Lab Manual- Azure Data Factory Provisioning and Data Ingestion Part1

Prepared for:

Date: 18th March 2022

Prepared by:

Document Name: Lab Manual

Document Number AZLabn990

Contributor:

Contents

1. Introduction

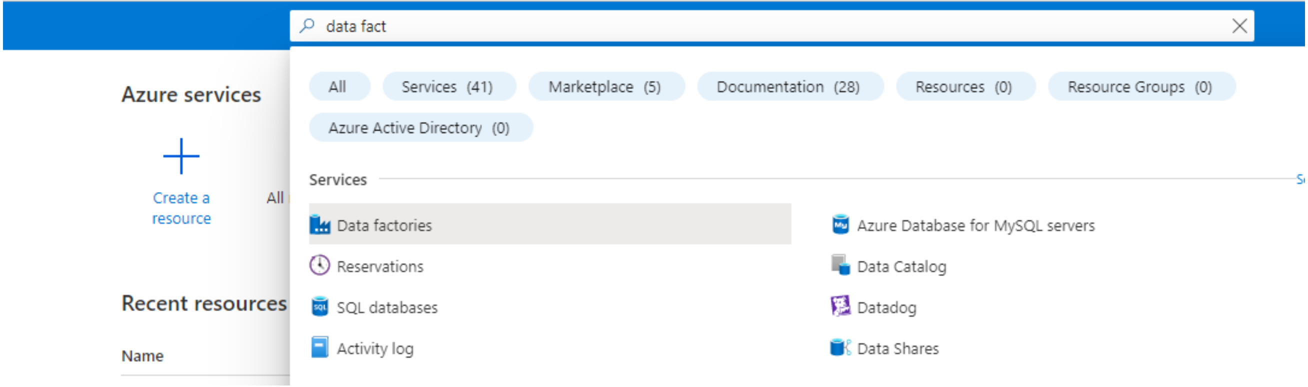
Azure Data Factory (ADF) is a data pipeline orchestrator and ETL tool that is part of the Microsoft Azure cloud ecosystem. ADF can pull data from the outside world (FTP, Amazon S3, Oracle), transform it, filter it, enhance it, and move it along to another destination. In my work for a health-data project we are using ADF to drive our data flow from raw ingestion to polished analysis that is ready to display.

- Stores data with the help of Azure Data Lake Storage
- Analyzes the data
- Transforms the data with the help of pipelines (a logical grouping of activities that together perform a task)
- Publishes the organized data
- Visualizes the data with third-party applications like Apache Spark or Hadoop.

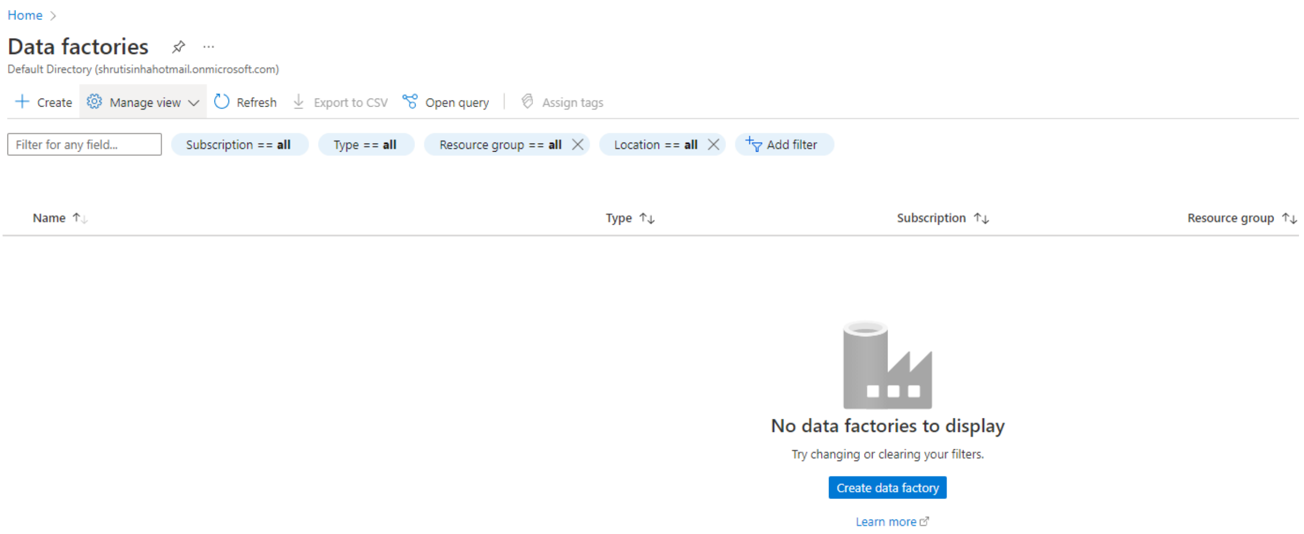
In this Azure Data Factory tutorial, you will learn about Azure Data Factory, its basic concepts and why do we need it. Also, you will learn the working process of Azure Data Factory and will be introduced to Azure Data Lake. Here, you will learn how to copy data from Azure SQL to Azure Data Lake,

2. Exercise 1 – Provisioning Azure Data Factory

1. Go to the Azure portal.
2. In the Azure portal search type Data Factory



3. Click **Create Data Factory** page,



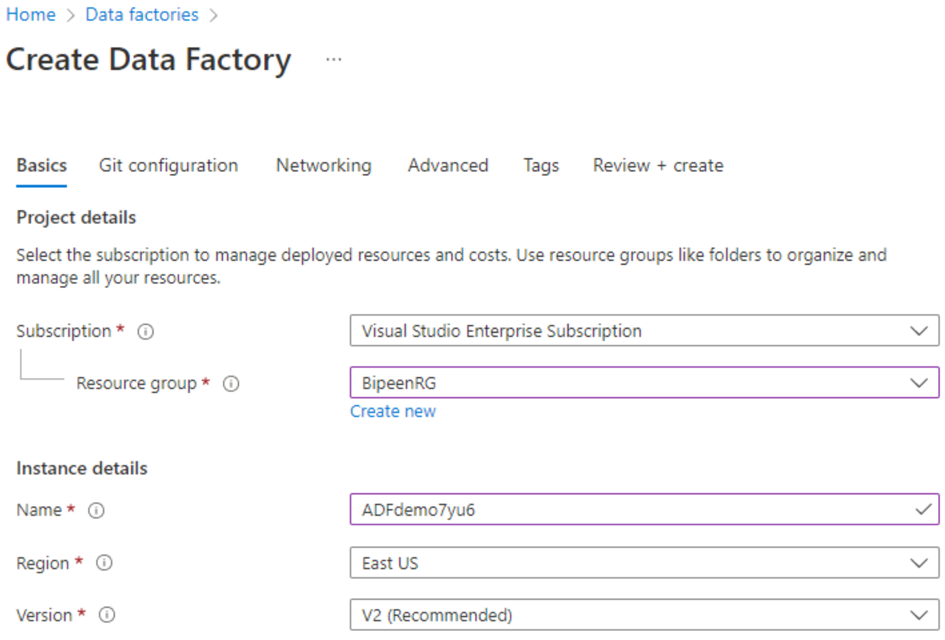
Data factories

Default Directory (shrutisinhahotmail.onmicrosoft.com)

No data factories to display

Try changing or clearing your filters.

4. Under **Basics** tab, select your Azure **Subscription** in which you want to create the data factory.
5. For **Resource Group** Select an existing resource group from the drop-down list.
6. For **Name**, enter **ADFDemo+number**



Create Data Factory

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : Visual Studio Enterprise Subscription

Resource group * : BipeenRG

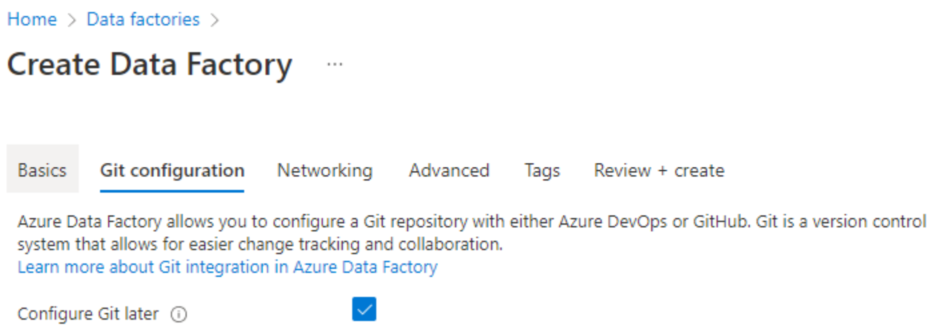
Instance details

Name * : ADFdemo7yu6

Region * : East US

Version * : V2 (Recommended)

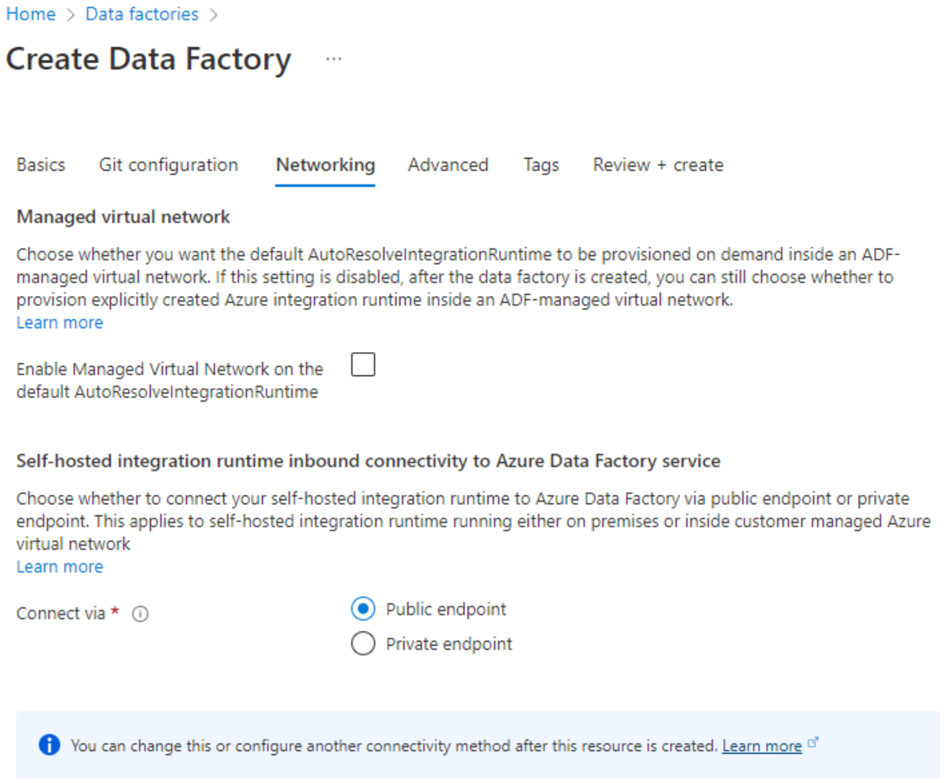
7. Select **Next: Git configuration**, and then select **Configure Git later** check box.



Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration.
Learn more about Git integration in Azure Data Factory

Configure Git later ⓘ ☑

8. Click Next in Networking



Managed virtual network

Choose whether you want the default AutoResolveIntegrationRuntime to be provisioned on demand inside an ADF-managed virtual network. If this setting is disabled, after the data factory is created, you can still choose whether to provision explicitly created Azure integration runtime inside an ADF-managed virtual network.
Learn more

Enable Managed Virtual Network on the default AutoResolveIntegrationRuntime □

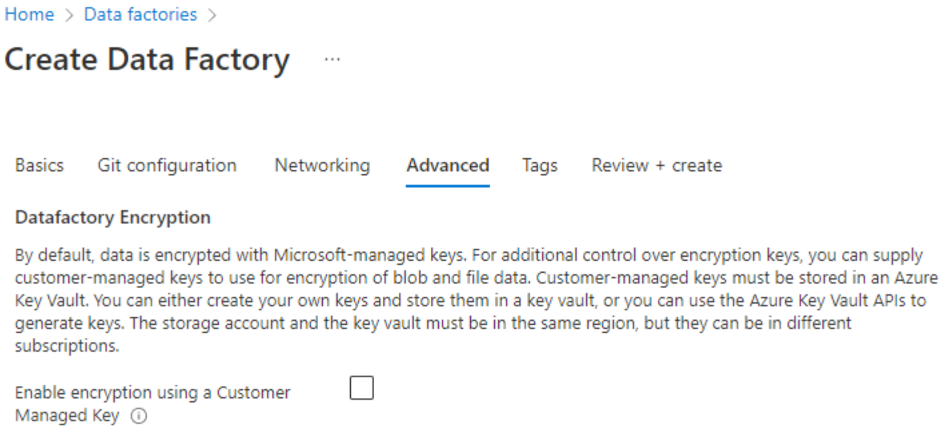
Self-hosted integration runtime inbound connectivity to Azure Data Factory service

Choose whether to connect your self-hosted integration runtime to Azure Data Factory via public endpoint or private endpoint. This applies to self-hosted integration runtime running either on premises or inside customer managed Azure virtual network
Learn more

Connect via * ⓘ ● Public endpoint
○ Private endpoint

ⓘ You can change this or configure another connectivity method after this resource is created. Learn more

9. Click Next in Advance

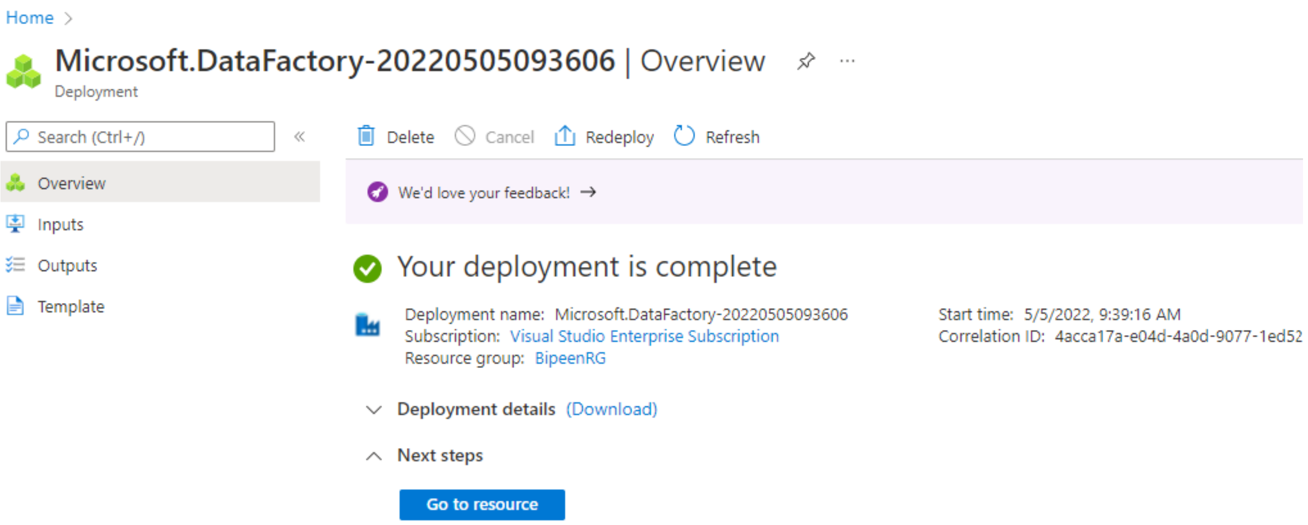


Datafactory Encryption

By default, data is encrypted with Microsoft-managed keys. For additional control over encryption keys, you can supply customer-managed keys to use for encryption of blob and file data. Customer-managed keys must be stored in an Azure Key Vault. You can either create your own keys and store them in a key vault, or you can use the Azure Key Vault APIs to generate keys. The storage account and the key vault must be in the same region, but they can be in different subscriptions.

Enable encryption using a Customer Managed Key ⓘ □

10. Select **Review + create**, and select **Create** after the validation is passed.



Your deployment is complete

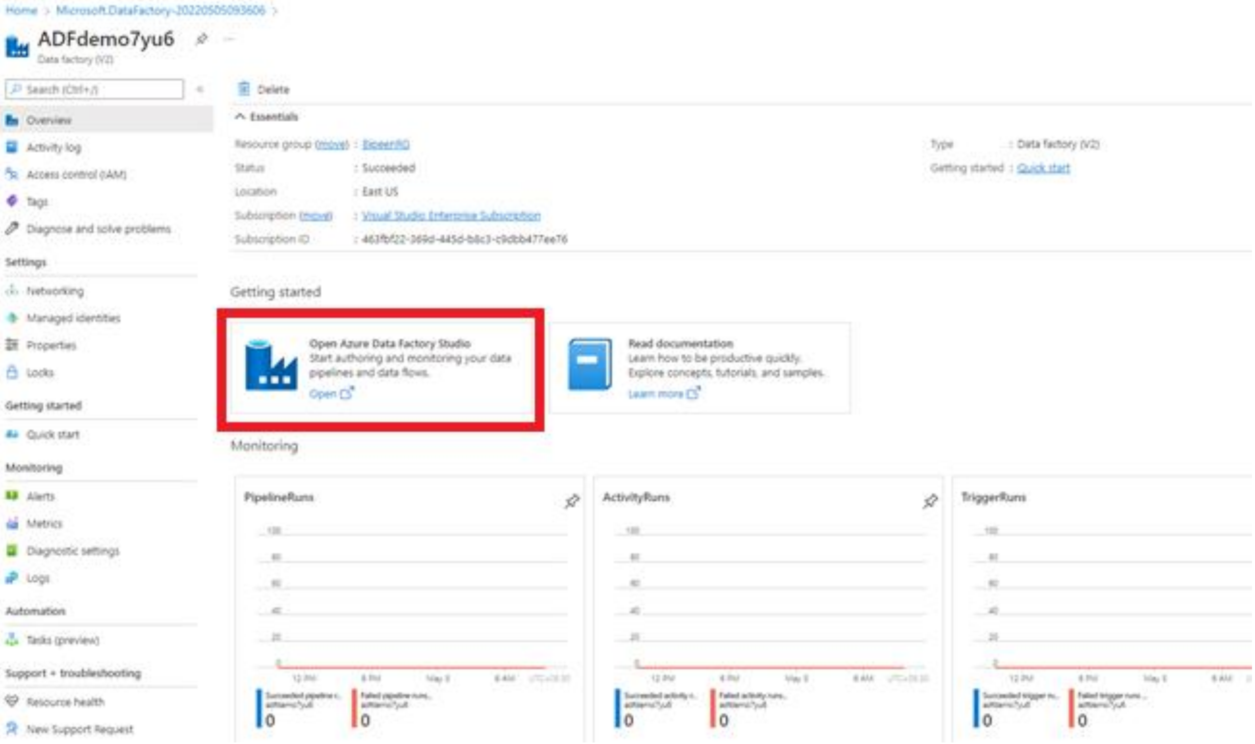
Deployment name: Microsoft.DataFactory-20220505093606
Subscription: Visual Studio Enterprise Subscription
Resource group: BipeenRG

Start time: 5/5/2022, 9:39:16 AM
Correlation ID: 4acca17a-e04d-4a0d-9077-1ed52

11. After the creation is complete, select **Go to resource** to navigate to the **Data Factory** page.

3. Exercise 2 – Launch Azure Data Factory Studio

1. Select Open on the Open **Azure Data Factory Studio** .

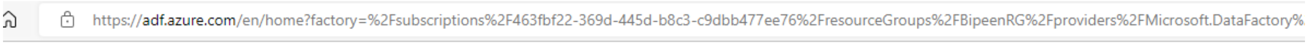


Resource group (move) : BipeenRG
Status : Succeeded
Location : East US
Subscription (move) : Visual Studio Enterprise Subscription
Subscription ID : 463fbf22-369d-445d-b8c3-c9dbb477ee76
Type : Data factory (V2)
Getting started : Quick start

Open Azure Data Factory Studio
Start authoring and monitoring your data pipelines and data flows.
Open

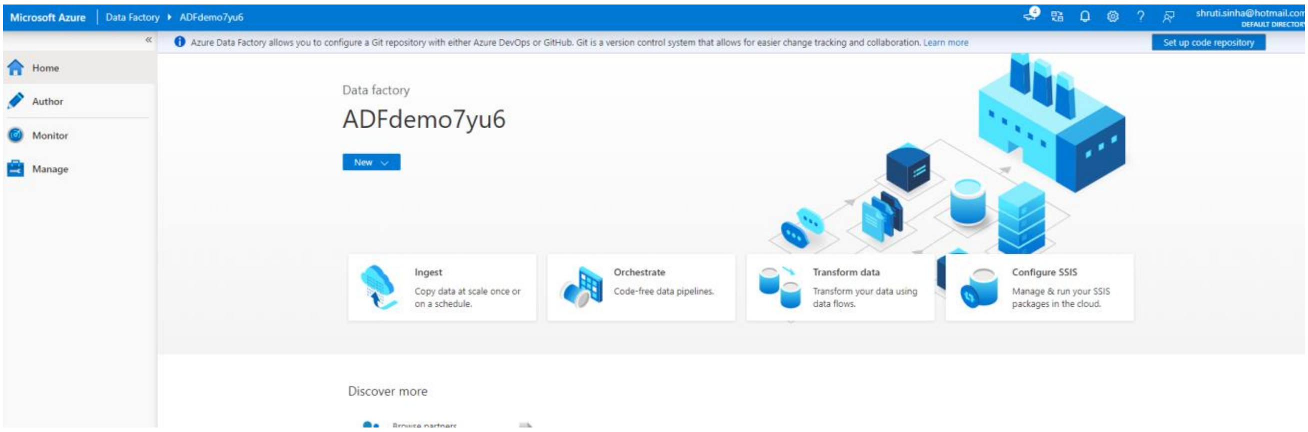
Read documentation
Learn how to be productive quickly.
Explore concepts, tutorials, and samples.
Learn more

2. It start the Azure Data Factory user interface (UI) application on a separate browser tab.



https://adf.azure.com/en/home?factory=%2Fsubscriptions%2F463fbf22-369d-445d-b8c3-c9dbb477ee76%2FresourceGroups%2FBipeenRG%2Fproviders%2FMicrosoft.DataFactory%

3. You Should be in Azure Data Factory Studio



Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration. Learn more

Data factory

ADFdemo7yu6

Ingest: Copy data at scale once or on a schedule.

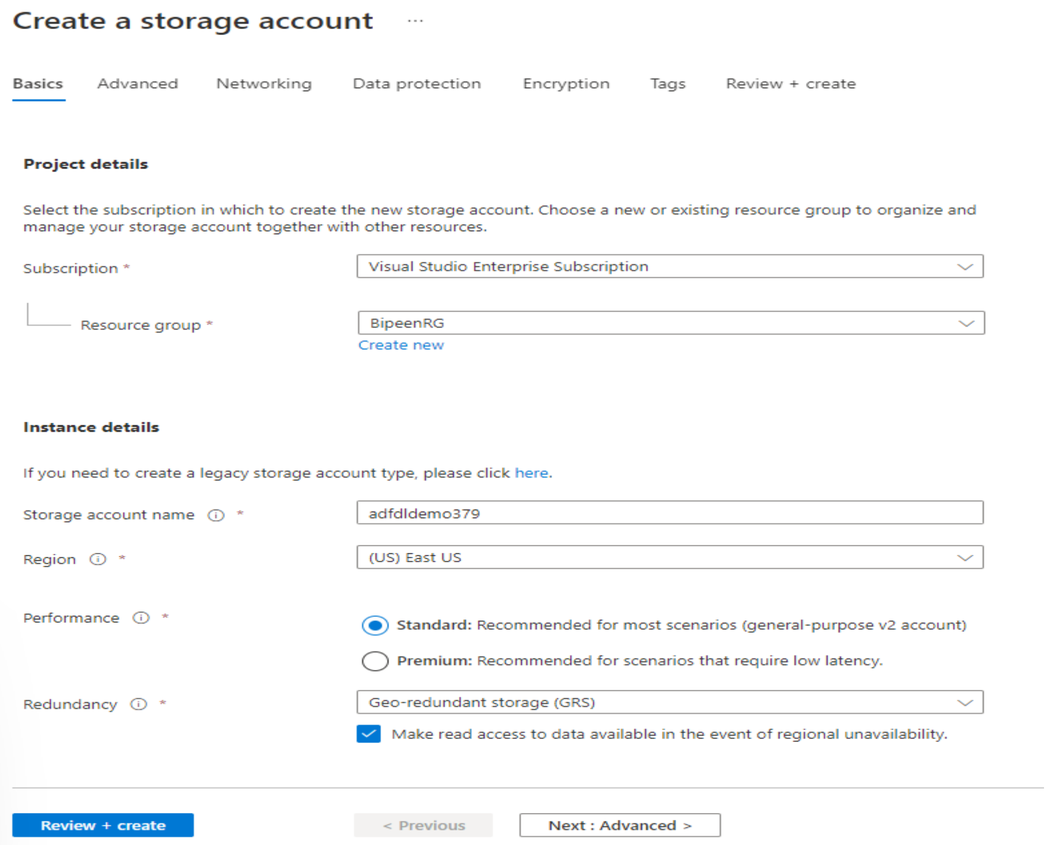
Orchestrate: Code-free data pipelines.

Transform data: Transform your data using data flows.

Configure SSIS: Manage & run your SSIS packages in the cloud.

4. Exercise 3 – Create a New Data Lake Storage

1. Create new storage as we did in our prevision exercise



Create a storage account ···

Basics | Advanced | Networking | Data protection | Encryption | Tags | Review + create

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * : Visual Studio Enterprise Subscription

Resource group * : BipeenRG
Create new

Instance details

If you need to create a legacy storage account type, please click here.

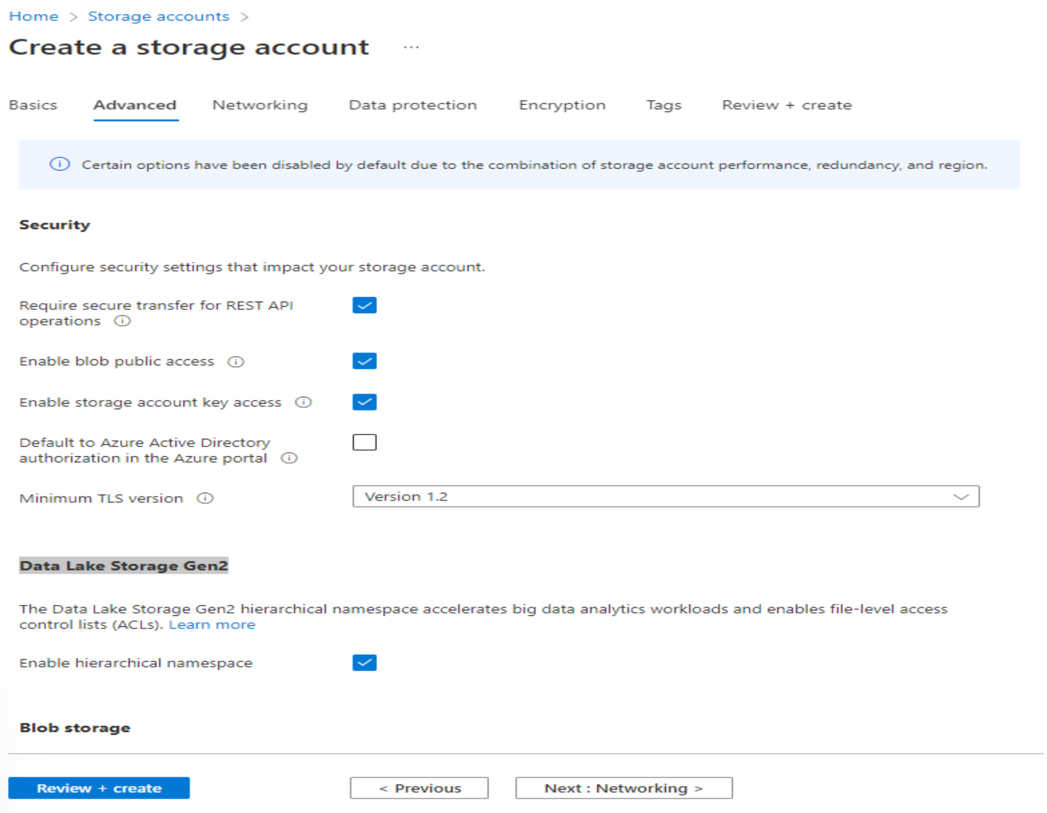
Storage account name ⓘ * : adfdldemo379

Region ⓘ * : (US) East US

Performance ⓘ * :
● **Standard:** Recommended for most scenarios (general-purpose v2 account)
○ **Premium:** Recommended for scenarios that require low latency.

Redundancy ⓘ * : Geo-redundant storage (GRS)
☑ Make read access to data available in the event of regional unavailability.

2. Enable Datalake and click Review and Create



Home > Storage accounts >

Create a storage account ···

Basics | Advanced | Networking | Data protection | Encryption | Tags | Review + create

ⓘ Certain options have been disabled by default due to the combination of storage account performance, redundancy, and region.

Security

Configure security settings that impact your storage account.

Require secure transfer for REST API operations ⓘ ☑

Enable blob public access ⓘ ☑

Enable storage account key access ⓘ ☑

Default to Azure Active Directory authorization in the Azure portal ⓘ □

Minimum TLS version ⓘ : Version 1.2

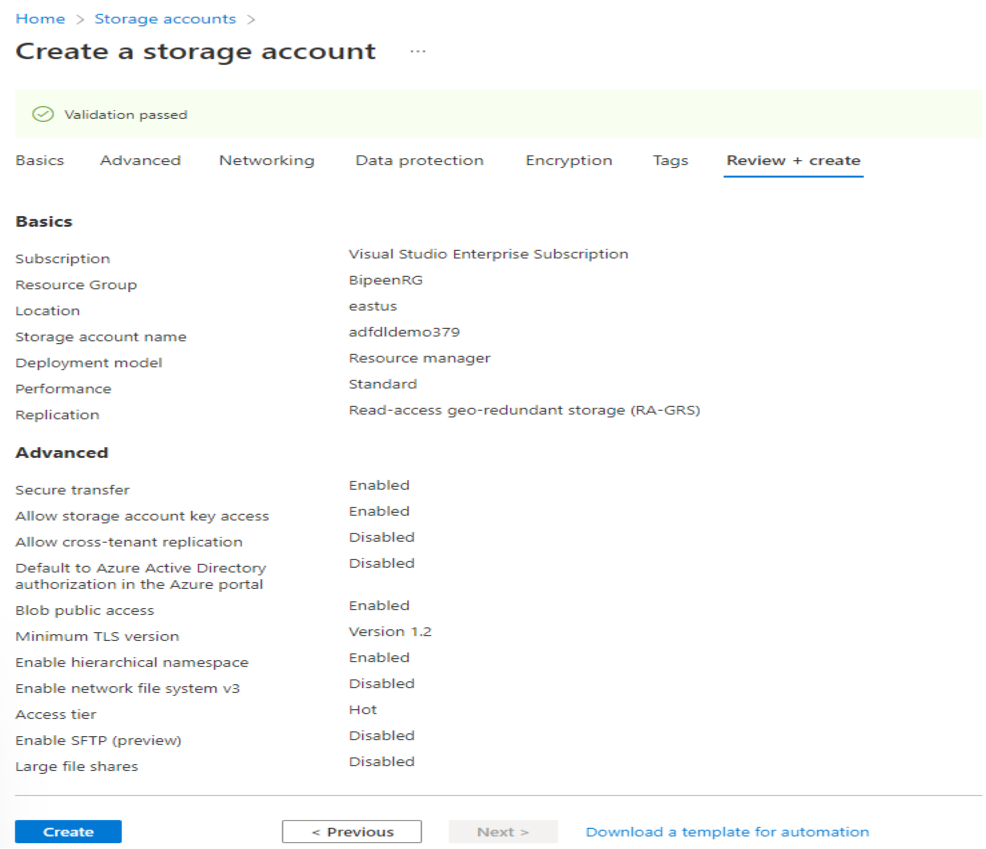
Data Lake Storage Gen2

The Data Lake Storage Gen2 hierarchical namespace accelerates big data analytics workloads and enables file-level access control lists (ACLs). Learn more

Enable hierarchical namespace ☑

Blob storage

3. Click **Create**



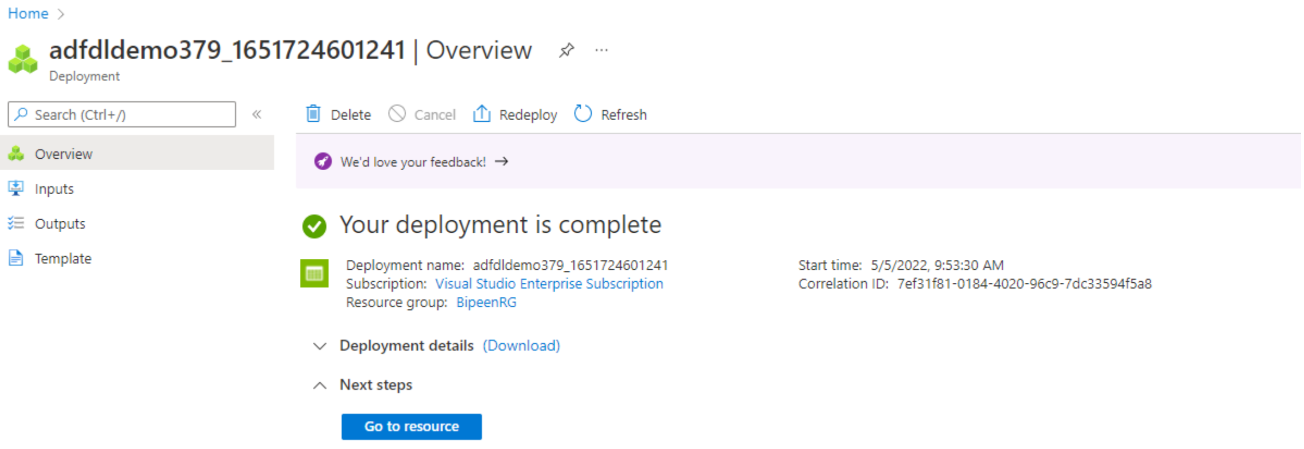
Basics

Subscription	Visual Studio Enterprise Subscription
Resource Group	BipeenRG
Location	eastus
Storage account name	adfdldemo379
Deployment model	Resource manager
Performance	Standard
Replication	Read-access geo-redundant storage (RA-GRS)

Advanced

Secure transfer	Enabled
Allow storage account key access	Enabled
Allow cross-tenant replication	Disabled
Default to Azure Active Directory authorization in the Azure portal	Disabled
Blob public access	Enabled
Minimum TLS version	Version 1.2
Enable hierarchical namespace	Enabled
Enable network file system v3	Disabled
Access tier	Hot
Enable SFTP (preview)	Disabled
Large file shares	Disabled

4. It should be created in few Minutes



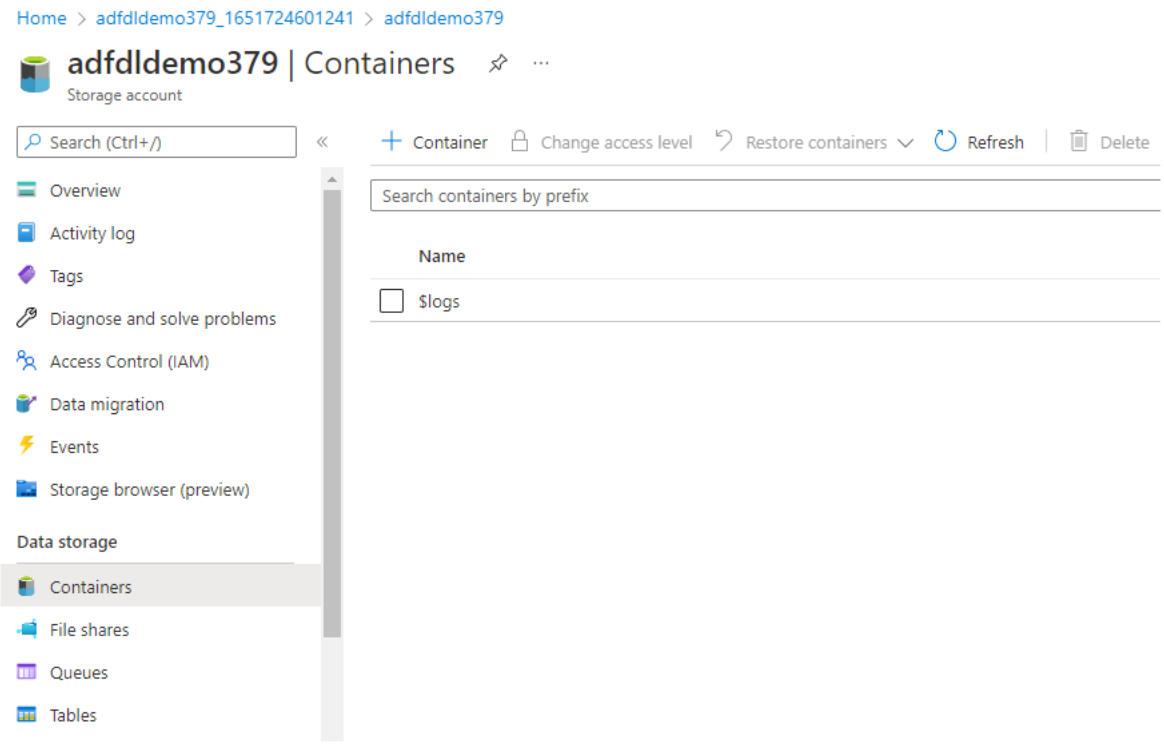
Your deployment is complete

Deployment name: adfdldemo379_1651724601241
Subscription: Visual Studio Enterprise Subscription
Resource group: BipeenRG

Start time: 5/5/2022, 9:53:30 AM
Correlation ID: 7ef31f81-0184-4020-96c9-7dc33594f5a8

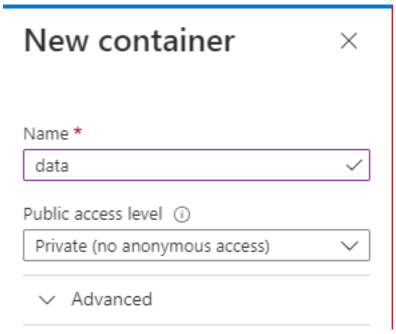
5. Exercise 4 – Create Data Lake Container and Upload Data

1. From the storage account page, select **Overview** > **Containers**.



	Name
□	$logs

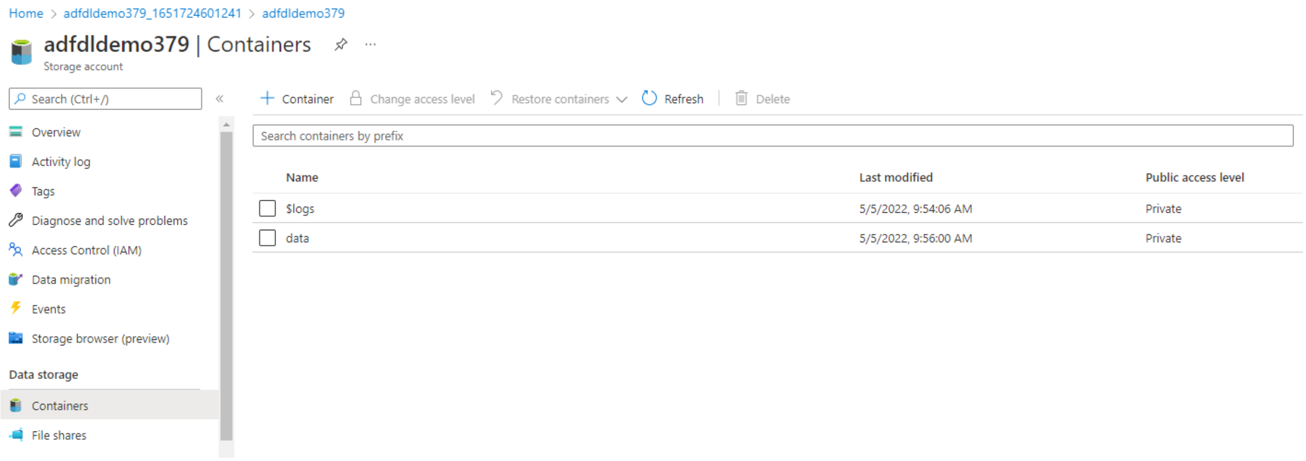
2. In the **New container** dialog box, enter **data** for the name, and then select **OK**



Name *: data

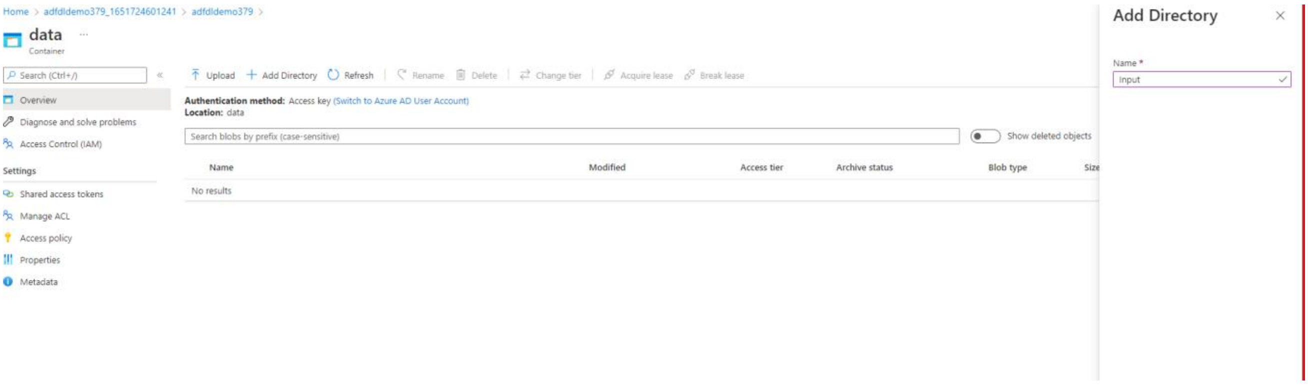
Public access level: Private (no anonymous access)

3. The **Containers** page is updated to include **data** in the list of containers.



	Name	Last modified	Public access level
□	$logs	5/5/2022, 9:54:06 AM	Private
□	data	5/5/2022, 9:56:00 AM	Private

4. Click **Add Directory** and type **Input** and click ok

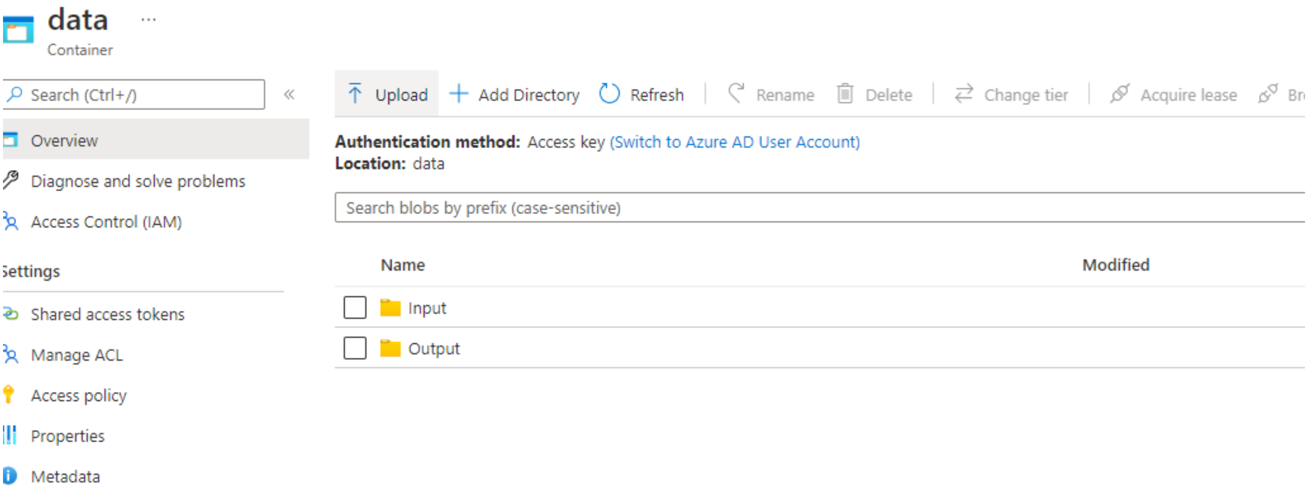


Authentication method: Access key (Switch to Azure AD User Account)
Location: data

No results

Name * Input

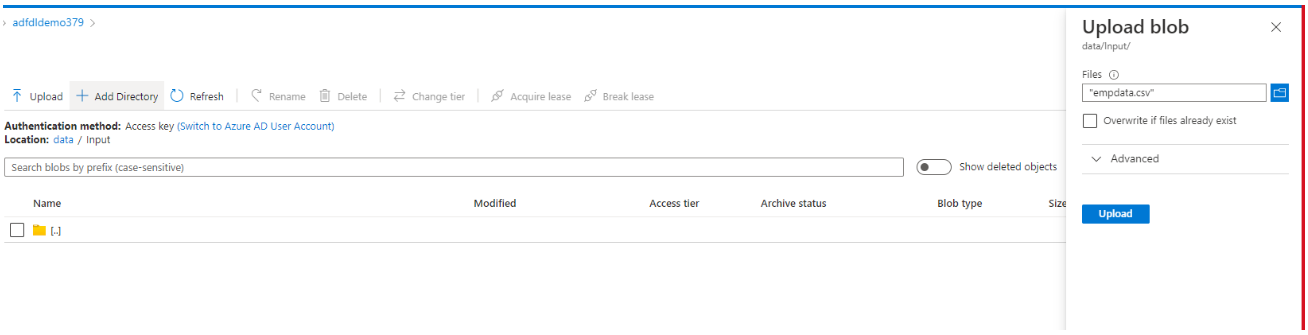
5. Click **Add Directory** and type **Output** and click ok



Authentication method: Access key (Switch to Azure AD User Account)
Location: data

□ Input
□ Output

6. Click the **Input Directory** and Click **Upload**.
7. In the **Upload blob** page, select the **Files** box, and then browse to and select the **empdata.CSV** file.



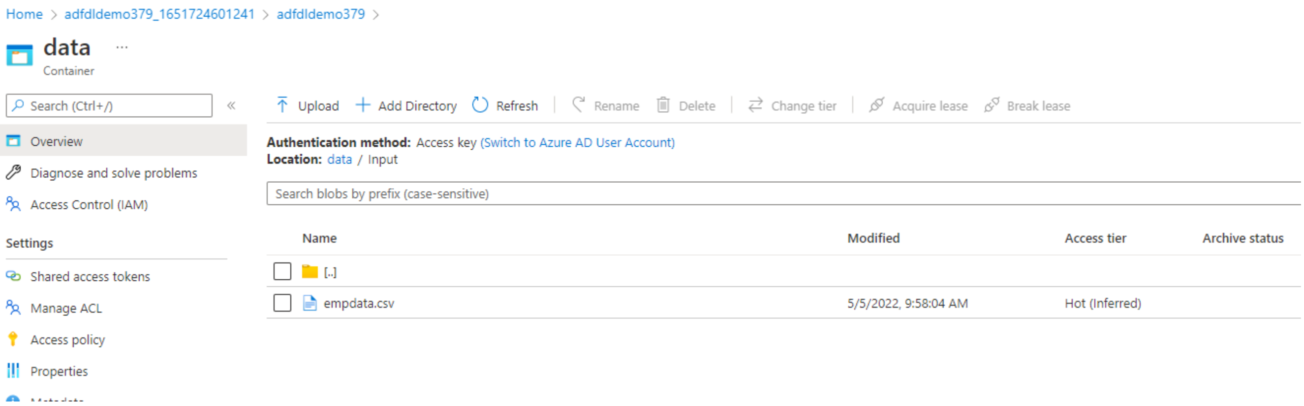
Authentication method: Access key (Switch to Azure AD User Account)
Location: data / Input

□ [..]

Upload blob
data/Input/

Files "empdata.csv"

□ Overwrite if files already exist

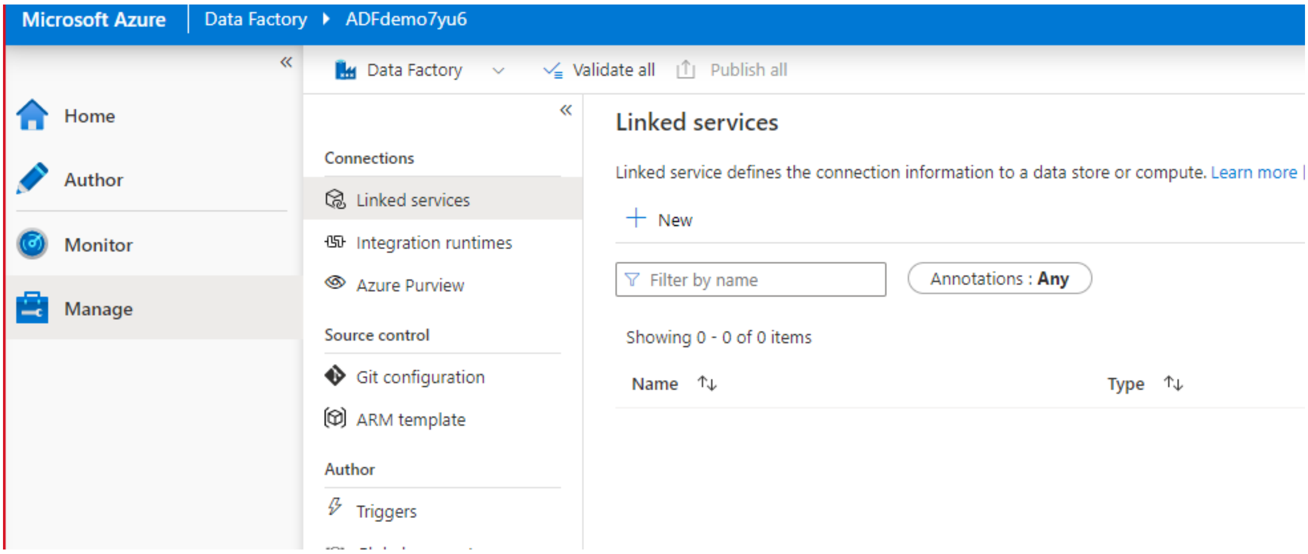


Authentication method: Access key (Switch to Azure AD User Account)
Location: data / Input

Name	Modified	Access tier	Archive status
☐ [..]			
☐ empdata.csv	5/5/2022, 9:58:04 AM	Hot (Inferred)	

6. Exercise 5 – Create Linked Services

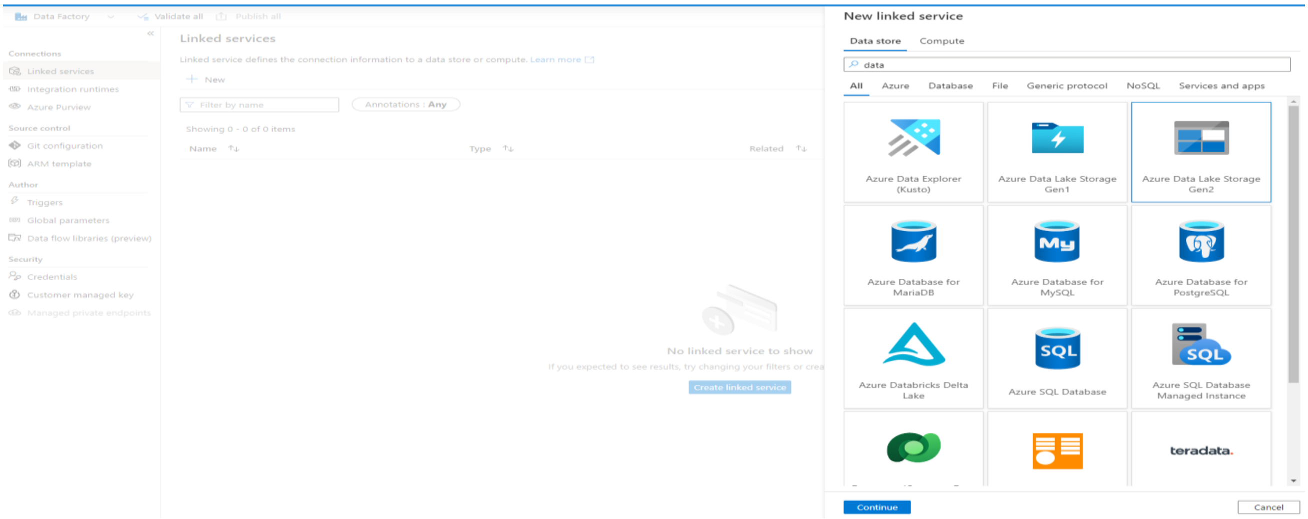
1. On the Azure Data Factory UI page, open **Manage** tab from the left pane.
2. On the Linked services page, select **+New** to create a new linked service.



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

Showing 0 - 0 of 0 items

3. On the **New Linked Service** page, select **Data Lake Verion2**, and then select **Continue**.



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

Showing 0 - 0 of 0 items

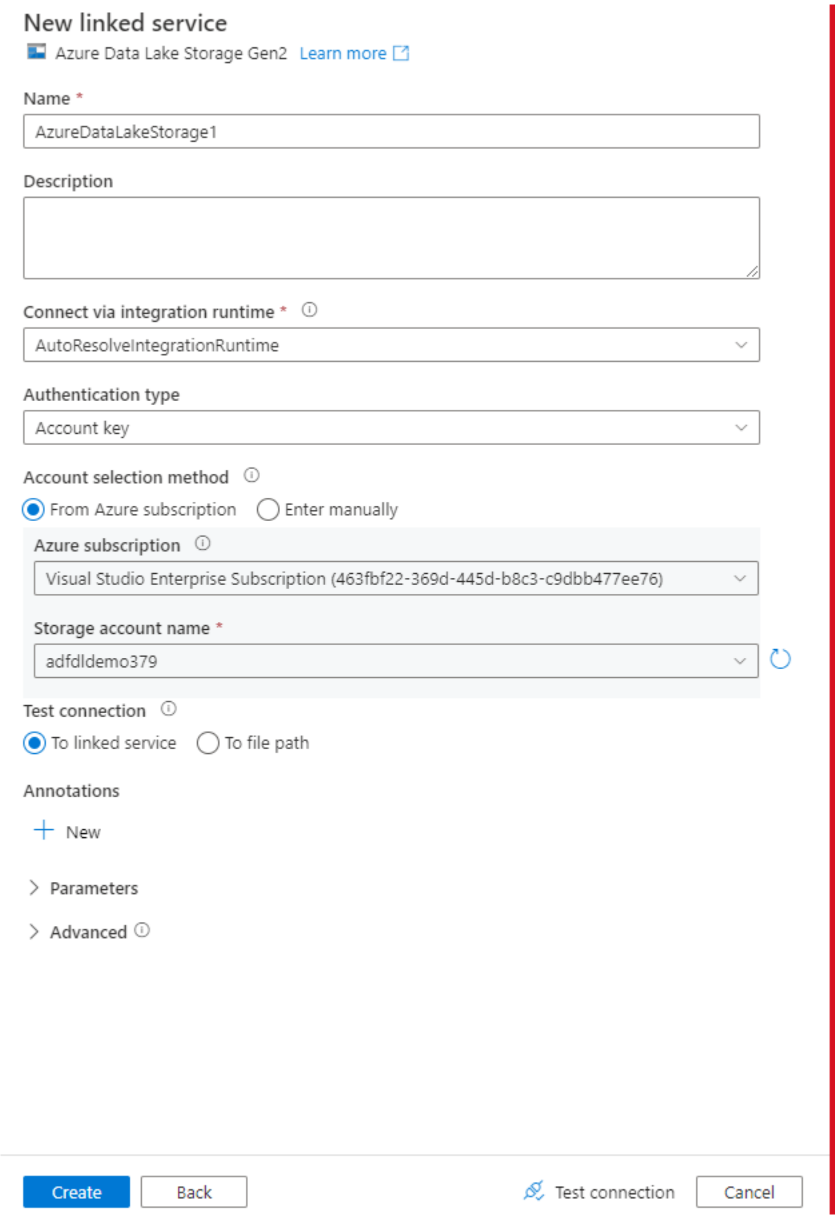
No linked service to show

If you expected to see results, try changing your filters or crea

4. On the New Linked Service (Azure Blob Storage) page, complete the following steps:

a. For **Name**, enter **AzureStorageLinkedService**.

b. For **Storage account name**, select the name of your Azure Storage account.



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

Visual Studio Enterprise Subscription (463fbf22-369d-445d-b8c3-c9dbb477ee76)

Storage account name *

adfdldemo379

Test connection

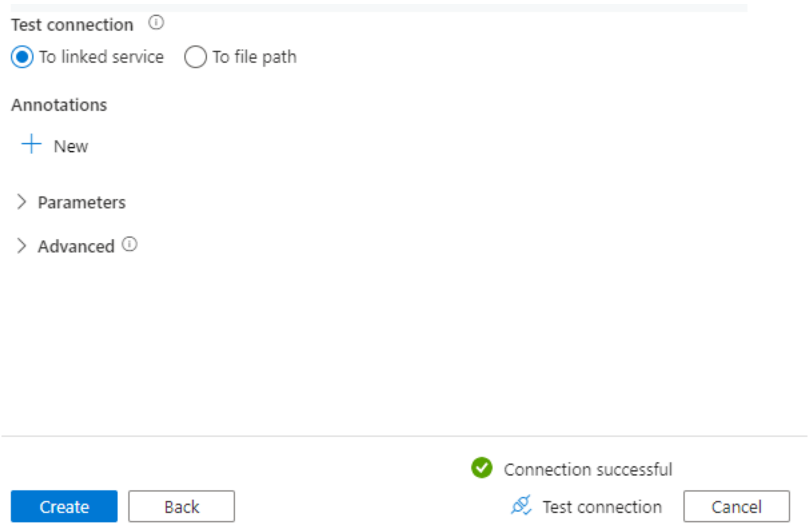
● To linked service ○ To file path

Annotations

+ New

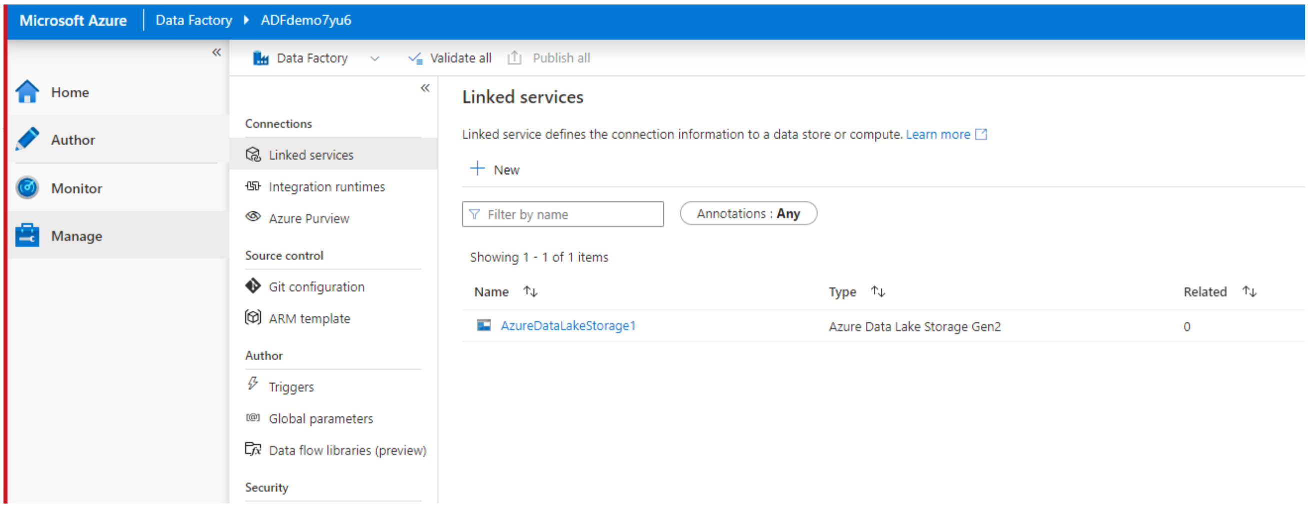
5. Select **Test connection** to confirm that the Data Factory service can connect to the storage account.

6. Select **Create** to save the linked service.


Test connection

● To linked service ○ To file path

Connection successful


Linked services

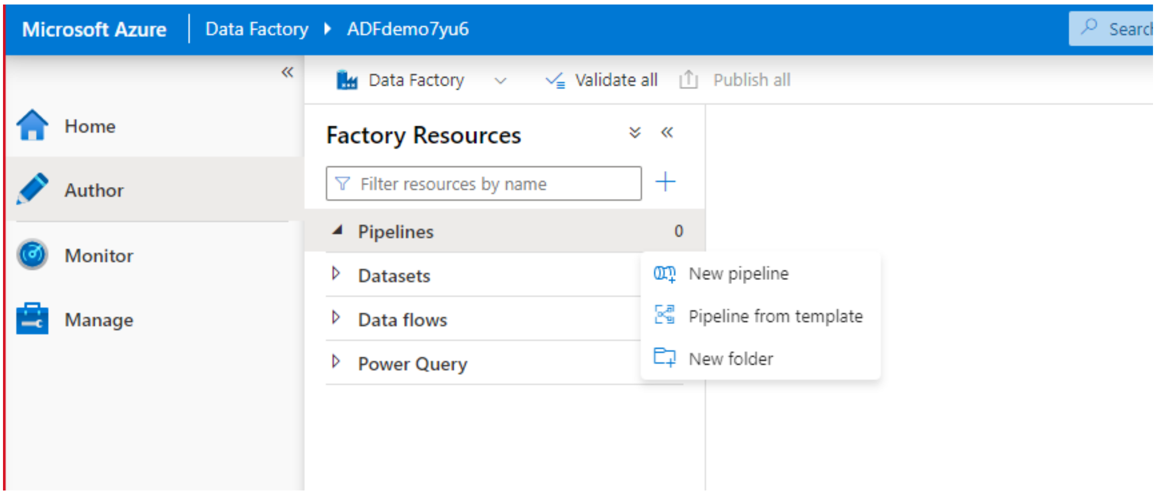
Linked service defines the connection information to a data store or compute. Learn more

Showing 1 - 1 of 1 items

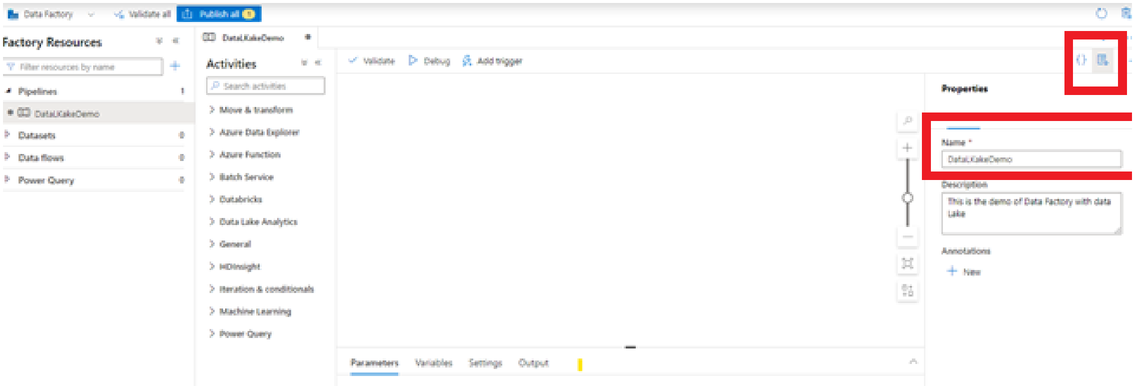
Name	Type	Related
AzureDataLakeStorage1	Azure Data Lake Storage Gen2	0

7. Exercise 1 – Create Pipeline and Dataset

1. Select the **+** (plus) button, and then select **Pipeline**.



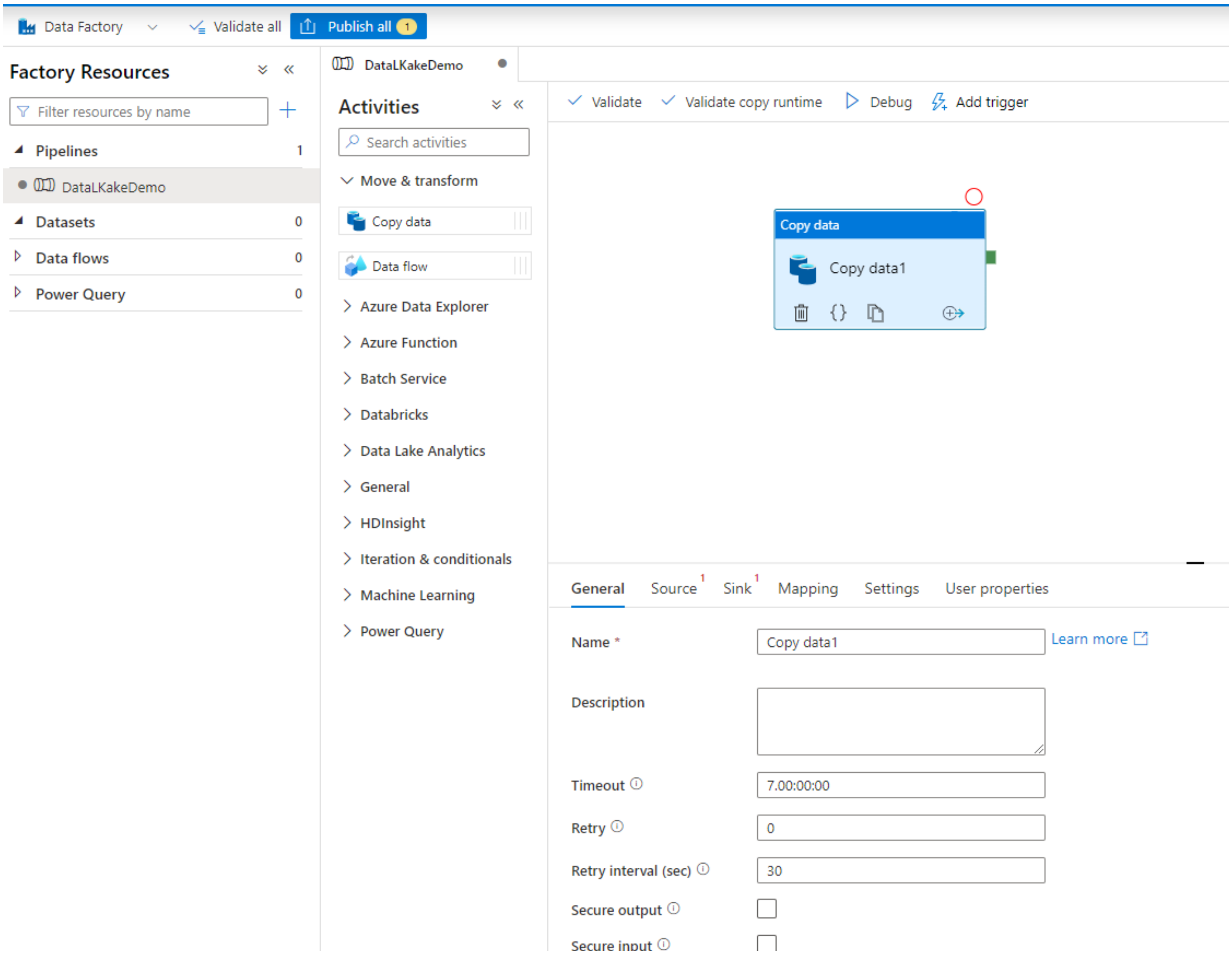
2. In the General panel under **Properties**, specify **CopyPipeline** for **Name**. Then collapse the panel by clicking the Properties icon in the top-right corner.



Name * DataLKakeDemo

Description This is the demo of Data Factory with data Lake

3. In the **Activities** toolbox, expand **Move & Transform**. Drag the **Copy Data** activity from the **Activities** toolbox to the pipeline designer surface.



Name * Copy data1

Description

Timeout 7.00:00:00

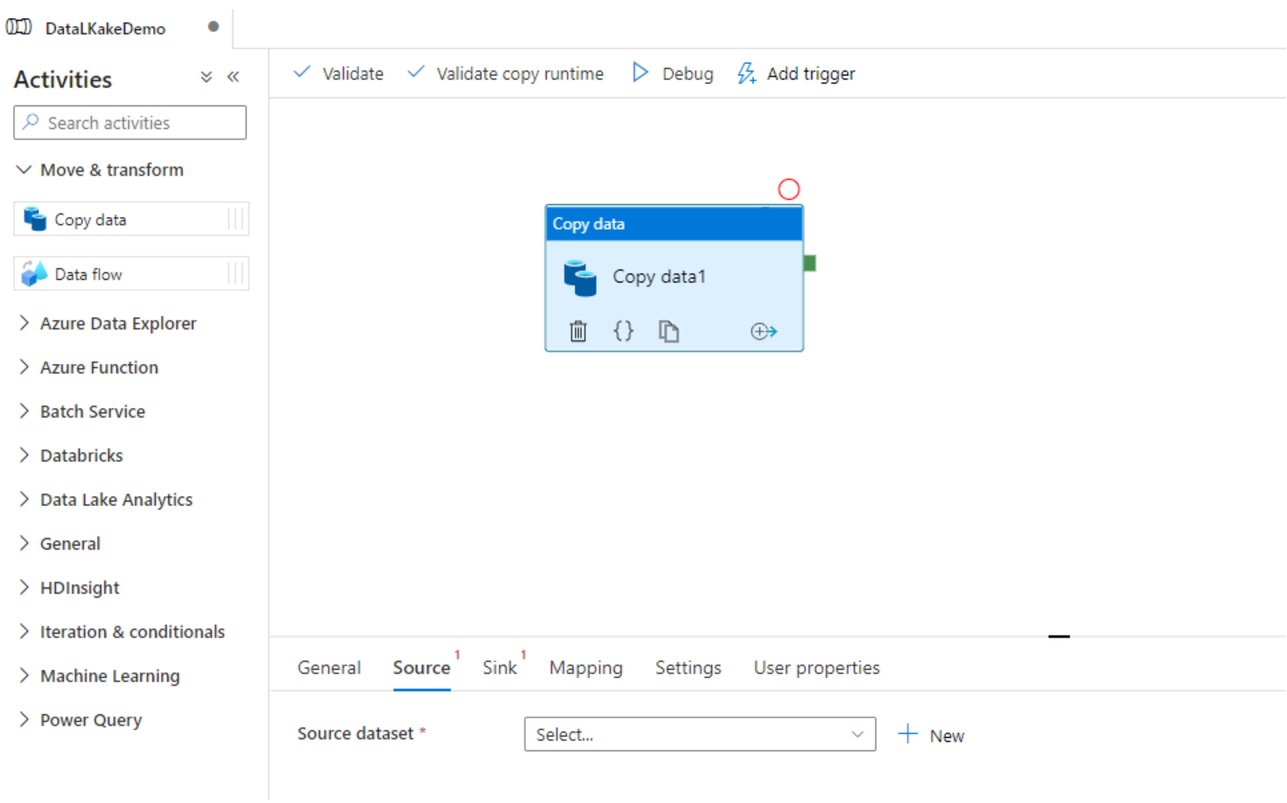
Retry 0

Retry interval (sec) 30

Secure output □

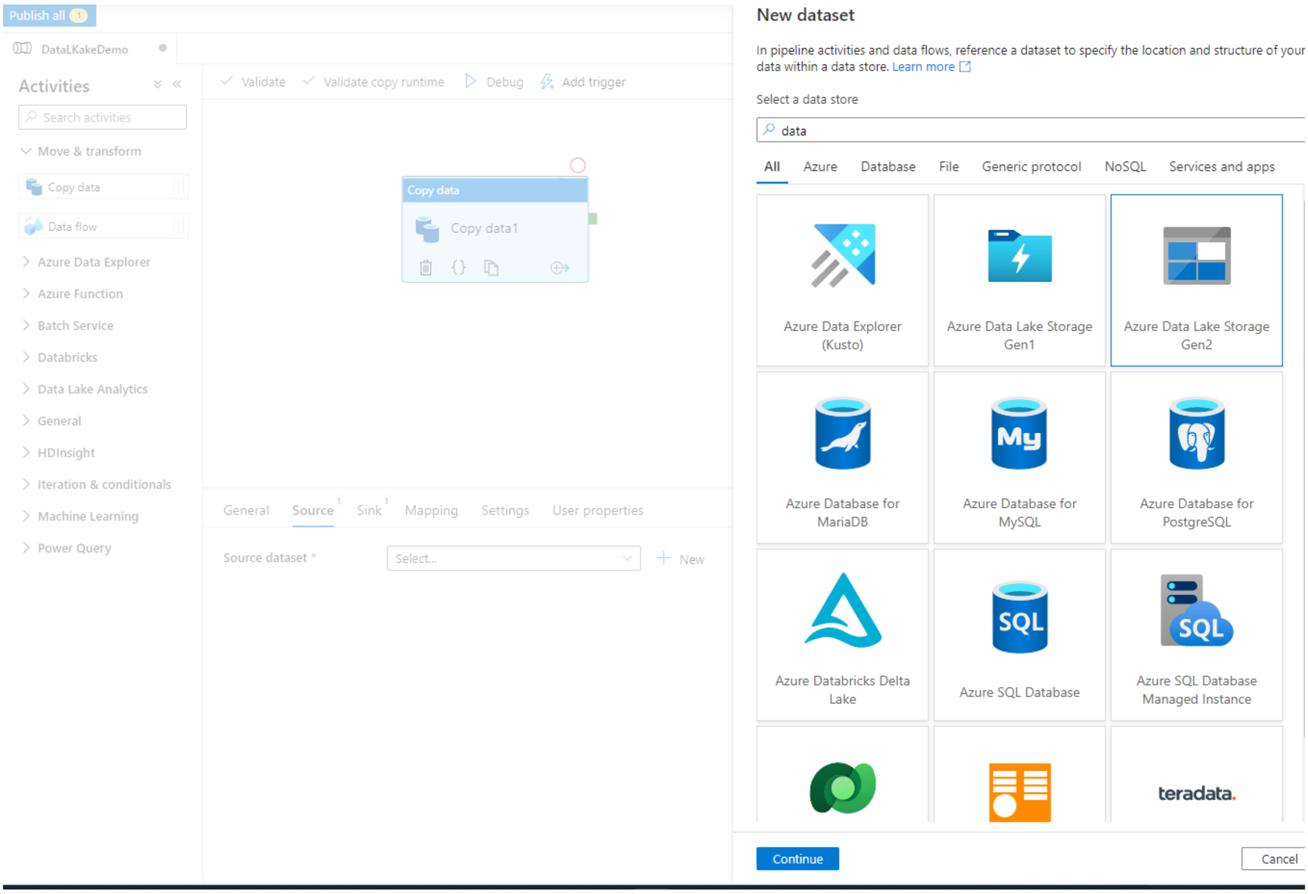
Secure input □

4. Switch to the **Source** tab in the copy activity settings, and Click +New



Source dataset * Select...

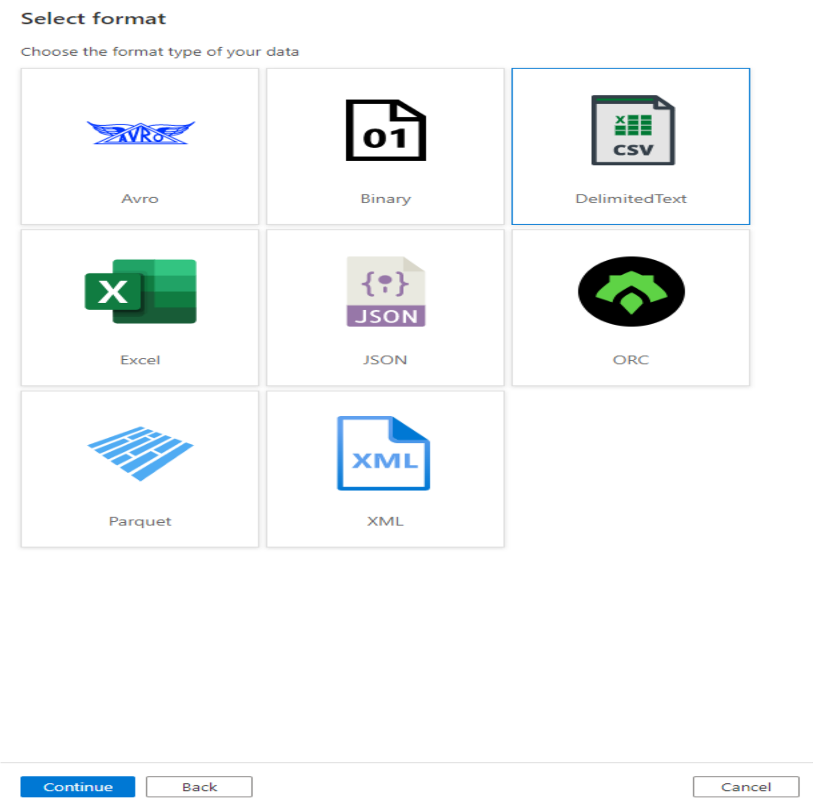
5. On the **New Dataset** page, select **Azure Data Lake Storage V2**, and then select **Continue**.



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: data

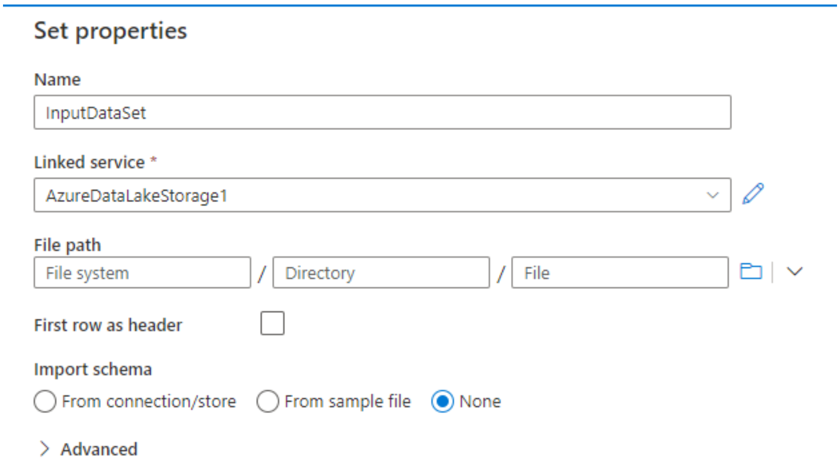
6. On the **Select Format** page, choose the format type of your data, and then select **Continue**. In this case, select **CSV** when copy files as-is without parsing the content.



7. On the **Set Properties** page, complete following steps:

a. Under **Name**, enter **InputDataset**.

b. For **Linked service**, select **AzureStorageLinkedService**.



Set properties

Name
InputDataSet

Linked service *
AzureDataLakeStorage1

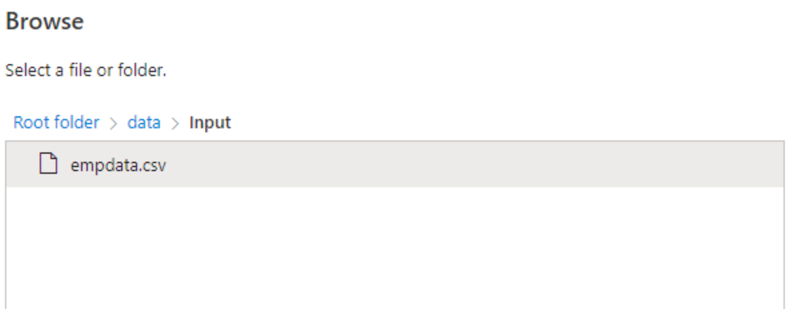
File path
File system / Directory / File

First row as header □

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

> Advanced

8. . For **File path**, select the **Browse** button.
9. In the **Choose a file or folder** window, browse to the **input** folder in the **data** container, select the **empdata.csv**, and then select **OK**.



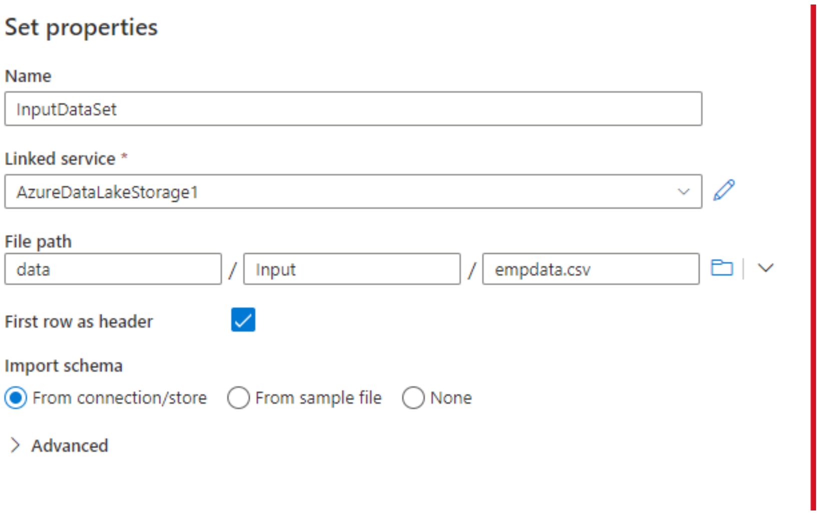
Browse

Select a file or folder.

Root folder > data > Input

empdata.csv

10. Select **OK**.



Set properties

Name: InputDataSet

Linked service *: AzureDataLakeStorage1

File path: data / Input / empdata.csv

First row as header ☑

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

11. Repeat the steps to create the output dataset:

a. Select the **+** (plus) button, and then select **Dataset**.

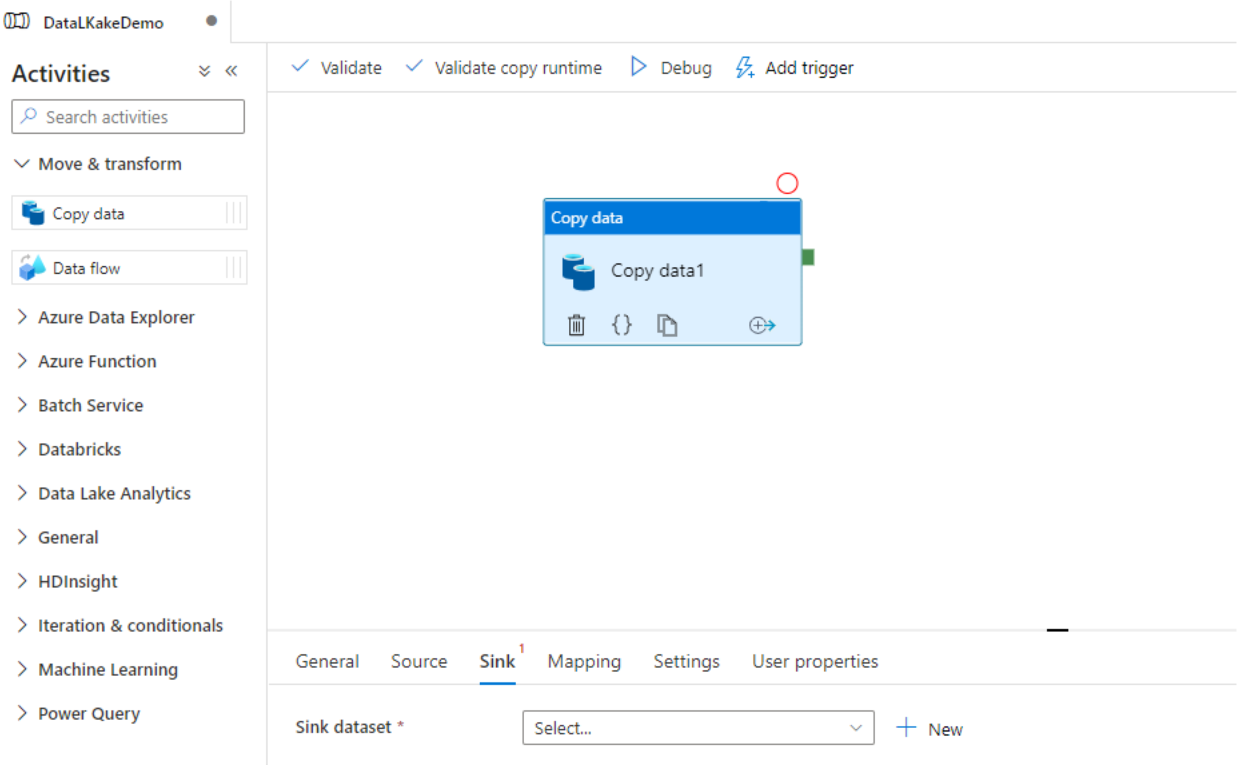
b. On the **New Dataset** page, select **Azure Data Lake Storage V2**, and then select **Continue**.

c. On the **Select Format** page, choose the format type of your data, and then select **Continue**.

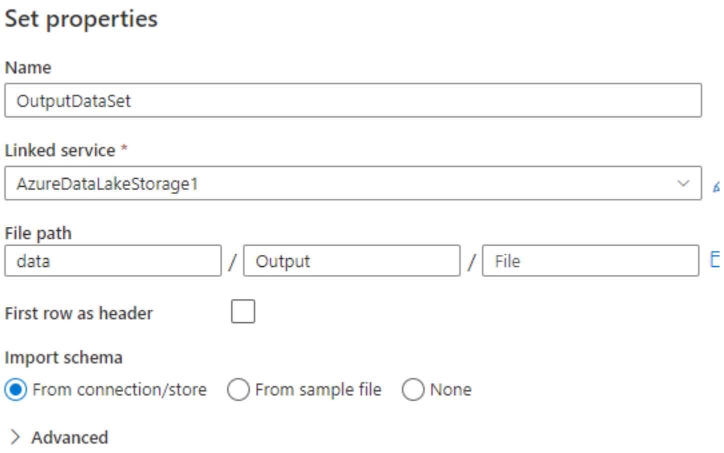
d. On the **Set Properties** page, specify **OutputDataset** for the name. Select **AzureStorageLinkedService** as linked service.

e. Under **File path**, enter **data/output**. If the **output** folder doesn't exist, the copy activity creates it at runtime.

f. Select **OK**.



Sink dataset * Select...



Set properties

Name

OutputDataSet

Linked service *

AzureDataLakeStorage1

File path

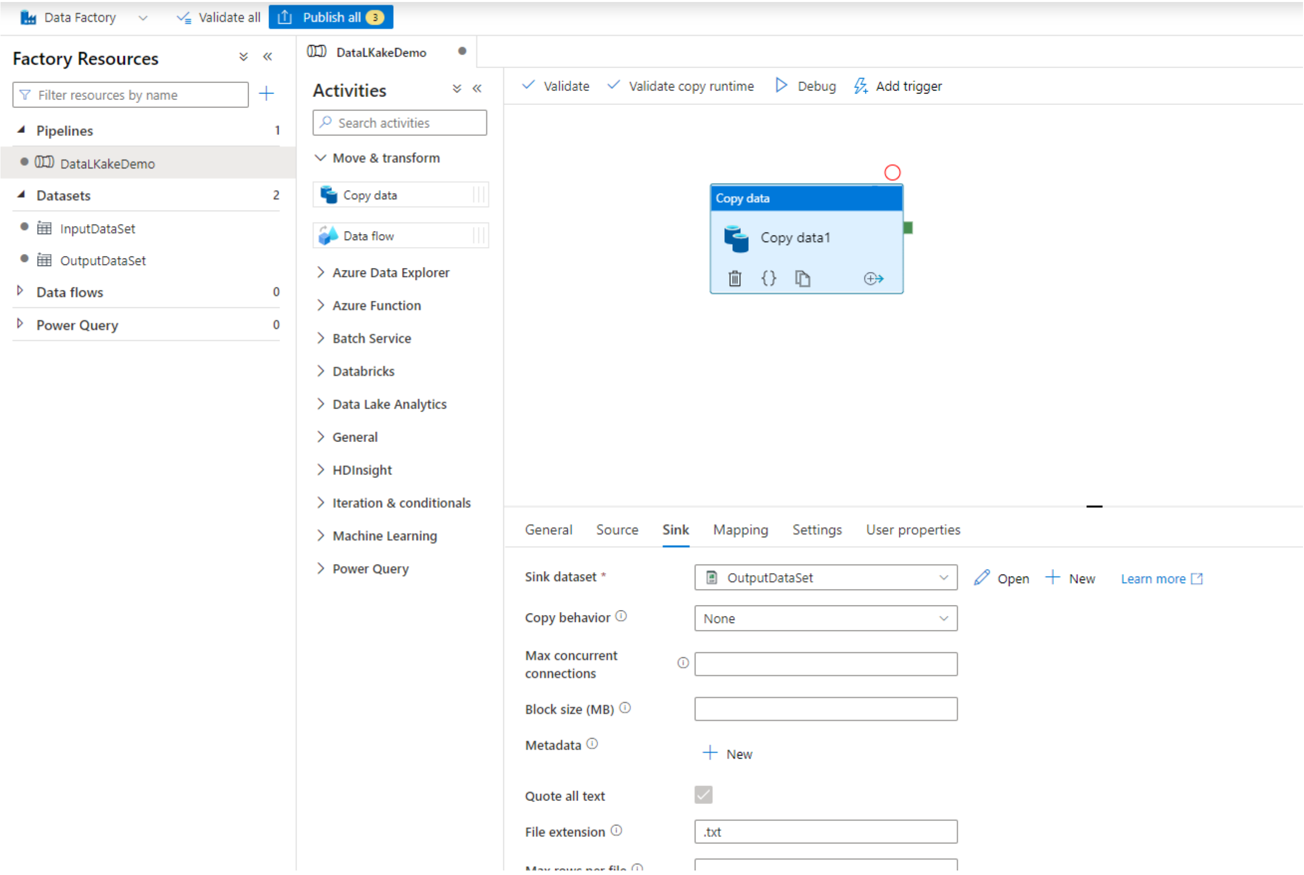
data / Output / File

First row as header □

Import schema

● From connection/store ○ From sample file ○ None

Advanced



Sink dataset * OutputDataSet

Copy behavior None

Max concurrent connections

Block size (MB)

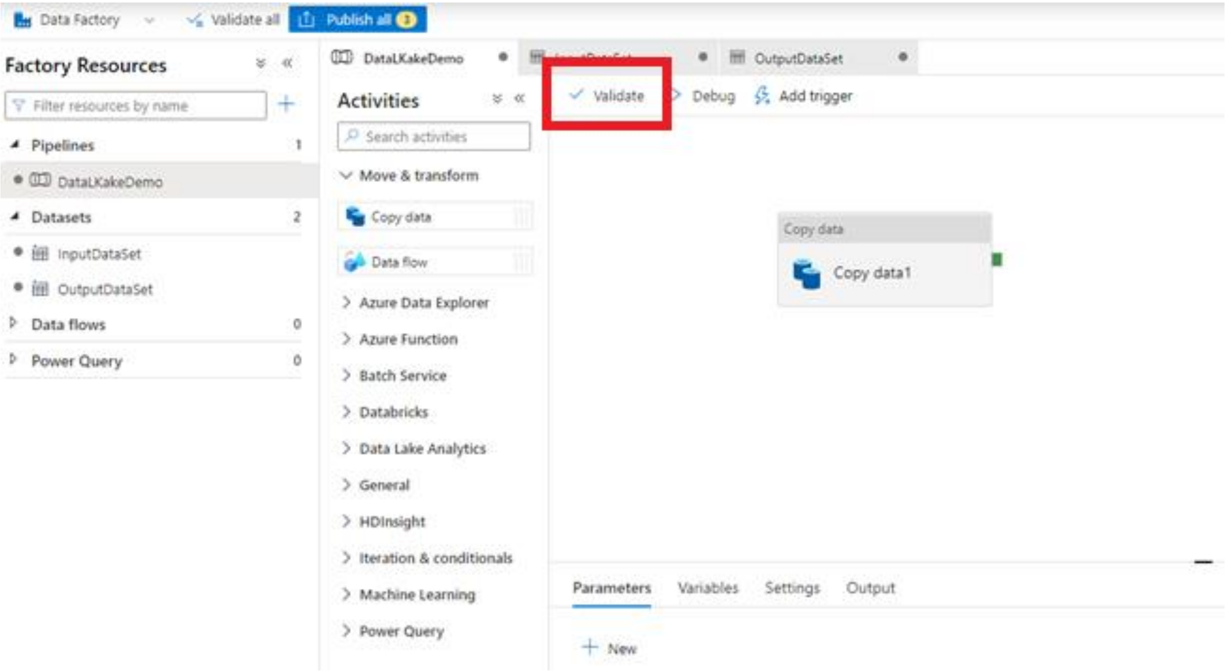
Metadata + New

Quote all text ☑

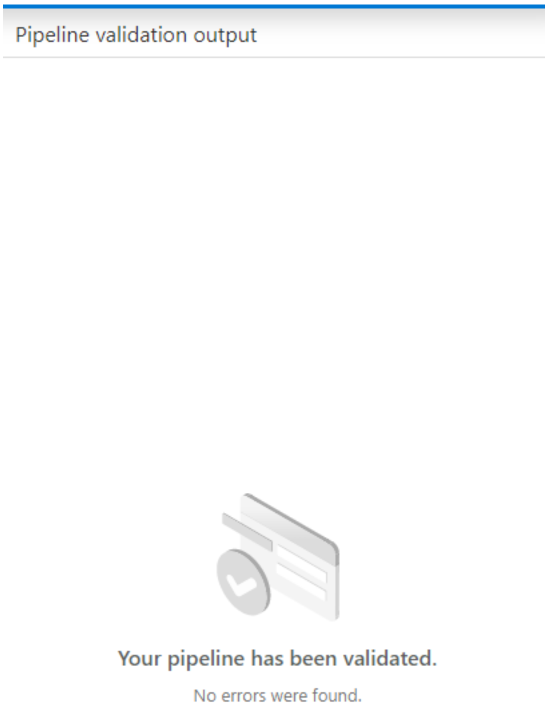
File extension .txt

8. Exercise 7 – Validate and Debug the Pipeline

1. In the Pipeline click **Validate** to Validate the Pipeline for any configuration error



2. It Should show the message Validation is Passed

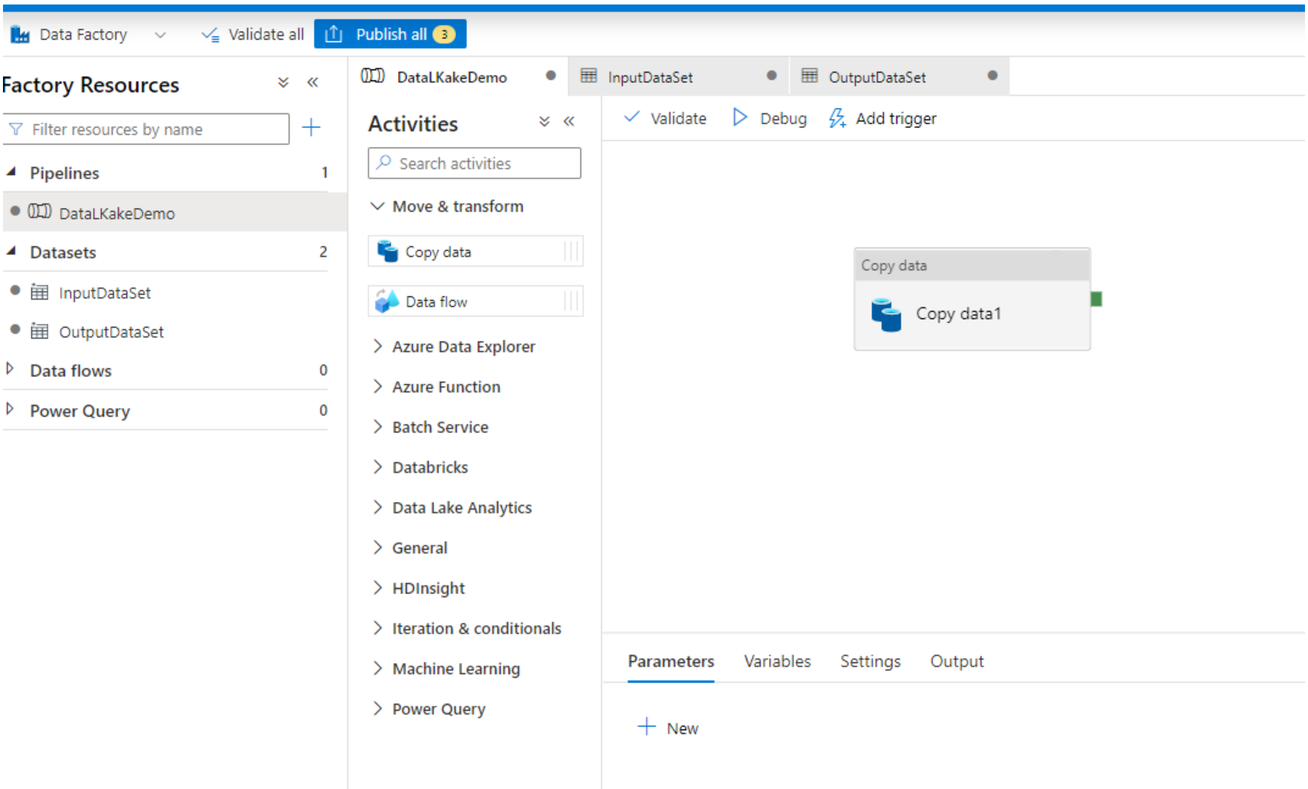


Pipeline validation output

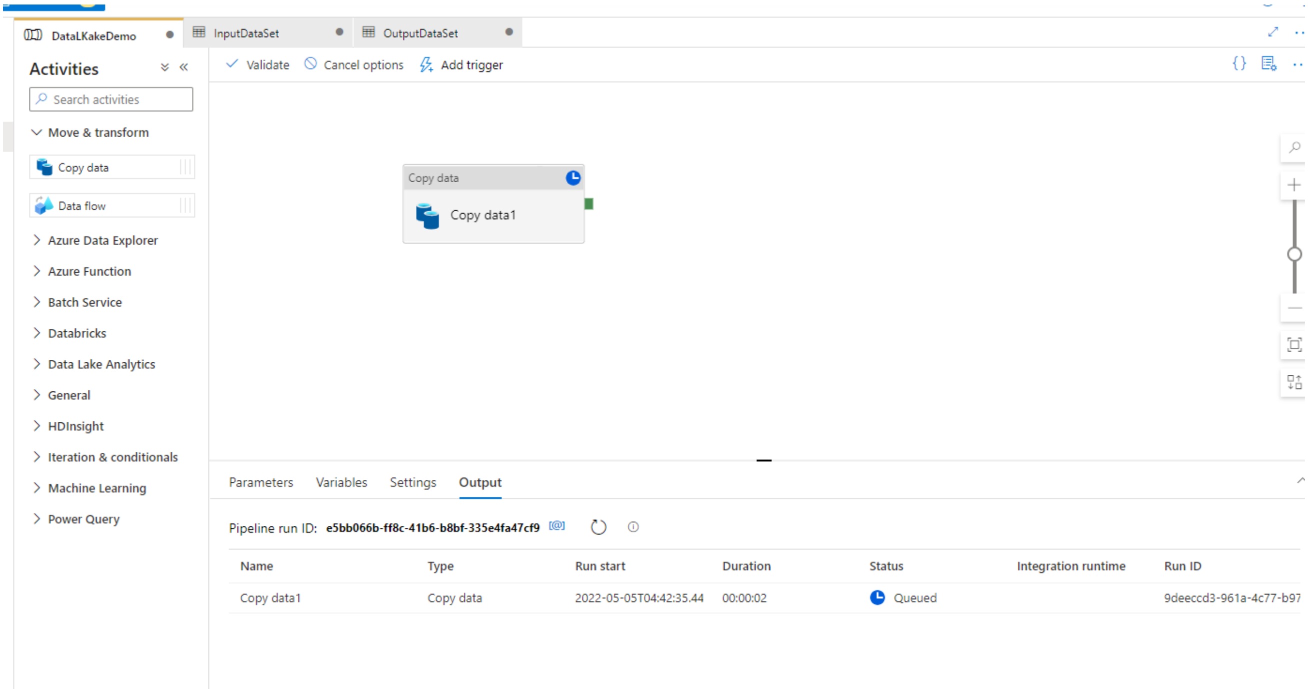
Your pipeline has been validated.

No errors were found.

3. On the pipeline toolbar above the canvas, click **Debug** to trigger a test run.



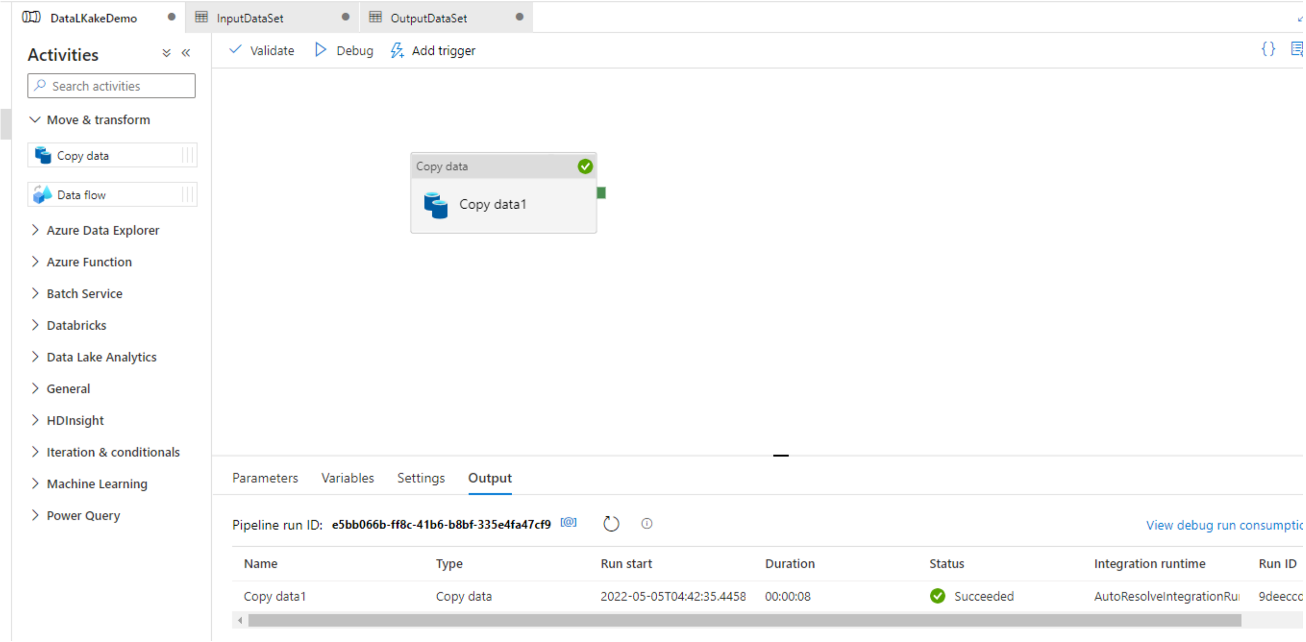
4. Confirm that you see the status of the pipeline run on the Output tab of the pipeline settings at the bottom.



Pipeline run ID: e5bb066b-ff8c-41b6-b8bf-335e4fa47cf9

Name	Type	Run start	Duration	Status	Integration runtime	Run ID
Copy data1	Copy data	2022-05-05T04:42:35.44	00:00:02	Queued		9deeccd3-961a-4c77-b97

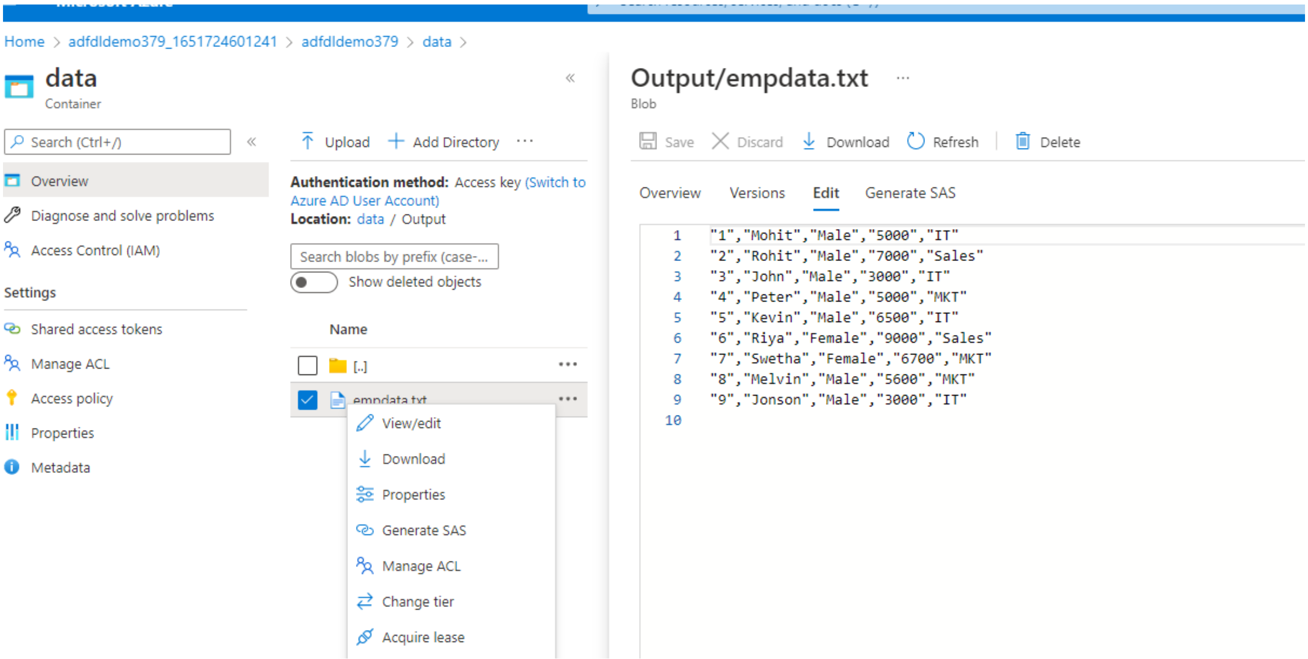
5. Confirm that you see that you succeed Message .



Pipeline run ID: e5bb066b-ff8c-41b6-b8bf-335e4fa47cf9

Name	Type	Run start	Duration	Status	Integration runtime	Run ID
Copy data1	Copy data	2022-05-05T04:42:35.4458	00:00:08	Succeeded	AutoResolveIntegrationRu	9deeccc

5. Go back to your storage and check output folder .

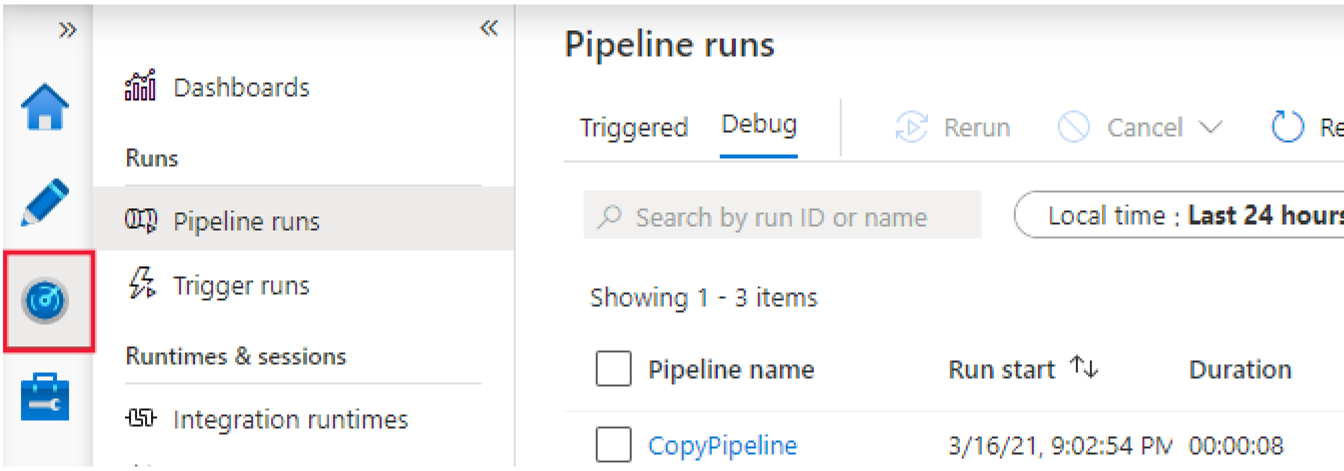


Authentication method: Access key (Switch to Azure AD User Account)
Location: data / Output

```
"1","Mohit","Male","5000","IT"
"2","Rohit","Male","7000","Sales"
"3","John","Male","3000","IT"
"4","Peter","Male","5000","MKT"
"5","Kevin","Male","6500","IT"
"6","Riya","Female","9000","Sales"
"7","Swetha","Female","6700","MKT"
"8","Melvin","Male","5600","MKT"
"9","Jonson","Male","3000","IT"
```

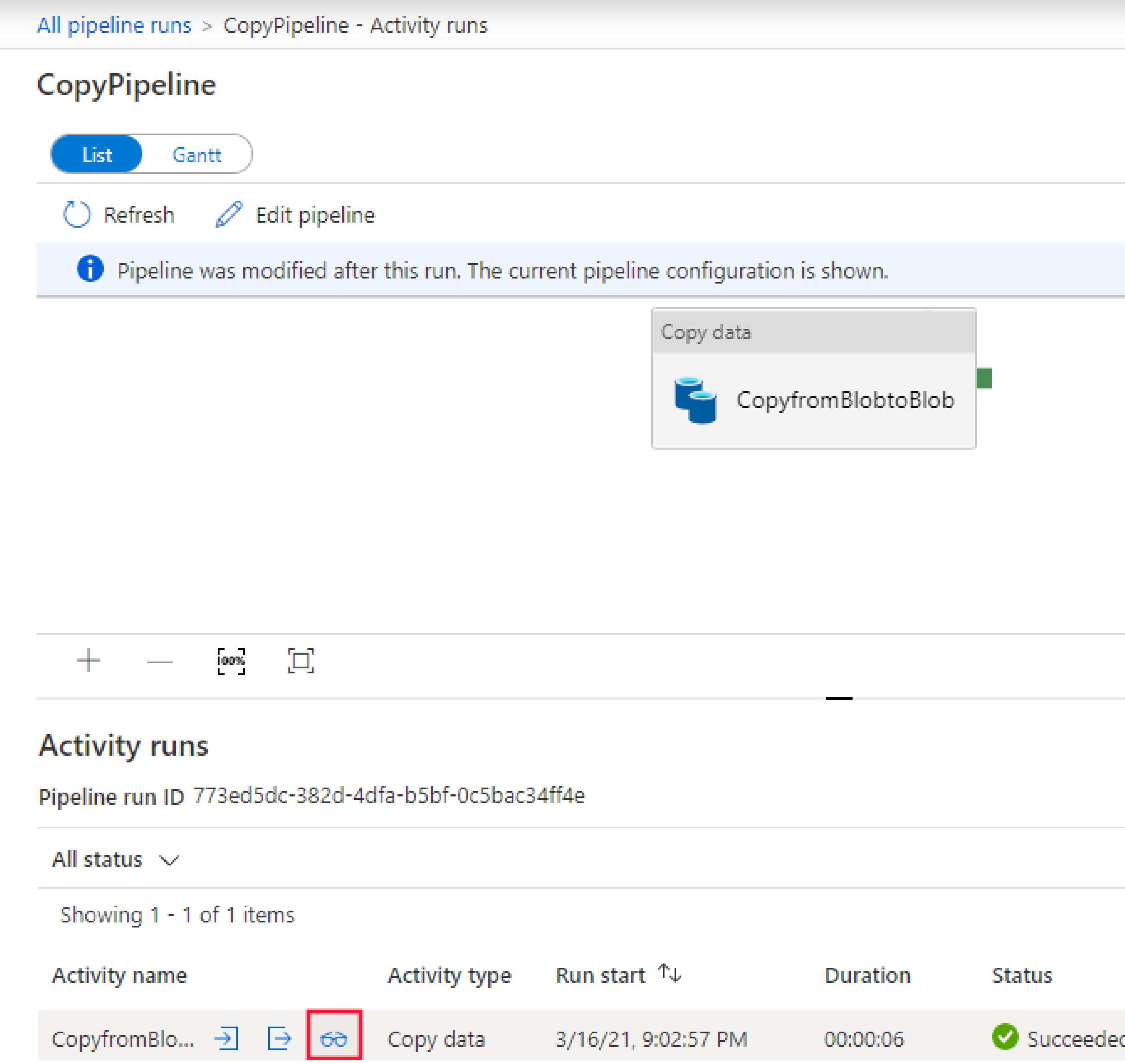
9. Monitor the pipeline

1. Switch to the **Monitor** tab on the left. Use the **Refresh** button to refresh the list.



□	Pipeline name	Run start	Duration
□	CopyPipeline	3/16/21, 9:02:54 PM	00:00:08

2. Select the **CopyPipeline** link, you'll see the status of the copy activity run on this page.
3. To view details about the copy operation, select the **Details** (eyeglasses image) link. For details about the properties, see Copy Activity overview.



Pipeline was modified after this run. The current pipeline configuration is shown.

Activity runs

Pipeline run ID 773ed5dc-382d-4dfa-b5bf-0c5bac34ff4e

All status

Showing 1 - 1 of 1 items

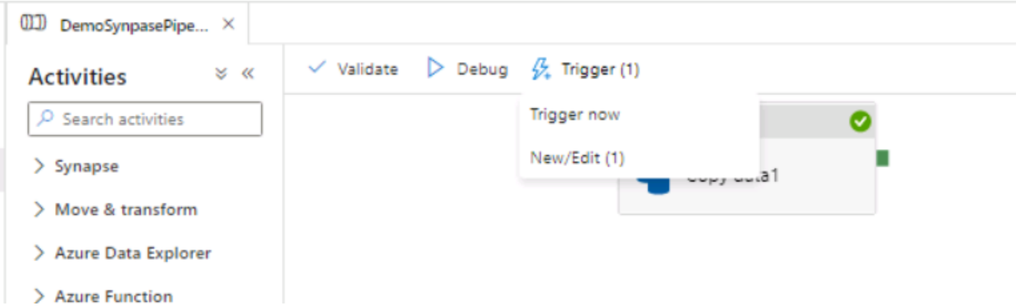
Activity name	Activity type	Run start	Duration	Status
CopyfromBlo...	Copy data	3/16/21, 9:02:57 PM	00:00:06	✓ Succeeded

4. Confirm that you see a new file in the **output** folder.
5. You can switch back to the **Pipeline runs** view from the **Activity runs** view by selecting the **All pipeline runs** link.

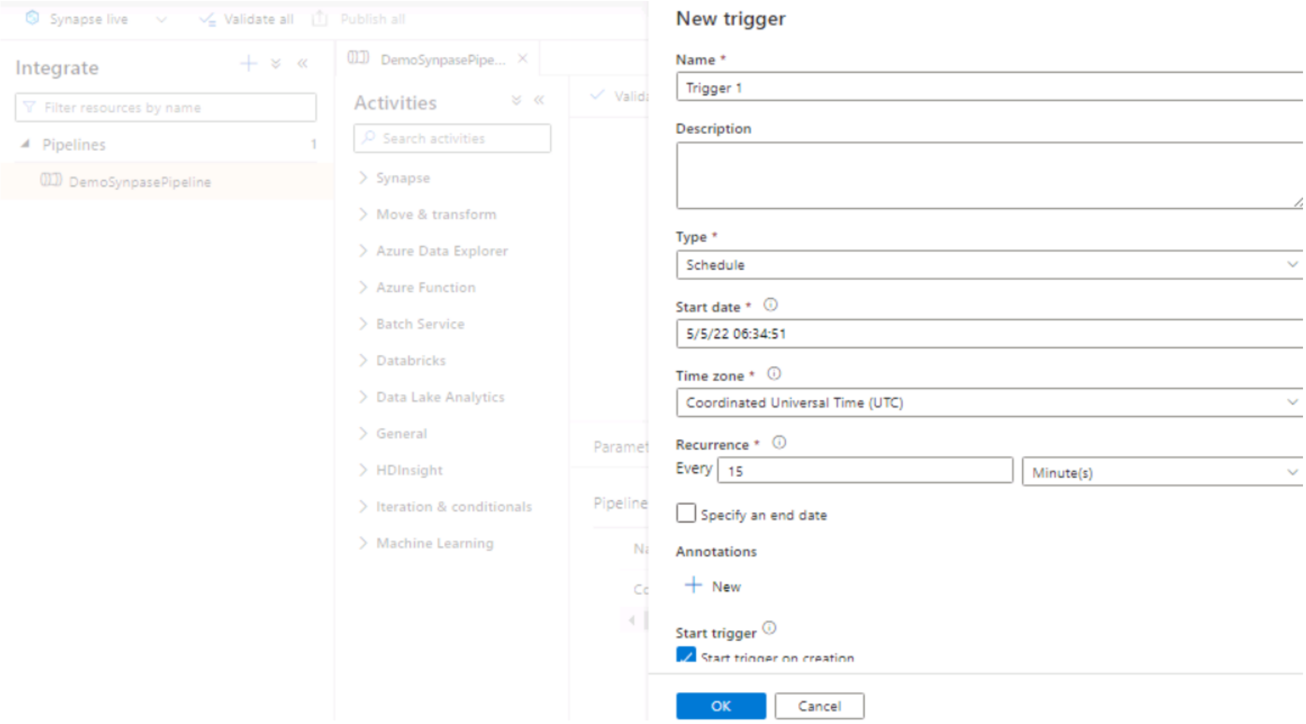
10. Trigger the pipeline on a schedule

1. Switch to the **Author** tab.

2. Go to your pipeline, select **Add Trigger** on the pipeline toolbar, and then select **New/Edit**.



3. On the **Add Triggers** page, select **Choose trigger**, and then select **New**.
4. On the **New Trigger** page, under **End**, select **On Date**, specify an end time a few minutes after the current time, and then select **OK**.



Name * Trigger 1

Type * Schedule

Start date * 5/5/22 06:34:51

Time zone * Coordinated Universal Time (UTC)

Recurrence * Every 15 Minute(s)

☐ Specify an end date

Start trigger

☑ Start trigger on creation