

## Formal Documentation of Azure Data Factory Pipeline – Training.

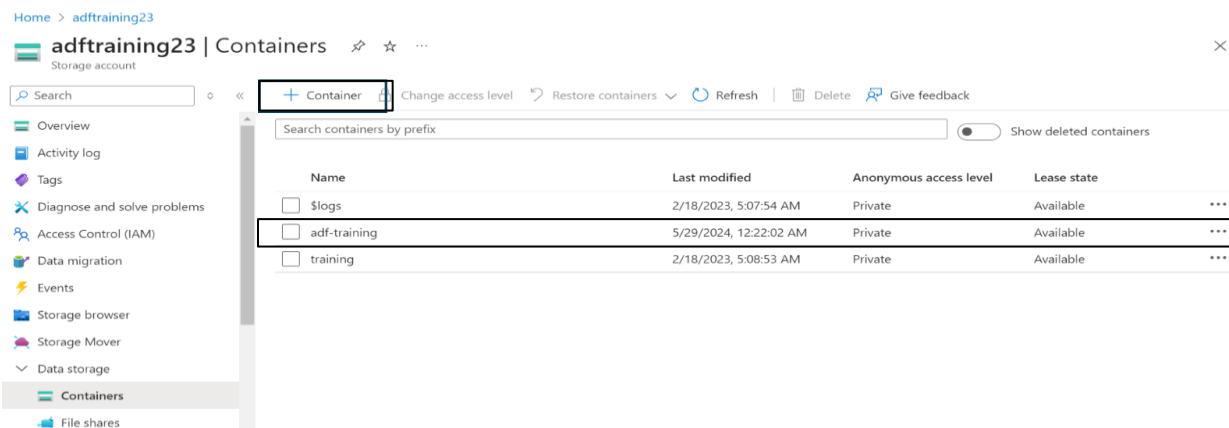
### Usecase:

This document outlines the steps taken to create and manage a data pipeline in Azure Data Factory (ADF). The pipeline, named “bt\_pl\_ff\_product\_data\_ld”, was designed to load data from a CSV file into a target dataset, add audit fields dynamically, and delete the source file to prevent duplication.

### Steps:

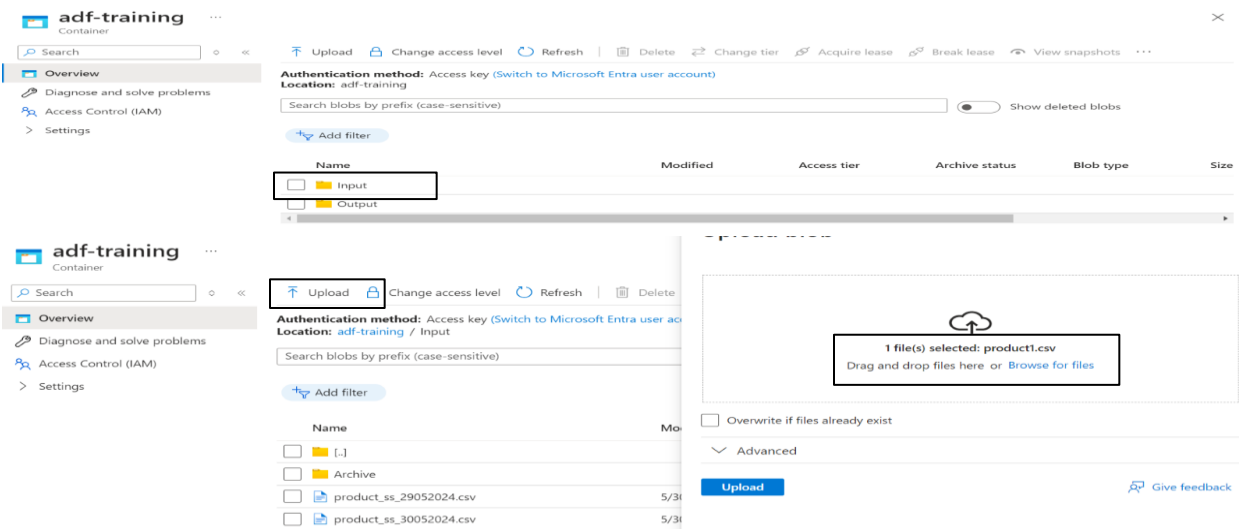
#### 1. Creating a Container in the storage account.

- Through the storage account, under data storage option, I opened the container tab.
- Then, a container named adf-training was created using the +Container option.
- It was created within the Azure Storage account to serve as the storage location for the data files.



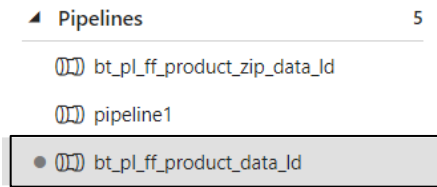
#### 2. Uploading the input file in the container.

- Within our container, I opened the input folder (refer image 1)
- Selecting the upload tab, I uploaded the product1.csv file in the input folder (refer image 2)



#### 3. Pipeline Creation

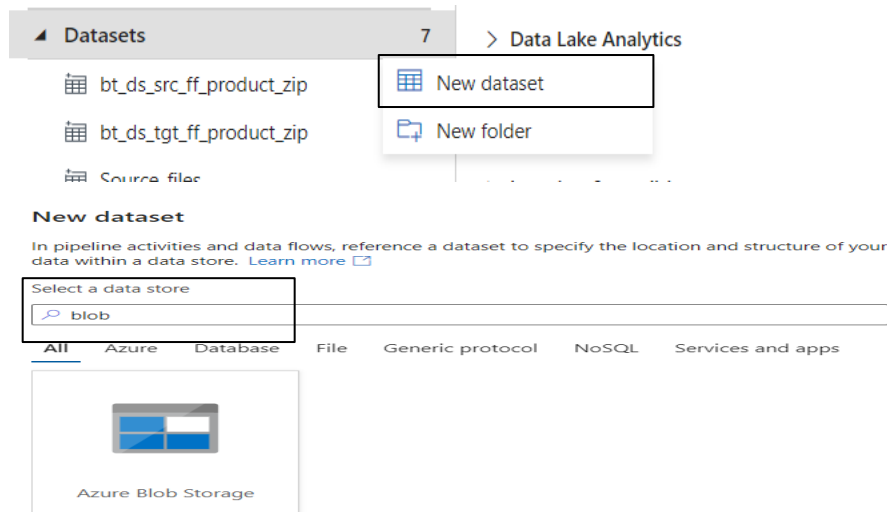
- A new pipeline named “bt\_pl\_ff\_product\_data\_ld” was created in ADF to orchestrate the data loading process from source to target.



## 4. Dataset Creation









### Source Dataset

- A new dataset was created for the source data file.
- Source Dataset Name: bt\_ds\_src\_ff\_product1
- Inside the Azure Data Factory, In the Author tab, I selected the Dataset option and clicked on "New Dataset." ( refer image 1)
- Then, I chose the Azure Blob Storage option. (refer image 2)
- Next, I selected the Delimited Text file format, which brought me to the properties page where I needed to define the dataset name and path. (refer image 3)
- I specified the dataset name, selected the linked service, and provided the path of my input file: adf-training/Input/product1.csv. (refer image 4)
- These steps created my source dataset.( refer image 5)
- I opened my source file and updated connection, updated the column delimiter option to pipe (|) because my CSV file is pipe-delimited, all other options remained unchanged. (refer image 6)
- Finally, I selected the "Preview Data" option and verified my input file, which was loaded successfully in proper format. (refer image 7)



## Select format

Choose the format type of your data

|  |   |  |
|--|---|--|
| <br>Avro    | <br>Binary | <br>DelimitedText |
| <br>Excel   | <br>JSON   | <br>ORC           |
| <br>Parquet | <br>XML    |  |

Continue

Back

Cancel

## Set properties

Name

bt\_ds\_src\_ff\_product1

Linked service \*

ADF\_Training

File path

adf-training

/ Input

/ product1.csv

First row as header



Import schema

☒ From connection/store ☐ From sample file ☐ None

▲ Datasets

9

bt\_ds\_src\_ff\_product1

bt\_ds\_src\_ff\_product\_zip

Connection Schema Parameters

Linked service \* ADF\_Training Test connection Edit + New Learn more

File path \* adf-training / Input / product1.csv

Compression type Select...

Column delimiter ⓘ Pipe (|)

Row delimiter ⓘ Default (\r\n, or \n)

Encoding ⓘ Default(UTF-8)

**Preview data**

Linked service: ADF\_Training

Object: product1.csv

|   | product_id | product_name | product_type |
|---|------------|--------------|--------------|
| 1 | 1          | Mobile       | Electronics  |
| 2 | 2          | Laptop       | Electronics  |
| 3 | 3          | Books        | Stationary   |
| 4 | 4          | Laptop       | Electronics  |
| 5 | 5          | Pens         | Stationary   |

## Target Dataset

- I followed the few similar steps for target dataset too, In the author tab, selected the datasets, clicked on new dataset, selected Azure Blob Storage, selected the Delimited text format which brought me to the properties page where I needed to define the dataset name and path.
- Assigned the path as adf-training/ output folder.
- Target Dataset name: bt\_ds\_tgt\_ff\_product1 (Refer image 1)
- I opened my source file and updated connection, updated the column delimiter option to pipe (|) because my CSV file is pipe-delimited, all other options remained unchanged. (refer image 2)

▲ Datasets 9

- bt\_ds\_src\_ff\_product1
- bt\_ds\_src\_ff\_product\_zip
- bt\_ds\_tgt\_ff\_product1**

Connection Schema Parameters

Linked service \* ADF\_Training Test connection Edit + New Learn more

File path \* adf-training / Output / File name Browse Preview data

Compression type Select...

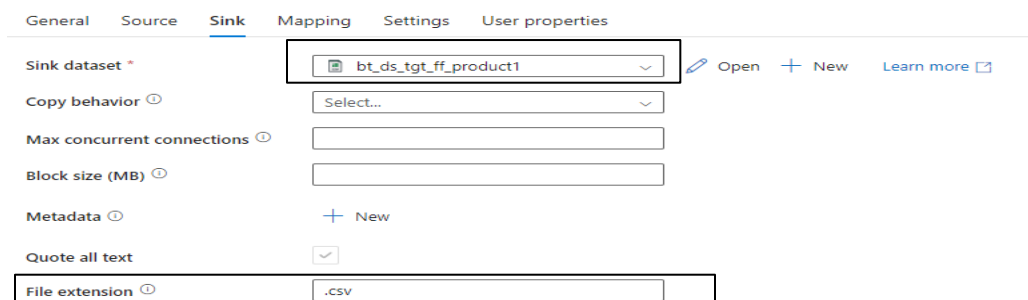
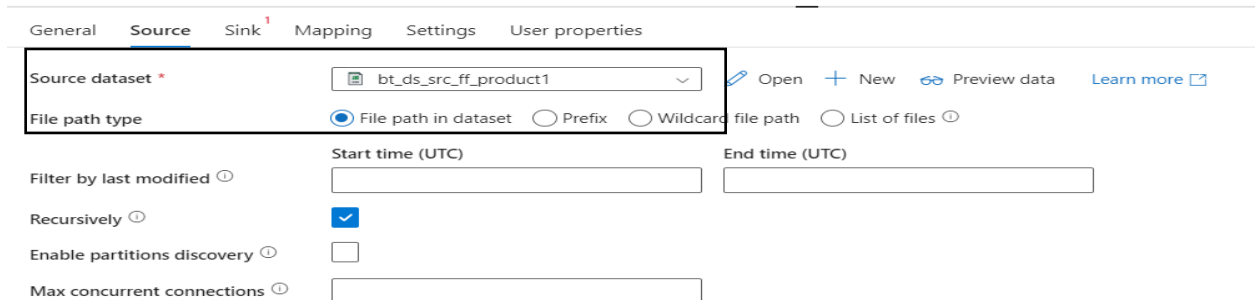
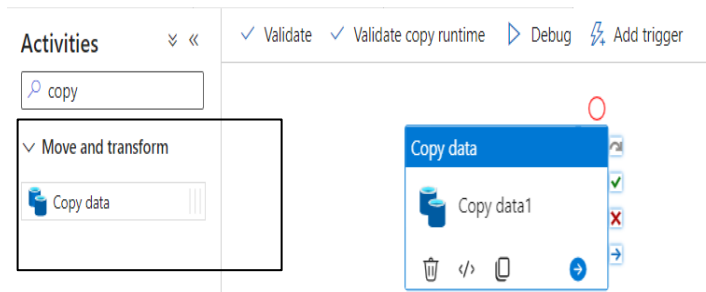
Column delimiter ⓘ Pipe (|)

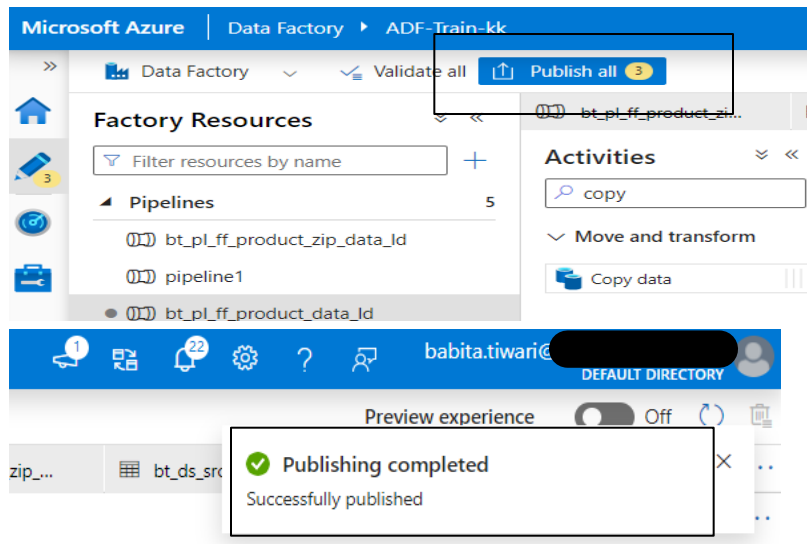
Row delimiter ⓘ Default (\r\n, or \n)

## 5. Copy Activity

- The Copy Data activity was used to transfer data from the source to the target dataset.
- Source: bt\_ds\_src\_ff\_product1
- Sink: bt\_ds\_tgt\_ff\_product1
- Selected the copy activity from the activities tab (refer image 1).
- In source option of copy activity, added the source dataset, and selected file path in dataset. I opted file path in dataset as it specifies a single file to copy from the source location. (refer image 2)

- There are other options too for the path and here are the importance of those:
  - Prefix: Specifies a prefix for the files to copy from the source location. For example, folder/subfolder/\* would copy all files in the subfolder directory.
  - Wildcard file path: Specifies a wildcard pattern to match files to copy from the source location. For example, \*.csv would copy all CSV files.
  - List of files: Specifies a list of files to copy from the source location.
- In the sink option, added the target dataset and changed the file extension from .txt to .csv(refer image 3)
- Then to save the activity, I clicked on publish all and the activity was successfully published (refer image 4 & 5).





## 6. Adding Audit Fields

- The data initially contained three columns. Additional audit fields were added using the Additional Columns feature in the Copy Data activity.
- Inserted additional columns into the existing product1 table adding details in the source tab (refer image 1)
- I specified the name and value as file name – filepath and date\_time-utcnow. Utcnow was selected by clicking on the add dynamic content.
- I have also added the images of how did I achieve utcnow. (refer image 2)
- I then moved to mapping tab and selected import schema, which successfully added the additional columns in the existing table. (image 3)
- I then published and saved my data as shown in the earlier images.
- I then debugged the copy activity and it was successfully executed.

GeneralSourceSinkMappingSettingsUser properties

Skip line count

Additional columns ⓘ

+ NewDelete

| <input type="checkbox"/> | Name      | Value        |
|--------------------------|-----------|--------------|
| <input type="checkbox"/> | file_name | \$\$FILEPATH |
| <input type="checkbox"/> | date_time | @utcnow()    |

Add dynamic content

Add dynamic content below using any combination of [expressions](#), [functions](#) and [system variable](#):

utcnow

Clear contents

ParametersSystem variablesFunctionsVariables

Search

+

GeneralSourceSinkMappingSettingsUser properties

> Type conversion settings

↶ Import schemas

↷ Preview source

+ New mapping

↺ Clear ⓘ

↺ Reset ⓘ

🗑 Delete

| <input type="checkbox"/> | Source                  | Type       | Destination    | Type       |   |   |
|--------------------------|-------------------------|------------|----------------|------------|---|---|
| <input type="checkbox"/> | product_id              | abc String | → product_id   | abc String | + | 🗑 |
| <input type="checkbox"/> | product_name            | abc String | → product_name | abc String | + | 🗑 |
| <input type="checkbox"/> | product_type            | abc String | → product_type | abc String | + | 🗑 |
| <input type="checkbox"/> | file_name<br>Additional | abc String | → file_name    | abc String | + | 🗑 |
| <input type="checkbox"/> | date_time<br>Additional | abc String | → date_time    | abc String | + | 🗑 |

ParametersVariablesSettingsOutput

Pipeline run ID: 1df74c9b-90ac-4f2d-b3b1-d4262c94143e ⓘ ⓘ ⓘ

Pipeline status ✔ Succeeded

[View debug run consumption](#)

All status ▾

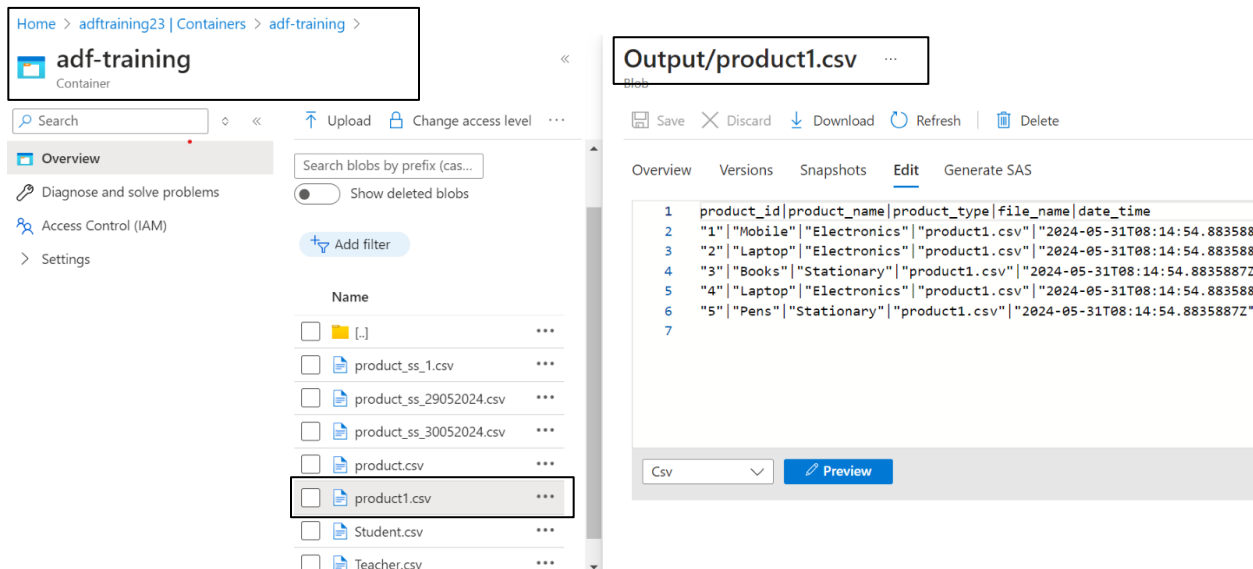
[Monitor in Azure Metrics](#) ⓘ [Export to CSV](#) | ▾

Showing 1 - 1 of 1 items

| Activity name | Activity status          | Activity type | Run start             | Duration | Integration runtime    | User properties | Activity run ID                      |
|---------------|--------------------------|---------------|-----------------------|----------|------------------------|-----------------|--------------------------------------|
| Copy data1    | <span>✔</span> Succeeded | Copy data     | 5/31/2024, 1:39:05 PM | 12s      | AutoResolveIntegration |                 | b6f72d74-ff89-45bb-a125-34b613ddb79a |

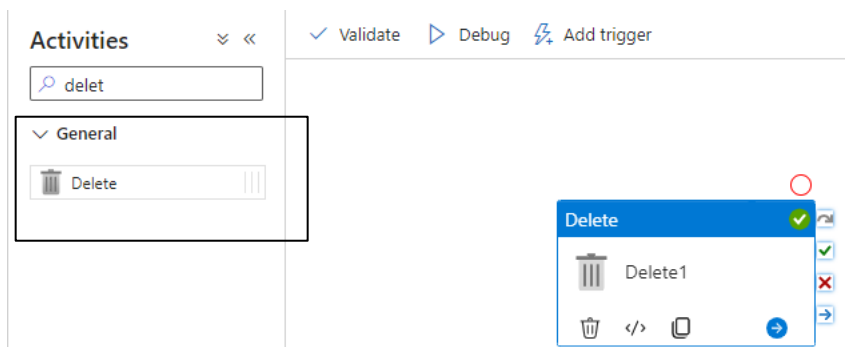
## 7. Copy data output.

- Since the copy data was executed successfully, i went to the adf-training container and checked the output folder which we assigned as an output location for our target file.
- The file was created in the target location and also additional columns were added successfully.



## 8. Deleting the Source File

- To avoid the creation of duplicate files, a Delete activity was performed to remove the source file after the data load.
- I selected the delete activity from the activities tab (refer image 1)
- Source File Location: as bt\_ds\_src\_ff\_product1.
- Selected the delete activity and added the source file in the source tab, rest other options remain unchanged (refer image 2)
- I then updated the logging settings, there is enable logging option which we need to disable. Enable logging setting captures the deleted file logs in the linked services (refer image 3)
- I disabled the loggings (refer image 4)
- Then I published and debugged the delete activity and it was executed successfully (refer image 5).
- I then visited the container adf-training again, checked the input folder and my source file was deleted from there successfully (refer image 6)





General

Source

Logging settings<sup>1</sup>

User properties

Dataset \* ⓘ

bt\_ds\_src\_ff\_product1

Open

New

Preview data

Learn more

File path type

☒ File path in dataset

☐ Wildcard file path

☐ Prefix

☐ List of files ⓘ

Filter by last modified ⓘ

Start time (UTC)

End time (UTC)

Recursively ⓘ

☒

Max concurrent connections ⓘ

General

Source

Logging settings<sup>1</sup>

User properties

Enable logging ⓘ

☒

Logging settings

Logging account linked service \* ⓘ

Select...

New

General

Source

Logging settings

User properties

Enable logging ⓘ

☐

All status ▾

Monitor in Azure Metrics

Export to CSV

Showing 1 - 2 of 2 items

| Activity name | Activity status | Activity type | Run start             | Duration | Integration runtime    | User properties | Activity run ID                      |
|---------------|-----------------|---------------|-----------------------|----------|------------------------|-----------------|--------------------------------------|
| Delete1       | Succeeded       | Delete        | 5/31/2024, 1:50:15 PM | 8s       | AutoResolveIntegration |                 | dbefd413-506b-4fa9-92ab-6ab8b6bc2dba |

adf-training

Container

Search

Upload

Change access level

Refresh

Delete

Change tier

Acquire lease

Break lease

View snapshots

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: adf-training / Input

Search blobs by prefix (case-sensitive)

Show deleted blobs

Add filter

| Name   | Modified              | Access tier    | Archive status | Blob type  | Size |
|--|-----------------------|----------------|----------------|------------|------|
| <input type="checkbox"/> .                       |                       |                |                |            |      |
| <input type="checkbox"/> Archive                 |                       |                |                |            |      |
| <input type="checkbox"/> product_ss_29052024.csv | 5/30/2024, 4:23:55 PM | Hot (Inferred) |                | Block blob | 115  |
| <input type="checkbox"/> product_ss_30052024.csv | 5/30/2024, 4:23:55 PM | Hot (Inferred) |                | Block blob | 116  |

Summary

- This document provides a detailed overview of the steps taken to create and manage the “bt\_pl\_ff\_product\_data\_ld” pipeline in Azure Data Factory.