



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

Cloud Data Fusion | Studio OPERATIONS HUB SYSTEM ADMIN Enterprise Edition

The preview of the pipeline "P01" has completed successfully.

Filter

- Source 22
- Transform 23
- Analytics 6
- Sink 21
  - Database
  - Dataplex (Preview)
  - Datastore
  - File
  - GCS
  - GCS Multi File
- Conditions and Actions 20
- Error Handlers and Alerts

```
graph LR; GCSFile[GCSFile 0.20.1] --> Wrangler[Wrangler 4.7.1]; GCSFile14[GCSFile14 0.20.1] --> Wrangler41[Wrangler41 4.7.1]; Wrangler --> Joiner[Joiner 2.9.1]; Wrangler41 --> Joiner; Joiner --> GCS3[GCS3 0.20.1];
```



© Copyright **Lituus Spa**

Este documento es propiedad de Lituus Spa y su contenido es confidencial. Este no puede ser reproducido, en su totalidad o parcialmente, ni mostrado a terceros, ni utilizado para otros propósitos que los que han originado su entrega, sin el previo permiso escrito de **Lituus Spa**.



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

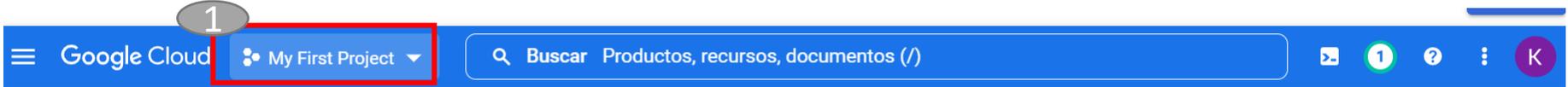
Nombre comun	Capital	Continente
Emiratos Árabes Unidos	Abu Dabi	Asia
Nigeria	Abuya	África
Ghana	Acra	África
Etiopía	Adís Abeba	África
Jordania	Amán	Asia
Andorra	Andorra la Vieja	Europa
Turquía	Ankara	Asia-Europa
Madagascar	Antananarivo	África
Argelia	Argel	África
Turkmenistán	Asjabad	Asia
Eritrea	Asmara	África
Paraguay	Asunción	América
Grecia	Atenas	Europa
Irak	Bagdad	Asia
Azerbaiyán	Bakú	Asia-Europa
Mali / Malí	Bamako	África
Brunéi	Bandar Seri Begawan	Asia
Tailandia	Bangkok	Asia
República Centroafricana	Bangui	África
Gambia	Banjul	África
San Cristóbal y Nieves	Basseterre	América
Líbano	Beirut	Asia
Serbia	Belgrado	Europa
Belice	Belmopán	América
Alemania	Berlín	Europa
Suiza	Berna	Europa
Kirguistán	Biskek	Asia

Nombre comun	Forma de Gobierno	Estatus ONU
Emiratos Árabes Unidos	Monarquía constitucional	el
Nigeria	República presidencialista	Miembro
Ghana	República presidencialista	Miembro
Etiopía	República parlamentaria	Miembro
Jordania	Monarquía constitucional	Miembro
Andorra	Monarquía constitucional	Miembro
Turquía	República parlamentaria	Miembro
Madagascar	República semipresidencialista	Mier
Argelia	República semipresidencialista	Mier
Turkmenistán	República presidencialista	Mier
Eritrea	República unipartidista	Miembro
Paraguay	República presidencialista	Miembro
Grecia	República parlamentaria	Miembro
Irak	República parlamentaria	Miembro
Azerbaiyán	República semipresidencialista	Mier
Mali / Malí	República semipresidencialista	Mier
Brunéi	Monarquía absoluta	Miembro
Tailandia	Monarquía constitucional	Miembro
República Centroafricana	República semipresider	
Gambia	República presidencialista	Miembro
San Cristóbal y Nieves	Monarquía constitucional	
Líbano	República parlamentaria	Miembro
Serbia	República parlamentaria	Miembro
Belice	Monarquía constitucional	Miembro
Alemania	República parlamentaria	Miembro
Suiza	República parlamentaria	Miembro
Kirguistán	República parlamentaria	Miembro

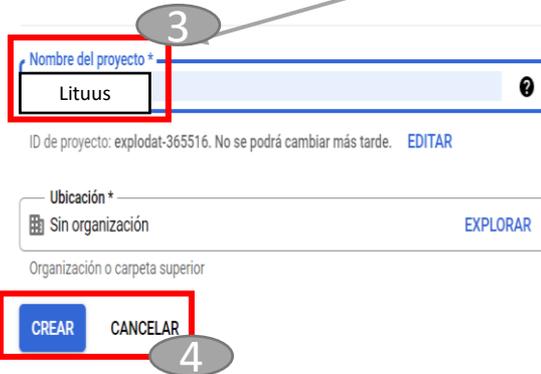


Para crear un proyecto, debemos seleccionar el menú de proyectos. Por defecto tiene el nombre del último proyecto en el que se trabajó:



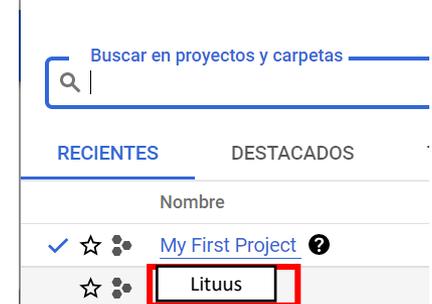
Luego se abre la ventana de selección de proyecto. Clic en el botón "Proyecto Nuevo", asignar un nombre (en este caso Lituus) y presionar el botón "Crear"

Seleccionar un proyecto



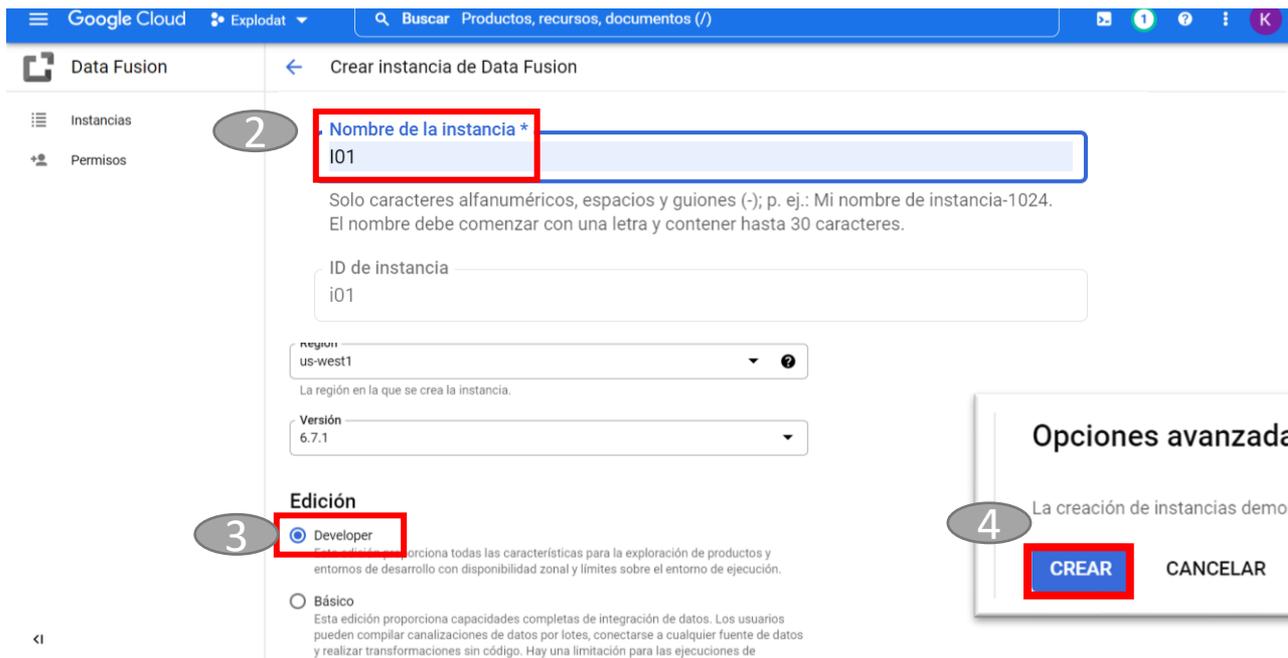
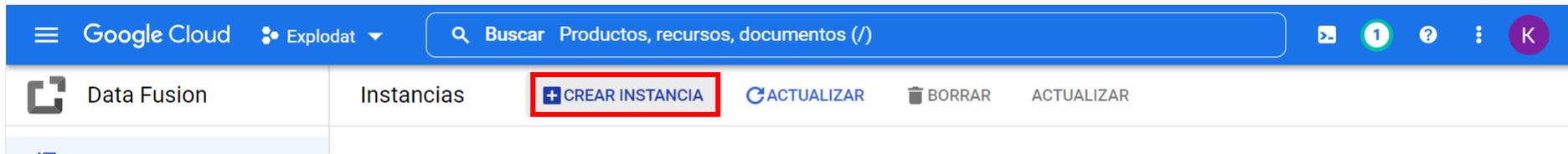
Una vez creado, lo seleccionaremos de la lista disponible para entrar en él

5 Seleccionar un proyecto





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.





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**.

Google Cloud **Lituus**

Data Fusion

Instancias [+ CREAR INSTANCIA](#) [ACTUALIZAR](#) [BORRAR](#) [ACTUALIZAR](#)

Selecciona la instancia de Cloud Data Fusion que deseas ver.

<input type="checkbox"/>	<input checked="" type="radio"/>	Nombre de la instancia	Acción	Edición	Región	Zona	Versión	Encriptación	Creado
<input type="checkbox"/>	<input checked="" type="radio"/>	<a href="#">Instancia01</a>	<a href="#">Ver instancia</a>	Developer	us-west1	us-west1-a	6.7.1 (latest version)	Administrada por Google	13 oct 2022 06:42:4
<input type="checkbox"/>	<input checked="" type="radio"/>	<a href="#">I01</a>	<a href="#">Ver instancia</a>	Enterprise	eu-west4	--	6.7.1 (latest version)	Administrada por Google	17 oct 2022 12:53:4

Cloud Data Fusion OPERATIONS HUB SYSTEM ADMIN [Developer Edition](#)

**2**

### Wrangle

Explore, transform, cleanse, and enrich data using a code-free environment.

[Wrangler](#)

### Integrate

Ingest and integrate data from on-prem, SaaS, or cloud sources using pipelines.

[Studio](#) [List](#)

### Discover and govern

Discover data using metadata. Perform root cause and impact analysis using lineage.

[Metadata](#)

### Monitor

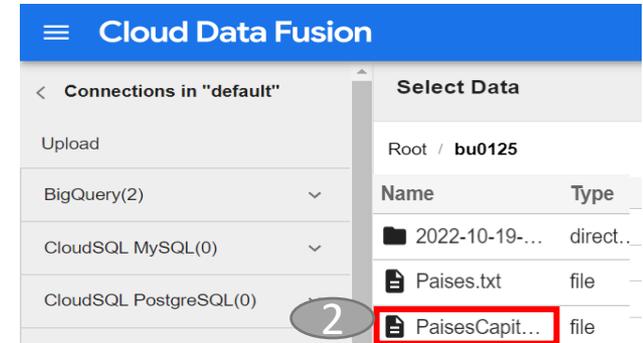
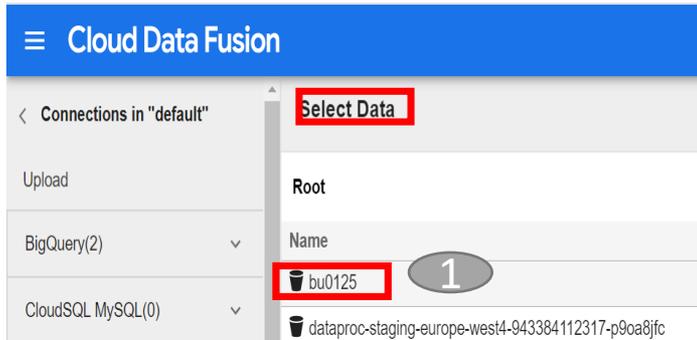
Centrally monitor applications for

### Manage

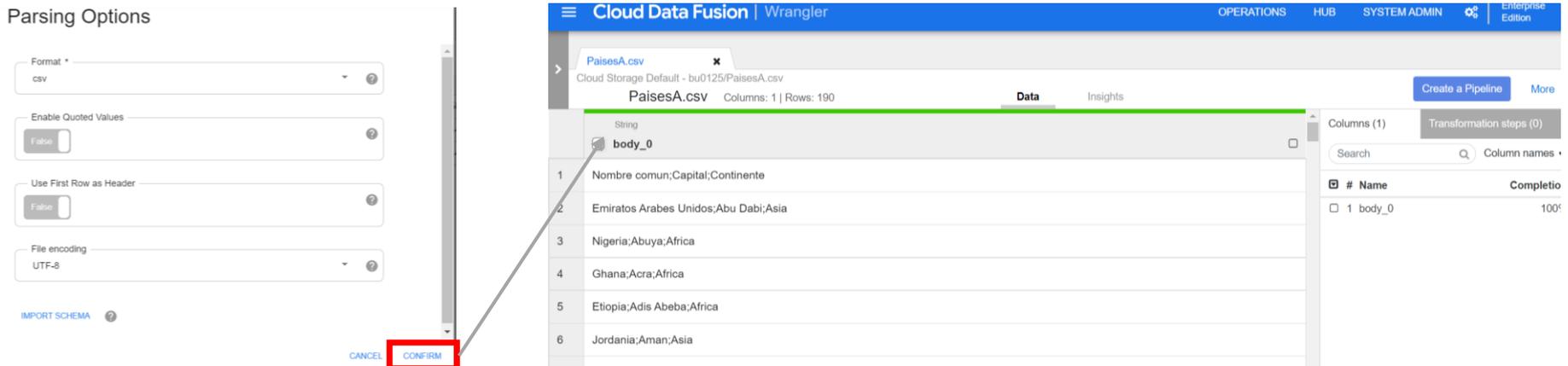
Manage system and namespace



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.





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.

The screenshot shows the Cloud Data Fusion Wrangler interface. The top navigation bar includes "Cloud Data Fusion | Wrangler", "OPERATIONS", "HUB", "SYSTEM ADMIN", and "Enterprise Edition". The main area displays a file named "PaisesA.csv" with "Columns: 1 | Rows: 190". A table with one column, "body\_0", is visible. A transformation menu is open, showing options like "Parse", "Set character encoding", "Change data type", "Format", "Calculate", "Custom transform", "Filter", and "Send to error". The "Parse" option is selected, and a "Parse as CSV" dialog box is open. The dialog box prompts the user to "Please select the delimiter" and lists options: Comma, Tab, Space, Pipe, ^A, ^D, and Custom delimiter. The "Custom delimiter" option is selected, and the delimiter is set to "|". The "[Deprecated] Set first row as header" checkbox is checked. The "Apply" button is highlighted with a callout number 4.

#	Name	Completion
1	body_0	100%

**Parse as CSV**

Please select the delimiter

- Comma
- Tab
- Space
- Pipe
- ^A
- ^D
- Custom delimiter

|

[Deprecated] Set first row as header

Apply Cancel



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

The screenshot shows the Cloud Data Fusion Wrangler interface. At the top, there's a blue header with 'Cloud Data Fusion | Wrangler' and navigation links for 'OPERATIONS', 'HUB', 'SYSTEM ADMIN', and 'Enterprise Edition'. Below the header, a breadcrumb trail shows 'PaísesCapitalContinente'. The main area is split into 'Data' and 'Insights' tabs. The 'Data' tab is active, displaying a table with columns 'Nombre\_comun', 'Capital', and 'Continente'. A 'Create a Pipeline' button is highlighted with a red box in the top right corner. To the right of the table, there's a sidebar with 'Columns (3)' and 'Transformation steps (2)'. The 'Columns' section shows a search bar and a dropdown for 'Column names'. The 'Transformation steps' section shows a table with columns '#', 'Name', and 'Completion'.

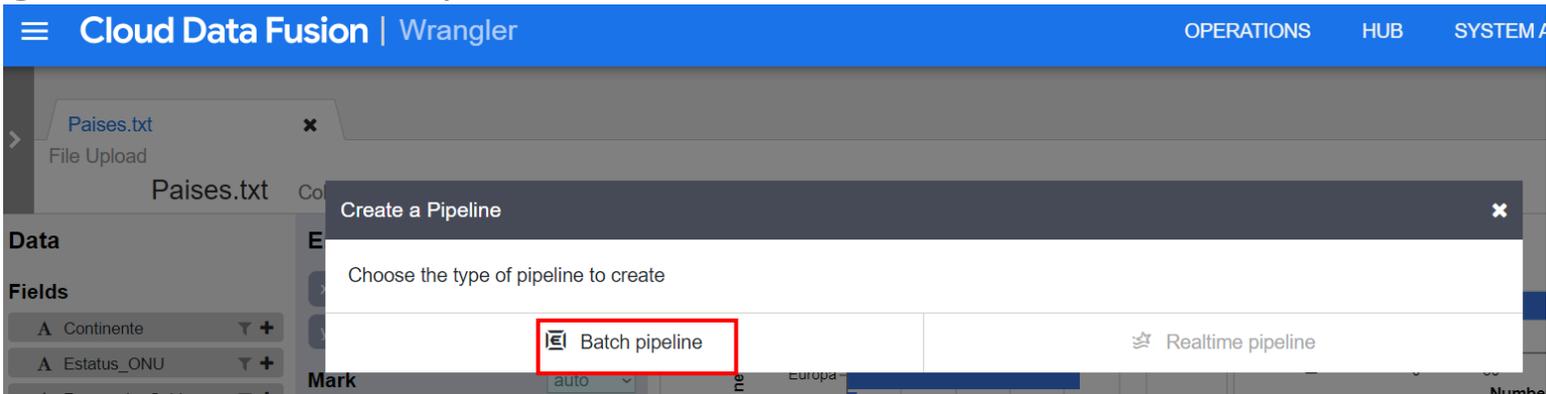
#	Name	Completion
<input type="checkbox"/>	1 Nombre_comun	100%
<input type="checkbox"/>	2 Capital	100%
<input type="checkbox"/>	3 Continente	100%

The screenshot shows the 'Data' and 'Insights' panels in the Cloud Data Fusion Wrangler interface. The 'Data' panel on the left shows a list of fields: 'Continente', 'Capital', 'Nombre\_comun', and '# COUNT'. The 'Insights' panel on the right shows a horizontal bar chart titled 'Univariate Summaries' for the field 'Continente'. The chart displays the number of records for each continent: Africa (50), Africa-Asia (35), America (35), Asia (40), Asia-Europa (5), Europa (45), Europa-Asia (5), and Oceania (15).

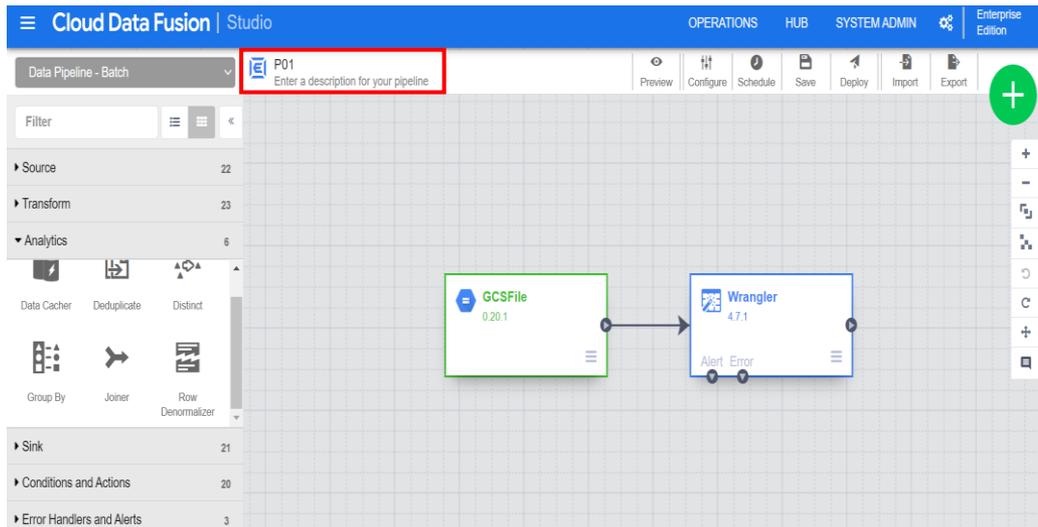
Continente	Number of Records
Africa	50
Africa-Asia	35
America	35
Asia	40
Asia-Europa	5
Europa	45
Europa-Asia	5
Oceania	15



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.



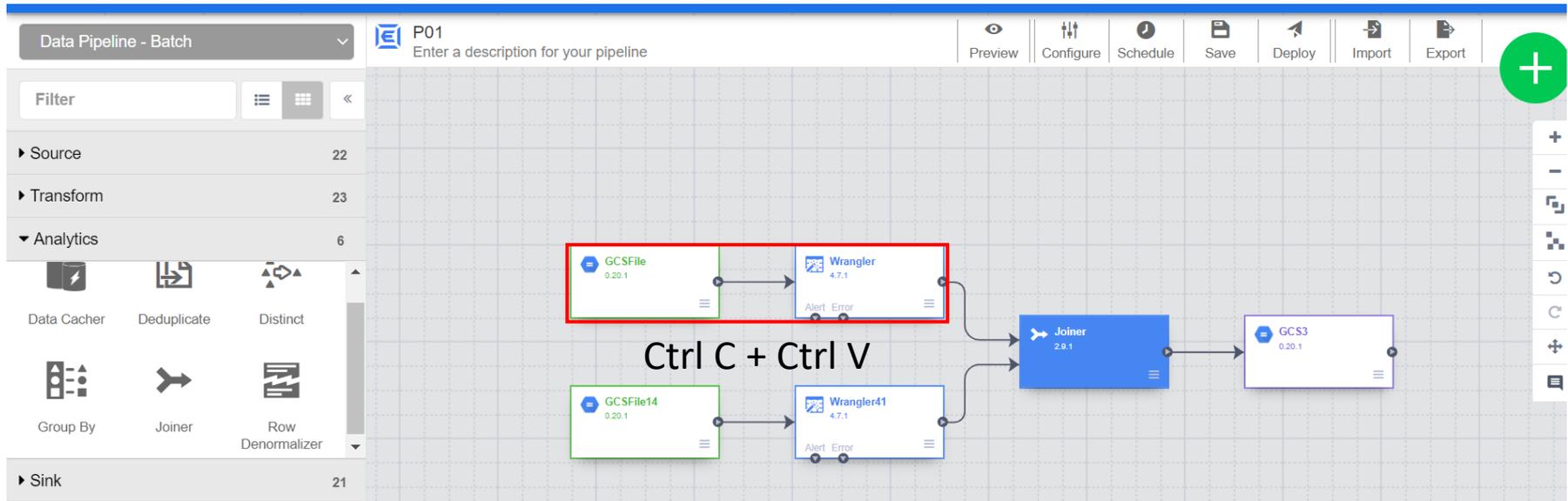
En un primer momento se observa el Wrangler y su archivo de entrada “GCSFile”, ambos preconfigurados.



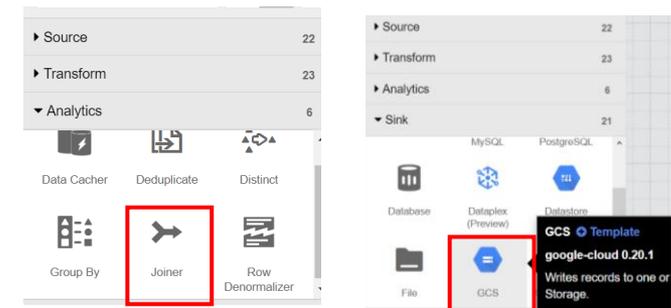
Nota: Debe asignar un nombre a la canalización, en este caso la llamaremos “P01”.

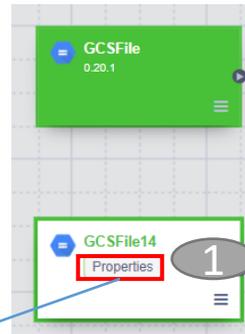


Recordar que consideramos dos orígenes de datos que deben ser estructurados con sus respectivos Wrangler. Para poder unir dichos archivos usaremos el objeto “**Joiner**”, el cual nos entregará como salida un archivo consolidado almacenado en Google Cloud Storage “**GCS3**”. Una vez lista la canalización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



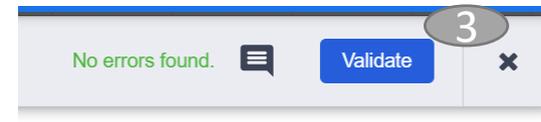
Para ahorrar tiempo podemos copiar y pegar los objetos “GCSFile y Wrangler” (1) generados por defecto al crear la canalización, para luego adaptarlos al archivo que nos falta por configurar.



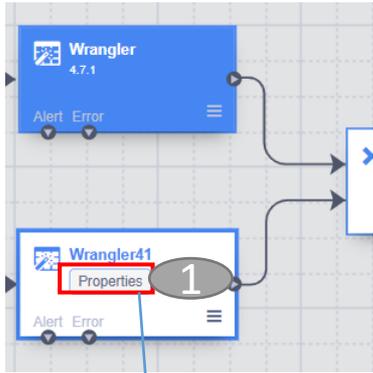


La primera entrada GCSFile ya está configurada con el archivo PaisesCapitalContinente.csv, por lo que ahora se debe indicar la ruta del segundo archivo. Debido al copy/paste el objeto quedo con el nombre GCSFile14 el cual puede ser cambiado en cualquier momento.

Indicaremos que el archivo a procesar se llama **PaisesGobiernoOnu.csv** al editar las opciones en “Propiedades”



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



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

## Directives

```
Recipe
1 parse-as-csv :body_0 ';' true
2 drop :body_0
```

**WRANGLE** 2

Root / bu0125 1 directory, 4 files

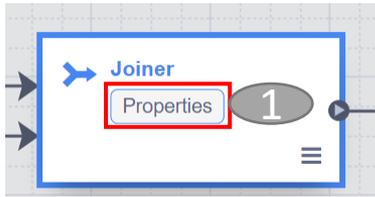
Name	Type	Last ...	Size	File T...	String
					<input checked="" type="checkbox"/> body_0 4
1					Nombre comun;Forr
2					Emiratos Arabes Un
3					Nigeria;Republica pr
4					Ghana;Republica pr

File list table:

Name	Type	Last ...	Size	File T...
2022-10-19-...	direct...	--	--	--
Países.txt	file	10-14...	11.82KB	text/pl...
PaísesCapit...	file	10-20...	5.24KB	text/csv
PaísesGobie...	file	10-20...	8.72KB	text/csv

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





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.

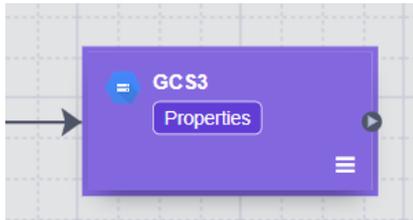
Se excluye de la salida para evitar redundancia

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.



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



Cloud Data Fusion | Studio

OPERATIONS HUB SYSTEM ADMIN

GCS Properties 0.20.1

Writes records to one or more files in a directory on Google Cloud Storage.

Validate

Properties Documentation

Input Schema

Field Name	Type	Required	Visible	Actions
Nombre_comun	string	Yes	Yes	+ -
Capital	string	Yes	Yes	+ -
Continente	string	Yes	Yes	+ -
Forma_de_Gobierno	string	Yes	Yes	+ -
Estatus_ONU	string	Yes	Yes	+ -

**Basic**

Reference Name \*  
Salida

BROWSE

Path \*  
bu0125

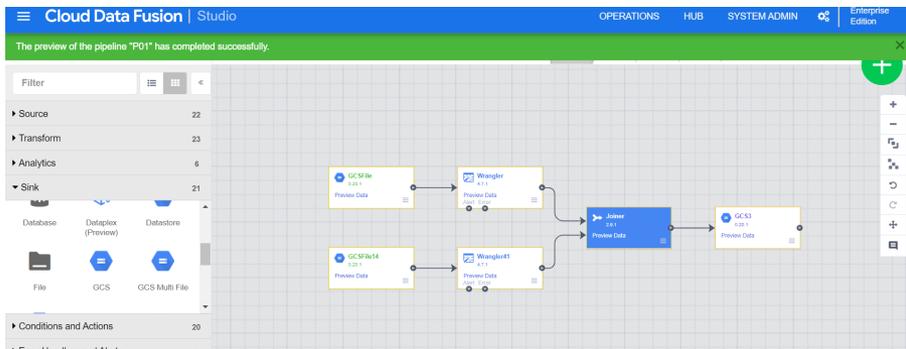
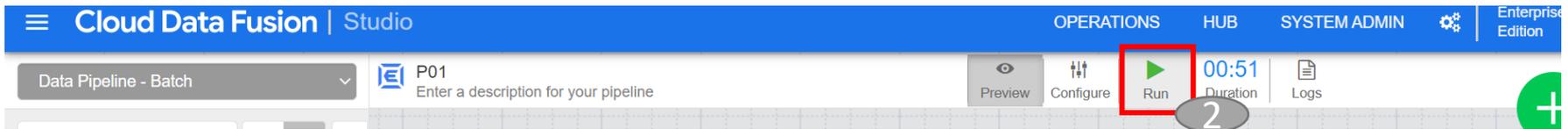
Path Suffix  
yyyy-MM-dd-HH-mm

Format \*  
csv

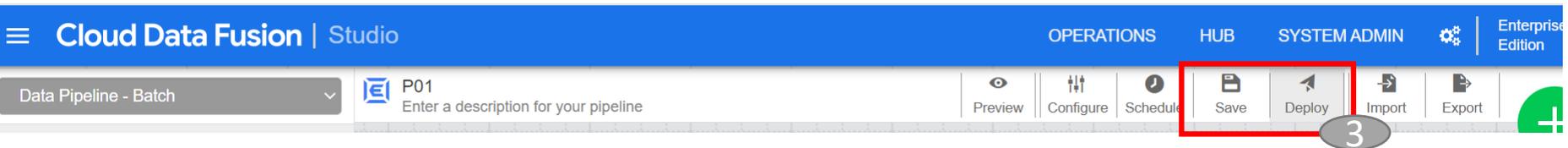
Una vez hecha la configuración no olvide Validar que el objeto no tenga errores.



# 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

The screenshot shows the Cloud Data Fusion Pipeline interface. At the top, there's a navigation bar with 'Cloud Data Fusion | Pipeline' and 'Enterprise Edition'. Below it, there's a toolbar with buttons for 'Configure', 'Schedule', 'Stop', 'Run' (highlighted in red), and 'Summary'. The main area displays a pipeline diagram with nodes: 'GCSFile' (0.20.1) and 'GCSFile14' (0.20.1) on the left; 'Wrangler' (4.7.1) and 'Wrangler41' (4.7.1) in the middle; 'Joiner' (2.9.1) on the right; and 'GCS3' (0.20.1) at the end. The 'Run' button is highlighted in red, indicating the start of the pipeline execution.

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

The screenshot shows the Cloud Data Fusion Pipeline interface after successful execution. The 'Run' button is now 'Run' (green). The 'Status' is 'Succeeded' (highlighted in red). The pipeline diagram shows the following output: 'GCSFile' (0.20.1) and 'GCSFile14' (0.20.1) both have 'Out 190 / Errors 0'; 'Wrangler' (4.7.1) and 'Wrangler41' (4.7.1) both have 'Out 189 / Errors 0'; 'Joiner' (2.9.1) has 'Out 189 / Errors 0'; and 'GCS3' (0.20.1) has 'In 189 / Errors 0'. The 'GCS3' node is highlighted in red, and the text 'Archivo consolidado' is written below it, indicating the location of the output file.



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

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

OBJETOS CONFIGURACIÓN

Depósitos > **bu0125** 1

SUBIR ARCHIVOS SUBIR CARPETA

BORRAR

Filtrar solo por prefijo de nombre ▾

<input type="checkbox"/>	Nombre
<input type="checkbox"/>	2022-10-19-12-20/
<input type="checkbox"/>	Países.txt
<input type="checkbox"/>	PaísesCapitalContinente.csv
<input type="checkbox"/>	PaísesGobiernoOnu.csv
<input type="checkbox"/>	<b>Salida.csv</b>

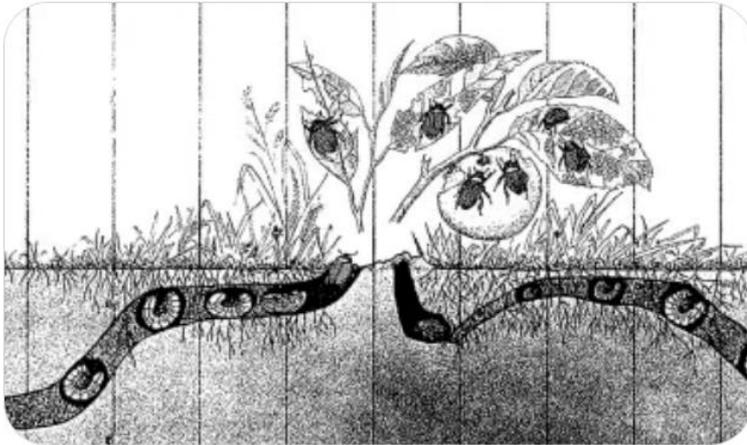
Depósitos > bu0125 > Salida.csv

Fecha y hora de creación	19 oct 2022 17:06:30
Última modificación	19 oct 2022 17:06:30
Clase de almacenamiento	Standard
Tiempo personalizado	—
URL pública	No aplicable
URL autenticada	<a href="https://storage.cloud.google.com/bu0125/Salida.csv">https://storage.cloud.google.com/bu0125/Salida.csv</a>
URI de gsutil	gs://bu0125/Salida.csv
Permisos	
Acceso público	No público
Protección	
Estado de conservación	Ninguno
Historial de versiones	—
Política de retención	Ninguno
Tipo de encriptación	Google-managed key

\*Salida: Bloc de notas 3

Archivo Editar Ver

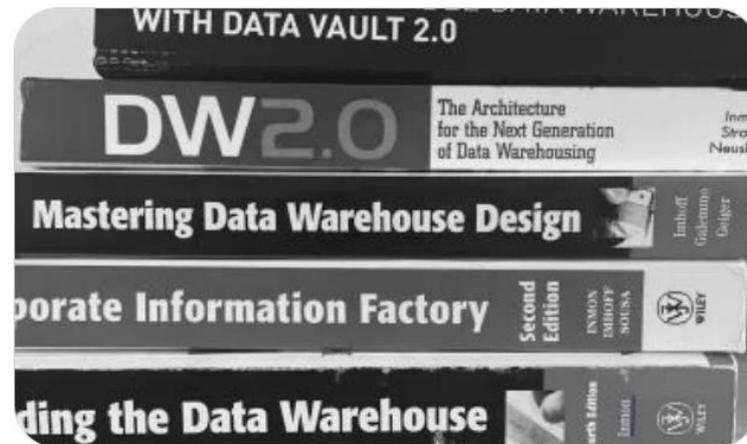
```
Nombre comun,Capital,Continente,Forma de Gobierno,Estatus ONU
Emiratos Arabes Unidos,Abu Dabi,Asia,Monarquia constitucional electiva,Miembro
Nigeria,Abuya,Africa,Republica presidencialista,Miembro
Ghana,Acra,Africa,Republica presidencialista,Miembro
Etiopia,Adis Abeba,Africa,Republica parlamentaria,Miembro
Jordania,Aman,Asia,Monarquia constitucional,Miembro
Andorra,Andorra la Vieja,Europa,Monarquia constitucional,Miembro
Turquia,Ankara,Asia-Europa,Republica parlamentaria,Miembro
Madagascar,Antananarivo,Africa,Republica semipresidencialista,Miembro
Argelia,Argel,Africa,Republica semipresidencialista,Miembro
Turkmenistan,Asjabad,Asia,Republica presidencialista,Miembro
Eritrea,Asmara,Africa,Republica unipartidista,Miembro
Paraguay,Asuncion,America,Republica presidencialista,Miembro
Grecia,Atenas,Europa,Republica parlamentaria,Miembro
Irak,Bagdad,Asia,Republica parlamentaria,Miembro
Azerbaiyan,Baku,Asia-Europa,Republica semipresidencialista,Miembro
Mali / Mali,Bamako,Africa,Republica semipresidencialista,Miembro
Brunei,Bandar Seri Begawan,Asia,Monarquia absoluta,Miembro
Tailandia,Bangkok,Asia,Monarquia constitucional,Miembro
Republica CentroAfricana,Bangui,Africa,Republica semipresidencialista,Miembro
Gambia,Banjul,Africa,Republica presidencialista,Miembro
San Cristobal y Nieves,Basseterre,America,Monarquia constitucional,Miembro
Libano,Beirut,Asia,Republica parlamentaria,Miembro
Serbia,Belgrado,Europa,Republica parlamentaria,Miembro
Belice,Belmopan,America,Monarquia constitucional,Miembro
Alemania,Berlin,Europa,Republica parlamentaria,Miembro
Suiza,Berna,Europa,Republica parlamentaria,Miembro
Kirguistan,Biskek,Asia,Republica parlamentaria,Miembro
Guinea-Bisau,Bisau,Africa,Republica semipresidencialista,Miembro
Colombia,Bogota,America,Republica presidencialista,Miembro
Brasil,Brasilia,America,Republica presidencialista,Miembro
Eslovaquia,Bratislava,Europa,Republica parlamentaria,Miembro
```



## ¿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 integrar en una sola fuente de la verdad nuestros dato



## Presentado por:

- Lenin González S.
- Director General, CEO
- +569 3072 9405
- [lgonzalez@lituus.cl](mailto:lgonzalez@lituus.cl)

