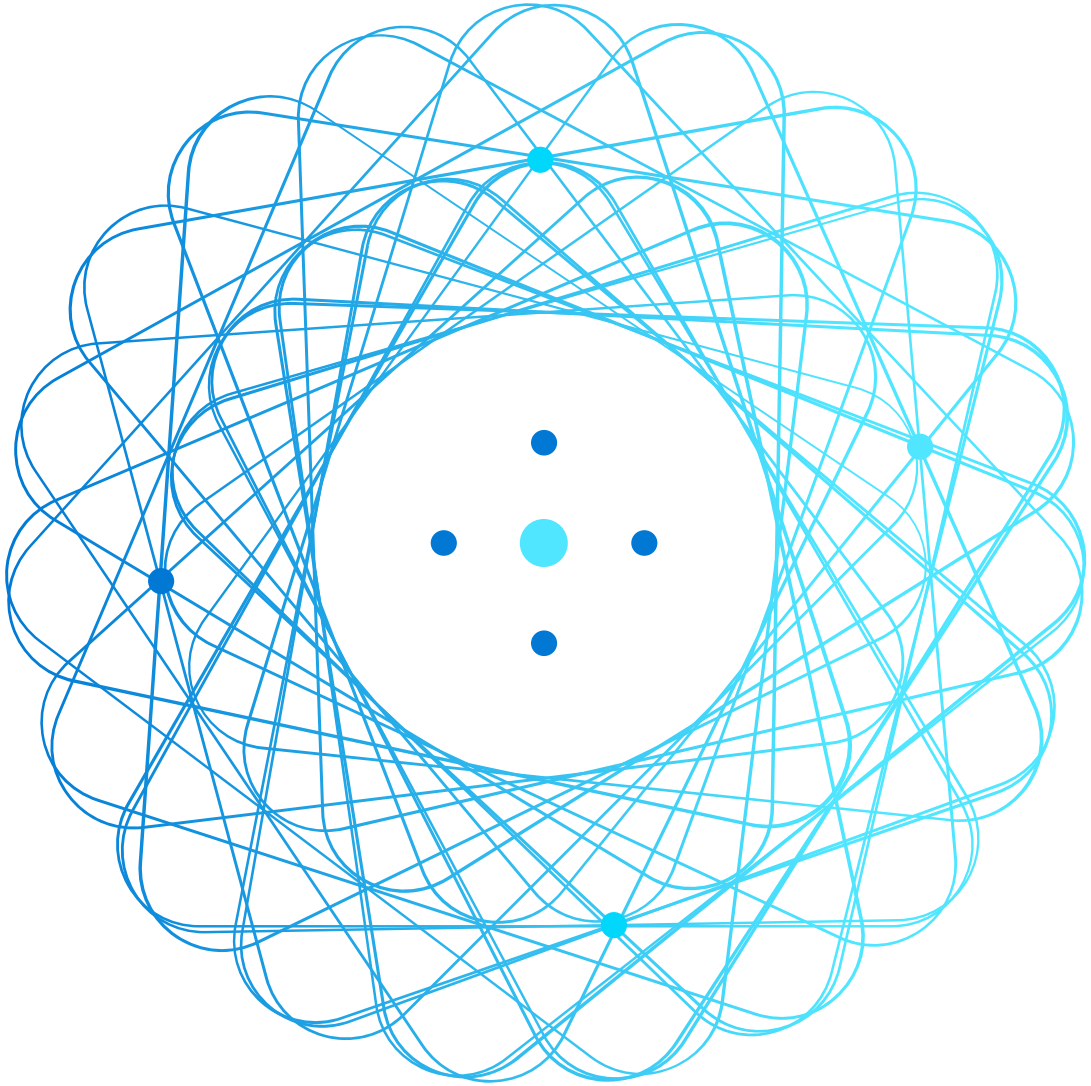
Microsoft Azure

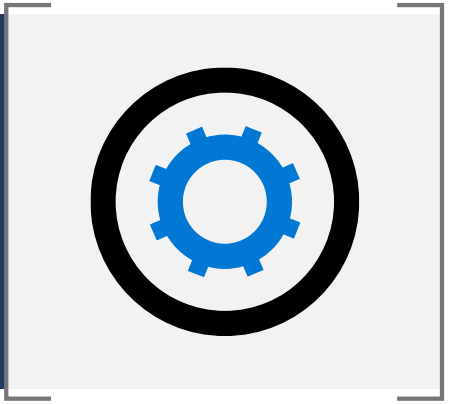
Make data available in Azure Machine Learning



Module Agenda

Make data available in Azure Machine Learning

Make data available in Azure Machine Learning



Understand URIs

A URI references the **location of your data**.

For Azure Machine Learning to connect to your data directly, you need to prefix the URI with the appropriate protocol.

There are three common protocols when working with data in the context of Azure Machine Learning:

- `http(s)` – Use for data stores publicly or privately in an Azure Blob Storage or publicly available http(s) location.
- `abfs(s)` – Use for data stores in an Azure Data Lake Storage Gen 2.
- `azureml` – Use for data stored in an Azure Machine Learning datastore.

Create a datastore(1)

You have the choice between two different authentication methods when creating a datastore with an existing storage account on Azure:

- **Credential-based**: Use a service principal, shared access signature (SAS) token or account key to authenticate access to your storage account.
- **Identity-based**: Use your Azure Active Directory identity or managed identity.

Azure Machine Learning supports the creation of datastores for multiple kinds of Azure data source, including:

- Azure Blob Storage
- Azure File Share
- Azure Data Lake (Gen 1)
- Azure Data Lake (Gen 2)

Create a datastore(2)

- Datastores are attached to workspaces and are used to store **connection information** to storage services
- You can create a datastore through:
  - The graphical user interface (studio)
  - The Azure command-line interface (CLI)
  - The Python software development kit (SDK)

Understand data assets

The benefits of using data assets are:

- You can share and reuse data with other members of the team such that they don't need to remember file locations.
- You can seamlessly access data during model training (on any supported compute type) without worrying about connection strings or data paths.
- You can version the metadata of the data asset.

Three main types of data assets you can use:

- **URI file**: Points to a specific file.
- **URI folder**: Points to a folder.
- **MLTable**: Points to a folder or file, and includes a schema to read as tabular data.

Create a URI file data asset

To create a URI file data asset, you can use the following code:

```
from azure.ai.ml.entities import Data
from azure.ai.ml.constants import AssetTypes
my_path = '<supported-path>'
my_data = Data(
    path=my_path,
    type=AssetTypes.URI_FILE,
    description="<description>",
    name="<name>",
    version="<version>"
)
ml_client.data.create_or_update(my_data)
```

Create a URI folder data asset

To create a URI folder data asset with the Python SDK, you can use the following code:

```
from azure.ai.ml.entities import Data
from azure.ai.ml.constants import AssetTypes
my_path = '<supported-path>'
my_data = Data(
    path=my_path,
    type=AssetTypes.URI_FOLDER,
    description="<description>",
    name="<name>",
    version='<version>'
)
ml_client.data.create_or_update(my_data)
```

Create a MLTable data asset

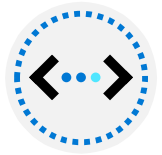
To create a MLTable data asset with the Python SDK, you can use the following code:

```
from azure.ai.ml.entities import Data
from azure.ai.ml.constants import AssetTypes
my_path = '<path-including-mltable-file>'
my_data = Data(
    path=my_path,
    type=AssetTypes.MLTABLE,
    description="<description>",
    name="<name>",
    version='<version>'
)
ml_client.data.create_or_update(my_data)
```

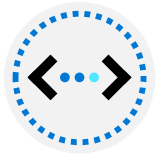
Exercise - Make data available

In this exercise, you will:

Task 1: Explore the default datastores



Task 2: Create a datastore



Task 3: Create data assets

Instructions

Follow these instructions to complete the exercise:

1. View the exercise repo at https://microsoftlearning.github.io/mslearn-azure-ml/.
2. Complete the **Make data available in Azure Machine Learning** exercise.

Knowledge check

A data scientist wants to read data stored in a publicly available GitHub repository. The data will be read in a Jupyter notebook in the Azure Machine Learning workspace for some quick experimentation. Which protocol should be used to read the data in the notebook?

- ❑ azureml
- ☑ http(s)
- ❑ abfs(s)

What type of data asset should someone create when the schema changes frequently and the data is used in many different jobs?

- ❑ URI file
- ❑ URI folder
- ☑ MLTable

Microsoft Azure