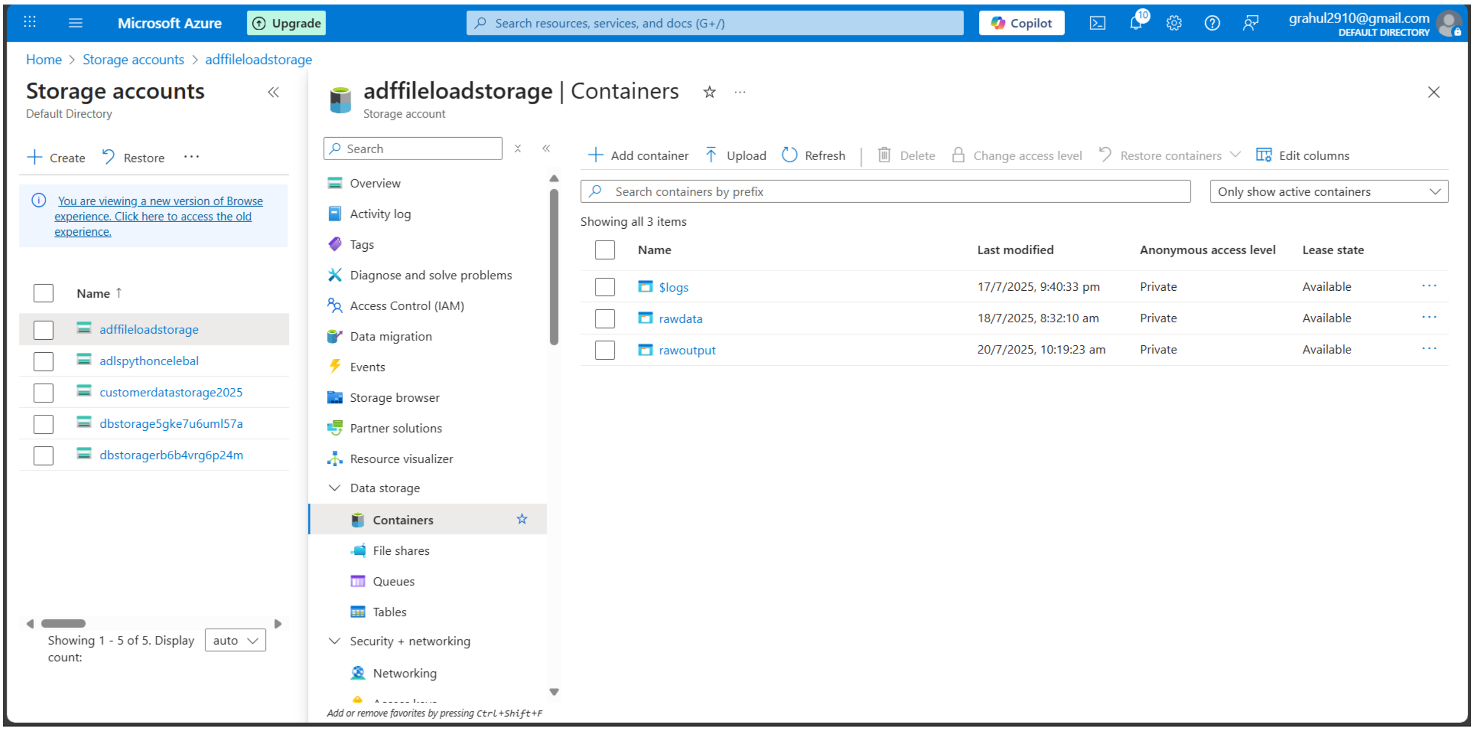
FILE COPY WITH TRANSFORMATION

1. Setup Resources

- Created Azure Data Lake Gen2 container and uploaded all types of test files.



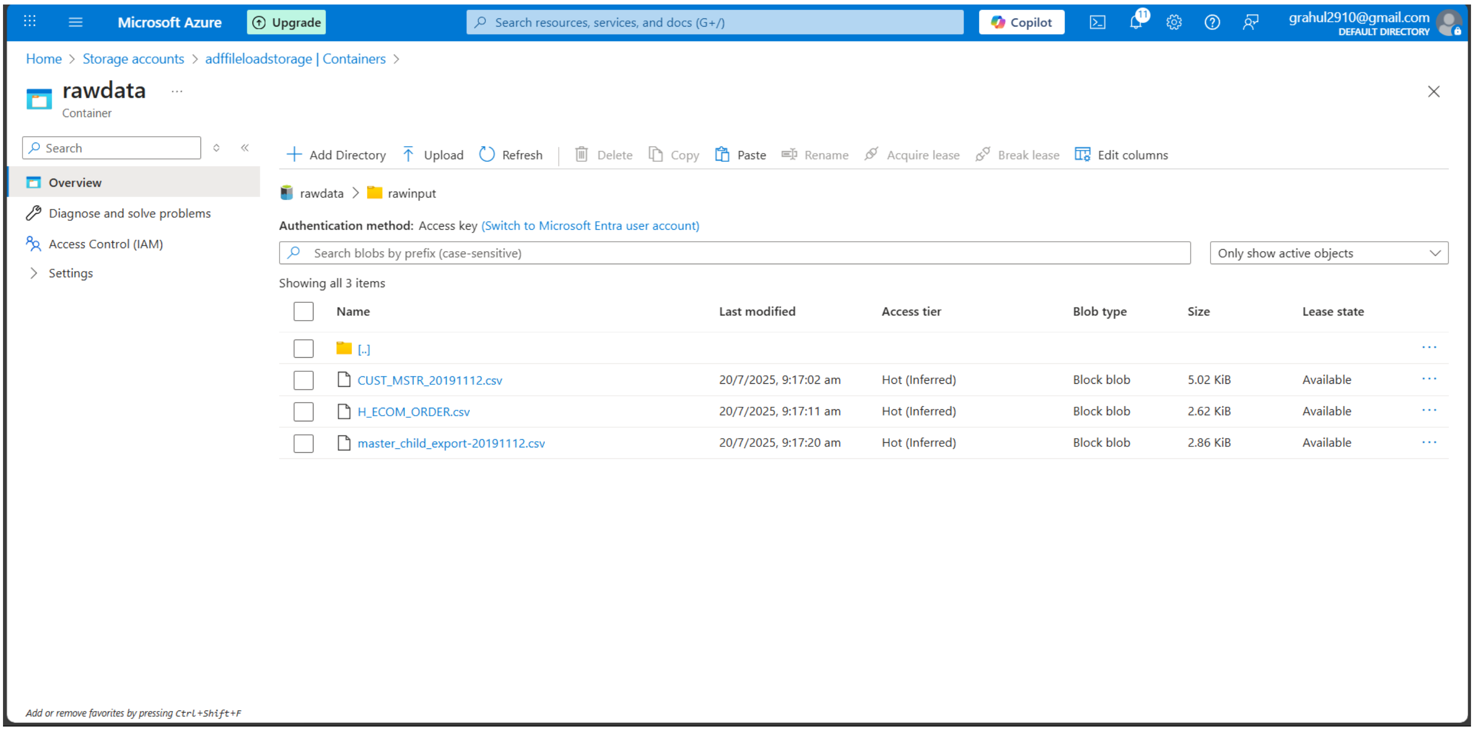
You are viewing a new version of Browse experience. Click here to access the old experience.

- adffileloadstorage
- adlspythoncelebal
- customerdatastorage2025
- dbstorage5gke7u6uml57a
- dbstoragerb6b4vrg6p24m

Showing all 3 items

Name	Last modified	Anonymous access level	Lease state
$logs	17/7/2025, 9:40:33 pm	Private	Available
rawdata	18/7/2025, 8:32:10 am	Private	Available
rawoutput	20/7/2025, 10:19:23 am	Private	Available

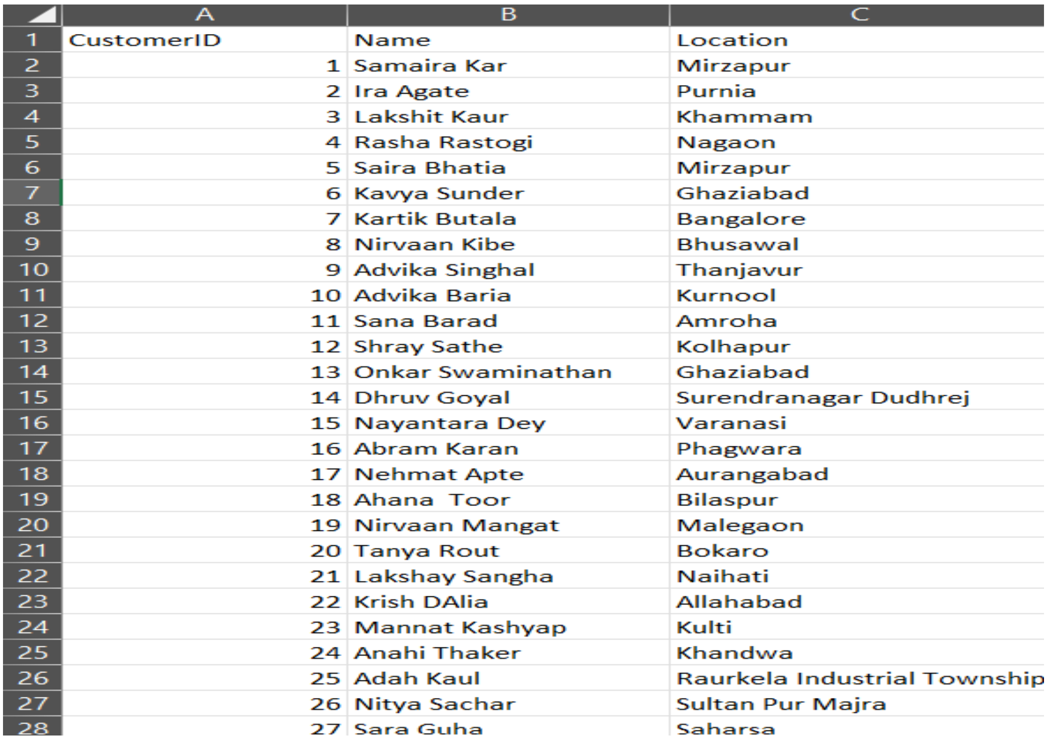
- Created Azure SQL Database with the three required tables:
  - CUST_MSTR(Date, ...)
  - master_child(Date, DateKey, ...)
  - H_ECOM_Orders(...)



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

Showing all 3 items

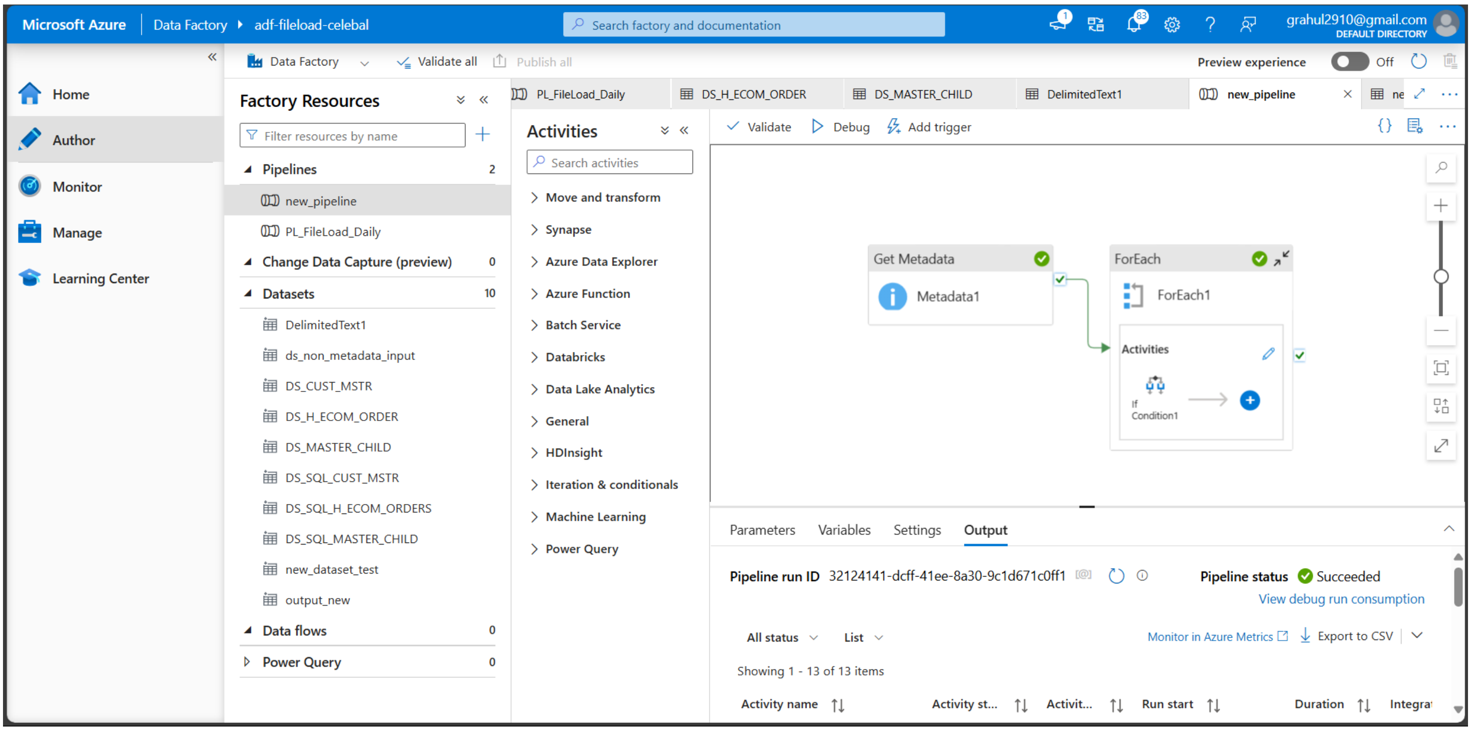
Name	Last modified	Access tier	Blob type	Size	Lease state
[..]					
CUST_MSTR_20191112.csv	20/7/2025, 9:17:02 am	Hot (Inferred)	Block blob	5.02 KiB	Available
H_ECOM_ORDER.csv	20/7/2025, 9:17:11 am	Hot (Inferred)	Block blob	2.62 KiB	Available
master_child_export-20191112.csv	20/7/2025, 9:17:20 am	Hot (Inferred)	Block blob	2.86 KiB	Available



	A	B	C
1	CustomerID	Name	Location
2	1	Samaira Kar	Mirzapur
3	2	Ira Agate	Purnia
4	3	Lakshit Kaur	Khammam
5	4	Rasha Rastogi	Nagaon
6	5	Saira Bhatia	Mirzapur
7	6	Kavya Sunder	Ghaziabad
8	7	Kartik Butala	Bangalore
9	8	Nirvaan Kibe	Bhusawal
10	9	Advika Singhal	Thanjavur
11	10	Advika Baria	Kurnool
12	11	Sana Barad	Amroha
13	12	Shray Sathe	Kolhapur
14	13	Onkar Swaminathan	Ghaziabad
15	14	Dhruv Goyal	Surendranagar Dudhrej
16	15	Nayantara Dey	Varanasi
17	16	Abram Karan	Phagwara
18	17	Nehmat Apte	Aurangabad
19	18	Ahana Toor	Bilaspur
20	19	Nirvaan Mangat	Malegaon
21	20	Tanya Rout	Bokaro
22	21	Lakshay Sangha	Naihati
23	22	Krish DAlia	Allahabad
24	23	Mannat Kashyap	Kulti
25	24	Anahi Thaker	Khandwa
26	25	Adah Kaul	Raurkela Industrial Township
27	26	Nitya Sachar	Sultan Pur Majra
28	27	Sara Guha	Saharsa

2. Created ADF Pipeline

- Designed a single pipeline with:
  - A Get Metadata activity to list all files from the container.
  - A ForEach activity to iterate through each file.



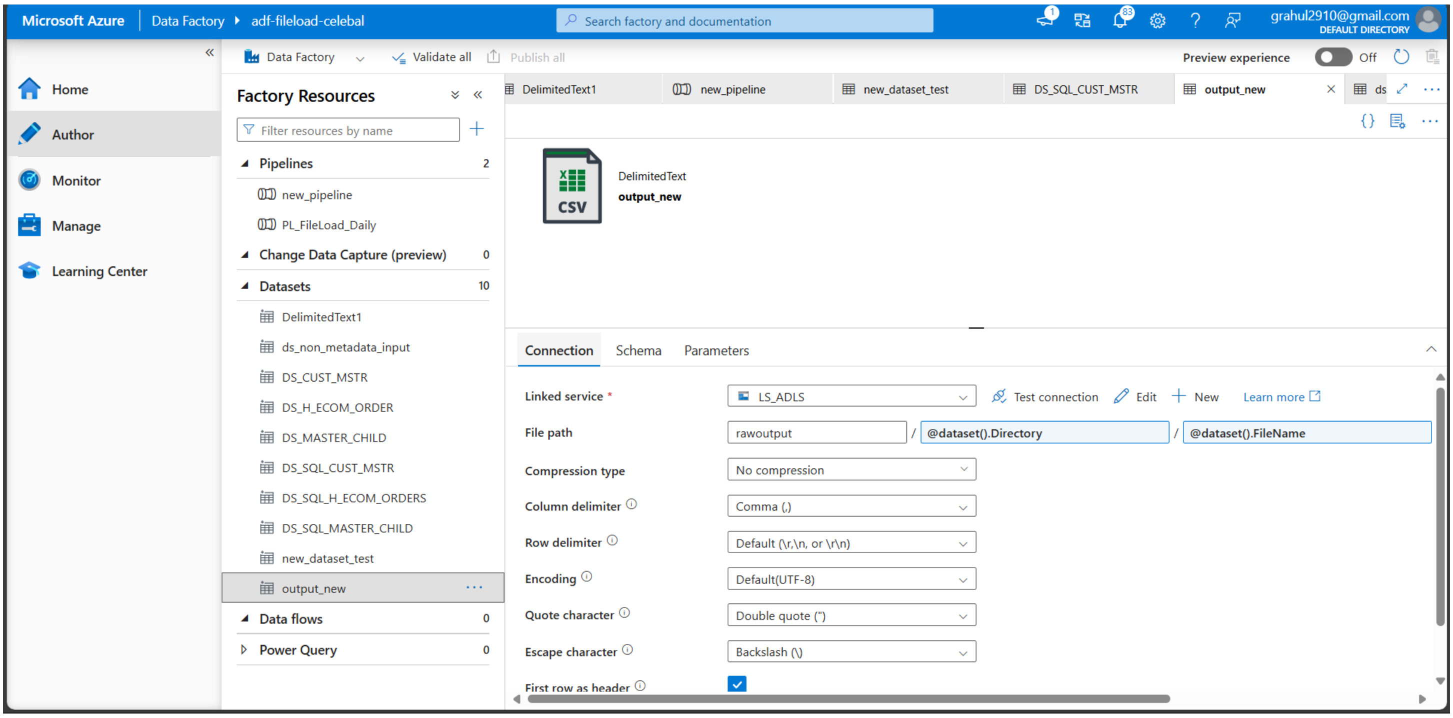
Pipeline run ID 32124141-dcff-41ee-8a30-9c1d671c0ff1

Pipeline status Succeeded

View debug run consumption

Showing 1 - 13 of 13 items

Make the datasets and make them parameterized, so that they can get values directly from the file name



Microsoft Azure | Data Factory ▸ adf-fileload-celebal

Factory Resources

- Pipelines 2
  - new_pipeline
  - PL_FileLoad_Daily
- Change Data Capture (preview) 0
- Datasets 10
  - DelimitedText1
  - ds_non_metadata_input
  - DS_CUST_MSTR
  - DS_H_ECOM_ORDER
  - DS_MASTER_CHILD
  - DS_SQL_CUST_MSTR
  - DS_SQL_H_ECOM_ORDERS
  - DS_SQL_MASTER_CHILD
  - new_dataset_test
  - output_new
- Data flows 0
- Power Query 0

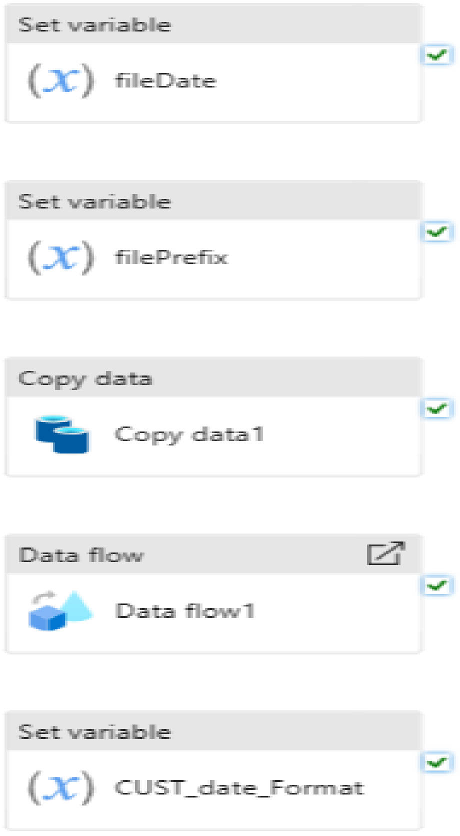
DelimitedText
output_new

Connection | Schema | Parameters

Linked service *	LS_ADLS
File path	rawoutput / @dataset().Directory / @dataset().FileName
Compression type	No compression
Column delimiter	Comma (,)
Row delimiter	Default (\r,\n, or \r\n)
Encoding	Default(UTF-8)
Quote character	Double quote (")
Escape character	Backslash (\)
First row as header	☑

3. Used Set Variable Activities

- Extracted fileName using @item().name.
- Created two variables:
  - file_date → Format YYYY-MM-DD
  - CUST_date_format → Format YYYYMMDD



- Used expressions like:

substring(replace(item().name, '.csv', ''), length(...) - 8, 8)

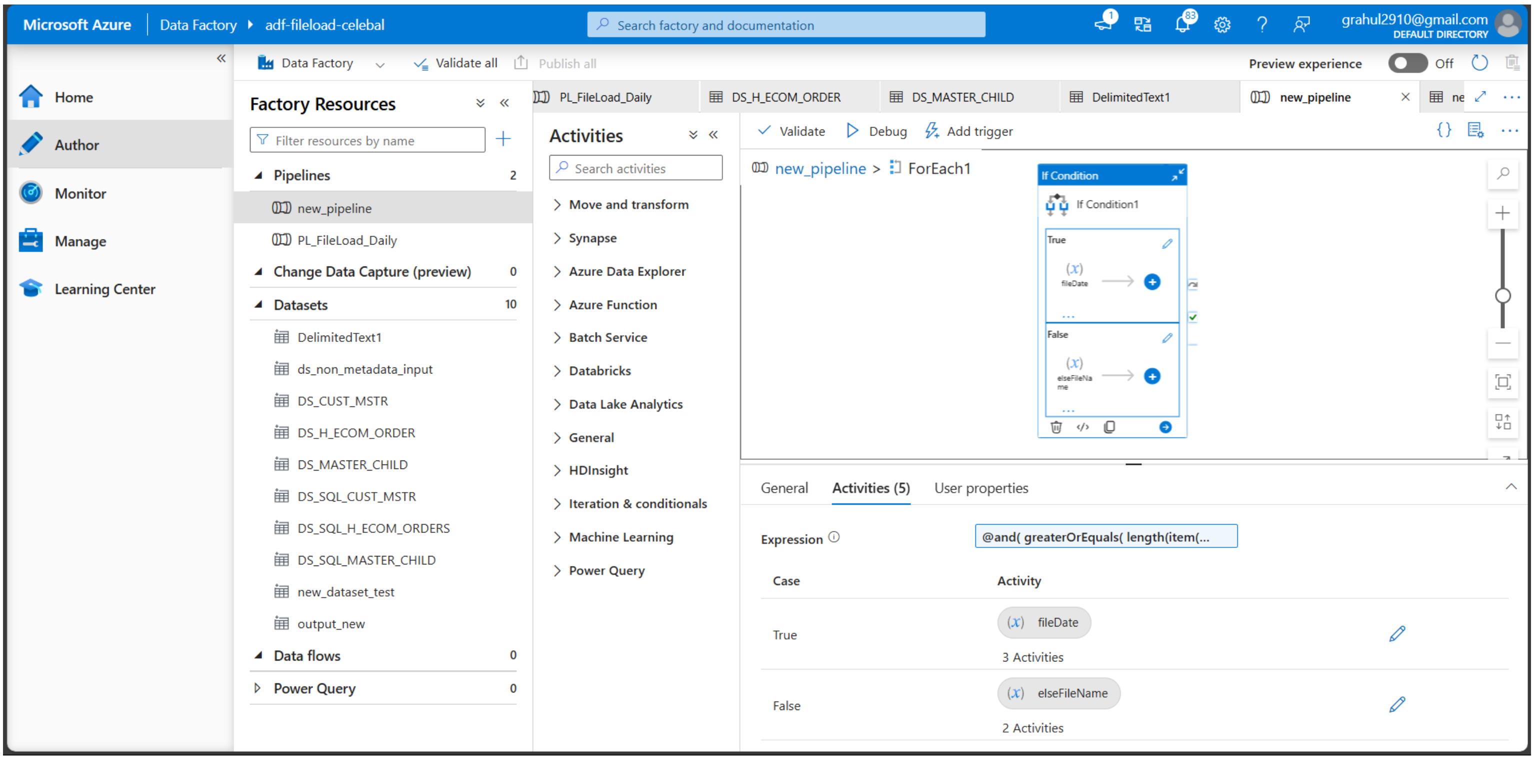
to extract the date from the filename.

- Transformed the extracted dateKey into proper date format using:

@concat(substring(variables(CUST_date_format '), 0, 4), '-', substring(variables(' CUST_date_format '), 4, 2), '-', substring(variables(' CUST_date_format '), 6, 2))

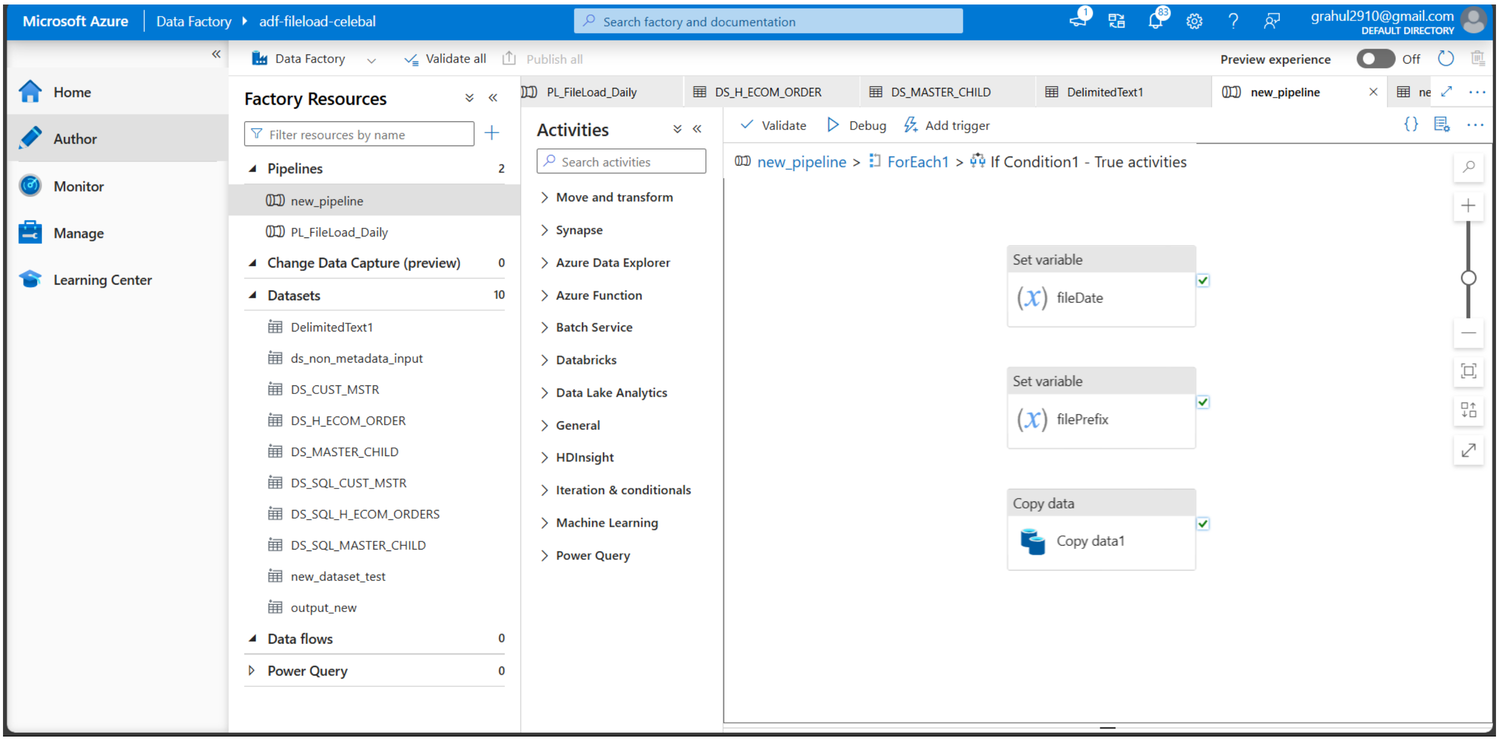
4. Used If Conditions

- Inside the ForEach, placed 3 If Conditions:
  - If filename starts with CUST_MSTR
  - If filename starts with master_child_export
  - If filename starts with H_ECOM_ORDER



Expression: @and(greaterOrEquals(length(item(...

Case	Activity
True	fileDate 3 Activities
False	elseFileName 2 Activities



5. Defined Data Flows

- For CUST_MSTR:
  - Used Data Flow to add one derived column: Date from file_date
  - Sink: Load to CUST_MSTR table (with truncate behavior)
- For master_child_export:
  - Used Data Flow to add two derived columns:
    - Date from file_date
    - DateKey from CUST_date_format
  - Sink: Load to master_child table
- For H_ECOM_ORDER:
  - Direct copy activity (or simple data flow) without any transformations.
  - Sink: Load to H_ECOM_Orders table

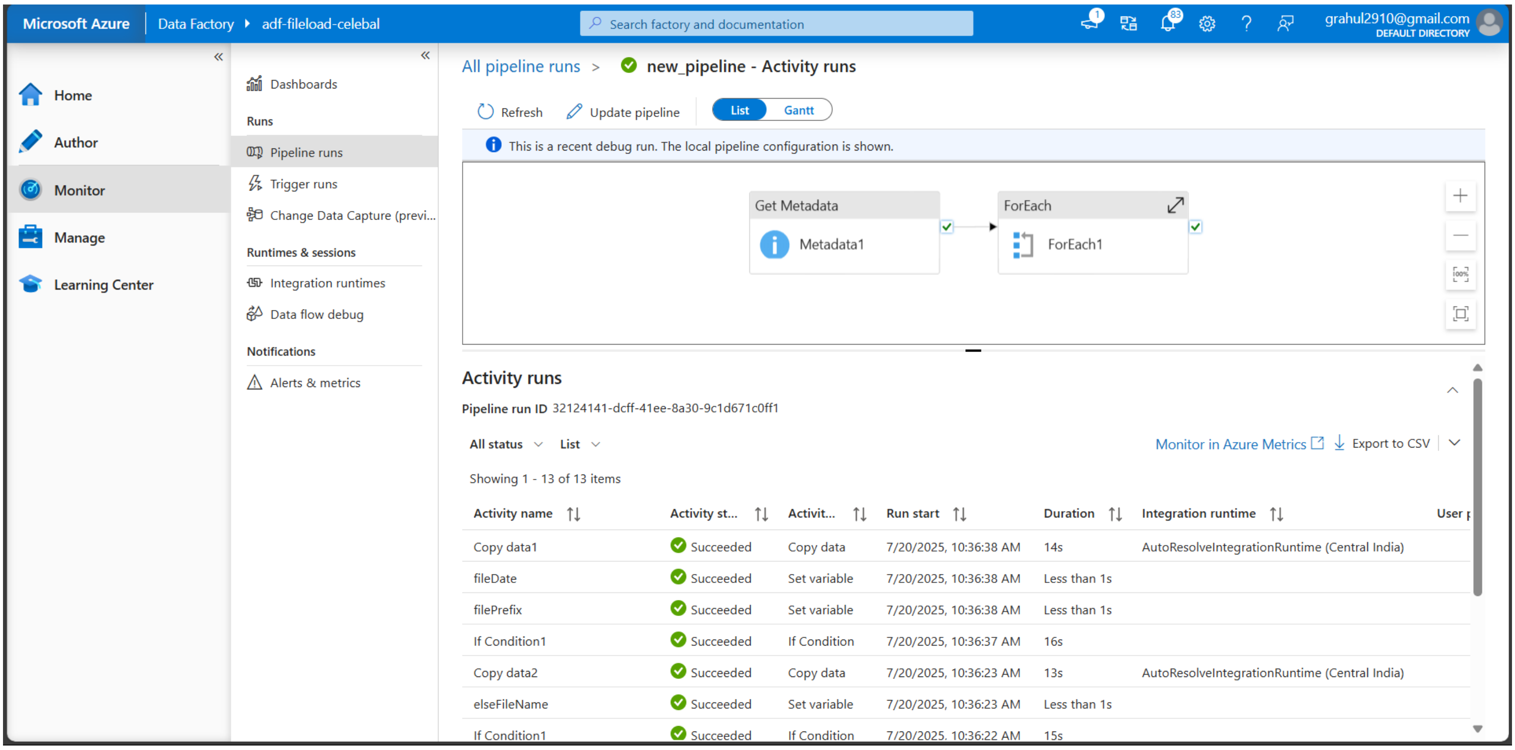
6. Truncate Load Logic

- Enabled truncate option in each sink to ensure tables are cleared before inserting new data.

7. Debug and Testing

- The pipeline should be debugged in ADF to:
  - Confirm file iteration logic
  - Validate column additions
  - Ensure successful data writes

- Monitor any errors or data issues



This is a recent debug run. The local pipeline configuration is shown.

Activity runs

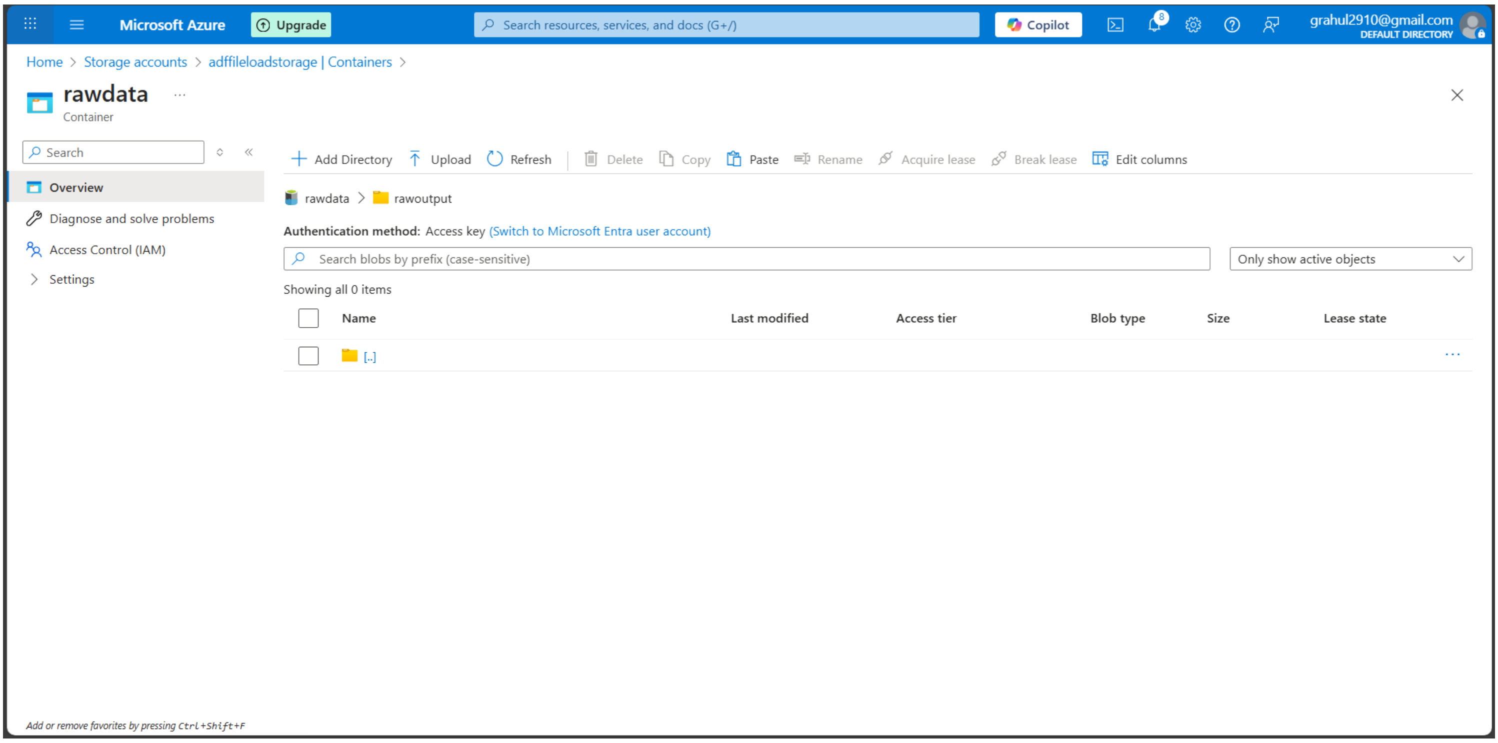
Pipeline run ID 32124141-dcff-41ee-8a30-9c1d671c0ff1

Showing 1 - 13 of 13 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User p
Copy data1	✓ Succeeded	Copy data	7/20/2025, 10:36:38 AM	14s	AutoResolveIntegrationRuntime (Central India)	
fileDate	✓ Succeeded	Set variable	7/20/2025, 10:36:38 AM	Less than 1s		
filePrefix	✓ Succeeded	Set variable	7/20/2025, 10:36:38 AM	Less than 1s		
If Condition1	✓ Succeeded	If Condition	7/20/2025, 10:36:37 AM	16s		
Copy data2	✓ Succeeded	Copy data	7/20/2025, 10:36:23 AM	13s	AutoResolveIntegrationRuntime (Central India)	
elseFileName	✓ Succeeded	Set variable	7/20/2025, 10:36:23 AM	Less than 1s		
If Condition1	✓ Succeeded	If Condition	7/20/2025, 10:36:22 AM	15s		

8. Output Folder Structure

- After the pipeline runs, the processed files are stored in the rawdata/rawoutput/ directory of the data lake.



rawdata

Container

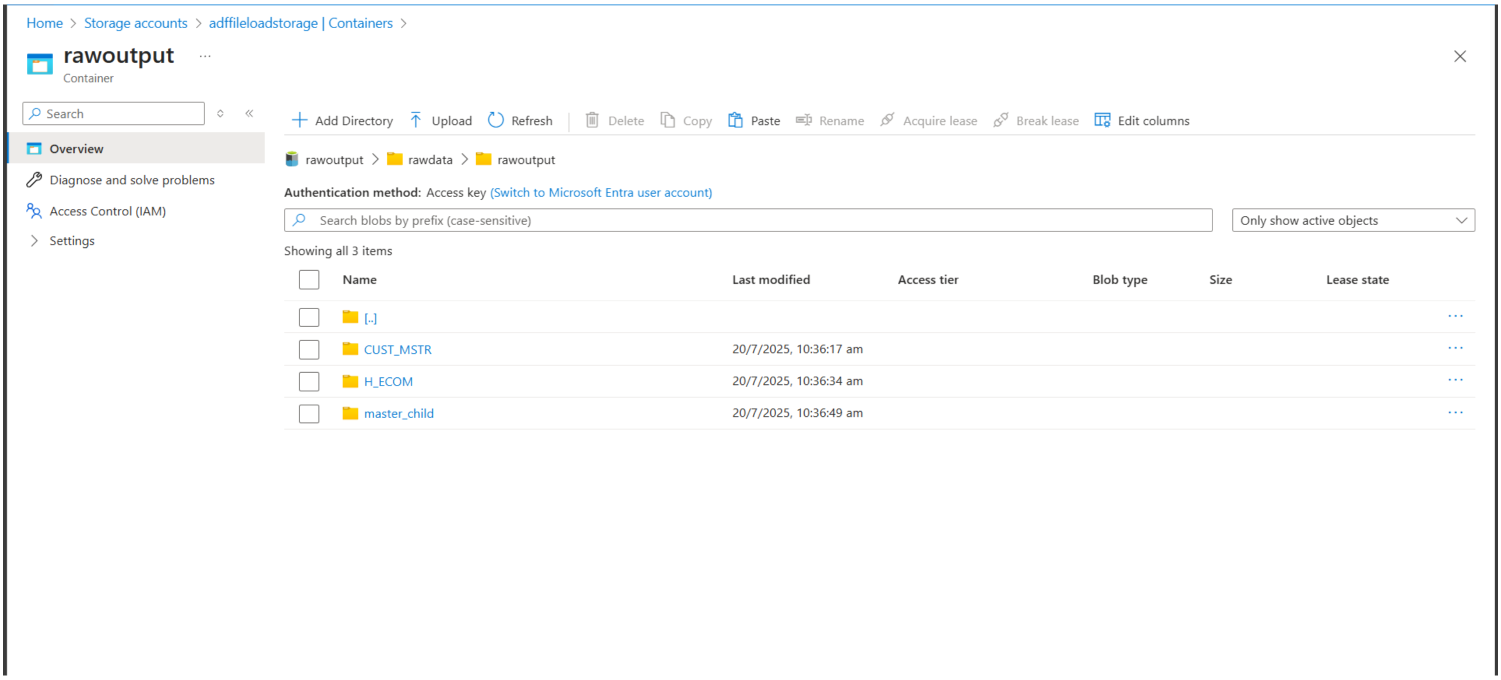
Authentication method: Access key (Switch to Microsoft Entra user account)

Showing all 0 items

□	Name	Last modified	Access tier	Blob type	Size	Lease state
□	[..]					

Add or remove favorites by pressing Ctrl+Shift+F

- New folders are automatically created for each file type inside rawoutput.



rawoutput > rawdata > rawoutput

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

Search blobs by prefix (case-sensitive)

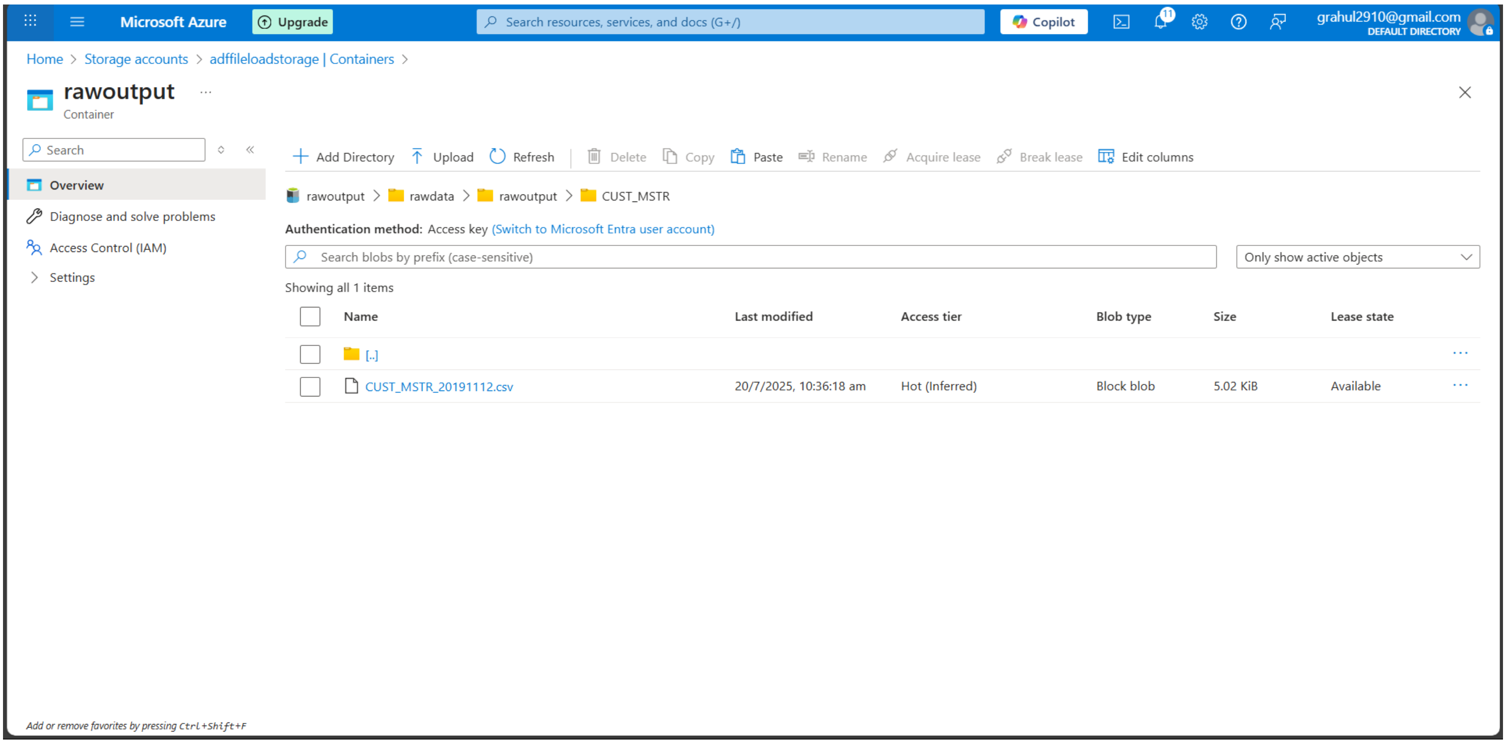
Showing all 3 items

□	Name	Last modified	Access tier	Blob type	Size	Lease state
□	[..]					
□	CUST_MSTR	20/7/2025, 10:36:17 am				
□	H_ECOM	20/7/2025, 10:36:34 am				
□	master_child	20/7/2025, 10:36:49 am				

- For example, if the file is master_child_export-20191112.csv, the file will be stored in:

rawdata/rawoutput/master_child/

- This folder naming is based on the first part of the file name (before the date or extension), such as:
  - CUST_MSTR → folder cust_mstr/
  - master_child_export → folder master_child/
  - H_ECOM_ORDER → folder h_ecom_order/



rawoutput > rawdata > rawoutput > CUST_MSTR

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

Search blobs by prefix (case-sensitive)

Showing all 1 items

□	Name	Last modified	Access tier	Blob type	Size	Lease state
□	[..]					
□	CUST_MSTR_20191112.csv	20/7/2025, 10:36:18 am	Hot (Inferred)	Block blob	5.02 KiB	Available

- CUST_MSTR Files
  These files will have one additional column:

  A Date column will be added based on the date extracted from the file name (in the format YYYY-MM-DD).

  All other columns will remain unchanged.

  There will be only one new column added in this case.

- master_child_export Files
  These files will have two new columns added:

  A Date column (format: YYYY-MM-DD)

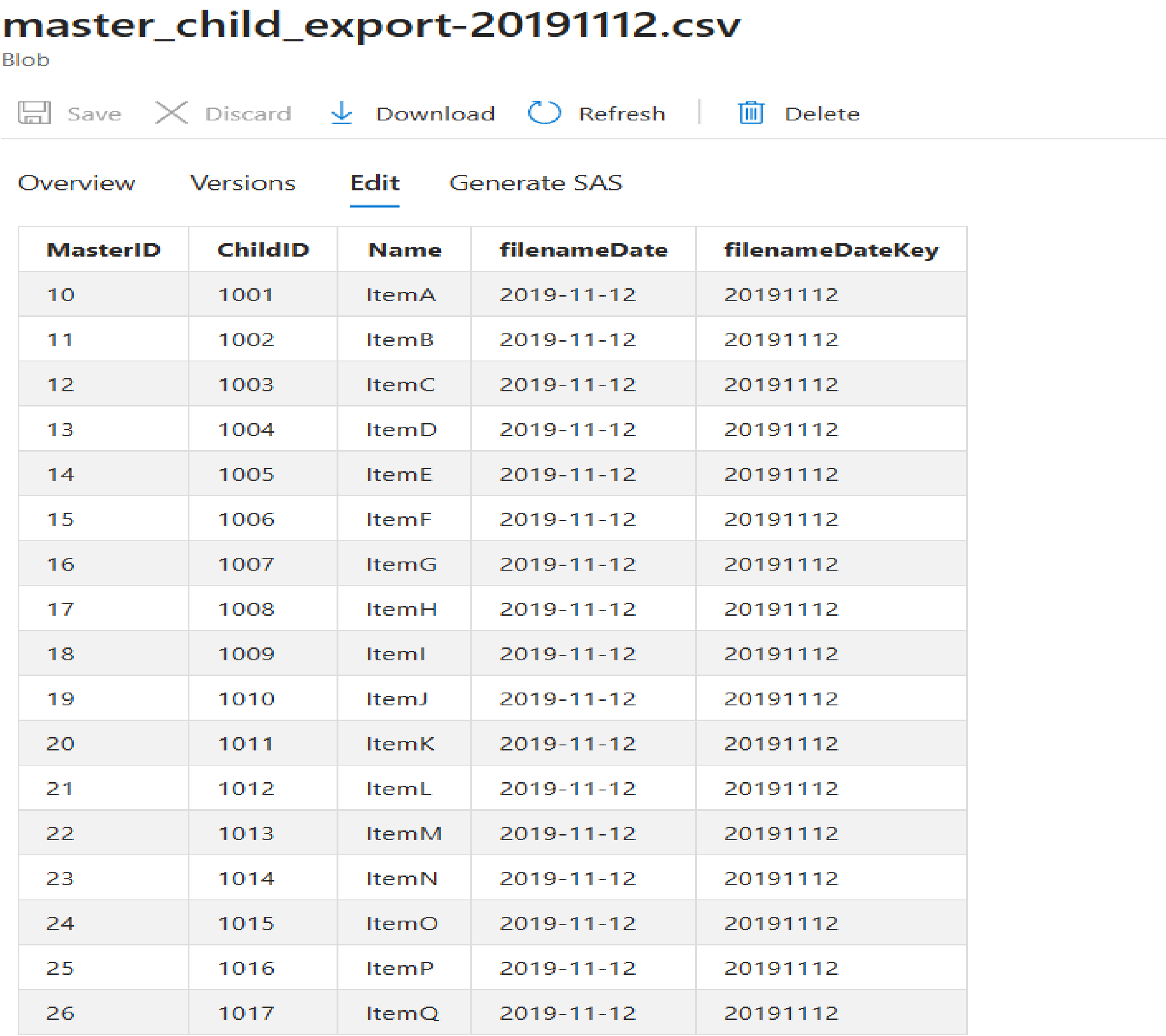
  A DateKey column (format: YYYYMMDD)

  These columns will be appended at the end of each row during transformation.

- H_ECOM_ORDER Files
  These files are loaded as-is, with no transformation or additional columns.

  The output will exactly match the structure of the input file.

  This is because these files do not contain any embedded date in the filename.



MasterID	ChildID	Name	filenameDate	filenameDateKey
10	1001	ItemA	2019-11-12	20191112
11	1002	ItemB	2019-11-12	20191112
12	1003	ItemC	2019-11-12	20191112
13	1004	ItemD	2019-11-12	20191112
14	1005	ItemE	2019-11-12	20191112
15	1006	ItemF	2019-11-12	20191112
16	1007	ItemG	2019-11-12	20191112
17	1008	ItemH	2019-11-12	20191112
18	1009	ItemI	2019-11-12	20191112
19	1010	ItemJ	2019-11-12	20191112
20	1011	ItemK	2019-11-12	20191112
21	1012	ItemL	2019-11-12	20191112
22	1013	ItemM	2019-11-12	20191112
23	1014	ItemN	2019-11-12	20191112
24	1015	ItemO	2019-11-12	20191112
25	1016	ItemP	2019-11-12	20191112
26	1017	ItemQ	2019-11-12	20191112