LAB - Incrementally copy new and changed files based on LastModifiedDate by using the Copy Data tool

In this tutorial, you'll use the Azure portal to create a data factory. You'll then use the Copy Data tool to create a pipeline that incrementally copies new and changed files only, from Azure Blob storage to Azure Blob storage. It uses `LastModifiedDate` to determine which files to copy.

After you complete the steps here, Azure Data Factory will scan all the files in the source store, apply the file filter by `LastModifiedDate`, and copy to the destination store only files that are new or have been updated since last time. Note that if Data Factory scans large numbers of files, you should still expect long durations. File scanning is time consuming, even when the amount of data copied is reduced.

In this tutorial, you'll complete these tasks:

- Create a data factory.
- Use the Copy Data tool to create a pipeline.
- Monitor the pipeline and activity runs.

Prerequisites

- **Azure subscription**: If you don't have an Azure subscription, create a free account before you begin.
- **Azure Storage account**: Use Blob storage for the source and sink data stores. If you don't have an Azure Storage account, follow the instructions in Create a storage account.

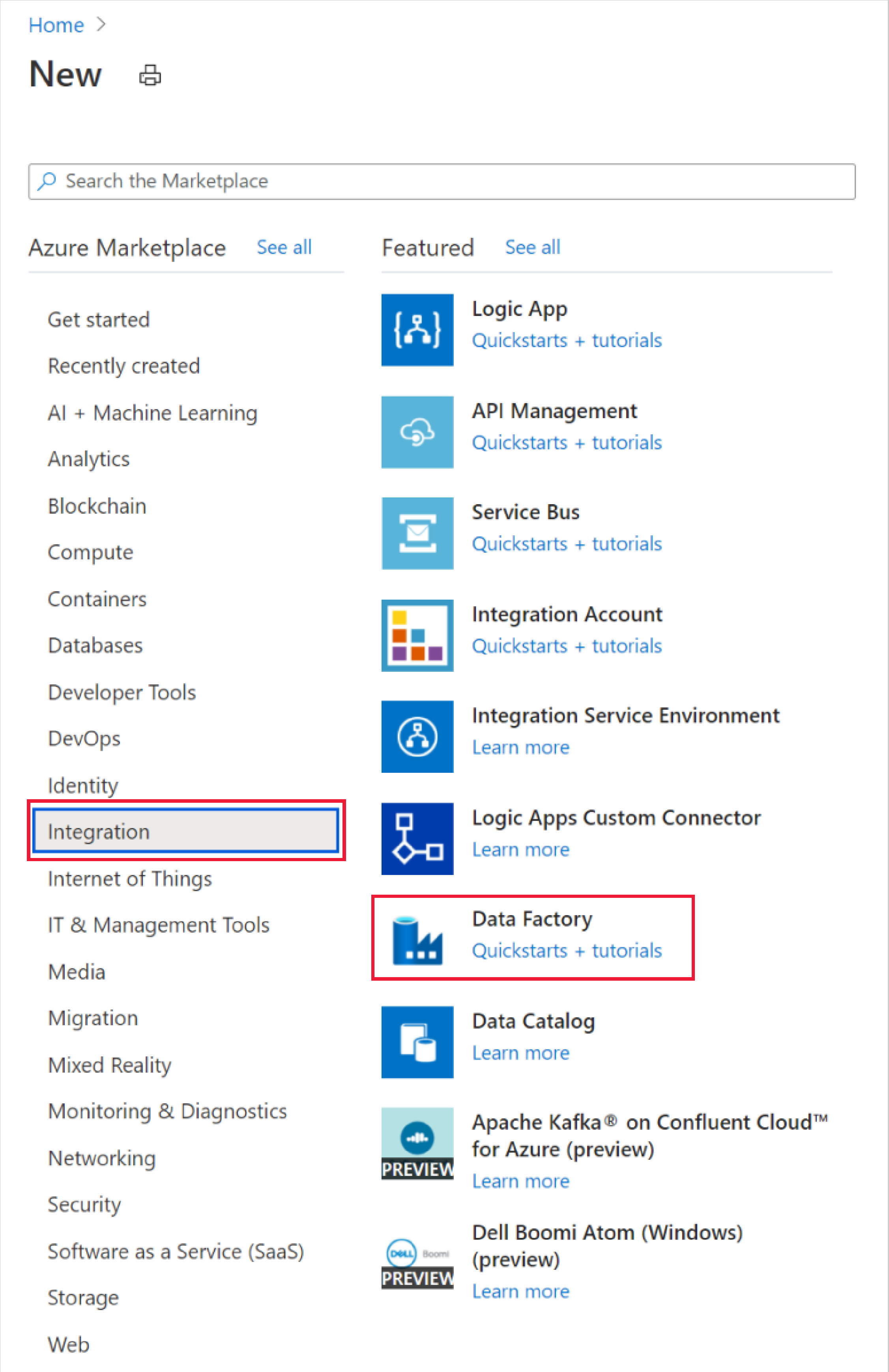
Create two containers in Blob storage

Prepare your Blob storage for the tutorial by completing these steps:

1. Create a container named **source**. You can use various tools to perform this task, like Azure Storage Explorer.
2. Create a container named **destination**.

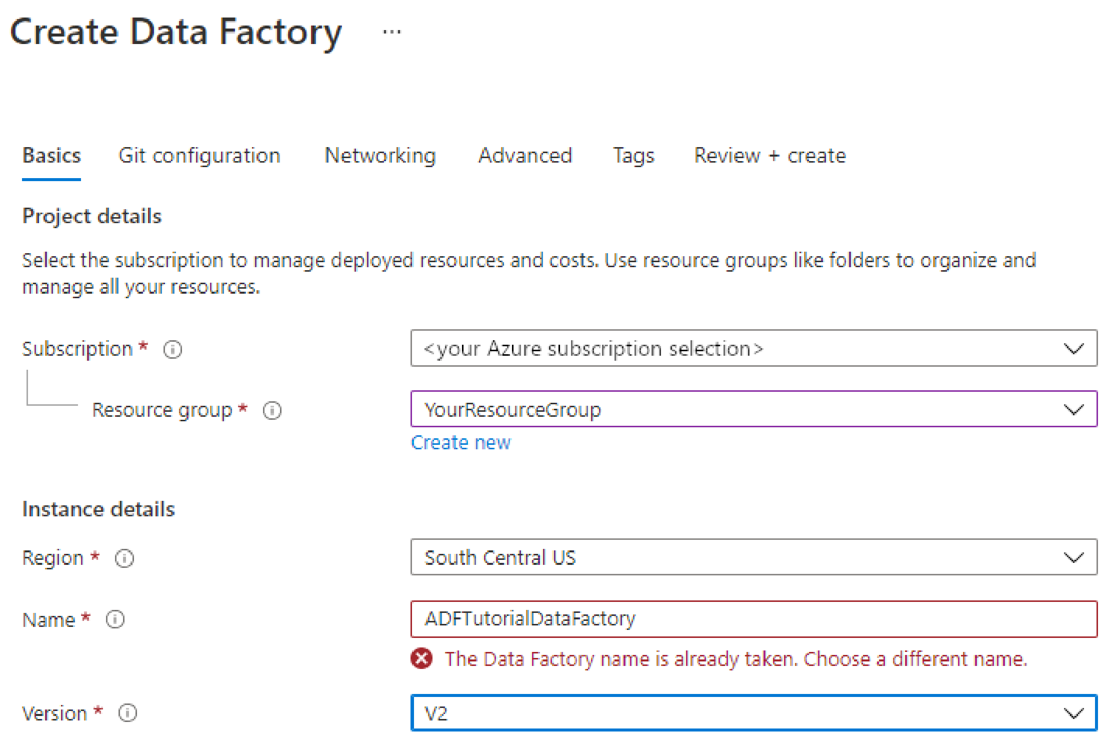
Create a data factory

1. In the left pane, select **Create a resource**. Select **Integration** > **Data Factory**:



2. On the **New data factory** page, under **Name**, enter **ADFTutorialDataFactory**.

   The name for your data factory must be globally unique. You might receive this error message:



Create Data Factory ...

Basics | Git configuration | Networking | Advanced | Tags | Review + create

Project details

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

Subscription * <your Azure subscription selection>

Resource group * YourResourceGroup

Create new

Instance details

Region * South Central US

Name * ADFTutorialDataFactory

The Data Factory name is already taken. Choose a different name.

Version * V2

If you receive an error message about the name value, enter a different name for the data factory. For example, use the name ***yourname*ADFTutorialDataFactory**. For the naming rules for Data Factory artifacts, see Data Factory naming rules.

3. Under **Subscription**, select the Azure subscription in which you'll create the new data factory.

4. Under **Resource Group**, take one of these steps:

   - Select **Use existing** and then select an existing resource group in the list.
   - Select **Create new** and then enter a name for the resource group.

   To learn about resource groups, see Use resource groups to manage your Azure resources.

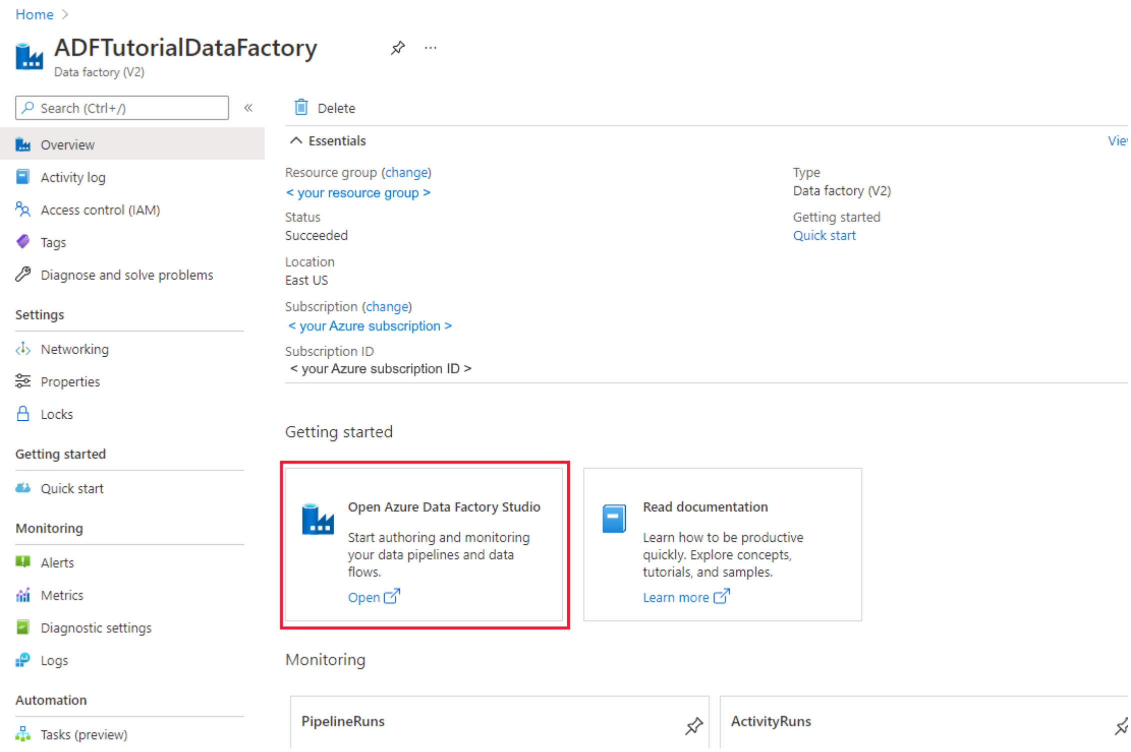
5. Under **Version**, select **V2**.

6. Under **Location**, select the location for the data factory. Only supported locations appear in the list. The data stores (for example, Azure Storage and Azure SQL Database) and computes (for example, Azure HDInsight) that your data factory uses can be in other locations and regions.

7. Select **Create**.

8. After the data factory is created, the data factory home page appears.

9. To open the Azure Data Factory user interface (UI) on a separate tab, select **Open** on the **Open Azure Data Factory Studio** tile:



Resource group (change)
< your resource group >

Status
Succeeded

Location
East US

Subscription (change)
< your Azure subscription >

Subscription ID
< your Azure subscription ID >

Type
Data factory (V2)

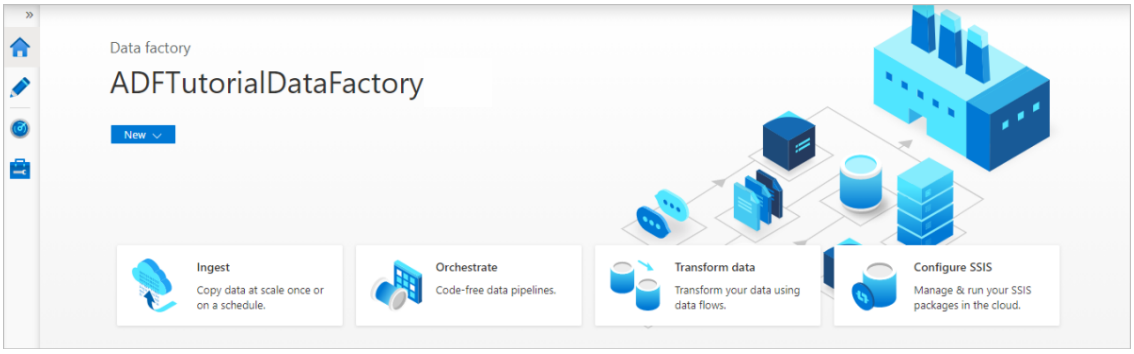
Getting started
Quick start

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

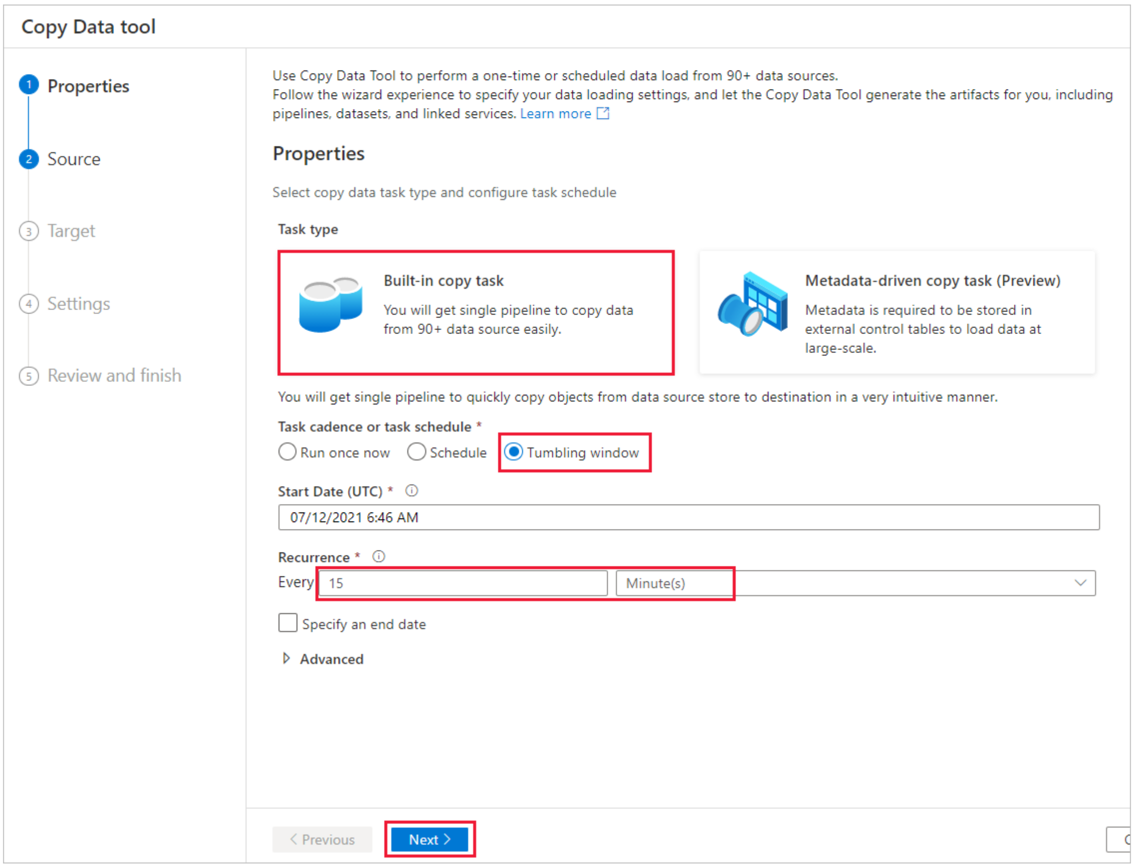
Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more

Use the Copy Data tool to create a pipeline

1. On the Azure Data Factory home page, select the **Ingest** tile to open the Copy Data tool:


Data factory
ADFTutorialDataFactory
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.

2. On the **Properties** page, take the following steps:
    1. Under **Task type**, select **Built-in copy task**.
    2. Under **Task cadence or task schedule**, select **Tumbling window**.
    3. Under **Recurrence**, enter **15 Minute(s)**.
    4. Select **Next**.



Copy Data tool

Use Copy Data Tool to perform a one-time or scheduled data load from 90+ data sources.
Follow the wizard experience to specify your data loading settings, and let the Copy Data Tool generate the artifacts for you, including pipelines, datasets, and linked services. Learn more

Properties

Select copy data task type and configure task schedule

Task type

Built-in copy task
You will get single pipeline to copy data from 90+ data source easily.

Metadata-driven copy task (Preview)
Metadata is required to be stored in external control tables to load data at large-scale.

You will get single pipeline to quickly copy objects from data source store to destination in a very intuitive manner.

Task cadence or task schedule *
○ Run once now ○ Schedule ● Tumbling window

Start Date (UTC) *
07/12/2021 6:46 AM

Recurrence *
Every 15 Minute(s)

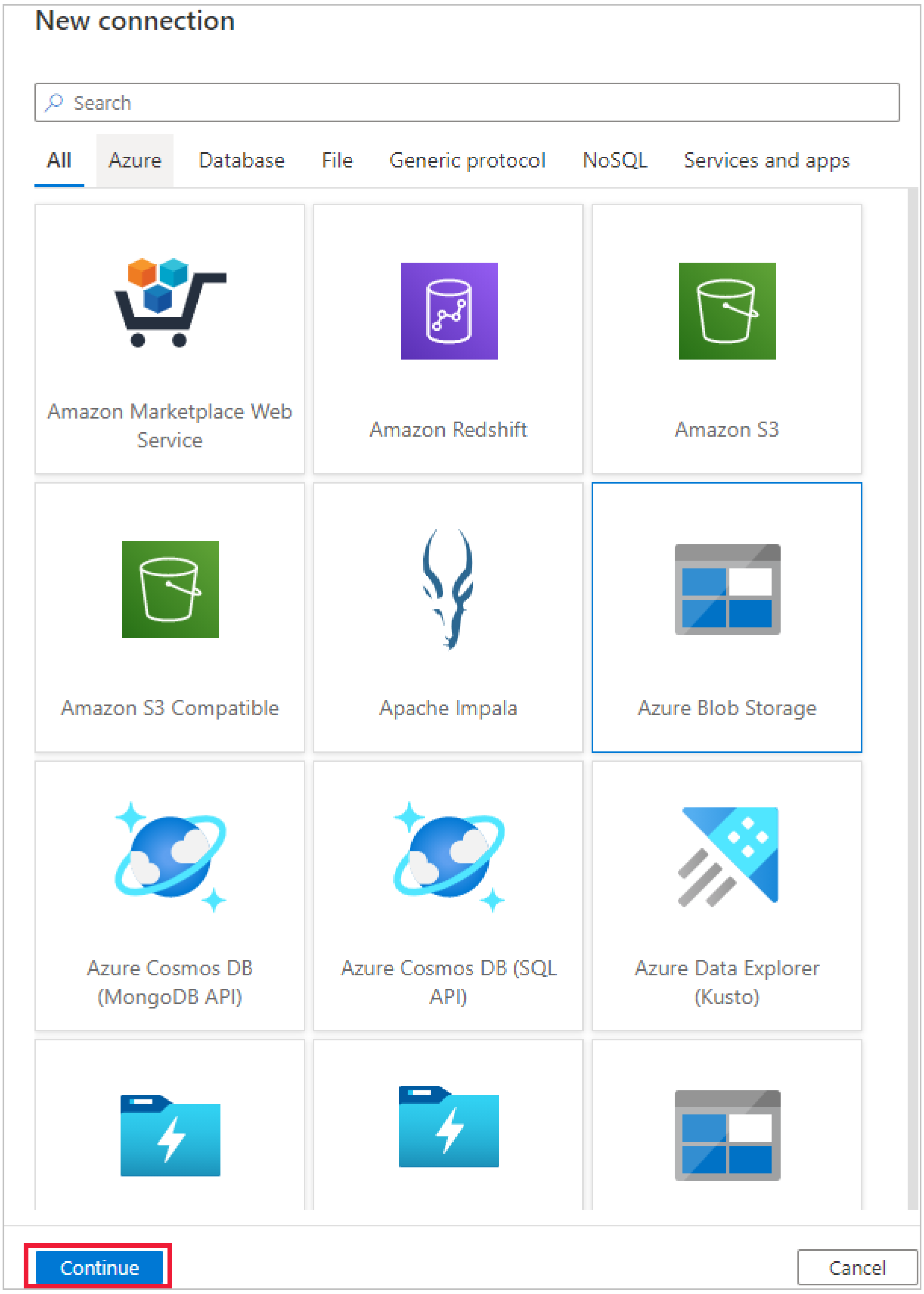
□ Specify an end date

Advanced

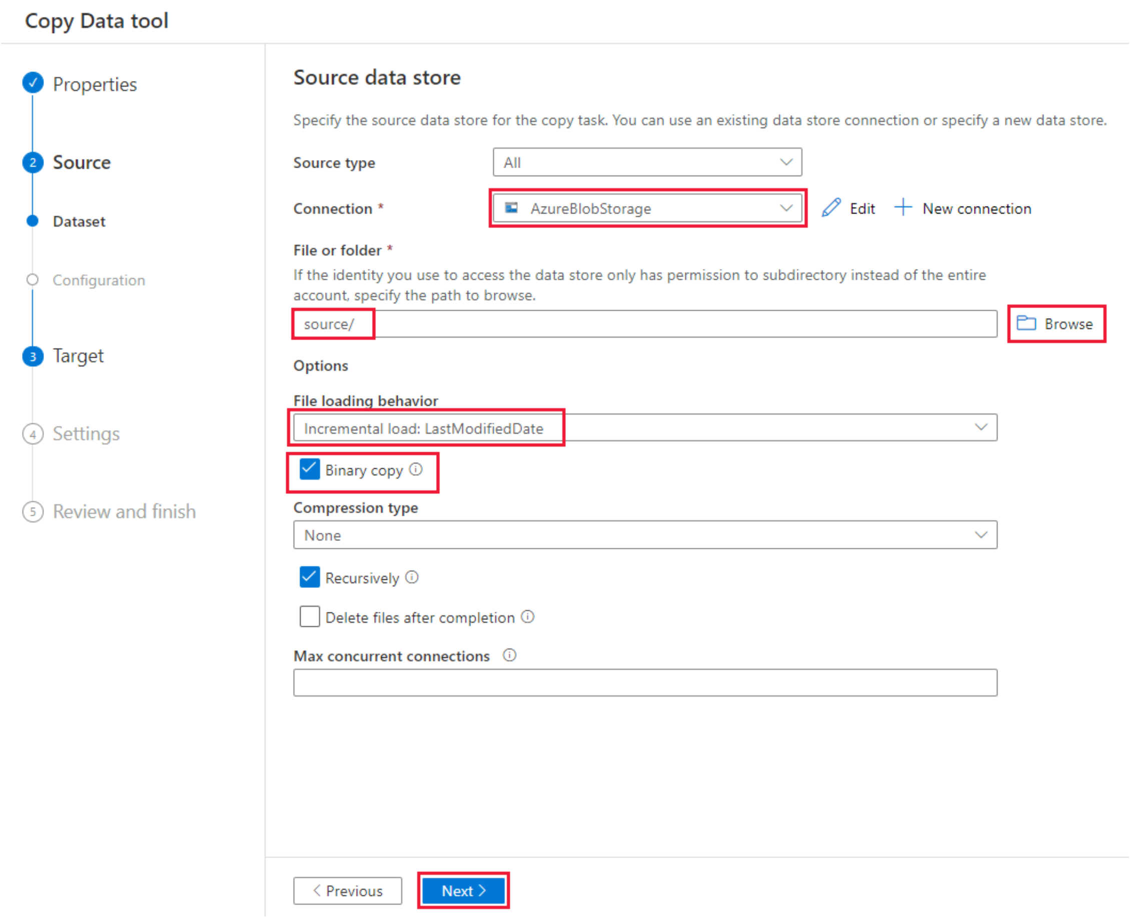
3. On the **Source data store** page, complete these steps:

   1. Select **+ New connection** to add a connection.

   2. Select **Azure Blob Storage** from the gallery, and then select **Continue**:



3. On the **New connection (Azure Blob Storage)** page, select your Azure subscription from the **Azure subscription** list and your storage account from the **Storage account name** list. Test the connection and then select **Create**.
4. Select the newly created connection in the **Connection** block.
5. In the **File or folder** section, select **Browse** and choose the **source** folder, and then select **OK**.
6. Under **File loading behavior**, select **Incremental load: LastModifiedDate**, and choose **Binary copy**.
7. Select **Next**.



Copy Data tool

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.

Source type: All

Connection *: AzureBlobStorage

File or folder *

If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

source/

Options

File loading behavior: Incremental load: LastModifiedDate

☑ Binary copy

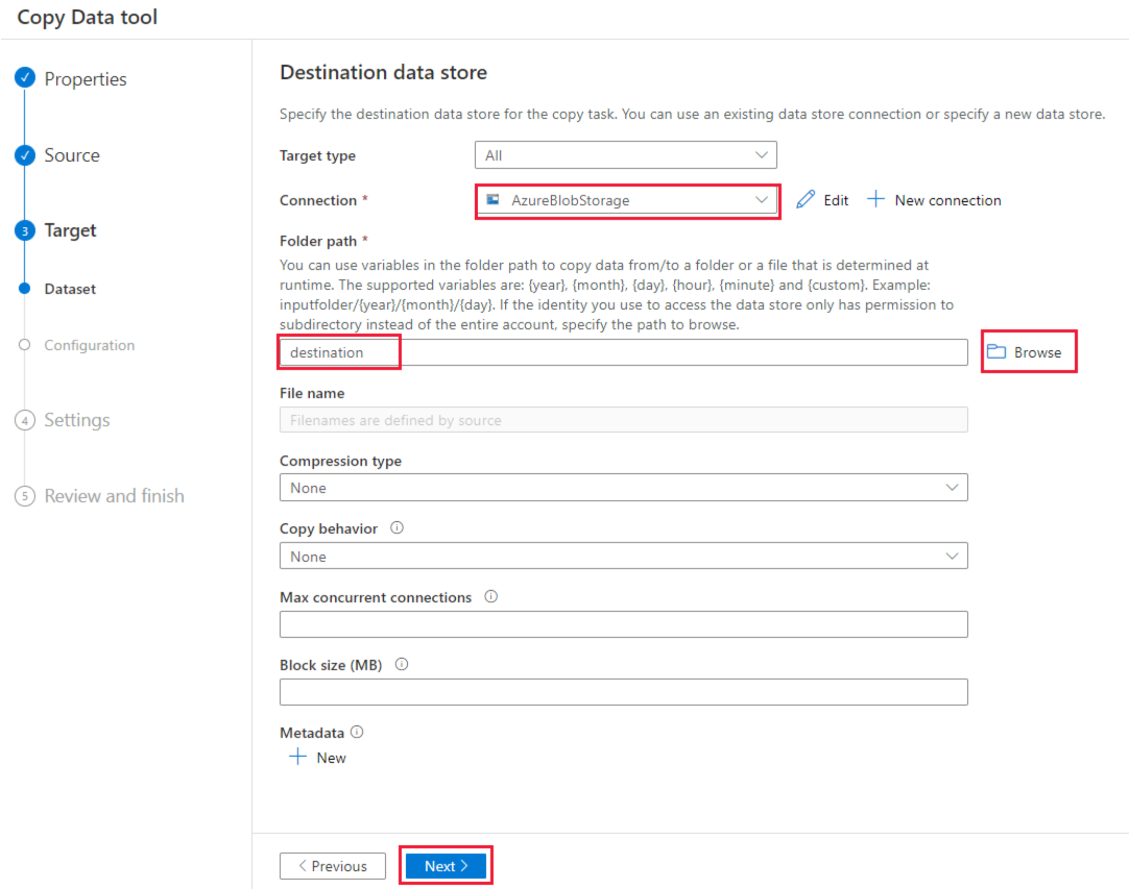
Compression type: None

☑ Recursively

☐ Delete files after completion

Max concurrent connections:

4. On the **Destination data store** page, complete these steps:
    1. Select the **AzureBlobStorage** connection that you created. This is the same storage account as the source data store.
    2. In the **Folder path** section, browse for and select the **destination** folder, and then select **OK**.
    3. Select **Next**.



Copy Data tool

Destination data store

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

Target type: All

Connection *: AzureBlobStorage

Folder path *

You can use variables in the folder path to copy data from/to a folder or a file that is determined at runtime. The supported variables are: {year}, {month}, {day}, {hour}, {minute} and {custom}. Example: inputfolder/{year}/{month}/{day}. If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

destination

File name: Filenames are defined by source

Compression type: None

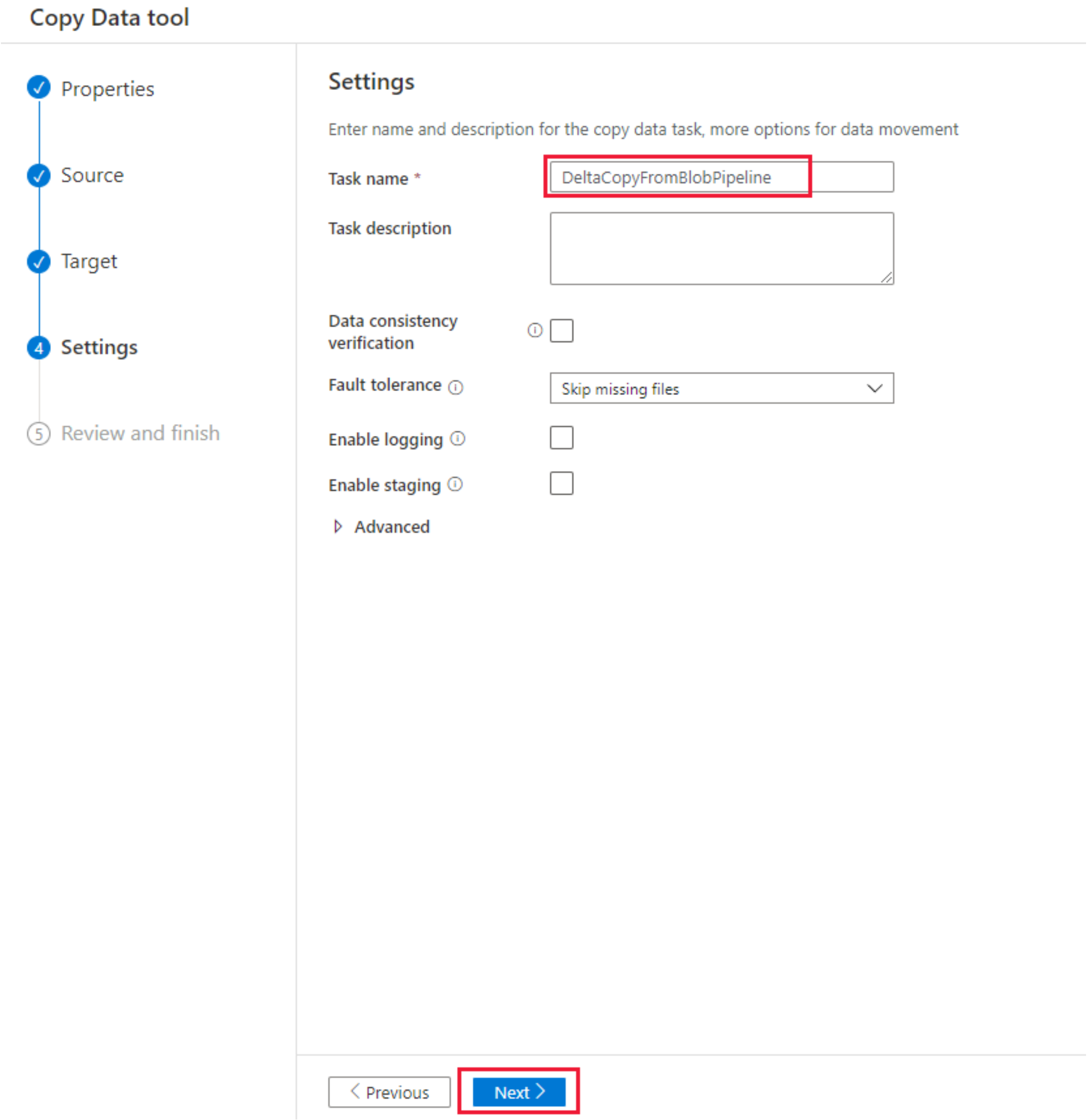
Copy behavior: None

Max concurrent connections:

Block size (MB):

Metadata

5. On the **Settings** page, under **Task name**, enter **DeltaCopyFromBlobPipeline**, then select **Next**. Data Factory creates a pipeline with the specified task name.



Copy Data tool

Settings

Enter name and description for the copy data task, more options for data movement

Task name * DeltaCopyFromBlobPipeline

Task description

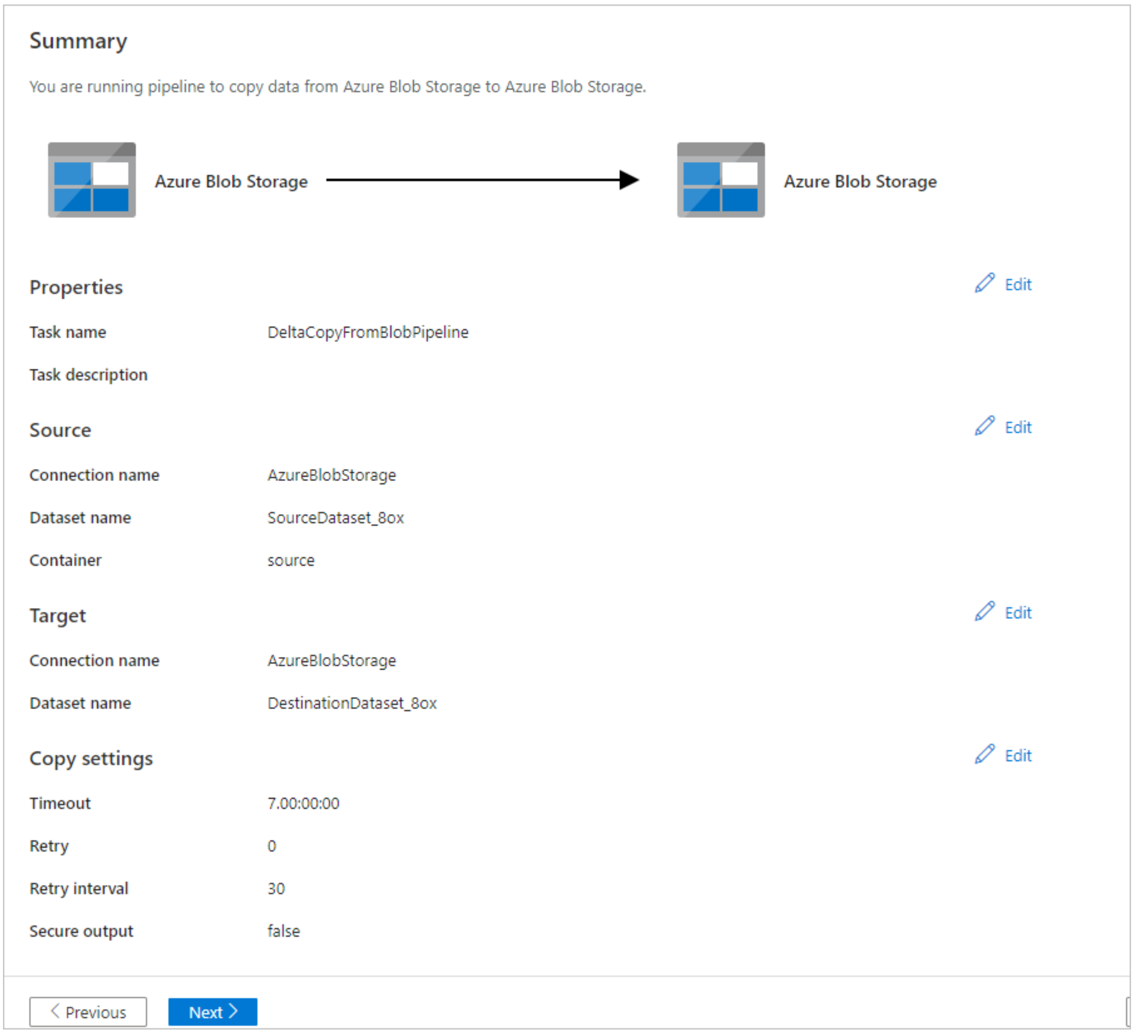
Data consistency verification □

Fault tolerance: Skip missing files

Enable logging □

Enable staging □

6. On the **Summary** page, review the settings and then select **Next**.



Summary

You are running pipeline to copy data from Azure Blob Storage to Azure Blob Storage.

Properties ✎ Edit

Task name	DeltaCopyFromBlobPipeline
Task description	

Source ✎ Edit

Connection name	AzureBlobStorage
Dataset name	SourceDataset_8ox
Container	source

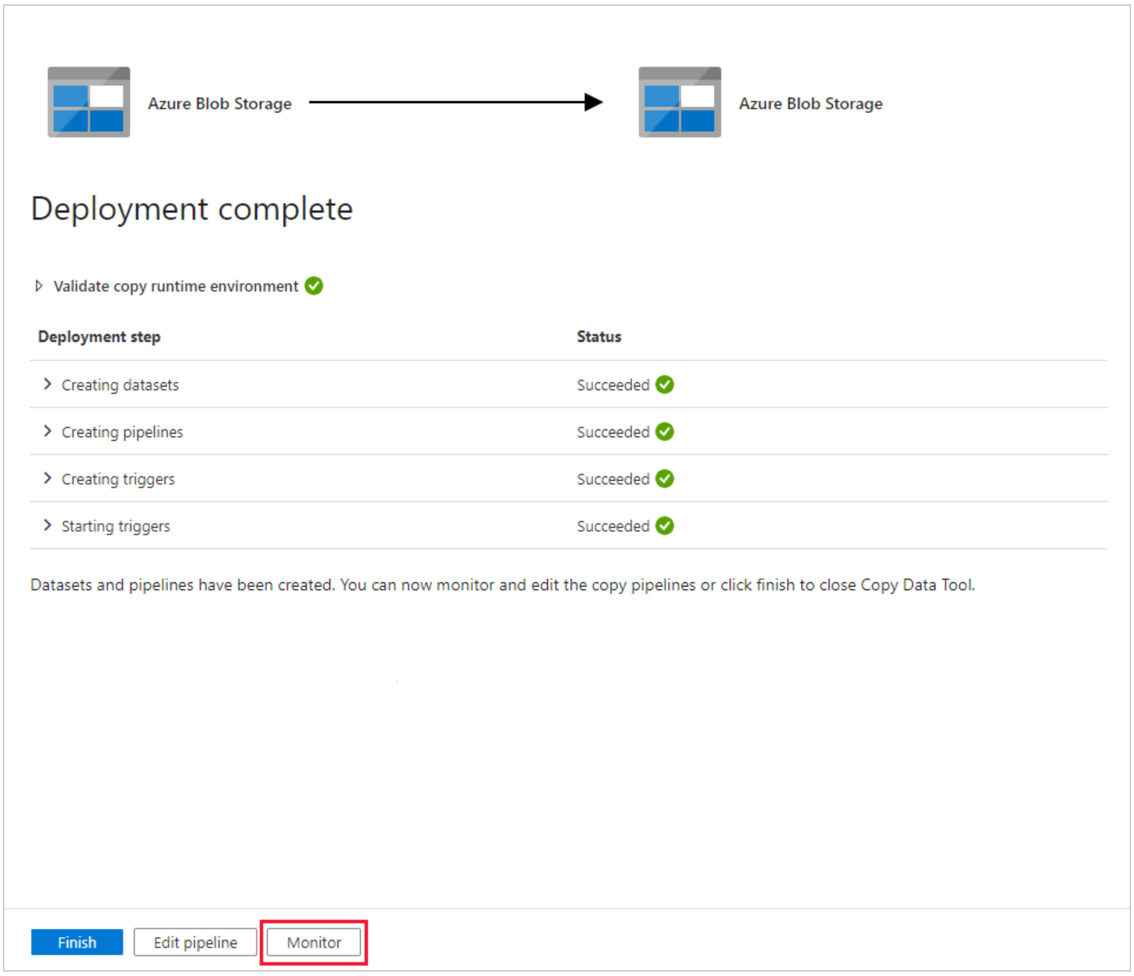
Target ✎ Edit

Connection name	AzureBlobStorage
Dataset name	DestinationDataset_8ox

Copy settings ✎ Edit

Timeout	7.00:00:00
Retry	0
Retry interval	30
Secure output	false

7. On the **Deployment** page, select **Monitor** to monitor the pipeline (task).



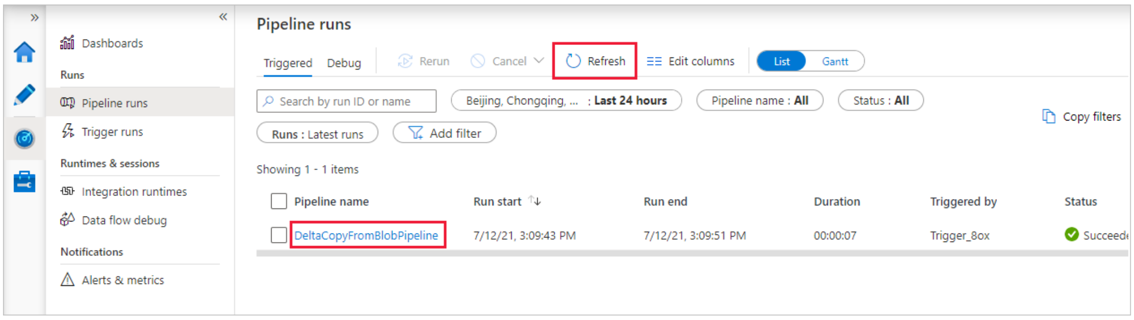
Deployment complete

▷ Validate copy runtime environment ✓

Deployment step	Status
› Creating datasets	Succeeded ✓
› Creating pipelines	Succeeded ✓
› Creating triggers	Succeeded ✓
› Starting triggers	Succeeded ✓

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

8. Notice that the **Monitor** tab on the left is automatically selected. The application switches to the **Monitor** tab. You see the status of the pipeline. Select **Refresh** to refresh the list. Select the link under **Pipeline name** to view activity run details or to run the pipeline again.



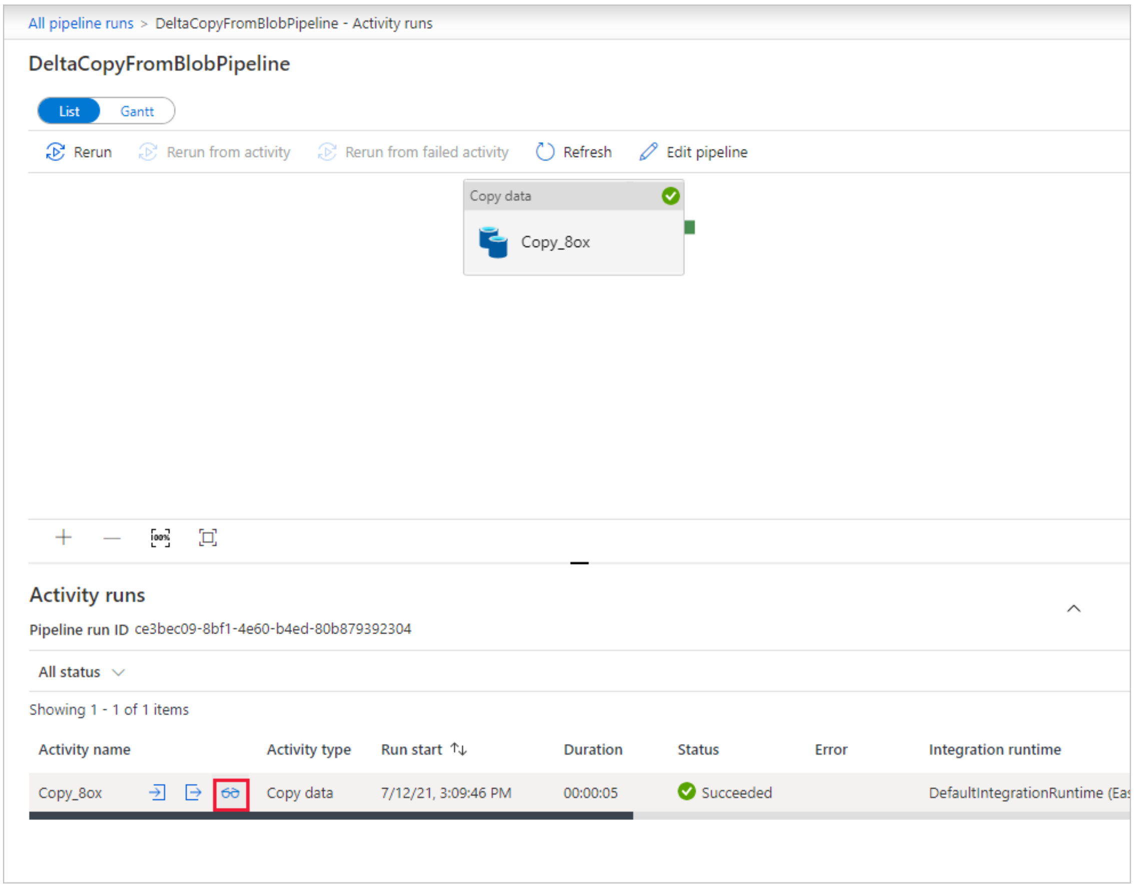
Pipeline runs

Beijing, Chongqing, ... : **Last 24 hours** | Pipeline name : **All** | Status : **All** | Runs : Latest runs

Showing 1 - 1 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status
DeltaCopyFromBlobPipeline	7/12/21, 3:09:43 PM	7/12/21, 3:09:51 PM	00:00:07	Trigger_8ox	Succeed

9. There's only one activity (the copy activity) in the pipeline, so you see only one entry. For details about the copy operation, on the **Activity runs** page, select the **Details** link (the eyeglasses icon) in the **Activity name** column. For details about the properties, see Copy activity overview.



All pipeline runs > DeltaCopyFromBlobPipeline - Activity runs

DeltaCopyFromBlobPipeline

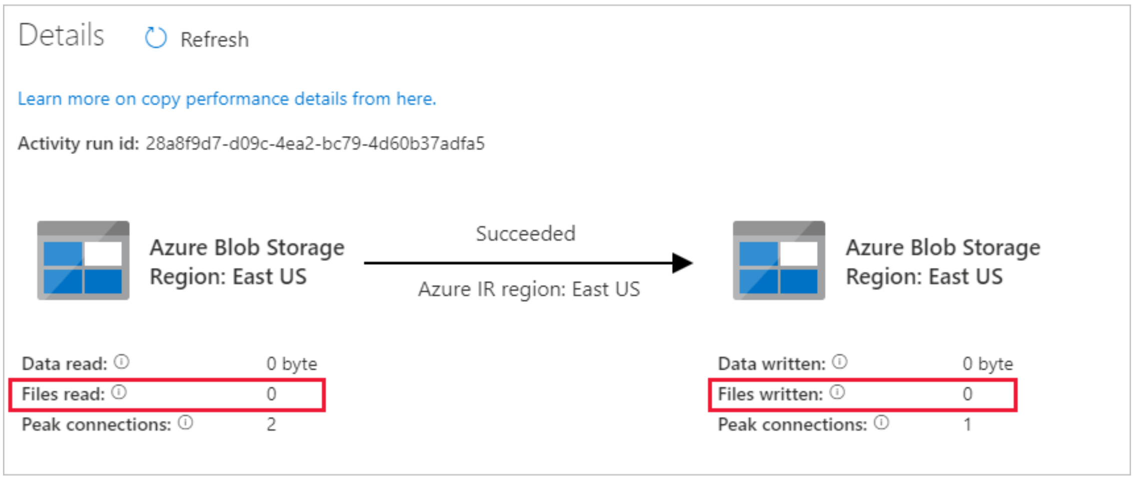
Activity runs

Pipeline run ID ce3bec09-8bf1-4e60-b4ed-80b879392304

Showing 1 - 1 of 1 items

Activity name	Activity type	Run start	Duration	Status	Error	Integration runtime
Copy_8ox	Copy data	7/12/21, 3:09:46 PM	00:00:05	Succeeded		DefaultIntegrationRuntime (Eas

Because there are no files in the source container in your Blob storage account, you won't see any files copied to the destination container in the account:



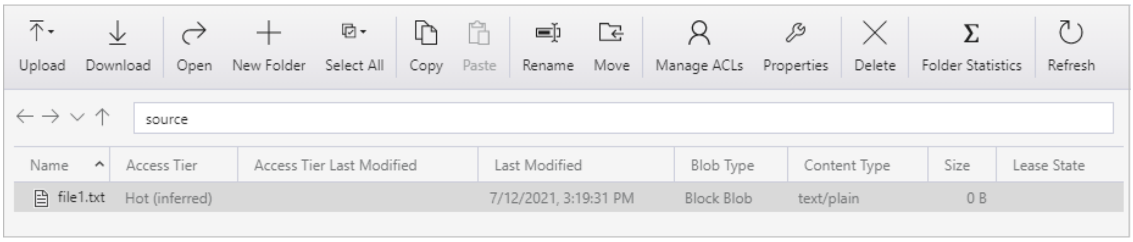
Learn more on copy performance details from here.

Activity run id: 28a8f9d7-d09c-4ea2-bc79-4d60b37adfa5

Data read:	0 byte
Files read:	0
Peak connections:	2

Data written:	0 byte
Files written:	0
Peak connections:	1

10. Create an empty text file and name it **file1.txt**. Upload this text file to the source container in your storage account. You can use various tools to perform these tasks, like Azure Storage Explorer.



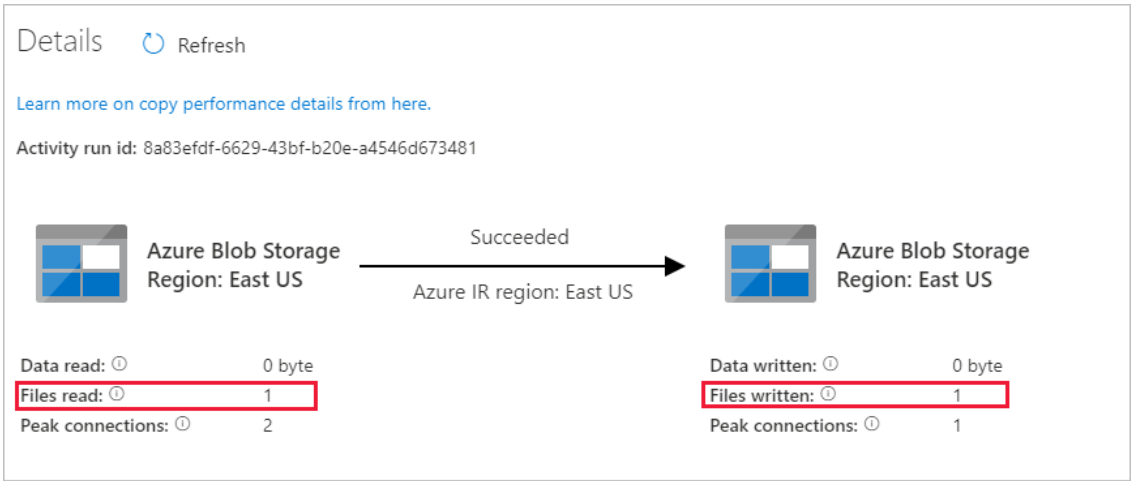
source

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Lease State
file1.txt	Hot (inferred)		7/12/2021, 3:19:31 PM	Block Blob	text/plain	0 B	

11. To go back to the **Pipeline runs** view, select **All pipeline runs** link in the breadcrumb menu on the **Activity runs** page, and wait for the same pipeline to be automatically triggered again.

12. When the second pipeline run completes, follow the same steps mentioned previously to review the activity run details.

    You'll see that one file (file1.txt) has been copied from the source container to the destination container of your Blob storage account:



Activity run id: 8a83efdf-6629-43bf-b20e-a4546d673481

Azure Blob Storage
Region: East US

Succeeded

Azure IR region: East US

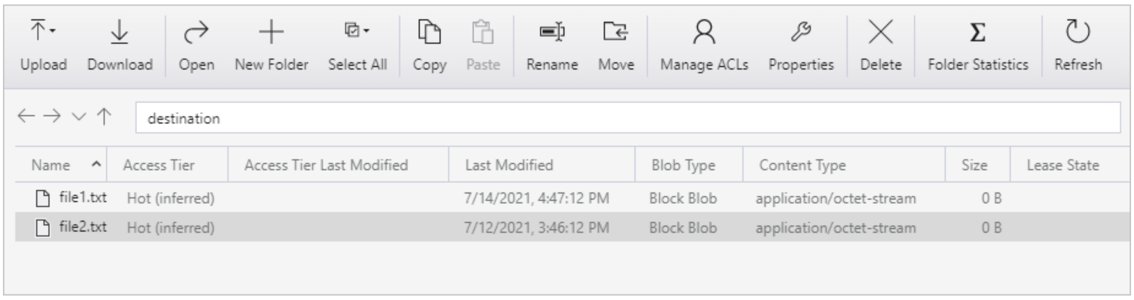
Azure Blob Storage
Region: East US

Data read:	0 byte	Data written:	0 byte
Files read:	1	Files written:	1
Peak connections:	2	Peak connections:	1

13. Create another empty text file and name it **file2.txt**. Upload this text file to the source container in your Blob storage account.

14. Repeat steps 11 and 12 for the second text file. You'll see that only the new file (file2.txt) was copied from the source container to the destination container of your storage account during this pipeline run.

    You can also verify that only one file has been copied by using Azure Storage Explorer to scan the files:



destination

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Lease State
file1.txt	Hot (inferred)		7/14/2021, 4:47:12 PM	Block Blob	application/octet-stream	0 B	
file2.txt	Hot (inferred)		7/12/2021, 3:46:12 PM	Block Blob	application/octet-stream	0 B	