Learn about how to connect to data from the Azure Machine Learning workspace. You're introduced to datastores and data assets.

**Learning objectives**

In this module, you learn how to:

* Access data by using Uniform Resource Identifiers (URIs).
* Connect to cloud data sources with datastores.
* Use data asset to access specific files or folders.

**Introducción**

Los datos son un elemento fundamental en cualquier carga de trabajo de machine learning. Se necesitan datos para entrenar un modelo, y también se generan datos al usar un modelo para realizar predicciones.

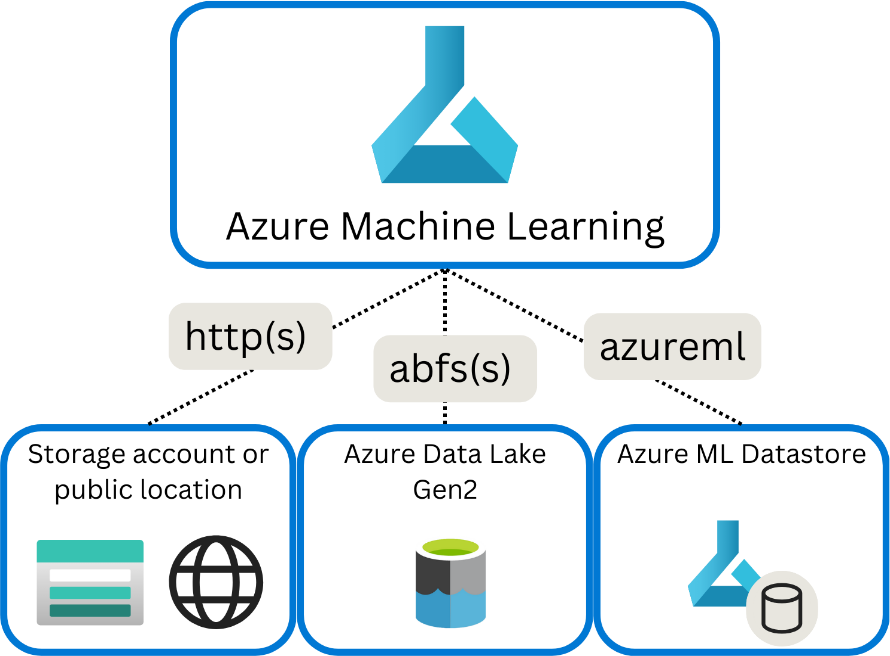
En Azure Machine Learning, puedes acceder a los datos usando Identificadores Uniformes de Recursos (URIs). Cuando trabajas de manera recurrente con una fuente de datos o un archivo o carpeta específicos, puedes crear *datastores* y *activos de datos* en el espacio de trabajo. Estos permiten almacenar de forma segura la información de conexión a tus datos.

En este módulo, aprenderás a crear y usar URIs, datastores y activos de datos en Azure Machine Learning.

**Comprender los URIs**

Puedes almacenar datos en tu dispositivo local o en la nube. Dondequiera que almacenes tus datos, es importante acceder a ellos para entrenar modelos de machine learning. En Azure Machine Learning, se utilizan *Identificadores Uniformes de Recursos* (URIs) para localizar y acceder a estos datos.

Un URI referencia la ubicación de tus datos, y para que Azure Machine Learning pueda conectarse a ellos, es necesario anteponer el URI con el protocolo adecuado. Existen tres protocolos comunes al trabajar con datos en Azure Machine Learning:



**Protocolos para URIs en Azure Machine Learning**

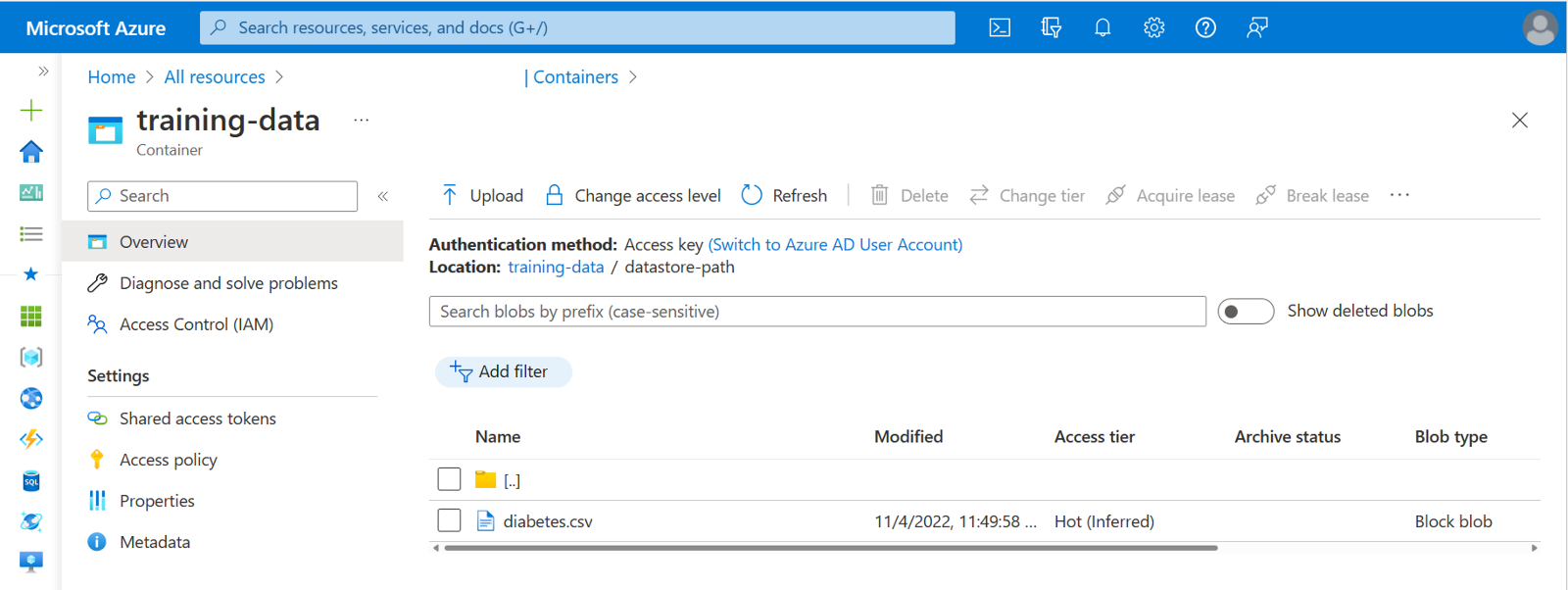
Existen tres protocolos principales para acceder a datos en Azure Machine Learning:

1. **http(s)**: Usado para datos almacenados públicamente o de forma privada en Azure Blob Storage o en una ubicación accesible públicamente mediante http(s).
2. **abfs(s)**: Usado para datos en Azure Data Lake Storage Gen 2.
3. **azureml**: Usado para datos almacenados en un *datastore* dentro de Azure Machine Learning.

**Ejemplo de Uso**

Imagina que has creado un almacenamiento en Azure Blob Storage, con un contenedor llamado training-data. Dentro de este contenedor, hay una carpeta datastore-path que contiene el archivo diabetes.csv.

* Si deseas acceder a estos datos directamente, puedes utilizar la ruta al archivo con el protocolo http(s). Si el contenedor es privado, deberás proporcionar algún tipo de autenticación, como una *firma de acceso compartido* (SAS).
* Al crear un *datastore* en Azure Machine Learning, la información de conexión y autenticación queda almacenada de manera segura en el espacio de trabajo. Luego, puedes usar el protocolo azureml para acceder al archivo dentro de la carpeta sin tener que autenticar cada vez.



**Datastores en Azure Machine Learning**

Un *datastore* es una referencia a una cuenta de almacenamiento existente en Azure, como Azure Blob Storage o Azure Data Lake Storage. Cuando se refiere a datos en un *datastore*, no es necesario autenticarse cada vez, ya que la información de conexión almacenada en el *datastore* se usa automáticamente.

**Mejores Prácticas**

Es recomendable evitar incluir datos sensibles, como información de autenticación, en el código. Siempre que sea posible, usa *datastores* y *activos de datos* en Azure Machine Learning para proteger la información de acceso. Sin embargo, en fases de experimentación en notebooks, es común conectar directamente a la ubicación de almacenamiento para evitar configuraciones adicionales.

**-- Crear un Datastore en Azure Machine Learning**

Los *datastores* en Azure Machine Learning son abstracciones para fuentes de datos en la nube. Encapsulan la información necesaria para conectarse a estas fuentes y almacenan esta información de forma segura, evitando la inclusión de datos de conexión en el código.

**Beneficios de Usar Datastores**

* Proporcionan URIs fáciles de usar para acceder a los datos.
* Facilitan la exploración de datos dentro de Azure Machine Learning.
* Almacenan la información de conexión de forma segura, sin exponer secretos ni claves a los científicos de datos.

**Métodos de Autenticación**

Al crear un *datastore* con una cuenta de almacenamiento existente en Azure, puedes elegir entre dos métodos de autenticación:

1. **Basado en credenciales**: Usa un *service principal*, token de *Shared Access Signature* (SAS) o clave de cuenta para autenticar el acceso.
2. **Basado en identidad**: Usa tu identidad de Microsoft Entra o una identidad administrada (Managed Identity).

Este enfoque asegura una conexión segura y simplifica el acceso a datos en proyectos de machine learning en Azure.

**Crear un Datastore para un Contenedor de Azure Blob Storage**

Cuando deseas crear un *datastore* que se conecte a un contenedor de Azure Blob Storage, puedes utilizar varios métodos de autenticación. Aquí te explico cómo hacerlo usando una **clave de cuenta** mediante el SDK de Python.

**Pasos para Crear un Datastore usando la Clave de Cuenta**

1. **Importa las librerías necesarias**: Asegúrate de tener instaladas e importadas las librerías para Azure Machine Learning.
2. **Define los detalles de conexión**: Incluye el nombre del contenedor, el nombre de la cuenta de almacenamiento y la clave de la cuenta. Esta información es fundamental para establecer la conexión.
3. **Autentica con el SDK**: Utiliza la clave de cuenta para autenticar y conectar el *datastore* al contenedor.

**Ejemplo de Código en Python usando el SDK de Azure Machine Learning**

Texto

Descripción generada automáticamenteAlternatively, you can create a datastore to connect to an Azure Blob Storage container by using a SAS token to authenticate:

Texto

Descripción generada automáticamente

--

**Data Assets en Azure Machine Learning**

Los *data assets* en Azure Machine Learning son referencias a ubicaciones de datos que simplifican el acceso a los mismos durante el entrenamiento de modelos. Estos permiten definir la ubicación, el método de acceso y metadatos relevantes, y ofrecen ventajas como:

* **Compartir y reutilizar datos** con otros miembros del equipo sin necesidad de recordar las ubicaciones exactas de los archivos.
* **Acceso sencillo** a los datos durante el entrenamiento en cualquier tipo de *compute* compatible, sin preocuparse por cadenas de conexión o rutas.
* **Versionado** del metadato del *data asset*, facilitando el seguimiento de cambios en los datos.

**Tipos Principales de Data Assets**

1. **URI file**: Apunta a un archivo específico.
2. **URI folder**: Apunta a una carpeta.
3. **MLTable**: Apunta a una carpeta o archivo, e incluye un esquema para leer los datos en formato tabular.

Estos tipos permiten organizar y acceder a los datos de manera eficiente y segura en Azure Machine Learning.

**Cuándo usar los *Data Assets* en Azure Machine Learning**

Los *data assets* son útiles cuando ejecutas tareas de machine learning como trabajos (jobs) en Azure Machine Learning. En un job, puedes ejecutar un script en Python que toma entradas y genera salidas, y los *data assets* pueden configurarse como entradas o salidas de estos trabajos.

**Tipos de *Data Assets* y cómo crearlos**

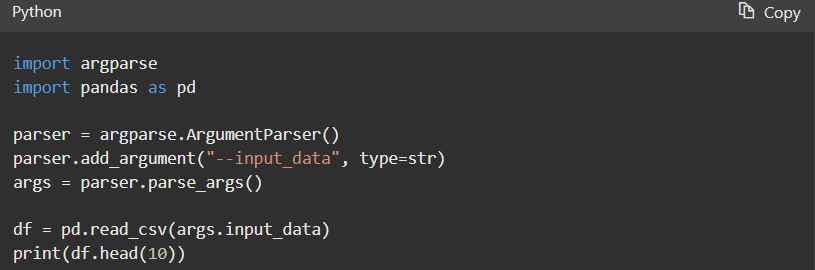
**URI File Data Asset**

Un *URI file data asset* apunta a un archivo específico. Azure Machine Learning solo almacena la ruta al archivo, lo que permite conectar cualquier tipo de archivo. Para utilizarlo, especificas cómo deseas leer los datos según el tipo de archivo.

**Rutas Soportadas para *URI File Data Assets***

* **Local**: ./<path>
* **Azure Blob Storage**: wasbs://<account\_name>.blob.core.windows.net/<container\_name>/<folder>/<file>
* **Azure Data Lake Storage (Gen 2)**: abfss://<file\_system>@<account\_name>.dfs.core.windows.net/<folder>/<file>
* **Datastore**: azureml://datastores/<datastore\_name>/paths/<folder>/<file>

**Nota Importante**: Si el *data asset* apunta a un archivo en tu dispositivo local, una copia se cargará en el *datastore* predeterminado (workspaceblobstore). Así, el archivo estará disponible desde el workspace, incluso si el dispositivo local no está accesible.



**--**

**Crear un URI Folder Data Asset en Azure Machine Learning**

Un *URI folder data asset* apunta a una carpeta específica, similar al *URI file data asset*, y admite las mismas rutas.

**Código para Crear un URI Folder Data Asset**

Usando el SDK de Python, puedes crear un *URI folder data asset* de la siguiente manera:

Texto

Descripción generada automáticamente**Utilizar un URI Folder Data Asset como Entrada en un Job**

Texto

Descripción generada automáticamenteSi usas el *data asset* en un job que apunta a múltiples archivos CSV, puedes leer y concatenar todos los archivos CSV en la carpeta. Por ejemplo:

Si trabajas con otro tipo de archivos, debes ajustar el código para usar el método adecuado según el tipo de datos.

--

**Crear un MLTable Data Asset en Azure Machine Learning**

Un *MLTable data asset* permite trabajar con datos tabulares en Azure Machine Learning, definiendo un esquema específico para leer los datos. Esto es útil para datos con esquemas complejos o cambiantes, ya que la definición del esquema se almacena en el data asset y no es necesario definir cómo leer los datos en cada script.

**Definición del Esquema en un Archivo MLTable**

Para definir el esquema, puedes incluir un archivo MLTable en la misma carpeta que los datos que deseas leer. Un ejemplo de archivo MLTable podría verse así:

Texto

Descripción generada automáticamente

**Código para Crear un MLTable Data Asset**

Texto

Descripción generada automáticamenteCon el SDK de Python, puedes crear un *MLTable data asset* de la siguiente manera:

**Utilizar un MLTable Data Asset en un Job**

Texto

Descripción generada automáticamenteCuando se usa un *MLTable data asset* en un script de Python que ejecutas como un job en Azure Machine Learning, puedes leer los datos con el siguiente código:

Texto

Descripción generada automáticamente