

Manual de conexión entre BigQuery (Google Cloud) y PowerBI (Azure Cloud) pasando por DataBricks.

Descripción del objetivo del manual

Visualizar en PowerBI los datos abiertos relacionados con el COVID-19 almacenados en Google Cloud (GCP)

Descripción general del proceso

El proceso a que será descrito de manera detallada a continuación, genera un flujo de datos que tiene origen en un conjunto de datos abiertos de google (<https://console.cloud.google.com/marketplace/product/bigquery-public-datasets/covid19-public-data-program>), el cuál es capturado en un proyecto de BigQuery herramienta de Google Cloud, desde dónde es extraído a través de la herramienta Azure Dataflow bajo autenticación OAuth 2.0 (Claves generadas desde la consola de desarrolladores), estos datos terminan almacenados en una Blob Storage de Azure (dentro de su respectivo contenedor). Desde esa ubicación son cargados a un cluster de Azure Databricks usando una conexión JDBC, posteriormente se les realizan algunos procesamientos y consultas para finalmente a través de un conector de PowerBI visualizar la información resultante.

¿Qué vamos a necesitar?

1. Cuenta de Google Cloud, se puede crear en: <https://cloud.google.com/free>
2. Cuenta de Microsoft Azure, se puede crear en: <https://my.visualstudio.com>
3. Cuenta de Postman, se puede crear en: <https://www.postman.com/>
4. Power Bi Desktop, se puede descargar de: <https://powerbi.microsoft.com/es-es/downloads/>

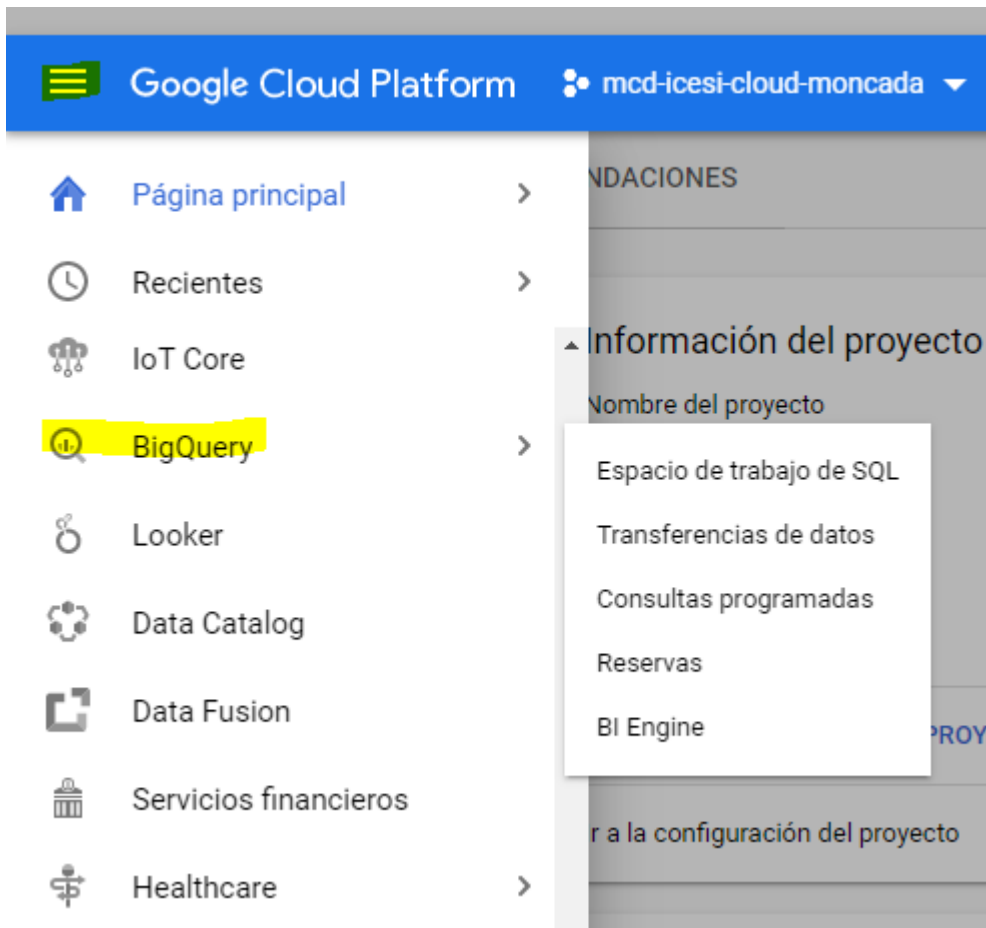
Si es la primera vez que se crean las cuentas relacionadas con las nubes (Azure y Google Cloud) se reciben créditos gratuitos para usar en los servicios.

Instrucciones

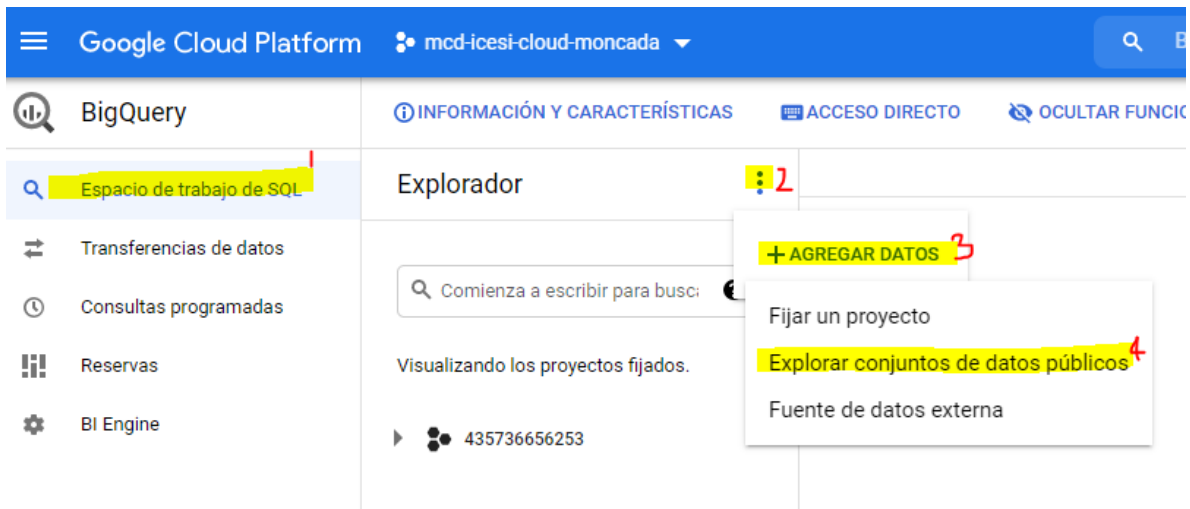
Estas instrucciones, asumen que las cuentas han sido creadas antes de iniciar el proceso. Las herramientas están configuradas en español por lo que si su versión está en inglés podrían diferir los nombres.

Parte 1 (Preparación de los datos en BigQuery)

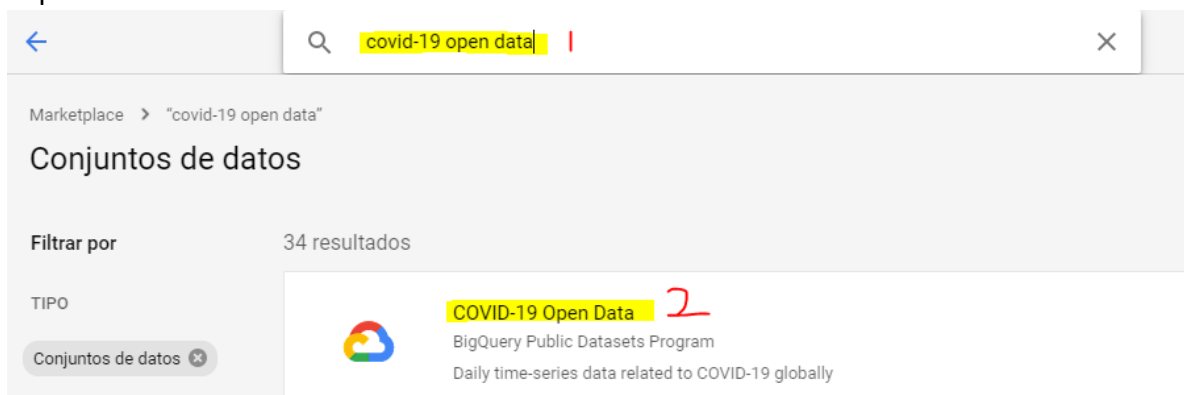
- a. Estando ubicado en la consola de google cloud (<https://console.cloud.google.com/>), procesa a realizar el acceso a la herramienta BigQuery, la imagen a continuación describe el ingreso (las zonas marcadas en amarillo, muestran los puntos de contacto).



- b. Una vez adentro de BigQuery, seleccione en el menú izquierdo la opción marcado como espacio de trabajo de SQL, una vez presionado se despliega un espacio marcado como Explorador, en los tres puntos verticales ubicados en el lado superior izquierdo del Explorador presione (+AGREGAR DATOS) y de ahí en la opción (Explorar conjuntos de datos públicos).



- c. La acción del paso (b) genera el despliegue de una nueva ventana, una vez ahí introduzca el texto: “covid-19 open data” en la barra de búsqueda y seleccione el resultado etiquetado como: “COVID-19 Open Data”, la imagen a continuación muestra el proceso:



- d. En la ventana desplegada tras la selección del paso (c), presione en el botón “Ver Conjunto de Datos”. Esto generará que regrese a la ventana anterior, pero en su explorador se habrá fijado un nuevo proyecto (ver paso e).



COVID-19 Open Data

BigQuery Public Datasets Program

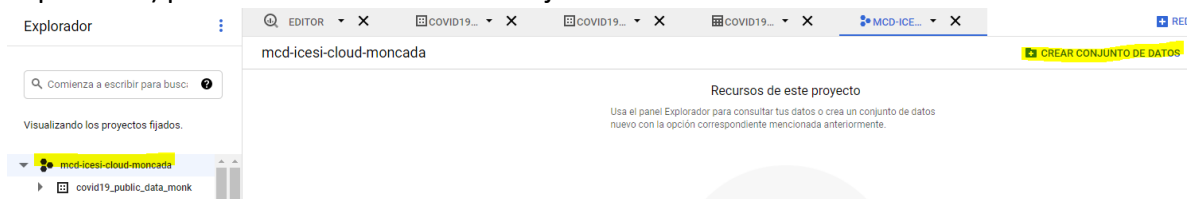
Daily time-series data related to COVID-19 globally

[VER CONJUNTO DE DATOS](#)

DESCRIPCIÓN GENERAL

EJEMPLOS

- e. Ahora vamos a crear un conjunto de datos en nuestro proyecto, sobre el cuál copiaremos los datos del proyecto de datos abiertos
- f. Seleccione su proyecto y posteriormente en el área de trabajo (a la derecha del explorador) presione el botón “Crear Conjunto de Datos”



- g. En la ventana lateral que despliega el botón, indique el nombre del conjunto de datos, marque predeterminada, sin vencimiento y clave administrada por google. Finalmente presione “Crear conjunto de datos” en la parte inferior.

Crear conjunto de datos

ID de conjunto de datos

Puede incluir letras, números y guiones bajos

Ubicación de los datos (Opcional) ?

Predeterminada

Vencimiento predeterminado de la tabla ?

☒ Nunca

☐ Cantidad de días después de la creación de la tabla:

Encriptación

Los datos se encriptan automáticamente. Selecciona una solución de administración de claves de encriptación.

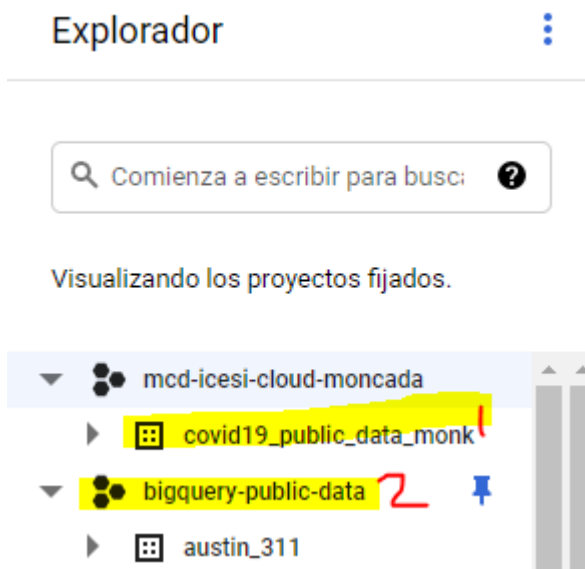
☒ Clave administrada por Google

No se requiere configuración

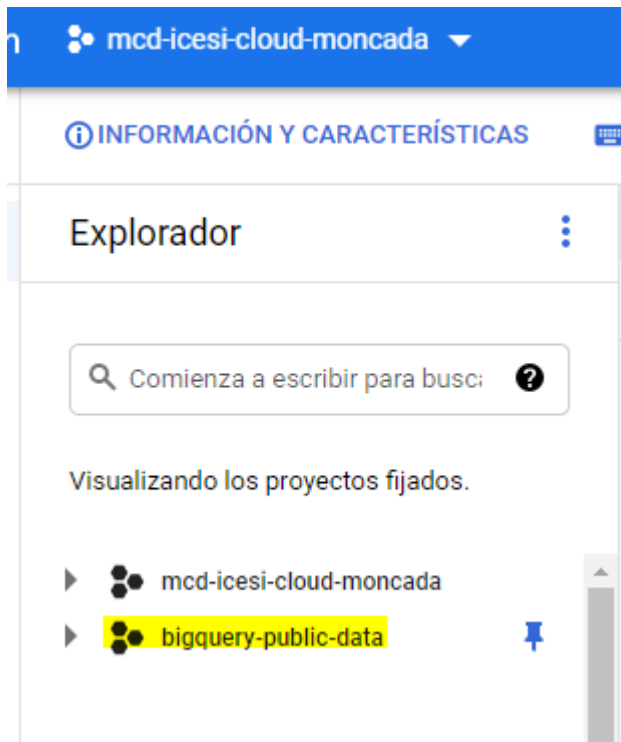
☐ Clave administrada por el cliente

Administrar mediante Google Cloud Key Management Service

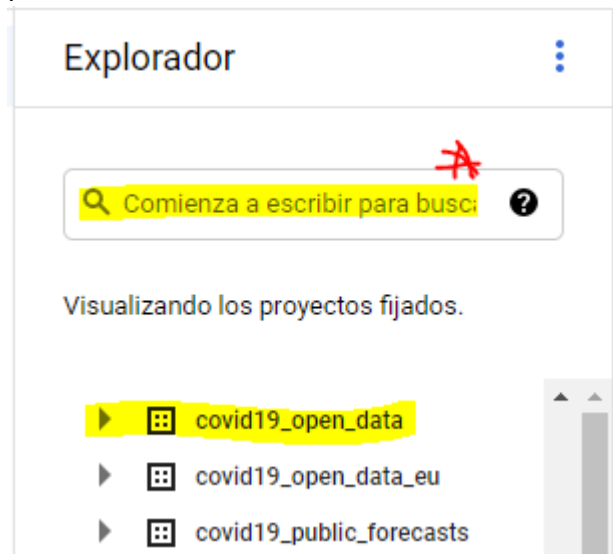
- h. La acción previa generó un conjunto de datos asociados a su proyecto (1 en la siguiente imagen), pero por ahora no tiene ninguna tabla, vamos ahora a traer los datos del proyecto de datos abiertos (2 en la siguiente imagen).



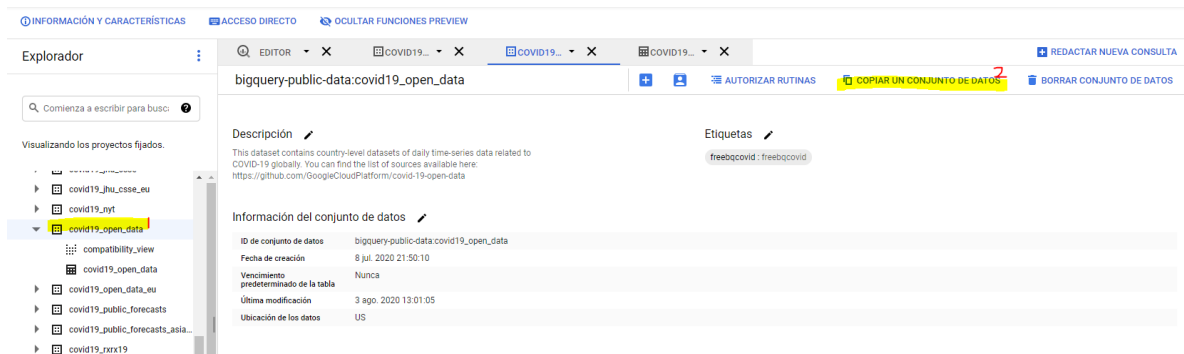
- i. En la ventana del explorador ahora usted tiene además de su proyecto un proyecto llamado: “bigquery-public-data” (agregado en el paso (d)). Al desplegar este usted verá los conjuntos de datos disponibles:



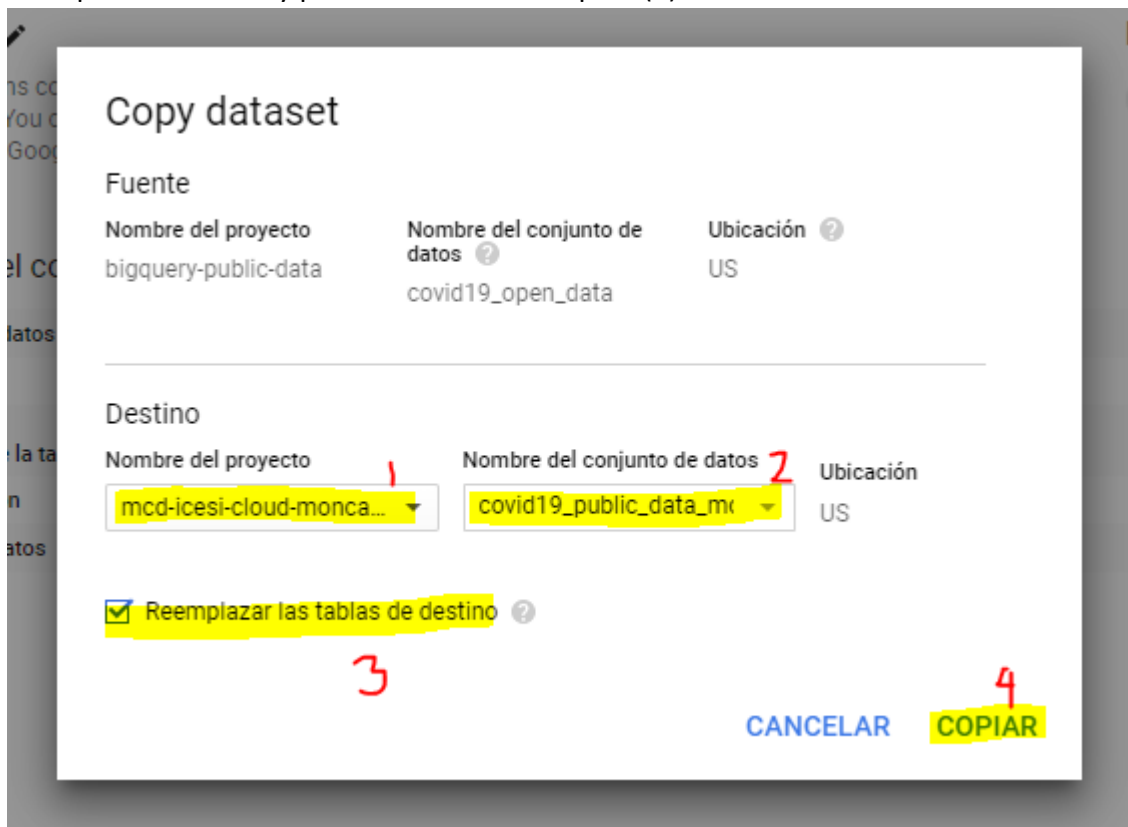
- j. Dentro de los datos desplegados en el paso (i), busque uno llamado: “covid_19_open_data”, si lo prefiere use este texto en el buscador para acelerar el proceso.



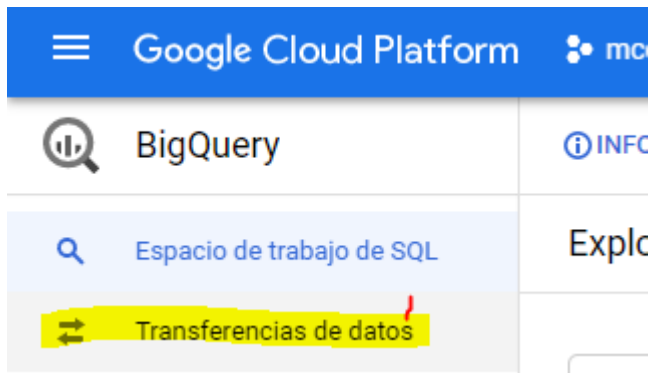
- k. Ahora necesitamos transferir estos datos del proyecto de datos abiertos al proyecto personal, para esto seleccione el conjunto de datos con el nombre indicado en el paso (j) y en la información desplegada en el área de trabajo (a la derecha del explorador), identifique y presione el botón con nombre “Copiar conjunto de datos” (2 en la imagen siguiente).



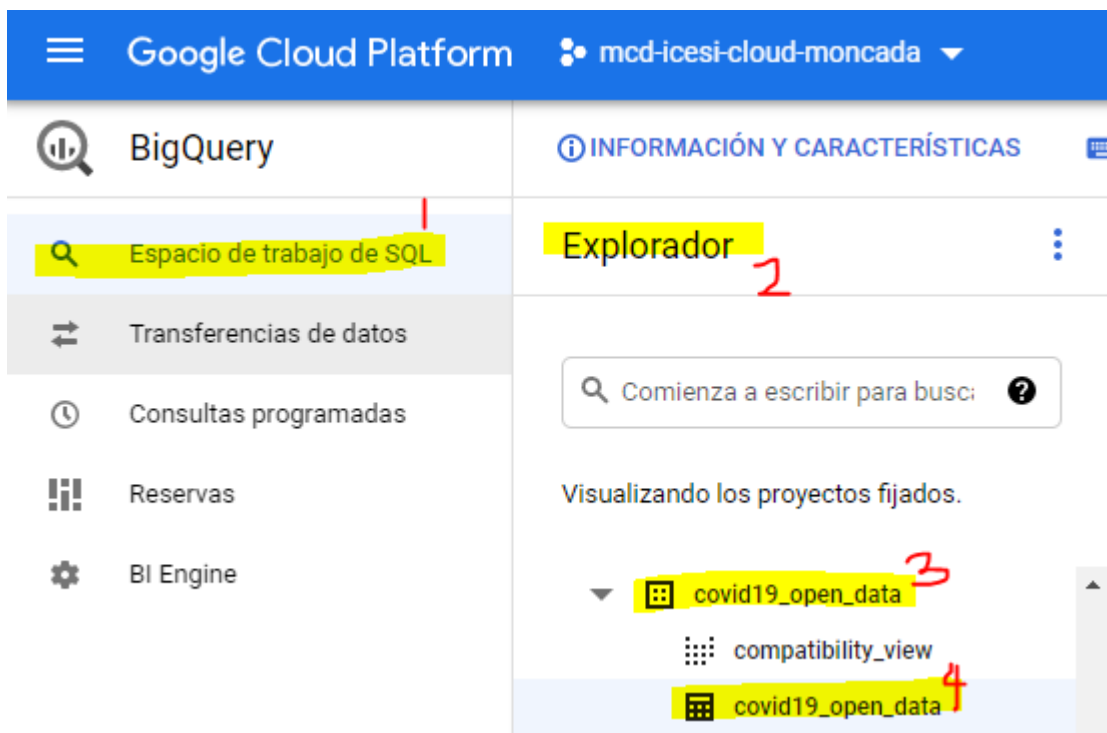
- l. El paso (k) ha desplegado una ventana emergente, seleccione en esta el nombre del proyecto(1) y el conjunto de datos en el cuál desea copiar(2), marque la opción de reemplazo de tablas y presione el botón “copiar”(4).



- m. Hecho el paso previo vamos a copiar la tabla de datos, pero ahora debemos activar las transferencias de datos, esto se hace presionando en la opción “Transferencia de datos” ubicada en el lado izquierdo debajo de la opción “Espacio de trabajo de SQL”



- n. Con esta opción de transferencia seleccionada, presione activar en el API de transferencias y una vez completado el proceso proceda de regreso a la opción “Espacio de trabajo SQL” para realizar la copia de la tabla. (no funciona el paso siguiente si no se activa la API)
- o. Estando ubicado en el “Espacio de trabajo SQL” (1), seleccione en el “Explorador” (2) el proyecto de datos abiertos, y busque nuevamente el conjunto de datos “covid19_open_data” (3), seleccione la tabla “covid19_open_data” (4). La imagen a continuación muestra los números.



- p. Con la tabla seleccionada, presione el botón “copiar tabla” disponible en el lado derecho del área de trabajo en la parte superior.

- q. En la ventana emergente lanzada por el botón seleccione la opción “buscar un proyecto” (1), después seleccione su proyecto personal (2) y el conjunto de datos que creamos antes en su proyecto (3), finalmente introduzca el nombre de la tabla que se creará en su proyecto personal (4) y para terminar presione el botón copiar (5).

Copiar tabla

Fuente

Nombre del proyecto	Nombre del conjunto de datos	Nombre de la tabla
bigquery-public-data	covid19_open_data	covid19_open_data

Destino

☒ Buscar un proyecto (1) ☐ Ingresa un nombre de proyecto

Nombre del proyecto (2): mcd-icesi-cloud-moncada

Nombre del conjunto de datos (3): covid19_public_data_monk

Nombre de la tabla (4): nueva-tabla-personal

Opciones avanzadas ▼

CANCELAR COPIAR (5)

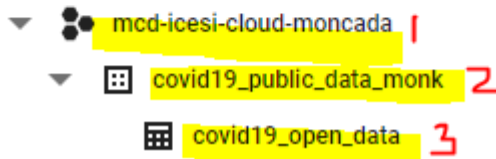
- r. La última acción debe dejar disponibles los datos dentro del proyecto personal (1), dentro de un conjunto de datos (2) y en la tabla con el nombre que le haya asignado (3).

Explorador



Comienza a escribir para buscar ?

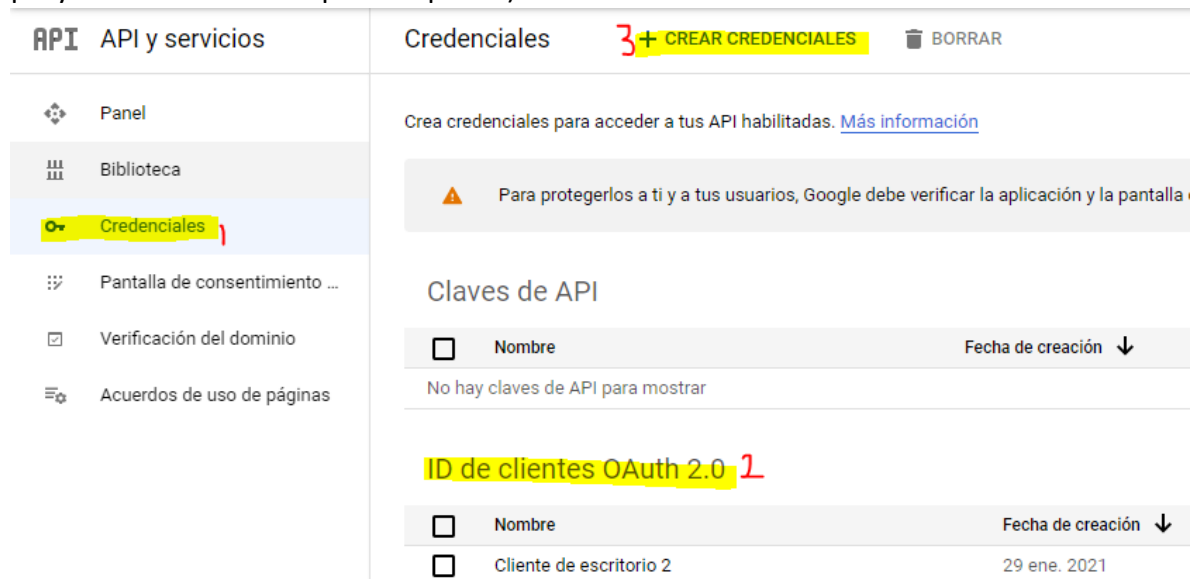
Visualizando los proyectos fijados.



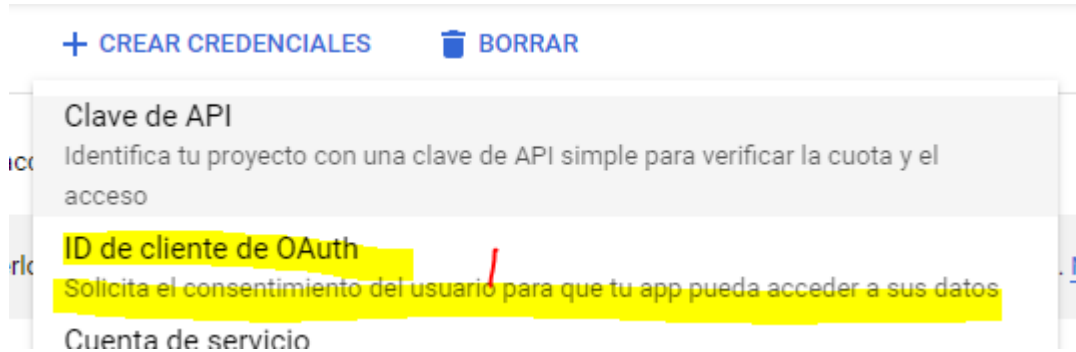
Parte 2 (Preparación de las credenciales de conexión entre GCP y Azure)

Durante esta parte vamos a usar la consola de desarrolladores de google (<https://console.developers.google.com/>) y una cuenta de Postman (<https://www.postman.com/>), el tutorial asume que ya han sido creadas las cuentas.

- Estando en la consola de google, en el menú lateral izquierdo seleccione la opción “Credenciales” (1), vamos a crear una credencial de tipo OAuth 2.0 (2) y para ellos debemos presionar en “+Crear Credenciales”. (Asegurarse de tener seleccionado el proyecto correcto en la parte superior)



- b. En el desplegable seleccione: "ID de cliente de OAuth"



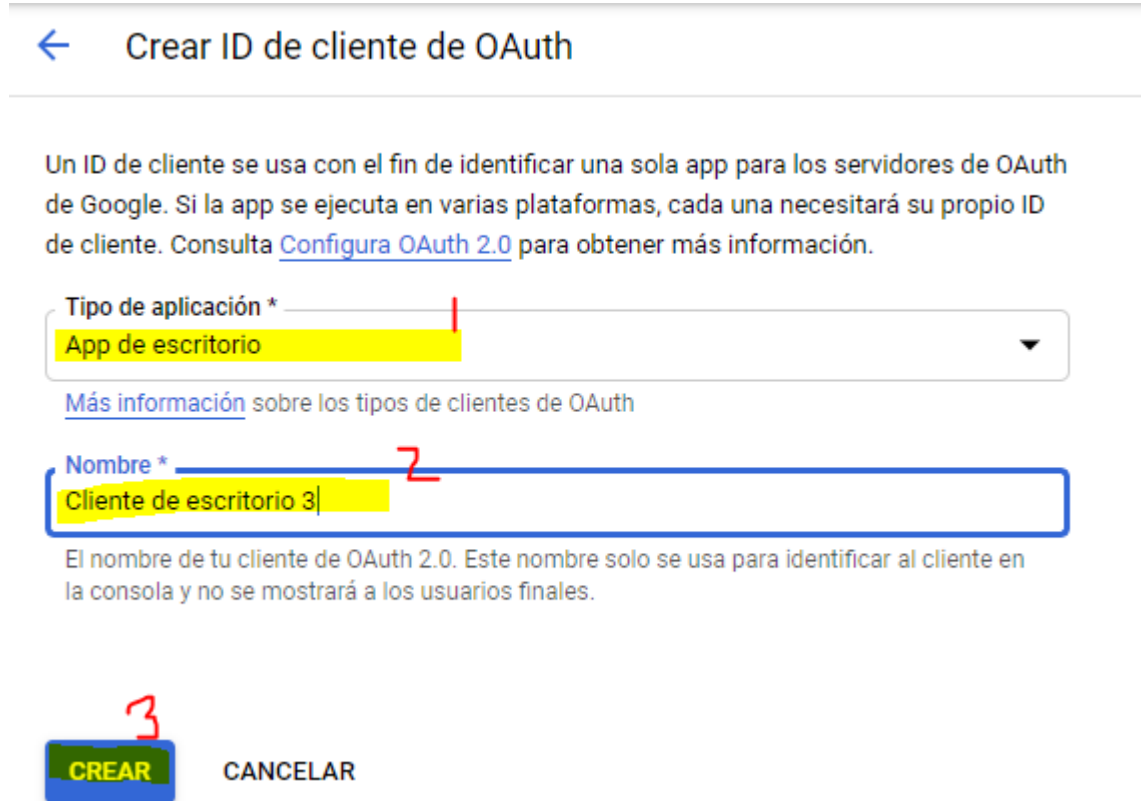
+ CREAR CREDENCIALES BORRAR

Clave de API
Identifica tu proyecto con una clave de API simple para verificar la cuota y el acceso

ID de cliente de OAuth
Solicita el consentimiento del usuario para que tu app pueda acceder a sus datos

Cuenta de servicio

- c. En la ventana seleccione como tipo de aplicación "App de escritorio" (1), después inserte el nombre que desee para la aplicación (2) y finalmente presione "crear" (3)



← Crear ID de cliente de OAuth

Un ID de cliente se usa con el fin de identificar una sola app para los servidores de OAuth de Google. Si la app se ejecuta en varias plataformas, cada una necesitará su propio ID de cliente. Consulta [Configura OAuth 2.0](#) para obtener más información.

Tipo de aplicación *
App de escritorio

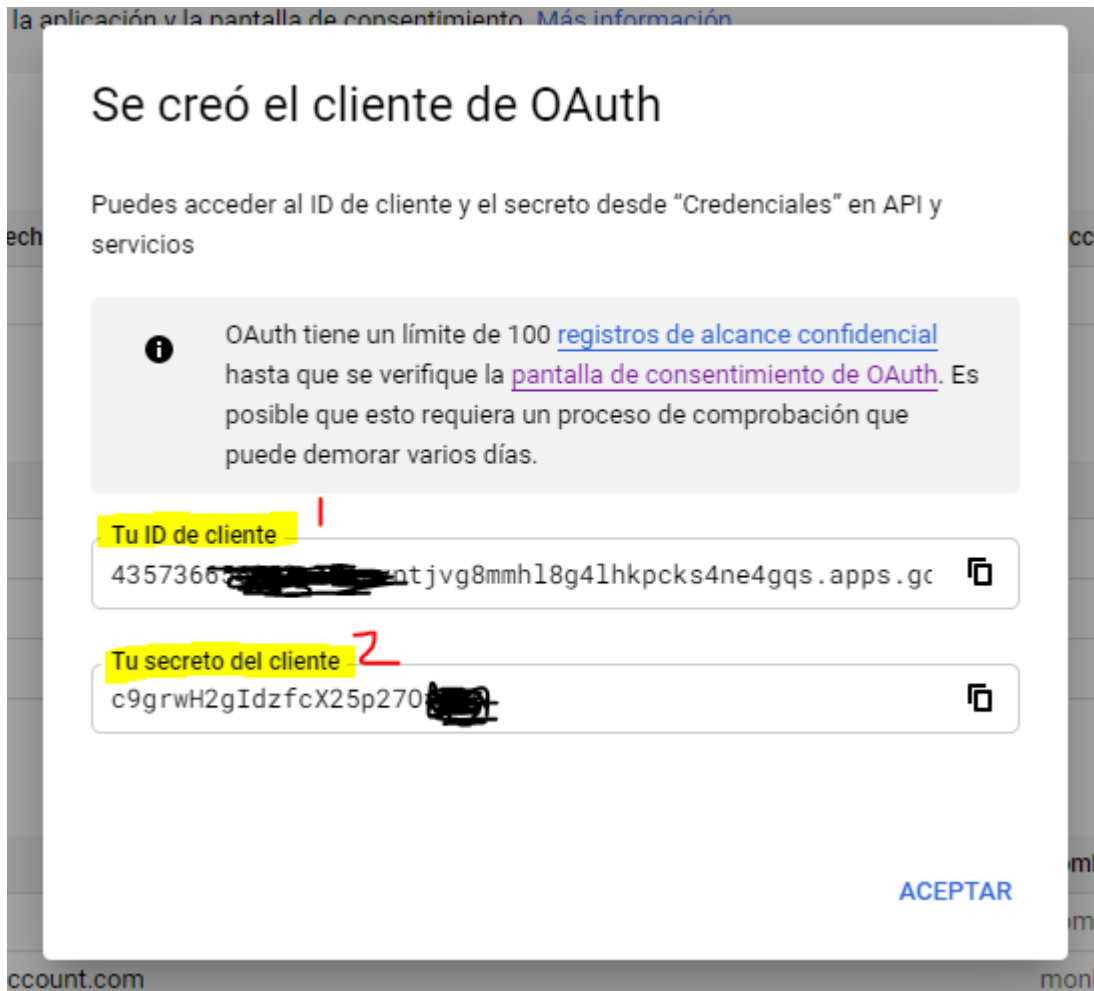
Más información sobre los tipos de clientes de OAuth

Nombre *
Cliente de escritorio 3

El nombre de tu cliente de OAuth 2.0. Este nombre solo se usa para identificar al cliente en la consola y no se mostrará a los usuarios finales.

CREAR CANCELAR

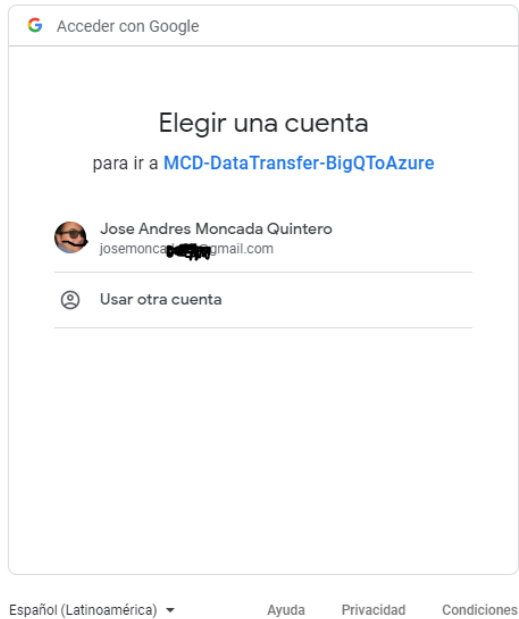
- d. En la ventana siguiente, debes copiar los datos de ID de cliente y secreto de cliente (déjalos en un bloc de notas), los vamos a necesitar más adelante. Si lo cerraste y no lo copiaste, puedes ver estos datos al presionar sobre el ID de cliente en la sección de credenciales.



- e. Con los códigos generados vamos a crear un token de autenticación, para ellos vamos a usar la siguiente dirección web, Ojo: Debes reemplazar: =<Your-client-Id> por el ID de cliente que acabas de copiar.

https://accounts.google.com/o/oauth2/v2/auth?client_id=<Your-client-Id>&redirect_uri=urn:ietf:wg:oauth:2.0:oob&state=GBQAUthTest&access_type=offline&scope=https://www.googleapis.com/auth/bigquery&response_type=code

- f. Al pegarlo en tu navegador, debes ver una entrada de cuenta, selecciona la cuenta Gmail que tiene el acceso a Google Cloud Platform (GCP). Accede con los datos de login.



- g. Si recibes un mensaje que dice “Google no verificó esta app”, debes presionar en configuración avanzada y en ir al app.



Google no verificó esta app¹

La app solicita acceso a información sensible de tu Cuenta de Google. Si el desarrollador (josemoncada87@gmail.com) aún no verificó esta app con Google, no deberías usarla.

Si eres el programador, envía una solicitud de verificación para quitar esta pantalla. [Más información](#)

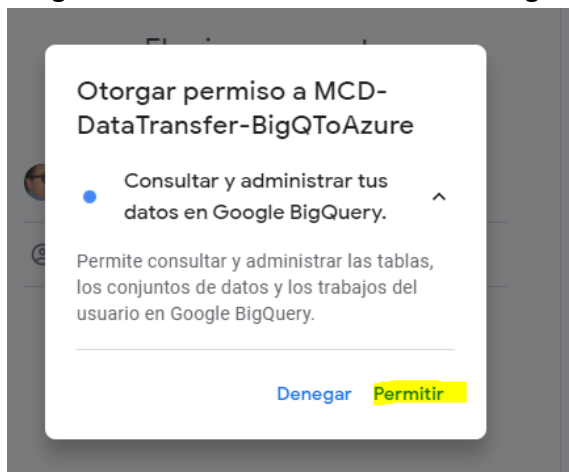
Ocultar configuración avanzada²

VOLVER A UN SITIO SEGURO

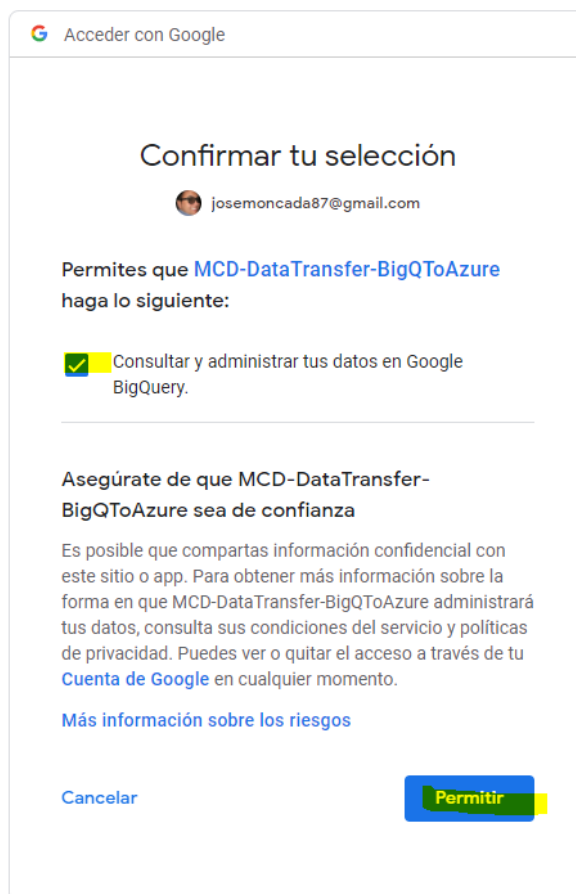
Continúa solo si entiendes los riesgos y confías en el desarrollador (josemoncada87@gmail.com).

Ir a MCD-DataTransfer-BigQToAzure (no seguro)³

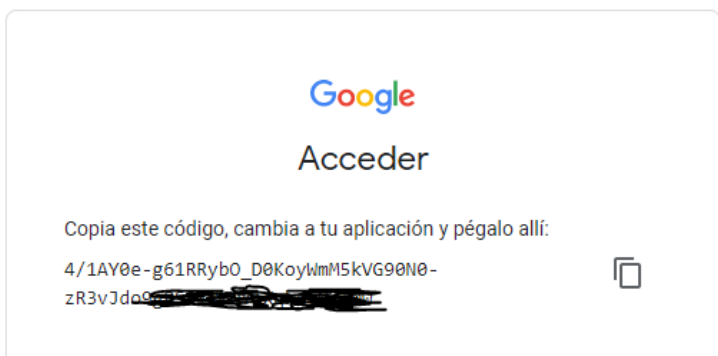
- h. Marcar permitir en la siguiente ventana, si los permisos de la imagen no coinciden asegúrate de haber creado la cuenta asignando los permisos al usuario.



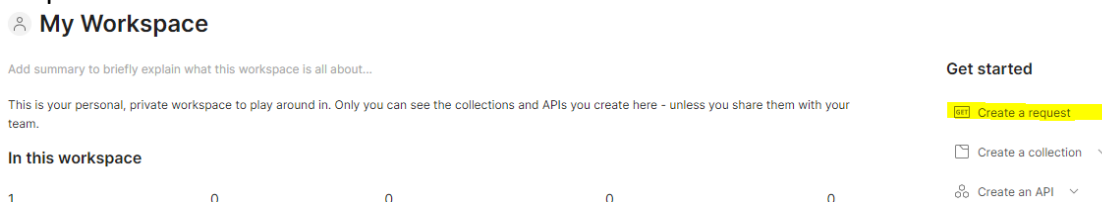
- i. Verificar la selección de: "Consultar y administrar tus datos en Google BigQuery" y después en permitir.



- j. Copia de la ventana siguiente el token correspondiente, este nos ayudará en la generación del código de refresco.

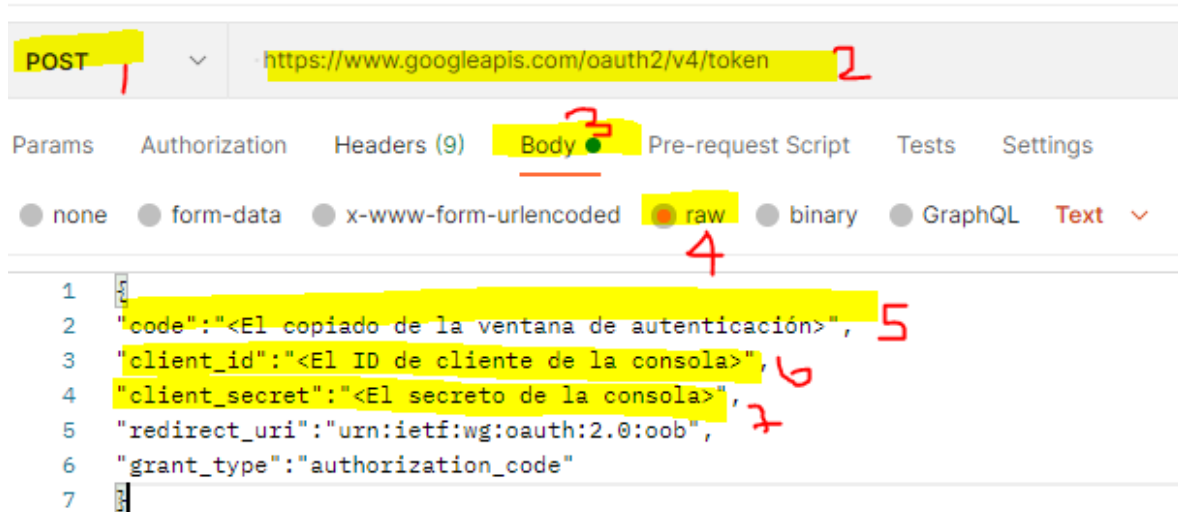


- k. Ahora vamos a abrir Postman, en la ventana principal seleccionar la opción “Create Request”:



- l. Para realizar la petición usaremos el método POST (1), con la dirección <https://www.googleapis.com/oauth2/v4/token> (2), seleccionamos la pestaña “Body” (3) en la opción “raw” (4) y completamos los datos (después de la imagen siguiente dejo la plantilla en texto), (5)(6)(7) son los datos que vienen de la consola de desarrolladores de google. Code es el de autenticación, id y secreto son los del cliente Auth2.0 creado (App de escritorio). Presionar “send” para terminar.

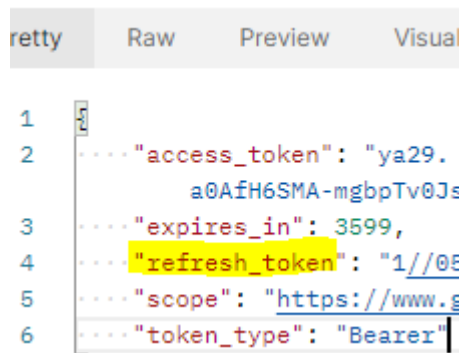
<https://www.googleapis.com/oauth2/v4/token>



Plantilla:

```
{
  "code": " <código> ",
  "client_id": " <Id> ",
  "client_secret": "< Secreto >",
  "redirect_uri": "urn:ietf:wg:oauth:2.0:oob",
  "grant_type": "authorization_code"
}
```

- m. En el resultado obtenido, copiar el “refresh_token”, si no se obtuvo y aparece un “bad_request” debe generar de nuevo el “code” haciendo el proceso de autenticación:



The screenshot shows a REST client interface with tabs for 'retty', 'Raw', 'Preview', and 'Visual'. The 'Raw' tab is selected, displaying a JSON response. The response contains the following fields: 'access_token' with a long alphanumeric string, 'expires_in' with the value 3599, 'refresh_token' with a value starting with '1//0', 'scope' with a URL, and 'token_type' with the value 'Bearer'.

```
1 {
2   "access_token": "ya29.
   a0AfH6SMA-mgbpTv0Js
3   "expires_in": 3599,
4   "refresh_token": "1//0
5   "scope": "https://www.g
6   "token_type": "Bearer"
```

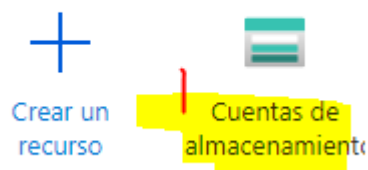
- n. Con esto completamos la información necesaria para ir a Azure Datafactory en el siguiente paso.

Parte 3 (Configuración del Azure DataFactory)

En este paso vamos a configurar un pipeline de Azure DataFactory, para que nos traiga los datos desde BigQuery, los datos quedarán alojados en un Blob Storage (también de Azure) dentro de su respectivo contenedor y asociado a una cuenta de almacenamiento.

- a. Vamos a crear una cuenta de almacenamiento, para eso vamos a la página principal de Azure <https://portal.azure.com/#home>, de ahí vamos a seleccionar la opción “Cuenta de almacenamiento”.






Servicios de Azure



- b. Dentro del gestor, presionamos “agregar”

Cuentas de almacenamiento

Universidad Icesi (@icesi.edu.co)

 Agregar  Administrar vista   Actualizar  Exp

Suscripción == **todo** Grupos

Mostrando de 1 a 2 de 2 registros.

- c. En la ventana que se lanza, vamos a insertar los datos. Seleccione primero una suscripción (1), después indique cuál es el grupo de recursos (2), si no tiene ninguno debe crear uno (2.1), después inserte el nombre de la cuenta de almacenamiento (3), seleccione la ubicación más cercana (4), marque Estándar en el rendimiento (5) y seleccione el tipo de cuenta en la versión más actual (6), seleccione el nivel de redundancia deseado (7) y presione “revisar y crear” (8).

Datos básicos

Redes

Protección de datos

Opciones avanzadas

Etiquetas

Revisar y crear

Azure Storage es un servicio administrado por Microsoft que proporciona almacenamiento en la nube altamente disponible, seguro, duradero, escalable y redundante. Azure Storage incluye Azure Blob (objetos), Azure Data Lake Storage Gen2, Azure Files, Azure Queues y Azure Tables. El costo de una cuenta de Storage depende del uso y de las opciones que elija a continuación. [Más información sobre las cuentas de almacenamiento de Azure](#)

Detalles del proyecto

Seleccione la suscripción para administrar recursos implementados y los costes. Use los grupos de recursos como carpetas para organizar y administrar todos los recursos.

Suscripción *

Azure subscription 1 1



Grupo de recursos *

2

Crear nuevo 2.1

Detalles de instancia

El modelo de implementación predeterminado es el de Resource Manager, que admite las últimas características de Azure. Como alternativa, puede elegir el modelo de implementación clásica. [Elegir el modelo de implementación clásica](#)

Nombre de la cuenta de almacenamiento

* ⓘ

3

Ubicación *

(US) Centro-Sur de EE. UU. 4

Rendimiento ⓘ

5



Estándar



Premium

Tipo de cuenta ⓘ

StorageV2 (uso general v2) 6

Replicación ⓘ

Almacenamiento con redundancia geográfica con acceso de lectura (RA-... 7

Revisar y crear 8

< Anterior




Siguiente: Redes >

- d. Dejamos los otros datos por defecto y al finalizar debemos ver la cuenta creada en el listado de cuentas de almacenamiento.

[Inicio](#) >

Cuentas de almacenamiento

Universidad Icesi (@icesi.edu.co)


[+ Agregar](#)  [Administrar vista](#)  [Actualizar](#) 

Filtrar por cualquier ca...

Suscripción == **todo**

Mostrando de 1 a 2 de 2 registros.



☐ **Nombre** ↑↓

☐  **cuentamonk**

- e. Ingresamos a la cuenta de almacenamiento (1) y nos disponemos a crear un contenedor para esto (3) en la información general (2). Presionamos sobre “contenedores”.

Cuentas de almace... <<

Universidad Icesi (@icesi.edu.co)

[+ Agregar](#)  [Administrar vista](#) 

Filtrar por cualquier campo...


Nombre ↑↓



 **cuentamonk** ...


 dbstorageeerb3x735djxqz ...


cuentamonk


Cuenta de almacenamiento





 **Información general** 


 Registro de actividad


 Etiquetas

 Diagnosticar y solucionar pro...

 Control de acceso (IAM)


 Transferencia de datos


 Eventos


 Explorador de Storage (versió...


Configuración




 Claves de acceso


 Replicación geográfica

 CORS

 Configuración

 Cifrado

 [Abrir en el Explorador](#)  [Mover](#) 

 Se ha anunciado la retirada de las alertas clásicas d información, consulte [Conservar las alertas con cui](#)

^ Información esencial

Grupo de recur... (cambiar) : [grupo-recursos-moi](#)



Estado : Principal: Disponible

Ubicación : Centro-Sur de EE. U

Suscripción (cambiar) : [Azure subscription 1](#)

Id. de suscripción : 10952968-3829-49f

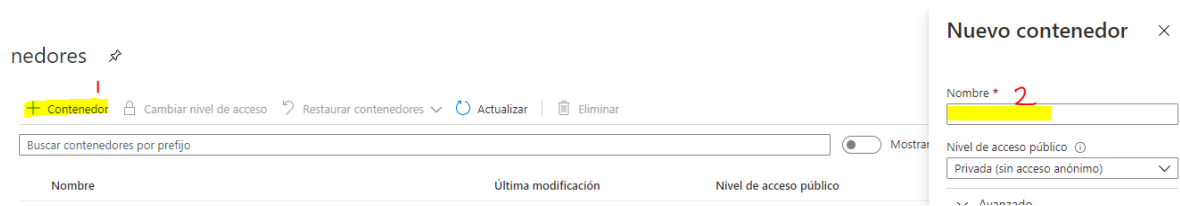
Etiquetas (cambiar) : [Haga clic aquí para .](#)

 **Contenedores** 

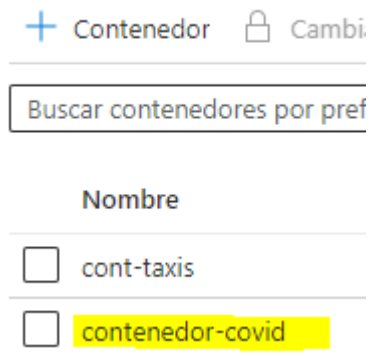
Almacenamiento escalable y rentable para datos no estructurados

[Más información](#)

- f. Vamos a crear un contendor, para esto presionamos en “+Contenedor” (1). En el costado derecho se despliega un lateral, para que insertemos el nombre del contenedor (2), el nivel de acceso puede quedar en “privado”. Finalizamos presionando el botón crear en la parte inferior del lateral.

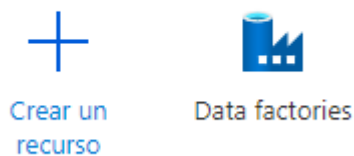


- g. Al finalizar este proceso debe verse un contenedor en el listado, con el nombre seleccionado.

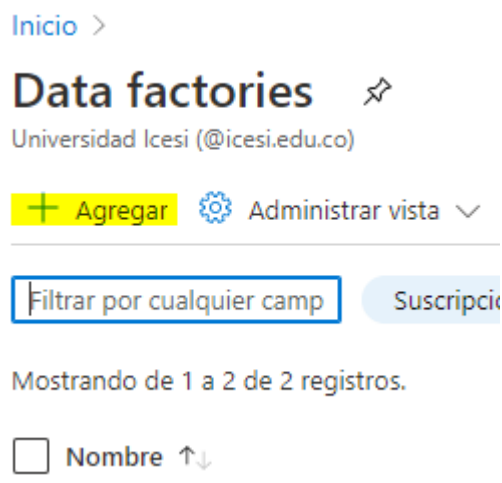


- h. Ahora que tenemos una cuenta de almacenamiento, un grupo de recursos y un contenedor. Vamos a comenzar con la configuración del pipeline. Para ellos vamos a DataFactory, para ingresar regresamos al home de Azure (<https://portal.azure.com/#home>) y de ahí seleccionamos DataFactories.

Servicios de Azure



- i. Al ingresar seleccionamos “+agregar” y eso nos lleva a la ventana de creación.



- j. En la ventana de creación, vamos a llenar los datos de la pestaña básico, seleccionamos la suscripción (1), el grupo de recursos que ya habíamos creado (2), asignamos una región (3) y un nombre y versión (4)(5).

Crear Data Factory

Básico Git configuration Networking Advanced Tags Revisar y crear

Detalles del proyecto

Seleccione la suscripción para administrar recursos implementados y los costes. Use los grupos de recursos como carpetas para organizar y administrar todos los recursos.

Suscripción * ⓘ Azure subscription 1 1

Grupo de recursos * ⓘ grupo-recursos-bigQueryToAzure 2
[Crear nuevo](#)

Detalles de la instancia

Región * ⓘ Centro-Sur de EE. UU. 3

Name * NombreDeDataFactory 4

Version * ⓘ V2 5

- k. Vamos a la pestaña de configuración de git y seleccionamos configurar después:

Básico **Git configuration** Networking Advanced Tags Revisar y crear

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration.

[Learn more about Git integration in Azure Data Factory](#)

Configure Git later ⓘ ☒ 2

- l. Presionamos revisar y crear, después nuevamente en el botón crear en la parte inferior.

[Inicio](#) > [Data factories](#) >

Crear Data Factory

✓ Validación superada

[Básico](#) [Git configuration](#) [Networking](#) [Advanced](#) [Tags](#) [Revisar y crear](#)

TÉRMINOS

Al hacer clic en "Crear", (a) acepto los términos legales y las declaraciones de privacidad relacionados con cada oferta de Marketplace que se enumeró previamente; (b) autorizo a Microsoft a facturar con mi método de pago actual las cuotas relacionadas con las ofertas, con la misma frecuencia de facturación que mi suscripción de Azure; y (c) autorizo a Microsoft a compartir mi información de contacto y los datos de transacción y uso con los proveedores de dichas ofertas. Microsoft no proporciona derechos sobre ofertas de terceros. Para obtener información adicional, consulte los [Términos de Azure Marketplace](#).

Básico

Suscripción	Azure subscription 1
Grupo de recursos	grupo-recursos-bigQueryToAzure
Región	Centro-Sur de EE. UU.
Name	NombreDeDataFactory
Version	V2

Networking

Connect via	Public endpoint
-------------	-----------------

Crear

< Anterior

Siguiente

[Descargar una plantilla para la automatización](#)

- m. La implementación tomará un tiempo, pero al finalizar debes presionar "ir al recurso".

✓ Se completó la implementación



Nombre de implementación: Microsoft.DataFactory-202101301935...

Suscripción: [Azure subscription 1](#)

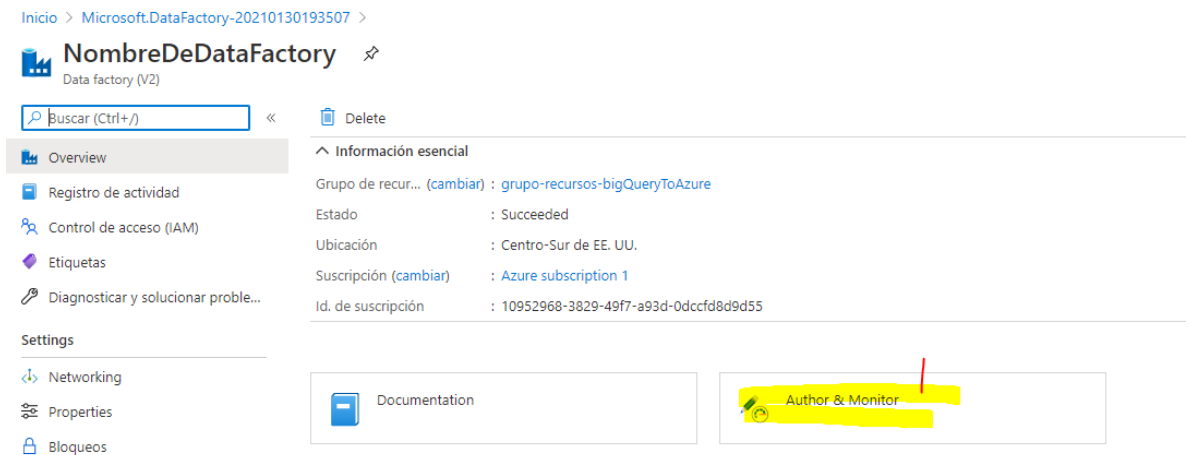
Grupo de recursos: [grupo-recursos-bigQueryToAzure](#)

∨ **Detalles de implementación** ([Descargar](#))

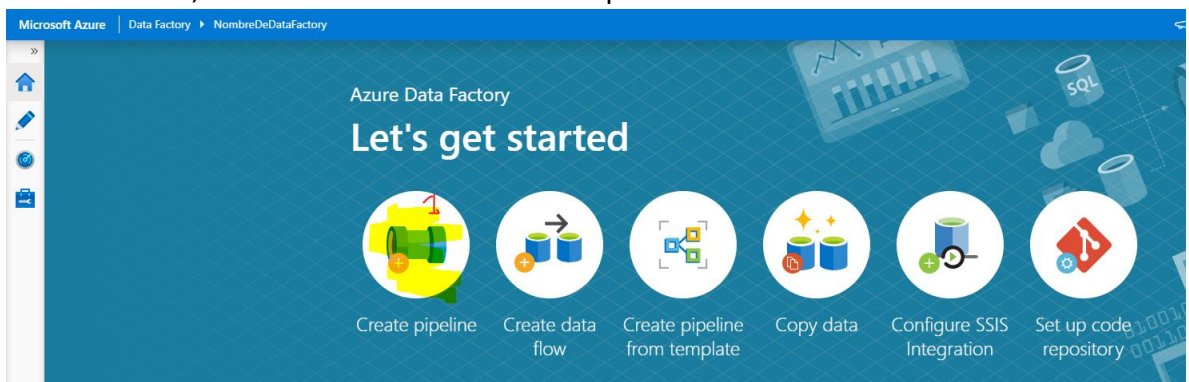
∧ **Pasos siguientes**

[Ir al recurso](#)

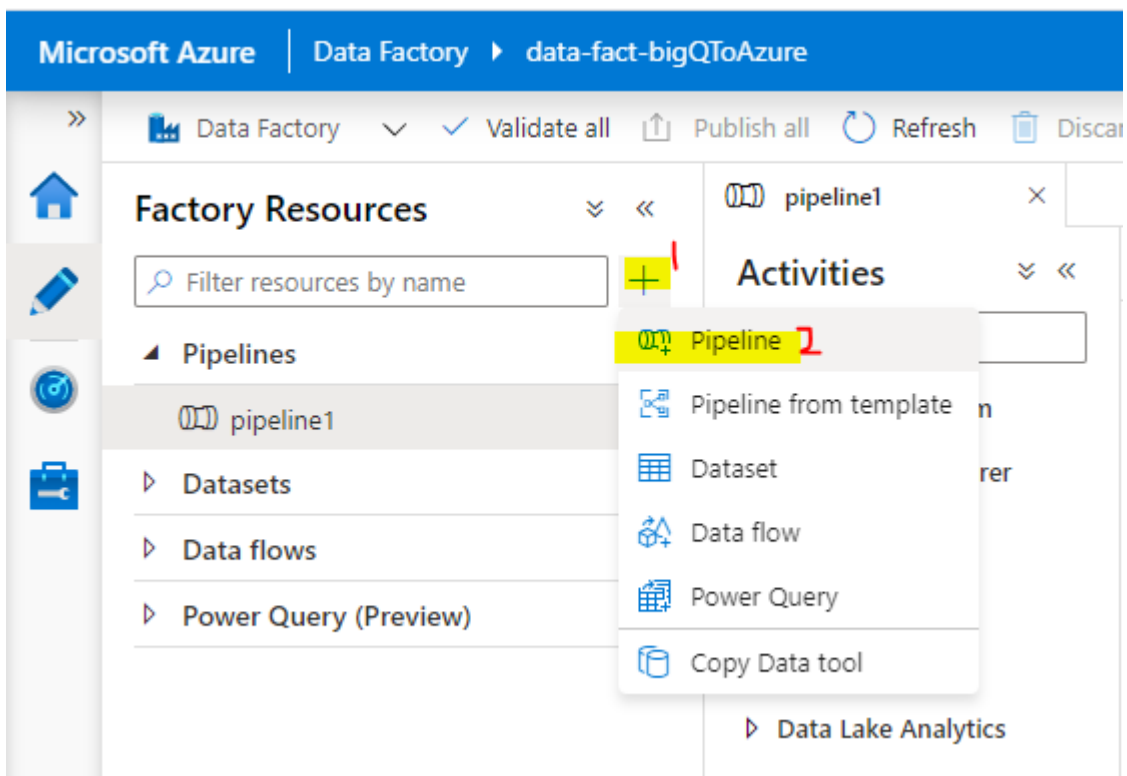
- n. Al entrar al recurso, debemos presionar en "Author & Monitor"



o. En la instancia, vamos a seleccionar “Create Pipeline”



p. En los recursos, presionamos + (1) y seleccionamos el pipeline (2)



- q. Después ingresamos las propiedades y listo.

Properties

General Related

Name *

Description

Concurrency ⓘ


Annotations
[+ New](#)


- r. Adicionamos un “CopyData” de las actividades (arrastrar al área de trabajo).

pipeline1 × pipeline2

Activities ≡ <<

▲ Move & transform

 Copy data

 Data flow

▶ Azure Data Explorer

▶ Azure Function

▶ Batch Service

▶ Databricks

▶ Data Lake Analytics


▶ General





▶ HDInsight

▶ Iteration & conditionals

Save as template ✓ Validate ✓ Validate copy runt

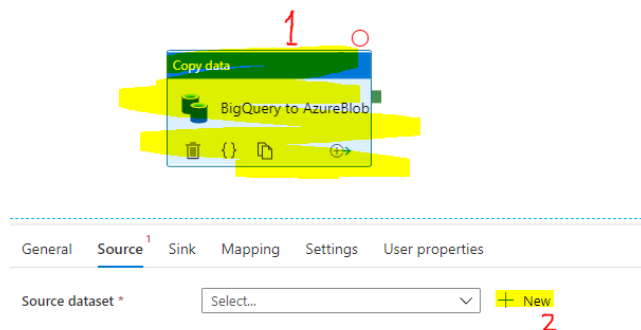
Copy data

 BigQuery to AzureBlob

Parte 3.1 (Creación de Source & Sink en el Pipeline del Data Factory)

- a. A partir de este punto vamos a construir el Origen(Source) de los datos y el destino (Sink). Para esto teniendo seleccionado el componente “copy data”, seleccione la pestaña “Source” y cree un nuevo “Source dataset” presionando sobre el “+New” (2).

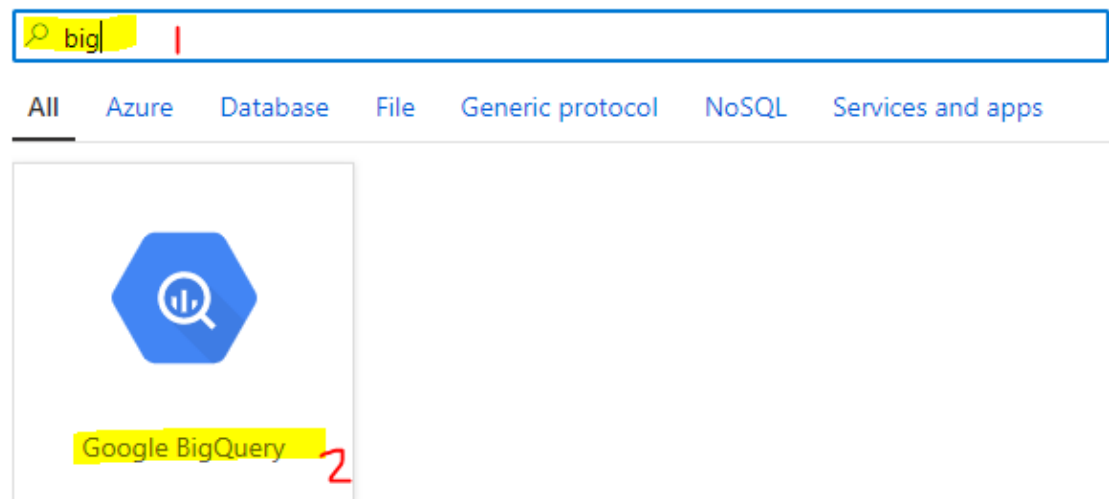


- b. En el lateral desplegable, escriba en la barra de búsqueda big (1) y seleccione el componente Google BigQuery (2) y después en el botón “continue”.

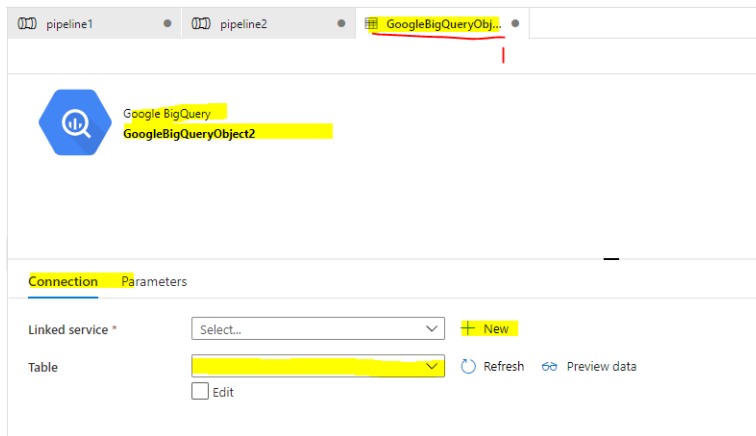
New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store



- c. Para completar la Edición seleccione el componente y presione “Open”, verifique que ha abierto el componente y en la pestaña de conexión en “Linked Service” presione el “+New”.



- d. Al presionar New, debemos configurar la conexión usando las cadenas que adquirimos durante el Paso de autenticación. Agregamos una descripción (1), después dejamos que de manera automática se seleccione el runtime de integración, sacamos el proyecto id del proyecto de BigQuery en Google Cloud Platform (3), marcamos como falso la solicitud de acceso a drive (4), Copiamos el Client-ID que nos dieron en la consola de desarrolladores de google (5) e introducimos el secret y el refresh token (7)(8). Y completamos el paso.

Edit linked service (Google BigQuery)

Name *

GoogleBigQuery1

Description

Servicio de conexión con Google bigQuery 1

Connect via integration runtime * ⓘ

AutoResolveIntegrationRuntime 2

Project ID *

mcd-icesi-cloud-moncada 3

Additional project IDs

Request access to Google Drive

☐ True ☒ False 4

Authentication type *

User authentication 5

Client ID

21sdq2jqal4.apps.googleusercontent.com 6

Client secret Azure Key Vault

Client secret 7

Refresh Token Azure Key Vault

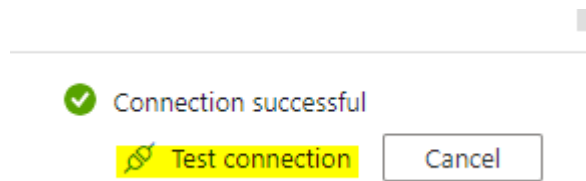
Refresh Token 8

Annotations

+ New

Advanced ⓘ

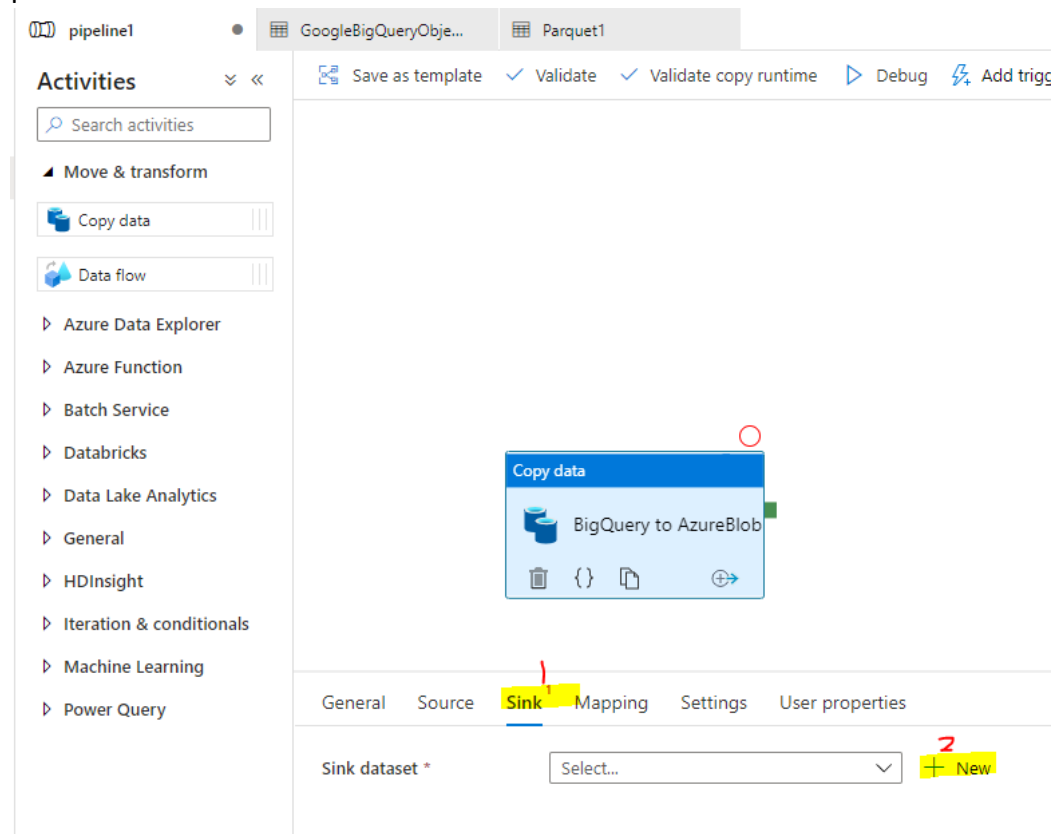
- e. Realizamos la prueba de conexión y finalizamos.



- f. Con la conexión probada, probada refrescamos (1) y seleccionamos la table (2). Este nombre de tabla es el mismo que tenemos en BigQuery.




- g. Con esto completamos la creación del Source. Y regresamos al Pipeline en la pestaña para crear el Sink(1), presionamos “+New” (2) para continuar en el proceso.



- h. En el lateral desplegado, escribir “blob” en la barra de búsqueda (1) y después seleccionar el dataset “Azure Blob Storage” (2).

New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#) 

Select a data store

All

Azure

Database

File

Generic protocol

Services and apps









Azure Blob Storage

- i. En las opciones seleccionar Parquet (ocupa menos espacio, pero se tarda un poco más en traerlo desde BigQuery).

Select format

Choose the format type of your data

 Avro	 Binary	 DelimitedText
 JSON	 ORC	 Parquet

- j. Asignamos un nombre y en el “linked Service” presionamos para crear uno nuevo.

Set properties

Name

Linked service *

Select...

Filter...

Select...

+ New

AzureBlobStorage1

+ New

- k. En el desplegable, incluimos los datos solicitados: nombre del servicio (1), descripción (2) , indicamos que sea de una suscripción de Azure (3) y marcamos la

conexión y la cuenta (4)(5), indicamos que el tipo de prueba de conexión sea “To linked service” (6) y probamos la conexión, una vez probada presionamos el botón “crear” (8).

New linked service (Azure Blob Storage)

Name * 1

Description 2

Connect via integration runtime * ⓘ

Authentication method

☒ Connection string ☐ Azure Key Vault

Account selection method ⓘ
☒ From Azure subscription 3 ☐ Enter manually

Azure subscription ⓘ
 4

Storage account name *
 5

Additional connection properties
[+ New](#)

Test connection ⓘ
☒ To linked service 6 ☐ To file path

Annotations
[+ New](#)

▸ Advanced ⓘ

8 7

- I. Marcamos las propiedades con el nombre del parquet (1), el servicio vinculado (2) y la ruta (seleccionando desde la carpeta (3)), indicamos que el esquema proviene de la conexión o esquema y finalizamos con el botón “OK/Create”.

Set properties

Name

Parquet2

Linked service *

AzureBlobStorage2

File path

Container

Directory

File

Import schema

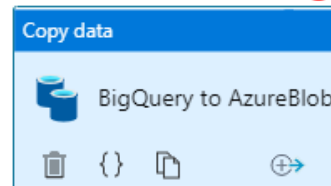
☐ From connection/store

☐ From sample file

☒ None

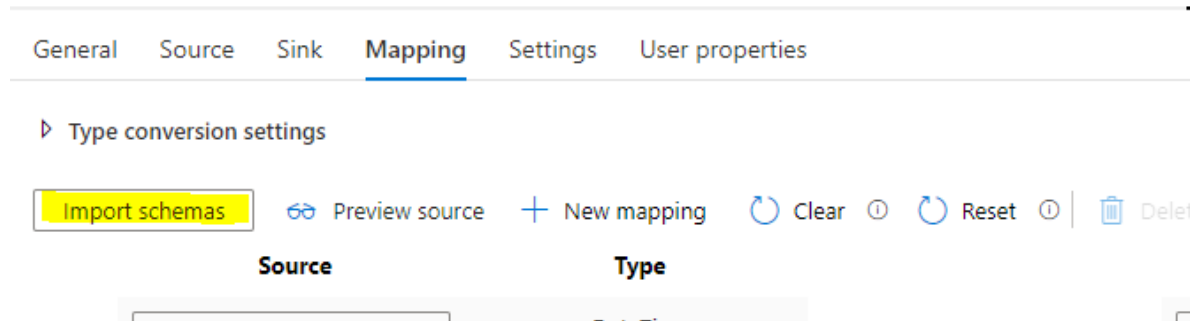
Advanced

- m. Dejamos en Copy behavior: None y pasamos a la pestaña de Mapping (Aquí se puede personalizar la forma de conversión de los campos de las tablas)

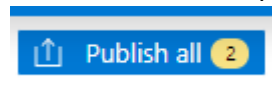


General	Source	Sink	Mapping	Settings	User properties
Sink dataset *		Parquet2	Open New Learn more		
Copy behavior ⓘ		None			
Max concurrent connections ⓘ					
Block size (MB) ⓘ					
Max rows per file ⓘ					

- n. Para traer los Esquemas desde la conexión presionamos “Import schemas” y listo.



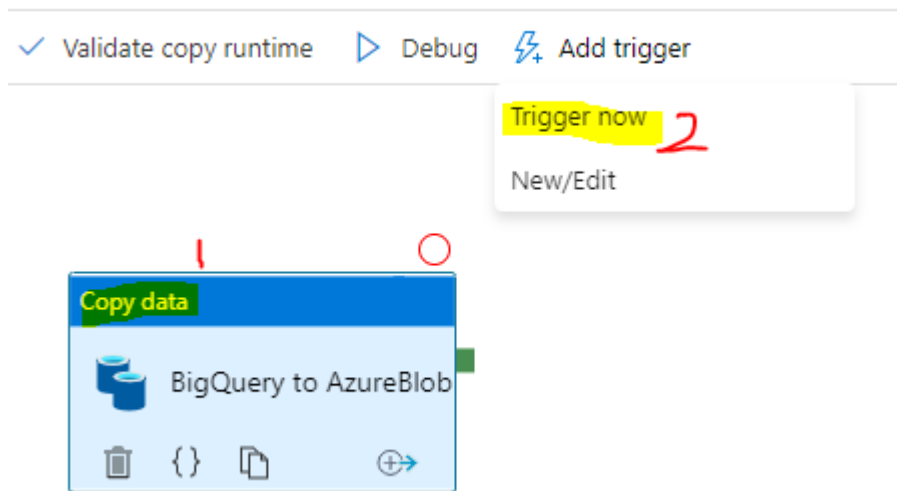
- o. Terminamos, las pestañas de Settings y User Properties quedan por defecto.
- p. Ahora vamos a publicar y para ello presionamos “publish all”, se debe ver un número 3. Y debe salir que no hay errores, de lo contrario deben corregirse con las recomendaciones que brinda la interfaz



Your Factory has been validated.

No errors were found.

- q. Finalmente vamos a ejecutar el pipeline, para ello se selecciona el pipeline y se presiona en “Add trigger” seguido de “Trigger now” (2).



- r. Al ejecutarlo, podemos ver en el monitor (1) el progreso y una vez finalizado aparecerá verde (2).

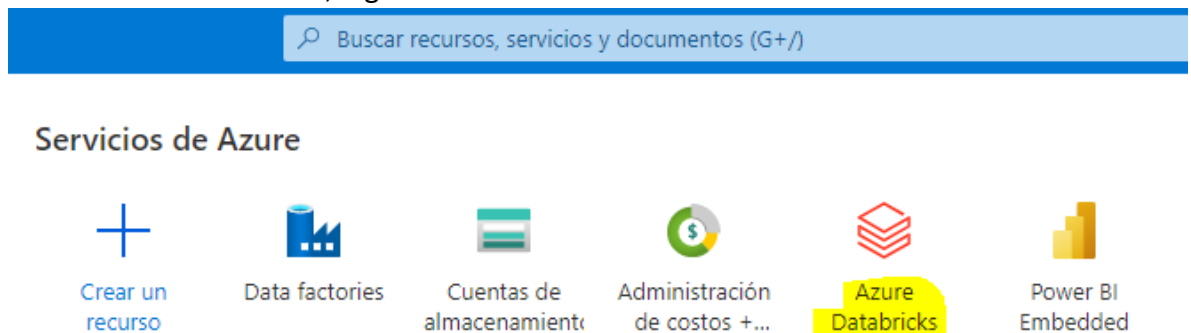
Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters
pipeline1	1/29/21, 1:14:20 PM	1/29/21, 1:18:37 PM	00:04:17	Manual trigger	Cancelled	Original	
pipeline1	1/29/21, 1:10:09 PM	1/29/21, 1:33:20 PM	00:23:11	Manual trigger	Succeeded	Original	

- s. Con esto se da por finalizado y ahora los datos están en Azure Blob Storage.

Parte 4 (Conexión con Azure Databricks).

Durante esta parte crearemos un cluster de Azure Databricks y probaremos algunas consultas básicas.

- a. Desde el home de Azure, ingresamos a Azure DataBricks.



- b. Presionamos “+ Agregar” para crear una nueva instancia.


The screenshot shows the Azure Databricks interface. At the top, there's a blue header with the Microsoft Azure logo and an 'Actualización' button. Below the header, the breadcrumb 'Inicio >' is visible. The main heading is 'Azure Databricks' with a star icon. Underneath, it says 'Universidad Icesi (@icesi.edu.co)'. A navigation bar contains a yellow '+ Agregar' button, a gear icon for 'Administrar vista', and a circular arrow icon for 'Actualizar'. Below this is a search bar with the placeholder 'Filtrar por cualquier camp' and a filter button set to 'Suscripción == todo'. A message states 'Mostrando de 1 a 1 de 1 registros.' Below this, there's a table with one row. The first column has a checkbox and the text 'Nombre ↑↓'. The second column has a checkbox, a red cube icon, and the text 'mcd-inst-databricks'.

- c. Procedemos a configurar la instancia, insertamos los datos básicos y seleccionamos los elementos de nuestra suscripción creada.


The screenshot shows the 'Datos básicos' tab of the Azure Databricks configuration page. The tabs are 'Datos básicos', 'Redes', 'Etiquetas', and 'Revisar y crear'. The section is titled 'Detalles del proyecto'. Below this, there's a text block: 'Seleccione la suscripción para administrar recursos implementados y los costes. Use los grupos de recursos como carpetas para organizar y administrar todos los recursos.' There are two dropdown menus: 'Suscripción *' with the value 'Azure subscription 1' and 'Grupo de recursos *' with the value 'grupo-recursos-bigQueryToAzure'. Below these is a link 'Crear nuevo'. The section 'Detalles de instancia' has three fields: 'Nombre del área de trabajo *' with the value 'instancia-prueba' and a green checkmark, 'Región *' with the value 'Centro-Sur de EE. UU.', and 'Plan de tarifa *' with the value 'Estándar (Apache Spark, seguro con Azure AD)'.

- d. Esperamos la implementación.


 Eliminar  Cancelar  Volver a implementar  Actualizar

 Nos encantaría recibir sus comentarios. →

... La implementación está en curso


 Nombre de implementación: grupo-recursos-bigQueryToAzure_Ins... Hora
Suscripción: [Azure subscription 1](#) Id. d
Grupo de recursos: [grupo-recursos-bigQueryToAzure](#)

^ Detalles de implementación [\(Descargar\)](#)

Recurso	Tipo
 Instancia-prueba	Microsoft.Databric

e. Al finalizar presionar en ir al recurso.

✓ Se completó la implementación

 Nombre de implementación: grupo-recursos-bigQueryToAzure_Ins...
Suscripción: [Azure subscription 1](#)
Grupo de recursos: [grupo-recursos-bigQueryToAzure](#)

✓ Detalles de implementación [\(Descargar\)](#)

^ Pasos siguientes

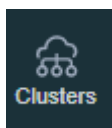
[Ir al recurso](#)

f. Presionar ir al recurso, y proceder a iniciar área de trabajo



[Iniciar área de trabajo](#)

g. Ahí en la barra izquierda seleccionamos los clusters



- h. Asignamos un nombre, seleccionamos un solo nodo y la versión 7.5 ML sin GPU.

Create Cluster

New Cluster

Cancel

Create Cluster

0 Workers: 0
1 Driver: 14.1

Cluster Name

icesi-cluster

Cluster Mode ?

Single Node

Pool ?

None

Databricks Runtime Version ?

[Learn more](#)

Runtime: 7.5 ML (Scala 2.12, Spark 3.0.1)

Autopilot Options

☒ Terminate after 120 minutes of inactivity ?

Node Type ?

Standard_DS3_v2

14.0 GB Memory, 4 Cores, 0.75 DBU

?

► Advanced Options

- i. Creamos un nuevo notebook.



Explore the Quickstart Tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Common Tasks



New Notebook

- j. Incluimos el código disponible en el html (databricks) usando los datos propios de su conexión.
- k. Revisar las consultas SQL.

Cmd 11

```
1 %sql
2 --Número global de casos
3 SELECT sum(new_confirmed)
4 from covidtable
5 where aggregation_level = 0;
```

► (2) Spark Jobs

	sum(new_confirmed) ▲	
1	100956257	

Showing all 1 rows.

```
1 %sql
2 -- Países más afectados
3 SELECT country_name, sum(new_confirmed)
4 from covidtable
5 where aggregation_level = 0
6 group by country_name
7 order by 2 desc;
```

► (2) Spark Jobs

	country_name ▲	sum(new_confirmed) ▲	
1	United States of America	25354044	
2	India	10720048	
3	Brazil	8937717	
4	Russia	3793810	
5	United Kingdom	3743733	
6	France	3107248	
7	Spain	2705189	
8	Italy	2518829	

Showing all 247 rows.

```

1 %sql
2 --Regiones más afectadas en Estados Unidos
3 SELECT country_name, subregion1_name, sum(new_confirmed)
4 from covidtable
5 where aggregation_level = 1 and country_code = 'US'
6 group by country_name, subregion1_name
7 order by 3 desc;

```

► (2) Spark Jobs

	country_name ▲	subregion1_name ▲	sum(new_confirmed) ▲
1	United States of America	California	3201774
2	United States of America	Texas	2089257
3	United States of America	Florida	1674681
4	United States of America	New York	1374480
5	United States of America	Illinois	1116371
6	United States of America	Ohio	883716
7	United States of America	Pennsylvania	824403
8	United States of America	Arizona	713230

Showing all 56 rows.

Cmd 14

```

1 %sql
2 -- Indice de letalidad
3 SELECT sum(new_deceased)/sum(new_confirmed) indice
4 from covidtable
5 where aggregation_level = 0;

```

► (2) Spark Jobs

	indice ▲
1	0.021450022656842357

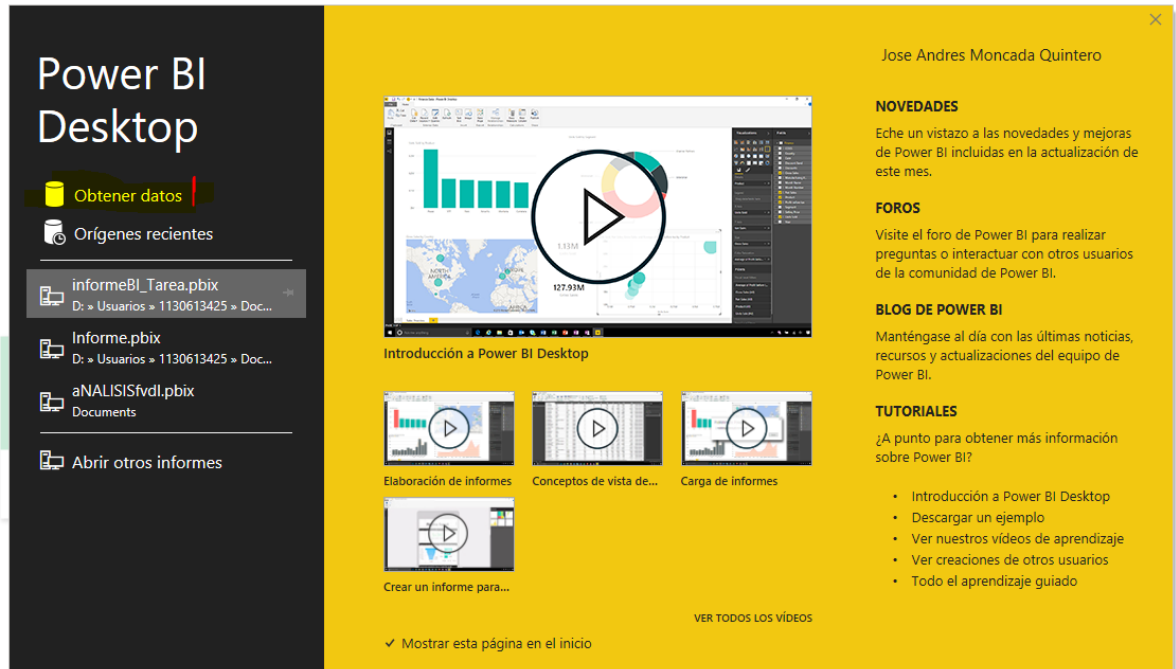
Showing all 1 rows.

- I. Al finalizar el proceso, los datos han sido creados en el cluster y pueden ser capturados desde PowerBI.

Parte 5 (Conexión con Power BI).

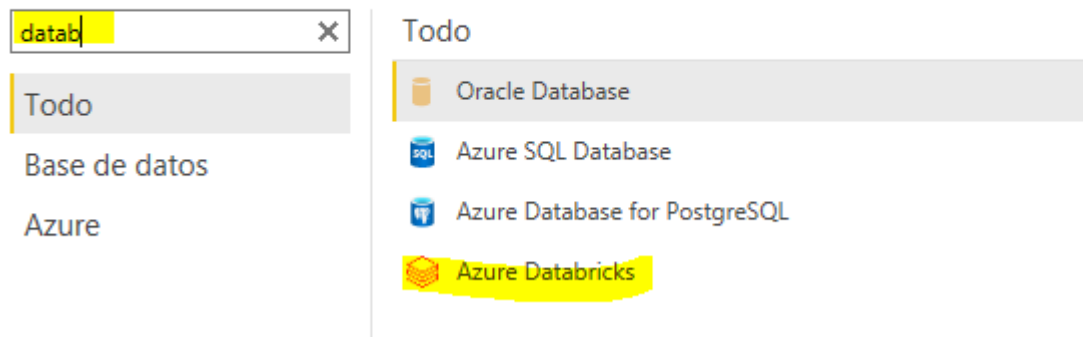
En esta última parte usaremos el conector que trae la versión más reciente de PowerBI para conectarse a Azure Databricks.

- a. Al iniciar Power BI, presione en obtener Datos:



- b. Use la barra de búsqueda y seleccione el conector de Azure Databricks

Obtener datos



- c. Capture los datos del cluster en la configuración, opciones avanzadas. Y complete la información solicitada por el conector.

Clusters / icesi-cluster

icesi-cluster ✨

[Edit](#) [Clone](#) [Restart](#) [Terminate](#) [Delete](#)

Configuration Notebooks (0) Libraries Event Log Spark UI Driver Logs Metrics Apps Spark Cluster UI - Master ▼

/ 5 ML (includes Apache Spark 3.0.1, Scala 2.12)

Autopilot Options

☒ Terminate after 120 minutes of inactivity ⓘ

Node Type ⓘ

Standard_DS3_v2 14.0 GB Memory, 4 Cores, 0.75 DBU

▼ **Advanced Options**

Azure Data Lake Storage Credential Passthrough ⓘ Available on Azure Databricks Premium [Learn more](#)

☐ Enable credential passthrough for user-level data access

[Spark](#) [Tags](#) [Logging](#) [Init Scripts](#) [JDBC/ODBC](#) [Permissions](#)

Server Hostname

adb-3884809831586967.7.azuredatabricks.net

Port

- d. Navegue en el árbol de carpetas hasta la tabla y selecciónela, para después presionar “cargar”

Navegador

Opciones de presentación ▼

- adb-3884809831586967.7.azuredatabricks.net...
- SPARK [1]
- default [1]
 - covidtable**

covidtable

Vista previa descargada el viernes

date	location_key	country_code	country_name
4/01/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
29/01/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
30/01/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
17/02/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
26/02/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
5/03/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
8/03/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
23/03/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
30/03/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
2/04/2020 12:00:00 a. m.	US_WY_56035	US	United States of America
23/04/2020 12:00:00 a. m.	US_WY_56035	US	United States of America

Los datos de la vista previa se han truncado debido a límites de tamaño.

- e. Crea algunos objetos visuales y con eso habremos finalizado el proceso.

The screenshot displays the Databricks workspace interface. On the left, there are three line charts showing COVID-19 data trends over time (from January 2020 to January 2021). The top chart is titled 'new_confirmed y new_recovered por date'. The middle chart is 'new_hospitalized_patients, new_intensive_care_patients y new_deceased por Mes'. The bottom chart is 'new_intensive_care_patients y new_deceased por Mes'. On the right, there is a sidebar with 'Filtros' (Filters) and 'Campos' (Fields) sections. The 'Campos' section shows a list of fields for the 'covidtable' table, including 'date', 'location_key', 'country_code', 'country_name', and various COVID-19 metrics like 'new_confirmed', 'new_recovered', 'cumulative_confirmed', etc.

