


Ingerir datos con una canalización en Microsoft Fabric

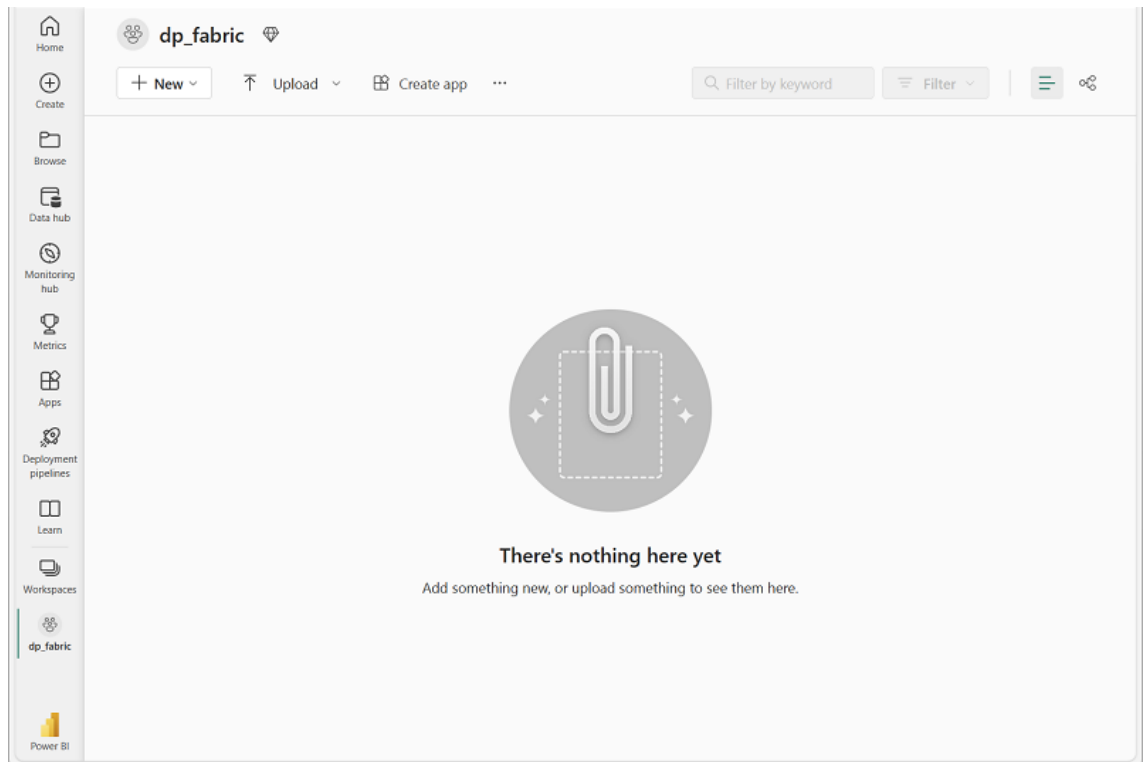
Un data lakehouse es un almacén de datos analíticos común para soluciones de análisis a escala de nube. Una de las tareas principales de un ingeniero de datos es implementar y gestionar la ingesta de datos de múltiples fuentes de datos operativos en la casa del lago. En Microsoft Fabric, puede implementar soluciones *de extracción, transformación y carga* (ETL) o *extracción, carga y transformación* (ELT) para la ingesta de datos mediante la creación de *canalizaciones*.

Fabric también es compatible con Apache Spark, lo que le permite escribir y ejecutar código para procesar datos a escala. Al combinar las capacidades de canalización y Spark en Fabric, puede implementar una lógica de ingesta de datos compleja que copia datos de fuentes externas en el almacenamiento OneLake en el que se basa Lakehouse y luego usa el código Spark para realizar transformaciones de datos personalizadas antes de cargarlos en tablas para análisis.

Crear un espacio de trabajo

Antes de trabajar con datos en Fabric, cree un espacio de trabajo con la versión de prueba de Fabric habilitada.

1. En la [página de inicio de Microsoft Fabric](#), seleccione **Ingeniería de datos de Synapse**.
2. En la barra de menú de la izquierda, seleccione **Espacios de trabajo** (el ícono se parece a .
3. Cree un nuevo espacio de trabajo con el nombre que elija y seleccione un modo de licencia que incluya capacidad de Fabric (*Prueba*, *Premium* o *Fabric*).
4. Cuando se abra su nuevo espacio de trabajo, debería estar vacío.



Crear una casa en el lago

Ahora que tiene un espacio de trabajo, es hora de crear un lago de datos en el que incorporará los datos.

1. En la página de inicio de **Synapse Data Engineering** , cree un nuevo **Lakehouse** con el nombre que elija.

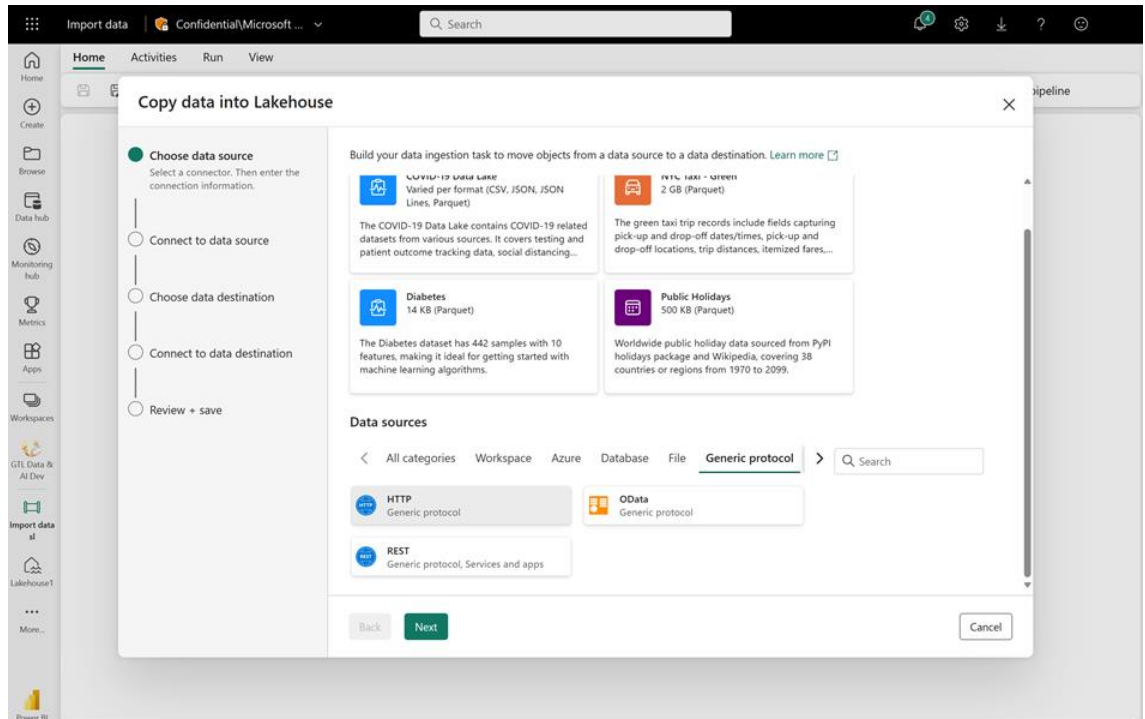
Después de aproximadamente un minuto, se creará una nueva casa en el lago sin **tablas** ni **archivos** .

2. En la pestaña **Vista del lago** en el panel de la izquierda, en el menú ... del nodo **Archivos** , seleccione **Nueva subcarpeta** y cree una subcarpeta llamada **new_data** .

Crear una canalización

Una forma sencilla de ingerir datos es utilizar una actividad **Copiar datos** en una canalización para extraer los datos de una fuente y copiarlos a un archivo en la casa del lago.

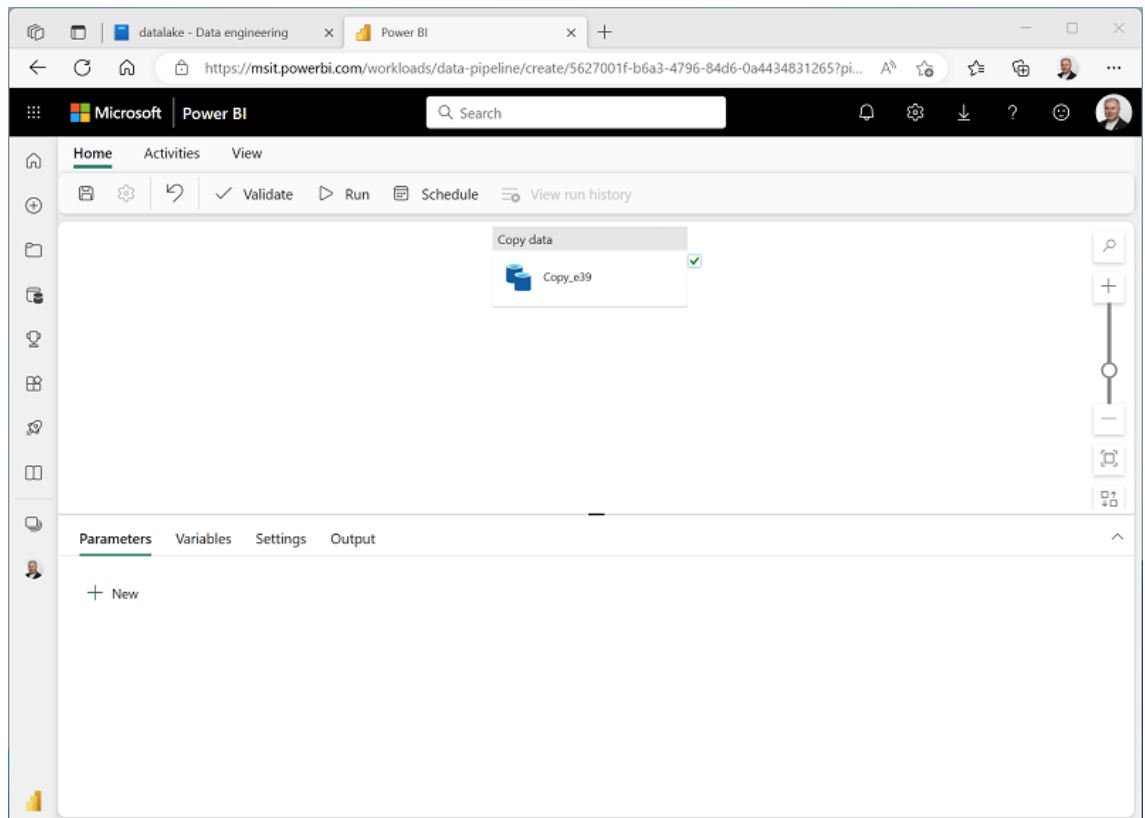
1. En la página **de inicio** de su casa en el lago, seleccione **Obtener datos** y luego seleccione **Nueva canalización de datos** y cree una nueva canalización de datos denominada **Ingestar datos de ventas**.
2. Si el asistente **Copiar datos** no se abre automáticamente, seleccione **Copiar datos** en la página del editor de canalización.
3. En el asistente **Copiar datos**, en la página **Elegir una fuente de datos**, en la sección **de fuentes de datos**, seleccione la pestaña **Protocolo genérico** y luego seleccione **HTTP**.




4. Seleccione **Siguiente** y luego seleccione **Crear nueva conexión** e ingrese las siguientes configuraciones para la conexión a su fuente de datos:
 - **URL** : <https://raw.githubusercontent.com/MicrosoftLearning/dp-data/main/sales.csv>
 - **Conexión** : Crear nueva conexión
 - **Nombre de conexión** : *especifique un nombre único*
 - **Puerta de enlace de datos** : (ninguna)
 - **Tipo de autenticación** : Básica (*Deje el nombre de usuario y la contraseña en blanco*)
5. Seleccione **Siguiente**. Luego asegúrese de que estén seleccionadas las siguientes configuraciones:
 - **URL relativa** : *dejar en blanco*
 - **Método de solicitud** : OBTENER
 - **Encabezados adicionales** : *dejar en blanco*
 - **Copia binaria** : no seleccionada
 - **Solicitar tiempo de espera** : *dejar en blanco*

- **Máximo de conexiones simultáneas** : *dejar en blanco*
- 6. Seleccione **Siguiente** , espere a que se muestren los datos y luego asegúrese de que estén seleccionadas las siguientes configuraciones:
 - **Formato de archivo** : texto delimitado
 - **Delimitador de columna** : coma (,)
 - **Delimitador de fila** : avance de línea (\n)
 - **Primera fila como encabezado** : Seleccionado
 - **Tipo de compresión** : Ninguno
- 7. Seleccione **Vista previa de datos** para ver una muestra de los datos que se ingerirán. Luego cierre la vista previa de datos y seleccione **Siguiente** .
- 8. En la página **Conectarse a destino de datos** , seleccione su casa en el lago existente. Luego seleccione **Siguiente** .
- 9. Configure las siguientes opciones de destino de datos y luego seleccione **Siguiente** :
 - **Carpeta raíz** : Archivos
 - **Nombre de ruta de carpeta** : new_data
 - **Nombre del archivo** : ventas.csv
 - **Comportamiento de copia** : Ninguno
- 10. Configure las siguientes opciones de formato de archivo y luego seleccione **Siguiente** :
 - **Formato de archivo** : texto delimitado
 - **Delimitador de columna** : coma (,)
 - **Delimitador de fila** : avance de línea (\n)
 - **Agregar encabezado al archivo** : Seleccionado
 - **Tipo de compresión** : Ninguno
- 11. En la página **Resumen de copia** , revise los detalles de su operación de copia y luego seleccione **Guardar + Ejecutar** .

Se crea una nueva canalización que contiene una actividad **Copiar datos** , como se muestra aquí:



12. Cuando la canalización comienza a ejecutarse, puede monitorear su estado en el panel **Salida** en el diseñador de canalizaciones. Utilice el icono  (*Actualizar*) para actualizar el estado y espere hasta que se realice correctamente.
13. En la barra de menú de la izquierda, seleccione su casa en el lago.
14. En la página **de inicio** , en el panel **del explorador de Lakehouse** , expanda **Archivos** y seleccione la carpeta **new_data** para verificar que se haya copiado el archivo **sales.csv** .

Crear un cuaderno

1. En la página **de inicio** de su casa en el lago, en el menú **Abrir libreta** , seleccione **Nueva libreta** .

Después de unos segundos, se abrirá una nueva libreta que contiene una sola *celda* . Los cuadernos se componen de una o más celdas que pueden contener *código* o *rebajas* (texto formateado).

2. Seleccione la celda existente en el cuaderno, que contiene un código simple, y luego reemplace el código predeterminado con la siguiente declaración de variable.

```
table_name = "sales"
```

3. En el menú ... de la celda (en la parte superior derecha), seleccione **Alternar celda de parámetro** . Esto configura la celda para que las variables declaradas en ella se traten como parámetros cuando se ejecuta el cuaderno desde una canalización.
4. Debajo de la celda de parámetros, use el botón + **Código** para agregar una nueva celda de código. Luego agregue el siguiente código:

```
from pyspark.sql.functions import *

# Read the new sales data
df =
spark.read.format("csv").option("header","true").load("Files/new_data/
*.csv")

## Add month and year columns
df = df.withColumn("Year", year(col("OrderDate"))).withColumn("Month",
month(col("OrderDate")))

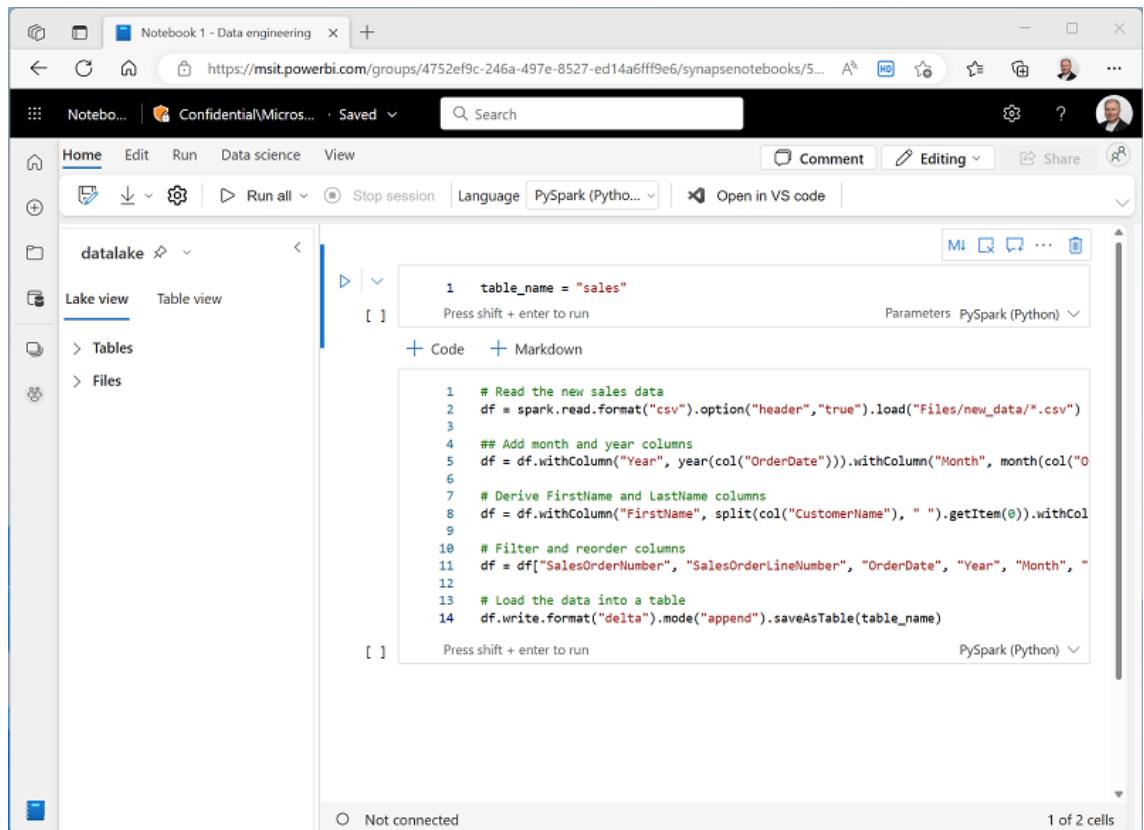
# Derive FirstName and LastName columns
df = df.withColumn("FirstName", split(col("CustomerName"), "
").getItem(0)).withColumn("LastName", split(col("CustomerName"), "
").getItem(1))

# Filter and reorder columns
df = df["SalesOrderNumber", "SalesOrderLineNumber", "OrderDate",
"Year", "Month", "FirstName", "LastName", "EmailAddress", "Item",
"Quantity", "UnitPrice", "TaxAmount"]

# Load the data into a table
df.write.format("delta").mode("append").saveAsTable(table_name)
```

Este código carga los datos del archivo sales.csv que fue ingerido por la actividad **Copiar datos** , aplica cierta lógica de transformación y guarda los datos transformados como una tabla, agregando los datos si la tabla ya existe.

5. Verifique que sus cuadernos se vean similares a este y luego use el botón ► **Ejecutar todo** en la barra de herramientas para ejecutar todas las celdas que contiene.



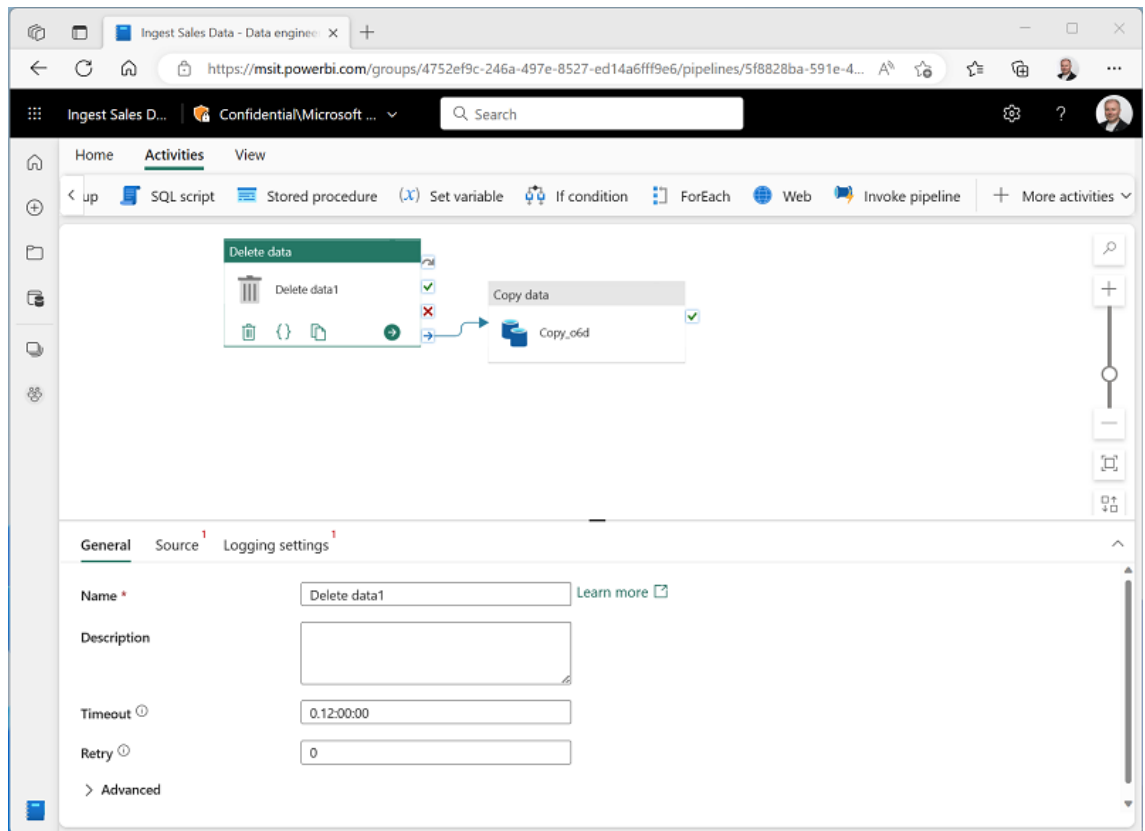
6. Cuando se haya completado la ejecución del cuaderno, en el panel **del explorador de Lakehouse** a la izquierda, en el menú **de Tablas**, seleccione **Actualizar** y verifique que se haya creado una tabla **de ventas**.
7. En la barra de menú del cuaderno, use el ícono **Configuración** para ver la configuración del cuaderno. Luego configure el **Nombre** del cuaderno para **Cargar ventas** y cierre el panel de configuración.
8. En la barra de menú central a la izquierda, seleccione su casa en el lago.
9. En el panel **del Explorador**, actualice la vista. Luego expanda **Tablas** y seleccione la tabla **de ventas** para ver una vista previa de los datos que contiene.

Modificar la canalización

Ahora que ha implementado un cuaderno para transformar datos y cargarlos en una tabla, puede incorporar el cuaderno a una canalización para crear un proceso ETL reutilizable.

1. En la barra de menú central de la izquierda, seleccione el canal **de ingesta de datos de ventas** que creó anteriormente.
2. En la pestaña **Actividades**, en la lista **Más actividades**, seleccione **Eliminar datos**. Luego coloque la nueva actividad **Eliminar**

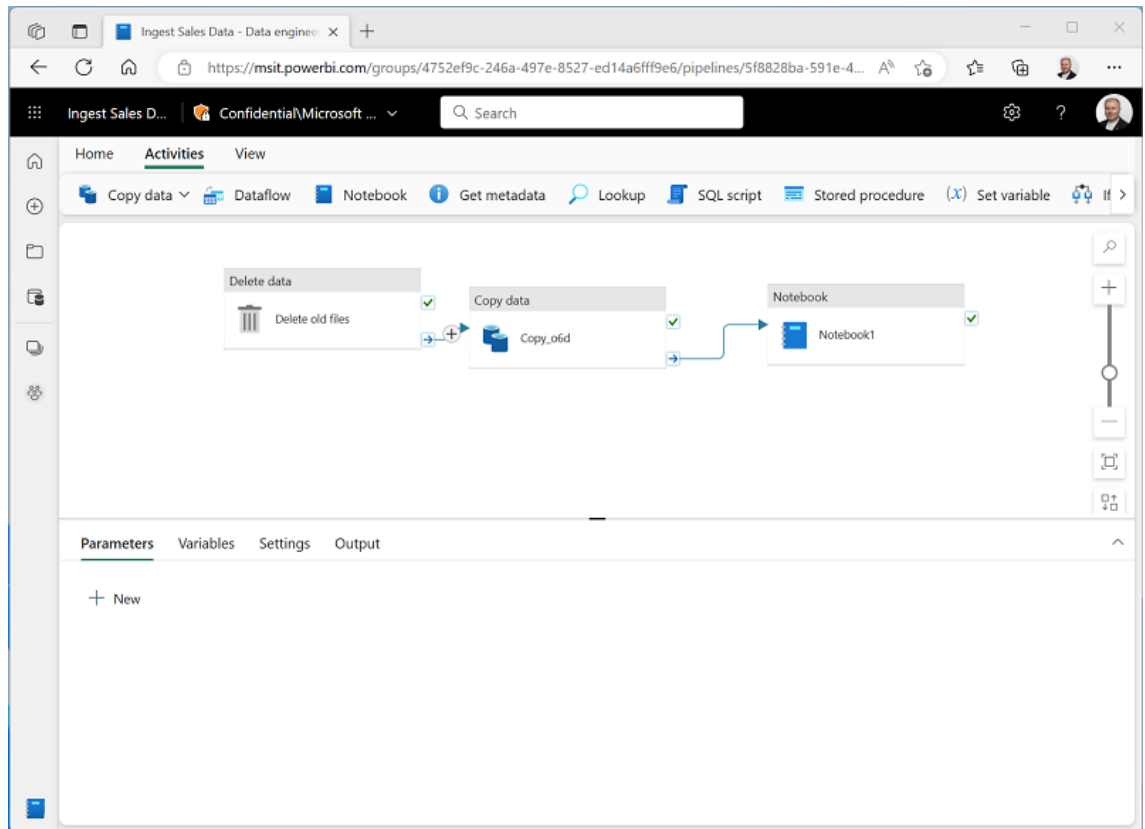
datos a la izquierda de la actividad **Copiar datos** y conecte su salida **Al finalizar a la actividad Copiar datos**, como se muestra aquí:



3. Seleccione la actividad **Eliminar datos** y, en el panel debajo del lienzo de diseño, establezca las siguientes propiedades:
 - **General :**
 - **Nombre :** Eliminar archivos antiguos
 - **Fuente**
 - **Tipo de almacén de datos :** espacio de trabajo
 - **Almacén de datos del espacio de trabajo :** *su casa en el lago*
 - **Tipo de ruta de archivo :** ruta de archivo comodín
 - **Ruta de la carpeta :** Archivos / **new_data**
 - **Nombre de archivo comodín :** *.csv
 - **Rekursivamente :** *Seleccionado*
 - **Configuración de registro :**
 - **Habilitar registro :** *no seleccionado*

Esta configuración garantizará que todos los archivos .csv existentes se eliminen antes de copiar el archivo **sales.csv**.


4. En el diseñador de canalizaciones, en la pestaña **Actividades** , seleccione **Notebook** para agregar una actividad **de Notebook** a la canalización.
5. Seleccione la actividad **Copiar datos** y luego conecte su salida **Al finalizar** a la actividad **Notebook como se muestra aquí:**

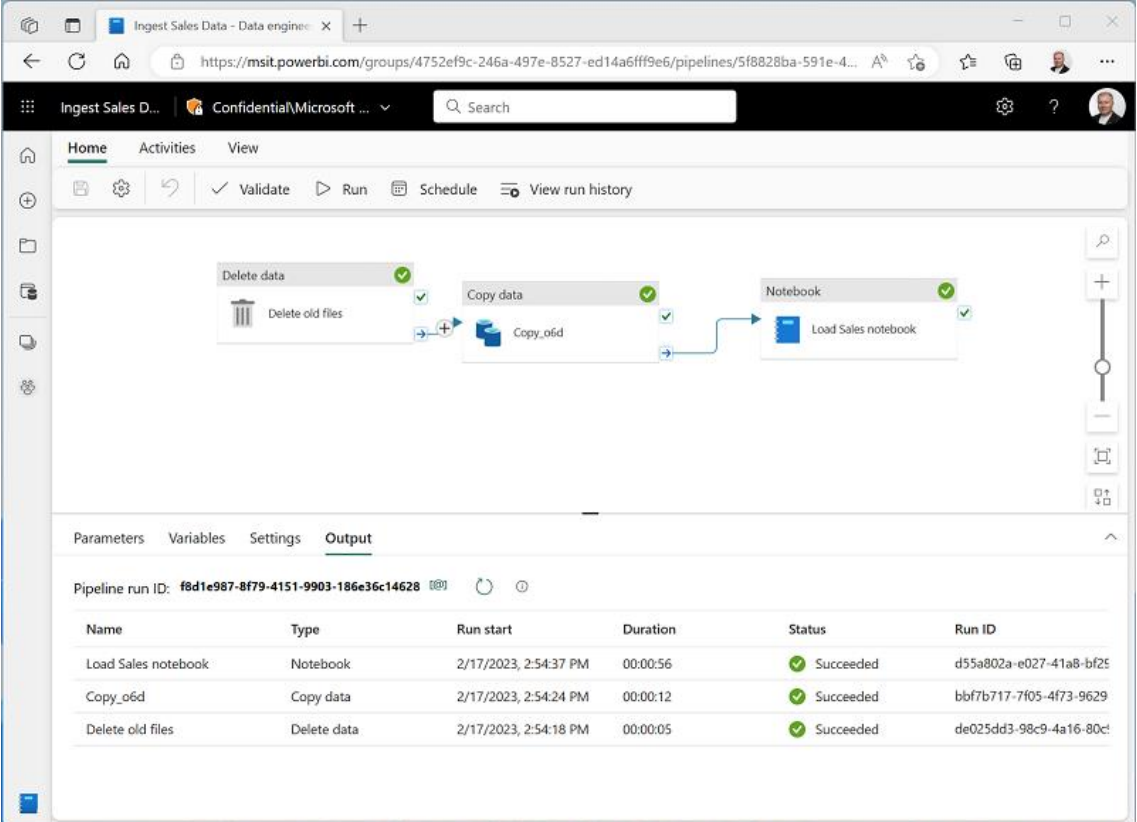


6. Seleccione la actividad **de Notebook** y luego, en el panel debajo del lienzo de diseño, establezca las siguientes propiedades:
 - **General :**
 - **Nombre :** Cargar cuaderno de ventas
 - **Ajustes :**
 - **Cuaderno :** Cargar Ventas
 - **Parámetros base :** *agregue un nuevo parámetro con las siguientes propiedades:*

Nombre	Tipo	
nombre de la tabla	Cadena	:

7. El parámetro **table_name** se pasará al cuaderno y anulará el valor predeterminado asignado a la variable **table_name** en la celda de parámetros.

- En la pestaña **Inicio** , use el ícono  (*Guardar*) para guardar la canalización. Luego use el botón **Ejecutar** para ejecutar la canalización y espere a que se completen todas las actividades.



The screenshot shows the Microsoft Power BI Data Engineer Portal interface. The top navigation bar includes 'Home', 'Activities', and 'View'. The main area displays a pipeline diagram with three activities: 'Delete data' (Delete old files), 'Copy data' (Copy_o6d), and 'Notebook' (Load Sales notebook). Below the diagram, the 'Output' tab is selected, showing a table of pipeline run history.

Name	Type	Run start	Duration	Status	Run ID
Load Sales notebook	Notebook	2/17/2023, 2:54:37 PM	00:00:56	✓ Succeeded	d55a802a-e027-41a8-bf25
Copy_o6d	Copy data	2/17/2023, 2:54:24 PM	00:00:12	✓ Succeeded	bbf7b717-7f05-4f73-9629
Delete old files	Delete data	2/17/2023, 2:54:18 PM	00:00:05	✓ Succeeded	de025dd3-98c9-4a16-80c1

- En la barra de menú central en el borde izquierdo del portal, seleccione su casa del lago.
- En el panel **Explorador** , expanda **Tablas** y seleccione la tabla **new_sales** para ver una vista previa de los datos que contiene. Esta tabla fue creada por el cuaderno cuando la ejecutó la canalización.

En este ejercicio, implementó una solución de ingesta de datos que utiliza una canalización para copiar datos a su casa del lago desde una fuente externa y luego usa un cuaderno Spark para transformar los datos y cargarlos en una tabla.