Azure Databricks workshop – TO Dos

Contents

[1. Validate access to Azure 2](#_Toc506825826)

[2. Create an Azure resource group using az CLI on cloud shell/bash 2](#_Toc506825827)

[3. Create an Azure storage account using az CLI on cloud shell/bash 2](#_Toc506825828)

[4. Create storage containers in the storage account using az CLI on cloud shell/bash 2](#_Toc506825829)

[5. Copy the workshop data to your storage account - storage container nyctaxi-staging 3](#_Toc506825830)

[6. Provision a blank Azure SQL database in the resource group you created 5](#_Toc506825831)

[7. Create tables in Azure SQL database from portal query explorer 6](#_Toc506825832)

[8. Provision Azure Databricks in your resource group, in US East 2 datacenter 7](#_Toc506825833)

[Appendix 9](#_Toc506825834)

[A1. Copy all the workshop data with a single command 9](#_Toc506825835)

# Validate access to Azure

Ensure you can access your Azure pass, and connect to the Azure subscription provided to you for the workshop. Note: There is just enough credit to do the workshop.

# Create an Azure resource group using az CLI on cloud shell/bash

Name: dew-rg  
Location: US East 2

az group create \

--name dew-rg \

--location eastus2

# Create an Azure storage account using az CLI on cloud shell/bash

Resource group: dew-rg  
Name: dew<yourinitials>sa

az storage account create \

--name <yourStorageAccount> \

--resource-group dew-rg \

--location eastus2 \

--sku Standard\_LRS

# Create storage containers in the storage account using az CLI on cloud shell/bash

* 1. Get storage account key listing

az storage account keys list \

--account-name <yourStorageAccount > \

--resource-group dew-rg \

--output table

* 1. Export storage account details

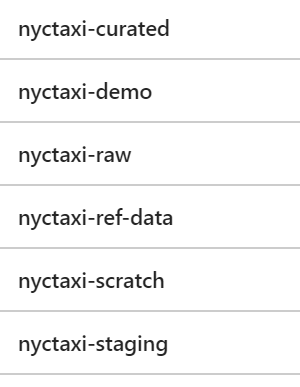
export AZURE\_STORAGE\_ACCOUNT="<YourStorageAccount>"

export AZURE\_STORAGE\_ACCESS\_KEY="<YourKey>"

* 1. Create storage containers

az storage container create --name nyctaxi-consumption  
az storage container create --name nyctaxi-curated  
az storage container create --name nyctaxi-demo  
az storage container create --name nyctaxi-raw  
az storage container create --name nyctaxi-ref-data  
az storage container create --name nyctaxi-staging  
az storage container create --name nyctaxi-scratch

This is what it should look like in the portal.



# Copy the workshop data to your storage account - storage container nyctaxi-staging

* 1. Export source storage credentials

These are the storage account details to copy data from. Do not modify this.

export SRC\_STORAGE\_ACCOUNT="nyctaxidew"

export SRC\_STORAGE\_ACCESS\_KEY="8pzBEhPSUJKcZ+Jol6wWq5wM2RtCQNk75r577kt2ed/LYxmOOiAZdxqN2wMHfN2duhlswqGbT3Pw/YdHtuqngw=="

* 1. Copy reference data

azcopy \

--source https://nyctaxidew.blob.core.windows.net/nyctaxi-staging/reference-data/ \

--destination https://<YourStorageAccount>.blob.core.windows.net/nyctaxi-staging/reference-data/ \

--source-key $SRC\_STORAGE\_ACCESS\_KEY \

--dest-key $AZURE\_STORAGE\_ACCESS\_KEY \

--sync-copy \

--recursive

* 1. Copy transactional data for 2015/2016/2017

azcopy \

--source https://nyctaxidew.blob.core.windows.net/nyctaxi-staging/transactional-data/year=2015/ \

--destination https://<YourStorageAccount>.blob.core.windows.net/nyctaxi-staging/transactional-data/year=2015/ \

--source-key $SRC\_STORAGE\_ACCESS\_KEY \

--dest-key $AZURE\_STORAGE\_ACCESS\_KEY \

--sync-copy \

--recursive

azcopy \

--source https://nyctaxidew.blob.core.windows.net/nyctaxi-staging/transactional-data/year=2016/ \

--destination https://<YourStorageAccount>.blob.core.windows.net/nyctaxi-staging/transactional-data/year=2016/ \

--source-key $SRC\_STORAGE\_ACCESS\_KEY \

--dest-key $AZURE\_STORAGE\_ACCESS\_KEY \

--sync-copy \

--recursive

azcopy \

--source https://nyctaxidew.blob.core.windows.net/nyctaxi-staging/transactional-data/year=2017/ \

--destination https://<YourStorageAccount>.blob.core.windows.net/nyctaxi-staging/transactional-data/year=2017/ \

--source-key $SRC\_STORAGE\_ACCESS\_KEY \

--dest-key $AZURE\_STORAGE\_ACCESS\_KEY \

--sync-copy \

--recursive

5.4. List data copied

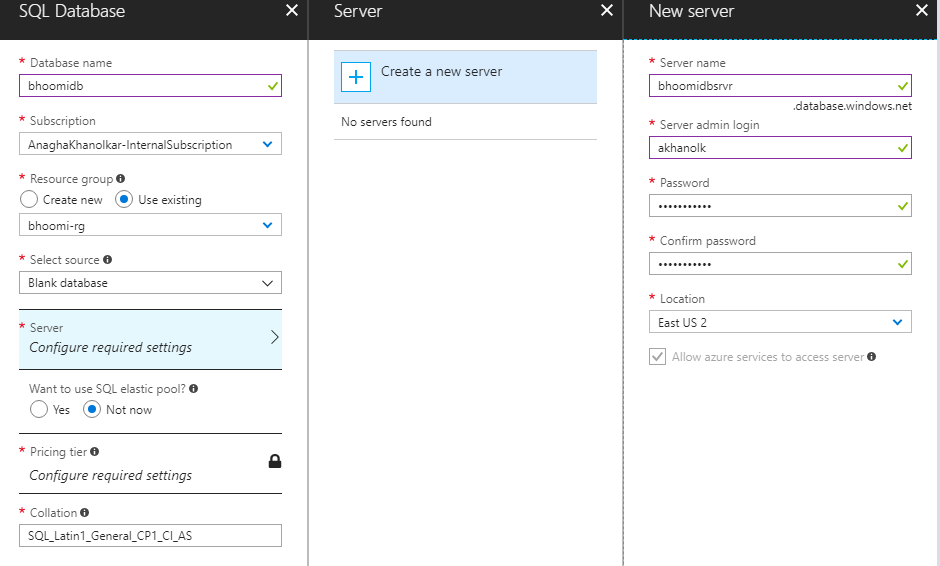
az storage blob list \

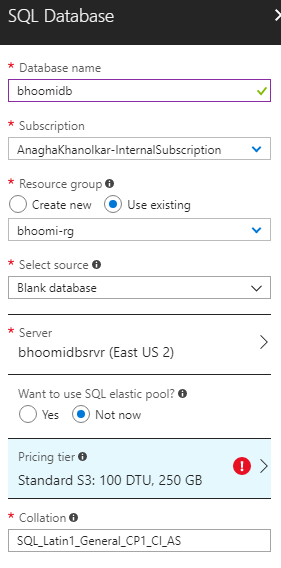
--container-name nyctaxi-staging \  
--output table

5.5. To copy all the data with a single command  
Refer appendix section A1

# Provision a blank Azure SQL database in the resource group you created

Use a database name and server name that is easy to type up and remember.  
Make a note of the credentials





# Create tables in Azure SQL database from portal query explorer

7.1. Trips by year

DROP TABLE IF EXISTS TRIPS\_BY\_YEAR;

CREATE TABLE TRIPS\_BY\_YEAR (

TAXI\_TYPE VARCHAR(10),

TRIP\_YEAR INT,

TRIP\_COUNT BIGINT

);

7.2. Trips by hour

DROP TABLE IF EXISTS TRIPS\_BY\_HOUR;

CREATE TABLE TRIPS\_BY\_HOUR (

TAXI\_TYPE VARCHAR(10),

TRIP\_YEAR INT,

TRIP\_HOUR INT,

TRIP\_COUNT BIGINT

);

7.3. Batch job history

DROP TABLE IF EXISTS BATCH\_JOB\_HISTORY;

CREATE TABLE BATCH\_JOB\_HISTORY

(

batch\_id int,

batch\_step\_id int,

batch\_step\_description varchar(50),

batch\_step\_status varchar(10),

batch\_step\_time varchar(25)

);

ALTER TABLE BATCH\_JOB\_HISTORY

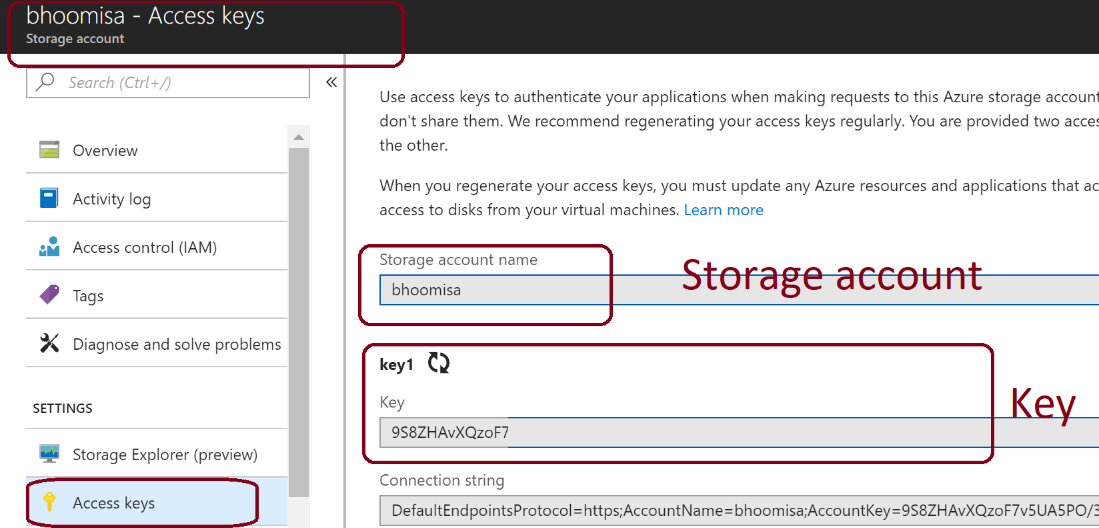
ADD CONSTRAINT batch\_step\_time\_def

DEFAULT CURRENT\_TIMESTAMP FOR batch\_step\_time;

# Provision Azure Databricks in your resource group, in US East 2 datacenter

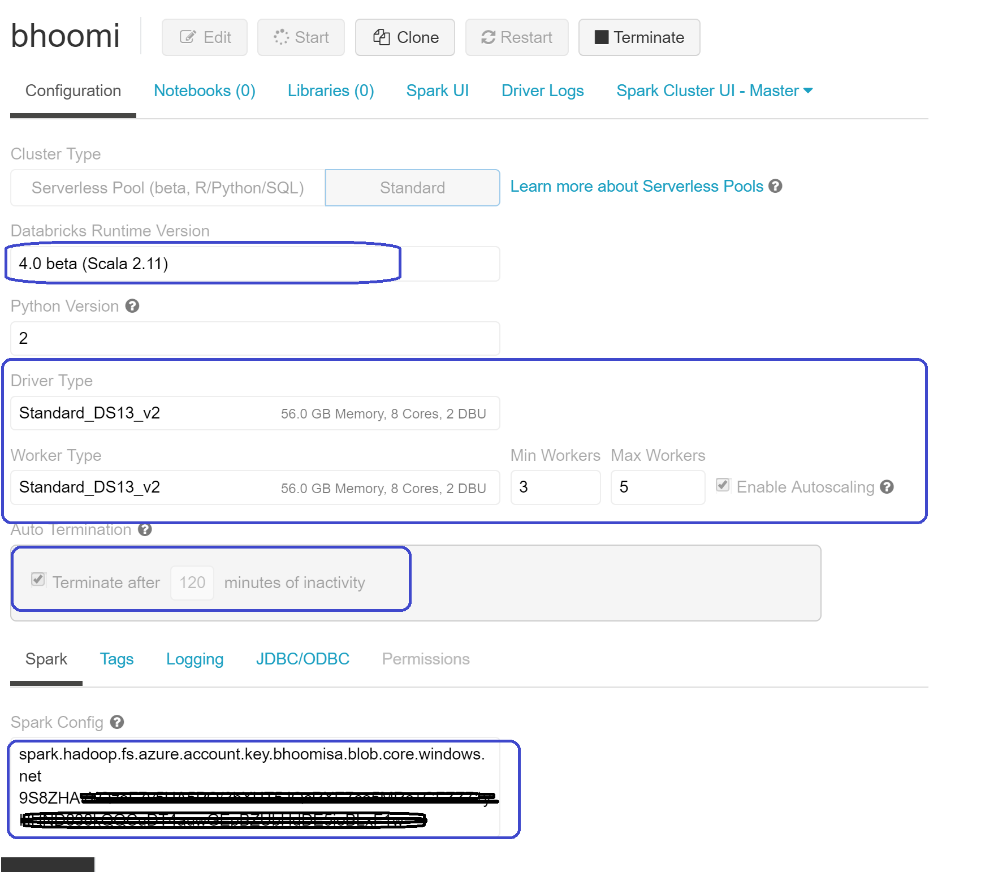
8.1. Storage account credentials

Make a note of the storage account credentials:  
Name of storage account: e.g. bhoomisa  
Key to storage account: 9S8ZHA……



* 1. Provision Databricks and set up a cluster

For the lab – we will use DS13\_v2 VM SKU for driver and executors.If you are running against all of the NYC taxi dataset, provision a cluster with DS13\_v2 as shown below till you can tune your work to operate with lower SKUs.



In the Spark config section paste a key value pair that is as follows:

spark.hadoop.fs.azure.account.key.<yourStorageAccount>.blob.core.windows.net <yourKey>

# Appendix

## A1. Copy all the workshop data with a single command using cloud shell

Replace 5.2 and 5.3 with this single command:

azcopy \

--source https://nyctaxidew.blob.core.windows.net/nyctaxi-staging/ \

--destination https://<YourStorageAccountKey>.blob.core.windows.net/nyctaxi-staging/ \

--source-key $SRC\_STORAGE\_ACCESS\_KEY \

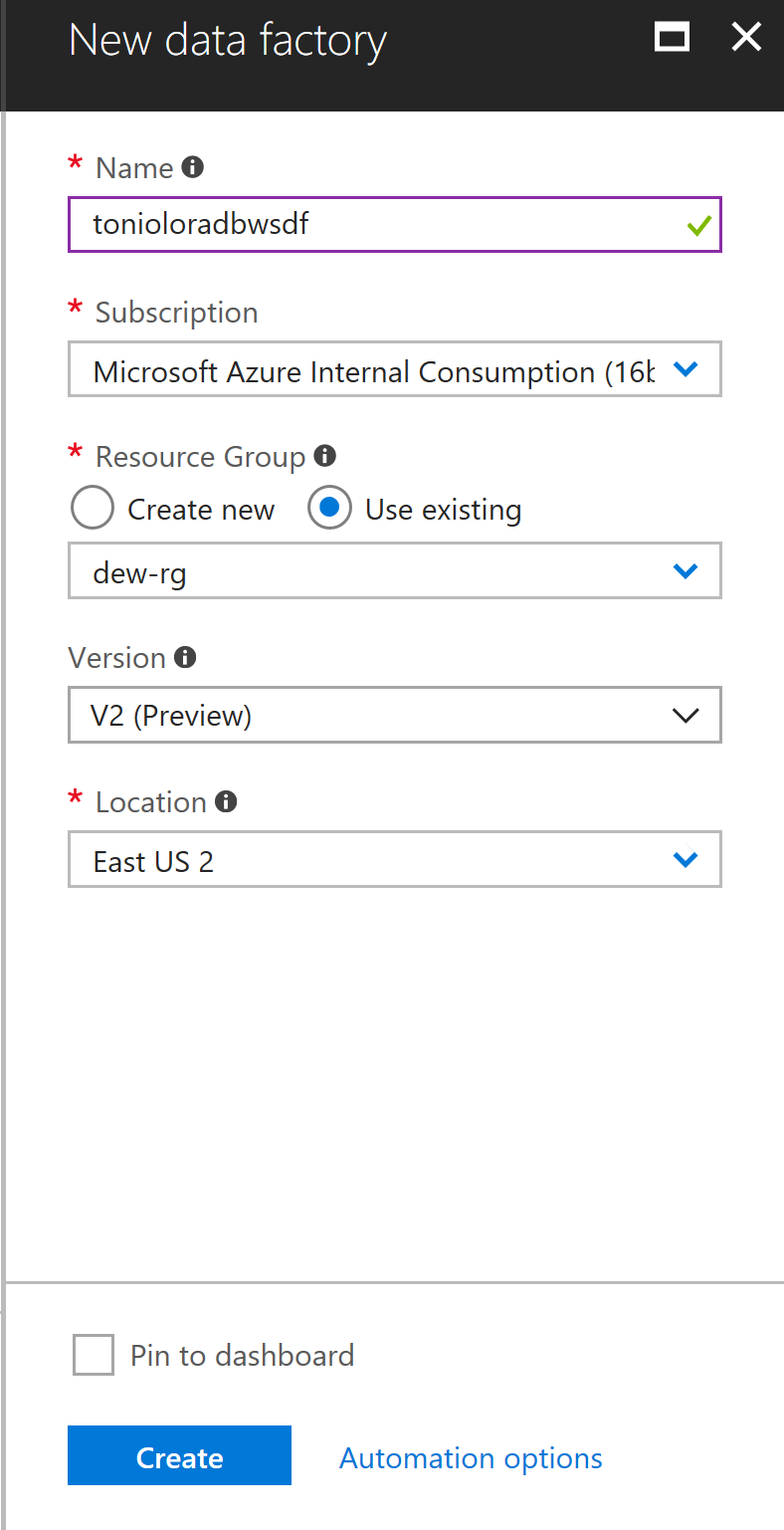
--dest-key $AZURE\_STORAGE\_ACCESS\_KEY \

--sync-copy \

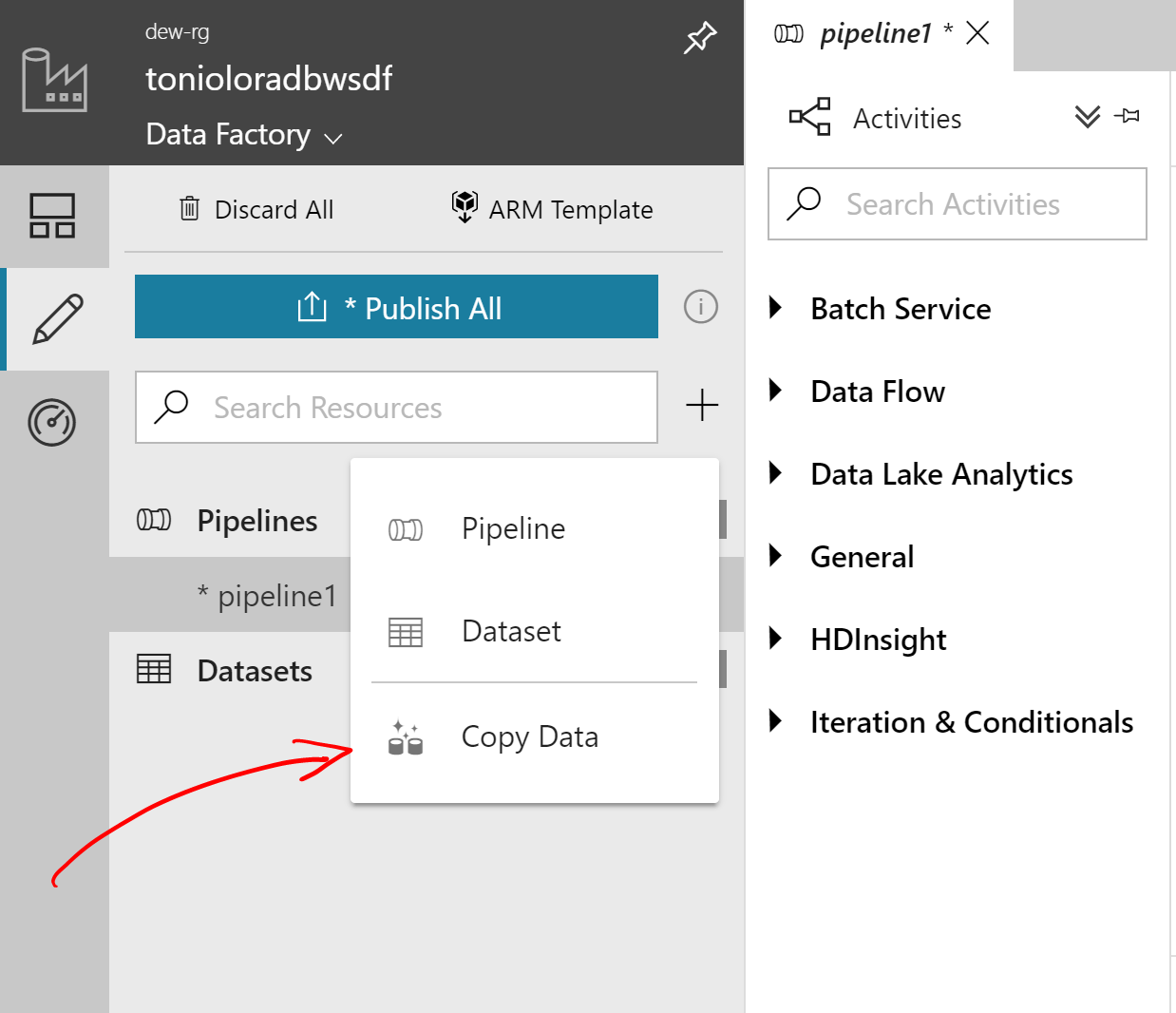
--recursive

## A2. Copy all the workshop data using Azure Data factory

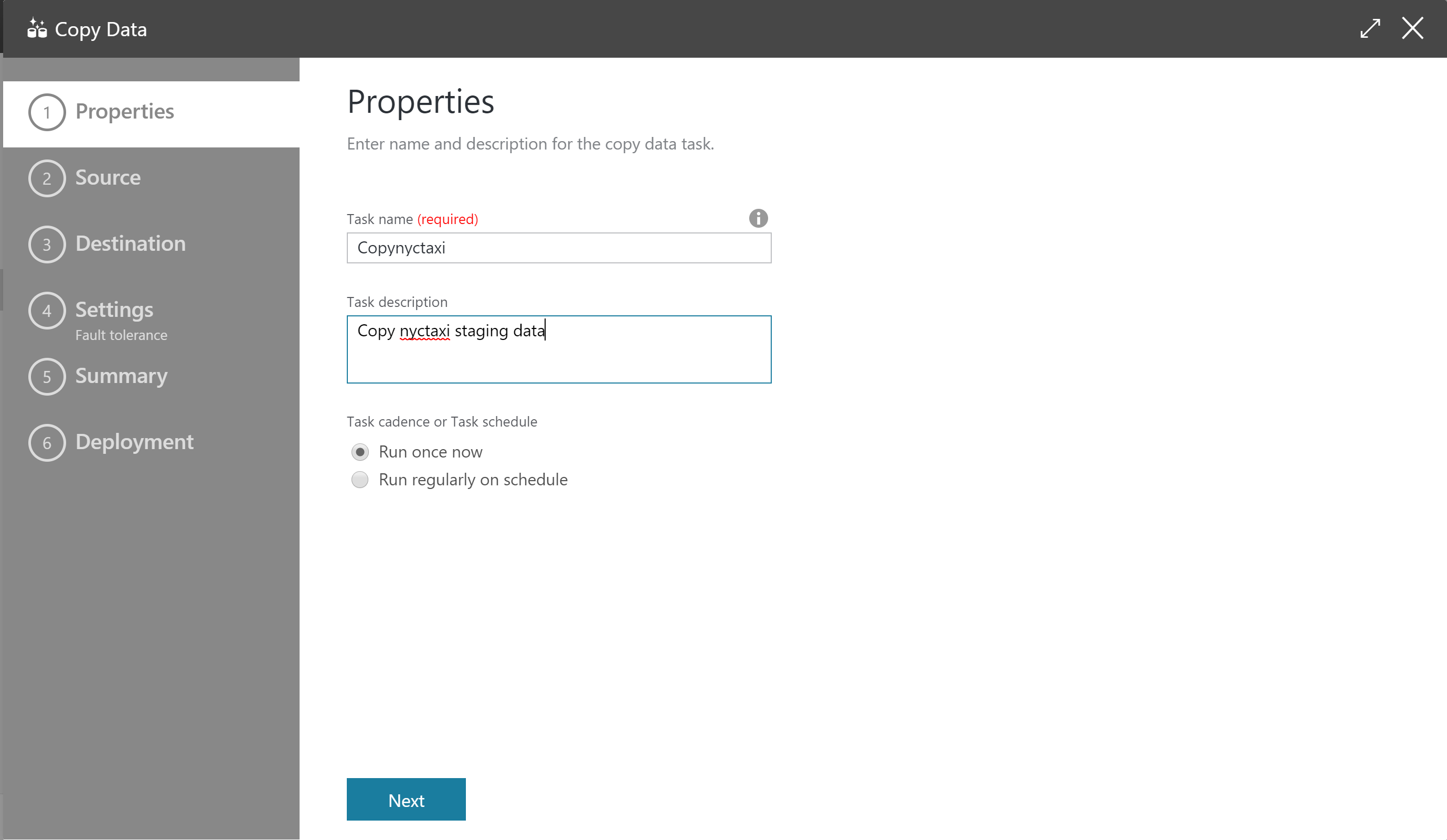
1. Create an Azure Data Factory



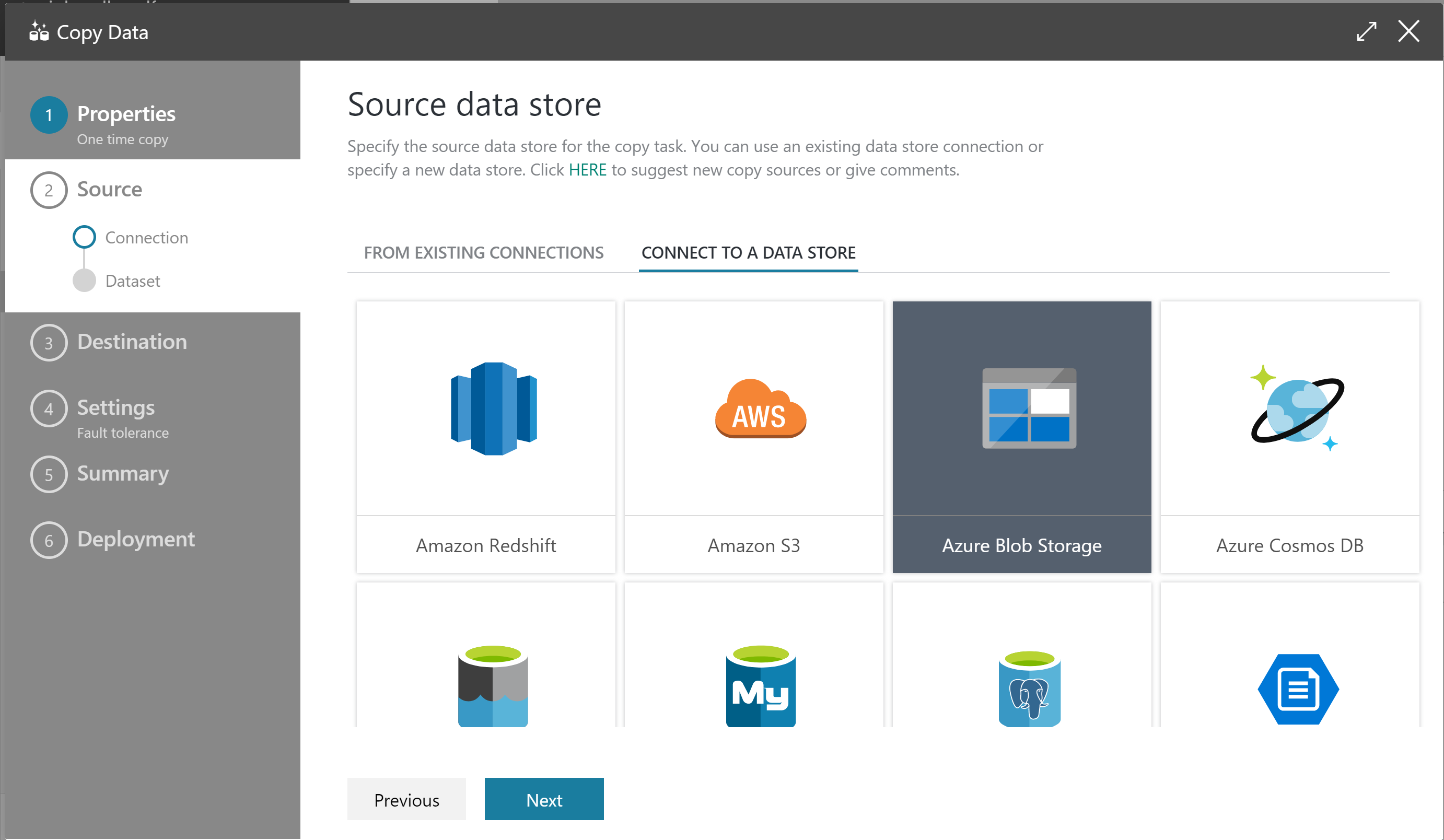
1. Click the Author and Monitor Quick Link
2. Select Create Pipeline Link
3. Create a new Copy Data pipeline



1. Enter name and description



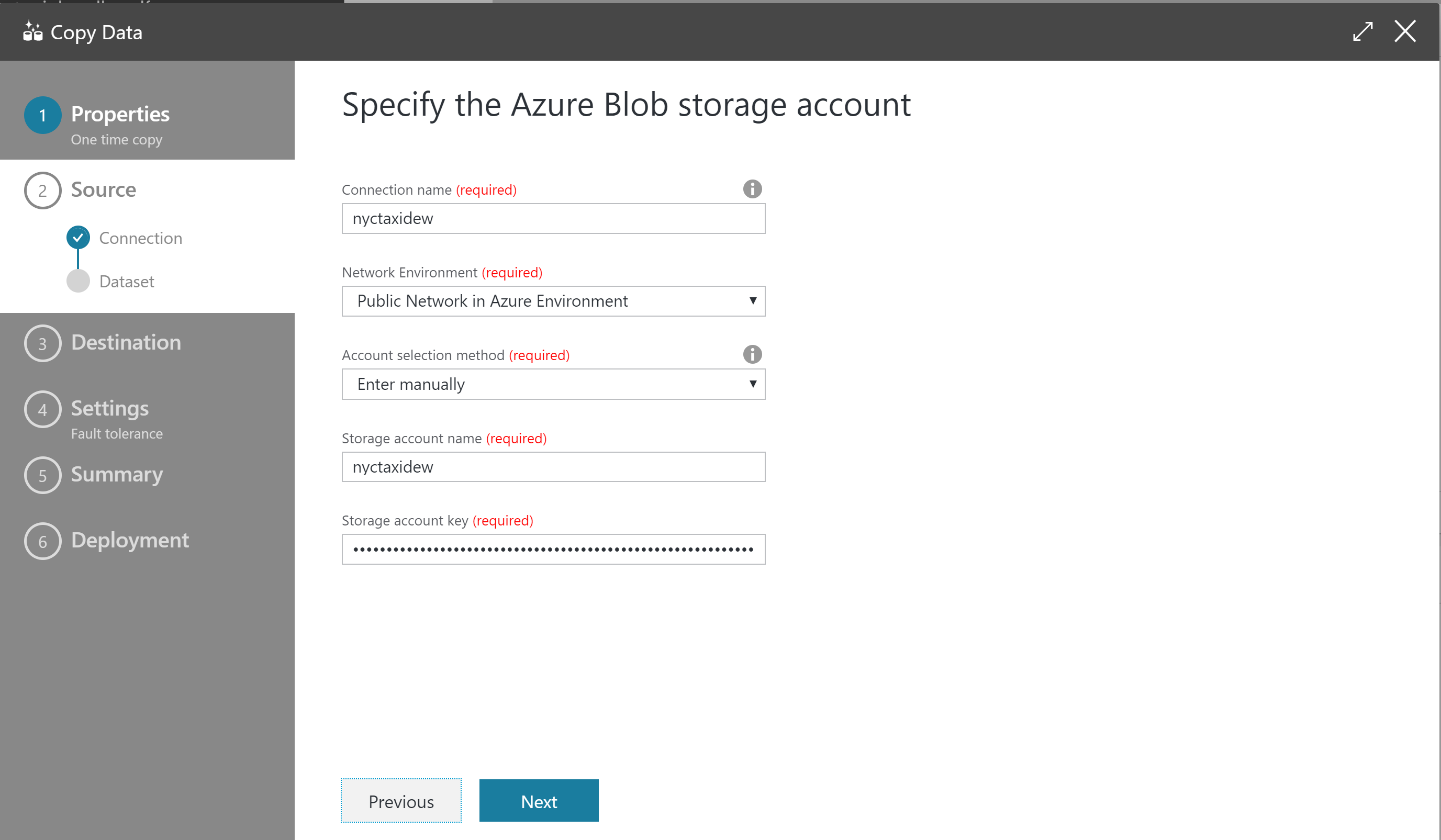
1. Select Azure Blob Storage as Source Data



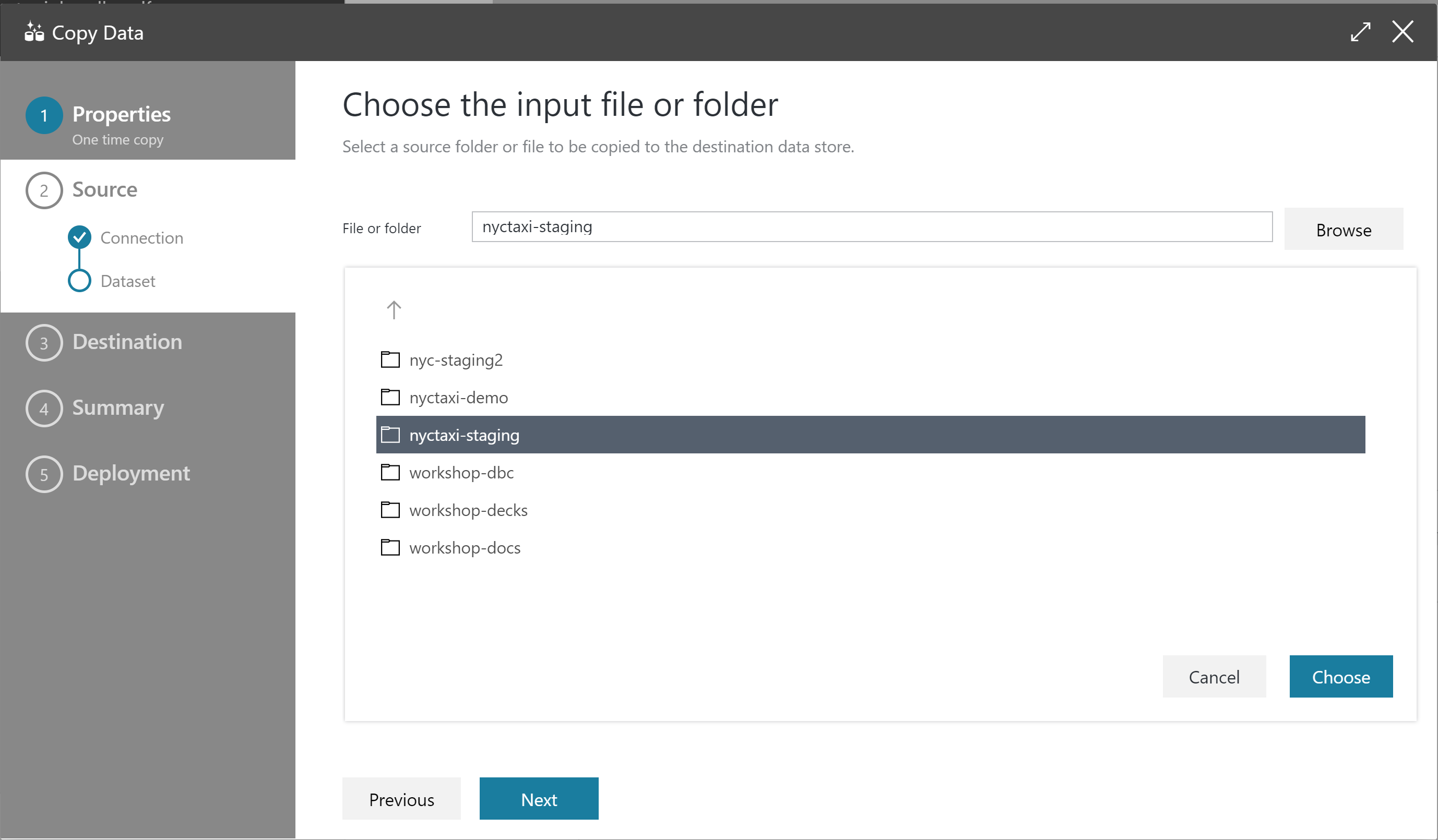
1. Enter the name and credentials for your source data:

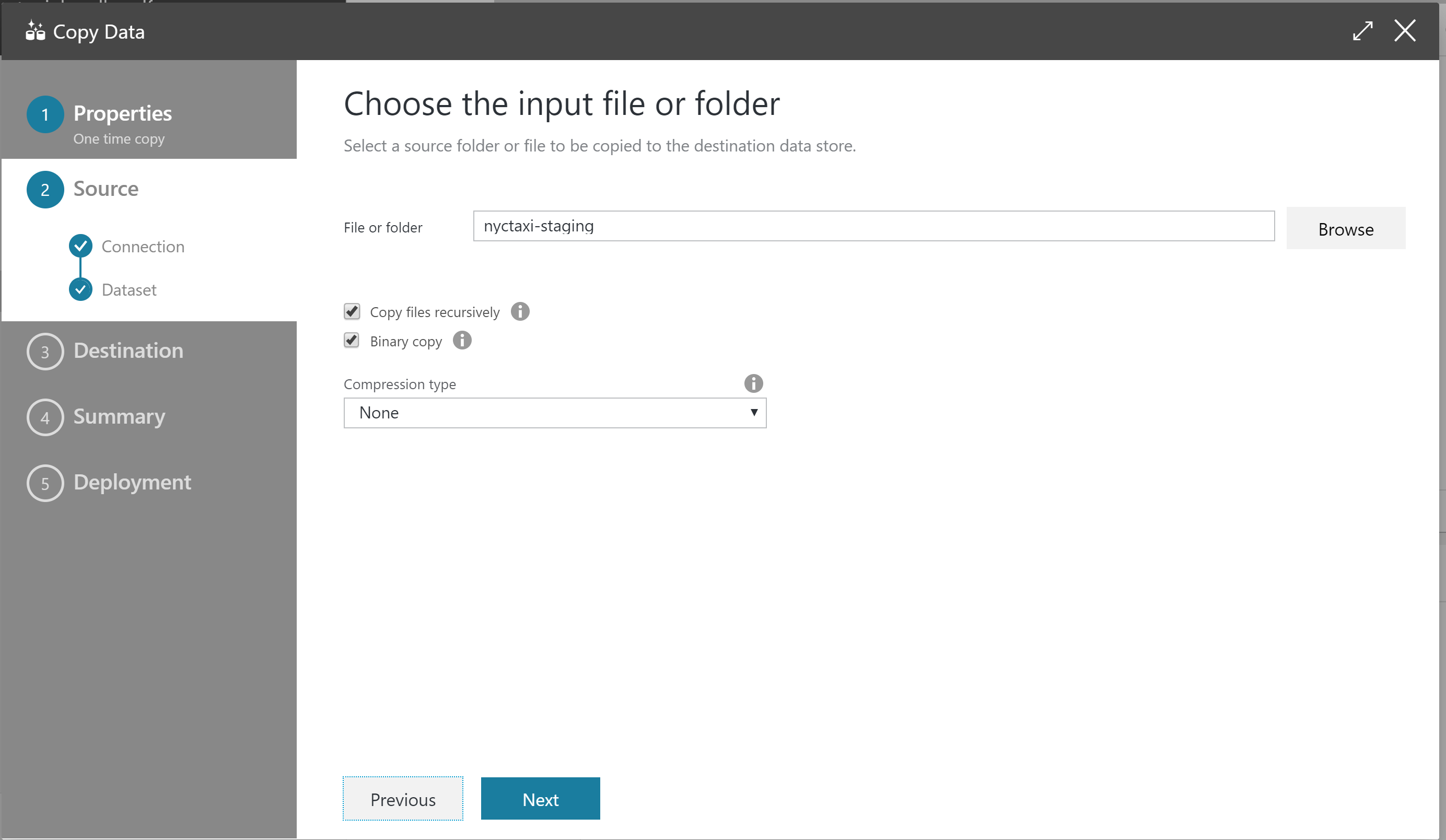
Storage Account Name: nyctaxidew

Storage Account Key: 8pzBEhPSUJKcZ+Jol6wWq5wM2RtCQNk75r577kt2ed/LYxmOOiAZdxqN2wMHfN2duhlswqGbT3Pw/YdHtuqngw==

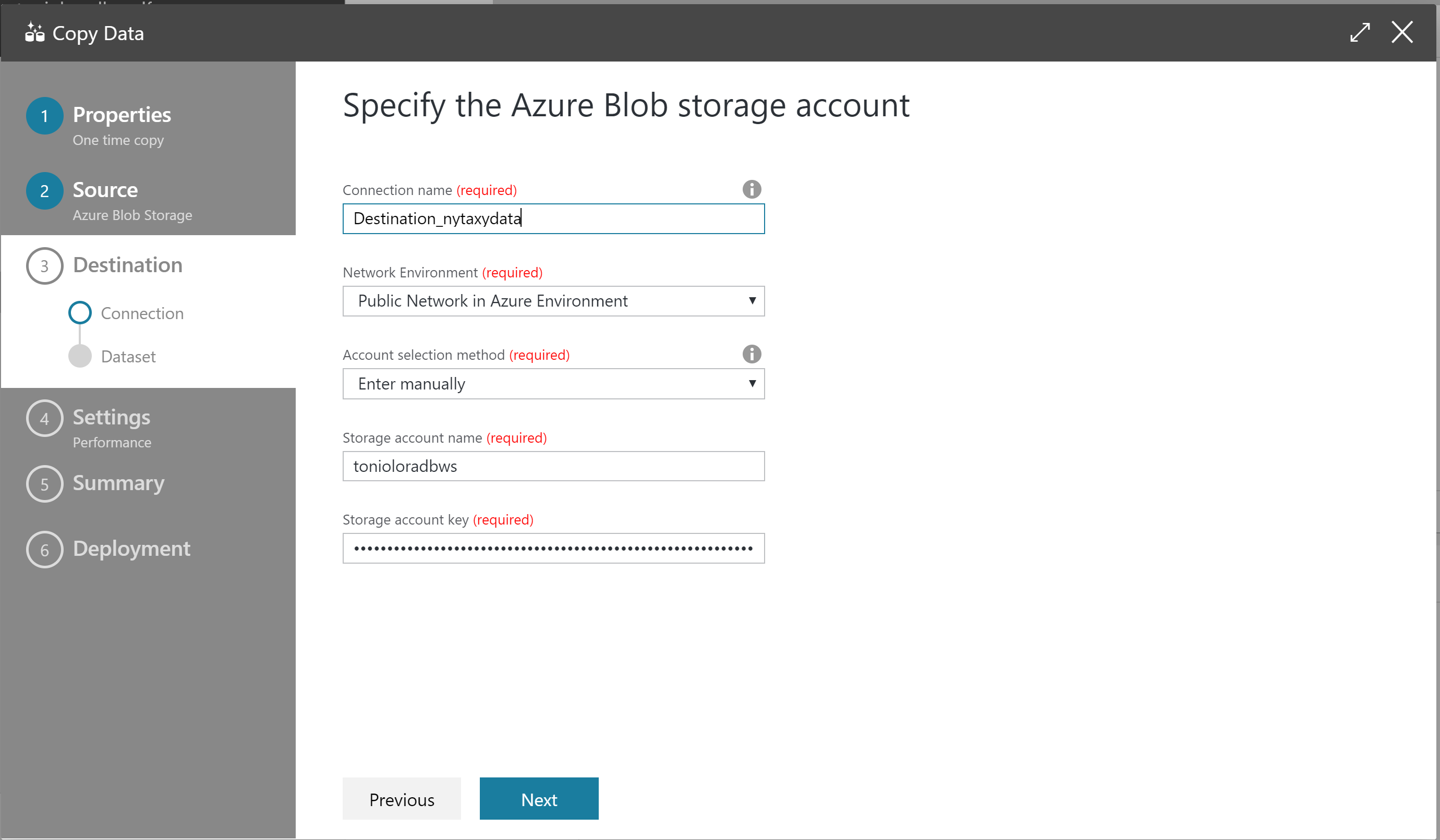


1. Choose the nyctaxi-staging folder and select both “Copy files recursively” and “Binary Copy”

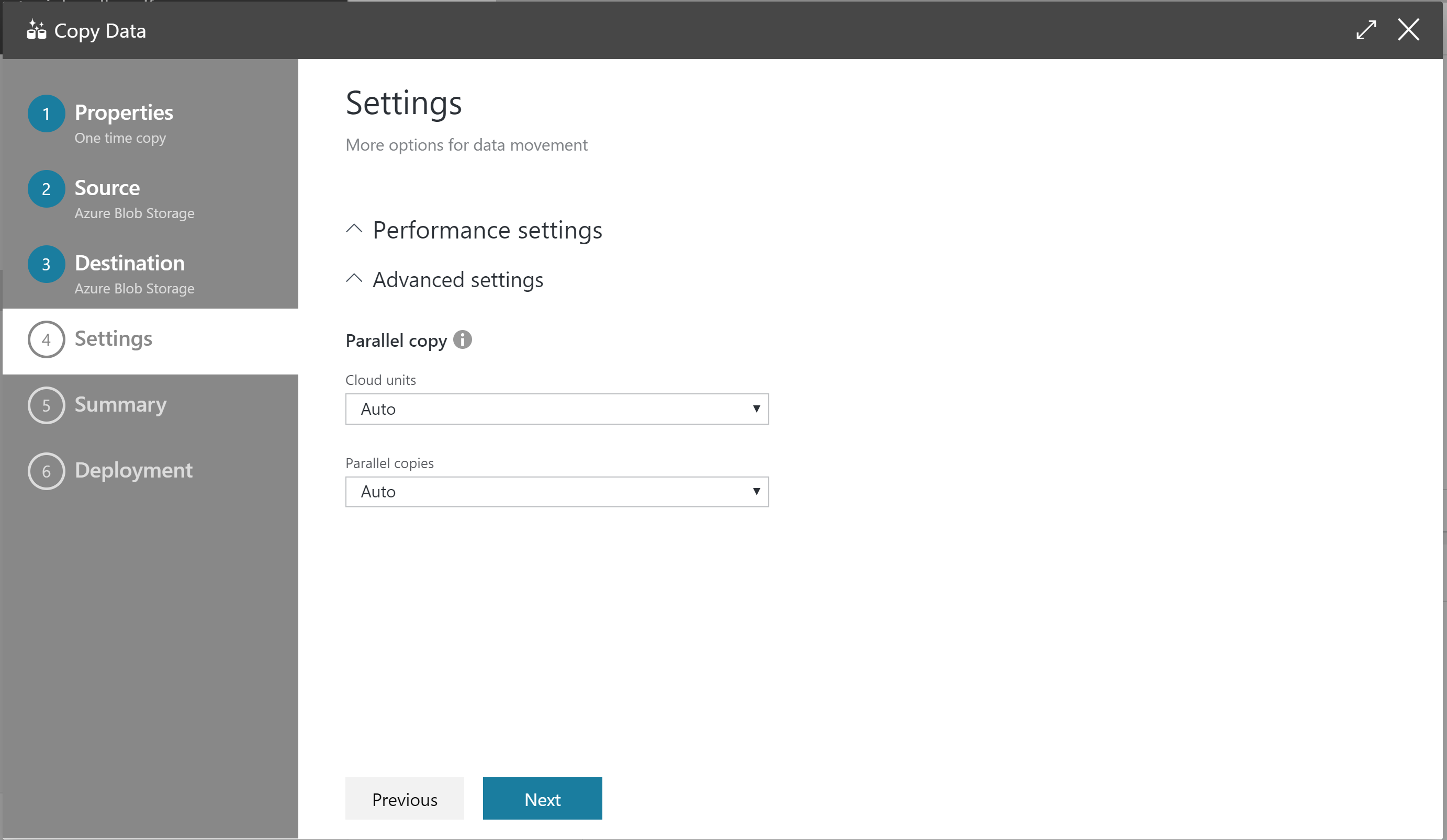




1. Select Azure Blob Storage as your Destination data store. Enter te account name and the key you saved earlier.



1. Choose the nyctaxi-staging target.



1. Continue until the end of the wizard and select Monitor Pipeline