Exercise - Use Azure Data Factory to copy data from Data Lake Storage Gen1 to Data Lake Storage Gen2

Aure Data Factory is a cloud-based data integration service that creates workflows in the cloud. These workflows orchestrate batch data movement and transformations. Use Data Factory to create and schedule workflows (called *pipelines*) to ingest data from various data stores. The data can then be processed and transformed with services like these:

- Azure HDInsight
- Spark
- Azure Data Lake
- Azure Machine Learning

Data Factory can orchestrate many data tasks. In this exercise, you'll use it to copy data from Azure Data Lake Storage Gen1 to Data Lake Storage Gen2.

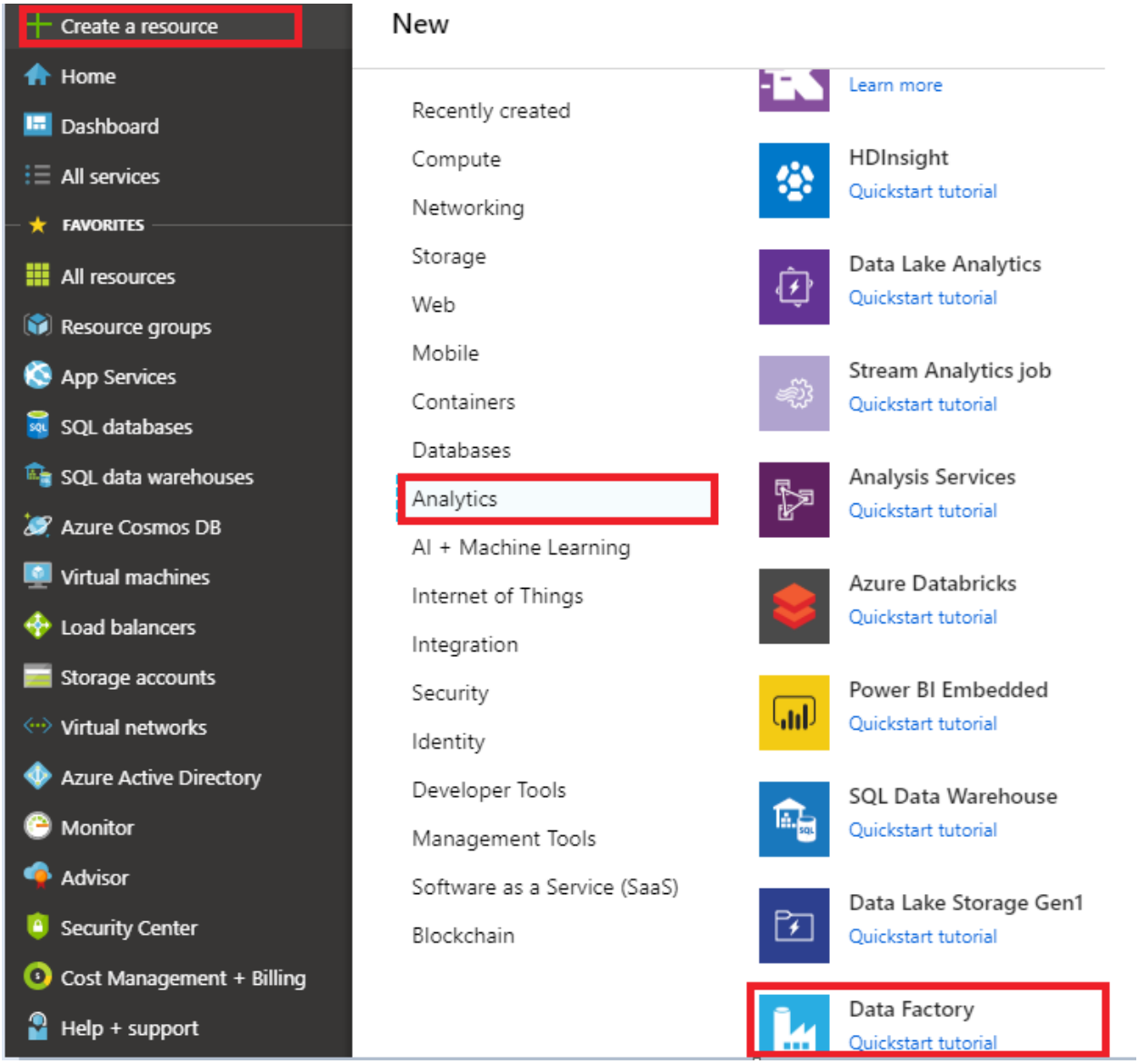
Note

If you don't have an Azure account or prefer not to do this exercise in your account, just read through the exercise to understand how to use Data Factory to copy data into a data lake.

Create a data factory

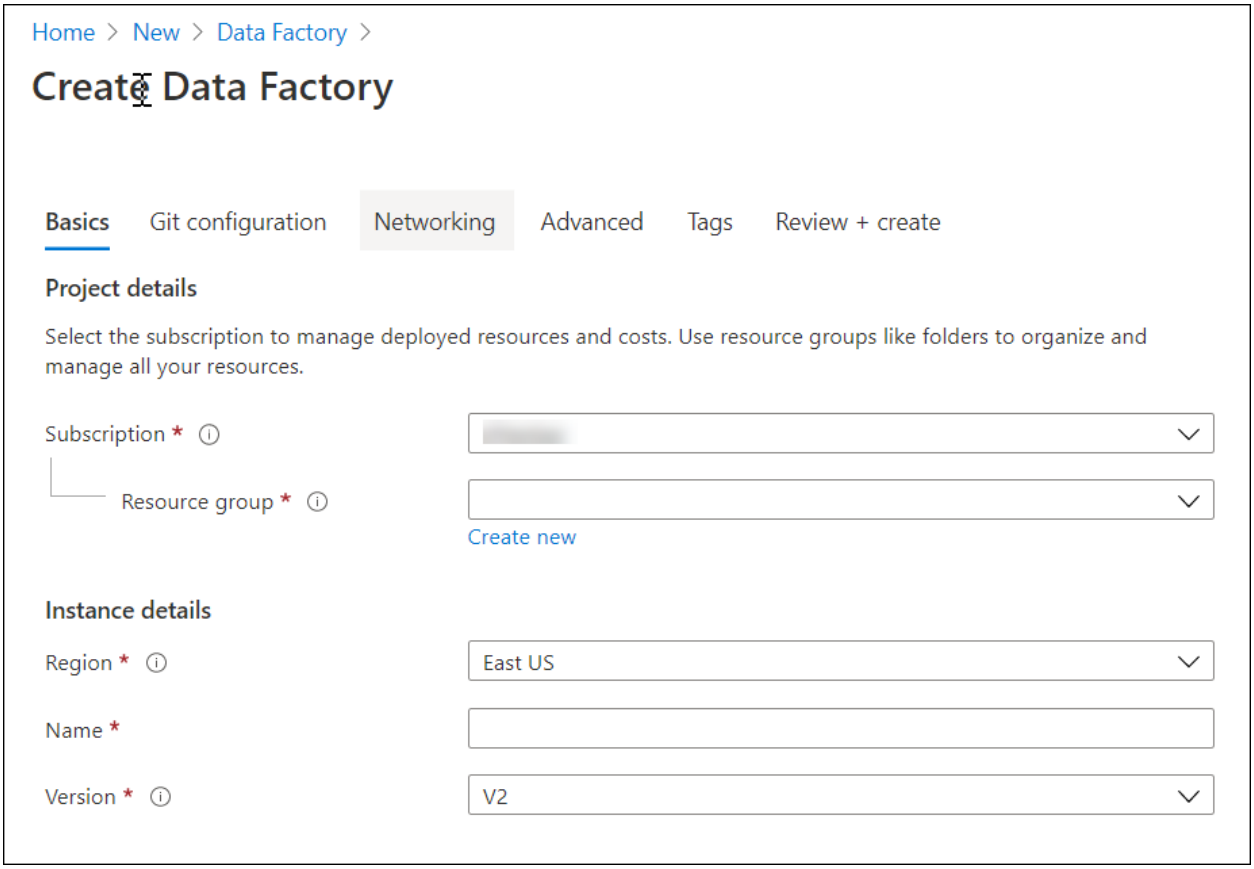
The first step is to provision a data factory in the Azure portal.

1. Sign in to the Azure portal.
2. On the left sidebar, select **+ Create Resource** > **Integration** > **Data Factory**.



3. On the **New data factory** page, provide values for each of the required fields:

- **Subscription**: The subscription in which the ADF instance is created
- **Resource group**: The resource group where the ADF instance will reside
- **Region**: The datacenter location in which the instance is stored
- **Name**: The name of the Azure Data Factory instance
- **Version**: select V2 for the latest features

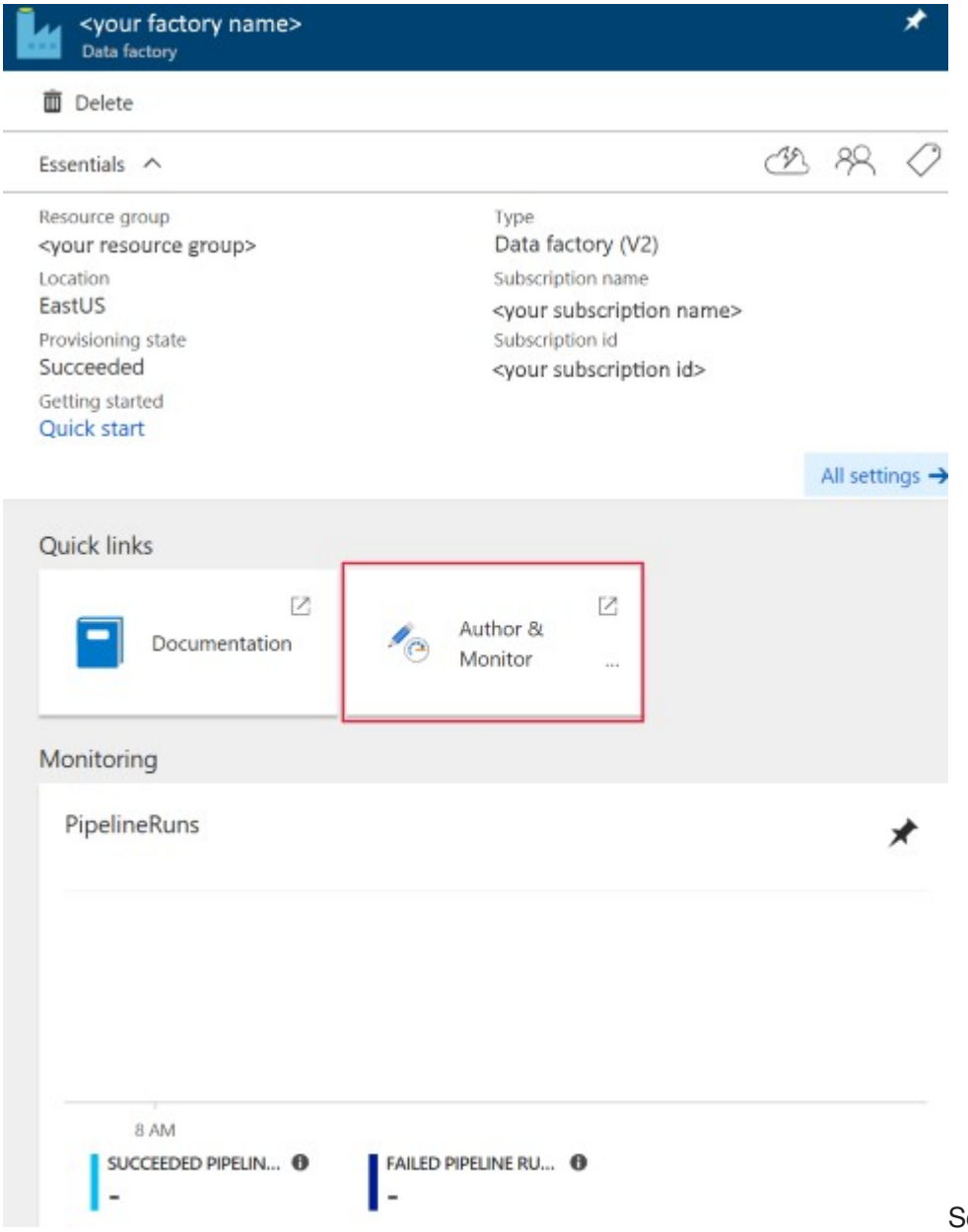


Screenshot showing how to set up a data factory

4. Click on the **Git configuration** tab, and choose to set it up later.

5. Select **Create**.

Now go to the newly created data factory. You should see the **Data factory** home page.



Screenshot of the new data factory.

Important

You will need a Data Lake Storage Gen1 account that contains data. If you don't have this, follow the steps in the next sections.

Create a Data Lake Storage Gen1 account

1. On the left, select **Create a new resource**.
2. On the **New** pane, select **Storage** > **Data Lake Storage Gen1**.
3. In the **Name** box, type **dlsgen1XXX**, but replace *XXX* with numbers that you choose. A green check mark indicates that the name is unique.
4. In the **Subscription** list, select your subscription.

5. In the **Resource Group** list, select **mslearn-datalake-test**.
6. Select a location. Typically, you'll want to select a region near where the data will be consumed. For this example, select a location near you.
7. Select **Create**.

Create a sample text file

You'll need some sample data to work with, so create a text file on your local computer. Name the file **salesUK.txt**. Then paste the following text into the file:

```
#salaries Details
#Company Information
#Fields : Date company employee Salaries
01-02-2019  d1   f1 8000
01-02-2019  d2   f2 9000
01-02-2019  d1   f3 2000
01-02-2019  d2   f4 3000
01-02-2019  d1   f5 4000
01-02-2019  d3   f6 5000
```

You'll upload this data file in various ways. Keep in mind that this is a *simple* example. Typically, you'll populate your data lake with much larger data samples from a variety of sources.

Upload a file into a Data Lake Storage Gen1 account

1. In the Azure portal, search for the Data Lake Storage Gen1 service you created (**dlsgen1XXX**).
2. On the **Overview** pane, select **Data Explorer**.
3. On the **Data Explorer** pane, select the **Upload** button.
4. On the **Upload file** pane, select the browse icon, go to the folder, and select **salesUK.txt**. Then select **Add selected files**. You'll know the file is uploaded when the **Status** column displays **Completed**.
5. Close the **Upload files** pane.

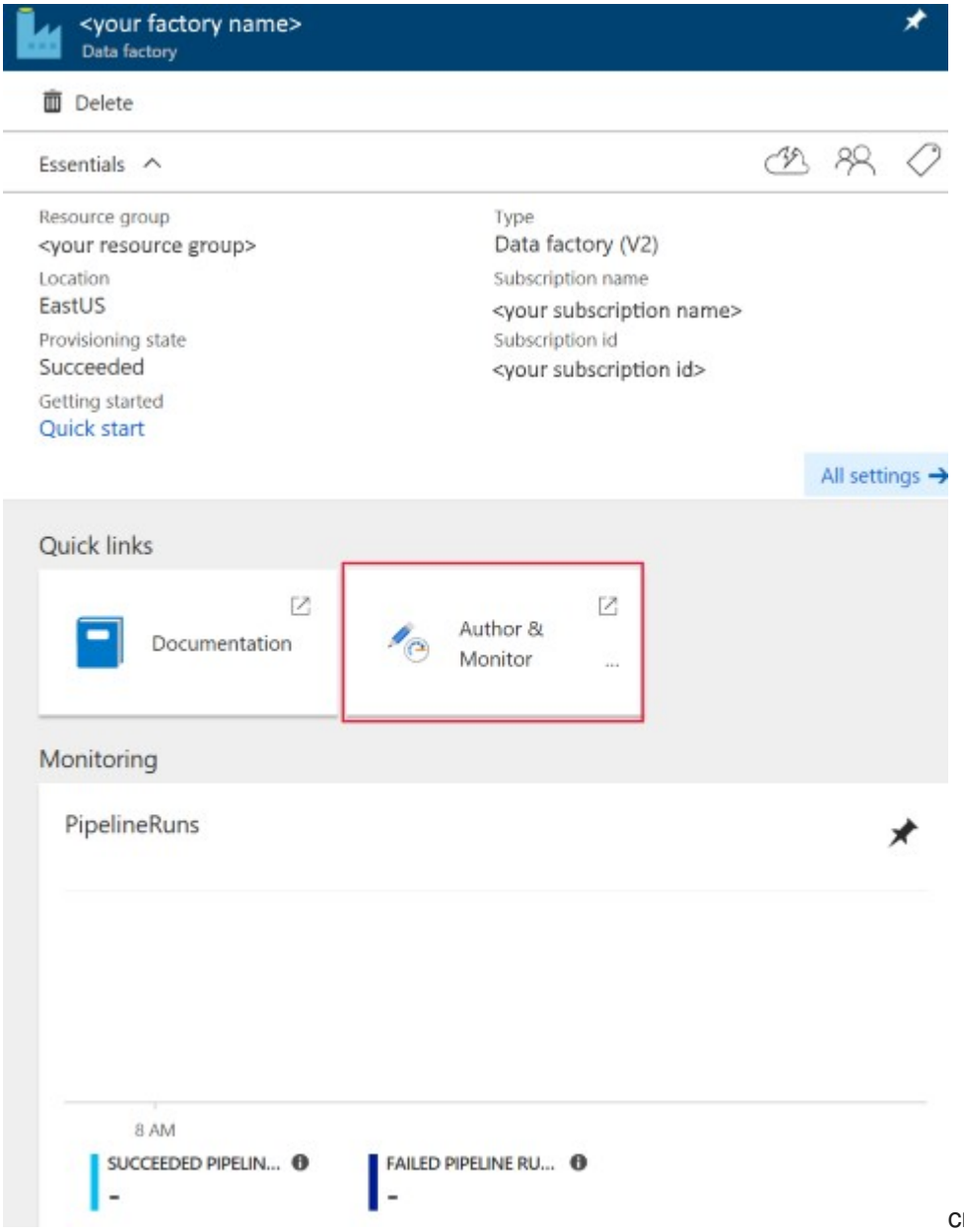
Set permissions for the Data Lake Storage Gen1 account

Set permissions to allow the data factory to access the data in your Data Lake Store Gen1 account.

1. In the Azure portal, search for your Data Lake Storage Gen1 service named **dlsgen1XXX**.
2. On the **Overview** pane, select **Access control (IAM)**.
3. On the **Access control (IAM)** pane, in the **Add Role Assignment** box, select **Add**.
4. On the **Add Role Assignment** pane, for the **Role**, select **Owner**.
5. Under **Select**, enter your data factory name.
6. Select **Save**.
7. Close the **Access control (IAM)** pane.

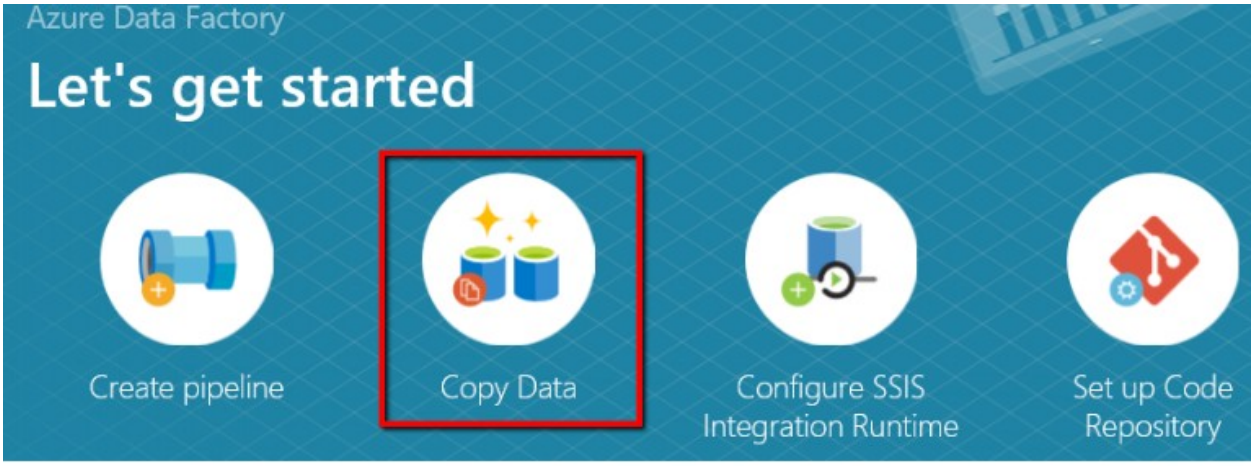
Load data into the Data Lake Storage Gen2 account

1. In the Azure portal, go to your data factory. You'll see the **Data factory** home page.
2. Select **Author & Monitor** to open the Data Integration application in a separate tab.



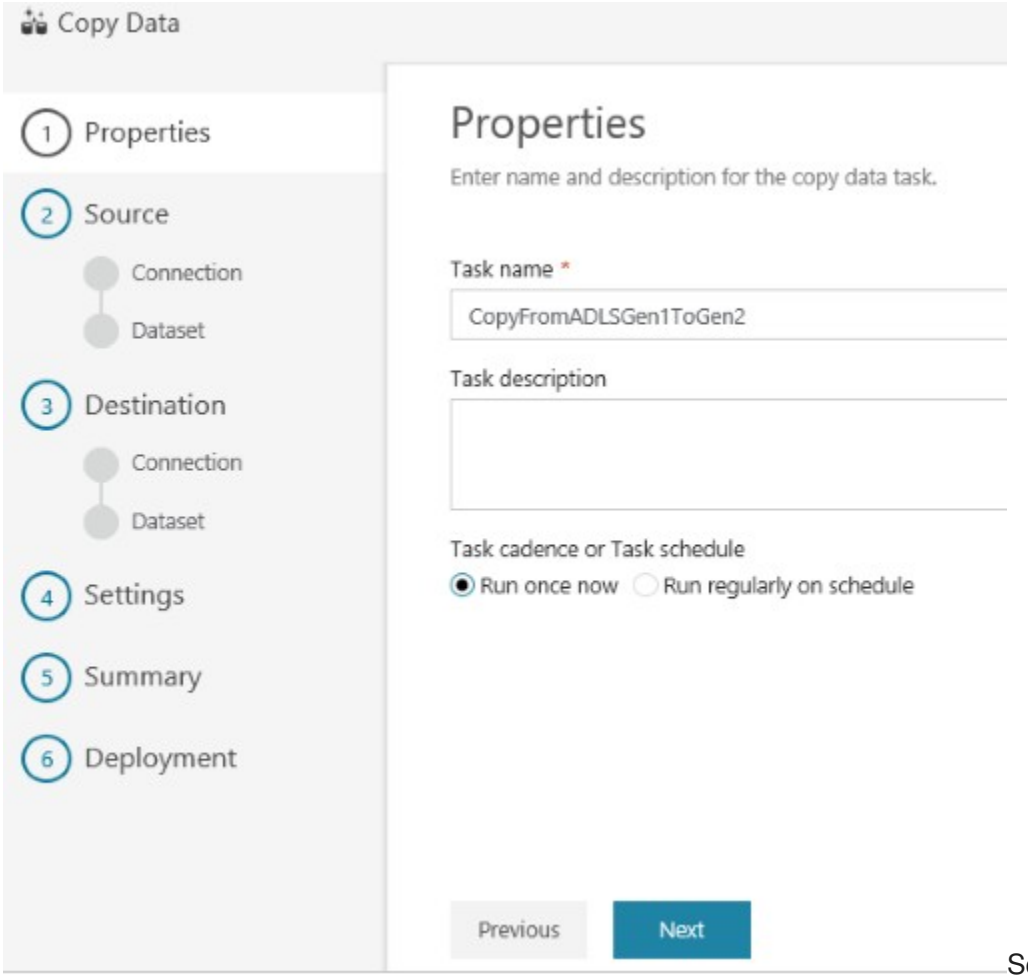
creenshot showing the Data factory home page, where Author & Monitor is selected.

3.Select **Ingest** to open the Copy Data tool.



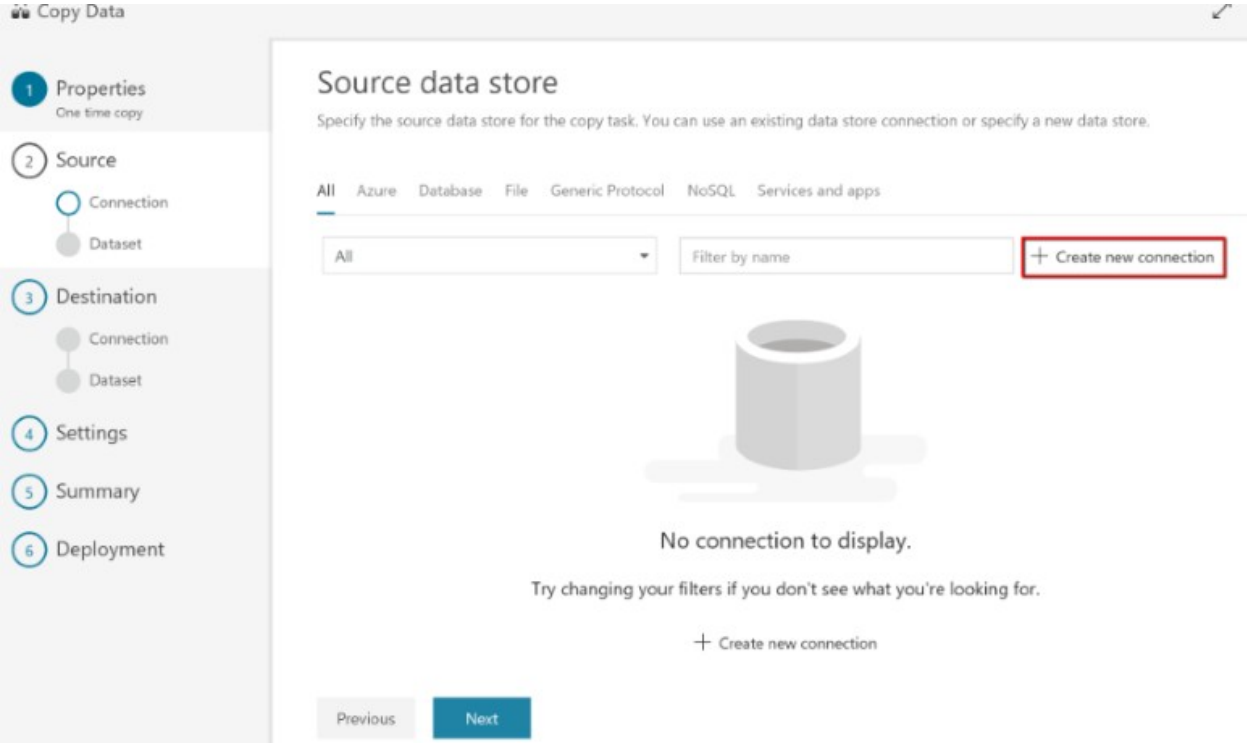
Screenshot showing how to open the Copy Data tool.

4.On the **Properties** page, under **Task type**, click **Built-in copy task**. Then set the task cadence to **Run once now**, and select **Next**.



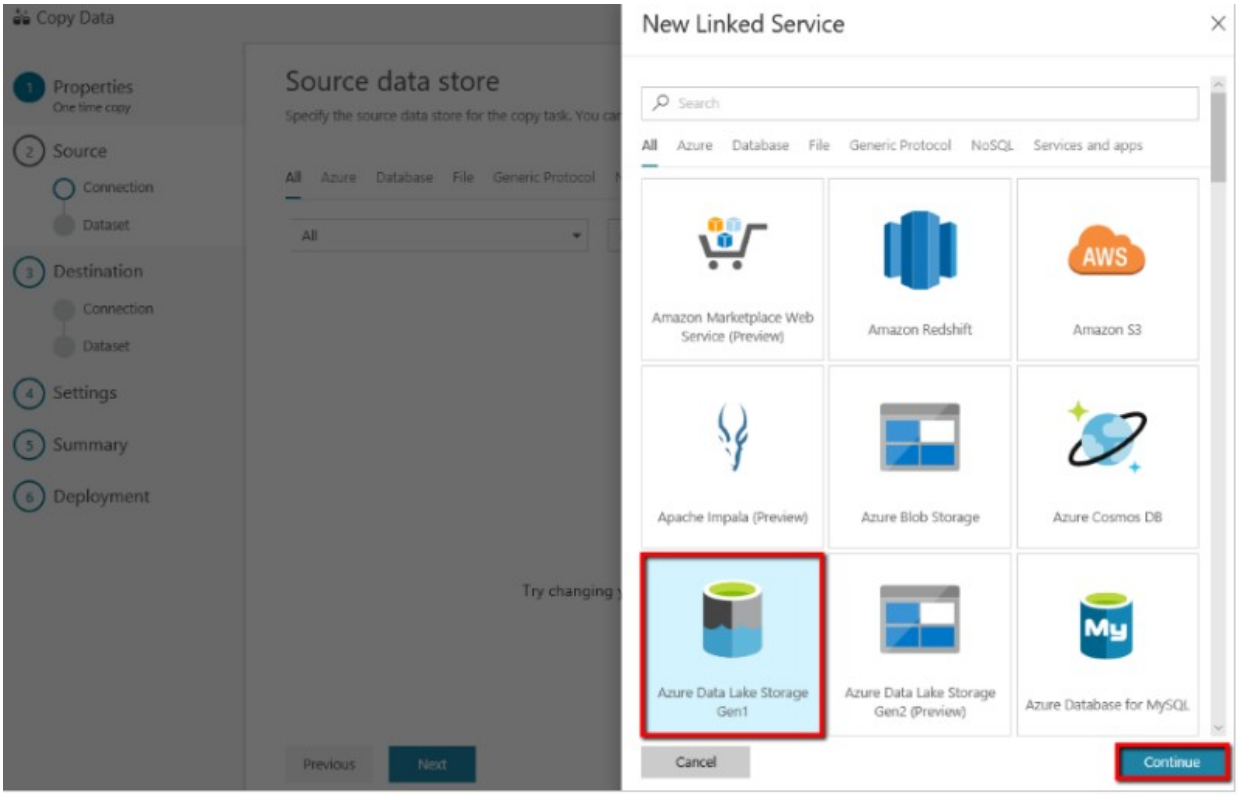
Screenshot showing the Properties page of the Copy Data tool

5.On the **Source data store** page, select **Create new connection**.



Screenshot showing the Source data store page, where Create new connection is selected.

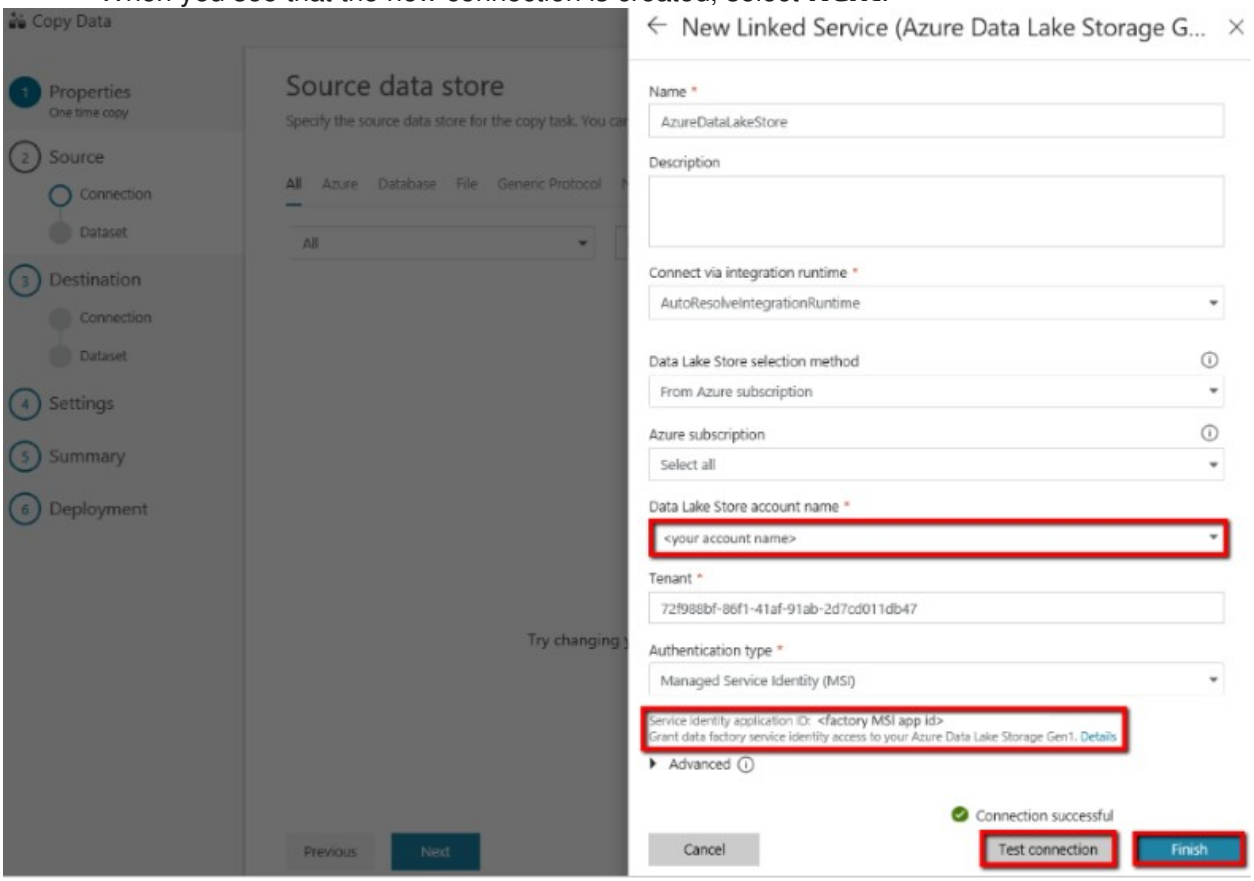
6. In the connector gallery, select **Azure Data Lake Storage Gen1** > **Continue**.



Screenshot showing selections in the connector gallery.

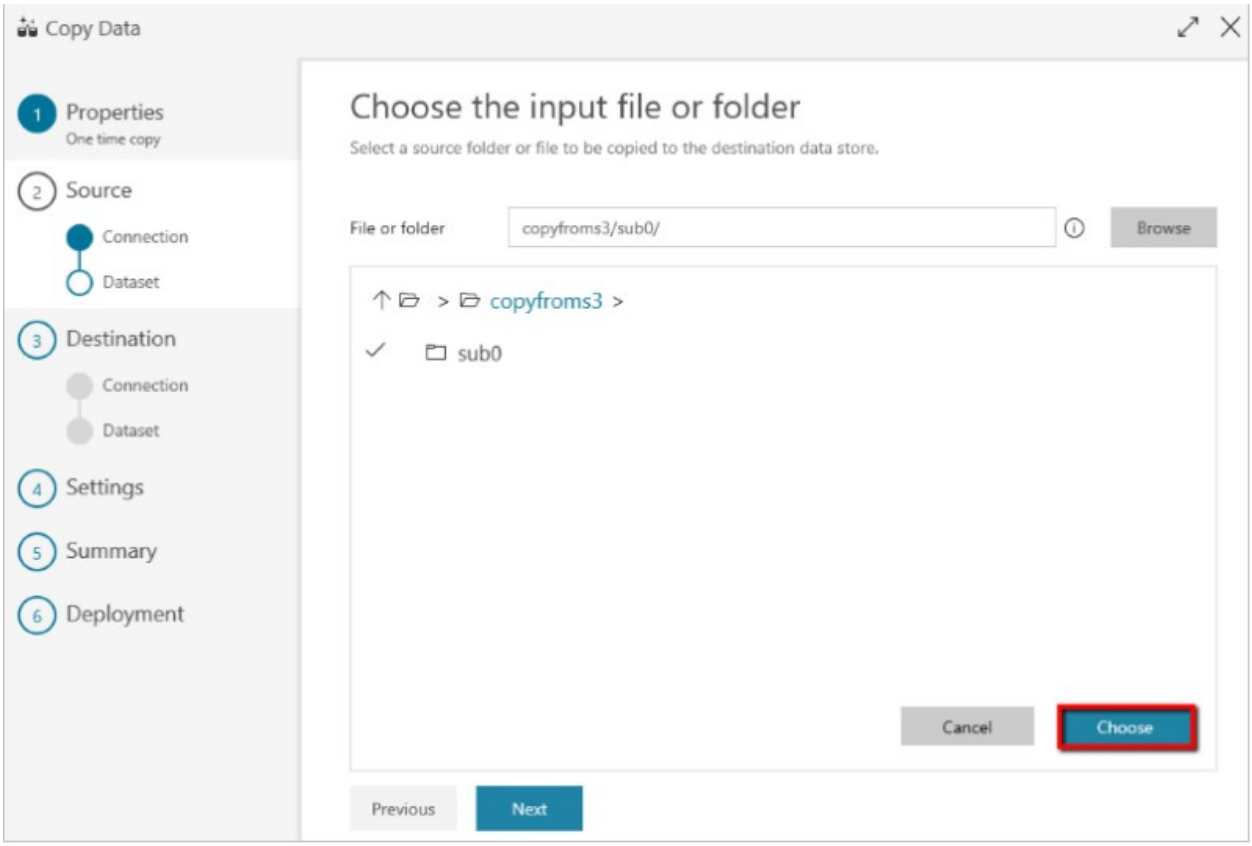
7. On the **New Connection (Azure Data Lake Storage Gen1)** page:

- Under **Data Lake Store Selection method**, select your Azure subscription.
- Under **Tenant**, specify or validate the tenant.
- To validate the settings, select **Test connection** > **Finish**.
- When you see that the new connection is created, select **Next**.



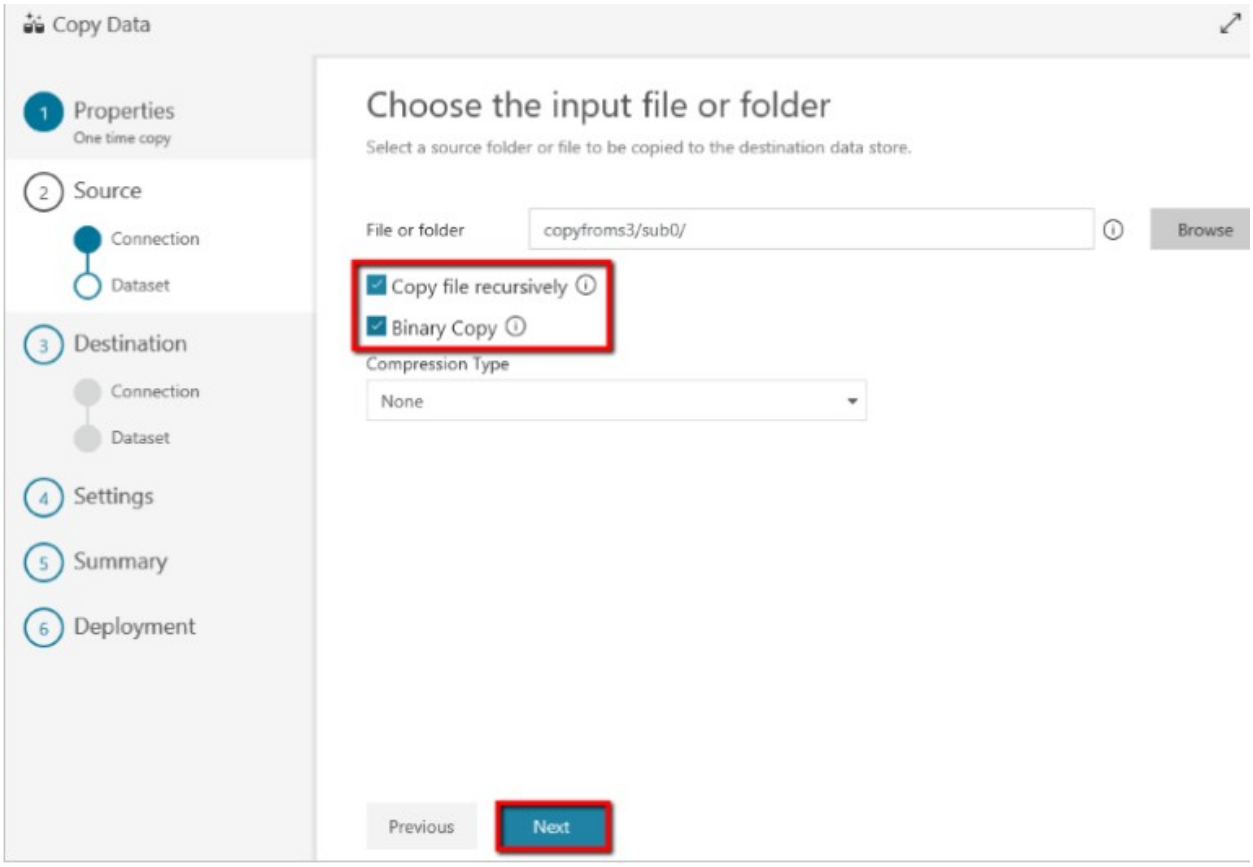
Screenshot showing how to create a linked service.

8. On the **Choose the input file or folder** page, go to the folder and file that you want to copy over. Select the folder or file, and then select **Choose**.



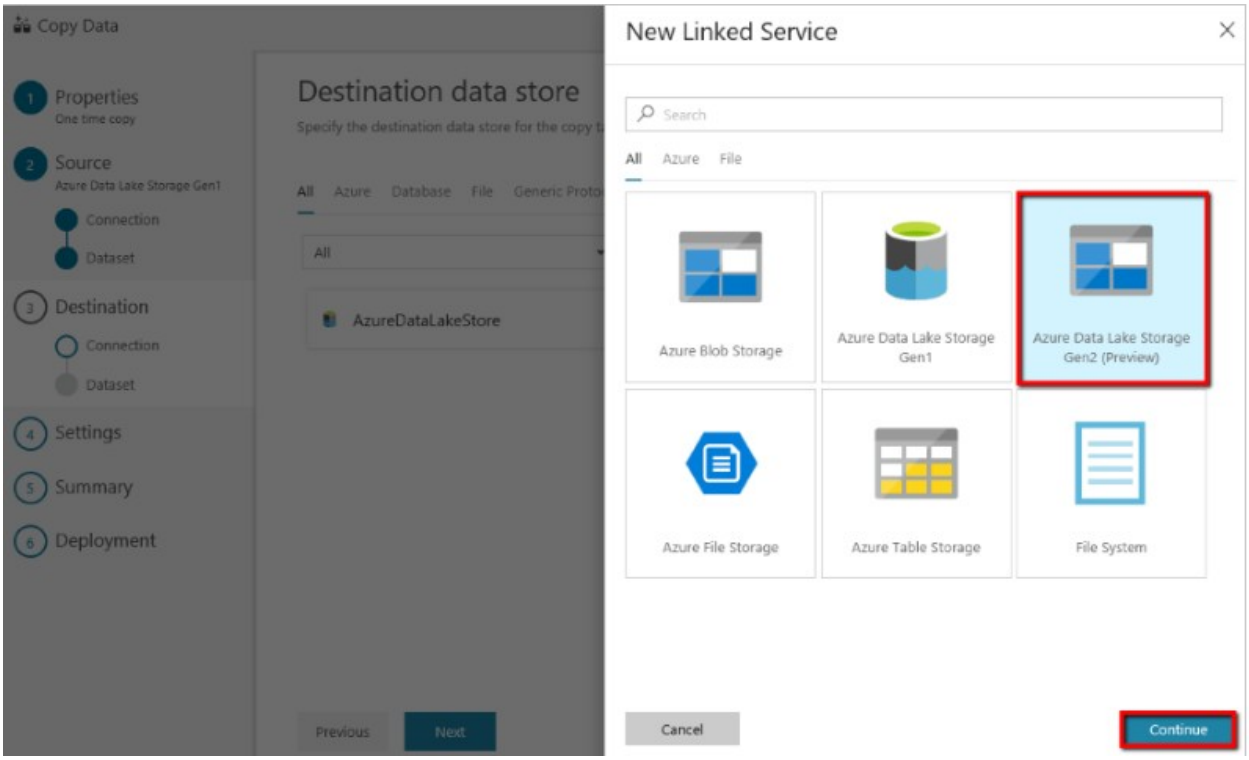
Screenshot showing how to select the input file or folder.

9. Specify the copy behavior by selecting **Copy files recursively** and **Binary Copy**. Then select **Next**.



Screenshot showing the two copy options on the Choose the input file or folder page.

10. On the **Destination data store** page, select **Create new connection** > **Azure Data Lake Storage Gen2 (Preview)** > **Continue**.

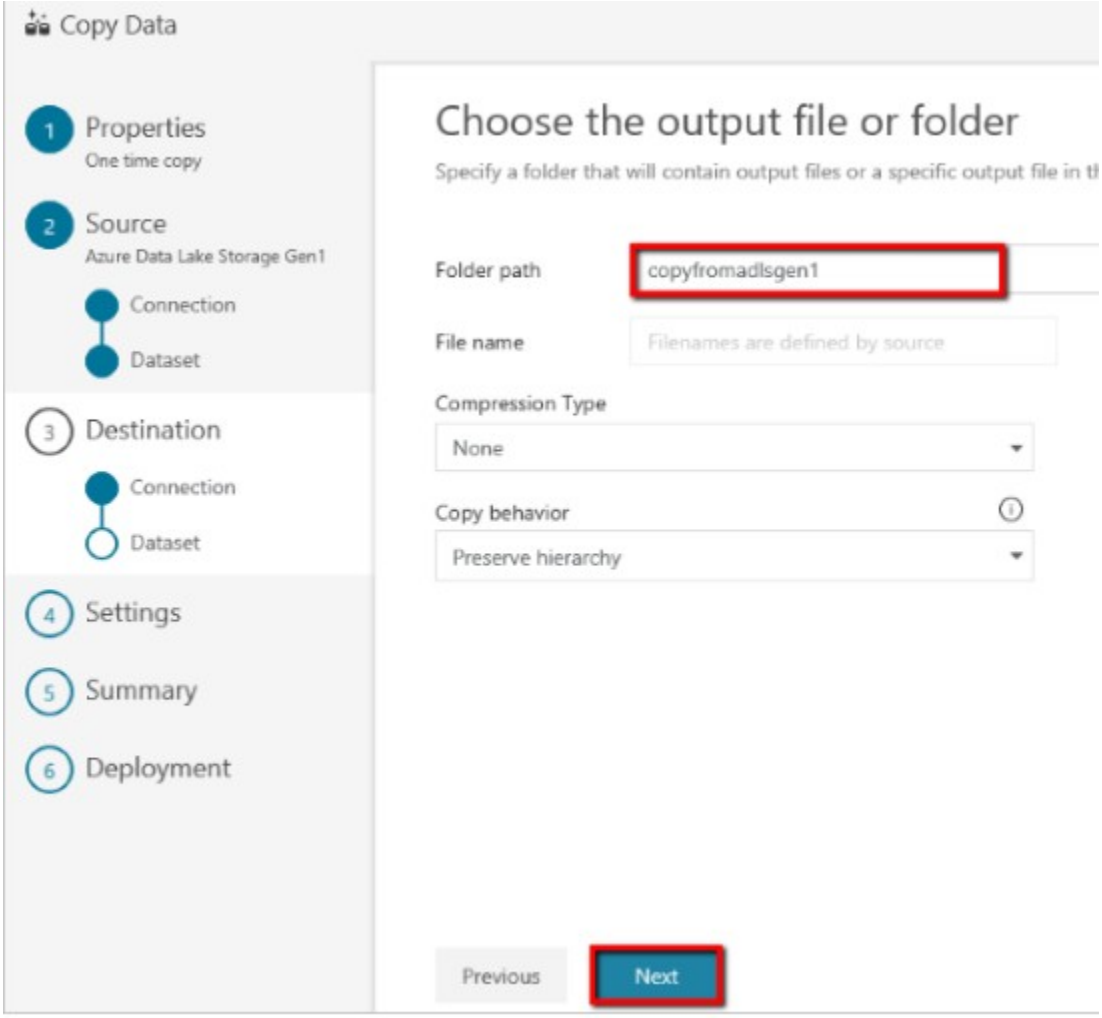


Screenshot showing how to select the destination.

11. On the **Specify Azure Data Lake Storage Gen2 connection** page:

- In the **Storage account name** list, select your Data Lake Storage Gen2 account, this will automatically populate the access key.
- To create the connection, select **Finish** > **Next**.

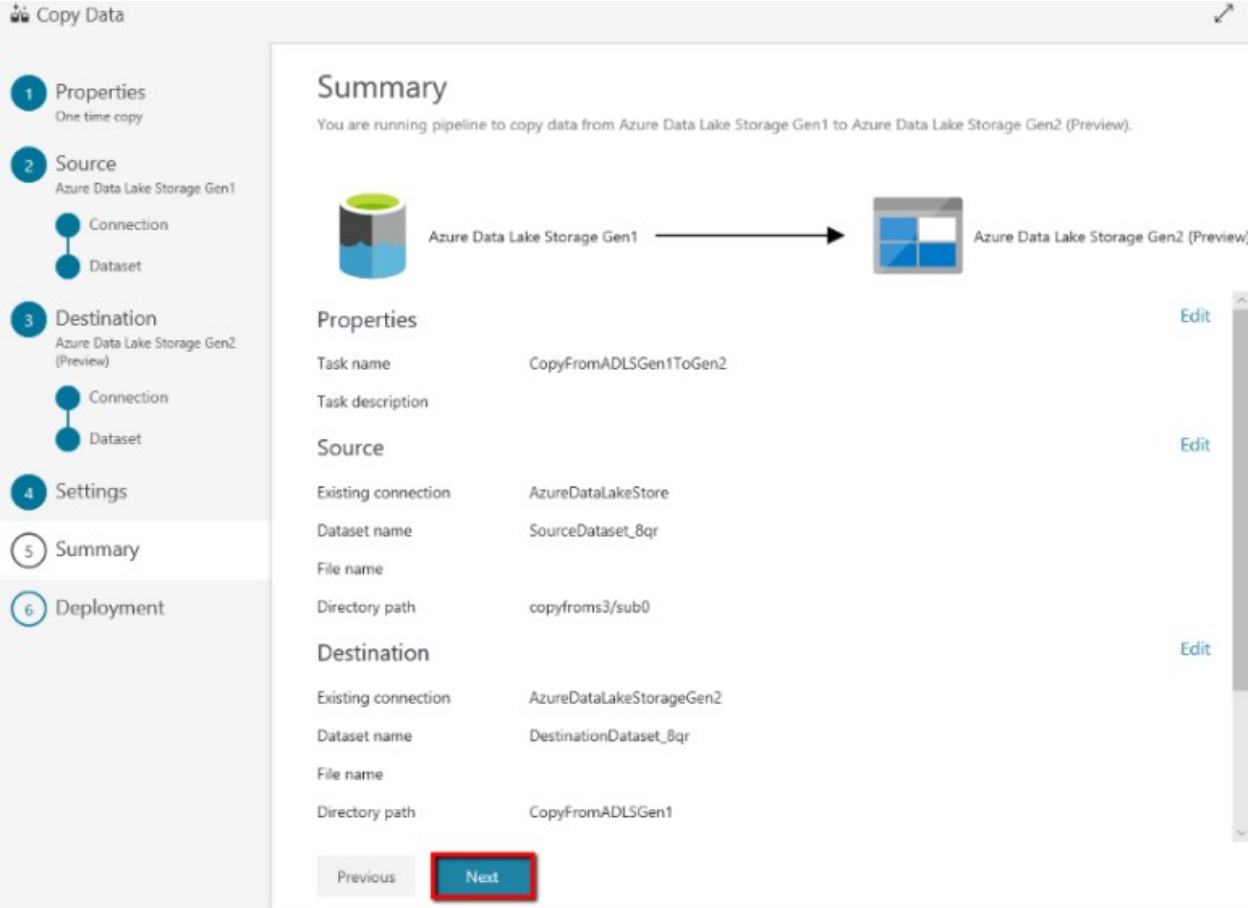
12. On the **Choose the output file or folder** page, next to **Folder path**, enter **copyfromadlsgen1**. Then select **Next**.



Screenshot showing where to enter the output folder path.

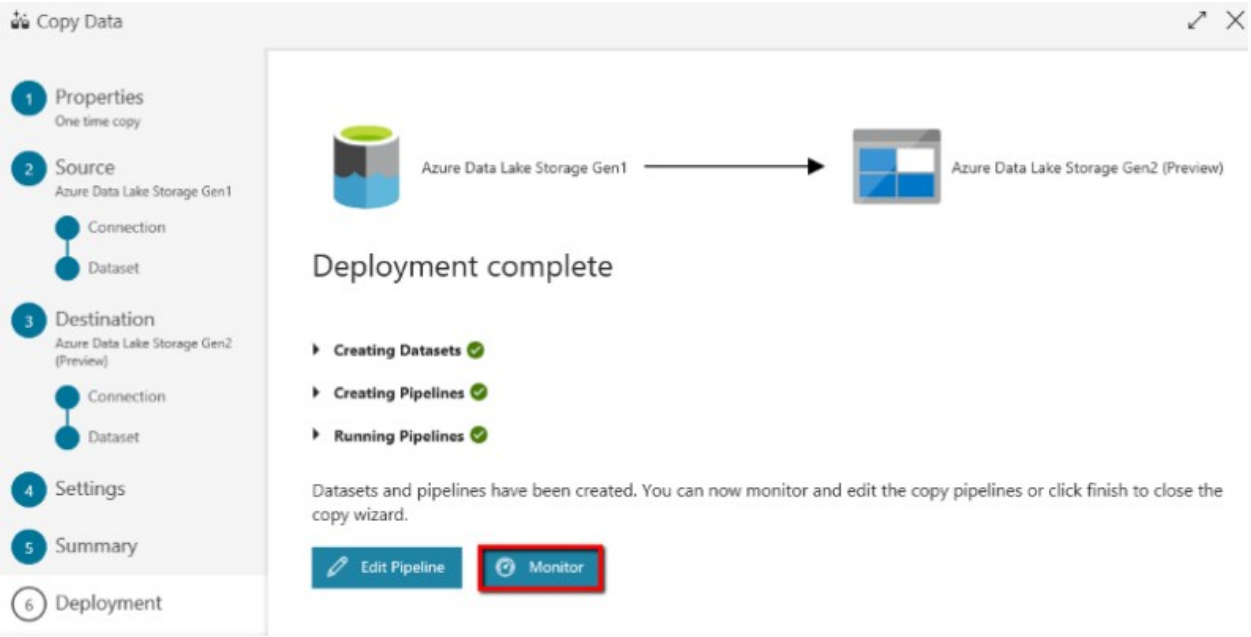
13. On the **Settings** page, select **Next** to use the default settings.

14. Review the settings on the **Summary** page, and select **Next**.



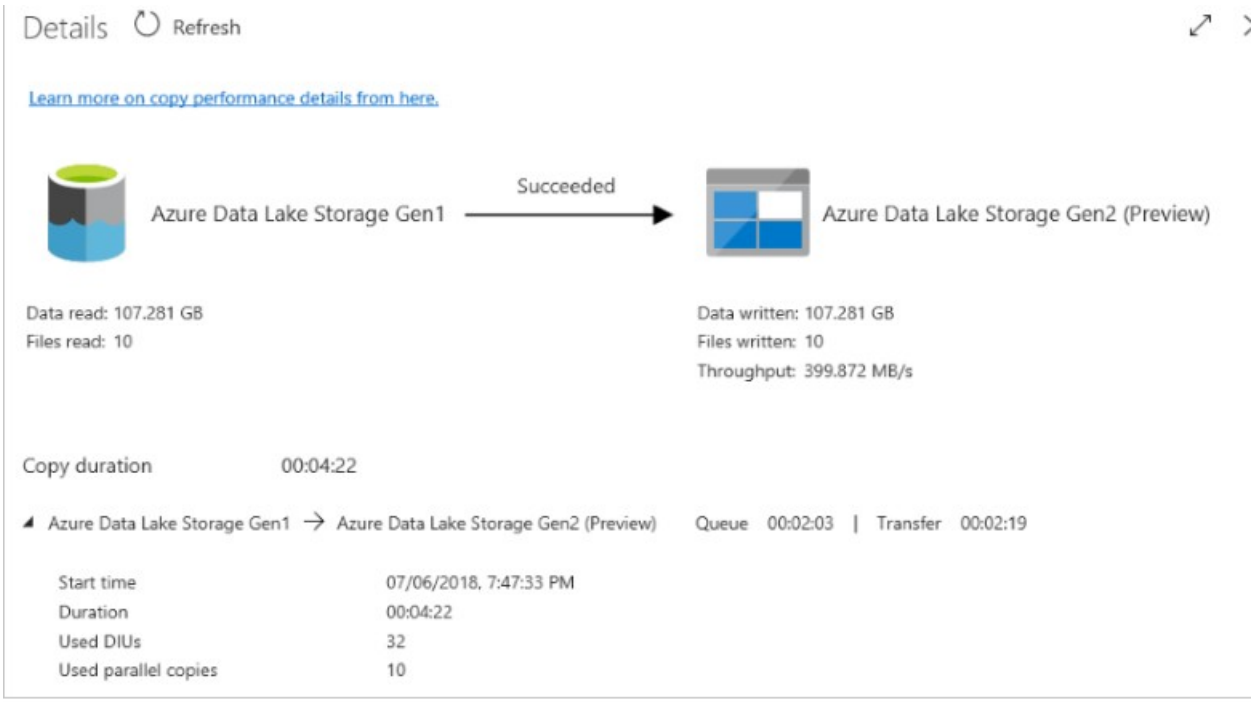
Screenshot of the Summary page.

15.To monitor the pipeline, on the deployment page, select **Monitor**.



Screenshot of the deployment page.

You can monitor details like how much data is copied from the source to the sink, data throughput, execution steps and their duration, and configurations.



Screenshot of the Details page.
After the transfer is complete, you can use Azure Storage Explorer to verify that the data has been copied into your Data Lake Storage Gen2 account.