😊 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.

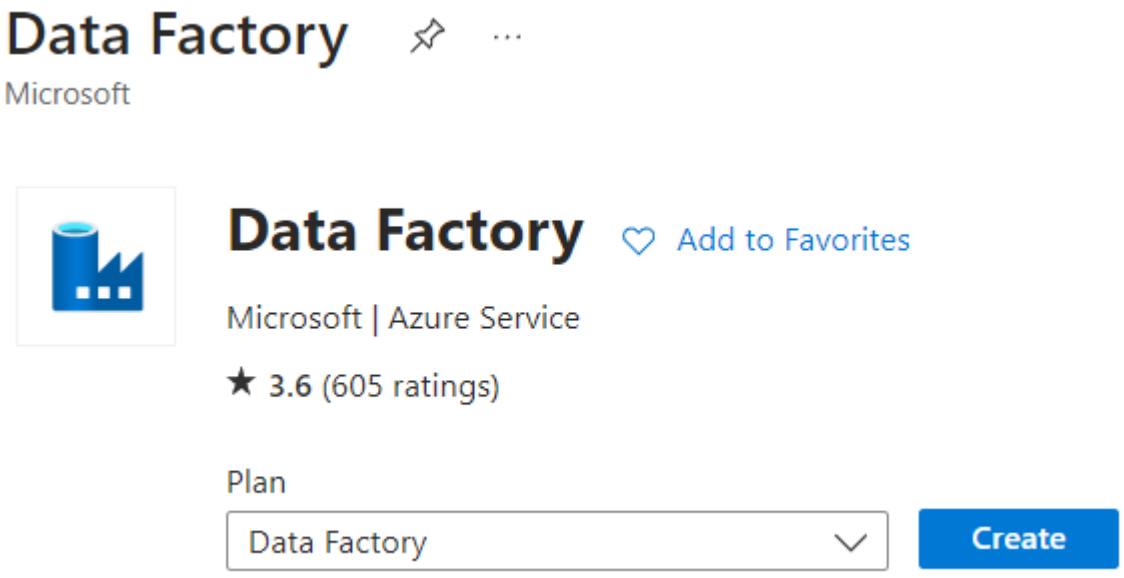
Azure Data Factory is a powerful tool for organizations aiming to streamline and modernize their data integration and processing workflows in a cloud-first world.

😄 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.

```
CREATE TABLE PoolActivityLog
(
    [Correlationid] varchar(200),
    [Operationname] varchar(300),
    [Status] varchar(100),
    [Eventcategory] varchar(100),
    [Level] varchar(100),
    [Time] varchar(100),
    [Subscription] varchar(200),
    [Eventinitiatedby] varchar(1000),
    [Resourcetype] varchar(300),
    [Resourcegroup] varchar(1000),
    [Resource] varchar(2000)
)
WITH
(
    DISTRIBUTION= HASH(Operationname)
)
```

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



8. 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 * : MSDN Platforms Subscription

Resource group * : NewRG

Create new

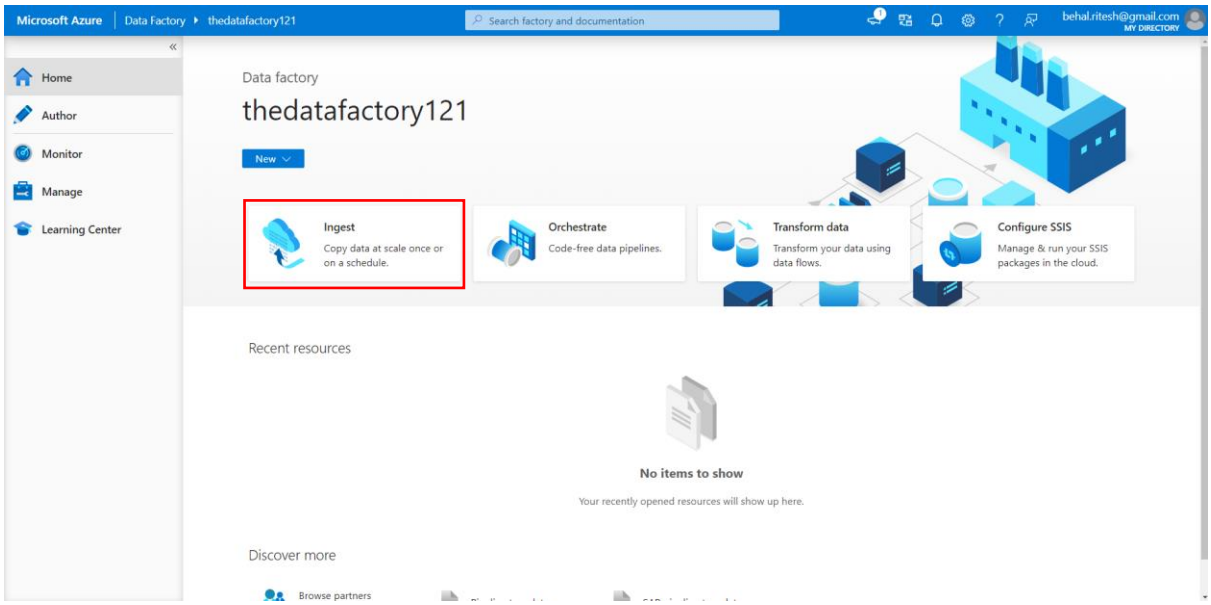
Instance details

Name * : thedatafactory121

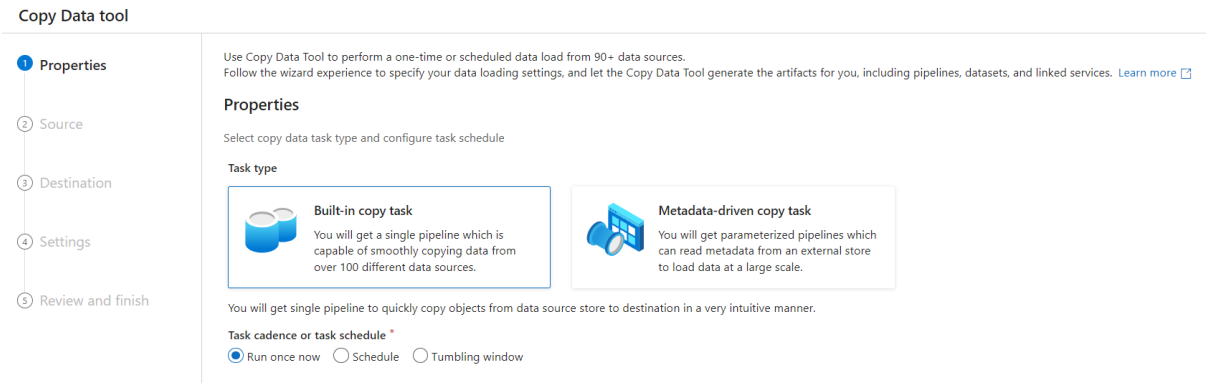
Region * : North Europe

Version * : V2

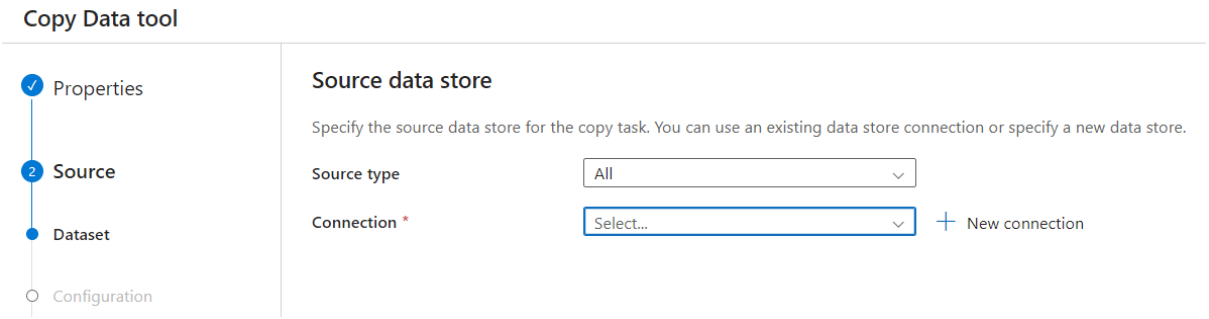
9. This is the dashboard for the Data factory. It is quite similar to what we have seen in the Synapse Analytics. We have to click on Ingest to directly create a pipeline.



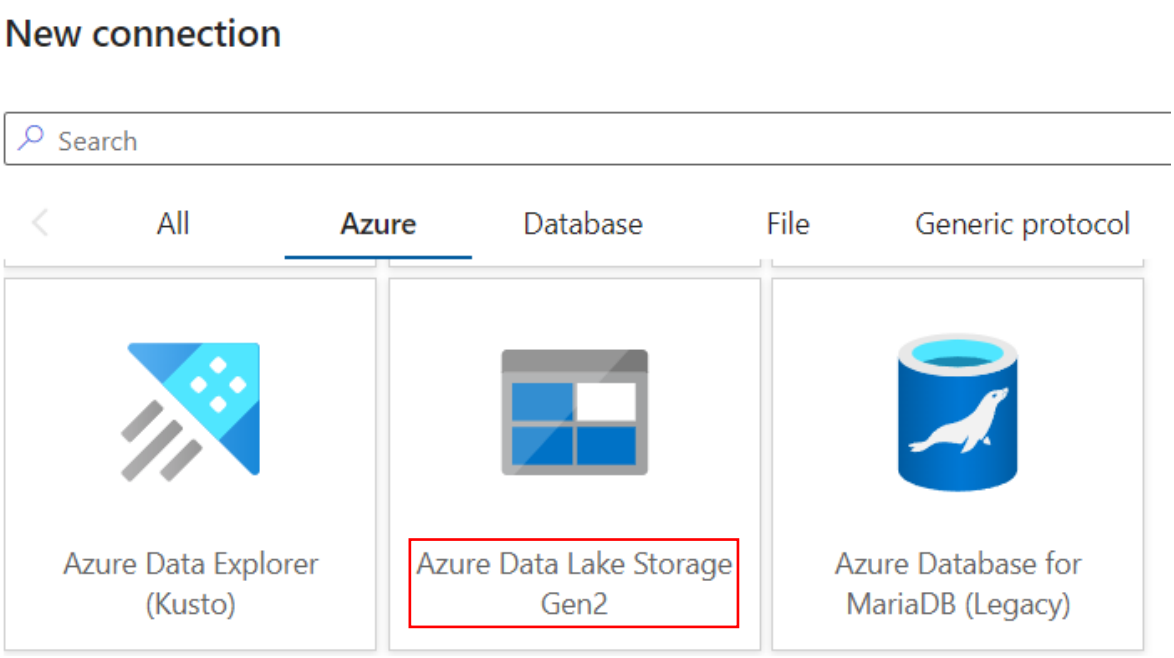
10. First, we will run the copy data tool, choose the built-in copy task and choose run once now for task schedule.



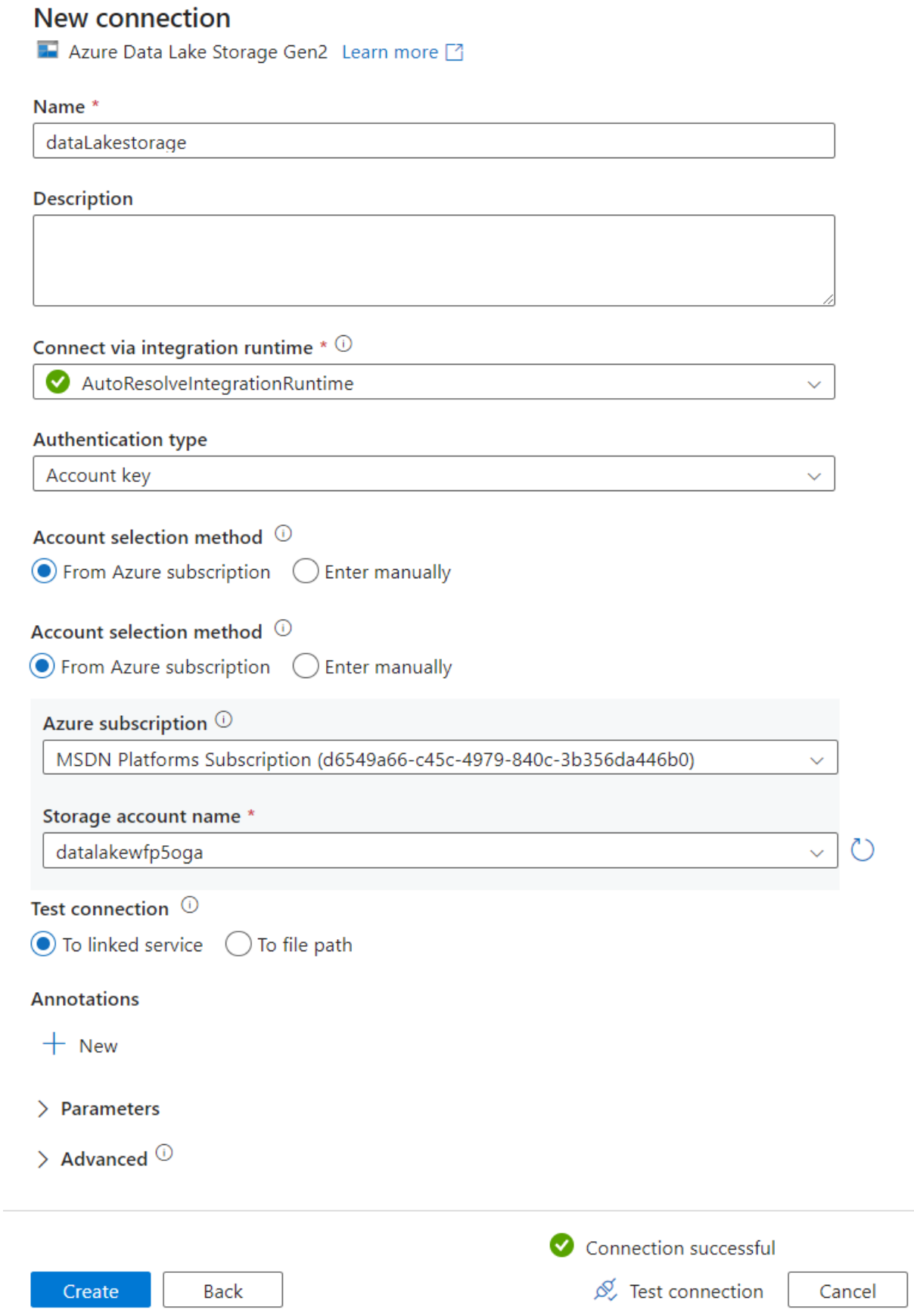
11. Now for the source we will create a new connection for our data lake. Click on new connection.



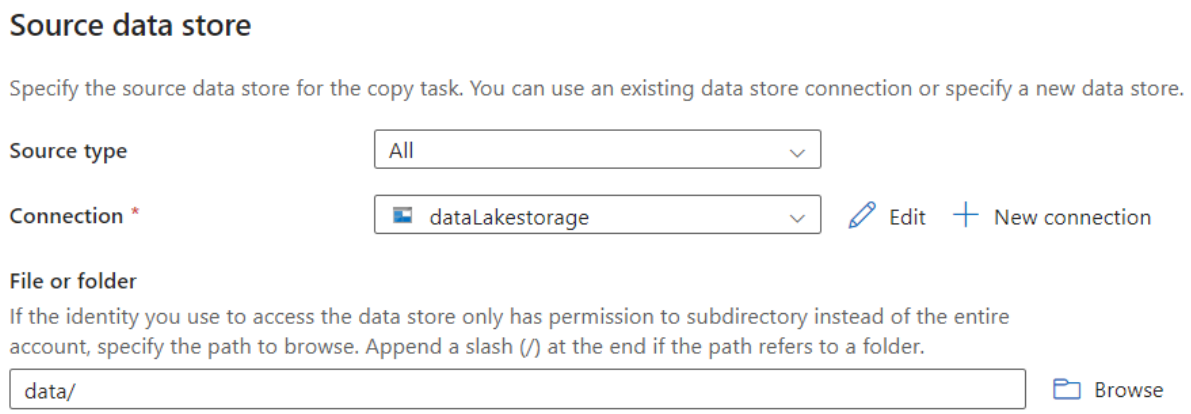
12. Choose the Azure Data Lake Gen2 and move to next page.



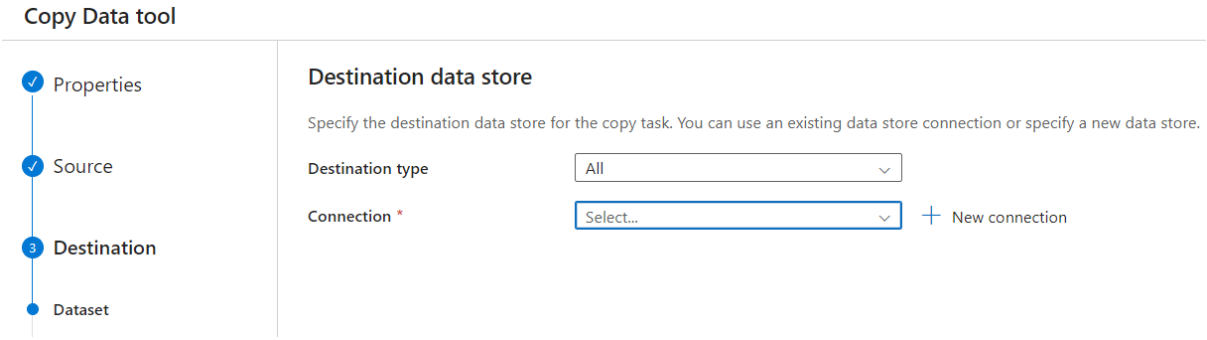
13. Then give a name to the connection and then choose your subscription and the storage account then test the connection. Click on create.



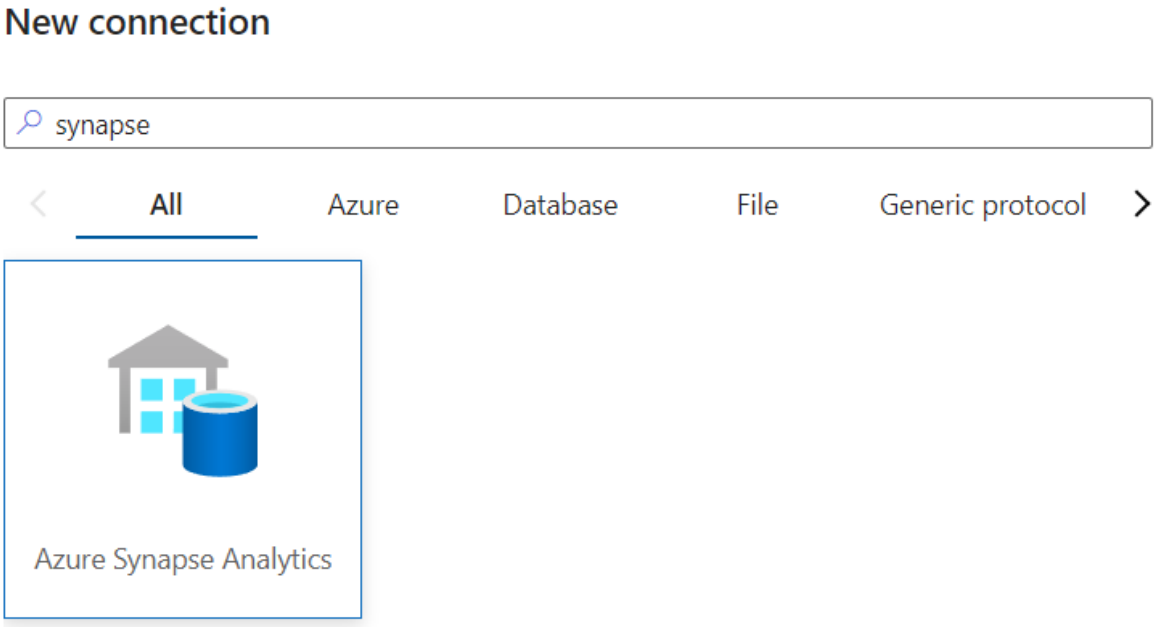
14. After that you have to choose the container and move to the next step.



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



16. For the new connection choose 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 *

synapseworkspace

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

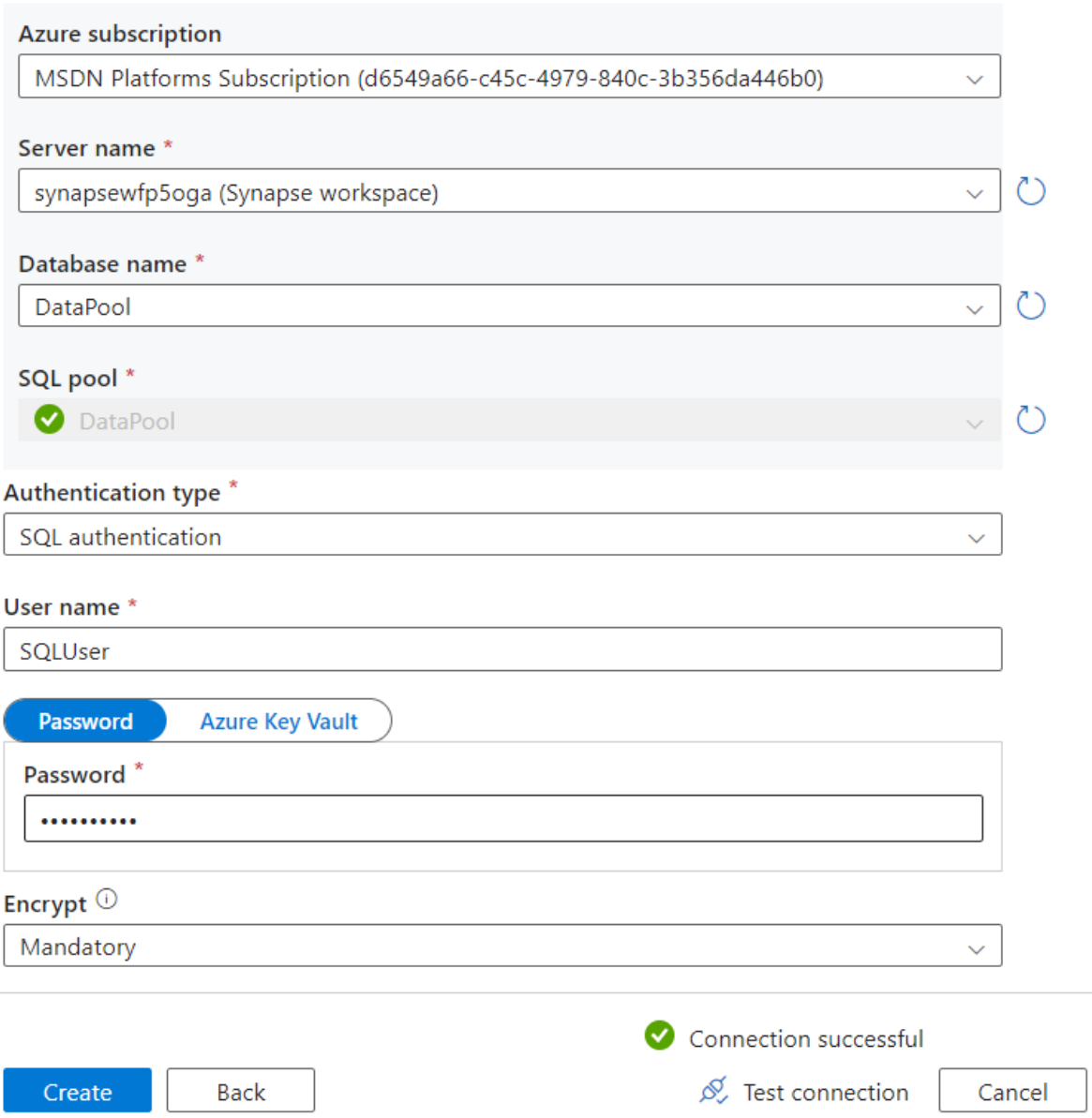
Version

(•) Recommended ( ) Legacy

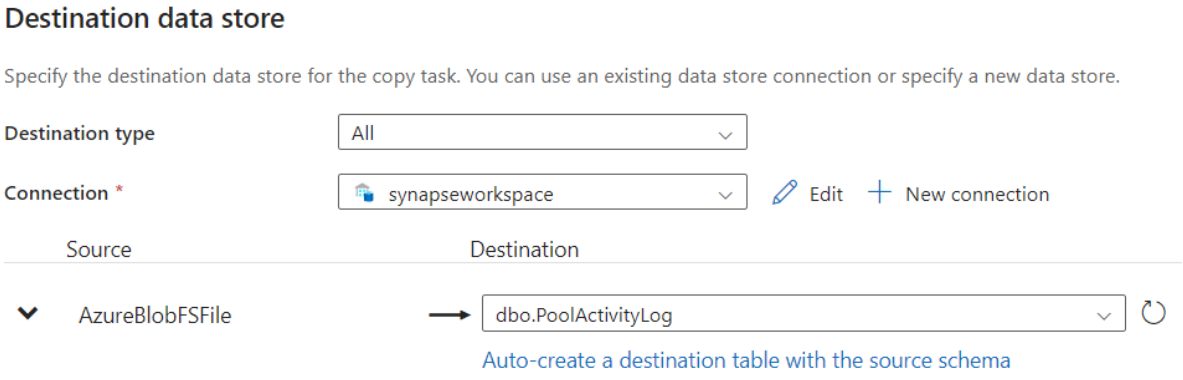
Import from connection string

Account selection method

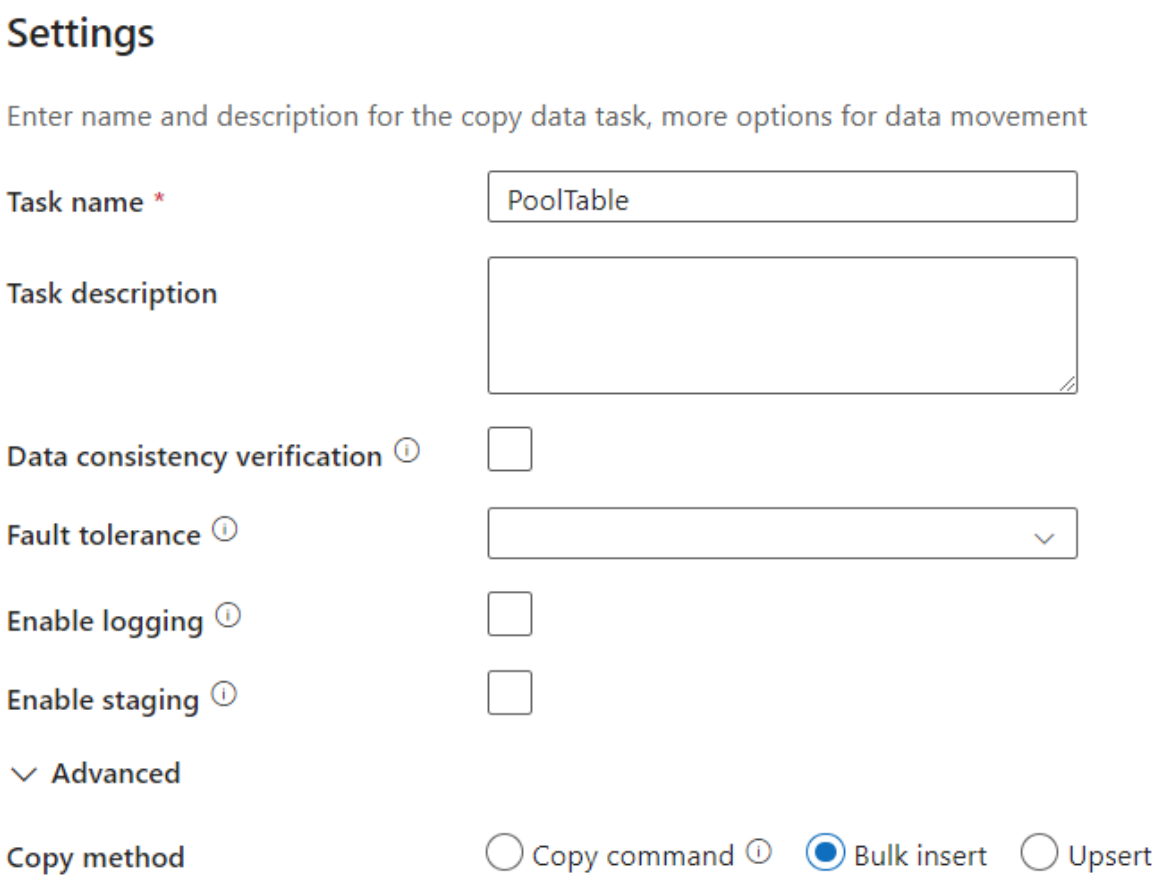
(•) From Azure subscription ( ) Enter manually



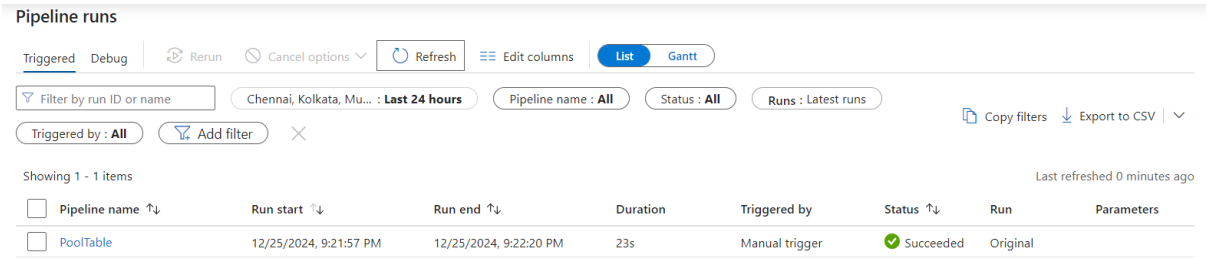
18. Then you have to choose the table you created in your dedicated SQL Pool.



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



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



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

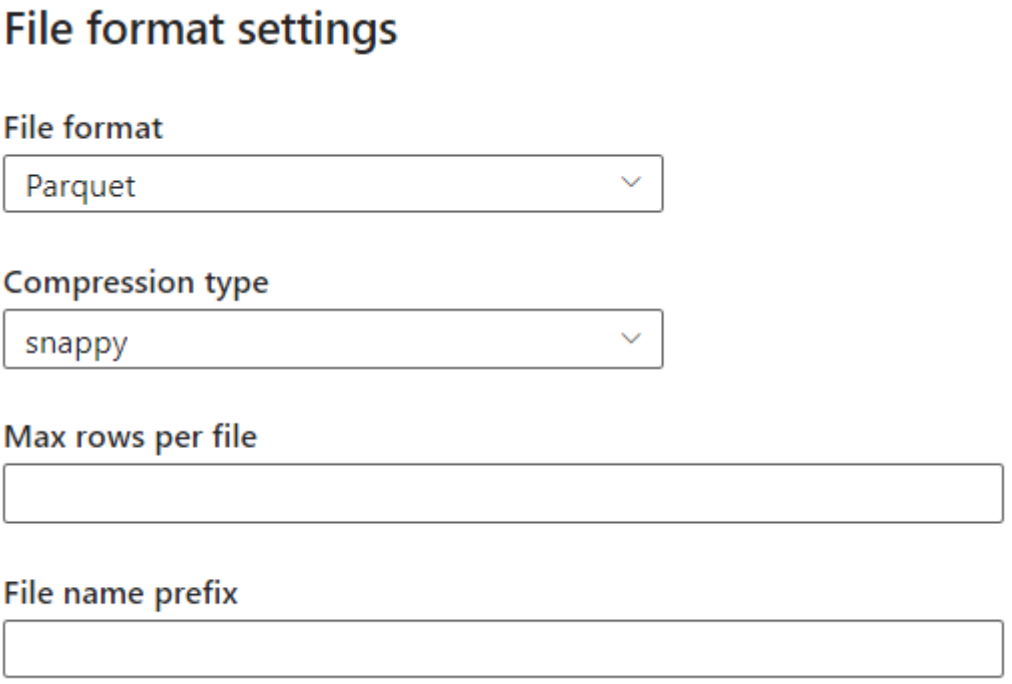
Connection *: dataLakestorage | Edit | New connection

Folder path

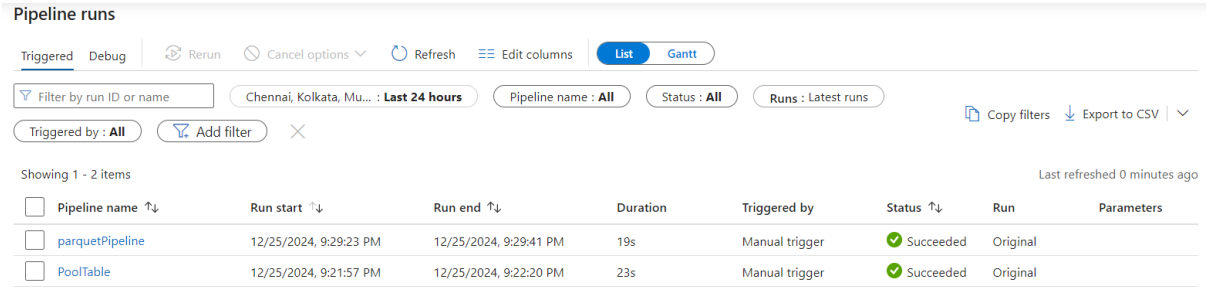
If the identity you use to access the data store only has permission to subdirectory instead of the entire account, specify the path to browse.

data/ | Browse

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



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.



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