Instructivo de Instalación

**Antes de comenzar el instructivo debemos iniciar sesión con nuestro correo institucional de DuocUC.**

Damos clic en la opción de arriba a la izquierda

Interfaz de usuario gráfica

Descripción generada automáticamente

Vamos a la opción “TODOS”Interfaz de usuario gráfica

Descripción generada automáticamente

Seleccionamos nuestro proyecto

Interfaz de usuario gráfica, Tabla

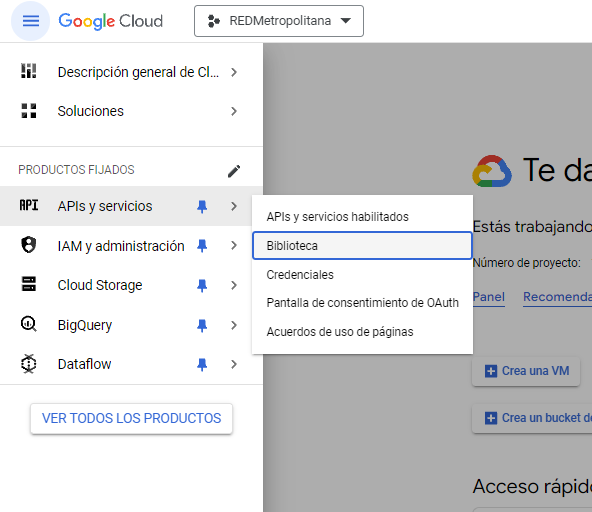
Descripción generada automáticamente

Ahora nos debería aparecer así:

Interfaz de usuario gráfica

Descripción generada automáticamente

Luego abrimos el panel lateral izquierda que despliega un menú de acceso directo a los diferentes productos de la consola de Google. En particular, debemos poner el mouse sobre el producto que dice “APIs y servicios” y luego hacer clic en donde dice “Biblioteca”



Y en el buscador vamos a buscar “Dataflow API” Texto

Descripción generada automáticamente

Hacemos clic en la APIInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Le damos clic en “HABILITAR” Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Una vez habilitada la primera API, nos devolvemos y repetimos el mismo proceso para las siguientes APIs:

- Cloud Logging API

- Compute Engine API

- Data Pipelines API

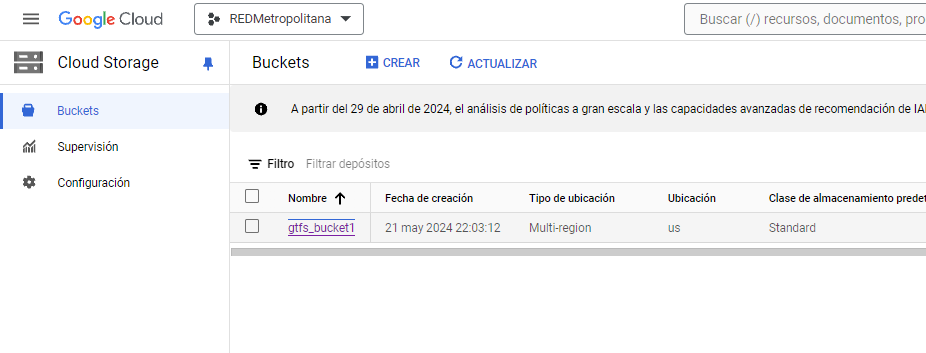
- Cloud Resource Manager API

Luego de habilitar las APIs necesarias, nos dirigimos a Cloud Storage, Buckets:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Y hacemos clic en el bucket llamado “gtfs\_bucket1”



Ahora, para subir los archivos .py que ejecutarán los pipelines, vamos a hacerle clic en donde dice “SUBIR ARCHIVOS”, tal como aparece en la siguiente imagen:

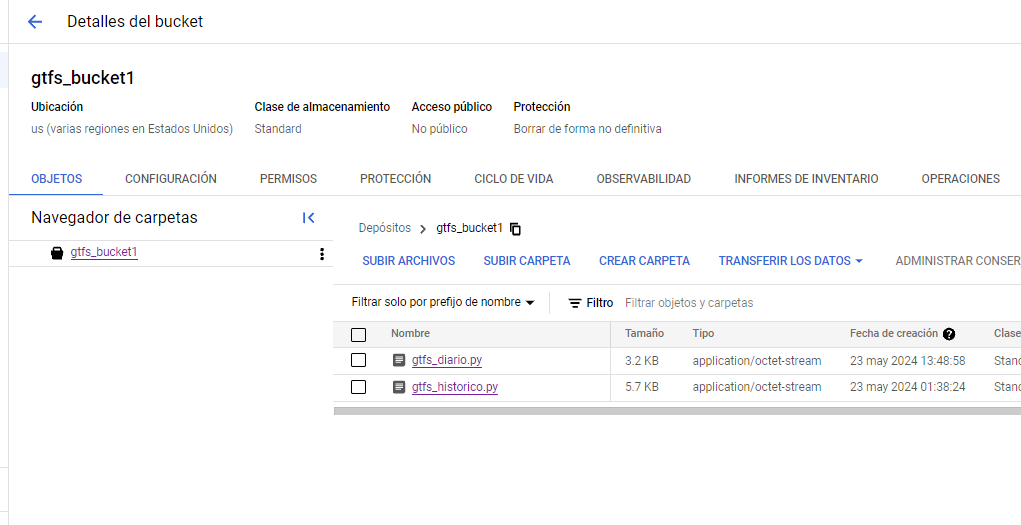
Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Seleccionamos los archivos para subirlos

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Y deberíamos ver algo así 

Bien, entonces ahora abrimos la terminal Cloud Shell para posteriormente ejecutar los comandos necesarios para ejecutar los pipelines.

Primero hacemos clic en este ícono de arriba a la derecha de nuestra consolaInterfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Y veremos nuestra terminal abierta, justo como se muestra en la siguiente imagen:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Y ahora deberá ejecutar la siguiente lista de comandos (uno por uno):**

====================================================================

*#configuración del entorno.*

gcloud auth login

gcloud config set project redmetropolitana-423718

gcloud config set compute/region us-central1

python3 --version

python3 -m venv myenv

source myenv/bin/actívate

====================================================================

*#instalaciones para el ambiente virtual.*

pip install --upgrade pip setuptools \

pip install apache\_beam \

pip install apache\_beam[gcp] \

pip install beautifulsoup4

====================================================================

*#copiar los archivos Python desde el bucket al ambiente virtual*

gsutil cp gs://gtfs\_bucket1/gtfs\_historico.py .

gsutil cp gs://gtfs\_bucket1/gtfs\_diario.py .

====================================================================

**Para ejecutar el primer pipeline, Datos históricos:**

python gtfs\_historico.py \

--project redmetropolitana-423718 \

--runner DataflowRunner \

--job\_name gtfs-datos-historicos \

--staging\_location gs://gtfs\_bucket1/staging \

--temp\_location gs://gtfs\_bucket1/temp \

--output\_prefix gs://gtfs\_bucket1/Datos\_Historicos \

--input\_url https://us-central1-duoc-bigdata-sc-2023-01-01.cloudfunctions.net/datos\_transporte\_et

====================================================================

**Para ejecutar el segundo pipeline, Datos Diarios:**

python gtfs\_diario.py