

Assignment

Utilize Azure Data Factory (ADF) to ingest Orders and Customers data, and execute fundamental transformations on the datasets.

Task -1: Ingest orders.csv file from external URL to ADLS Gen2

Dataset orders.csv can be downloaded from below link

https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv

Implementation:

1. I have created resource group.

The screenshot shows the Azure portal interface for 'Resource groups'. The title bar indicates 'Default Directory (tsrao999outlook.onmicrosoft.com)'. Below the title bar are buttons for '+ Create', 'Manage view', 'Refresh', 'Export to CSV', 'Open query', and 'Assign tags'. A search bar is present with the text 'Filter for any field...'. Below the search bar, there are filters: 'Subscription equals all' and 'Location equals all'. A table shows one record: 'Mega-sales-rg' with 'Free Trial' subscription and 'Central India' location.

Name	Subscription	Location
Mega-sales-rg	Free Trial	Central India

2. I have create a Storage Account (Enable Hierarchical namespace to make it as data lake storage and not just the blob storage)

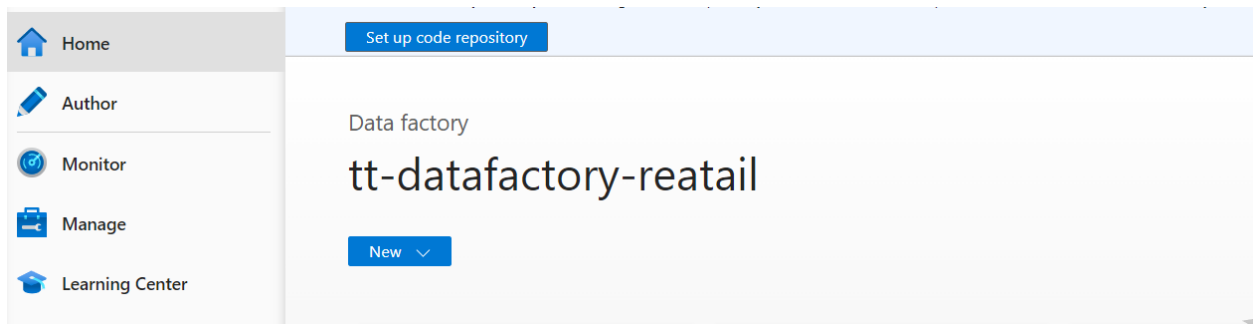
The screenshot shows the Azure portal interface for a Storage Account named 'reatailstorageaccount'. The title bar indicates 'Storage account'. Below the title bar are buttons for 'Upload', 'Open in Explorer', 'Delete', 'Move', 'Refresh', 'Open in mobile', 'CLI / PS', and 'Feedback'. A search bar is present with the text 'Search'. Below the search bar, there are tabs: 'Activity log', 'Tags', and 'Diagnose and solve problems'. The main content area shows 'Essentials' with 'Resource group (move)' and 'Performance' (Standard).

3. I have created container with directory inside storage account.

The screenshot shows the Azure portal interface for a Storage Container named 'data'. The title bar indicates 'Container'. Below the title bar are buttons for 'Upload', 'Add Directory', 'Refresh', 'Rename', 'Delete', 'Change tier', 'Acquire lease', 'Break lease', and 'Give feedback'. A search bar is present with the text 'Search'. Below the search bar, there are tabs: 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Settings', 'Shared access tokens', and 'Manage ACL'. The main content area shows 'Authentication method: Access key (Switch to Microsoft Entra user account)' and 'Location: data'. A search bar is present with the text 'Search blobs by prefix (case-sensitive)'. A table shows one record: 'input_data'.

Name	Modified	Access tier	Archive status	Blob type
input_data				

4. I have created a Resource - Azure Data Factory:



5. Create a Linked Service for Source (choose HTTP connector): orders.csv can be downloaded from below link.

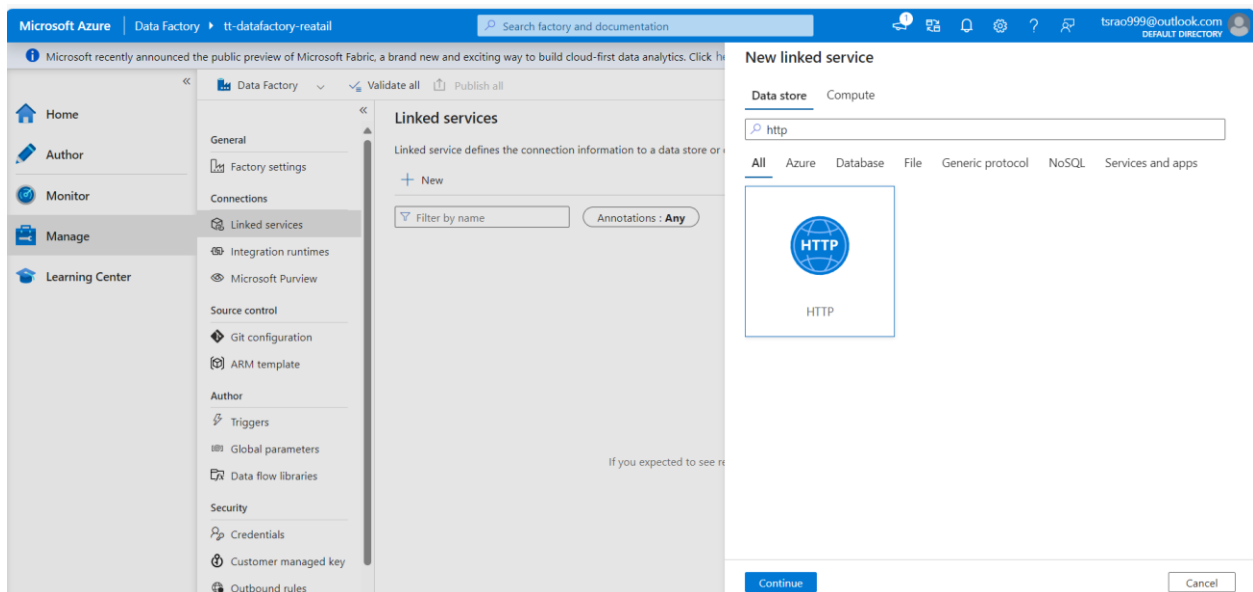
“https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv”

a) **Base URL** - <https://files.cdn.thinkific.com>

b) **Relative URL** - [file_uploads/349536/attachments/c28/5fb/25b/orders.csv](https://files.cdn.thinkific.com/file_uploads/349536/attachments/c28/5fb/25b/orders.csv)

I have launched the **Azure Data Factory** to create **the Linked service** and **datasets** and pipeline

Check below screenshot for creating linked service for our source i.e. **Azure Data lake Gen2** Go to **Monitor => Linked Service => New => search Http server** and select **Azure data Lake Gen2** => **Continue**



New linked service

HTTP [Learn more](#)

Name *
ls_orders_Http

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Base URL *
https://files.cdn.thinkific.com
 ⚠ Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server certificate validation
☒ Enable ☐ Disable

Authentication type *
Anonymous

Auth headers
 + New

Annotations

[Create](#) [Back](#) [Test connection](#) [Cancel](#)

I have create a Linked Service for Sink (choose Data Lake Gen 2 connector):

- Check below screenshot for creating linked service for our sink to ADLS gen storage

tt-datafactory-reatail Search factory and documentation tsrao999@outlook.com

the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here

Data Factory Validate all Publish all

Linked services

Linked service defines the connection information to a data store or...

+ New

Filter by name Annotations: Any

Showing 1 - 1 of 1 items

Name	Type
ls_orders_Http	HTTP

New linked service

Data store Compute

azure

All Azure Database File Generic protocol NoSQL Services and apps

Azure AI Search Azure Blob Storage Azure Cosmos DB for MongoDB

Azure Cosmos DB for NoSQL Azure Data Explorer (Kusto) **Azure Data Lake Storage Gen2**

Continue Cancel

New linked service

Azure Data Lake Storage Gen2 [Learn more](#)

Connect via integration runtime * ⓘ
 AutoResolveIntegrationRuntime

Authentication type
 Account key

Account selection method ⓘ
☒ From Azure subscription ☐ Enter manually

Azure subscription ⓘ
 Free Trial (f871fa9c-f2ac-451c-ab89-6945feafcf06)

Storage account name *
 reatailstorageaccount

Test connection ⓘ
☒ To linked service ☐ To file path

Annotations
 + New

> Parameters

Connection successful

Create Back Test connection Cancel

After creating new linked services in Azure Data Factory, be sure to **publish** these changes to make them active and available for use in your data workflows.

Validate all Publish all

Linked services

Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New

Filter by name Annotations : Any

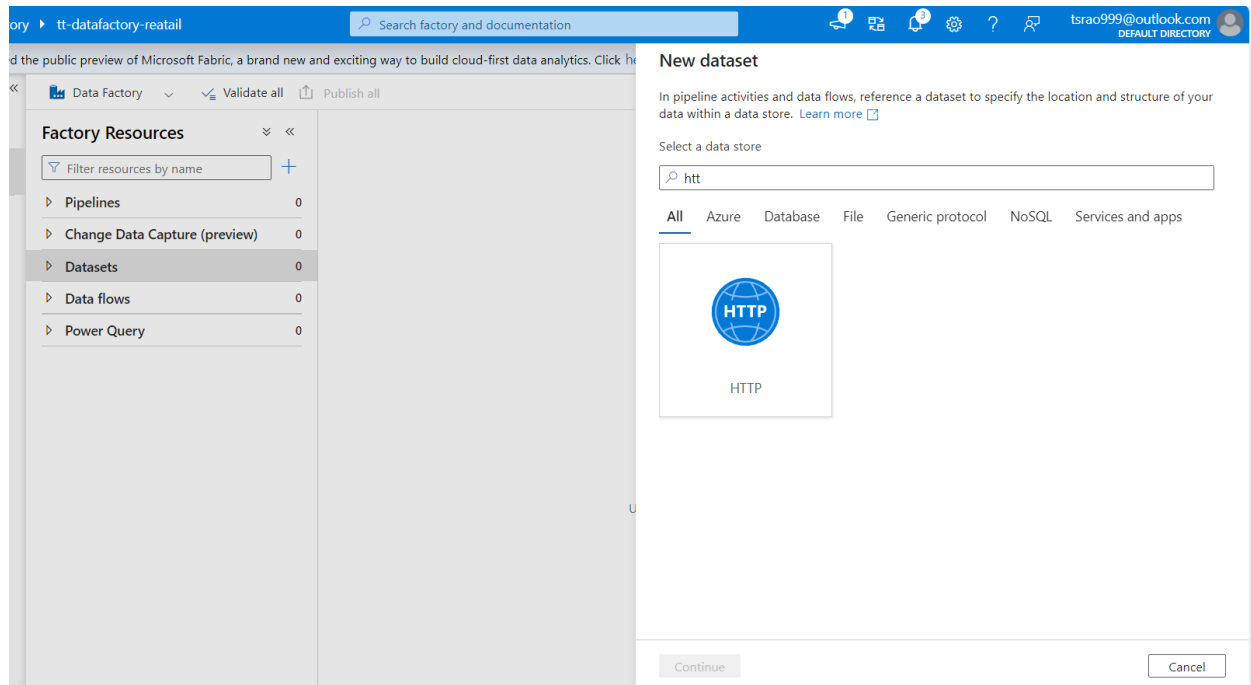
Showing 1 - 2 of 2 items

Name ↑↓	Type ↑↓	Related ↑↓	Annotations ↑↓
Is_orders_AdlsGen2	Azure Data Lake Storage Gen2	0	
Is_orders_Http	HTTP	0	

Publishing completed
 Successfully published

I have created **dataset for Source** (choose CSV format and also provide the relative URL as it is for the HTTP Linked Service)

- To create dataset click on Author => Datasets => New datasets



The screenshot shows the 'Set properties' dialog for a new dataset. The dialog has the following fields and options:

- Name:** ds_orders_http
- Linked service ***: ls_orders_Http (with a dropdown arrow and a pencil icon)
- Relative URL:** file_uploads/349536/attachments/c28/5fb/25b/orders.csv
- First row as header:** ☒
- Import schema:** ☒ From connection/store, ☐ From sample file, ☐ None
- Request method:** GET (with a dropdown arrow)
- Additional headers:** (empty text area)
- Request body:** (empty text area)

At the bottom, there are three buttons: 'OK', 'Back', and 'Cancel'.

I have created dataset for Sink:

the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here

Data Factory Validate all Publish all

Factory Resources

Filter resources by name

- Pipelines 0
- Change Data Capture (preview) 0
- Datasets 1**
 - ds_orders_http
- Data flows 0
- Power Query 0

ds_orders_http

DelimitedText
ds_orders_http

Connection Schema Parameters

file_uploads/349536/attachments/c28/5f...

Compression type Select...

Column delimiter Comma (,)

Row delimiter Default (\r\n)

Encoding Default(UTF-8)

Quote character Double quote

Escape character Backslash (\)

First row as header ☒

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

azure

All Azure Database File Generic protocol NoSQL Services and apps

- Azure AI Search
- Azure Blob Storage
- Azure Cosmos DB for MongoDB
- Azure Cosmos DB for NoSQL
- Azure Data Explorer (Kusto)
- Azure Data Lake Storage Gen2

Continue Cancel

Select format

Choose the format type of your data

- Avro
- Binary
- DelimitedText**
- Excel
- JSON
- ORC
- Parquet
- XML

Continue Back Cancel

Set properties

Name

ds_orders_adlagen2

Linked service *

ls_orders_AdlsGen2

File path

data

/ input_data

/ File name

First row as header



Import schema

☒ From connection/store ☐ From sample file ☐ None

OK

Back

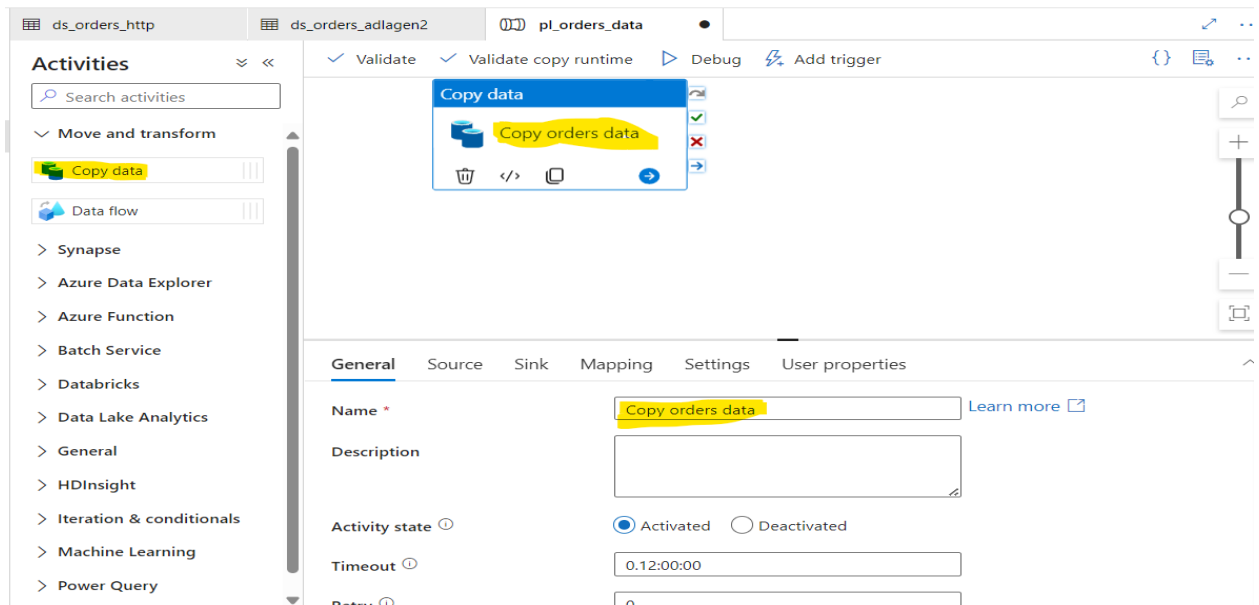
Cancel

After creating new datasets in Azure Data Factory, be sure to **publish** these changes to make them active and available for use in your data workflows.

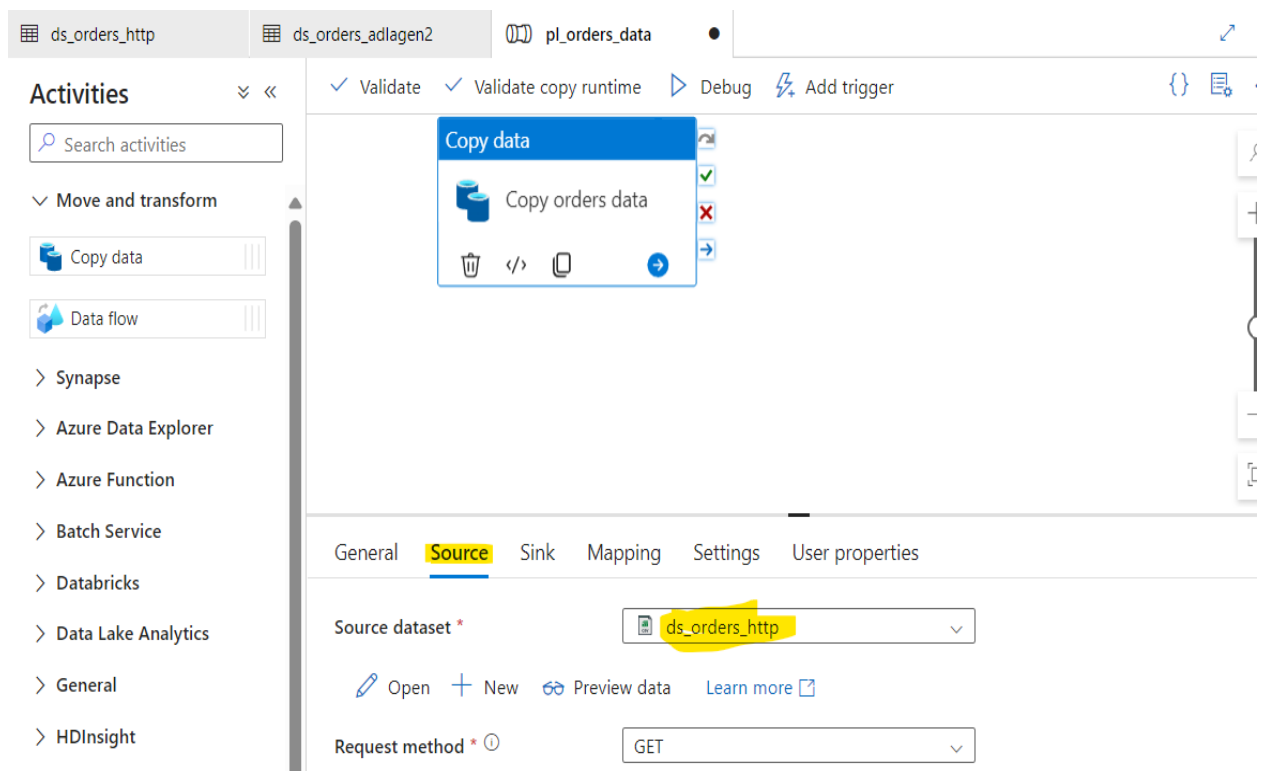
I have created a data Pipeline and Create a Copy activity within the Pipeline:

Now click on “Move and transform” and drag copy activity in the pipeline as shown below. Rename the pipeline copy data activity as shown in screenshot.

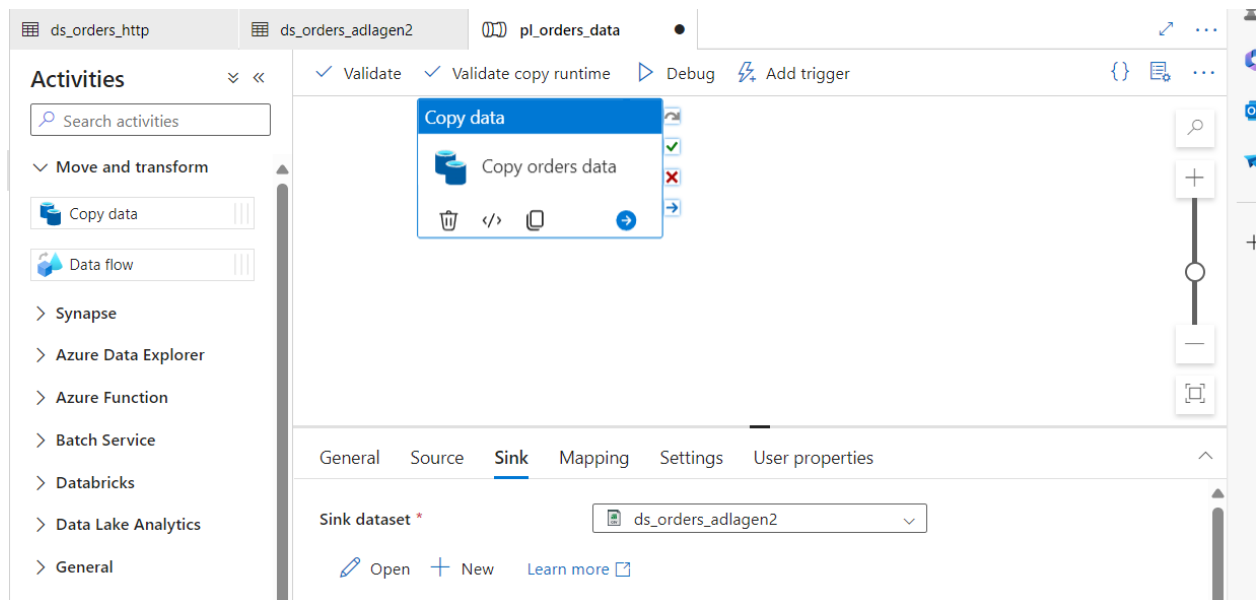
The screenshot displays the Azure Data Factory (ADF) interface. On the left, the 'Factory Resources' pane shows a tree view with 'Pipelines' expanded, containing 'pl_orders_data'. The 'Activities' pane on the right lists various activities, including 'Move and transform', 'Synapse', 'Azure Data Explorer', 'Azure Function', 'Batch Service', 'Databricks', 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', 'Machine Learning', and 'Power Query'. The main canvas shows a pipeline diagram with a single activity named 'pl_orders_data'. The 'Properties' pane on the right is open, showing the 'General' tab with the 'Name' field set to 'pl_orders_data' and a 'Description' field. The 'Parameters' tab is also visible, showing a '+ New' button.



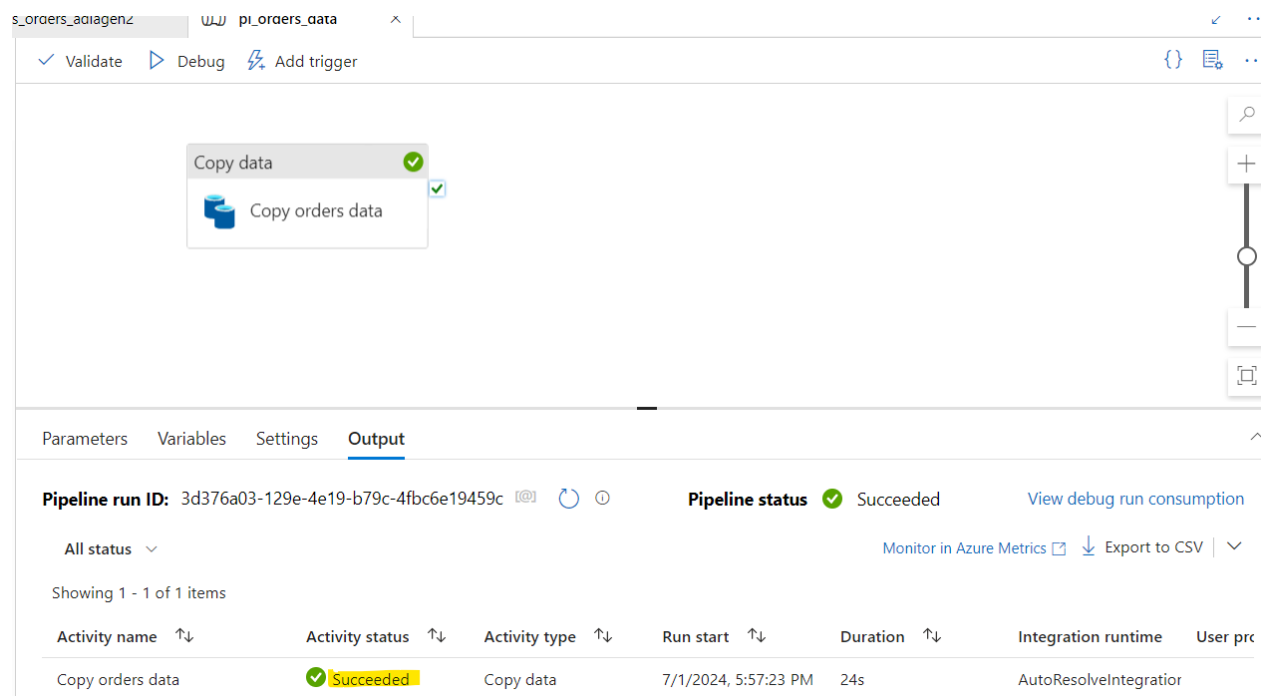
Now in **source** select dataset here “ds_orders_http”



Now in **sink** select dataset for sink here “ds_orders_adlsngen2”



Debug and validate the pipeline, and upon successful validation, proceed to publish the pipeline:



And if this is successful then publish the pipeline.

After creating new pipelines in Azure Data Factory, be sure to publish these changes to make them active and available for use in your data workflows.

Home > All resources > **retailstorageaccount** | Containers >

data ...
Container

Search [] x << Upload + Add Directory Refresh | Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

- Diagnose and solve problems
- Access Control (IAM)
- Settings
 - Shared access tokens
 - Manage ACL
 - Access policy
 - Properties

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

Location: data / input_data

Search blobs by prefix (case-sensitive) ☐ Show deleted objects

	Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/>	ordersdata.csv	7/1/2024, 5:57:46 PM	Hot (Inferred)		Block blob

Task -2: Ingest customers.csv file from ADLS Gen2 to Azure SQL Database.

1. I have uploaded customer's data in ADLS Gen2 storage account.

Home > All resources > **retailstorageaccount** | Containers >

customersdata ...
Container

Search [] x << Upload + Add Directory Refresh | Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

- Diagnose and solve problems
- Access Control (IAM)
- Settings
 - Shared access tokens
 - Manage ACL
 - Access policy
 - Properties
 - Metadata

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

Location: customersdata / customers

Search blobs by prefix (case-sensitive) ☐ Show deleted objects

	Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/>	[.]					
<input type="checkbox"/>	customers_final.csv	7/1/2024, 6:11:54 PM	Hot (Inferred)		Block blob	93

I have Create a Linked Service for Source (Adls gen2):

New linked service
Azure Data Lake Storage Gen2 [Learn more](#)

Name *
ls_customers_adlsqen2

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Authentication type
Account key

Account selection method
☒ From Azure subscription ☐ Enter manually

Azure subscription
Free Trial (f871fa9c-f2ac-451c-ab89-6945feafcf06)

Storage account name *
reatailstorageaccount

Test connection
☒ To linked service ☐ To file path

Create **Back** **Test connection** **Cancel**

Connection successful

I have created a SQL Database for sink system:

I have Create a Linked Service for Sink (Azure SQL Database):

New linked service

Data store **Compute**

sql

All **Azure** **Database** **File** **Generic protocol** **NoSQL** **Services and apps**

Amazon RDS for SQL Server

Azure Cosmos DB for NoSQL

Azure Database for MySQL



Azure Database for PostgreSQL

Azure SQL Database

Azure SQL Database Managed Instance

Continue **Cancel**

Edit linked service

 Azure SQL Database [Learn more](#) 

Name *

ls_customers_sqldata

Description

Connect via integration runtime * 

AutoResolveIntegrationRuntime

Version

☒ Recommended ☐ Legacy

Account selection method 

☐ From Azure subscription ☒ Enter manually

Fully qualified domain name *

customer360x.database.windows.net

Database name *


customers-server

Authentication type *

SQL authentication




Apply

Cancel


 Test connection

To create dataset click on Author => Datasets => New datasets for source system:


Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to learn more.

Data Factory   Validate all  Publish all

Factory Resources


Filter resources by name 



- Pipelines 1
- Change Data Capture (preview) 0
- Datasets 2**
 - ds_orders_adlagen2
 - ds_orders_http
- Data flows 0
- Power Query 0

Use the resource 

Set properties

Name

Linked service * 

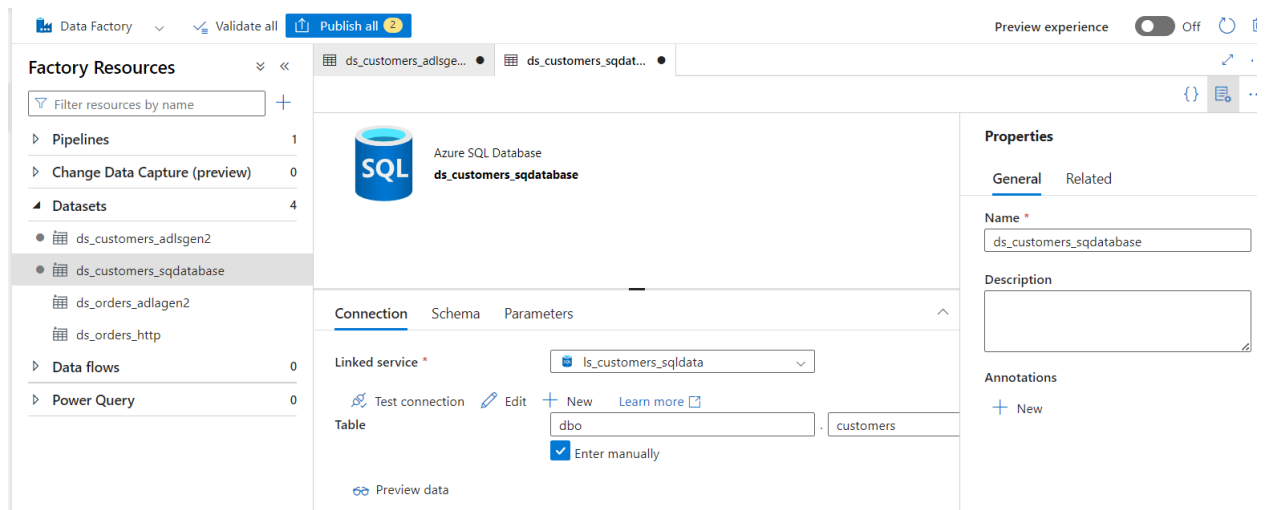
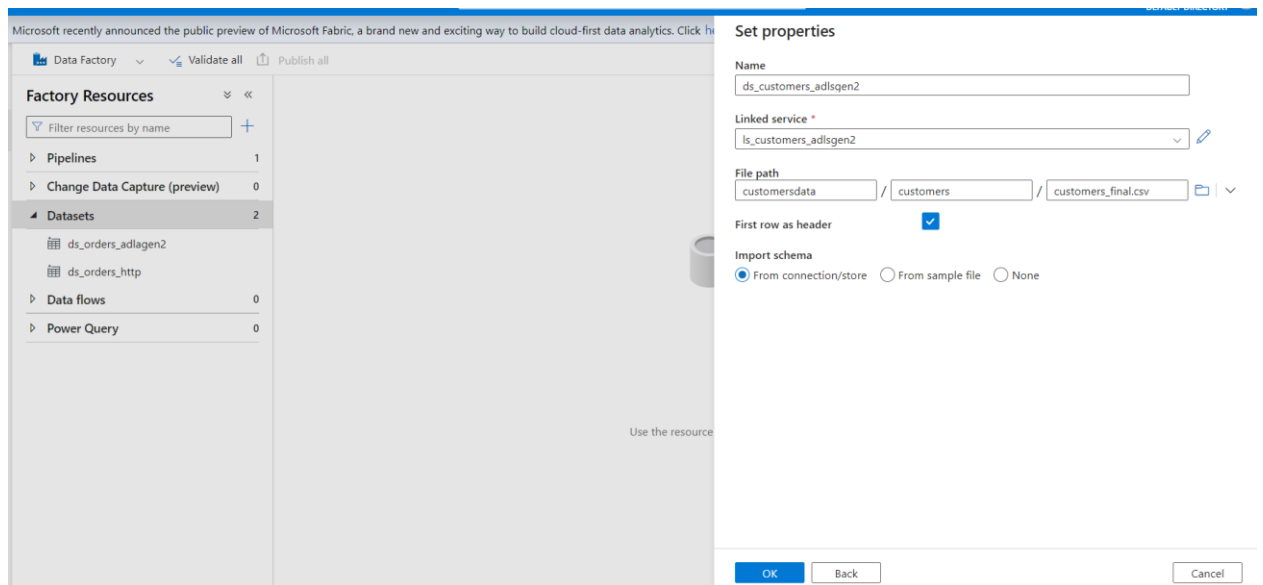
File path / /  

First row as header ☒

Import schema ☒ From connection/store ☐ From sample file ☐ None

OK Back Cancel

Create Dataset for Sink (SQL database) :



After that publish all the datasets in azure data factory.

I have created a data Pipeline and Create a Copy activity within the Pipeline:

Now click on “Move and transform” and drag copy activity in the pipeline as shown below. Rename the pipeline copy data activity as shown in screenshot.

The screenshot displays the Azure Data Factory (ADF) interface. On the left, the 'Factory Resources' pane shows a list of pipelines, including 'pl_customers_data'. The main canvas shows a pipeline with a single activity named 'Copy data'. The 'Properties' pane on the right shows the 'General' tab with the activity name 'pl_customers_data'. Below the canvas, the 'Output' tab shows the pipeline run ID '3e9b9079-1c76-43a4-ac1a-a6b5d48ef41f' and the pipeline status 'Succeeded'. A table below shows the activity status for 'Copicustomersdata', which is 'Succeeded'.

Activity name	Activity status	Activity type	Run start
Copicustomersdata	Succeeded	Copy data	7/1/2024, 1

After this activity checking customers table in SQL database and I have attached below screenshot

The screenshot displays the Azure Data Studio interface. The left pane shows the 'customers-server (tsrao999)' database with a table explorer showing the 'dbo.customers' table. The main editor shows a SQL query: 'select * from [dbo].[customers];'. The 'Results' pane shows the query output as a table with 5 columns: 'customer_id', 'customer_fname', 'customer_lname', 'customer_email', and 'customer_zipcode'. The results show 3 rows of data.

customer_id	customer_fname	customer_lname	customer_email	customer_zipcode
1	Richard	Hernandez	XXXXXXXXXX	XXXXXXXXXX
2	Mary	Barrett	XXXXXXXXXX	XXXXXXXXXX
3	Ann	Smith	XXXXXXXXXX	XXXXXXXXXX

Mapping the datasets using dataflow in 2 sources:

1) Azure SQL database

The screenshot shows the 'Source settings' tab for a dataset source named 'sourcecustomers'. The source is connected to 'ds_customers_sqdatabase'. It has 9 columns in total. The 'Source type' is set to 'Dataset'. The 'Options' section includes 'Allow schema drift' (checked) and 'Infer drifted column types' (unchecked). The 'Data flow debug' toggle is turned on.

Source settings | Source options | Projection | Optimize | Inspect | Data preview

Source type * | Dataset | Inline

Dataset * | ds_customers_sqdatabase | Connection | Test connection

Options | ☒ Allow schema drift ⓘ | ☐ Infer drifted column types ⓘ

2) ADLS Gen2 storage Account

The screenshot shows the 'Source settings' tab for a dataset source named 'sourceorders'. The source is connected to 'ds_orders_adlagen2'. It has 4 columns in total. The 'Source type' is set to 'Dataset'. The 'Options' section includes 'Allow schema drift' (checked) and 'Infer drifted column types' (unchecked). The 'Data flow debug' toggle is turned on.

Source settings | Source options | Projection | Optimize | Inspect | Data preview

Source type * | Dataset | Inline

Dataset * | ds_orders_adlagen2 | Connection | Test connection

Options | ☒ Allow schema drift ⓘ | ☐ Infer drifted column types ⓘ

I have mapped two sources to do some transformations in dataflow below.

Join Transformation: I am using inner join the two sources based on common field from customer source in **customer_id** and orders source in **order_customer_id** using to create one join transformation.

Reference: 1
Columns: 9 total

joinreataildata
Columns: 13 total

selectreataildata
Renaming joinreataildata to selectreataildata with columns 'customer_id, customer_fname, customer_lname'

filtercity
Filtering rows: expressions c 'customer_cit'

sourceorders
Import data from ds_orders_sqdatabase?

Join settings Optimize Inspect Data preview ●

Left stream *
sourcecustomers

Right stream *
sourceorders

Join type *
Full outer Inner Left outer Right outer Custom (cross)

Use fuzzy matching ⓘ
☐

Join conditions *
Left: sourcecustomers's column Right: sourceorders's column
abc customer_id == abc order_customer_id + -

Select Transformation: incoming stream data (joinreataildata) using to select transform activity through remove the 3 fields in selectreataildata.

sourcecustomers
Import data from ds_customers_sqdatabase

joinreataildata
Inner join on 'sourcecustomers' and 'sourceorders'

selectreataildata
Columns: 10 total

Select settings Optimize Inspect Data preview ●

Output stream name *
selectreataildata [Learn more](#) ⓘ

Description
Renaming joinreataildata to selectreataildata with columns 'customer_id, customer_fname, customer_lname' [Reset](#)

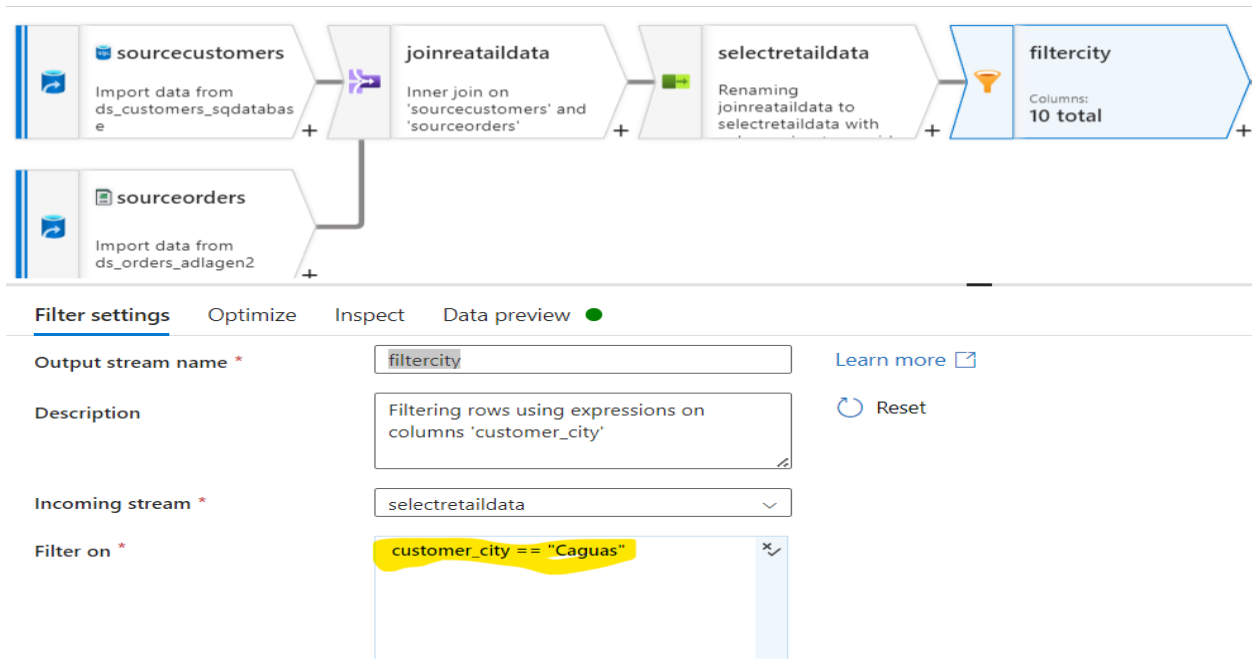
Incoming stream *
joinreataildata

Options
☒ Skip duplicate input columns ⓘ
☒ Skip duplicate output columns ⓘ

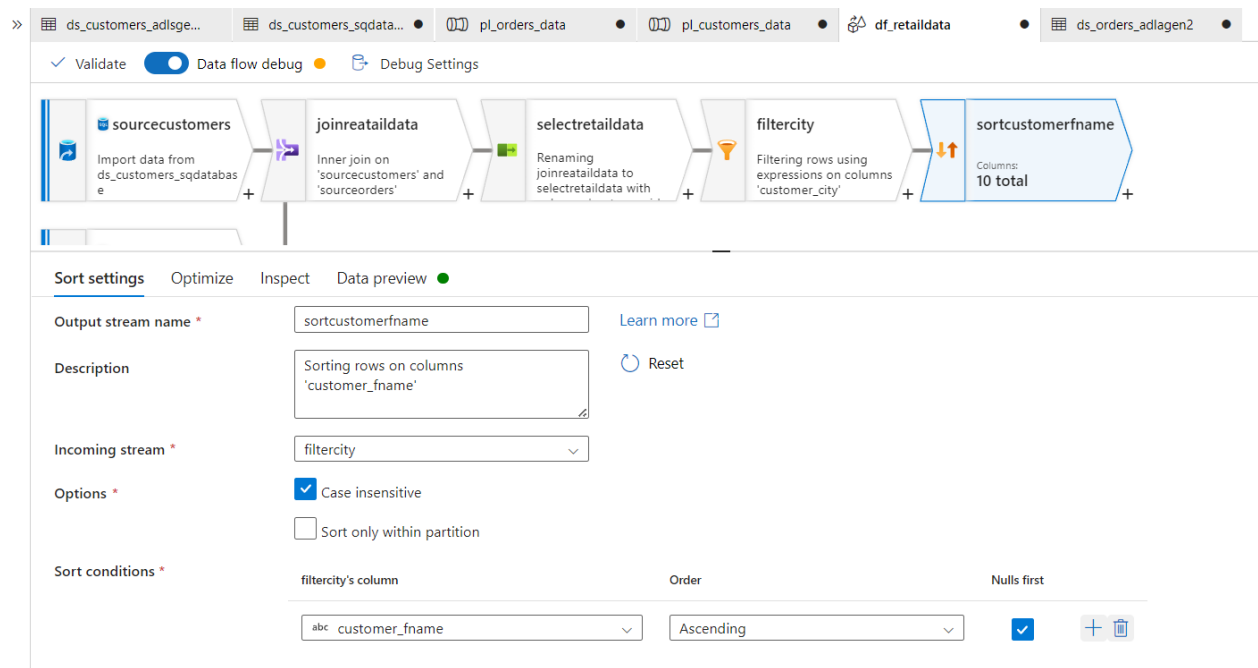
Input columns *
☐ joinreataildata's column [Reset](#) + Add mapping - Delete

Name a

Filter Transformation: incoming stream data (selectretaildata) using to filter on city is Caguas to transformed output stream (filtercity) and I have attached screen shot below.



Sort Transformation: I have sorted the ascending order the customer fname based on filtercity incoming stream and I have attached screenshot to check it once.



Aggregate Transformation: I am using aggregate transform to found total number of customer id's based on order status.

Now select the group by the "order status" column and Click on "Aggregates" mention order_id column and click on "Open expression builder"

Now mention count (order_id) in "Expression" refer attached screenshot and click "save and finish"

Aggregate settings | Optimize | Inspect | Data preview ●

Output stream name: aggregateetaildata

Description: Aggregating data by 'order_status' producing columns 'customer_id'

Incoming stream: sortcustomerfname

Group by: **Aggregates**

Grouped by: order_status

+ Add | Clone | Delete | Open expression builder

Column	Expression
customer_id	count(customer_id)

After this click on "Data Preview" and refresh it to see all the change

Aggregate settings Optimize Inspect **Data preview**

Refresh

Typecast

Modify

Map d

<div><div></div><div></div></div>	order_status	abc	<div><div></div><div></div></div>	customer_id
<div><div></div><div></div></div>	ON_HOLD			1
<div><div></div><div></div></div>	PENDING			4
<div><div></div><div></div></div>	PROCESSING			5
<div><div></div><div></div></div>	COMPLETE			12
<div><div></div><div></div></div>	PENDING_PAYMENT			6
<div><div></div><div></div></div>	SUSPECTED_FRAUD			1
<div><div></div><div></div></div>	PAYMENT_REVIEW			1
<div><div></div><div></div></div>	CLOSED			3

We have to Add the sink but before proceeding ahead we will create the “result” directory in the “data” container in ADLS gen2 storage “**reatailstorageaccount**” that we have created.

Home > All resources > reatailstorageaccount | Containers >

customersdata
Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

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

Location: customersdata

Search blobs by prefix (case-sensitive) Show deleted objects

Name	Modified	Access tier	Archive status	Blob type
<input type="checkbox"/> customers				
<input type="checkbox"/> results				

Also create dataset “ds_retail_json_output” for storing “output_data.json” data in this “result” directory

Avro

Binary

DelimitedText

Excel

JSON

ORC

Parquet

XML

Continue Back Cancel



JSON
ds_retail_json_output

Connection Schema Parameters

ls_customers_adlsgen2

Test connection

Edit

+ New

Learn more

customersdata

/ results

/ output_data.json

Browse

Preview data

None

Default(UTF-8)

For “df_retaildata” click on “Optimize” option and select “Single partitioning” to store the complete output in a single file.

aggregateretailda...
Aggregating data by 'order_status' producing columns

sinkretaildata
Columns: 2 total

Sink Settings Errors Mapping **Optimize** Inspect Data preview

This sink currently has Single partition set in Optimize. This will make your data flow execution longer. The recommended setting is Use current partitioning.

Partition option *
☐ Use current partitioning ☒ Single partition ☐ Set partitioning

Properties
General Related
Name *
df_retaildata
Description

Note: Before proceeding ahead. Be sure to publish these changes to make them active and available for use in my data workflows.

I have created the new pipeline “pl_retail_output” and will drag this dataflow in the pipeline “pl_retail_output”.

Data Factory | Validate all | Publish all | Preview experience | Off

Factory Resources

- Pipelines (3)
 - pl_retail_output
 - pl_customers_data
 - pl_orders_data
- Change Data Capture (preview) (0)
- Datasets (5)
 - ds_customers_adlsgen2
 - ds_customers_sqdatabase
 - ds_orders_adlagen2
 - ds_orders_http
 - ds_retail_json_output
- Data flows (1)
 - df_retaildata

pl_retail_output | Validate | Debug | Add trigger | Data flow debug

Data flow

df_retaildata

Properties

General | Related

Name * pl_retail_output

Description

Annotations

+ New

df_retaildata

General | Settings | Parameters | User properties

Name * df_retaildata

Description

Debug the pipeline and validate it. i can see this running pipeline in the monitor tab (Debug) and publish it.

All pipeline runs > pl_retail_output - Activity runs > df_retaildata

✓ df_retaildata

Cluster startup time: 4s 113ms Number of transformations: 8 Data flow status: Success

Refresh Auto refresh On Edit dataflow

Sinks All streams

Sink	Status	Processing time	Highest processing ti	Rows written	Stages	Lineage
sinkretaildata	✓ Succeeded	3s 830ms	911ms	9		

Pipeline runs

Triggered | Debug | Rerun | Cancel options | Refresh | Edit columns | List | Gantt

Filter by run ID or name Chennai, Kolkata, Mu... Last 24 hours Pipeline name : All

Status : All Add filter

Showing 1 - 8 items Last refreshed 0 minutes ago

Pipeline name	Run start	Run end	Duration	Status	Triggered by	Run ID
pl_retail_output	7/2/2024, 1:53:56 AM	7/2/2024, 1:55:07 AM	1m 11s	✓ Succeeded	Manual trigger	e33f20f1-ac26-4c8d-a
pipeline1customers	7/1/2024, 11:02:41 PM	7/1/2024, 11:02:56 PM	15s	✓ Succeeded	Manual trigger	3e9b9079-1c76-43a4-

I have opened my storage account and see the output data in results container.

Home > All resources > retailstorageaccount | Containers >

customersdata

Container

Search

UploadAdd DirectoryRefreshRenameDeleteChange tierAcquire leaseBreak leaseGive feedback

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

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

Location: customersdata / results

Search blobs by prefix (case-sensitive)

Show deleted objects

	Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/>	[.]					
<input type="checkbox"/>	part-00000-1fb14e15-2ed1-4eed-ac18-de370cdaa54...	7/2/2024, 1:54:14 AM	Hot (Inferred)		Block blob	43i