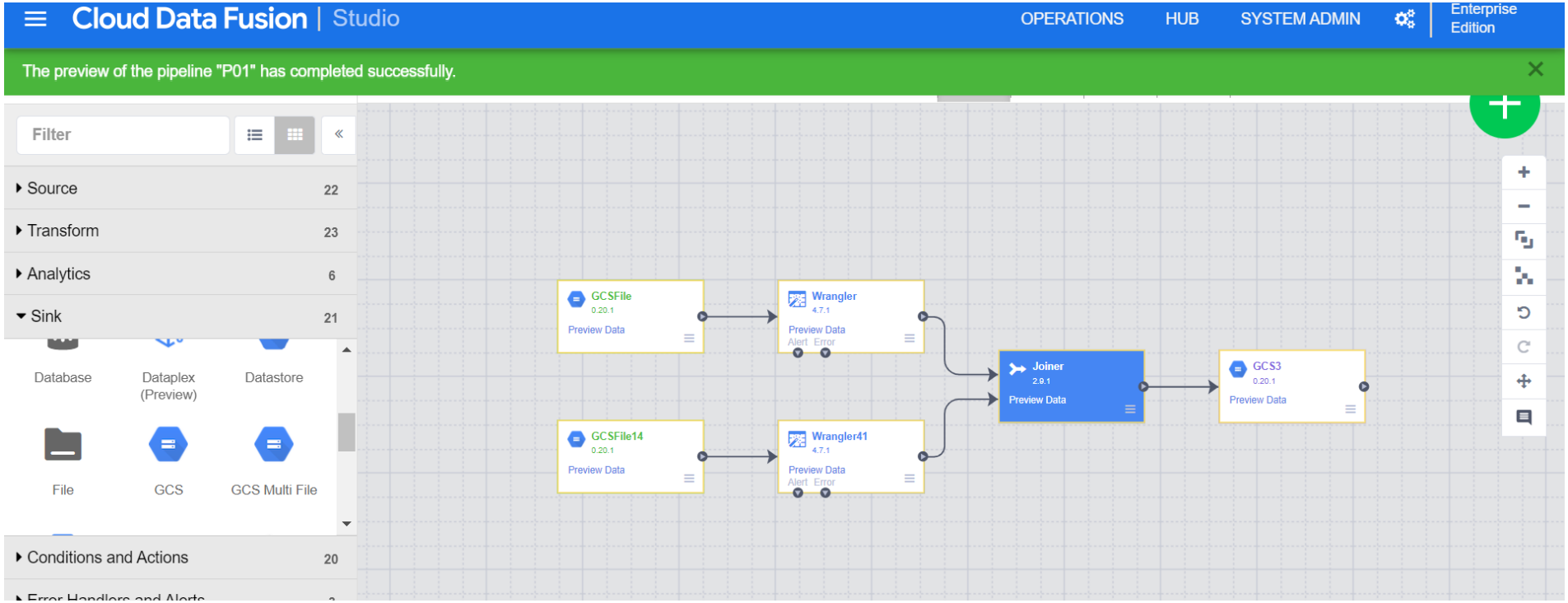
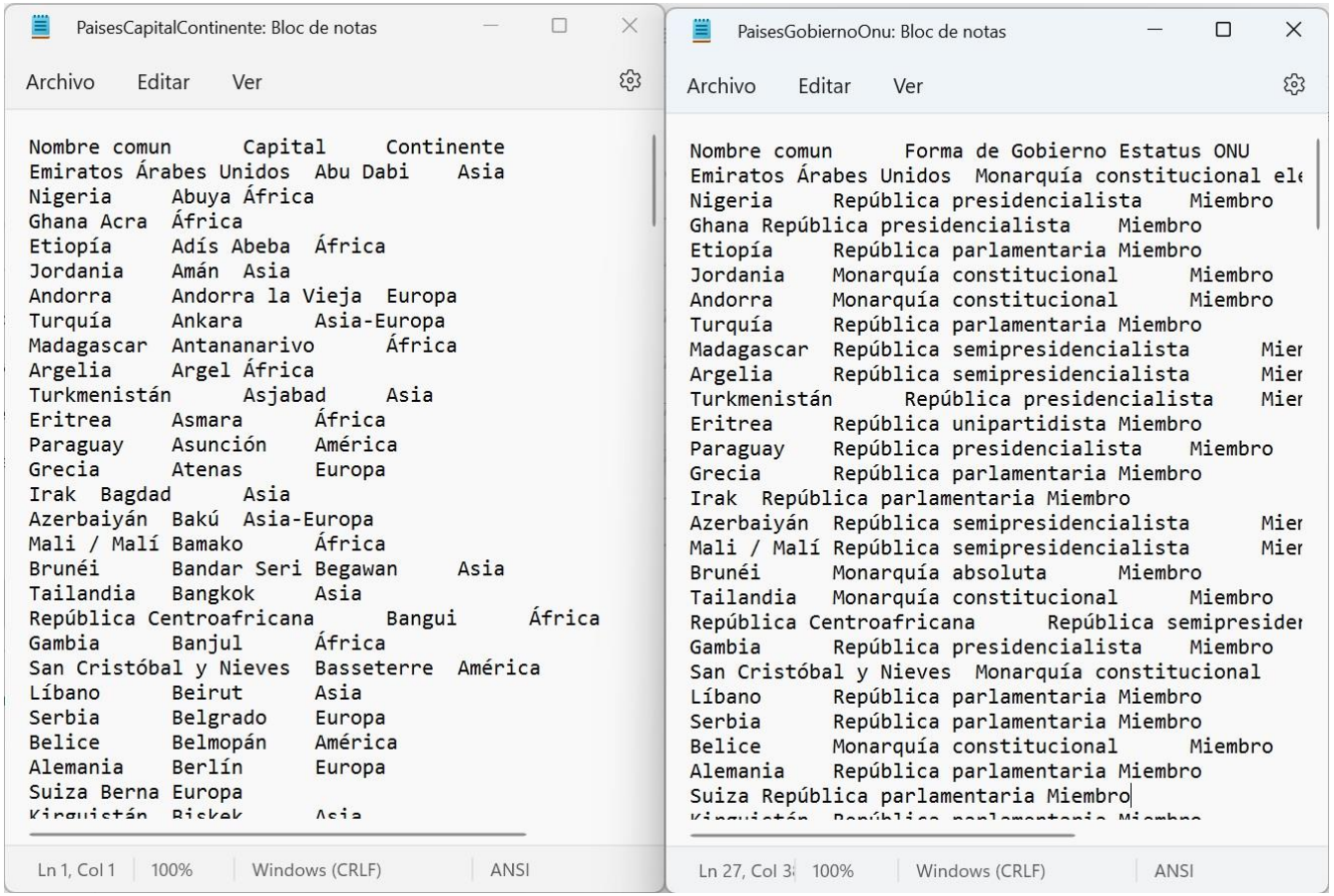
Ejemplo de ETL paso a paso con la herramienta Google Cloud Data Fusion

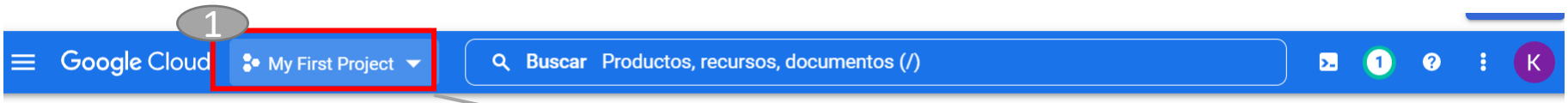


Contamos con 2 archivos que consolidaremos en 1 archivo: PaisesCapitalContinente.csv y PaisesGobiernoOnu.csv

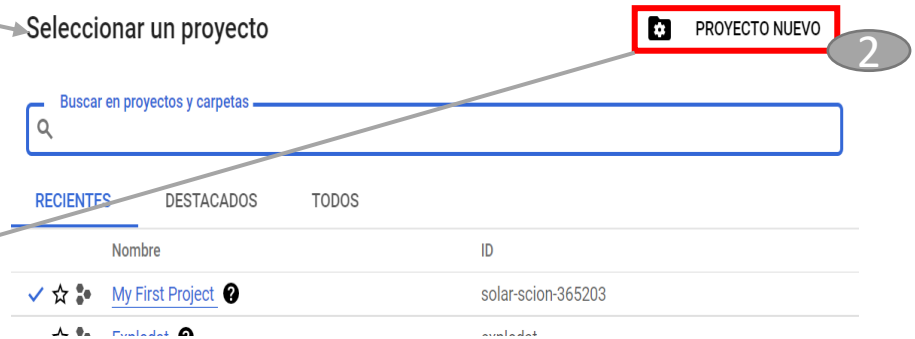


CREAR UN PROYECTO

Para crear un proyecto, debemos seleccionar el menú de proyectos. Por defecto tiene el nombre del último proyecto en el que se trabajó:



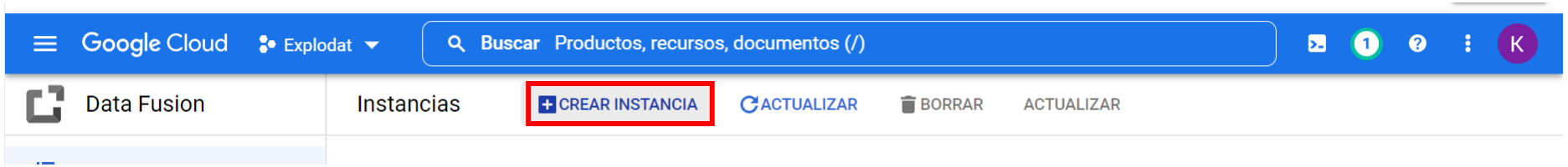
Luego se abre la ventana de selección de proyecto. Clic en el botón “Proyecto Nuevo”, asignar un nombre (en este caso Lituus) y presionar el botón “Crear”

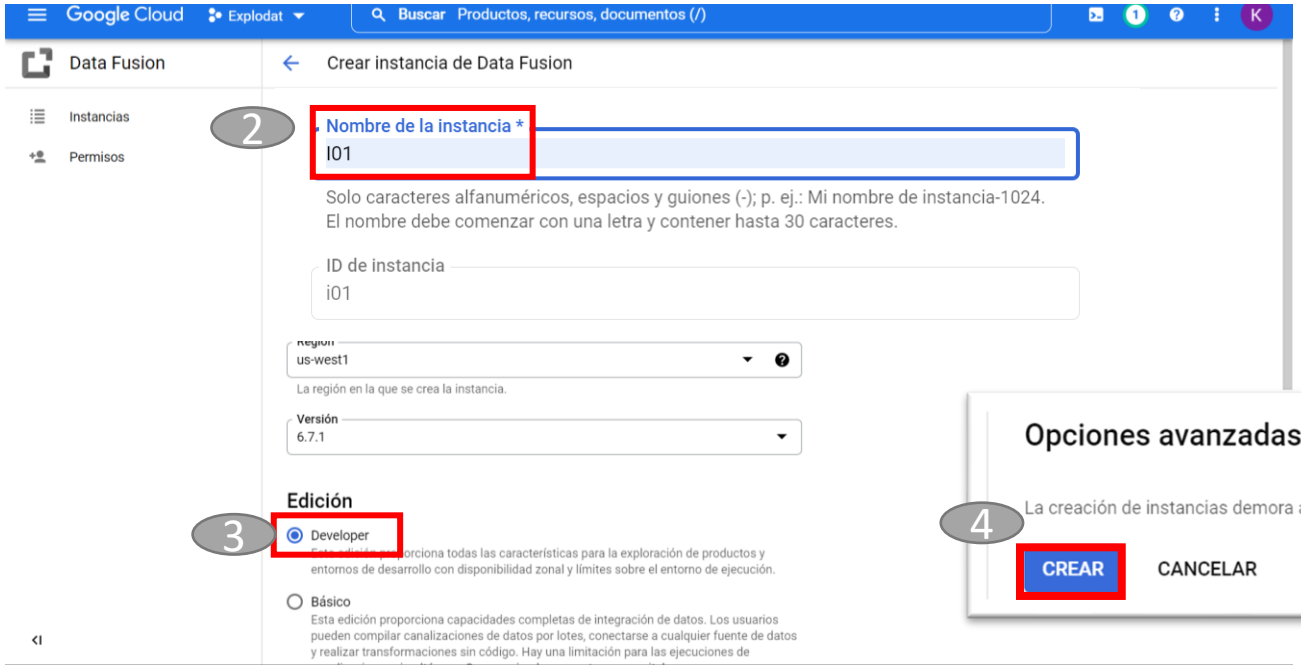


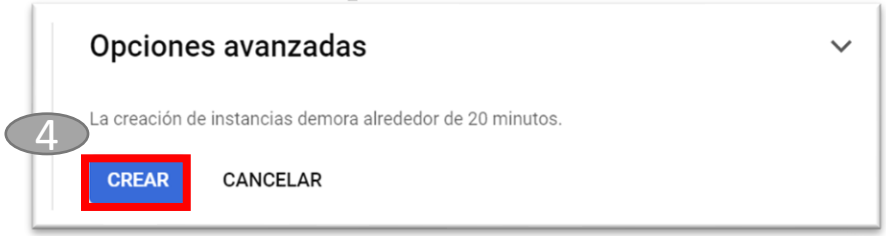
Una vez creado, lo seleccionaremos de la lista disponible para entrar en él

Crear una instancia

Una vez dentro del proyecto, crearemos una Instancia, la cual se encargará de realizar el procesamiento. **En este caso la llamaremos I01.**
Atención: Este proceso puede demorar varios algunos minutos.

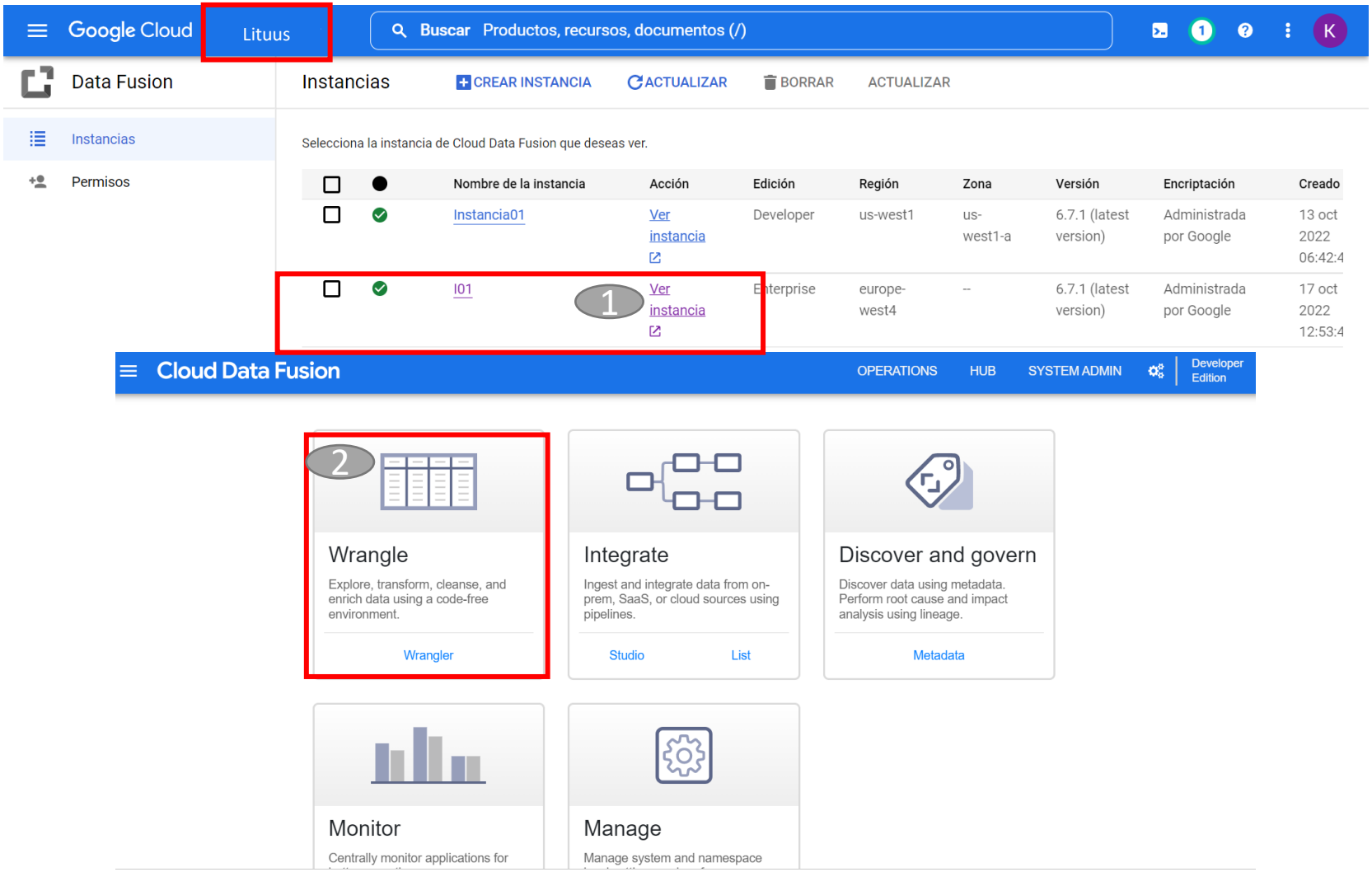






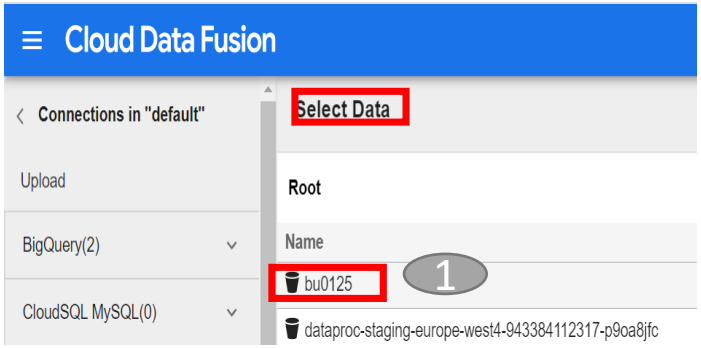
Vista dentro de la instancia

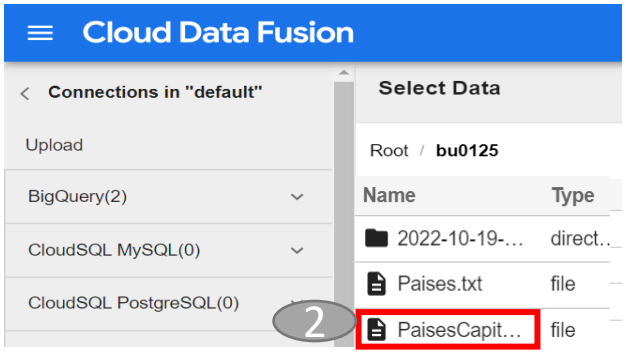
Una vez creada la instancia, ya podemos usarla para desarrollar. Seleccionamos Ver Instancia. Nos aparecerán 5 herramientas de trabajo de Data Fusion, en esta ocasión, ocuparemos **Wrangle.**



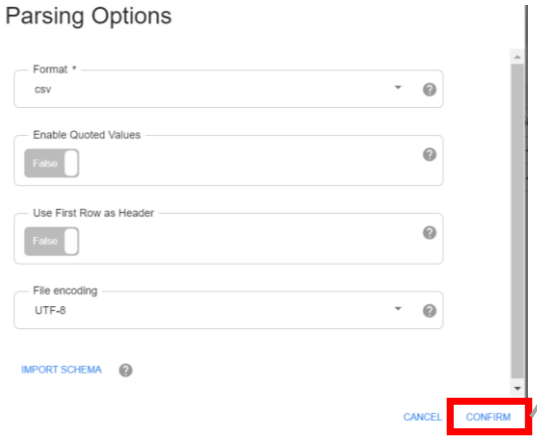
Estructurar archivo con Wrangler

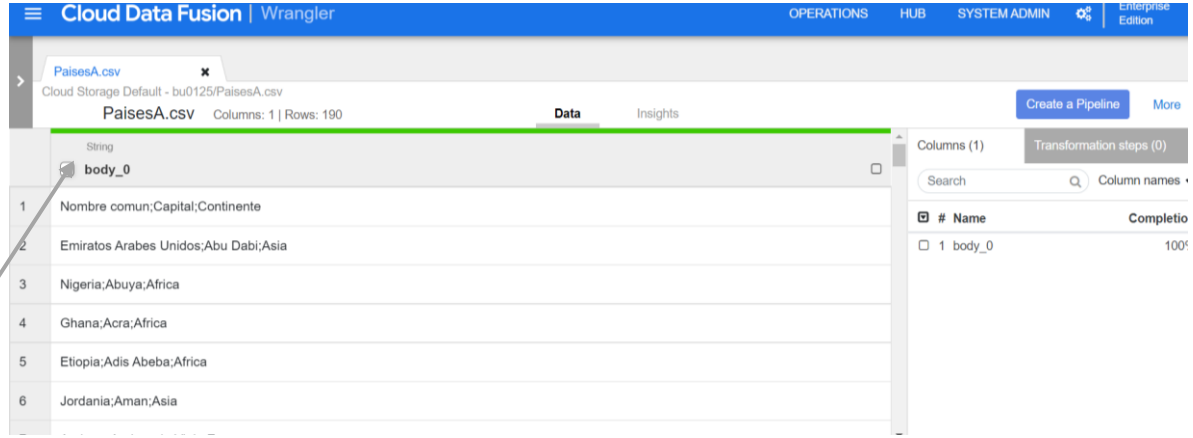
Dentro de Wrangler debemos indicar la fuente de información a procesar, en este caso comenzaremos con el archivo PaisesCapitalContinente.csv guardado previamente en un bucket en Google Cloud Storage (GCS).
Seleccionamos nuestro bucket (para el ejemplo bu0125) y luego el archivo de interés.





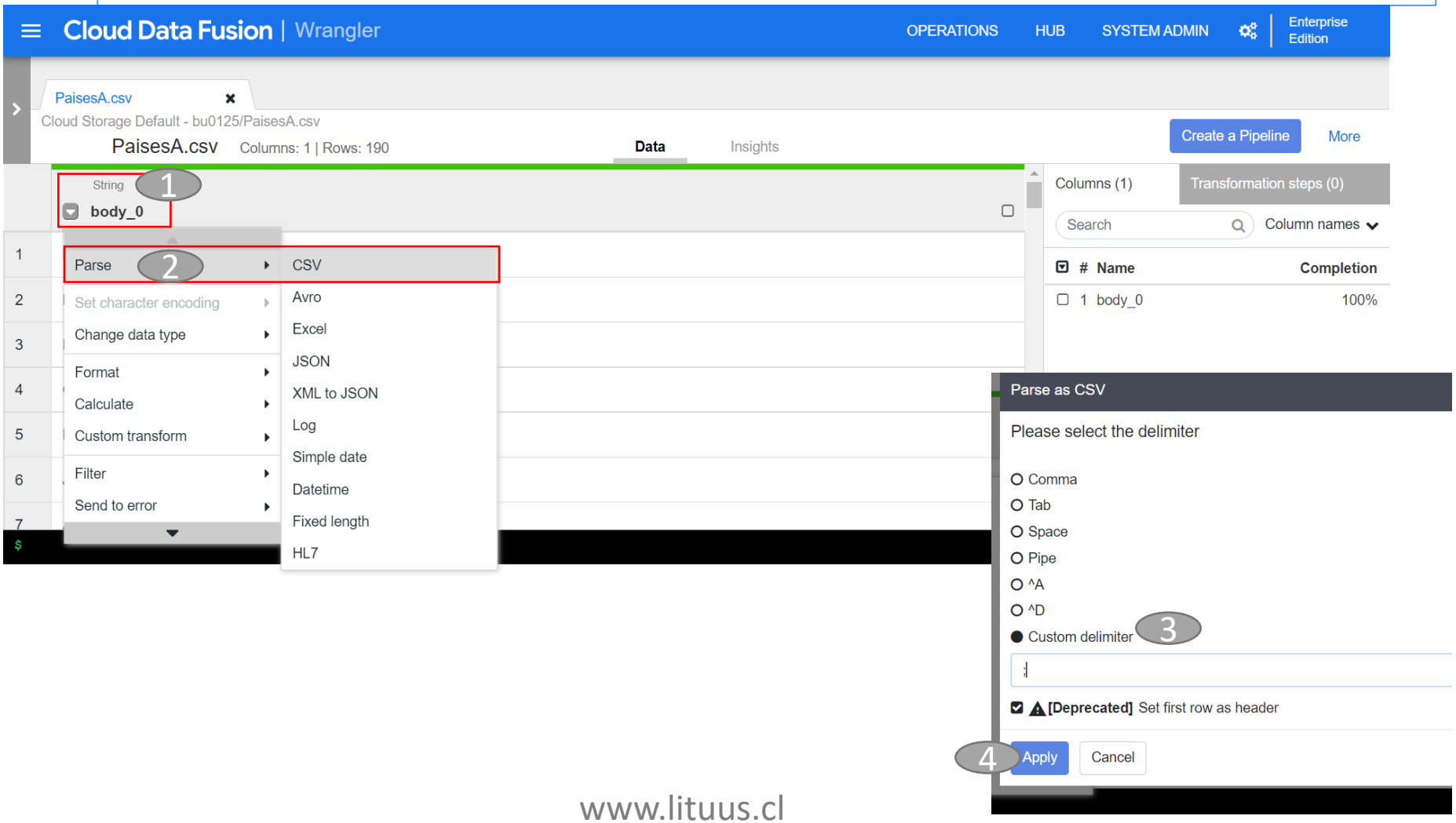
Luego de confirmar las opciones de Parseo (Parsing), se puede trabajar con Wrangler el cual detecta el contenido del archivo (body) pero aún no reconoce las columnas.





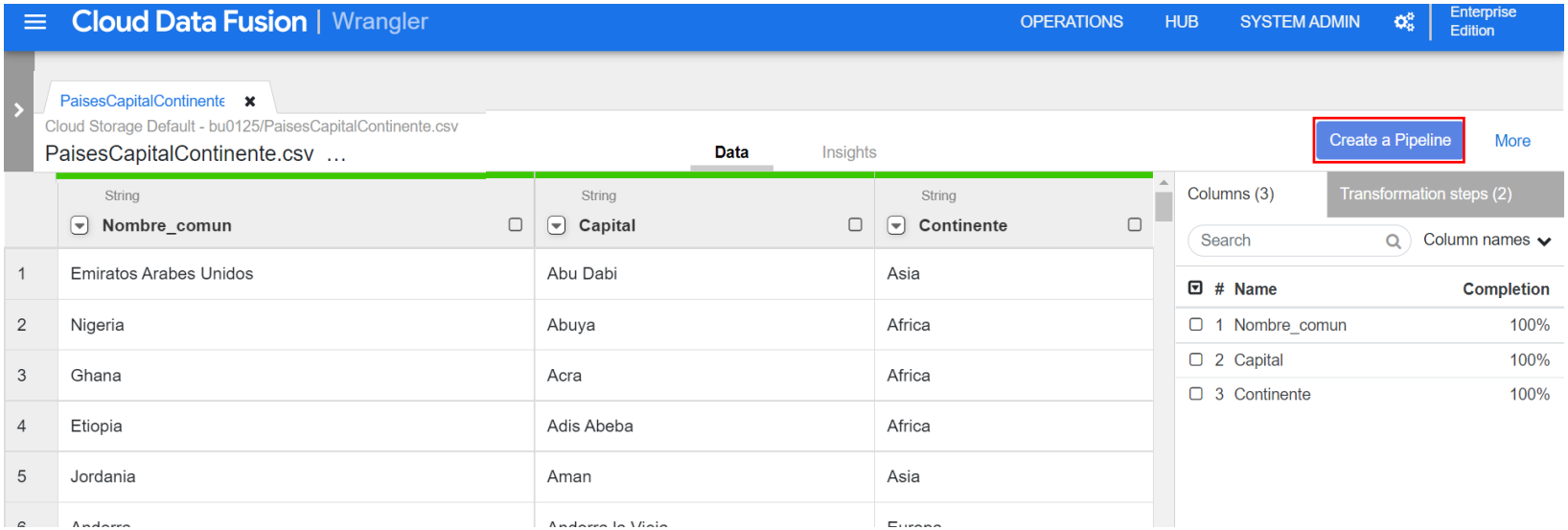
Detección de columnas

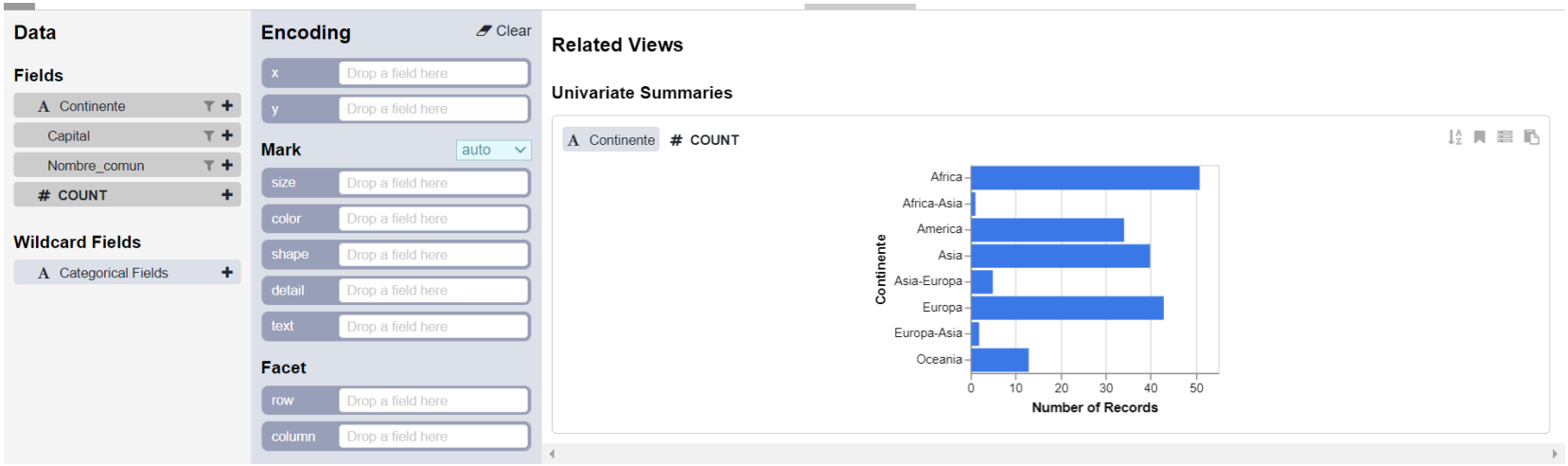
Para indicar cuales son las columnas del archivo csv, seleccionamos “body_0” , luego “parse” y finalmente “CSV”. En la ventana “Parse as CSV” se debe indicar el delimitador que en nuestro caso es “|”, además indicaremos que la primera fila contiene los encabezados.



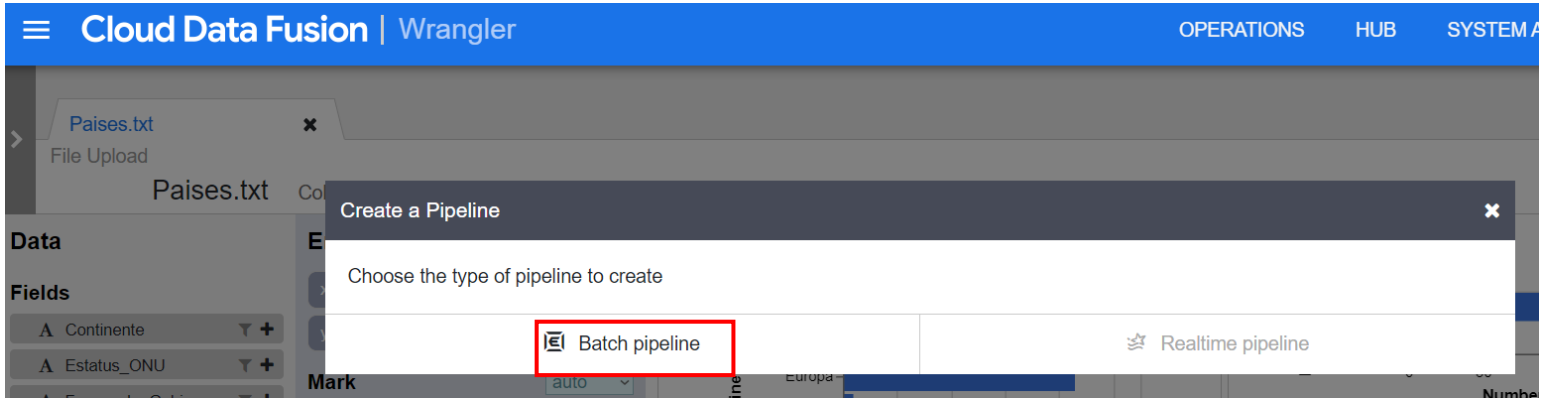
Primera vista de datos

Dentro del Wrangler, podemos observar los datos en forma de tabla y también con gráficos (Denominado Insights). **Seleccionamos Create a Pipeline**

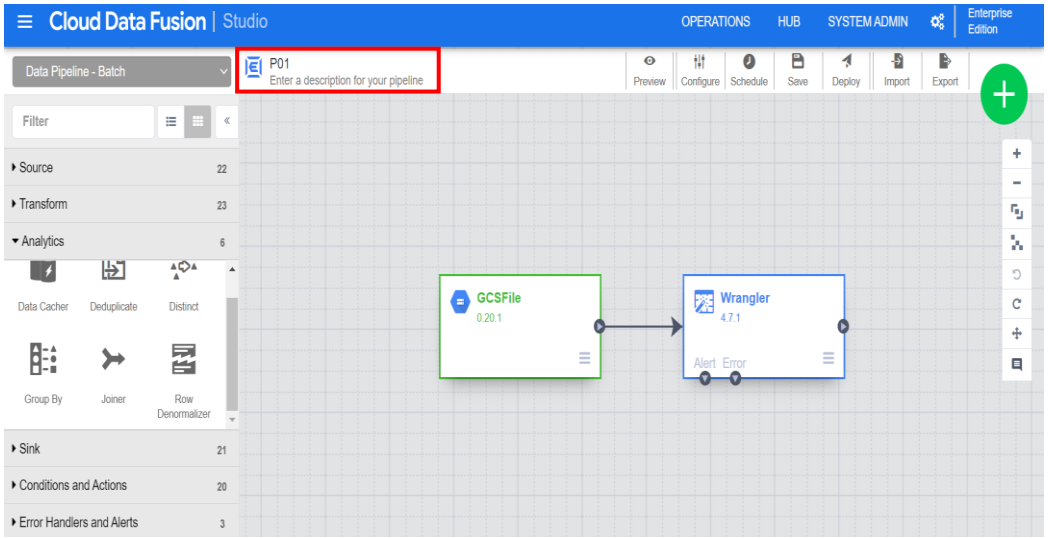




Luego de presionar *Create a Pipeline*, seleccionaremos *Batch pipeline*. Esto nos redirigirá al canvas llamado Studio. Esta es la zona donde realizaremos gráficamente nuestro proceso de ETL.



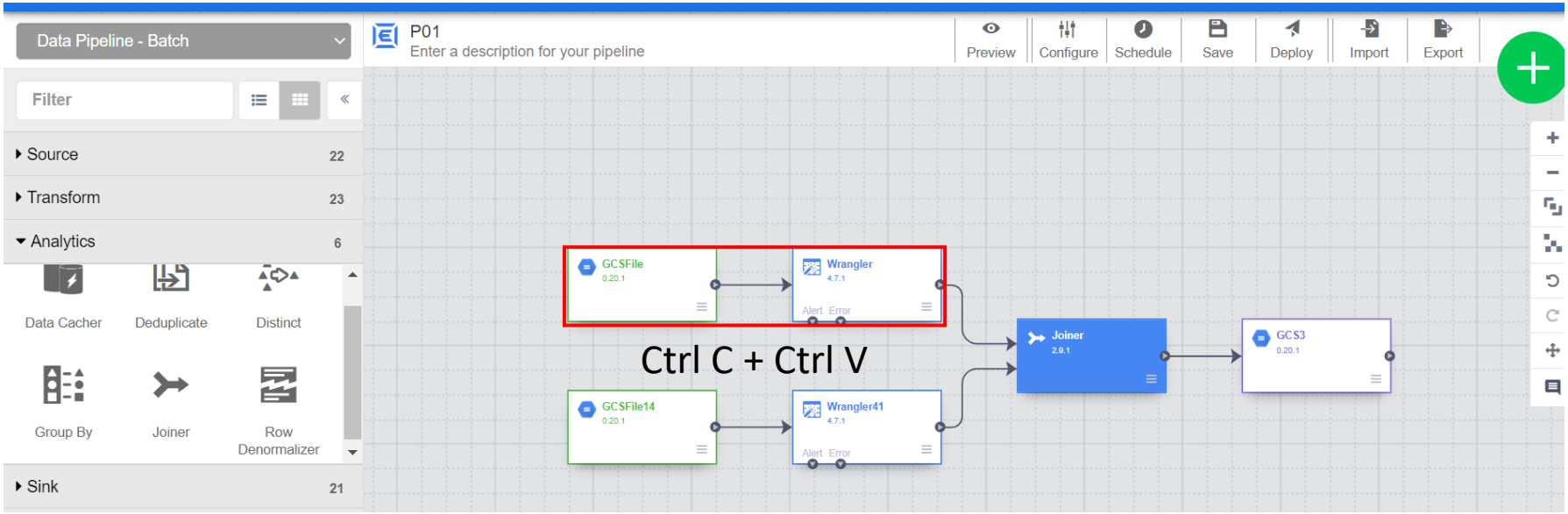
En un primer momento se observa el Wrangler y su archivo de entrada “GCSFile”, ambos preconfigurados.



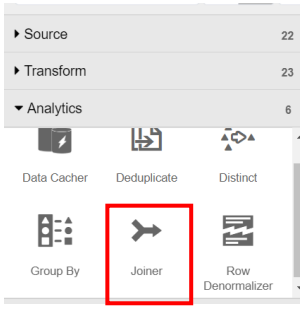
Nota: Debe asignar un nombre a la canalización, en este caso la llamaremos “P01”.

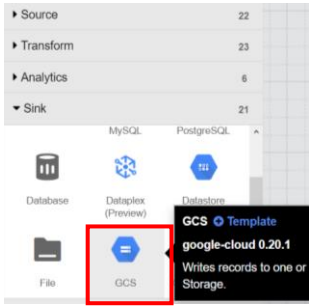
Desarrollo de Canalización

Recordar que consideramos dos orígenes de datos que deben ser estructurados con sus respectivos Wrangler. Para poder unir dichos archivos usaremos el objeto "**Joiner**", el cual nos entregará como salida un archivo consolidado almacenado en Google Cloud Storage "**GCS3**". Una vez lista la canalización procederemos a configurar y validar cada uno de sus elementos.

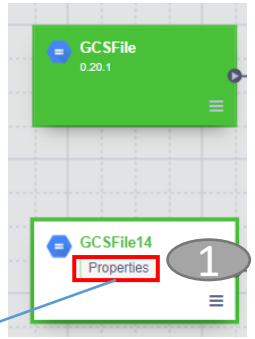


Para ahorrar tiempo podemos copiar y pegar los objetos "GCSFile y Wrangler" (1) generados por defecto al crear la canalización, para luego adaptarlos al archivo que nos falta por configurar.

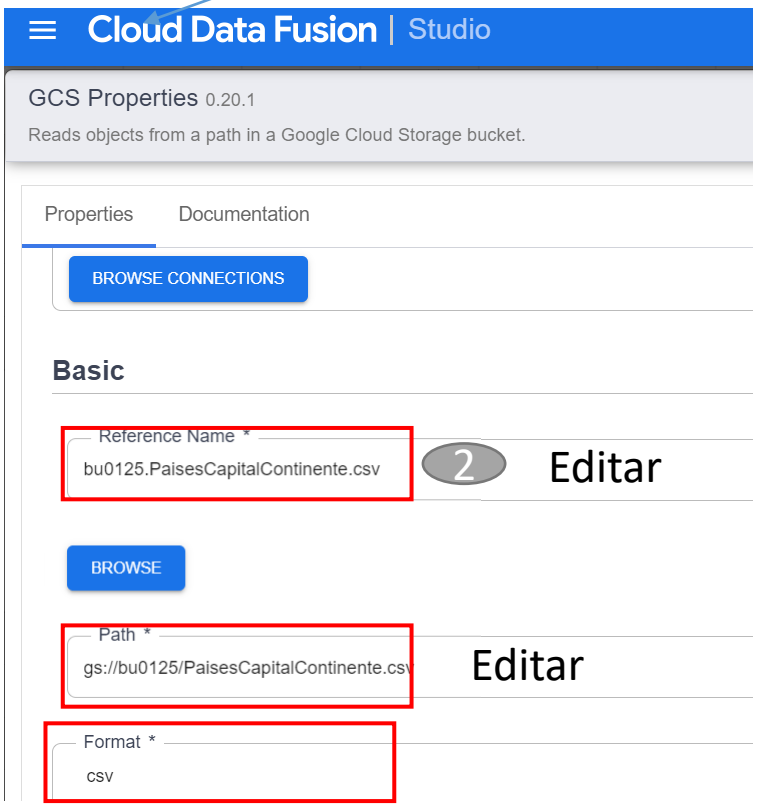


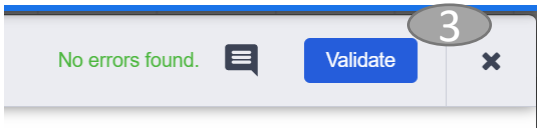


Segundo Archivo .csv

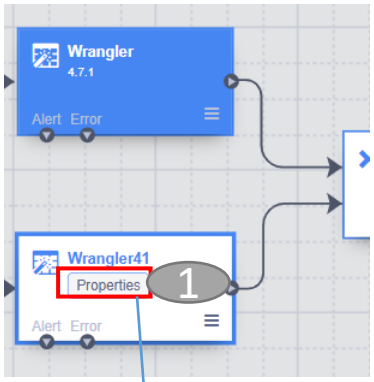


La primera entrada GCSFile ya está configurada con el archivo PaisesCapitalContinente.csv, por lo que ahora se debe indicar la ruta del segundo archivo. Debido al copy/paste el objeto quedo con el nombre GCSFile14 el cual puede ser cambiado en cualquier momento.
Indicaremos que el archivo a procesar se llama **PaisesGobiernoOnu.csv** al editar las opciones en “Propierties”

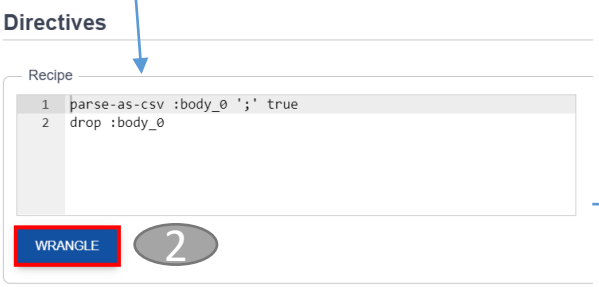


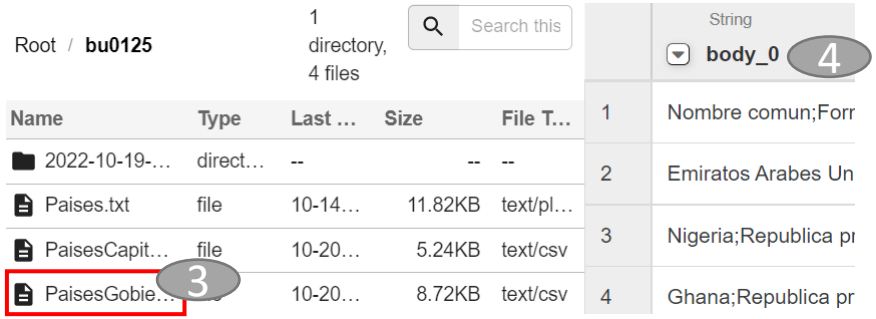


Cuando ya estén ingresados los datos, debe “Validar” que el objeto no tenga errores.



El primer elemento Wrangler, ya está configurado con el archivo PaisesCapitalContinente.csv, por lo que ahora se debe configurar el segundo elemento Wrangler41 para el archivo **PaisesGobiernoOnu.csv**. Entrar a las propiedades del Wrangler y realizar una configuración equivalente a la descrita en diapositivas anteriores.

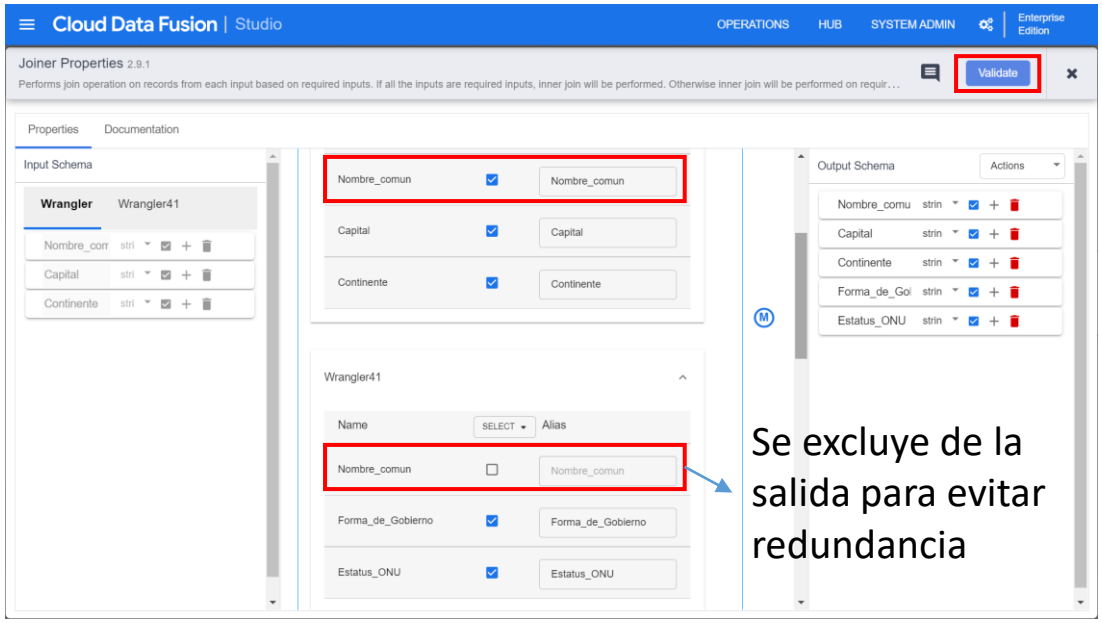




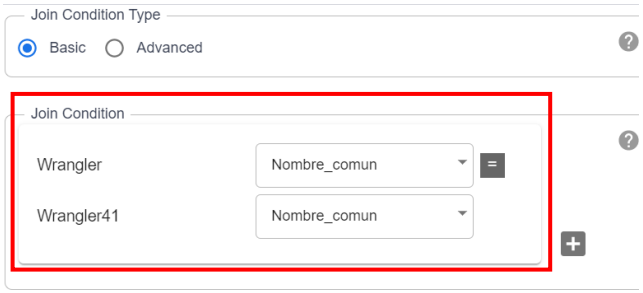
Una vez hecha la configuración no olvide “Validar” que el objeto no tenga errores.

Inner Join con Joiner

A continuación, uniremos los dos archivos (Join) usando el objeto llamado “**Joiner** ”. Entrar a sus propiedades y elegir la(s) columna(s) que deben estar incluidas en el archivo de salida y bajo que criterio se unirán los archivos.



Se excluye de la salida para evitar redundancia

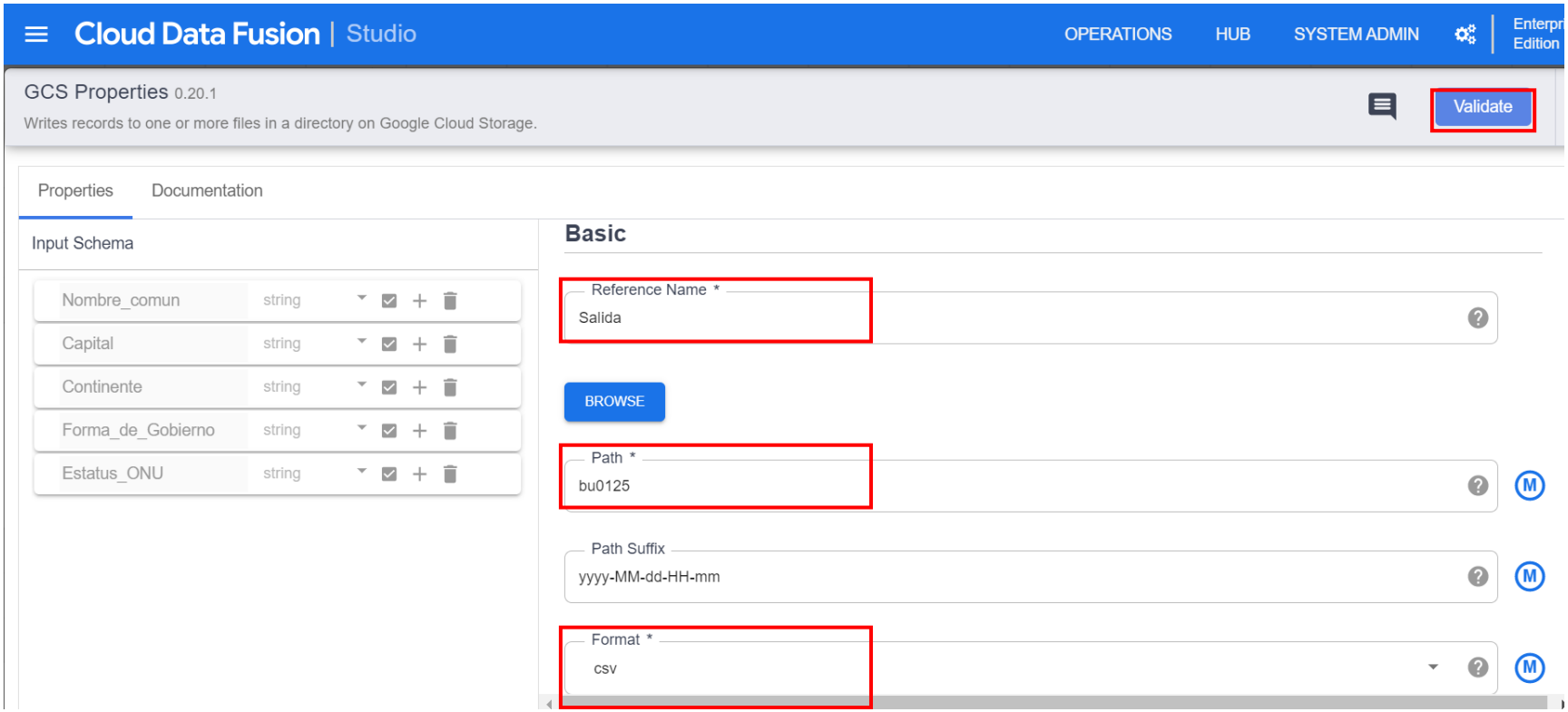


Se realiza un join básico o directo considerando la columna común en ambos archivos. Es equivalente a un inner join

Una vez hecha la configuración no olvide “Validar” que el objeto no tenga errores.

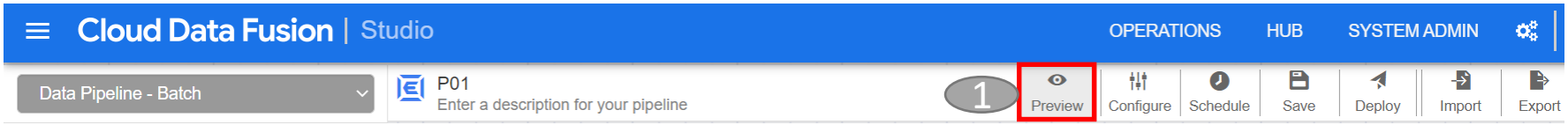
Ubicación Archivo Salida

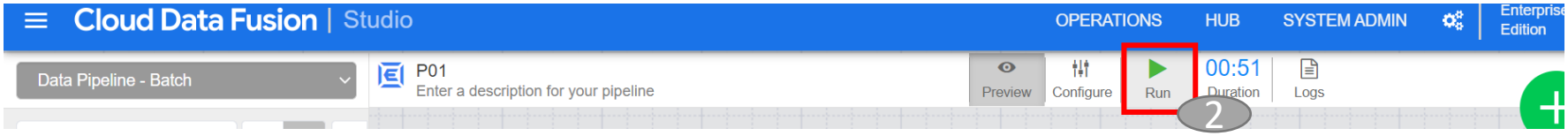
Nuestro último paso es indicar donde quedará almacenado el archivo consolidado. Usaremos nuevamente un Google Cloud Storage (En Data Fusion está en el apartado Sink). Entrar a sus propiedades e indicar un nombre, el bucket donde quedará almacenado y que tipo de archivo será, en este caso un .csv

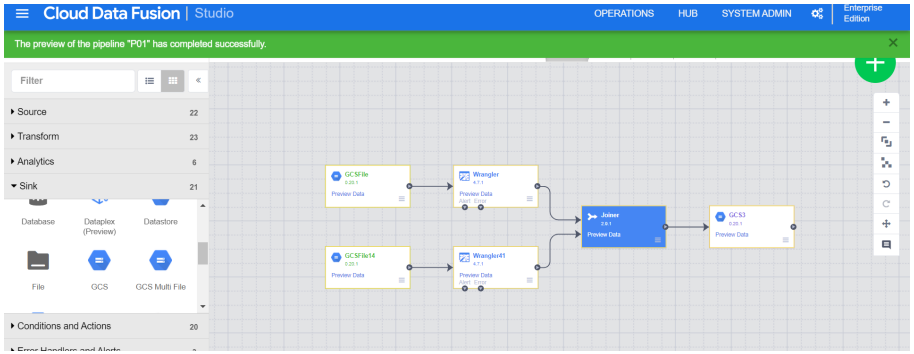


Una vez hecha la configuración no olvide Validar que el objeto no tenga errores.

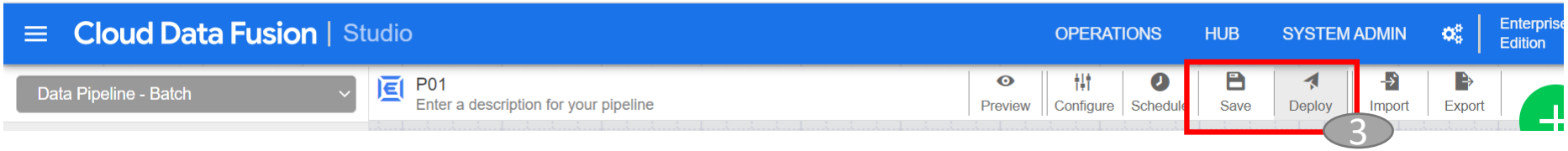
Preview: Resultados Preliminares





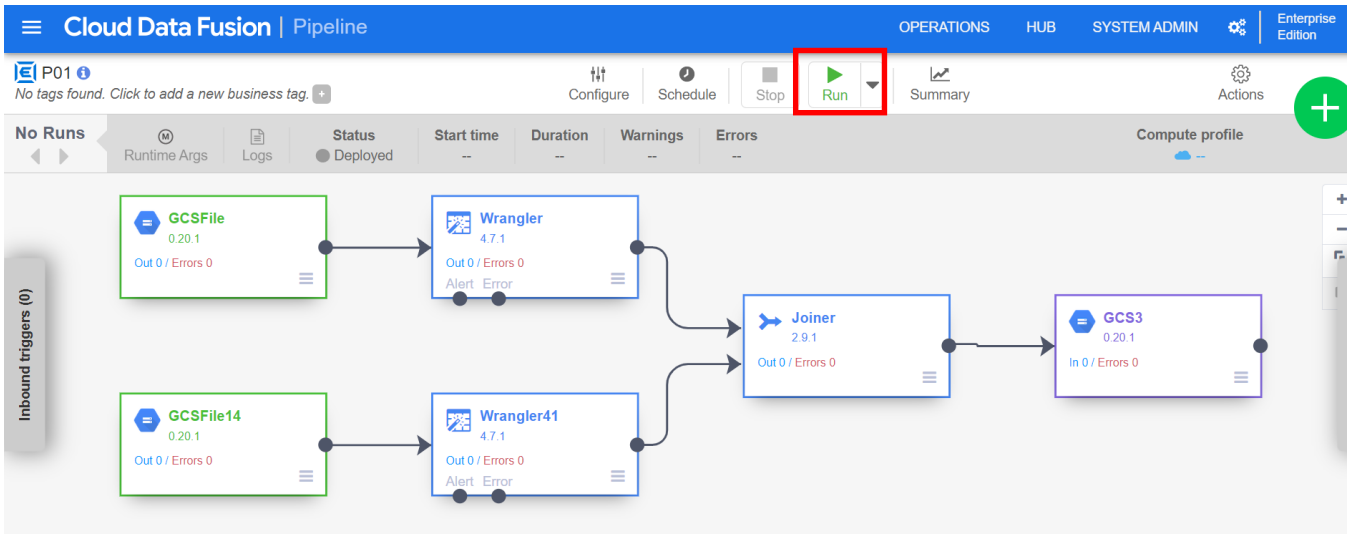


Presionamos el botón “Run” (demora unos minutos). Si el resultado es “Exitoso” (Successfully), debe presionar nuevamente el botón “Preview” y guardar la canalización antes de ejecutarla definitivamente. Finalmente presionar el botón “Deploy”.

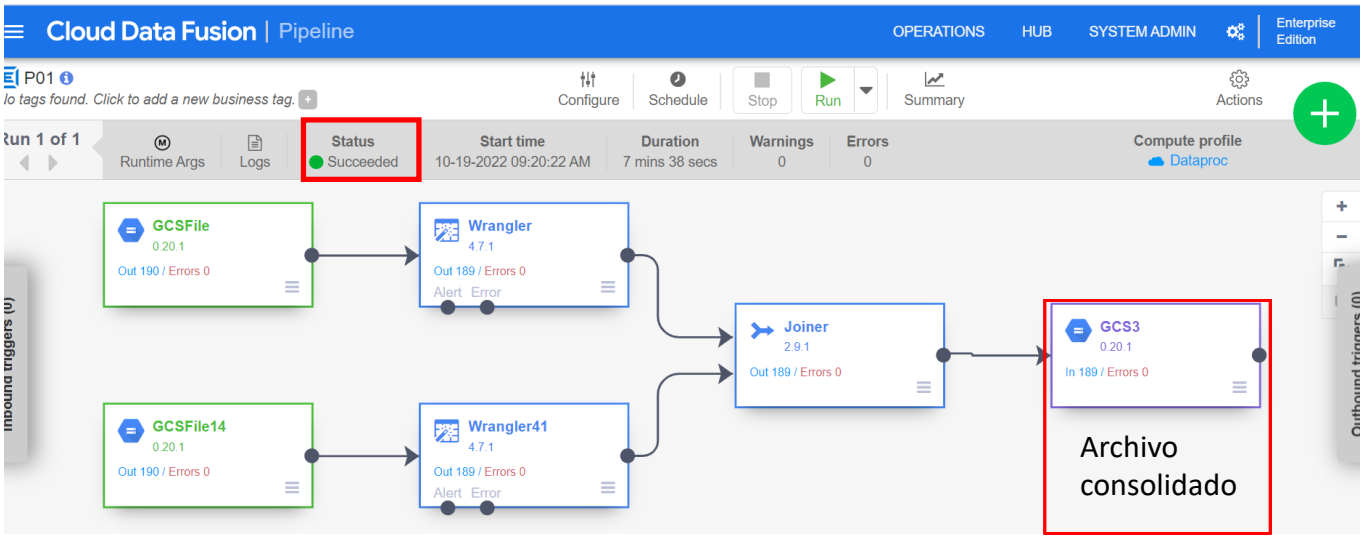


Ejecutar Canalización

Para ejecutar la canalización clic en el botón “Run”, puede tomar un par de minutos



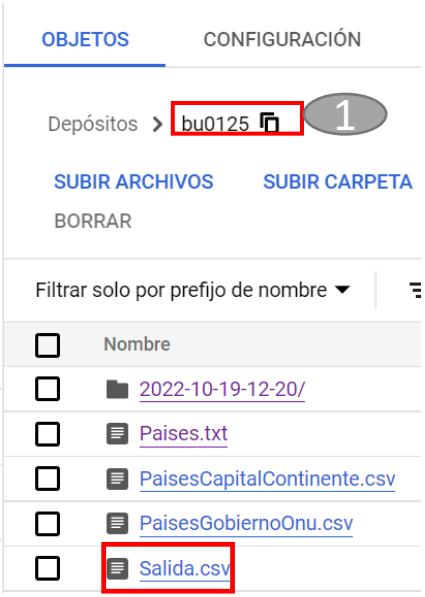
Si el Status de la canalización cambió a “Succeded”(Existosa), podremos revisar el archivo resultante en la ruta asignada en la configuración GCS3 (bucket)

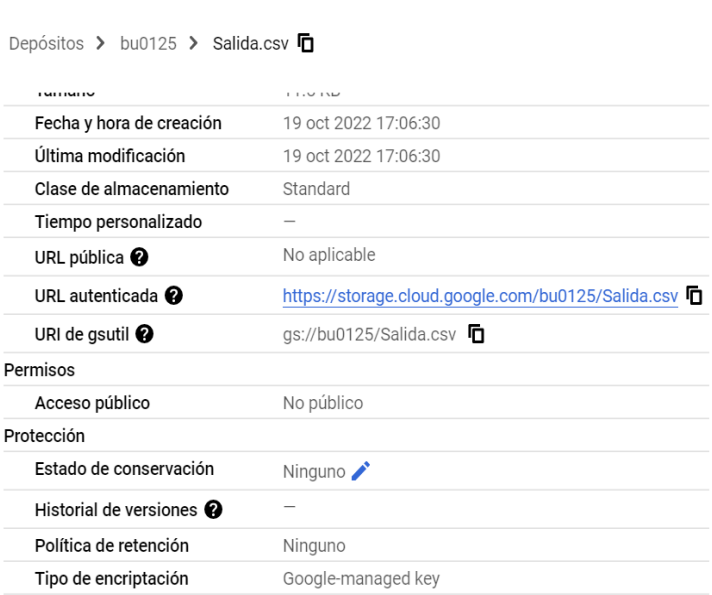


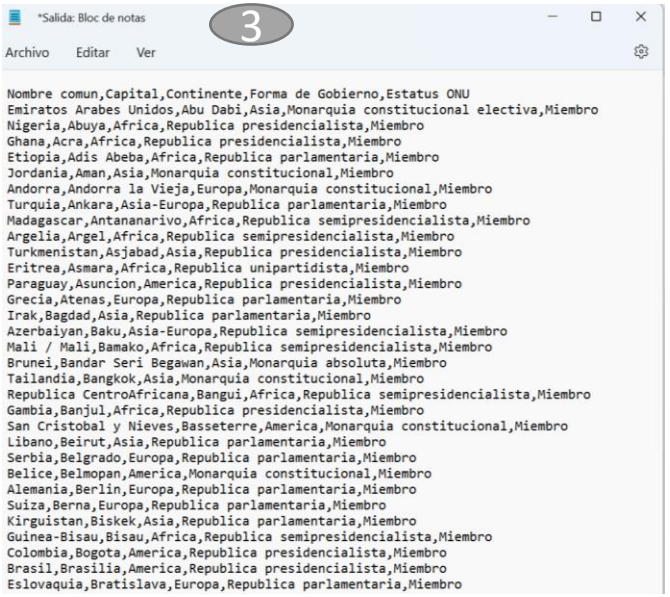
Archivo Salida en Bucket

En el interior del bucket se encuentra el nuevo archivo llamado Salida.csv. Este archivo contiene tanto las columnas del archivo PaisesCapitalContinente.csv como las columnas del archivo PaisesGobiernoOnu.csv.

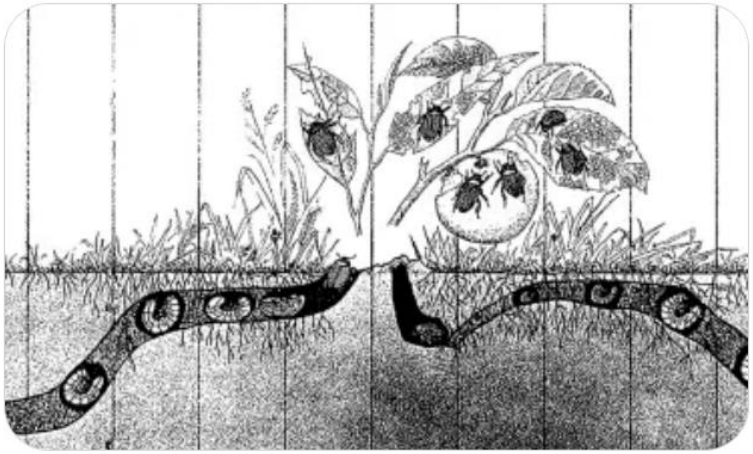
Nota: Este archivo puede ser descargado del bucket al pc sin problema.







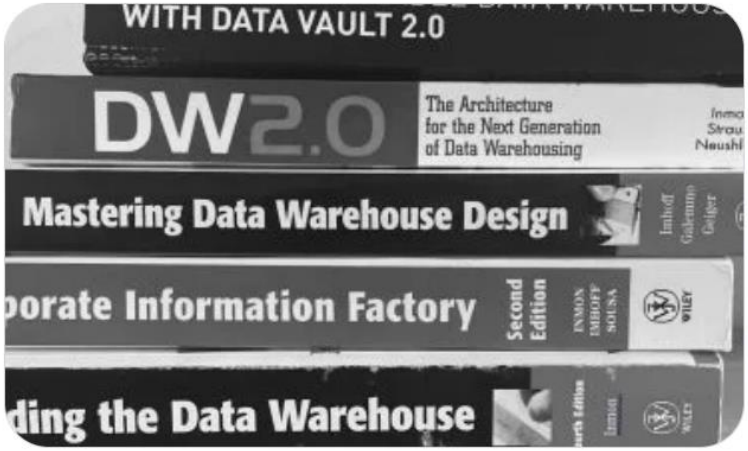
Más Contenido



¿Qué es la visualización de datos cuantitativos y por qué es importante?

11/03/2021 Lenin González

Intentaré convencerte de por qué debemos aprender visualización de datos cuantitativos. Verás ejemplos, un poco de historia y cuatro recomendaciones prácticas que puedes aplicar desde hoy mismo.



¿Qué es un Data Warehouse y Para Qué Sirve?

18/01/2021 Lenin González

En la medida que las empresas requieren ser más competitivas suelen incorporar Software que les permita gestionar sus procesos de negocios de forma tal que se vuelvan predecibles y optimizables. Cada proceso de negocio tiene su complejidad y vida propia, por lo que sin quererlo ni beberlo, las organizaciones comienzan adquirir más Software o Módulos encargados de administrar sus procesos de forma atómica, durable y concurrente. Suele suceder que dichos sistemas/módulos no usan las mismas tecnologías, puede que no sean del mismo proveedor o sus arquitecturas sean incompatibles. Es Aquí cuando un Data Warehouse puede aportar credibilidad en la toma de decisiones al integradar en una sola fuente de la verdad nuestros dato

Contacto

Lituus

Presentado por:

- Lenin González S.
- Director General, CEO
- +569 3072 9405
- lgonzalez@lituus.cl