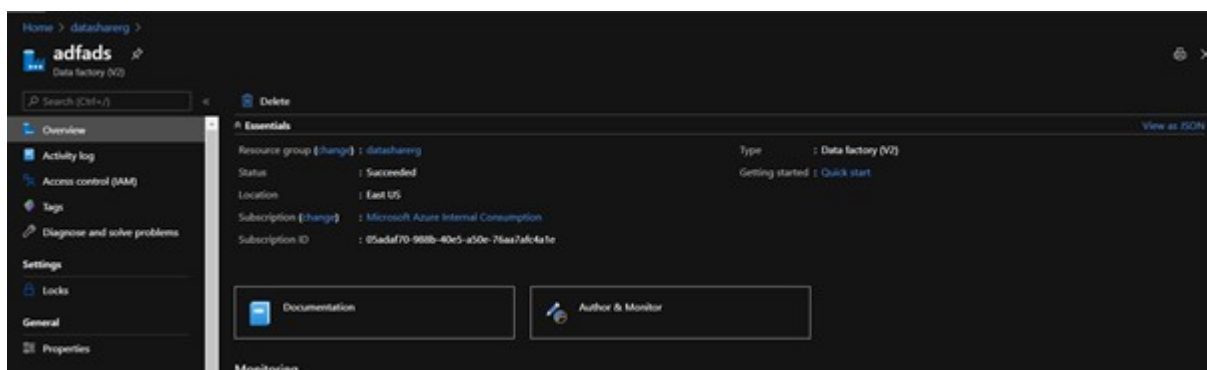


Exercise: Ingest data into Azure Data Lake Storage Gen 2 with Azure Data Factory

Open the Azure Data Factory UX

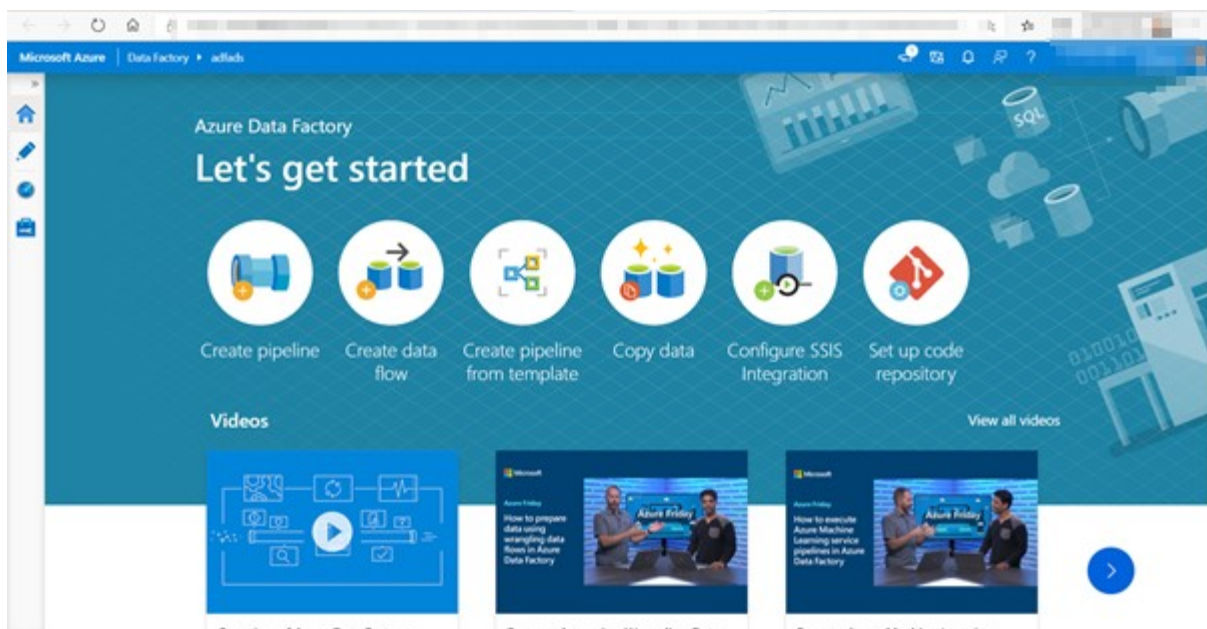
Open the Azure portal through a browser.

Navigate to the resource group in which you have deployed the Azure Data Factory and select, you'll be redirected to the following page.



ure Data Factory Home Page

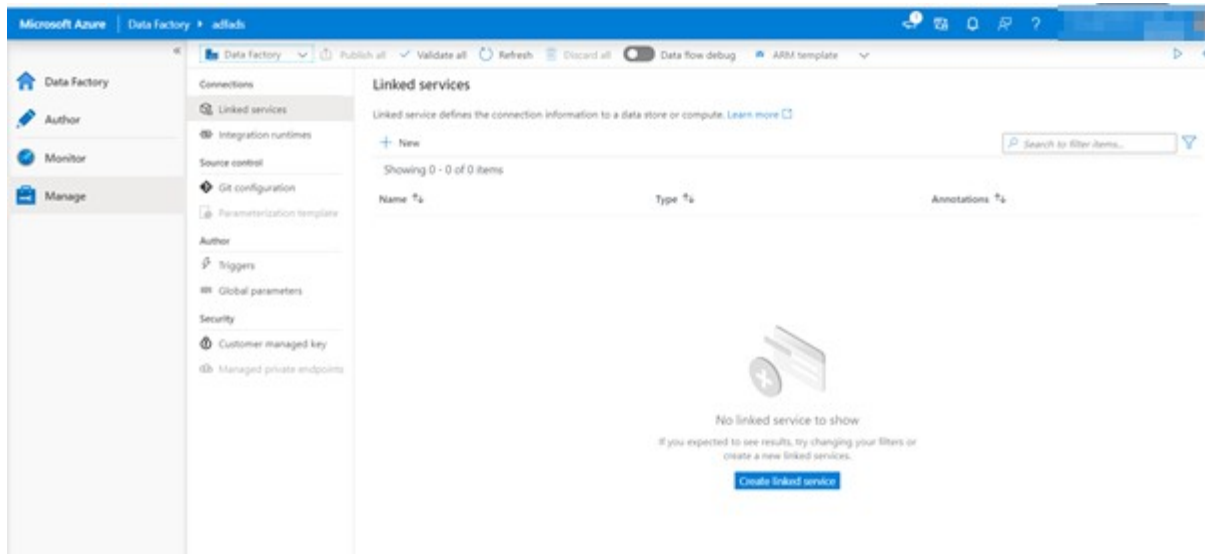
Select 'Author & Monitor' which will redirect you to the following page:



reate Azure Data Lake Storage Gen2 as Linked Service

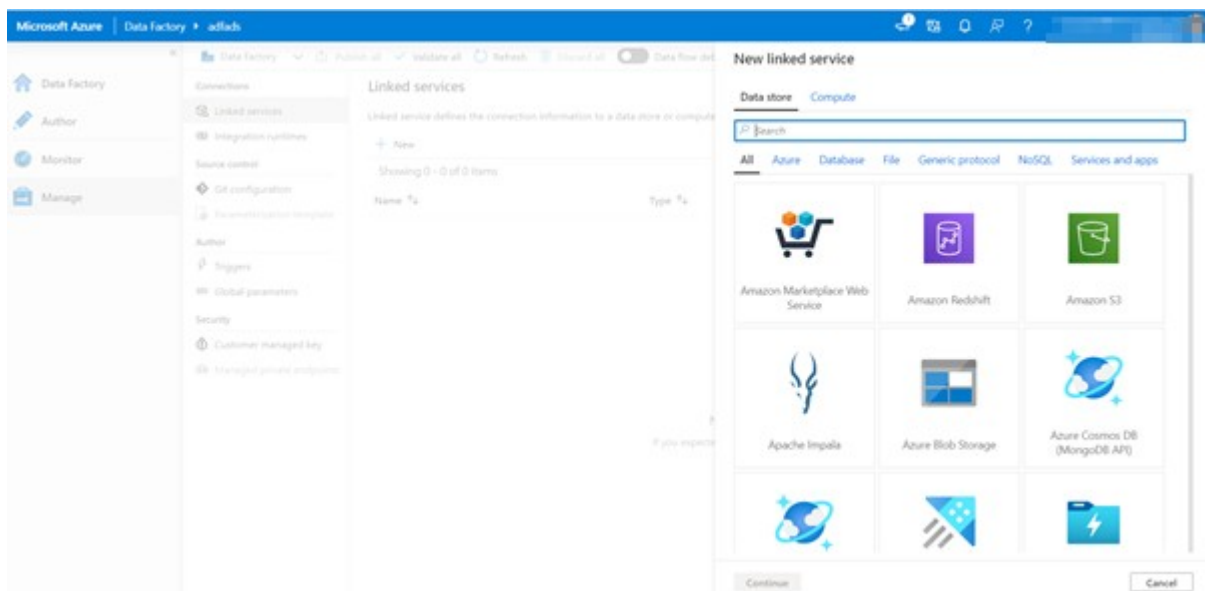
Create an Azure SQL database linked service

The authoring page is where you create data factory resources such as pipelines, datasets, data flows, triggers and linked services. To create a linked service, click on the Manage button



create Azure SQL Database as Linked Service

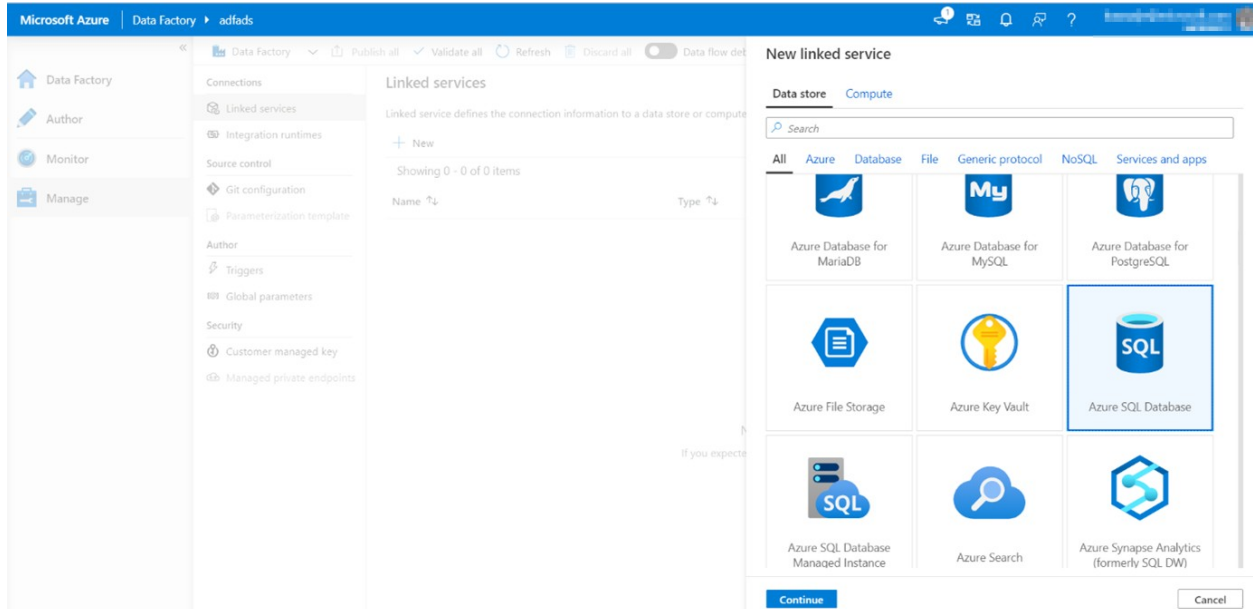
Click New to add a new linked service and you'll be directed to the following page:



create new Linked Service

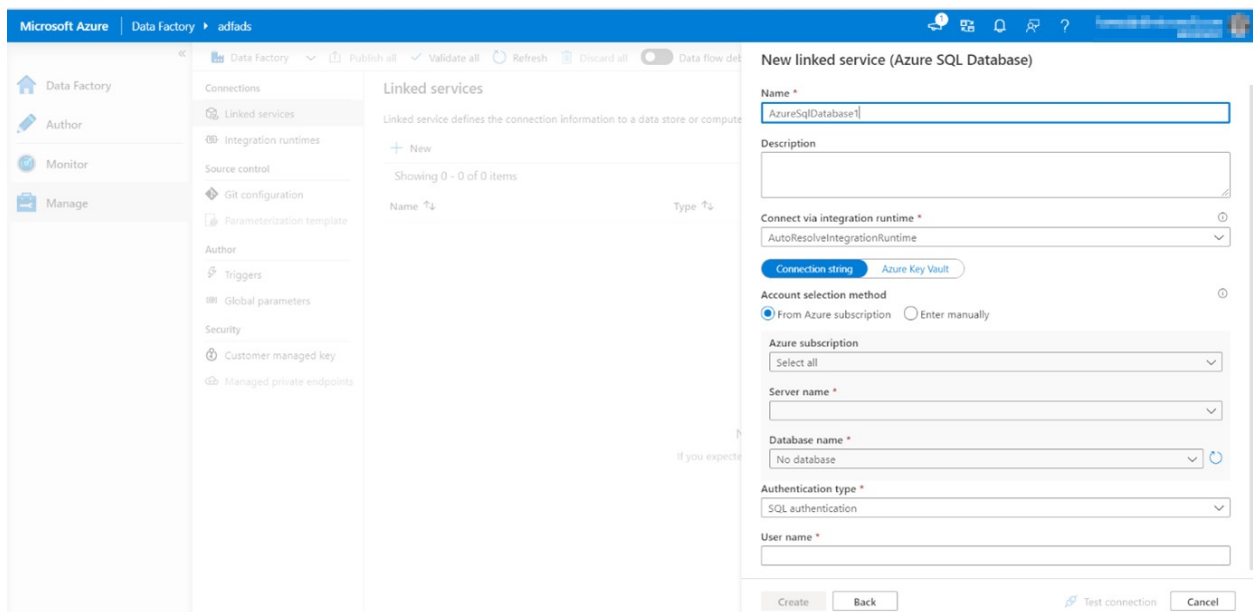
The first linked service you will configure is an Azure SQL DB. You can use the search bar to filter the data store list.

Click on the Azure SQL Database tile and click continue.



Create Azure SQL Database as Linked Service

When you click continue you'll get the following page in which you need to fill out some settings of your SQL Database:



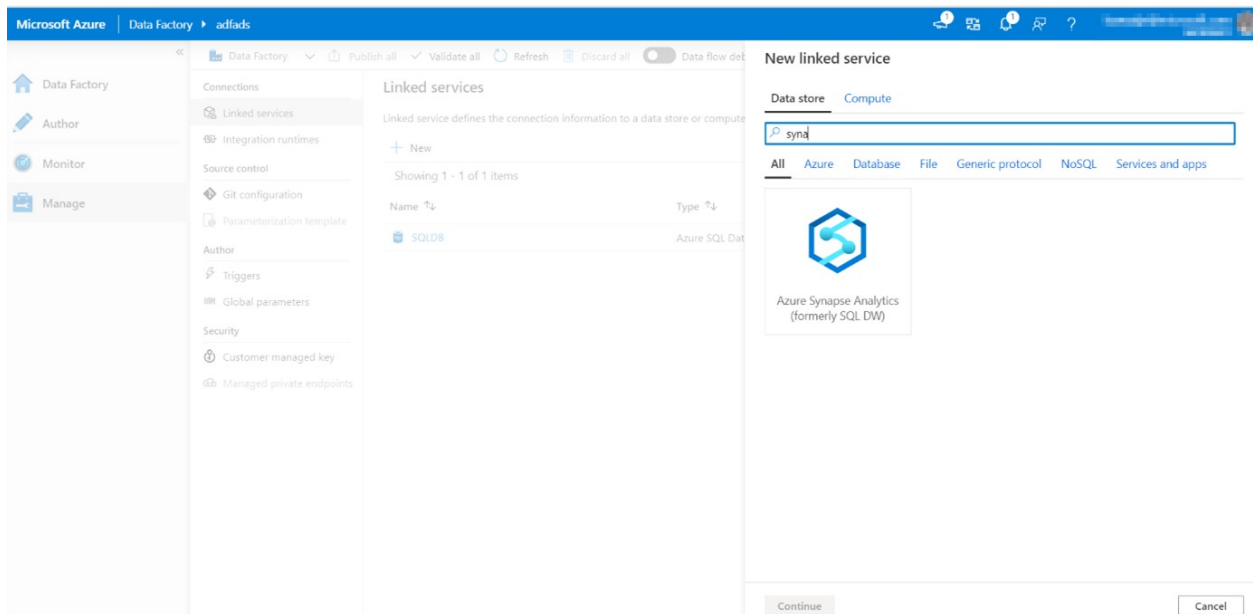
Specification Azure SQL Database as Linked Service

In the SQL DB configuration pane, enter 'SQLDB' as your linked service name. Enter your credentials to allow data factory to connect to your database.

If you're using SQL authentication, enter in the server name, the database, your user name and password. You can verify your connection information is correct by clicking Test connection. Click Create when finished.

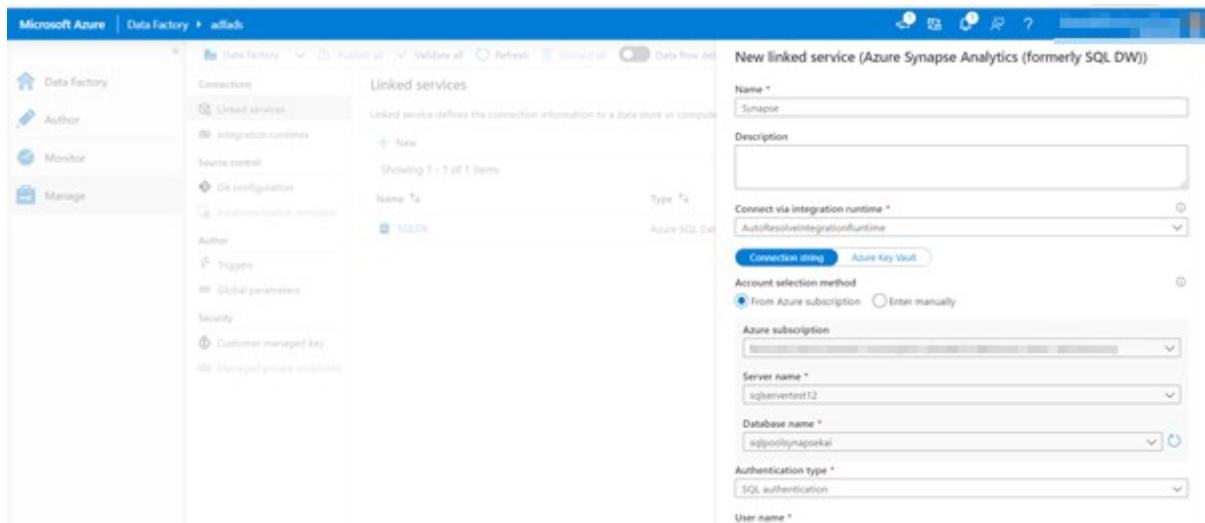
Create an Azure Synapse Analytics linked service

Repeat the same process to add an Azure Synapse Analytics linked service. In the connections tab, click New. Select the Azure Synapse Analytics (formerly SQL DW) tile and click continue.



Create Azure Synapse Analytics as Linked Service

Select Synapse Analytics and you'll be redirected to the following screen:



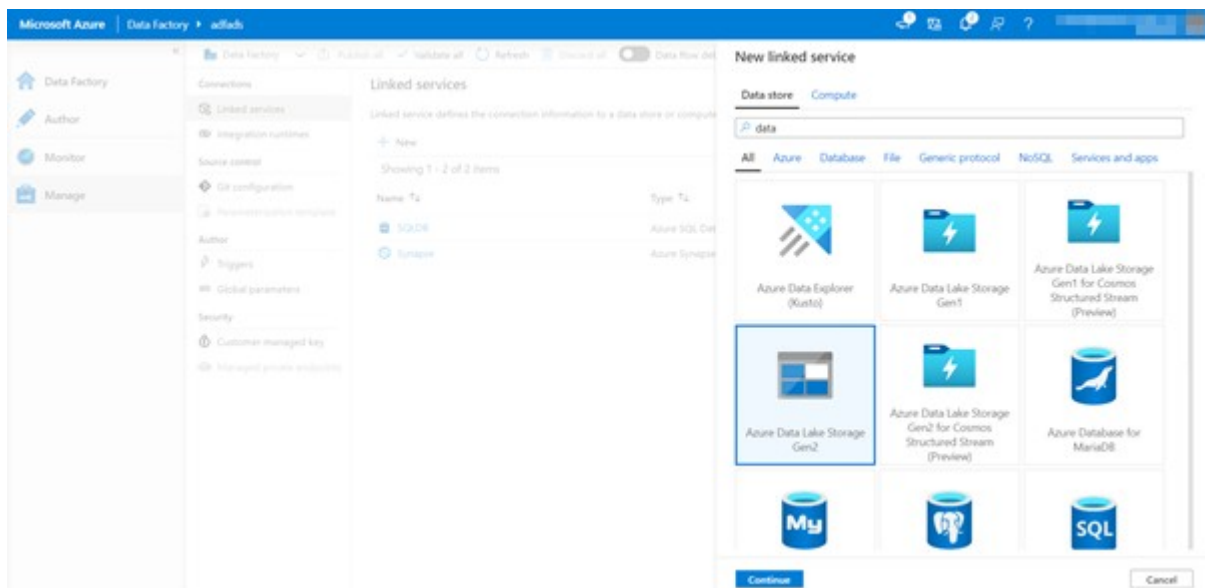
Sp

ification Azure Synapse Analytics as Linked Service

Please fill out the settings and click create. The Linked service connection has then been established for the Synapse Analytics resource.

Create an Azure Data Lake Storage Gen2 linked service

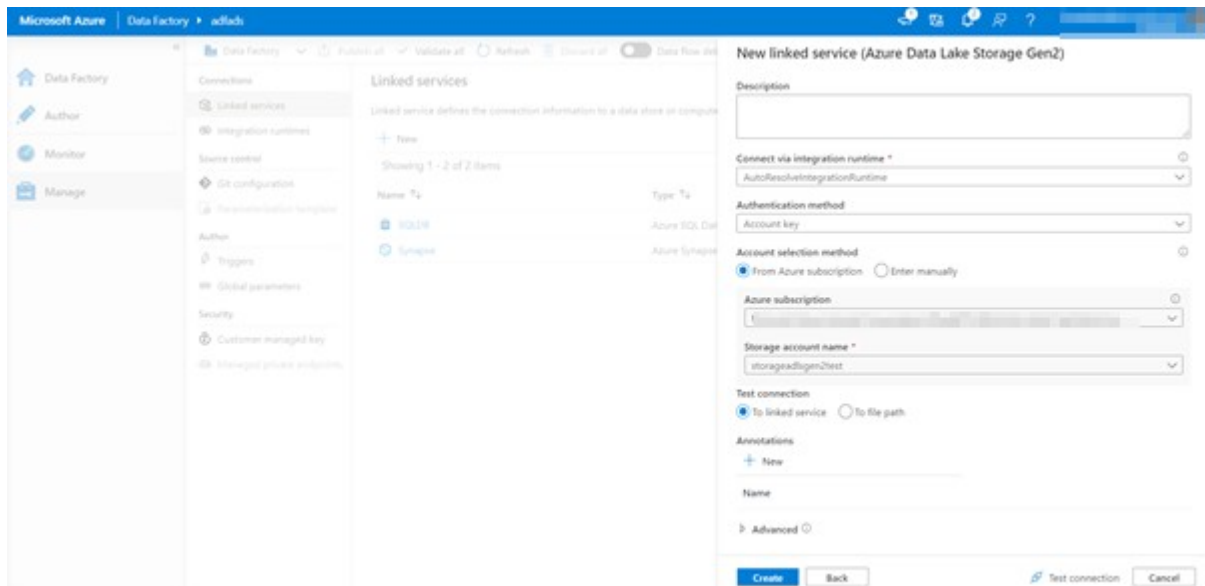
The last linked service needed is an Azure Data Lake Storage gen2. In the connections tab, click New. Select the Azure Data Lake Storage Gen2 tile and click continue.



Cr

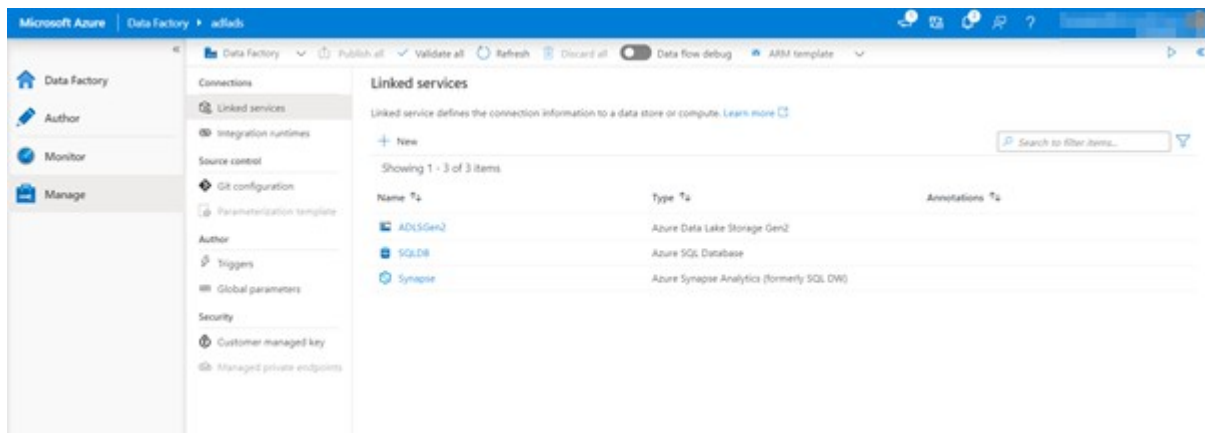
ate Azure Data Lake Storage Gen2 as Linked Service

You'll be redirected to the following screen:



create Azure Data Lake Storage Gen2 as Linked Service

Once you hit create you will be redirected to the following screen:



Linked Services

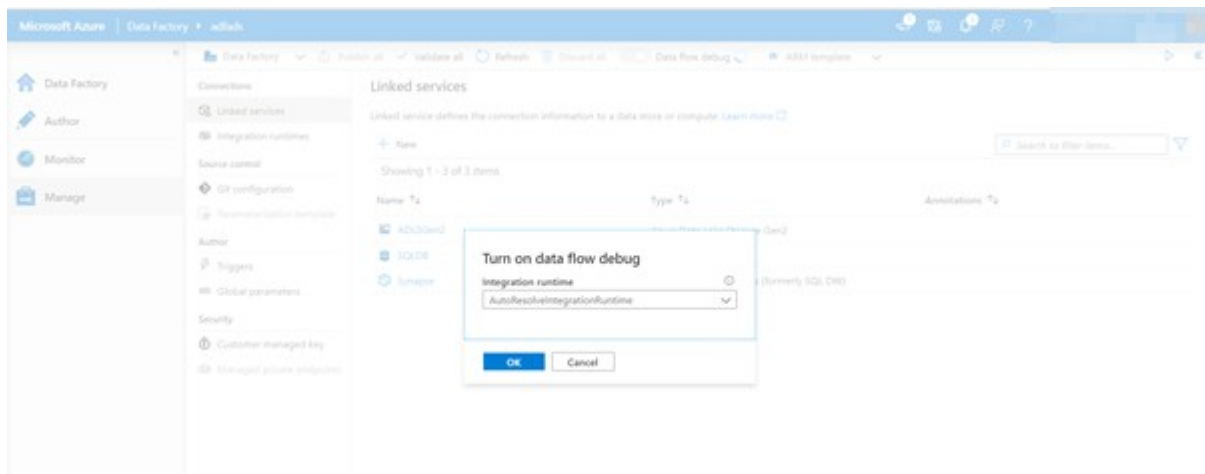
Select Data flow debug to be on.

Turn on data flow debug mode

Now we are building a mapping data flow. A best practice before building mapping data flows is to turn on debug mode which allows you to test transformation logic in seconds on an active spark cluster.

To turn on debug, click the Data flow debug slider in the factory top bar.

Click ok when the confirmation dialog pop-ups. The cluster will take about 5-7 minutes to start-up.



Turn on data flow debug

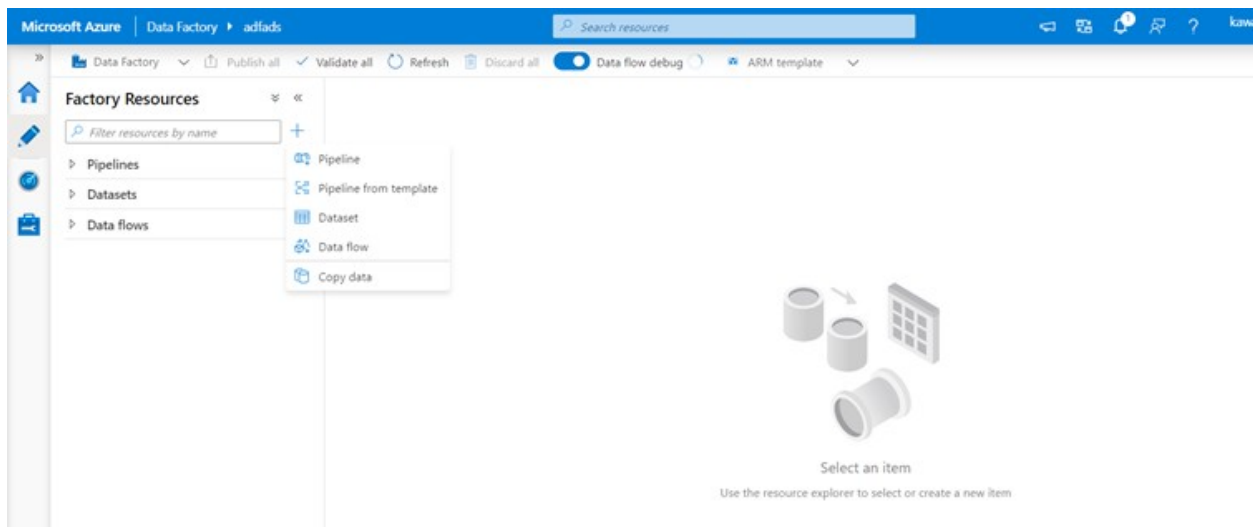
Tur

Ingest data from Azure SQL DB into ADLS gen2 using the copy activity

Now you will create a pipeline with a copy activity that ingests one table from an Azure SQL DB into an ADLS gen2 storage account by adding a pipeline, configure a dataset and debug a pipeline via the ADF UX.

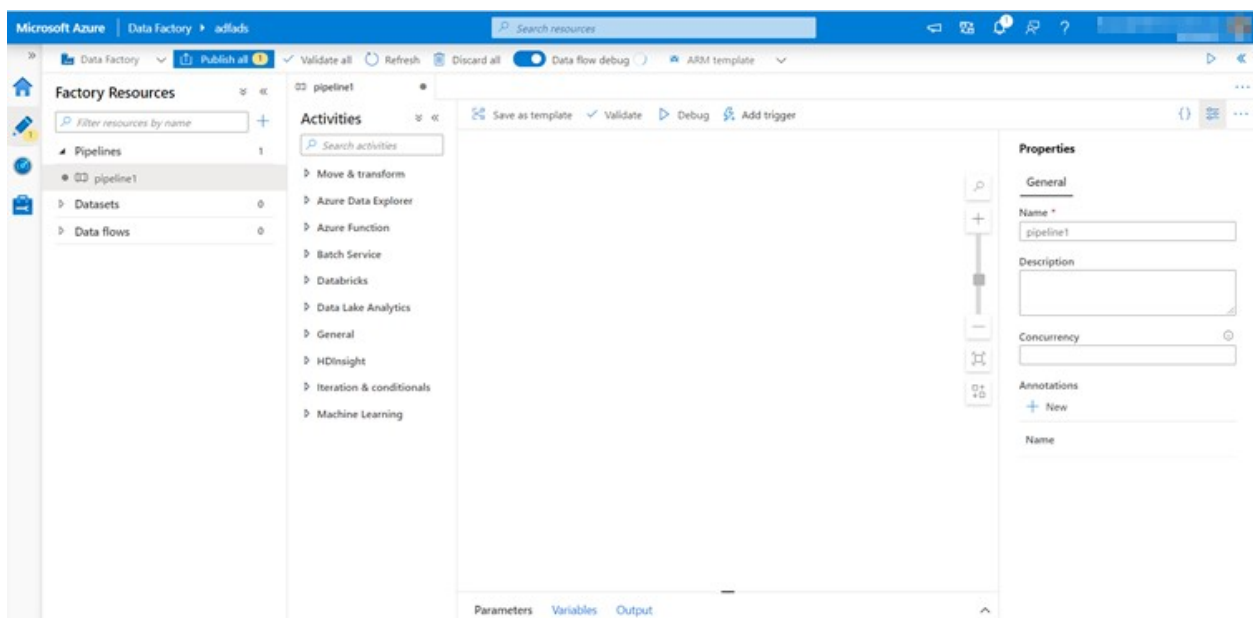
Create a pipeline with a copy activity

In the factory resources pane, click on the plus icon to open the new resource menu. Select Pipeline.



Create a pipeline

You'll be redirected to the following screen:

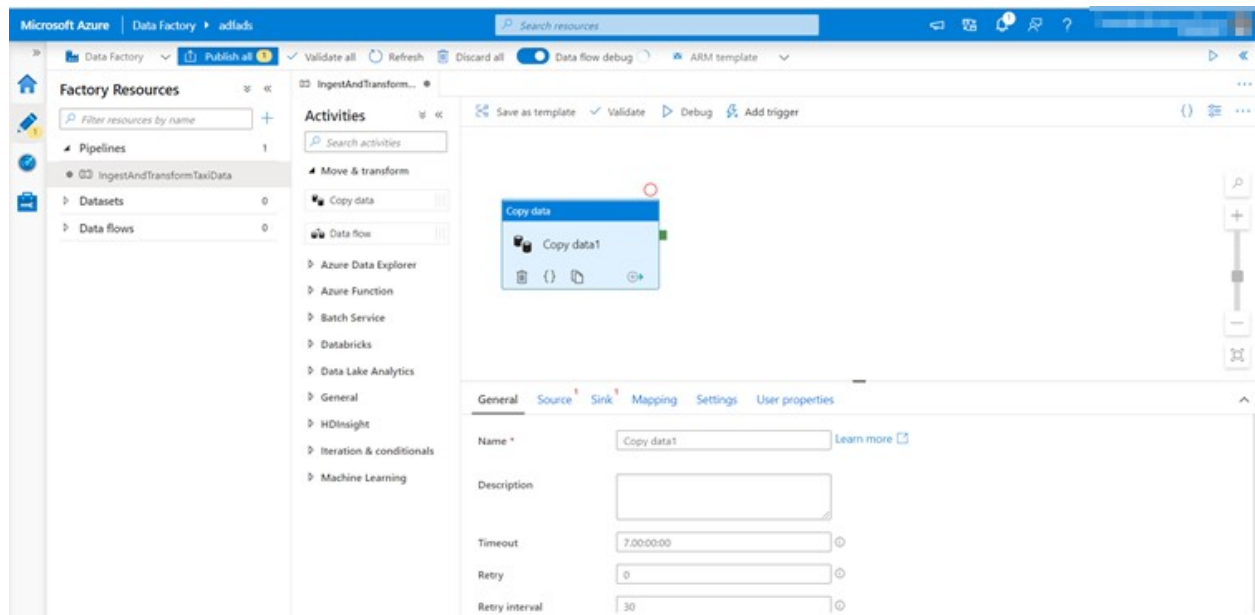


Properties of Pipeline

Give the pipeline a name and save.

In the activities pane of the pipeline canvas, open the Move and Transform accordion and drag the Copy data activity onto the canvas.

Give the copy activity a descriptive name such as 'IngestIntoADLS'.



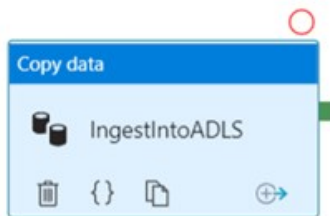
Copy Data Pipeline

Configure Azure SQL DB source dataset

Click on the Source tab of the copy activity. To create a new dataset, click New.

Your source will be the table 'dbo.TripData' located in the linked service 'SQLDB' that we configured in the previous exercise.

 Save as template  Validate  Debug  Add trigger




General **Source¹** Sink¹ Mapping Settings User properties

Source dataset *  New


Source Copy Data Pipeline

Search for Azure SQL Database 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

 sql

All

Azure










Database

File

Generic protocol

NoSQL

Services and apps

| | | |
|--|--|--|
|  Azure Cosmos DB (SQL API) |  Azure Database for MySQL |  Azure Database for PostgreSQL |
|  Azure SQL Database |  Azure SQL Database Managed Instance |  Azure Synapse Analytics (formerly SQL DW) |
|  |  |  |

Continue

Cancel

New SQL Dataset Source

Call your dataset 'TripData'.

Select 'SQLDB' as your linked service.

Select table name 'dbo.TripData' from the table name dropdown.

Import the schema From connection/store.

Click OK when finished.

Set properties

Name

TripData

Linked service *

SQLDB

Table name

dbo.TripData

☐ Edit

Import schema

☒ From connection/store ☐ None

▸ Advanced

OK

Back

Cancel

Set properties SQL Dataset Source

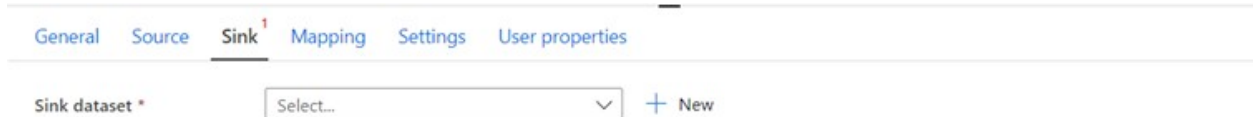
You have successfully created your source dataset.

Make sure in the source settings, the default value Table is selected in the use query field.

Configure ADLS Gen 2 sink dataset

Click on the Sink tab of the copy activity.


To create a new dataset, click New.




Select ADLS Gen2 as Sink Dataset

Please select ADLS Gen 2 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

 data l

All

Azure

Database

File

Generic protocol

Services and apps



Azure Data Lake Storage
Gen1



Azure Data Lake Storage
Gen2

Continue

Cancel







ADLS Gen2 as Sink Dataset

In the select format pane, select DelimitedText as you are writing to a csv file.

Click continue.

Select format

Choose the format type of your data

| | | |
|---|---|--|
|  Avro |  Binary |  DelimitedText |
|  Json |  ORC |  Parquet |

Continue

Back

Cancel

Select format of Sink Dataset

Name your sink dataset 'TripDataCSV'.

Select 'ADLSGen2' as your linked service.

Enter where you want to write your csv file. For example, you can write your data to file trip-data.csv in container staging-container.

Set First row as header to true as you want your output data to have headers.

Since no file exists in the destination yet, set Import schema to None.

Click OK when finished.

Set properties

Name

TripDataCSV

Linked service *

ADLSGen2

File path

staging-container

/

Directory

/

trip-data.csv



First row as header



Import schema



From connection/store



From sample file



None

▸ Advanced

OK

Back

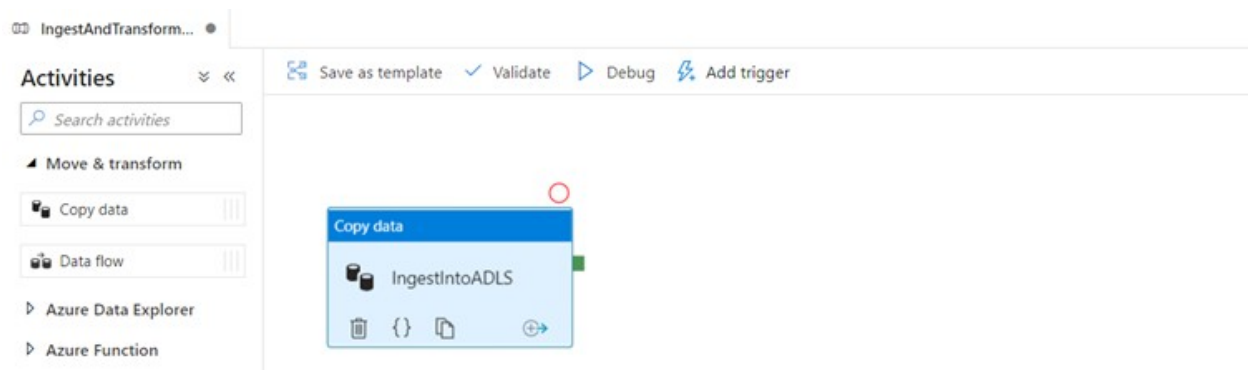
Cancel

Set Properties of Sink Dataset

Test the copy activity with a pipeline debug run

To verify your copy activity is working correctly, click Debug at the top of the pipeline canvas to execute a debug run. A debug run allows you

to test your pipeline either end-to-end or until a breakpoint before publishing it to the data factory service.

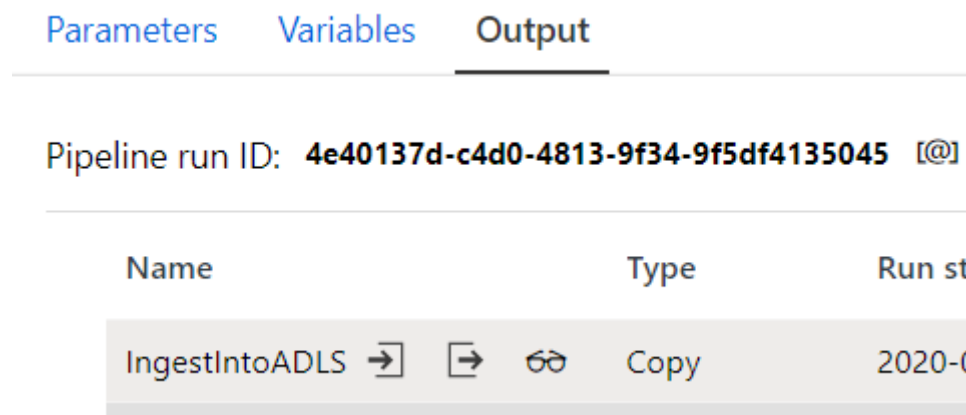


Debug Copy pipeline

To monitor your debug run, go to the Output tab of the pipeline canvas.

The monitoring screen will auto-refresh every 20 seconds or when you manually click the refresh button.

The copy activity has a special monitoring view which can be access by clicking the eye-glasses icon in the Actions column.



pipeline

Output of Debug Copy

If you click on the eyeglass you'll be redirected to the following screen.

The copy monitoring view gives the activity's execution details and performance characteristics. You can see information such as data read/written, rows read/written, files read/written, and throughput.

Details Refresh



[Learn more on copy performance details from here.](#)

Activity run id: e3f41d84-83d7-4089-b17f-a07cb20f2b07



Azure SQL Database
Region: East US

Succeeded



Azure Data Lake Storage Gen2
Region: East US

Data read: 14.393 MB
Rows read: 49,999
Peak connections: 2

Data written: 9.152 MB
Files written: 1
Rows written: 49,999
Peak connections: 1
Throughput: 2.879 MB/s

Copy duration 00:00:05

▲ Azure SQL Database → Azure Data Lake Storage Gen2

Start time 8/31/20, 3:02:37 PM
Used DIUs 4
Used parallel copies 1

▲ Duration 00:00:05

| Details | Working duration | Total duration |
|---------|------------------|----------------|
| Queue | | 00:00:02 |

Time to first byte 00:00:00

Monitoring Specification of Copy pipeline

It's suggested that you publish your changes to the data factory service by clicking Publish all in the factory top bar. Azure Data Factory supports full git integration. Git integration allows for version control, iterative saving in a repository, and collaboration on a data factory. For more information, see source control in Azure Data Factory.

Microsoft Azure | Data Factory > adfads

Search resources

» Data Factory Publish all 3 Validate all Refresh Discard all Data flow debug ARM template

Factory Resources

Filter resources by name

- Pipelines 1
 - IngestAndTransformTaxiData
- Datasets 2
- Data flows 0

Activities

Search activities

- Move & transform
 - Copy data
 - Data flow
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning

Copy data

IngestIntoADLS

Parameters Variables Output


Pipeline run ID: 4e40137d-c4d0-4813-9f34-9f5df4135045

| Name | Type | Run s |
|----------------|------|-------|
| IngestIntoADLS | Copy | 2020- |

Publish all changes of pipeline

If you click Publish All you'll be redirected to the following screen to confirm:

Publish all

You are about to publish all pending changes to the live environment. [Learn more](#) 

Pending changes (3)

| NAME | CHANGE | EXISTING |
|---|--------|----------|
| ▲ Pipelines | | |
|  IngestAndTransformTaxiDa... (New) | | - |
| ▲ Datasets | | |
|  TripData | (New) | - |
|  TripDataCSV | (New) | - |

Publish

Cancel

Confirmation of Publish all changes of pipeline

Select Publish, and the pipeline will be published.