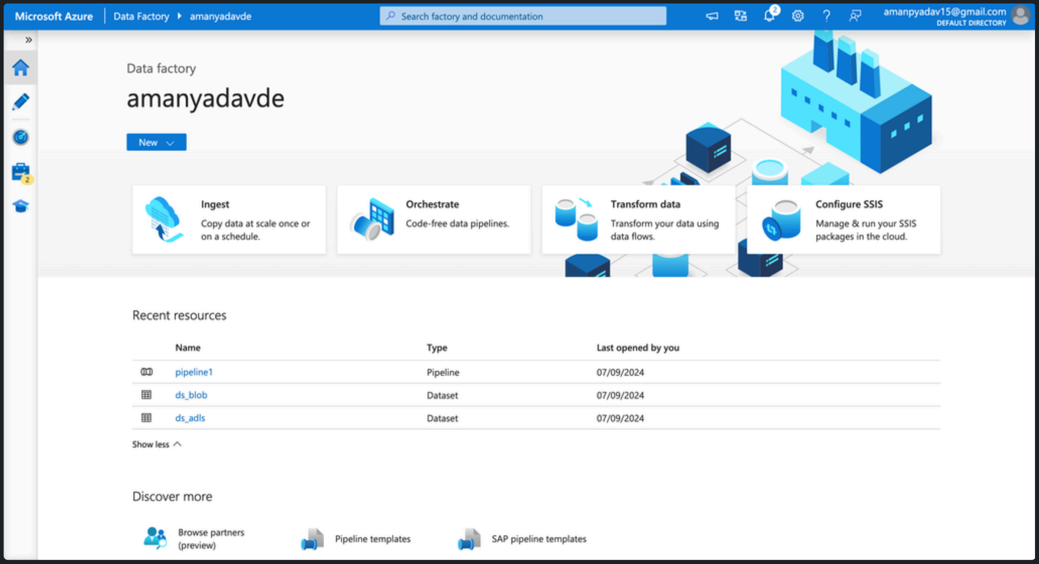
Creation of a data Pipeline

Setting Up the Data Factory & Creating Storage Accounts

1. Create and Setup Azure Data Factory.


Microsoft Azure | Data Factory ▸ amanyadavde
Search factory and documentation
amanpyadav15@gmail.com
DEFAULT DIRECTORY
Data factory
amanyadavde
New
Ingest
Copy data at scale once or on a schedule.
Orchestrate
Code-free data pipelines.
Transform data
Transform your data using data flows.
Configure SSIS
Manage & run your SSIS packages in the cloud.
Recent resources
Show less
Discover more
Browse partners (preview)
Pipeline templates
SAP pipeline templates

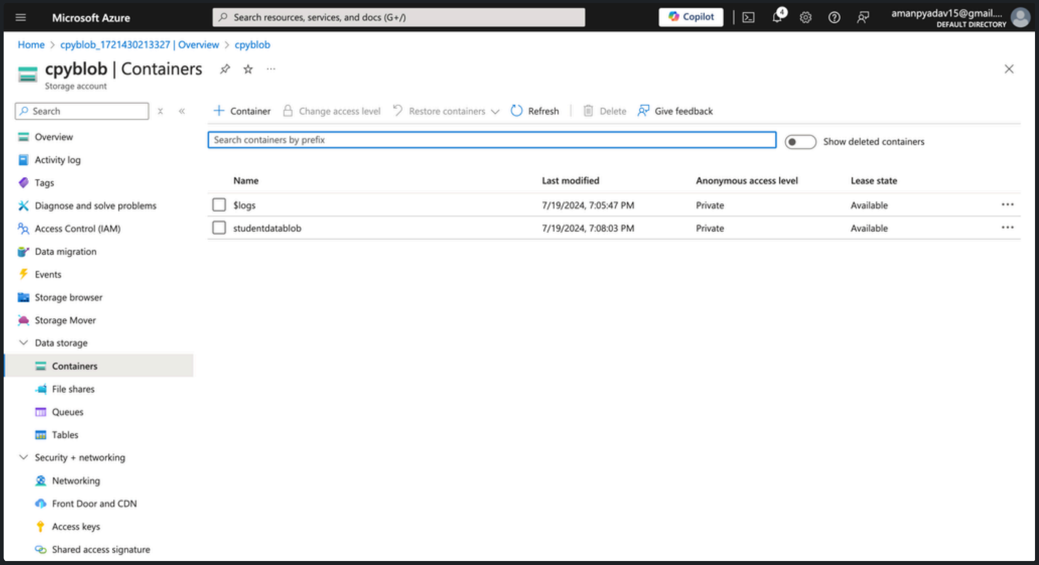
Name	Type	Last opened by you
pipeline1	Pipeline	07/09/2024
ds_blob	Dataset	07/09/2024
ds_adls	Dataset	07/09/2024

Azure Data Factory

2. Create Blob Storage Account.

Blob Storage

- We'll create first blob storage account in the name **cpyblob**. Inside blob storage account we'll create a container (studentdatablob) and upload our student dataset as a .csv format.

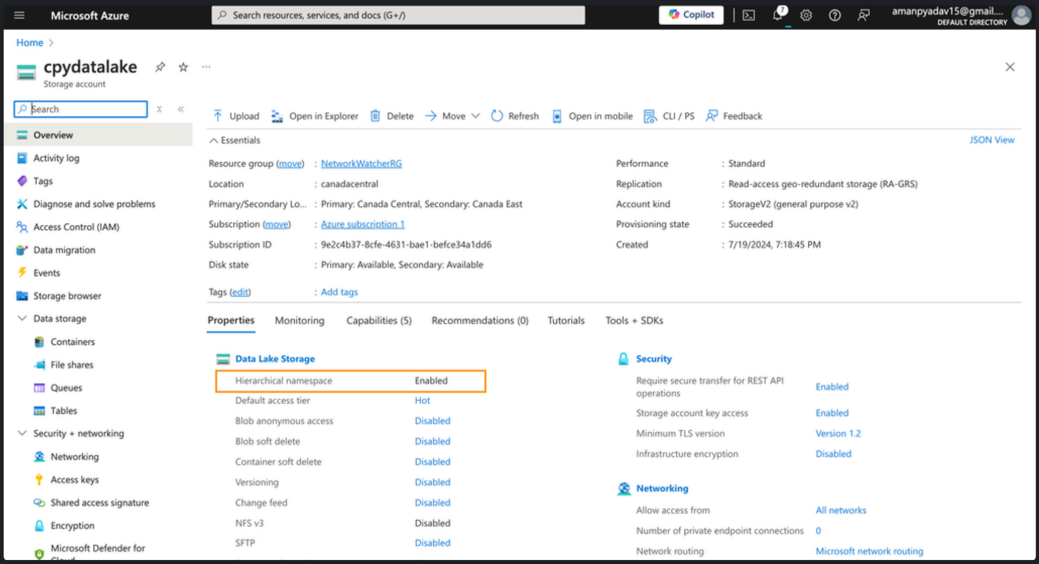


Name	Last modified	Anonymous access level	Lease state
$logs	7/19/2024, 7:05:47 PM	Private	Available
studentdatablob	7/19/2024, 7:08:03 PM	Private	Available

Blob Storage Account

Azure Data Lake Storage Account

- We'll follow the same step as we perform in creating a blob storage account but this time we'll just **enable Hierarchical namespaces**



cpydatalake
Storage account

Essentials

Resource group (move) : NetworkWatcherRG
Location : canadacentral
Primary/Secondary Lo... : Primary: Canada Central, Secondary: Canada East
Subscription (move) : Azure subscription 1
Subscription ID : 9e2c4b37-8cfe-4631-bae1-befce34a1dd6
Disk state : Primary: Available, Secondary: Available
Tags (edit) : Add tags

Performance : Standard
Replication : Read-access geo-redundant storage (RA-GRS)
Account kind : StorageV2 (general purpose v2)
Provisioning state : Succeeded
Created : 7/19/2024, 7:18:45 PM

Data Lake Storage

Hierarchical namespace	Enabled
Default access tier	Hot
Blob anonymous access	Disabled
Blob soft delete	Disabled
Container soft delete	Disabled
Versioning	Disabled
Change feed	Disabled
NFS v3	Disabled
SFTP	Disabled

Security

Require secure transfer for REST API operations	Enabled
Storage account key access	Enabled
Minimum TLS version	Version 1.2
Infrastructure encryption	Disabled

Networking

Allow access from	All networks
Number of private endpoint connections	0
Network routing	Microsoft network routing

Azure Storage Account

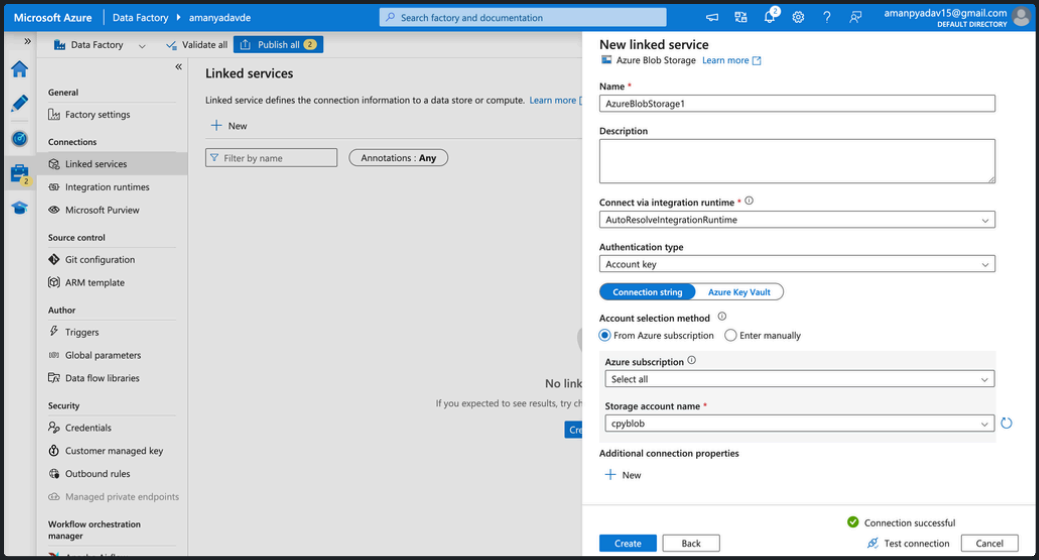
Pipeline 1

Overview

We have a student dataset and we'll be performing dataflow from Blob Storage to Azure Data Lake and perform some operations.

Creating Linked Services

Step 1. After launching the Azure Data Factory. We'll navigate to Manage and create a Blob linked service & Data Lake Linked Services. While creating that we'll make sure to test connection before creating it to make sure it's been linked properly with our blob storage.



New linked service
Azure Blob Storage Learn more

Name *
AzureBlobStorage1

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Authentication type
Account key

Connection string | Azure Key Vault

Account selection method
● From Azure subscription ○ Enter manually

Azure subscription
Select all

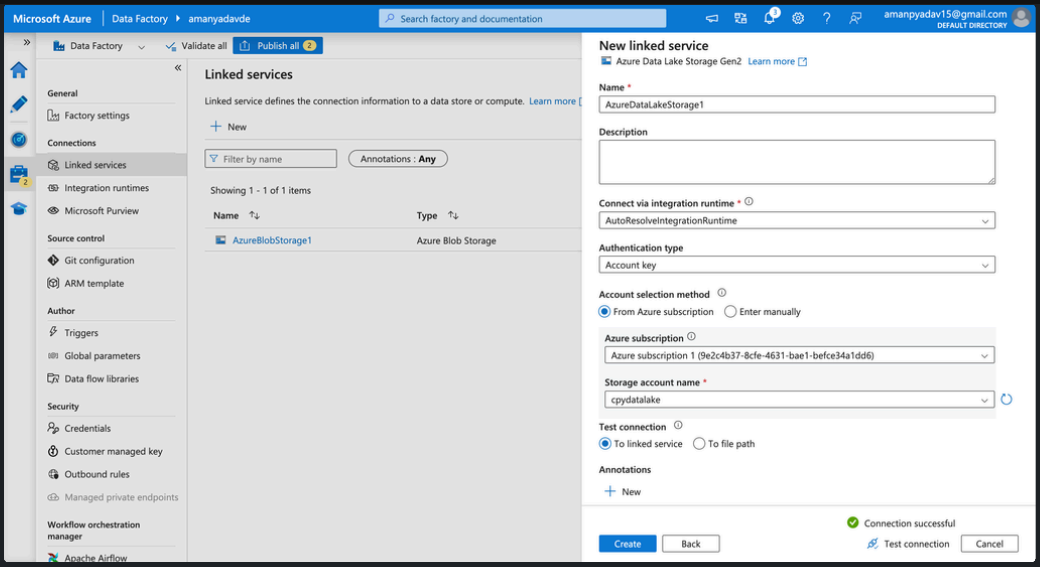
Storage account name *
cpyblob

Additional connection properties
+ New

✓ Connection successful

Create | Back | Test connection | Cancel

Blob Linked Service



Linked services

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

+ New

Filter by name | Annotations : Any

Showing 1 - 1 of 1 items

Name	Type
AzureBlobStorage1	Azure Blob Storage

New linked service

Azure Data Lake Storage Gen2 Learn more

Name *: AzureDataLakeStorage1

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication type: Account key

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription: Azure subscription 1 (9e2c4b37-8cfe-4631-bae1-befce34a1dd6)

Storage account name *: cpydatalake

Test connection: ● To linked service ○ To file path

Annotations

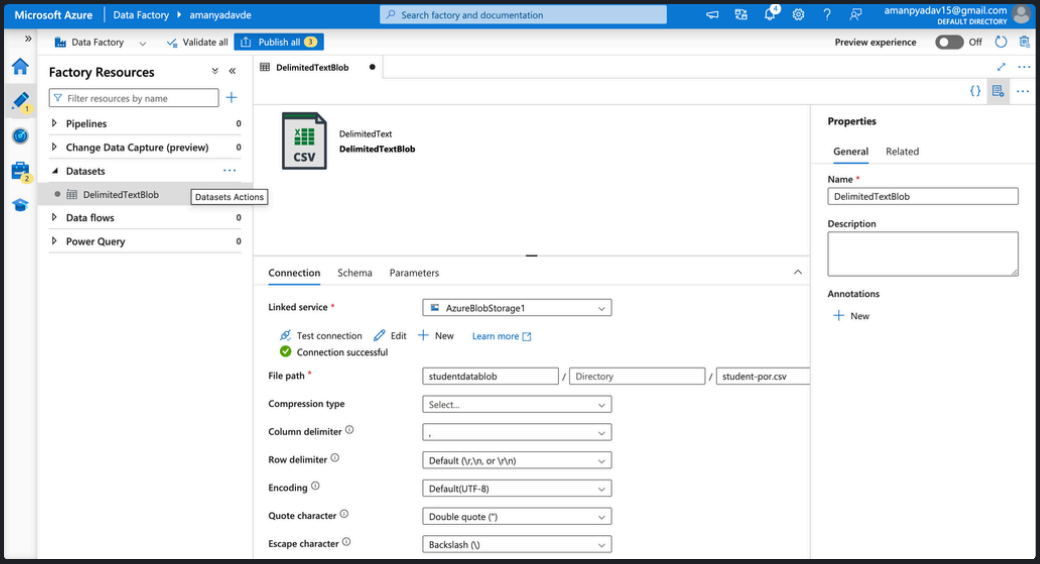
+ New

Connection successful

Data Lake Linked Service

Creating Dataset

Step 2. We'll create Dataset for both Blob Storage and Azure Data Lake Storage. To Create Dataset navigate to Author → Datasets → Create New Datasets.



Factory Resources

Pipelines	0
Change Data Capture (preview)	0
Datasets	
DelimitedTextBlob	
Data flows	0
Power Query	0

DelimitedTextBlob

DelimitedText
DelimitedTextBlob

Connection | Schema | Parameters

Linked service *: AzureBlobStorage1

Test connection | Edit | + New | Learn more

Connection successful

File path *: studentdatablob / Directory / student-por.csv

Compression type: Select...

Column delimiter: ,

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

Encoding: Default(UTF-8)

Quote character: Double quote (")

Escape character: Backslash (\)

Properties

General | Related

Name *: DelimitedTextBlob

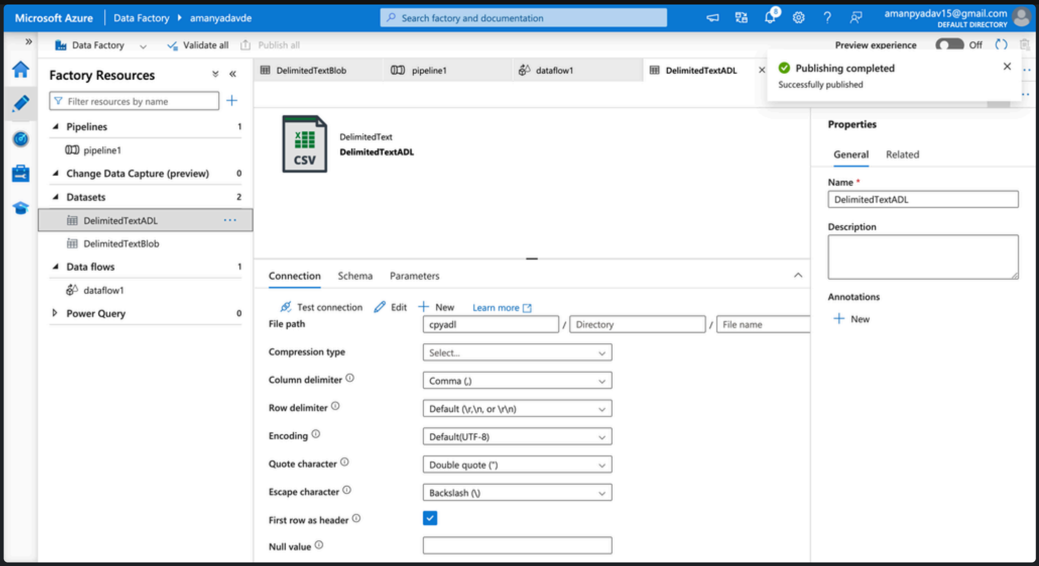
Description:

Annotations

+ New

Blob Datasets

In Azure Data Lake Dataset we'll be using Deilmited Text as our format.



Publishing completed
Successfully published

File path: cpyadl / Directory / File name

Compression type: Select...

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: ☑

Null value:

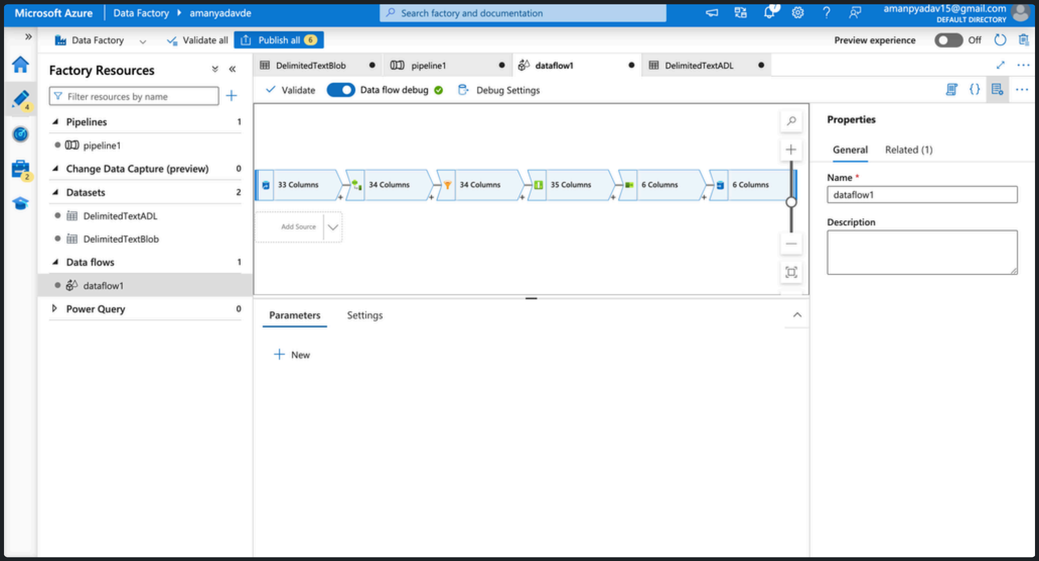
Name *: DelimitedTextADL

Description:

Azure Data Lake Dataset

Creating Pipeline

Step 3. We'll remain under author section and we'll navigate to Pipelines → Create New Pipeline.

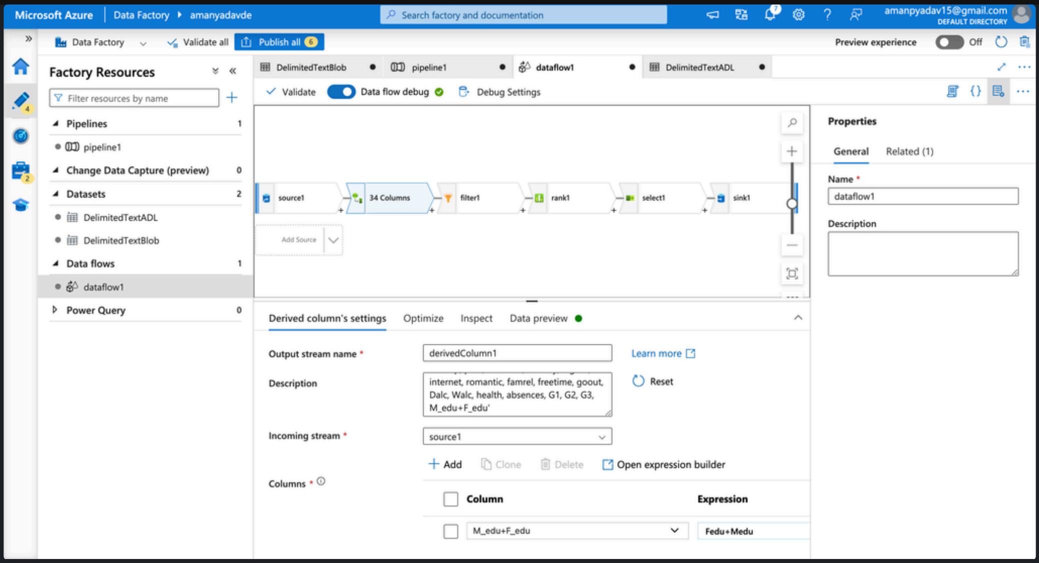


Name *: dataflow1

Description:

Pipeline 1

In Pipeline I have Defined my source as **DelimitedTextBlob**. I've linked then DerivedColumn in which I've used added the values of Fedu and Medu.



Name * dataflow1

Description

Output stream name * derivedColumn1

Description: internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3, M_edu+F_edu'

Incoming stream * source1

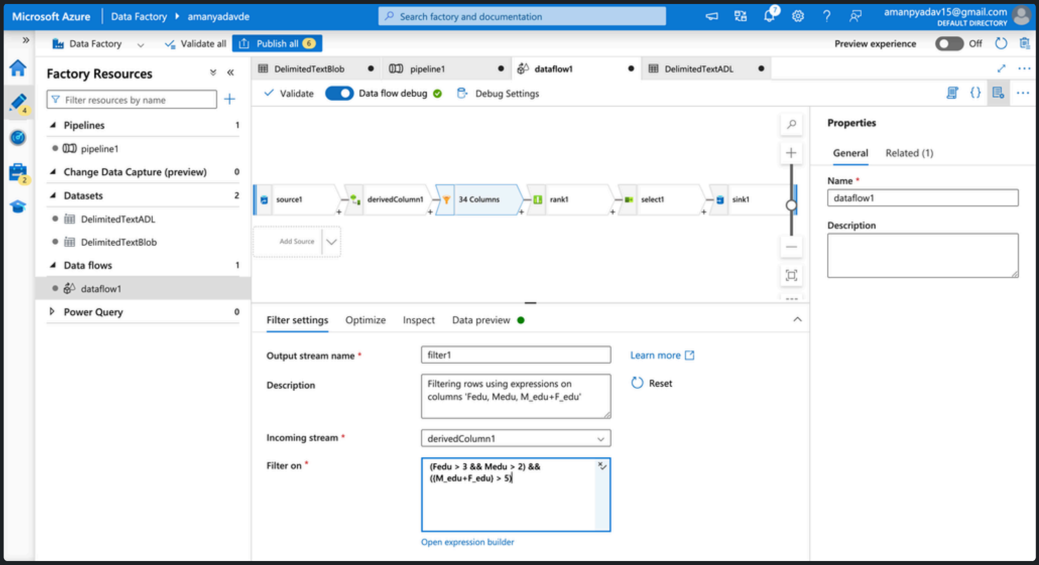
Columns *

Column	Expression
M_edu+F_edu	Fedu+Medu

Derived Column - Pipeline 1

Then I've filtered the data based on the following conditions:

> (Fedu > 3 && Medu > 2) && ({M_edu+F_edu} > 5)



Name * dataflow1

Description

Output stream name * filter1

Description: Filtering rows using expressions on columns 'Fedu, Medu, M_edu+F_edu'

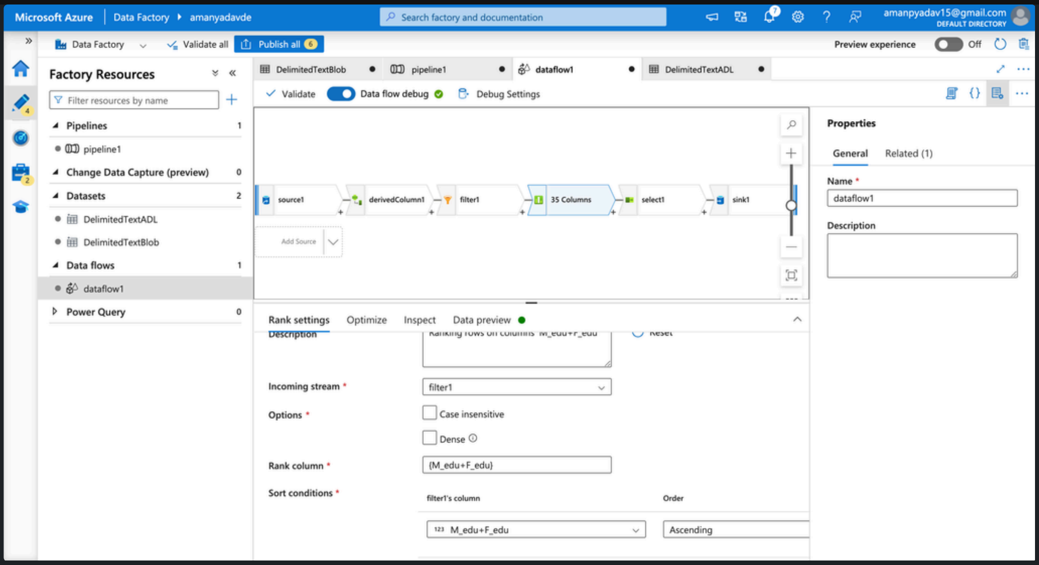
Incoming stream * derivedColumn1

Filter on *

```
(Fedu > 3 && Medu > 2) &&
({M_edu+F_edu} > 5)
```

Filtered - Pipeline 1

Then order the column based on M_Edu+F_edu



Name * dataflow1

Description

Description: Ranking rows on columns M_edu+F_edu

○ Reset

Incoming stream * filter1

Options * □ Case insensitive

□ Dense

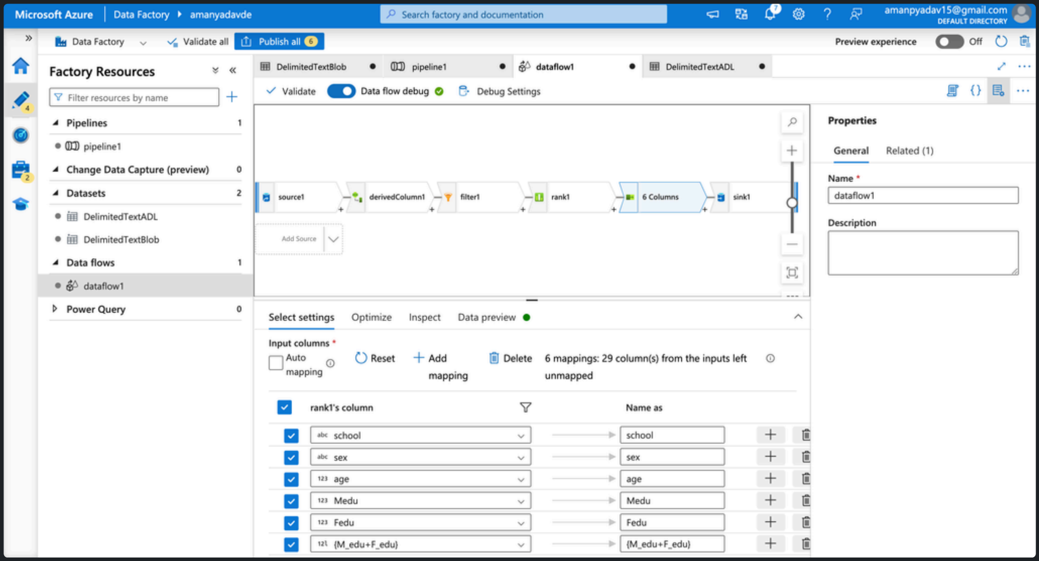
Rank column * {M_edu+F_edu}

Sort conditions *

filter1's column	Order
123 M_edu+F_edu	Ascending

Ranking - Pipeline

Then I selected only School, Sex, Age, Medu, Fedu and {M_edu+F_edu}.



Name * dataflow1

Description

Input columns *

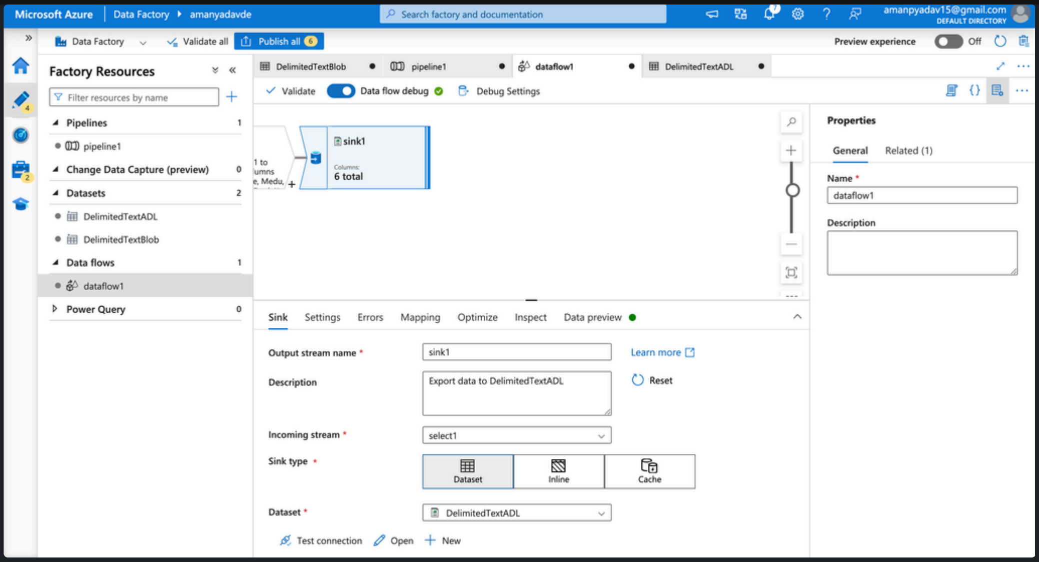
□ Auto mapping

6 mappings: 29 column(s) from the inputs left unmapped

☑	rank1's column	Name as
☑	abc school	school
☑	abc sex	sex
☑	123 age	age
☑	123 Medu	Medu
☑	123 Fedu	Fedu
☑	12l {M_edu+F_edu}	{M_edu+F_edu}

Selection - Pipeline 1

Then sinked into DelimetedTextADL



Factory Resources

- Pipelines 1
  - pipeline1
- Change Data Capture (preview) 0
- Datasets 2
  - DelimitedTextADL
  - DelimitedTextBlob
- Data flows 1
  - dataflow1
- Power Query 0

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

Output stream name *: sink1

Description: Export data to DelimitedTextADL

Incoming stream *: select1

Sink type *: Dataset | Inline | Cache

Dataset *: DelimitedTextADL

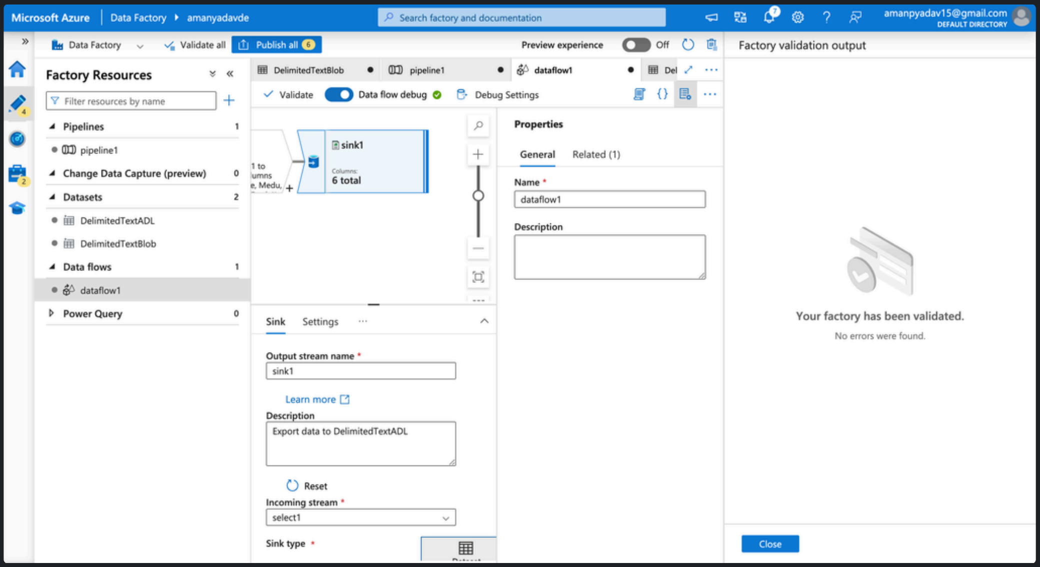
Properties

Name *: dataflow1

Description:

Sink - Pipeline 1

Once done with it I clicked on “Validate All”



Factory validation output

Your factory has been validated.

No errors were found.

Output stream name *: sink1

Description: Export data to DelimitedTextADL

Incoming stream *: select1

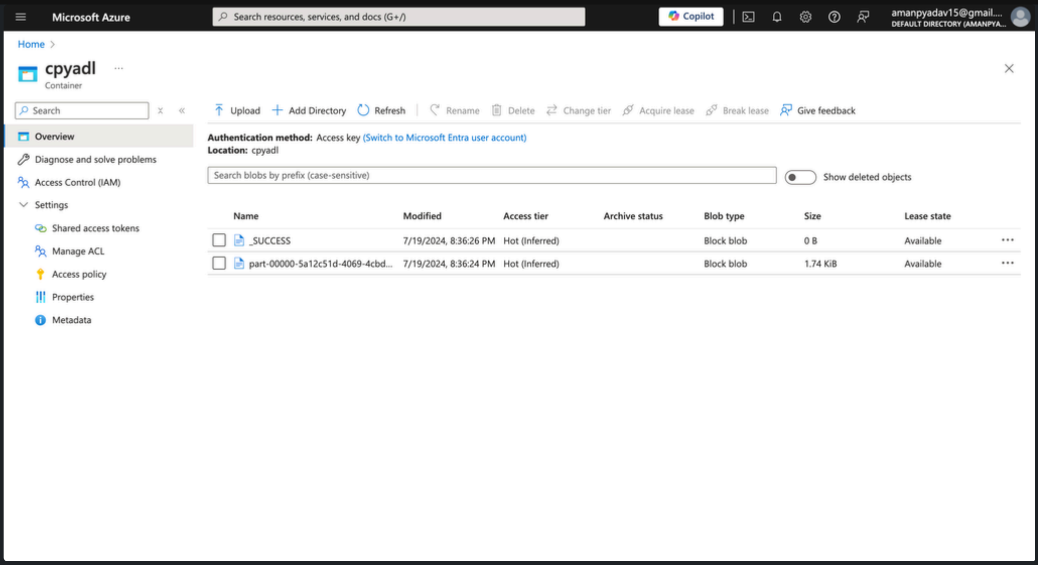
Properties

Name *: dataflow1

Validating & Publishing

Once there were no errors found I’ve published it.

Checking on our destination dataset.

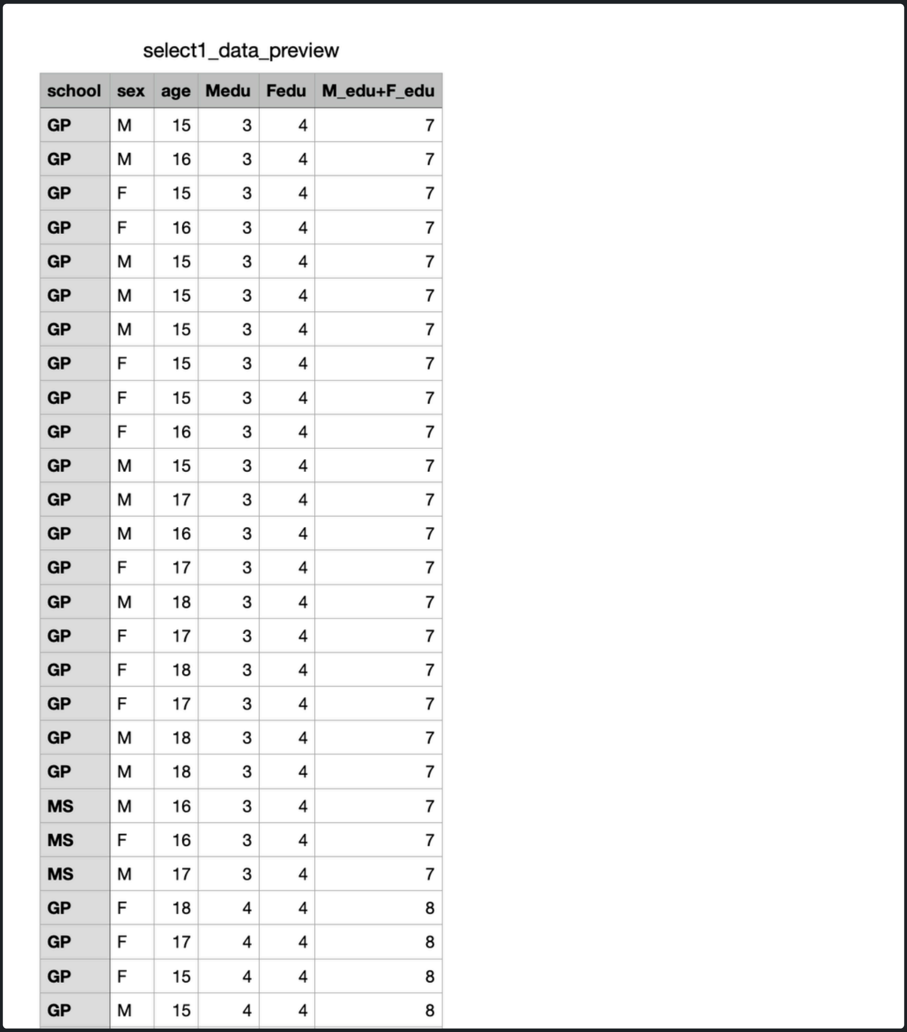


Authentication method: Access key (Switch to Microsoft Entra user account)
Location: cpyadl

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
_SUCCESS	7/19/2024, 8:36:26 PM	Hot (Inferred)		Block blob	0 B	Available
part-00000-5a12c51d-4069-4cbd...	7/19/2024, 8:36:24 PM	Hot (Inferred)		Block blob	1.74 KiB	Available

Desitnation Directory

Output Data:



select1_data_preview

school	sex	age	Medu	Fedu	M_edu+F_edu
GP	M	15	3	4	7
GP	M	16	3	4	7
GP	F	15	3	4	7
GP	F	16	3	4	7
GP	M	15	3	4	7
GP	M	15	3	4	7
GP	M	15	3	4	7
GP	F	15	3	4	7
GP	F	15	3	4	7
GP	F	16	3	4	7
GP	M	15	3	4	7
GP	M	17	3	4	7
GP	M	16	3	4	7
GP	F	17	3	4	7
GP	M	18	3	4	7
GP	F	17	3	4	7
GP	F	18	3	4	7
GP	F	17	3	4	7
GP	M	18	3	4	7
GP	M	18	3	4	7
MS	M	16	3	4	7
MS	F	16	3	4	7
MS	M	17	3	4	7
GP	F	18	4	4	8
GP	F	17	4	4	8
GP	F	15	4	4	8
GP	M	15	4	4	8

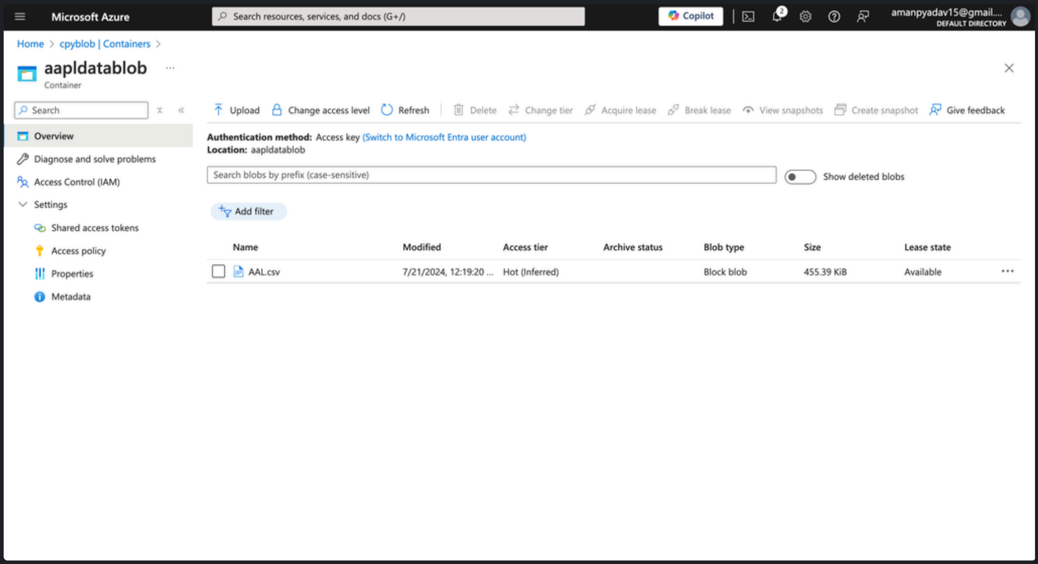
Pipeline 2

Dataset: AAL stocks

Objective

To analyze the market trend of Apple.

We will first upload our csv files in our blob storage account which we previously created.

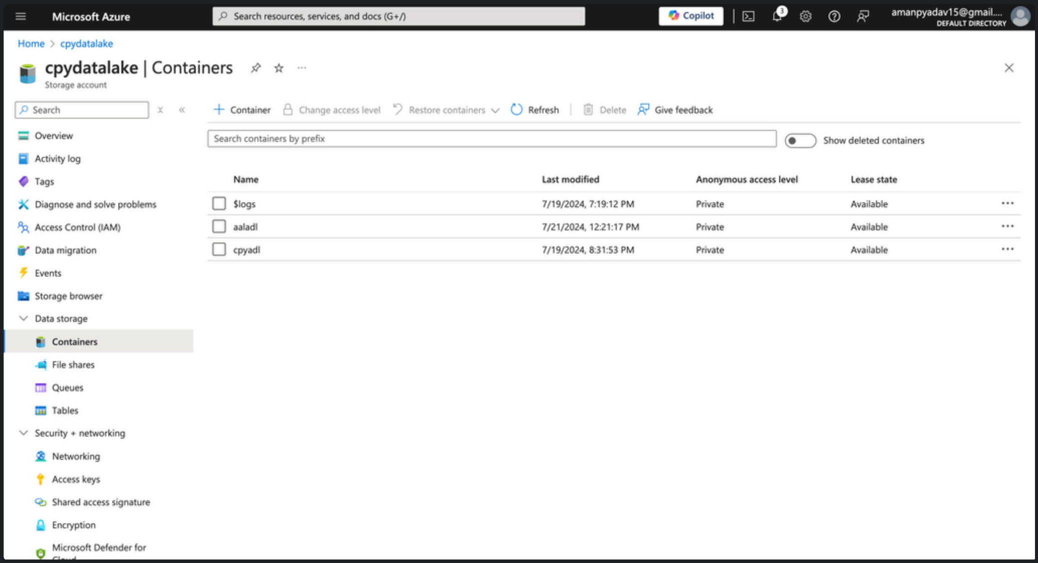


Authentication method: Access key (Switch to Microsoft Entra user account)
Location: aapldatablob

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
AAL.csv	7/21/2024, 12:19:20 ...	Hot (Inferred)		Block blob	455.39 KiB	Available

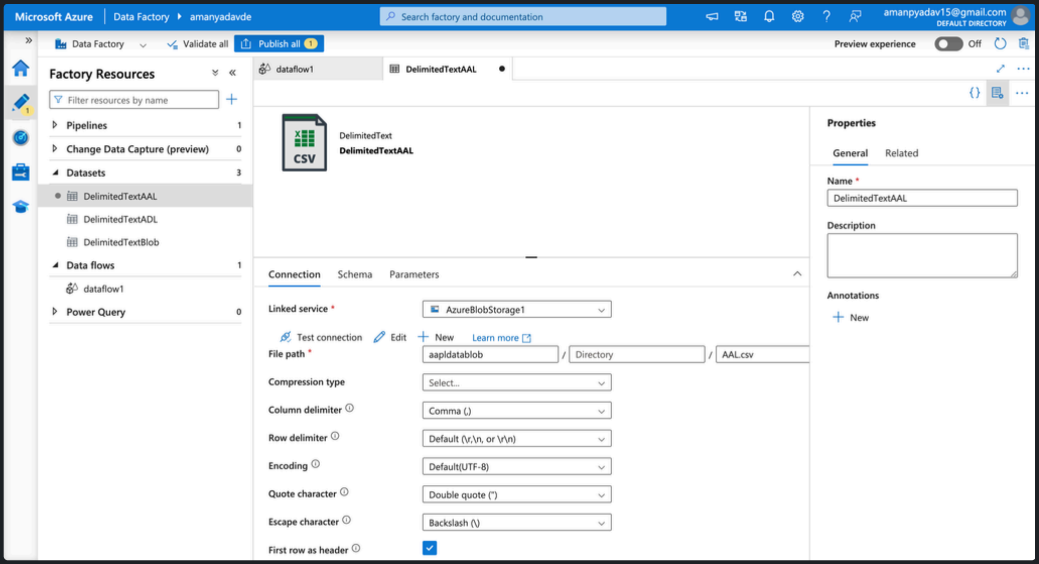
Blob Storage Account

Then will just create a new container at our Data lake Storage Account as **aaladl** .



Name	Last modified	Anonymous access level	Lease state
$logs	7/19/2024, 7:19:12 PM	Private	Available
aaladl	7/21/2024, 12:21:17 PM	Private	Available
cpyadl	7/19/2024, 8:31:53 PM	Private	Available

Then in our azure data factory studio we will create a new dataset for our blob storage and we will just be using the linked services which we created previously.



Name * DelimitedTextAAL

Description

Linked service * AzureBlobStorage1

File path * aapldatablob / Directory / AAL.csv

Compression type: Select...

Column delimiter: Comma (,)

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

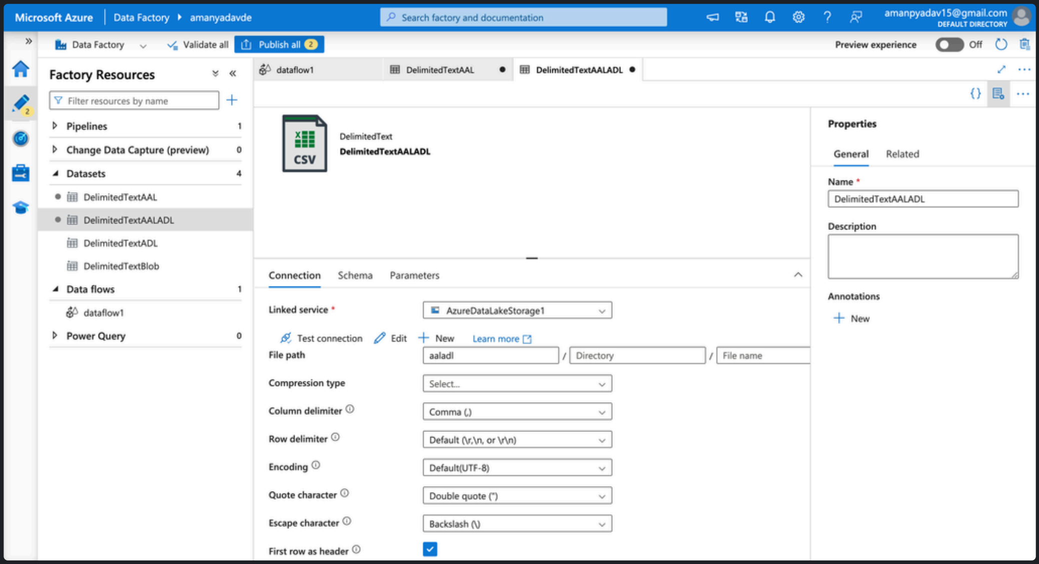
Encoding: Default(UTF-8)

Quote character: Double quote (")

Escape character: Backslash (\)

First row as header ☑

Creating a dataset for our data lake storage as well.



Name * DelimitedTextAALADL

Description

Linked service * AzureDataLakeStorage1

File path aaladl / Directory / File name

Compression type: Select...

Column delimiter: Comma (,)

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

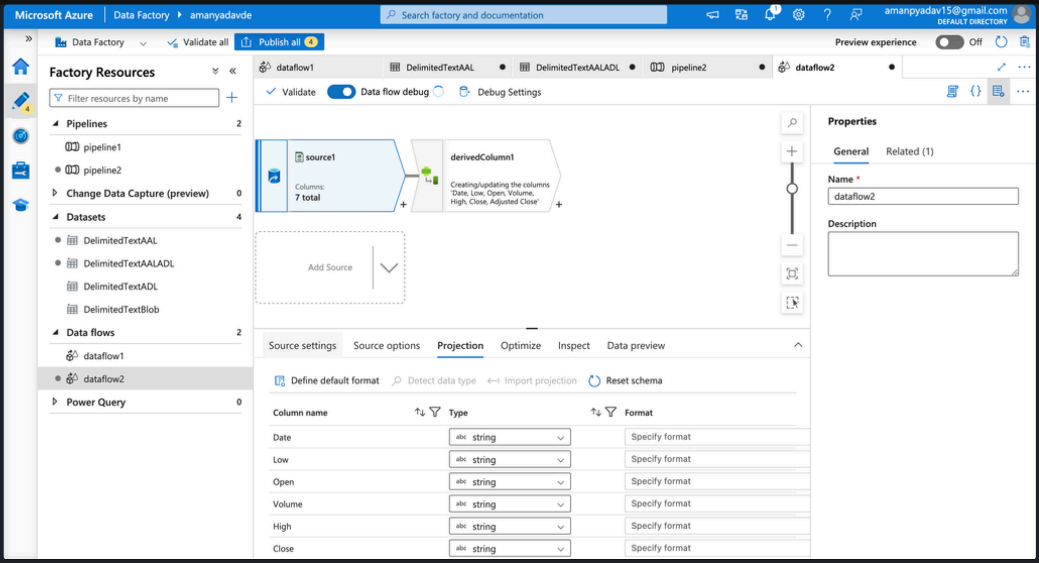
Encoding: Default(UTF-8)

Quote character: Double quote (")

Escape character: Backslash (\)

First row as header ☑

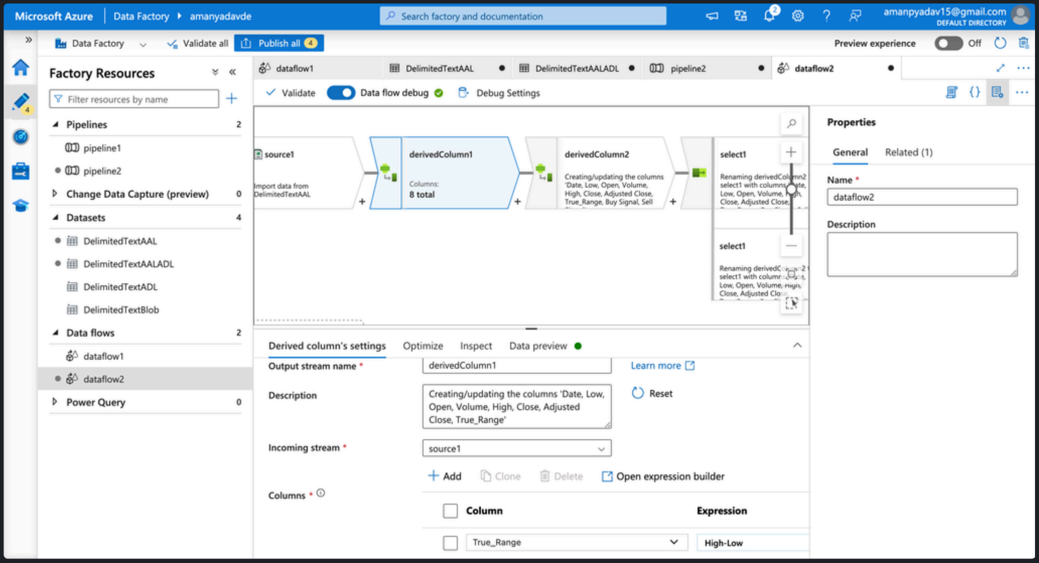
Creating a new pipeline 2 with dataflow2.



Name *: dataflow2

Column name	Type	Format
Date	abc string	Specify format
Low	abc string	Specify format
Open	abc string	Specify format
Volume	abc string	Specify format
High	abc string	Specify format
Close	abc string	Specify format

Added Derived column in which the column name is **True_Range** and expression is High - Low



Name *: dataflow2

Output stream name *: derivedColumn1

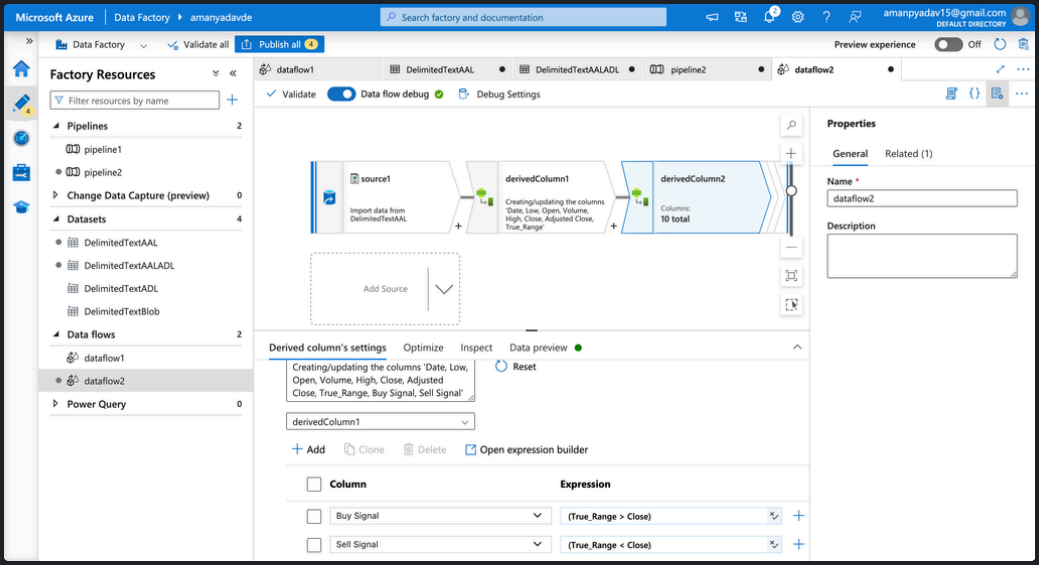
Description: Creating/updating the columns 'Date, Low, Open, Volume, High, Close, Adjusted Close, True_Range'

Incoming stream *: source1

Columns *

Column	Expression
True_Range	High-Low

Added another derived column in which I've added two more columns name **BUY SIGNAL** and **SELL SIGNAL**



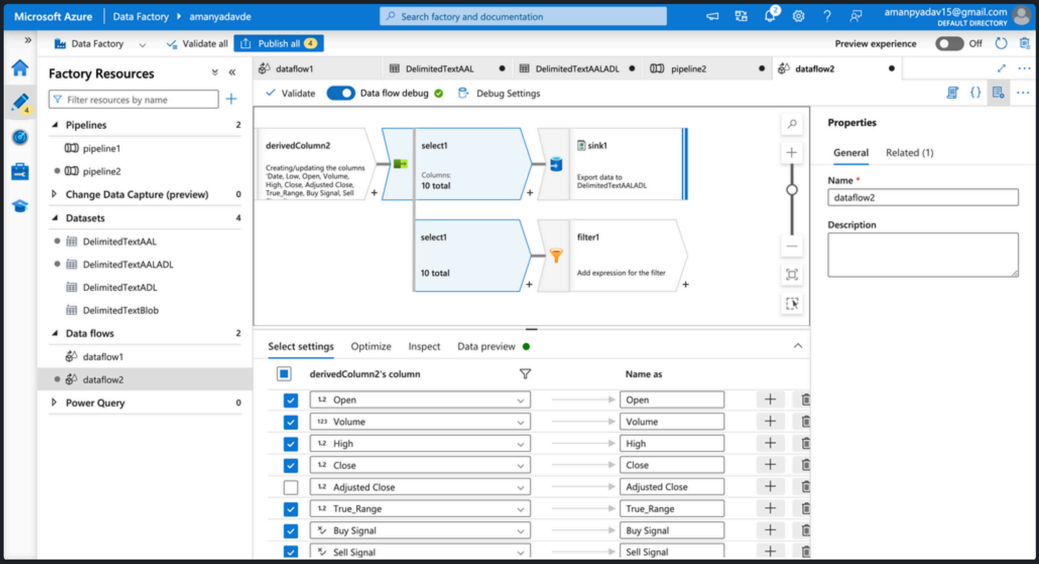
Name *: dataflow2

Creating/updating the columns 'Date, Low, Open, Volume, High, Close, Adjusted Close, True_Range, Buy Signal, Sell Signal'

derivedColumn1

Column	Expression
Buy Signal	(True_Range > Close)
Sell Signal	(True_Range < Close)

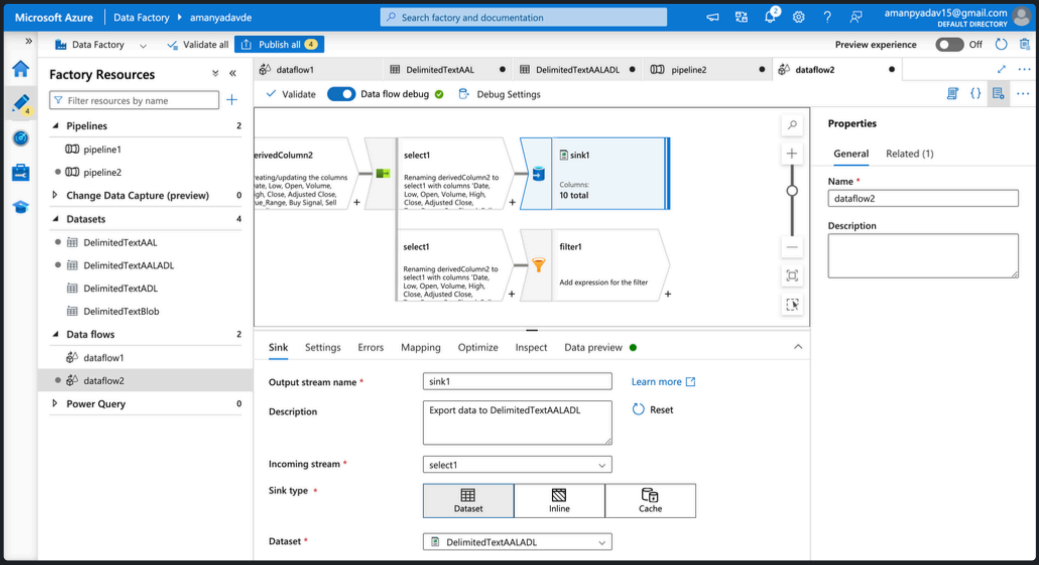
Then selected the column below:



Name *: dataflow2

	derivedColumn2's column	Name as
☑	1.2 Open	Open
☑	123 Volume	Volume
☑	1.2 High	High
☑	1.2 Close	Close
☐	1.2 Adjusted Close	Adjusted Close
☑	1.2 True_Range	True_Range
☑	Buy Signal	Buy Signal
☑	Sell Signal	Sell Signal

Then sinked in to our destination storage account i.e. Data Lake Storage.



Name *: dataflow2

Output stream name *: sink1

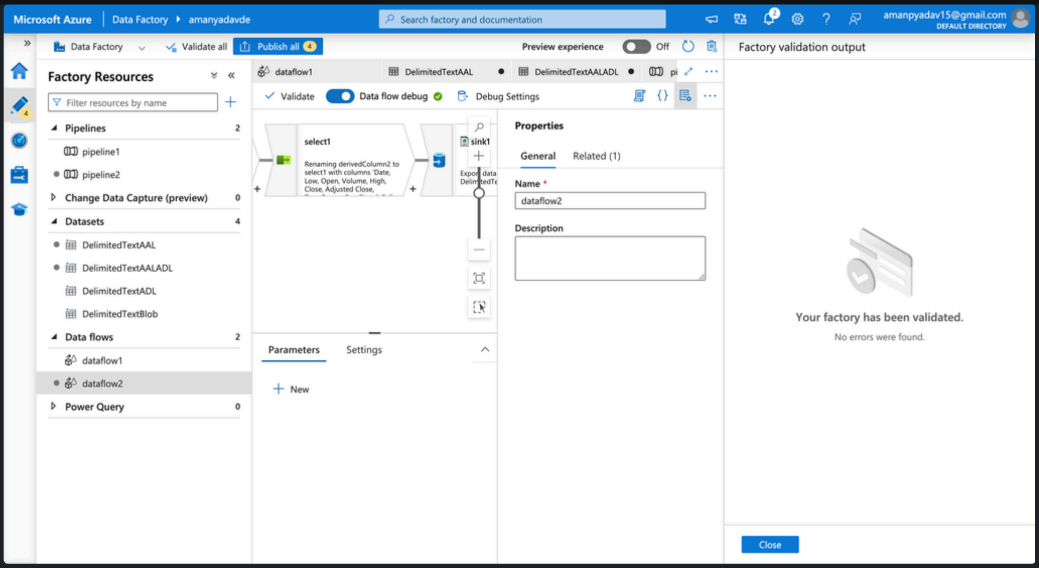
Description: Export data to DelimitedTextAALADL

Incoming stream *: select1

Sink type *: [✓ Dataset] Inline Cache

Dataset *: DelimitedTextAALADL

Validated everything to see if there’s any error.



Renaming derivedColumn2 to select1 with columns 'Date, Low, Open, Volume, High, Close, Adjusted Close,

Name * dataflow2

Description

Your factory has been validated.

No errors were found.

Published it, and let’s check in our destination folder whether we got the file or not.

Output Data: