

**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 3](#_Toc102722629)

[2. Exercise 1 – Provisioning Azure Data Factory 3](#_Toc102722630)

[3. Exercise 2 – Launch Azure Data Factory Studio 7](#_Toc102722631)

[4. Exercise 3 – Create a New Data Lake Storage 8](#_Toc102722632)

[5. Exercise 4 – Create Data Lake Container and Upload Data 10](#_Toc102722633)

[6. Exercise 5 – Create Linked Services 13](#_Toc102722634)

[7. Exercise 1 – Create Pipeline and Dataset 15](#_Toc102722635)

[8. Exercise 7 – Validate and Debug the Pipeline 21](#_Toc102722636)

[9. Monitor the pipeline 24](#_Toc102722643)

[10. Trigger the pipeline on a schedule 25](#_Toc102722644)

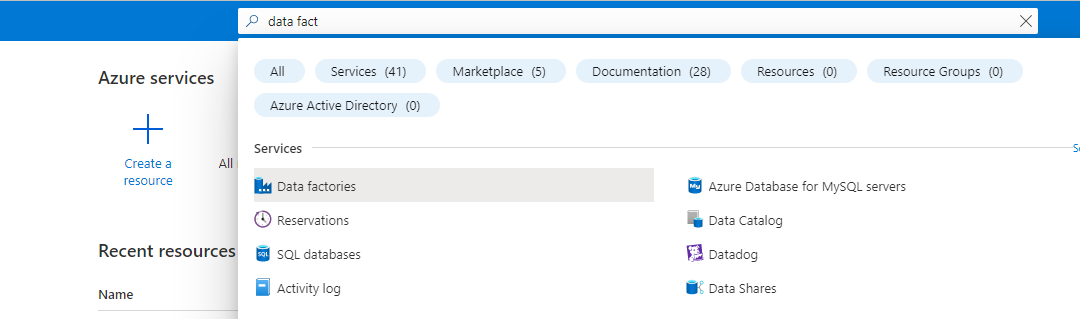
# Introduction

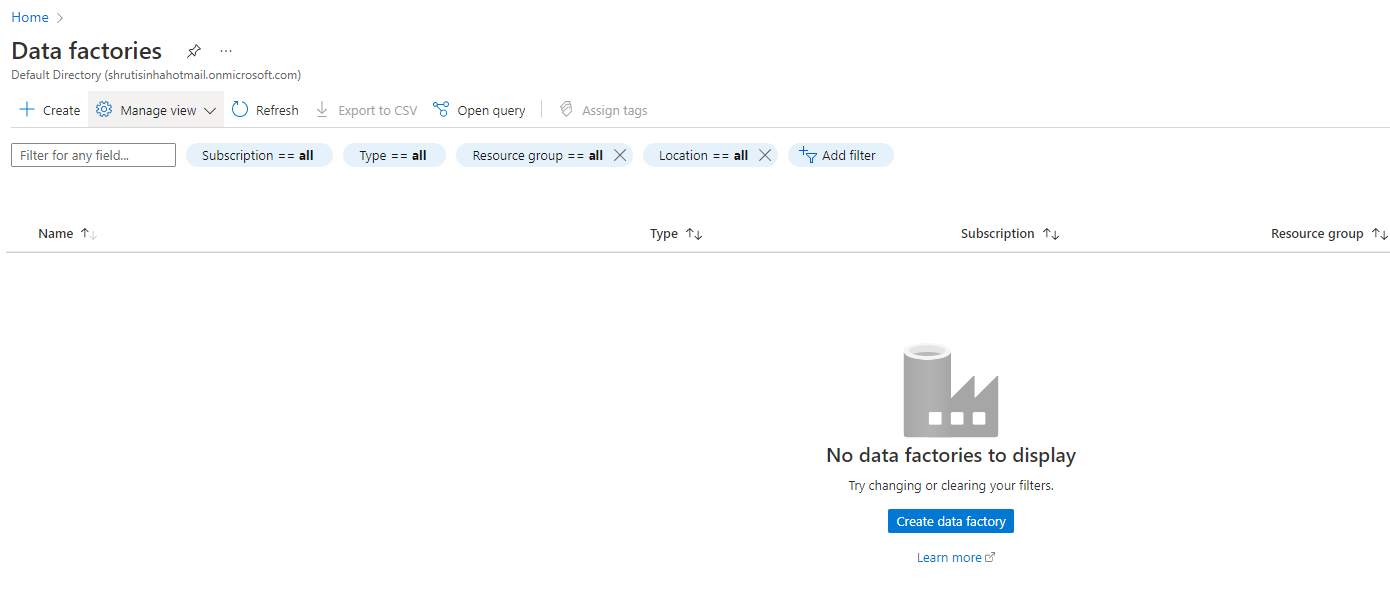
Azure Data Factory (ADF) is a data pipeline orchestrator and ETL tool that is part of the Microsoft Azure cloud ecosystem. ADF can pull data from the outside world (FTP, Amazon S3, Oracle), transform it, filter it, enhance it, and move it along to another destination. In my work for a health-data project 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](https://intellipaat.com/blog/tutorial/spark-tutorial/) or [Hadoop](https://intellipaat.com/blog/tutorial/hadoop-tutorial/).

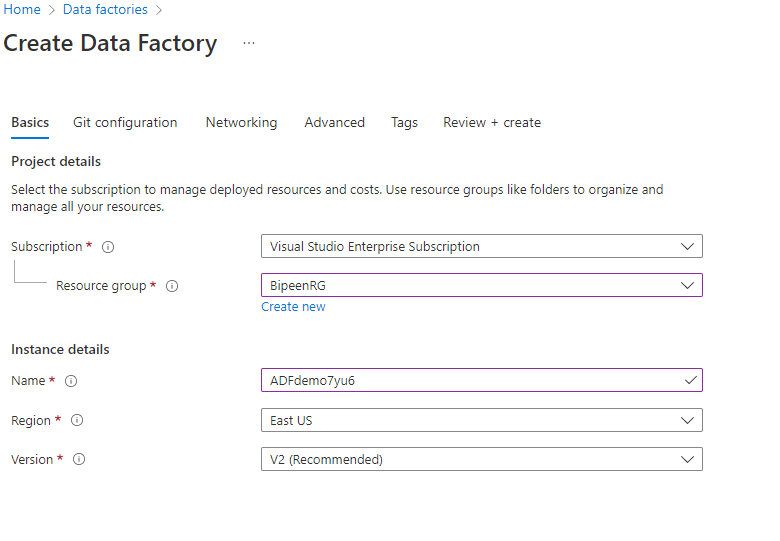
In this Azure Data Factory tutorial, you will learn about Azure Data Factory, its basic concepts and why do we need it. Also, you will learn the working process of Azure Data Factory and will be introduced to Azure Data Lake. Here, you will learn how to copy data from Azure SQL to Azure Data Lake,

# Exercise 1 – Provisioning Azure Data Factory

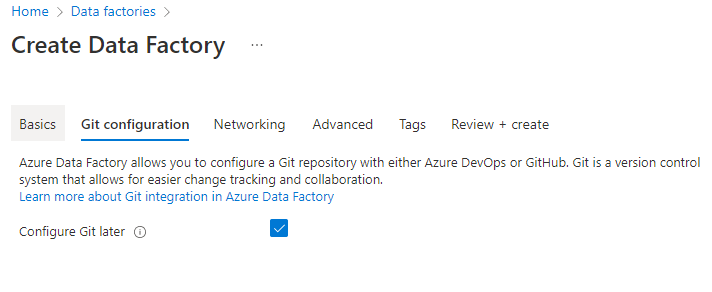
1. Go to the [Azure portal](https://portal.azure.com/).
2. In the Azure portal search type Data Factory 
3. Click  **Create Data Factory** page,



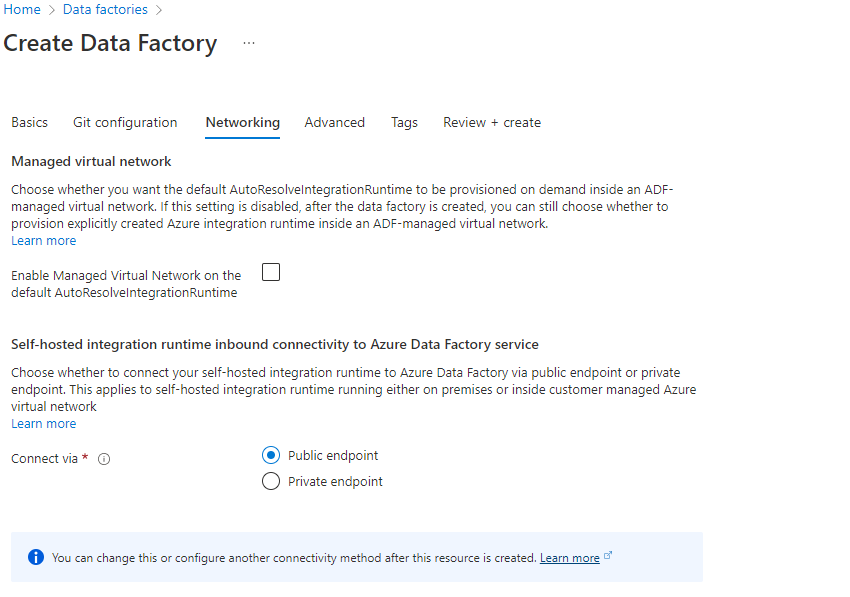
1. Under **Basics** tab, select your Azure **Subscription** in which you want to create the data factory.
2. For **Resource Group** Select an existing resource group from the drop-down list.
3. For **Name**, enter **ADFDemo+number**



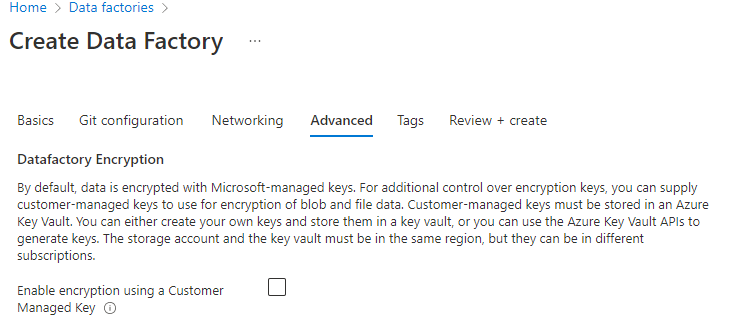
1. Select **Next: Git configuration**, and then select **Configure Git later** check box.



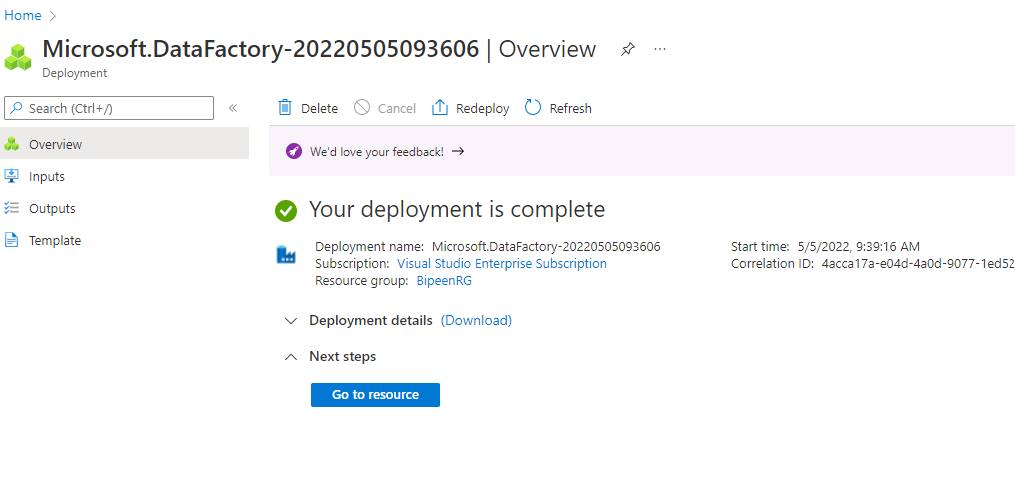
1. Click Next in Networking



1. Click Next in Advance

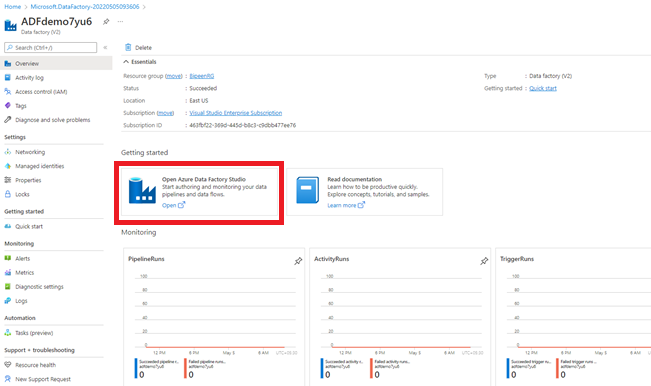


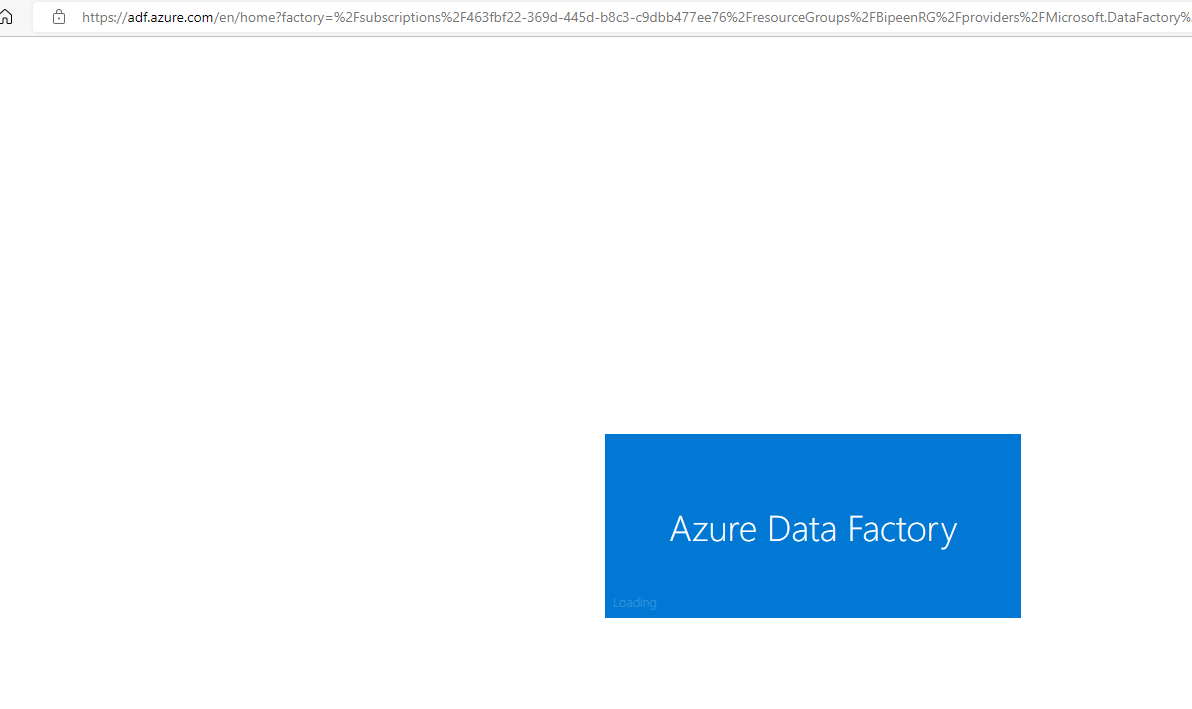
1. Select **Review + create**, and select **Create** after the validation is passed.



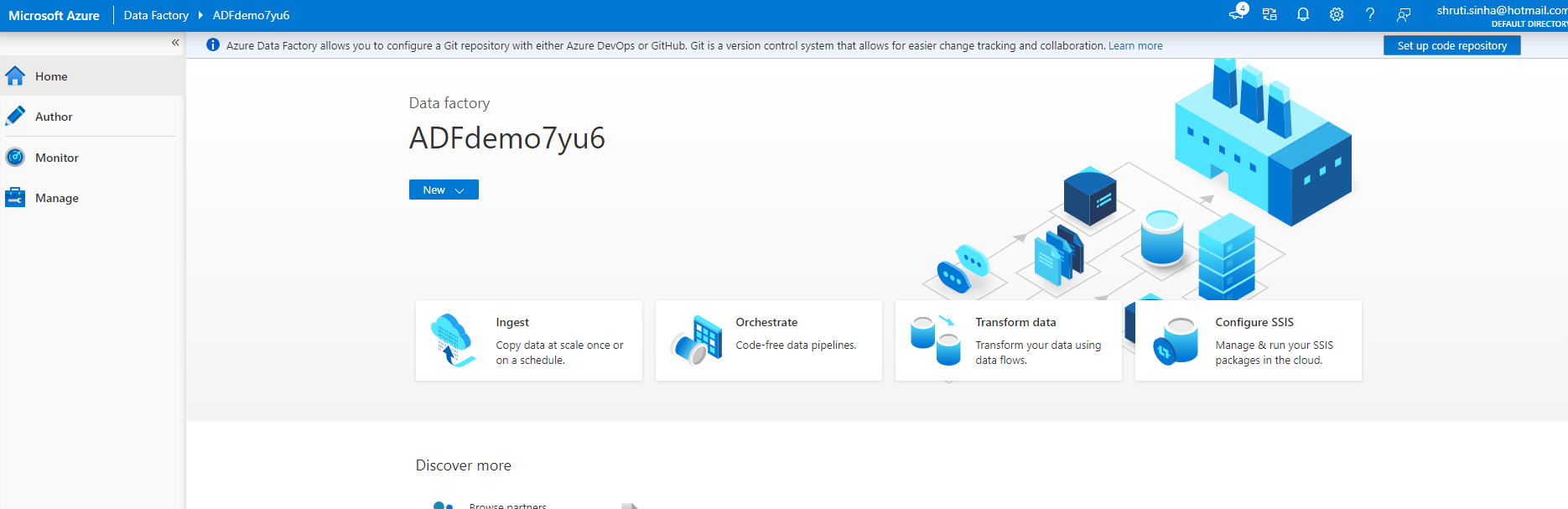
1. After the creation is complete, select **Go to resource** to navigate to the **Data Factory** page.

# Exercise 2 – Launch Azure Data Factory Studio

1. Select Open on the Open **Azure Data Factory Studio** . 
2. It start the Azure Data Factory user interface (UI) application on a separate browser tab.

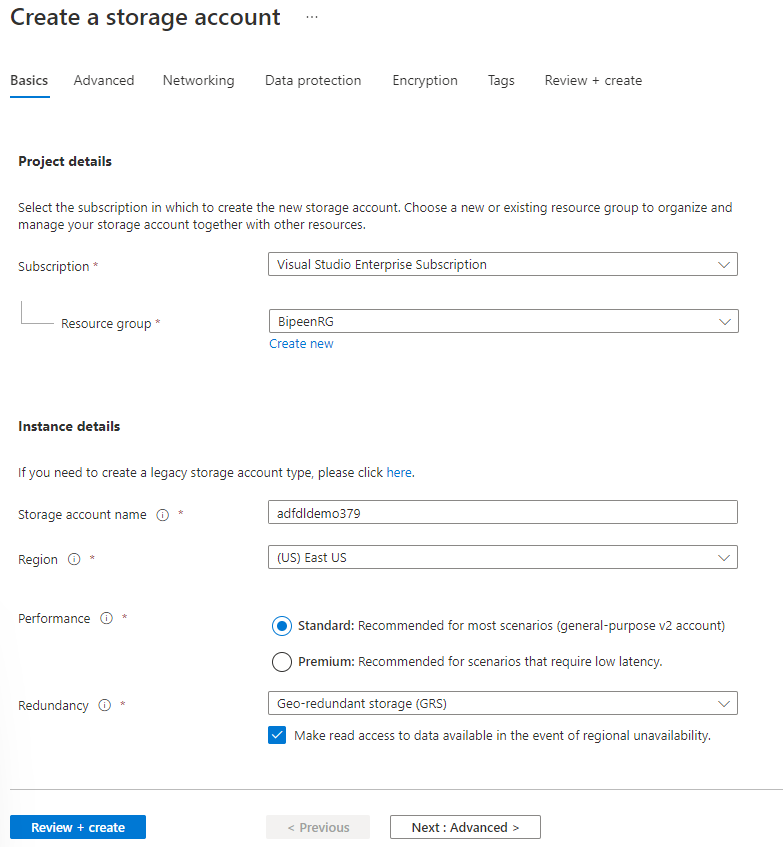


1. You Should be in Azure Data Factory Studio

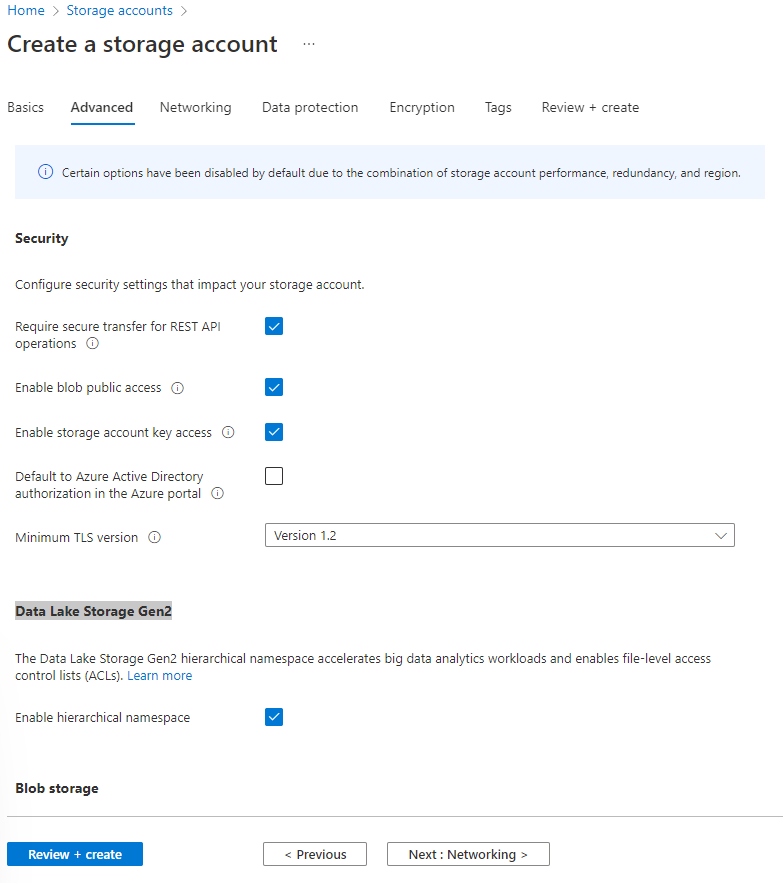


# Exercise 3 – Create a New Data Lake Storage

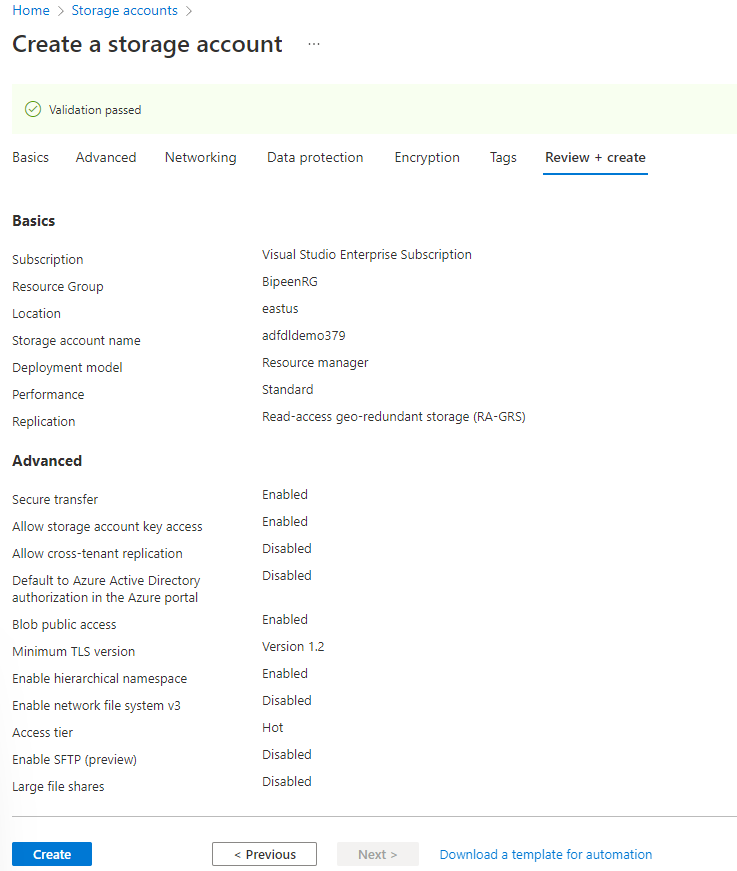
1. Create new storage as we did in our prevision exercise



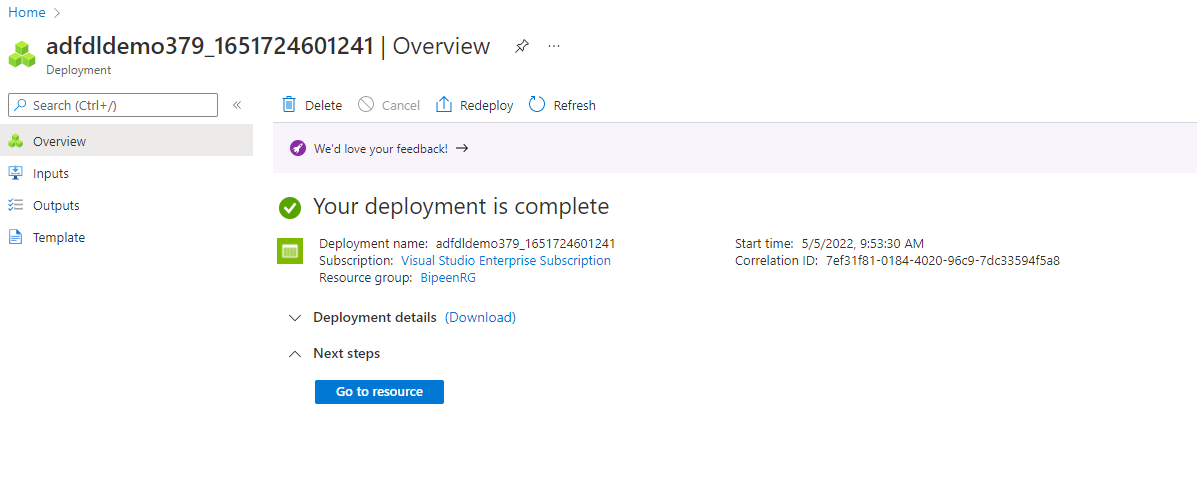
1. Enable Datalake and click Review and Create



1. Click **Create**

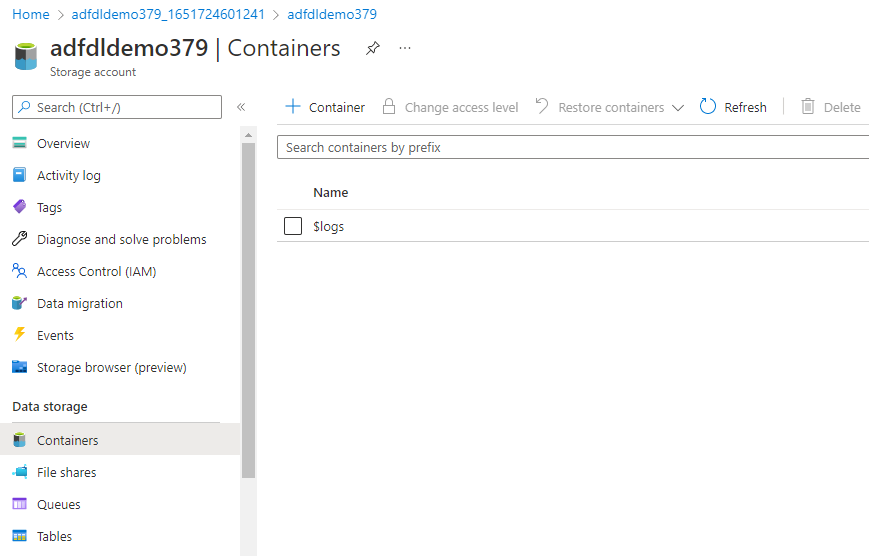


1. It should be created in few Minutes

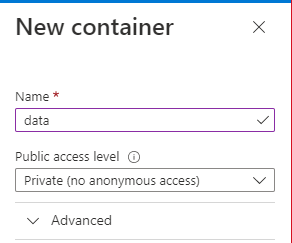


# Exercise 4 – Create Data Lake Container and Upload Data

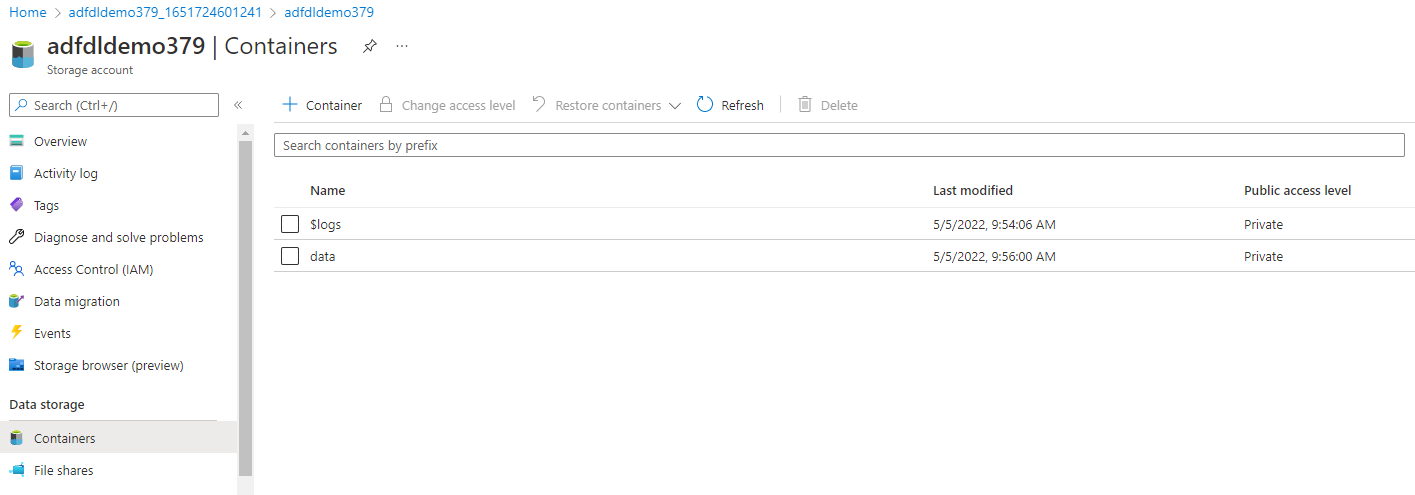
1. From the storage account page, select **Overview** > **Containers**.



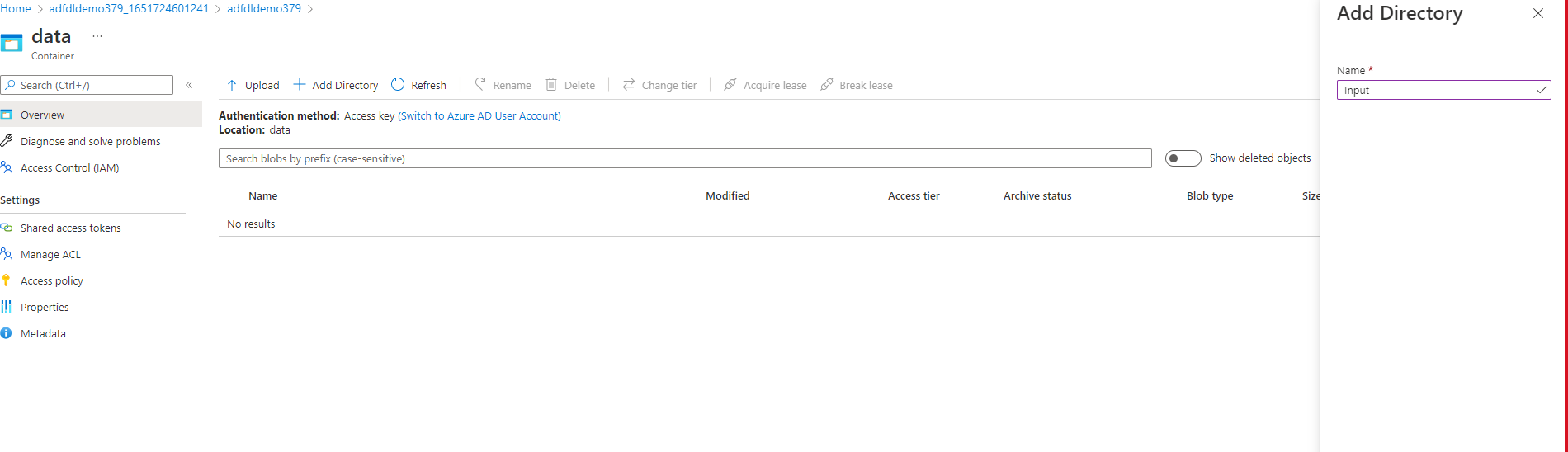
1. In the **New container** dialog box, enter **data** for the name, and then select **OK**



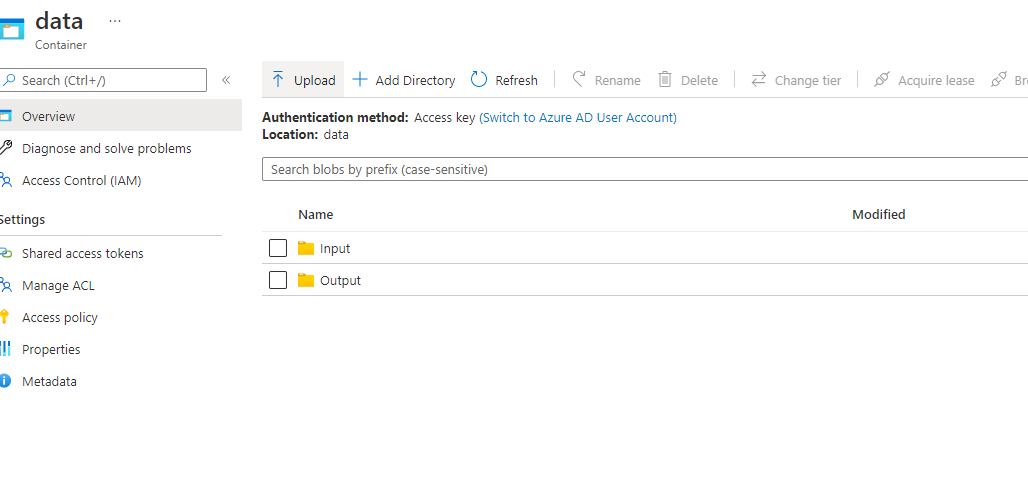
1. The **Containers** page is updated to include **data** in the list of containers.



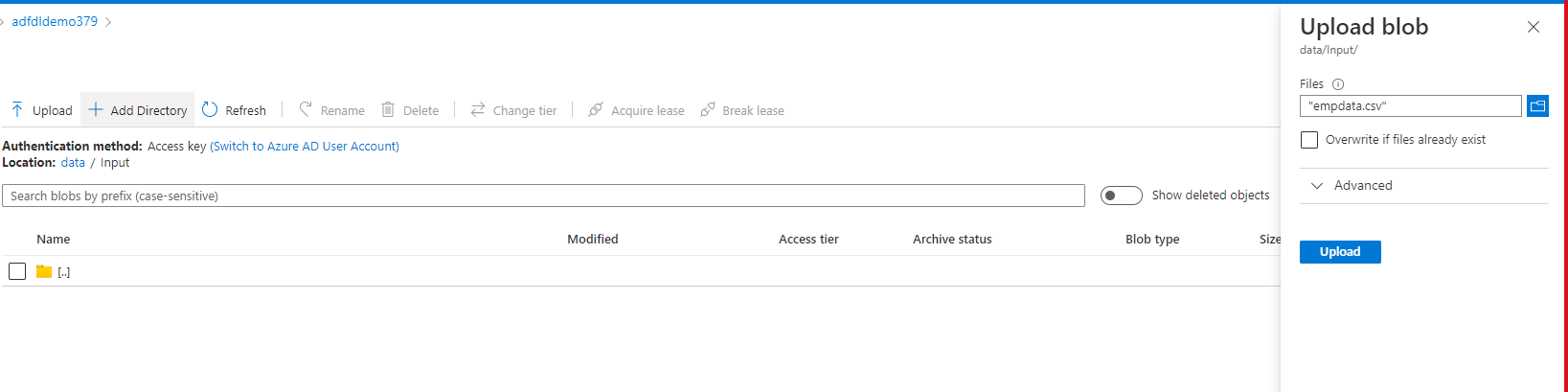
1. Click **Add Directory** and type **Input** and click ok

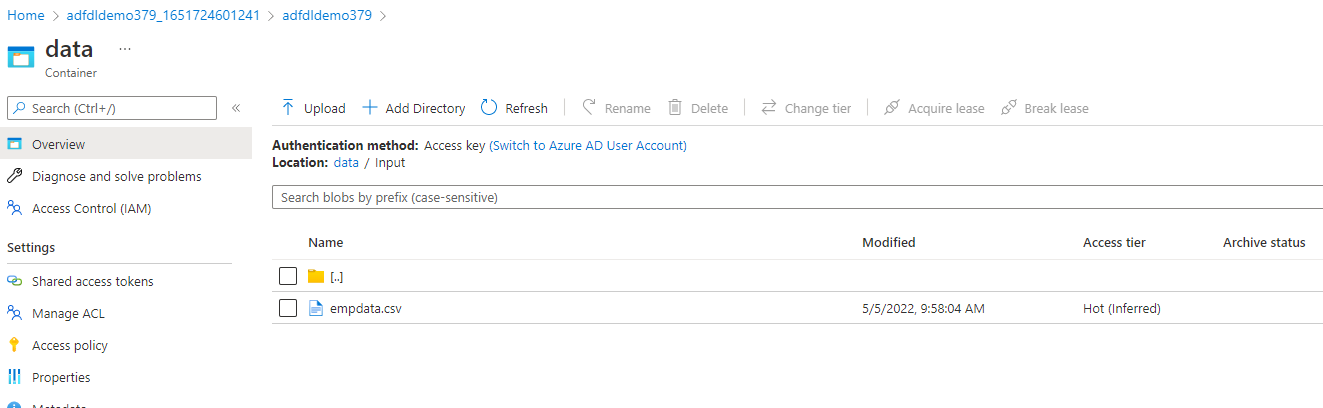
****

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

****

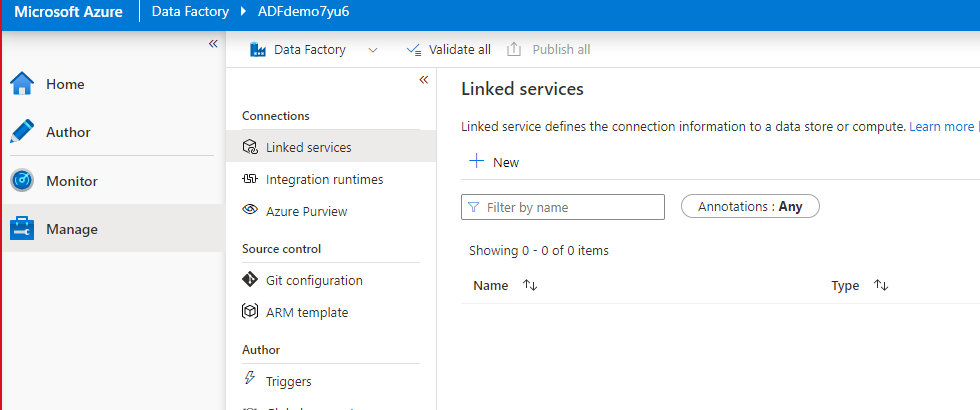
1. Click the **Input Directory** and Click **Upload**.
2. In the **Upload blob** page, select the **Files** box, and then browse to and select the **empdata.CSV** file.

****

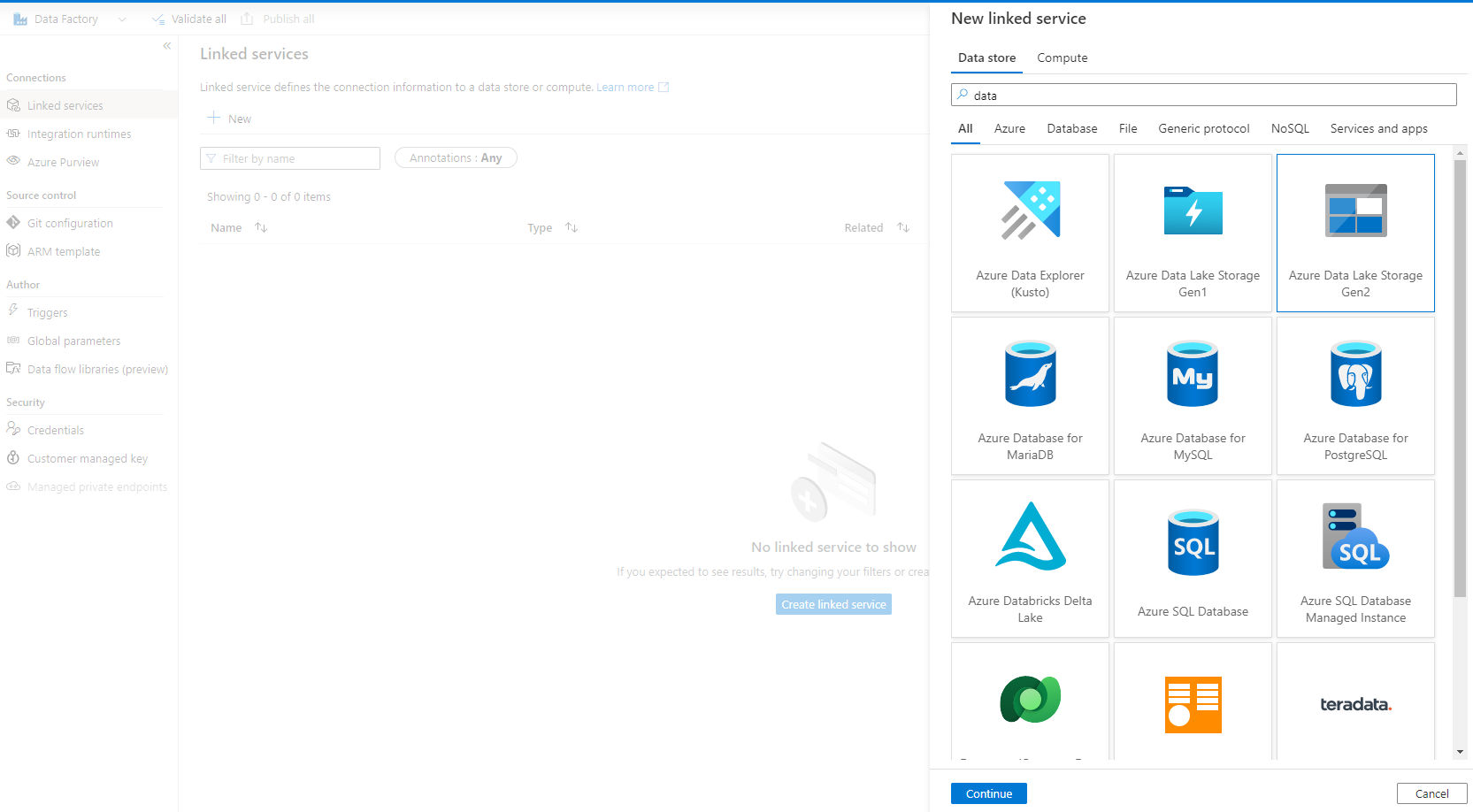
****

# Exercise 5 – Create Linked Services

1. On the Azure Data Factory UI page, open [**Manage**](https://docs.microsoft.com/en-us/azure/data-factory/author-management-hub) tab from the left pane.
2. On the Linked services page, select **+New** to create a new linked service.

****

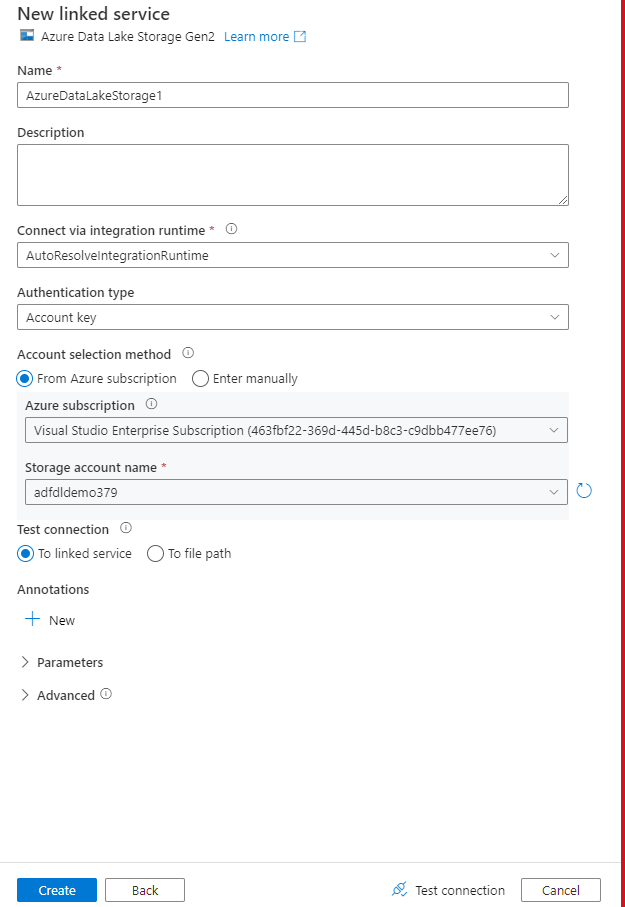
1. On the **New Linked Service** page, select **Data Lake Verion2**, and then select **Continue**.

****

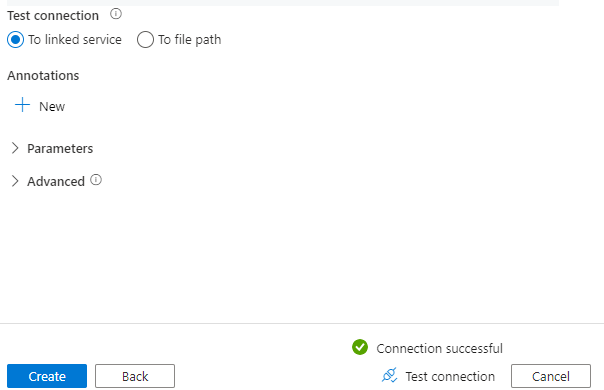
1. On the New Linked Service (Azure Blob Storage) page, complete the following steps:

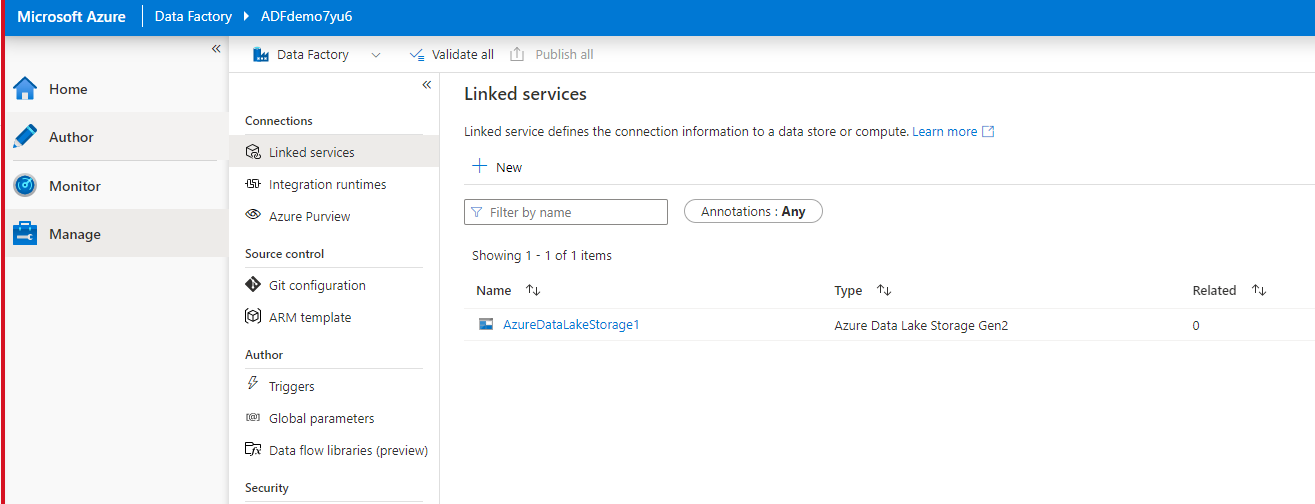
a. For **Name**, enter **AzureStorageLinkedService**.

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

****

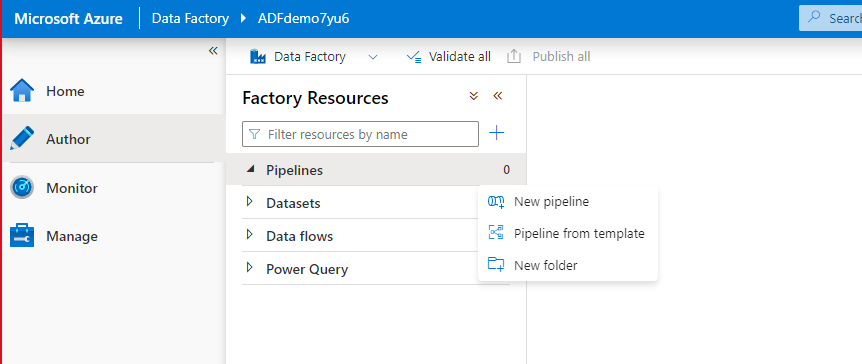
1. Select **Test connection** to confirm that the Data Factory service can connect to the storage account.
2. Select **Create** to save the linked service.

****

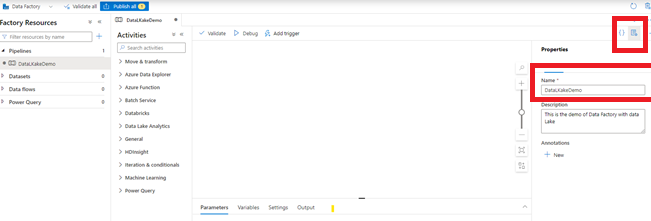
****

# Exercise 1 – Create Pipeline and Dataset

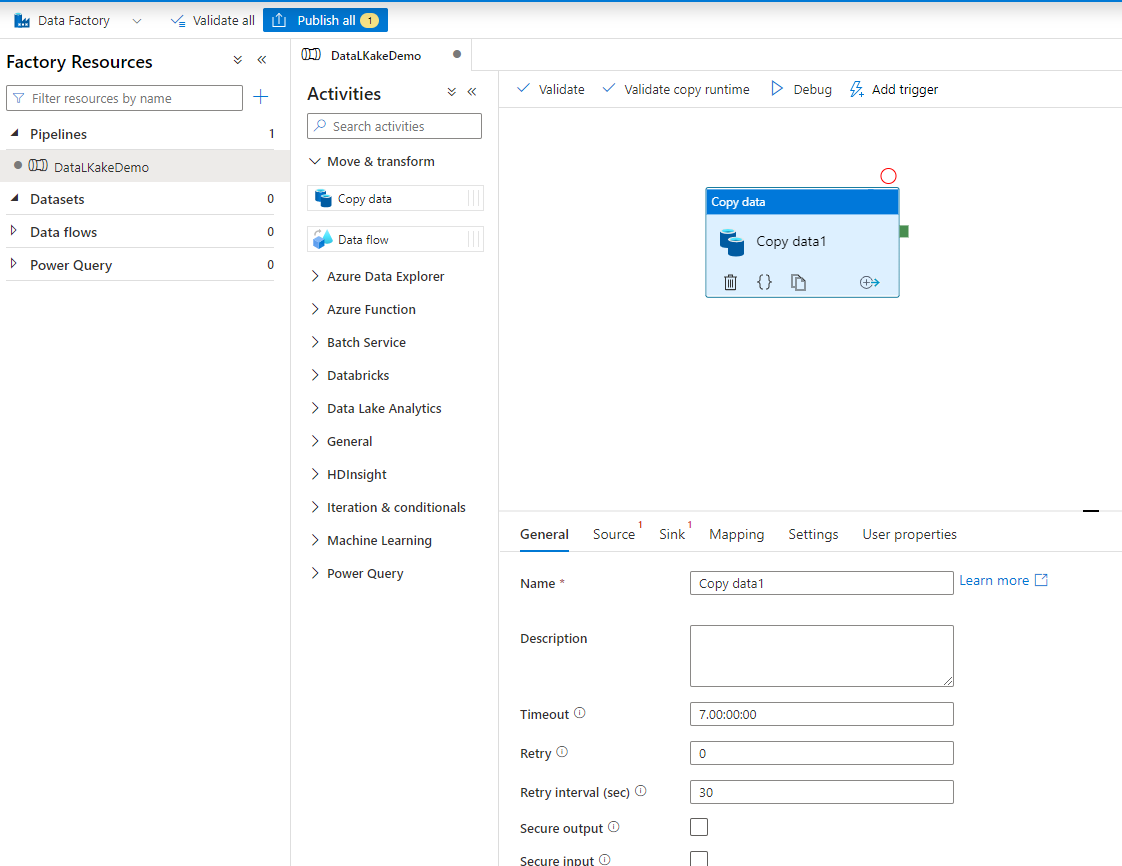
1. Select the **+** (plus) button, and then select **Pipeline**.

****

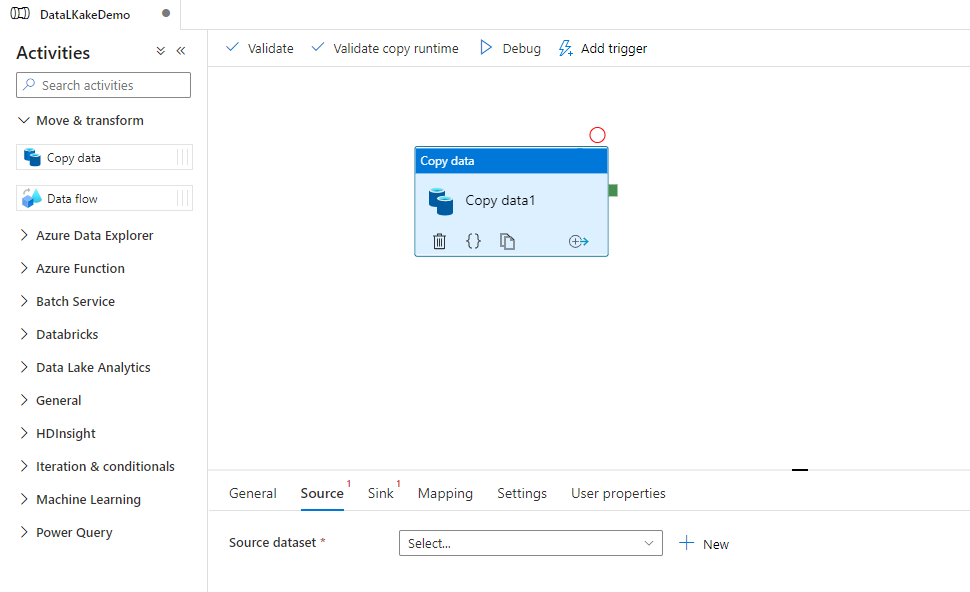
1. In the General panel under **Properties**, specify **CopyPipeline** for **Name**. Then collapse the panel by clicking the Properties icon in the top-right corner.



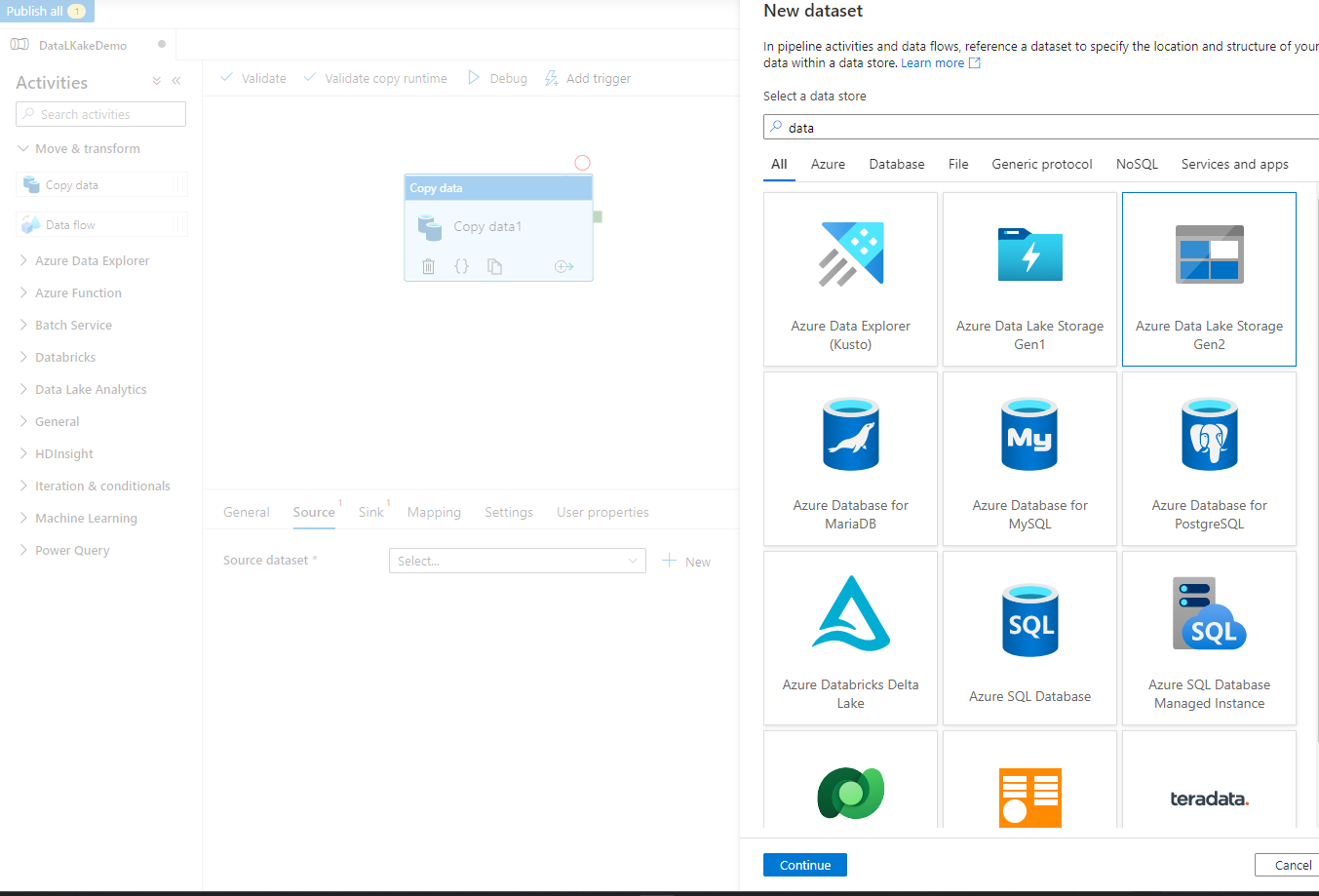
1. In the **Activities** toolbox, expand **Move & Transform**. Drag the **Copy Data** activity from the **Activities** toolbox to the pipeline designer surface.

****

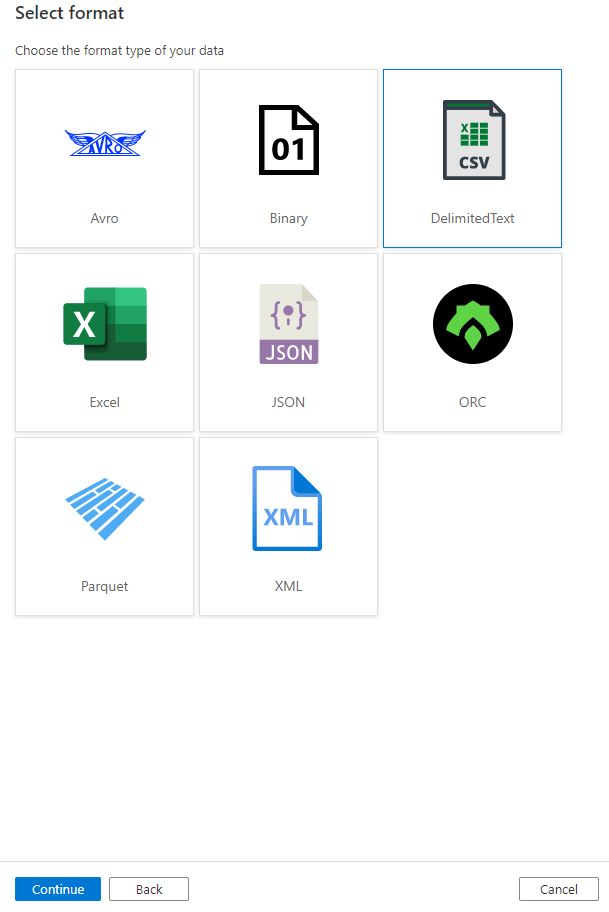
1. Switch to the **Source** tab in the copy activity settings, and Click +New

****

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

****

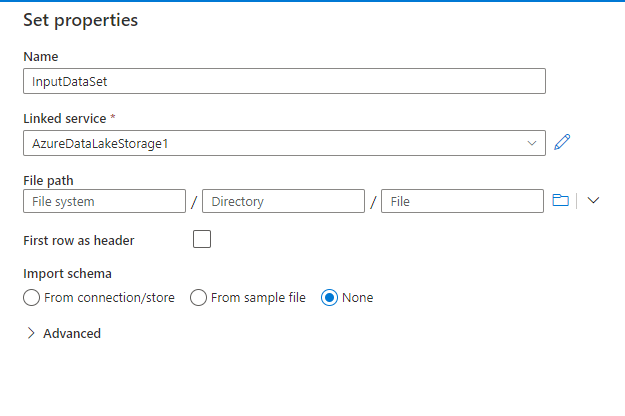
1. On the **Select Format** page, choose the format type of your data, and then select **Continue**. In this case, select **CSV** when copy files as-is without parsing the content.

****

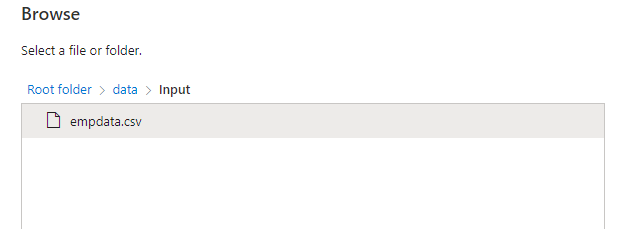
1. On the **Set Properties** page, complete following steps:

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

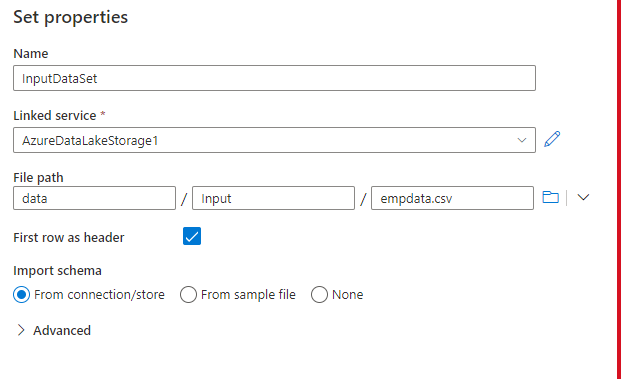
b. For **Linked service**, select **AzureStorageLinkedService**.

****

1. . For **File path**, select the **Browse** button.
2. 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**.

****

1. Select **OK**.

****

1. Repeat the steps to create the output dataset:

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

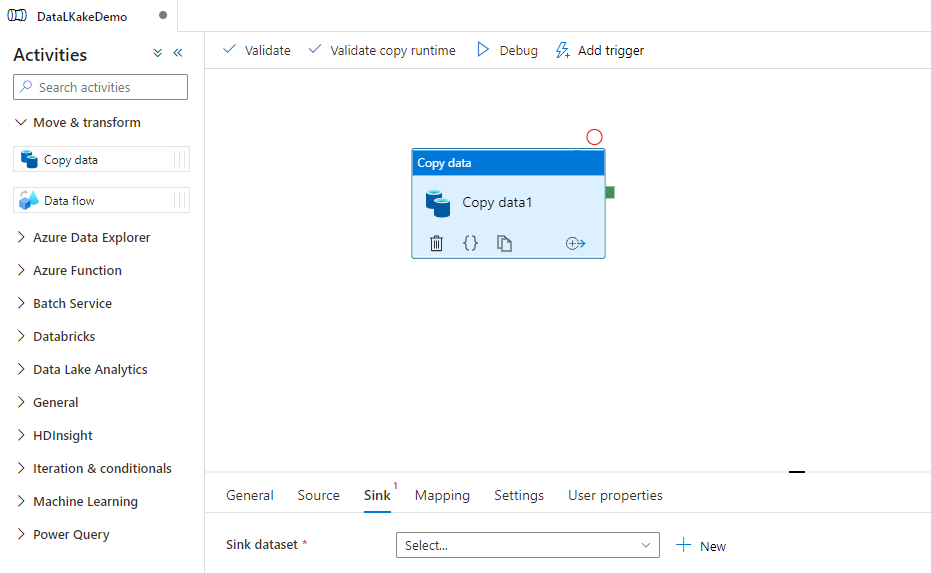
b. On the **New Dataset** page, select **Azure Data Lake Storage V2**, and then select **Continue**.

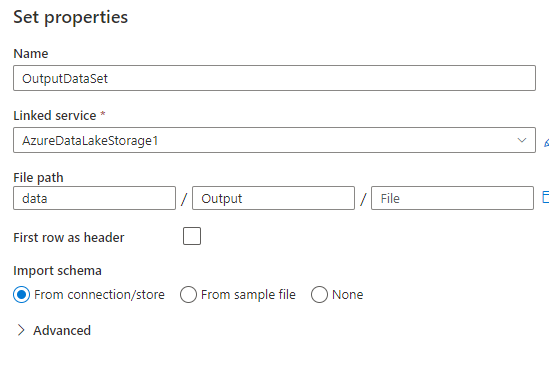
c. On the **Select Format** page, choose the format type of your data, and then select **Continue**.

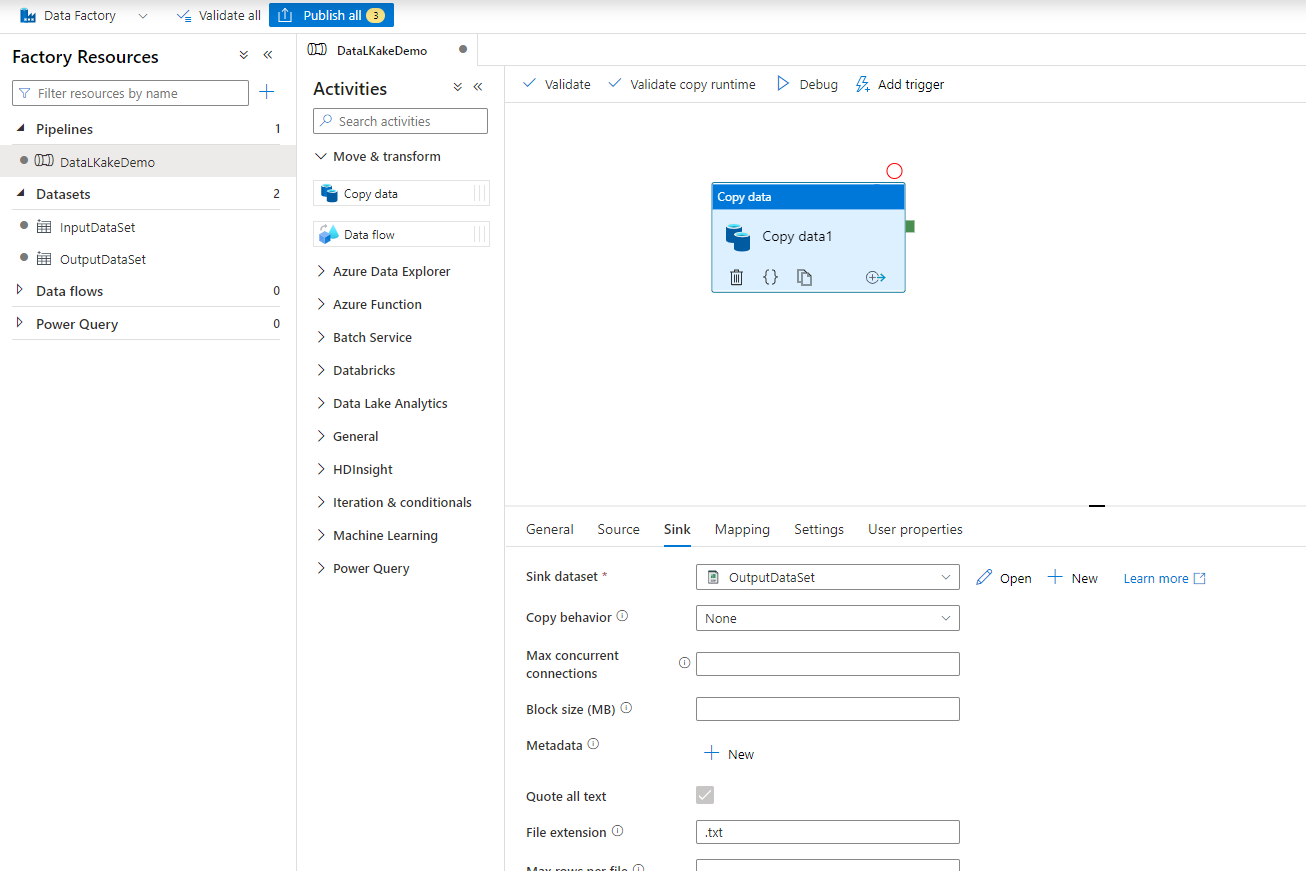
d. On the **Set Properties** page, specify **OutputDataset** for the name. Select **AzureStorageLinkedService** as linked service.

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

f. Select **OK**.

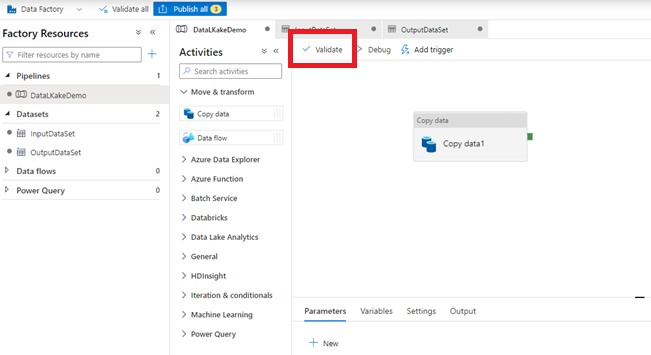
****

****

****

# Exercise 7 – Validate and Debug the Pipeline

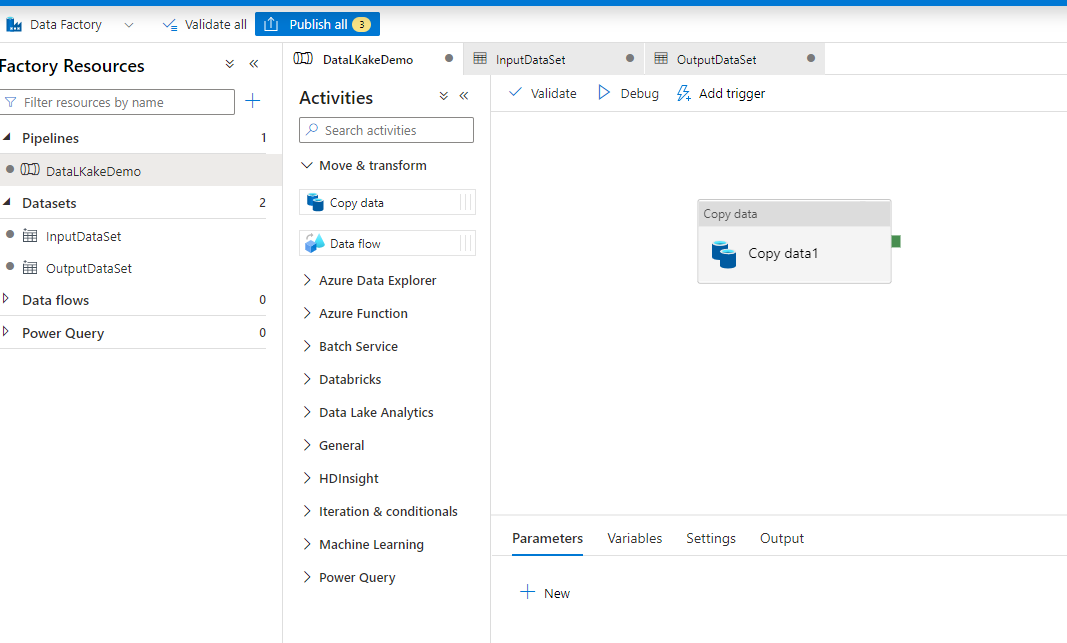
## In the Pipeline click Validate to Validate the Pipeline for any configuration error



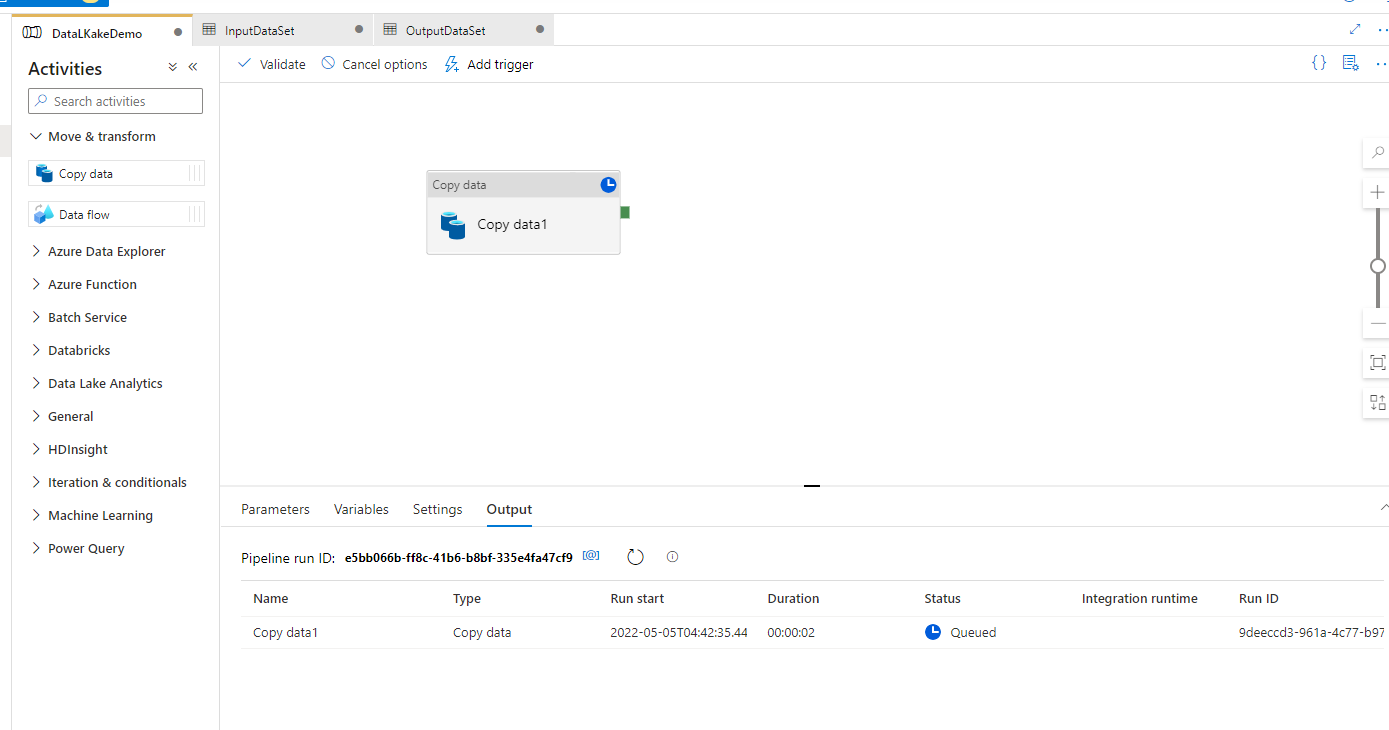
## It Should show the message Validation is Passed

****

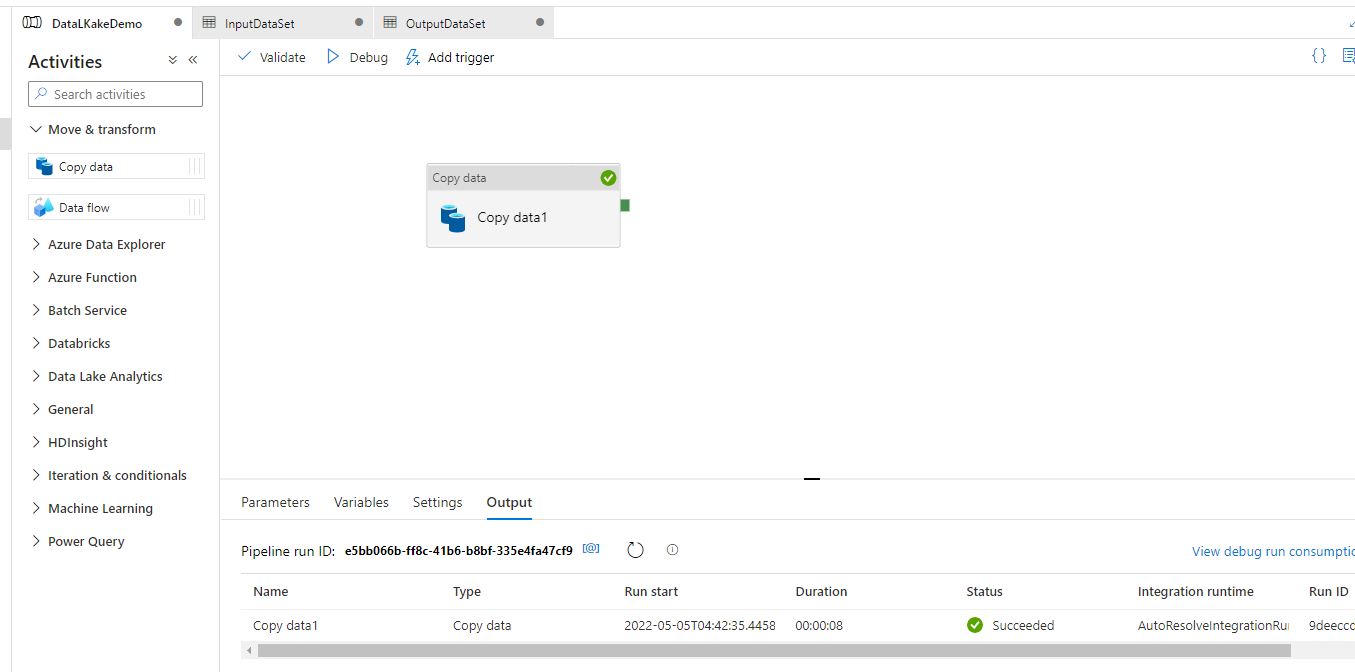
## On the pipeline toolbar above the canvas, click Debug to trigger a test run.

****

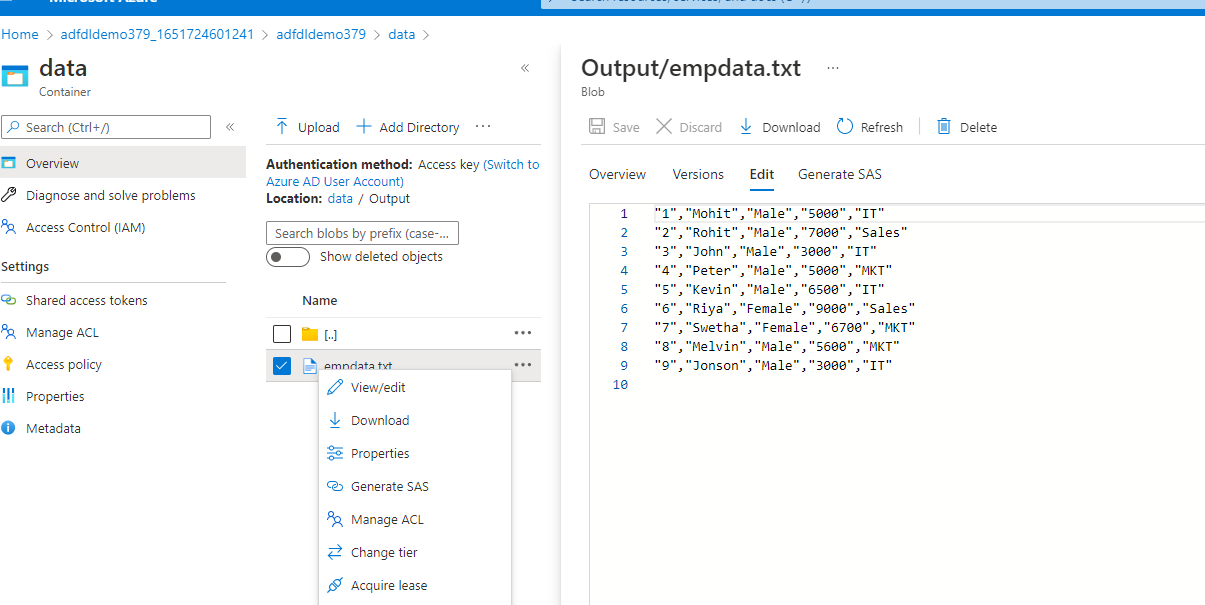
## Confirm that you see the status of the pipeline run on the Output tab of the pipeline settings at the bottom.

****

## Confirm that you see that you succeed Message .

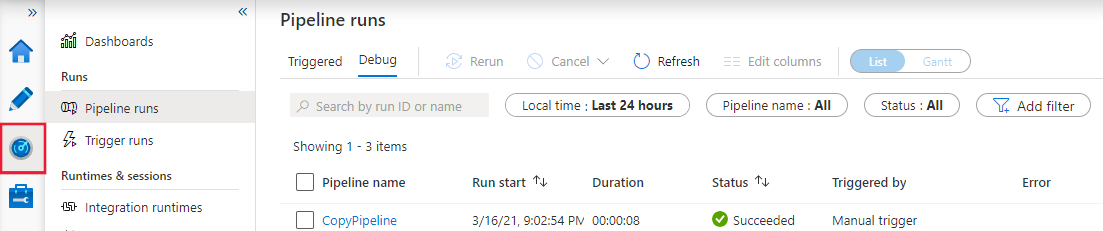
****

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

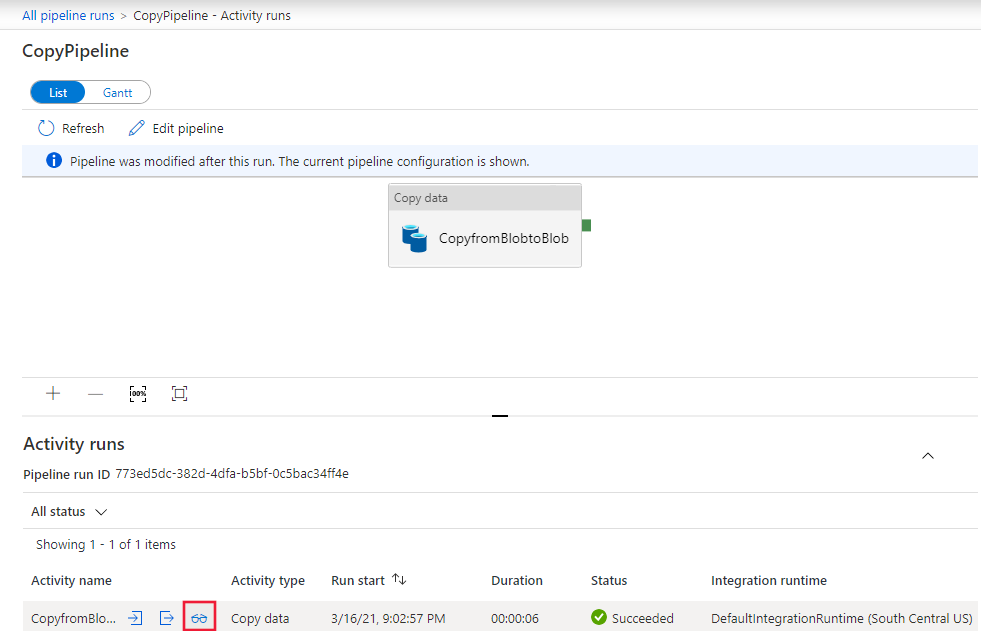
****

# Monitor the pipeline

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

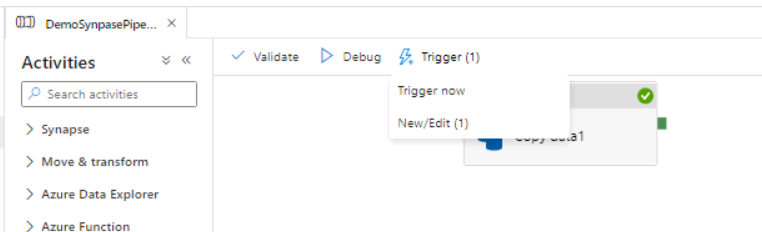


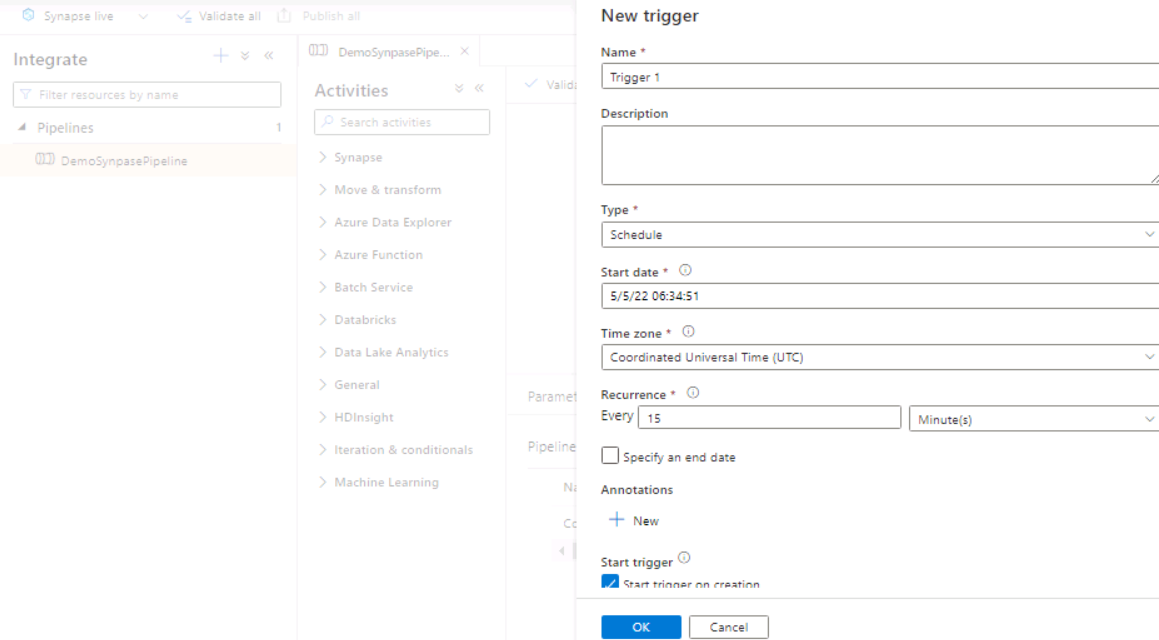
1. Select the **CopyPipeline** link, you'll see the status of the copy activity run on this page.
2. To view details about the copy operation, select the **Details** (eyeglasses image) link. For details about the properties, see [Copy Activity overview](https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview).



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

# 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**.

****