Highlight Note

Creating and Registering Datasets

You can create a dataset and work with it immediately, and you can then *register* the dataset in the workspace to make it available for use in experiments and data processing pipelines later.

You can create datasets by using the visual interface in Azure Machine Learning studio, or you can use the Azure Machine Learning SDK.

Creating and Registering Tabular Datasets

To create a tabular dataset using the SDK, use the **from_delimited_files** method of the **Dataset.Tabular** class, like this:

```
from azureml.core import Dataset

blob_ds = ws.get_default_datastore()
csv_paths = [(blob_ds, 'data/files/current_data.csv'),
             (blob_ds, 'data/files/archive/*.csv')]
tab_ds = Dataset.Tabular.from_delimited_files(path=csv_paths)
tab_ds = tab_ds.register(workspace=ws, name='csv_table')
```

The dataset in this example includes data from two file paths within the default datastore:

- The **current_data.csv** file in the **data/files** folder.
- All .csv files in the **data/files/archive/** folder.

After creating the dataset, the code registers it in the workspace with the name **csv_table**.

Creating and Registering File Datasets

To create a file dataset using the SDK, use the **from_files** method of the **Dataset.File** class, like this:

```
from azureml.core import Dataset

blob_ds = ws.get_default_datastore()
file_ds = Dataset.File.from_files(path=(blob_ds, 'data/files/images/*.jpg'))
file_ds = file_ds.register(workspace=ws, name='img_files')
```

The dataset in this example includes all .jpg files in the **data/files/images** path within the default datastore:

After creating the dataset, the code registers it in the workspace with the name **img_files**.

Retrieving a Registered Dataset

After registering a dataset, you can retrieve it by using any of the following techniques:

- The **datasets** dictionary attribute of a **Workspace** object.
- The **get_by_name** or **get_by_id** method of the **Dataset** class.

Both of these techniques are shown in the following code:

```
import azureml.core
from azureml.core import Workspace, Dataset

# Load the workspace from the saved config file
ws = Workspace.from_config()

# Get a dataset from the workspace datasets collection
ds1 = ws.datasets['csv_table']

# Get a dataset by name from the datasets class
ds2 = Datasets.get_by_name(ws, 'img_files')
```