

Create and version a dataset

Lab Overview

To access your data in your storage account, Azure Machine Learning offers datastores and datasets. Create an Azure Machine Learning dataset to interact with data in your datastores and package your data into a consumable object for machine learning tasks. Register the dataset to your workspace to share and reuse it across different experiments without data ingestion complexities.

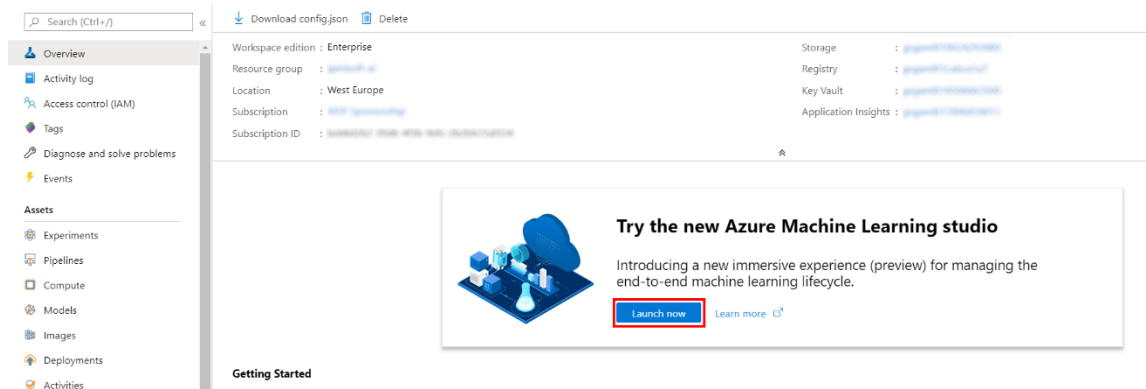
Datasets can be created from local files, public urls, Azure Open Datasets, or specific file(s) in your datastores. To create a dataset from an in memory pandas dataframe, write the data to a local file, like a csv, and create your dataset from that file. Datasets aren't copies of your data, but are references that point to the data in your storage service, so no extra storage cost is incurred.

In this lab, we are using a subset of NYC Taxi & Limousine Commission - green taxi trip records available from [Azure Open Datasets](#) to show how you can register and version a Dataset using the AML designer interface. In the first exercises we use a modified version of the original CSV file, which includes collected records for five months (January till May). The second exercise demonstrates how we can create a new version of the initial dataset when new data is collected (in this case, we included records collected in June in the CSV file).

Exercise 1: Register Dataset with Azure Machine Learning studio

Task 1: Upload Dataset from web file

1. In [Azure portal](#), open the available machine learning workspace.
2. Select **Launch now** under the **Try the new Azure Machine Learning studio** message.



3. When you first launch the studio, you may need to set the directory and subscription. If so, you will see this screen:

Welcome to the studio!

Select a subscription and a workspace to get started or go to the [Azure Portal](#) to create your subscription and workspace. You can switch subscriptions and workspaces at any time. [Learn more.](#)

Switch directory

Subscription

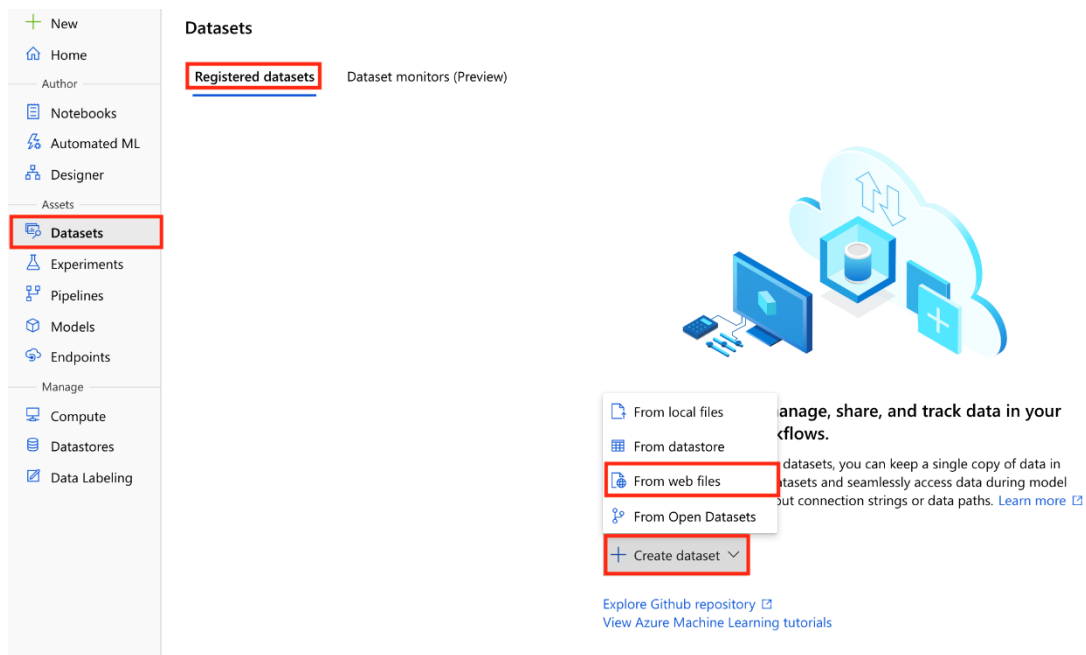
Machine learning workspace

quick-starts-ws-190124	▼
quick-starts-ws-190124	southcentralus
aml-quickstarts-190124	

Get started

For the directory, select **Udacity** and for the subscription, select **Azure Sponsorship**. For the machine learning workspace, you may see multiple options listed. **Select any of these** (it doesn't matter which) and then click **Get started**.

4. From the studio, select **Datasets**, + **Create dataset**, **From web files**. This will open the **Create dataset from web files** dialog on the right.



5. Provide the following information and then select **Next**:

1. Web URL: `https://introtomlsampledatablob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-5months.csv`
2. Name: `nyc-taxi-sample-dataset`

Create dataset from web files ×

Basic info

Settings and preview

Schema

Confirm details

Basic info

Web URL *

https://introtomlsampledatablob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-5months.csv

Name * 👁 Dataset version

nyc-taxi-sample-dataset 1

Dataset type * ⓘ

Tabular

Description

Dataset description

☐ Skip data validation ⓘ

Back

Next

Cancel

Task 2: Preview Dataset

1. On the Settings and preview panel, set the **Column headers** drop down to **All files have same headers**.
2. Scroll the data preview to right to observe the target column: **totalAmount**. After you are done reviewing the data, select **Next**

Create dataset from web files ×

☒ Basic info

☒ Settings and preview

☐ Schema

☐ Confirm details

Settings and preview

These settings were automatically detected. Please verify that the selections were made correctly or update

File format

Delimited ▼

Delimiter Comma ▼ **Example** Field1,Field2,Field3

Encoding UTF-8 ▼

Column headers All files have same headers ▼

Skip rows None ▼

0.0 snowDepth	0.0 precipTime	0.0 precipDepth	0.0 temperature	0.0 totalAmount
29.05882353	24	3	6.185714286	44.3
0	6	0	4.571929825	44.8
0	1	0	4.384090909	18.96
29.05882353	24	3	6.185714286	16.3
0	1	0	3.846428571	5.3
0	6	0	0.159459459	16.3

Back Next Cancel

Task 3: Select Columns

1. Select columns from the dataset to include as part of your training data. Leave the default selections and select **Next**

Create dataset from web files ×

☒ Basic info

☒ Settings and preview

☒ **Schema**

☐ Confirm details

Include	Column name	Properties	Type
<input type="checkbox"/>	Path	Not applicable to select...	String
<input checked="" type="checkbox"/>	vendorID	Not applicable to select...	Integer
<input checked="" type="checkbox"/>	passengerCount	Not applicable to select...	Integer
<input checked="" type="checkbox"/>	tripDistance	Not applicable to select...	Decimal
<input checked="" type="checkbox"/>	hour_of_day	Not applicable to select...	Integer
<input checked="" type="checkbox"/>	day_of_week	Not applicable to select...	Integer
<input checked="" type="checkbox"/>	day_of_month	Not applicable to select...	Integer
<input checked="" type="checkbox"/>	month_num	Not applicable to select...	Integer
<input checked="" type="checkbox"/>	normalizeHolidayName	Not applicable to select...	String
<input checked="" type="checkbox"/>	isPaidTimeOff	Not applicable to select...	Boolean

Back

Next

Cancel

Task 4: Create Dataset

1. Confirm the dataset details and select **Create**

Create dataset from web files ×

- ✓ Basic info
- ✓ Settings and preview
- ✓ Schema
- **Confirm details**

Confirm details

Basic info

Name
nyc-taxi-sample-dataset

Dataset version
1

Dataset type
Tabular

Web URL
https://introtojsonsampledata.blob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-5months.csv

File settings

File format
Delimited

Delimiter
Comma

Encoding
UTF-8

Column headers
All files have same headers

Skip rows
None

☐ Profile this dataset after creation

Back Create Cancel

Exercise 2: Create a version of the existing Dataset

Task 1: Register new dataset version

1. From the [Azure Machine Learning studio](#), select **Datasets** and select the `nyc-taxi-sample-dataset` dataset created in the first exercise. This will open the `Dataset details` page.
2. Select **New version, From web files** to open the same `Create dataset from web files` dialog you already entered in the first exercise.

New

Home

Author

Notebooks

Automated ML

Designer

Assets

Datasets

Experiments

Pipelines

Models

Endpoints

Manage

Compute

Datastores

Data Labeling

nyc-taxi-sample-dataset

Version 1 (latest) ▾

Details

Consume

Explore

Models

Refresh

Generate profile

Unregister

New version ▾

From local files

From datastore

From web files

From Open Datasets

Attributes

Properties

Tabular

Description

--

Created by

Web Url

https://introtomlsampledatablob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-5months.csv

Profile

No profile generated

Current version

1

Latest version

1

Created time

Jun 16, 2020 3:57 PM

Modified time

Jun 16, 2020 3:57 PM

3. This time, the **Name** and **Dataset version** fields are already filled in for you. Provide the following information and select **Next** to move on to the next step:

1. Web URL: `https://introtomlsampledatablob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-6months.csv`

Create dataset from web files ×

● Basic info

○ Settings and preview

○ Schema

○ Confirm details

Basic info

Web URL *

https://introtomlsampledatablob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-6months.csv

Name *

nyc-taxi-sample-dataset

Dataset version

2

Dataset type *

Tabular

Description

Dataset description

☐ Skip data validation

Back

Next

Cancel

4. Select **All files have the same headers** in the **Column headers** drop-down and move on to the schema selection step.
5. On the **Schema** page, let's suppose you decided to exclude some columns from your dataset. Exclude columns: **snowDepth**, **precipTime**, **precipDepth**. Select **Next** to move on to the final step.

Create dataset from web files ×

Basic info

Settings and preview

Schema

Confirm details

Schema

Include	Column name	Properties	Type	Format settings a
<input checked="" type="checkbox"/>	hour_of_day	Not applicable to selecte...	Integer	15, 13, 23
<input checked="" type="checkbox"/>	day_of_week	Not applicable to selecte...	Integer	2, 4, 4
<input checked="" type="checkbox"/>	day_of_month	Not applicable to selecte...	Integer	27, 15, 8
<input checked="" type="checkbox"/>	month_num	Not applicable to selecte...	Integer	1, 1, 1
<input checked="" type="checkbox"/>	normalizeHolidayName	Not applicable to selecte...	String	None, None, Non
<input checked="" type="checkbox"/>	isPaidTimeOff	Not applicable to selecte...	Boolean	false, false, false
<input type="checkbox"/>	snowDepth	Not applicable to selecte...	Decimal	29.058823529411
<input type="checkbox"/>	precipTime	Not applicable to selecte...	Decimal	24, 6, 1
<input type="checkbox"/>	precipDepth	Not applicable to selecte...	Decimal	3, 0, 0
<input checked="" type="checkbox"/>	temperature	Not applicable to selecte...	Decimal	6.1857142857142
<input checked="" type="checkbox"/>	totalAmount	Not applicable to selecte...	Decimal	44.3, 44.8, 18.96

Back

Next

Cancel

6. Notice the **Dataset version** value in the basic info section. Select **Create** to close the new version confirmation page.

Create dataset from web files ×

- Basic info
- Settings and preview
- Schema
- Confirm details**

Confirm details

Basic info

Name
nyc-taxi-sample-dataset

Dataset version
2

Dataset type
Tabular

Web URL
https://introtoisampledatablob.core.windows.net/data/nyc-taxi/nyc-taxi-sample-data-6months.csv

File settings

File format
Delimited

Delimiter
Comma

Encoding
UTF-8

Column headers
All files have same headers

Skip rows
None

☐ Profile this dataset after creation

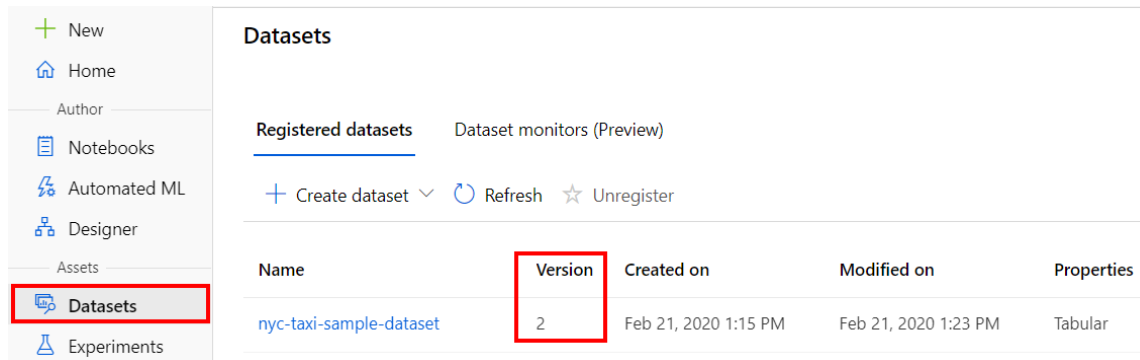
Back

Create

Cancel

Task 2: Review both versions of the dataset

1. Back to the **Datasets** page, in the **Registered datasets** list, notice the version value for the `nyc-taxi-sample-dataset` dataset.



Datasets				
Registered datasets Dataset monitors (Preview)				
+ Create dataset Refresh Unregister				
Name	Version	Created on	Modified on	Properties
nyc-taxi-sample-dataset	2	Feb 21, 2020 1:15 PM	Feb 21, 2020 1:23 PM	Tabular

2. Select the `nyc-taxi-sample-dataset` dataset link to open the dataset details page, where **Version 2(latest)** is automatically selected. Go to the **Explore** section to observe the structure and content of the new version. Notice the columns and rows structure in the dataset preview pane:
 - **Number of columns:** 11
 - **Number of rows:** 10000
 - Scroll right to check that the three excluded columns are missing (**snowDepth**, **prcipTime**, **precipDepth**)

nyc-taxi-sample-dataset

Version 2 (latest)

Details Consume **Explore** Models

Refresh Generate profile Unregister New version

Profile: This is the quick profile generated by sampled data. Please generate a profile from the action bar to view the full profile.

Preview Profile

Number of columns: 11

Number of rows: 50 (of 10000)

passengerCount	tripDistance	hour_of_day	day_of_week	day_of_month	month_num	normalizeHoli...	isPaidTimeOff	temperature	totalAmount
1	9.4	15	2	27	1	None	false	6.18571428571429	44.3
5	14.75	13	4	15	1	None	false	4.571929824561403	44.8
1	3.35	23	4	8	1	None	false	4.384090909090913	18.96
1	3.33	18	2	27	1	None	false	6.18571428571429	16.3
1	0.47	17	6	3	1	None	false	3.846428571428569	5.3
1	3.07	9	1	12	1	None	false	0.1594594594594597	16.3
1	0.92	23	4	22	1	None	false	-2.999107142857142	8.97
1	1.9	12	4	8	1	None	false	4.384090909090913	11.8
1	0.77	0	1	19	1	None	false	-5.393749999999998	7.3
1	2.35	2	6	10	1	None	false	10.943654822335034	14.16
1	8.3	18	3	21	1	None	false	-0.040000000000000...	34.3
2	4.28	18	0	18	1	Martin Luther King, J...	true	-2.3351145038167944	18.96
1	10.77	2	2	27	1	None	false	6.18571428571429	31.3
1	1.75	17	3	14	1	None	false	-1.9500000000000008	14.3
1	3.75	2	4	1	1	New Year's Day	true	5.197345132743359	19.3
1	5.79	14	6	3	1	None	false	3.846428571428569	33.55
5	1.06	19	4	29	1	None	false	3.3651785714285696	8.3
1	5.7	11	2	13	1	None	false	-2.06875	29.1
1	3.26	10	3	14	1	None	false	-1.9500000000000008	23.34

3. Select **Version 1** from the drop-down near the dataset name title and notice the changing values for:

- **Number of columns:** 14 (since the previous version still contains the three excluded columns)
- **Number of rows:** 9776 (since the previous version contains only data for 5 months)

nyc-taxi-sample-dataset

Version 1

Details

Consume

Explore

Models

Refresh

Generate profile

Unregister

New version

Profile: This is the quick profile generated by sampled data. Please generate a profile from the action bar to view the full profile.

Preview

Profile

Number of columns: 14

Number of rows: 50 (of 9776)

our_of_day	day_of_week	day_of_month	month_num	normalizeHoli...	isPaidTimeOff	snowDepth	precipTime	precipDepth	temperature	tr
2	27	1	None	false	29.05882353	24	3	6.185714286	44.3	-
4	15	1	None	false	0	6	0	4.571929825	44.8	-
4	8	1	None	false	0	1	0	4.384090909	18.96	-
2	27	1	None	false	29.05882353	24	3	6.185714286	16.3	-
6	3	1	None	false	0	1	0	3.846428571	5.3	-
1	12	1	None	false	0	6	0	0.159459459	16.3	-
4	22	1	None	false	0	1	0	-2.999107143	8.97	-
4	8	1	None	false	0	1	0	4.384090909	11.8	-
1	19	1	None	false	0	1	0	-5.39375	7.3	-
6	10	1	None	false	0	24	254	10.94365482	14.16	-
3	21	1	None	false	0	1	0	-0.04	34.3	-
0	18	1	Martin Luther King, J...	true	3	24	13	-2.335114504	18.96	-
2	27	1	None	false	29.05882353	24	3	6.185714286	31.3	-
3	14	1	None	false	0	6	0	-1.95	14.3	-
4	1	1	New Year's Day	true	0	1	0	5.197345133	19.3	-
6	3	1	None	false	0	1	0	3.846428571	33.55	-
4	29	1	None	false	15.64705882	6	0	3.365178571	8.3	-
2	13	1	None	false	0	6	0	-2.06875	29.1	-
3	14	1	None	false	0	6	0	-1.95	23.34	-

Next Steps

Congratulations! You have now explored a first simple scenario for dataset versioning using the Azure Machine Learning studio. You found out how you can create and version a simple dataset when new training data is available. You can continue to experiment in the environment but are free to close the lab environment tab and return to the Udacity portal to continue with the lesson.