

An abstract graphic in the top right corner of the cover. It features a dark, semi-circular shape resembling a globe or a data sphere. Inside this shape, there are numerous thin, glowing lines in shades of purple, pink, and blue. These lines radiate from a central point and connect to various nodes, creating a network-like structure. The lines have a slight motion blur effect, giving the impression of data flowing or being processed.

Azure Data Factory by Example

Practical Implementation for
Data Engineers

—
Richard Swinbank

Apress®

Contenido

1. Creación de una instancia de Azure Data Factory	3
1.1. Empezar en Azure	3
1.1.1. Crear una cuenta gratuita de Azure	3
1.1.2. Explorar el portal Azure	4
1.2. Crear un grupo de recursos	6
1.3. Crear una Azure Data Factory	9
1.4. Explore la experiencia de usuario de Azure Data Factory	11
1.4.1. Barra de cabecera de navegación	11
1.4.2. Barra lateral de navegación	13
1.5. Enlace a un repositorio Git	14
1.5.1. Crear un repositorio Git en Azure Repos	14
1.5.2. Vincular la fábrica de datos al repositorio Git	18
1.6. La UX del ADF como un IDE basado en la web	21
Revisión del capítulo	23
Conceptos clave	23
Para los desarrolladores de SSIS	25

1. Creación de una instancia de Azure Data Factory

Una de las principales responsabilidades del ingeniero de datos es el desarrollo y la gestión de la extracción, transformación y carga (ETL) y otras cargas de trabajo de integración de datos. Las cargas de trabajo de integración en tiempo real procesan los datos a medida que se generan, por ejemplo, una transacción que se registra en un terminal de punto de venta o un sensor que mide la temperatura en un centro de datos. En cambio, las cargas de trabajo de integración por lotes se ejecutan a intervalos, normalmente procesando los datos producidos desde la anterior ejecución por lotes.

Azure Data Factory (ADF) es el servicio nativo en la nube de Microsoft para gestionar las cargas de trabajo de integración de datos por lotes. ADF es un ejemplo de servicio en la nube sin servidor: lo utilizas para crear tus propias aplicaciones ETL, pero no tienes que preocuparte de la infraestructura, como sistemas operativos o servidores, ni de cómo gestionar los cambios en la demanda. El acceso al servicio se consigue mediante una instancia de fábrica de datos (a menudo llamada simplemente "fábrica de datos"). La mayor parte de este libro trata de la creación y gestión de pipelines de ADF - data integration workload units escritas y ejecutadas en una instancia de ADF.

Para crear pipelines, primero hay que tener acceso a una instancia de ADF. En este capítulo, creará una nueva instancia de ADF, lista para empezar a crear pipelines en el Capítulo 2. Para empezar, no necesitará más que una conexión a Internet y el navegador web Microsoft Edge o Google Chrome.

Nota Es posible que utilice variaciones de ETL como extraer, cargar y transformar (ELT) o extraer, cargar, transformar y cargar (ELTL). ADF puede utilizarse en cualquiera de estos escenarios de integración de datos, y utilizo el término ETL de forma imprecisa para incluir cualquiera de ellos.

1.1. Empezar en Azure

Para acceder a los servicios en la nube de Microsoft Azure, necesitas una suscripción a Azure. Mi objetivo es que te pongas en marcha sin coste alguno: en las siguientes secciones, te explico cómo crear una suscripción de prueba gratuita a Azure que podrás utilizar a lo largo de este libro, y luego te presento el portal de Azure para interactuar con él.

1.1.1. Crear una cuenta gratuita de Azure

Muchos de los ejercicios del libro requieren permisos de acceso elevados en Azure. Puedes optar por omitir esta sección si ya tienes una suscripción a Azure que prefieres utilizar, pero asegúrate de que te concede el acceso suficiente para crear y modificar recursos.

1. En su navegador web, vaya a <https://azure.microsoft.com> e inicie sesión. Si aún no tiene una cuenta en línea de Microsoft, primero deberá crear una. La experiencia de usuario de Azure Data Factory (introducida más adelante en el capítulo) solo es compatible con Microsoft Edge o Google Chrome, por lo que deberá utilizar uno de esos dos navegadores web.
2. Haz clic en el enlace Cuenta gratuita de la parte superior derecha y, en la siguiente página, haz clic en Iniciar gratis.
3. Sigue el proceso de cuatro pasos para configurar tu cuenta. Durante la configuración de la cuenta, se le pedirá que proporcione información de facturación, pero no se cargará en su tarjeta de crédito a menos que se actualice a una suscripción de pago.

Una vez creada la cuenta con éxito, aparecerá el botón "Ir al portal"; haga clic en él. Si no ve el botón, puede navegar al portal directamente utilizando su URL: <https://portal.azure.com>.

1.1.2. Explorar el portal Azure

El portal de Azure es el lugar donde se gestionan todos los recursos de Azure. Utilizará el portal con frecuencia, por lo que es una buena idea marcar esta página. La página de inicio del portal se parece a la Figura 1-1. Digo "algo parecido" porque puede ver diferentes herramientas, recomendaciones, enlaces u otros mensajes de vez en cuando. Tres o cuatro características están siempre presentes:

1. Si está utilizando una suscripción con límite, aparece brevemente una notificación sobre su crédito restante cuando abre el portal por primera vez. El crédito restante se muestra en la moneda local de su cuenta. El crédito gratuito que se incluye con la suscripción de prueba de Azure está limitado a 30 días.
2. En la página de inicio, encontrará el botón Crear un recurso (icono del signo más). Esta opción también está disponible en el menú del portal, al que se accede mediante el botón de la parte superior izquierda.
3. En la parte superior derecha aparece la dirección de correo electrónico con la que se ha registrado.
4. Inmediatamente debajo de su dirección de correo electrónico se encuentra su directorio actual. Si está utilizando una suscripción de prueba gratuita, dirá DIRECTORIO POR DEFECTO.

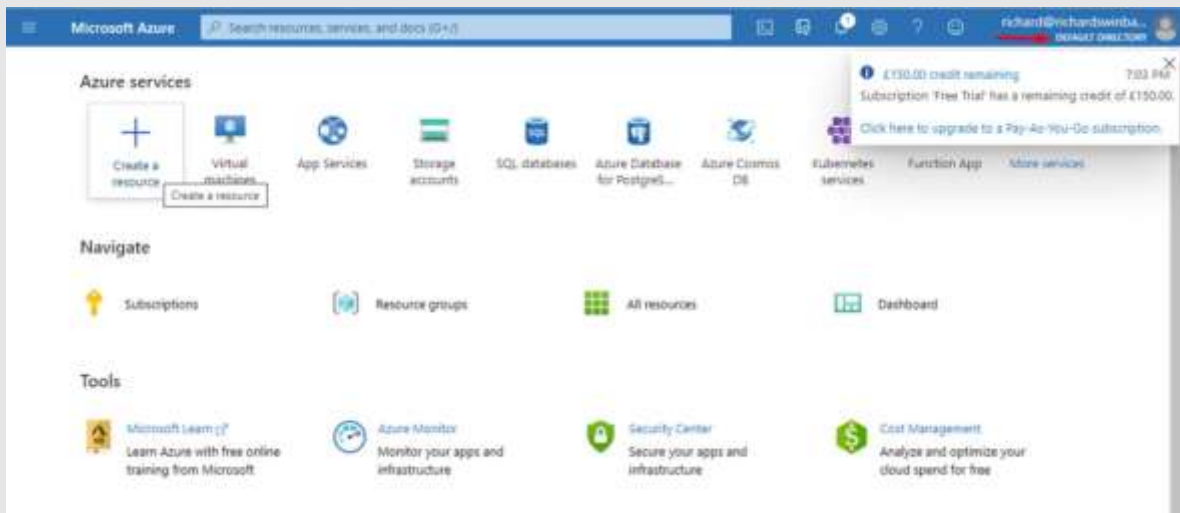


Figura 1-1 Página de inicio del portal de Azure

Su directorio, comúnmente llamado tenant, es una instancia de Azure Active Directory (AAD). "Directorio predeterminado" es el nombre por defecto de un nuevo tenant. Si ya está utilizando Azure en su trabajo, probablemente estará utilizando un tenant que representa a su empresa u organización - a menudo, todos los recursos y usuarios de Azure de una organización están definidos en el mismo tenant.

Un tenant contiene una o más suscripciones. Una suscripción identifica un medio de pago para los servicios de Azure - el coste de usar cualquier recurso de Azure se factura a la suscripción a la que está asociado. Una suscripción de prueba de Azure incluye una cantidad de crédito gratuito limitado en el tiempo, y si quiere gastar más, puede hacerlo actualizando a una suscripción de pago. Su organización puede tener varias suscripciones, tal vez identificando a distintos titulares del presupuesto responsables de pagar por distintos recursos.

El registro de una suscripción de prueba a Azure crea una serie de cosas, entre ellas

- Un tenant de Azure
- Su cuenta de usuario de Azure, con permisos AAD de nivel de administrador dentro del tenant
- Una suscripción a Azure en el tenant con cierto crédito gratuito limitado en el tiempo para su uso

1.2. Crear un grupo de recursos

Las instancias de los servicios de Azure se denominan generalmente recursos. Una instancia de Azure Data Factory es un ejemplo de recurso. Los recursos que pertenecen a una suscripción se organizan además en grupos de recursos. Un grupo de recursos es un contenedor lógico que se utiliza para reunir recursos relacionados, por ejemplo, todos los recursos que pertenecen a una plataforma de almacenamiento de datos o de análisis.

La Figura 1-2 ilustra la agrupación lógica de recursos en Azure. En esta sección, creará un grupo de recursos para contener una instancia de ADF y otros recursos que serán necesarios en capítulos posteriores.

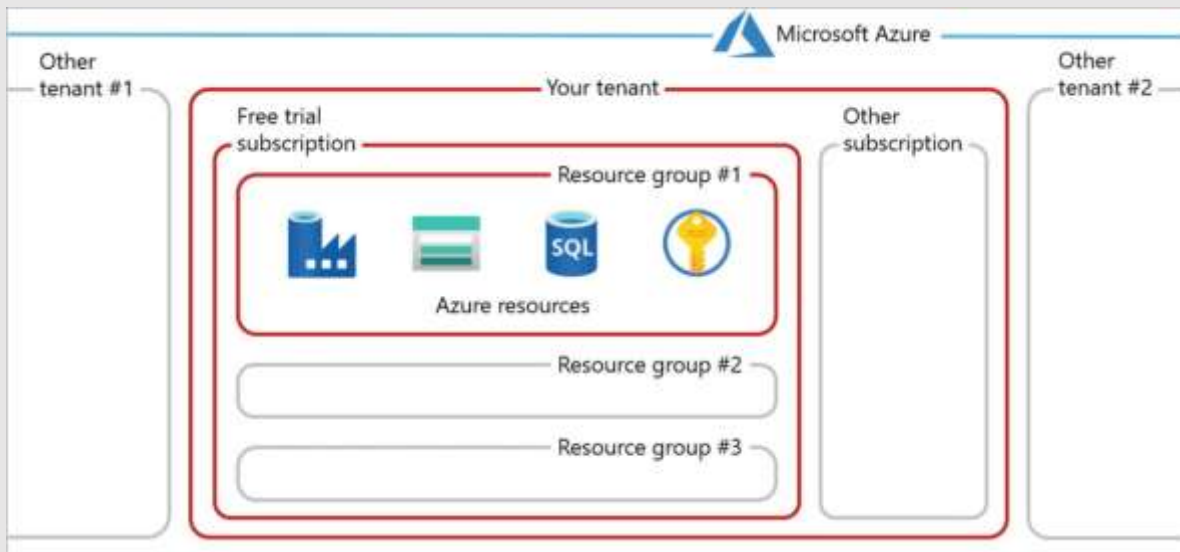


Figura 1-2 Agrupación lógica de recursos en Azure

1. Haga clic en Crear un recurso, utilizando el botón de la página de inicio del portal o el botón del menú en la parte superior izquierda.
2. Las páginas del portal de Azure se denominan blades: el nuevo blade de recursos se muestra en la Figura 1-3. Puede examinar los servicios disponibles utilizando los menús Azure Marketplace o Popular, o puede utilizar la función Buscar en Marketplace. En el cuadro de búsqueda, comience a escribir "grupo de recursos" (sin las comillas). A medida que escriba, aparecerá un menú desplegable filtrado. Cuando vea la opción de menú "Grupo de recursos", haga clic en ella. Esto le llevará a la hoja de resumen del grupo de recursos.

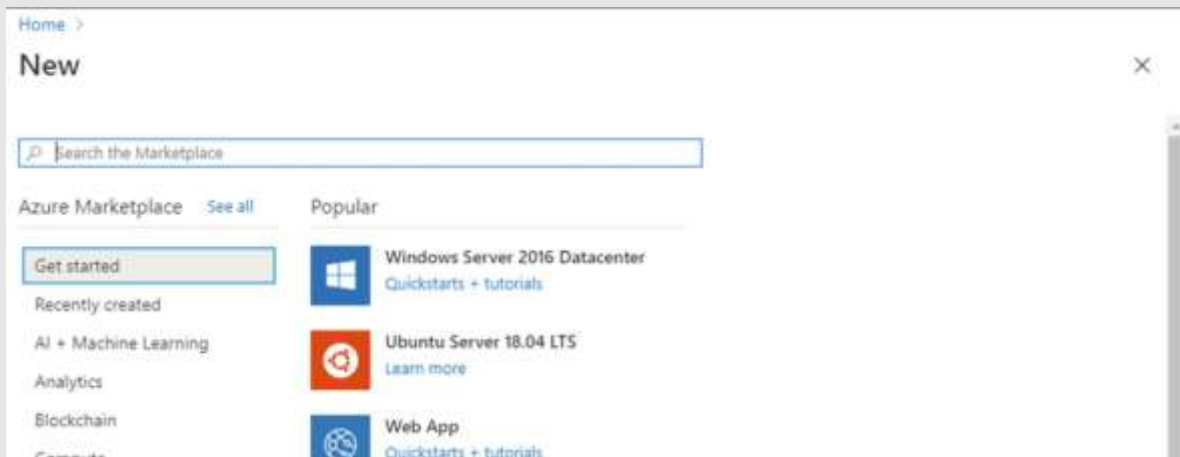


Figura 1-3 Nueva hoja de recursos

3. La hoja de resumen de grupos de recursos ofrece una descripción de los grupos de recursos y un botón de creación. Haga clic en el botón para empezar a crear un nuevo grupo de recursos.
4. Complete los campos de la hoja Crear un grupo de recursos, que se muestra en la Figura 1-4. Asegúrese de que su suscripción de prueba está seleccionada en el campo Suscripción y proporcione un nombre para el nuevo grupo de recursos. Yo utilizo nombres de grupos de recursos que terminan en "-rg" para que sea fácil ver de qué tipo de recurso Azure se trata. Elige una región geográficamente cercana a ti - la mía es "(Europa) Reino Unido Sur", pero la tuya puede ser diferente. Cuando esté listo, haga clic en Revisar + crear.

Figura 1-4 Crear una hoja de grupo de recursos

5. En la pestaña Revisar + crear que aparece a continuación, comprueba los datos que has introducido y haz clic en Crear.

Nota Se dará cuenta de que he omitido la pestaña Etiquetas. En un entorno empresarial, las etiquetas son útiles para etiquetar los recursos de diferentes maneras, por ejemplo, asignando recursos a centros de coste dentro de una suscripción o marcando los recursos sólo de desarrollo para permitir que se detengan automáticamente durante la noche y los fines de semana. No utilizaré etiquetas en este libro, pero su empresa puede utilizar una política de etiquetado de recursos para satisfacer requisitos como estos.

1.3. Crear una Azure Data Factory

El grupo de recursos que ha creado en la sección anterior es un contenedor para los recursos de Azure de cualquier tipo. En esta sección, crearás el primer recurso nuevo del grupo: una instancia de Azure Data Factory.

1. Vuelve a la página de inicio del portal de Azure y haz clic en Crear un recurso, de la misma manera que hiciste al crear tu grupo de recursos.
2. En el cuadro Search the Marketplace de la hoja de recursos nuevos, introduzca "data factory". Cuando "Data Factory" aparezca como un elemento en el menú desplegable, selecciónalo y, a continuación, en la hoja de resumen de la fábrica de datos, haz clic en Crear.
3. Se muestra la pestaña Básica de la hoja Crear fábrica de datos, como se muestra en la Figura 1-5. Seleccione el grupo de Suscripción y Recursos que creó anteriormente y, a continuación, elija la Región más cercana geográficamente.
4. Elija un nombre para su instancia de ADF. Los nombres de las fábricas de datos sólo pueden contener caracteres alfanuméricos y guiones, y deben ser únicos a nivel mundial: el nombre que elija no estará disponible si alguien ya lo está utilizando. Yo uso nombres de fábricas de datos que terminan en "-adf" para que sea fácil ver de qué tipo de recurso de Azure se trata.
5. Establezca la versión como "V2". (Este libro se refiere exclusivamente a Azure Data Factory V2 - ADF V1 sigue estando disponible únicamente para dar soporte a las implementaciones heredadas).

The screenshot shows the 'Create Data Factory' wizard in the Azure portal. The breadcrumb navigation at the top reads 'Home > New > Data Factory >'. The title of the wizard is 'Create Data Factory' with a close button (X) in the top right corner. Below the title, there are tabs for 'Basics', 'Git configuration', 'Networking', 'Advanced', 'Tags', and 'Review + create'. The 'Basics' tab is selected. Under the 'Project details' section, there is a description: 'Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.' Below this, there are two dropdown menus: 'Subscription *' with 'Free Trial' selected, and 'Resource group *' with 'adfbexample-rg' selected. A link 'Create new' is visible below the resource group dropdown. Under the 'Instance details' section, there are three more dropdown menus: 'Region *' with 'UK South' selected, 'Name *' with 'adfbexample-adf' selected (marked with a green checkmark), and 'Version *' with 'V2' selected. At the bottom of the wizard, there are three buttons: 'Review + create' (highlighted in blue), '< Previous', and 'Next : Git configuration >'. The entire form is enclosed in a light gray border.

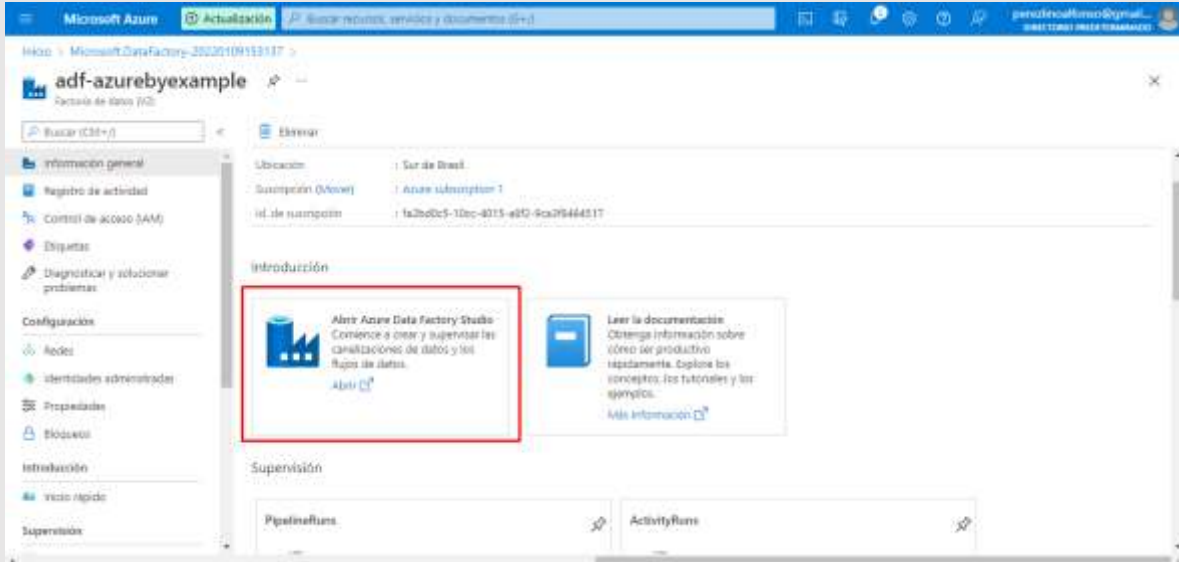
Figura 1-5 Crear hoja de Data Factory

6. Haga clic en el botón **Next: Configuración de Git** y, en la pestaña Configuración de Git, marque la casilla **Configurar Git más tarde**.
7. Por último, haga clic en **Revisar + crear**, compruebe la configuración de fábrica que proporcionó en los pasos 3 a 6, y luego haga clic en Crear para iniciar el despliegue. (Estoy omitiendo a propósito las tres pestañas restantes -Red, Avanzado y Etiquetas- y aceptando sus valores por defecto).

Cuando se inicia el despliegue, aparece una nueva hoja con el mensaje El despliegue está en curso. La creación de una nueva instancia del ADF no suele tardar más de 30 segundos, tras los cuales se mostrará el mensaje Su despliegue ha finalizado. Haga clic en Ir al recurso para inspeccionar su nueva fábrica de datos.

La hoja del portal que se muestra al hacer clic en **Ir al recurso** proporciona una visión general de su instancia de fábrica de datos. Contiene controles de acceso y otras herramientas estándar de recursos de Azure, junto con información de supervisión y detalles básicos sobre la fábrica, por ejemplo, su suscripción, grupo de recursos y ubicación. El portal no proporciona herramientas para trabajar dentro del ADF.

Debajo de los detalles básicos de la fábrica, encontrará dos mosaicos: Documentación y Autor y Monitor. Haga clic en el mosaico Author & Monitor para iniciar la experiencia de usuario de Azure Data Factory. Aquí es donde pasará la mayor parte de su tiempo cuando trabaje con ADF.



1.4. Explore la experiencia de usuario de Azure Data Factory

La experiencia de usuario de Azure Data Factory (ADF UX) proporciona un entorno de desarrollo integrado (IDE) sin código para crear pipelines de ADF, publicarlos y luego programar y supervisar su ejecución. Utilizará la UX de ADF con frecuencia, por lo que es una buena idea marcar esta página.

La Figura 1-6 muestra la página de visión general de ADF UX. Dentro de la UX, puede volver a esta página haciendo clic en el botón de visión general de Data Factory (icono de inicio) en la barra lateral de navegación.

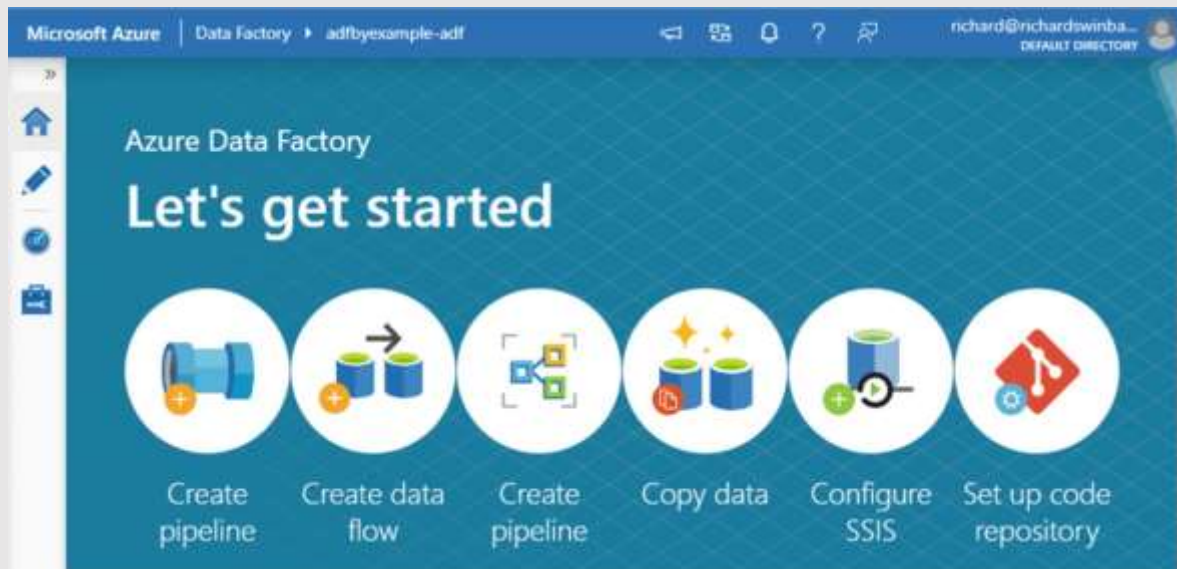


Figura 1-6 Página de visión general de la fábrica de datos del ADF UX

La página de visión general tiene tres regiones:

- Una barra de cabecera de navegación
- Una barra lateral de navegación ampliable
- Un panel de contenido, que actualmente muestra la visión general de Data Factory.

La barra de cabecera de navegación y la barra lateral están visibles en todo momento, dondequiera que se encuentre en el ADF UX. El panel de contenido muestra diferentes cosas, dependiendo de la parte de la UX que esté utilizando.

1.4.1. Barra de cabecera de navegación

La Figura 1-7 muestra la UX del ADF con la barra lateral de navegación expandida y las funciones de la barra de cabecera de navegación etiquetadas. Para mayor claridad, se ha eliminado el panel de contenido de la captura de pantalla.

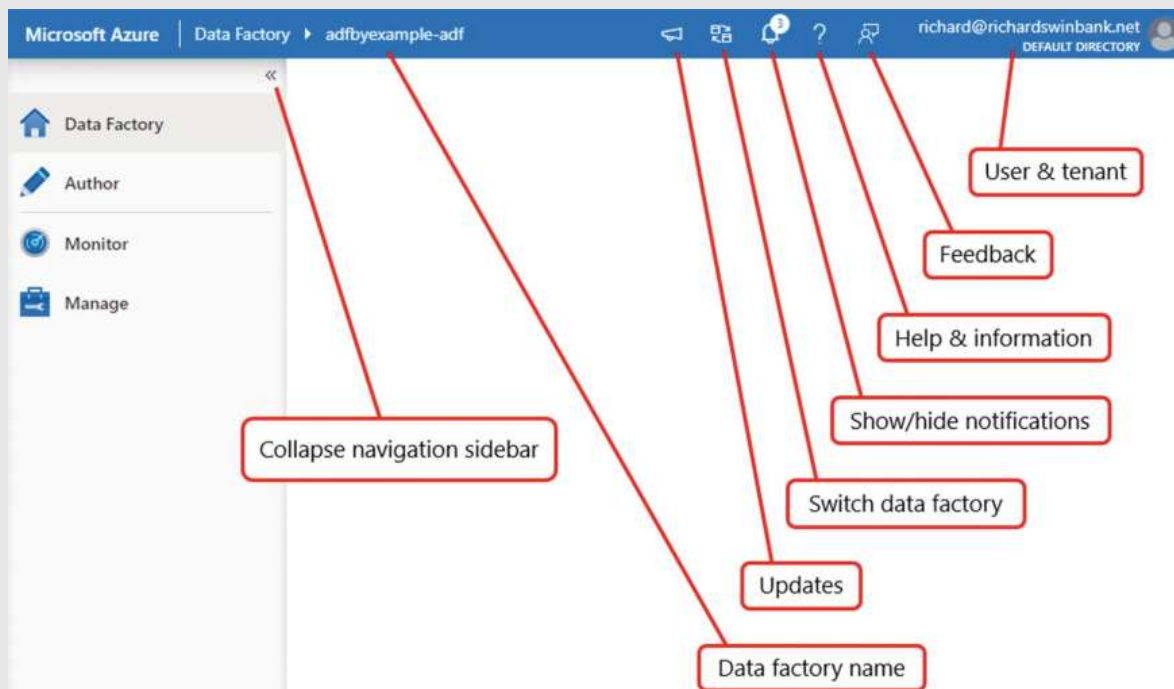


Figura 1-7 Barra de cabecera de navegación de ADF UX etiquetada

Hacia su extremo izquierdo, la barra de cabecera de navegación indica el nombre de la instancia de fábrica de datos a la que está conectada la UX de ADF. En su otro extremo, identifica el usuario y el tenant actuales, de la misma manera que en el portal de Azure. Entre ambos hay una fila de cinco botones:

- **Actualizaciones:** Muestra las actualizaciones recientes del servicio Azure Data Factory. ADF está en constante desarrollo y evolución: aquí se anuncian los cambios en el servicio a medida que se producen.
- **Cambiar de Data Factory:** Le permite desconectarse de la instancia actual de ADF y conectarse a una diferente.

Nota Cuando se abre el ADF UX desde la hoja de fábrica de datos del portal de Azure, se conecta automáticamente a la nueva fábrica. De hecho, el ADF UX siempre está conectado a una instancia de ADF. Si accede a ella directamente (utilizando la URL <https://adf.azure.com/>), se le pedirá que seleccione una fábrica de datos antes de que se abra el ADF UX.

- **Mostrar notificaciones:** El ADF UX le notifica automáticamente los eventos que ocurren durante su sesión - este botón alterna la visualización de esas notificaciones. El "3" marcado con un círculo en la captura de pantalla indica que actualmente hay tres notificaciones sin leer.
- **Ayuda/información:** Proporciona enlaces a soporte e información adicional sobre el ADF.

- **Comentarios:** Si desea proporcionar a Microsoft comentarios sobre su experiencia con Azure Data Factory, puede hacerlo aquí.

1.4.2. Barra lateral de navegación

La barra lateral de navegación proporciona acceso a diferentes partes de la UX de ADF, cambiando lo que se muestra en el panel de contenido. El icono del cheurón en la parte superior de la barra lateral cambia su estado entre colapsado y expandido - en la Figura 1-6, la barra lateral está colapsada, mientras que la Figura 1-7 la muestra expandida.

- El botón de visión general de Data Factory (icono de inicio) le devuelve a la página de visión general. Esta página contiene enlaces rápidos a una serie de herramientas para apoyar las tareas comunes de ADF, junto con enlaces a vídeos, tutoriales y otros recursos de aprendizaje. Usted utilizará una de las herramientas aquí en el Capítulo 2.
- El botón Autor (icono del lápiz) carga el espacio de trabajo de autoría del ADF. El espacio de trabajo de autoría proporciona un editor visual para construir pipelines de ADF. Dado que este libro trata principalmente de la creación de pipelines, pasará mucho tiempo aquí.
- El botón de monitorización (icono de un medidor) proporciona acceso a las herramientas de monitorización visual. Aquí podrá ver las ejecuciones de pipelines ADF ejecutadas en la instancia de la fábrica y profundizar en los detalles de la ejecución. El capítulo 12 examina la experiencia de monitorización con más detalle.
- El botón Manage (icono de la caja de herramientas) carga el centro de gestión de ADF. Esto incluye una variedad de características como las conexiones al almacenamiento de datos externo y los recursos de computación, junto con la configuración Git de la instancia de ADF, introducida en la siguiente sección. Volverá al centro de gestión en varias ocasiones a lo largo de este libro.

1.5. Enlace a un repositorio Git

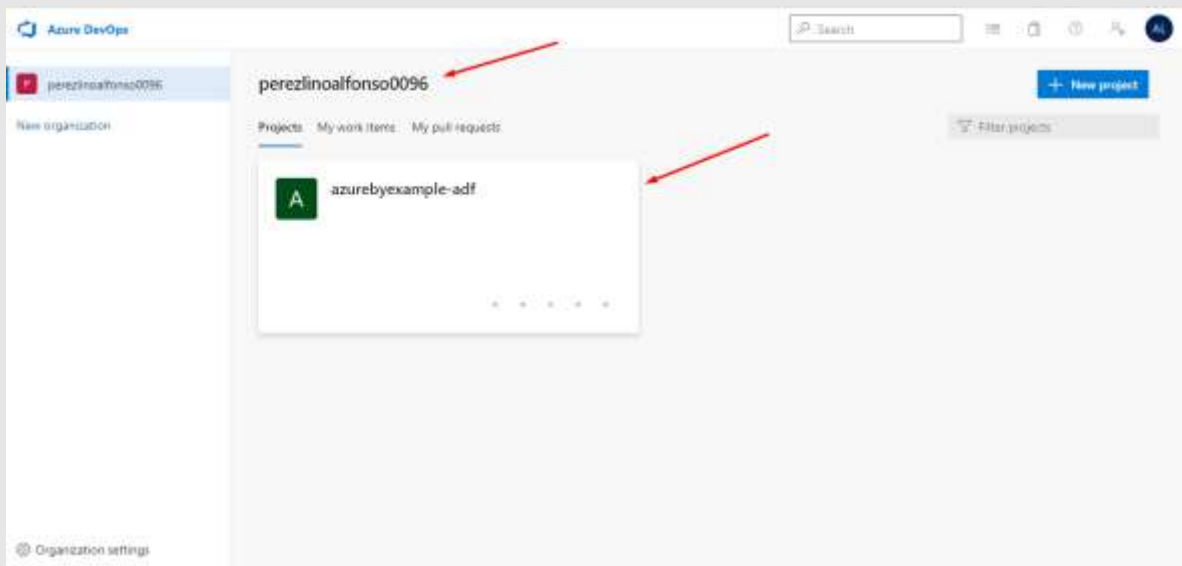
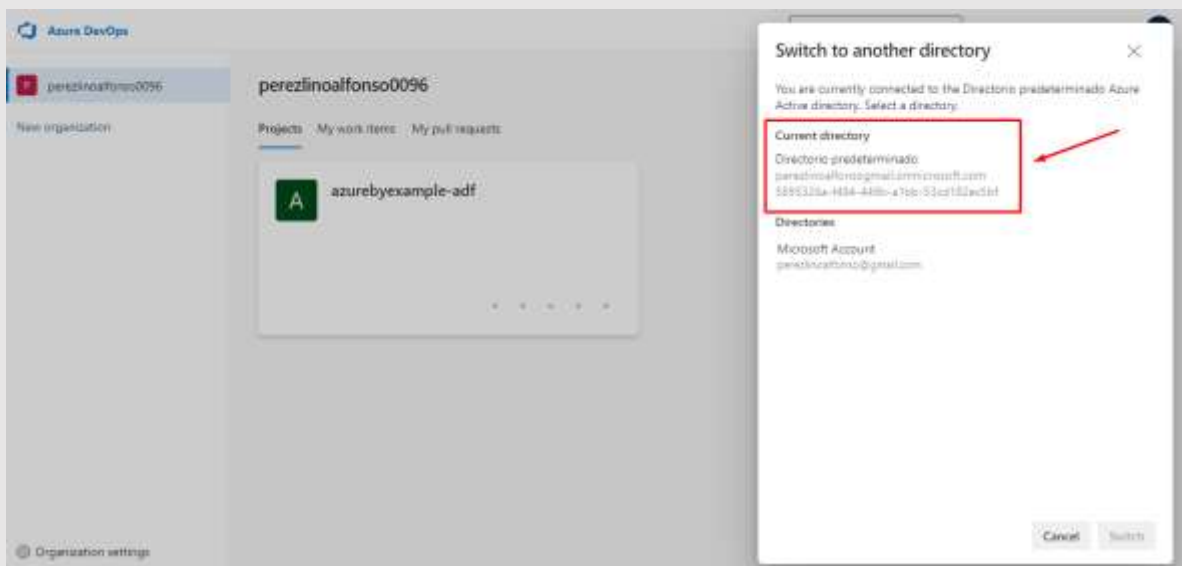
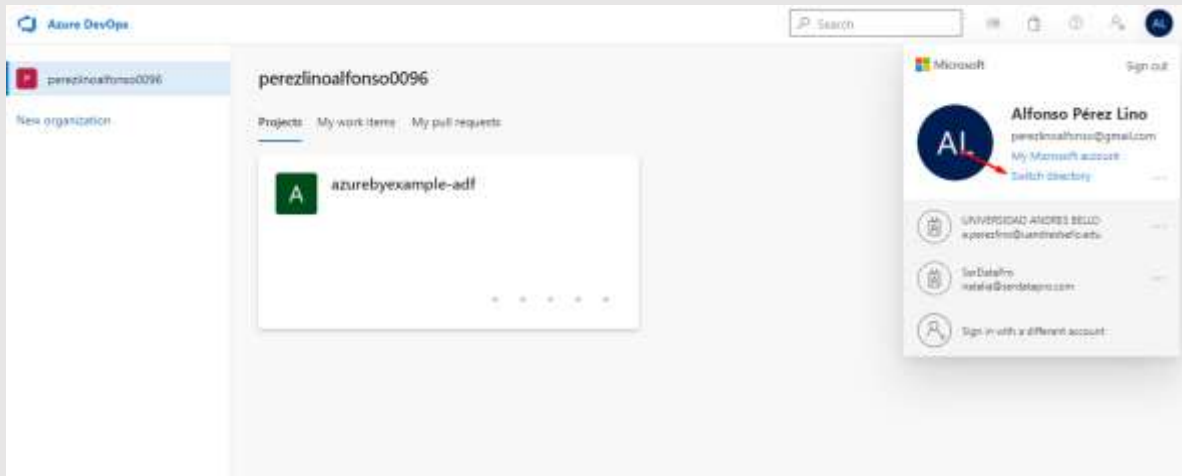
Una instancia de fábrica de datos puede ponerse bajo control de código fuente vinculándola a un repositorio Git basado en la nube. Aunque es posible emprender el trabajo de desarrollo en ADF sin vincular su fábrica de datos a un repositorio Git, hay muchas desventajas de hacerlo - sin un repositorio vinculado, incluso guardar el trabajo en curso es difícil. Antes de empezar a trabajar en su nueva instancia de ADF, deberá vincularla a un repositorio Git.

Consejo Es más fácil configurar el repositorio Git de una fábrica de datos desde la UX de ADF que desde el portal de Azure - esta es la razón por la que elegiste la opción Configurar Git más tarde cuando creaste tu fábrica de datos.

1.5.1. Crear un repositorio Git en Azure Repos

Antes de vincular una fábrica de datos a un repositorio Git, necesitas un repositorio Git al que se pueda vincular. El soporte para los diferentes proveedores de servicios Git varía entre los diferentes servicios de Azure - actualmente, una instancia de ADF puede ser vinculada a un repositorio Git proporcionado por Azure Repos o GitHub. Azure Repos es una de las herramientas para desarrolladores nativos en la nube proporcionadas por Azure DevOps Services. Los repositorios Git (y otras instancias de servicio) proporcionados por Azure DevOps se agrupan en proyectos - en esta sección, crearás una organización gratuita de Azure DevOps para alojar un proyecto, y luego inicializarás un repositorio Git en el nuevo proyecto.

1. Navega hasta <https://microsoft.com/devops> e inicia sesión, utilizando la misma cuenta que utilizaste para crear tu tenant de Azure. Haz clic en Iniciar gratis.
2. Aparece la página Get started with Azure DevOps, como se muestra en la Figura 1-8. Cerca de la parte superior del cuadro de diálogo se muestra la dirección de correo electrónico con la que ha iniciado sesión y un enlace Switch directory (indicado en la figura). Esto indica el directorio Azure (tenant) que su nueva organización Azure DevOps estará conectada. Utilice el enlace Cambiar de directorio para verificar que el tenant seleccionado es el que contiene su fábrica de datos y, a continuación, haga clic en Continuar.



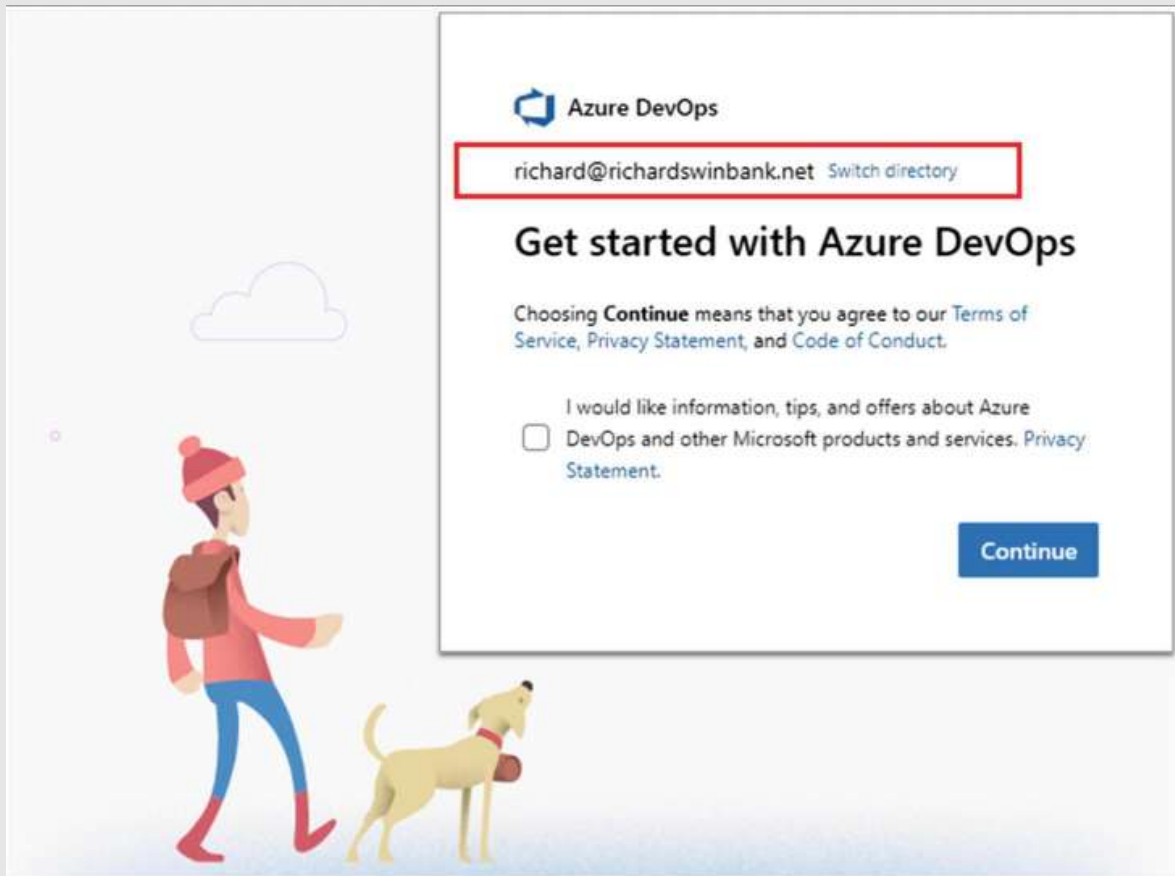


Figura 1-8 Diálogo de inicio indicando el tenant de Azure a enlazar

Consejo Crear su instancia ADF y su repositorio Git en el mismo tenant no es esencial, pero hacerlo simplifica la integración entre ellos.

3. Azure DevOps crea una nueva organización para usted - si se le pide, proporcione un nombre para ella - y luego muestra el panel Crear un proyecto para empezar. Elige un nombre para tu proyecto e introdúcelo en el campo Nombre del proyecto. Establezca la Visibilidad del proyecto en "Privado", y luego haga clic en + Crear proyecto.
4. Aparece la página de bienvenida del nuevo proyecto, como se muestra en la Figura 1-9. Elige comenzar con el servicio Azure Repos, ya sea haciendo clic en el botón Repos de la página de bienvenida o seleccionando Repos (botón rojo con icono de rama) en la barra lateral de navegación.

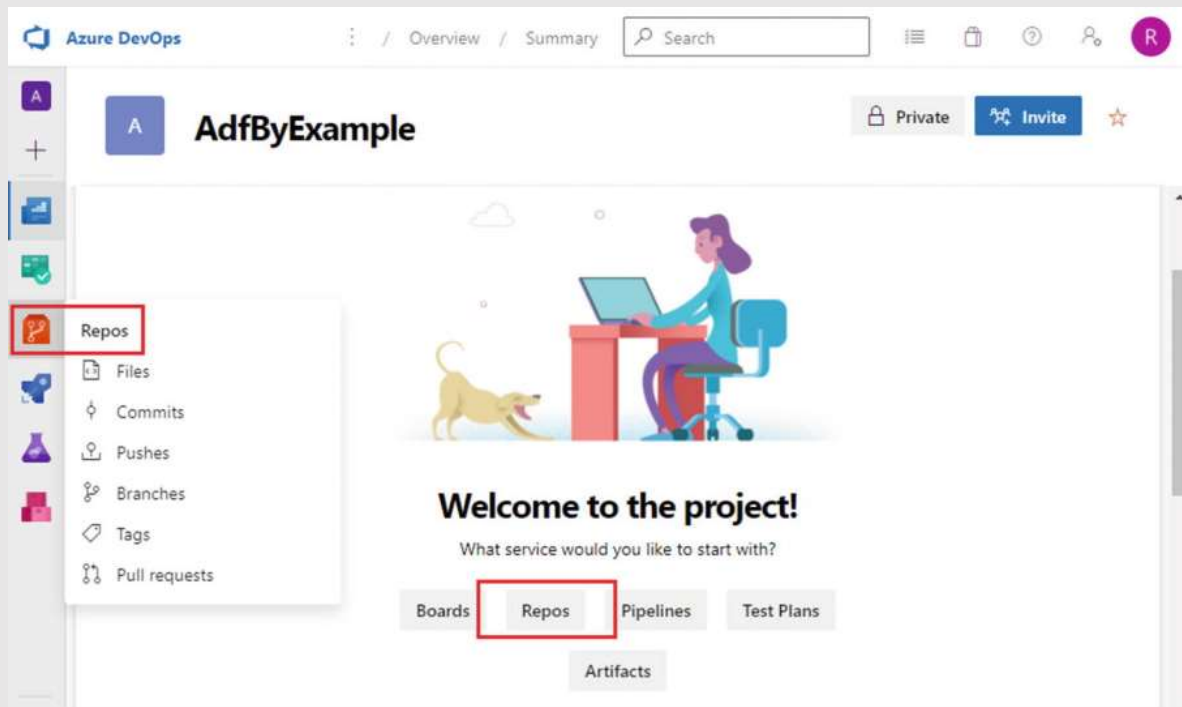
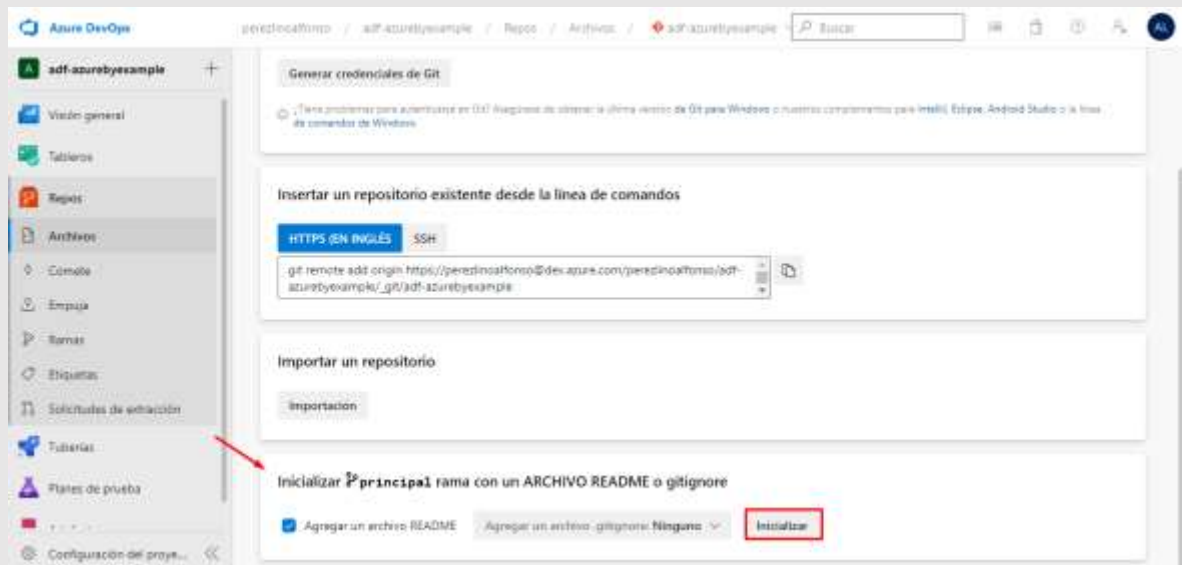


Figura 1-9 Página de bienvenida del proyecto Azure DevOps

- Debido a que aún no existen repositorios, Azure DevOps indica que su proyecto está vacío. Desplácese hacia abajo hasta el encabezado Initialize main branch with a README or gitignore, y haga clic en Initialize para crear un nuevo repositorio con el mismo nombre que su proyecto.



Puedes elegir vincular una fábrica de datos a un repositorio Git proporcionado por Azure Repos o por GitHub. He elegido un repositorio de Azure Repos porque al hacerlo la integración con otros servicios de Microsoft es ligeramente más sencilla y porque más adelante en el libro utilizarás otro servicio proporcionado por Azure DevOps.

1.5.2. Vincular la fábrica de datos al repositorio Git

Vincular la fábrica de datos al repositorio Git

En esta sección, enlazará su instancia de ADF con su nuevo repositorio Git.

1. Vuelva a la UX de ADF y abra el centro de gestión haciendo clic en Gestionar (icono de caja de herramientas) en la barra lateral de navegación.
2. En la sección de control de fuentes del menú del centro de gestión, haga clic en Configuración de Git.
3. El panel de contenido indica que no hay ningún repositorio configurado, como se muestra en la Figura 1-10. Haga clic en el botón central Configurar para conectar la instancia de fábrica a su repositorio Git.

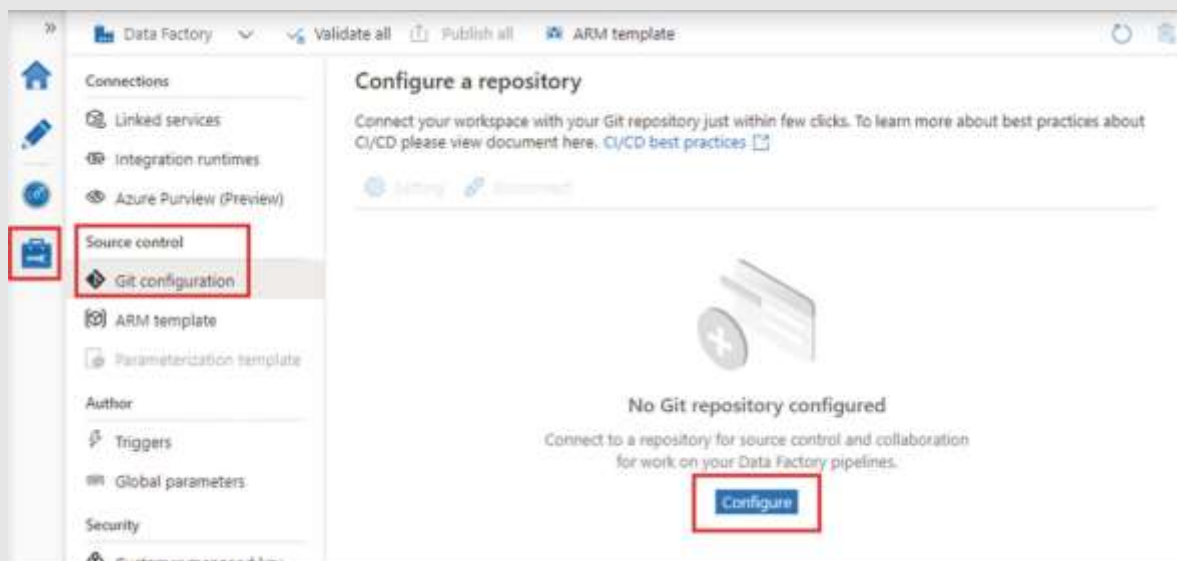


Figura 1-10 Configurar un repositorio Git en el hub de gestión de ADF UX

4. Se abre la hoja Configurar un repositorio. Elija "Azure DevOps Git" en el menú desplegable de tipo de repositorio. Al hacerlo, aparecen más listas desplegables: selecciona tu tenant de Azure en la lista de Azure Active Directory, y luego elige la **organización Azure DevOps** que creaste en la sección anterior en el desplegable Azure DevOps Account.

5. Cuando aparezcan más opciones, selecciona el proyecto Azure DevOps que creaste en la sección anterior en el desplegable **Nombre del proyecto**, luego en **Nombre del repositorio**, selecciona "Use existing." Elige tu repositorio recién creado en la lista desplegable.
6. Establezca la **Rama de Colaboración** de la fábrica como "main" y acepte el valor por defecto de "adf_publish" para la **Rama de Publicación**. 7. Establezca el valor de la **Carpeta raíz** a "/data-factory-resources". Es una buena práctica almacenar los recursos de la fábrica en una subcarpeta del repositorio (en lugar de en la propia raíz del repositorio), ya que permite segregar los archivos gestionados por ADF de cualquier otro archivo almacenado en el mismo repositorio Git.
7. El formulario correctamente rellenado, incluyendo los valores por defecto para el resto de ajustes, se muestra en la Figura 1-11. 8. Haga clic en Aplicar para vincular la fábrica de datos al repositorio Git.

Configure a repository

Specify the settings that you want to use when connecting to your repository.

Repository type * ⓘ
Azure DevOps Git

Azure Active Directory ⓘ
Default Directory (795c19da-59e7-4d09-a8b1-25b33f7ca145)

Azure DevOps Account * ⓘ
richard0343

Project name * ⓘ
AdfByExample

Repository name * ⓘ
☐ Create new ☒ Use existing
AdfByExample

Collaboration branch * ⓘ
main

Publish branch * ⓘ
adf_publish

Root folder * ⓘ
/data-factory-resources

Import existing resource
☒ Import existing resources to repository

Import resource into this branch * ⓘ
☒ Use Collaboration ☐ Create new ☐ Use existing
main

Apply **Cancel**

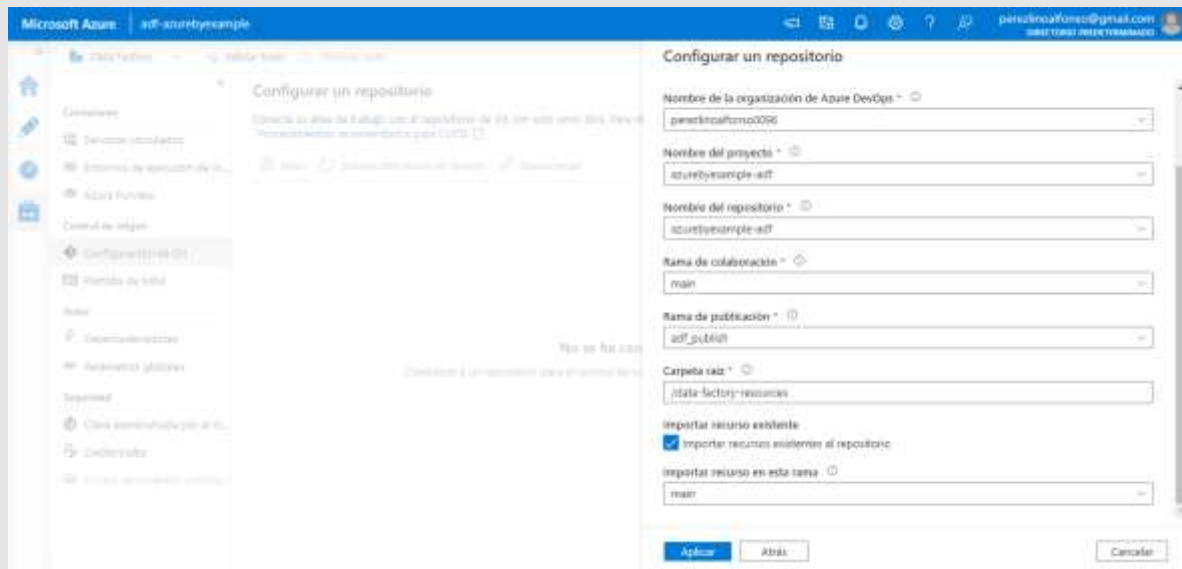


Figura 1-11 Vinculación de un repositorio Git de Azure DevOps a una fábrica de datos

Cuando una instancia de ADF está vinculada a un repositorio Git, el logotipo y la etiqueta de "Data Factory" en la parte superior izquierda de la UX de ADF (visible en la Figura 1-11) se sustituyen por el logotipo del servicio de repositorio Git seleccionado. Inmediatamente a su derecha, se muestra el nombre de su rama de trabajo, por defecto la rama de colaboración del repositorio.

1.6. La UX del ADF como un IDE basado en la web

Si tienes experiencia con casi cualquier otro tipo de trabajo de desarrollo, entonces la relación entre una instancia de fábrica de datos, Git y la UX del ADF puede parecer extraña. En un modelo de desarrollo "tradicional", podría utilizar una herramienta instalada localmente como Visual Studio para crear desarrollos en su propio ordenador. Visual Studio le permite depurar su trabajo utilizando la potencia de cálculo local de su propia máquina y almacena la configuración del repositorio Git localmente para apoyar el control de origen.

En esta situación hipotética, cuando se completa un trabajo de desarrollo, los cambios se despliegan en los servidores o servicios de destino. Puede haber herramientas adicionales para supervisar el rendimiento del entorno publicado: el portal de Azure ofrece una funcionalidad de este tipo para muchos servicios de Azure. La Figura 1-12 muestra la disposición de alto nivel de los componentes en este modelo. Muestra dos posibles rutas para la publicación de cambios en el servicio - ya sea directamente desde el entorno de desarrollo o, como es cada vez más común, a través de despliegues automatizados desde el repositorio de control de origen.

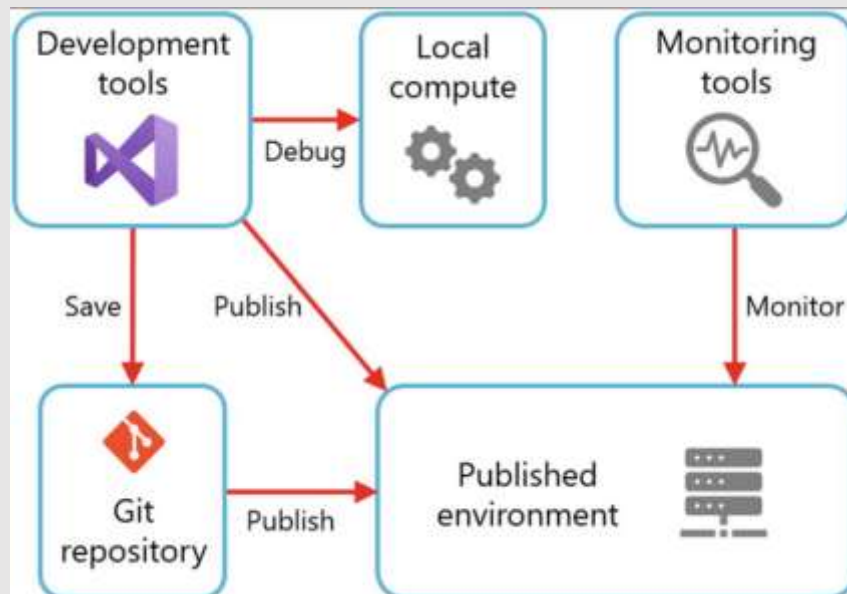


Figura 1-12 Componentes de alto nivel en un modelo de desarrollo "tradicional"

Para los desarrolladores de SSIS Esta disposición de componentes será familiar para los usuarios de SQL Server Integration Services (SSIS). Normalmente, los paquetes SSIS se crean en proyectos SSIS de Visual Studio y se envían al control de origen utilizando un plugin instalado. Cuando están listos, los proyectos SSIS se publican en un catálogo SSIS, donde las herramientas de generación de informes proporcionadas por SQL Server Management Studio (SSMS) permiten supervisar el comportamiento y el rendimiento del paquete.

Al desarrollar para Azure Data Factory, una diferencia significativa es que su IDE - el ADF UX - está basado en la web. Esto significa que el entorno de desarrollo no cuenta con un ordenador local propio. Para poder depurar los pipelines, el ADF UX debe estar conectado a la computación basada en la nube, por lo que el ADF UX siempre está conectado a una instancia de Data Factory. El ADF UX tampoco tiene almacenamiento propio - toda su información de configuración proviene de la fábrica

de datos conectada - por lo que la instancia conectada es el único lugar disponible para almacenar la configuración del repositorio Git. Aunque es la instancia de fábrica de datos la que está vinculada a un repositorio Git, el objetivo subyacente es en realidad vincular el entorno de desarrollo de ADF UX con el control de fuentes a través de la instancia de fábrica conectada.

La Figura 1-13 muestra la disposición equivalente de los componentes de Azure Data Factory. Los recursos de computación de la instancia de fábrica -descrita aquí informalmente como "computación de fábrica"- se utilizan tanto en el entorno de depuración (mientras se desarrolla en ADF UX) como en el entorno de publicación (después de que se hayan publicado los pipelines). Las herramientas de desarrollo y gestión, frecuentemente separadas en el modelo de desarrollo tradicional, están unificadas para Azure Data Factory por el ADF UX.

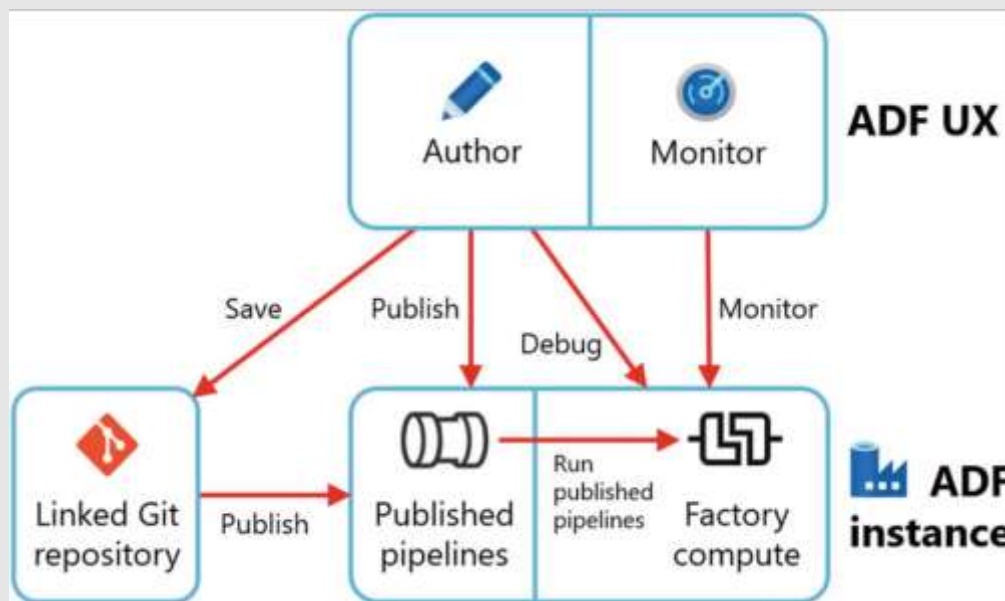


Figura 1-13 Componentes de alto nivel en el modelo de desarrollo de ADF

Revisión del capítulo

En este capítulo, has creado una suscripción de Azure y un grupo de recursos dentro de ella. A continuación, has creado una instancia de Azure Data Factory habilitada para Git, utilizando un repositorio alojado en Azure DevOps Repos. En el próximo capítulo, empezarás a utilizar tu nueva fábrica de datos creando tu primer pipeline de ADF.

Conceptos clave

- **Azure tenant:** Un tenant (o directorio) es una instancia de Azure Active Directory (AAD). Identifica a los usuarios de Azure y les permite acceder a los servicios de Azure.
- **Recurso:** Una instancia de un servicio de Azure, que permite que los usuarios de un tenant accedan al servicio.
- **Suscripción:** Especifica un medio de pago para los servicios de Azure. Un tenant puede contener una o más suscripciones. Cada recurso de Azure está asociado a una suscripción a la que se pueden facturar los cargos por consumo.
- **Grupo de recursos:** Una subagrupación lógica de recursos de Azure dentro de una suscripción.
- **Portal de Azure:** Consola basada en el navegador para crear, supervisar y gestionar los recursos de Azure.
- **Hoja (Blade):** Las páginas del portal de Azure basado en la web se denominan blades.
- **Software como servicio (SaaS):** Modelo de entrega en el que las aplicaciones de usuario final se entregan como servicios en la nube. Las aplicaciones de Microsoft Office 365 son ejemplos de servicios SaaS.
- **Plataforma como servicio (PaaS):** Modelo de entrega en el que los servicios de aplicaciones y las herramientas para desarrolladores se entregan como servicios en la nube. Los sistemas de gestión de bases de datos que no requieren la gestión de la infraestructura del servidor subyacente -por ejemplo, Azure SQL Database- son ejemplos de servicios PaaS.
- **Servicio sin servidor (Serverless service):** Al igual que los PaaS, los servicios en la nube sin servidor proporcionan aplicaciones que normalmente no son utilizadas directamente, excepto por los especialistas en TI, pero con características adicionales de gestión automatizada - por ejemplo, la capacidad de auto-escalar en respuesta a los cambios en la demanda. Azure Data Factory es un ejemplo de servicio sin servidor.
- **Infraestructura como servicio (IaaS):** Modelo de prestación de servicios en la nube en el que los componentes de la infraestructura informática -por ejemplo, los servidores virtuales y las redes- se proporcionan sin necesidad de gestionar el hardware físico correspondiente.

Las máquinas virtuales le dan la libertad de elegir su propio sistema operativo y software de aplicación, pero con el coste de gestionar usted mismo la instalación y las actualizaciones.

- **Azure Data Factory (ADF):** El servicio nativo en la nube y sin servidor de Microsoft para gestionar el ETL por lotes y otras cargas de trabajo de integración de datos.
- **Instancia de fábrica de datos (Data factory instance):** Una instancia del servicio Azure Data Factory, a menudo referida simplemente como "una fábrica de datos". Una instancia de ADF es un ejemplo de recurso de Azure.
- **ADF V2:** La versión 2 del servicio ADF es la opción más reciente y preferida para las instancias de Azure Data Factory. La versión 1 de ADF sólo permanece disponible para dar soporte a las implementaciones heredadas.
- **Entorno de desarrollo integrado (IDE):** Aplicación utilizada para el desarrollo de software que incluye funciones de apoyo como la edición, la depuración y las capacidades de punto de interrupción. Visual Studio es un ejemplo de IDE que puede utilizarse con muchas tecnologías de Microsoft.
- **Experiencia de usuario de Azure Data Factory (ADF UX):** IDE sin código que permite la autoría visual, la depuración y la publicación de pipelines de ADF. Además, la UX de ADF incluye herramientas para programar las canalizaciones publicadas y supervisar su ejecución.
- **Espacio de trabajo de autoría:** Editor visual en la UX de ADF donde se crean, modifican y depuran los pipelines de ADF. Utilice el botón de autor (icono del lápiz) en la barra lateral de navegación para acceder al espacio de trabajo.
- **Experiencia de supervisión (Monitoring experience):** Característica de la UX del ADF que le permite inspeccionar el historial de ejecución de las canalizaciones. Utilice el botón Monitor (icono de calibre) en la barra lateral de navegación para acceder a esta herramienta.
- **Centro de gestión (Management hub):** Área de la UX del ADF que se utiliza para gestionar diversas características de la fábrica. En este capítulo, ha utilizado el centro de gestión para editar la configuración Git de la instancia de ADF, y volverá a él en capítulos posteriores. Utilice el botón Manage (icono de la caja de herramientas) en la barra lateral de navegación para abrir el centro de gestión.
- **Habilitado para Git (Git-enabled):** Una instancia de ADF vinculada a un repositorio Git se dice que está "habilitada para Git". La vinculación de una fábrica de datos a un repositorio Git permite consignar el trabajo de desarrollo de la UX de ADF en el control de origen.
- **Rama de colaboración (Collaboration branch):** Rama de Git, normalmente la rama por defecto del repositorio, donde se realizan los cambios destinados a la publicación en ADF.

- **Rama de publicación (Publish branch):** Rama reservada utilizada por el ADF para dar soporte a los procesos de publicación del pipeline. La publicación en el ADF es el tema del capítulo 10.
- **Carpeta raíz (Root folder):** Ruta en un repositorio Git donde una instancia de ADF vinculada guarda sus archivos. Elegir una carpeta raíz de ADF por debajo del nivel de la raíz del repositorio es una buena práctica, porque permite gestionar otros recursos de la plataforma de datos en el mismo repositorio Git.
- **Entorno publicado (Published environment):** Se refiere al entorno en el que se publican los pipelines de ADF después del desarrollo; el entorno en el que se realiza el trabajo real de integración de datos.
- **Entorno de depuración (Debugging environment):** Se refiere al uso de recursos de fábrica por parte de ADF UX para ejecutar pipelines en desarrollo con fines de depuración.

Para los desarrolladores de SSIS

La UX de ADF proporciona una colección de herramientas con propósitos familiares para los usuarios de SQL Server Integration Services, pero vinculadas entre sí de una manera nueva:

- El espacio de trabajo de autoría proporciona un IDE visual para el desarrollo de tuberías, equivalente al uso de Visual Studio para desarrollar un proyecto SSIS.
- La experiencia de monitorización le permite inspeccionar el resultado de las canalizaciones ADF publicadas, de forma similar a la utilización de los informes del catálogo SSIS en SSMS para ver el historial de ejecución de los paquetes.

Utilizar la UX de ADF es como editar un proyecto SSIS emparejado permanentemente con un único servidor de Integration Services. La conexión de la UX de ADF a una fábrica de datos es necesaria porque la UX de ADF basada en la web no tiene potencia de cálculo ni almacenamiento propios.