Data Engineering in the Cloud

Azure Data Factory

Xuemao Zhang
East Stroudsburg University

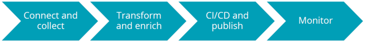
January 18, 2025

Outline

- Overview of Azure Data Factory
- Data Integration
- Lab

Overview of Azure Data Factory

- Once the data is loaded, it is important to perform serverless transformation and integration of the data.
- For data transformation, you can take the help of Azure Data Factory to create data transformation workflows.
- Azure Data Factory is an extract load transform (ETL) service that creates data driven workflows on cloud.
- The following is the working of Data Factory by a series of interconnected system:



Overview of Azure Data Factory

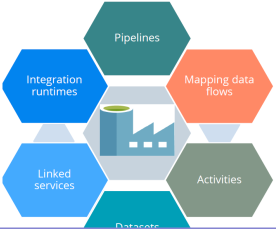
Connect and Collect

- Connect Data Sources: All important data and processing sources, such as SaaS providers, databases, and other systems, are connected to build an information production system.
  - For example, Connect to relational databases.
- Centralize Data: Move the data to a central location for further processing.
  - Data Lake: Use Azure Data Lake Storage to store raw data in its native format.
- Data Movement with Data Factory: Use Azure Data Factory to create a copy activity in a pipeline.
  - Create a pipeline in Azure Data Factory to copy data from the source to the central location.
    - The source and central location in an Azure Data Factory (ADF) pipeline can both be containers in Azure Data Lake Storage.
- Data Analysis: After centralizing the data, perform further analysis as needed.
  - For example, use Azure Data Factory to transform data as needed (e.g., cleaning, aggregating).

Overview of Azure Data Factory

Transform and Enrich: Azure Data Factory (ADF) offers robust capabilities for data transformation using mapping data flows, as well as integration with services like Apache Spark and Azure Machine Learning

- Azure Data Factory Mapping Data Flows: Mapping data flows in ADF allow you to design visually-driven transformations.
- We can use the drag-and-drop interface to build complex transformation logic.
- Services such as Spark and Machine Learning can transform the data with the help of the Azure Data Factory.
  - Azure Synapse Analytics integrates with Apache Spark, enabling large-scale data processing and transformation.
  - Azure Data Factory can orchestrate data processing and machine learning workflows by integrating with Azure Machine Learning.



What is Data Integration?

- Data integration involves combining data from different sources to provide a unified view, making it easier to manage and analyze.
- Data integration can be performed with Azure Data Factory, which enables for the development of data driven processes for data transformation at scale.
- It may also be used to build and organize data driven processes that allow for the consumption of data from multiple sources.
  - ADF supports integration with various data sources such as databases, file systems, APIs, and cloud services.
- In this process, complex ETL processes to modify data visually may also be created with the use of data flows or computing services like Azure Databricks, Azure Synapse, and so on.

Transformation Data Using Mapping Data Flow

- Mapping Data Flows provides a visual platform for creating a variety of data transformations without needing to use code.
- The data flows are built and then executed on scaled out Apache Spark clusters that are deployed automatically when the Mapping Data Flow is executed.
- It also enables the monitoring of the transformation's execution to observe their progress and comprehend any difficulties that may develop.

Transformation Data Using Compute Resources

- Azure Data Factory can transform data by using the computation resources and a platform service that is better suited for the task.
- For example, Azure Data Factory may build a pipeline to an analytical data platform, such as Spark pools on an Azure Synapse Analytics instance, to perform a complex python computation.
- Another example is sending data to an Azure SQL Database instance in order to run a Transact SQL stored procedure.

Types of Azure Data Factory Transformation

A range of transformation types are available in Mapping Data Flows to allow data to be changed. They are divided into the following categories:

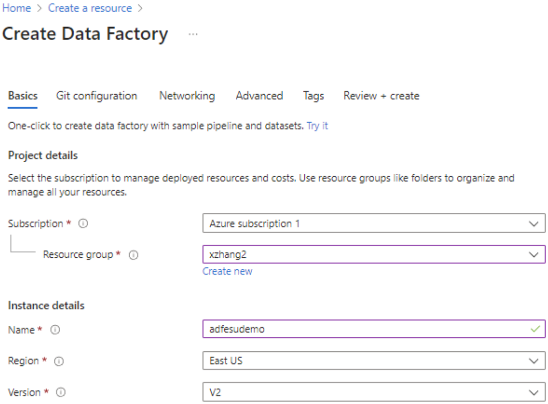
- Schema Modifier Transformations allow you to change the structure of your data by adding, removing, or modifying columns.
- Row Modifier Transformations are used to modify the data within each row. These transformations enable you to clean, aggregate, and manipulate your data on a row-by-row basis, like sort, filter, aggregate
- Multiple Inputs/Outputs Transformations allow you to handle data from multiple sources or direct data to multiple destinations. These transformations are useful for complex data processing scenarios involving multiple datasets.

Prerequisites

- Prerequisites (5 minutes):
  - Create a datalake (storage account with hierarchical namespace enabled)
    - To lower the cost, you may choose `Redundancy` as `LRS`
  - Create a container in the storage: `raw`
    - Upload the data set `CoursesData.csv` to the container
  - Create an Azure Synapse Workspace
    - Create a dedicated SQL pool

Lab

- Click on `create a resource` and search for a `Data Factory`
- Go to `Data Factory` and click on `create`
- Set the `Basics` as shown in the below image



One-click to create data factory with sample pipeline and datasets. Try it

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * : Azure subscription 1

Resource group * : xzhang2

Create new

Instance details

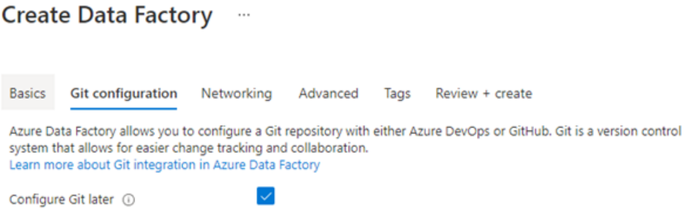
Name * : adfesudemo

Region * : East US

Version * : V2

Lab

- Click on "Next: Git configuration" and "Enable configure Git later"



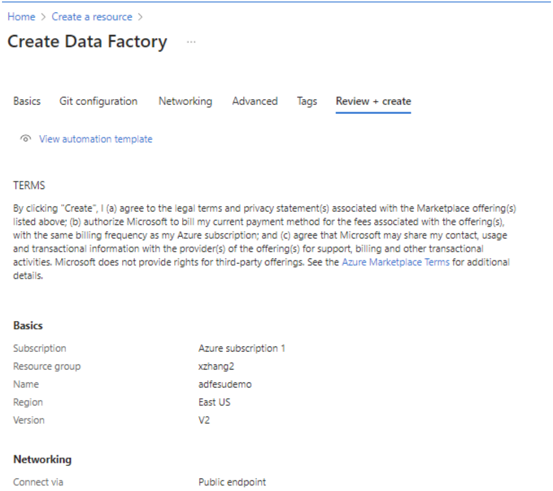
Create Data Factory ···

Azure Data Factory allows you to configure a Git repository with either Azure DevOps or GitHub. Git is a version control system that allows for easier change tracking and collaboration.
Learn more about Git integration in Azure Data Factory

Configure Git later ⓘ ☑

Lab

- Now leave the other configurations as it is and go to `review + create`
- Once the configuration is validated, click on the "create" button



Home > Create a resource >

Create Data Factory ...

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the Azure Marketplace Terms for additional details.

Basics

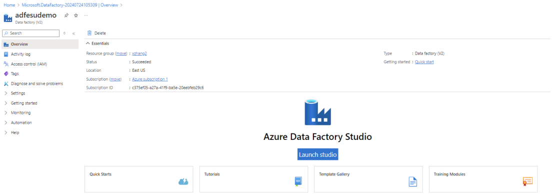
Subscription	Azure subscription 1
Resource group	xzhang2
Name	adfesudemo
Region	East US
Version	V2

Networking

Connect via	Public endpoint

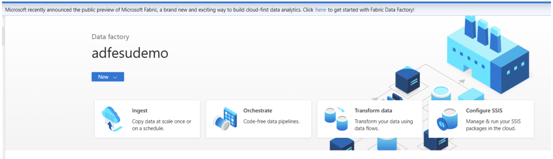
Lab

- Click `Go to resource` go to the `Data Factory` and lunch `Azure Data Factory Studio`



Resource group (move) : xzhang2
Status : Succeeded
Location : East US
Subscription (move) : Azure subscription 1
Subscription ID : c375ef05-a27a-41f9-ba5e-20eebfeb29c6
Type : Data factory (V2)
Getting started : Quick start

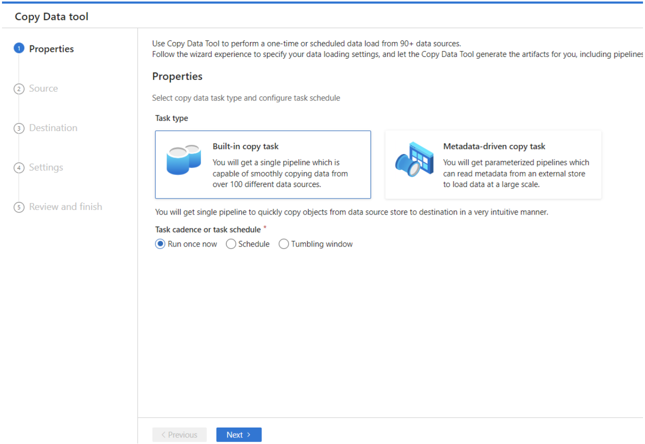
- Now go to the home section and click on `Ingest`


Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Data factory
adfesudemo
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.

Lab

- Choose Task type as `Built-in copy task`
- Choose “Run once now” and click on “next”



Copy Data tool

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

Properties

Select copy data task type and configure task schedule

Task type

Built-in copy task
You will get a single pipeline which is capable of smoothly copying data from over 100 different data sources.

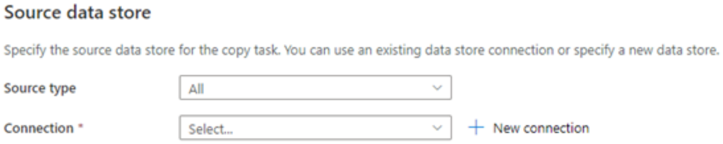
Metadata-driven copy task
You will get parameterized pipelines which can read metadata from an external store to load data at a large scale.

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

Lab

- Click on “+ New connection”



Source data store

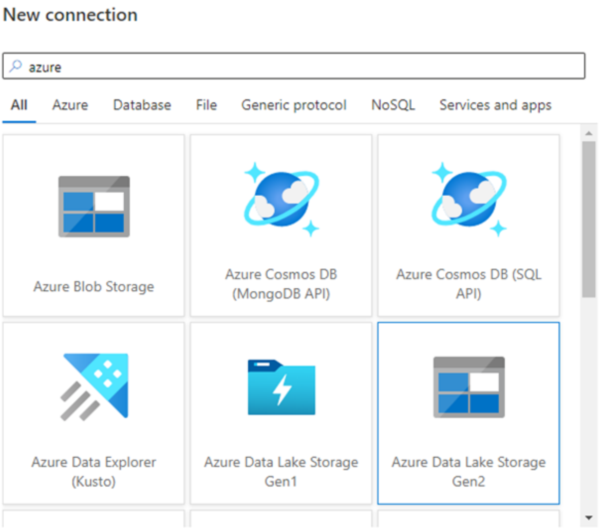
Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: All

Connection *: Select...

Lab

- Select **Azure Data Lake Storage Gen-2** and click on `continue`. You are choosing this because we have a dataset stored in the Gen-2 already.



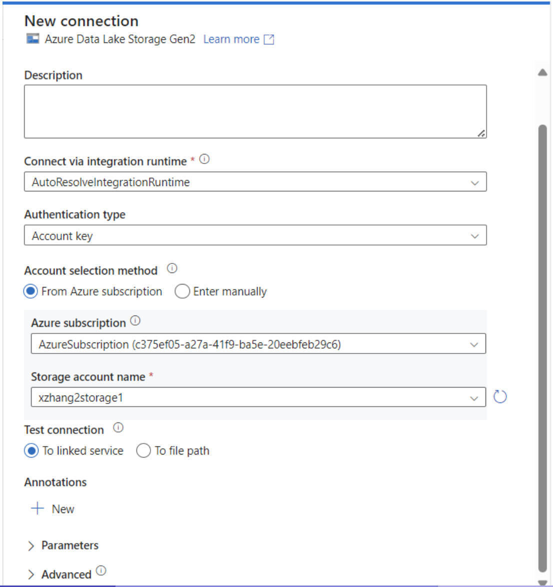
New connection

azure

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

Lab

- Enter the details, *test connection*, and click on the `create` button



New connection

Azure Data Lake Storage Gen2 Learn more

Description

Connect via integration runtime * : AutoResolveIntegrationRuntime

Authentication type: Account key

Account selection method

● From Azure subscription ○ Enter manually

Azure subscription: AzureSubscription (c375ef05-a27a-41f9-ba5e-20eebfeb29c6)

Storage account name * : xzhang2storage1

Test connection

● To linked service ○ To file path

Annotations

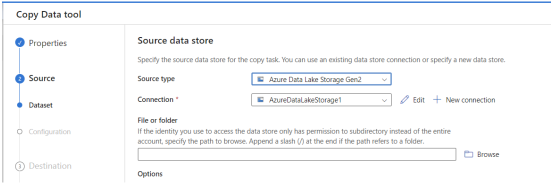
+ New

> Parameters

> Advanced

Lab

- Select the source type as `Azure Data Lake Storage Gen-2`, now click on the `Browse` button



Copy Data tool

Source data store

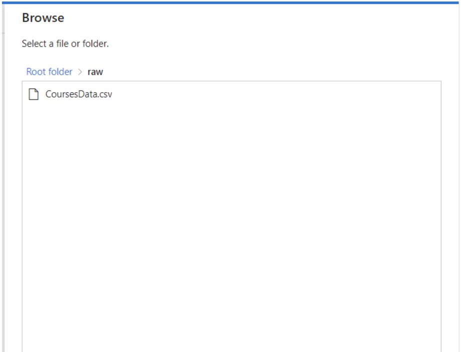
Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store.

Source type: Azure Data Lake Storage Gen2

Connection *: AzureDataLakeStorage1

File or folder

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. Append a slash (/) at the end if the path refers to a folder.



Browse

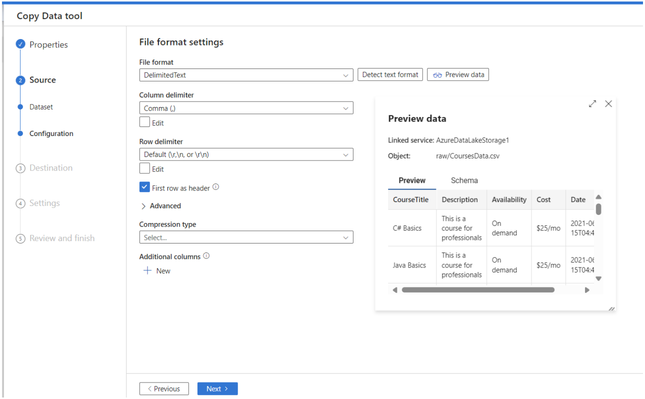
Select a file or folder.

Root folder > raw

CoursesData.csv

Lab

- Untick the `recursively` option
- Click on `Next` and click on “Preview data” to check the connection



Copy Data tool

File format settings

File format: DelimitedText

Column delimiter: Comma (,)

□ Edit

Row delimiter: Default (\r,\n, or \r\n)

□ Edit

☑ First row as header ⓘ

> Advanced

Compression type: Select...

Additional columns ⓘ

Preview data

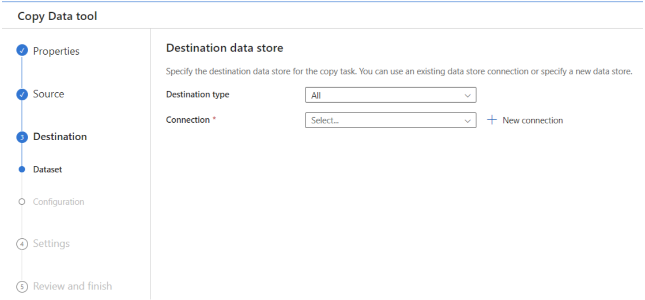
Linked service: AzureDataLakeStorage1

Object: raw/CoursesData.csv

CourseTitle	Description	Availability	Cost	Date
C# Basics	This is a course for professionals	On demand	$25/mo	2021-06-15T04:4
Java Basics	This is a course for professionals	On demand	$25/mo	2021-06-15T04:4

Lab

- Click on `Next`
- Now let us specify our "new target". Click on the "new connection"



Copy Data tool

Destination data store

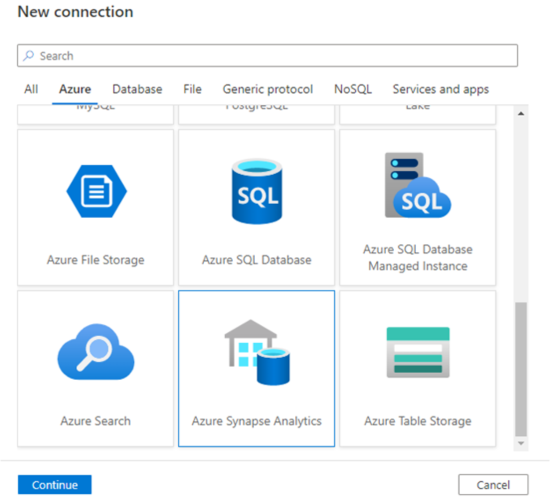
Specify the destination data store for the copy task. You can use an existing data store connection or specify a new data store.

Destination type: All

Connection *: Select...

Lab

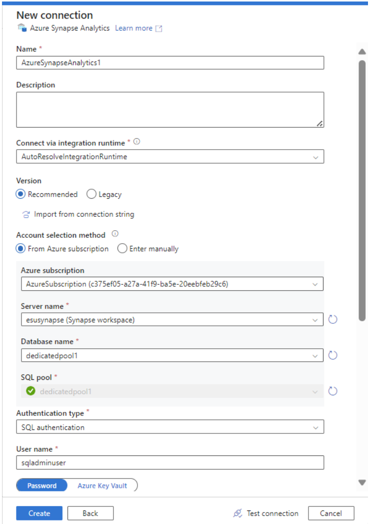
- Select `Azure Synapse Analytics` and click on the `Continue` button



New connection

Lab

- Enter the details, *test connection*, and click on the `Create` button



New connection

Azure Synapse Analytics Learn more

Name *
AzureSynapseAnalytics1

Description

Connect via integration runtime *
AutoResolveIntegrationRuntime

Version
● Recommended ○ Legacy

Import from connection string

Account selection method
● From Azure subscription ○ Enter manually

Azure subscription
AzureSubscription (c375ef05-a27a-41f9-ba5e-20eebfeb29c6)

Server name *
esusynapse (Synapse workspace)

Database name *
dedicatedpool1

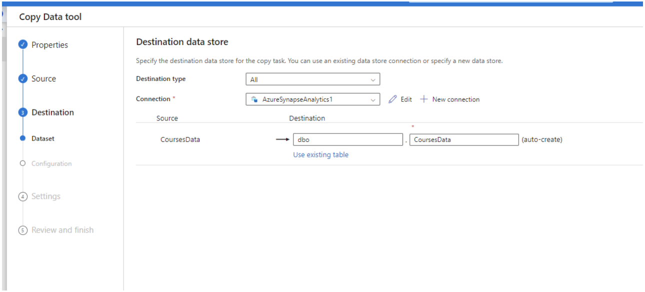
SQL pool *
✓ dedicatedpool1

Authentication type *
SQL authentication

User name *
sqladminuser

Lab

- Click on "Use existing table" and click `Auto-create a destination table with the source schema`



Copy Data tool

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.

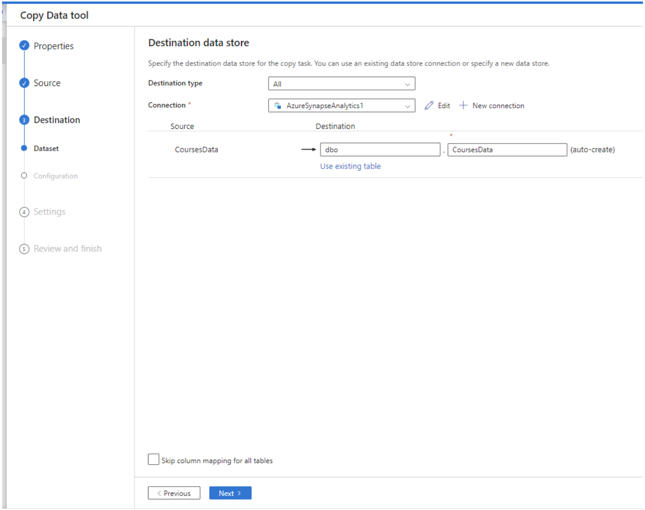
Destination type: All

Connection *: AzureSynapseAnalytics1

Source	Destination
CoursesData	dbo . CoursesData (auto-create)

Lab

- Click on "Next"



Copy Data tool

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.

Destination type: All

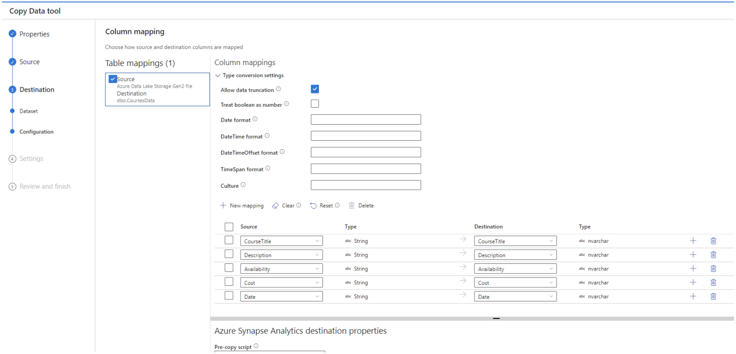
Connection *: AzureSynapseAnalytics1

Source	Destination		
CoursesData	dbo	CoursesData	(auto-create)

□ Skip column mapping for all tables

Lab

- Keep the “conversion settings”.



Copy Data tool

Column mapping

Choose how source and destination columns are mapped

Table mappings (1)

☑ Source
Azure Data Lake Storage Gen2 file
Destination
dbo.CoursesData

Column mappings

Type conversion settings

Allow data truncation	☑
Treat boolean as number	□
Date format	
DateTime format	
DateTimeOffset format	
TimeSpan format	
Culture	

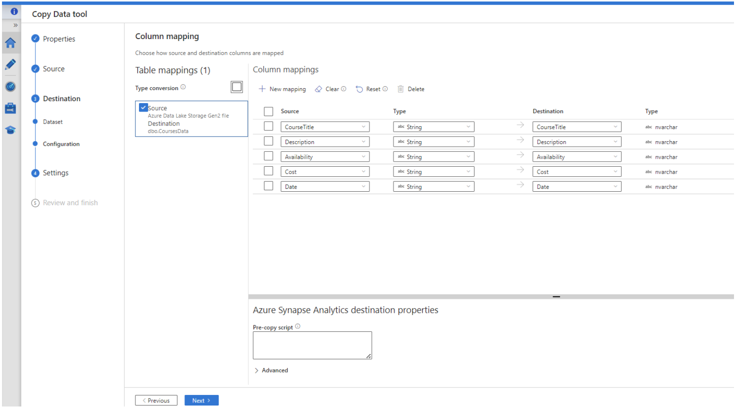
□	Source	Type		Destination	Type
□	CourseTitle	abc String	→	CourseTitle	abc nvarchar
□	Description	abc String	→	Description	abc nvarchar
□	Availability	abc String	→	Availability	abc nvarchar
□	Cost	abc String	→	Cost	abc nvarchar
□	Date	abc String	→	Date	abc nvarchar

Azure Synapse Analytics destination properties

Pre-copy script

Lab

- Have a look at the “Column mappings”, make sure **Untick** the type “Type conversion” option, and click on `Next`



Copy Data tool

- Properties
- Source
- Destination
- Dataset
- Configuration
- Settings
- Review and finish

Column mapping

Choose how source and destination columns are mapped

Table mappings (1)

Type conversion □

☑ Source
Azure Data Lake Storage Gen2 file
Destination
dbo.CoursesData

Column mappings

New mapping · Clear · Reset · Delete

□	Source	Type		Destination	Type
□	CourseTitle	abc String	→	CourseTitle	abc nvarchar
□	Description	abc String	→	Description	abc nvarchar
□	Availability	abc String	→	Availability	abc nvarchar
□	Cost	abc String	→	Cost	abc nvarchar
□	Date	abc String	→	Date	abc nvarchar

Azure Synapse Analytics destination properties

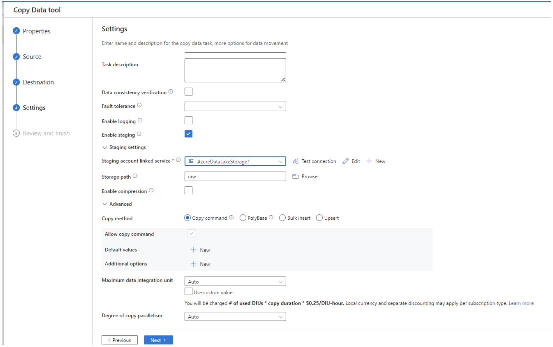
Pre-copy script

> Advanced

< Previous · Next >

Lab

- Give a name to the task and the storage path. Specify the staging account linked service as well



Settings

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

Task description:

Data consistency verification: □

Fault tolerance:

Enable logging: □

Enable staging: ☑

Staging settings

Staging account linked service *: AzureDataLakeStorage1

Storage path: raw

Enable compression: □

Advanced

Copy method: ● Copy command ○ PolyBase ○ Bulk insert ○ Upsert

Allow copy command: ☑

Default values

Additional options

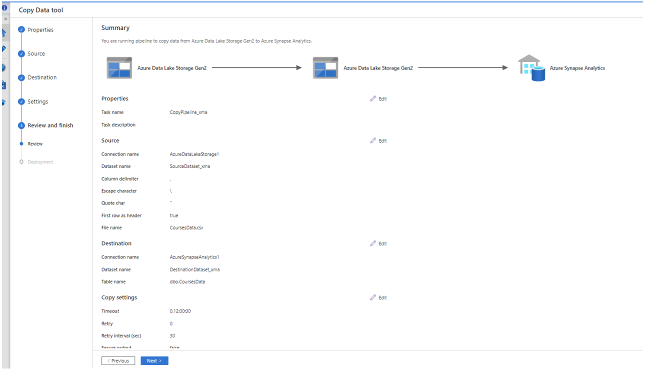
Maximum data integration unit: Auto

□ Use custom value

You will be charged **# of used DIUs * copy duration * $0.25/DIU-hour.** Local currency and separate discounting may apply per subscription type.

Degree of copy parallelism: Auto

Lab


Copy Data tool

Summary

You are running pipeline to copy data from Azure Data Lake Storage Gen2 to Azure Synapse Analytics.

Properties ✎ Edit

Task name	CopyPipeline_xma
Task description	

Source ✎ Edit

Connection name	AzureDataLakeStorage1
Dataset name	SourceDataset_xma
Column delimiter	,
Escape character	\
Quote char	"
First row as header	true
File name	CoursesData.csv

Destination ✎ Edit

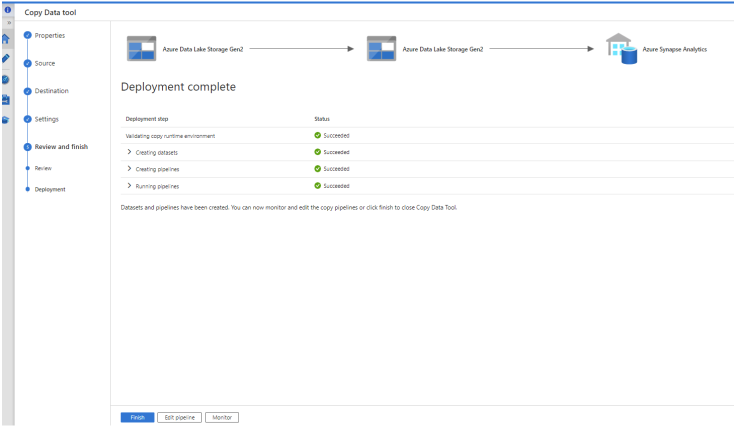
Connection name	AzureSynapseAnalytics1
Dataset name	DestinationDataset_xma
Table name	dbo.CoursesData

Copy settings ✎ Edit

Timeout	0.12:00:00
Retry	0
Retry interval (sec)	30
Secure output	false

- Click on `Next` again

Lab


Copy Data tool

Deployment complete

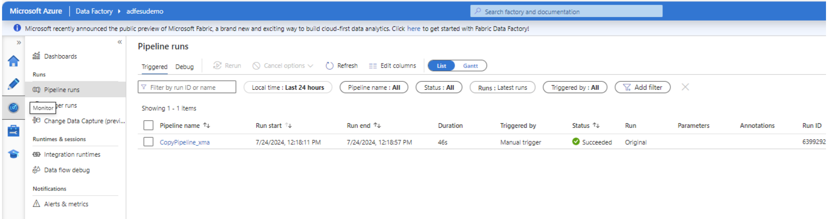
Deployment step	Status
Validating copy runtime environment	✓ Succeeded
› Creating datasets	✓ Succeeded
› Creating pipelines	✓ Succeeded
› Running pipelines	✓ Succeeded

Datasets and pipelines have been created. You can now monitor and edit the copy pipelines or click finish to close Copy Data Tool.

- Click on the “Finish” button

Lab

- Now go to the “monitor” section and you should be able to see the pipeline in succeeded state or progress state



Microsoft recently announced the public preview of Microsoft Fabric, a brand new and exciting way to build cloud-first data analytics. Click here to get started with Fabric Data Factory!

Pipeline runs

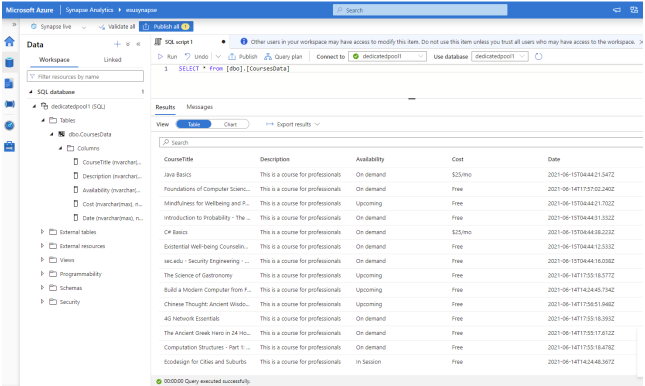
Showing 1 - 1 items

□	Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Run ID
□	CopyPipeline_xma	7/24/2024, 12:18:11 PM	7/24/2024, 12:18:57 PM	46s	Manual trigger	✓ Succeeded	Original			6399292

Lab

- Now go to or open the “Azure Synapse studio”,
  - and check the database or
  - create a SQL script and run the following command and you must be able to see the data fetched in Azure Synapse

```
SELECT * from [dbo].[CoursesData]
```



Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.

```
SELECT * from [dbo].[CoursesData]
```

CourseTitle	Description	Availability	Cost	Date
Java Basics	This is a course for professionals	On demand	$25/mo	2021-06-15T04:44:21.547Z
Foundations of Computer Scienc...	This is a course for professionals	On demand	Free	2021-06-14T17:57:02.240Z
Mindfulness for Wellbeing and P...	This is a course for professionals	Upcoming	Free	2021-06-15T04:44:21.702Z
Introduction to Probability - The ...	This is a course for professionals	On demand	Free	2021-06-15T04:44:31.332Z
C# Basics	This is a course for professionals	On demand	$25/mo	2021-06-15T04:44:38.223Z
Existential Well-being Counselin...	This is a course for professionals	On demand	Free	2021-06-15T04:44:12.533Z
sec.edu - Security Engineering - ...	This is a course for professionals	On demand	Free	2021-06-15T04:44:16.038Z
The Science of Gastronomy	This is a course for professionals	Upcoming	Free	2021-06-14T17:55:18.577Z
Build a Modern Computer from F...	This is a course for professionals	Upcoming	Free	2021-06-14T14:24:45.734Z
Chinese Thought: Ancient Wisdo...	This is a course for professionals	Upcoming	Free	2021-06-14T17:56:51.948Z
4G Network Essentials	This is a course for professionals	On demand	Free	2021-06-14T17:55:18.393Z
The Ancient Greek Hero in 24 Ho...	This is a course for professionals	On demand	Free	2021-06-14T17:55:17.612Z
Computation Structures - Part 1: ...	This is a course for professionals	On demand	Free	2021-06-14T17:55:18.478Z
Ecodesign for Cities and Suburbs	This is a course for professionals	In Session	Free	2021-06-14T14:24:48.367Z

00:00:00 Query executed successfully.

lab

- Delete everything we have done in Azure

License