**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. El siguiente diagrama describe los pasos del flujo de datos:

[Diagrama general de la solución]

**¿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/>

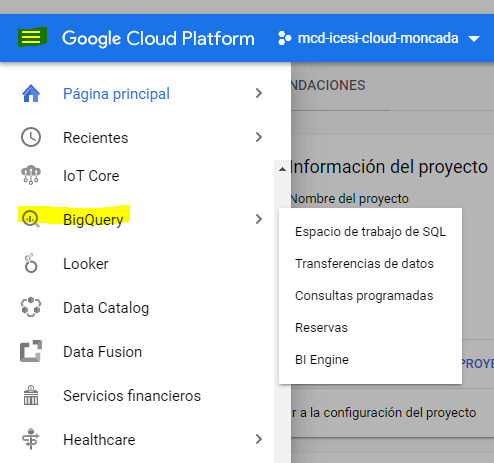
Si es la primera vez que se crean las cuentas relacionadas con las nubes (Azure y Google Cloud) se reciben créditos gratuitos para usar en los servicios.

**Instrucciones**

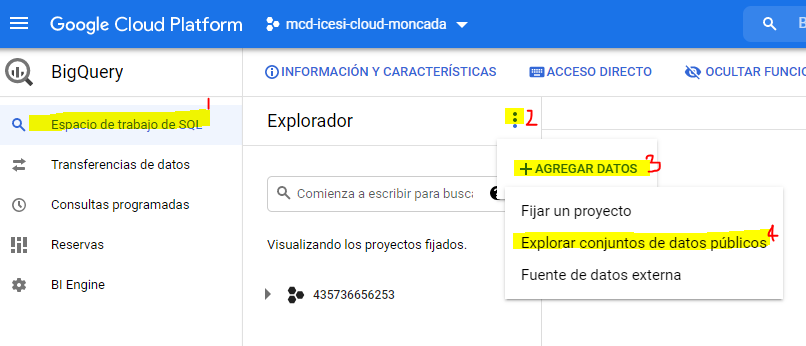
Estas instrucciones, asumen que las cuentas han sido creadas antes de iniciar el proceso. Las herramientas están configuradas en español por lo que si su versión está en inglés podrían diferir los nombres.

**Parte 1 (Preparación de los datos en BigQuery)**

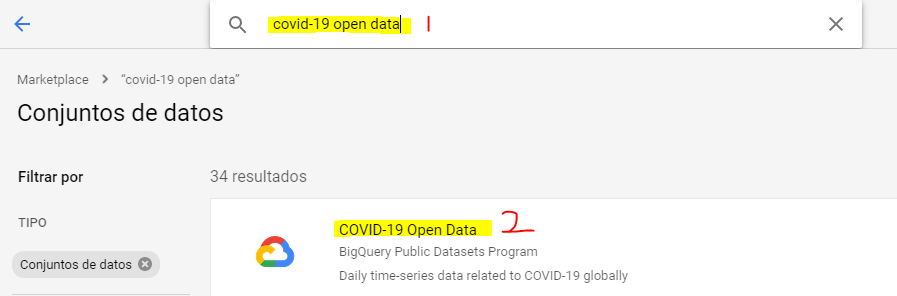
1. Estando ubicado en la consola de google cloud (<https://console.cloud.google.com/>), procesa a realizar el acceso a la herramienta BigQuery, la imagen a continuación describe el ingreso (las zonas marcadas en amarillo, muestran los puntos de contacto).



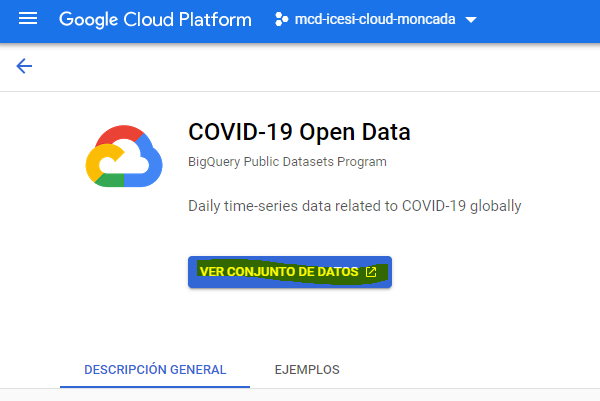
1. Una vez adentro de BigQuery, seleccione en el menú izquierdo la opción marcado como espacio de trabajo de SQL, una vez presionado se despliega un espacio marcado como Explorador, en los tres puntos verticales ubicados en el lado superior izquierdo del Explorador presione (+AGREGAR DATOS) y de ahí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



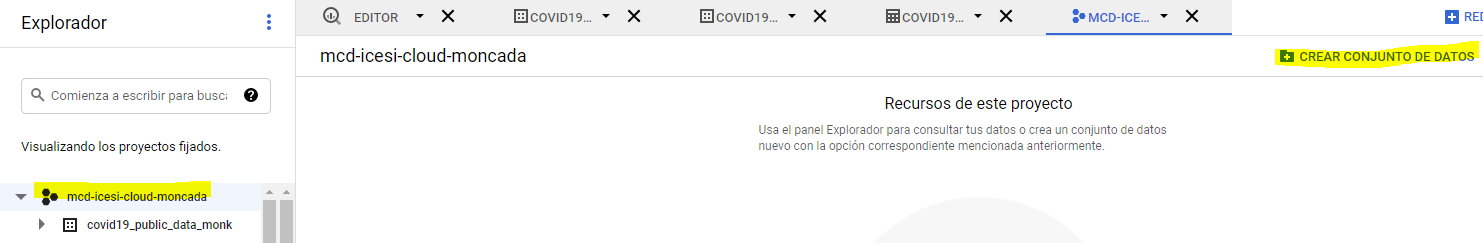
1. La acción del paso (b) genera el despliegue de una nueva ventana, una vez ahí introduzca el texto: “covid-19 open data” en la barra de búsqueda y seleccione el resultado etiquetado como: “COVID-19 Open Data”, la imagen a continuación muestra el proceso:



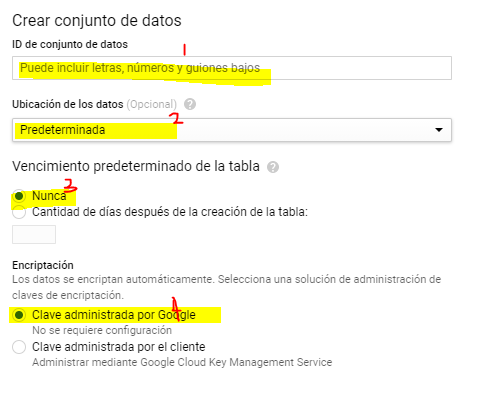
1. En la ventana desplegada tras la selección del paso (c), presione en el botón “Ver Conjunto de Datos”. Esto generará que regrese a la ventana anterior, pero en su explorador se habrá fijado un nuevo proyecto (ver paso e).



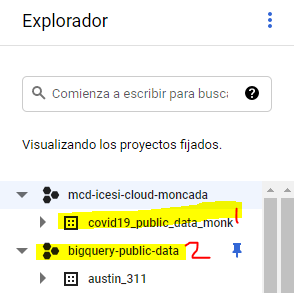
1. Ahora vamos a crear un conjunto de datos en nuestro proyecto, sobre el cuál copiaremos los datos del proyecto de datos abiertos
2. Seleccione su proyecto y posteriormente en el área de trabajo (a la derecha del explorador) presione el botón “Crear Conjunto de Datos”



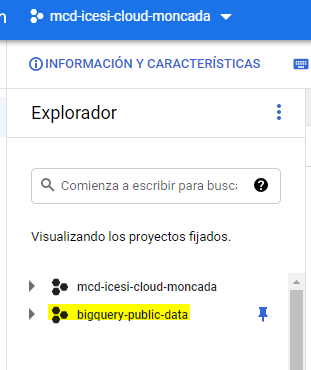
1. En la ventana lateral que despliega el botón, indique el nombre del conjunto de datos, marque predeterminada, sin vencimiento y clave administrada por google. Finalmente presione “Crear conjunto de datos” en la parte inferior.



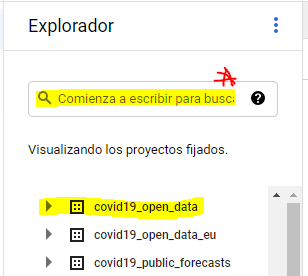
1. La acción previa generó un conjunto de datos asociados a su proyecto (1 en la siguiente imagen), pero por ahora no tiene ninguna tabla, vamos ahora a traer los datos del proyecto de datos abiertos (2 en la siguiente imagen).



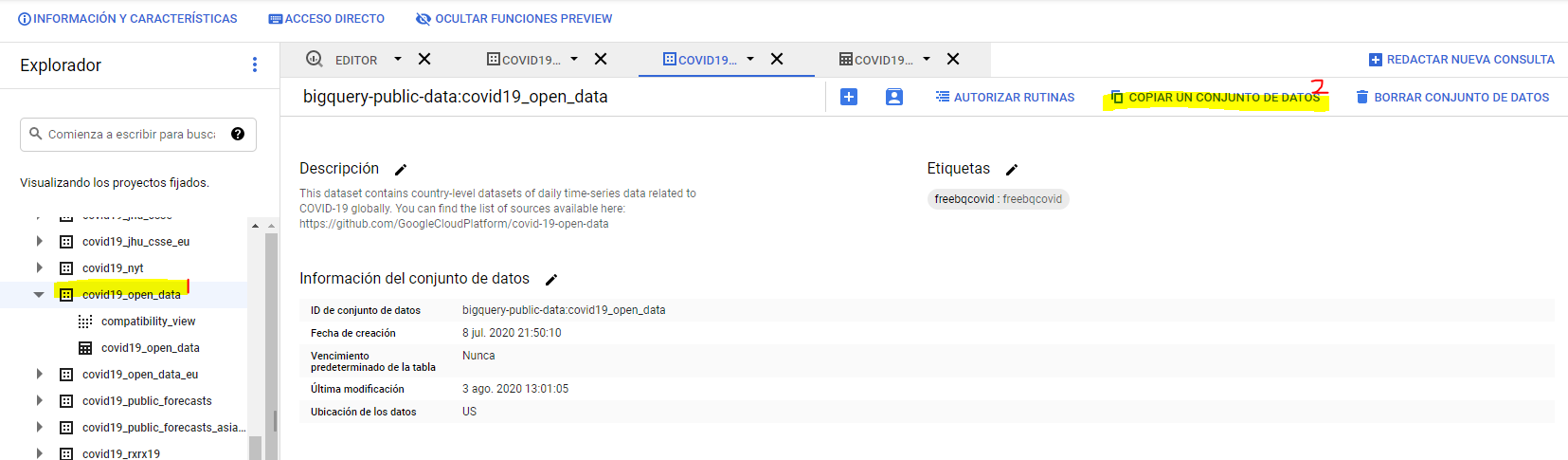
1. En la ventana del explorador ahora usted tiene además de su proyecto un proyecto llamado: “bigquery-public-data” (agregado en el paso (d)). Al desplegar este usted verá los conjuntos de datos disponibles:



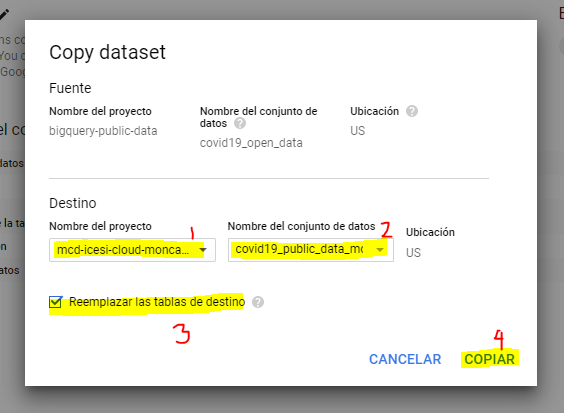
1. Dentro de los datos desplegados en el paso (i), busque uno llamado: “covid\_19\_open\_data”, si lo prefiere use este texto en el buscador para acelerar el proceso.



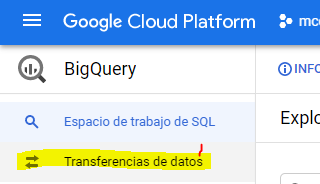
1. Ahora necesitamos transferir estos datos del proyecto de datos abiertos al proyecto personal, para esto seleccione el conjunto de datos con el nombre indicado en el paso (j) y en la información desplegada en el área de trabajo (a la derecha del explorador), identifique y presione el botón con nombre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



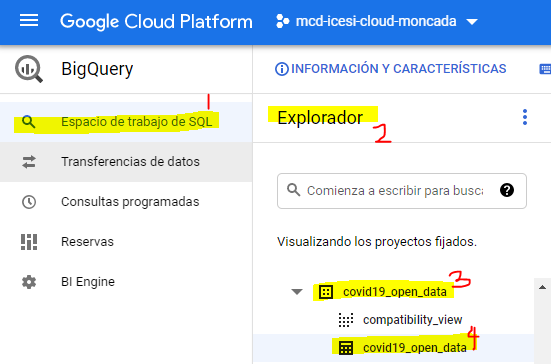
1. El paso (k) ha desplegado una ventana emergente, seleccione en esta el nombre del proyecto(1) y el conjunto de datos en el cuál desea copiar(2), marque la opción de reemplazo de tablas y presione el botón “copiar”(4).



1. 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”



1. 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)
2. 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.



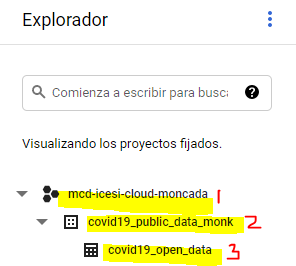
1. Con la tabla seleccionada, presione el botón “copiar tabla” disponible en el lado derecho del área de trabajo en la parte superior.



1. 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).



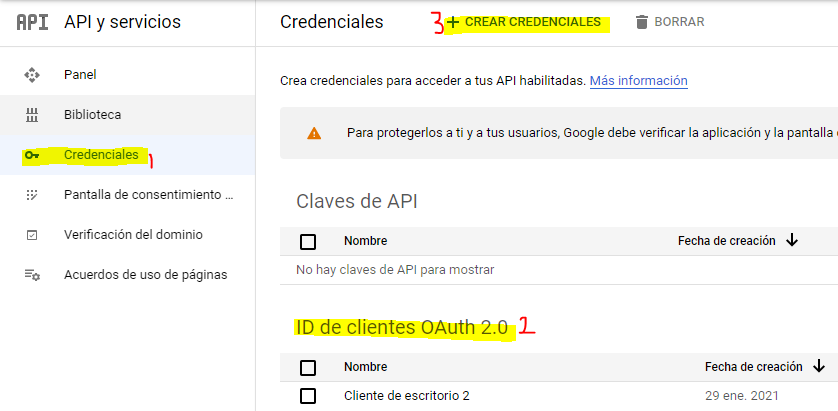
1. La última acción debe dejar disponibles los datos dentro del proyecto personal (1), dentro de un conjunto de datos (2) y en la tabla con el nombre que le haya asignado (3).



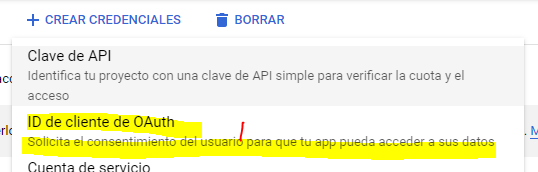
**Parte 2 (Preparación de las credenciales de conexión entre GCP y Azure)**

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.

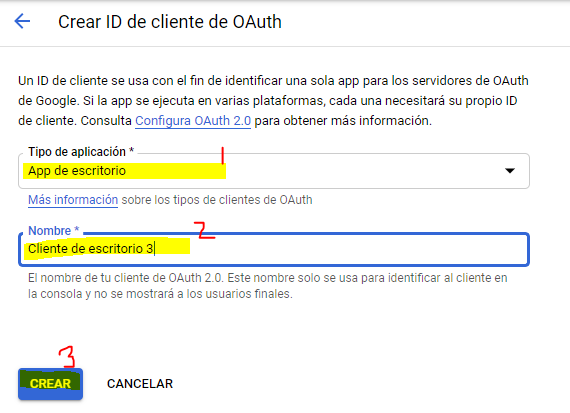
1. Estando en la consola de google, en el menú lateral izquierdo seleccione la opción “Credenciales” (1), vamos a crear una credencial de tipo OAuth 2.0 (2) y para ellos debemos presionar en “+Crear Credenciales”. (Asegurarse de tener seleccionado el proyecto correcto en la parte superior)



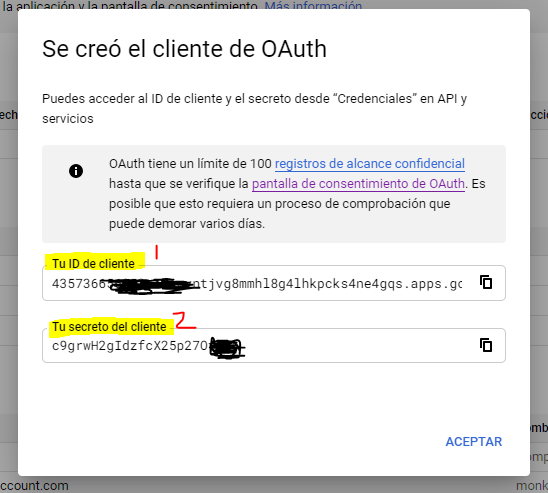
1. En el desplegable seleccione: “ID de cliente de OAuth”



1. 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)



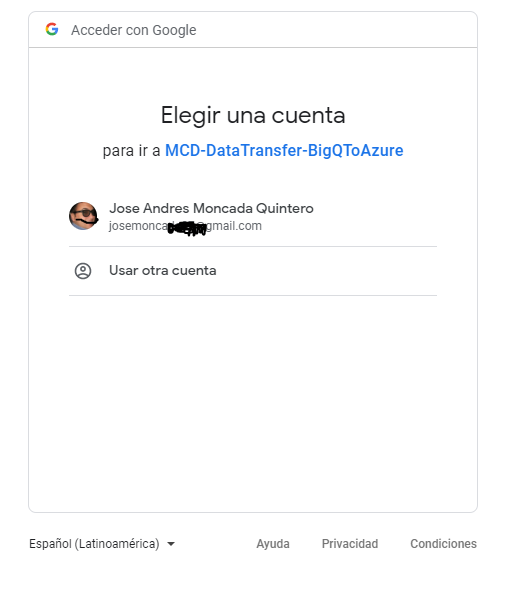
1. En la ventana siguiente, debes copiar los datos de ID de cliente y secreto de cliente (déjalos en un bloc de notas), los vamos a necesitar más adelante. Si lo cerraste y no lo copiaste, puedes ver estos datos al presionar sobre el ID de cliente en la sección de credenciales.



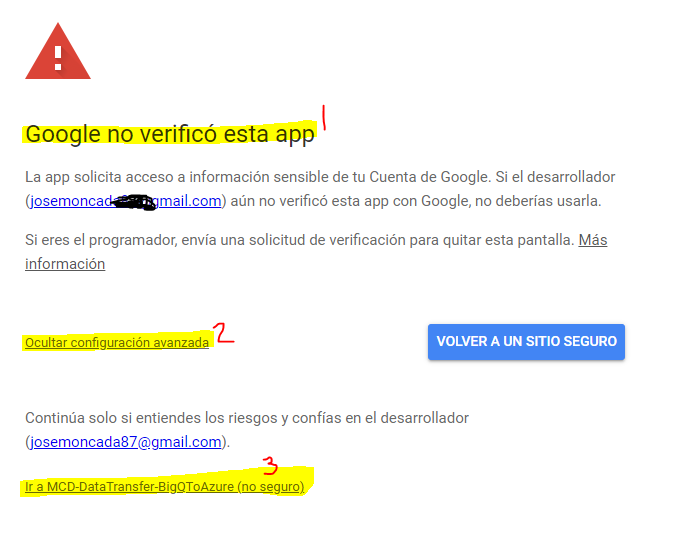
1. Con los códigos generados vamos a crear un token de autenticación, para ellos vamos a usar la siguiente dirección web, Ojo: Debes reemplazar: =<Your-client-Id> por el ID de cliente que acabas de copiar.

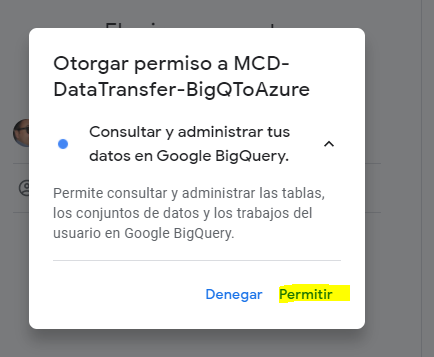
https://accounts.google.com/o/oauth2/v2/auth?client\_id=<Your-client-Id>&redirect\_uri=urn:ietf:wg:oauth:2.0:oob&state=GBQAUthTest&access\_type=offline&scope=https://www.googleapis.com/auth/bigquery&response\_type=code

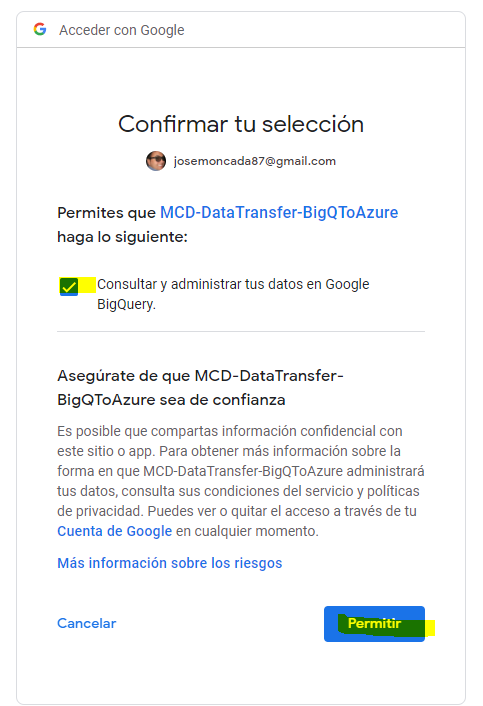
1. Al pegarlo en tu navegador, debes ver una entrada de cuenta, selecciona la cuenta Gmail que tiene el acceso a Google Cloud Platform (GCP). Accede con los datos de login.



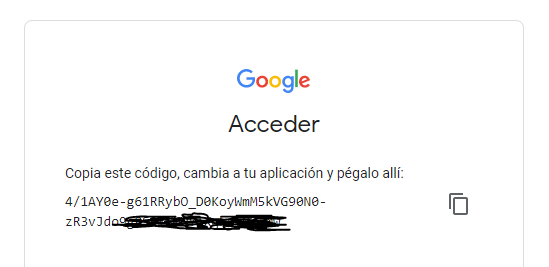
1. Si recibes un mensaje que dice “Google no verificó esta app”, debes presionar en configuración avanzada y en ir al app.



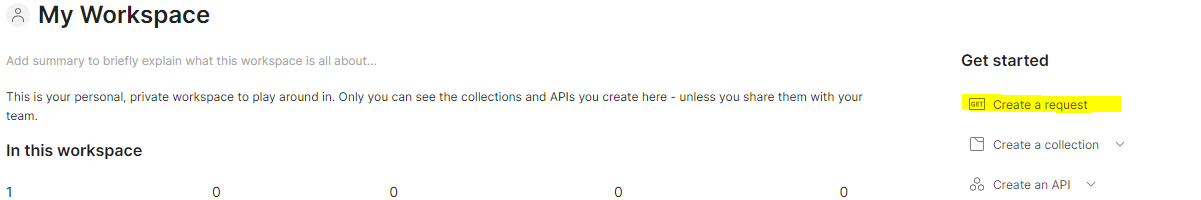
1. Marcar permitir en la siguiente ventana, si los permisos de la imagen no coinciden asegúrate de haber creado la cuenta asignando los permisos al usuario.
2. Verificar la selección de: “Consultar y administrar tus datos en Google BigQuery” y después en permitir.



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



1. Ahora vamos a abrir Postman, en la ventana principal seleccionar la opción “CReate Request”:



1. 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.



Plantilla:

{

"code":" <código> ",

"client\_id":" <Id> ",

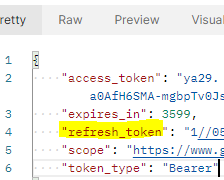
"client\_secret":"< Secreto >",

"redirect\_uri":"urn:ietf:wg:oauth:2.0:oob",

"grant\_type":"authorization\_code"

}

1. 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:

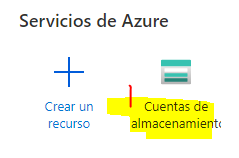


1. Con esto completamos la información necesaria para ir a Azure Datafactory en el siguiente paso.

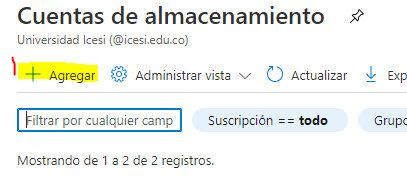
**Parte 3 (Configuración del Azure DataFactory)**

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.

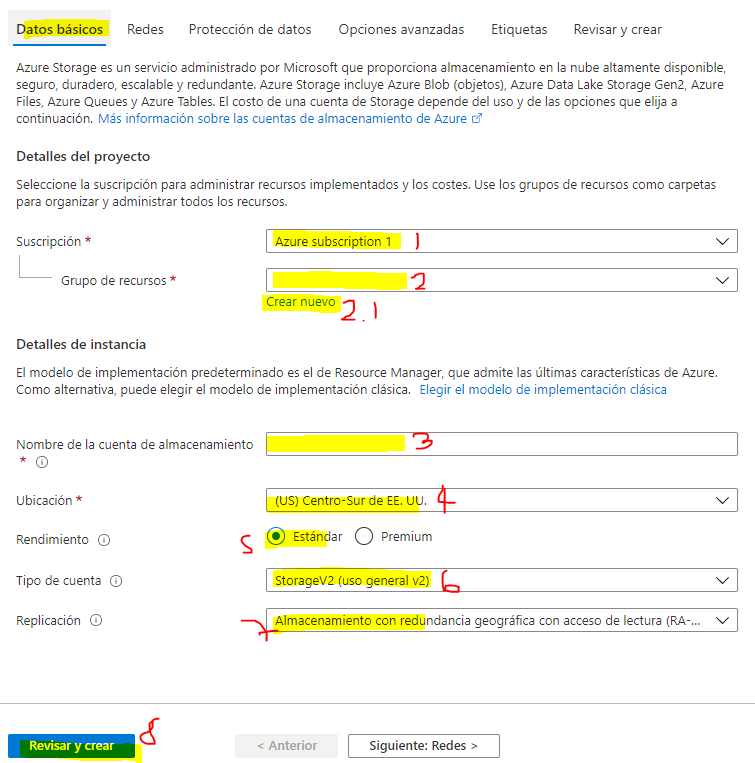
1. Vamos a crear una cuenta de almacenamiento, para eso vamos a la página principal de Azure <https://portal.azure.com/#home>, de ahí vamos a seleccionar la opción “Cuenta de almacenamiento”.



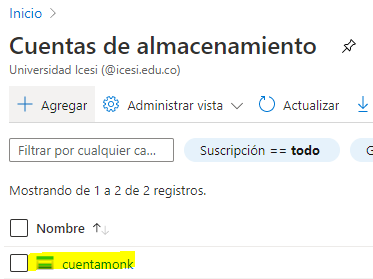
1. Dentro del gestor, presionamos “agregar”



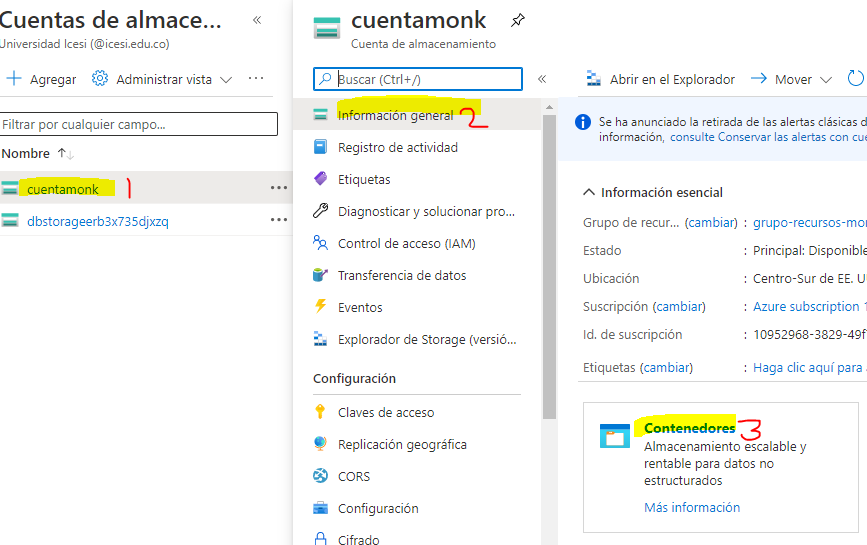
1. En la ventana que se lanza, vamos a insertar los datos. Seleccione primero una suscripción (1), después indique cuál es el grupo de recursos (2), si no tiene ninguno debe crear uno (2.1), después inserte el nombre de la cuenta de almacenamiento (3), seleccione la ubicación más cercana (4), marque 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).



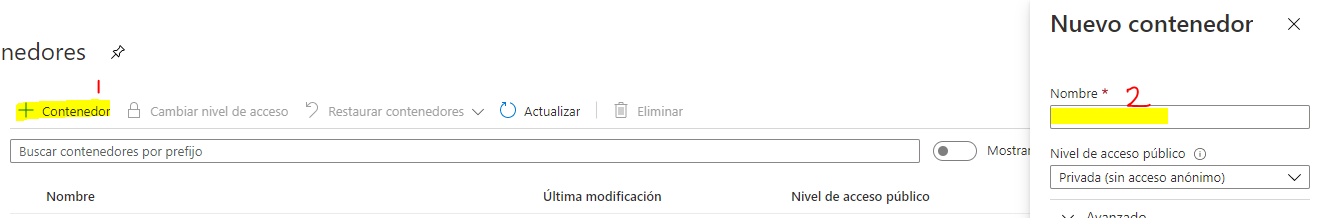
1. Dejamos los otros datos por defecto y al finalizar debemos ver la cuenta creada en el listado de cuentas de almacenamiento.



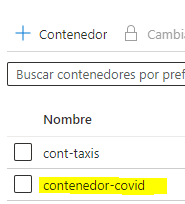
1. Ingresamos a la cuenta de almacenamiento (1) y nos disponemos a crear un contenedor para esto (3) en la información general (2). Presionamos sobre “contenedores”.



1. 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.



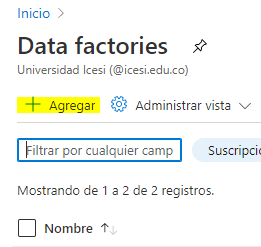
1. Al finalizar este proceso debe verse un contendor en l listado, con el nombre seleccionado.



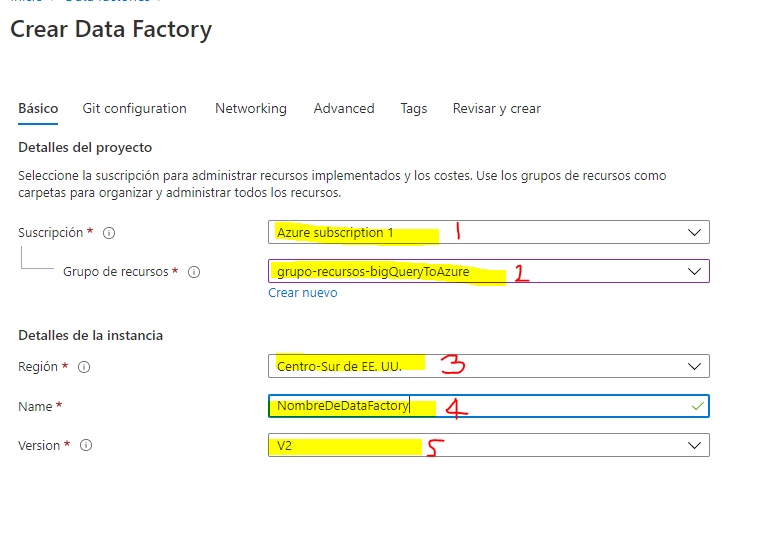
1. Ahora que tenemos una cuenta de almacenamiento, un grupo de recursos y un contenedor. Vamos a comenzar con la configuración del pipeline. Para ellos vamos a DataFactory, para ingresar regresamos al home de Azure (<https://portal.azure.com/#home>) y de ahí seleccionamos DataFactories.



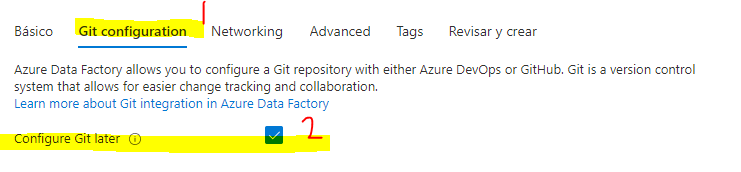
1. Al ingresar seleccionamos “+agregar” y eso nos lleva a la ventana de creación.



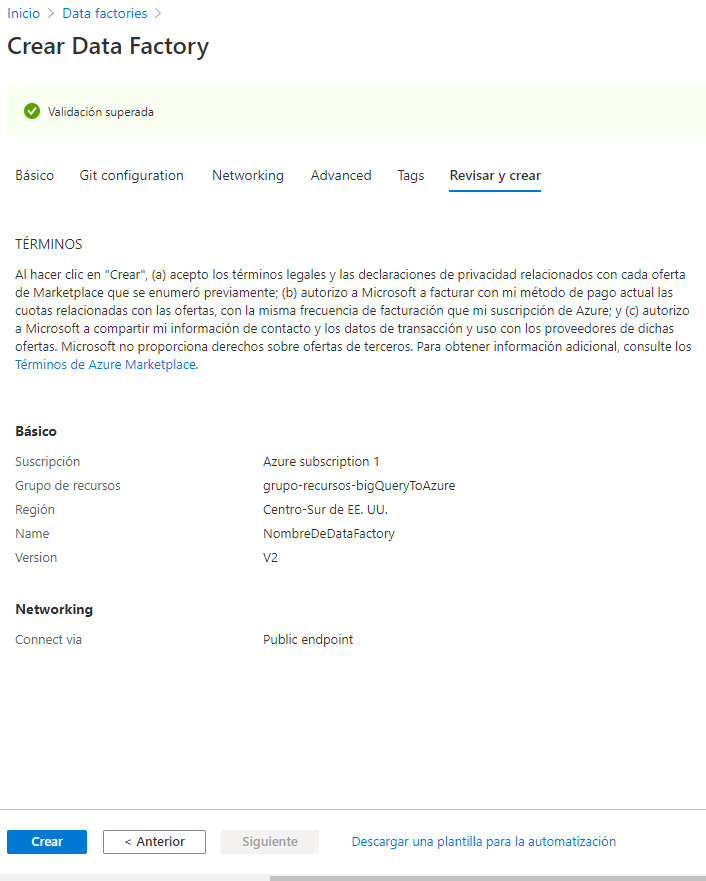
1. En la ventana de creación, vamos a llenar los datos de la pestaña básico, seleccionamos la suscripción (1), el grupo de recursos que ya habíamos creado (2), asignamos una región (3) y un nombre y versión (4)(5).



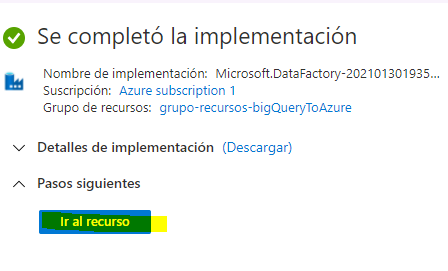
1. Vamos a la pestaña de configuración de git y seleccionamos configurar después:



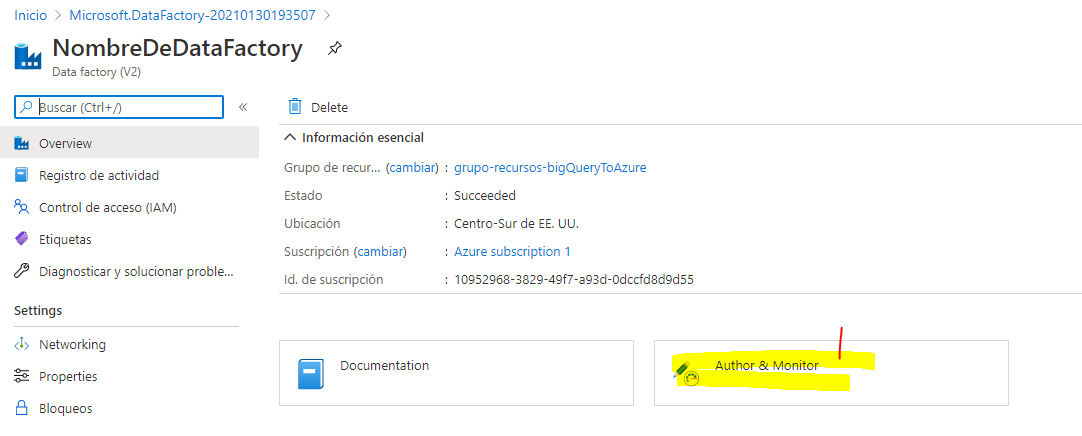
1. Presionamos revisar y crear, después nuevamente en el botón crear en la parte inferior.



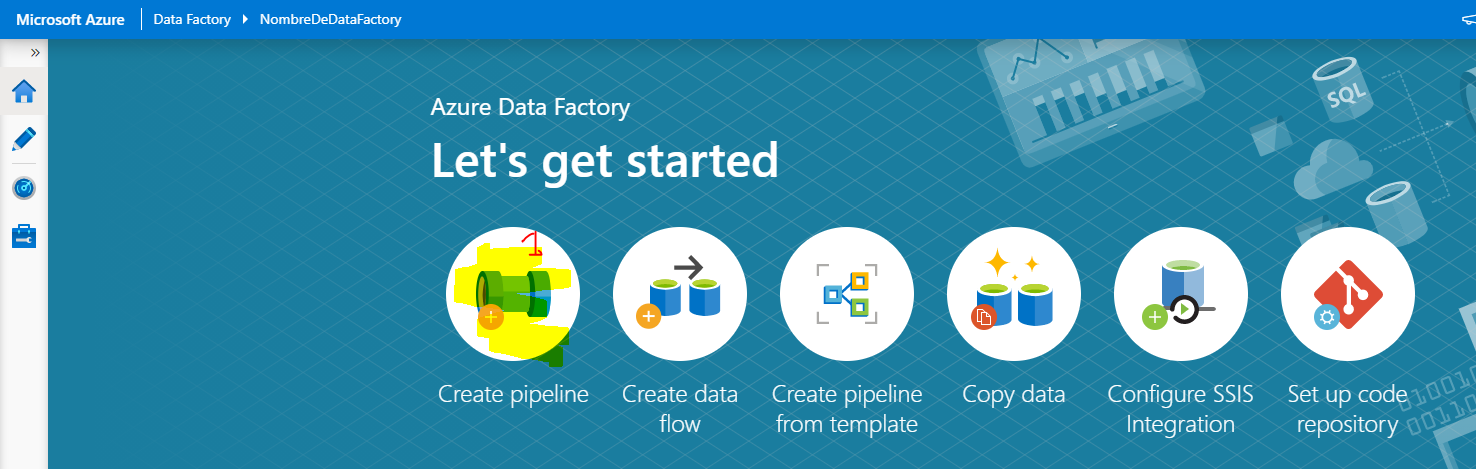
1. La implementación tomará un tiempo, pero al finalizar debes presionar “ir al recurso”.



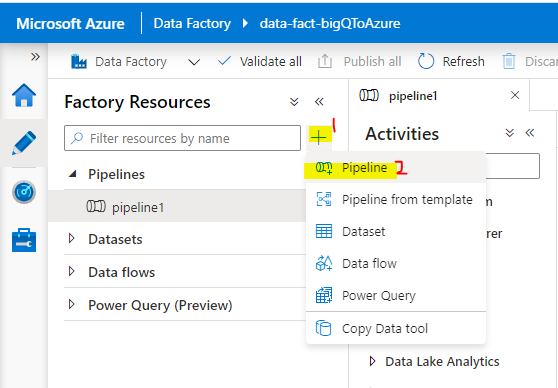
1. Al entrar al recurso, debemos presionar en “Author & Monitor”



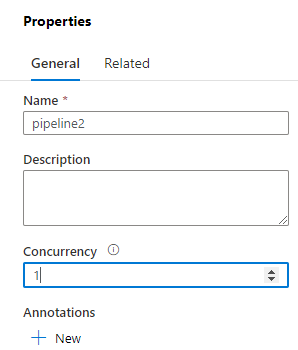
1. En la instancia, vamos a seleccionar “Create PipeLine”



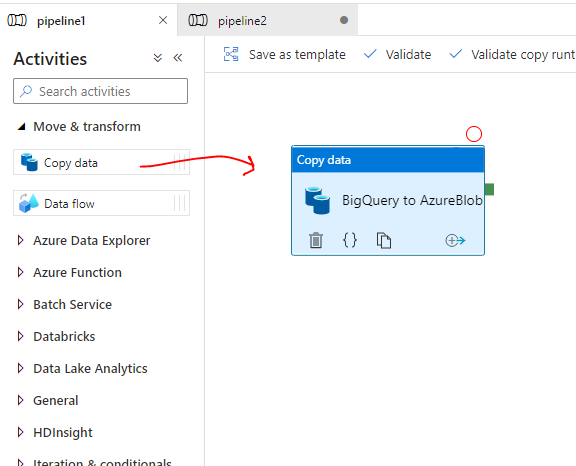
1. En los recursos, presionamos + (1) y seleccionamos el pipeline (2)



1. Después ingresamos las propiedades y listo.



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



1. A partir de este punto vamos a construir el Origen(Source) de los datos y el destino (Sink).