😊 Loading Data using Pipelines – Storage Account

Pipelines in Azure Synapse Analytics

A **pipeline** in **Azure Synapse Analytics** is a collection of activities that automate data movement and transformation workflows. Pipelines are part of **Azure Synapse Pipelines**, which offer **data integration capabilities** to orchestrate Extract, Transform, and Load (ETL) or Extract, Load, and Transform (ELT) processes. These pipelines can manage complex workflows, integrate various data sources, and process large-scale data effectively.

What is the Copy Data Tool in Azure Synapse?

The **Copy Data Tool** is a feature in Azure Synapse that simplifies the process of copying data from one source to a destination. It is designed for **data migration, ingestion, and integration** tasks, enabling users to efficiently move data between different storage systems and databases.

Key Features of the Copy Data Tool

1. **Ease of Use:**
    - Provides a simple, wizard-based interface for creating data copy pipelines without requiring deep technical knowledge.
2. **Support for Multiple Data Sources:**
    - Supports copying data from diverse sources, including:
        - Azure Blob Storage
        - Azure Data Lake Storage
        - SQL Server
        - Azure SQL Database
        - REST APIs
        - SaaS applications (e.g., Salesforce)
        - On-premises systems (via integration runtime)
3. **Scalability:**
    - Optimized for handling large-scale data ingestion with parallel processing.
4. **Flexible Scheduling:**
    - Allows you to schedule data copy operations on a **one-time** or **recurring** basis.
5. **Data Transformation:**
    - Basic transformations like column mapping and data format conversion can be applied during data movement.

6. **Integration with Synapse Pipelines:**
    - Copy Data Tool workflows are built into Synapse Pipelines and can be extended to include additional activities.
7. **Monitoring:**
    - Provides detailed monitoring and logging for tracking data copy progress, success, and errors.

Steps to Use the Copy Data Tool

1. **Open the Copy Data Tool:**
    - In Azure Synapse Studio, navigate to the **Integrate** hub and select **Copy Data**.
2. **Configure Source Data:**
    - Choose the data source from a list of supported connectors.
    - Specify the connection details, such as server, database, and credentials.
3. **Configure Destination:**
    - Select the destination where the data will be copied.
    - Define storage options and target schema if applicable.
4. **Define Data Mapping:**
    - Map columns between the source and destination.
    - Optionally, perform data transformations during the mapping step.
5. **Set Up Scheduling:**
    - Choose whether the copy operation will run immediately, on a schedule, or on-demand.
6. **Run and Monitor:**
    - Execute the copy pipeline and monitor its progress through Synapse's monitoring tools.

Use Cases for the Copy Data Tool

1. **Data Ingestion:**
    - Load raw data from various sources into Azure Data Lake or Dedicated SQL Pools for analytics.
2. **Data Migration:**
    - Migrate data from on-premises systems to Azure cloud storage or databases.
3. **ETL/ELT Pipelines:**
    - Quickly implement the **Extract** and **Load** phases of ETL or ELT workflows.

4. **Backup and Archiving:**
    - Copy data from operational systems to long-term storage for backup or compliance purposes.
5. **Hybrid Data Integration:**
    - Synchronize data between on-premises and cloud environments.

Benefits of the Copy Data Tool

- **Simplified Data Movement:** No-code or low-code approach to data integration.
- **Speed:** Parallel data transfer capabilities for large-scale ingestion.
- **Flexibility:** Supports a wide range of data formats and storage systems.
- **Cost-Efficient:** Automates data movement without requiring additional tools or infrastructure.

Example Pipeline Using the Copy Data Tool

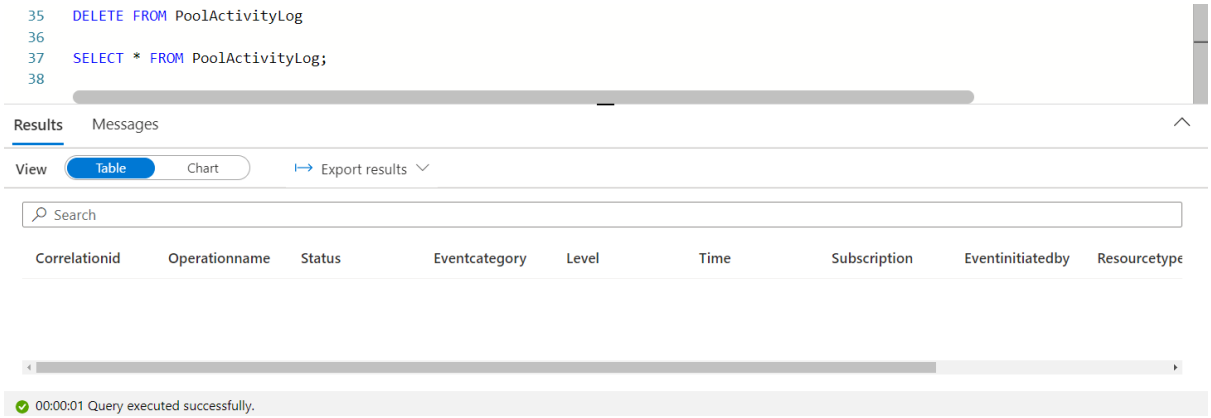
Scenario: Ingest sales data from an on-premises SQL Server database into Azure Data Lake for analysis.

1. **Source:** SQL Server (on-premises).
2. **Destination:** Azure Data Lake Storage Gen2.
3. **Steps:**
    - Create a linked service to SQL Server using a self-hosted integration runtime.
    - Use the Copy Data Tool to define the source query and map it to a destination folder in Azure Data Lake.
    - Set a daily schedule for the pipeline to run automatically.
    - Monitor the pipeline to ensure successful data ingestion.

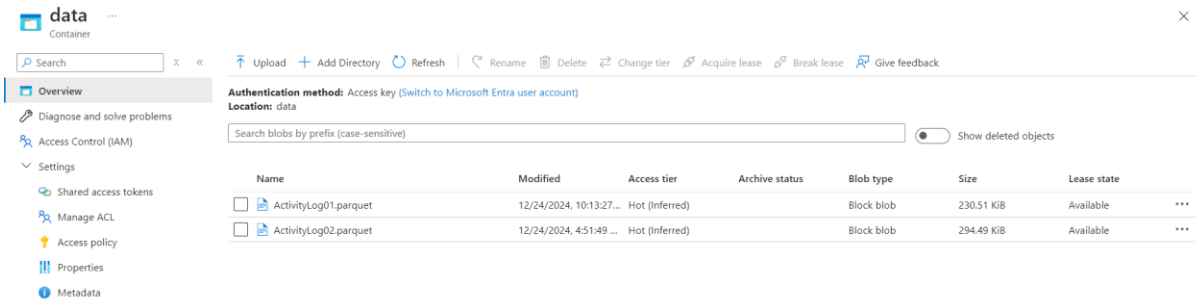
The **Copy Data Tool** is an essential feature of Azure Synapse Analytics for orchestrating and automating data integration workflows, ensuring efficient and reliable data movement across platforms.

😄 To begin with the Lab

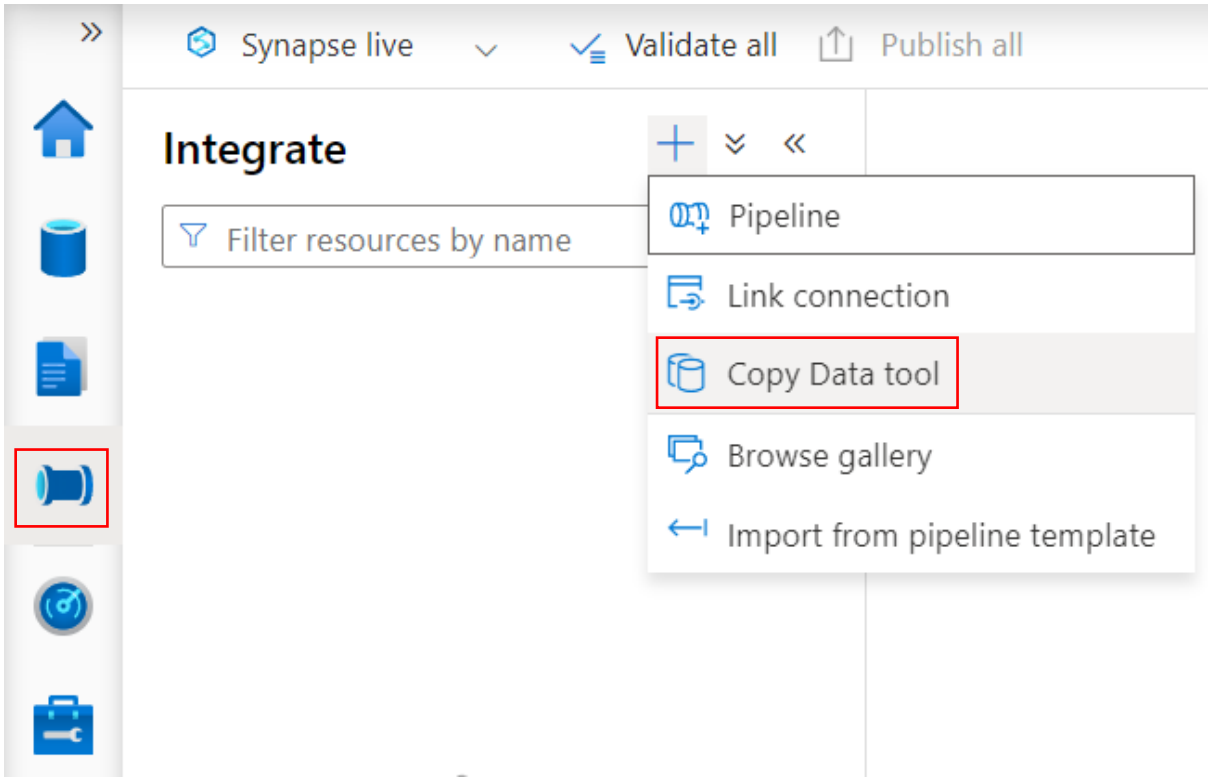
1. In this lab, we are going to use the pipelines to load the data into the tables. So, first, we will run the Delete command to delete the data in the table and then check that our table should be empty.



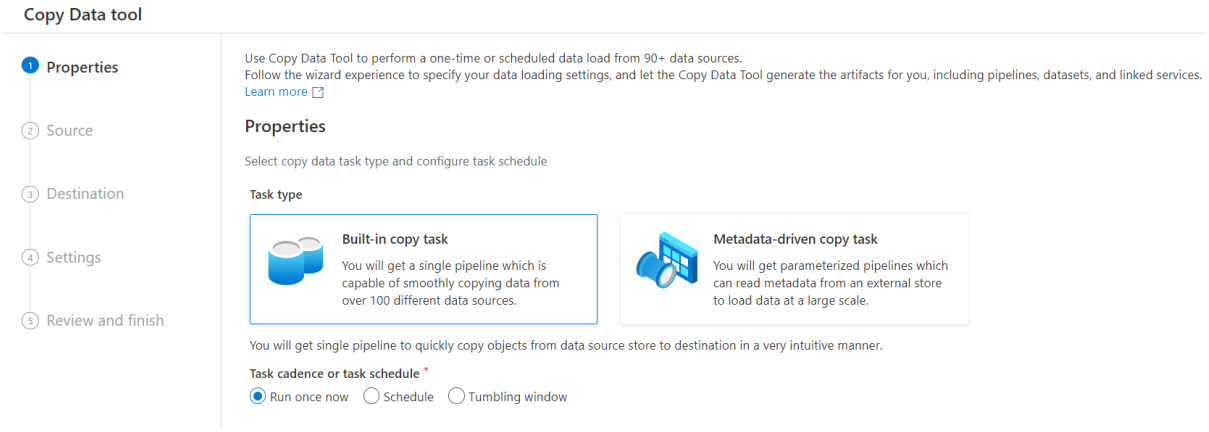
2. Also, we have updated our container and here you can see that we only two parquet files.



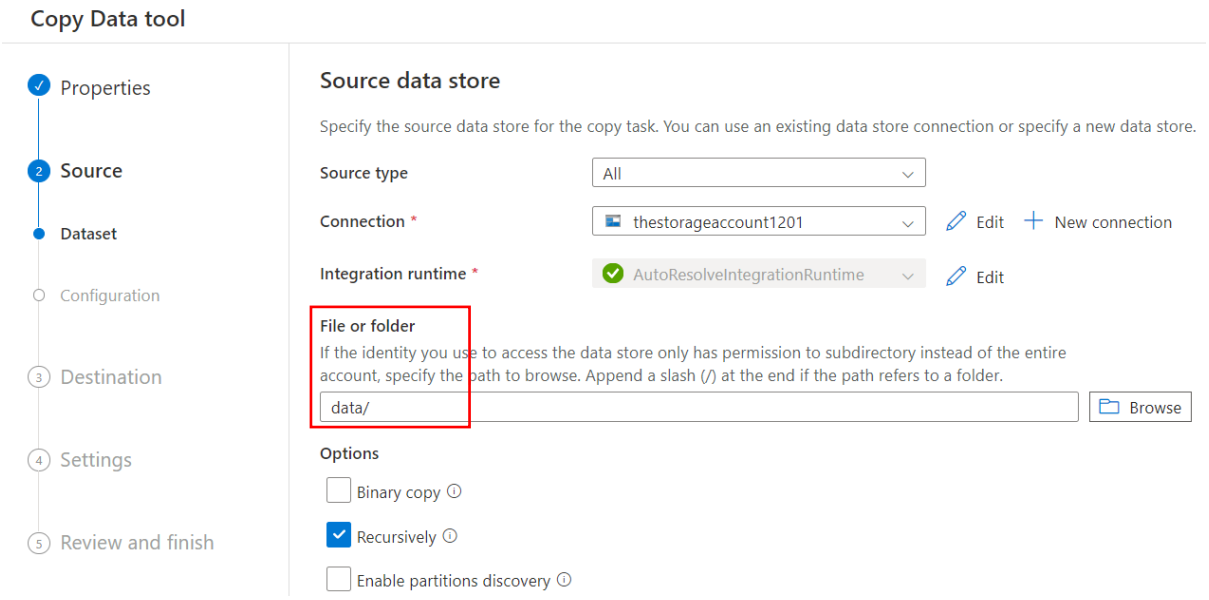
3. Now, we need to go to the integrate tab, click on the Plus icon, and choose the Copy Data tool.



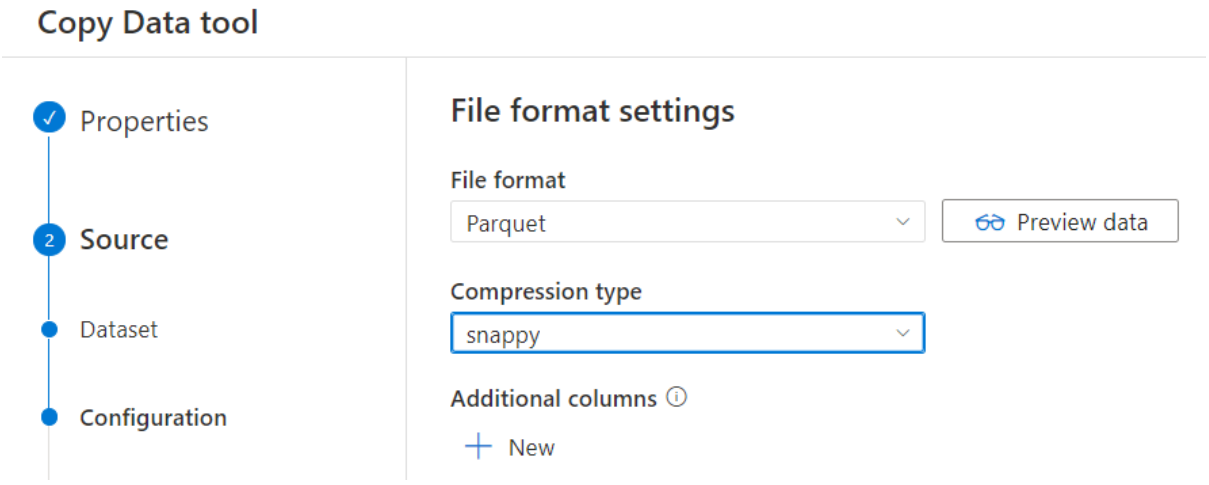
4. In our Copy data tool, we will first choose the built-in copy task, run once now, and click on next.



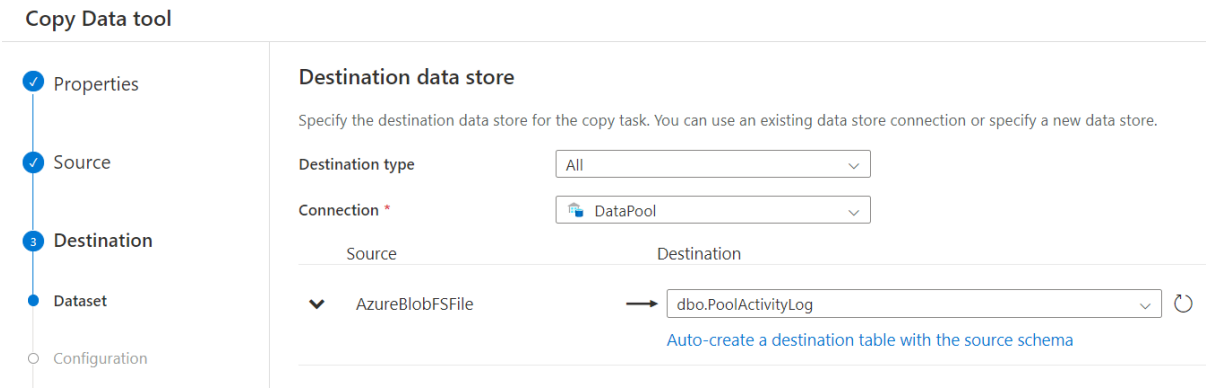
5. Our second step is to choose our source, which is our storage account or Data Lake here will choose our container as a source object and click on next.



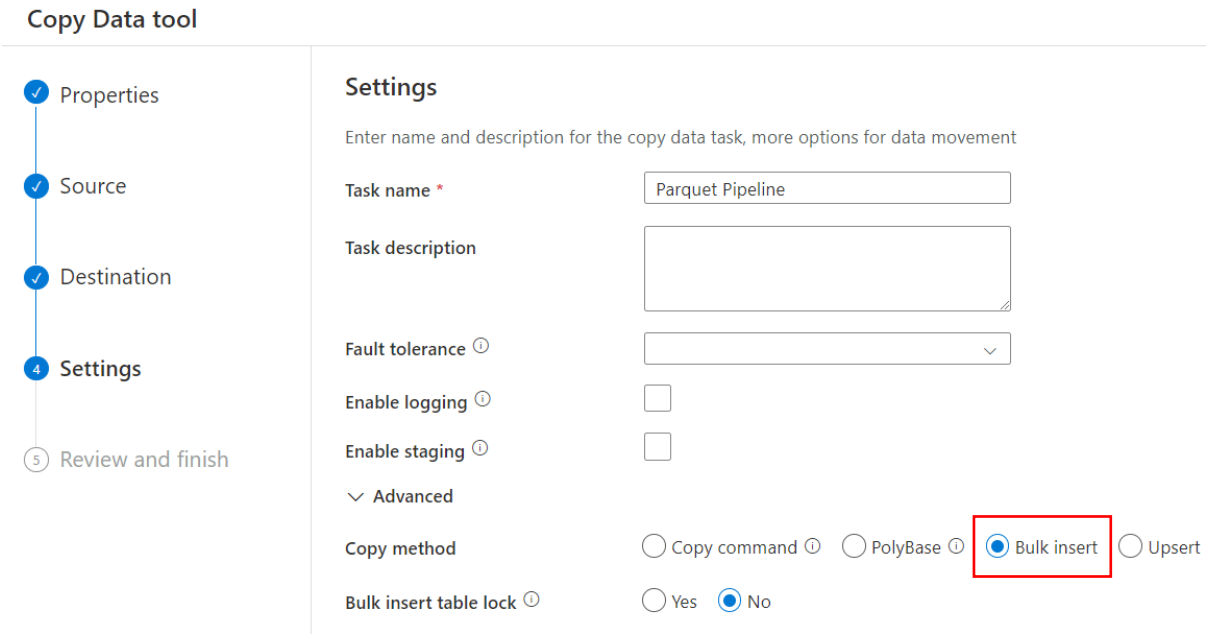
6. Then it will ask for the file format settings you have to choose Parquet and choose compression type to snappy. Click on next.



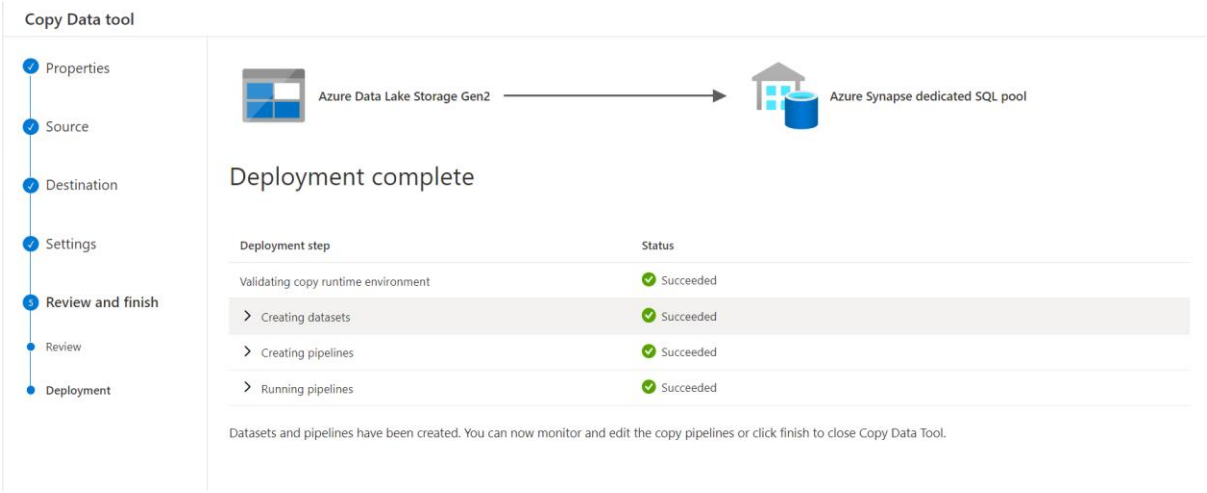
7. For the third step, which is the destination we will choose our Dedicated SQL Pool and then choose our existing table which is Pool Activity Log. Click on Next.



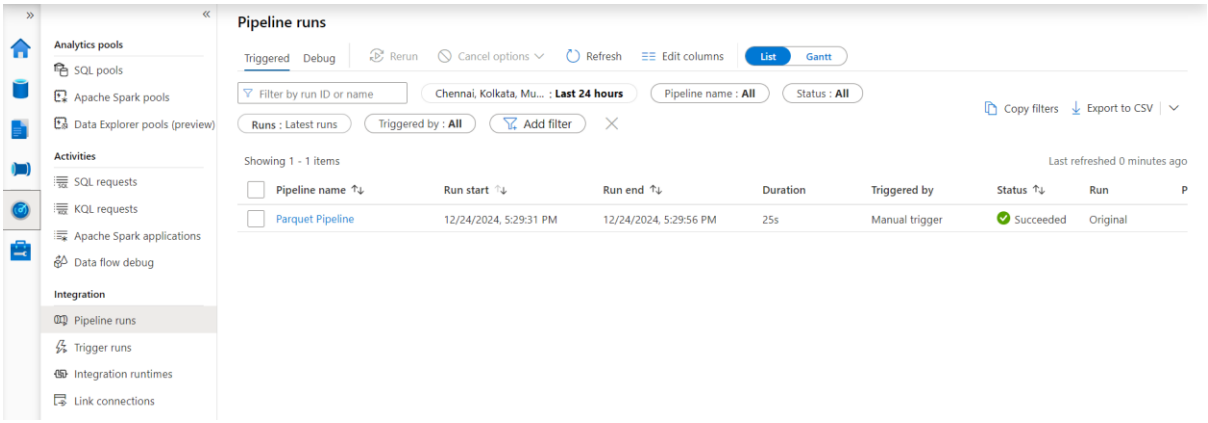
8. Then for our last step, we will give a name to the pipeline and choose Bulk insert for Copy method.



9. We will review the pipeline and then deploy it. Once it is done then we can just click on finish.



10. If we go to the monitor tab, we will see our pipeline whose status is succeeded.



11. Now if we go and run the Select command we will see the data inside the table.