Azure Data Factory

Lab : Ingest Data from HTTP source to Azure Storage:

Pre-requisites:

- Azure Pass subscription
- Azure Data Lake Storage Gen2 storage account
- Azure Data Factory

Lab Objective:

After completing this lab, you will be able to:

- Ingest data using the Copy Activity
- Create Linked service and Dataset

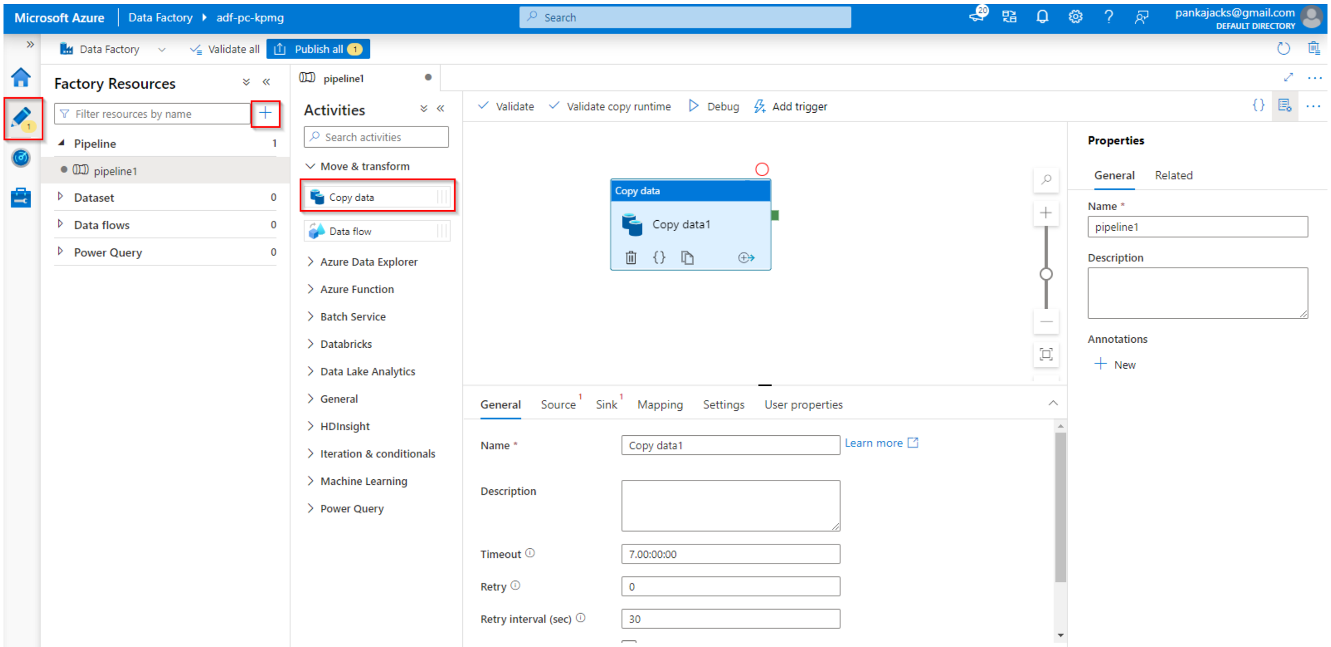
Exercise : Ingest Data using copy activity.

The task for this exercise are as follows:

1. Add the Copy Activity to the pipeline area
2. Create a new HTTP dataset to use as a source
3. Create a new ADLS Gen2 sink
4. Test the Copy Activity

Task 1: Add the Copy Activity to the designer

1. Launch data factory studio from Azure Portal.
2. **Open the authoring canvas** If coming from the ADF homepage, click on the **pencil icon** on the left sidebar and select the **+ pipeline button** to open the authoring canvas and create a pipeline.
3. **Add a copy activity** In the Activities pane, open the **Move and Transform** accordion and drag the **Copy data** activity onto the pipeline canvas.



Name * pipeline1

Description

Name * Copy data1

Description

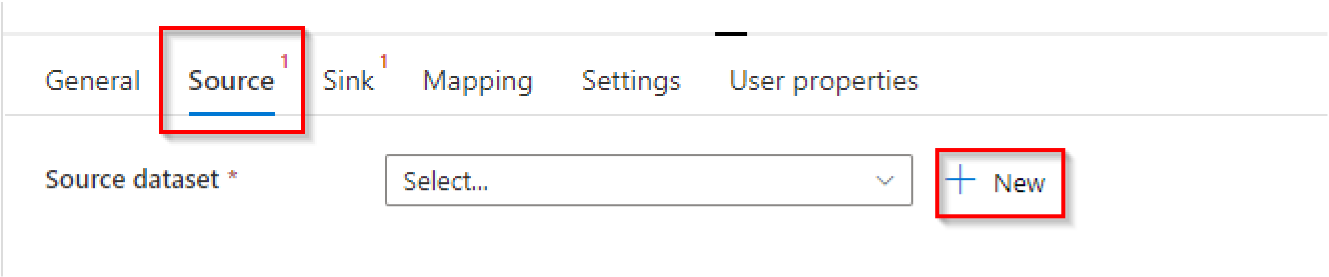
Timeout 7.00:00:00

Retry 0

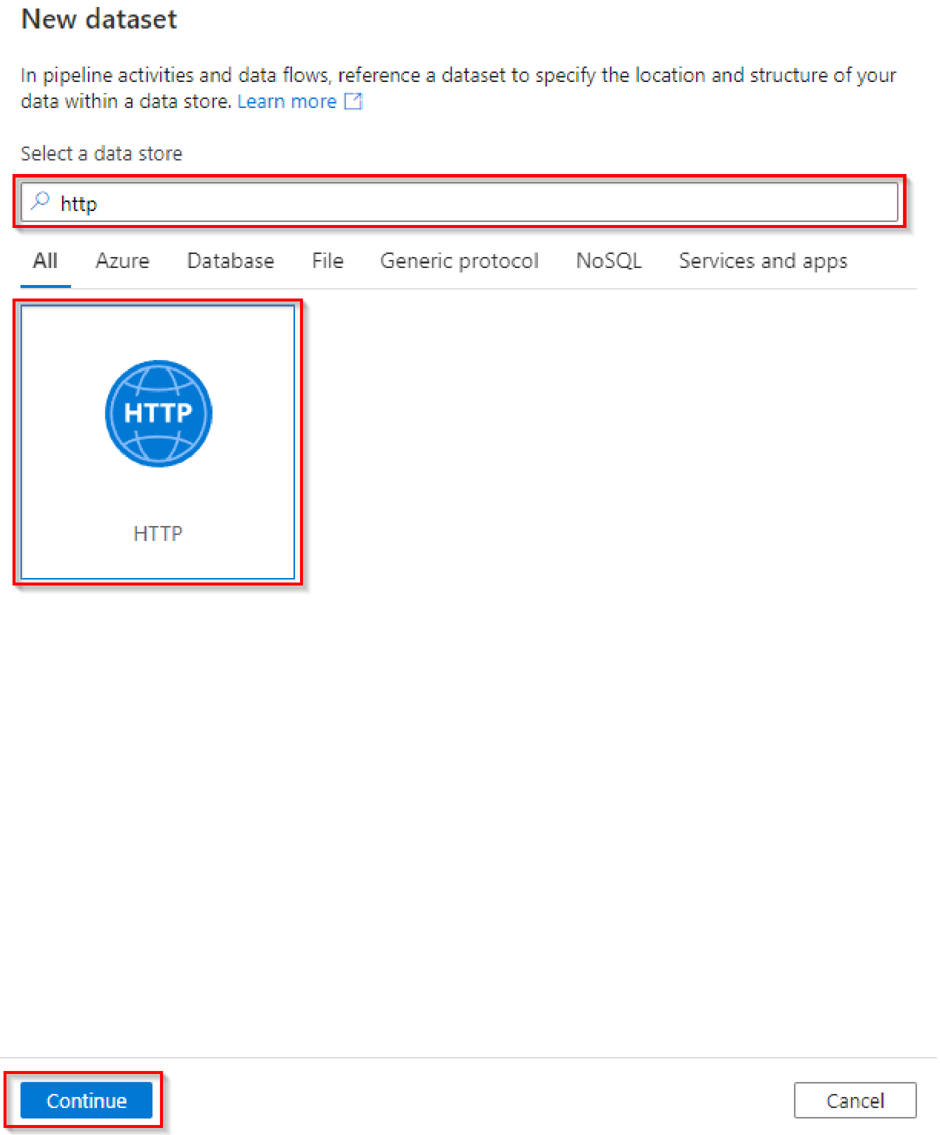
Retry interval (sec) 30

Task 2: Create a new HTTP dataset to use as a source

1. In the Source tab of the Copy activity settings, click **+ New**



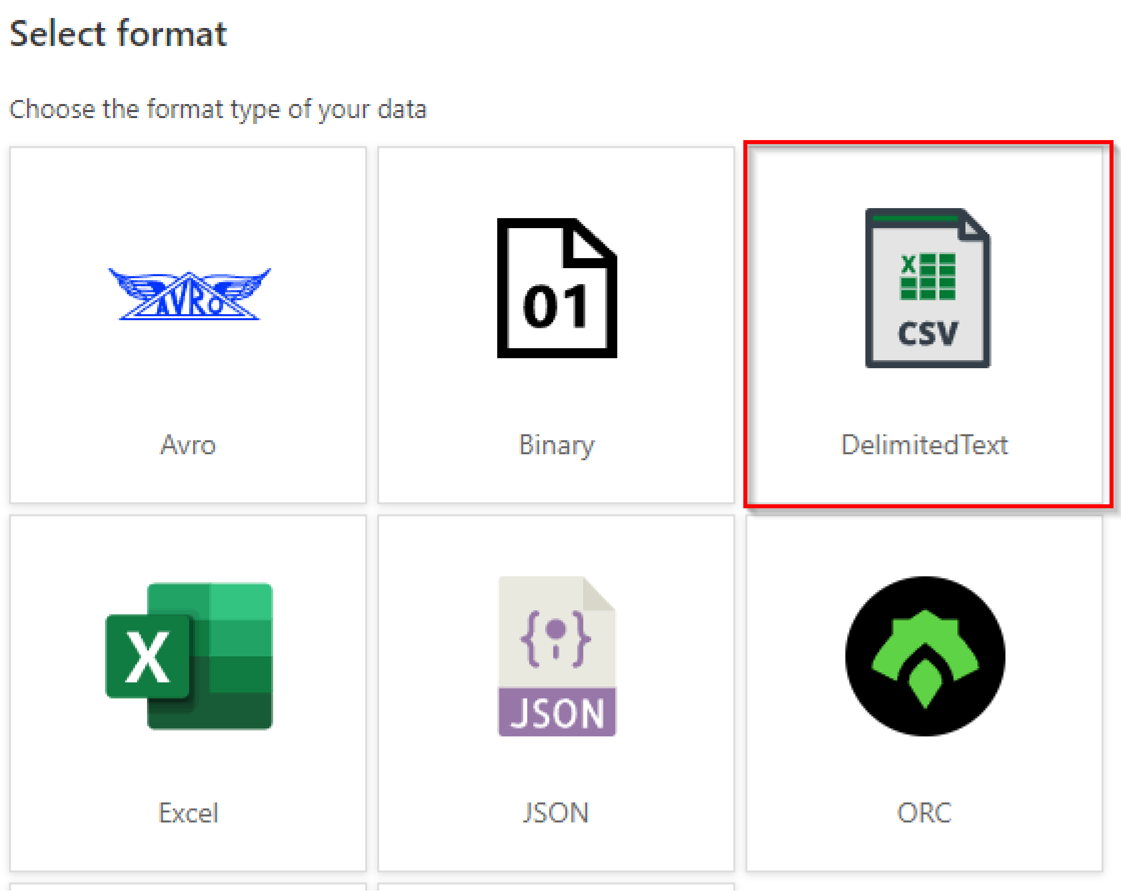
2. In the data store list, select the **HTTP** tile and click **continue**


New dataset

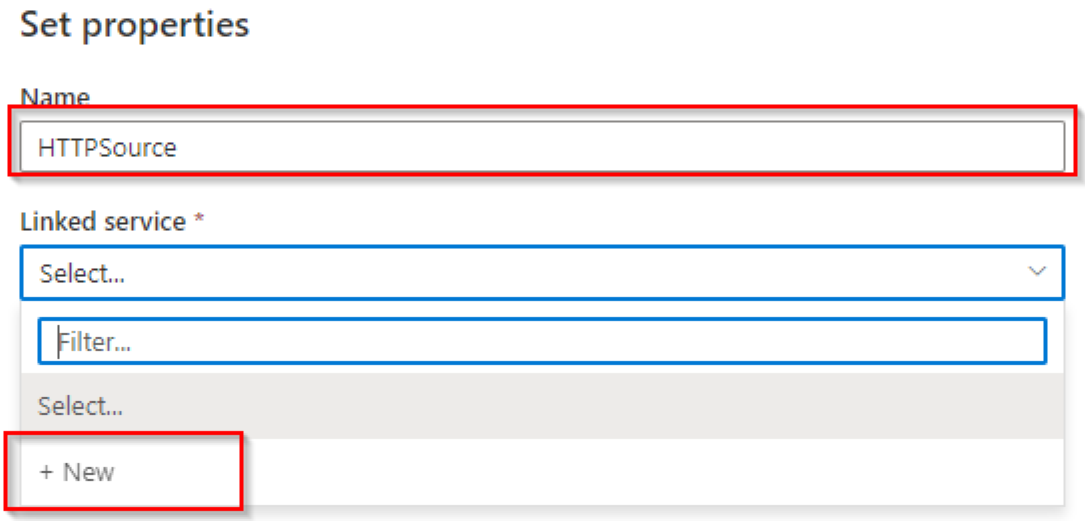
In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more

Select a data store: http

3. In the file format list, select the **DelimitedText** format tile and click **continue**



4. In **Set properties** blade, give your dataset an understandable name such as **HTTPSource** and click on the **Linked Service** dropdown. Create your HTTP Linked Service, select **New**.



Set properties

Name

HTTPSource

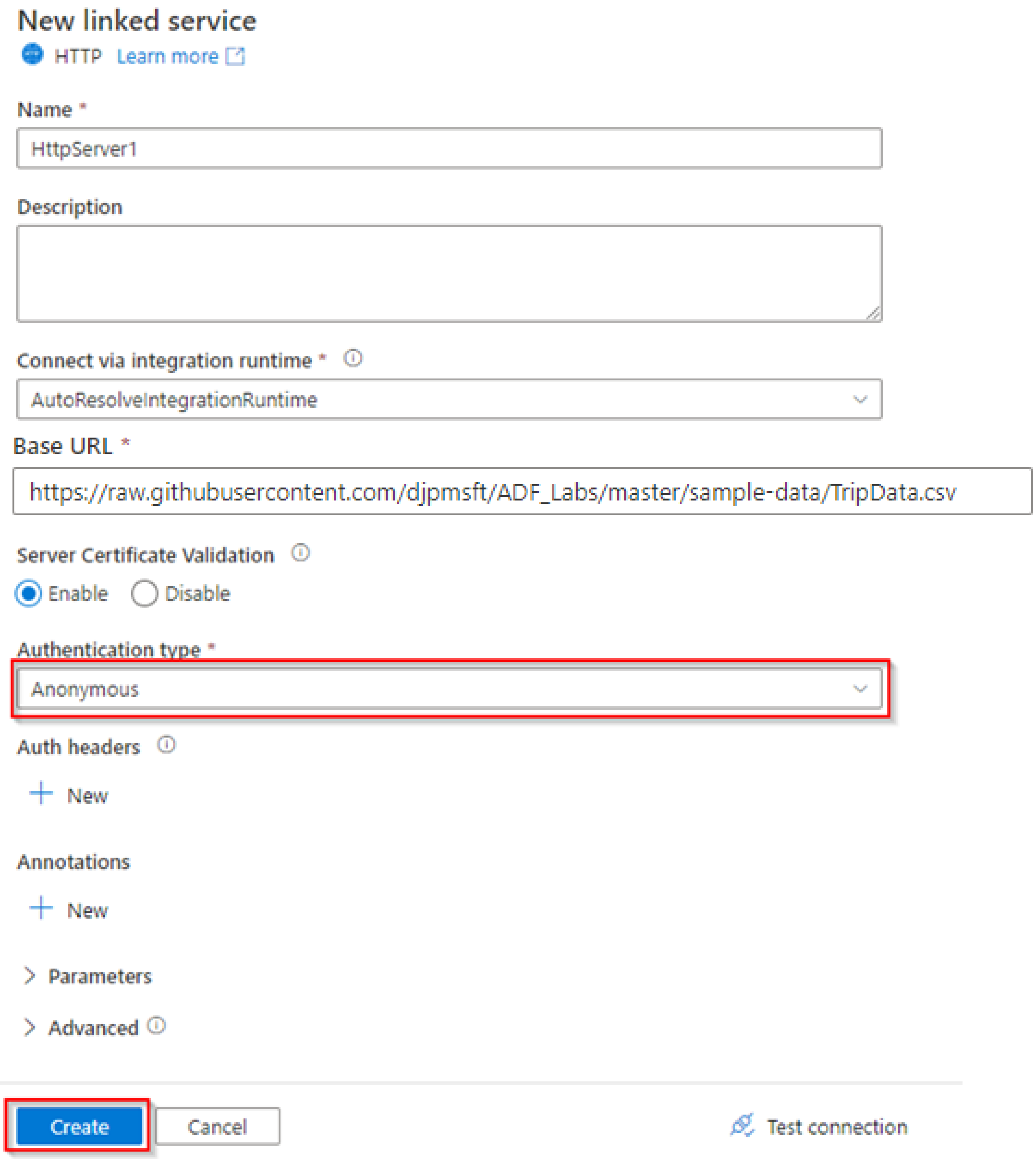
Linked service *

Select...

5. In the **New Linked Service (HTTP)** screen, copy the URL of the TripData csv file below in the **Base URL** textbox. You can access the data with no authentication required using the following endpoint:

   https://raw.githubusercontent.com/djpmsft/ADF_Labs/master/sample-data/TripData.csv

6. In the **Authentication type** drop down, select **Anonymous**. and click on **Create**.



New linked service

HTTP

Name *

HttpServer1

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Base URL *

https://raw.githubusercontent.com/djpmsft/ADF_Labs/master/sample-data/TripData.csv

Server Certificate Validation

● Enable ○ Disable

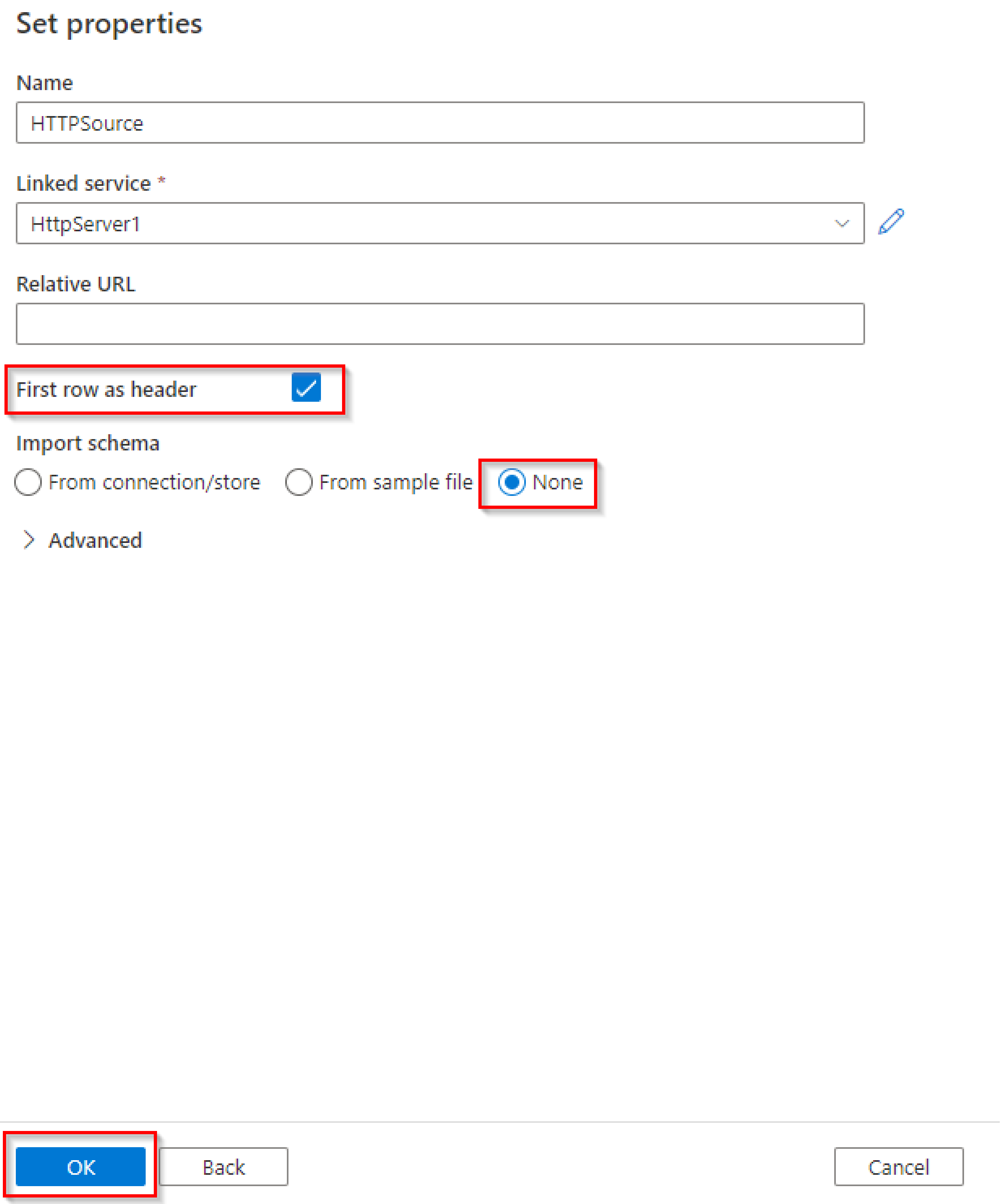
Authentication type *

Anonymous

Auth headers

Annotations

- Once you have created and selected the linked service, specify the rest of your dataset settings. These settings specify how and where in your connection we want to pull the data. As the url is pointed at the file already, no relative endpoint is required. As the data has a header in the first row, set **First row as header** to be true and select **None** in Import schema and click **OK.**



Set properties

Name: HTTPSource

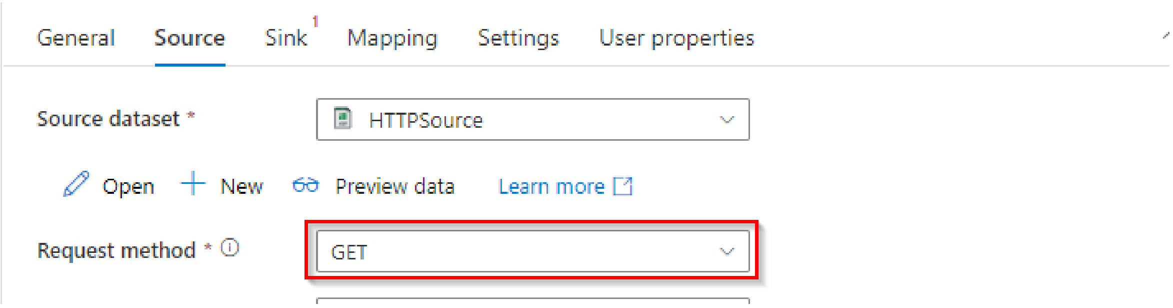
Linked service *: HttpServer1

Relative URL:

First row as header ☑

Import schema: ○ From connection/store ○ From sample file ● None

- Select **Get** as the request method. You will see the following screen

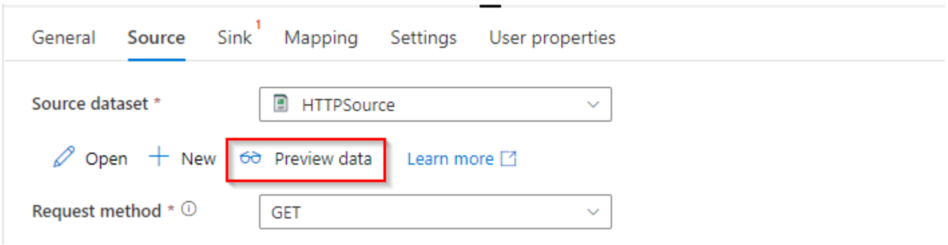


Source dataset * HTTPSource

Request method * GET

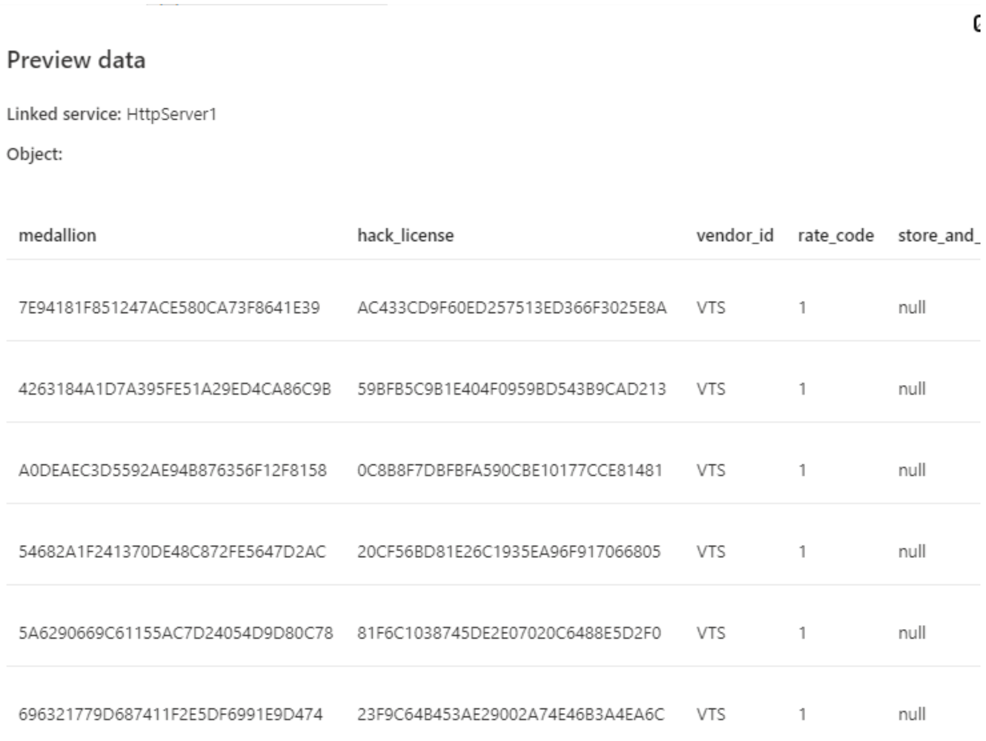
- Click **OK** once completed.

To verify your dataset is configured correctly, click **Preview data** in the Source tab of the copy activity to get a small snapshot of your data.



Source dataset * HTTPSource

Request method * GET



Preview data

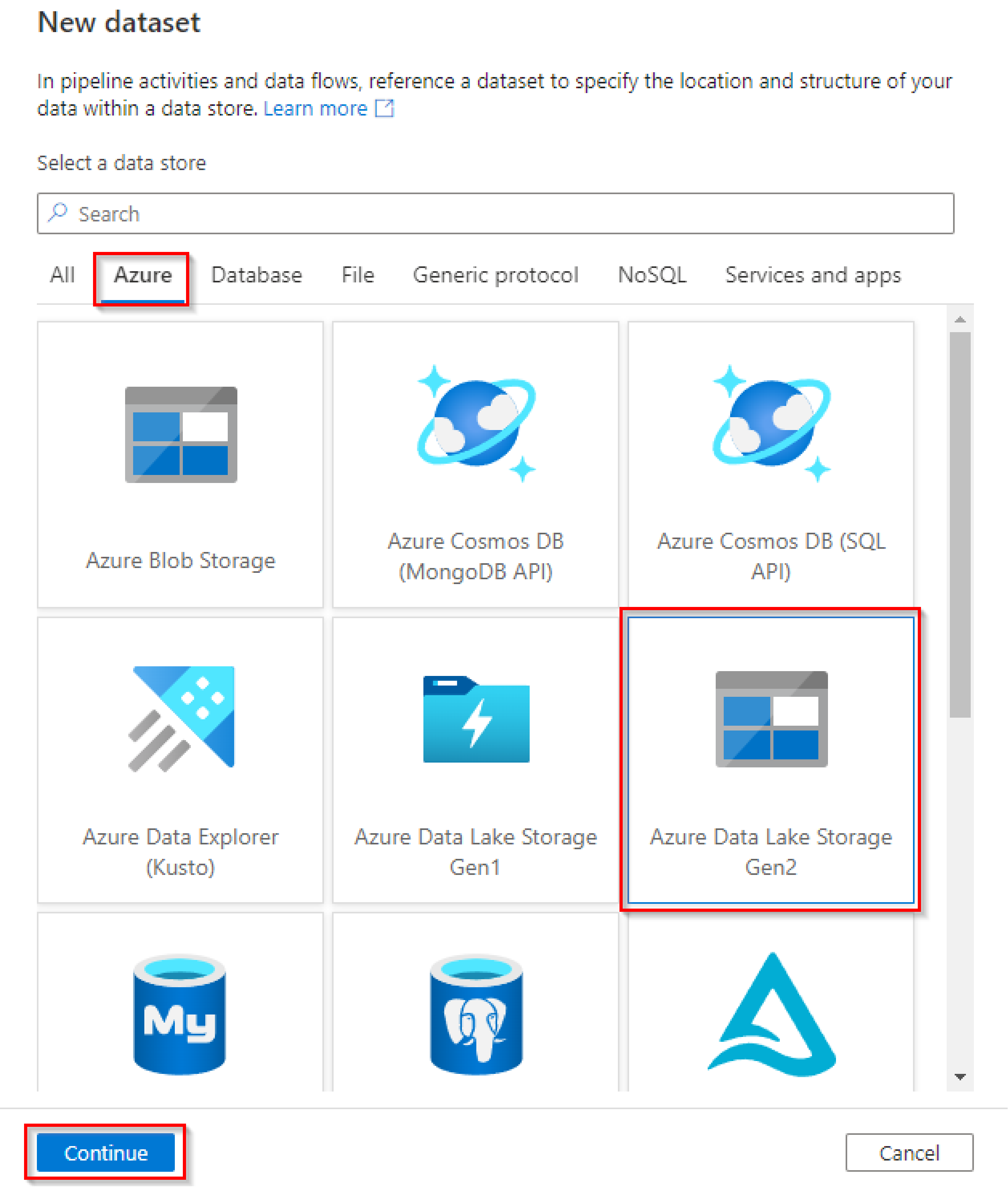
Linked service: HttpServer1

Object:

medallion	hack_license	vendor_id	rate_code	store_and_
7E94181F851247ACE580CA73F8641E39	AC433CD9F60ED257513ED366F3025E8A	VTS	1	null
4263184A1D7A395FE51A29ED4CA86C9B	59BFB5C9B1E404F0959BD543B9CAD213	VTS	1	null
A0DEAEC3D5592AE94B876356F12F8158	0C8B8F7DBFBFA590CBE10177CCE81481	VTS	1	null
54682A1F241370DE48C872FE5647D2AC	20CF56BD81E26C1935EA96F917066805	VTS	1	null
5A6290669C61155AC7D24054D9D80C78	81F6C1038745DE2E07020C6488E5D2F0	VTS	1	null
696321779D687411F2E5DF6991E9D474	23F9C64B453AE29002A74E46B3A4EA6C	VTS	1	null

Task 3: Create a new ADLS Gen2 dataset sink

1. Click on the **Sink tab**, and the click **+ New**
2. Select the **Azure Data Lake Storage Gen2** tile and click **Continue**.
3. Select the **DelimitedText** format tile and click **Continue**.



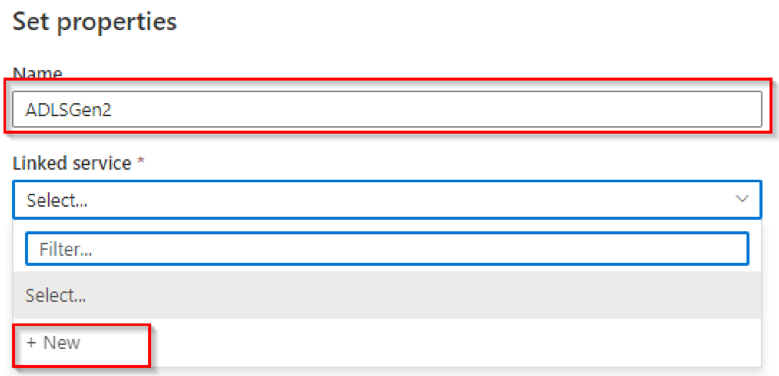
New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. Learn more

Select a data store

Search
All
Azure
Database
File
Generic protocol
NoSQL
Services and apps
Azure Blob Storage
Azure Cosmos DB (MongoDB API)
Azure Cosmos DB (SQL API)
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2
Continue
Cancel

4. In Set Properties blade, give your dataset an understandable name such as **ADLSGen2** and click on the **Linked Service** dropdown. If you have not created your ADLS Linked Service, select **New**.



Set properties

Name

ADLSGen2

Linked service *

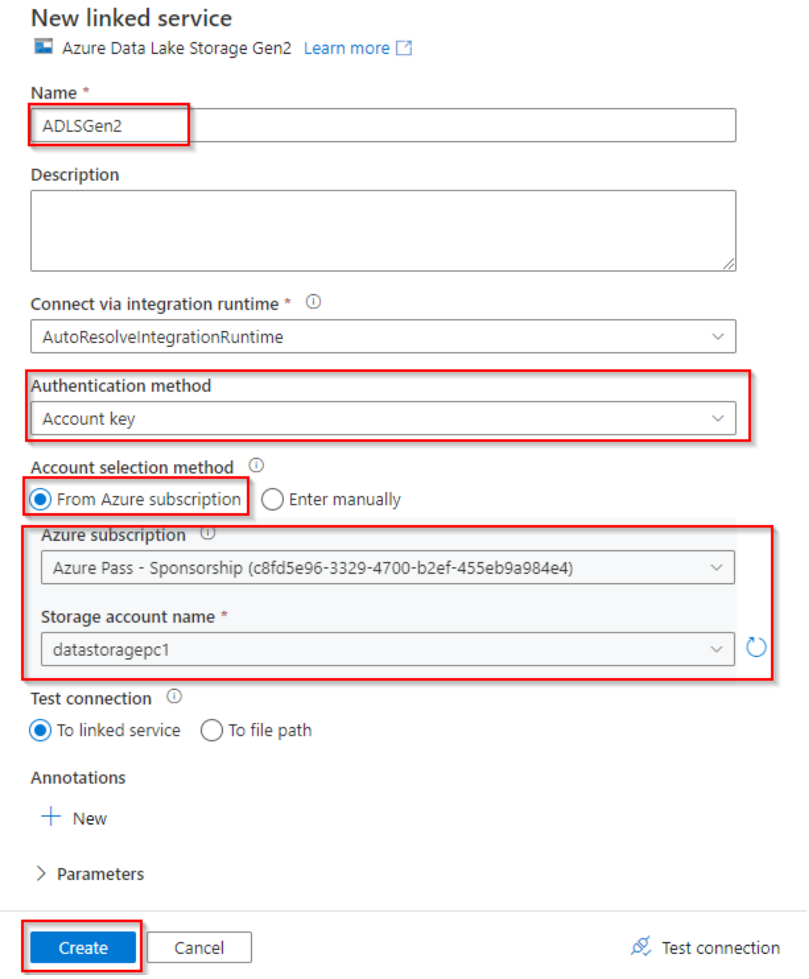
Select...

Filter...

Select...

+ New

5. In the New linked service (Azure Data Lake Storage Gen2) blade, select your authentication method as **Account key**, select your **Azure Subscription** and select your Storage account name of **datastoragexx.**



New linked service

Azure Data Lake Storage Gen2 Learn more

Name *

ADLSGen2

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Authentication method

Account key

Account selection method

● From Azure subscription ○ Enter manually

Azure subscription

Azure Pass - Sponsorship (c8fd5e96-3329-4700-b2ef-455eb9a984e4)

Storage account name *

datastoragepc1

Test connection

● To linked service ○ To file path

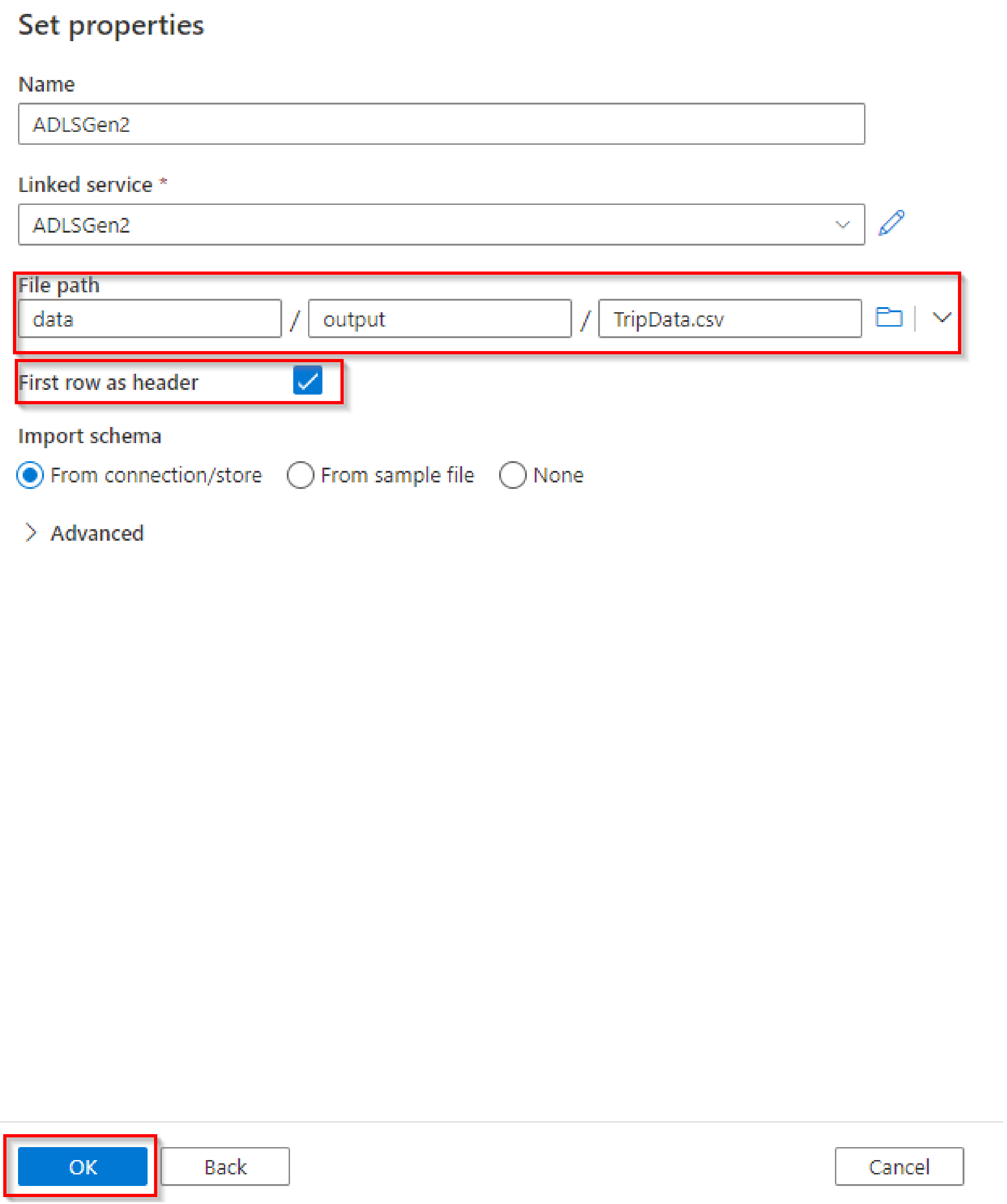
Annotations

+ New

> Parameters

6. Click on **Create**

7. Once you have configured your linked service, you enter the set properties blade. As you are writing to this dataset, you want to point the folder where you want moviesDB.csv copied to. In the example below, I am writing to folder **output** in the **data container**. While the folder can be dynamically created, the file system must exist prior to writing to it. Set **First row as header** to be true and Import schema **None.**



Set properties

Name
ADLSGen2

Linked service *
ADLSGen2

File path
data / output / TripData.csv

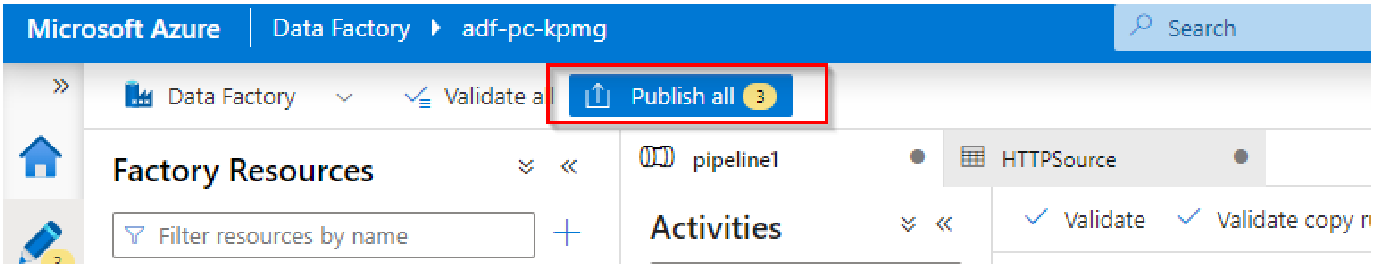
First row as header ☑

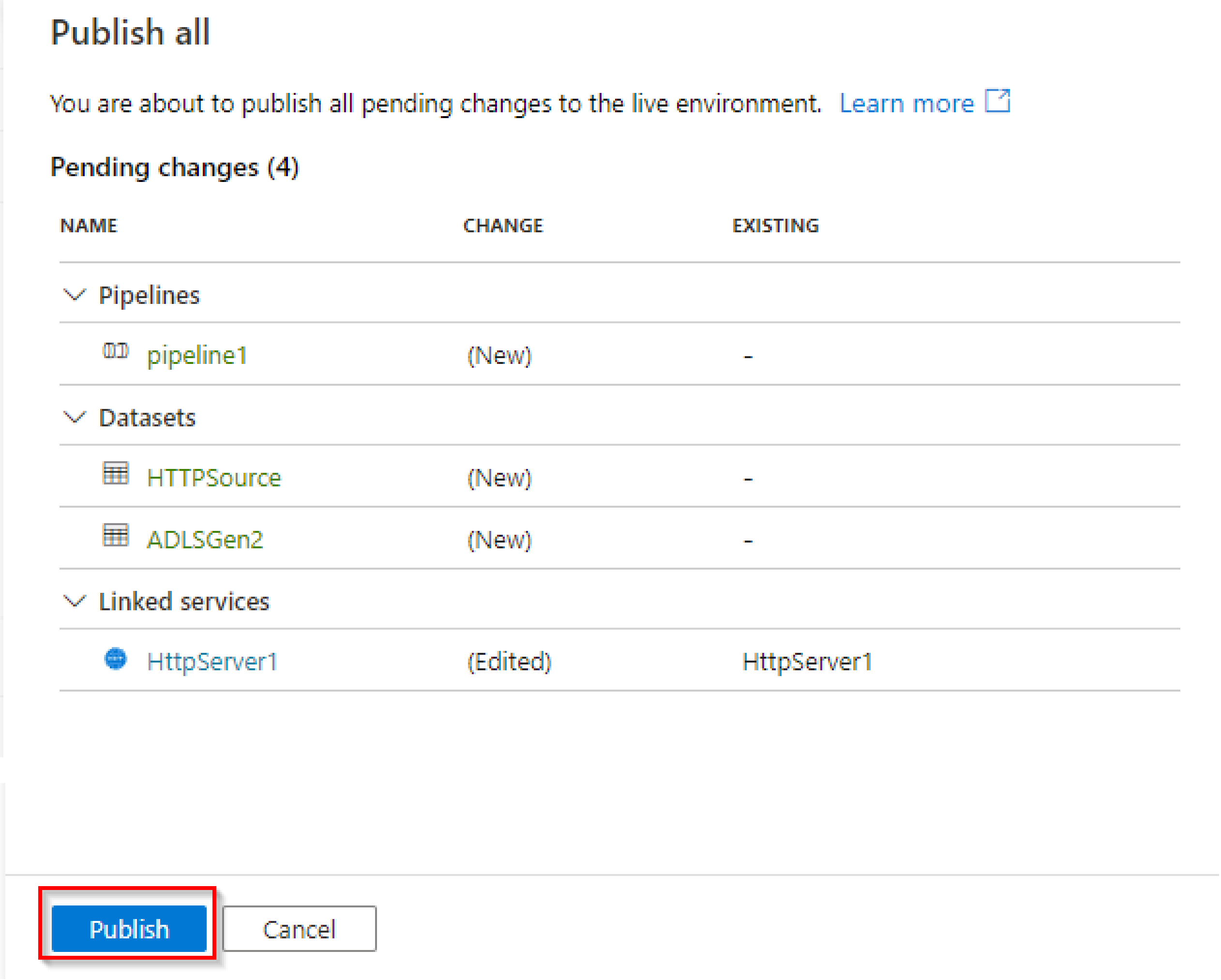
Import schema
● From connection/store ○ From sample file ○ None

8. Click **OK** once completed.

Task 4: Test the Copy Activity

At this point, you have fully configured your copy activity. To test it out, click on the **Publish all** button at the top of the pipeline canvas. This will save the pipeline.





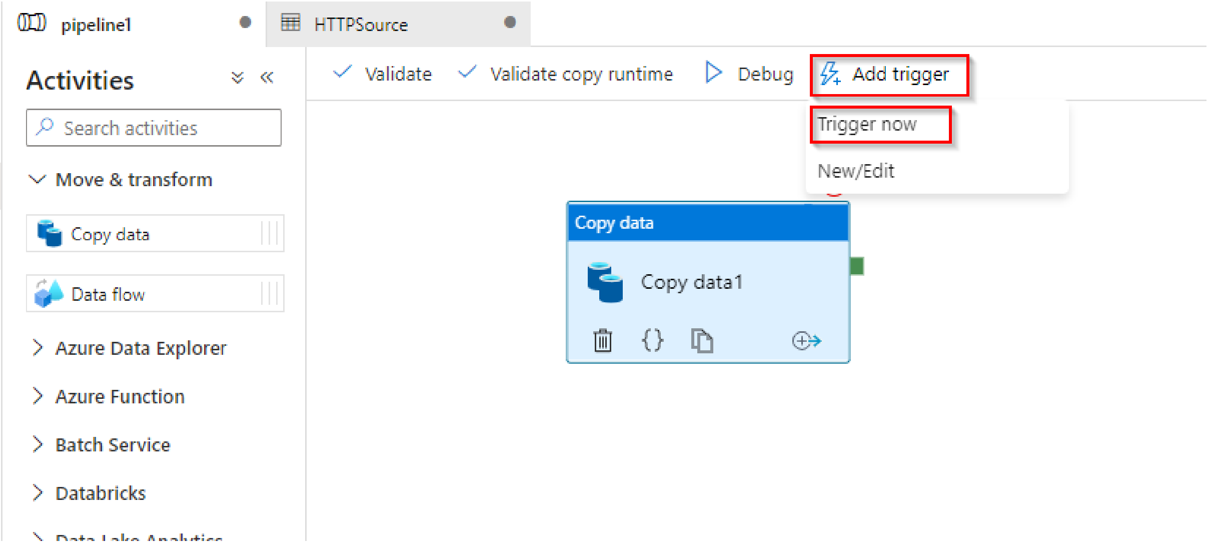
Publish all

You are about to publish all pending changes to the live environment. Learn more

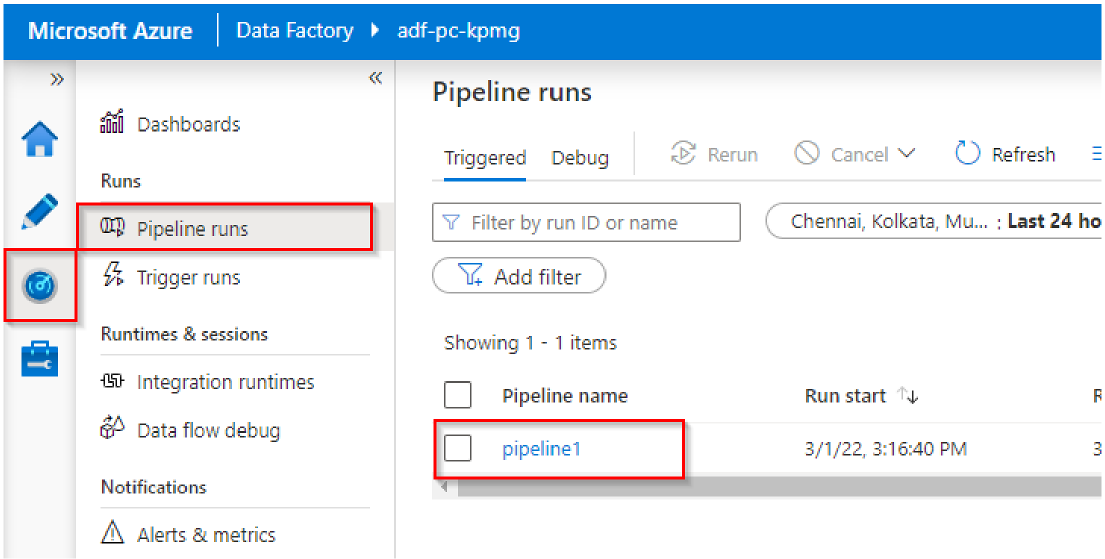
Pending changes (4)

NAME	CHANGE	EXISTING
Pipelines		
pipeline1	(New)	-
Datasets		
HTTPSource	(New)	-
ADLSGen2	(New)	-
Linked services		
HttpServer1	(Edited)	HttpServer1

1. To run the pipeline click on Add trigger->Trigger now->Ok.



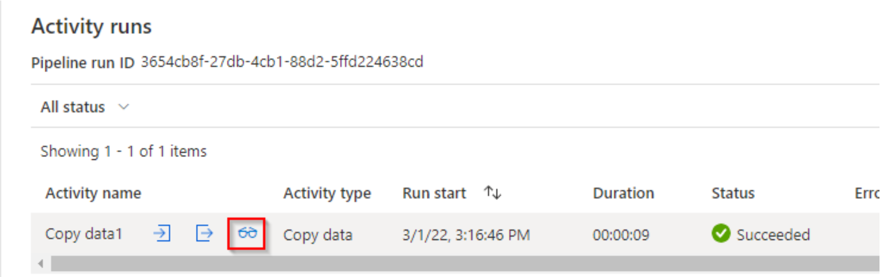
2. To monitor the progress of a pipeline debug run, click on the **Output** tab of the pipeline



Showing 1 - 1 items

Pipeline name	Run start	R
pipeline1	3/1/22, 3:16:40 PM	3

3. To view a more detailed description of the activity output, click on the eyeglasses icon. This will open up the copy monitoring screen which provides useful metrics such as Data read/written, throughput and in-depth duration statistics.

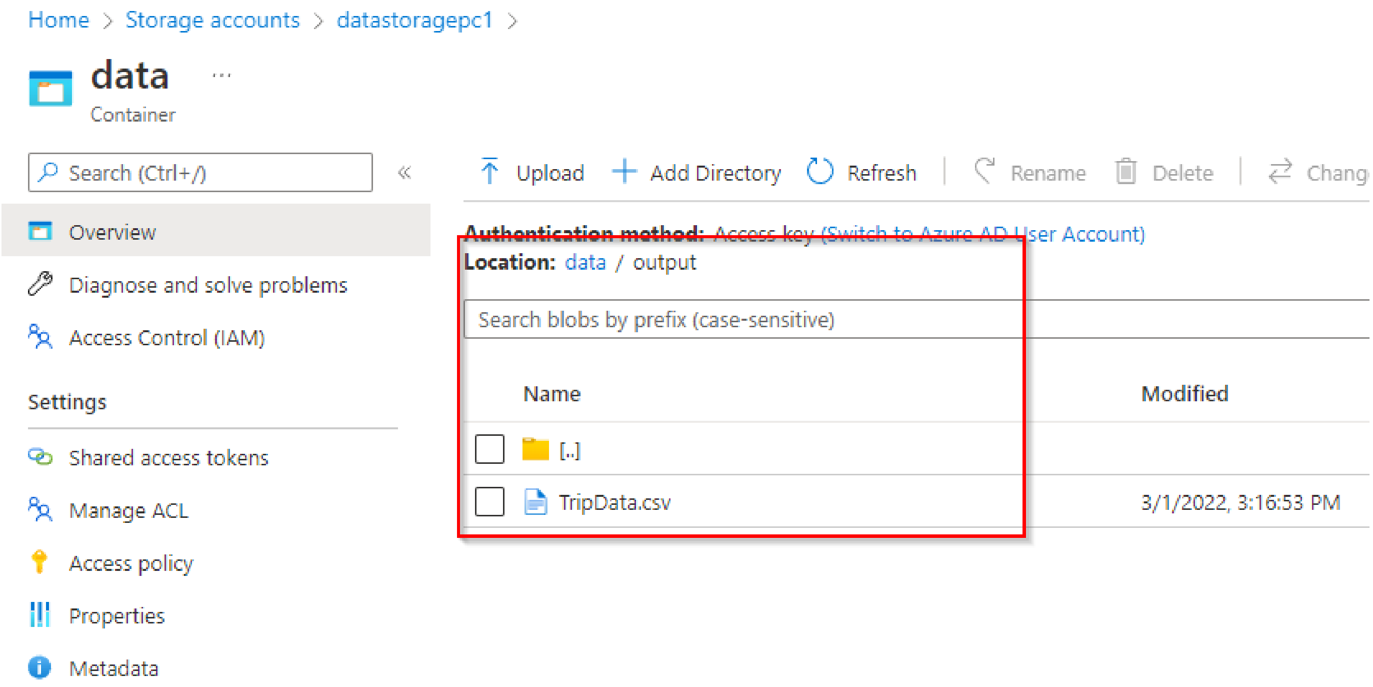


Pipeline run ID 3654cb8f-27db-4cb1-88d2-5ffd224638cd

Showing 1 - 1 of 1 items

Activity name	Activity type	Run start	Duration	Status	Errc
Copy data1	Copy data	3/1/22, 3:16:46 PM	00:00:09	Succeeded	

4. To verify the copy worked as expected, open up your ADLS gen2 storage account and check to see your file was written as expected



Authentication method: Access key (Switch to Azure AD User Account)

Location: data / output

	Name	Modified
□	[..]	
□	TripData.csv	3/1/2022, 3:16:53 PM