



# Azure Data Factory by Example

Practical Implementation for  
Data Engineers

—  
Richard Swinbank

Apress®

## Contenido

<b>3. La actividad de copia de datos .....</b>	<b>4</b>
<b>3.1. Preparar una base de datos Azure SQL.....</b>	<b>4</b>
3.1.1. Crear la base de datos .....	4
3.1.2. Crear objetos de base de datos.....	9
<b>3.2. Importar datos estructurados a Azure SQL DB.....</b>	<b>11</b>
3.2.1. Crear el pipeline básico .....	11
3.2.2. Crear el linked service de la base de datos y el dataset.....	11
3.2.3. Crear un conjunto de datos de archivos de texto delimitado.....	17
3.2.4. Crear y ejecutar el pipeline .....	20
3.2.5. Verificación de los resultados .....	23
3.2.6. Procesar Múltiples Archivos.....	25
3.2.7. Truncar antes de cargar .....	27
<b>3.3. Asignar los esquemas de origen y de destino .....</b>	<b>30</b>
3.3.1. Crear un nuevo dataset de origen.....	30
3.3.2. Crear una nueva pipeline .....	31
3.3.3. Configurar la asignación de esquemas.....	33
<b>3.4. Importar datos semiestructurados a Azure SQL DB .....</b>	<b>38</b>
3.4.1. Crear un conjunto de datos de archivos JSON .....	38
3.4.2. Crear el pipeline .....	40
3.4.3. Configurar la asignación de esquemas.....	41
3.4.4. Establezca la referencia de la colección .....	44
3.4.5. El efecto de la desviación del esquema .....	46
3.4.6. Comprender la conversión de tipos .....	48
<b>3.5. Transformar archivos JSON en Parquet.....</b>	<b>49</b>
3.5.1. Crear un nuevo conjunto de datos JSON .....	50
3.5.2. Crear un conjunto de datos Parquet.....	50
3.5.3. Cree y ejecute el pipeline de transformación .....	51
<b>3.6. Configuración de rendimiento .....</b>	<b>55</b>
3.6.1. Unidad de integración de datos .....	55
3.6.2. Grado de Paralelismo de Copia .....	55
<b>Revisión del capítulo .....</b>	<b>57</b>
Conceptos clave .....	57

Experiencia de usuario de Azure Data Factory (ADF UX) ..... 58

Para los desarrolladores de SSIS ..... 59

## 3. La actividad de copia de datos

---

Las tareas de integración de datos pueden dividirse en dos grupos: las de movimiento de datos y las de transformación de datos. En el capítulo 2, creó un pipeline de Azure Data Factory que copiaba datos de un contenedor de almacenamiento de blob a otro, un movimiento de datos simple utilizando la actividad Copiar datos. La actividad de datos Copy es la herramienta principal de Azure Data Factory para mover datos de un lugar a otro, y este capítulo explora su aplicación con mayor detalle.

El movimiento de datos realizado en el capítulo 2 fue no estructurado. Al elegir la opción de **copia binaria (binary copy)** cuando utilizó la herramienta Copiar datos, le dijo explícitamente a la actividad Copiar datos que no tuviera en cuenta la estructura interna de los archivos. No definiste ni intentaste inferir ninguna información sobre las estructuras de datos dentro de los archivos individuales, simplemente los copiaste de un lugar a otro como blobs.

Una de las potentes características de la actividad Copiar datos es su capacidad para inferir y persistir estructuras de datos - esquemas - para los archivos. En combinación con las capacidades de manejo de múltiples archivos, esto proporciona un soporte simple pero poderoso para cargar datos basados en archivos en almacenes de datos estructurados como bases de datos relacionales. En la siguiente sección, añadirás una instancia de Azure SQL Database a tu grupo de recursos Azure para poder explorar esta funcionalidad.

### 3.1. Preparar una base de datos Azure SQL

Una base de datos SQL de Azure (o Azure SQL DB) es uno de los servicios de SQL Server basados en Azure. Proporciona un motor de base de datos de plataforma como servicio (PaaS), totalmente gestionado para que no tenga que preocuparse de los requisitos de administración, como copias de seguridad, parches y actualizaciones. Se basa en el sistema de gestión de bases de datos SQL Server de Microsoft, por lo que puedes seguir interactuando con él utilizando herramientas cliente conocidas como SQL Server Management Studio o Azure Data Studio.

#### 3.1.1. Crear la base de datos

Para crear tu base de datos Azure SQL:

1. En el portal de Azure, crea un nuevo recurso de tipo Azure SQL.
2. Cuando se le pregunte ¿Cómo piensa utilizar el servicio?, elija "Base de datos única" en el menú desplegable Tipo de recurso en el mosaico de bases de datos SQL. Haz clic en el botón Crear debajo del desplegable en el mismo mosaico.

3. Seleccione la suscripción y el grupo de recursos que contiene su instancia de ADF, y luego elija un nombre para su base de datos. Debajo del cuadro desplegable Servidor, haga clic en Crear nuevo.
4. Aparece la hoja de servidor nuevo (Figura 3-1). Debe proporcionar un nombre de servidor único a nivel mundial, un nombre de usuario de administrador del servidor y una contraseña. 5. Localice su servidor en la misma región que su instancia del ADF y haga clic en OK.

The image shows two overlapping windows from the Azure portal. The background window is titled 'Create SQL Database' and is part of the 'New > SQL Database' path. It contains the following fields: 'Subscription' (Free Trial), 'Resource group' (adfbexample-rg), 'Database name' (AdfByExample), 'Server' (Select a server), 'Want to use SQL elastic pool?' (No), and 'Compute + storage' (Please select a server first). The foreground window is titled 'New server' and is a modal for creating a new logical server. It contains the following fields: 'Server name' (adfbexample-sql), 'Server admin login' (adfbexample-admin), 'Password' (masked), 'Confirm password' (masked), and 'Location' ((Europe) UK South). Both windows have 'OK' buttons at the bottom right.

Figura 3-1 Crear una base de datos SQL y nuevas hojas de servidor

El resultado de este proceso no es una instancia tradicional de SQL Server, sino un servidor lógico de SQL Server. El servidor lógico es un contenedor para un grupo de una o más bases de datos Azure SQL y proporciona servicios para todas las bases de datos del grupo, incluyendo la configuración del firewall y la administración de usuarios de AAD.

5. Asegúrese de que la opción ¿Desea utilizar el grupo elástico de SQL? esté establecida en "No" y, a continuación, en Compute + Storage, haga clic en el enlace Configure database.
6. La hoja Configurar para su base de datos controla cómo se asignan los recursos a su base de datos SQL y en qué cantidad. En Compute tier, asegúrese de que el mosaico Serverless está seleccionado, y luego haga clic en Apply. El nivel de cómputo sin servidor no reserva recursos de cómputo por adelantado y se factura sólo en función del uso: es una opción de bajo coste para el aprendizaje y el desarrollo.

Microsoft Azure Actualización Buscar recursos, servicios y documentos (0+)

Inicio > Crear un recurso >

## Crear base de datos SQL

Microsoft

Nombre de la base de datos \*

sqlfradbyexample

Servidor \*

(nueva) sqlfradbyexample2022 (Brazil South)

Crear nuevo

¿Quieres usar un grupo elástico de SQLT? ☐ SI ☒ NO

Proceso y almacenamiento \*

Use general  
Sin servidor, Gen5, 1 vCore, Almacenamiento: 32 GB  
Configurar base de datos

Redundancia del almacenamiento de copias de seguridad

Elige el modo de replicación de las copias de seguridad de PITR y LTR. La restauración geográfica o la posibilidad de recuperación tras una interrupción regional solo están disponibles si se ha seleccionado el almacenamiento con redundancia geográfica.

Redundancia de almacenamiento de copia de seguridad ☒ Almacenamiento de copias de seguridad con redundancia local  
☐ Almacenamiento de copias de seguridad con redundancia de zona  
☐ Almacenamiento de copias de seguridad con redundancia regional

Revisar y crear Siguiendo: Redes >

Microsoft Azure Actualización Buscar recursos, servicios y documentos (0+)

Inicio > Crear un recurso > Crear base de datos SQL >

## Configurar

Comentarios

Nivel de servicio y proceso

Seleccione entre los niveles disponibles en función de las necesidades de la carga de trabajo. El modelo de núcleo virtual proporciona una amplia gama de controles de configuración y ofrece Hiperescape y Sin servidor para escalar automáticamente la base de datos en función de las necesidades de la carga de trabajo. Como alternativa, el modelo de DTU proporciona paquetes de precio y rendimiento establecidos entre los que elegir para facilitar la configuración. [Más información](#)

Nivel de servicio

Use general (Opciones de proceso y almacenamiento escalables)  
Comparar niveles de servicio

Nivel de proceso

☐ Aproximado - Los recursos de proceso están preasignados. Facturación por hora según los núcleos virtuales configurados.

☒ Sin servidor - Los recursos de proceso se escalan automáticamente. Facturación por segundo según los núcleos virtuales usados.

Hardware de proceso

Seleccione la configuración de hardware en función de los requisitos de la carga de trabajo. La disponibilidad del hardware de proceso optimizado, optimizado para memoria y computación confidencial depende de la región, el nivel de servicio y el nivel de proceso.

Resumen de costo

Gen5 - Use general (GP, L, Gen5, L)	
Costo por vCore (en USD)	0.22
Más storage (incluido en GB)	x 41.6
COSTO DE ALMACENAMIENTO ESTIMADO POR MES: 5.09 USD	
COSTO DE PROCESO POR NÚCLEO VIRTUAL POR SEGUNDO: 0.000275 USD	

Aplicar

Microsoft Azure Actualización Buscar recursos, servicios y documentos (0+)

Inicio > Crear un recurso > Crear base de datos SQL >

## Configurar

Comentarios

Número mínimo de núcleos virtuales

0.1 Núcleos virtuales

2.02 GB MEMORIA MÍNIMA 3 GB MEMORIA MÁXIMA

Retraso de pausa automática

La base de datos se pausa automáticamente si está inactiva durante el período de tiempo especificado aquí y se reanuda automáticamente cuando vuelve la actividad de la base de datos. Como alternativa, se puede desactivar la pausa automática.

☒ Habilitar pausa automática

Días: 2 Horas: 1 Minutos: 0

Tamaño máximo de datos (GB)

12

9.6 GB ESPACIO DE REGISTRO ASIGNADO

Aplicar

**Nota** Por defecto, las bases de datos en el nivel de computación sin servidor se ponen en pausa después de una hora de inactividad. Esto mantiene los costes de funcionamiento bajos, pero significa que, cuando vuelvas a la base de datos después de un período inactivo, es posible que tengas que esperar unos minutos antes de poder conectarte con éxito.

7. Cuando haya especificado los detalles del servidor y de la base de datos, la hoja Crear base de datos SQL debería tener un aspecto similar al de la Figura 3-2. Haga clic en Revisar + crear, y después de la validación haga clic en Crear. (Estoy omitiendo a propósito las tres pestañas restantes -Red, Configuración adicional y Etiquetas- y aceptando sus valores por defecto).

The screenshot shows the 'Create SQL Database' page in the Azure portal. The page has a breadcrumb trail: Home > New > SQL Database >. The title is 'Create SQL Database' with the Microsoft logo. Below the title are tabs: Basics, Networking, Additional settings, Tags, and Review + create. The 'Basics' tab is active. The page content includes a description: 'Create a SQL database with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. Learn more'. There are two main sections: 'Project details' and 'Database details'. In 'Project details', 'Subscription' is set to 'Visual Studio Enterprise Subscription' and 'Resource group' is 'adfbexample-rg'. In 'Database details', 'Database name' is 'AdfByExample', 'Server' is '(new) adfbexample-sql (UK South)', and 'Want to use SQL elastic pool?' is set to 'No'. Under 'Compute + storage', the 'General Purpose' option is selected, showing 'Serverless, Gen5, 1 vCore, 32 GB storage'. At the bottom, there are two buttons: 'Review + create' and 'Next : Networking >'. The 'Review + create' button is highlighted in blue.

Figura 3-2 Hoja de creación de base de datos SQL completada

8. El proceso de despliegue crea tres nuevos recursos en su grupo de recursos: un SQL Server lógico, una cuenta de almacenamiento dedicada para los registros de la base de datos y su nueva base de datos Azure SQL Server. Se muestra un mensaje de notificación cuando se completa el despliegue, incluyendo un botón de ir al recurso. Haga clic en el botón para abrir la hoja de base de datos SQL del portal (Figura 3-3).



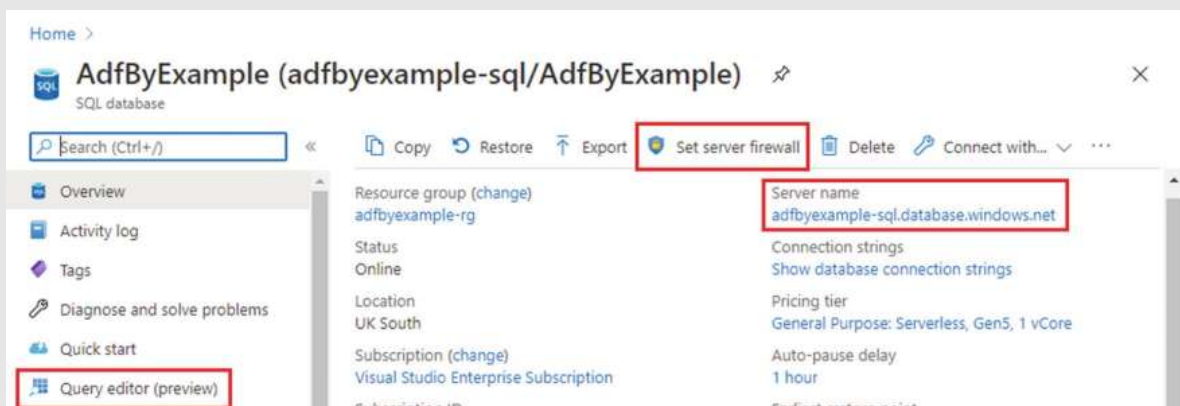
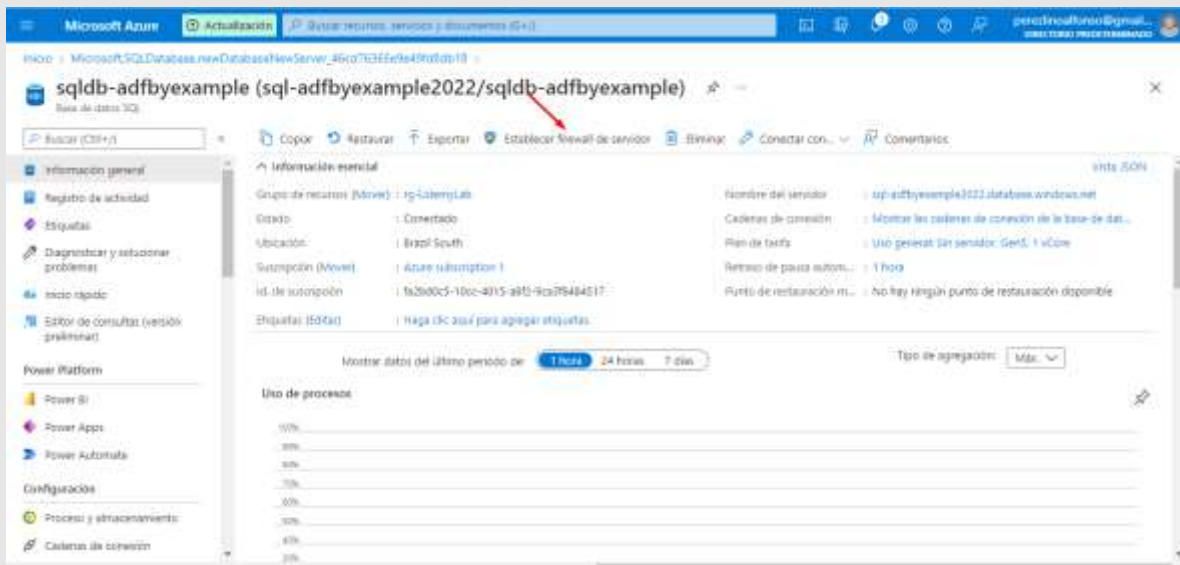
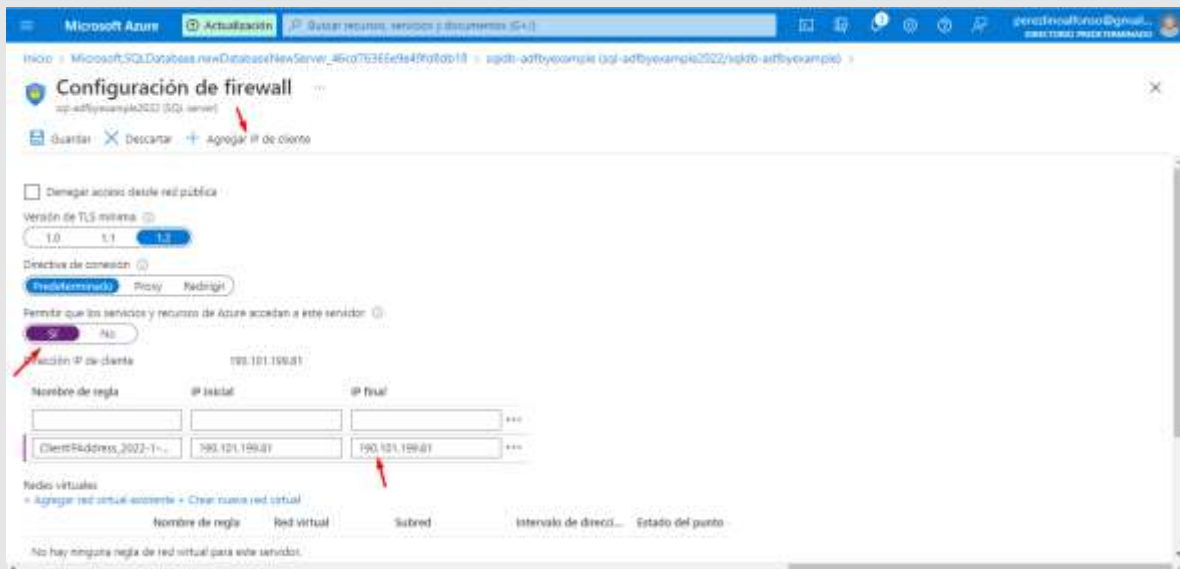


Figura 3-3 Hoja de la base de datos SQL del portal de Azure, indicando las características importantes

9. Su base de datos ya está creada, pero no podrá acceder a ella hasta que haya configurado una regla de firewall del servidor que le permita entrar. En la parte superior de la hoja de la base de datos SQL, haga clic en el botón Set server firewall (indicado en la Figura 3-3).
10. En la parte superior de la hoja de configuración del Firewall, haga clic en + Agregar IP de cliente. Más abajo en la hoja, establezca el conmutador Permitir que los servicios y recursos de Azure accedan a este servidor en "Sí" - esto permitirá que su instancia de ADF acceda al servidor. En la parte superior de la hoja, haga clic en Guardar.





### 3.1.2. Crear objetos de base de datos

La conexión a su instancia de Azure SQL DB es posible utilizando una serie de herramientas cliente, por ejemplo

- SQL Server Management Studio (SSMS) instalado en su ordenador
- Azure Data Studio (ADS) instalado en su ordenador
- El editor de consultas en línea de SQL DB

Para utilizar SSMS o ADS, tendrá que conectarse al servidor, utilizando su nombre de dominio completo, con el nombre de usuario y la contraseña que configuró en la sección anterior. La ubicación del nombre del servidor en la hoja de la base de datos SQL se indica en la Figura 3-3 - en este ejemplo, es **adfbvexample-sql.database.windows.net**. Para utilizar el editor de consultas en línea, haga clic en Editor de consultas (vista previa) (también indicado en la Figura 3-3) e inicie sesión con las mismas credenciales.

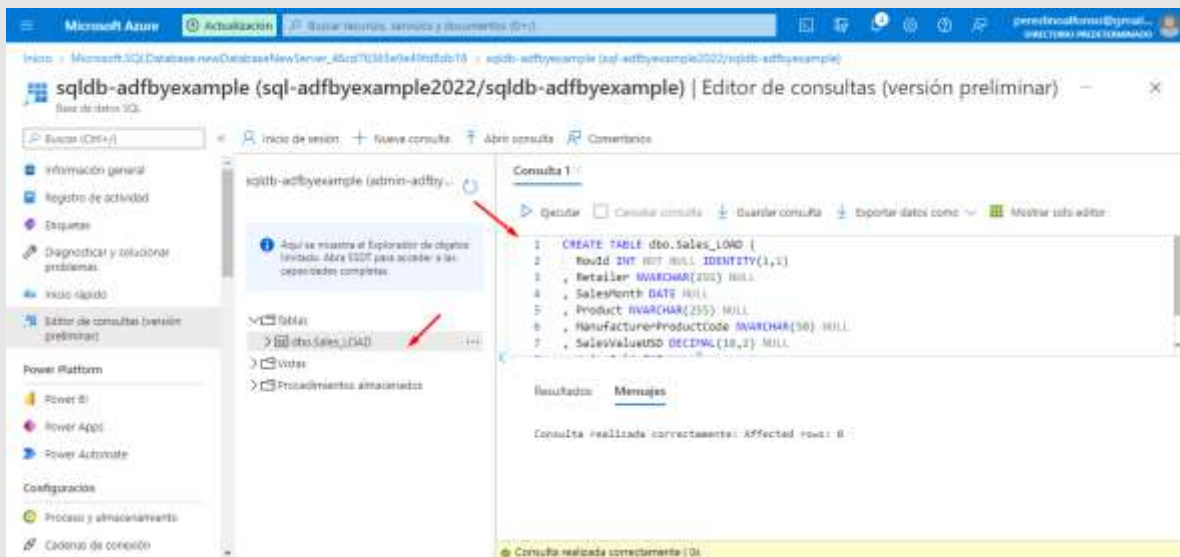
Una vez que se haya conectado con éxito, podrá crear objetos de base de datos. El listado 3-1 proporciona el código SQL para crear una nueva tabla en tu base de datos - los ejemplos de código fuente utilizados en este libro están disponibles en el repositorio GitHub del libro, ubicado en <https://github.com/Apress/azure-data-factory-by-example>. Copie y pegue la sentencia SQL en la herramienta cliente de su elección y ejecute la consulta.

```

CREATE TABLE dbo.Sales_LOAD (
    RowId INT NOT NULL IDENTITY(1,1)
,   Retailer NVARCHAR(255) NULL
,   SalesMonth DATE NULL
,   Product NVARCHAR(255) NULL
,   ManufacturerProductCode NVARCHAR(50) NULL
,   SalesValueUSD DECIMAL(18,2) NULL
,   UnitsSold INT NULL
,   CONSTRAINT PK__dbo_Sales_LOAD PRIMARY KEY (RowId)
);

```

Listado 3-1 Script de creación de tabla para dbo.Sales\_LOAD



La Figura 3-4 muestra el editor de consultas en línea después de ejecutar la consulta, con el nodo Tables de la barra lateral del explorador de objetos expandido para mostrar el resultado.

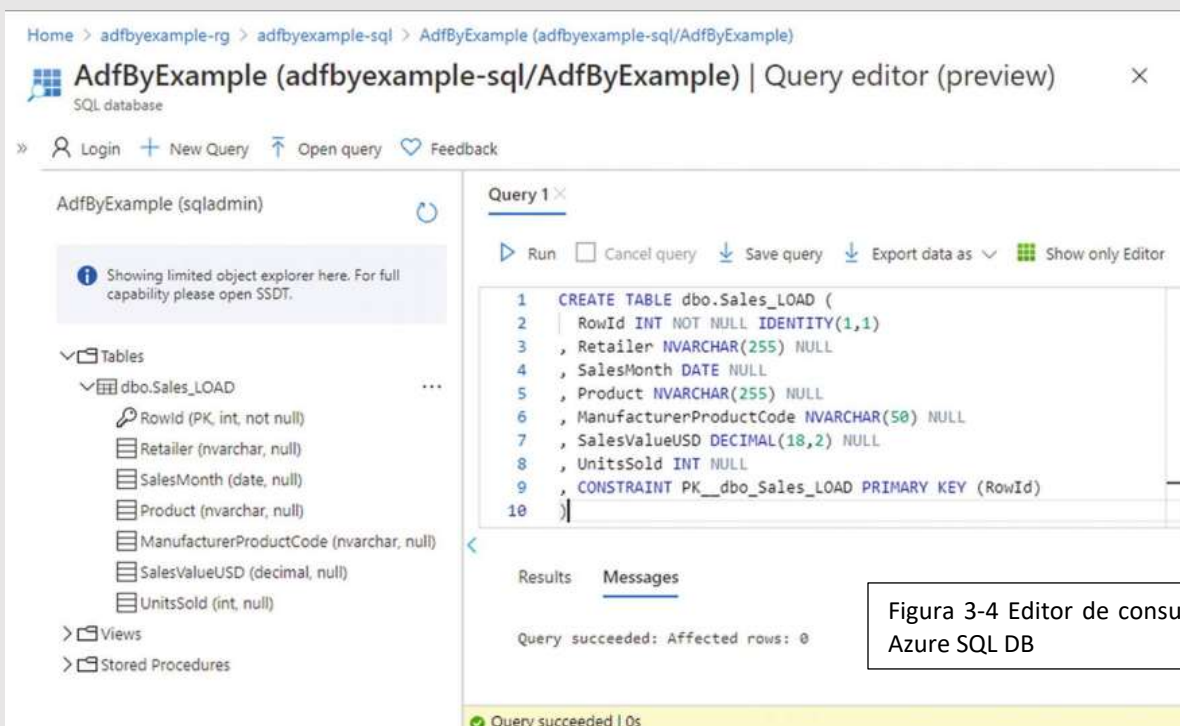


Figura 3-4 Editor de consultas en línea de Azure SQL DB

## 3.2. Importar datos estructurados a Azure SQL DB

Ya ha creado todos los recursos que necesita para empezar a utilizar los pipelines de ADF para mover los datos entre los archivos del almacenamiento blob y las tablas de la base de datos SQL. En esta sección, aprenderá a crear pipelines que copien datos de archivos de datos estructurados. Los formatos de datos estructurados son esencialmente tabulares: un archivo contiene un número de filas, cada una de las cuales contiene el mismo número de campos. Los archivos de valores separados por comas (CSV) son un formato de datos estructurados común.

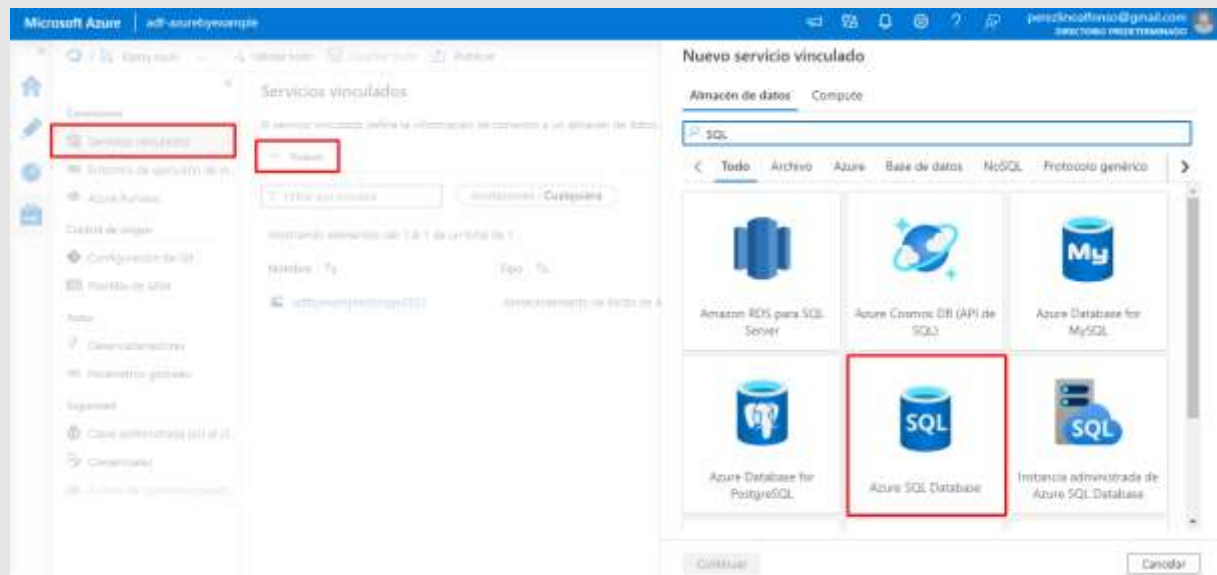
### 3.2.1. Crear el pipeline básico

Ahora creará un pipeline para copiar los datos del archivo CSV en su base de datos Azure SQL. Recuerde que en el Capítulo 2 se requiere un dataset y un linked service en cada extremo del movimiento de datos y una actividad Copy data para conectar ambos.

### 3.2.2. Crear el linked service de la base de datos y el dataset

Comience por crear un linked service para su nueva Azure SQL DB, como sigue:

1. Abra la UX del ADF y navegue hasta el hub de gestión. En Conexiones en la barra lateral del hub, seleccione Servicios vinculados.
2. Haga clic en el botón + Nuevo en la parte superior de la página de servicios vinculados y seleccione Azure SQL Database. Haga clic en Continuar.
3. Complete la hoja Nuevo servicio vinculado (Azure SQL Database) (Figura 3-5). Utilice el método de selección de cuentas "Desde la suscripción de Azure" y, a continuación, seleccione la suscripción, el nombre del servidor y el nombre de la base de datos. Elija el tipo de autenticación "SQL authentication" y proporcione el nombre de usuario y la contraseña de su servidor de base de datos.



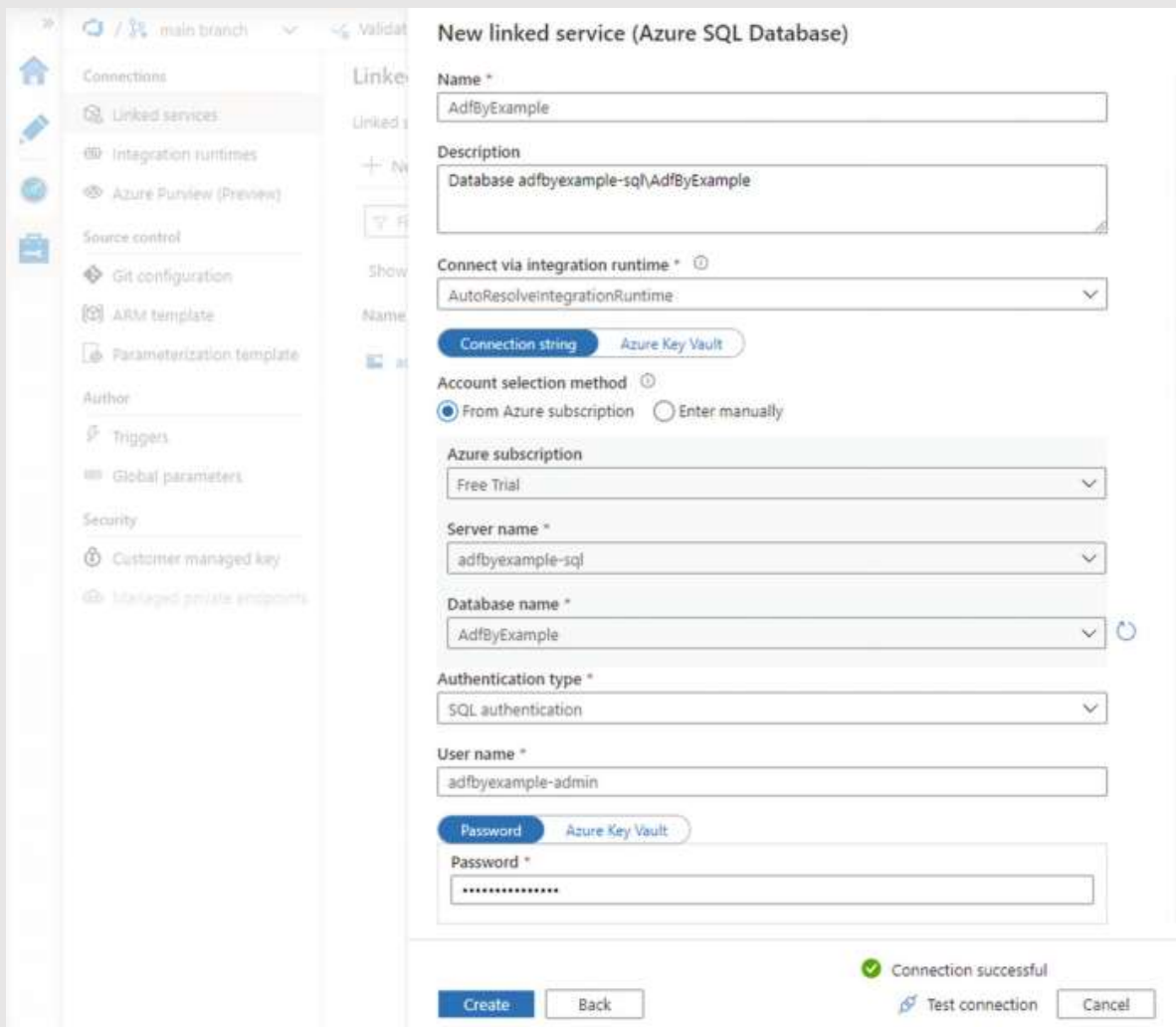
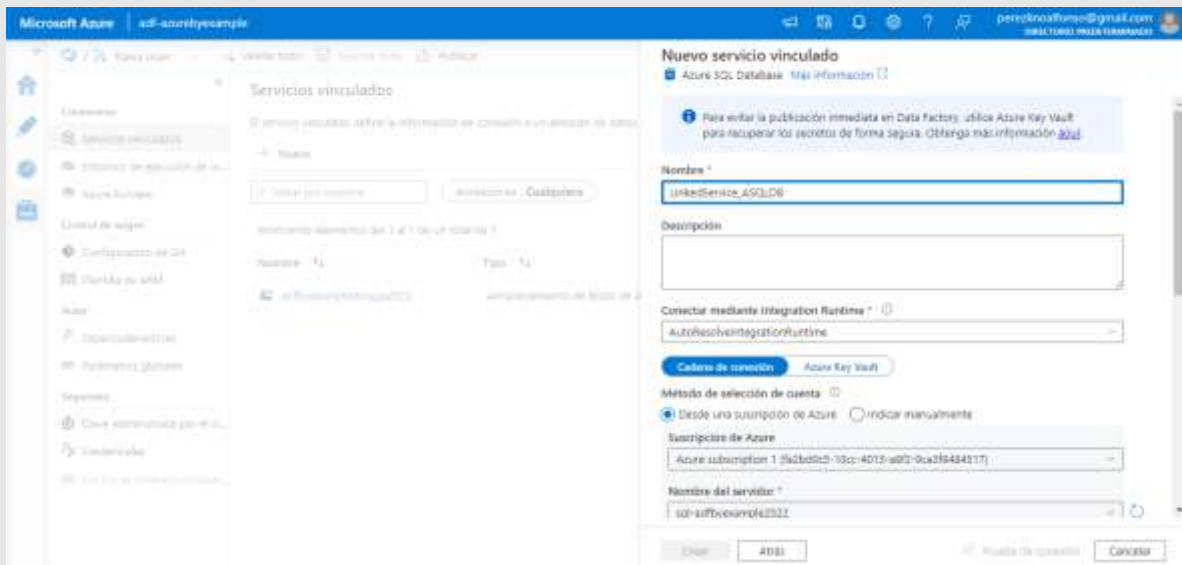
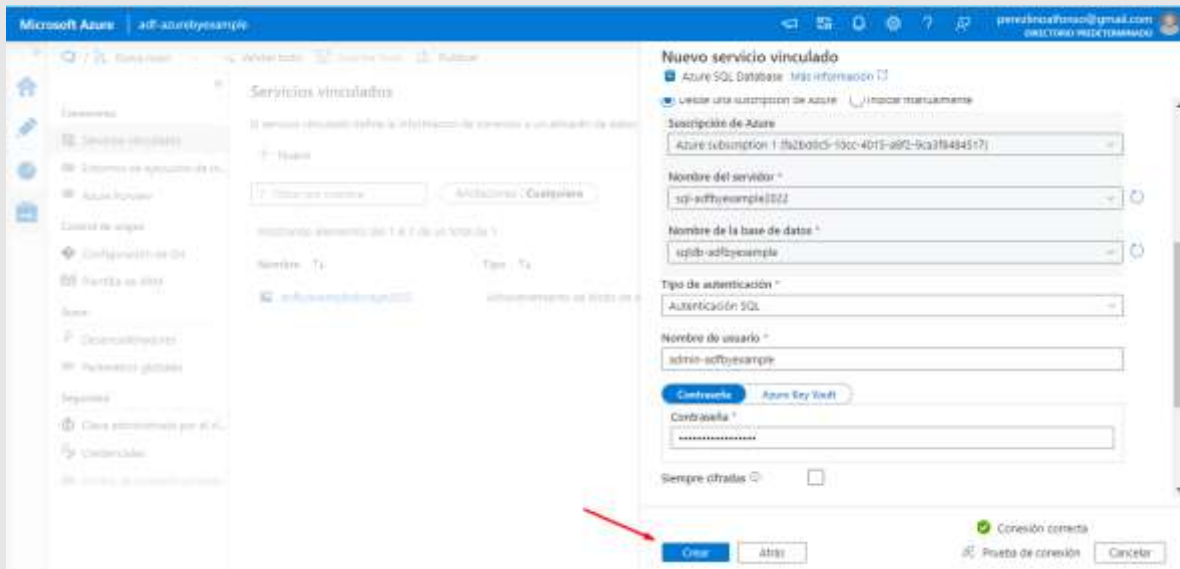


Figura 3-5 Diálogo de nuevo servicio vinculado (Azure SQL Database)



4. Utilice el botón Probar conexión para verificar los valores introducidos y, a continuación, haga clic en Crear. Al igual que cuando creó su cuenta de almacenamiento, el servicio vinculado a la base de datos SQL se publicará inmediatamente, para permitir que el ADF almacene la contraseña del servidor de forma segura.

Ahora cree un dataset para representar la tabla dbo.Sales\_LOAD que creó en la sección anterior:

1. Navegue al espacio de trabajo de creación. El explorador de Recursos de Fábrica enumera los recursos definidos en su sesión de ADF UX - son las definiciones de recursos cargadas desde su rama de colaboración en Git. La cabecera de cada tipo de recurso incluye un recuento del número de recursos (por ejemplo, 1 pipeline, 2 datasets). Pasa el ratón por encima del recuento de datasets para que aparezca un botón de Acciones con elipsis (indicado en la Figura 3-6).

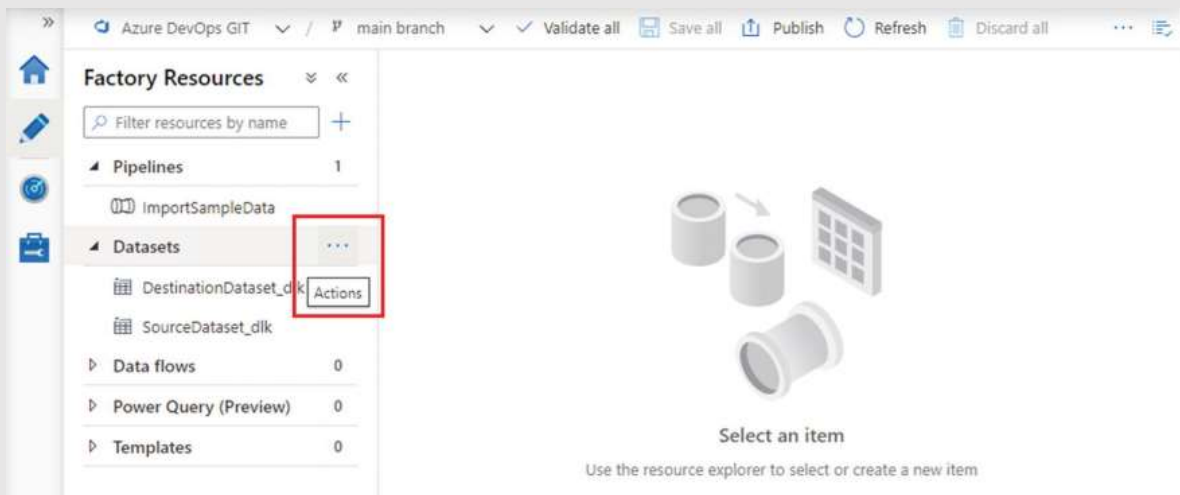
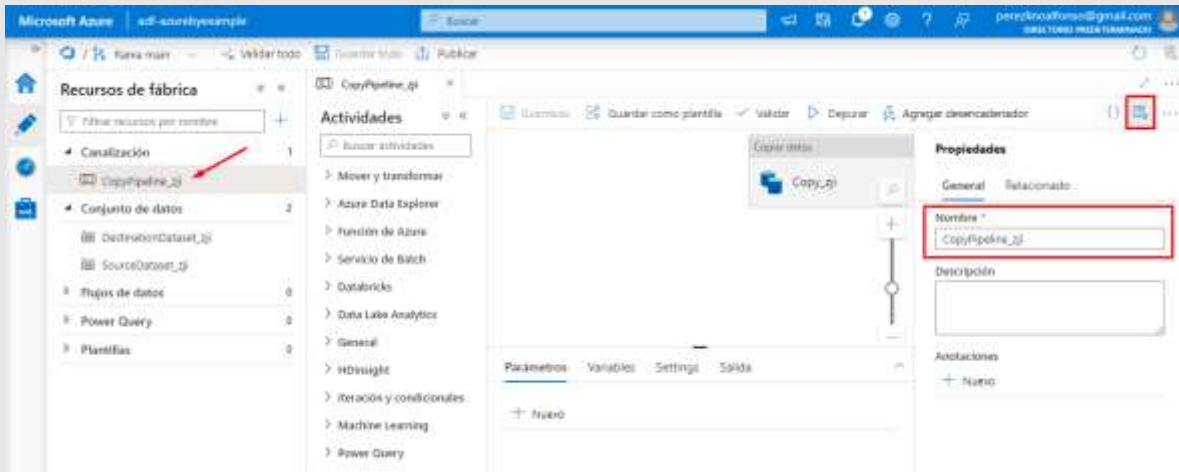
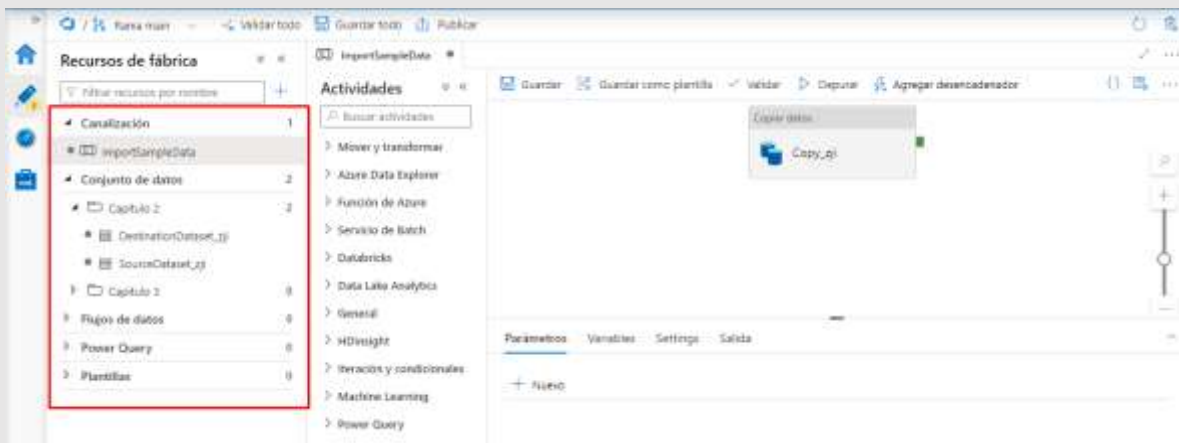


Figura 3-6 Botón de acciones de conjuntos de datos en el explorador de recursos de la fábrica

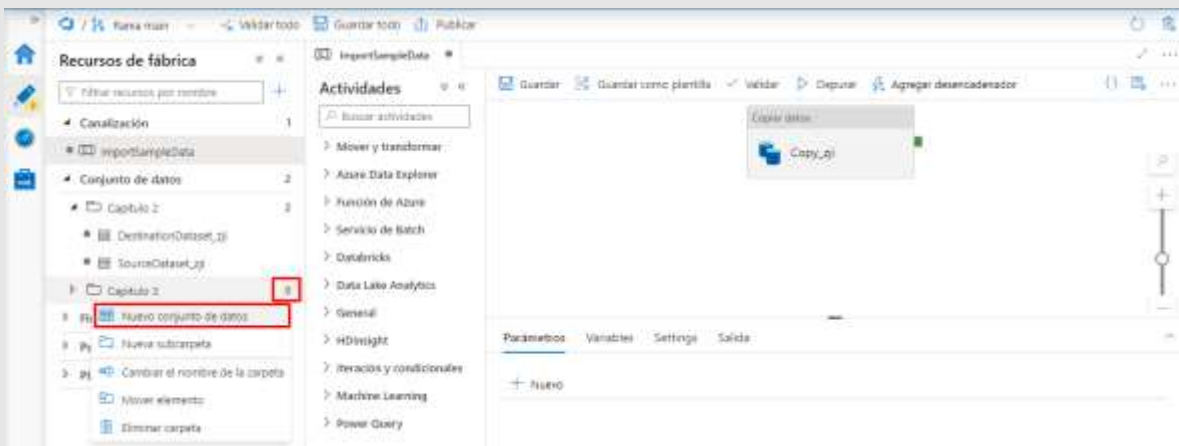


Aquí cambiamos el nombre del Pipeline

- Haga clic en el botón Acciones para acceder al menú Acciones. Contiene dos elementos: Nuevo conjunto de datos y Nueva carpeta. Las carpetas ayudan a organizar los recursos dentro de la fábrica y pueden anidarse. Cree una nueva carpeta y nómbrela "Capítulo3". (También puede crear una carpeta "Capítulo2" para contener los conjuntos de datos creados por la herramienta Copiar datos; si lo hace, puede simplemente arrastrar los conjuntos de datos existentes y soltarlos en esa carpeta).

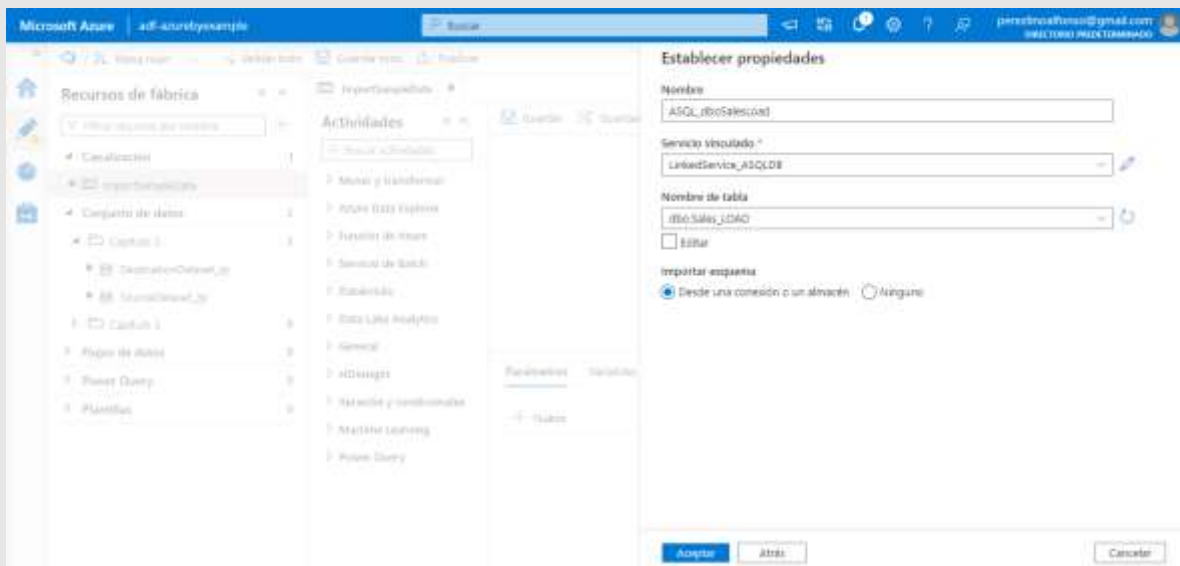


- La carpeta "Capítulo3" también muestra un recuento de recursos (actualmente cero) y el botón del menú Acciones. Elija Nuevo conjunto de datos en el menú Acciones y, a continuación, seleccione Azure SQL Database. Haga clic en Continuar.



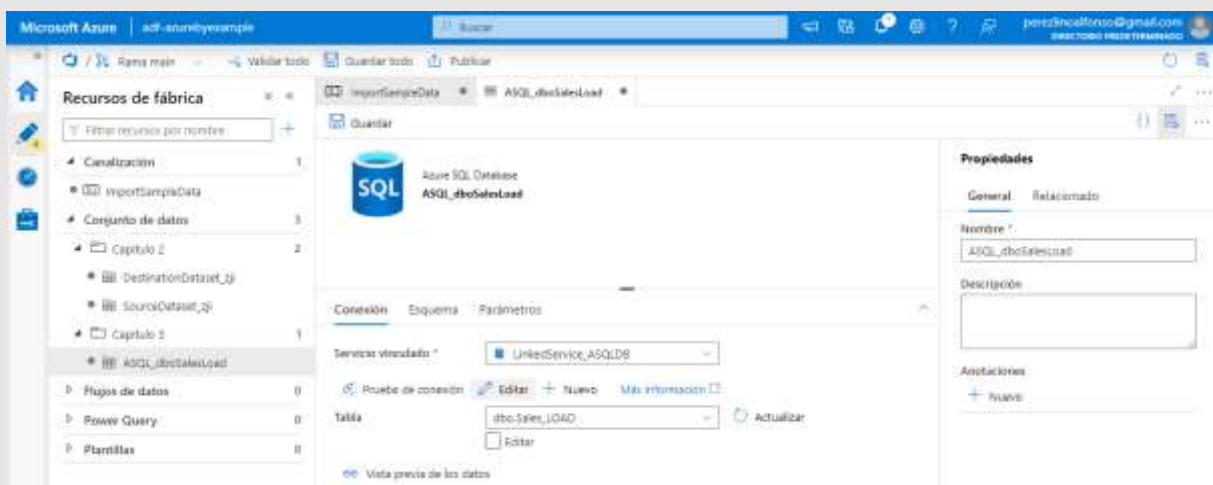


4. Nombre el conjunto de datos "ASQL\_dboSalesLoad" y seleccione el nuevo servicio vinculado a la base de datos SQL en el menú desplegable. Aparecerá un menú desplegable con el nombre de la tabla: seleccione la tabla dbo.Sales\_LOAD y haga clic en Aceptar.



**Sugerencia** Es posible que desee adoptar una convención de nomenclatura para los recursos de la fábrica. Mi preferencia es evitar prefijos que indiquen el tipo de un recurso, porque los tipos de recursos suelen estar claros en sus contextos. Sin embargo, considero que los prefijos que indican el tipo de almacenamiento de un conjunto de datos son útiles (por ejemplo, "ASQL\_" para los conjuntos de datos de Azure SQL DB).

5. El nuevo conjunto de datos se abre en una nueva pestaña en el lienzo de creación (Figura 3-7). En el lado derecho, verá la hoja de propiedades con pestañas, que muestra el nombre y la descripción del conjunto de datos. La pestaña de la hoja de propiedades le permite ver qué conductos se refieren a un dataset, una característica que también está disponible para otros recursos de la fábrica. Puede alternar la visibilidad de la hoja utilizando el botón de Propiedades (icono deslizando) que se encuentra inmediatamente encima de ella; utilice el botón para cerrar la hoja.





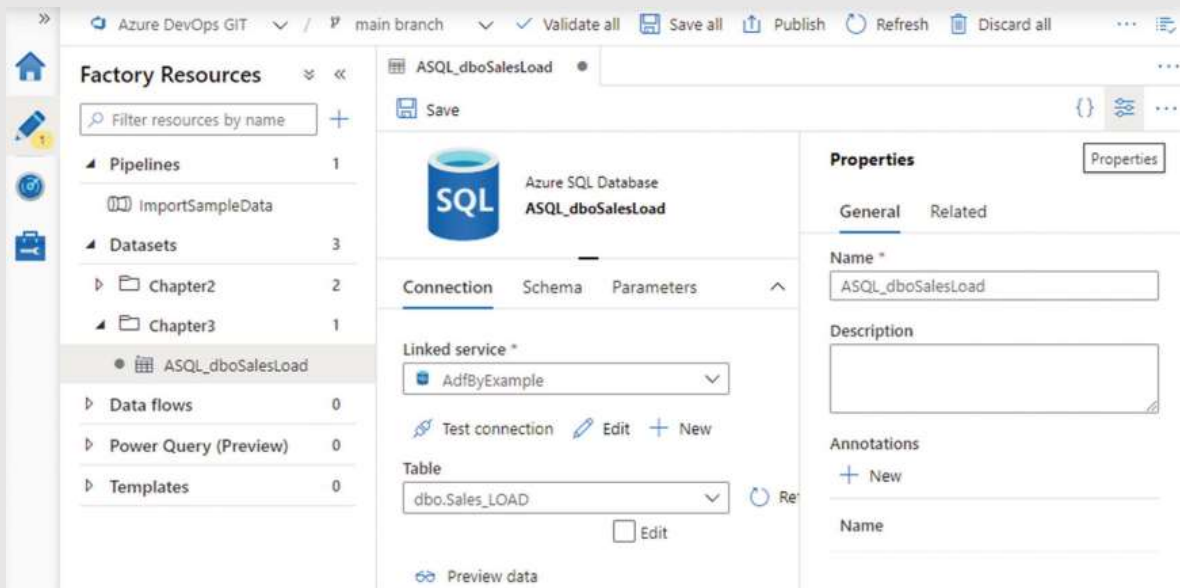
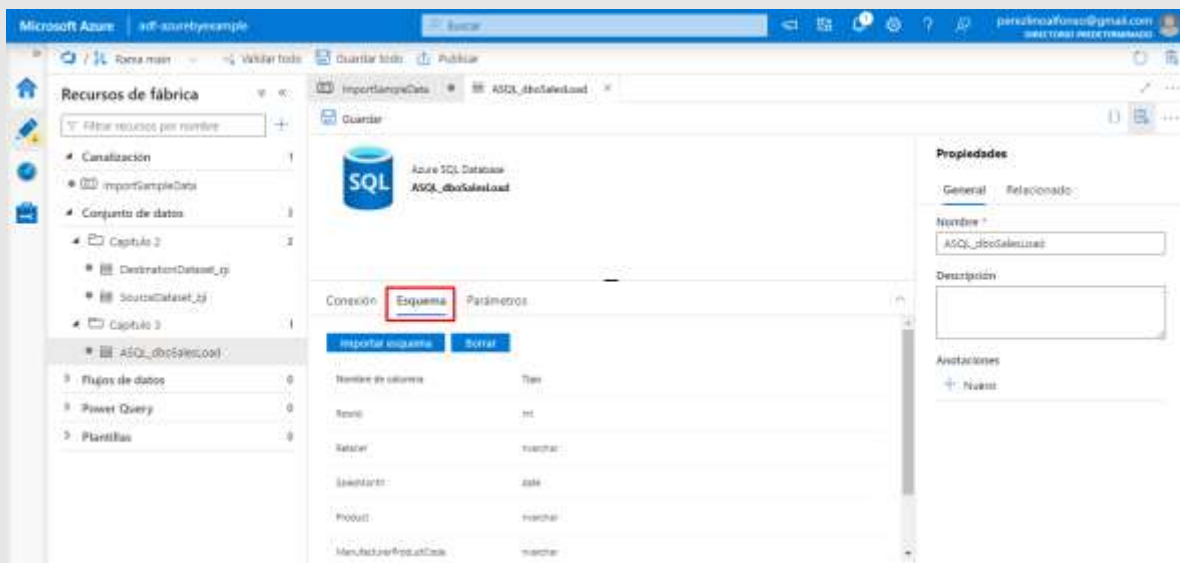


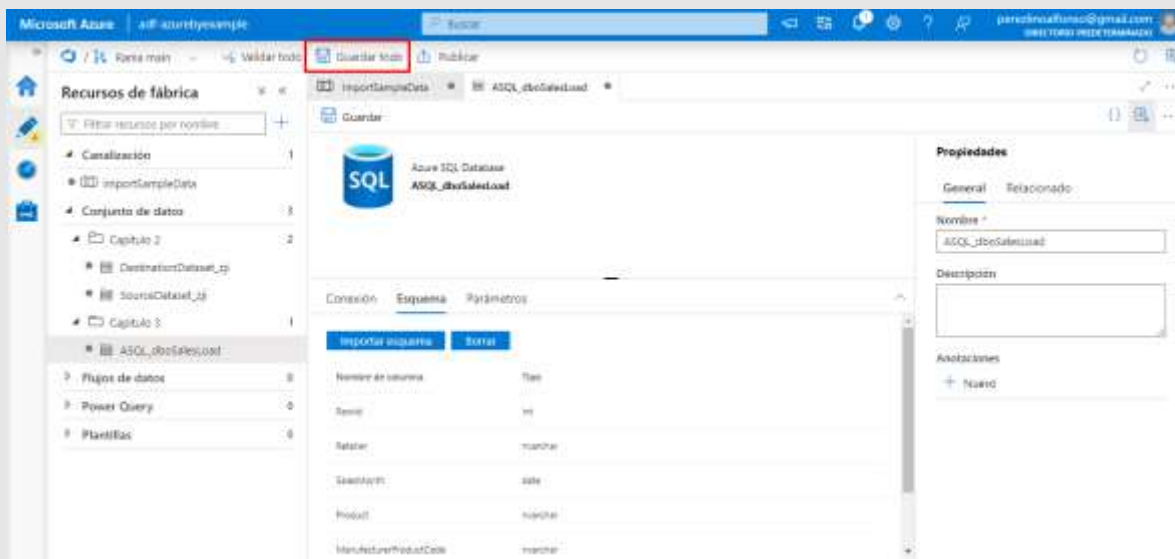
Figura 3-7 Edición de un conjunto de datos en el espacio de trabajo de creación

6. A la izquierda del botón de Propiedades, el botón de Código (llaves o icono {}) le permite inspeccionar -y si lo desea editar- la definición de un recurso de fábrica como JSON. Haga clic en el botón para ver el código, y luego haga clic en Cancelar para cerrar la hoja.
7. El panel de configuración con pestañas situado debajo del lienzo de creación contiene información de configuración para el recurso seleccionado, en este caso, el nuevo conjunto de datos. Las pestañas presentes en el panel de configuración varían según los tipos de recursos. Seleccione la pestaña **Esquema** para inspeccionar el esquema de la tabla `dbo.Sales_LOAD`.



- La UX del ADF indica los cambios no guardados de varias maneras. Aparece un punto gris a la izquierda del nombre del conjunto de datos en el explorador de Recursos de fábrica y a la derecha de su nombre en la pestaña del lienzo de creación. En la barra lateral de navegación, el número "1" en un círculo amarillo debajo del icono del lápiz indica que su sesión contiene un total de cambios sin guardar. Haga clic en Guardar en el panel de pestañas del conjunto de datos - esto confirmará su cambio en el conjunto de datos y lo empujará a su repositorio Git - o haga clic en Guardar todo en la barra de cabecera de la fábrica para confirmar y empujar todos los cambios. Asegúrese de guardar sus cambios regularmente.

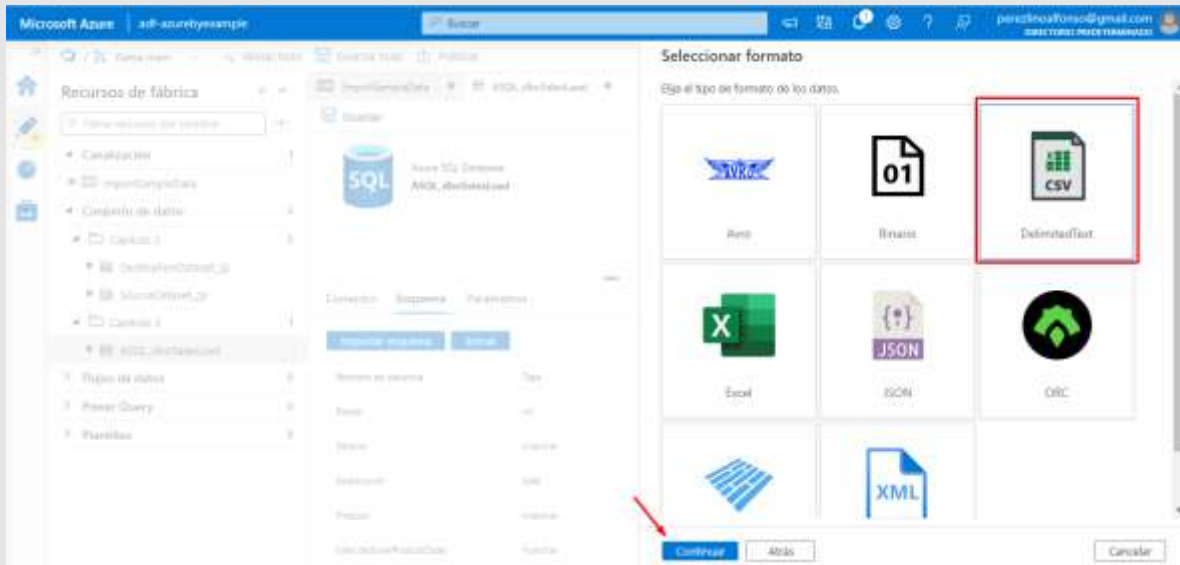
**Nota** La UX del ADF está más estrechamente vinculada a Git que en otros entornos de desarrollo. Una sesión de ADF UX no tiene almacenamiento permanente propio, por lo que los cambios se guardan directamente en su repositorio Git. Cada guardado que haga en el ADF UX es un commit de Git que se empuja inmediatamente a tu repositorio alojado.



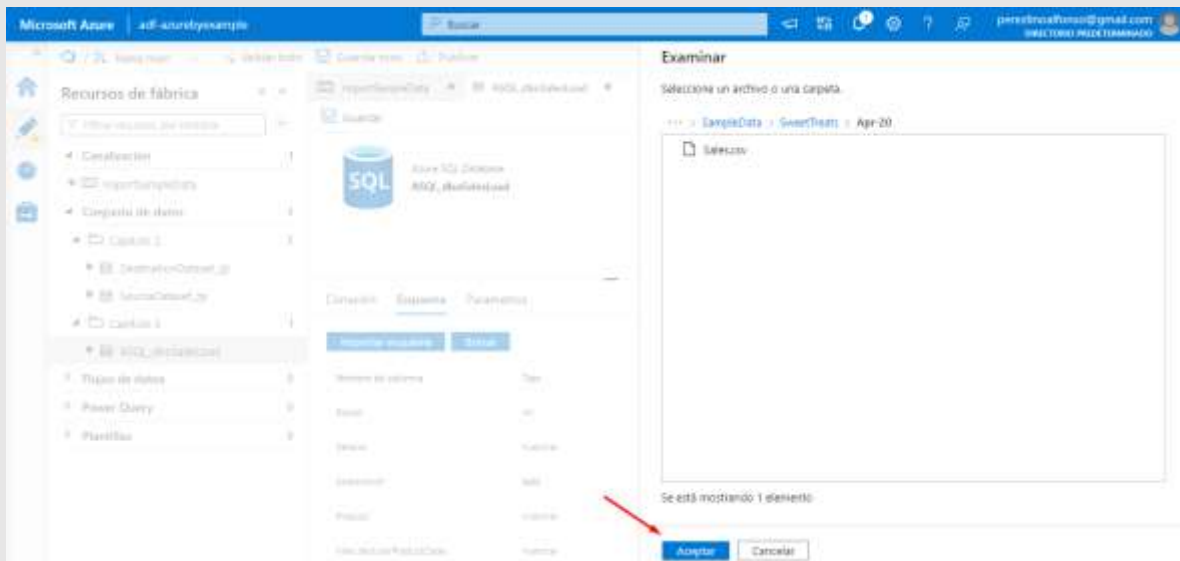
### 3.2.3. Crear un conjunto de datos de archivos de texto delimitado

El pipeline que creas cargará los datos CSV de muestra de tu cuenta de almacenamiento blob en una tabla de Azure SQL Database. Usted creó un servicio vinculado para conectarse a la cuenta de almacenamiento en el Capítulo 2 - puede reutilizarlo para crear un conjunto de datos que represente un archivo CSV de origen.

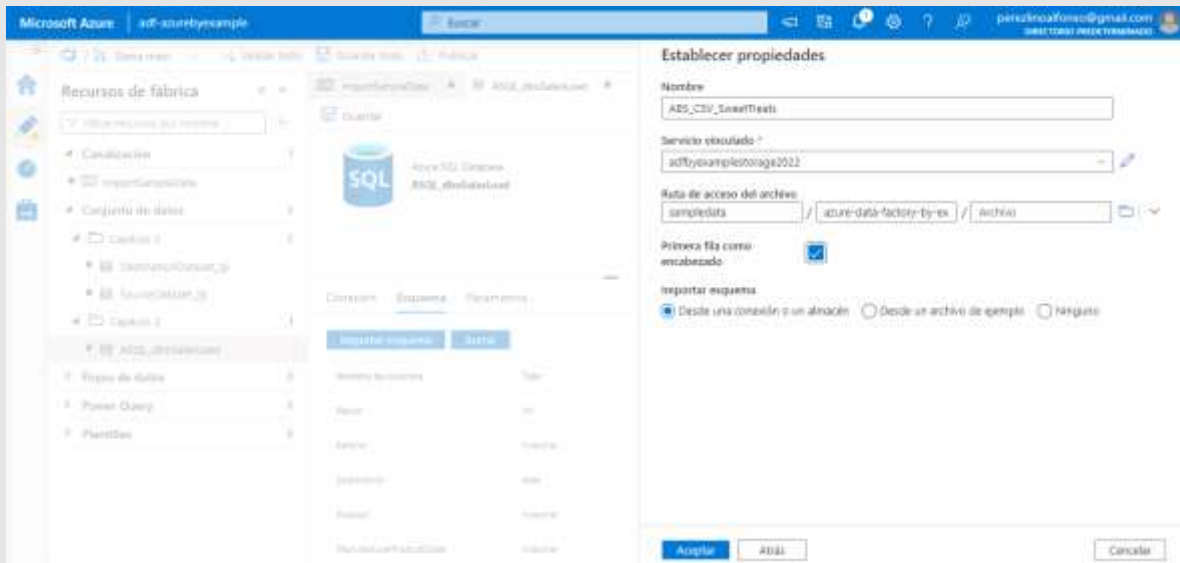
- Cree un nuevo conjunto de datos en la carpeta "Capítulo 3" seleccionando Nuevo conjunto de datos en el menú Acción de la carpeta y, a continuación, seleccione Azure Blob Storage. Haz clic en Continuar.
- En la hoja Seleccionar formato, elija Texto delimitado y haga clic en Continuar.



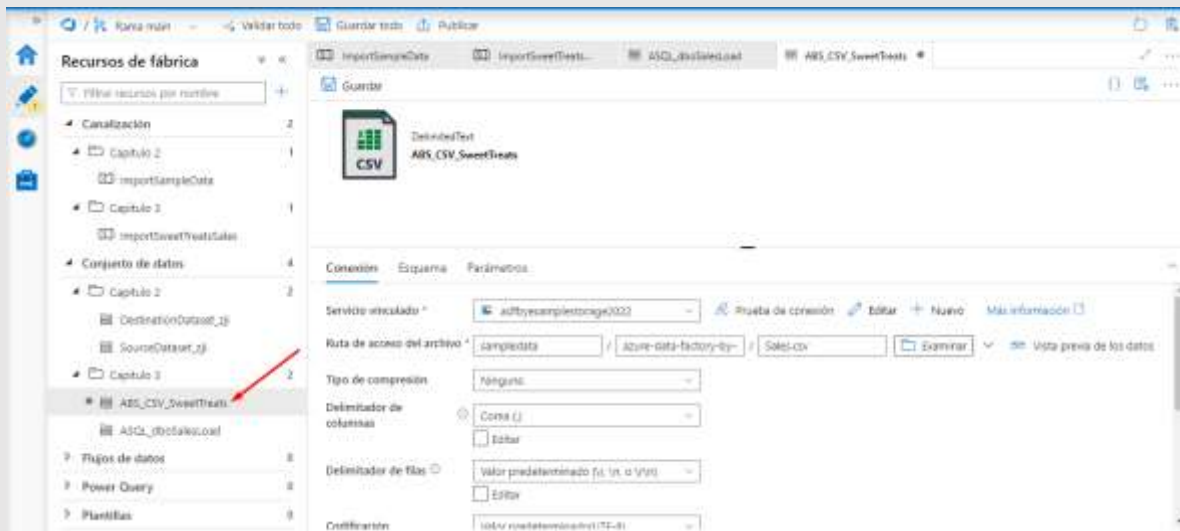
3. Nombre el conjunto de datos "ABS\_CSV\_SweetTreats" - el archivo que va a importar es un archivo CSV en Azure Blob Storage (ABS) y contiene datos de ventas para un minorista llamado Sweet Treats. Seleccione su servicio vinculado de Blob storage existente y, a continuación, haga clic en el icono de la carpeta situado a la derecha de los campos de la ruta del archivo.
4. Navegue por el contenedor "sampledata" hasta encontrar la carpeta "SampleData". Navegue hasta la ruta de la subcarpeta "SweetTreats/Apr-20" y seleccione el archivo "Sales.csv". Haga clic en Aceptar para seleccionar el archivo y desechar el selector de archivos.



5. Marque la casilla Primera fila como cabecera y haga clic en Aceptar para crear el conjunto de datos.



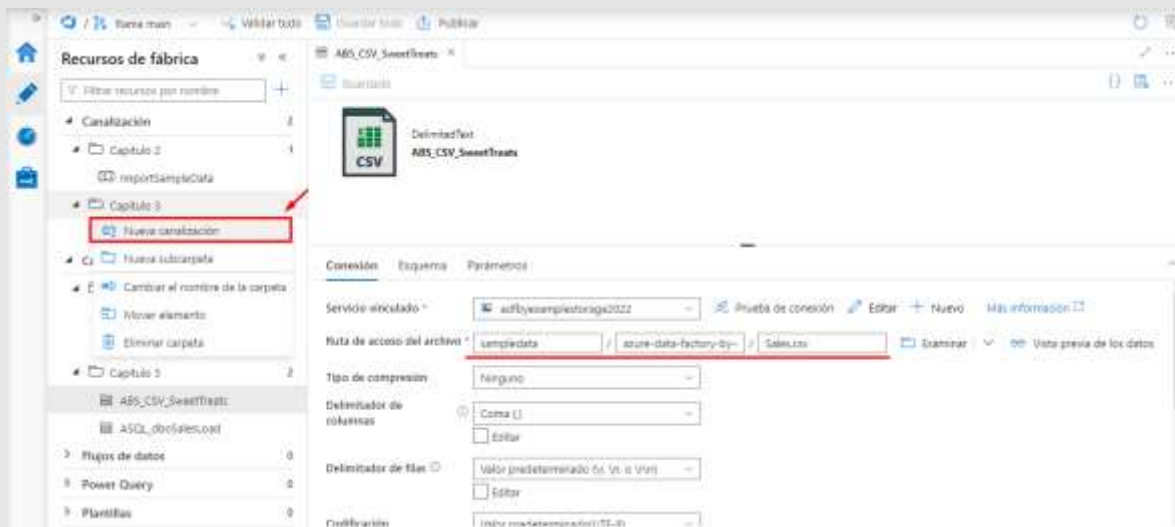
6. El nuevo conjunto de datos se abre en otra pestaña nueva en el lienzo de creación - haga clic en Guardar para confirmar y enviar la definición del conjunto de datos a Git.



### 3.2.4. Crear y ejecutar el pipeline

Ahora está listo para crear el pipeline. Antes de empezar, es posible que desee mover el pipeline del Capítulo 2 a una carpeta de pipelines del "Capítulo 2".

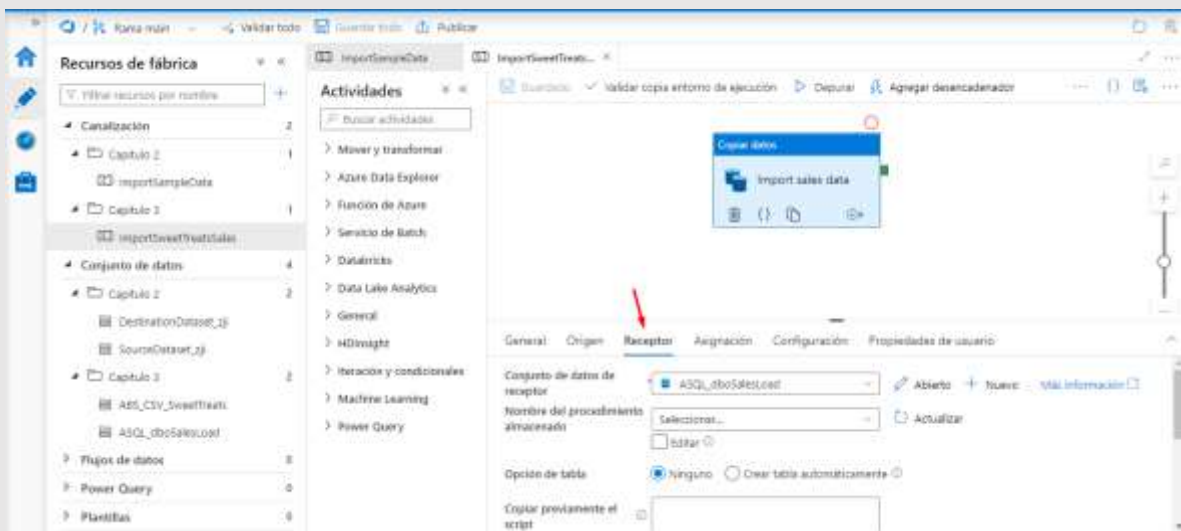
1. En el explorador de Recursos de la Fábrica, crea una carpeta "Capítulo 3" en Pipelines, y luego selecciona Nuevo pipeline en el menú Acciones de la carpeta.



2. Se abre una nueva pestaña de pipeline con la hoja de Propiedades. El nuevo pipeline tiene un nombre por defecto como "pipeline1" - cambie el nombre a "ImportSweetTreatsSales" y añada una descripción, luego cierre la hoja.
3. Si necesita más espacio en el lienzo de creación, contraiga el explorador de Recursos de fábrica haciendo clic en el botón de los galones (") a la derecha del encabezamiento Recursos de fábrica.
4. A la izquierda del lienzo de creación se encuentra la caja de herramientas de actividades, titulada Actividades. Despliegue el grupo Mover y transformar, luego arrastre una actividad **Copiar datos** desde la caja de herramientas y suéltela en el lienzo. El panel de configuración debajo del lienzo se expande automáticamente para permitirle configurar la actividad.
5. En la pestaña General del panel de configuración, cambie el nombre de la actividad Copiar datos a "Import sales data".

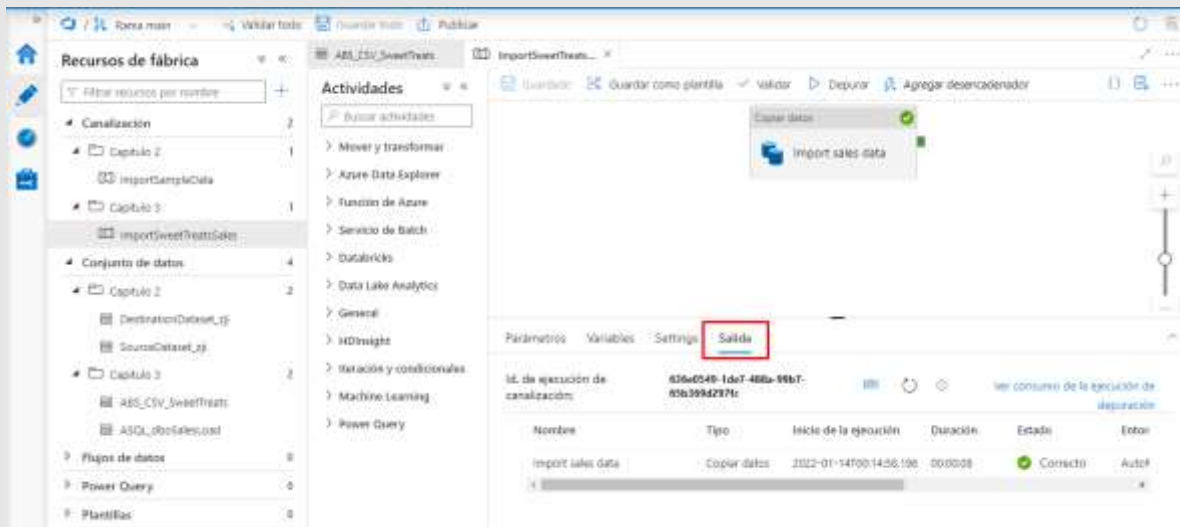
**Nota** Los nombres de las actividades pueden contener caracteres alfanuméricos, guiones, guiones bajos y espacios, pero los nombres no pueden empezar ni terminar con un espacio. Los nombres de las actividades son validados por la UX del ADF a medida que se escribe, por lo que cuando se añade un carácter de espacio, se recibe un error de validación inmediato; éste desaparece cuando se introduce el siguiente carácter válido que no sea un espacio.

6. En la pestaña Origen, seleccione el conjunto de datos "ABS\_CSV\_SweetTreats", y en la pestaña Receptor (Sink), elija el conjunto de datos "ASQL\_dboSalesLoad". Azure Data Factory se refiere a los destinos de datos de pipeline como receptores.



7. En la barra de herramientas del lienzo (encima del lienzo de creación), haga clic en Guardar para confirmar y enviar los cambios, y luego en Depurar para ejecutar el pipeline.





La información de la ejecución del pipeline se muestra en la pestaña Salida del panel de configuración del pipeline (mostrado en la Figura 3-8). Al pasar el ratón por encima del nombre de la actividad Copiar datos en la pestaña Salida, se muestran **tres iconos** para acceder a la información sobre la **Entrada**, la **Salida** y los **Detalles** de la actividad.

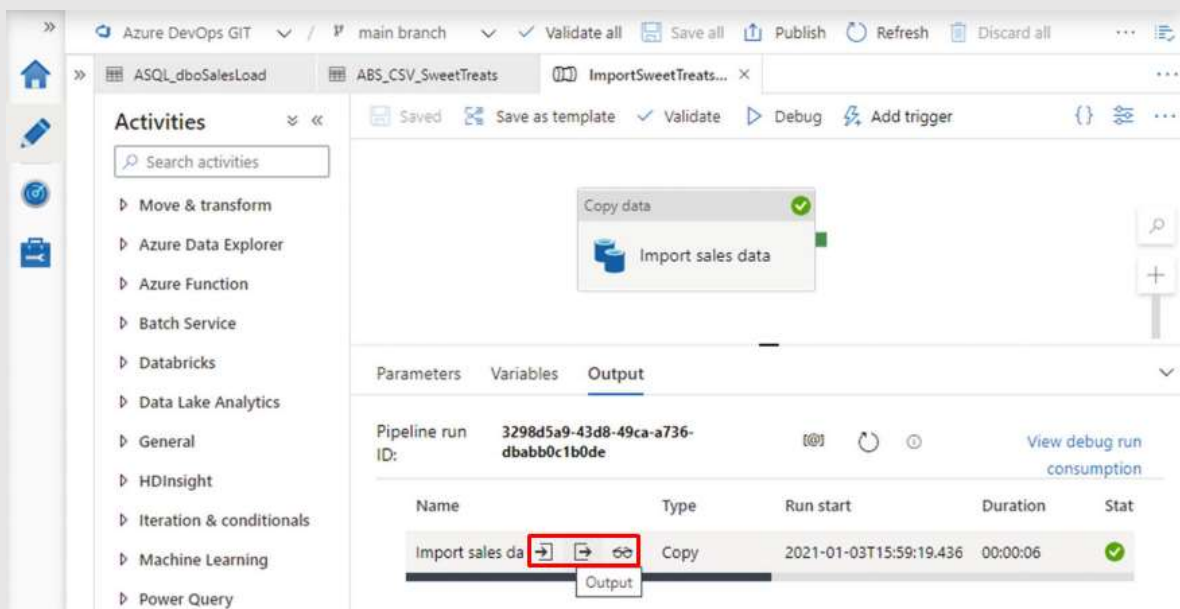


Figura 3-8 Lienzo de creación con la pestaña Salida del panel de configuración de la canalización

La salida de la actividad Copiar datos es un objeto JSON que contiene datos sobre la actividad ejecutada - la primera parte de un ejemplo de salida se muestra en la Figura 3-9. Observe especialmente los atributos

**filesRead:** El número de archivos leídos por la ejecución de la actividad *Copiar datos* (en este ejemplo, un archivo)

**rowsRead:** El número total de filas de origen leídas por la actividad ejecutada



**rowsCopied:** El número de filas del receptor escritas por la ejecución de la actividad

The screenshot displays the Azure Data Factory portal. At the top, the 'Salida' (Output) tab is selected for the 'Copiar datos' (Copy data) activity. The activity's execution details are shown in a table:

Nombre	Tipo	Inicio de la ejecución	Duración	Estado	Enton
Import sales data	Copiar datos	2022-01-14T00:14:58.196	00:00:08	Correcto	AutoF

A red box highlights the 'Copiar datos' activity. Below the table, a 'Salida' (Output) window is open, displaying the following JSON output:

```
{  "dataRead": 6993,  "dataWritten": 10238,  "filesRead": 1,  "sourcePeakConnections": 1,  "sinkPeakConnections": 2,  "rowsRead": 117,  "rowsCopied": 117,  "copyDuration": 5,  "throughput": 1.366,  "errors": []}
```

Figura 3-9 Primera parte del objeto JSON de salida de una actividad de copia de datos

### 3.2.5. Verificación de los resultados

Vuelva al cliente SQL de su elección y verifique que la tabla `dbo.Sales_LOAD` contiene ahora el número de filas informadas por la ejecución de la actividad Copiar datos.

También puede inspeccionar el contenido del archivo "Sales.csv" de abril de 2020 para verificar que su contenido se ha cargado correctamente. Las primeras filas del archivo se muestran en el Listado 3-2. En particular, fíjese en los nombres de los campos de la primera fila del archivo: coinciden exactamente con los nombres de las columnas de la tabla de la base de datos. Durante la creación del pipeline, **usted no proporcionó ninguna asignación de los campos CSV a las columnas de la tabla, pero la actividad Copiar datos dedujo la asignación automáticamente utilizando los nombres de las columnas coincidentes.**

```

SalesMonth,Retailer,Product,SalesValueUSD,UnitsSold
"01-Apr-2020","Sweet Treats","Schnoogles 8.81oz",3922.31,409
"01-Apr-2020","Sweet Treats","Creamies 10.57oz",3057.18,502
"01-Apr-2020","Sweet Treats","Caramax 6.59oz",1147.37,443

```

Listado 3-2 Las primeras filas del archivo Sales.csv

SalesMonth	Retailer	Product	SalesValueUSD	UnitsSold
"01-Apr-2020"	"Sweet Treats"	"Schnoogles 8.81oz"	3922.31	409
"01-Apr-2020"	"Sweet Treats"	"Creamies 10.57oz"	3057.18	502
"01-Apr-2020"	"Sweet Treats"	"Caramax 6.59oz"	1147.37	443
"01-Apr-2020"	"Sweet Treats"	"Salt Chocolate 10.57oz"	2788.00	874
"01-Apr-2020"	"Sweet Treats"	"Chocolatey Nougat 8.81oz"	666.81	279
"01-Apr-2020"	"Sweet Treats"	"The Original 4.23oz"	886.38	663
"01-Apr-2020"	"Sweet Treats"	"Mellows 28.25oz"	4950.46	254
"01-Apr-2020"	"Sweet Treats"	"Zoots 3.37oz"	48.65	33
"01-Apr-2020"	"Sweet Treats"	"Sweetverse 7.05oz"	1149.29	281
"01-Apr-2020"	"Sweet Treats"	"Humburs 2.47oz"	1007.40	460
"01-Apr-2020"	"Sweet Treats"	"Xero 4.33oz"	11.96	10
"01-Apr-2020"	"Sweet Treats"	"Zigzags 5.84oz"	1231.77	463
"01-Apr-2020"	"Sweet Treats"	"YumYums 22.91oz"	3453.22	198
"01-Apr-2020"	"Sweet Treats"	"Zigzags 3.52oz"	1522.20	705
"01-Apr-2020"	"Sweet Treats"	"Bitterbutter 3.88oz"	899.94	476
"01-Apr-2020"	"Sweet Treats"	"Dr Sweets 9.52oz"	1364.22	858
"01-Apr-2020"	"Sweet Treats"	"Paranails 8.25oz"	1709.86	254
"01-Apr-2020"	"Sweet Treats"	"Whippersnappers 4.41oz"	1535.67	493
"01-Apr-2020"	"Sweet Treats"	"YumYums 5.71oz"	490.14	126
"01-Apr-2020"	"Sweet Treats"	"Minotaur 13.39oz"	4550.28	814
"01-Apr-2020"	"Sweet Treats"	"Elysian Fudge 4.41oz"	1954.93	777

Inicio > SQL Database > sqlidb-adfbyexample (sql-adfbyexample2022/sqlidb-adfbyexample)

sqlidb-adfbyexample (sql-adfbyexample2022/sqlidb-adfbyexample) | Editor de consultas (versión preliminar)

Consulta 1: `SELECT * FROM Sales_LOAD`

RowId	Retailer	SalesMonth	Product	ManufacturerProduct
1	Sweet Treats	2020-04-01T00:00:00	Schnoogles 8.81oz	
2	Sweet Treats	2020-04-01T00:00:00	Creamies 10.57oz	
3	Sweet Treats	2020-04-01T00:00:00	Caramax 6.59oz	

Consulta realizada correctamente | 0s

**Para los desarrolladores de SSIS:** En el Capítulo 2, comparé la actividad Copiar datos con una File System Task de SSIS, pero aquí proporciona la funcionalidad de una simple Data Flow Task que contiene dos componentes: una fuente de datos de archivo plano y un destino de SQL Server.

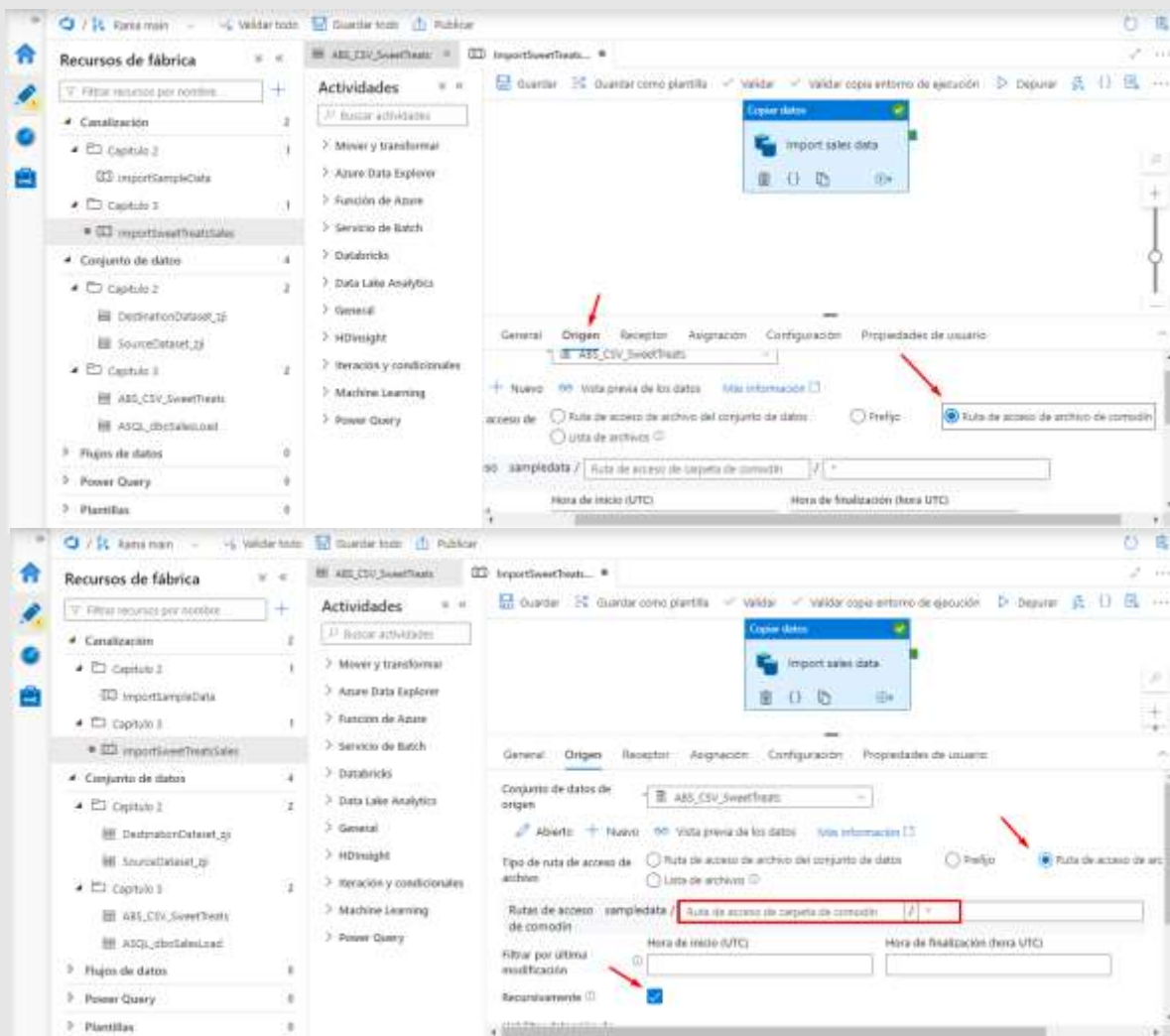
### 3.2.6. Procesar Múltiples Archivos

El sample dataset incluye seis meses de datos de ventas para Sweet Treats. El archivo de cada mes tiene el mismo nombre - "Sales.csv" - y como el archivo de abril de 2020 se encuentra en una carpeta con el nombre del mes al que se refiere. El conjunto de datos que creó en la sección anterior identifica específicamente el archivo de datos de abril, pero **la actividad Copiar datos admite un comportamiento de comodín que le permite especificar varios archivos para su procesamiento.**

1. En el lienzo de creación del ADF UX, seleccione la actividad Copiar datos de su pipeline sales data import. El panel de configuración de la actividad se expande automáticamente debajo del lienzo.

**Consejo** Cuando se selecciona una actividad en el lienzo de creación, se muestran funciones adicionales específicas de la actividad, como Eliminar, Clonar y Ver código fuente. La función Agregar salida se presenta en el capítulo 6.

2. Seleccione la pestaña Origen y cambie el Tipo de ruta de archivo a "**Ruta de archivo comodín**". Esto significa que ya no se utilizará la ruta de archivo original especificada en el conjunto de datos. Se muestran los campos para las rutas comodín (Figura 3-10).



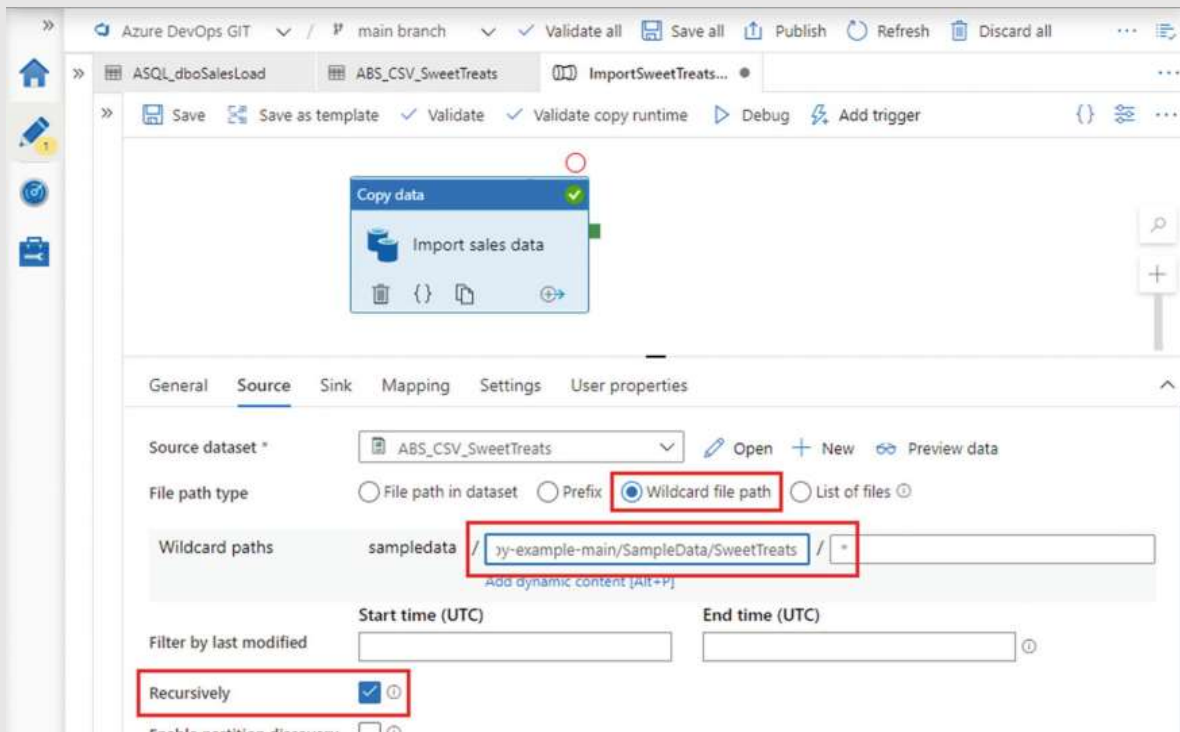
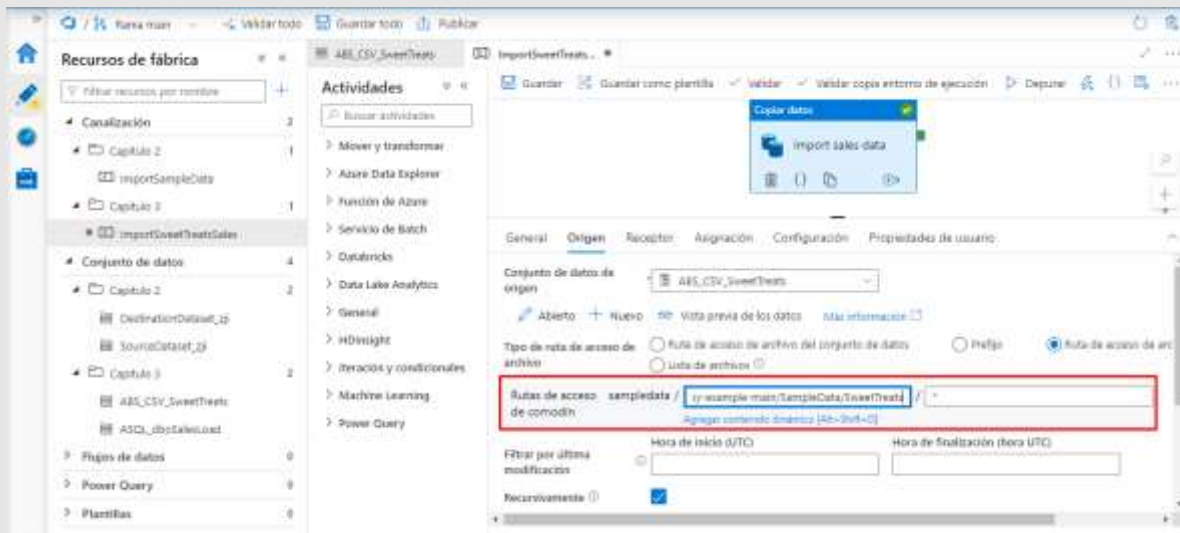


Figura 3-10 Ruta de archivos con comodines en el origen de la actividad de copia de datos

3. La parte del contenedor de la ruta de almacenamiento blob se rellena automáticamente a partir del conjunto de datos "ABS\_CSV\_SweetTreats". En el campo de ruta de la carpeta Wildcard, introduzca **"azure-data-factory-by-example-main/SampleData/SweetTreats"** (la ruta de la carpeta "SweetTreats"). La ruta termina con "SweetTreats" y no tiene ningún carácter de barra inicial o final.

**Consejo** A diferencia de los entornos de Windows, las rutas de los archivos de Azure Blob Storage distinguen entre mayúsculas y minúsculas, y las rutas se delimitan con el carácter de la barra diagonal ("/").

4. En Nombre de archivo comodín, introduzca un asterisco ("\*"): esto indica que se importarán todos los archivos que se encuentren en la carpeta "SweetTreats".
5. Desplácese hacia abajo y asegúrese de que la casilla **"Recursivamente"** está marcada. Esto indica que se deben incluir los archivos que coincidan con la ruta comodín en cualquier subcarpeta anidada de "SweetTreats".



Guarde los cambios y, a continuación, haga clic en Depurar para ejecutar el pipeline. Cuando termine, mire el JSON de salida de la actividad Copiar datos. Esta vez, notará que la actividad leyó un total de 6 archivos y 684 filas - todos los archivos "Sales.csv" para los seis meses entre abril y septiembre de 2020.

### 3.2.7. Truncar antes de cargar

Vuelva al cliente SQL de su elección y ejecute la consulta del Listado 3-3 para contar el número de registros cargados para cada mes.

```
SELECT
    Retailer,
    SalesMonth,
    COUNT(*) AS [Rows]
FROM Sales_LOAD
GROUP BY
    Retailer,
    SalesMonth;
```

Listado 3-3 Recuento de registros cargados por mes

sqladb-adfbyexample (admin-adfby...)

Aquí se muestra el Explorador de objetos limitado. Abra SSOT para acceder a las capacidades completas.

Tablas

- dbo.Sales\_LOAD
  - Rowid (PK, int, not null)
  - Retailer (nvarchar, null)
  - SalesMonth (date, null)
  - Product (nvarchar, null)
  - ManufacturerProductCode (nvarchar, null)
  - SalesValueUSD (decimal, null)
  - UnitsSold (int, null)

Vistas

Procedimientos almacenados

Consulta 1

Consulta 2

Ejecutar

Cancelar consulta

Guardar consulta

Exportar datos como

Mostrar solo editor

```
1 SELECT
2   Retailer,
3   SalesMonth,
4   COUNT(*) AS [Rows]
5 FROM Sales_LOAD
6 GROUP BY
7   Retailer,
8   SalesMonth;
```

Resultados

Mensajes

Buscar en elementos de filtro...

Retailer	SalesMonth	Rows
Sweet Treats	2020-04-01T00:00:00.0000000	234

Consulta realizada correctamente | 0s

Sin el comodín



Consulta 2

Ejecutar Cancelar consulta Guardar consulta Exportar datos como Mostrar todo

Resultados Mensajes

Buscar en elementos de filtro...

Retailer	SalesMonth	Rows
Sweet Treats	2020-04-01T00:00:00.0000000	351
Sweet Treats	2020-06-01T00:00:00.0000000	114
Sweet Treats	2020-05-01T00:00:00.0000000	113
Sweet Treats	2020-09-01T00:00:00.0000000	112
Sweet Treats	2020-08-01T00:00:00.0000000	111
Sweet Treats	2020-07-01T00:00:00.0000000	117

Consulta realizada correctamente | 0s

Con el comodín

Como se esperaba, ahora hay filas presentes para cada uno de los seis meses, pero observe que el conteo de filas para abril es aproximadamente el doble que el de los otros meses. Esto se debe a que ha cargado los datos de abril dos veces - una vez cuando ejecutó su pipeline por primera vez, y luego una segunda vez mientras cargaba los datos de los seis meses. Para utilizar la tabla `dbo.Sales_LOAD` correctamente, debe truncarse antes de cada actividad de carga.

Este requisito puede ser soportado usando la funcionalidad de la actividad de Copiar datos provista para los receptores de la base de datos SQL. En la pestaña Receptor de la actividad en la UX del ADF, añada este comando SQL al campo Pre-copy script (Figura 3-11):

TRUNCATE TABLE `dbo.Sales_LOAD`

Receptor

Conjunto de datos de receptor: ASQL\_dboSalesLoad

Nombre del procedimiento almacenado: Seleccionar...

Opción de tabla: Ninguno

Copiar previamente el script: TRUNCATE TABLE `dbo.Sales_LOAD`

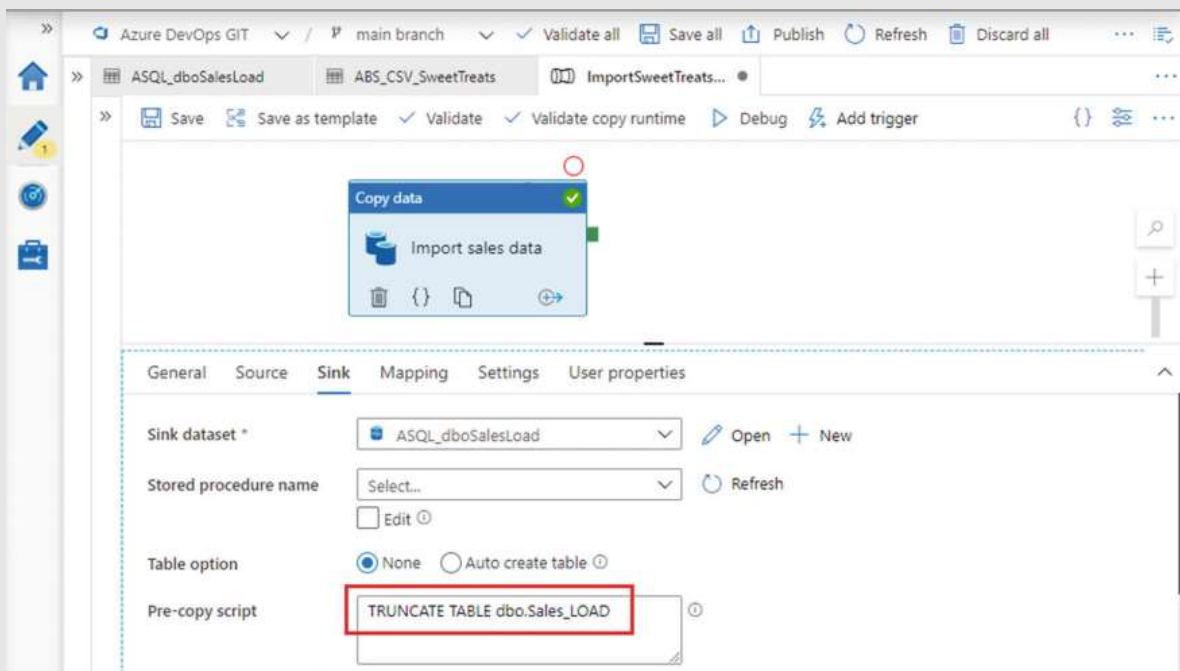
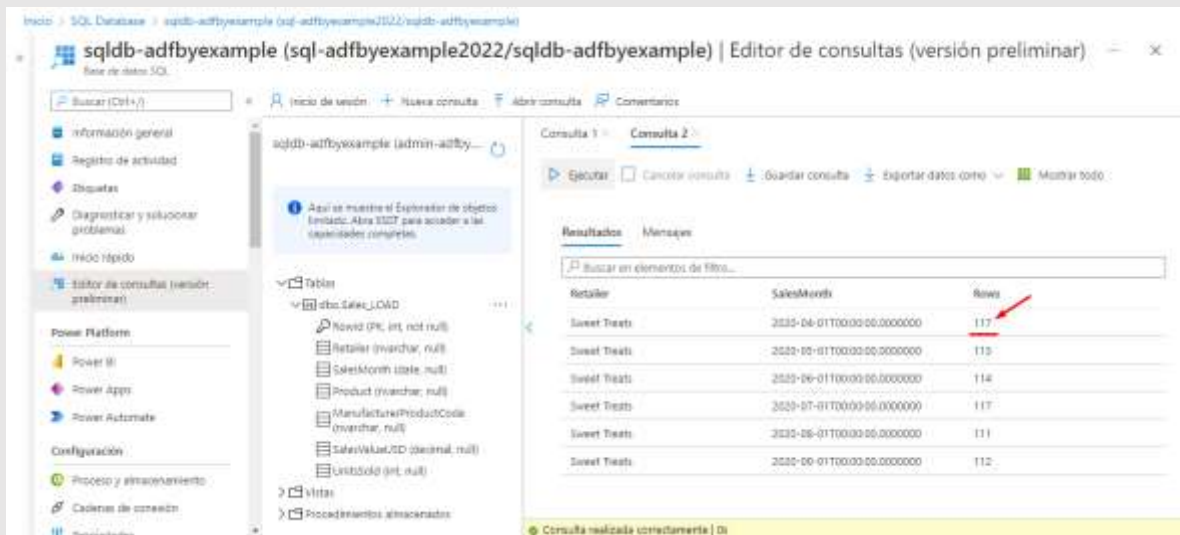


Figura 3-11 Script previo a la copia en el sumidero de la actividad Copiar datos

Este comando SQL se ejecutará en la base de datos que contiene el conjunto de datos del sumidero cada vez que se ejecute la actividad Copiar datos, antes de que se copie cualquier dato.

Ejecute la canalización de nuevo haciendo clic en Depuración (observe que el modo de depuración le permite ejecutar la canalización revisada sin tener que guardar los cambios) y, a continuación, compruebe el recuento de filas de la tabla utilizando el script del Listado 3-3. Esta vez, encontrará que se reporta un número similar de filas para cada uno de los seis meses.



### 3.3. Asignar los esquemas de origen y de destino

El pipeline que creó en la sección anterior se basa en la capacidad de la actividad Copiar datos para inferir mapeos de columnas mediante la coincidencia de nombres en los archivos de origen y la tabla de destino. En esta sección, descubrirá qué sucede cuando los nombres de las columnas de origen y de destino no coinciden y cómo manejar esta situación.

#### 3.3.1. Crear un nuevo dataset de origen

El pipeline que va a crear cargará los datos de un minorista diferente llamado Candy Shack. Sus datos de origen también se suministran en archivos CSV, por lo que puede basar su conjunto de datos de origen en "ABS\_CSV\_SweetTreats" clonándolo y editando el clon.

1. La Figura 3-12 muestra la UX del ADF con el explorador de Recursos de Fábrica expandido. Al pasar el ratón por encima del extremo derecho del conjunto de datos "ABS\_CSV\_SweetTreats" (debajo del recuento de pipeline de la carpeta "Capítulo 3") aparece el botón de Acciones de la elipsis - haga clic en él para acceder al menú de Acciones del conjunto de datos, y luego seleccione Clonar.

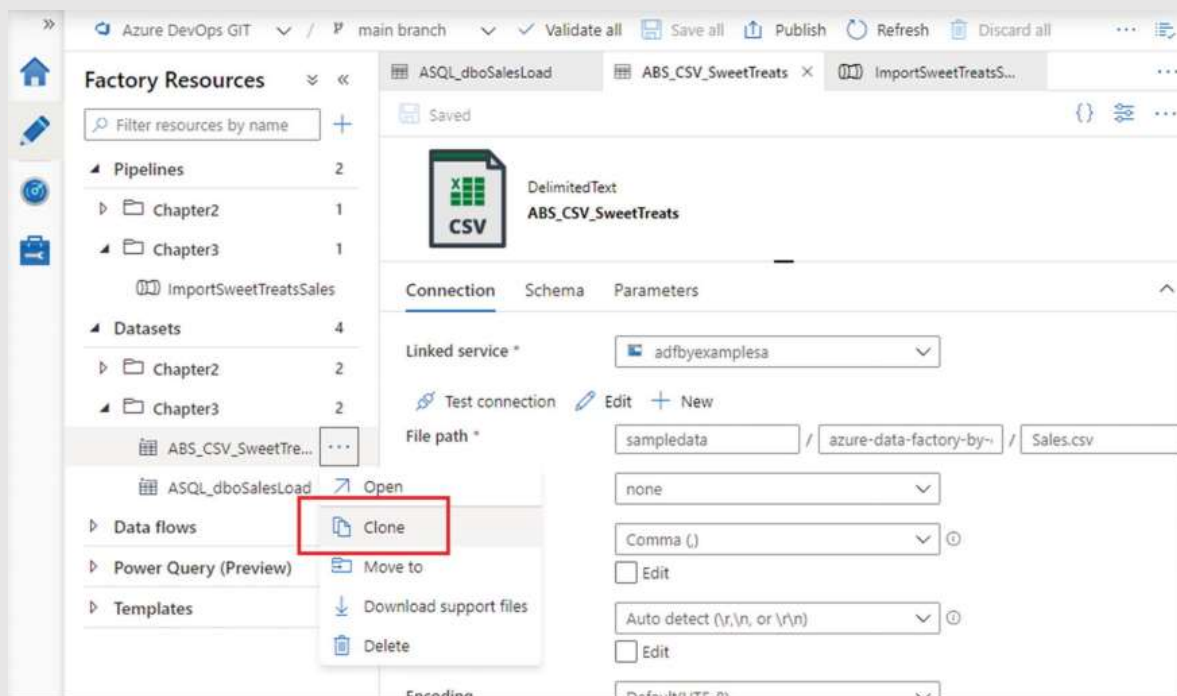


Figura 3-12 Menú de acciones del conjunto de datos

2. Clonar crea una copia del conjunto de datos original y lo abre en la UX del ADF. El nuevo conjunto de datos recibe automáticamente un nombre único basado en el original; cuando se abre, aparece la hoja de propiedades para que pueda cambiarle el nombre. Cambie su nombre a "ABS\_CSV\_CandyShack" y cierre la hoja.

3. Edite la ruta del archivo del conjunto de datos navegando hasta la carpeta "CandyShack" anidada bajo el contenedor "sampledata" y seleccionando el archivo "2020-04.csv".

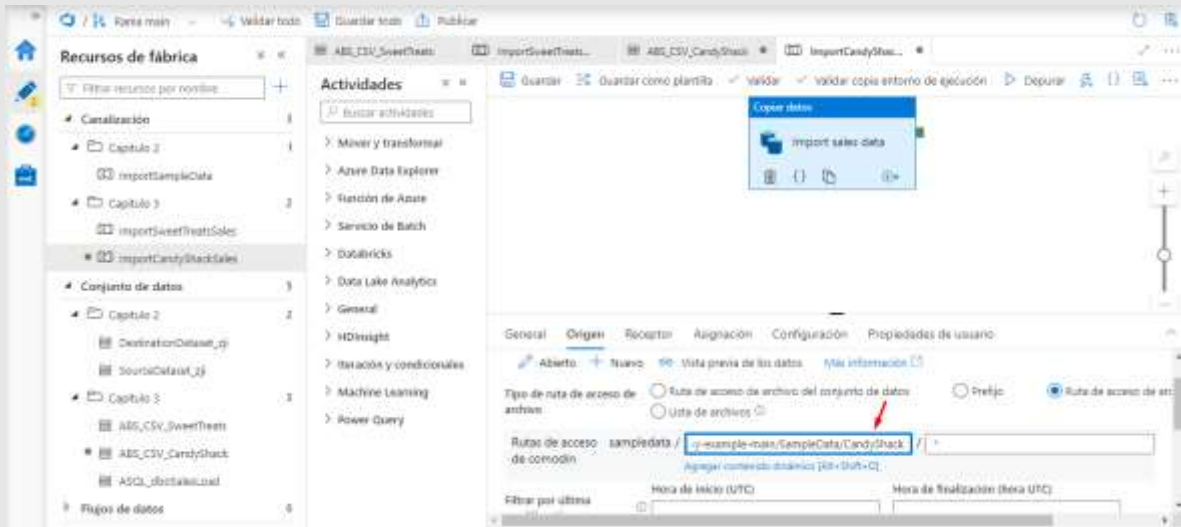
4. Haga clic en Guardar todo para guardar el nuevo conjunto de datos y cualquier cambio no guardado.



### 3.3.2. Crear una nueva pipeline

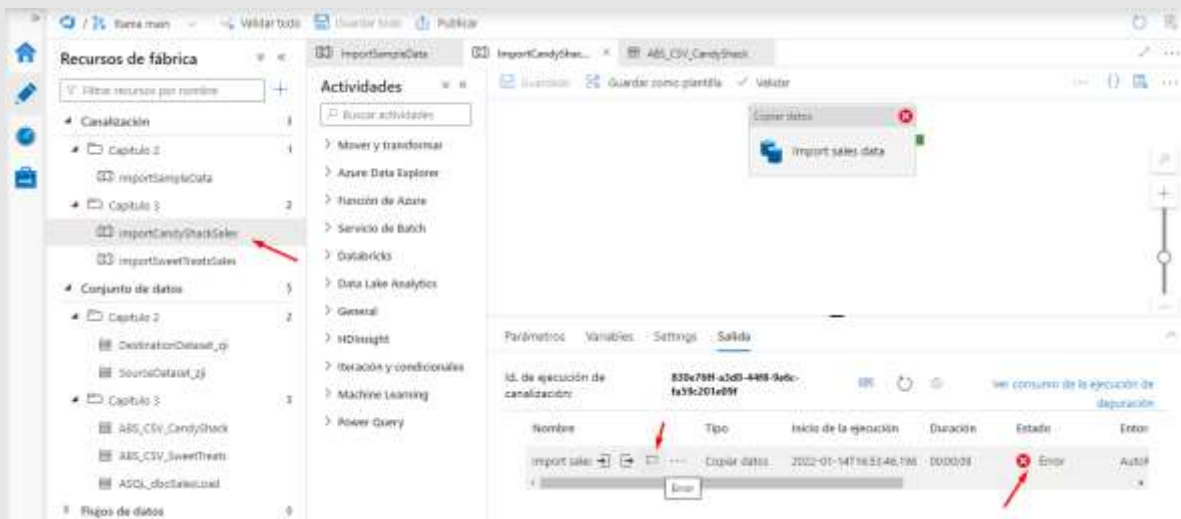
El proceso en sí es muy similar al que creó para cargar los datos de Sweet Treats. Clone el pipeline "ImportSweetTreatsSales" de la misma manera que clonó el conjunto de datos, y modifíquelo como sigue:

1. Nombre el nuevo pipeline "ImportCandyShackSales".
2. Cambie el conjunto de datos de origen de la actividad Copiar datos; utilizando la lista desplegable, seleccione "ABS\_CSV\_CandyShack".
3. Todavía en la pestaña Origen, edite las rutas Wildcard, sustituyendo la referencia a la carpeta "SweetTreats" por "CandyShack" (asegurándose de dejar el resto de la ruta intacta).



4. Haga clic en Debug para ejecutar el pipeline. Cuando la ejecución del pipeline se detenga, descubrirá que la ejecución de la actividad de copia ha fallado.
5. Como descubrió en el Capítulo 2, la información de la ejecución del pipeline aparece en la pestaña Salida del panel de configuración del pipeline. Debido a que la ejecución de la actividad Copiar datos ha fallado, ahora aparece un botón adicional de Error (icono de burbuja de diálogo) al pasar el ratón por encima de su nombre. Haga clic en Error para ver la información sobre el motivo del fallo de la actividad.

La ventana emergente de error que describe el fallo de la actividad se muestra en la Figura 3-13. Observe especialmente el texto "La columna SkuCode no se encuentra en la parte de destino". Esto indica que se encontró un campo llamado "SkuCode" en los datos de origen, pero que la tabla de destino no contiene el campo correspondiente con el mismo nombre.



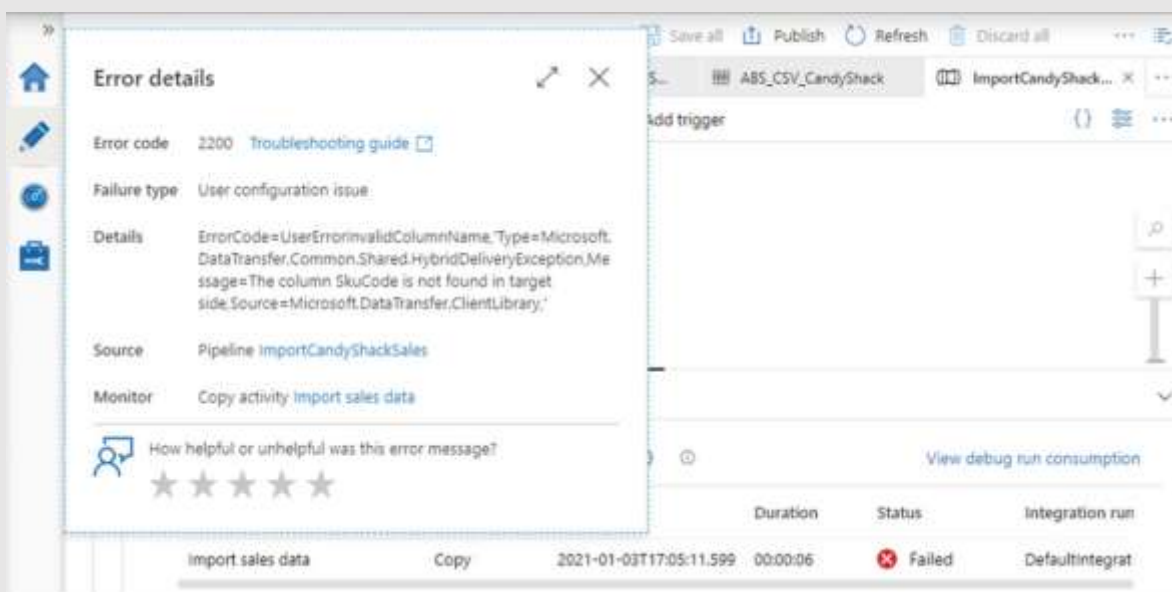
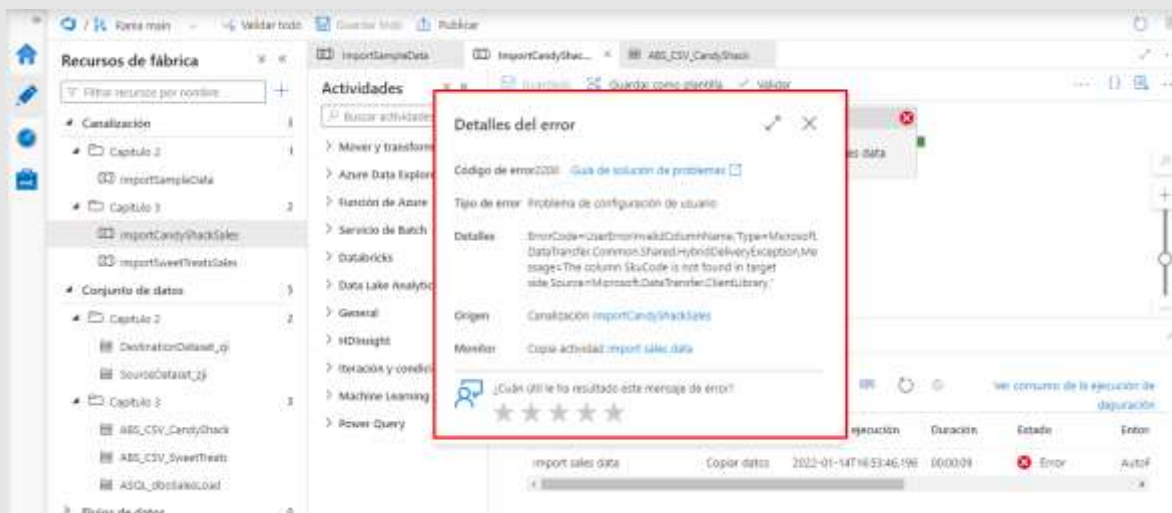
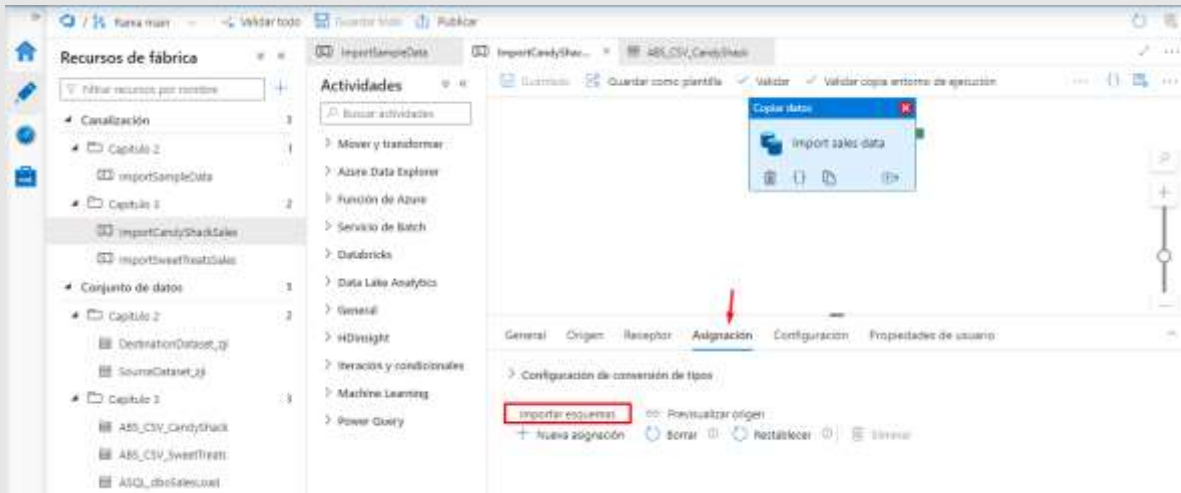


Figura 3-13 Detalles del error de la actividad de copia de datos

### 3.3.3. Configurar la asignación de esquemas

Para que esta actividad de copia de datos tenga éxito, primero debe asignar las columnas de origen a las columnas de destino de forma explícita, como se indica a continuación:

1. Seleccione la pestaña **Asignación** (Mapping) en el panel de configuración de la actividad Copiar datos y haga clic en el botón Importar esquemas. Espere unos instantes mientras el ADF UX recupera la información de los esquemas de un archivo de origen de Candy Shack y de la tabla de destino de la base de datos.



2. La UX del ADF intenta asignar automáticamente las columnas de origen y de destino importadas, mostrando mensajes de error o advertencia cuando sea necesario. La Figura 3-14 muestra los resultados iniciales de la importación de los dos esquemas.

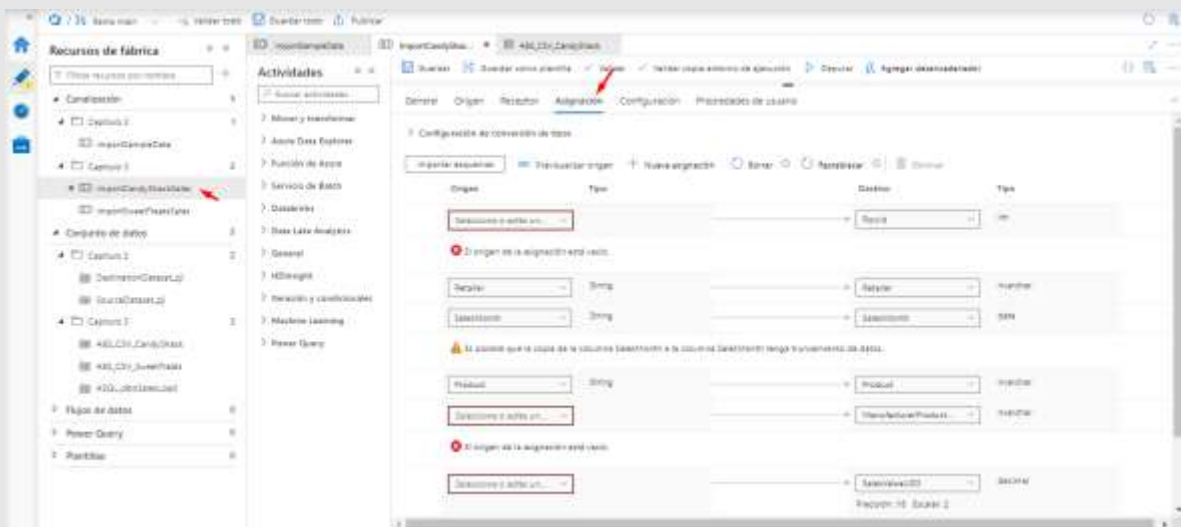
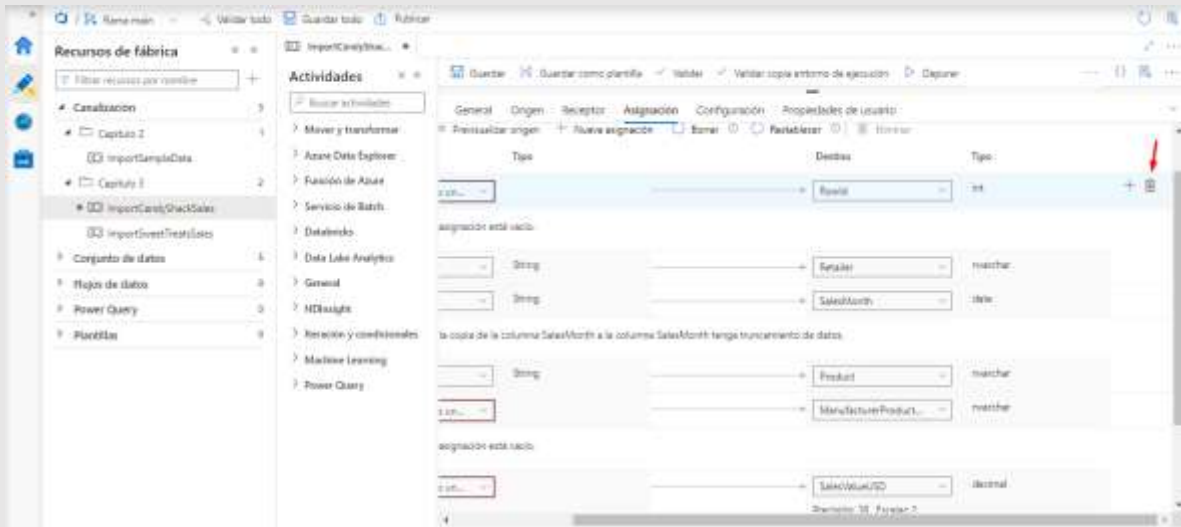
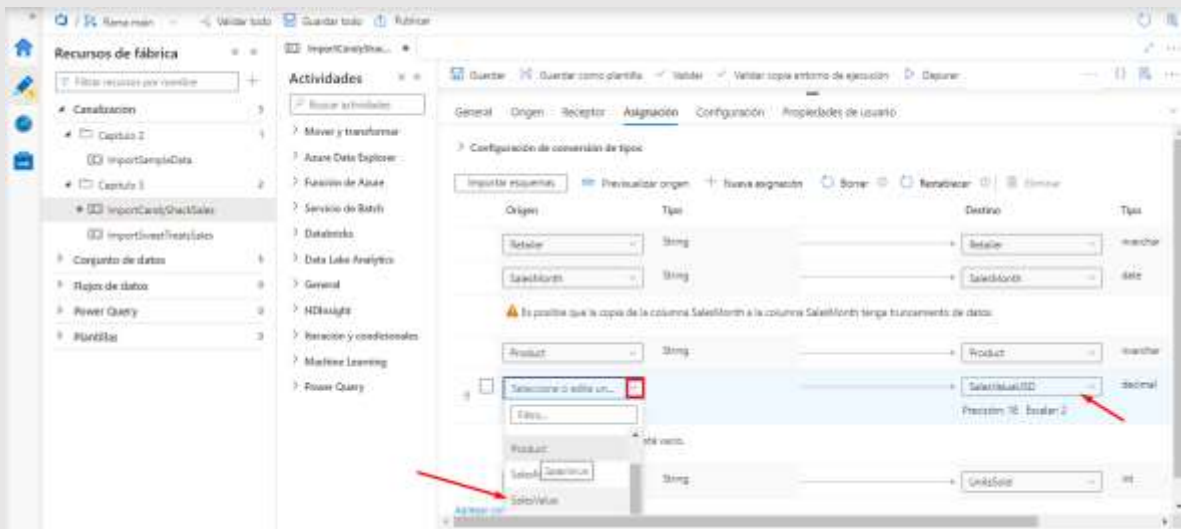


Figura 3-14 Esquemas importados de origen de Candy Shack y del receptor de la base de datos

3. Al pasar el ratón por encima de un mapeo de columna, aparecen controles adicionales a la derecha, que permiten borrarlo o añadir un nuevo mapeo. En este ejemplo, la columna de la base de datos de destino [RowId] es una identidad entera generada automáticamente, por lo que el mapeo puede ser eliminado - utilice el icono de la papelera a la derecha del mapeo para hacerlo. El archivo de origen no contiene ningún campo correspondiente a [ManufacturerProductCode], así que elimine también esta asignación.

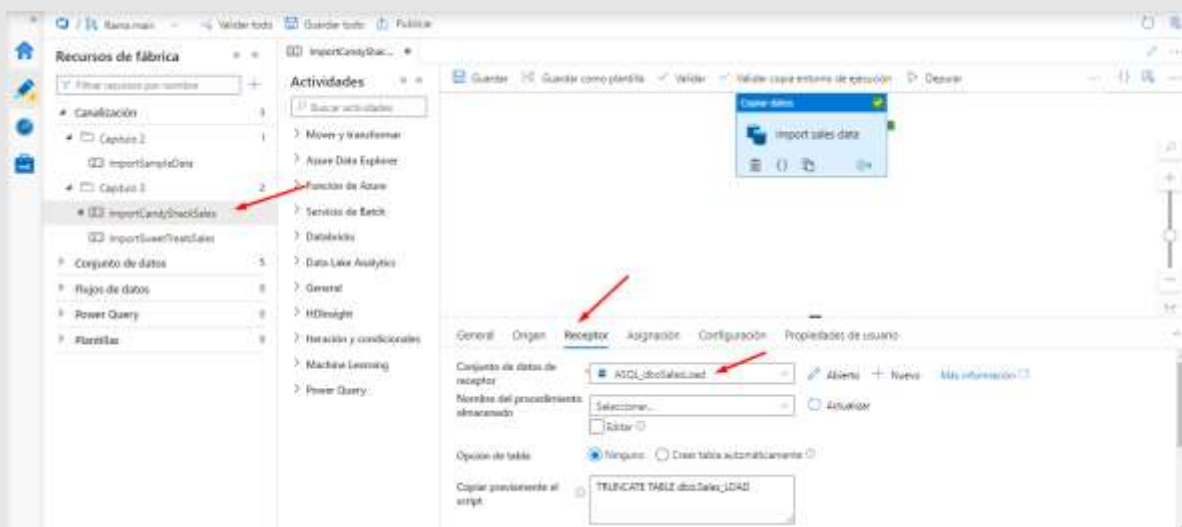
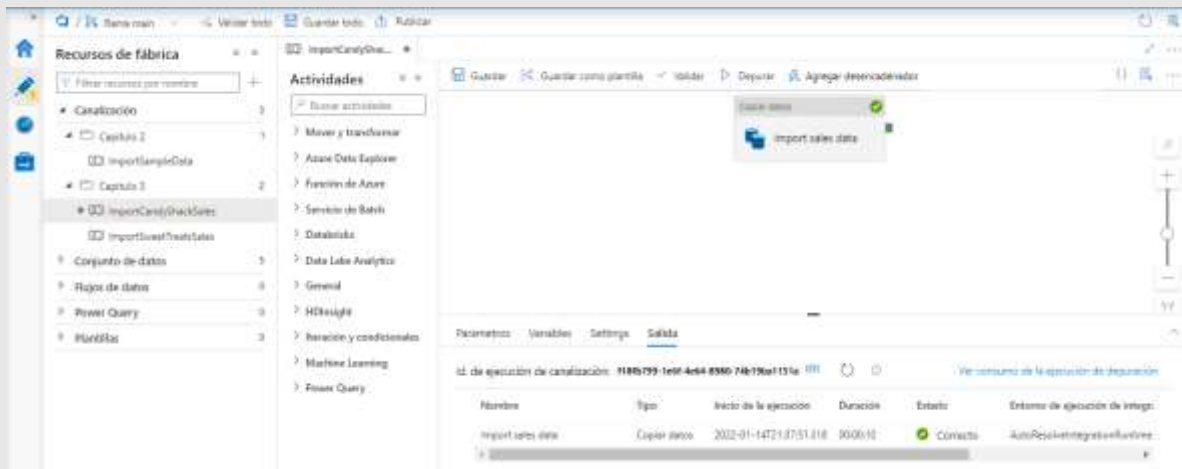


- Las listas desplegables para cada mapeo de campo le permiten seleccionar los mapeos de nombres de campo de origen y destino participantes. El campo de origen correspondiente a la columna de destino [SalesValueUSD] se llama simplemente "SalesValue" - selecciónelo de la lista.

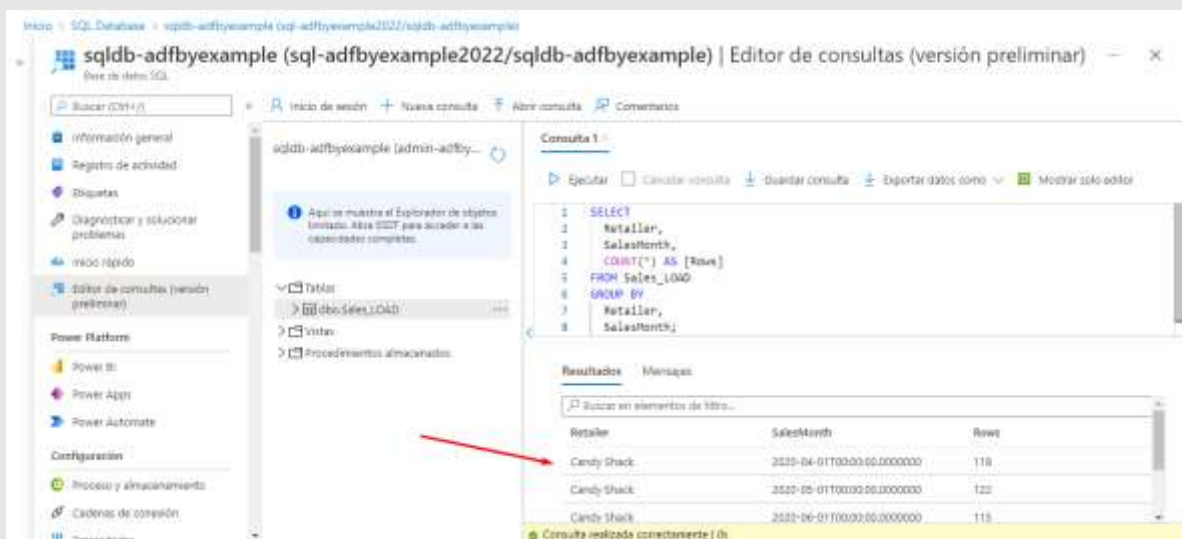


- Haga clic en Depurar para volver a ejecutar la canalización, esta vez con éxito. Utilice el Listado 3-3 en su cliente SQL para verificar que los datos de ventas de seis meses de Candy Shack se han cargado en dbo.Sales\_LOAD. Observe que los datos de Sweet Treats cargados anteriormente han sido eliminados por el script de Pre-copia de la actividad.



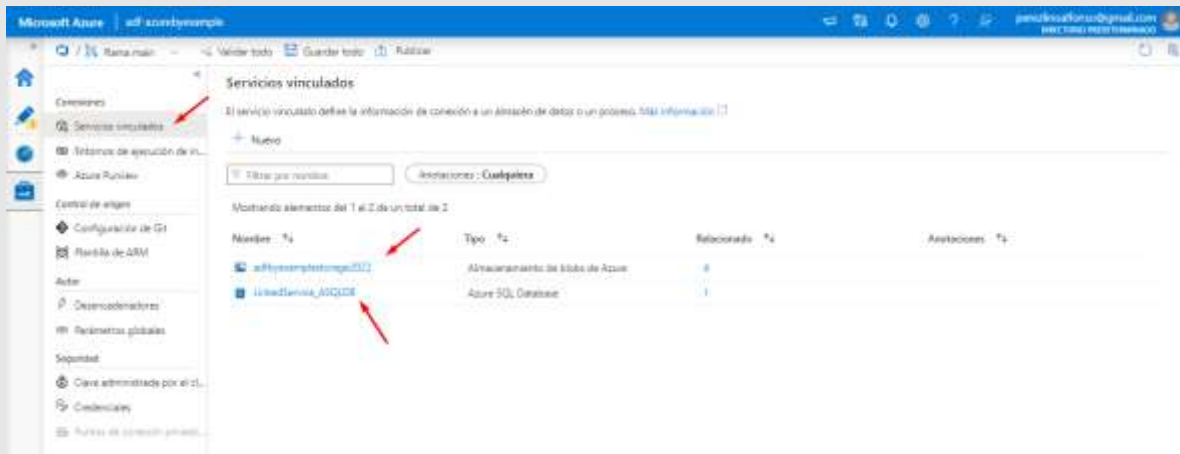


Los datos de Candy Shack reemplazan a los de Sweet Treats cargados anteriormente.





Recordar que nuestros Linked service (Servicios vinculados) son la Cuenta de almacenamiento Blob y la base de datos Azure SQL.



## 3.4. Importar datos semiestructurados a Azure SQL DB

Los archivos de datos estructurados y tabulares, como los archivos CSV descritos en la sección anterior, se asemejan a las tablas de las bases de datos en cuanto a su estructura. En cambio, los formatos de datos no tabulares o semiestructurados utilizan metadatos incrustados para identificar los elementos de datos y suelen contener componentes anidados. Los formatos de datos semiestructurados más comunes son JSON y XML; en esta sección, creará un pipeline utilizando las características de la actividad Copiar datos para manejar datos JSON.

Los datos de muestra para cargar son una colección de informes de datos de ventas mensuales para un minorista de confitería llamado **Sugar Cube**. El formato del informe de ventas de Sugar Cube tiene las siguientes características, como se muestra en el Listado 3-4:

El documento JSON tiene un campo Mes que identifica el mes al que se refiere el informe, un campo Empresa que identifica a Sugar Cube y un campo Ventas que contiene una lista de resúmenes de ventas de productos individuales.

Cada objeto de resumen de ventas de productos identifica el producto por su nombre y el código del fabricante, e informa de la cantidad vendida (Unidades) y de los ingresos totales por ventas (Valor).

```
{
  "Month": "01-apr-2020",
  "Company": "Sugar Cube",
  "Sales": [
    {
      "Product": "Schnoogles 8.81oz",
      "ManufacturerProductCode": "CS-20147-0250",
      "Units": 745,
      "Value": 6995.55
    },
  ],
}
```

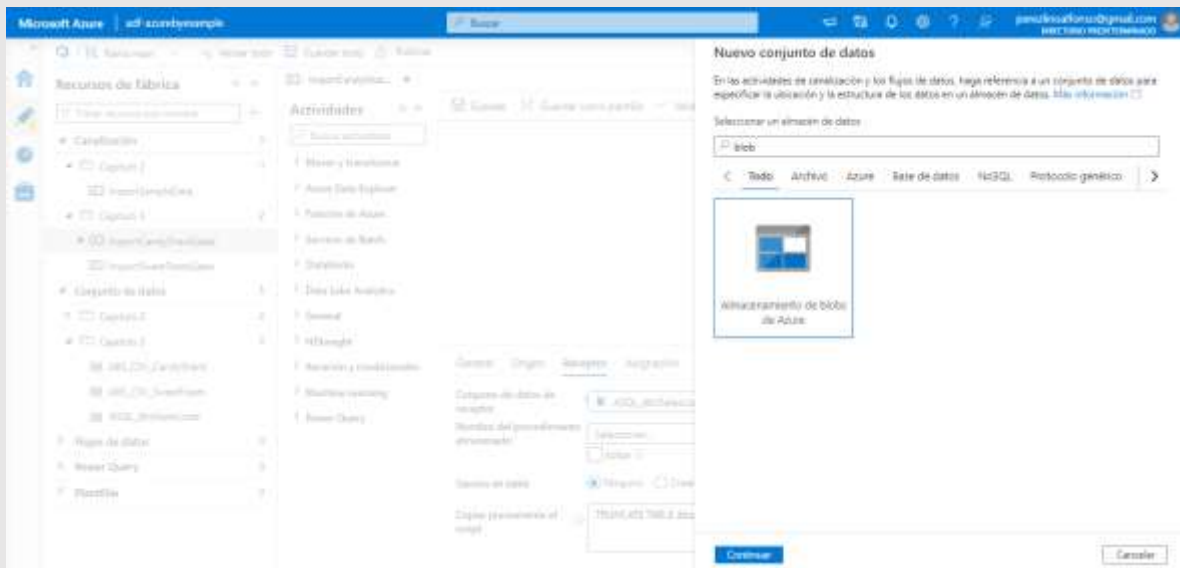
Listado 3-4 Inicio de un archivo de informe de ventas de Sugar Cube

### 3.4.1. Crear un conjunto de datos de archivos JSON

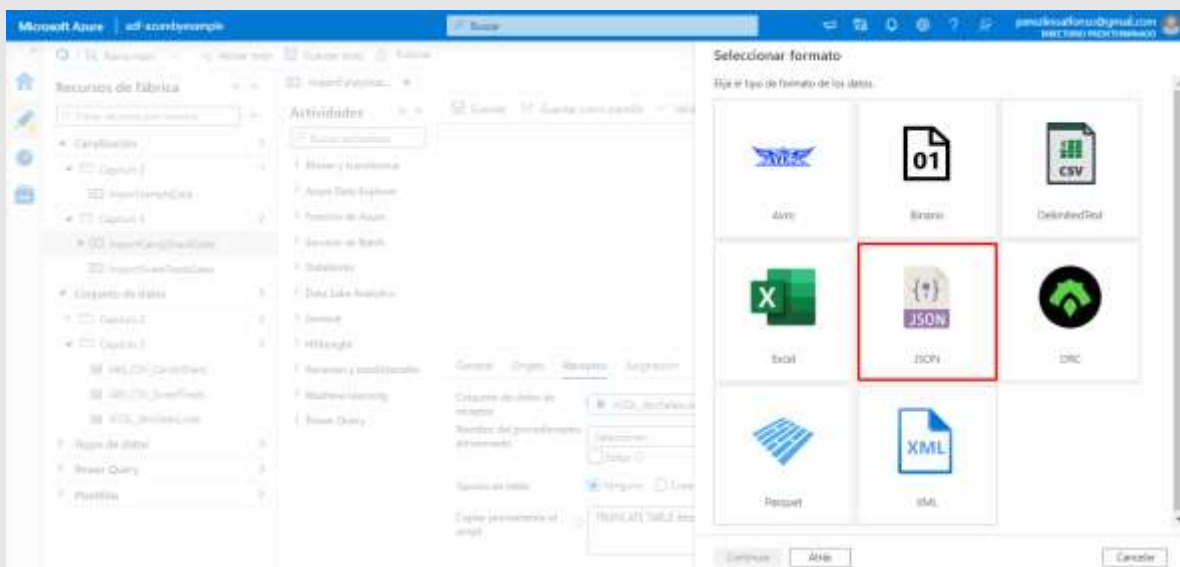
El nuevo pipeline cargará los datos JSON de muestra de su cuenta de almacenamiento blob en Azure SQL DB. En la sección anterior, creó el conjunto de datos "ABS\_CSV\_CandyShack" clonando "ABS\_CSV\_SweetTreats", pero en este caso debe crear un nuevo dataset desde cero. El tipo de archivo de un conjunto de datos sólo puede especificarse cuando se crea el conjunto de datos: para cargar datos JSON, primero debe crear un dataset de almacenamiento blob JSON.

1. En el menú Acciones de la carpeta "Capítulo 3" del explorador de Recursos de la Fábrica, haga clic en Nuevo conjunto de datos.

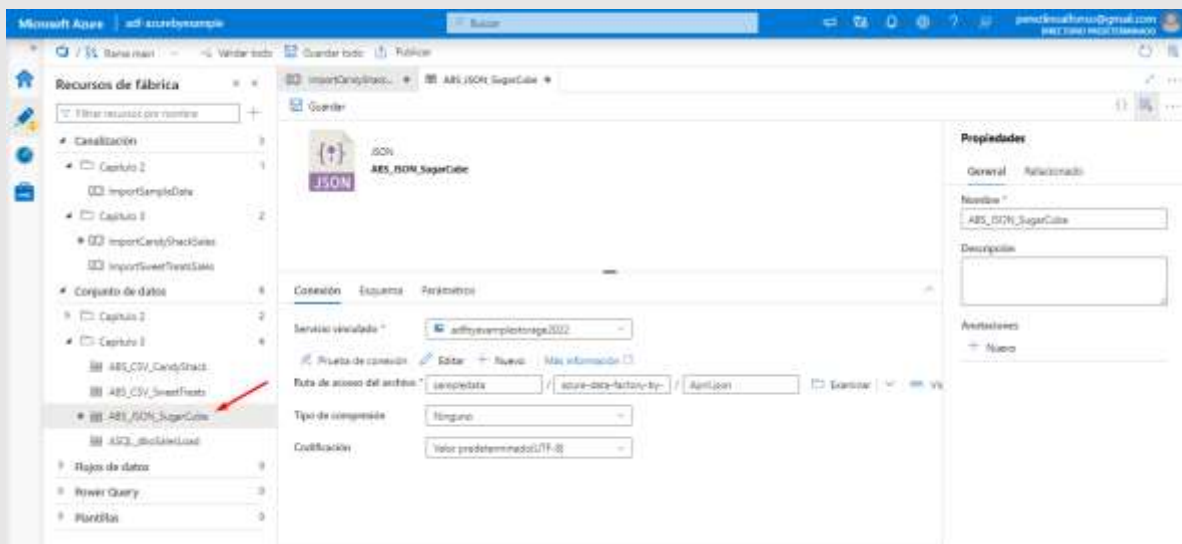
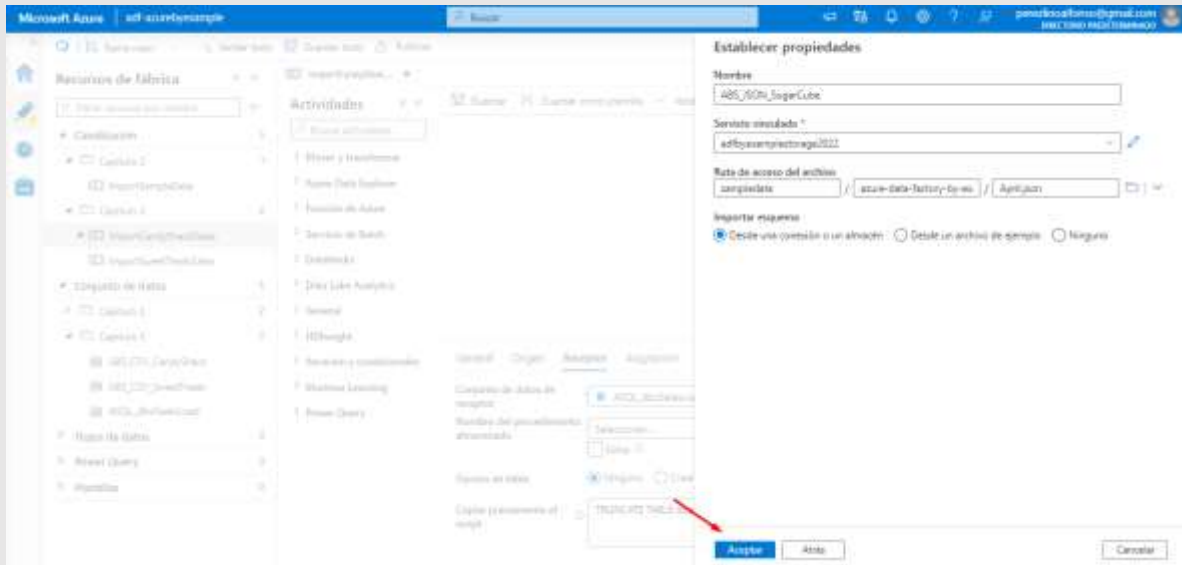
2. Seleccione el almacén de datos Azure Blob Storage y haga clic en Continuar.



3. En la hoja Seleccionar formato, elija Json. Haz clic en Continuar.



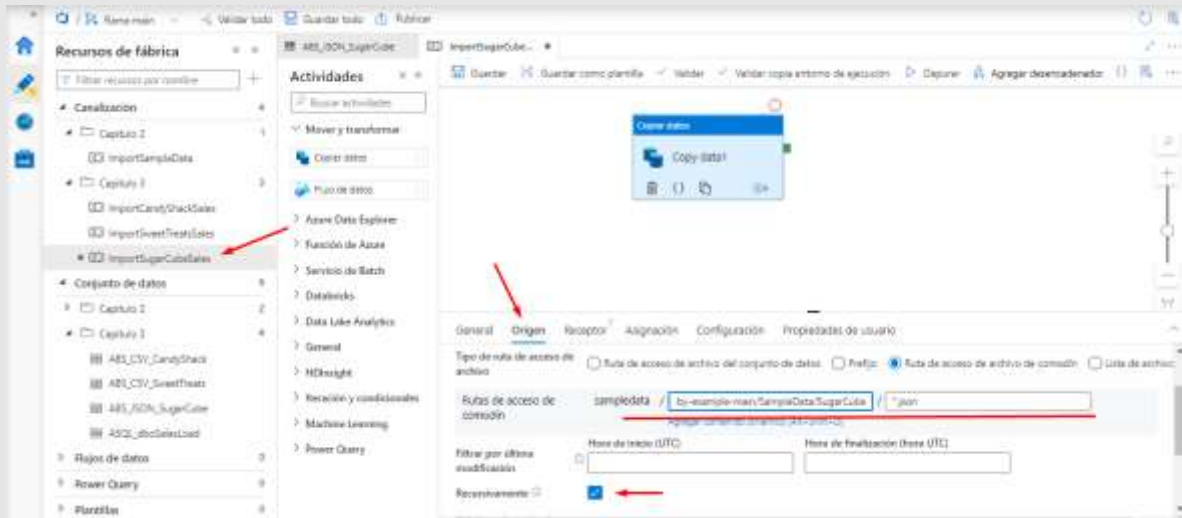
4. Nombre el conjunto de datos "ABS\_JSON\_SugarCube" y seleccione el servicio vinculado a su cuenta de almacenamiento existente. Cuando aparezca la opción de ruta de acceso al archivo, busque la subcarpeta "SugarCube" en el contenedor "sampledata" y encuentre el archivo "April.json". Haz clic en Aceptar para seleccionar el archivo.
5. Haz clic en Aceptar para crear el nuevo conjunto de datos y guárdalo.



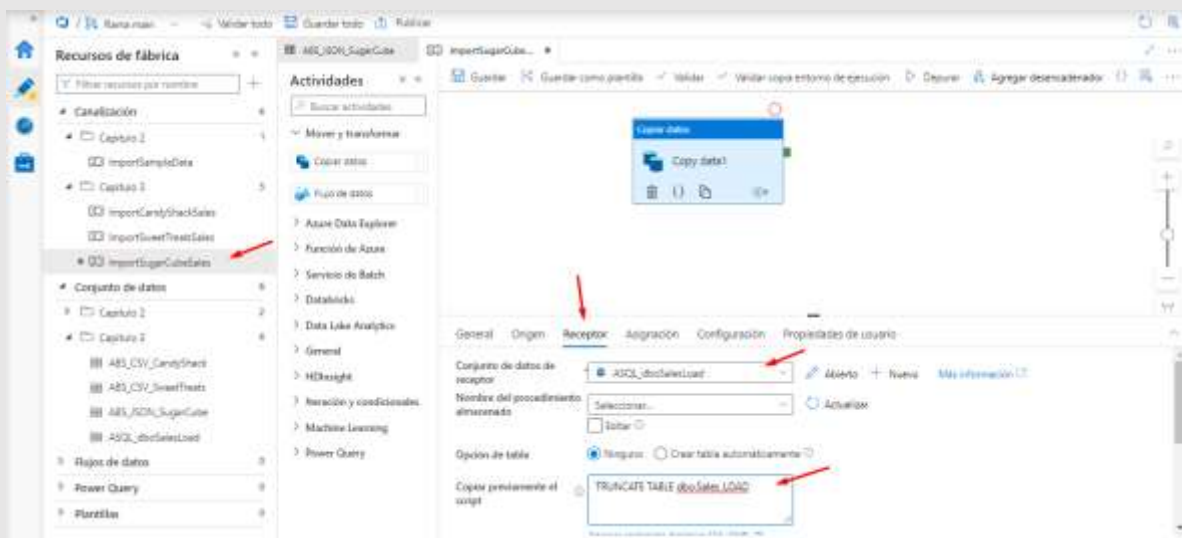
### 3.4.2. Crear el pipeline

Cree un nuevo pipeline llamado "ImportSugarCubeSales", ya sea clonando uno de sus pipelines "Capítulo 3" existentes o desde cero. Añada una actividad de copia de datos al pipeline y configure el origen y el destino de la actividad como se indica a continuación:

**Origen:** Elija el conjunto de datos "ABS\_JSON\_SugarCube" y seleccione el tipo de ruta de archivo "Wildcard file path". En Wildcard paths, establezca la ruta de la carpeta Wildcard en "**azure-data-factory-by-example-main/SampleData/SugarCube**" (omitiendo las subcarpetas de año y trimestre) y especifique "**\*.json**" como nombre de archivo Wildcard. Asegúrese de que la casilla de verificación "**Recursivamente**" esté marcada.



**Receptor:** Elija su conjunto de datos de Azure SQL DB y asegúrese de que el script de precopia esté configurado para truncar la tabla de la base de datos de destino.



### 3.4.3. Configurar la asignación de esquemas

Seleccione la pestaña de configuración de mapeo de la actividad Copiar datos y haga clic en Importar esquemas. La Figura 3-15 muestra el mapeo inicial, generado automáticamente por el ADF UX. **El esquema de origen importado incluye información sobre objetos JSON anidados.** "Sales" se identifica correctamente como una matriz de objetos con cuatro campos: "Producto", "ManufacturerProductCode", "Unidades" y "Valor".

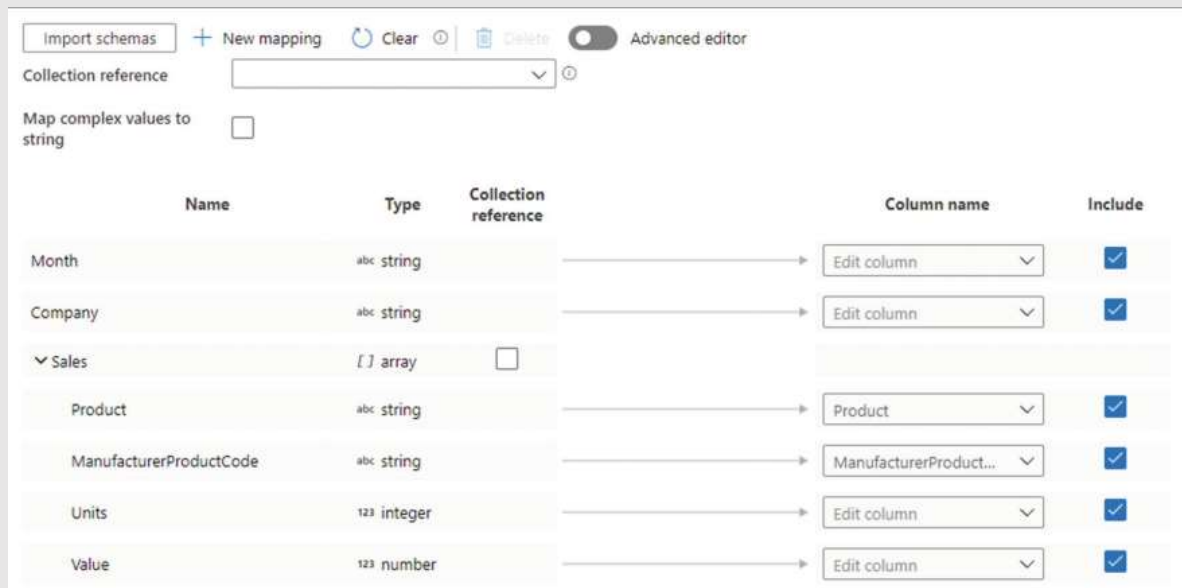
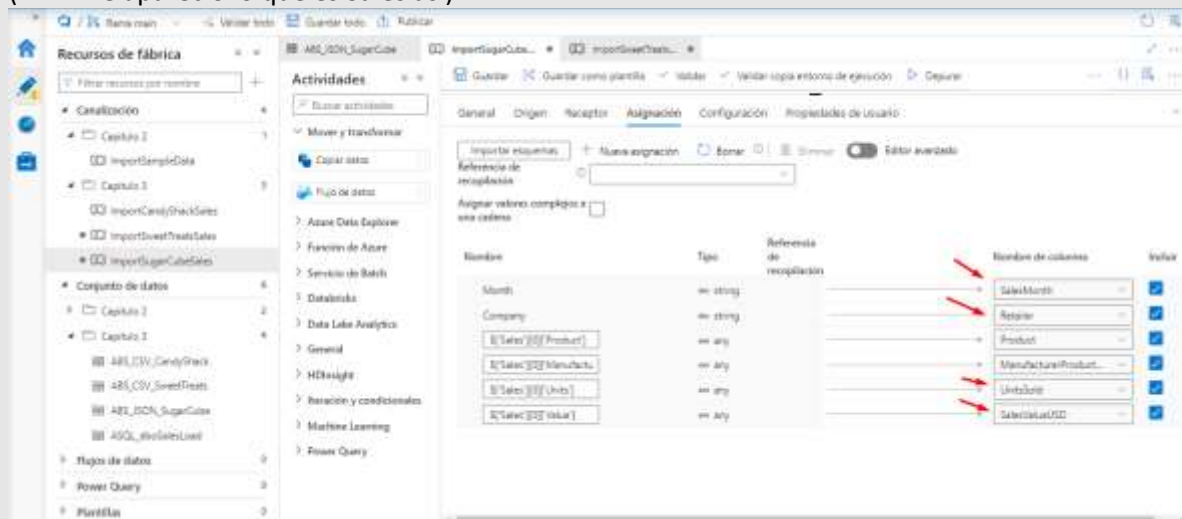


Figura 3-15 Imported Sugar Cube source and database sink schemas

Los campos fuente "Product" y "ManufacturerProductCode" ya han sido asignados a los nombres de columna de destino correspondientes. Asigne los nombres de campo JSON restantes a sus correspondientes columnas de la tabla de la base de datos:

- Month → SalesMonth
- Company → Retailer (empresa)
- Units → UnitsSold (Unidades vendidas)
- Value → SalesValueUSD

(A mi me apareció lo que es Sales así)



Ahora ejecute el pipeline. Cuando inspeccione los resultados en su cliente SQL, descubrirá que ha cargado exactamente seis filas.



Recursos de fábrica

- Canalización
  - Capítulo 2
    - ImportSampleData
    - Capítulo 2
      - ImportCarryShedData
      - ImportSweetHeatData
      - ImportSugarCubeData
- Conjunto de datos
  - Capítulo 2
    - Capítulo 2
      - AB1\_CSV\_CandyShed
      - AB1\_CSV\_SweetHeat
      - AB1\_CSV\_SugarCube
      - AB2\_CSV\_SalesLoad
- Flujos de datos
  - Power Query
  - Flujo de datos

Actividades

- Buscar actividades
- Mover y transformar
  - Copiar datos
  - Flujo de datos
- Asas de datos
- Función de Azure
- Servicio de Batch
- Datos de Azure
- Data Lake Analytics
- General
- HDInsight
- Recación y configuración
- Machine Learning
- Power Query

Copy data1

Parámetros Variables Settings Salida

Id. de ejecución de canalización: 87654321-4567-8901-2345-678901234567

Nombre	Tipo	Inicio de la ejecución	Duración	Estado	Entorno de ejecución de integración
Copy data1	Copiar datos	2022-01-14T12:57:14.333	00:00:10	Completado	AutoDataWarehouseRuntime

Inicio > SQL Database > sqldb-adfbyexample (sql-adfbyexample2022/sqldb-adfbyexample)

sqldb-adfbyexample (sql-adfbyexample2022/sqldb-adfbyexample) | Editor de consultas (versión preliminar)

Base de datos SQL

Buscar (Ctrl+F)

Inicio de sesión + Nueva consulta + Abrir consulta + Comentarios

Consulta 1

```

1 SELECT
2   Retailer,
3   SalesMonth,
4   COUNT(*) AS [Rows]
5 FROM Sales_LOAD
6 GROUP BY
7   Retailer,
8   SalesMonth
  
```

Resultados Mensajes

Retailer	SalesMonth	Rows
Sugar Cube	2020-04-01T00:00:00.0000000	1
Sugar Cube	2020-05-01T00:00:00.0000000	1
Sugar Cube	2020-06-01T00:00:00.0000000	1
Sugar Cube	2020-07-01T00:00:00.0000000	1
Sugar Cube	2020-08-01T00:00:00.0000000	1
Sugar Cube	2020-09-01T00:00:00.0000000	1

Consulta realizada correctamente | 0s

Inicio > SQL Database > sqldb-adfbyexample (sql-adfbyexample2022/sqldb-adfbyexample)

sqldb-adfbyexample (sql-adfbyexample2022/sqldb-adfbyexample) | Editor de consultas (versión preliminar)

Base de datos SQL

Buscar (Ctrl+F)

Inicio de sesión + Nueva consulta + Abrir consulta + Comentarios

Consulta 1

Resultados Mensajes

Buscar en elementos de filtro...

Retailer	SalesMonth	Rows
Sugar Cube	2020-04-01T00:00:00.0000000	1
Sugar Cube	2020-05-01T00:00:00.0000000	1
Sugar Cube	2020-06-01T00:00:00.0000000	1
Sugar Cube	2020-07-01T00:00:00.0000000	1
Sugar Cube	2020-08-01T00:00:00.0000000	1
Sugar Cube	2020-09-01T00:00:00.0000000	1

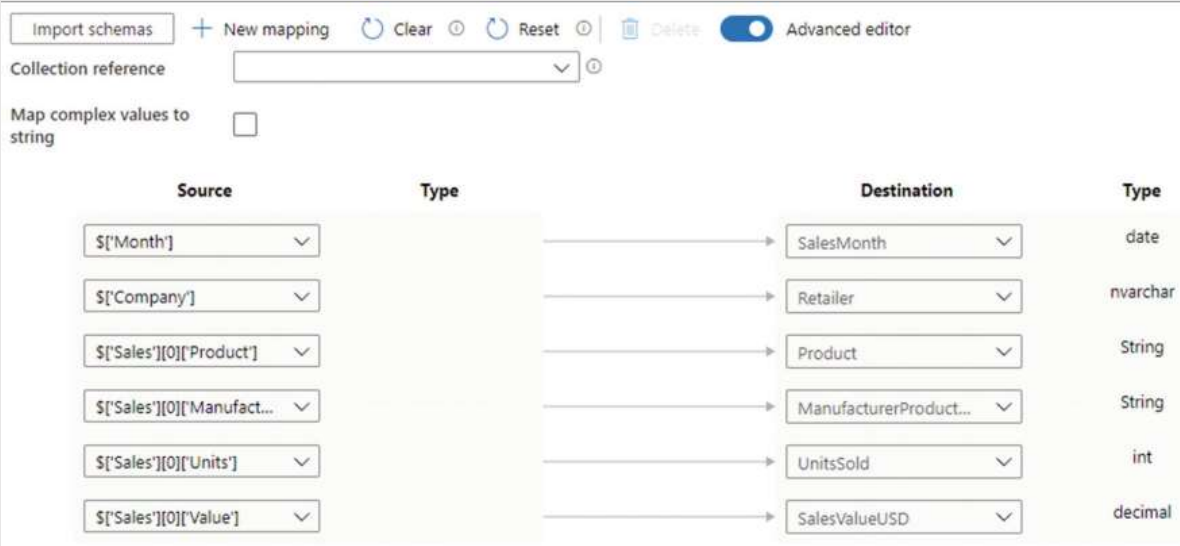
### 3.4.4. Establezca la referencia de la colección

La estructura anidada de los objetos JSON les permite representar estructuras de datos normalizadas. El formato de informe de ventas de Sugar Cube contiene las siguientes relaciones:

- Cada archivo de informe contiene exactamente un documento JSON de informe de ventas mensual.
- Cada informe de ventas mensual contiene muchos resúmenes de ventas de productos.

La propiedad **Referencia de la recopilación (Collection reference)** en la pestaña de Asignación de la actividad Copiar datos controla lo que representa una fila de datos de salida: en este caso, un resumen de ventas de productos individual o todo el documento del informe mensual.

Por defecto, la referencia de recopilación es la raíz del documento JSON. Las seis filas cargadas al ejecutar el pipeline corresponden a los seis archivos de informe del Sugar Cube, pero notarás que algunos datos de nivel inferior (producto, detalles de ventas) también fueron cargados en las seis filas. Para entender esto, activa el editor avanzado en la pestaña de Asignación (Figura 3-16) - esto muestra las expresiones de ruta JSON subyacentes utilizadas para extraer los valores de los campos de salida.



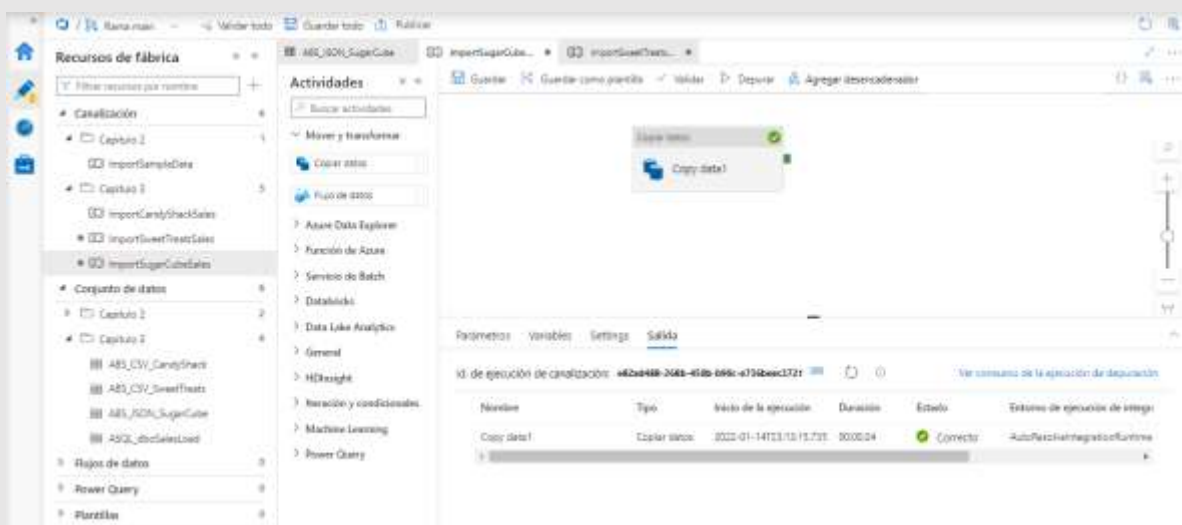
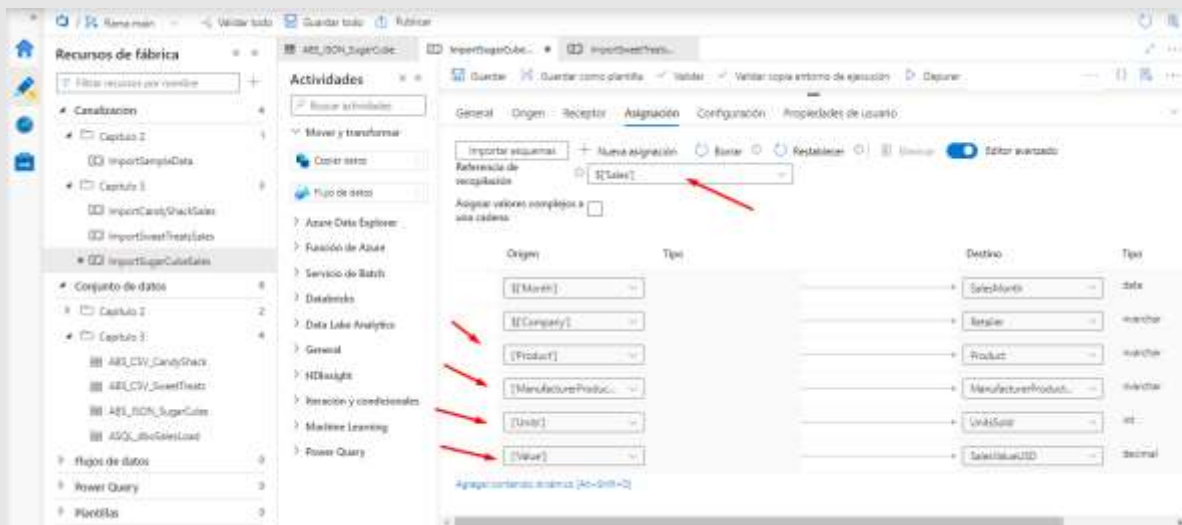
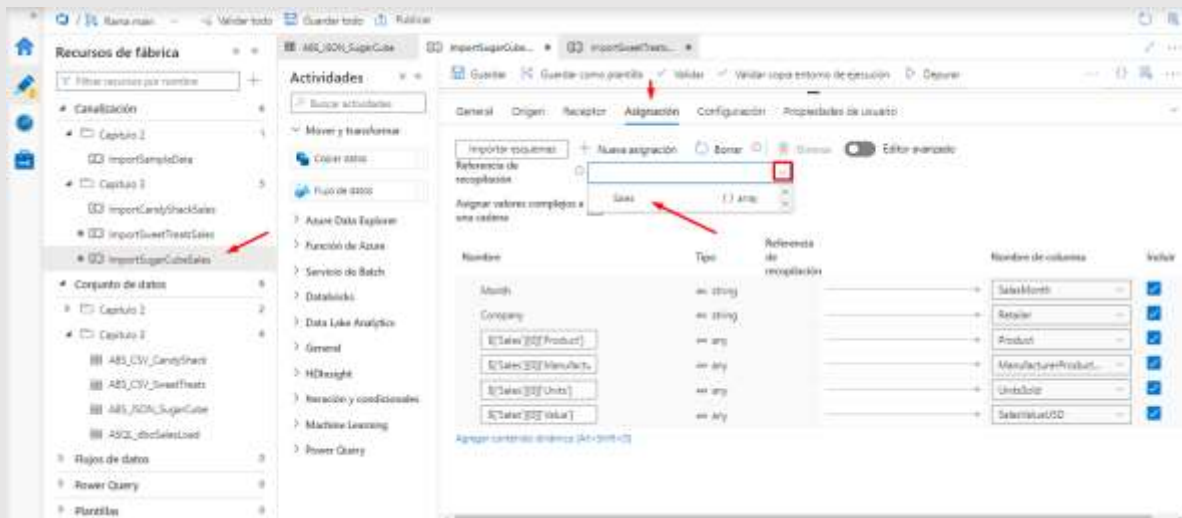
The screenshot shows the 'Advanced editor' tab in the mapping tool. An orange arrow points to the 'Collection reference' dropdown menu. Below it, a table lists six mappings from source JSON paths to destination fields.

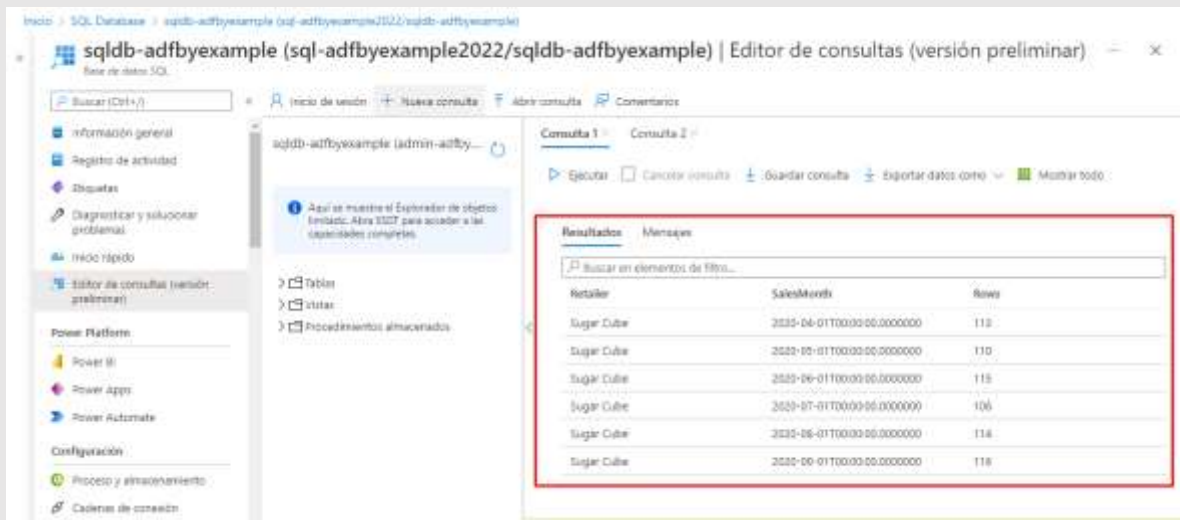
Source	Type	Destination	Type
<code>['Month']</code>		SalesMonth	date
<code>['Company']</code>		Retailer	nvarchar
<code>['Sales'][0]['Product']</code>		Product	String
<code>['Sales'][0]['Manufact...']</code>		ManufacturerProduct...	String
<code>['Sales'][0]['Units']</code>		UnitsSold	int
<code>['Sales'][0]['Value']</code>		SalesValueUSD	decimal

Figura 3-16 Editor de mapeo avanzado que muestra las expresiones de ruta JSON

La expresión para el campo "UnitsSold" es `['Sales'][0]['Units']`, lo que significa "utilizar el campo Units del primer elemento de resumen de ventas de productos de la matriz Sales".

Para especificar la referencia de recopilación correcta, desactive el editor avanzado y marque la casilla de verificación de la fila "Ventas" (o seleccione "Sales" en el menú desplegable de la referencia de recopilación situado encima de la lista de asignaciones de columnas). Vuelva a ejecutar el pipeline y, a continuación, utilice su cliente SQL para verificar que se han cargado entre 100 y 120 filas para cada mes.



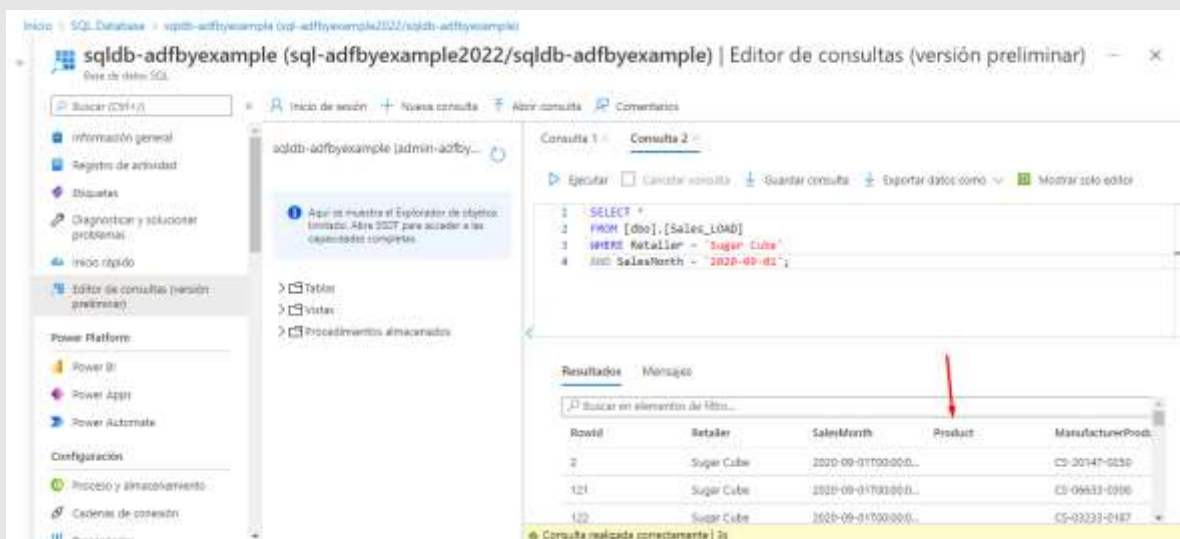


### 3.4.5. El efecto de la desviación del esquema

Utilice su cliente SQL para ver los datos del Sugar Cube cargados para septiembre de 2020 (utilizando una consulta como la que se muestra en el Listado 3-5).

```
SELECT *
FROM [dbo].[Sales_LOAD]
WHERE Retailer = 'Sugar Cube'
AND SalesMonth = '2020-09-01';
```

Listado 3-5 Seleccionar los datos del Sugar Cube de septiembre



Base de datos SQL

sqldb-adfbyexample (sql-adfbyexample2022/sqldb-adfbyexample) | Editor de consultas (versión preliminar)

Base de datos SQL

Información general  
Registro de actividad  
Etiquetas  
Diagnosticar y solucionar problemas  
Inicio rápido

Editor de consultas (versión preliminar)

Power Platform  
Power BI  
Power Apps  
Power Automate

Configuración  
Proceso y almacenamiento  
Cadenas de conexión

Inicio de sesión | Nueva consulta | Abrir consulta | Comentarios

Consulta 1 | Consulta 2

Ejecutar | Cancelar consulta | Guardar consulta | Exportar datos como | Mostrar todo

Resultados | Mensajes

Buscar en elementos de filtro...

Rowid	Retailer	SalesMonth	Product	ManufacturerProd.
2	Sugar Cube	2020-09-01T00:00:00...		CS-20147-0250
121	Sugar Cube	2020-09-01T00:00:00...		CS-06893-0300
122	Sugar Cube	2020-09-01T00:00:00...		CS-03233-0167
123	Sugar Cube	2020-09-01T00:00:00...		CS-12166-0055
124	Sugar Cube	2020-09-01T00:00:00...		CS-19197-0300
125	Sugar Cube	2020-09-01T00:00:00...		CS-04960-0230
126	Sugar Cube	2020-09-01T00:00:00...		CS-21792-0128
127	Sugar Cube	2020-09-01T00:00:00...		CS-15134-0800

Consulta realizada correctamente | 20

Una inspección detallada de los resultados de la consulta revela que todos los datos de septiembre carecen de un valor de nombre en la columna [Producto] de la tabla de la base de datos. Esto ha sido causado por un pequeño cambio en la estructura JSON del informe de ventas, introducido en ese mes. El listado 3-6 muestra el inicio del archivo de septiembre, en el que se puede ver que el campo del nombre del producto, **antes llamado "Product", se llama ahora "Item"**.

```
{
  "Month": "01-sep-2020",
  "Company": "Sugar Cube",
  "Sales": [
    {
      "Item": "Schnoogles 8.81oz",
      "ManufacturerProductCode": "CS-20147-0250",
      "Units": 643,
      "Value": 6745.07
    },
  ],
}
```

Listado 3-6 Inicio del archivo del informe de ventas del azucarero de septiembre de 2020

El resultado de cambiar el nombre del campo "Product" es que parece que simplemente se ha omitido, lo que el ADF ha aceptado sin error. El manejo de JSON por parte de la actividad de copia de datos es necesariamente tolerante con los campos omitidos, porque los elementos opcionales pueden no estar siempre presentes en un objeto JSON.

No hay una solución sencilla en este caso, pero ilustra el hecho de que **no se puede confiar en los mapeos del esquema JSON para detectar la deriva del esquema**. A diferencia de lo que ocurre cuando se carga un archivo de datos estructurados, la desviación del esquema no puede provocar un fallo. En el Capítulo 6, se añadirá una comprobación básica de la calidad de los datos cargados para mejorar la solidez de la cadena.

### 3.4.6. Comprender la conversión de tipos

Cuando se mueven o transforman datos, las tareas de integración deben mediar entre las diferencias en los sistemas de tipos utilizados por los almacenes de datos de origen y de destino. Los tipos de campo de origen mostrados en la Figura 3-15 son tipos de datos JSON, mientras que los tipos de campo de destino correspondientes, visibles en la Figura 3-16, son nativos de SQL Server. La conversión de datos entre estos dos sistemas de tipos tiene lugar a través de un tercero: **los tipos de datos interinos (IDT: interim data types)** de Azure Data Factory.

La actividad Copiar datos realiza una conversión de tipos en tres pasos para traducir los datos de un sistema de tipos fuente al sistema de tipos receptores:

1. Convertir el tipo de origen nativo en un tipo interino de ADF.
2. Convertir el tipo interino en un tipo interino posiblemente diferente compatible con el tipo de receptor.
3. Convertir el tipo interino compatible con el receptor en un tipo receptor nativo.

Este conocimiento no es esencial para utilizar la actividad de copia de datos, pero puede ayudarle a comprender el comportamiento de las asignaciones de esquemas entre diferentes formatos. **Por ejemplo, los archivos CSV no tienen un sistema de tipos consistente, por lo que ADF trata todos los campos entrantes como strings y los convierte automáticamente al tipo de datos interino String** (como puede verse en los tipos de origen mostrados en la Figura 3-14).

Los IDTs permiten a Azure Data Factory soportar una gama cada vez mayor de almacenes de datos soportados de forma escalable. Sin los IDT, añadir soporte para un nuevo tipo de almacén de datos requeriría que Microsoft especificara conversiones de tipo entre el nuevo almacén y cualquier otro tipo de almacén de datos existente. Este requisito crecería en tamaño con cada adición a la lista de almacenes de datos soportados. El uso de IDTs significa que, para extender el soporte a nuevos tipos de almacenes de datos, ADF sólo necesita ser capaz de convertir los tipos de datos del nuevo almacén a y desde su propio sistema de tipos de datos interinos.

**Para los desarrolladores de SSIS:** el enfoque de ADF para mediar en las conversiones de tipos a través de un sistema de tipos intermedios les resultará familiar por las SSIS Data Flow Tasks. Los tipos de datos de Integration Services (por ejemplo, DT\_STR, DT\_WSTR y DT\_I4) desempeñan el mismo papel en SSIS, proporcionando un soporte extensible para emparejamientos arbitrarios de sistemas de tipos de origen y destino.



### 3.5. Transformar archivos JSON en Parquet

Hasta ahora, en este capítulo, ha utilizado la capacidad de ADF para leer datos de archivos CSV y JSON y su capacidad para escribir datos en tablas de Azure SQL DB. El hecho de que el origen y el destino de la actividad Copiar datos se describan utilizando conjuntos de datos de ADF significa que puede extraer datos de una base de datos SQL y escribirlos en archivos CSV o JSON con la misma facilidad. **Del mismo modo, se puede utilizar la actividad Copiar datos para leer datos de archivos CSV y escribirlos como JSON - o de hecho para leer datos de cualquier tipo de conjunto de datos soportado y escribirlos en cualquier otro.**

Apache Parquet es un formato de almacenamiento de datos estructurados comprimido y orientado a columnas. Las aplicaciones analíticas que procesan un gran número de filas se benefician de un formato orientado a columnas, ya que no es necesario leer los datos de las columnas que están fuera del ámbito del análisis. Las columnas individuales a menudo contienen unos pocos valores distintos, lo que permite que los datos de las columnas se compriman en gran medida para un almacenamiento y una recuperación eficientes. Por estas razones, Parquet suele ser el formato elegido para un data lake storage de datos tabulares.

El minorista de confitería Handy Candy informa de los datos de ventas en forma de mensajes JSON de transacciones de ventas individuales. El listado 3-7 contiene uno de estos objetos de mensaje. En esta sección, usted creará un pipeline para ingerir estos mensajes y darles salida como Parquet. Este patrón de ingesta se utiliza con frecuencia para integrar nuevos datos en un data lake, pero en este caso simplemente emitirá el archivo Parquet al almacenamiento blob de Azure.

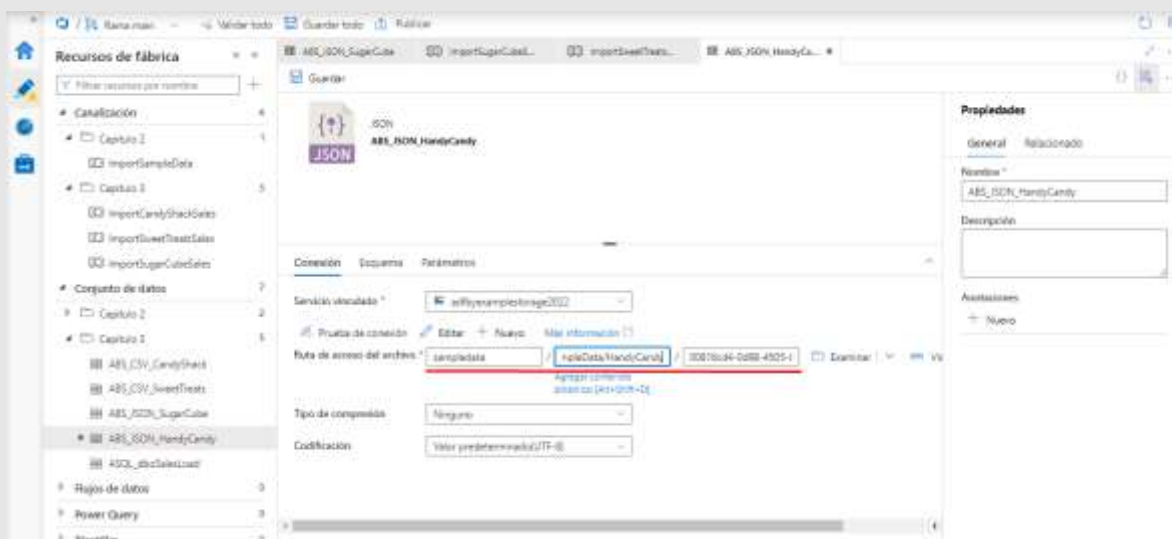
```
{
  "TransactionId": "0C90CC54-392B-4322-BB23-B4B34CE403D9",
  "TransactionDate": "2020-06-08",
  "StoreId": "114",
  "Items": [
    {
      "Producto": "Boho 2.82oz",
      "Precio": 2.89
    }
  ]
}
```

Listado 3-7 Mensaje de transacción de venta de Candy

### 3.5.1. Crear un nuevo conjunto de datos JSON

Para poder importar el esquema del archivo de mensajes de Handy Candy, necesita un conjunto de datos JSON que haga referencia a un archivo de mensajes en la subcarpeta "HandyCandy" de la carpeta SampleData.

1. Cree un nuevo conjunto de datos JSON utilizando su cuenta de almacenamiento blob o clone "ABS\_JSON\_SugarCube". Nombre el nuevo conjunto de datos "ABS\_JSON\_HandyCandy".
2. Establezca la ruta del archivo del conjunto de datos seleccionando uno de los archivos de mensajes en la subcarpeta "HandyCandy" del contenedor "sampledata".



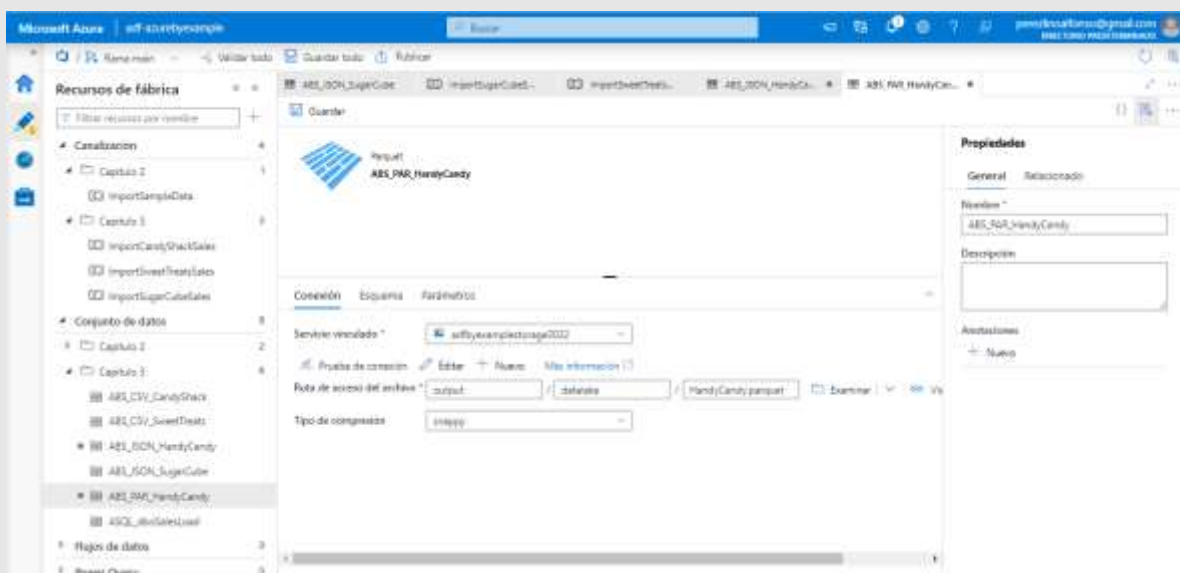
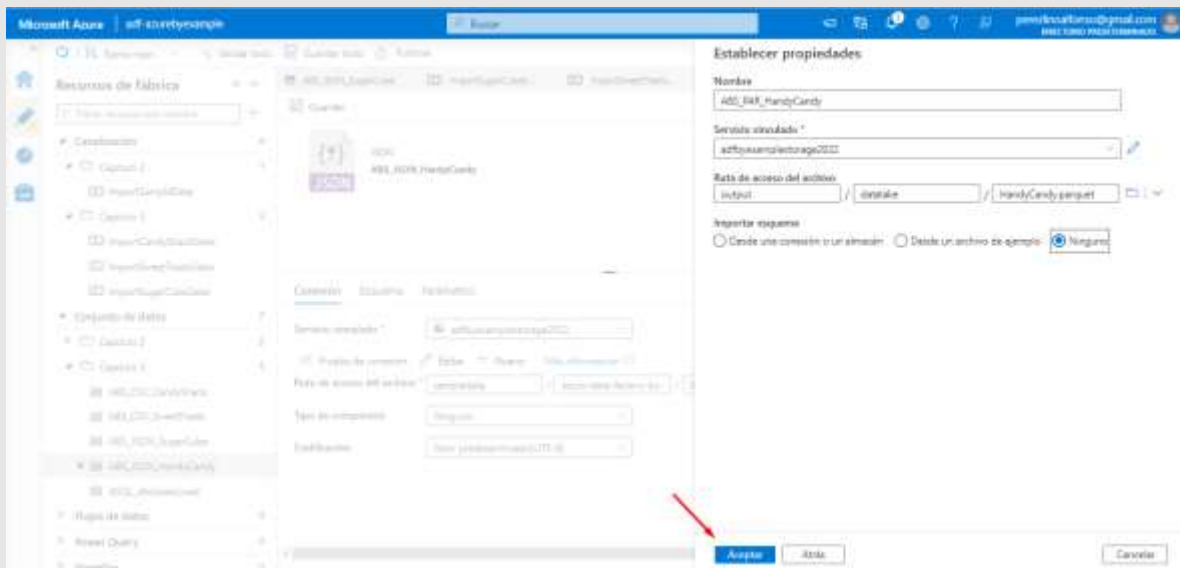
### 3.5.2. Crear un conjunto de datos Parquet

Cree un nuevo conjunto de datos en la carpeta de conjuntos de datos "Capítulo 3" con las siguientes propiedades:

- **Data store:** Azure Blob Storage.
- **Formato de archivo:** Parquet.
- **Nombre:** Introduzca "ABS\_PAR\_HandyCandy".
- **Servicio vinculado:** Elija su servicio vinculado de almacenamiento de blob existente.
- **Ruta del archivo:** Especifique el contenedor "output", el directorio "datalake" y el archivo "HandyCandy.parquet".

Antes de hacer clic en Aceptar en la hoja de propiedades, asegúrese de que el **esquema de importación** esté establecido en "Ninguno".

**Consejo** Si no existe ningún contenedor de "salida" cuando se ejecuta el pipeline, éste creará uno automáticamente.

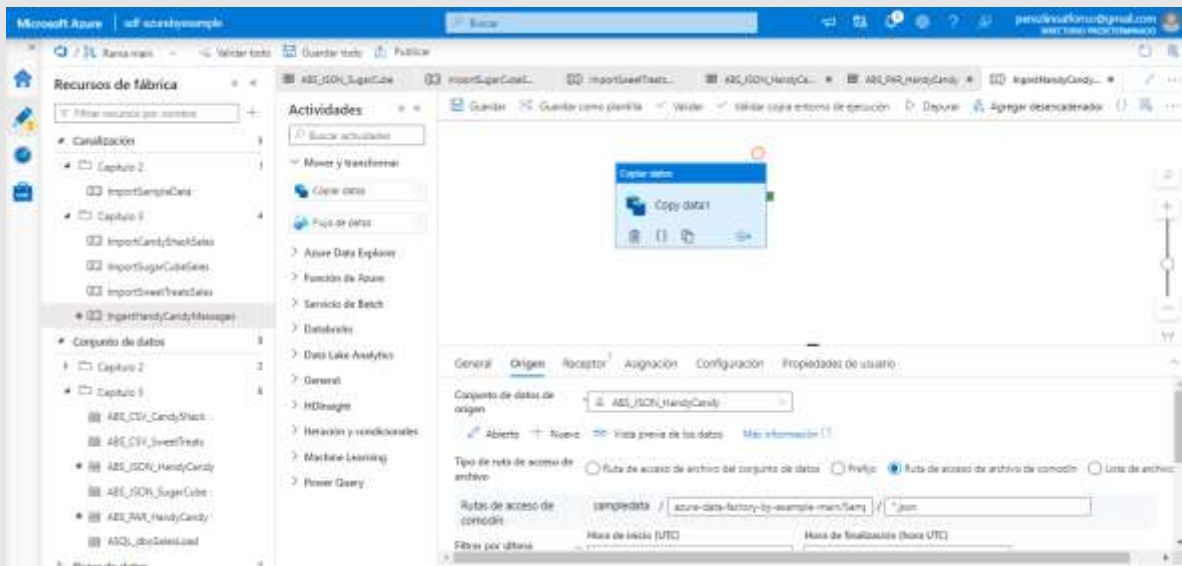


### 3.5.3. Cree y ejecute el pipeline de transformación

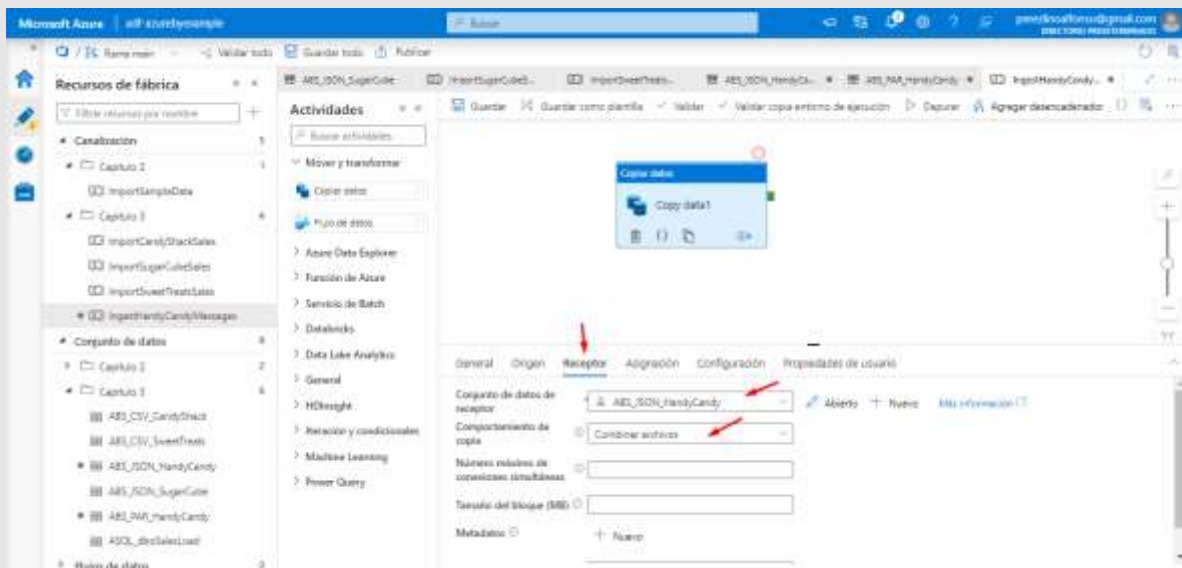
Cree un nuevo pipeline llamado "IngestHandyCandyMessages" en la carpeta de pipelines "Capítulo 3", luego arrastre una actividad de Copiar datos en el lienzo de creación. Configure la actividad como sigue:

**Origen:** Elija el conjunto de datos "ABS\_JSON\_HandyCandy" y seleccione el tipo de ruta de archivo "Wildcard file path". En Rutas comodín, establezca la ruta de la carpeta comodín como "**azure-**

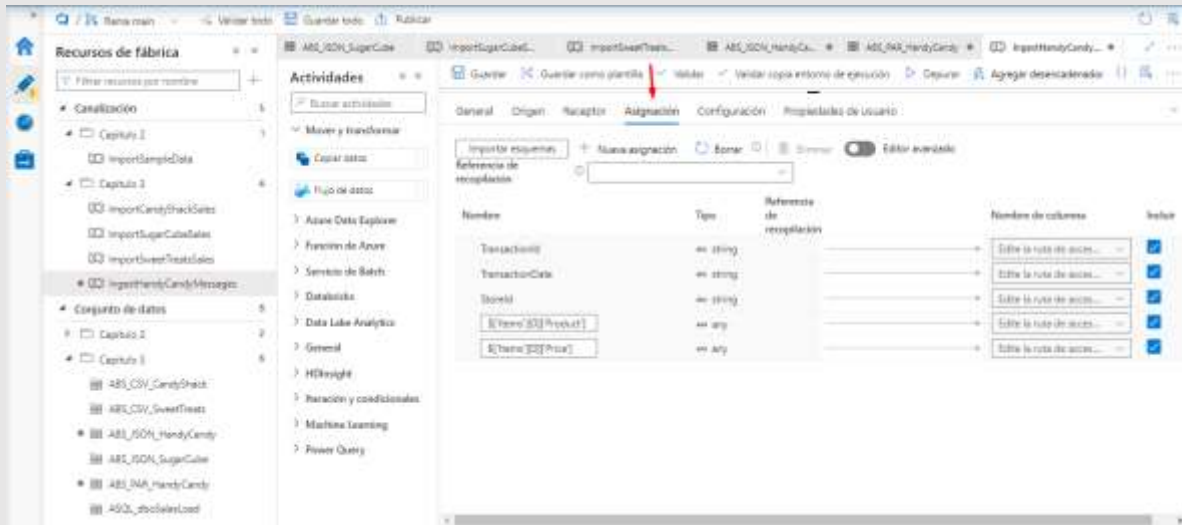
data-factory-by-example-main/SampleData/HandyCandy" y el nombre del archivo comodín como "\*.json".



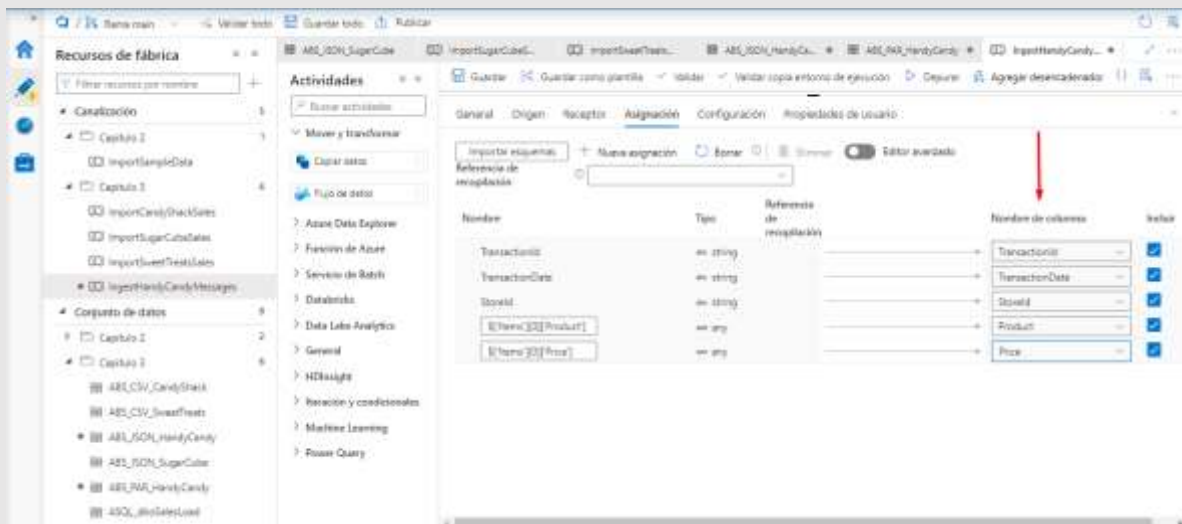
**Receptor:** Elija el conjunto de datos "ABS\_PAR\_HandyCandy" y establezca el **comportamiento de copia** como **"Combinar archivos (Merge files)"**. Esto indica al ADF que combine los archivos de mensajes JSON entrantes en un único archivo de salida Parquet.

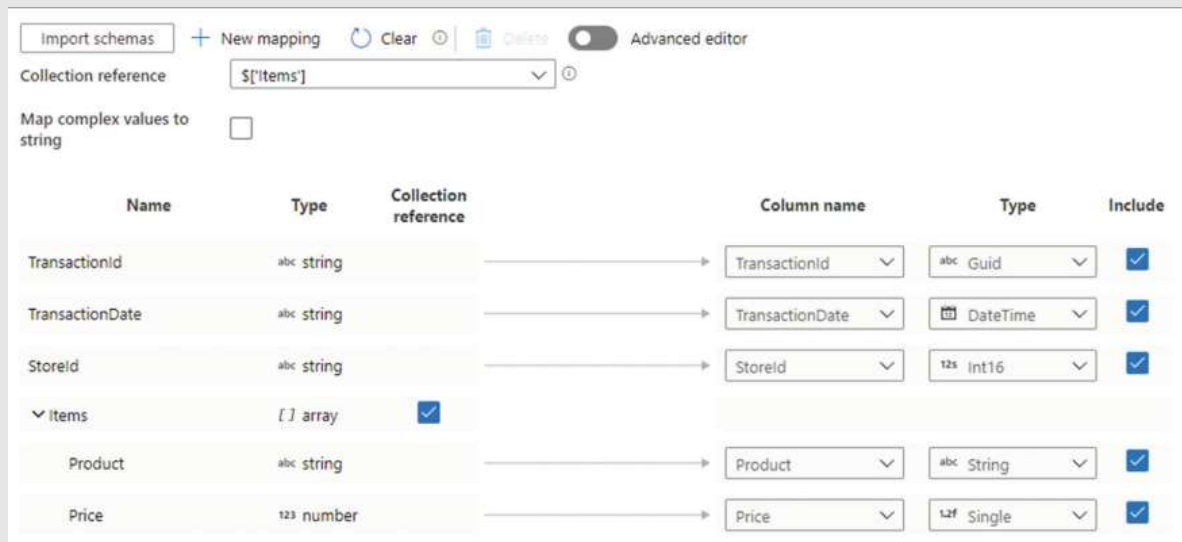


**Asignación:** En la pestaña Asignación, haga clic en **Importar esquemas**. Compruebe que el esquema importado coincide con la estructura de mensajes mostrada en el Listado 3-7. Seleccione el campo **"Items"** como **referencia de recopilación**, y observe que el ADF UX ha proporcionado automáticamente los nombres de los campos para el nuevo archivo - no es necesario realizar más cambios a menos que quiera refinar los nombres o tipos de los campos del fregadero. La Figura 3-17 muestra una asignación actualizada en la que se han conservado los nombres de campo automáticos pero se ha proporcionado información adicional sobre el tipo.



Escribi de manera manual los nombres de columnas y no me aparecieron los tipos de dato de las columnas destino.

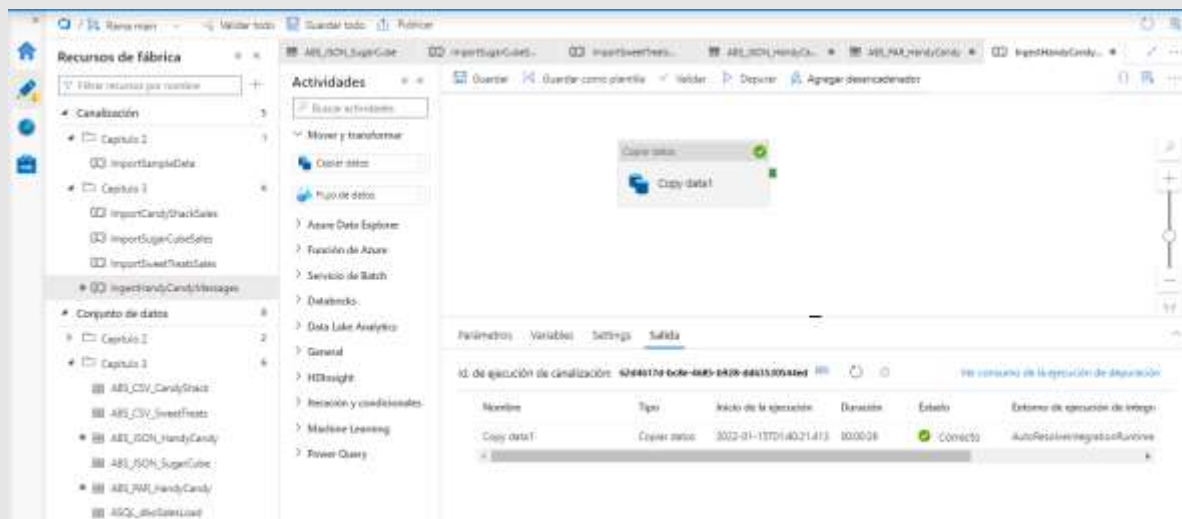




Name	Type	Collection reference	Column name	Type	Include
TransactionId	abc string		TransactionId	abc Guid	<input checked="" type="checkbox"/>
TransactionDate	abc string		TransactionDate	📅 DateTime	<input checked="" type="checkbox"/>
StoreId	abc string		StoreId	12s Int16	<input checked="" type="checkbox"/>
▼ Items	[ ] array	<input checked="" type="checkbox"/>			
Product	abc string		Product	abc String	<input checked="" type="checkbox"/>
Price	12s number		Price	12f Single	<input checked="" type="checkbox"/>

Figura 3-17 Mapeo preparado para la transformación de JSON a Parquet

**Ejecute el pipeline.** Cuando la ejecución esté completa, inspeccione el contenedor de salida de su cuenta de almacenamiento para verificar que el archivo Parquet está presente. La compresión de los datos de las columnas significa que el archivo de salida es aproximadamente un tercio del tamaño total de los archivos JSON de entrada.



Actividades: Copiar datos

Copy data1

Id. de ejecución de canalización: 62644174-b04e-4483-8428-6443330541e1

Nombre	Tipo	Inicio de la ejecución	Duración	Estado	Extremo de ejecución de integración
Copy data1	Copiar datos	2022-01-11T01:40:21.413	00:00:08	CÓRRECTO	AutoRelatividadIntegración

Si desea verificar el contenido del archivo Parquet, cree un pipeline que contenga una actividad de copia de datos con el conjunto de datos "ABS\_PAR\_HandyCandy" como origen. Puede inspeccionar los datos de origen copiándolos en una tabla de base de datos o en un receptor de archivos CSV o utilizando la opción Vista previa de datos en la pestaña de configuración de origen de la actividad Copiar datos.



## 3.6. Configuración de rendimiento

Además de las pestañas que ya ha explorado, el panel de configuración de la actividad Copiar datos presenta una pestaña de Configuración. Dos de estos ajustes están relacionados con el rendimiento de la actividad; los demás se tratan más adelante en el libro.

### 3.6.1. Unidad de integración de datos

El enfoque principal de este capítulo ha sido cómo copiar datos utilizando la actividad Copiar datos, sin prestar mucha atención al rendimiento. Esto no ha sido un problema porque los conjuntos de datos de muestra son pequeños, pero **cuando se trabaja con conjuntos de datos más grandes del mundo real, se puede mejorar el rendimiento de la copia ajustando las características de rendimiento predeterminadas.**

La potencia de la actividad de copia de datos se mide en **unidades de integración de datos (DIU)**, una medida única que combina el uso de la CPU, la memoria y la red. Puede aumentar la potencia asignada a una actividad de datos de copia incrementando el número de DIUs disponibles para ella, especificado mediante la opción Unidad de integración de datos en la pestaña Configuración.

**La configuración por defecto de la unidad de integración de datos es "Auto"**, lo que significa que el número de DIUs asignadas se determina automáticamente, basándose en la información sobre sus almacenes de datos de origen y de destino. Si lo desea, puede aumentar el número de DIUs asignadas por encima del valor por defecto, pero tenga en cuenta que al hacerlo aumentará el coste financiero de la ejecución de la actividad. **Por el contrario, el valor por defecto de los DIUs para muchos escenarios es de cuatro, por lo que puede querer reducir su coste de ejecución - particularmente en escenarios de aprendizaje y desarrollo - limitando el número de DIUs a su valor mínimo de dos.**

Puedes ver el número de DIUs utilizados por la ejecución de una actividad de copia de datos en el campo `usedDataIntegrationUnits` de la salida JSON de la ejecución (como se muestra en la Figura 3-9). Puedes encontrar más información sobre las DIUs y el rendimiento de la actividad de datos Copy en <https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance-features>.

### 3.6.2. Grado de Paralelismo de Copia

La opción Grado de paralelismo de la copia permite anular el grado de paralelismo por defecto de la actividad de datos de copia. El grado de paralelismo realmente utilizado en tiempo de ejecución aparece en el campo `usedParallelCopies` de la salida JSON de la ejecución. Puede ver cómo esto varía incluso dentro de sus pipelines de datos de muestra: el pipeline de ingestión de Handy Candy tiene que procesar muchos más archivos y muestra un mayor grado de paralelismo. Este tipo de comportamiento de escalado automático es una ventaja clave de los servicios sin servidor como Azure Data Factory.

El grado de paralelismo por defecto también se determina automáticamente y varía en función de la información sobre sus almacenes de datos de origen y de destino. Aunque puede parecer tentador anular el valor predeterminado, **el consejo de Microsoft es que el mejor rendimiento de los datos se suele conseguir con el comportamiento predeterminado**. En cuanto a los DIUs, un caso de uso más común podría ser el de limitar el paralelismo, para evitar la sobrecarga de un almacén de datos de origen.

## Revisión del capítulo

Al principio de este capítulo, describí la actividad Copiar datos como la herramienta principal para el movimiento de datos en Azure Data Factory. Si bien esto es cierto, ahora habrá obtenido una apreciación de algunas de las capacidades de transformación de datos que también proporciona la actividad, permitiéndole convertir conjuntos de datos entre formatos de almacenamiento.

La conversión entre formatos de datos tiene tres requisitos:

- La capacidad de leer datos de un formato de origen
- La capacidad de escribir datos en un formato de destino
- Soporte para asignar elementos de los datos de origen al formato de recepción

Azure Data Factory proporciona funcionalidad de lectura/escritura para un gran número de formatos y servicios de almacenamiento mediante su rica biblioteca de servicios vinculados y tipos de conjuntos de datos. El concepto abstracto de un conjunto de datos permite que la actividad Copiar datos transforme los datos entre cualquier emparejamiento de conjunto de datos fuente/sumidero, utilizando el sistema de tipos de datos intermedios de ADF para gestionar las conversiones de tipos entre ambos.

## Conceptos clave

El concepto clave de este capítulo es la actividad Copiar datos. Esta potente actividad soporta el movimiento y la transformación de datos entre una amplia variedad de formatos y servicios de almacenamiento. Los conceptos relacionados incluyen:

- **Archivo no estructurado:** Un archivo tratado como si no tuviera una estructura de datos interna: un blob. La actividad Copiar datos trata los archivos como no estructurados cuando se especifica una copia binaria.
- **Archivo estructurado:** Un archivo con una estructura de datos tabular como CSV o Parquet.
- **Archivo Parquet:** Un formato de archivo estructurado orientado a columnas y comprimido que admite el almacenamiento y la consulta eficientes de grandes volúmenes de datos.
- **Archivo semiestructurado:** Un archivo con una estructura de datos no tabular y frecuentemente anidada, como XML o JSON.
- **Colección de referencia:** Las estructuras de datos anidadas pueden representar varias colecciones de datos simultáneamente. En un mapeo de esquema de actividad de datos de copia, la referencia de la colección indica cuál de las colecciones se está transformando.
- **Receptor:** Azure Data Factory se refiere a los destinos de la canalización de datos como sumideros.
- **Tipo de datos intermedios:** La actividad Copiar datos convierte los valores de datos entrantes desde sus tipos de origen a tipos de datos intermedios de ADF, y luego los

convierte al tipo de sistema de sumidero correspondiente. Esto facilita y agiliza la ampliación del ADF para admitir nuevos conjuntos de datos.

- **Unidad de integración de datos (DIU):** Una DIU es una medida de potencia informática que incorpora el uso de la CPU, la memoria y la red. La potencia se asigna a las ejecuciones de actividades de datos de copia como un número de DIUs; el coste de una ejecución se determina por la duración para la que se asignaron esas DIUs.
- **Grado de paralelismo (DoP):** Una actividad de copia de datos puede realizarse en paralelo utilizando varios hilos para leer diferentes archivos simultáneamente. El número máximo de hilos utilizados durante la ejecución de una actividad es su grado de paralelismo; el número puede establecerse manualmente para la actividad, pero no se aconseja.
- **Azure SQL DB:** servicio de SQL Server basado en Azure, PaaS.
- **Servidor SQL lógico:** Agrupación lógica de bases de datos SQL de Azure para su gestión colectiva.
- **Editor de consultas en línea:** Editor de consultas basado en la web disponible para su uso con Azure SQL DB (y otras plataformas de bases de datos de Azure).

## Experiencia de usuario de Azure Data Factory (ADF UX)

Muchas de las características de ADF UX le resultarán familiares ahora como resultado de haber trabajado en este capítulo. La Figura 3-18 muestra el espacio de trabajo de autoría del ADF UX subdividido en varias regiones:

- La **región 1** es el explorador de Recursos de la Fábrica. Desde aquí, usted utilizó los diversos menús de Acciones para crear y clonar tuberías y conjuntos de datos y para organizarlos en carpetas.
- La **región 2** es la caja de herramientas de Actividades. Contiene la actividad Copiar datos, utilizada ampliamente a lo largo de este capítulo, entre otras.
- La **región 3** contiene pestañas para los recursos abiertos, como las definiciones de pipelines o conjuntos de datos.
- La **región 4** es la barra de herramientas del lienzo. Sus controles incluyen el botón de depuración para ejecutar pipelines y el interruptor del panel de propiedades (icono del deslizador). El editor de código (icono de llaves) permite editar directamente las definiciones JSON de los recursos de fábrica.
- La **región 5** es el lienzo de autoría, un editor visual utilizado para interactuar con las actividades de los pipelines. Las herramientas de búsqueda, zoom y alineación automática se encuentran en el extremo derecho.
- La **región 6** es el panel de configuración del lienzo de autoría.

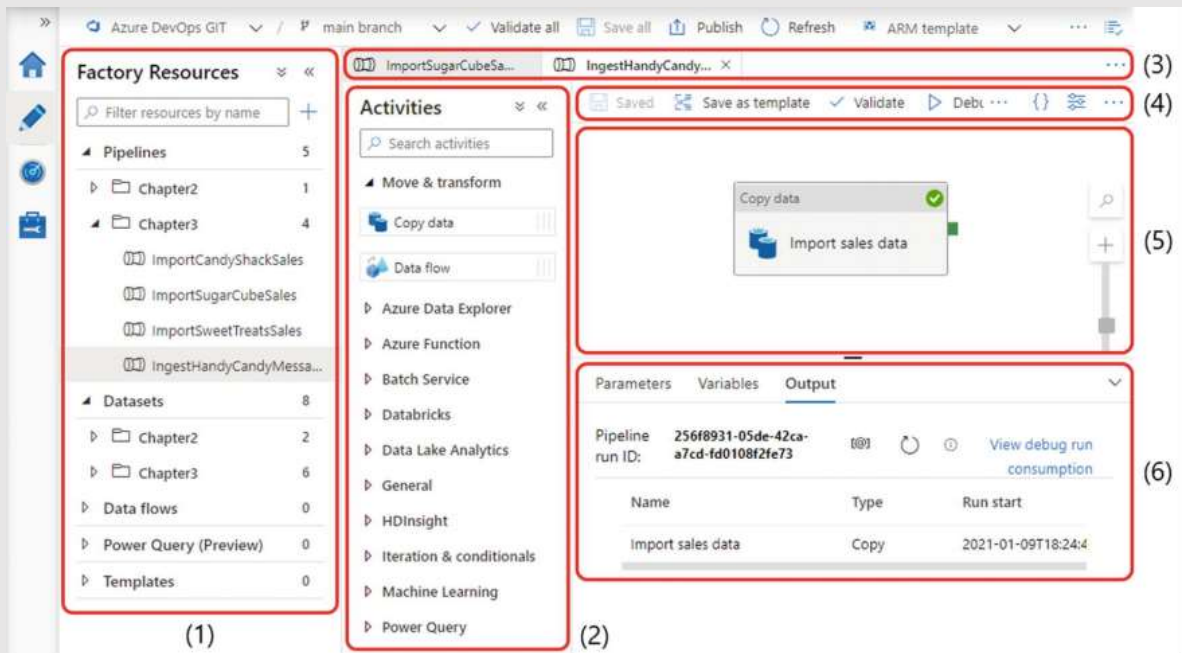


Figura 3-18 Regiones del espacio de trabajo de creación de ADF UX

El panel de configuración contiene pestañas de configuración específicas para el objeto actualmente seleccionado, por ejemplo, una canalización, un conjunto de datos o una actividad. Éstas varían ampliamente dependiendo del objeto seleccionado. En este capítulo, ha visto las opciones de configuración para

- Conexiones de conjuntos de datos de Azure SQL DB y de archivos CSV (Figuras 3-7 y 3-12)
- Conjuntos de datos CSV como fuentes (Figura 3-10)
- Conjuntos de datos SQL DB como sumideros (Figura 3-11)
- Asignación de esquemas estructurados y semiestructurados (figuras 3-14 a 3-17)
- Salida de ejecución de depuración de tuberías (Figura 3-8)

La amplia gama de conectores disponibles hace que la lista de opciones de configuración de conjuntos de datos sea muy amplia. Mi intención aquí no es ofrecer una introducción exhaustiva a todos los tipos de conectores, sino presentar algunos conjuntos de datos comunes y proporcionarle las herramientas para que encuentre su propio camino.

## Para los desarrolladores de SSIS

La actividad Copiar datos fue introducida en el Capítulo 2, en un rol similar al de una Tarea de Sistema de Archivos SSIS. Este capítulo ha extendido su uso a algo parecido a una Tarea de Flujo de Datos rudimentaria. Abstrayendo las fuentes y los sumideros como conjuntos de datos, la actividad es capaz de actuar como cualquier tipo de componente de origen y destino, pero no soporta las transformaciones intermedias por fila que son familiares en la superficie de flujo de datos de SSIS. La verdadera funcionalidad de las tareas de flujo de datos se introducirá en los capítulos 7 y 9.

Al igual que SSIS, ADF consigue una transformación flexible entre los sistemas de tipos de origen y destino utilizando un sistema de tipos intermedios propio. Al igual que con los tipos de datos de Integration Services, esto significa que los nuevos conectores sólo tienen que preocuparse de las conversiones de tipos hacia y desde los tipos de datos intermedios de ADF.