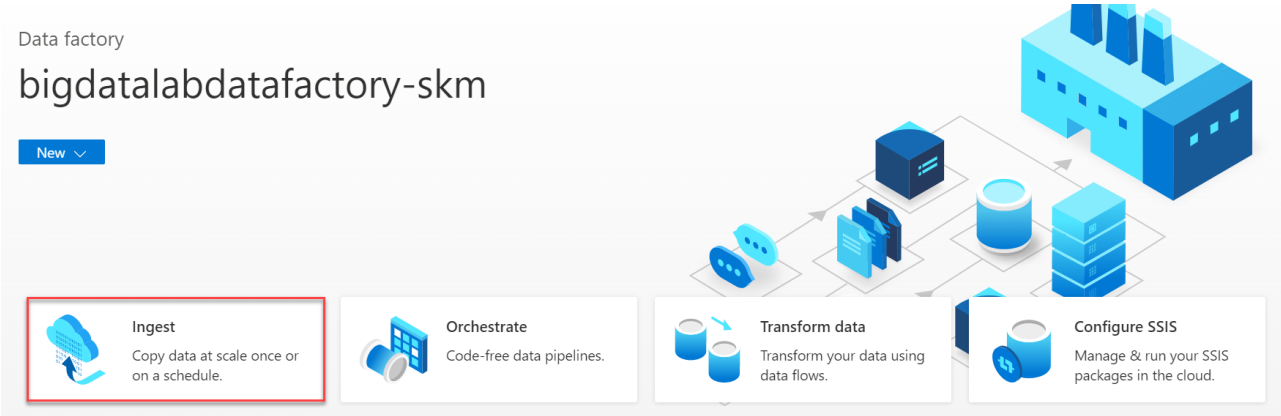
Challenge 4: Develop a data factory pipeline for data movement

Duration: 20 minutes

In this exercise, you will create an Azure Data Factory pipeline to copy data (.CSV files) from an on-premises server (your machine) to Azure Blob Storage. The goal of the exercise is to demonstrate data movement from an on-premises location to Azure Storage (via the Integration Runtime).

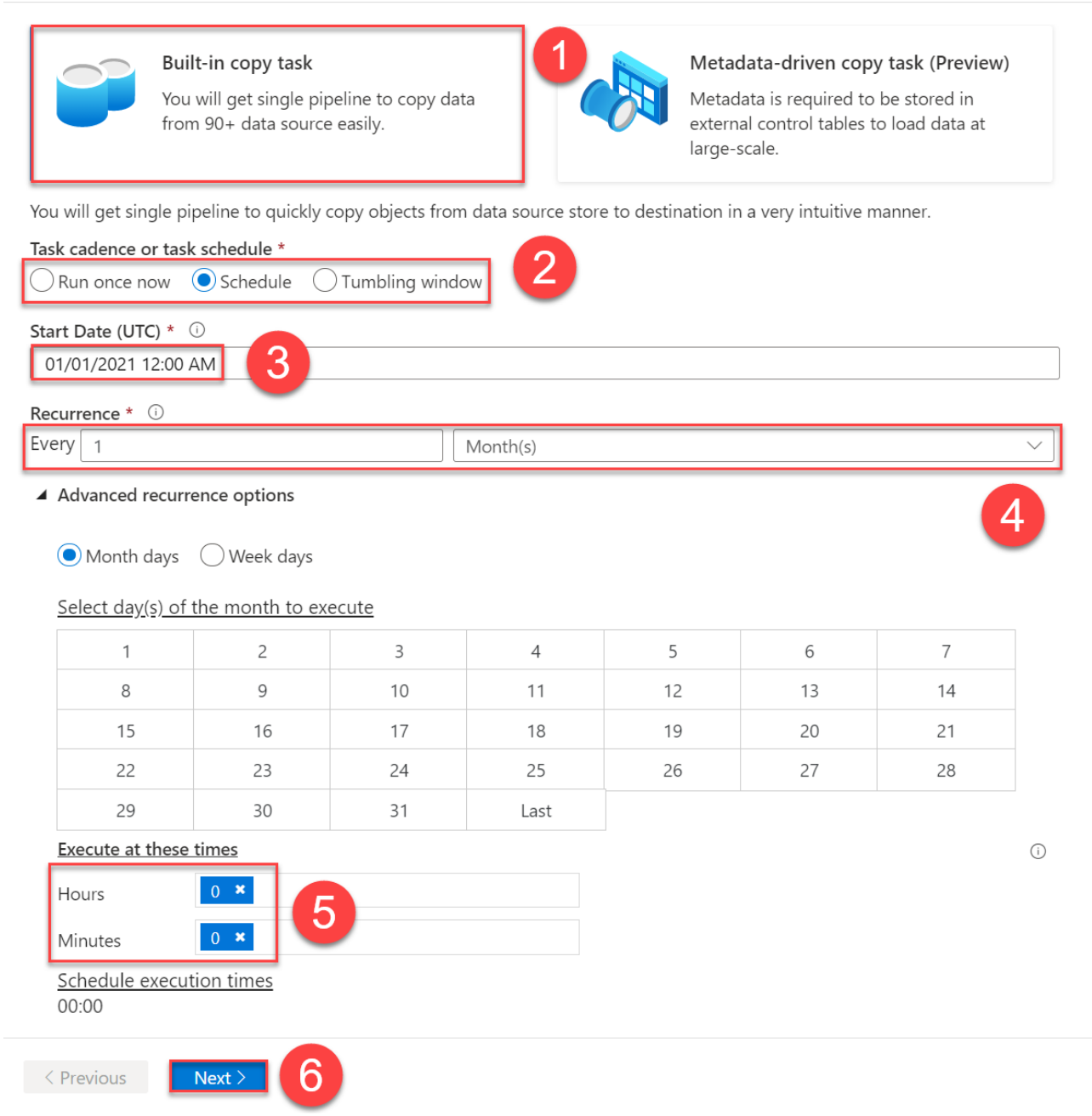
Task 1: Create copy pipeline using the Copy Data Wizard

1. Within the Azure Data Factory overview page, select **Ingest**.



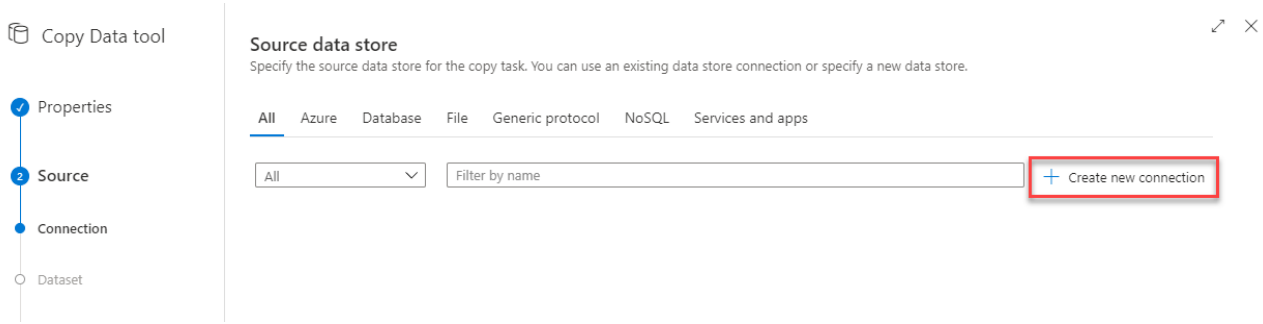
2. Enter the **Properties** page

    - Select **Built-in copy task (1)**
    - Select **Schedule** below **Task cadence or task schedule (2)**
    - Set the **Start Date (UTC)** to **01/01/2021 12:00 AM (3)**
    - Set the **Recurrence** to **Every 1 month (4)**
    - Below **Advanced recurrence options**, set **Hours** and **Minutes** to 0 **(5)**.

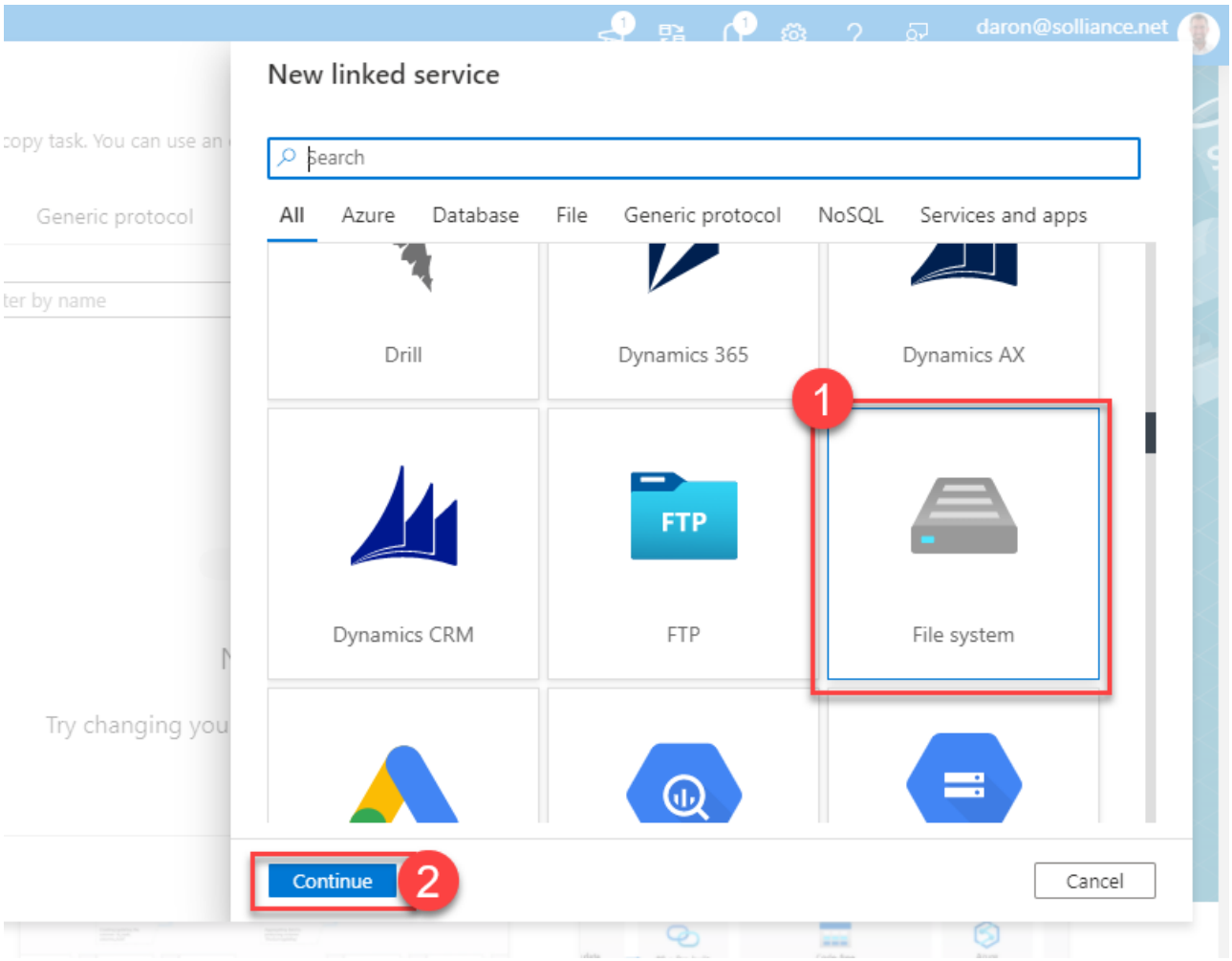


3. Select **Next (6)**.

4. On the Source data store screen, select **+ Create new connection**.

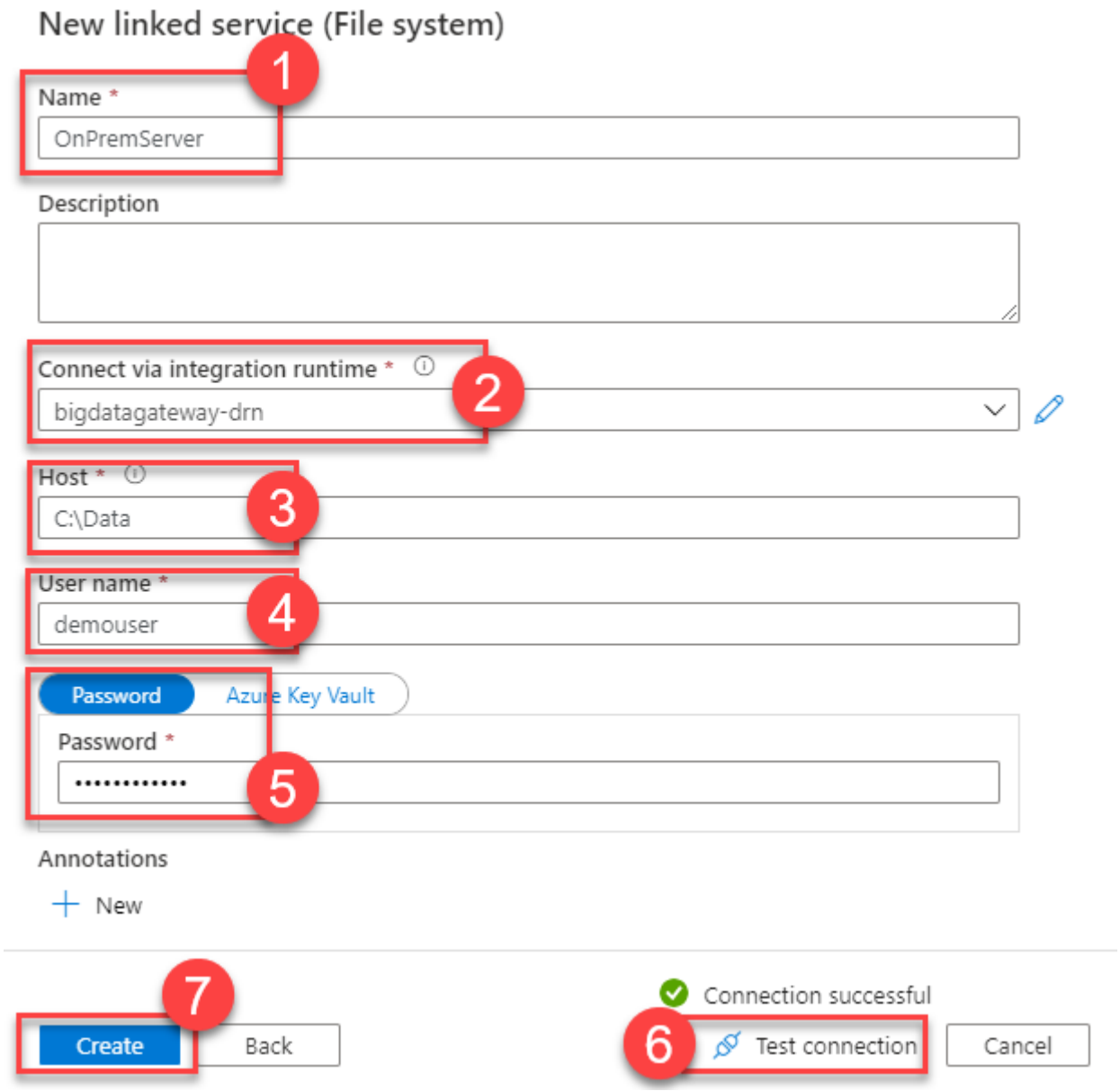


5. Scroll through the options and select **File System (1)**, then select **Continue (2)**.



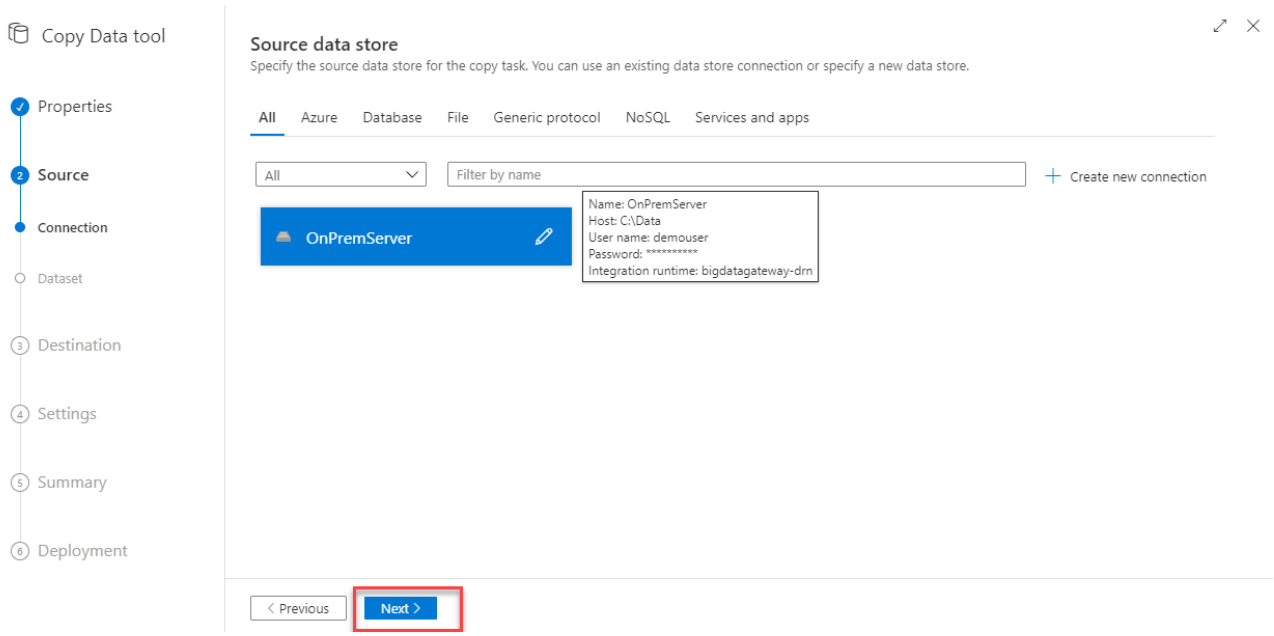
6. In the New Linked Service form, enter the following:

    - **Name (1)**: `OnPremServer`
    - **Connect via integration runtime (2)**: Select the Integration runtime created previously in this exercise.
    - **Host (3)**: **C:\Data**
    - **User name (4)**: Use your machine's login user name.
    - **Password (5)**: Use your machine's login password.

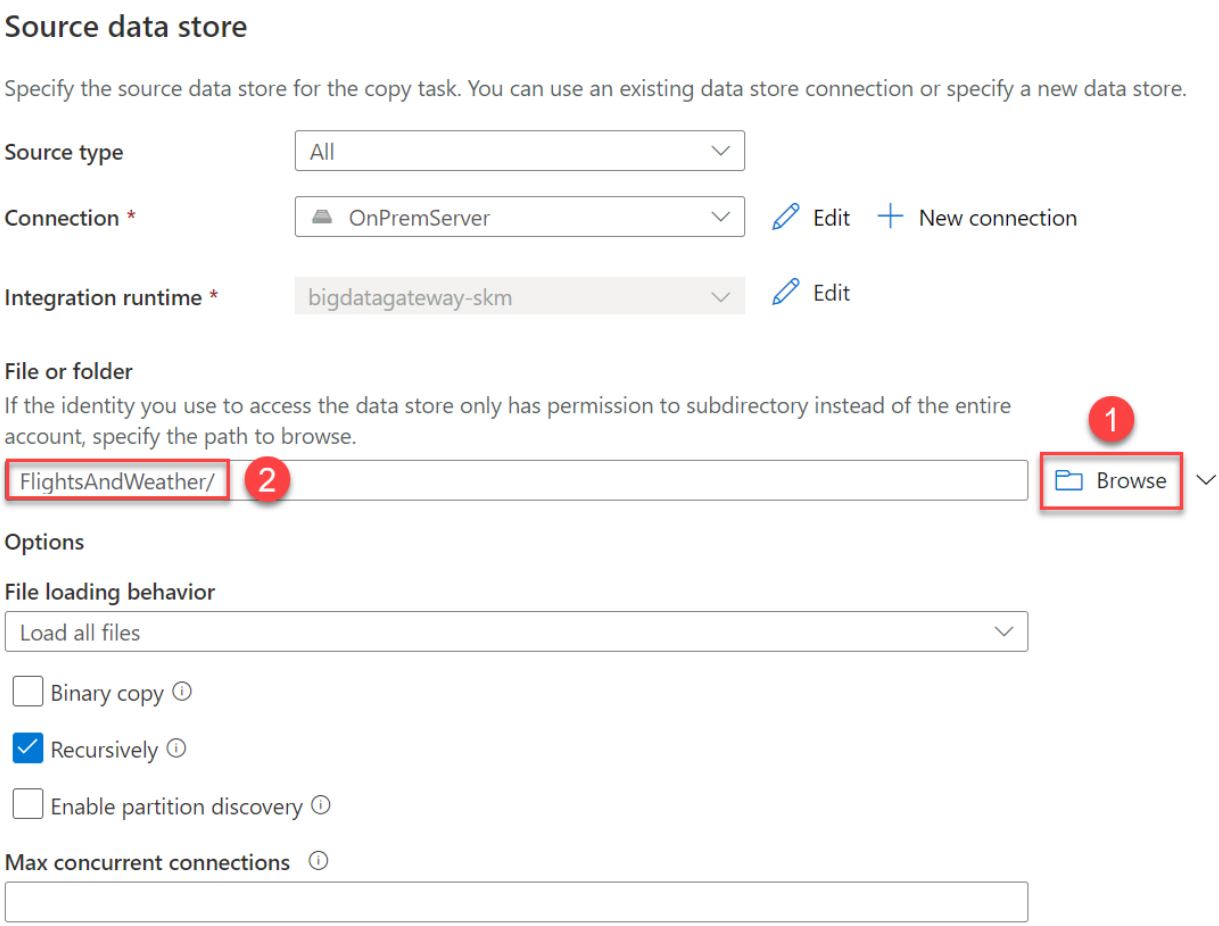


7. Select **Test connection (6)** to verify you correctly entered the values. Finally, select **Create (7)**.

8. On the Source data store page, select **Next**.

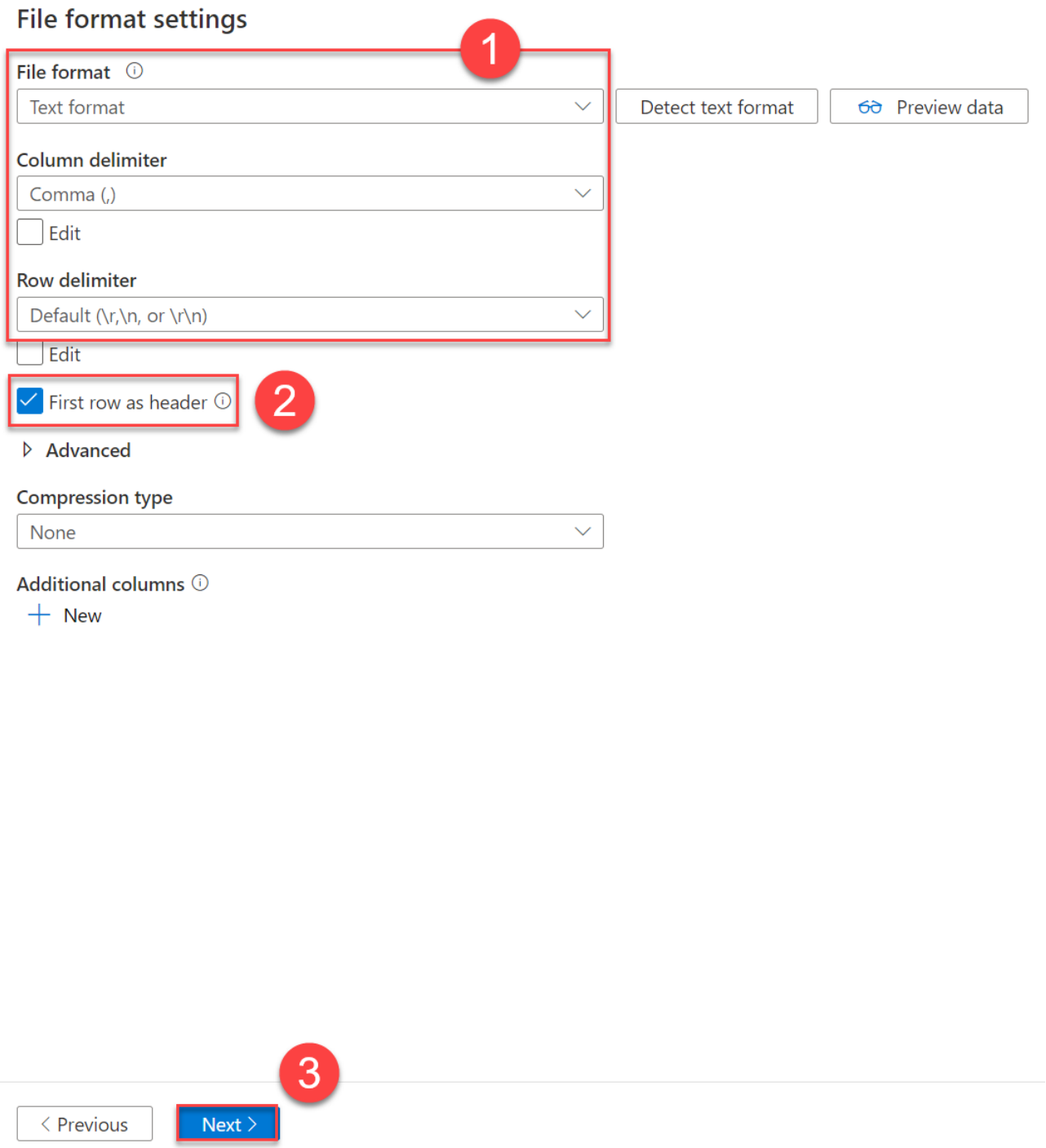


9. On the **Source data store** screen, select **Browse (1)**, then select the **FlightsAndWeather (2)** folder. Next, select **Load all files** under file loading behavior, check **Recursively**, then select **Next**.



10. On the File format settings page, select the following options:

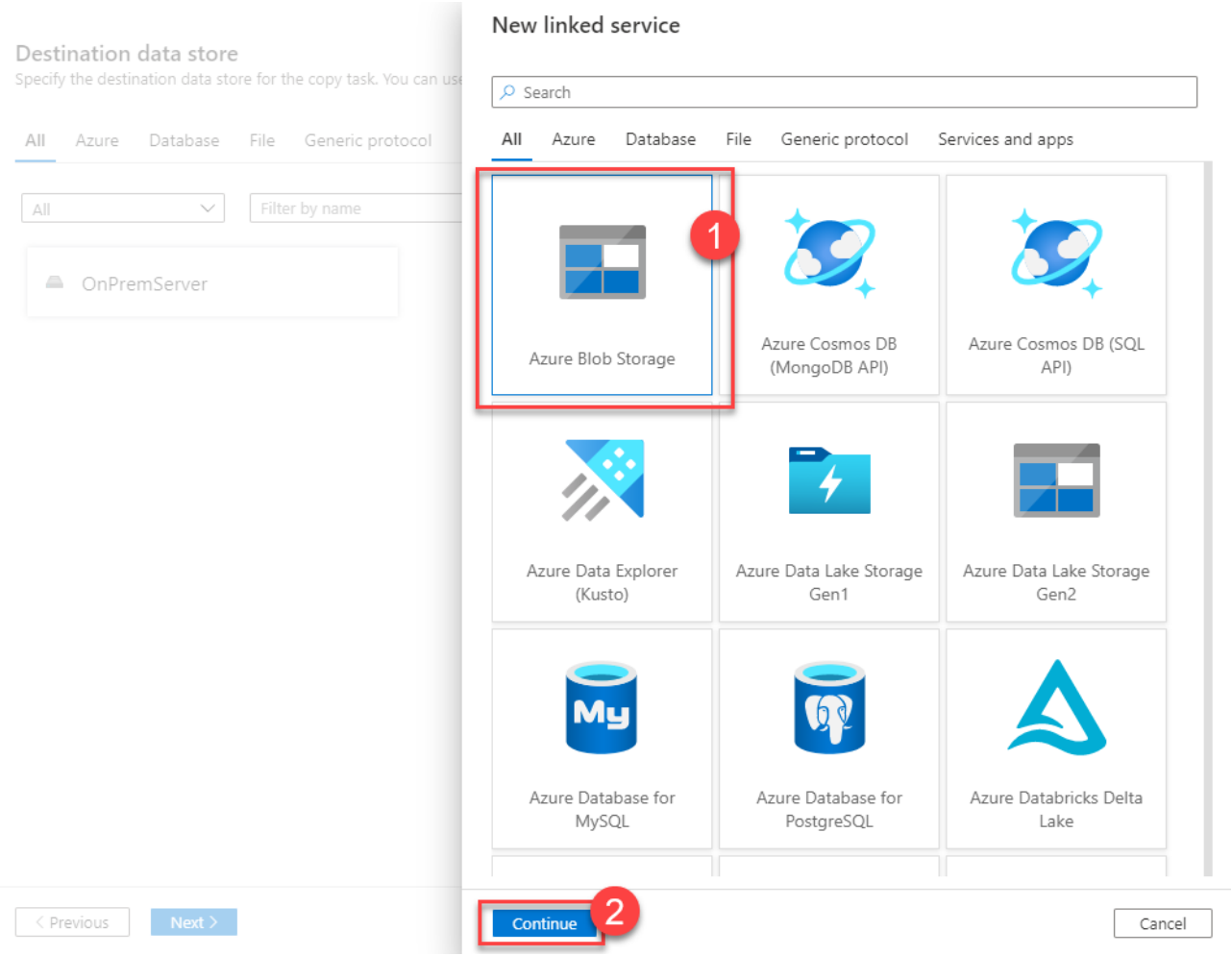
    - **File format (1)**: **Text format**
    - **Column delimiter**: **Comma (,)**
    - **Row delimiter**: **Default (\r, \n, or \r\n)**
    - **First row as header (2)**: **Checked**



11. Select **Next (3)**.

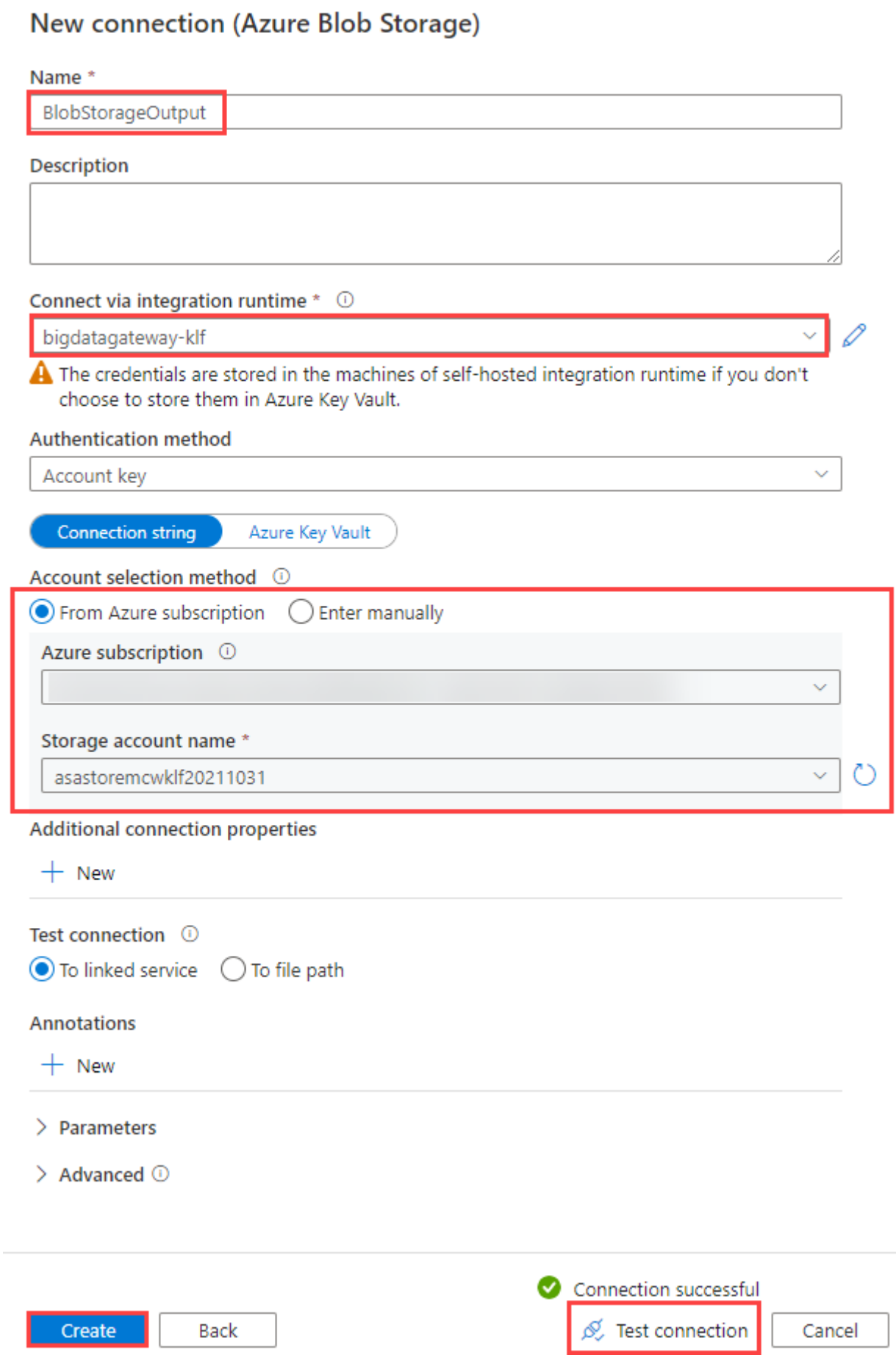
12. On the Destination data store screen, select **+ New connection**.

13. Select **Azure Blob Storage (1)** within the New Linked Service blade, then select **Continue (2)**.



14. On the New Linked Service (Azure Blob Storage) account screen, enter the following, test your connection **(4)**, and then select **Create (5)**.

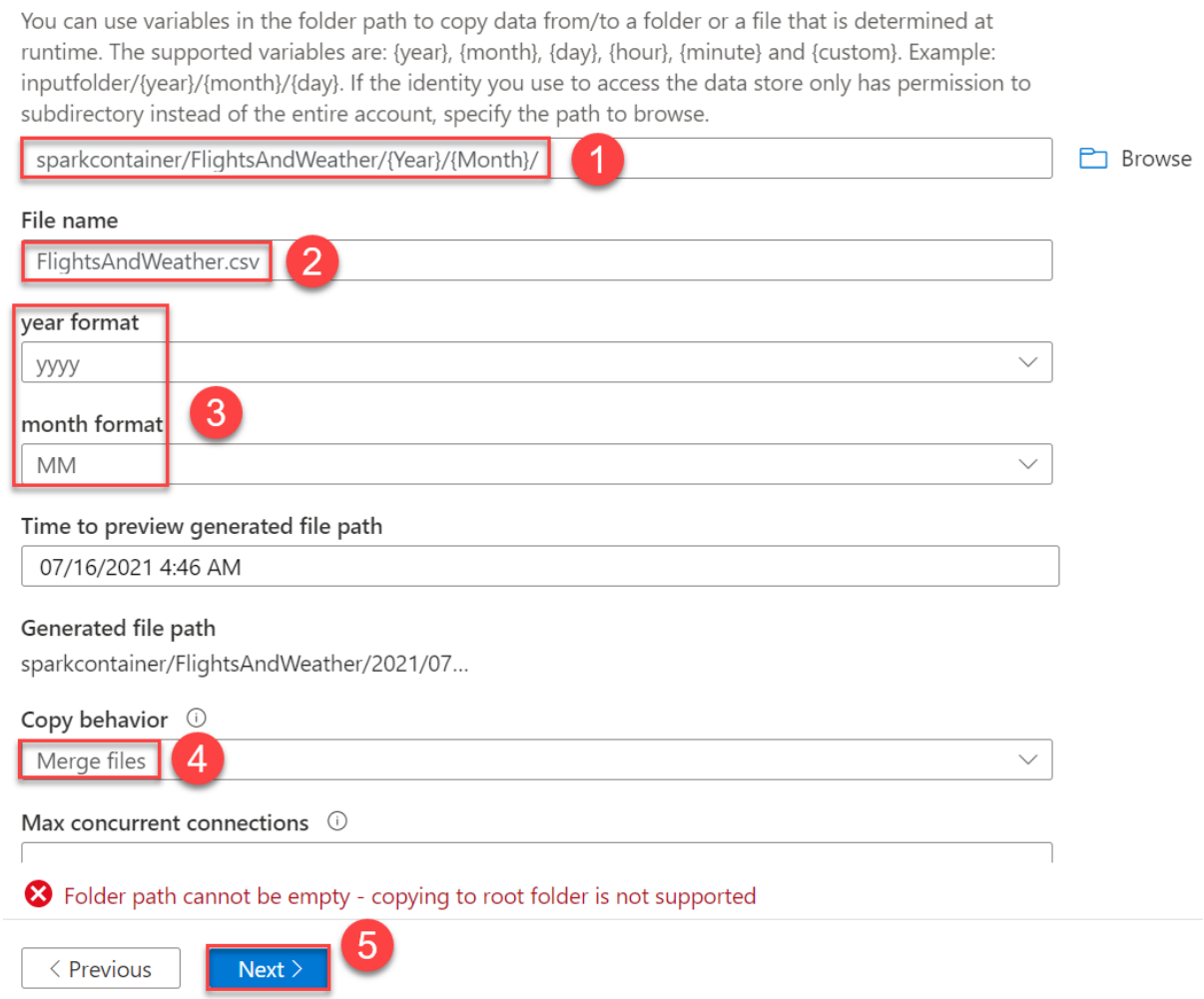
    - **Name**: `BlobStorageOutput`
    - **Connect via integration runtime**: Select your Integration Runtime.
    - **Authentication method**: Select **Account key**
    - **Account selection method**: **From Azure subscription**
    - **Storage account name**: Select the blob storage account you provisioned in the before-the-lab section. It will begin with **asastoremcw**.



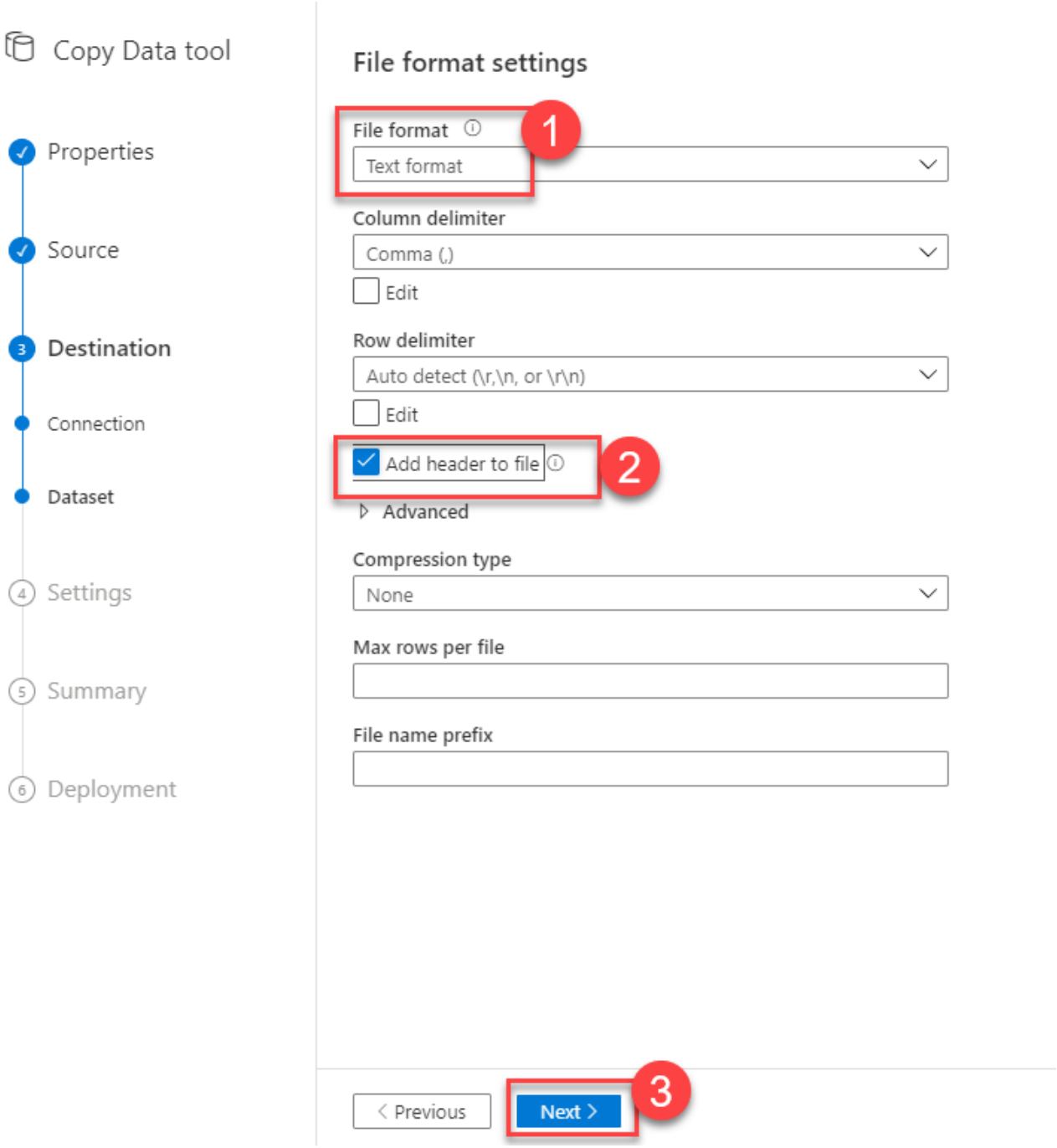
15. On the Destination data store page, configure the Blob Storage output path.

    - **Folder path (1)**: `sparkcontainer/FlightsAndWeather/{Year}/{Month}/`

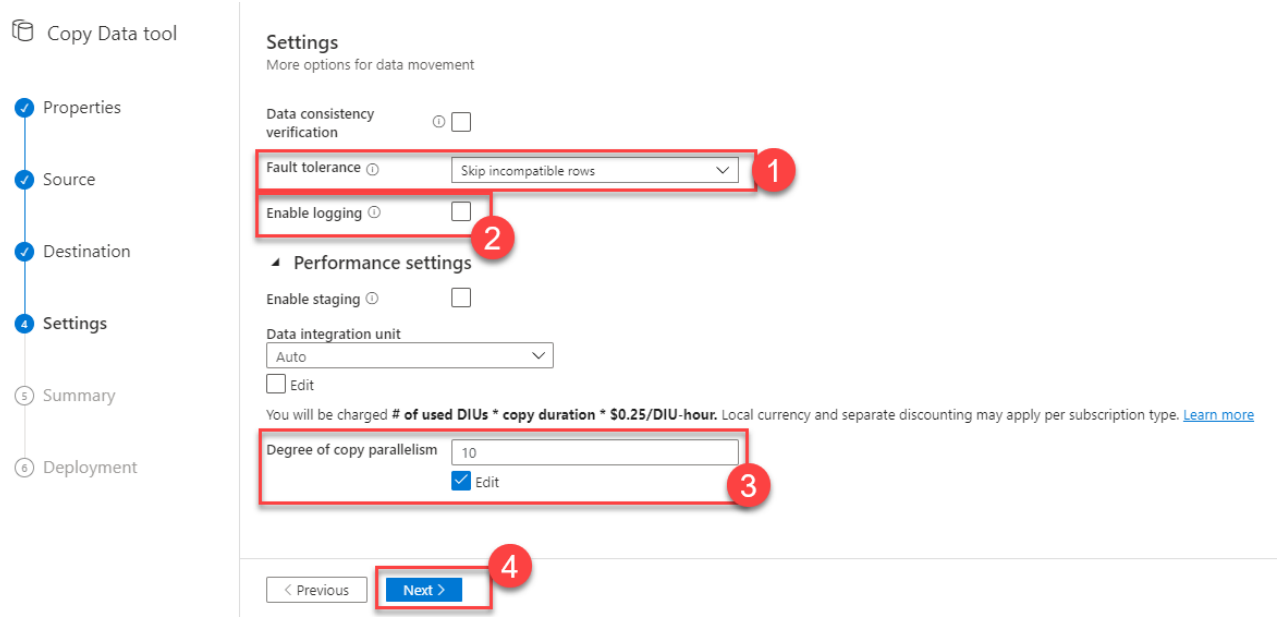
- **Filename (2)**: `FlightsAndWeather.csv`
- **Year (3)**: **yyyy**
- **Month (3)**: **MM**
- **Copy behavior (4)**: **Merge files**
- Select **Next (5)**.



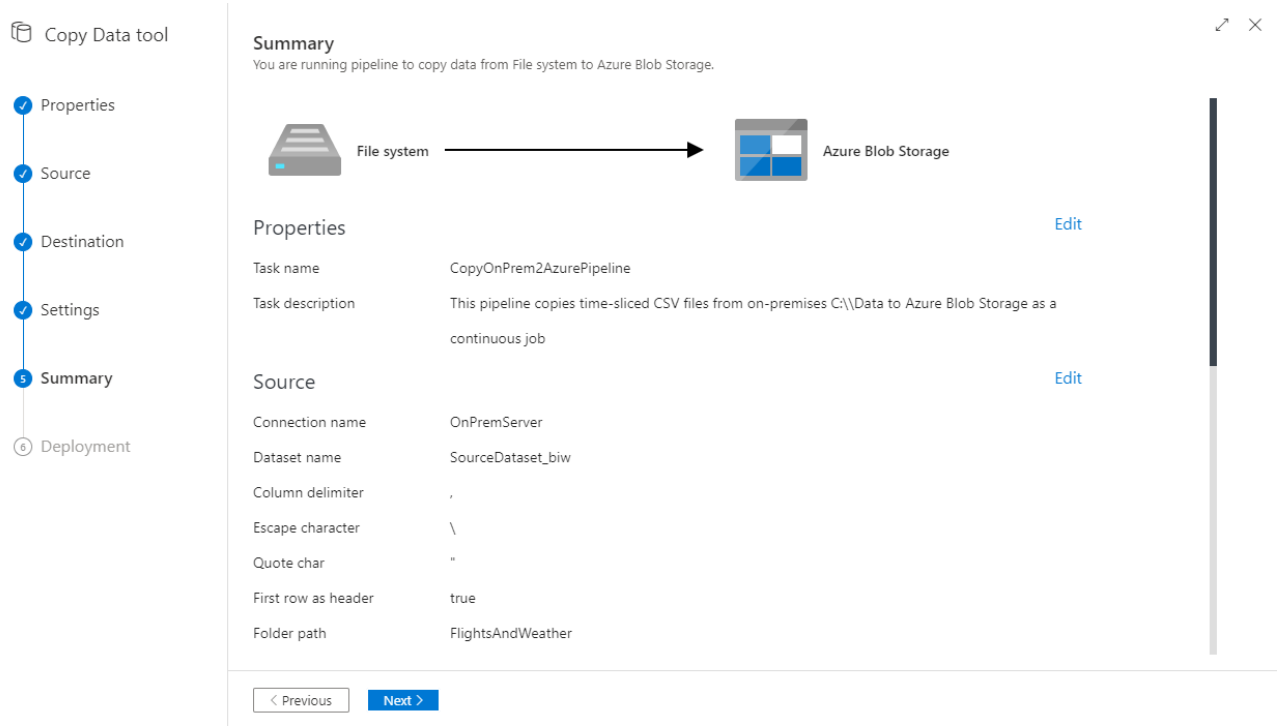
16. On the File format settings screen, select the **Text format (1)** file format, and check the **Add header to file (2)** checkbox, then select **Next (3)**. If present, leave **Max rows per file** and **File name prefix** at their defaults.



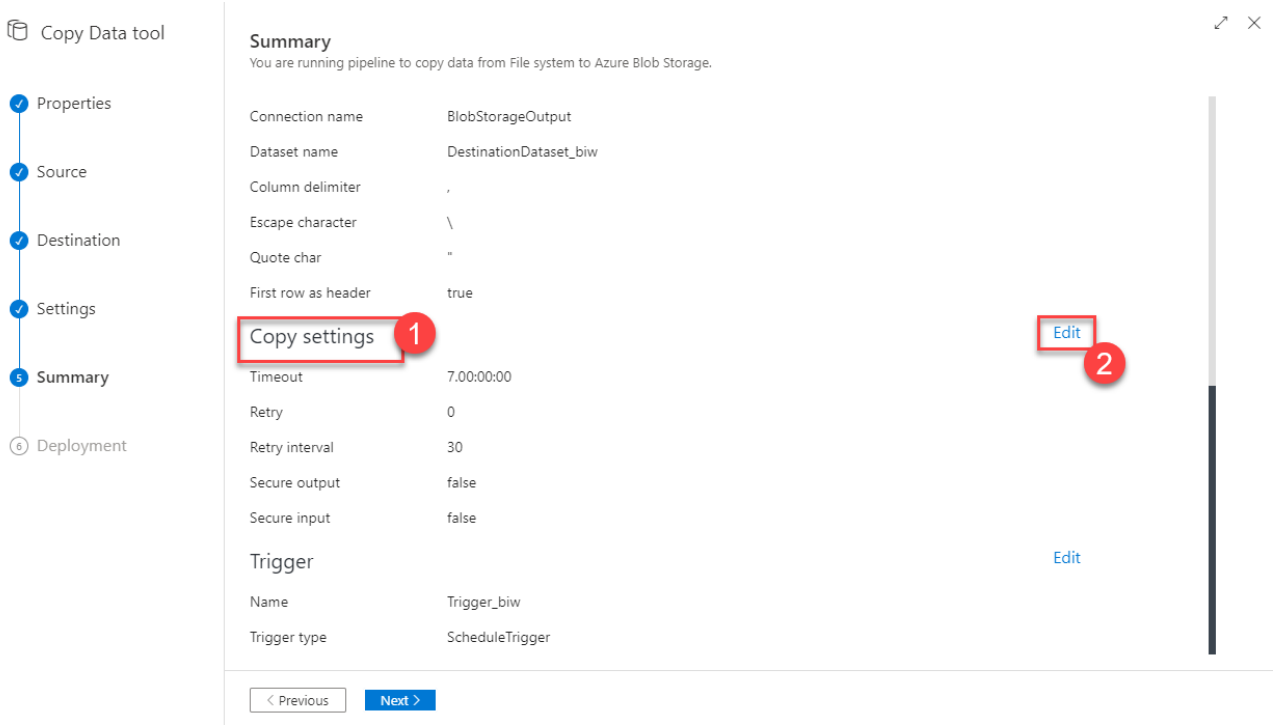
17. On the **Settings** screen, select **Skip incompatible rows (1)** under Fault tolerance, and uncheck **Enable logging (2)**. If present, keep **Data consistency verification** unchecked. Expand Advanced Settings and set Degree of copy parallelism to `10` **(3)**, then select **Next (4)**.



18. Review settings on the **Summary** tab, but **DO NOT choose Next**.

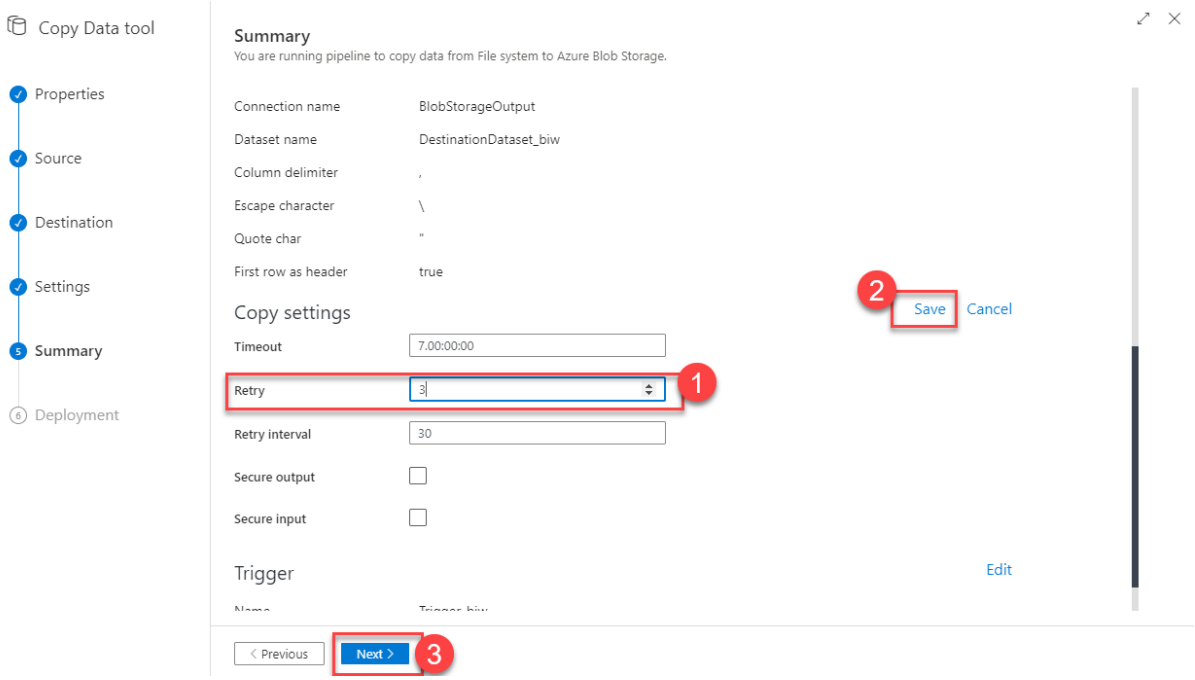


19. Scroll down on the summary page until you see the **Copy Settings (1)** section. Select **Edit (2)** next to **Copy Settings**.



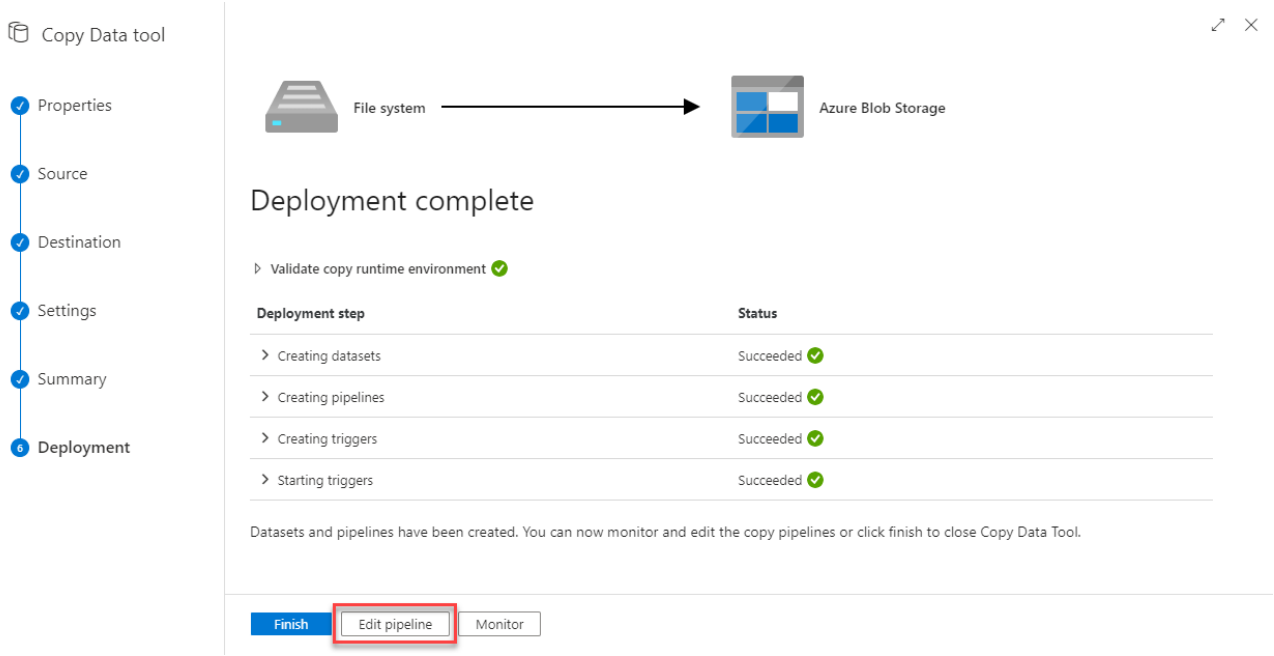
20. Change the following Copy setting:

    - **Retry (1)**: 3
    - Select **Save (2)**.



21. After saving the Copy settings, select **Next (3)** on the Summary tab.

22. On the **Deployment** screen, you will see a message that the deployment is in progress, and after a minute or two, the deployment is completed. Select **Edit Pipeline** to close out of the wizard and navigate to the pipeline editing blade.


Copy Data tool
Properties
Source
Destination
Settings
Summary
Deployment
File system
Azure Blob Storage
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.
Finish
Edit pipeline
Monitor