LAB - Incrementally copy new files based on time partitioned file name by using the Copy Data tool

In this tutorial, you use the Azure portal to create a data factory. Then, you use the Copy Data tool to create a pipeline that incrementally copies new files based on time partitioned file name from Azure Blob storage to Azure Blob storage.

In this tutorial, you perform the following steps:

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

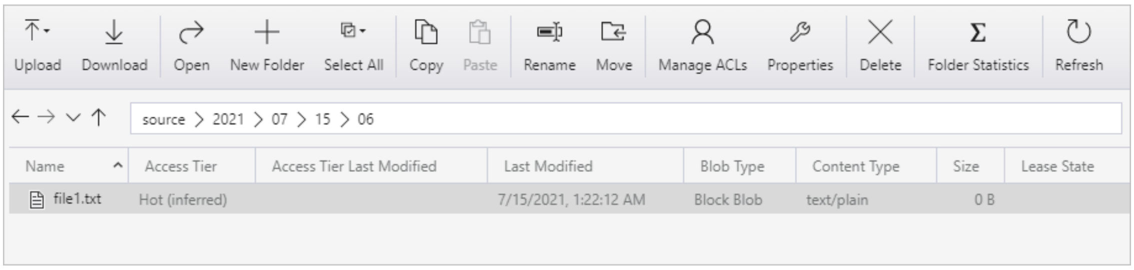
Prerequisites

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

Create two containers in Blob storage

Prepare your Blob storage for the tutorial by performing these steps.

1. Create a container named **source**. Create a folder path as **2021/07/15/06** in your container. Create an empty text file, and name it as **file1.txt**. Upload the file1.txt to the folder path **source/2021/07/15/06** in your storage account. You can use various tools to perform these tasks, such as Azure Storage Explorer.

Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Lease State
file1.txt	Hot (inferred)		7/15/2021, 1:22:12 AM	Block Blob	text/plain	0 B	

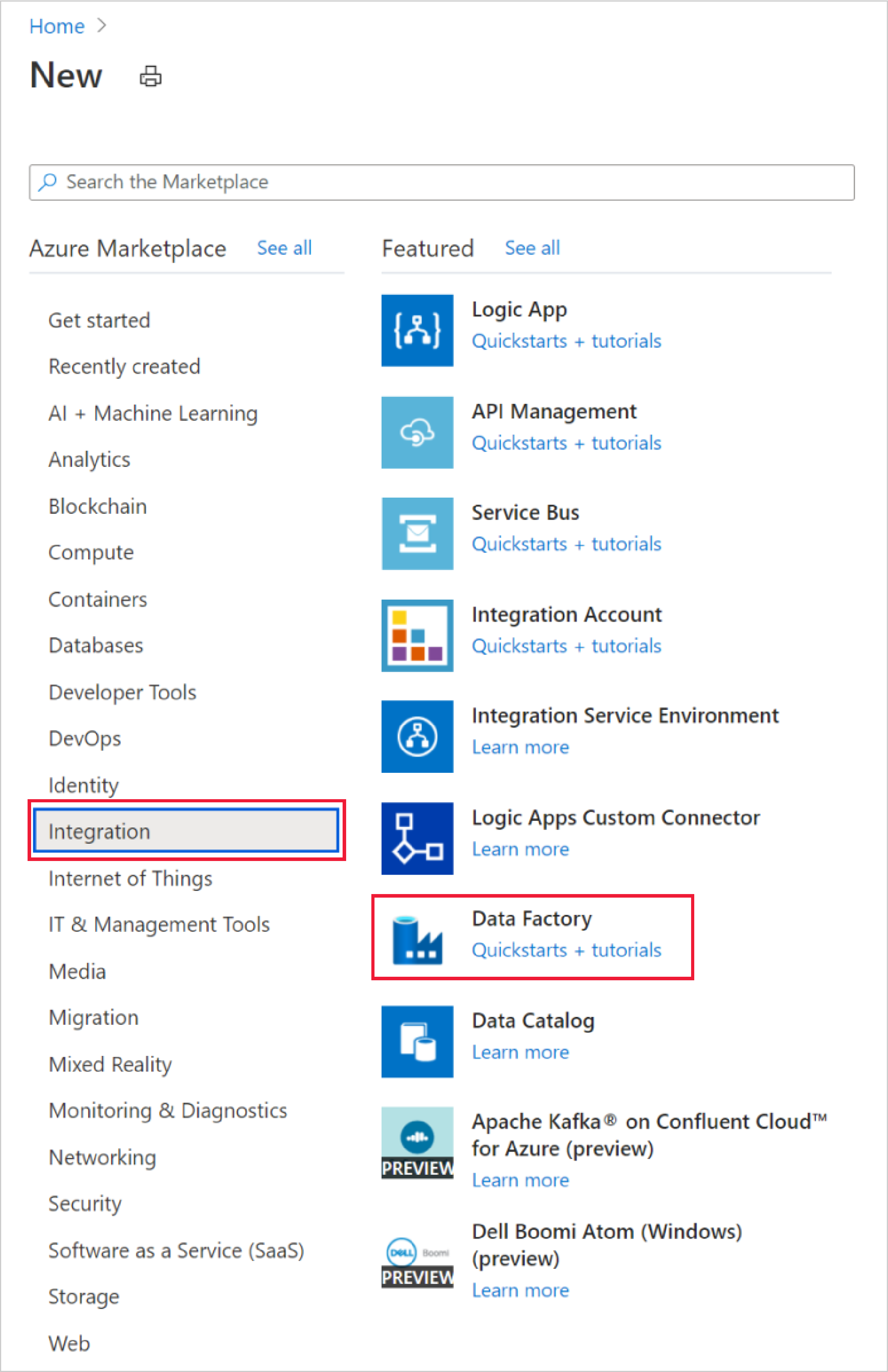
   Note

   Please adjust the folder name with your UTC time. For example, if the current UTC time is 6:10 AM on July 15, 2021, you can create the folder path as **source/2021/07/15/06/** by the rule of **source/{Year}/{Month}/{Day}/{Hour}/**.

2. Create a container named **destination**. You can use various tools to perform these tasks, such as Azure Storage Explorer.

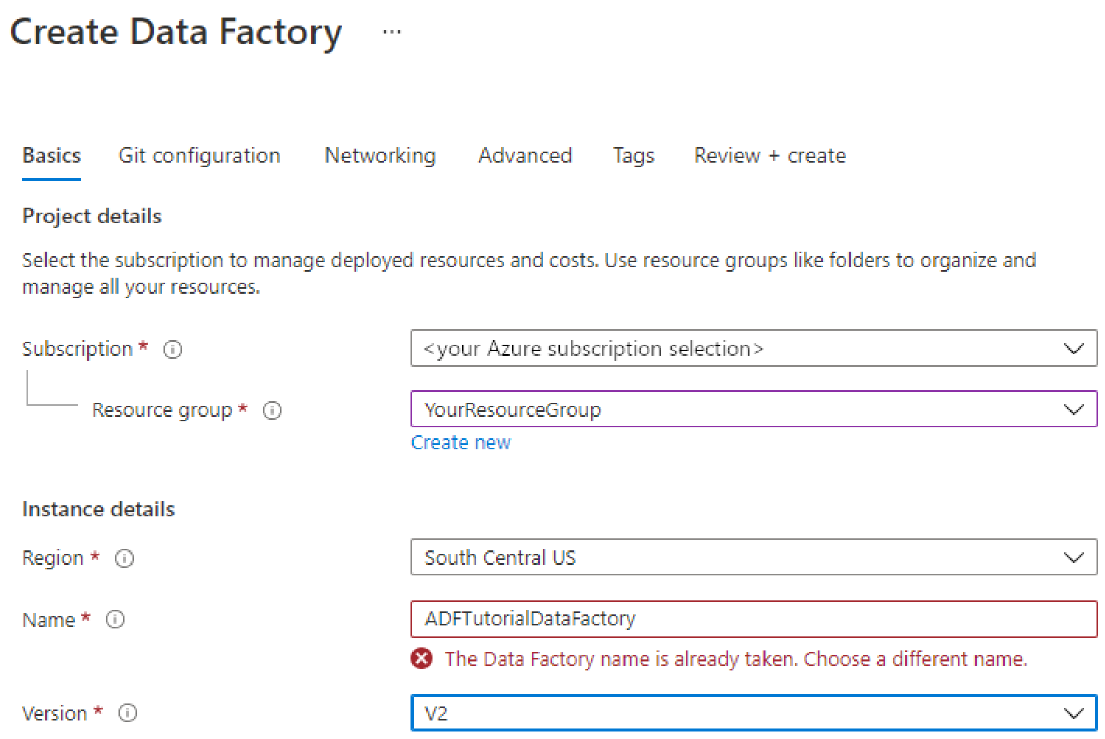
Create a data factory

1. On the left menu, select **Create a resource** > **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 the following error message:



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 *	<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.

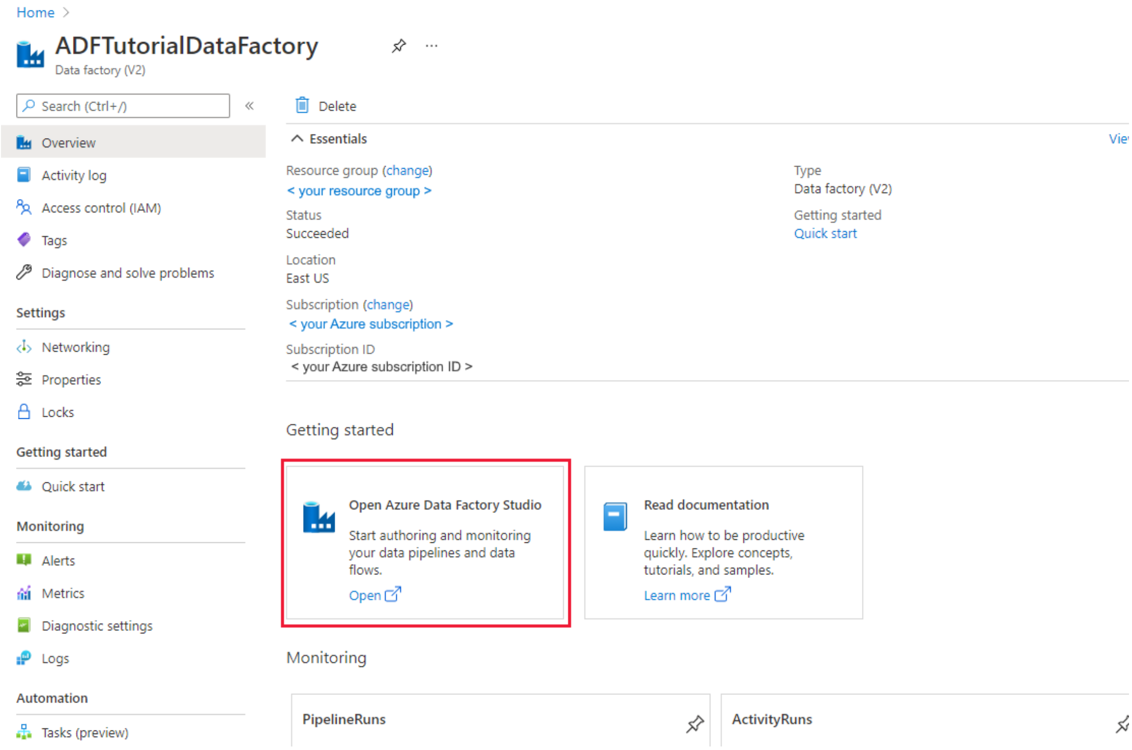
3. Select the Azure **subscription** in which to create the new data factory.
4. For **Resource Group**, take one of the following steps:

   a. Select **Use existing**, and select an existing resource group from the drop-down list.

   b. Select **Create new**, and enter the name of a resource group.

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

5. Under **version**, select **V2** for the version.
6. Under **location**, select the location for the data factory. Only supported locations are displayed in the drop-down list. The data stores (for example, Azure Storage and SQL Database) and computes (for example, Azure HDInsight) that are used by your data factory can be in other locations and regions.
7. Select **Create**.
8. After creation is finished, the **Data Factory** home page is displayed.
9. To launch the Azure Data Factory user interface (UI) in a separate tab, select **Open** on the **Open Azure Data Factory Studio** tile.



Home >

ADFTutorialDataFactory

Data factory (V2)

Search (Ctrl+/)

Overview

Activity log

Access control (IAM)

Tags

Diagnose and solve problems

Settings

Networking

Properties

Locks

Getting started

Quick start

Monitoring

Alerts

Metrics

Diagnostic settings

Logs

Automation

Tasks (preview)

Delete

Essentials

Vie

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

Getting started

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

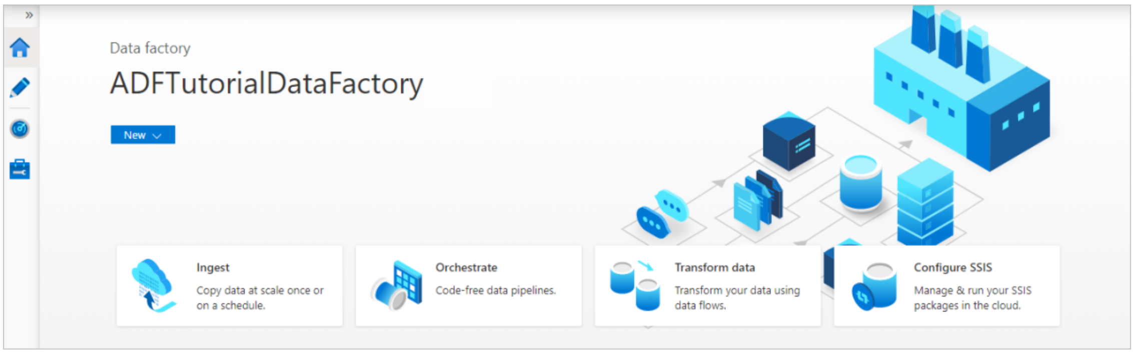
Monitoring

PipelineRuns

ActivityRuns

Use the Copy Data tool to create a pipeline

1. On the Azure Data Factory home page, select the **Ingest** title to launch the Copy Data tool.



Data factory

ADFTutorialDataFactory

New

Ingest
Copy data at scale once or on a schedule.

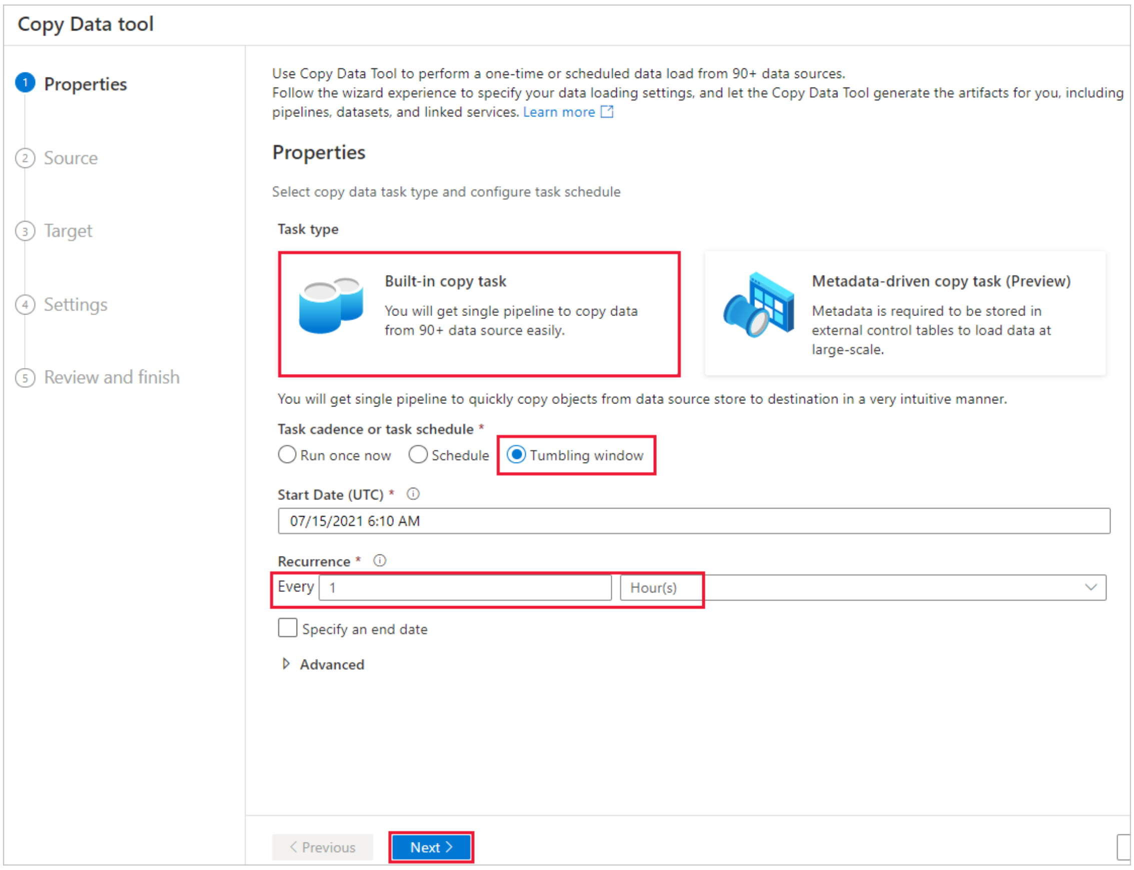
Orchestrate
Code-free data pipelines.

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**, choose **Built-in copy task**.
    2. Under **Task cadence or task schedule**, select **Tumbling window**.
    3. Under **Recurrence**, enter **1 Hour(s)**.
    4. Select **Next**.



Copy Data tool

1 Properties

2 Source

3 Target

4 Settings

5 Review and finish

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) *
07/15/2021 6:10 AM

Recurrence *
Every 1 Hour(s)

□ Specify an end date

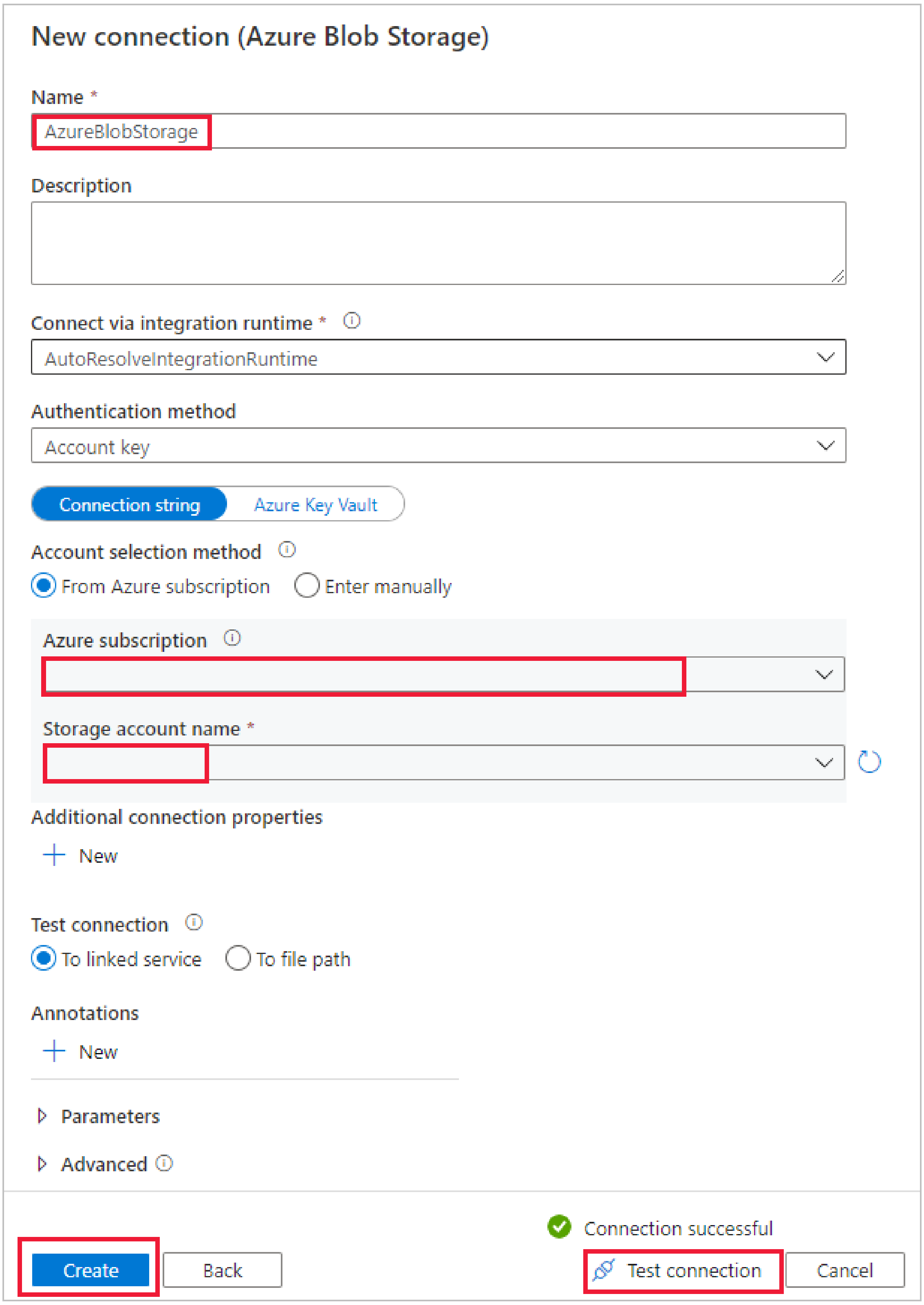
Advanced

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

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

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

   c. On the **New connection (Azure Blob Storage)** page, enter a name for the connection. Select your Azure subscription, and select your storage account from the **Storage account name** list. Test connection and then select **Create**.



New connection (Azure Blob Storage)

Name *: AzureBlobStorage

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Authentication method: Account key

Connection string | Azure Key Vault

Account selection method: ● From Azure subscription ○ Enter manually

Azure subscription:

Storage account name *:

Additional connection properties

Test connection: ● To linked service ○ To file path

Annotations

✓ Connection successful

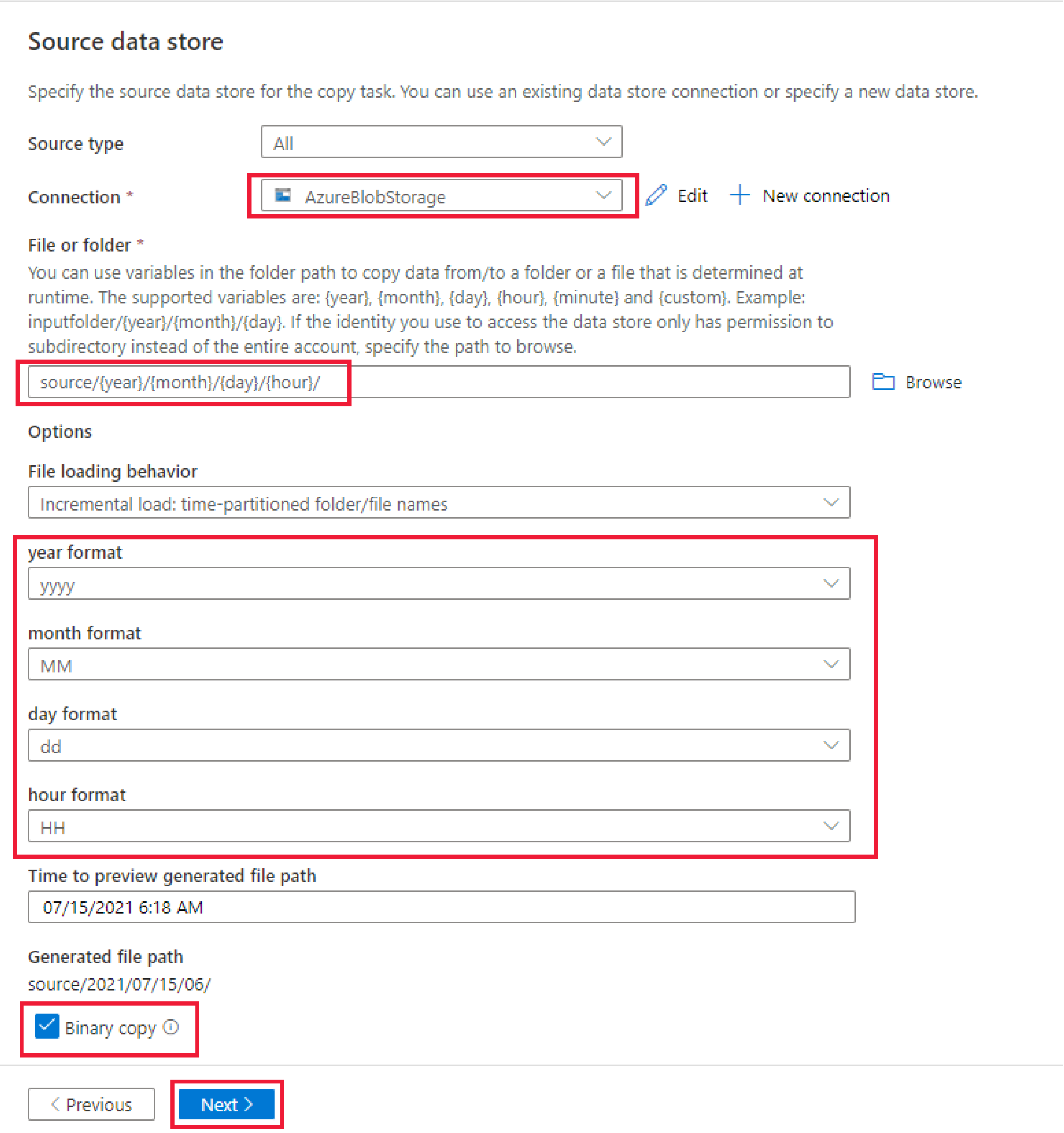
d. On the **Source data store** page, select the newly created connection in the **Connection** section.

e. In the **File or folder** section, browse and select the **source** container, then select **OK**.

f. Under **File loading behavior**, select **Incremental load: time-partitioned folder/file names**.

g. Write the dynamic folder path as **source/{year}/{month}/{day}/{hour}/**, and change the format as shown in the following screenshot.

h. Check **Binary copy** and select **Next**.



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 *

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.

source/{year}/{month}/{day}/{hour}/

Options

File loading behavior: Incremental load: time-partitioned folder/file names

year format: yyyy

month format: MM

day format: dd

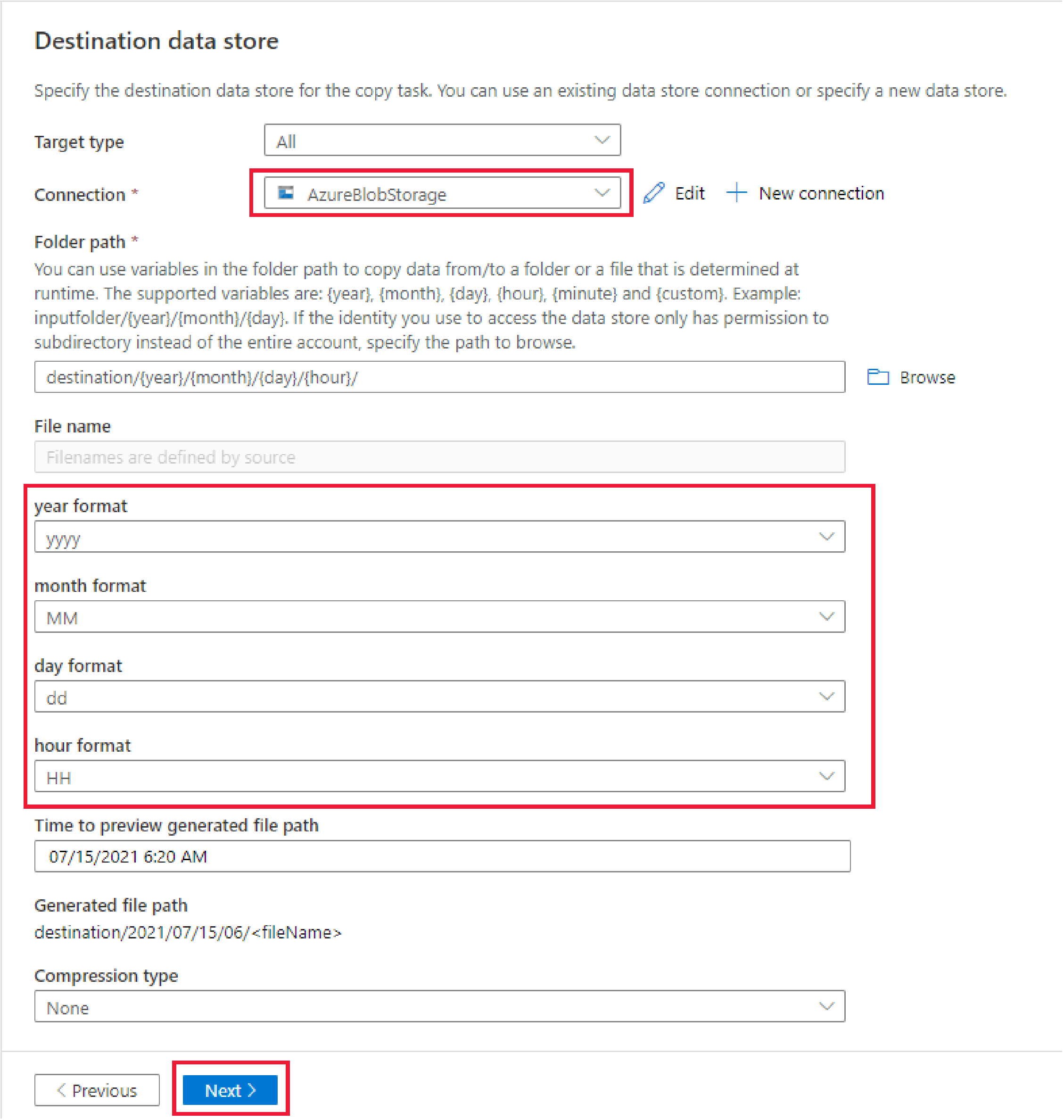
hour format: HH

Time to preview generated file path: 07/15/2021 6:18 AM

Generated file path: source/2021/07/15/06/

☑ Binary copy

4. On the **Destination data store** page, complete the following steps:
   1. Select the **AzureBlobStorage**, which is the same storage account as data source store.
   2. Browse and select the **destination** folder, then select **OK**.
   3. Write the dynamic folder path as **destination/{year}/{month}/{day}/{hour}/**, and change the format as shown in the following screenshot.
   4. Select **Next**.



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/{year}/{month}/{day}/{hour}/

File name: Filenames are defined by source

year format: yyyy

month format: MM

day format: dd

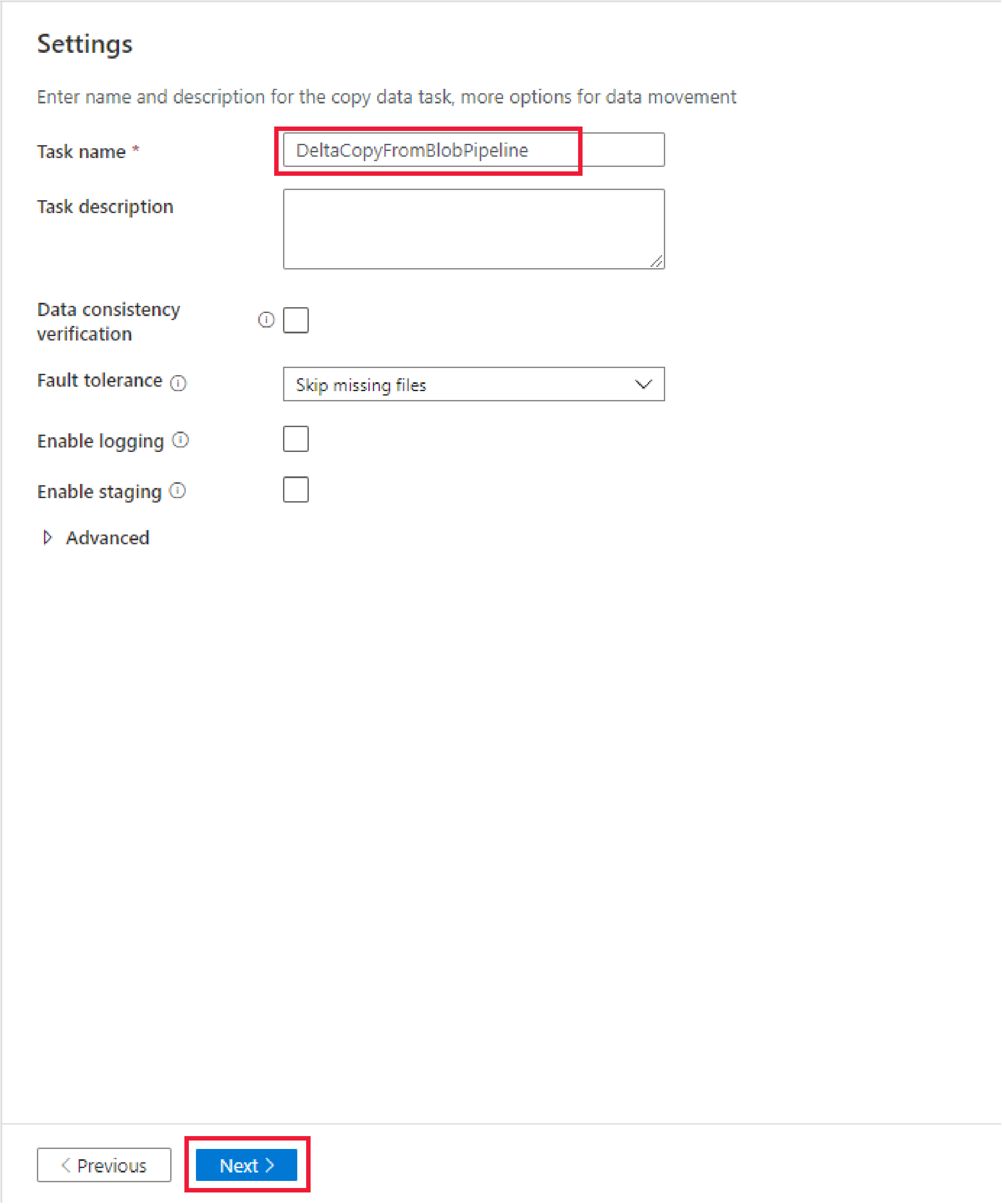
hour format: HH

Time to preview generated file path: 07/15/2021 6:20 AM

Generated file path
destination/2021/07/15/06/<fileName>

Compression type: None

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



Settings

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

Task name * DeltaCopyFromBlobPipeline

Task description

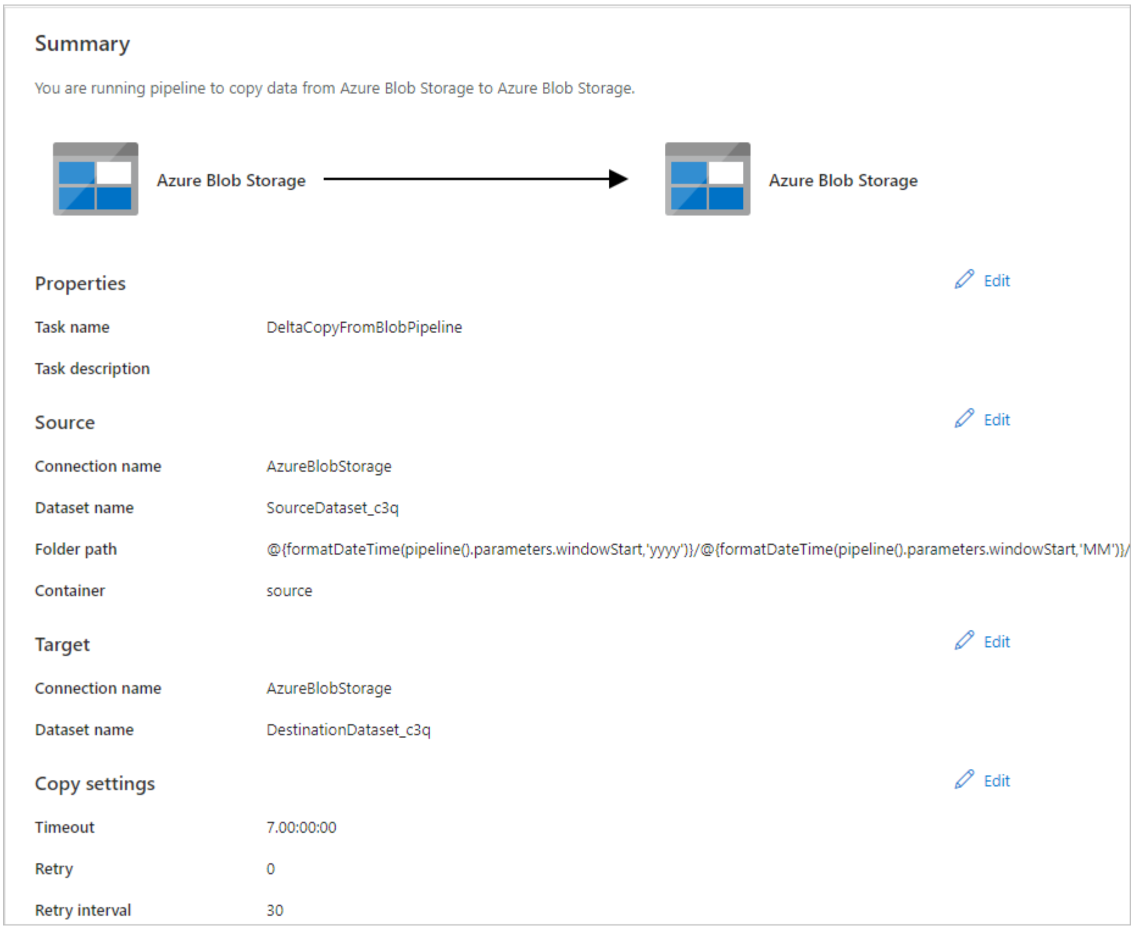
Data consistency verification □

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

Task name	DeltaCopyFromBlobPipeline
Task description	

Source

Connection name	AzureBlobStorage
Dataset name	SourceDataset_c3q
Folder path	@{formatDateTime(pipeline().parameters.windowStart,'yyyy')}/@{formatDateTime(pipeline().parameters.windowStart,'MM')}/
Container	source

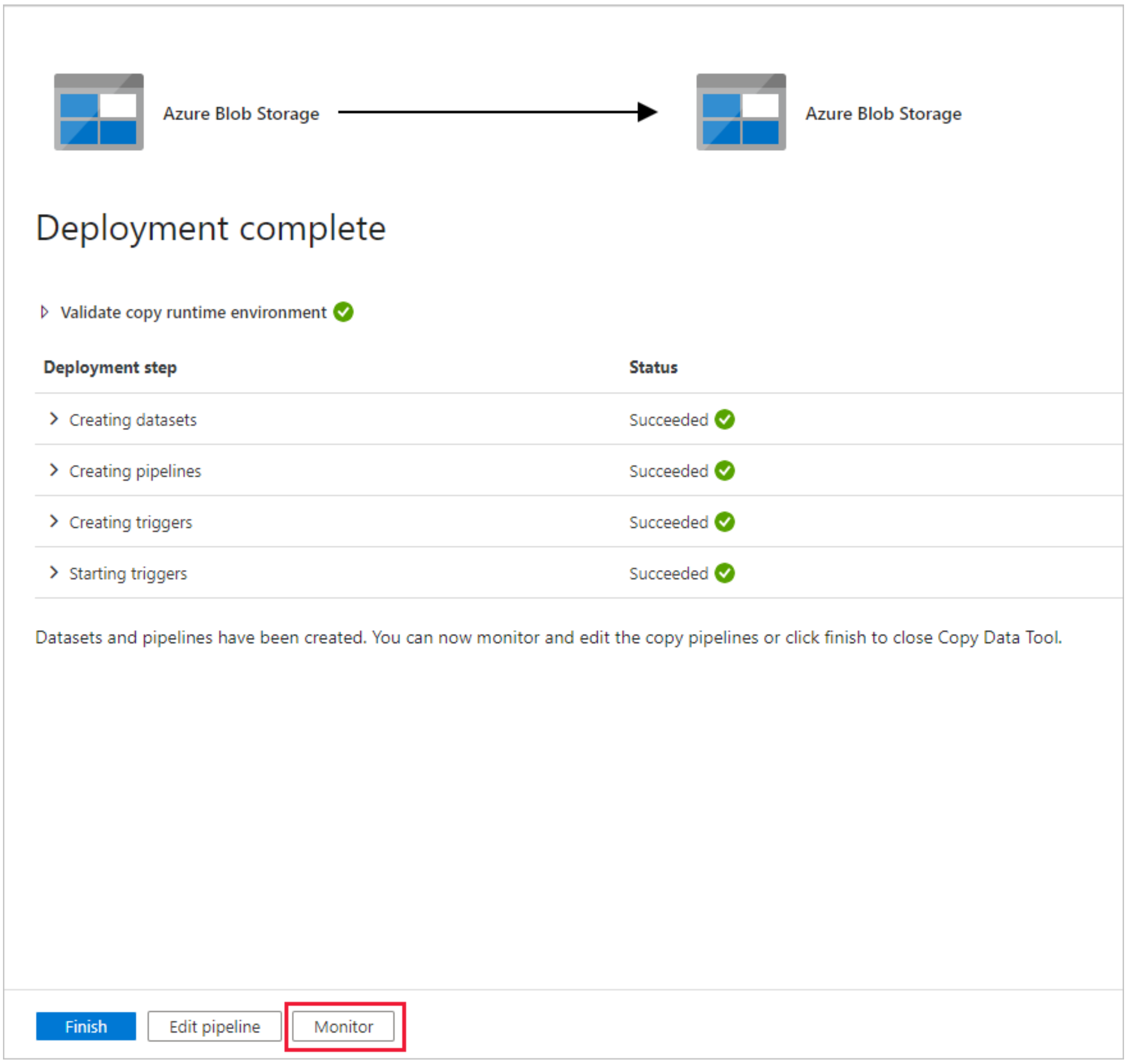
Target

Connection name	AzureBlobStorage
Dataset name	DestinationDataset_c3q

Copy settings

Timeout	7.00:00:00
Retry	0
Retry interval	30

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



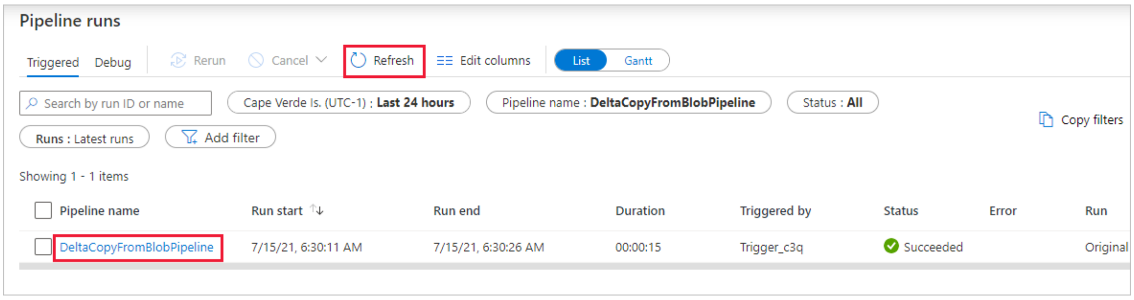
Deployment complete

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. You need wait for the pipeline run when it is triggered automatically (about after one hour). When it runs, select the pipeline name link **DeltaCopyFromBlobPipeline** to view activity run details or rerun the pipeline. Select **Refresh** to refresh the list.

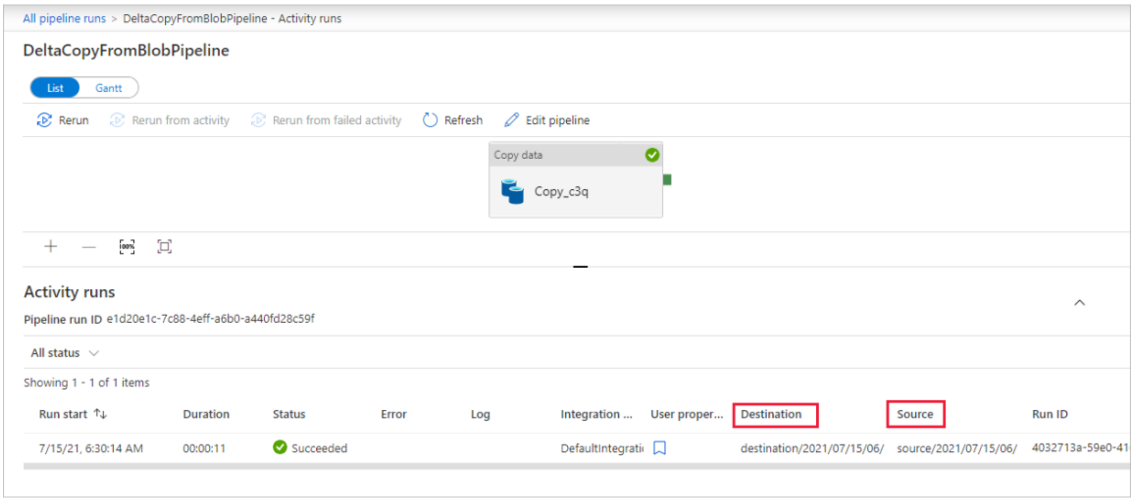

Pipeline runs

Cape Verde Is. (UTC-1) : **Last 24 hours** | Pipeline name : **DeltaCopyFromBlobPipeline** | Status : **All** | Runs : Latest runs

Showing 1 - 1 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Error	Run
DeltaCopyFromBlobPipeline	7/15/21, 6:30:11 AM	7/15/21, 6:30:26 AM	00:00:15	Trigger_c3q	✓ Succeeded		Original

9. There's only one activity (copy activity) in the pipeline, so you see only one entry. Adjust the column width of the **Source** and **Destination** columns (if necessary) to display more details, you can see the source file (file1.txt) has been copied from *source/2021/07/15/06/* to *destination/2021/07/15/06/* with the same file name.


All pipeline runs > DeltaCopyFromBlobPipeline - Activity runs

DeltaCopyFromBlobPipeline

Activity runs

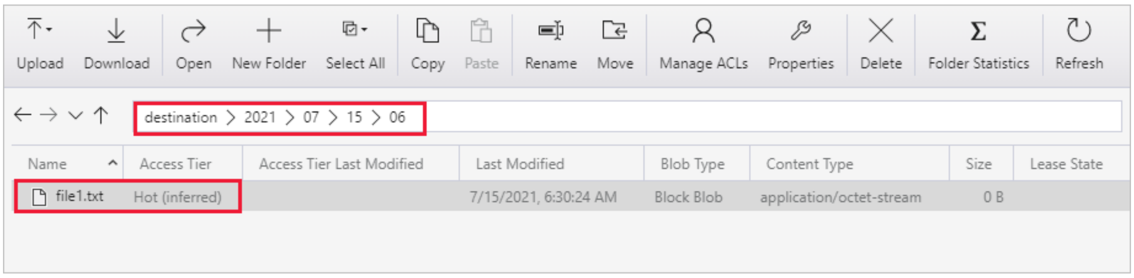
Pipeline run ID e1d20e1c-7c88-4eff-a6b0-a440fd28c59f

All status

Showing 1 - 1 of 1 items

Run start	Duration	Status	Error	Log	Integration ...	User proper...	Destination	Source	Run ID
7/15/21, 6:30:14 AM	00:00:11	✓ Succeeded			DefaultIntegrati		destination/2021/07/15/06/	source/2021/07/15/06/	4032713a-59e0-41

You can also verify the same by using Azure Storage Explorer (https://storageexplorer.com/) to scan the files.



destination > 2021 > 07 > 15 > 06

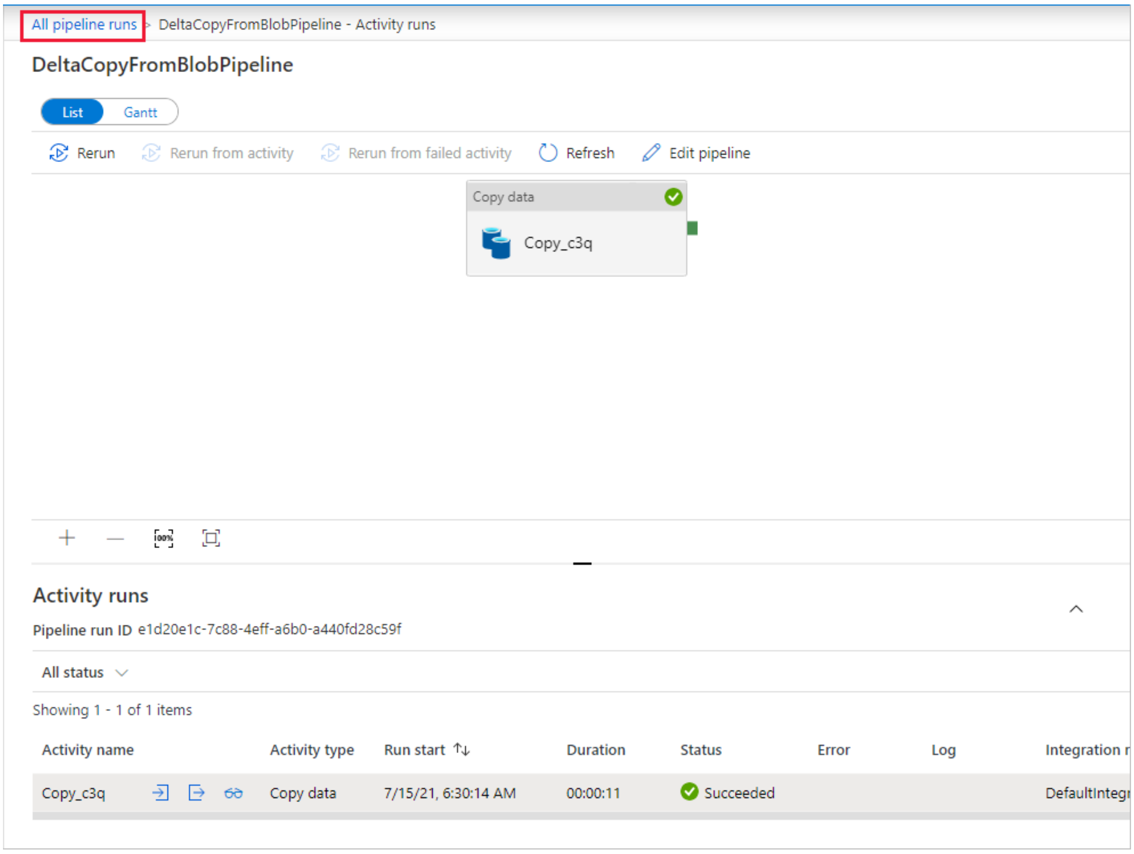
Name	Access Tier	Access Tier Last Modified	Last Modified	Blob Type	Content Type	Size	Lease State
file1.txt	Hot (inferred)		7/15/2021, 6:30:24 AM	Block Blob	application/octet-stream	0 B	

10. Create another empty text file with the new name as **file2.txt**. Upload the file2.txt file to the folder path **source/2021/07/15/07** in your storage account. You can use various tools to perform these tasks, such as Azure Storage Explorer.

Note

You might be aware that a new folder path is required to be created. Please adjust the folder name with your UTC time. For example, if the current UTC time is 7:30 AM on July. 15th, 2021, you can create the folder path as **source/2021/07/15/07/** by the rule of **{Year}/{Month}/{Day}/{Hour}/**.

11. To go back to the **Pipeline runs** view, select **All pipelines runs**, and wait for the same pipeline being triggered again automatically after another one hour.



DeltaCopyFromBlobPipeline

Activity runs

Pipeline run ID e1d20e1c-7c88-4eff-a6b0-a440fd28c59f

All status

Showing 1 - 1 of 1 items

Activity name	Activity type	Run start ↑↓	Duration	Status	Error	Log	Integration r
Copy_c3q	Copy data	7/15/21, 6:30:14 AM	00:00:11	✓ Succeeded			DefaultIntegr

12. Select the new **DeltaCopyFromBlobPipeline** link for the second pipeline run when it comes, and do the same to review details. You will see the source file (file2.txt) has been copied from **source/2021/07/15/07/** to **destination/2021/07/15/07/** with the same file name. You can also verify the same by using Azure Storage Explorer (https://storageexplorer.com/) to scan the files in **destination** container.