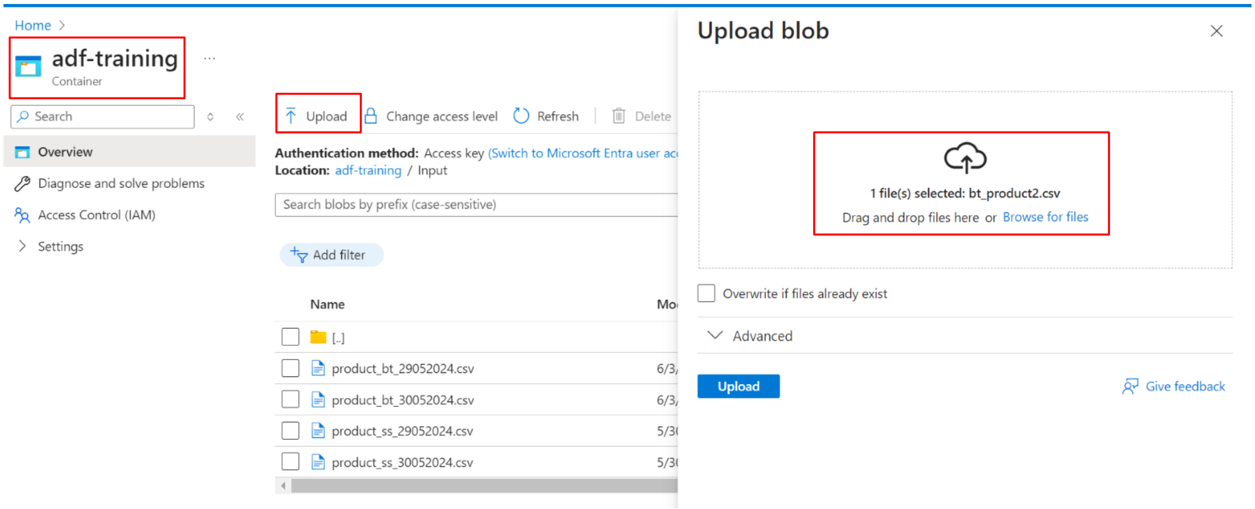
Formal Documentation of Azure Data Factory Pipeline – Training

Use case: Copying data to the output folder and archiving it with the current date and IST time in the archive folder.

Steps:

1. Loading the file into the container:

- A container named adf-training was previously created in the Azure storage account.
- For this pipeline, the files were uploaded into the container, in the input folder (refer to image 1)
- File names: bt_product2.csv
- Once the upload was successful, the file was added to the input folder.



Authentication method: Access key (Switch to Microsoft Entra user ac

Location: adf-training / Input

Name	Mo
[..]	
product_bt_29052024.csv	6/3,
product_bt_30052024.csv	6/3,
product_ss_29052024.csv	5/3(
product_ss_30052024.csv	5/3(

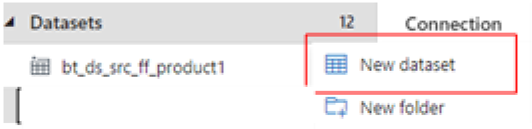
1 file(s) selected: bt_product2.csv

Drag and drop files here or Browse for files

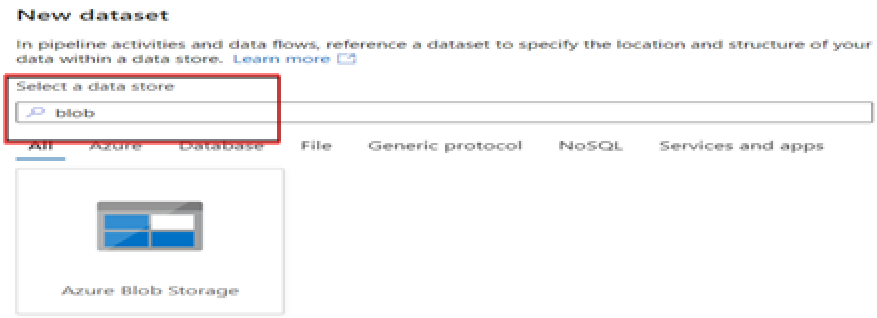
2. Dataset creation.

Source Dataset

- A new source dataset was created for the source data files.
- Source dataset name: bt_ds_src_task_data_ld
- Inside Azure Data Factory, in the Author tab, I selected the Dataset option and clicked on "New Dataset" (refer to image 1).
- I chose the Azure Blob Storage option (refer to image 2,3). Next, I selected the Delimited Text file format, which brought me to the properties page where I defined the dataset name and path
- I specified the dataset name, selected the linked service, and provided the path of my input file.
- These steps created my source dataset.
- I opened my source file and updated the connection, changing the column delimiter option to pipe (|) because my CSV file is pipe delimited (refer image 5).



- Image 1



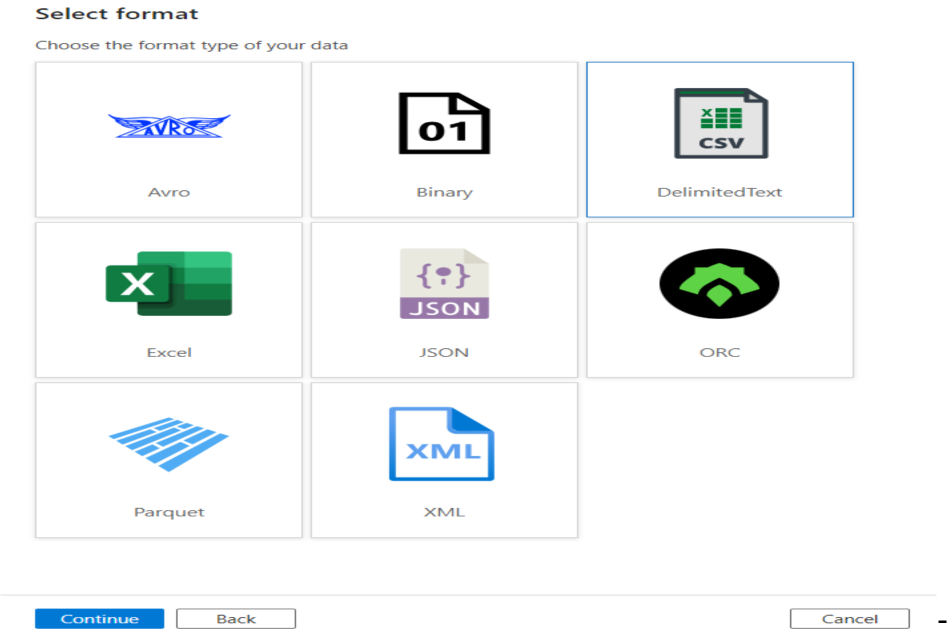
New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more

Select a data store

blob

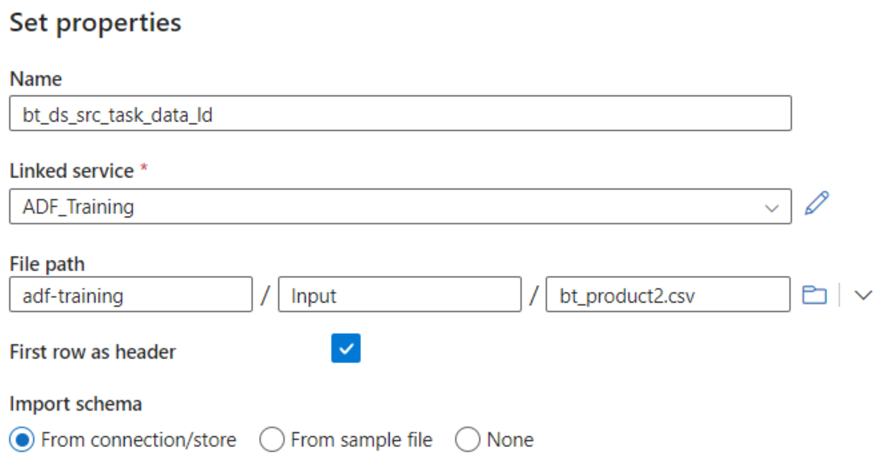
- Image 2



Select format

Choose the format type of your data

- Image 3



Set properties

Name: bt_ds_src_task_data_ld

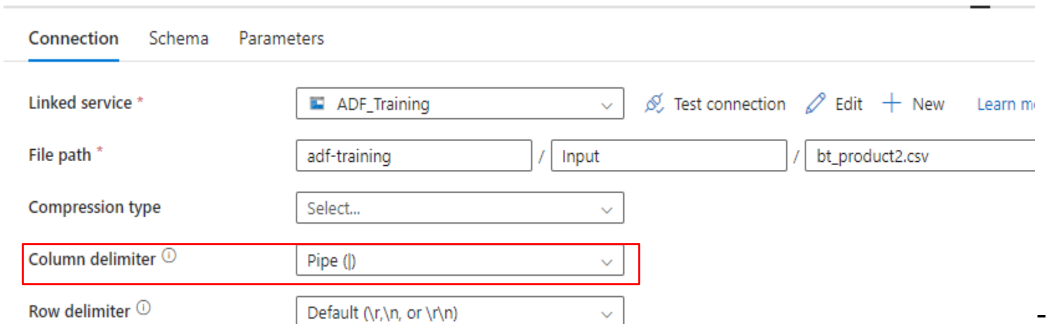
Linked service *: ADF_Training

File path: adf-training / Input / bt_product2.csv

First row as header ☑

Import schema: ● From connection/store ○ From sample file ○ None

- Image 4



Linked service *: ADF_Training

File path *: adf-training / Input / bt_product2.csv

Compression type: Select...

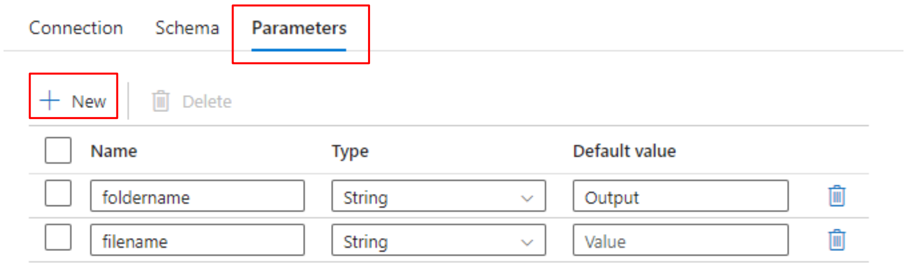
Column delimiter: Pipe (|)

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

- Image 5

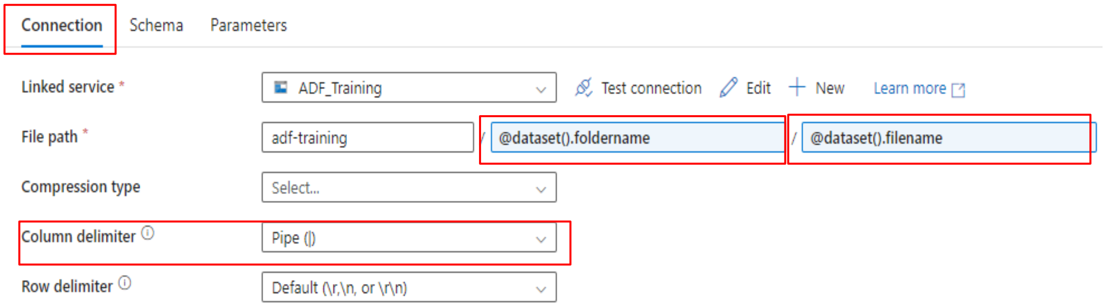
Target dataset.

- I followed similar steps for the target dataset. In the Author tab, I selected the Datasets option, clicked on "New Dataset," selected Azure Blob Storage, and then selected the Delimited Text format, which brought me to the properties page where I defined the dataset name and path.
- Assigned path: adf-training/output
- Target dataset name: bt_ds_tgt_task_data_ld
- I opened my target dataset and updated the connection, changing the column delimiter option to pipe (|) because my CSV file is pipe delimited.
- Then add the parameters, I added parameters in the sink dataset while copying data in Azure to make the dataset dynamic, reusable, and adaptable to different data destinations and environments. I added parameters for the folder and file name (refer image 1)
- Then in connection settings, by clicking on add dynamic content on folder and file, I added an expression (refer image 2,3,4).



	Name	Type	Default value
□	foldername	String	Output
□	filename	String	Value

- Image 1



Linked service * ADF_Training

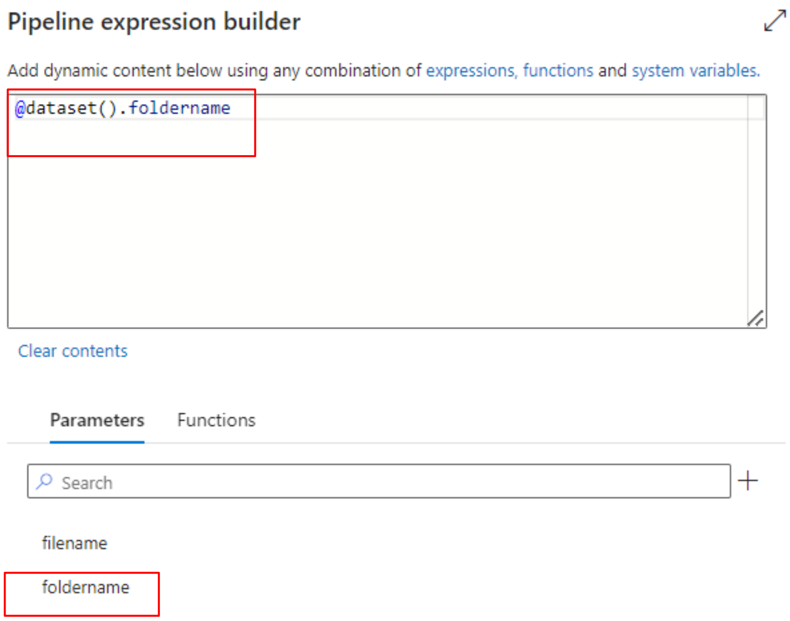
File path * adf-training / @dataset().foldername / @dataset().filename

Compression type Select...

Column delimiter Pipe (|)

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

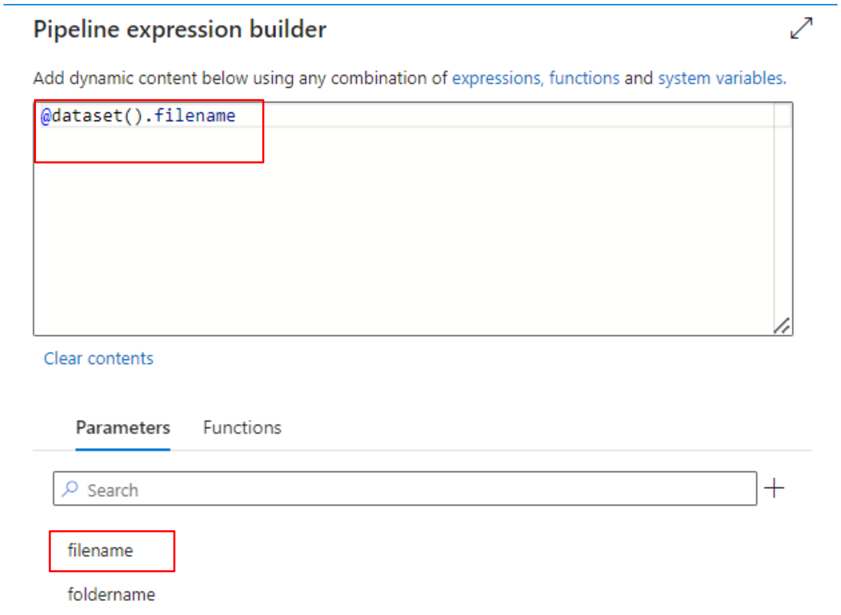
- Image 2


Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@dataset().foldername
```

- Image 3



Pipeline expression builder

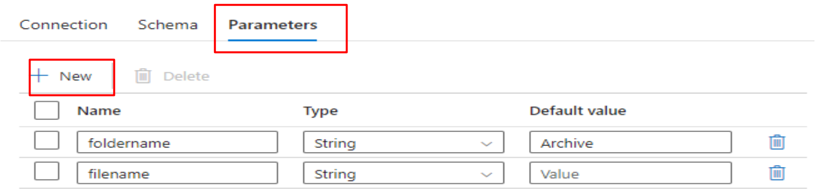
Add dynamic content below using any combination of expressions, functions and system variables.

```
@dataset().filename
```

- Image 4

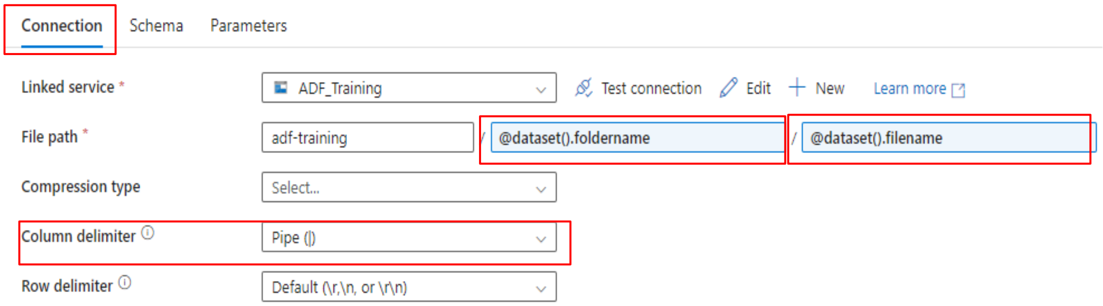
Archive dataset.

- I followed similar steps for the archive dataset. In the Author tab, I selected the Datasets option, clicked on "New Dataset," selected Azure Blob Storage, and then selected the Delimited Text format, which brought me to the properties page where I defined the dataset name and path.
- Assigned path: adf-training/Archive.
- Target dataset name: bt_ds_arc_task_data_ld
- I opened my archive dataset and updated the connection, changing the column delimiter option to pipe (|) because my CSV file is pipe delimited.
- Then add the parameters, I added parameters in the sink dataset while copying data in Azure to make the dataset dynamic, reusable, and adaptable to different data destinations and environments. I added parameters for the folder and file name. (refer image 1)
- Then in connection settings, by clicking on add dynamic content on folder and file, I added an expression.



□	Name	Type	Default value
□	foldername	String	Archive
□	filename	String	Value

- Image 1



Linked service * ADF_Training

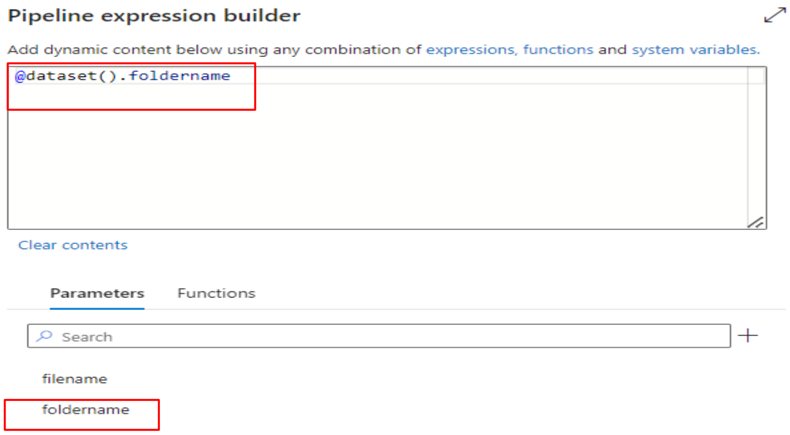
File path * adf-training / @dataset().foldername / @dataset().filename

Compression type Select...

Column delimiter Pipe (|)

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

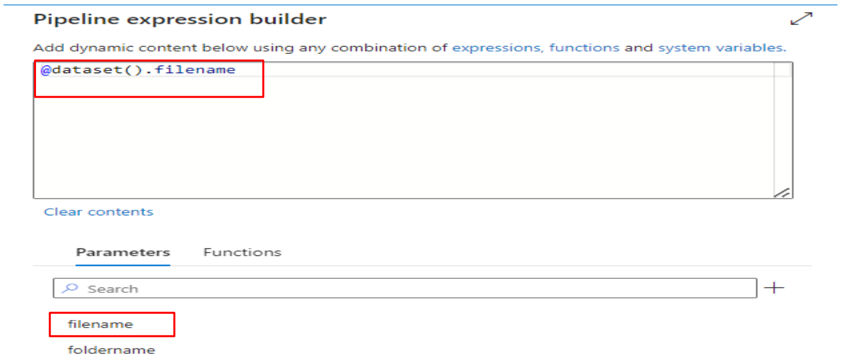
- Image 2


Pipeline expression builder

Add dynamic content below using any combination of expressions, functions and system variables.

```
@dataset().foldername
```

- Image 3


Pipeline expression builder

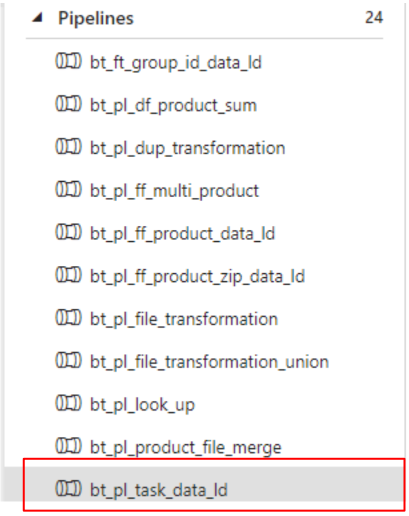
Add dynamic content below using any combination of expressions, functions and system variables.

```
@dataset().filename
```

- Image 4

3. Pipeline creation.

- A new pipeline named bt_pl_task_data_ld was created in Azure Data Factory (ADF) to load the file in the output and archive folder (refer image 1).



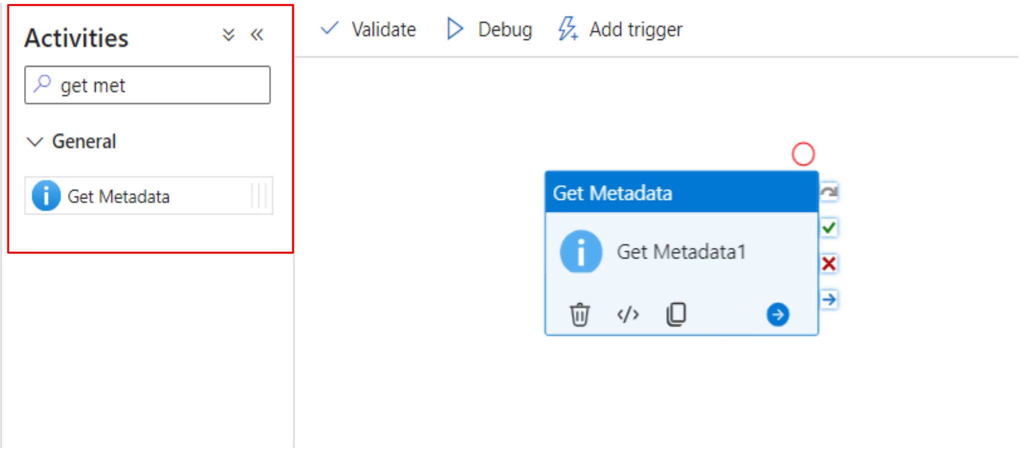
Pipelines 24

- bt_ft_group_id_data_Id
- bt_pl_df_product_sum
- bt_pl_dup_transformation
- bt_pl_ff_multi_product
- bt_pl_ff_product_data_Id
- bt_pl_ff_product_zip_data_Id
- bt_pl_file_transformation
- bt_pl_file_transformation_union
- bt_pl_look_up
- bt_pl_product_file_merge
- bt_pl_task_data_Id

- Image 1

4. Get Metadata Activity

- I used the Get Metadata activity to get all the files/subfolders in the given path (refer image 1)
- Added the source files and selected exist and items names from the field list.
- We use "exist" and "item names" in the Get Metadata activity in Azure to check the existence of a file and retrieve its metadata attributes dynamically (refer image 2).



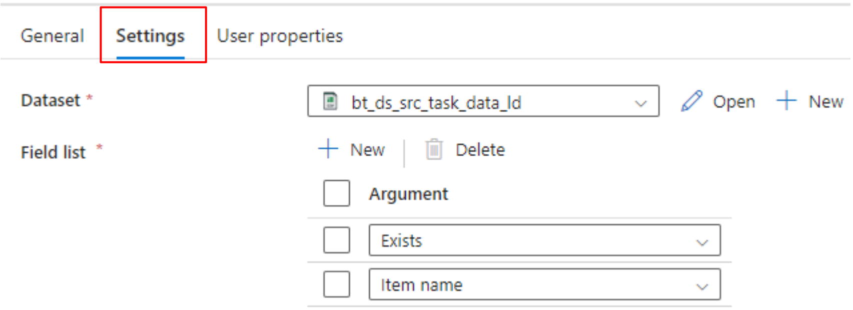
Activities

get met

General

Get Metadata

- Image 1



Dataset * bt_ds_src_task_data_Id

Field list *

	Argument
□	Exists
□	Item name

- Image 2

5. If Condition Activity.

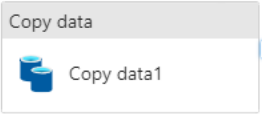
I then used If condition activity to execute different actions based on whether a specified condition is true or false.

True condition.

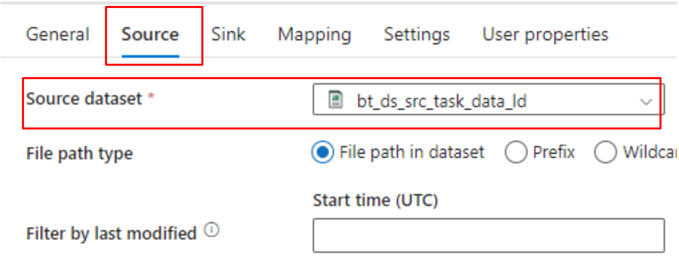
A. Copy Activity 1

- Inside if condition, I used copy activity to copy the file from source to target (output folder).
- Added the source file in source tab (refer image 2).
- Added sink file in the sink tab. Since I added parameters in our target dataset, dataset properties appeared here, and I assigned the folder and file name (refer image 3).

bt_pl_task_data_ld > If Condition1 - True activities



- Image 1.



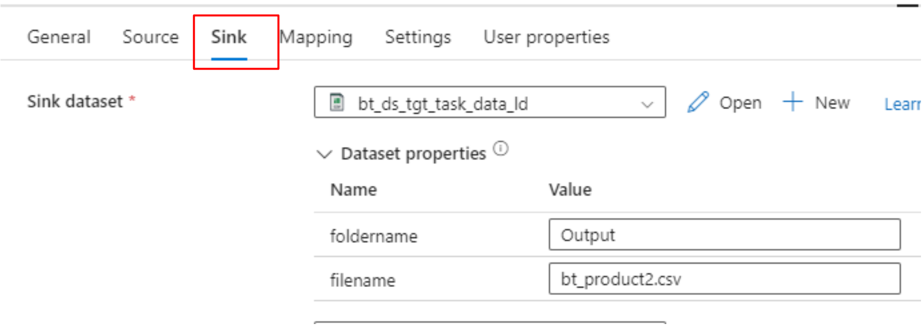
Source dataset * bt_ds_src_task_data_ld

File path type ● File path in dataset ○ Prefix ○ Wildca

Filter by last modified ⓘ Start time (UTC)

- Image 2

.



Sink dataset * bt_ds_tgt_task_data_ld

Dataset properties ⓘ

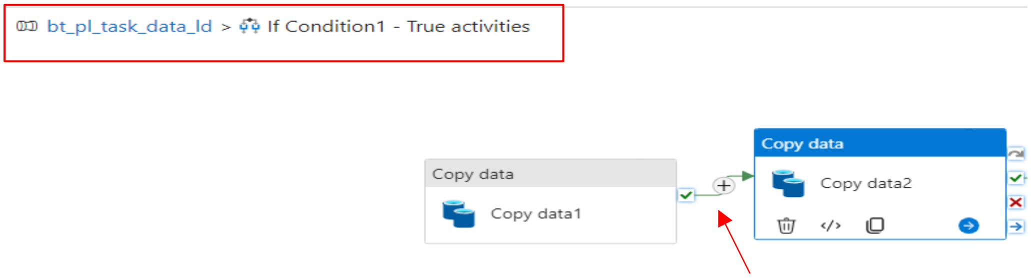
Name	Value
foldername	Output
filename	bt_product2.csv

- Image 3

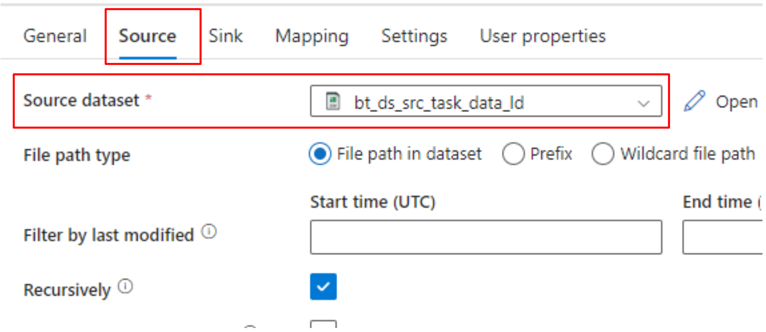
B. Copy Activity 2

- I used another copy activity to copy the file from source to target (Archive folder) (refer image 1).

- I wanted the file name to include the current date and IST time when the pipeline is executed.
- Added the source file in source tab (refer image 2).
- Added sink file in the sink tab. Since I added parameters in our archive dataset, dataset properties appeared here, and I assigned the folder name.
- In the file name, I added a dynamic expression that will print the file name along with the date and IST time (refer image 3,4)
- @concat(replace(activity('Get Metadata1').output.itemName,'.csv',''), '_',formatDateTime(convertFromUtc(utcnow(), 'India Standard Time')),'.csv') - this expression renames a file by removing the ".csv" extension, appending the current date and time in India Standard Time, and adding ".csv" back at the end.
- I connected CopyActivity1 with the CopyActivty2 on success as I want the archive file to print after the file is printed in the output folder.



- Image 1



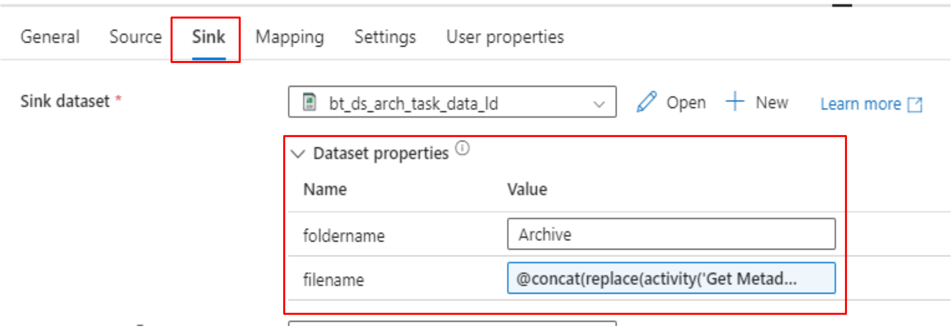
Source dataset * bt_ds_src_task_data_ld

File path type ● File path in dataset ○ Prefix ○ Wildcard file path

Filter by last modified Start time (UTC) End time

Recursively ☑

- Image 2

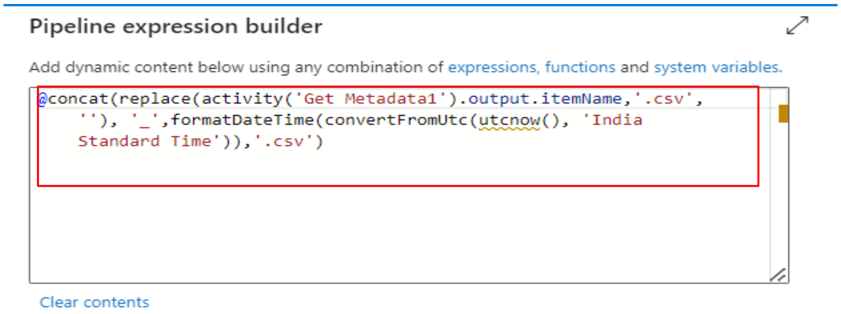


Sink dataset * bt_ds_arch_task_data_ld

Dataset properties

Name	Value
foldername	Archive
filename	@concat(replace(activity('Get Metad...

- Image 3



Pipeline expression builder

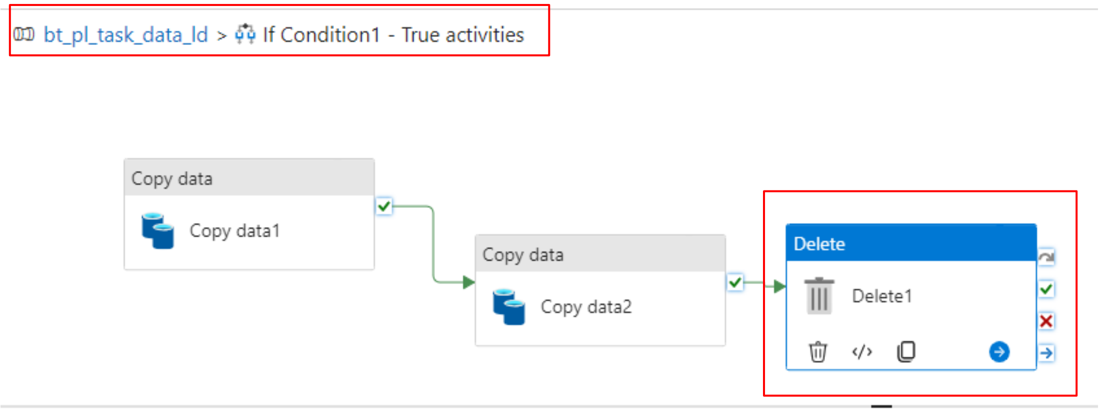
Add dynamic content below using any combination of expressions, functions and system variables.

```
@concat(replace(activity('Get Metadata1').output.itemName,'.csv',
    ''), '_',formatDateTime(convertFromUtc(utcnow(), 'India
    Standard Time')),'.csv')
```

- Image 4

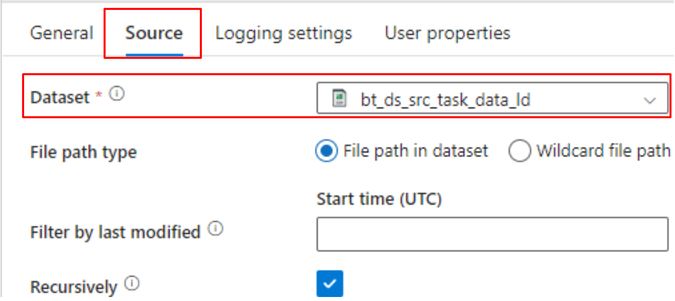
__C. Delete Activity__

- I used delete activity to delete the source file from the input folder in order to avoid any duplication (refer image 1,2).



bt_pl_task_data_ld > If Condition1 - True activities

-Image 1



Dataset * ⓘ bt_ds_src_task_data_ld

File path type ● File path in dataset ○ Wildcard file path

Start time (UTC)

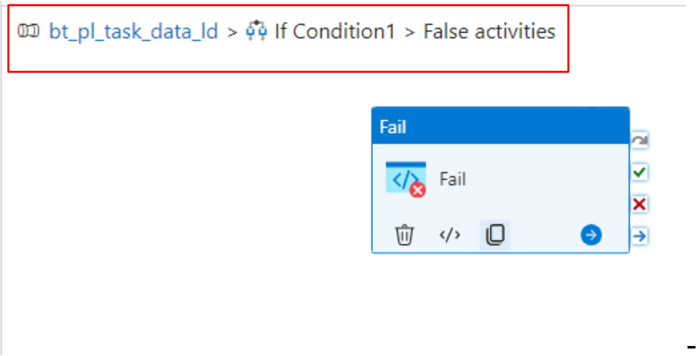
Filter by last modified ⓘ

Recursively ⓘ ☑

-Image 2

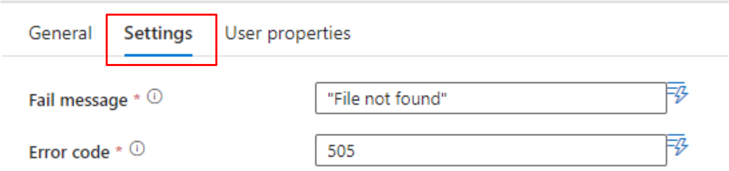
False condition.

- I added a fail activity inside the false condition from the Activities bar to explicitly handle and signal failures, ensuring proper error handling and logging even if the true condition is expected to succeed.
- In the settings, I added the error message and error code (refer image 1,2).



bt_pl_task_data_ld > If Condition1 > False activities

-Image 1



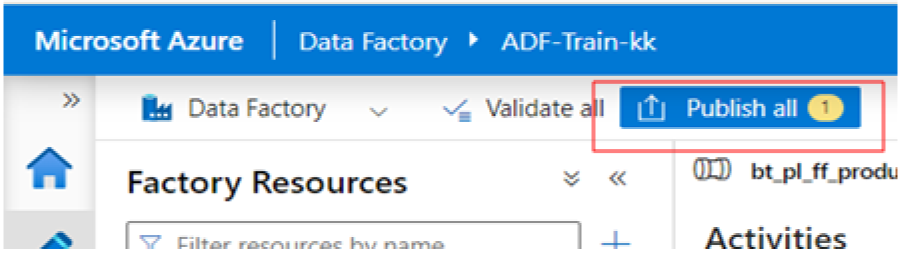
Fail message * "File not found"

Error code * 505

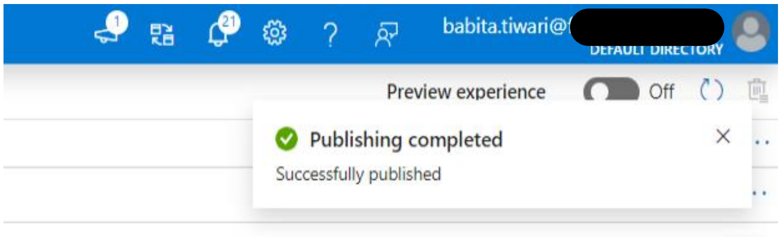
-Image 2

6. Publishing and Executing the Pipeline.

The activities were saved/ published, and all were successfully published and executed (Image 1,2,3,4)



- Image 1

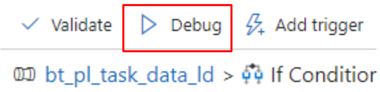


[redacted]

Publishing completed

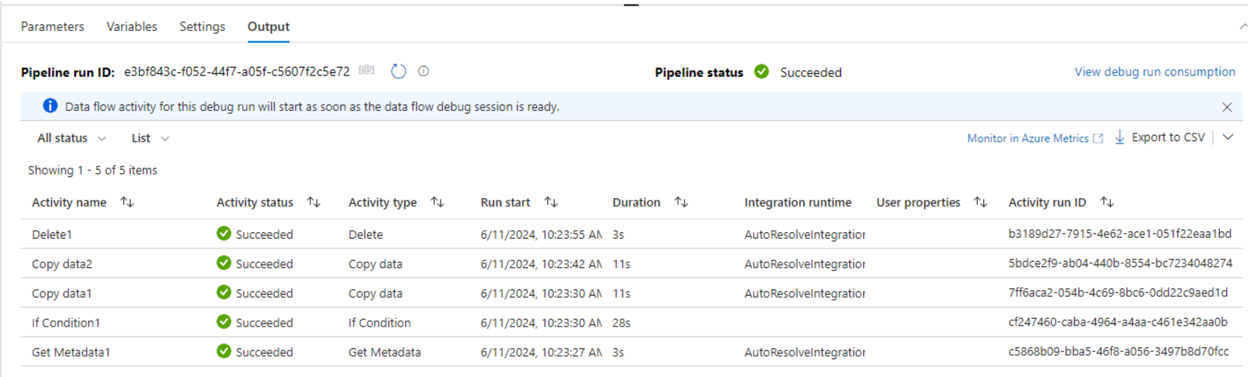
Successfully published

- Image 2



bt_pl_task_data_ld > If Conditior

-Image 3



Pipeline run ID: e3bf843c-f052-44f7-a05f-c5607f2c5e72

Pipeline status ✓ Succeeded

Data flow activity for this debug run will start as soon as the data flow debug session is ready.

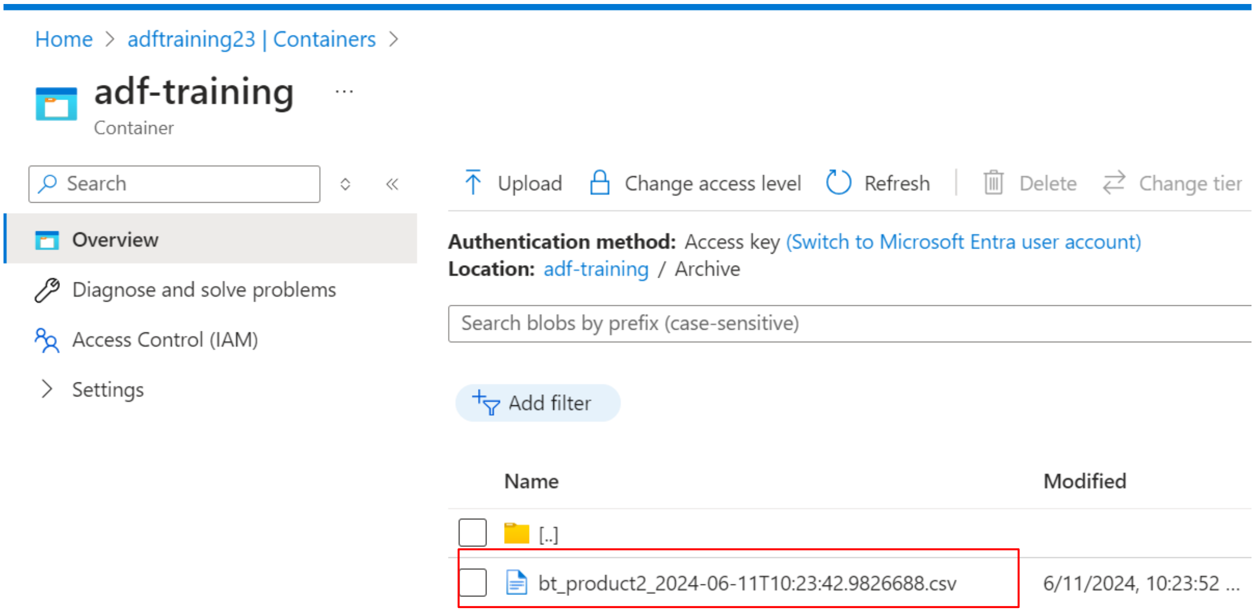
Showing 1 - 5 of 5 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Delete1	✓ Succeeded	Delete	6/11/2024, 10:23:55 AM	3s	AutoResolveIntegratior		b3189d27-7915-4e62-ace1-051f22eaa1bd
Copy data2	✓ Succeeded	Copy data	6/11/2024, 10:23:42 AM	11s	AutoResolveIntegratior		5bdce2f9-ab04-440b-8554-bc7234048274
Copy data1	✓ Succeeded	Copy data	6/11/2024, 10:23:30 AM	11s	AutoResolveIntegratior		7ff6aca2-054b-4c69-8bc6-0dd22c9aed1d
If Condition1	✓ Succeeded	If Condition	6/11/2024, 10:23:30 AM	28s			cf247460-caba-4964-a4aa-c461e342aa0b
Get Metadata1	✓ Succeeded	Get Metadata	6/11/2024, 10:23:27 AM	3s	AutoResolveIntegratior		c5868b09-bba5-46f8-a056-3497b8d70fcc

-Image 4

Verifying the output.

- I checked the output folder inside the adf-training container and my file was successfully added there.
- Also, I checked the archive folder, and the file was printed along with the data and time (refer image 1).


Home > adftraining23 | Containers >

adf-training
Container

Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Upload
Change access level
Refresh
Delete
Change tier

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: adf-training / Archive

Search blobs by prefix (case-sensitive)
Add filter

	Name	Modified
□	[..]	
□	bt_product2_2024-06-11T10:23:42.9826688.csv	6/11/2024, 10:23:52 ...

-Image 1

Summary:

The pipeline was created to copy data into the output folder and archive it with the current date and IST time.