



Azure Data Factory by Example

Practical Implementation for
Data Engineers

—
Richard Swinbank

Apress®

Contenido

2. Tu primer pipeline	3
2.1. Trabajar con Azure Storage	3
2.1.1. Crear una cuenta de Azure Storage	3
2.1.2. Explorar Azure Storage	6
2.1.3. Carga de datos de muestra (Upload Sample Data)	7
2.2. Utilice la herramienta de copia de datos (Copy Data Tool)	8
2.3. Explore su pipeline	15
2.3.1. Servicios vinculados (Linked Services)	15
2.3.2. Datasets	16
2.3.3. Pipelines	18
2.3.4. Actividades	19
2.3.5. Tiempos de ejecución de integración (Integration Runtimes)	19
2.3.6. Recursos de fábrica en Git (Factory Resources in Git)	22
2.4. Depurar tu pipeline	24
2.4.1. Ejecutar el pipeline en modo de depuración (debug mode)	24
2.4.2. Inspeccionar los resultados de la ejecución	25
Revisión del capítulo	26
Conceptos clave	26
Para los desarrolladores de SSIS	27

2. Tu primer pipeline

Las cargas de trabajo ETL se implementan en Azure Data Factory en unidades llamadas pipelines. Utilizando la instancia de Azure Data Factory que creó en el capítulo 1, en este capítulo creará una canalización utilizando la herramienta Copiar datos (Copy data tool), un asistente de creación de canalizaciones que va creando los distintos componentes que conforman una canalización. Después, podrá examinar el pipeline en detalle para comprender cómo se construye.

La herramienta Copiar Datos (Copy Data tool) le guía a través de la construcción de pipelines con el propósito de copiar datos de un lugar a otro. Antes de poder hacerlo, necesita algunos datos para copiar. En la primera sección de este capítulo, creará una cuenta de Azure Storage y cargará algunos datos de muestra para trabajar con ellos.

2.1. Trabajar con Azure Storage

Azure Storage es la plataforma de almacenamiento en la nube gestionada por Microsoft. Los datos almacenados mediante los servicios de Azure Storage están encriptados, replicados y se puede acceder a ellos de forma segura desde cualquier parte del mundo. La escalabilidad de la capacidad y la velocidad del servicio lo convierten en una buena opción para muchos escenarios de almacenamiento y procesamiento de datos.

2.1.1. Crear una cuenta de Azure Storage

Para utilizar los servicios de Azure Storage, primero debe crear una cuenta de Azure Storage. Cree su cuenta de almacenamiento de la siguiente manera:

1. En el portal de Azure, cree un nuevo recurso de tipo Cuenta de almacenamiento. Observará que el menú desplegable Buscar servicios y marketplace está limitado a cinco entradas: muchos nombres de servicios de Azure contienen la palabra "almacenamiento", por lo que es posible que tenga que introducir más texto antes de ver la opción de cuenta de almacenamiento.
2. Completa la pestaña Datos básicos del formulario Crear una cuenta de almacenamiento (Figura 2-1). En Detalles del proyecto, seleccione el grupo de Suscripción y Recursos que utilizó para crear su fábrica de datos en el Capítulo 1.

Home > New > Storage account >

Create a storage account

Basics | Advanced | Networking | Data protection | Tags | Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Resource group * [Create new](#)

Instance details

The default template offers choices that must be made at create time. You may choose to deploy using the original Resource Manager which offers all Azure feature, including legacy options. [Choose the original resource manager](#)

Storage account name *

Location *

Performance * ☒ Standard: GPV2 account recommended for most scenarios
☐ Premium: Recommended for scenarios that need minimal retrieval delays

Redundancy *

[Review + create](#) [< Previous](#) [Next : Advanced >](#)

Figura 2-1 Pestaña básica de la hoja Crear una cuenta de almacenamiento

Nota: te sugiero que utilices el mismo grupo de recursos porque te ayudará a hacer un seguimiento de los recursos que crees mientras utilizas este libro. Definitivamente no es un requisito para Azure Data Factory - ¡tu instancia de ADF puede conectarse a recursos prácticamente en cualquier lugar! Estos podrían ser recursos en otros grupos de recursos, suscripciones o tenants de Azure; recursos en plataformas de nube de la competencia como Amazon Web Services (AWS) o Google Cloud Platform (GCP); o incluso sus propios sistemas locales.

3. Especifique un nombre de cuenta de almacenamiento único a nivel mundial. Yo uso nombres que terminan en "sa" (los nombres de las cuentas de almacenamiento solo pueden contener caracteres alfanuméricos en minúsculas).

4. Elija la ubicación más cercana a usted geográficamente - la que creó su fábrica de datos.

Consejo La elección de una ubicación cercana a usted reduce la latencia de la recuperación de datos. La elección de la misma ubicación que su fábrica de datos reduce el coste, ya que el traslado de datos de una región de Azure a otra incurre en un cargo de ancho de banda (a veces denominado cargo de salida).

5. Para el Rendimiento, seleccione "Estándar". Los niveles de rendimiento para el almacenamiento están vinculados al tipo de hardware subyacente. El almacenamiento Premium utiliza discos de estado sólido y es más caro.
6. Seleccione la opción de redundancia "Almacenamiento redundante local (LRS)". Esta es la más barata de las opciones disponibles porque los datos sólo se replican dentro del mismo centro de datos. LRS le protege contra los fallos de hardware pero no contra la interrupción o pérdida del centro de datos - esto es suficiente para fines de aprendizaje o desarrollo, pero en los entornos de producción, es probable que se requiera un mayor nivel de resiliencia.
7. El último paso en la pestaña Datos básicos es hacer clic en Revisar + crear, y después de la validación hacer clic en Crear. (Estoy omitiendo a propósito las cuatro pestañas restantes - Avanzado, Redes, Protección de datos y Etiquetas- y aceptando sus valores por defecto).
8. Una vez completada la implementación, se muestra un mensaje de notificación que incluye un botón de ir al recurso. Haga clic en él para abrir la hoja de la cuenta de almacenamiento del portal (mostrada en la Figura 2-2).

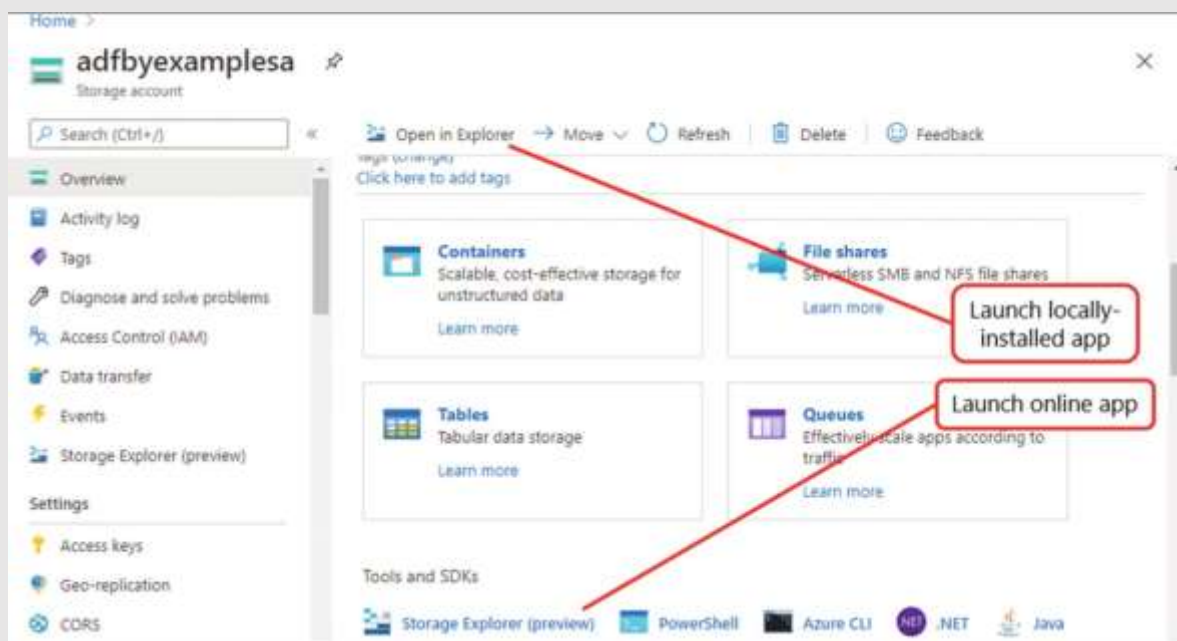


Figura 2-2 Hoja de la cuenta de almacenamiento del portal de Azure

Consejo Puede navegar a cualquier recurso desde el menú del portal (arriba a la izquierda) o la página de inicio. Utilice la opción Grupos de recursos para mostrar todos sus grupos de recursos y, a continuación, seleccione uno de la lista para explorar los recursos que contiene.

2.1.2. Explorar Azure Storage

Hay una variedad de herramientas disponibles para interactuar con las cuentas de almacenamiento. Una opción fácil de usar es Azure Storage Explorer, disponible como una aplicación descargable (disponible para Windows y otros sistemas operativos) o en línea, alojada dentro del portal. Puede iniciar tanto la aplicación en línea como su equivalente instalada localmente directamente desde el portal: para iniciar la aplicación en línea, desplácese hacia abajo y haga clic en Storage Explorer (vista previa) en la hoja de cuentas de almacenamiento del portal (Figura 2-2).

La Figura 2-3 muestra la aplicación online Storage Explorer con la barra lateral de navegación del portal colapsada. La barra lateral del explorador muestra los cuatro tipos de almacenamiento soportados por la cuenta de almacenamiento: contenedores blob (almacenamiento blob), archivos compartidos, colas y tablas. En la siguiente sección, añadirá archivos al almacenamiento blob.

Nota El término blob se utiliza para referirse a un archivo sin tener en cuenta su estructura de datos interna. Esto no implica que los archivos descritos como blobs no tengan estructura - simplemente significa que la estructura no es importante para la tarea en cuestión. El nombre "almacenamiento de blobs" refleja el hecho de que el servicio proporciona un almacén de archivos de uso general, sin restricciones en cuanto a los tipos de archivos que puede contener.

El almacenamiento blob se divide en un solo nivel de contenedores blob - los contenedores no pueden anidarse. Haga clic con el botón derecho del ratón en el elemento BLOB CONTAINERS y utilice la opción Create blob container del menú emergente para crear dos contenedores blob privados llamados "landing" y "sampledata", como se muestra en la Figura 2-3.

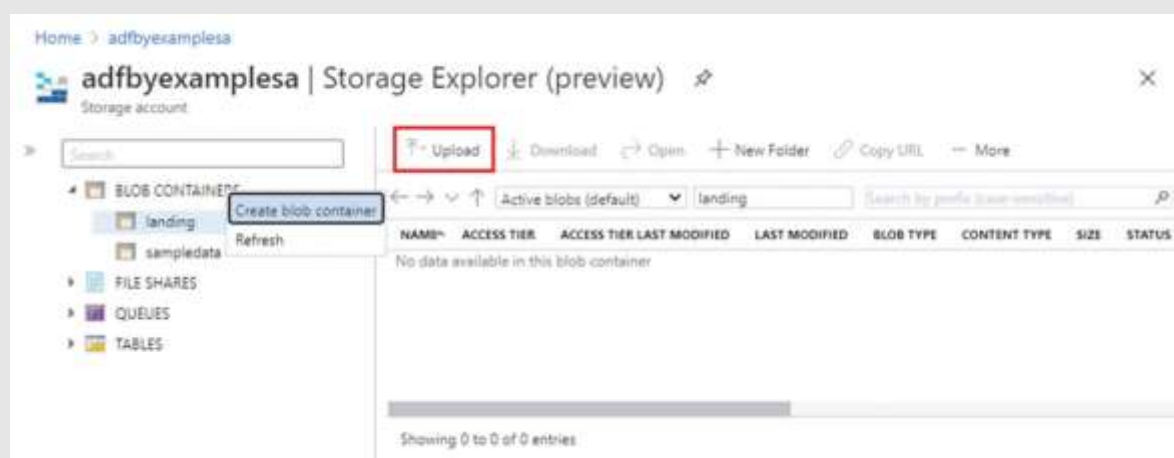


Figura 2-3 Azure Storage Explorer (versión en línea)

2.1.3. Carga de datos de muestra (Upload Sample Data)

Los archivos de datos de muestra utilizados en este libro están disponibles en el repositorio GitHub del libro, ubicado en <https://github.com/Apress/azure-data-factory-by-example>.

1. Descargue el repositorio como un archivo zip, para que pueda transferir los datos de la muestra a su cuenta de almacenamiento. Esta opción está disponible desde el botón verde del menú Código en la página principal del repositorio.
2. Seleccione el contenedor de "landing" en el Explorador de almacenamiento de Azure en línea y haga clic en Cargar en la barra de herramientas situada encima de la lista de contenidos del contenedor (indicada en la Figura 2-3).
3. Se muestra la hoja de carga de blob. En el campo Archivos, haga clic en Seleccionar un archivo y, a continuación, busque y seleccione azure-data-factory-by-example-main.zip, el archivo zip que ha descargado.
4. De nuevo en la hoja de carga de blob, haga clic en Cargar y, a continuación, cierre la hoja.
5. Ahora aparece una entrada para el archivo zip en la lista de contenidos del contenedor de "landing". (Si no la ve, intente seleccionar Actualizar en la colección de elementos del menú Más en la cinta).

Los archivos de datos de muestra contienen datos de ventas de productos elaborados por una multinacional ficticia de confitería, Acme Boxed Confectionery (ABC). El fabricante no vende directamente a los consumidores, sino a una serie de minoristas que informan de la actividad de ventas mensual a ABC. Los informes de ventas suelen elaborarse con los propios sistemas de gestión de datos de los minoristas y se suministran en una gran variedad de formatos de archivo. El manejo de estos formatos le permitirá conocer muchas de las funciones de transformación de datos del ADF en los próximos capítulos.

2.2. Utilice la herramienta de copia de datos (Copy Data Tool)

La herramienta Copy Data de Azure Data Factory proporciona una experiencia de estilo asistente para crear un pipeline con un propósito específico: copiar datos de un lugar a otro. En esta sección, utilizará la herramienta Copiar datos para crear un pipeline que copie el archivo zip del contenedor "landing" en su cuenta de almacenamiento de Azure, lo descomprima y luego escriba su contenido en el contenedor "sampledata". Se trata de una tarea de movimiento de datos muy sencilla, ya que la implementación de una tarea sencilla le permite centrarse en los detalles de la configuración del pipeline de ADF.

La herramienta Copiar Datos se encuentra en la página de resumen de la Fábrica de Datos de ADF UX, a la que se accede haciendo clic en el icono de inicio en la barra lateral de navegación. Bajo el título Comencemos hay una serie de burbujas - haga clic en la burbuja Copiar datos para comenzar.

La herramienta inicia un proceso guiado de varios pasos para crear un pipeline - la página del primer paso aparece en la Figura 2-4.

Copy Data tool

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources. Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines, datasets, and linked services. [Learn more](#)

Properties
Enter name and description for the copy data task.

Task name *
ImportSampleData

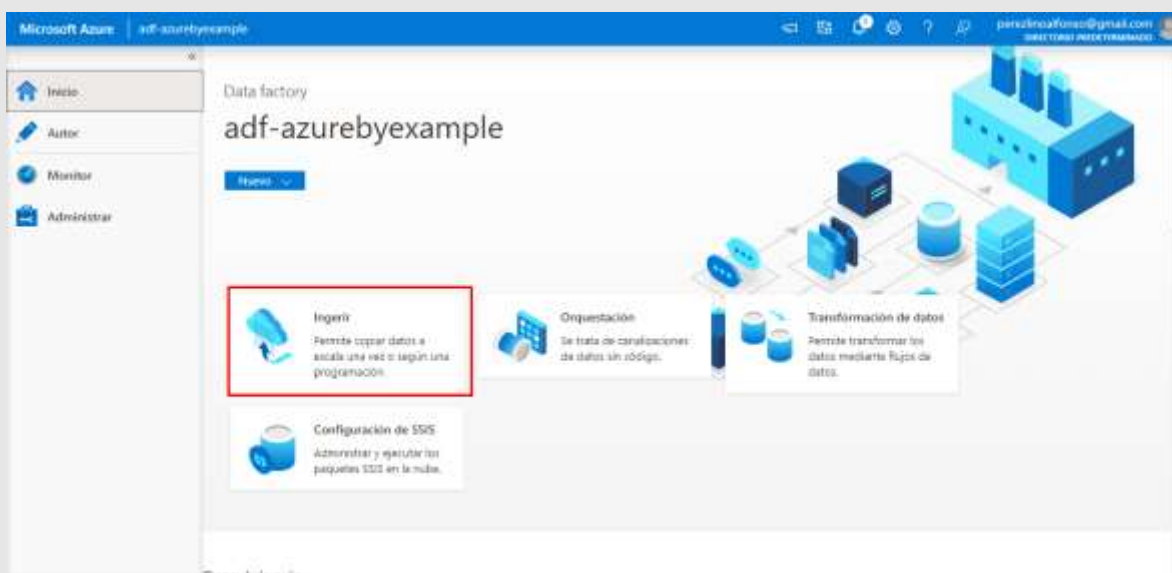
Task description
Unzip file in Azure blob storage 'landing' container and copy contents to 'sampledata' container

Task cadence or task schedule *
☒ Run once now ☐ Schedule ☐ Tumbling window

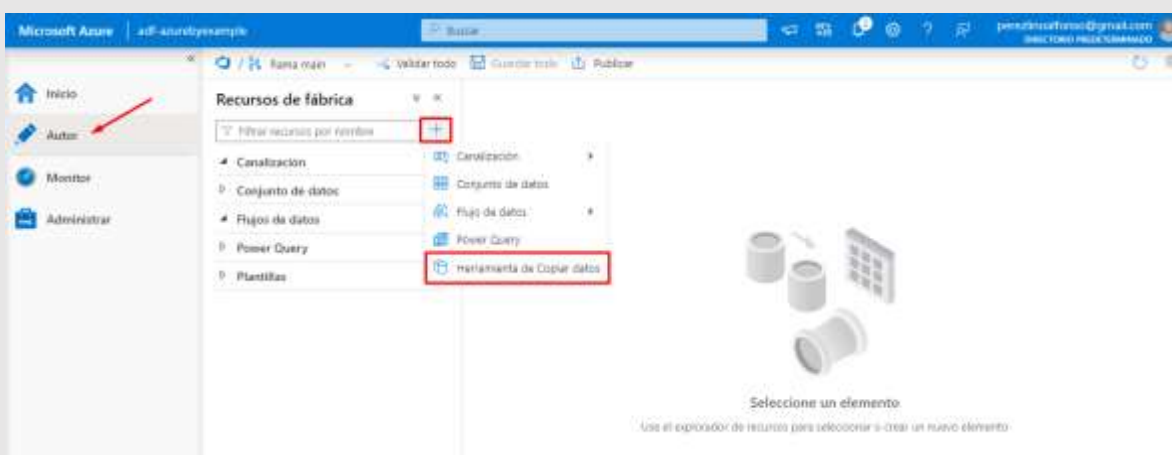
< Previous Next >

Figura 2-4 Primer paso de la herramienta Copiar Datos

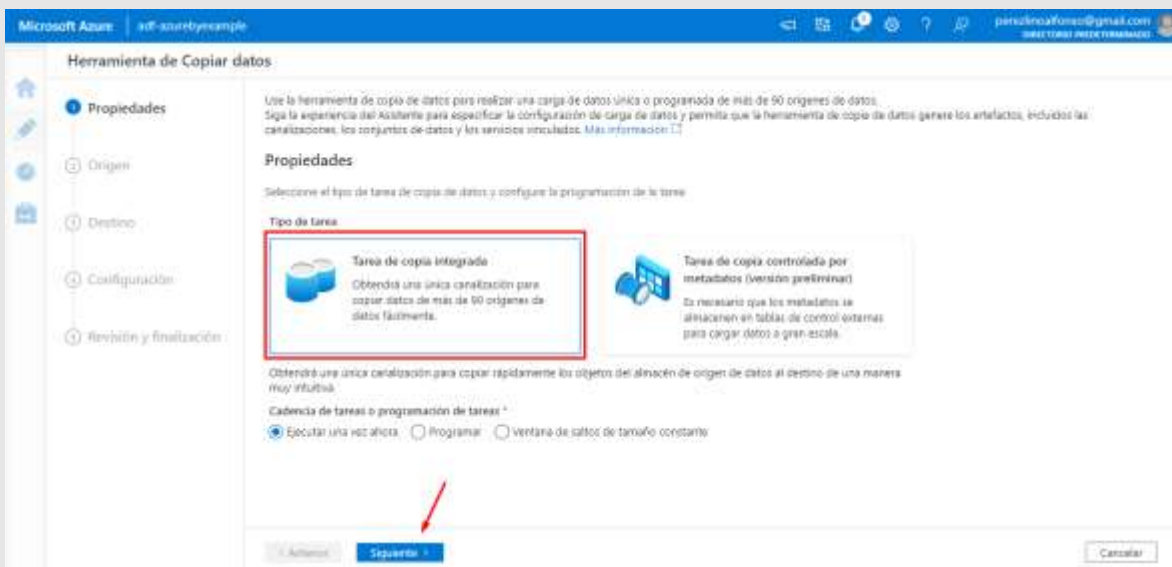
Mi proceso



Otra forma



Y llegamos a esta ventana



Sugerencia Si el proceso mostrado en la Figura 2-4 no se inicia, haga clic en el icono del lápiz para acceder al espacio de trabajo de autoría y, a continuación, busque el campo de búsqueda debajo del título del explorador de recursos de fábrica. Haga clic en el botón con el símbolo más a la derecha del cuadro de búsqueda y, a continuación, seleccione la herramienta Copiar datos en el menú emergente.

Complete el proceso de la siguiente manera:

1. En la página de Propiedades, establezca el nombre de la Tarea como "ImportSampleData" - este será el nombre de su pipeline- y proporcione una descripción de la Tarea. Las descripciones se pueden buscar, por lo que añadir una a cada nuevo pipeline facilita la gestión de una gran fábrica de datos. Haga clic en Siguiente.
2. En la página del almacén de datos de **origen**, haga clic en **+ Crear nueva conexión**. Elija el tipo de servicio vinculado (linked service) Azure Blob Storage y haga clic en Continuar.
3. Utilice la hoja Nuevo servicio vinculado (Azure Blob Storage) (Figura 2-5) para crear una conexión con su cuenta de Azure Storage. Proporcione un nombre y una descripción -he reutilizado el nombre de la cuenta de almacenamiento subyacente- y, a continuación, en Método de selección de la cuenta, asegúrese de que esté seleccionada la opción "Desde la suscripción de Azure". Elija la suscripción y la cuenta de almacenamiento pertinentes que creó anteriormente. En la parte inferior de la hoja, haz clic en Probar conexión para comprobar que funciona, y luego haz clic en Crear.

The screenshot shows the 'New linked service (Azure Blob Storage)' dialog in the Copy Data tool. The left sidebar shows the 'Source' tab selected. The main form has the following fields and options:

- Name ***: adfbyexamplesa
- Description**: Blob storage in 'adfbyexamplesa' storage account
- Connect via integration runtime ***: AutoResolveIntegrationRuntime
- Authentication method**: Account key
- Account selection method**: From Azure subscription (selected), Enter manually
- Azure subscription**: Free Trial (25a779fb-306e-4d88-82c6-801dd2d546c0)
- Storage account name ***: adfbyexamplesa

At the bottom, there is a 'Create' button, a 'Back' button, a 'Test connection' button, and a 'Cancel' button. A green checkmark and the text 'Connection successful' are displayed above the 'Test connection' button.

Figura 2-5 Diálogo de nuevo servicio vinculado (Azure Blob Storage)

Nota La UX del ADF utiliza una clave de almacenamiento (parte de la configuración de su cuenta de almacenamiento) para autorizar la conexión a su cuenta de almacenamiento. La razón por la que no necesita especificar una clave explícitamente es que la UX utiliza su identidad conectada para recuperar su valor.

- De vuelta en la página del almacén de datos de origen, verá un mosaico para su nuevo servicio vinculado - asegúrese de que está seleccionado, y luego haga clic en Siguiente.
- En la página Elegir el archivo o la carpeta de entrada, haga clic en el icono Examinar situado a la derecha del cuadro de texto Archivo o carpeta. Busque en el contenedor "landing", seleccione el archivo zip cargado y haga clic en Elegir.

Microsoft Azure | adf-by-example

Herramienta de Copiar datos

Propiedades

Origen

Conjunto de datos

Configuración

Destino

Configuración

Revisión y finalización

Almacén de datos de origen

Especifique el almacén de datos de origen para la tarea de copia. Puede utilizar una conexión de almacén de datos existente o especificar un nuevo almacén de datos.

Tipo de origen: Almacenamiento de blobs de Azure

Conexión: adfbyexamplestorage2022

Editar Nueva conexión

Archivo o carpeta

Si la identidad que usa para acceder al almacén de datos solo tiene permisos para el subdirectorio en lugar de tenerlos para toda la cuenta, especifique la ruta de acceso a su recurso.

landing.azure-data-factory-by-example-main.zip Examinar

Opciones

☒ Copia binaria

Tipo de compresión: ZipDeflate

Nivel de compresión: Óptimo

☐ Conservar el nombre del archivo .zip como carpeta

☒ Recursivamente

☐ Eliminar archivos tras la finalización

Número máximo de conexiones simultáneas

Filtrar por última modificación

Hora de inicio (UTC): Hora de finalización (hora UTC)

Anterior Siguiente Cerrar

- Marque la casilla Copia binaria (Binary Copy), seleccione Tipo de compresión "ZipDeflate" y desmarque Conservar el nombre del archivo zip como carpeta. Haga clic en Siguiente.

Microsoft Azure | adf-by-example

Herramienta de Copiar datos

Propiedades

Origen

Conjunto de datos

Configuración

Destino

Configuración

Revisión y finalización

Almacén de datos de origen

Especifique el almacén de datos de origen para la tarea de copia. Puede utilizar una conexión de almacén de datos existente o especificar un nuevo almacén de datos.

Tipo de origen: Almacenamiento de blobs de Azure

Conexión: adfbyexamplestorage2022

Editar Nueva conexión

Nivel de compresión: Óptimo

☐ Conservar el nombre del archivo .zip como carpeta

☒ Recursivamente

☐ Eliminar archivos tras la finalización

Número máximo de conexiones simultáneas: 1

Filtrar por última modificación

Hora de inicio (UTC): Hora de finalización (hora UTC)

Anterior Siguiente Cerrar

- En la página Elegir el archivo o carpeta de salida, haga clic en el icono Examinar, seleccione el contenedor "sampledata" y haga clic en Elegir. Deje los demás ajustes sin modificar y haga clic en Siguiente.

Microsoft Azure | adf-sample

Herramienta de Copiar datos

Propiedades

Origen

Destino

Configuración

Revisión y finalización

Almacén de datos de destino

Especifique el almacén de datos de destino para la tarea de copia. Puede utilizar una conexión de almacén de datos existente o especificar un nuevo almacén de datos.

Tipo de destino: Almacenamiento de blobs de Azure

Conexión: adfsamplestorage2022

Ruta de acceso de la carpeta: sampledata

Nombre de archivo:

Tipo de compresión: Ninguno

Comportamiento de copia: Ninguno

Número máximo de conexiones simultáneas:

Anterior Siguiente Cancelar

- Vuelva a hacer clic en Siguiente en la página Configuración, y luego inspeccione los detalles que ha proporcionado en la página Resumen. Cuando esté listo, haga clic en Siguiente para iniciar la creación del pipeline.

Microsoft Azure | adf-sample

Herramienta de Copiar datos

Propiedades

Origen

Destino

Configuración

Revisión y finalización

Configuración

Especifique el nombre y la descripción de la tarea de copia de datos, más opciones de movimiento de datos.

Nombre de la tarea: CopyPipeline_ad

Descripción de la tarea:

Habilitar el registro: ☒

Habilitar el almacenamiento provisional: ☒

Avanzadas

Anterior Siguiente Cancelar

La última etapa del proceso de la herramienta de copia de datos es la página de despliegue (Deployment page). Ésta se ejecutará rápidamente a través de la creación de recursos antes de que aparezca el botón Finalizar, como en la Figura 2-6, indicando que ha creado con éxito su pipeline. Haga clic en Finalizar para cerrar la herramienta.

Microsoft Azure | aif-azurebyexample

Herramienta de Copiar datos

Propiedades

Origen

Destino

Configuración


Revisión y finalización

Revisar

Implementación

Resumen

Está ajustando la canalización para copiar datos de Almacenamiento de blobs de Azure a Almacenamiento de blobs de Azure.



Propiedades

Nombre de la tarea: CopyPipeline_zp

Descripción de la tarea:

Origen

Nombre de conexión: aifbyexamplestorage0022

Nombre del conjunto de datos: SourceCatalog_zp

Nombre de archivo: azure-data-factory-by-example-main.zip

Contenedor: landing

Anterior

Siguiente

Cancelar

Microsoft Azure | aif-azurebyexample

Herramienta de Copiar datos

Propiedades

Origen

Destino

Configuración

Revisión y finalización

Revisar

Implementación

Implementación completado

Paso de implementación

Estado

> Creando conjuntos de datos: Correcto

> Creando canalizaciones: Correcto

Finalizar

Editar la canalización

Monitor

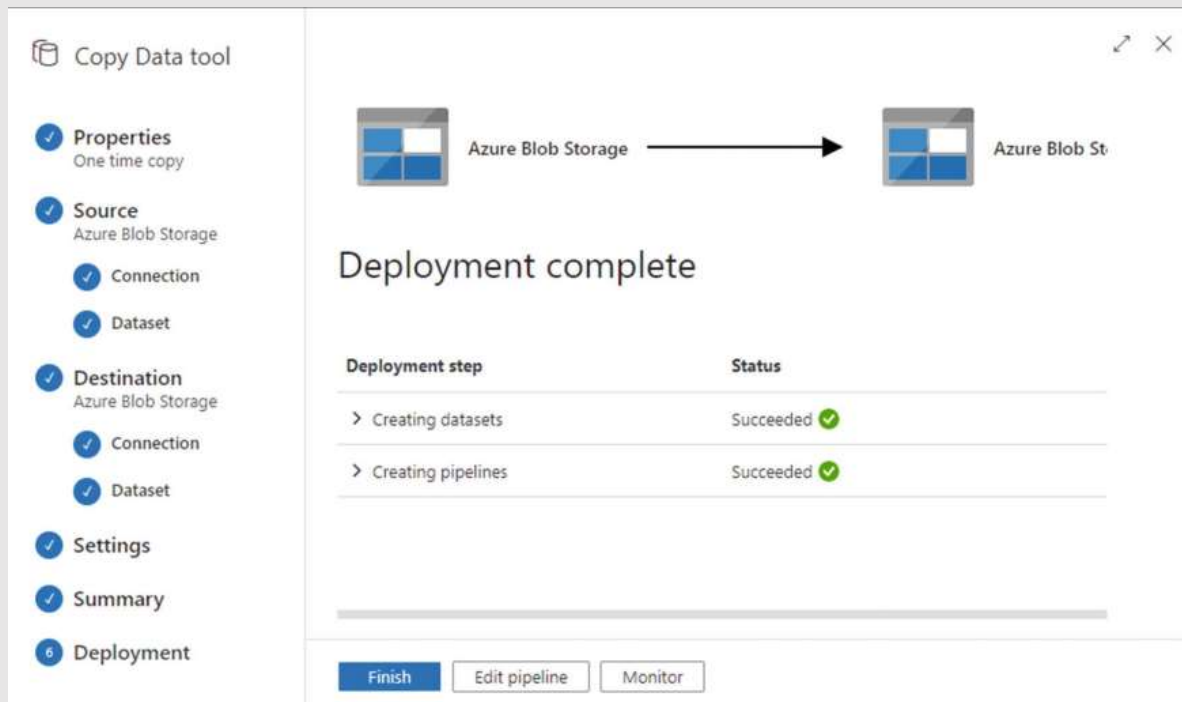


Figura 2-6 Página de despliegue de la herramienta Copiar Datos

2.3. Explore su pipeline

En la sección anterior, utilizó la herramienta Copiar Datos para crear su primer pipeline de ADF. Para lograrlo, la UX del ADF realizó las siguientes acciones en su nombre:

- Creó una conexión de servicio vinculada (linked service) a su cuenta de almacenamiento
- Publicó la conexión de servicio vinculado
- Creó conjuntos de datos que representan las entradas y salidas del proceso de copia
- Creó un pipeline de ADF para realizar el proceso de copia
- Se ha confirmado y enviado el servicio vinculado, los conjuntos de datos y la canalización a su repositorio Git.

Nota Para evitar la exposición, la clave de almacenamiento utilizada para autorizar la conexión de ADF a su cuenta de almacenamiento no se consigna en Git. En su lugar, la clave se guarda directamente en el ADF publicando la conexión del servicio vinculado inmediatamente. En esta sección, examinará los distintos recursos creados por la herramienta de copia de datos.

2.3.1. Servicios vinculados (Linked Services)

Una forma común de definir los recursos para cualquier operación de datos es en términos de almacenamiento y computación:

- El **almacenamiento (storage)** se refiere a la retención de datos, sin que se realice ningún procesamiento o movimiento adicional (excepto, por ejemplo, el movimiento que se produce dentro del sistema de almacenamiento para la replicación del almacenamiento).
- La **computación (compute)** describe la potencia de cálculo utilizada para mover los datos almacenados, transformarlos y analizarlos.

La separación de los servicios de almacenamiento y de computación en plataformas de nube como Azure es muy común. Añade flexibilidad al permitir que ambos se amplíen y reduzcan de forma independiente a medida que aumenta o disminuye la demanda.

Azure Data Factory no tiene recursos de almacenamiento propios, pero las instancias de la fábrica pueden acceder y utilizar recursos de almacenamiento y computación externos a través de servicios vinculados (linked services). Un servicio vinculado puede proporcionar una conexión a un sistema de almacenamiento -por ejemplo, una cuenta de almacenamiento de Azure o una base de datos- o puede permitir el acceso a recursos informáticos externos como una Azure Function App o un clúster de Databricks.

Utilizando la herramienta Copiar Datos, has creado un servicio vinculado de Azure Blob Storage para conectarte a tu cuenta de almacenamiento de Azure. Los servicios vinculados se definen en el centro de gestión de ADF UX, al que se accede haciendo clic en el icono de la caja de herramientas en la barra lateral de navegación. La Figura 2-7 muestra la página de servicios vinculados del centro de gestión, que contiene la conexión de Azure Blob Storage creada anteriormente.

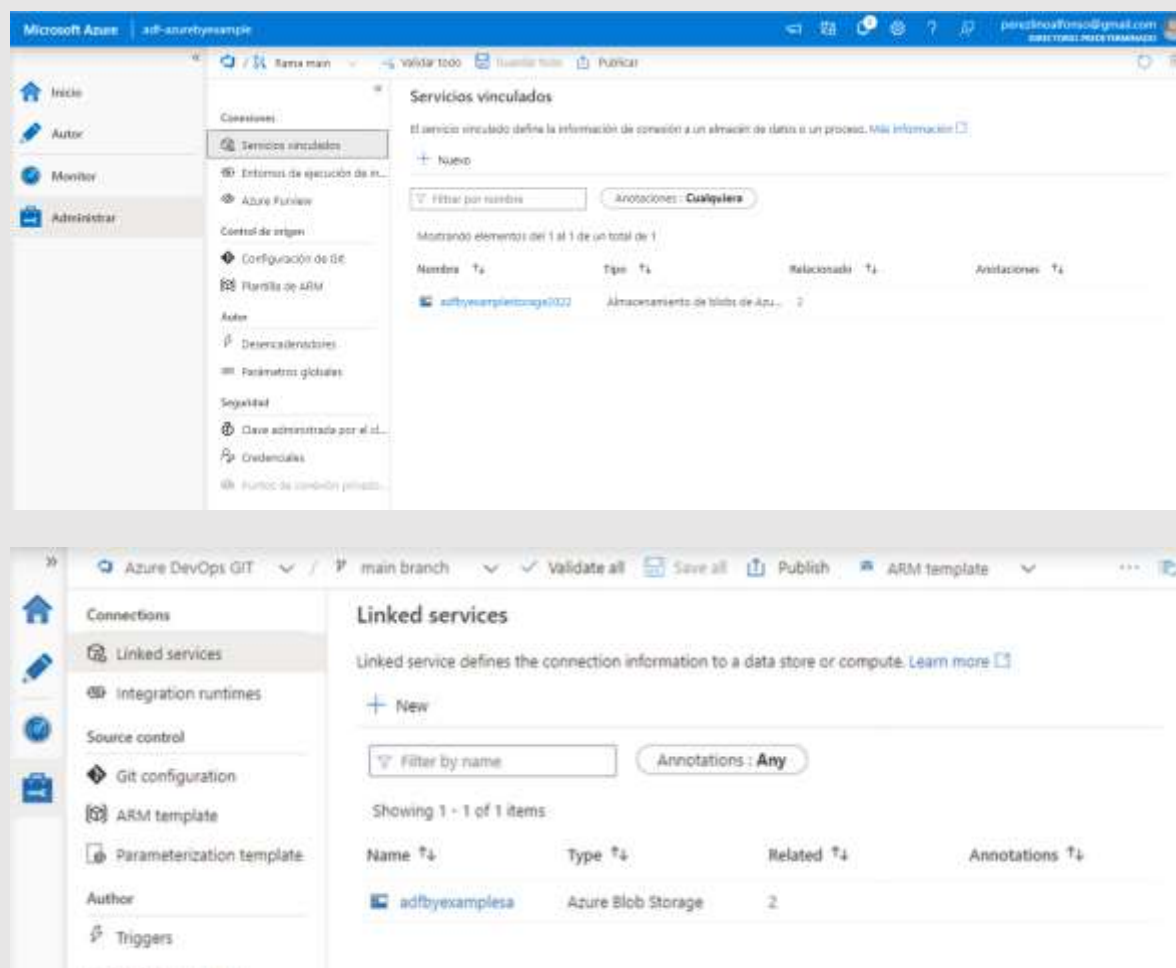


Figura 2-7 Servicios vinculados en el centro de gestión de ADF UX

2.3.2. Datasets

Una conexión de almacenamiento de servicios vinculados proporciona la información necesaria para conectarse a un almacén de datos externo, pero nada más. En el caso de un sistema de base de datos, esto podría consistir en una cadena de conexión a la base de datos, incluyendo detalles del servidor, la base de datos y las credenciales, pero sin información sobre las tablas de la base de datos. Del mismo modo, el servicio vinculado que creó en la sección anterior contiene metadatos para localizar la cuenta de Azure Blob Storage correspondiente, pero no tiene información sobre los archivos almacenados allí. Azure Data Factory almacena metadatos para representar objetos dentro de los sistemas de almacenamiento externos como conjuntos de datos.

Los conjuntos de datos (datasets) se definen en el espacio de trabajo de creación de ADF UX (al que se accede haciendo clic en el icono del lápiz en la barra lateral de navegación). Aquí encontrará dos conjuntos de datos creados por la herramienta Copiar datos: uno representa el archivo zip de origen de la actividad de copia y el otro el contenedor de destino. El espacio de trabajo tiene tres regiones, como se muestra en la Figura 2-8:

- Una barra de cabecera de fábrica (debajo de la barra de cabecera de navegación principal), que contiene varios controles e indica que la fábrica está vinculada a un repositorio Git de Azure DevOps
- El explorador de recursos de la fábrica, que enumera los pipelines, conjuntos de datos y otros recursos de la fábrica
- Un lienzo de autoría con pestañas, que muestra los detalles de los recursos de fábrica seleccionados

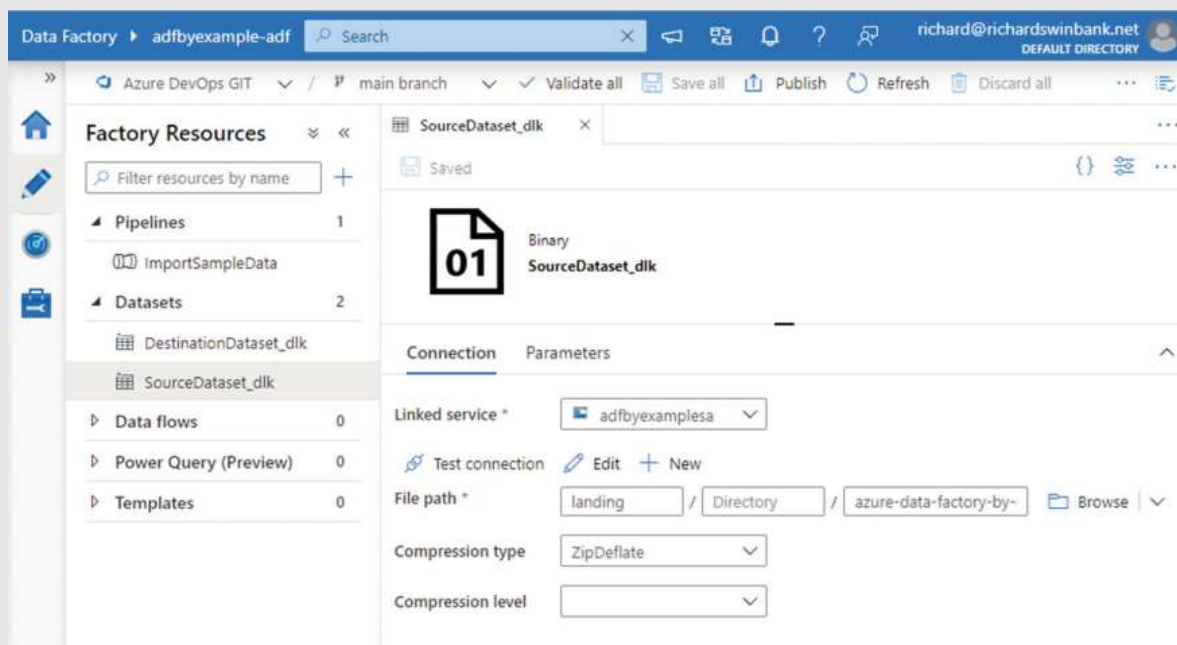


Figura 2-8 Configuración del conjunto de datos en el espacio de trabajo de creación de ADF UX

Consejo El término recurso se utiliza en el portal de Azure para describir diferentes instancias de los servicios de Azure (por ejemplo, una cuenta de almacenamiento o una fábrica de datos) y se utiliza dentro de ADF para describir varios componentes de fábrica como pipelines y conjuntos de datos. La reutilización de la terminología puede ser confusa, pero debería quedar claro por el contexto si me estoy refiriendo a recursos de Azure o a recursos de ADF.

Fíjate también en que, en el espacio de trabajo de autoría, la barra de cabecera de navegación cuenta con un cuadro de búsqueda. Esto le permite buscar definiciones de recursos de fábrica, incluyendo descripciones de texto como la que creó para el pipeline "ImportSampleData".

El lienzo de creación de la Figura 2-8 muestra el conjunto de datos de origen "SourceDataset_dlk". (El nombre del conjunto de datos fue generado automáticamente por la herramienta Copiar Datos). La pestaña Conexión del panel de configuración con pestañas en la mitad inferior de la pantalla muestra los detalles de la ruta de archivo seleccionada: el contenedor "landing", el archivo "azure-data-factory-by-example-main.zip" y el tipo de compresión "ZipDeflate".

Para los desarrolladores de SSIS: Los servicios vinculados se comportan de forma similar a los gestores de conexión con ámbito de proyecto en SSIS, pero a diferencia de algunos gestores de conexión (como el archivo plano), no contienen metadatos de esquema. Los metadatos de esquema pueden definirse por separado en un conjunto de datos (dataset) de ADF, aunque no siempre son necesarios (como en el caso de la copia de archivos sin esquema que realizó con la herramienta Copiar datos).

2.3.3. Pipelines

Los pipelines son el corazón de Azure Data Factory. Un pipeline es una colección de actividades de movimiento y transformación de datos, agrupadas para lograr una tarea de integración de datos de alto nivel. La Figura 2-9 muestra el espacio de trabajo de creación de ADF UX con el pipeline "ImportSampleData". Cuando se crean pipelines, el espacio de trabajo contiene además una caja de herramientas de Actividades, un menú de actividades de pipelines disponibles. Las actividades pueden arrastrarse desde la caja de herramientas de Actividades y soltarse en el lienzo de creación.

En el lienzo de creación de la Figura 2-9, puede ver que el pipeline contiene una única actividad de copia de datos, llamada "Copy_dlk". (El nombre de la actividad también fue generado automáticamente por la herramienta Copiar datos).

Para los desarrolladores de SSIS: Un pipeline de fábrica de datos es equivalente a un paquete SSIS, y el lienzo de autoría mostrado en la Figura 2-9 proporciona una funcionalidad comparable a la superficie de flujo de control de SSIS. Este sencillo pipeline es como un paquete SSIS que contiene una única File System Task para copiar archivos de una ubicación a otra - pero con la capacidad adicional de descomprimir archivos sobre la marcha.

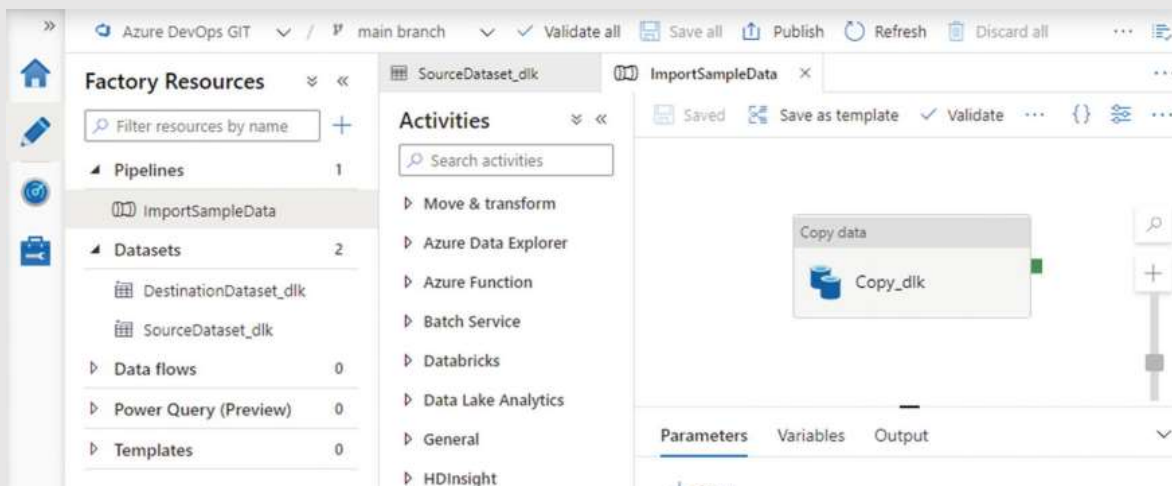


Figura 2-9 Configuración del pipeline en el lienzo de autoría de ADF UX

2.3.4. Actividades

La caja de herramientas de Actividades proporciona una variedad de tipos de actividades disponibles para su uso en un pipeline de ADF. Las actividades están disponibles para el movimiento nativo y la transformación de datos dentro de ADF, así como para orquestar el trabajo realizado por recursos externos como Azure Databricks y servicios de Machine Learning.

Este sencillo pipeline contiene sólo una actividad, pero en el Capítulo 4 empezará a ver cómo se pueden enlazar múltiples actividades dentro de un pipeline para orquestar flujos de trabajo ETL progresivamente más complejos. ADF define una biblioteca de más de 30 tipos de actividades, incluida la posibilidad de escribir sus propias actividades personalizadas, lo que hace que el alcance de las tareas que puede realizar un pipeline sea prácticamente ilimitado.

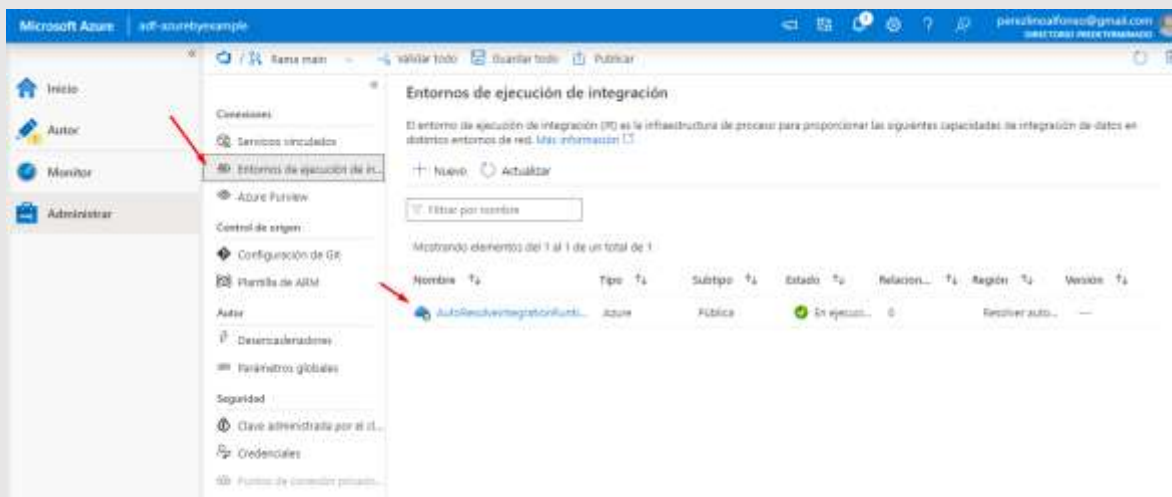
2.3.5. Tiempos de ejecución de integración (Integration Runtimes)

Azure Data Factory accede a la computación de dos maneras:

- Externamente, creando una conexión de servicio vinculada a un servicio de computación independiente como Azure Databricks o HDInsight
- Internamente, utilizando un servicio de computación gestionado por ADF llamado tiempo de ejecución de integración (integration runtime)

Las transformaciones y los movimientos de datos nativos -como la actividad Copiar datos- utilizan recursos informáticos proporcionados por un tiempo de ejecución de integración. Esto es lo que describí en el capítulo 1 como "computación de fábrica".

Los tiempos de ejecución de integración aparecen en el hub de gestión de ADF UX, inmediatamente debajo de Servicios vinculados en el menú del hub. Cada instancia de ADF incluye automáticamente un tiempo de ejecución de integración, llamado AutoResolveIntegrationRuntime. Esto proporciona acceso a los recursos informáticos de Azure en una ubicación geográfica que se elige automáticamente, dependiendo de la tarea que se está realizando. En determinadas circunstancias, es posible que desee crear tiempos de ejecución de integración propios; el capítulo 8 vuelve a tratar esta cuestión con más detalle.



El cálculo requerido para la actividad de copia de datos de su pipeline es proporcionado por el AutoResolveIntegrationRuntime. Esto no se especifica como parte de la actividad en sí, sino como parte de los servicios vinculados a la cuenta de almacenamiento que utiliza. Si revisa la Figura 2-5, notará que la opción Conectar vía integration runtime tiene el valor "AutoResolveIntegrationRuntime".

Para los desarrolladores de SSIS: El paralelo más cercano a un tiempo de ejecución de integración en SSIS es el servicio de Windows de los Servicios de Integración (Integration Services Windows) - proporciona acceso a los recursos de computación del servidor para el movimiento y transformación de datos. El concepto es menos visible en SSIS, simplemente porque sólo hay un tiempo de ejecución, utilizado por todas las tareas en todos los paquetes.

La Figura 2-10 ilustra la relación entre los servicios vinculados, los conjuntos de datos, las actividades, los tiempos de ejecución de integración y su pipeline. Las flechas indican la dirección del data flow desde el contenedor "landing" hasta el contenedor "sampledata".

En la figura, se puede ver que:

- El conjunto de datos "SourceDataset_dlk" utiliza el servicio vinculado "adfbvexamplesa" para conectarse al contenedor "landing" en la cuenta de almacenamiento del mismo nombre.

- El conjunto de datos "DestinationDataset_dlk" utiliza el servicio vinculado "adfbyexamplesa" para conectarse al contenedor "sampledata" en la misma cuenta de almacenamiento.
- El pipeline "ImportSampleData" contiene una única actividad de copia de datos, "Copy_dlk", que utiliza el AutoResolveIntegrationRuntime para copiar los datos de "SourceDataset_dlk" a "DestinationDataset_dlk".

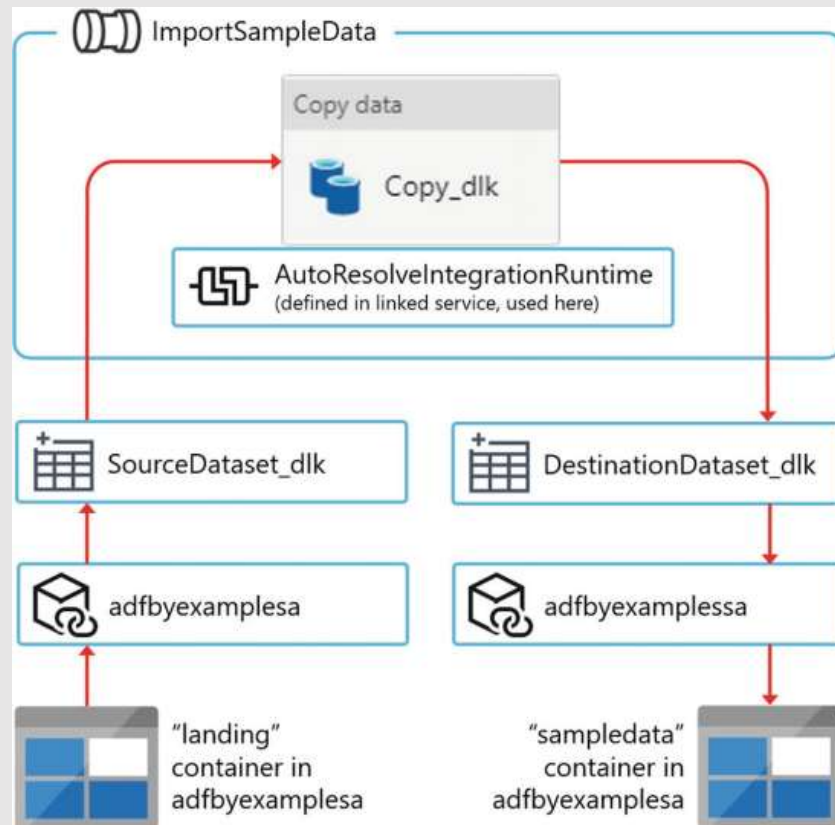


Figura 2-10 Relación entre los recursos de Azure Data Factory

Las características de la implementación del pipeline ilustrada por la Figura 2-10 son muy comunes:

- El acceso a los datos identificados en el almacenamiento externo es proporcionado por un conjunto de datos a través de una conexión de servicio vinculada.
- Las actividades del pipeline utilizan un tiempo de ejecución de integración de ADF para mover y transformar los datos entre los conjuntos de datos.

2.3.6. Recursos de fábrica en Git (Factory Resources in Git)

Los servicios vinculados, los conjuntos de datos y los pipelines son ejemplos de recursos de fábrica. Además de crear definiciones de recursos en la UX de ADF, la herramienta Copy Data también los guarda, al confirmarlos y empujarlos a la rama de colaboración de tu repositorio Git.

Mira tu repositorio Azure DevOps de nuevo, y verás una estructura similar a la de la Figura 2-11. La captura de pantalla muestra la estructura de carpetas de mi repositorio "AdfByExample", que ahora contiene cuatro carpetas: "dataset", "factory", "linkedService" y "pipeline". Cada carpeta contiene definiciones para los recursos del tipo correspondiente, almacenados como archivos JSON - el panel de contenido de la derecha muestra la definición JSON de mi conjunto de datos "SourceDataset_dlk".

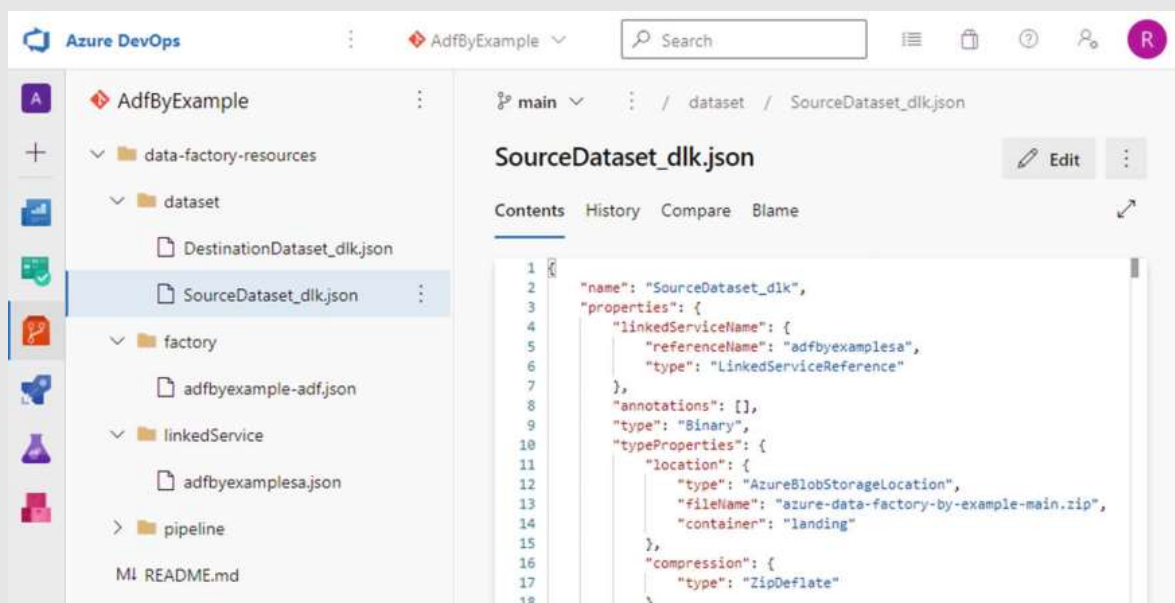


Figura 2-11 Contenido del repositorio Git después de crear los recursos de la fábrica

Trabajar con -y confirmar directamente en- tu rama de colaboración es característico de un flujo de trabajo Git centralizado. Los flujos de trabajo de desarrollo más sofisticados que utilizan las ramas de colaboración de Git también son compatibles con la UX del ADF. El menú desplegable de la rama, indicado en la Figura 2-12, le permite cambiar entre ramas Git, crear nuevas y abrir solicitudes de extracción. Los workflows de desarrollo están fuera del alcance de este libro - durante los siguientes capítulos, deberías continuar confirmando los cambios directamente en la rama de colaboración de la fábrica, main.

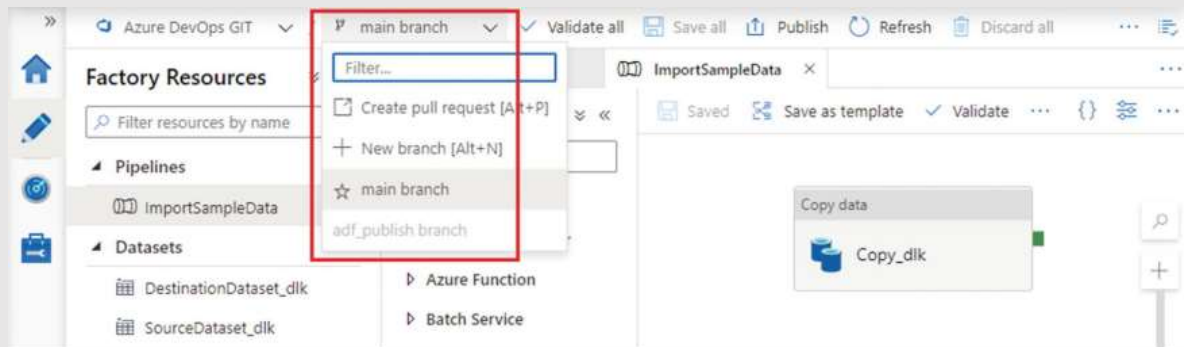


Figura 2-12 Despliegue de selección de rama Git

2.4. Depurar tu pipeline

Con el trabajo que has realizado con la herramienta Copiar Datos, ahora tienes definiciones de recursos en varios lugares diferentes:

- Su repositorio Git contiene guardados servicios vinculados, datasets y pipelines definidos.
- Esas definiciones se cargan en su sesión de ADF UX, donde puede editarlas.
- El entorno publicado de ADF contiene la definición de servicio vinculado (porque la herramienta Copy Data la publicó para guardar la clave de la cuenta de almacenamiento de forma segura).

Para ejecutar el pipeline en el entorno publicado, primero habría que publicar todos sus recursos relacionados. La publicación de los recursos de la fábrica es el tema del Capítulo 10. Hasta entonces, ejecutará los pipelines de forma interactiva en la UX del ADF, utilizando su modo de Depuración (Debug mode) - siempre que diga "ejecutar el pipeline" a partir de ahora, quiero decir "Haga clic en Depuración para ejecutar el pipeline". (Es posible que haya encontrado la opción Trigger now en el menú Add trigger sobre el lienzo de creación - esto ejecuta pipelines publicados y también será examinado en el Capítulo 10).

Tenga en cuenta que "Depurar" significa simplemente "ejecutar la definición del pipeline en mi sesión de ADF UX, sin publicarlo". Una ejecución de depuración de pipeline accede y modifica los recursos externos exactamente de la misma manera que un pipeline publicado.

2.4.1. Ejecutar el pipeline en modo de depuración (debug mode)

Para ejecutar el pipeline "ImportSampleData" en modo de depuración, abra el lienzo de autoría y seleccione el pipeline en el explorador de recursos de fábrica. La barra de herramientas situada inmediatamente encima del lienzo de creación contiene un botón de depuración, como se muestra en la Figura 2-13 - haga clic en él para ejecutar el canal.

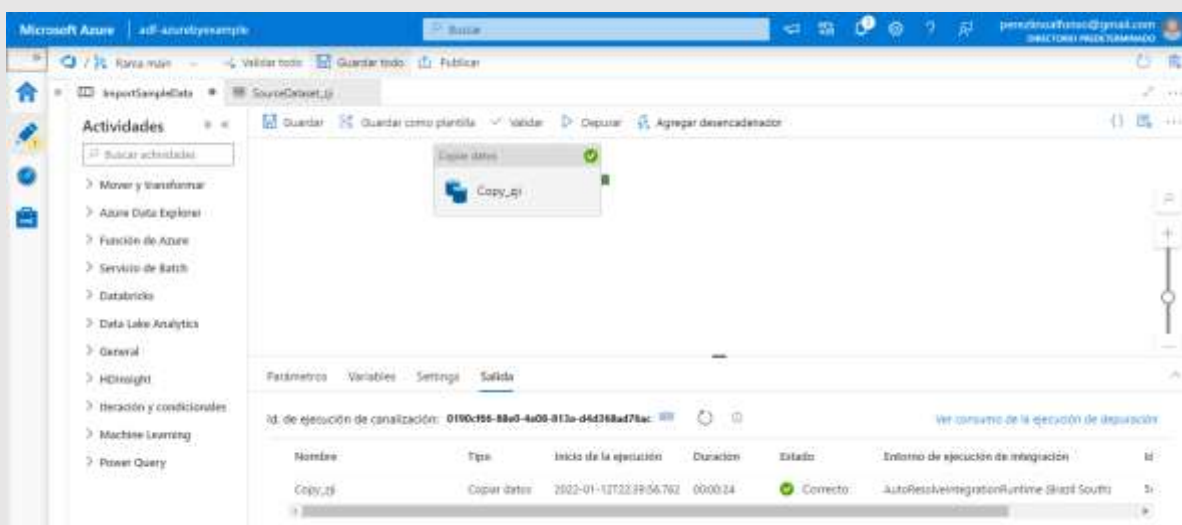




Figura 2-13 Controles de depuración en el lienzo de creación

Tan pronto como haga clic en Debug, el panel de configuración con pestañas debajo del lienzo se expande automáticamente, con la pestaña Output seleccionada. La Figura 2-13 muestra la pestaña después de que la ejecución de la depuración se haya completado con éxito - mientras el pipeline está todavía en ejecución, puede hacer clic en el botón Actualizar para actualizar la pestaña con la última información de estado.

En la parte inferior de la pestaña hay una lista de ejecuciones de actividades realizadas durante la ejecución del pipeline. En este caso, sólo hay una: la actividad Copiar datos. Puede utilizar los iconos situados a la derecha del nombre de una actividad para ver sus entradas, salidas e información de rendimiento. Puede ver la información de consumo de la ejecución del pipeline en su conjunto haciendo clic en **Ver consumo de la ejecución de depuración**.

2.4.2. Inspeccionar los resultados de la ejecución

Un pipeline no publicado sigue conectado a todos los mismos recursos externos que su equivalente publicado, y la ejecución del pipeline accede y modifica esos mismos recursos. La ejecución de su pipeline en modo de depuración ha realizado una copia de datos real desde el contenedor "landing" al contenedor "sampledata".

Vuelva al Azure Storage Explorer para inspeccionar el contenedor "sampledata" y verifique que ahora contiene una estructura de carpetas anidada, extraída del archivo zip en el contenedor "landing". En los siguientes capítulos utilizará los datos descomprimidos en este contenedor.

Revisión del capítulo

En este capítulo, creaste una cuenta de Azure Storage y utilizaste un pipeline de ADF para descomprimir y copiar archivos de un contenedor a otro.

La actividad Copy data utilizada por el pipeline es el caballo de batalla del movimiento de datos en ADF. En su pipeline, la actividad trata los archivos como blobs no estructurados, pero en el Capítulo 3 explorará su manejo de formatos de archivos de texto estructurados y semiestructurados, junto con otros conjuntos de datos estructurados.

Conceptos clave

En este capítulo se han introducido cinco conceptos clave de Azure Data Factory:

- **Pipeline:** Una unidad de carga de trabajo de integración de datos en Azure Data Factory. Una agrupación lógica de actividades reunidas para ejecutar un proceso de integración de datos concreto.
- **Actividad:** Realiza una tarea dentro de un pipeline, por ejemplo, copiar datos de un lugar a otro.
- **Dataset (conjunto de datos):** Contiene metadatos que describen un conjunto específico de datos mantenidos en un sistema de almacenamiento externo. Las actividades del pipeline utilizan conjuntos de datos para interactuar con datos externos.
- **Linked service (servicio vinculado):** Representa una conexión con un sistema de almacenamiento externo o un recurso informático externo.
- **Integration runtime (tiempo de ejecución de la integración):** Proporciona acceso a un recurso informático interno dentro de Azure Data Factory. ADF no tiene recursos de almacenamiento internos.

La Figura 2-10 ilustra la interacción entre estos componentes. Otros conceptos encontrados en este capítulo incluyen

- **Debug (depuración):** Se puede ejecutar un pipeline de forma interactiva desde la UX de ADF utilizando el modo "Debug". Esto significa que la definición del pipeline desde la sesión de ADF UX se ejecuta - no necesita ser publicada en la instancia de fábrica conectada. Durante una ejecución de depuración, un pipeline trata los recursos externos exactamente de la misma manera que en las ejecuciones de pipeline publicadas.
- **Copy Data tool (Herramienta de copia de datos):** Una experiencia de estilo asistente en la UX de ADF que crea un pipeline para copiar datos de un lugar a otro. La he presentado

en este capítulo como una forma rápida de empezar a explorar la estructura de los pipelines, pero en la práctica es poco probable que uses la herramienta muy a menudo.

- **Azure Storage:** La plataforma de almacenamiento gestionada en la nube de Microsoft.
- **Cuenta de almacenamiento:** Se crea una cuenta de almacenamiento para poder utilizar los servicios de Azure Storage.
- **Storage key (Clave de almacenamiento):** Las claves de almacenamiento son tokens utilizados para autorizar el acceso a una cuenta de almacenamiento. Puede gestionar las claves de una cuenta en el portal de Azure.
- **Blob storage (Almacenamiento de blobs):** Almacenamiento de archivos de propósito general (blob), uno de los tipos de almacenamiento que ofrece Azure Storage. Otros tipos de almacenamiento soportados (no descritos aquí) incluyen archivos compartidos, colas y tablas.
- **Contenedor:** Los archivos en el almacenamiento blob se almacenan en contenedores, subdivisiones del almacenamiento blob de una cuenta de almacenamiento. El almacenamiento blob se divide en contenedores sólo en el nivel raíz - no pueden ser anidados.
- **Azure Storage Explorer:** Una app utilizada para gestionar las cuentas de Azure Storage, disponible online y como aplicación de escritorio.
- **Bandwidth (Ancho de banda):** Término utilizado por Microsoft para describir el movimiento de datos hacia y desde los centros de datos de Azure. Los movimientos de datos salientes conllevan una tarifa, a veces denominada tarifa de salida.

Para los desarrolladores de SSIS

La mayoría de los conceptos básicos de Azure Data Factory tienen paralelos familiares en SSIS.

ADF Concept	Equivalent in SSIS
Pipeline	Package
Activity	Task (on control flow surface)
Copy data activity	Used in this chapter like a File System Task. In Chapter 3, you will explore more advanced behavior where it acts like a basic Data Flow Task
Linked service	Project-scoped connection manager (with no schema metadata)
Dataset	Schema metadata, stored in various different places (such as a flat file connection manager or OLE DB data flow source)
Integration runtime	SSIS Windows service
Set of pipelines in ADF UX	SSIS project open in Visual Studio
Set of pipelines published to an ADF instance	SSIS project deployed to SSIS catalog