Highlight Note

What are Datasets?

Datasets are versioned packaged data objects that can be easily consumed in experiments and pipelines. Datasets are the recommended way to work with data, and are the primary mechanism for advanced Azure Machine Learning capabilities like data labeling and data drift monitoring.

Types of Dataset

Datasets are typically based on files in a datastore, though they can also be based on URLs and other sources. You can create the following types of dataset:

- **Tabular**: The data is read from the dataset as a table. You should use this type of dataset when your data is consistently structured and you want to work with it in common tabular data structures, such as Pandas dataframes.
- **File**: The dataset presents a list of file paths that can be read as though from the file system. Use this type of dataset when your data is unstructured, or when you need to process the data at the file level (for example, to train a convolutional neural network from a set of image files).

You can create datasets from individual files or multiple file paths. The paths can include wildcards (for example, */files/*.csv*) making it possible to encapsulate data from a large number of files in a single dataset.