



Azure Data Factory

Azure Data Factory (ADF) is a cloud-based data integration service offered by Microsoft Azure. It enables organizations to create, manage, and orchestrate workflows for data movement and data transformation across various data sources and destinations. ADF is particularly useful for building data pipelines to process and prepare data for analytics, machine learning, or reporting.

Key Features of Azure Data Factory:

1. Data Integration:

- Supports integration with a wide variety of data sources, including Azure services, on-premises databases, cloud platforms, and third-party services.
- Connectors for services like Azure Blob Storage, SQL Server, Amazon S3, and Google BigQuery.

2. Data Movement:

- Facilitates the transfer of data between different sources and destinations securely and efficiently.
- Can move data at scale using Azure's global infrastructure.

3. Data Transformation:

- Supports transforming data using tools like:
 - Mapping Data Flows: A visual, code-free interface for creating transformations.
 - Custom transformations using Azure Databricks, Azure HDInsight, or Azure Functions.
 - SQL transformations for structured data.

4. Workflow Orchestration:

- Allows you to define and schedule workflows (pipelines) to automate data processes.
- Supports branching, looping, and error-handling logic.

5. Scalability and Performance:

- Automatically scales to handle large data volumes.
- Offers monitoring and management tools to track pipeline performance and resolve issues.

6. Hybrid Data Integration:

- Can integrate data from on-premises and cloud sources using a self-hosted integration runtime.

7. Security:

- Supports authentication through Azure Active Directory.
- Provides data encryption and role-based access control (RBAC).

Common Use Cases:

1. **Data Ingestion:** Importing raw data into a data lake or data warehouse.
2. **Data Preparation:** Cleaning, transforming, and enriching data for analytics.
3. **ETL/ELT Workflows:** Extracting, transforming, and loading (ETL) or extracting, loading, and transforming (ELT) data.
4. **Data Migration:** Moving data from legacy systems to modern platforms.
5. **Big Data Processing:** Integrating with big data platforms like Databricks or HDInsight for processing.

Benefits:

- **Ease of Use:** Intuitive UI and code-free capabilities.
- **Flexibility:** Supports a wide range of data operations and sources.
- **Cost-Effective:** Pay-as-you-go pricing model.
- **Integration with Azure Ecosystem:** Works seamlessly with other Azure services like Azure Synapse Analytics, Azure Machine Learning, and Power BI.

Azure Data Factory is a powerful tool for organizations aiming to streamline and modernize their data integration and processing workflows in a cloud-first world.

😊 To begin with the Lab

1. There are some prerequisites for this lab and they are, that you should have resources from previous labs.
2. You should have an external Gen2 storage account in which you should only have the CSV file.
3. Then you should have a Synapse Workspace running with a Dedicated SQL Pool or data warehouse.
4. Also, you should have an SQL Database on which sample data should be loaded while creating.
5. Once you have created all the prerequisite resources then we can move ahead with this lab.
6. Open the Synapse studio from the overview page and then create a new SQL Script. Then create a new SQL in the dedicated SQL Pool.

```

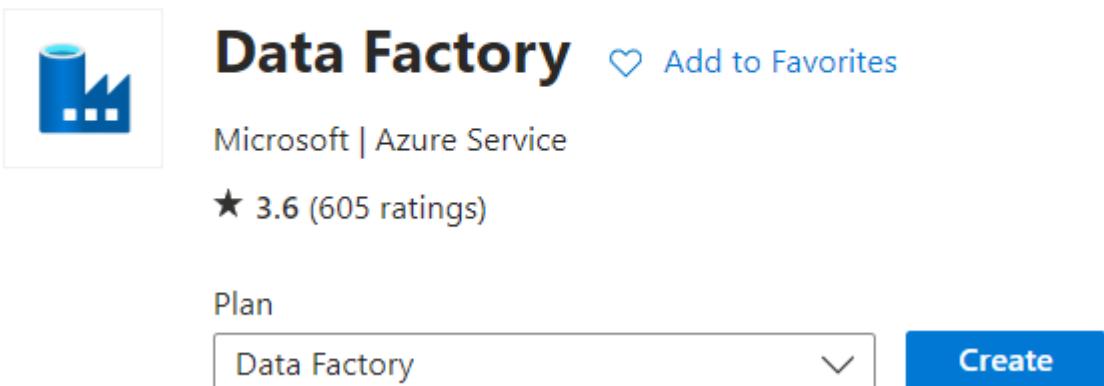
1   CREATE TABLE PoolActivityLog
2   (
3       [Correlationid] varchar(200),
4       [Operationname] varchar(300),
5       [Status] varchar(100),
6       [Eventcategory] varchar(100),
7       [Level] varchar(100),
8       [Time] varchar(100),
9       [Subscription] varchar(200),
10      [Eventinitiatedby] varchar(1000),
11      [Resourcetype] varchar(300),
12      [Resourcegroup] varchar(1000),
13      [Resource] varchar(2000)
14  )
15  WITH
16  (
17      DISTRIBUTION= HASH(Operationname)
18  )

```

- Now we are going to create the Azure Data Factory. Search and navigate to the creation for it.

Data Factory

Microsoft



- Choose your resource group and give it a name then move to the review page to create your data factory.

Basics Git configuration Networking Advanced Tags Review + create

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 *	①	MSDN Platforms Subscription	▼
Resource group *		NewRG	▼
Create new			

Instance details

Name *	①	thedatafactory121	✓
Region *	①	North Europe	▼
Version *	①	V2	▼

9. This is the dashboard for the Data factory. It is quite similar to what we have seen in the Synapse Analytics. We have to click on Ingest to directly create a pipeline.

The screenshot shows the Azure Data Factory dashboard for the factory named 'thedatafactory121'. On the left, there's a navigation sidebar with options: Home, Author, Monitor, Manage, and Learning Center. The main area is titled 'Data factory' and shows the factory name 'thedatafactory121'. Below the title, there are four main buttons: 'Ingest' (highlighted with a red box), 'Orchestrate' (Code-free data pipelines), 'Transform data' (Transform your data using data flows), and 'Configure SSIS' (Manage & run your SSIS packages in the cloud). Underneath these buttons, there's a section titled 'Recent resources' which says 'No items to show'. At the bottom, there are links for 'Discover more' and 'Browse partners'.

10. First, we will run the copy data tool, choose the built-in copy task and choose run once now for task schedule.

Copy Data tool

① Properties
② Source
③ Destination
④ Settings
⑤ Review and finish

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

11. Now for the source we will create a new connection for our data lake. Click on new connection.

Copy Data tool

① Properties
② Source
③ Dataset
④ Configuration

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... [New connection](#)

12. Choose the Azure Data Lake Gen2 and move to next page.

New connection

Search

All Azure Database File Generic protocol

Azure Data Explorer (Kusto)	Azure Data Lake Storage Gen2	Azure Database for MariaDB (Legacy)

13. Then give a name to the connection and then choose your subscription and the storage account then test the connection. Click on create.

New connection

 Azure Data Lake Storage Gen2 [Learn more](#) 

Name *

Description

Connect via integration runtime * 

AutoResolveIntegrationRuntime 

Authentication type

Account selection method 

From Azure subscription Enter manually

Account selection method 

From Azure subscription Enter manually

Azure subscription 

Storage account name *

Test connection 

To linked service To file path

Annotations

 New

Parameters

Advanced 

 Connection successful

Create

Back

 Test connection

Cancel

14. After that you have to choose the container and move to the next step.

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

Connection *

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.

15. Now we will create a new connection for destination which will be our dedicated SQL Pool.

Copy Data tool

Properties

Source

Destination

Dataset

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

Connection *

16. For the new connection choose Azure Synapse Analytics.

New connection

All Azure Database File Generic protocol



Azure Synapse Analytics

17. Give name to your connection, choose your subscription, server name and the database name, then you have to give the username and password of the synapse analytics. Test the connect and click on create.

New connection

 Azure Synapse Analytics [Learn more](#) 

Name *

Description

Connect via integration runtime * 

AutoResolveIntegrationRuntime 

Version

Recommended Legacy

 Import from connection string

Account selection method 

From Azure subscription Enter manually

Azure subscription

MSDN Platforms Subscription (d6549a66-c45c-4979-840c-3b356da446b0) 

Server name *

synapsewfp5oga (Synapse workspace) 

Database name *

DataPool 

SQL pool *

DataPool 

Authentication type *

SQL authentication 

User name *

SQLUser 

Password **Azure Key Vault**

Password *

..... 

Encrypt 

Mandatory 

 Connection successful

Create **Back**  **Test connection** **Cancel**

18. Then you have to choose the table you created in your dedicated SQL Pool.

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 *	<input data-bbox="547 1684 960 1718" type="button" value="synapseworkspace"/>  Edit  New connection
Source	Destination
▼ AzureBlobFSFile	 <input data-bbox="706 1808 1341 1841" type="button" value="dbo.PoolActivityLog"/>  Auto-create a destination table with the source schema

19. In the end give the name to your Pipeline and choose bulk insert and create your pipeline.

Settings

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

Task name *	<input type="text" value="PoolTable"/>
Task description	<input type="text"/>
Data consistency verification ⓘ	<input type="checkbox"/>
Fault tolerance ⓘ	<input type="text"/>
Enable logging ⓘ	<input type="checkbox"/>
Enable staging ⓘ	<input type="checkbox"/>
▼ Advanced	
Copy method	<input type="radio"/> Copy command ⓘ <input checked="" type="radio"/> Bulk insert <input type="radio"/> Upsert

20. Here you can see that our pipeline run has been completed.

Pipeline runs							
Triggered		Debug		Rerun		Cancel options	
Triggered by	All	Filter by run ID or name	Chennai, Kolkata, Mu... : Last 24 hours	Pipeline name	All	Status	All
Run start	↑↓	Run end	↑↓	Duration	Triggered by	Status	↑↓
Parameters	Run	Last refreshed 0 minutes ago					
<input type="checkbox"/> Pipeline name ↑↓		12/25/2024, 9:21:57 PM	12/25/2024, 9:22:20 PM	23s	Manual trigger	<input checked="" type="checkbox"/> Succeeded	Original
<input type="checkbox"/> PoolTable							

21. Also, run the Select command in the Synapse Analytics, you will see that the table has the data now.

```
20  SELECT * FROM PoolActivityLog
```

Results Messages

View

Table

Chart

Export results

Search

Correlationid	Operationname	Status	Eventcategory	Level	Time
8365a1d0-bc84...	Create or Upda...	Started	Administrative	Informational	2024-05-25T13:...
8365a1d0-bc84...	Create or Upda...	Started	Administrative	Informational	2024-05-25T13:...
ff9a8973-d0ea...	Create or Upda...	Succeeded	Administrative	Informational	2024-05-23T08:...
1bca2daa-9de0...	Get namespace...	Succeeded	Administrative	Informational	2024-05-28T05:...
5fe0bef1-b882...	Delete Redis Ca...	Succeeded	Administrative	Informational	2024-05-10T08:...

00:00:03 Query executed successfully.

😃 Create Parquet file

1. Now we will create the parquet file using the copy data tool in Azure data factory using the CSV file.
2. Again, choose the Ingest method to create the Copy data tool. Choose your source as the data lake and the container.

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 *

dataLakestorage

Edit

+ New connection

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.

data/

 Browse

3. For the destination again choose the data lake and the container.

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 *

dataLakestorage

Edit

+ New connection

Folder path

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.

data/

 Browse

4. Then for the file format you have to choose Parquet as your file format and snappy for compression type.

File format settings

File format

Parquet

Compression type

snappy

Max rows per file

File name prefix

5. Move to the last step of the pipeline and give a name to your pipeline and create it. Below you can see that our pipeline has been created.

Pipeline runs							
Triggered		Debug		List		Gantt	
<input type="button"/> Filter by run ID or name		Chennai, Kolkata, Mu... : Last 24 hours		Pipeline name : All		Status : All	
<input type="button"/> Triggered by : All		<input type="button"/> Add filter		<input type="button"/> Runs : Latest runs		<input type="button"/> Copy filters <input type="button"/> Export to CSV	
Showing 1 - 2 items	Last refreshed 0 minutes ago						
<input type="checkbox"/> Pipeline name ↑↓	Run start ↑↓	Run end ↑↓	Duration	Triggered by	Status ↑↓	Run	Parameters
<input type="checkbox"/> parquetPipeline	12/25/2024, 9:29:23 PM	12/25/2024, 9:29:41 PM	19s	Manual trigger	<input checked="" type="checkbox"/> Succeeded	Original	
<input type="checkbox"/> PoolTable	12/25/2024, 9:21:57 PM	12/25/2024, 9:22:20 PM	23s	Manual trigger	<input checked="" type="checkbox"/> Succeeded	Original	

6. If you go back to the storage account and refresh it you will see the parquet file there.

Storage account							
Upload		Add Directory		Refresh		Actions	
Authentication method: Access key (Switch to Microsoft Entra user account)		Location: data					
Search blobs by prefix (case-sensitive)							
Name	Modified	Access tier	Archive status	Blob type	Size	Lease state	
<input type="checkbox"/> ActivityLog01.csv	12/25/2024, 9:05:28 ...	Hot (Inferred)		Block blob	1.91 MiB	Available	***
<input type="checkbox"/> ActivityLog01.parquet	12/25/2024, 9:29:40 ...	Hot (Inferred)		Block blob	230.78 Kib	Available	***