Create an Azure Data Lake Storage Gen2

Upload Data into Azure Data Lake Storage Gen2

Azure Data Lake Storage Generation 2 (Gen2) is a data lake solution explicitly designed for enterprises to run large scale analytical workloads in the cloud. It takes the core capabilities from Azure Data Lake Storage Gen1 including file system semantics and security and scale and combines it with the low-cost, highly available capabilities of Azure Blob Storage.

You want to show Contoso the different ways in which you can upload data into Azure Data Lake Storage Gen2. This will include examples of how you can perform ad-hoc data loads to integrating the upload capability into applications. You will also demonstrate how you can use Azure Data Factory to copy data into Azure Data Lake Gen 2.

Learning Objectives

In this module you will:

- Create an Azure Data Lake Gen2 Store using Powershell
- Upload data into the Data Lake Storage Gen2 using Azure Storage Explorer
- Copy data from an Azure Data Lake Store Gen1 to an Azure Data Lake Store Gen2

Create an Azure Data Lake Storage Gen2 Account

Before uploading or transferring data into a data lake, you need to create one. Using the Azure portal, you are able to provision an Azure Data Lake Storage Gen2 within minutes.

NOTE

If you don't have an Azure account, or prefer not to do the exercise in your account, you can read through the following instructions to understand the steps involved in creating an Azure Data Lake Store.

Create a new Resource Group

Open up the PowerShell console within the Azure Portal and type out the following:

```
$resourceGroup = "mslearn-datalake-test"
$location = "westus2"
New-AzResourceGroup -Name $resourceGroup -Location $location
```

Create a Data Lake Storage Account Gen2

In the PowerShell console within the Azure Portal and type out the following:

```
$location = "westus2"

New-AzStorageAccount -ResourceGroupName $resourceGroup `
  -Name "dlakedata001" `
  -Location $location `
  -SkuName Standard_LRS `
```

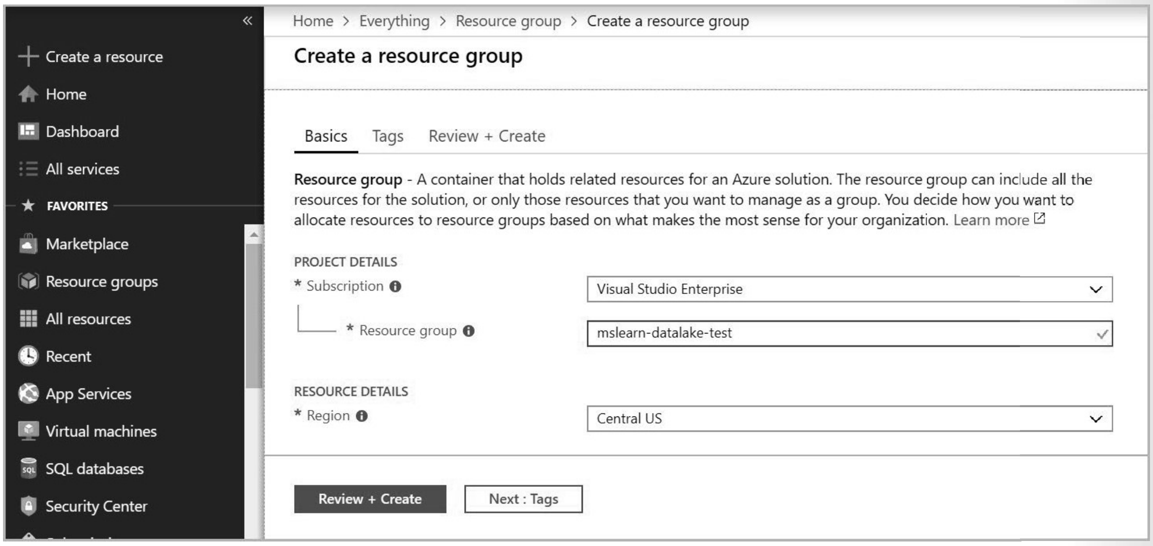
```
-Kind StorageV2 `
-EnableHierarchicalNamespace $True
```

If these steps fail, use the following steps in the Azure Portal:

Create a new Resource Group

First, create a new resource group to hold the data lake storage. A resource group will let you administer related services and applications together. It also makes it easier to clean up resources when you are done with this module. To create a resource group in the Azure portal, follow these steps:

1. Sign in to the **Azure portal**[3] using your account.
2. Select **Create a resource** from the left sidebar.
3. In the search box, type "**Resource**" and select **Resource group** from the results.
4. Click the **Create** button to add a new resource group.
5. In the **Basics** tab, select the appropriate subscription you want to work in.
6. Set the name of the resource group to **"mslearn-datalake-test"** without the quotes.
7. Choose the region (location) for the resource group - you typically want to choose a location close to you or to the data you are going to work with.



Create a resource group

Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization.

PROJECT DETAILS

* Subscription: Visual Studio Enterprise

* Resource group: mslearn-datalake-test

RESOURCE DETAILS

* Region: Central US

8. Click the **Review + Create** button and then **Create** on the review screen.

Resource group creation is fast, you can pin the resource group to your dashboard to make it easy to find later if you like.

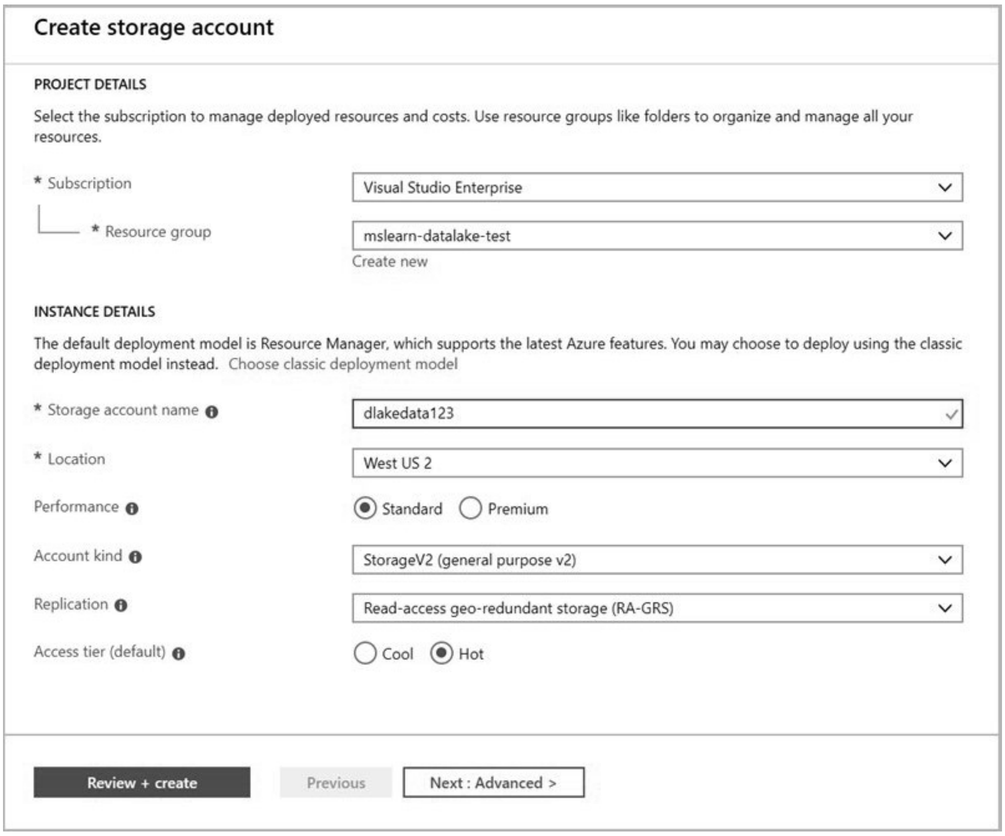
Create a Data Lake Storage Account Gen2

Creating an Azure Data Lake Storage Account Gen2 is the same as creating an Azure Blob Store, there's just one setting that is different. To create the data lake, perform the following steps:

1. In the Azure portal, choose **Create a resource** from the left sidebar.

[3] https://portal.azure.com?azure-portal=true

2. Select **Storage**, and choose **Storage account**.
3. Select your **Subscription** and the **Resource group** you created earlier (**mslearn-datalake-test**).
4. Enter a unique name for your storage account. It must be unique across all of Azure, so for example use the prefix “dlakedata” with some numbers. You might have to try a few variations to find a unique name. The portal will display a green checkmark next to the name when you have a valid entry.
5. Select a location - you typically want to select a region near where the data consumption will occur. Since this is an example, just select a location near you.
6. Make sure the Account kind is **StorageV2 (general-purpose V2)**. The rest of the values can be left as their defaults.



Create storage account

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

* Resource group: mslearn-datalake-test

Create new

INSTANCE DETAILS

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. Choose classic deployment model

* Storage account name: dlakedata123 ✓

* Location: West US 2

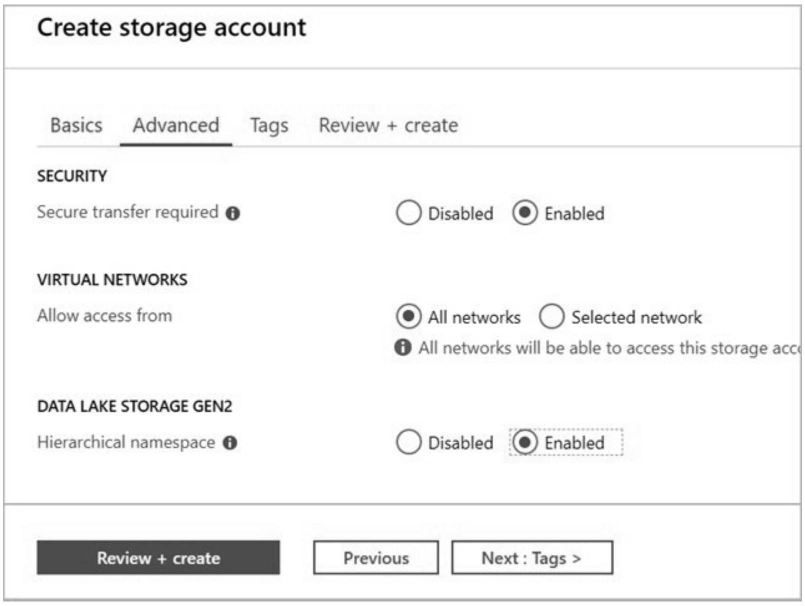
Performance: ● Standard ○ Premium

Account kind: StorageV2 (general purpose v2)

Replication: Read-access geo-redundant storage (RA-GRS)

Access tier (default): ○ Cool ● Hot

7. Select **Next: Advanced >**
8. In the **Data Lake Storage Gen2 (preview)** section set **Hierarchical namespace** to **Enabled**.
9. Click **Review + Create** to create the storage account.



Create storage account

SECURITY

Secure transfer required ○ Disabled ● Enabled

VIRTUAL NETWORKS

Allow access from ● All networks ○ Selected network

All networks will be able to access this storage acc

DATA LAKE STORAGE GEN2

Hierarchical namespace ○ Disabled ● Enabled

10. Once the creation details have been validated, click the **Create** button to start deployment.

Wait for a few moments for the deployment to complete, once you receive the message "Your deployment is complete", click **Go to resource** to confirm the deployment.

Upload Data using Azure Storage Explorer

If you need to perform ad-hoc data transfers into an Azure Data Lake Store, you can use the **Azure Storage Explorer** to upload your files.

Azure Storage Explorer is a free application available for Windows, macOS, and Linux. The app is designed to manage unstructured data in Azure such as tables, blobs, queues, and files. It also supports data in Azure Cosmos DB and Azure Data Lake Storage, which is what we'll use it for here.

NOTE
If you don't have an Azure account, or prefer not to do the exercise in your account, you can read through the instructions to understand the steps involved to install and use the Azure Storage Explorer tool.

Download and Install Azure Storage Explorer

Start by installing the **Azure Storage Explorer**[4].

Using Azure Storage Explorer

Once installed, you can use Azure Storage Explorer to perform several operations against data in your Azure Storage account including your data lake. Here are some features of the tool.

- You can upload files or folders from your local computer into Azure Storage.
- You can download cloud-based data to your local computer.
- You can copy or move files and folders around in the storage account.

[4] https://azure.microsoft.com/features/storage-explorer

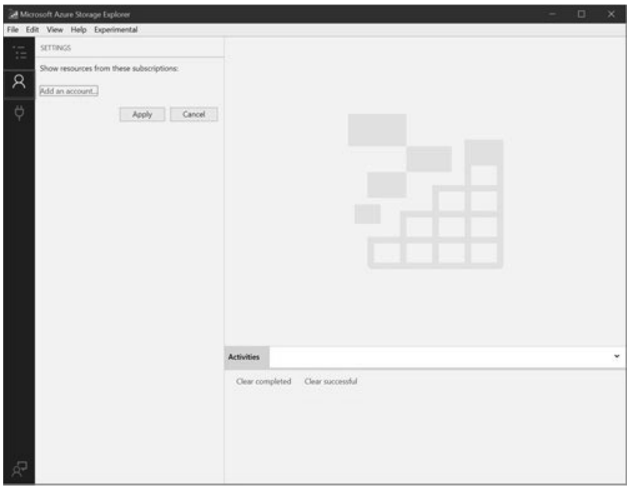
- You can delete data from the storage account.

Let's look at some of these capabilities.

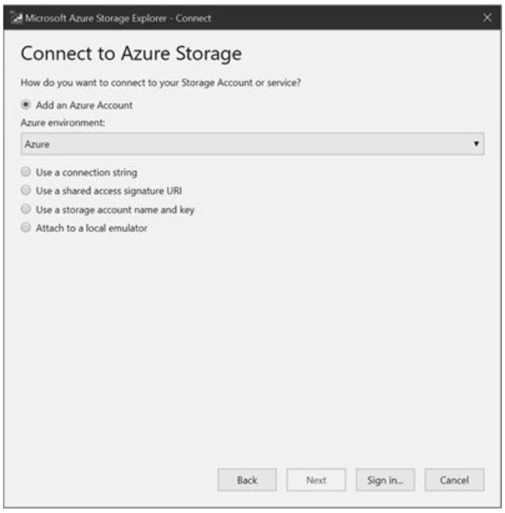
Connect the Azure Storage Explorer to your Azure account

Start by adding your Azure account.

1. Click on the Account button icon in the left sidebar.


Microsoft Azure Storage Explorer
File Edit View Help Experimental
SETTINGS
Show resources from these subscriptions:
Add an account...
Apply
Cancel
Activities
Clear completed
Clear successful

2. There are multiple options for connecting to your storage account.
   - Sign in with your Azure account to provide access to all your subscriptions.
   - Use a connection string to access a specific Azure Storage account.
   - Use a storage account name and access key.



Connect to Azure Storage

How do you want to connect to your Storage Account or service?

● Add an Azure Account

Azure environment: Azure

○ Use a connection string

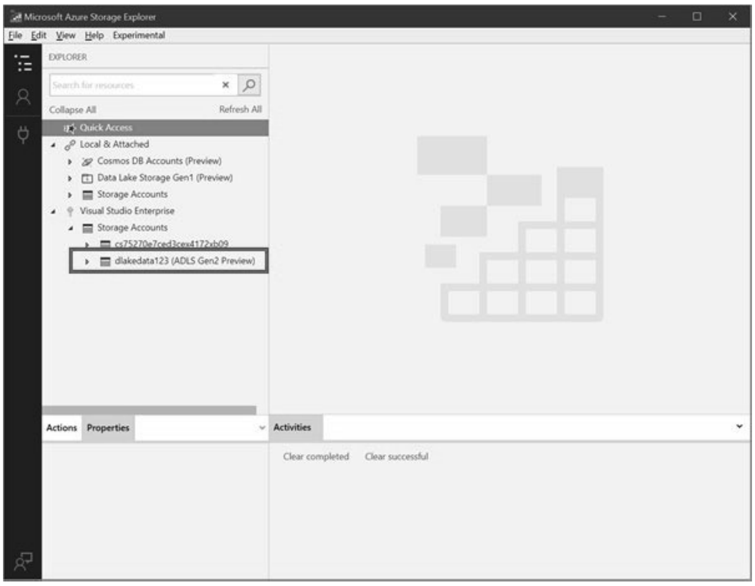
○ Use a shared access signature URI

○ Use a storage account name and key

○ Attach to a local emulator

3. Once you sign in, you can select the subscriptions you want to work with. Make sure to select the one you created the Azure Storage account in.

The app then shows a tree of storage areas you can work with from your subscriptions. You should see your Azure Storage account in the list.



Create a filesystem using Azure Storage explorer

Blobs are always uploaded into folders. This allows you to organize groups of blobs much like you organize files on your computer.

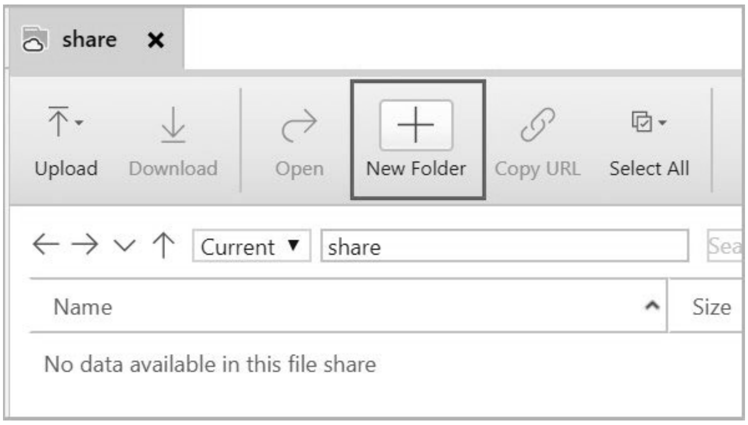
When working with Azure Data Lake, you start by creating a *filesystem*. This defines the specific container in Blob storage that will hold your data lake. You can then create folders and files within this dedicated area.

1. In Azure Storage Explorer, expand your subscription, and then expand storage accounts.
2. Expand the storage account that you have created in the previous unit, and click on **Blob Containers**.
3. Right-click **Blob Containers** and click **Create Blob Container**.
4. In the text box that appears below **Blob Containers**, type **salesdata**.
5. Once created, click on **salesdata**.

Create a folder in Storage Container using Azure Storage Explorer

Adding a folder provides a hierarchical structure for managing your data. You can create multiple levels in the account. However, you must ensure that parent folders exist before you create children.

1. Select the **New Folder** button from the menu running across the top.



2. For the folder name, enter **“sample”** without the quotes, and then select **OK** to create the directory.

   You may get a message box in Azure Storage Explorer that states. "Your view may be out of date. Do you want to refresh?". If so, click **Yes**.

3. Double-click on the new folder in the UI - this will traverse into the folder, which should be empty.
4. Create another folder named **“data”**.

Create a sample text file

To provide some sample data to work with, create a local text file on your computer named ”sales.txt” and paste the following text into the file.

```
#salaries Details
#Company Information
#Fields : Date company employee Salaries
01-01-2019  c1   e1 1000
01-01-2019  c2   e2 2000
01-01-2019  c1   e3 4000
01-01-2019  c2   e4 2000
01-01-2019  c1   e5 5000
01-01-2019  c3   e6 7000
```

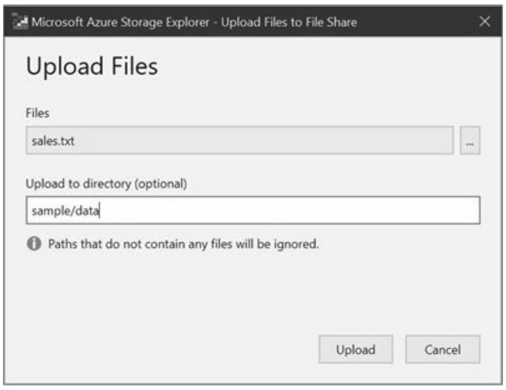
We'll upload this data file in various ways. Keep in mind that this is a *simple* example - you would typically be populating your data lake with much larger data samples from a variety of sources.

Upload a file

You can upload files and folders from your local machine to directories in your file share right from the tool.

1. In Azure Storage Explorer, double-click the folder named **data**.
2. In the top menu, select **Upload**. This gives you the option to upload a folder or a file.
3. Select **Upload Files**.
4. Select the ”sales.txt” file you created earlier as the file to upload

5. In **Upload to a directory** dialog box, ensure the Destination directory states **“sample/data”**, and then select **Upload**.



Microsoft Azure Storage Explorer - Upload Files to File Share

Upload Files

Files

sales.txt

Upload to directory (optional)

sample/data

Paths that do not contain any files will be ignored.

When you are finished, the file appears in the list.

Download a file

To download a copy of a file from your file share, right-click the file, and then select **Download**. Choose where you want to put the file on your local machine, and then select **Save**. The progress of the download appears in the **Activities** pane at the bottom of the window.

Copy Data from Data Lake Storage Gen1 to Data Lake Storage Gen2

Azure Data Factory is a cloud-based data integration service that creates workflows in the cloud for orchestrating batch data movement and transformations. Using Azure Data Factory, you can create and schedule workflows (called *pipelines*) to ingest data from disparate data stores. The data can then be processed and transformed with services such as:

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

There are many data orchestration tasks that can be conducted using Azure Data Factory. In this exercise, we'll copy data from Azure Data Lake Storage Gen1 to Azure Data Lake Storage Gen2.

NOTE
If you don't have an Azure account, or prefer not to do the exercise in your account, you can read through the following instructions to understand the steps involved using Azure Data Factory to copy data into an Azure Data Lake.

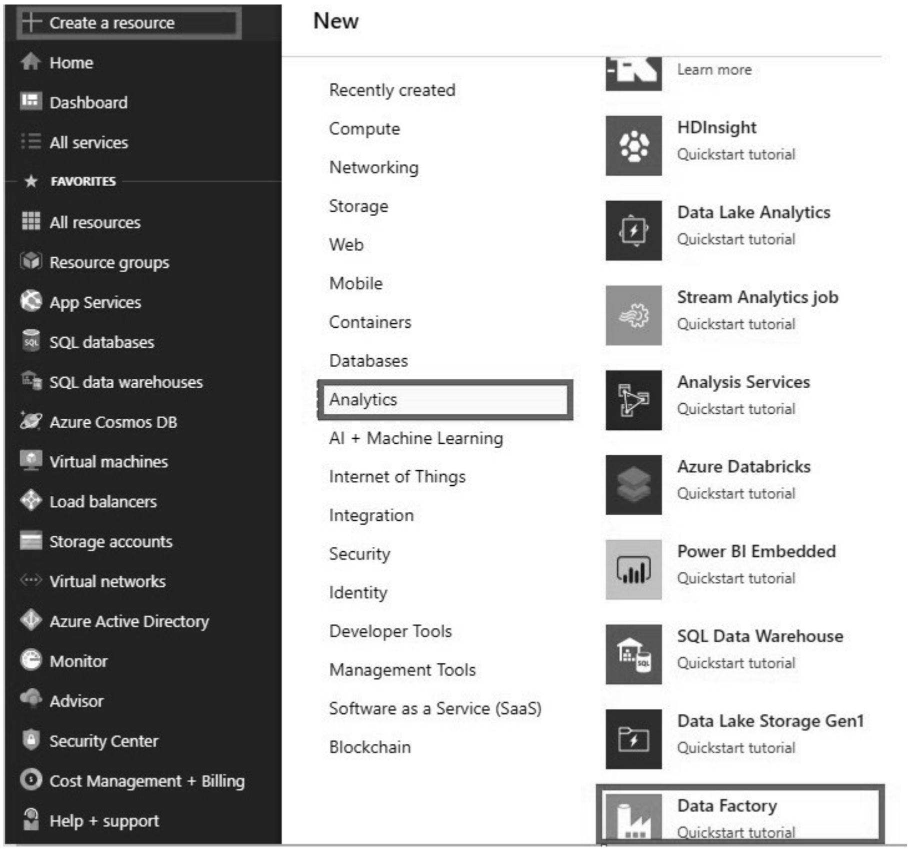
Create an Azure Data Factory instance

The first step is to provision an instance of Azure Data Factory in the Azure portal.

1. Sign into the **Azure portal**[5].

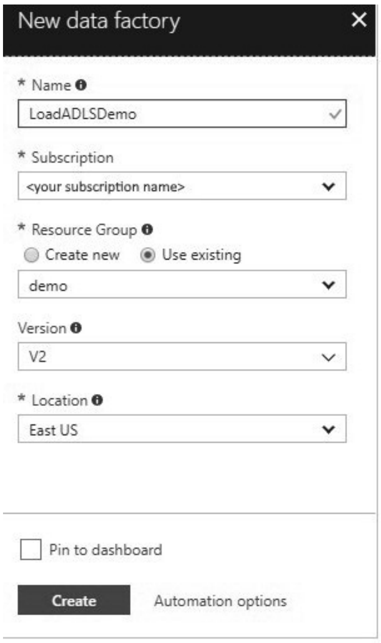
[5] https://portal.azure.com?azure-portal=true

2. On the left menu, select **New** > **Data + Analytics** > **Data Factory**:



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

- Enter a globally unique name for your Azure data factory. Try using your initials with a suffix such as **ADFTutorialDataFactory**.
- In the Subscription drop-down list, click on your subscription
- In the resource group drop-down list, select **mslearn-datalake-test**
- Select **V2** for the version.
- Select the location for the data factory. Only supported locations are displayed in the drop-down list. The data stores that are used by the data factory can be in other locations and regions.



New data factory

* Name
LoadADLSDemo

* Subscription
<your subscription name>

* Resource Group
○ Create new ● Use existing
demo

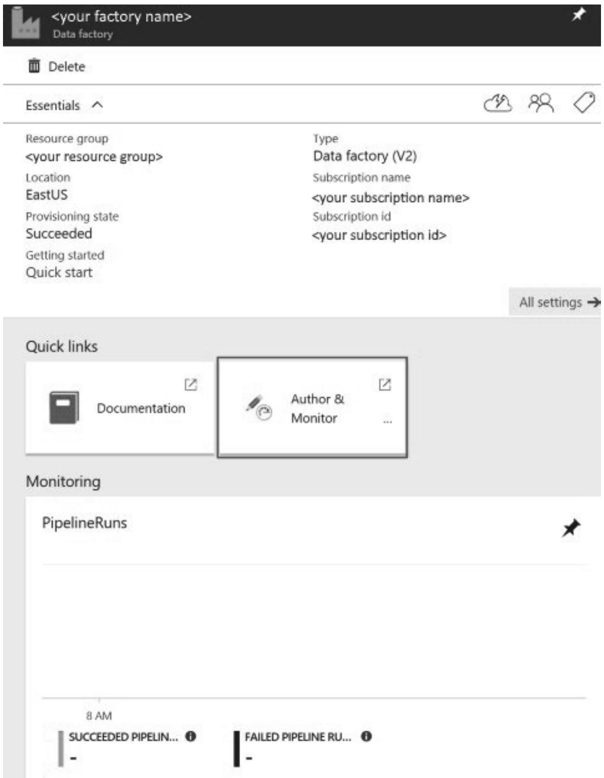
Version
V2

* Location
East US

□ Pin to dashboard

4. Select **Create**.

After creation is complete, navigate to the new data factory. You should see the **Data Factory** home page.



<your factory name>
Data factory

Resource group
<your resource group>
Location
EastUS
Provisioning state
Succeeded
Getting started
Quick start

Type
Data factory (V2)
Subscription name
<your subscription name>
Subscription id
<your subscription id>

[!IMPORTANT]
You will need an Azure Data Lake Storage Gen1 account with data in it. If you do not have this perform the following steps.

Create a Data Lake Storage Gen 1 Account

1. In the left sidebar, click on **+ Create a new resource**.
   In the **New** blade, click **Storage** and then click **Data Lake Storage Gen1**.
2. In the Name text box, type **dlsgen1XXX** replace "XXX" with numbers of your choice. A green tick should appear confirming that the name is unique.
3. In the Subscription drop-down list, click on your subscription.
4. In the resource group drop-down list, select **mslearn-datalake-test**.
5. Select a location - you typically want to select a region near where the data consumption will occur. In this example, select a location near you.
6. Click **Create**.

Create a second sample text file

To provide some sample data to work with, create a local text file on your computer named **salesUK.txt** and 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
```

We'll upload this data file in various ways. Keep in mind that this is a *simple* example - you would typically be populating your data lake with much larger data samples from a variety of sources.

Upload a file into data lake storage Gen 1 account

1. In the Azure portal, search for the Data Lake Storage Gen1 service you created (**dlsgen1XXX**).
2. In the overview blade, click **Data Explorer**
3. In the Data Explorer blade, click on the **Upload** icon.
4. In the Upload file blade, click on the browse icon, browse to the folder, and select your **salesUK.txt** file, click on the button **Add selected files**. Conformation that the upload has completed is when the states column states **completed**.
5. Close the Upload files blade.

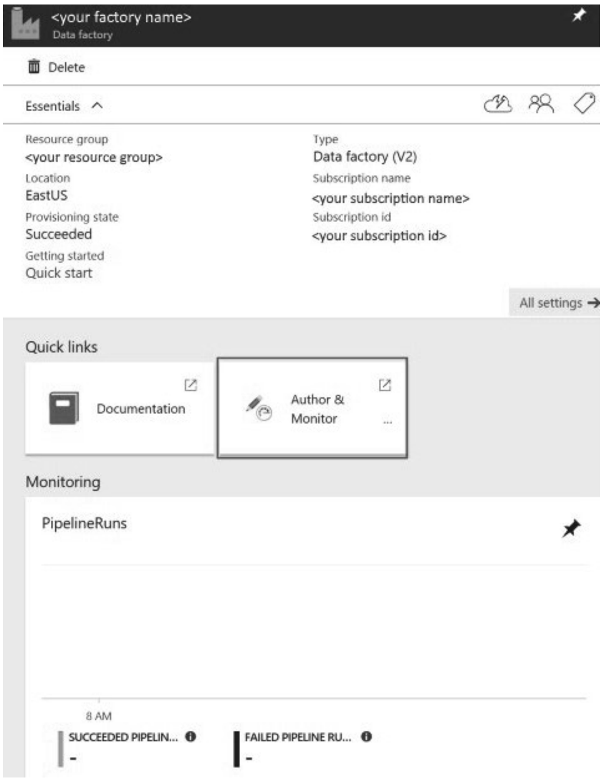
Setting permissions on the data lake storage Gen 1 account

Next, you need to set permissions to enable the Azure Data Factory instance to access the data in your Data Lake Store Gen 1.

1. In the Azure portal, search for the Data Lake Storage Gen1 named **dlsgen1XXX** that you created.
2. In the overview blade, click **Access control (IAM)**
3. In the Access Control (IAM) blade, click on the **+ Add Role Assignment** button.
4. In the Add Role Assignment blade, select **Owner** for the **Role**.
5. Select the text box under **Select** and type in the Azure Data Factory instance name you created.
6. Click **Save**.
7. Close the **Access control (IAM)** blade.

Load data into Azure Data Lake Storage Gen2

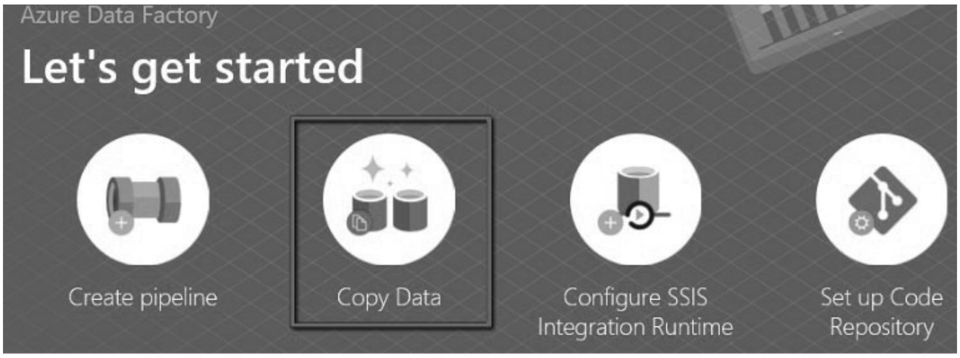
1. In the Azure portal, go to your data factory. You see the **Data Factory** home page.
2. Select the **Author & Monitor** tile to launch the Data Integration Application in a separate tab.



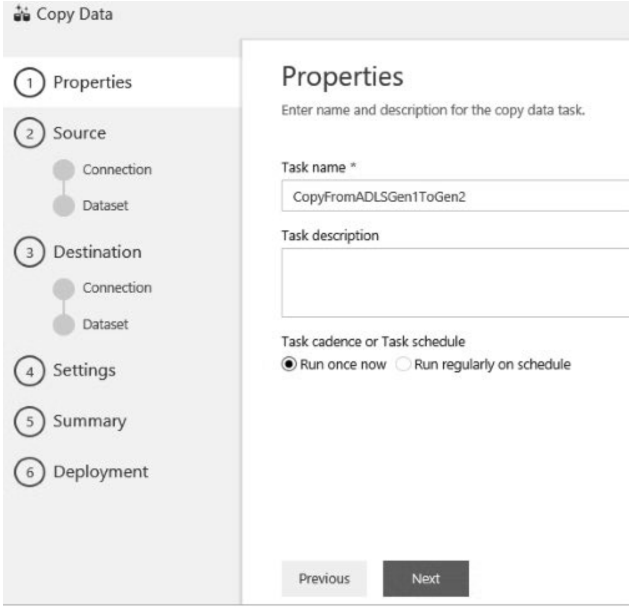
<your factory name>
Data factory

Resource group: <your resource group>
Location: EastUS
Provisioning state: Succeeded
Getting started: Quick start
Type: Data factory (V2)
Subscription name: <your subscription name>
Subscription id: <your subscription id>

3. In the **Get started** page, select the **Copy Data** tile to launch the Copy Data tool.



4. In the **Properties** page, specify **CopyFromADLSGen1ToGen2** for the **Task name** field, set the cadence to "once", and select **Next**:



Properties

Enter name and description for the copy data task.

Task name *

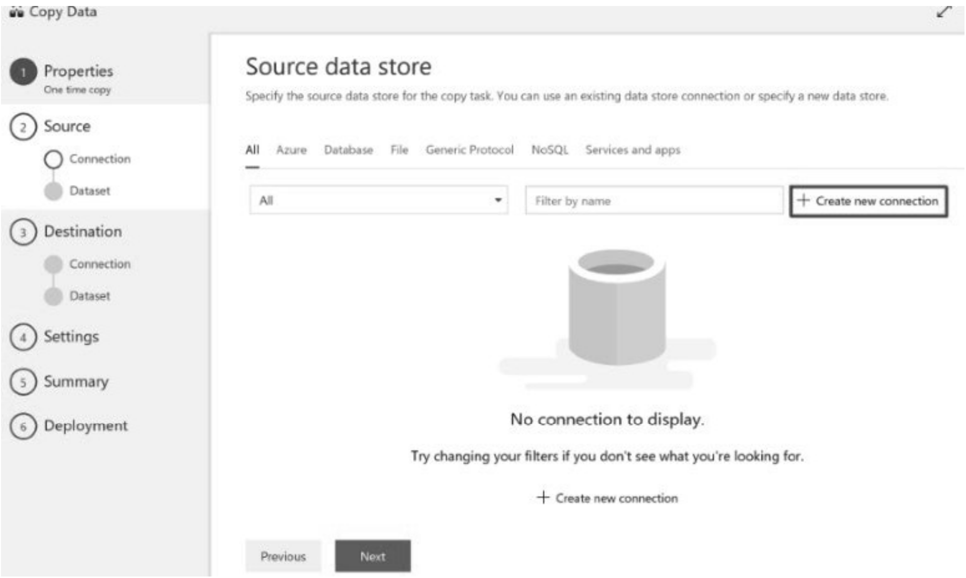
CopyFromADLSGen1ToGen2

Task description

Task cadence or Task schedule

● Run once now ○ Run regularly on schedule

5. In the **Source data store** page, click **+ Create new connection**.



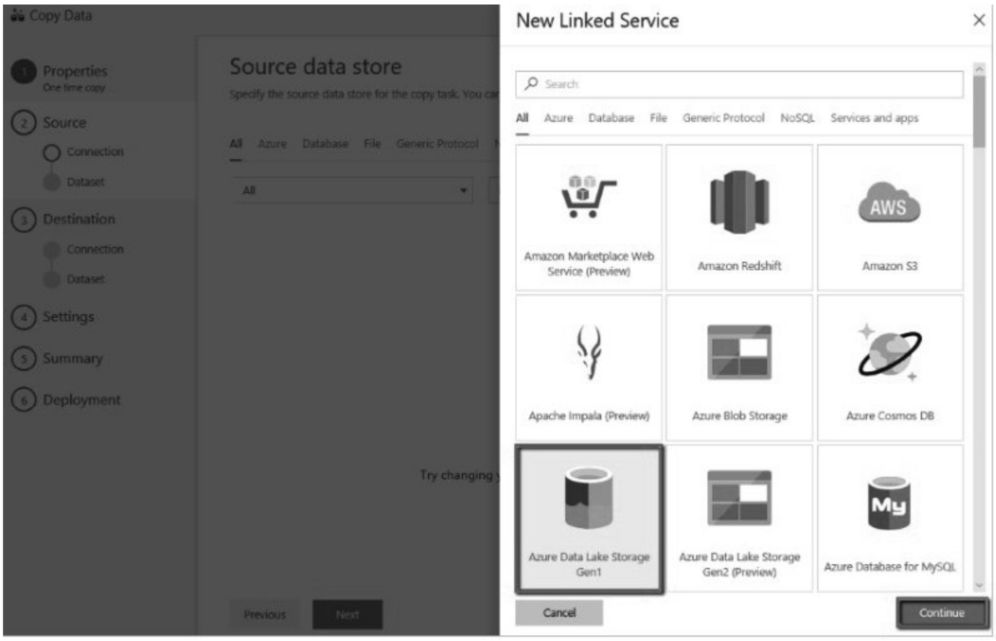
Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

No connection to display.

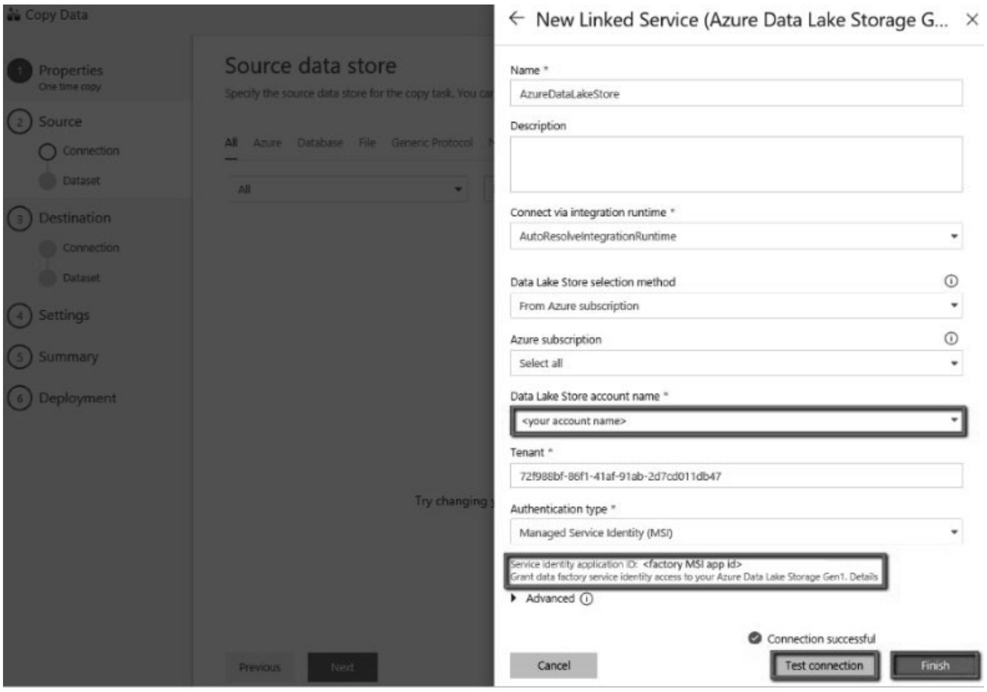
Try changing your filters if you don't see what you're looking for.

6. Select **Azure Data Lake Storage Gen1** from the connector gallery, and select **Continue**.



7. In the **Specify Azure Data Lake Storage Gen1 connection** page, do the following steps:

- Select your Data Lake Storage Gen1 for the name and specify or validate the Tenant.
- Click **Test connection** to validate the settings, then select **Finish**.
- You'll see a new connection gets created. Select **Next**.



New Linked Service (Azure Data Lake Storage G...

Name *: AzureDataLakeStore

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Data Lake Store selection method: From Azure subscription

Azure subscription: Select all

Data Lake Store account name *: <your account name>

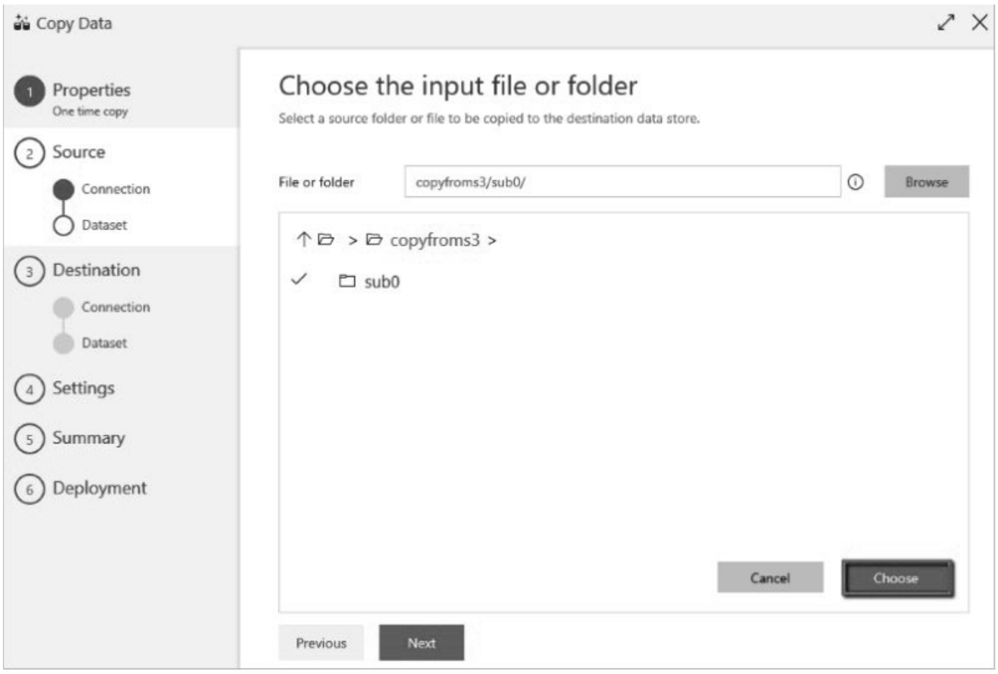
Tenant *: 72f988bf-86f1-41af-91ab-2d7cd011db47

Authentication type *: Managed Service Identity (MSI)

Service identity application ID: <factory MSI app id>
Grant data factory service identity access to your Azure Data Lake Storage Gen1. Details

Connection successful

8. In the **Choose the input file or folder** page, browse to the folder and file that you want to copy over. Select the folder/file, select **Choose**.

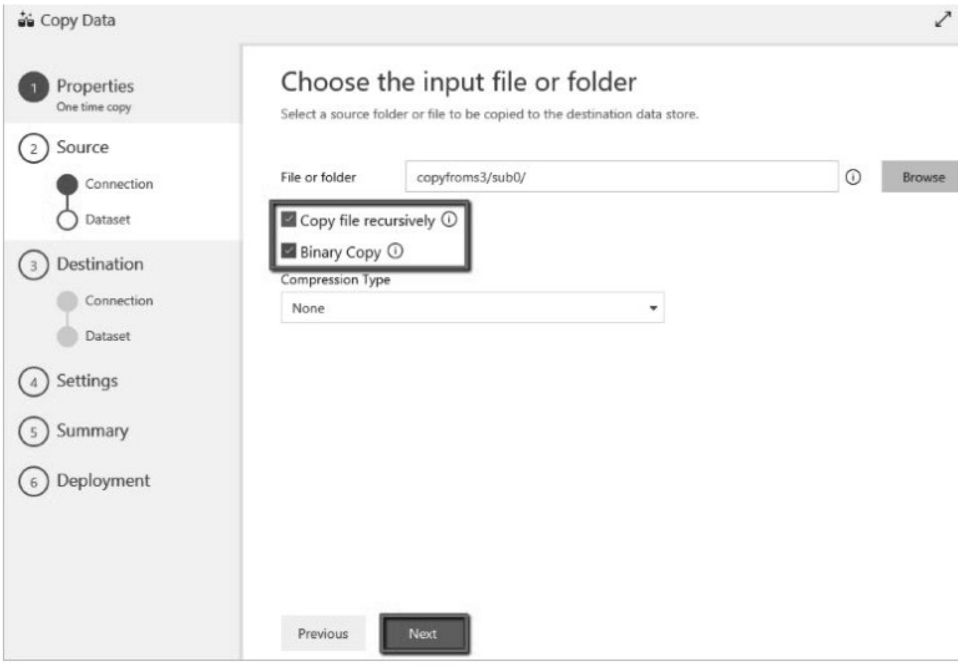


Choose the input file or folder

Select a source folder or file to be copied to the destination data store.

File or folder: copyfroms3/sub0/

9. Specify the copy behavior by checking the **Copy files recursively** and **Binary copy** options. Select **Next**:



Choose the input file or folder

Select a source folder or file to be copied to the destination data store.

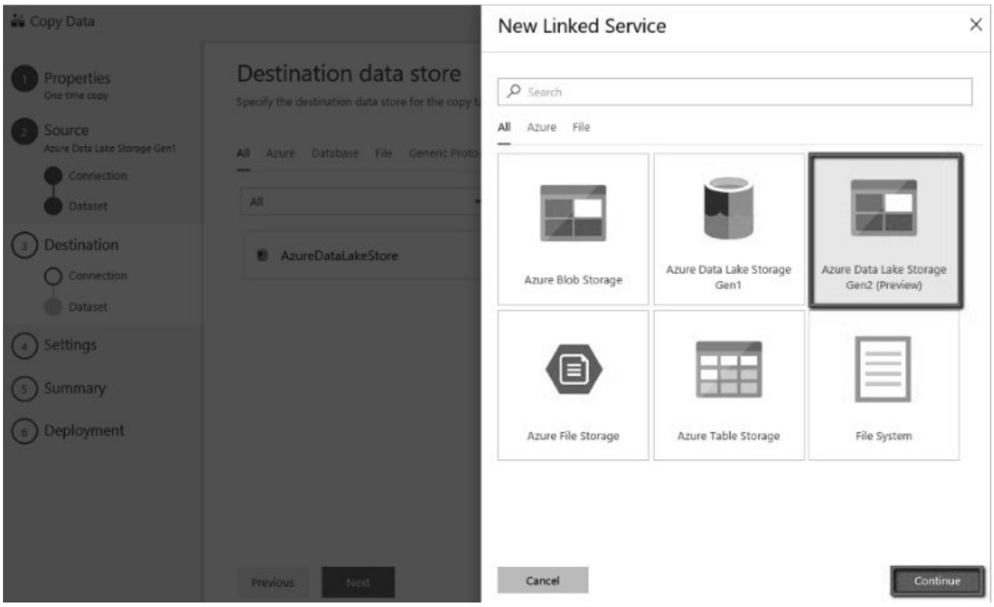
File or folder: copyfroms3/sub0/

☑ Copy file recursively

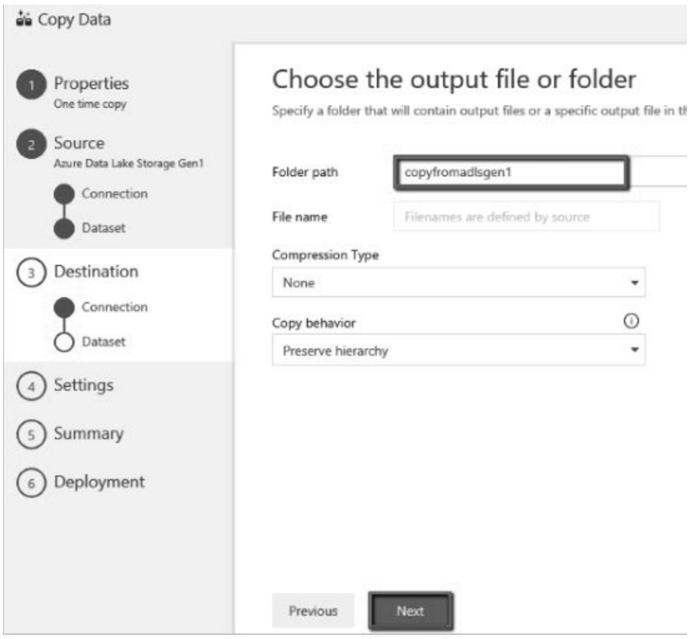
☑ Binary Copy

Compression Type: None

10. In the **Destination data** store page, click **+ Create new connection**, select **Azure Data Lake Storage Gen2 (Preview)**, and click **Continue**.



11. In the **Specify Azure Data Lake Storage Gen2 connection** page, do the following steps:
    - Select your Data Lake Storage Gen2 capable account from the "Storage account name" drop-down list.
    - Select **Finish** to create the connection. Then select **Next**.
12. In the **Choose the output file or folder** page, enter **copyfromadlsgen1** as the output folder name, and select **Next**.



Choose the output file or folder

Specify a folder that will contain output files or a specific output file in th

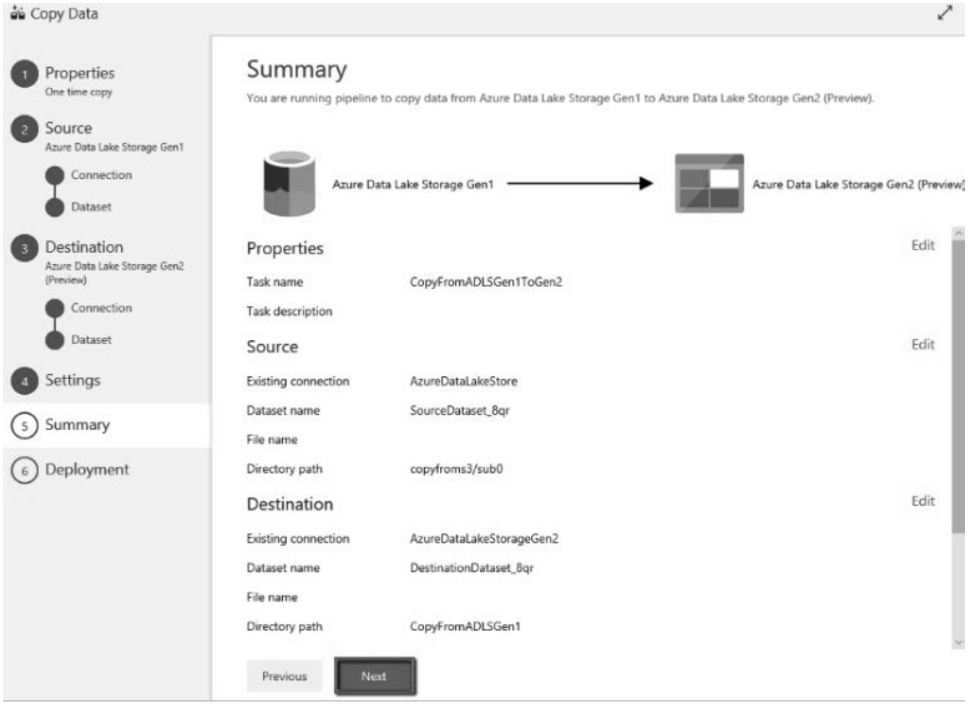
Folder path: copyfromadlsgen1

File name: Filenames are defined by source

Compression Type: None

Copy behavior: Preserve hierarchy

13. In the **Settings** page, select **Next** to use the default settings.
14. Review the settings, and select **Next**.



Summary

You are running pipeline to copy data from Azure Data Lake Storage Gen1 to Azure Data Lake Storage Gen2 (Preview).

Properties

Task name: CopyFromADLSGen1ToGen2

Task description:

Source

Existing connection: AzureDataLakeStore

Dataset name: SourceDataset_8qr

File name:

Directory path: copyfroms3/sub0

Destination

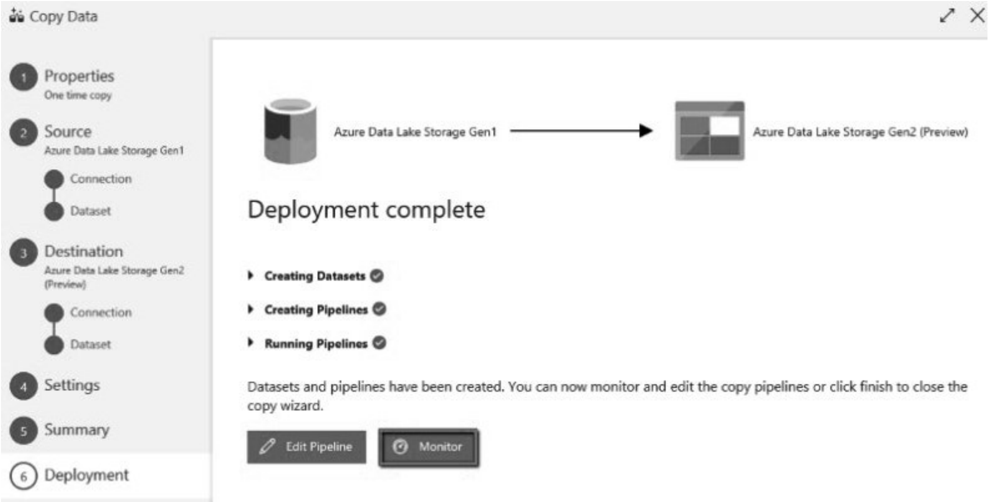
Existing connection: AzureDataLakeStorageGen2

Dataset name: DestinationDataset_8qr

File name:

Directory path: CopyFromADLSGen1

15. In the **Deployment page**, select **Monitor** to monitor the pipeline.

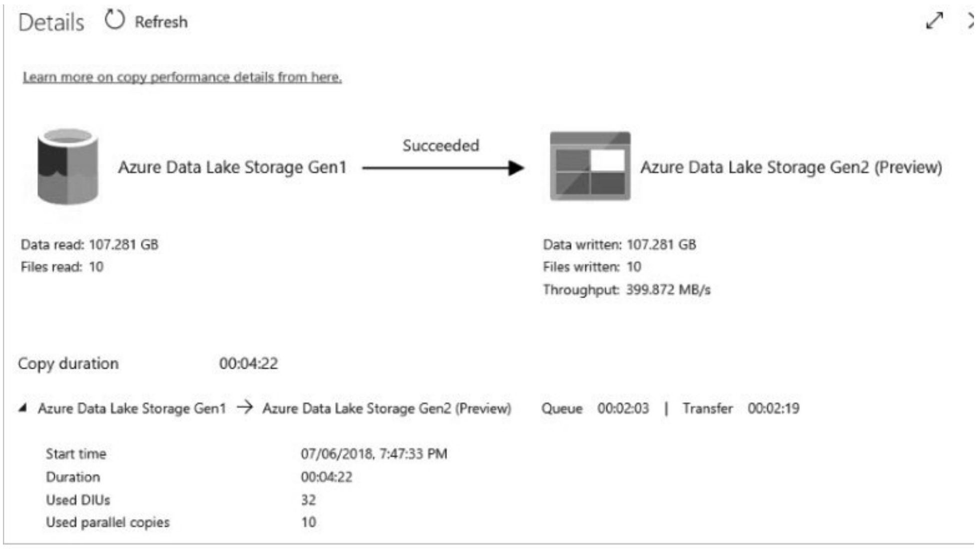


Deployment complete

- Creating Datasets ✓
- Creating Pipelines ✓
- Running Pipelines ✓

Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close the copy wizard.

You can monitor details like the volume of data copied from the source to the sink, data throughput, execution steps with the corresponding duration, and used configurations.


Details
Refresh
Learn more on copy performance details from here.
Azure Data Lake Storage Gen1
Succeeded
Azure Data Lake Storage Gen2 (Preview)

Data read: 107.281 GB
Files read: 10

Data written: 107.281 GB
Files written: 10
Throughput: 399.872 MB/s

Copy duration 00:04:22

Azure Data Lake Storage Gen1 → Azure Data Lake Storage Gen2 (Preview) Queue 00:02:03 | Transfer 00:02:19

Start time	07/06/2018, 7:47:33 PM
Duration	00:04:22
Used DIUs	32
Used parallel copies	10

Once the transfer is complete, you can verify the data has been copied into your Data Lake Storage Gen2 account with the Azure Storage Explorer.

Summary

Azure Data Lake Storage Gen2 provides a cloud storage service that is highly available, secure, durable, scalable, and redundant. It's the most comprehensive data lake solution available in market today. Azure Data Lake Storage brings new efficiencies to process big data analytics workloads and can provide data to a multitude of compute technologies including HDInsight and Azure Databricks without needing to move the data around. Creating an Azure Data Lake Storage Gen2 data store should be one of your go-to tools in building a big data analytics solution.

Important