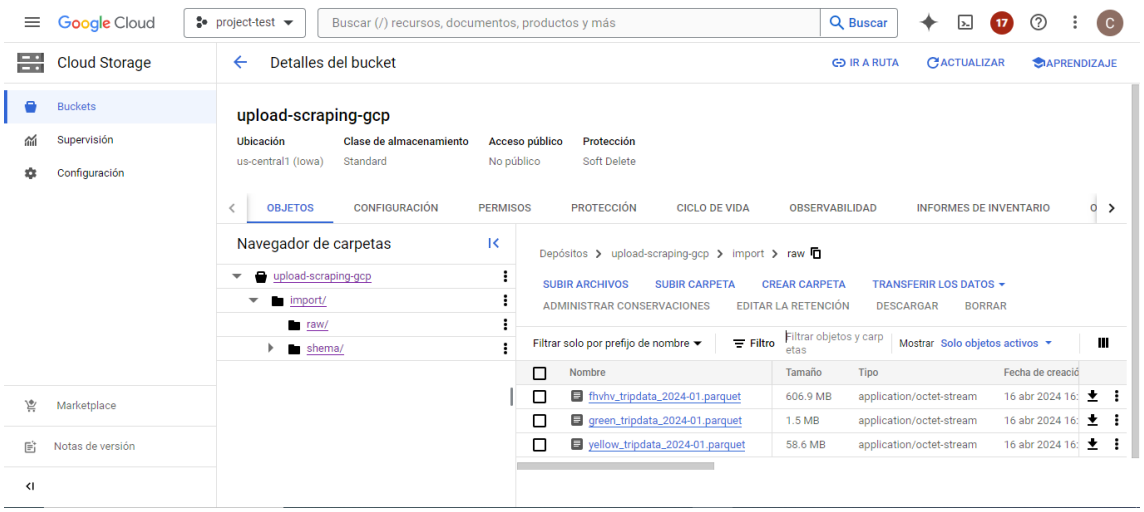
Después de crear la lógica creamos un Workspace en Databricks donde creamos un nuevo notebook y se hizo las configuraciones necesarias para la producción del código.




Tarea programada




Archivos subidos a GCS

## Justificación del uso de la nube

La decisión de utilizar los servicios en la nube de Google Cloud Platform (GCP), específicamente Google Cloud Storage, Cloud Functions y BigQuery, dentro de nuestro proyecto de Data Science ha sido fundamentada en una serie de consideraciones estratégicas y prácticas.

En primer lugar, la flexibilidad y escalabilidad inherentes a la nube son aspectos importantes para un proyecto de esta naturaleza. Al optar por servicios en la nube, podemos adaptar fácilmente nuestros recursos según las demandas cambiantes del proyecto, sin incurrir en costos adicionales significativos por infraestructura subutilizada o la necesidad de adquirir hardware adicional. Esta flexibilidad nos permite concentrar nuestros esfuerzos en la eficiencia y la optimización de los procesos de ingeniería de datos, en lugar de preocuparnos por la gestión y mantenimiento de la infraestructura subyacente.

Aprovechamos el servicio de Google Cloud Storage como almacenamiento temporal de datos no estructurados y crudos para después aplicar un preprocesamiento que permitiera aplicar ETL lo mismo como destino de los datos de scraping.

Además, el uso de Cloud Functions nos permite automatizar el procesamiento de datos en tiempo real de manera eficiente. Al recibir archivos Parquet en nuestro almacenamiento, podemos activar eventos que desencadenen procesos de transformación (ETL) y carga de datos en BigQuery de manera ágil y eficaz. Esto garantiza una actualización constante de nuestra base de datos, lo que es fundamental para la toma de decisiones informada y el análisis en tiempo real.

Por otro lado, la elección de BigQuery como nuestro almacén de datos principal se debe a su capacidad para gestionar conjuntos de datos a gran escala con un rendimiento excepcional. Con la capacidad de crear y manipular dataframes como el proyecto_taxis, que integra diversas tablas relacionales como climas, economía de autos, modelos aprobados, entre otros, podemos realizar consultas complejas de manera eficiente y

rápida. Esto es fundamental para la generación de insights y la construcción de modelos de machine learning, ya que nos permite acceder y analizar grandes volúmenes de datos de forma oportuna y efectiva.

Además, al aprovechar las herramientas y servicios integrados de GCP, podemos construir pipelines de datos robustos y confiables que automatizan el flujo de datos desde su origen hasta su destino final. Esto no solo aumenta la eficiencia operativa, sino que también garantiza la integridad y la calidad de los datos, lo que es crucial para la generación de informes precisos y la toma de decisiones fundamentadas.