Microsoft

Microsoft Partner Project Ready

Data Integration and Transformation with Data Factory in Microsoft Fabric

Day 01 of 02

Technofocus

Course Plan and Learning Objectives

Day 1

240 mins

Module 1 - Introduction to Data Factory in Microsoft Fabric

- Microsoft Fabric - The unified data platform for AI transformation
- Data Factory in Microsoft Fabric

Module 2 - Ingest data with Data Factory in Microsoft Fabric

- Data ingestion with Pipelines
- Copy activity with Pipelines
- Use parameters and expressions in Pipelines
- Data ingestion with Copy Job
- Mirroring databases in Microsoft Fabric
- Execute, monitor and troubleshoot Pipelines
- Data Pipeline storage event triggers
- REST API capabilities and CI/CD for Pipelines
- Fabric Pipelines vs. ADF/Synapse Pipelines

Hands-on labs

- Use case 01: Implementing Medallion Architecture with Data Factory in Microsoft Fabric for scalable data processing

Day 2

240 mins

Module 3 - Data Transformation with Dataflows Gen2

- Dataflows Gen2 in Microsoft Fabric
- Fast Copy in Dataflows Gen2
- Dataflow Gen2 with CI/CD and Git integration support
- Monitor your Dataflows
- Copilot for Data Factory in Microsoft Fabric
- Workflow Orchestration with Apache Airflow job

Module 4 - Migrate to Data Factory in Microsoft Fabric

- Plan your migration from Azure Data Factory to Data Factory in Microsoft Fabric
- Migrate from Dataflow Gen1 to Dataflow Gen2
- Move queries from Dataflow Gen1 to Dataflow Gen2
- Ingest data into Microsoft Fabric using the Azure Data Factory Copy Activity

Hands-on labs

- Use Case 02: Data Factory solution for moving and transforming data with dataflows and data pipelines

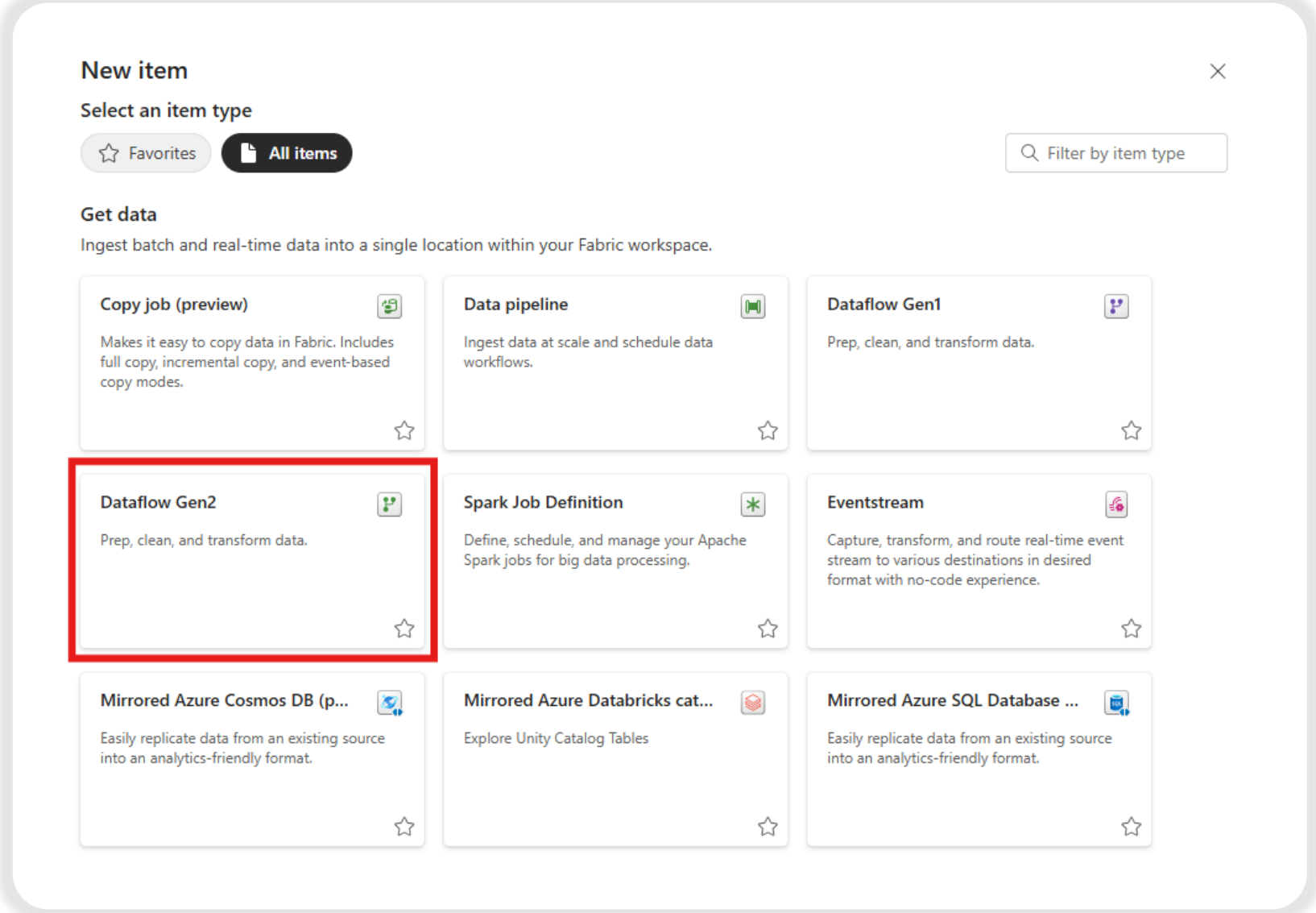
03
Data Transformation with Dataflows Gen2

Medallion Architecture in Microsoft Fabric Data Factory

Dataflows Gen2 in Microsoft Fabric

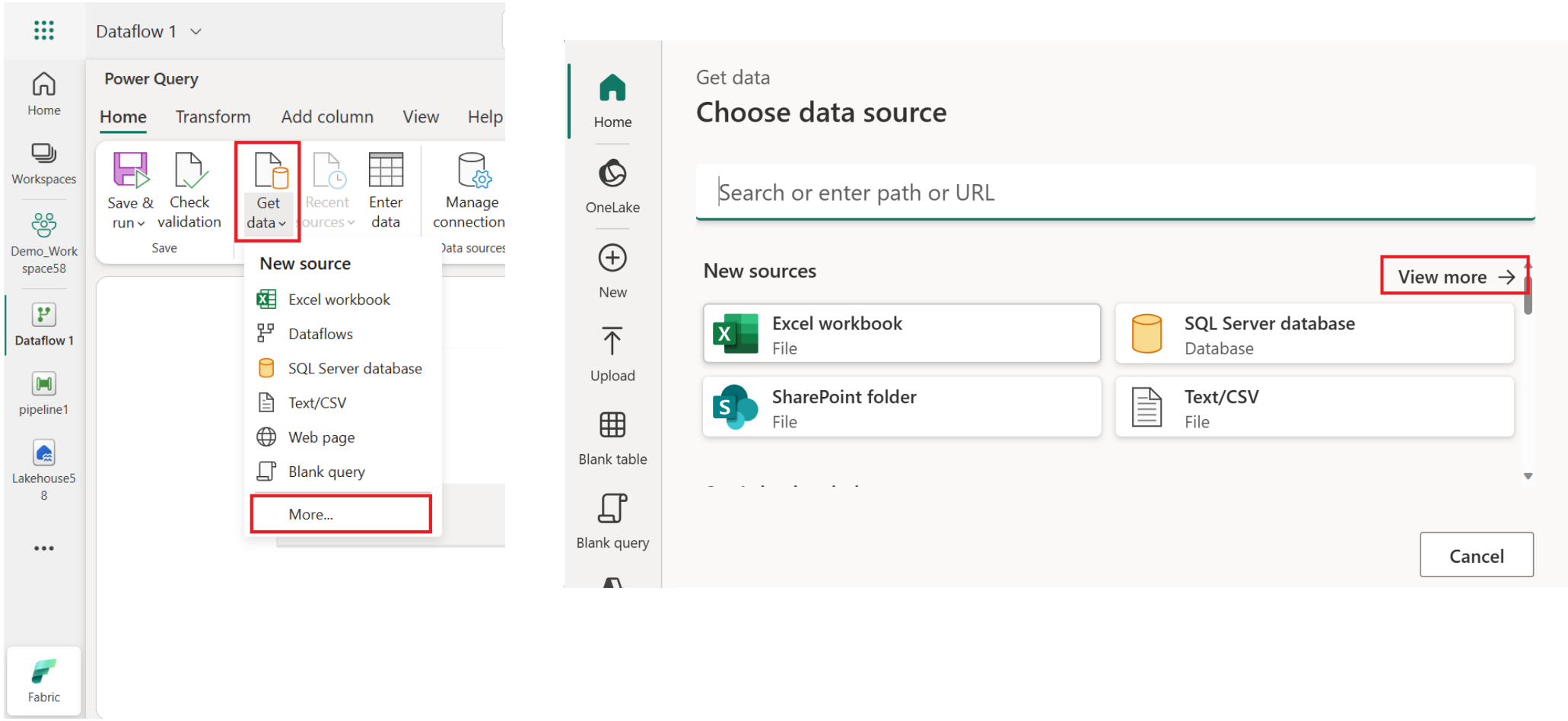
Create a dataflow

- Create a dataflow from your Microsoft Fabric workspace
- Get data
- Apply transformations
- Configure destination settings for transformed data



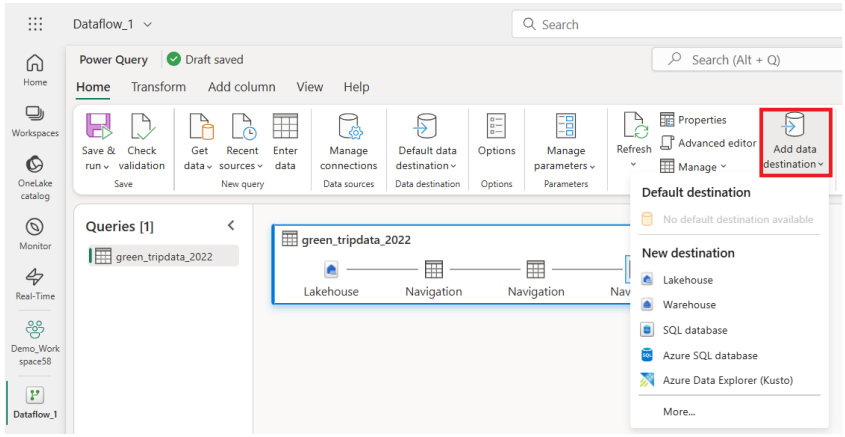
Get data for your dataflow

Dataflows (Gen2) connect to various data sources and perform transformations in Power Query Online

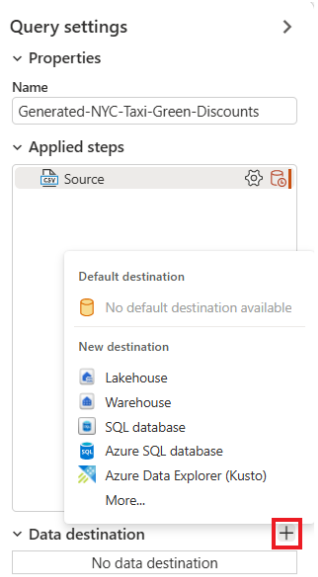


Entry points to specify the data destination

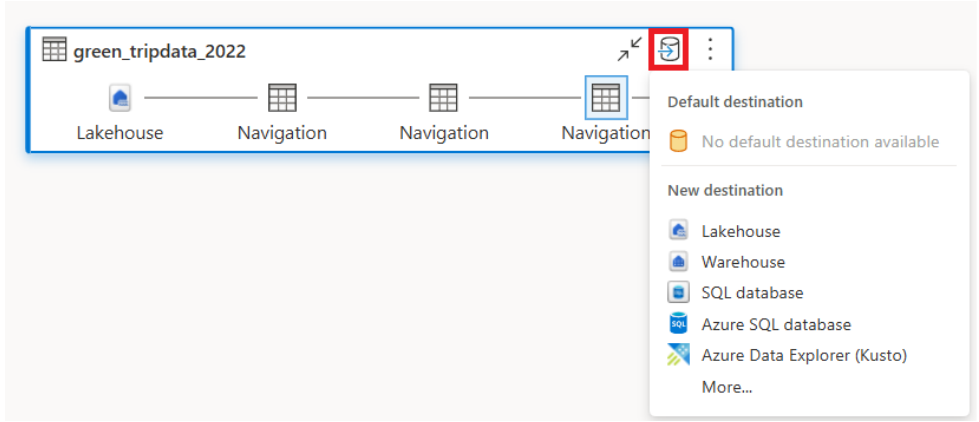
To specify the data destination



Through the top ribbon



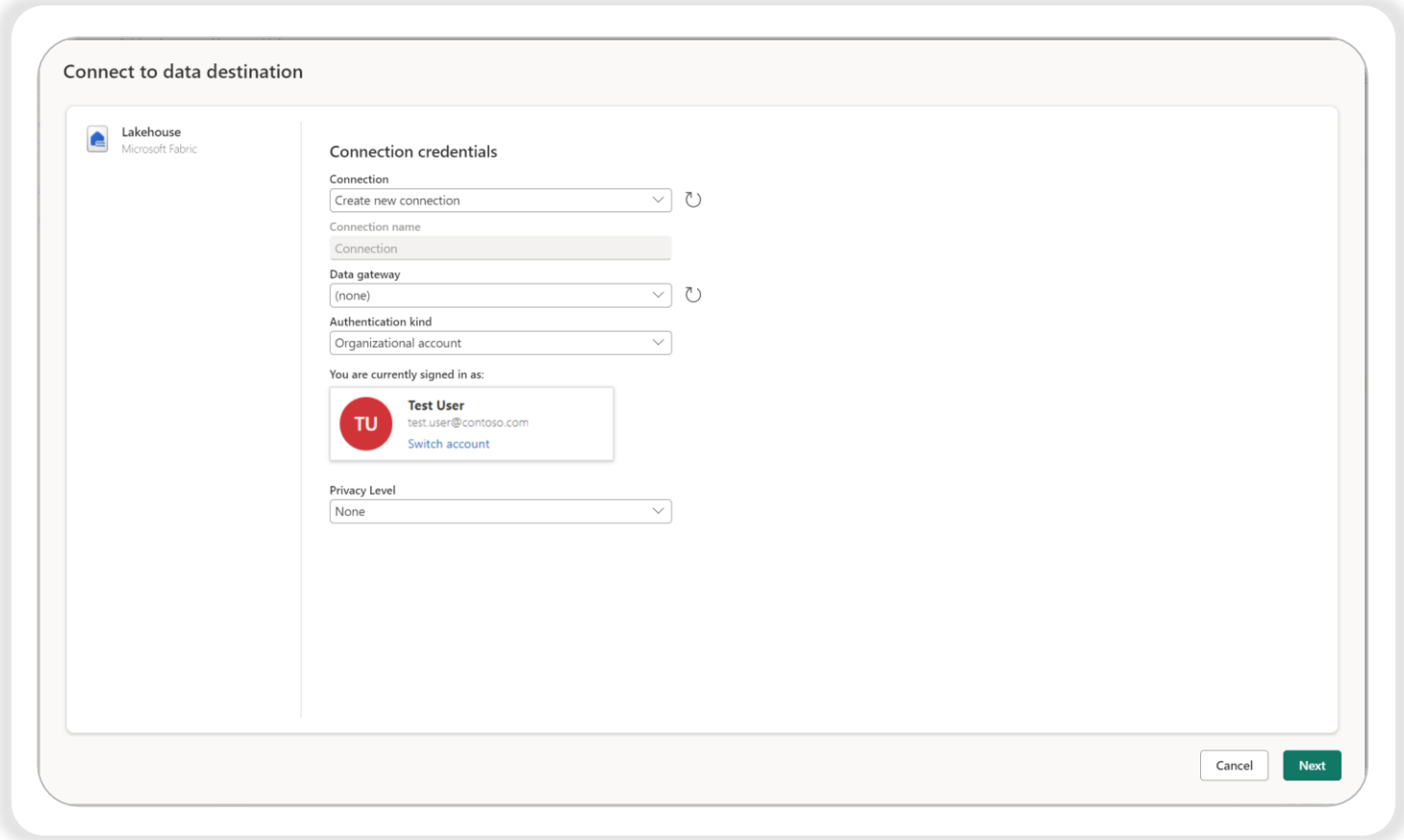
Through query settings



Through the diagram view

Connect to the data destination

- Connecting to the data destination is similar to connecting to a data source
- Connections can be used for both reading and writing your data



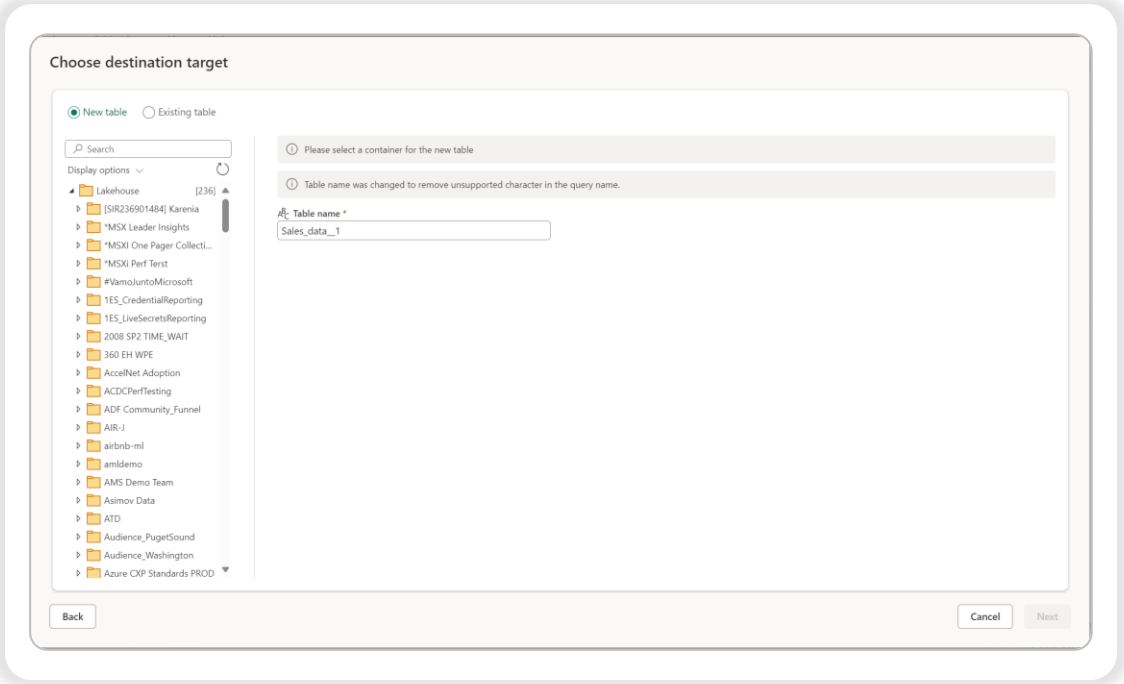
Load into your data destination

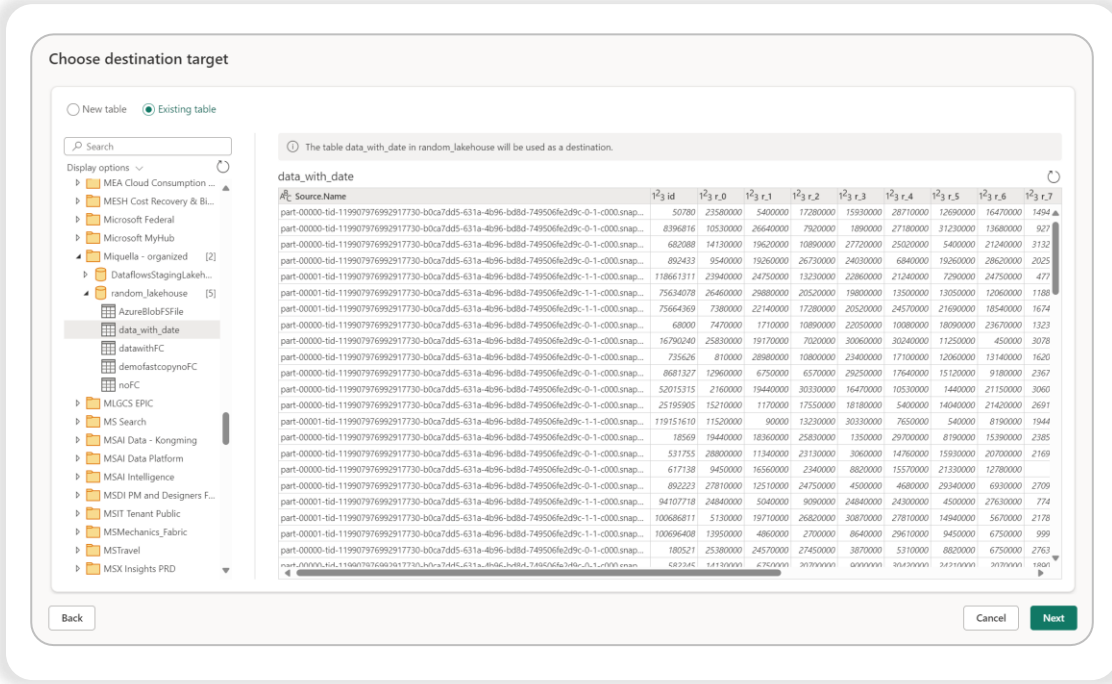
- **Create a new table**

  New table is created in your data destination

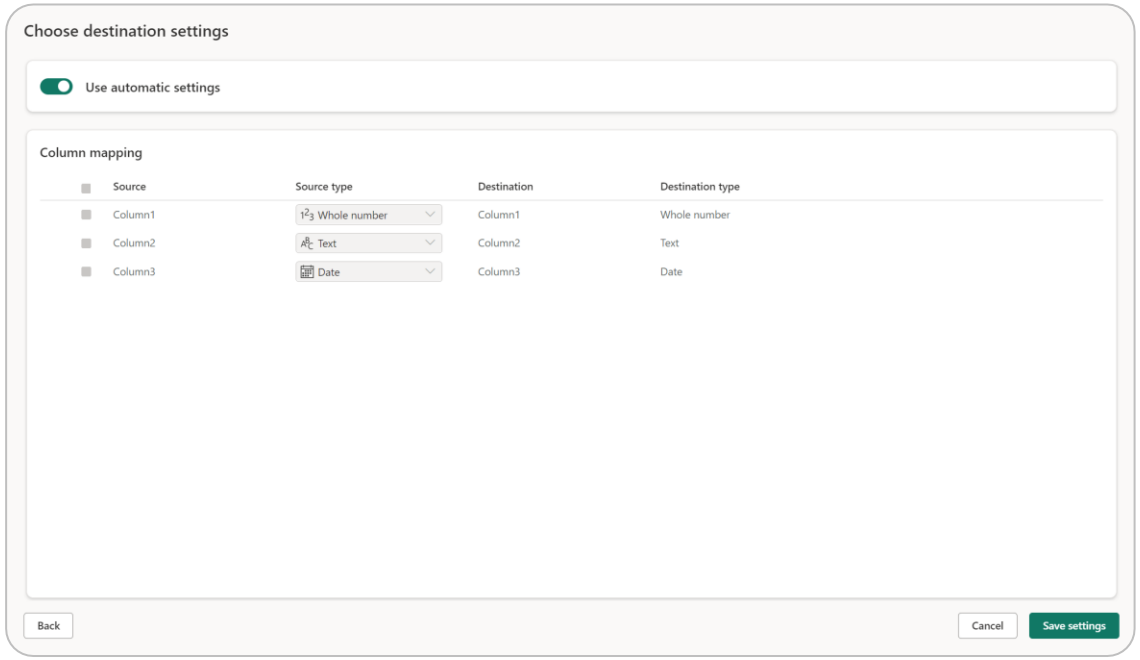
- **Pick an existing table**

  Table can't be recreated in any scenario

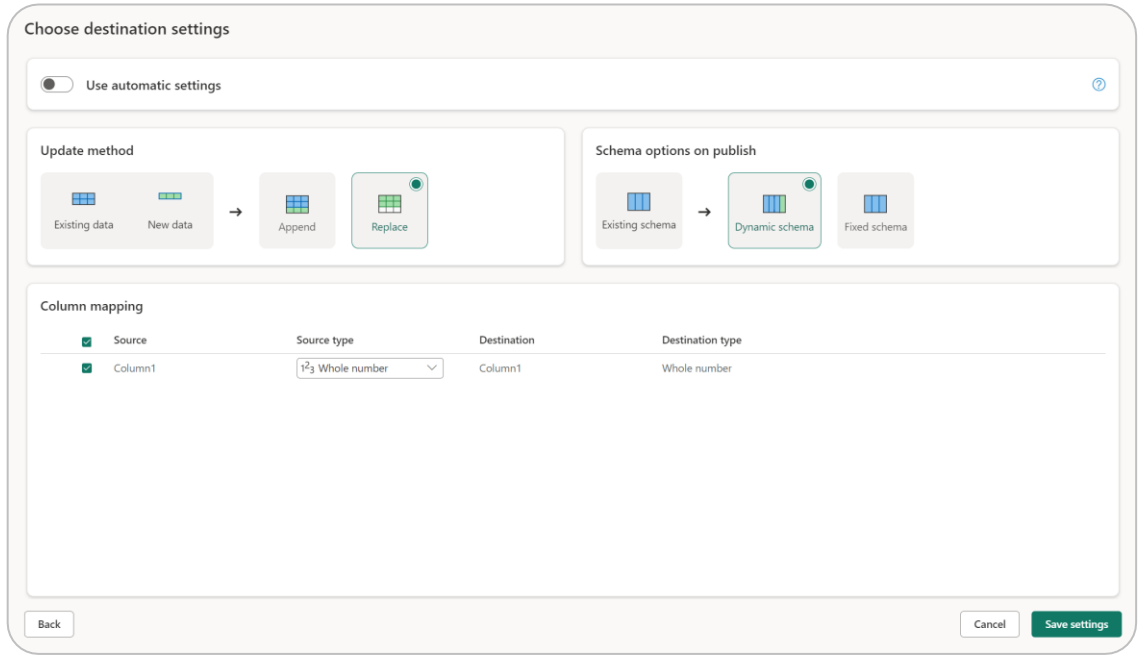




Settings for new tables



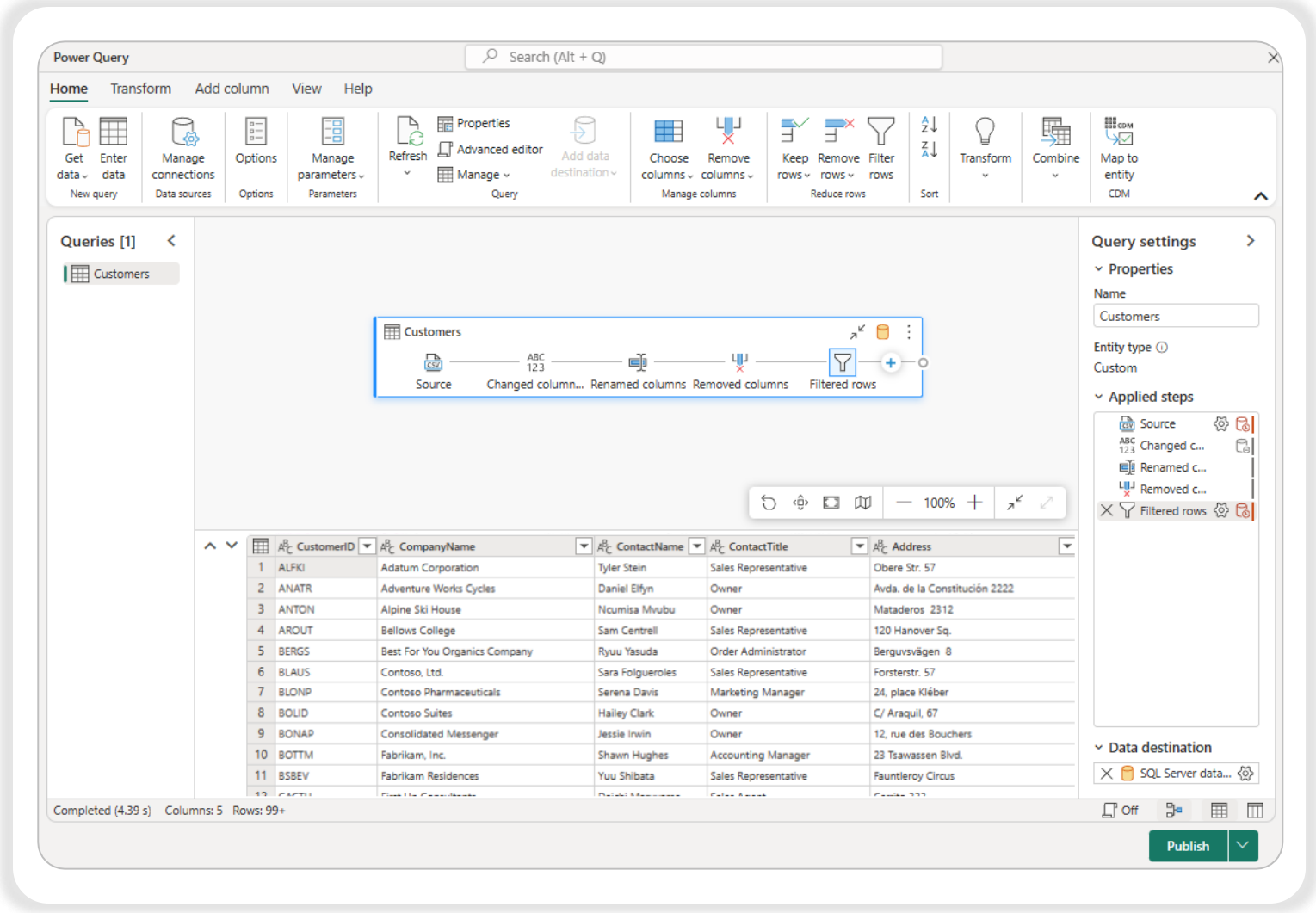
The automatic settings
Update method replace
Managed mapping
Drop and recreate table



Manual settings
Get full control over how to load your data
Update methods: Replace and Append
Schema options on publish

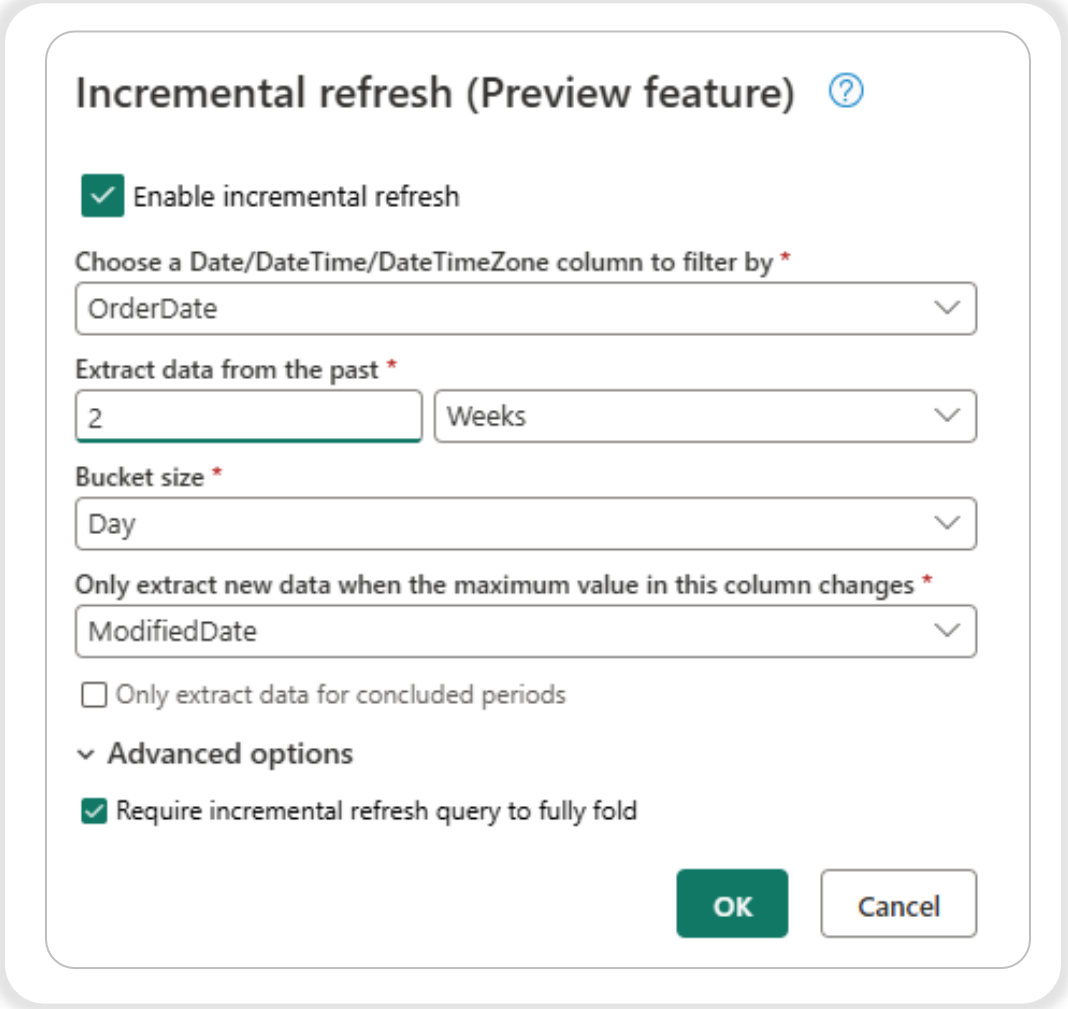
Dataflow refresh

- On-demand
- By setting up a refresh schedule



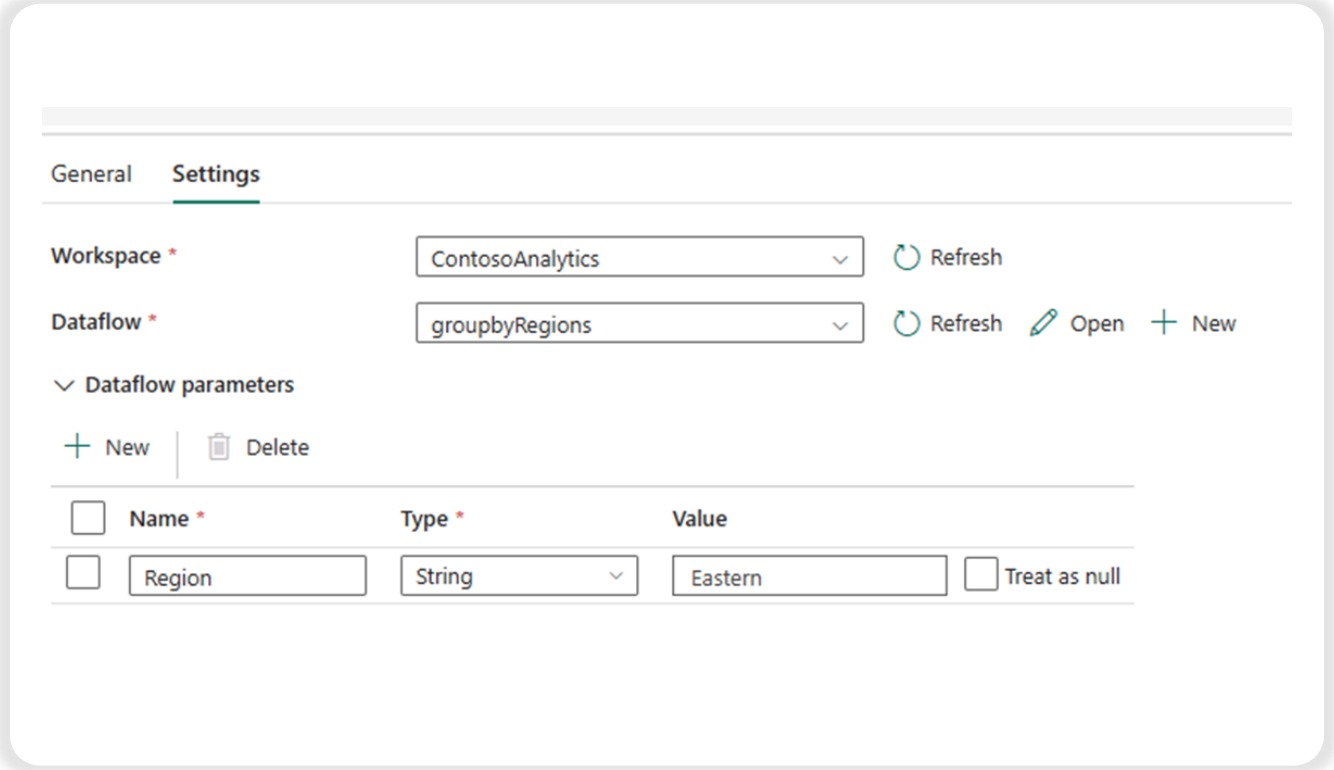
Incremental refresh in Dataflow Gen2 (Preview)

- Refresh only new or updated data
- Allows you to
  - Reduce refresh times
  - Enhance reliability
  - Minimize resource usage
- Supported data destinations supported
  - Fabric Warehouse
  - Azure SQL Database
  - Azure Synapse Analytics



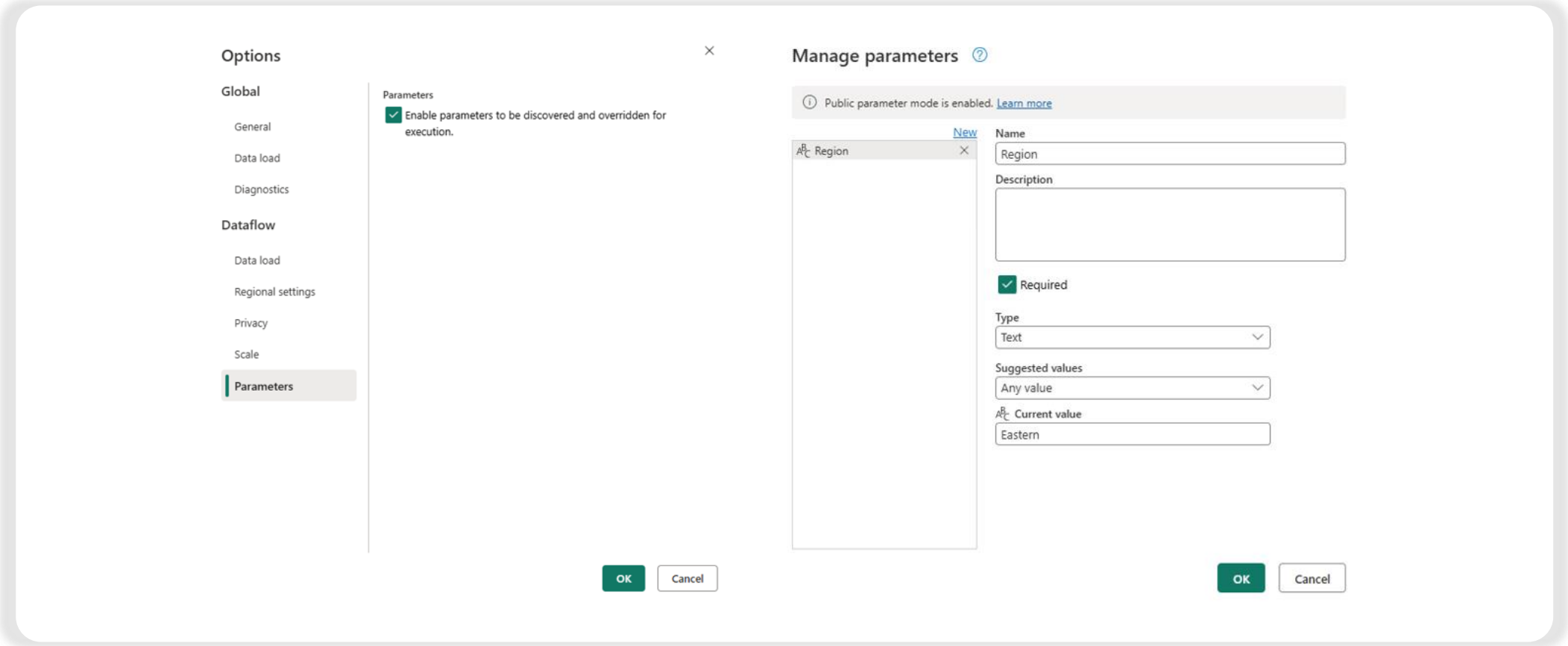
Use public parameters in Dataflow Gen2 (Preview)

- Dynamically control and customize dataflows
- Refresh dataflow by passing parameter values outside of the Power Query editor
- Pass parameters through the Fabric REST API or through native Fabric experiences



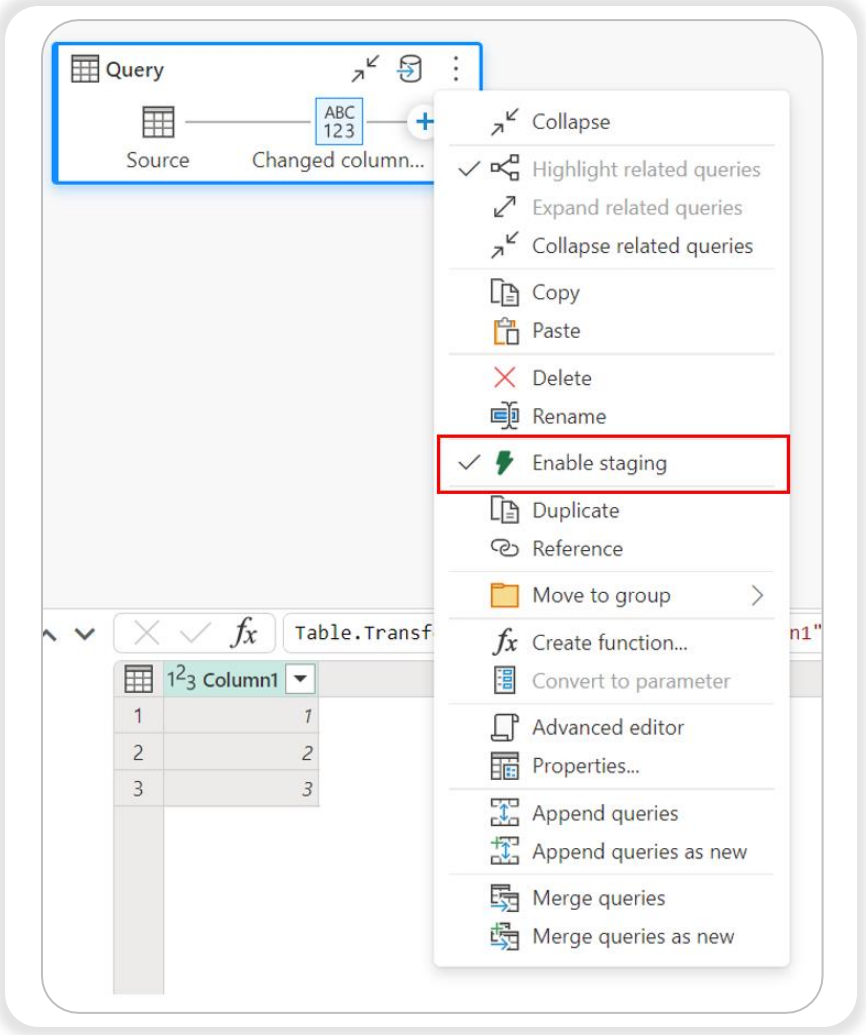
Pass custom parameter values for refresh

- Ensure the public parameter mode is enabled for the dataflow item
- Use the updated Dataflow refresh activity to pass parameters via Data Pipelines



Dataflow Gen2 staging artifacts

- DataflowsStagingLakehouse
- DataflowsStagingWarehouse
- Use the dataflow connector in Power BI, Excel, or other dataflows, if you need to access data

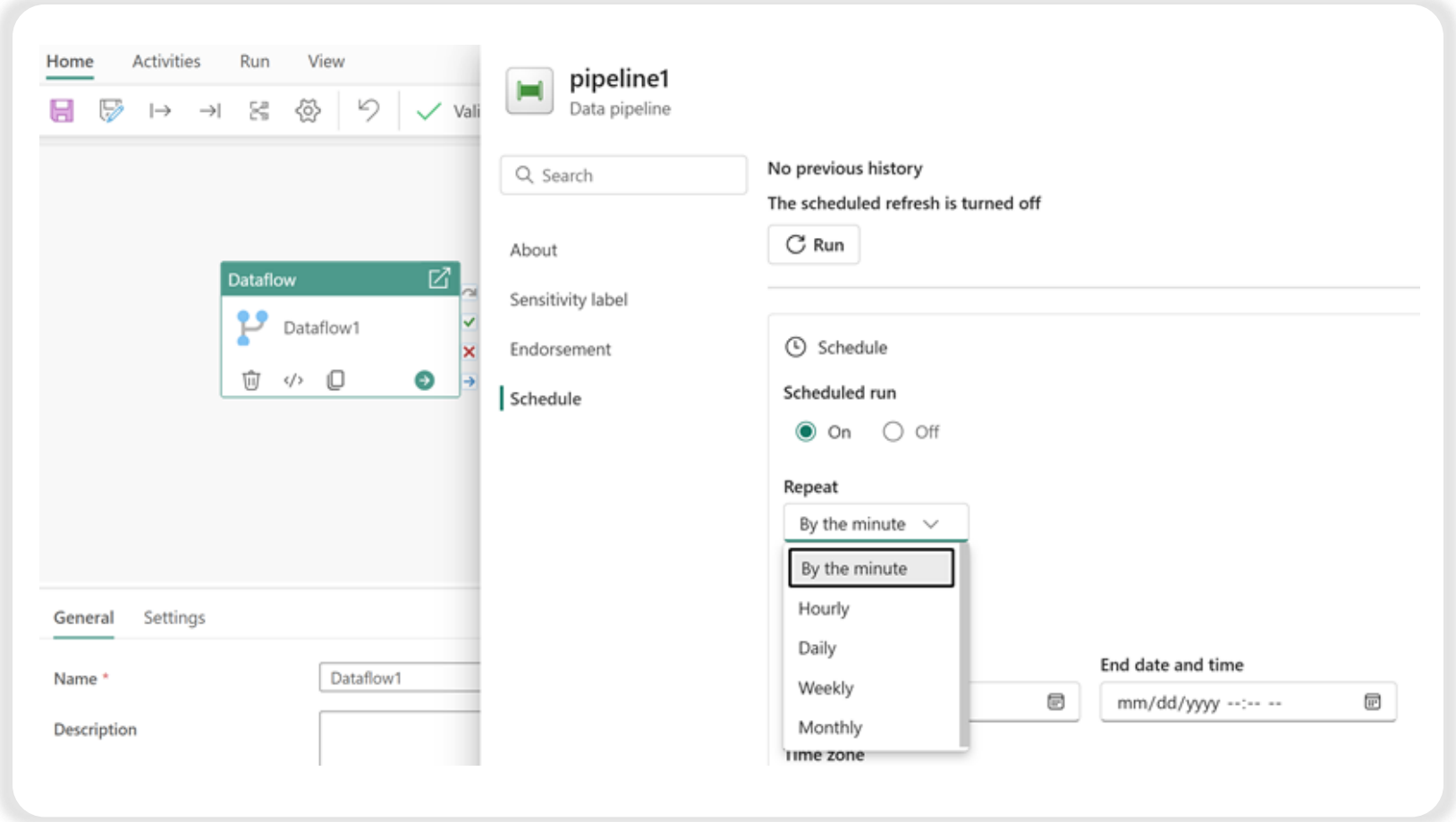


Decision Guide: Copy Activity vs Dataflow vs Spark

	Pipeline Copy Activity	Dataflow Gen 2	Spark Notebook Job Definition
Use case	Data lake and data warehouse migration, data ingestion Light weight transformation	Data transformation, data wrangling, data profiling Data ingestion	Data transformation, data processing, data profiling, data profiling Data ingestion
Primary developer persona	Data engineer, data integrator	Data engineer, data integrator, business analyst	Data engineer data scientist, data developer
Primary developer skillset	ETL, SQL, JSON	ETL, M, SQL	Spark (Scala, PySpark, Spark SQL, R)
Code written	No code Low code	No code Low code	Code
Data volume	Low to High	Low to High	High
Development interface	Wizard Canvas	Power Query	Notebook Spark job definition
Sources	30+ connectors	150+ connectors	Hundreds of Spark libraries
Destinations	18+ connectors	Lakehouse Azure SQL database Azure Data explorer Azure Synapse analytics	Hundreds of Spark libraries
Transformation complexity	Low: Lightweight - type conversion, column mapping, merge/split files, flatten hierarchy	Low to high: 300+ transformation functions	Low to high: Support for native Spark and open-source libraries

Integrate Dataflows Gen2 and Pipelines in Microsoft Fabric

- Useful when you need to perform additional operations on the transformed data
- Some common activities in Data engineering and Data Factory:
  - Copy data
  - Incorporate Dataflow
  - Add Notebook
  - Get metadata
  - Execute a script or stored procedure



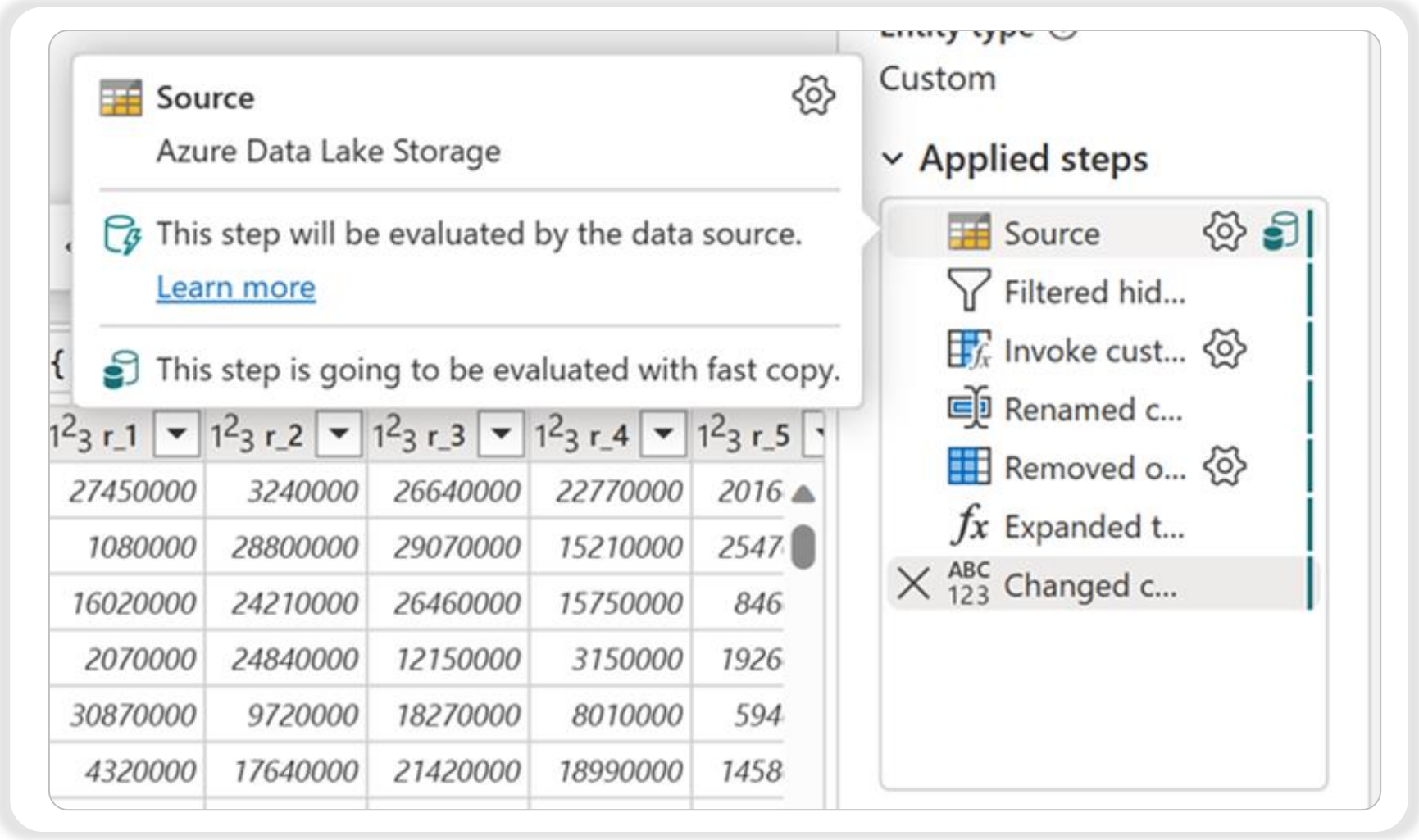
Demo

Use a dataflow in a pipeline

Fast Copy in Dataflows Gen2

Introduction to Fast copy

Ingest terabytes of data with the easy experience of dataflows



Fast copy | Supported Dataflow Gen2 connectors

- ADLS Gen2

- Blob storage

- Azure SQL DB

- Lakehouse

- PostgreSQL

- On premise SQL Server

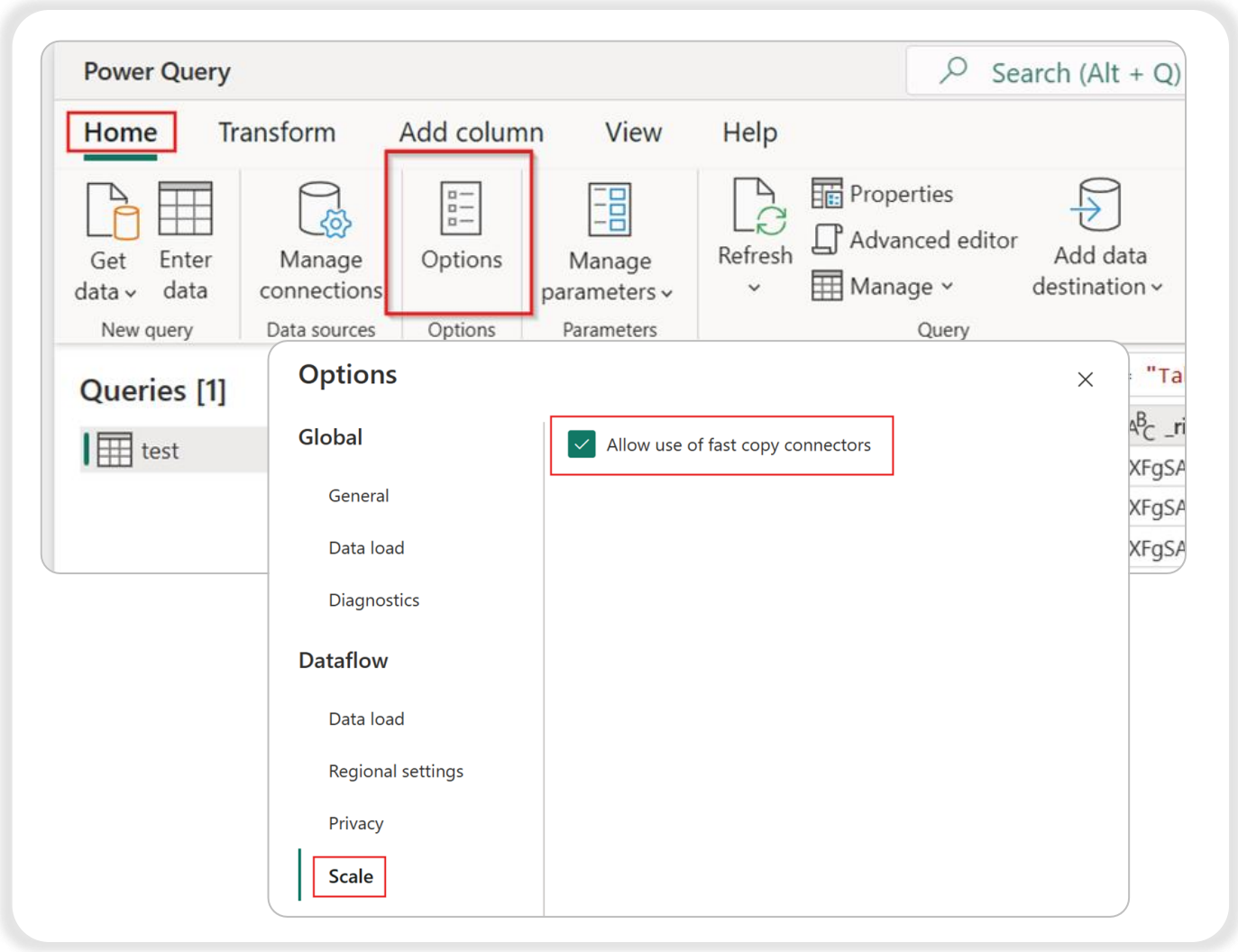
- Warehouse

- Oracle

- Snowflake

Use fast copy

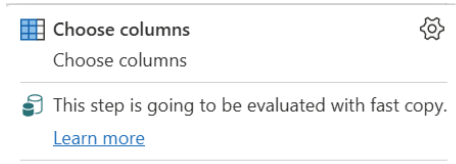
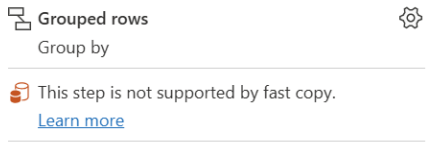
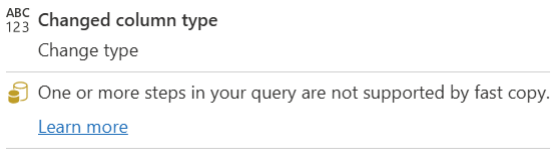
- Navigate to the appropriate Fabric endpoint then to a premium workspace
- Create a dataflow Gen2
- On the **Home** tab of the new dataflow, select **Options**
- Select **Scale** and allow use of fast copy connectors



Split your query to leverage fast copy

Part 1 - Data ingestion to staging

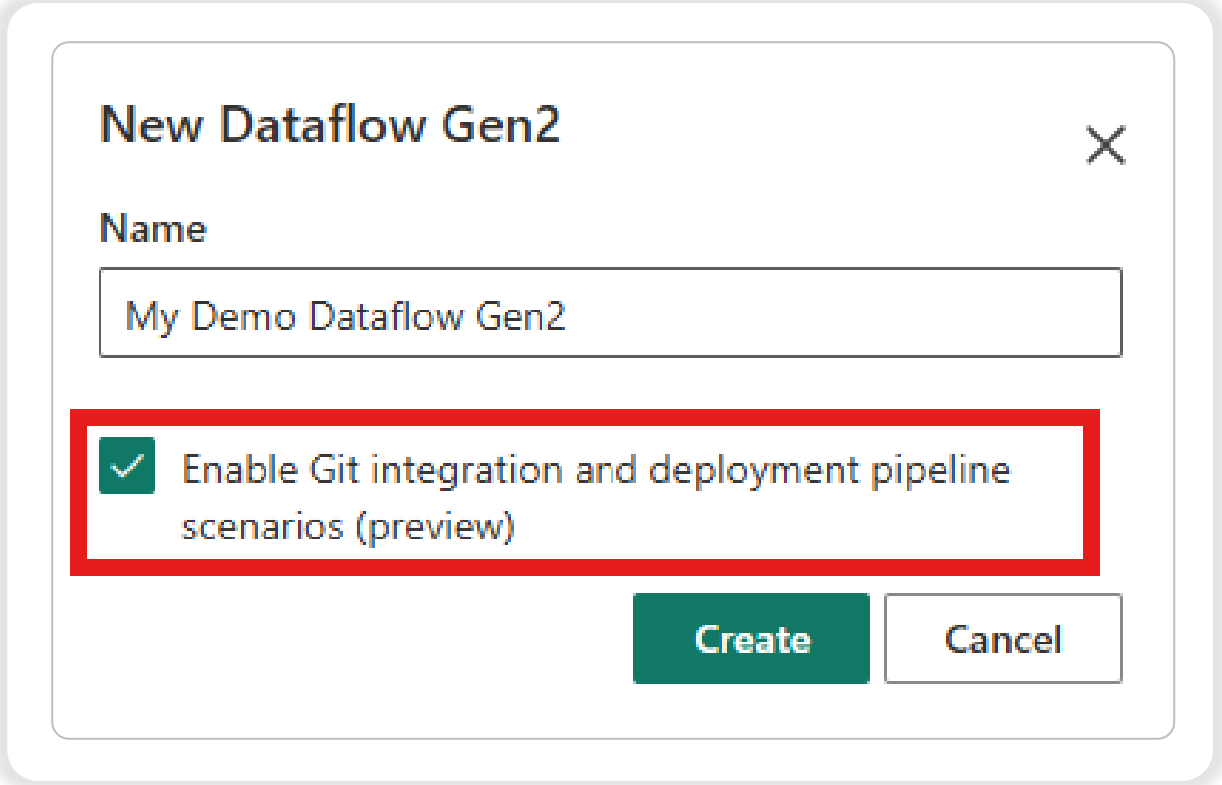
Part 2 - Large-scale transformation with SQL DW compute

Indicator	Icon	Description
This step is going to be evaluated with fast copy		The query up to this step supports fast copy
This step is not supported by fast copy		This step doesn't support Fast Copy
One or more steps in your query are not supported by fast query		Some steps in this query support Fast Copy, while others don't

Dataflow Gen2 with CI/CD and Git integration support

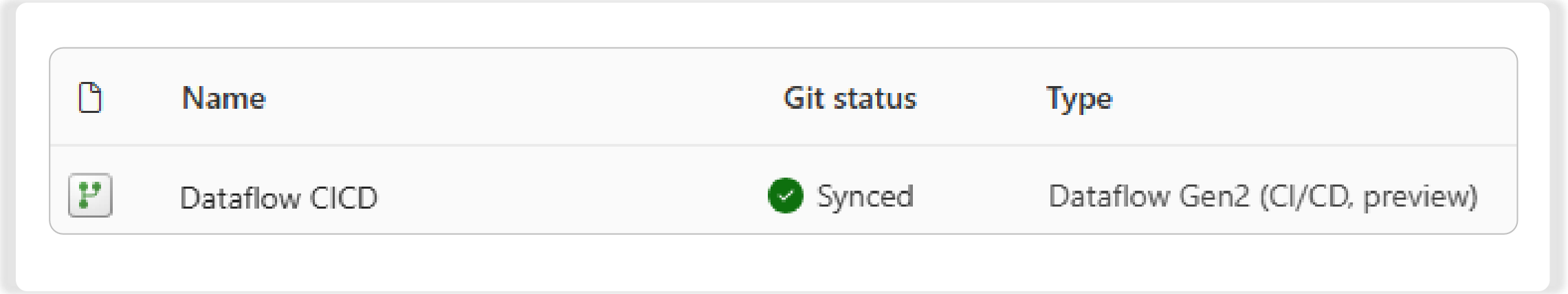
What you can do with Dataflow Gen2 (CI/CD preview)

- Create, edit, and manage dataflows in a Git repository connected to your fabric workspace
- Automate the deployment of dataflows from your workspace to other workspaces
- Use the Fabric settings and scheduler to refresh and edit settings for Dataflow Gen2



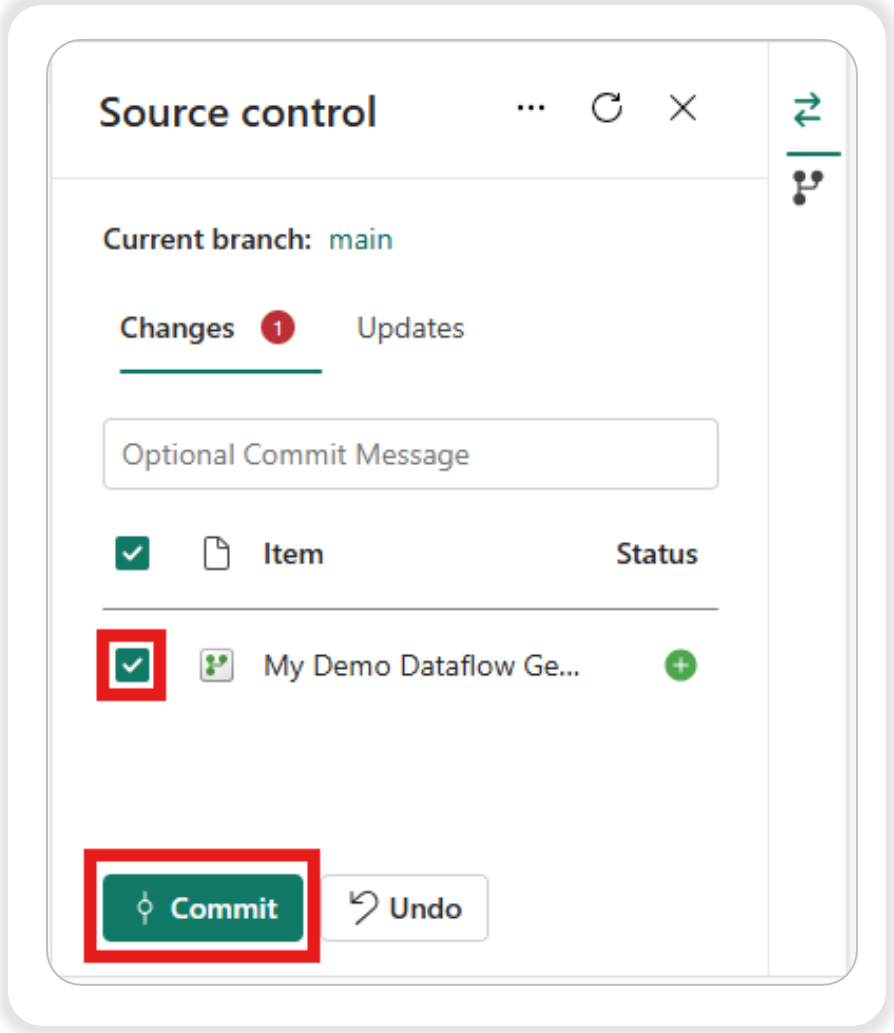
Benefits of integration

- Automated Deployments
- Version Control
- Collaboration



Set up Dataflow Gen2 with CI/CD and GIT integration

- Enable GIT integration in your workspace settings before proceeding
- When creating a new Dataflow Gen2, there is now an option to enable CI/CD and GIT integration
- Once enabled, you will be using the new Dataflow Gen2 with CI/CD and GIT integration

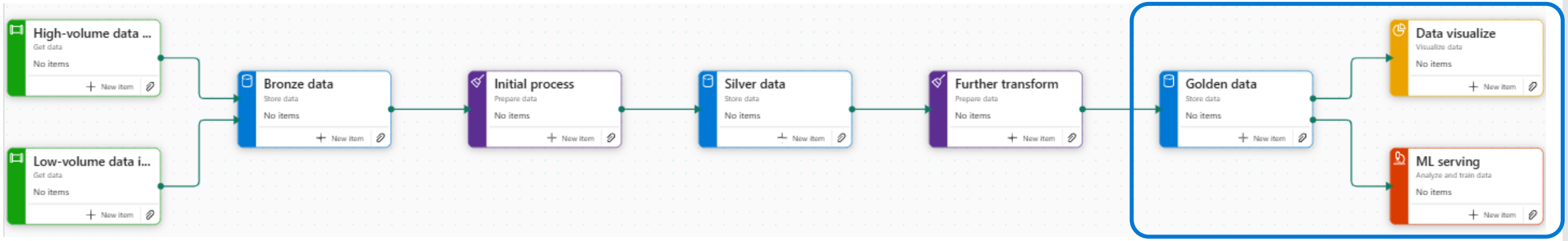


Steps to move existing dataflows over to the new Dataflows Gen2 with CI/CD and GIT integration

- Open the Dataflow Gen2 you would like to migrate
- Copy the queries or export a PQT template of the Dataflow Gen2
- Create a new Dataflow Gen2 with CI/CD and GIT integration enabled
- Import the queries or PQT template into the new Dataflow Gen2

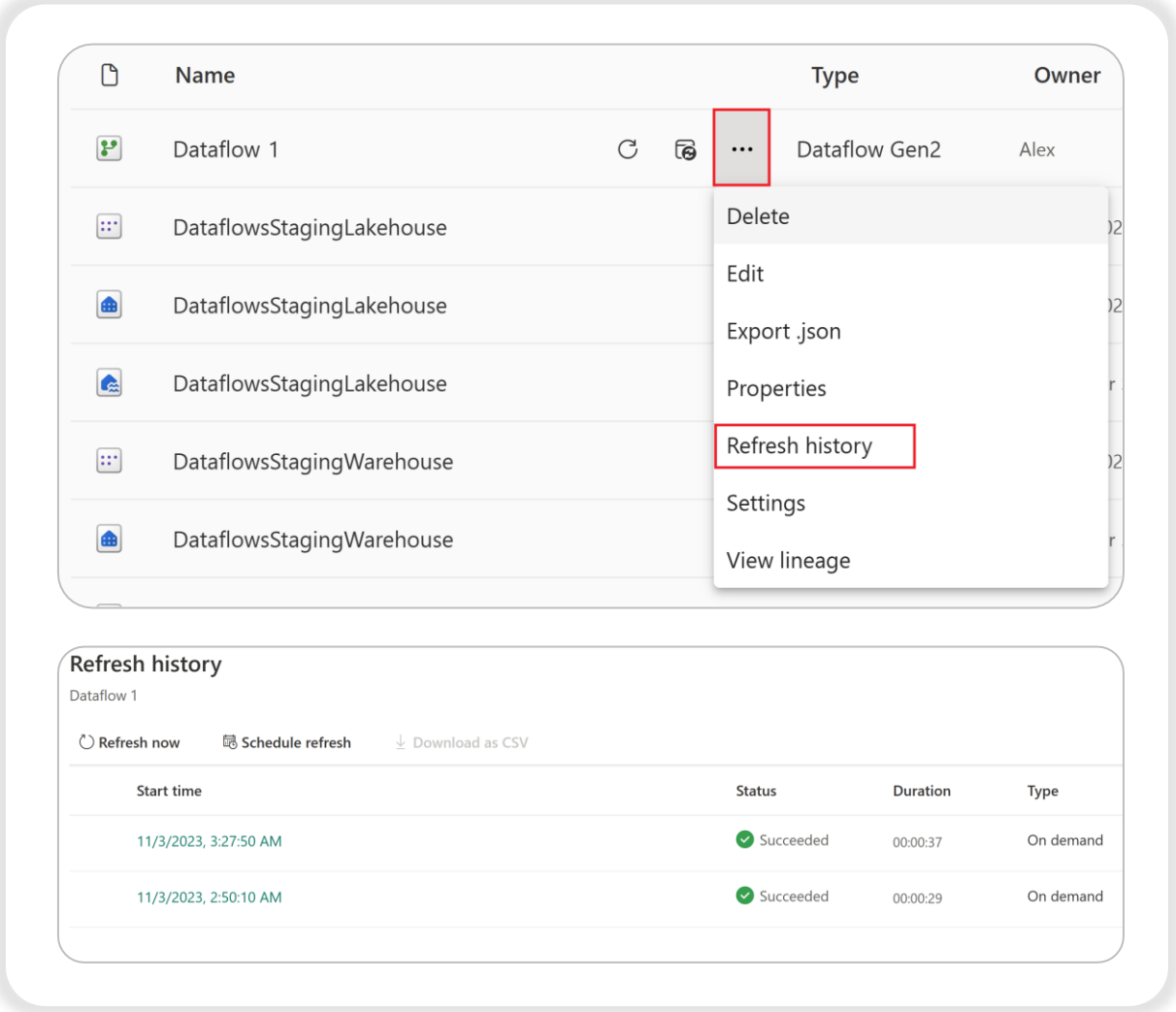
Monitor your Dataflows

Medallion Architecture in Microsoft Fabric Data Factory



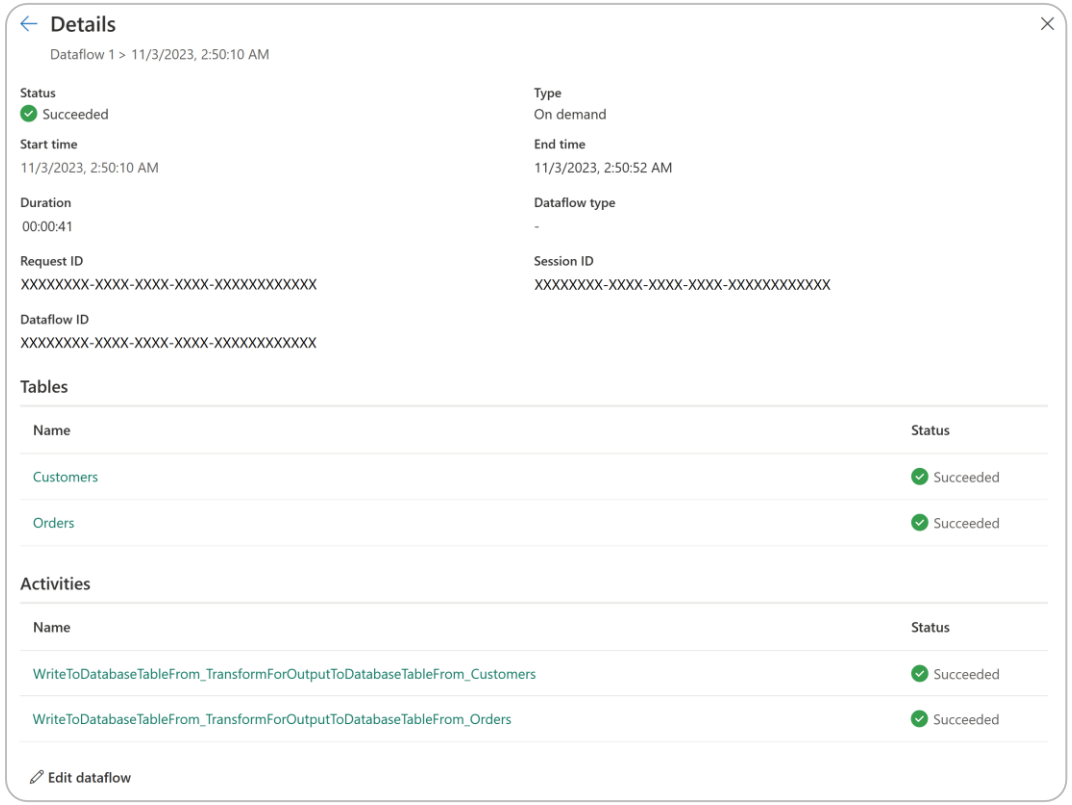
Refresh history

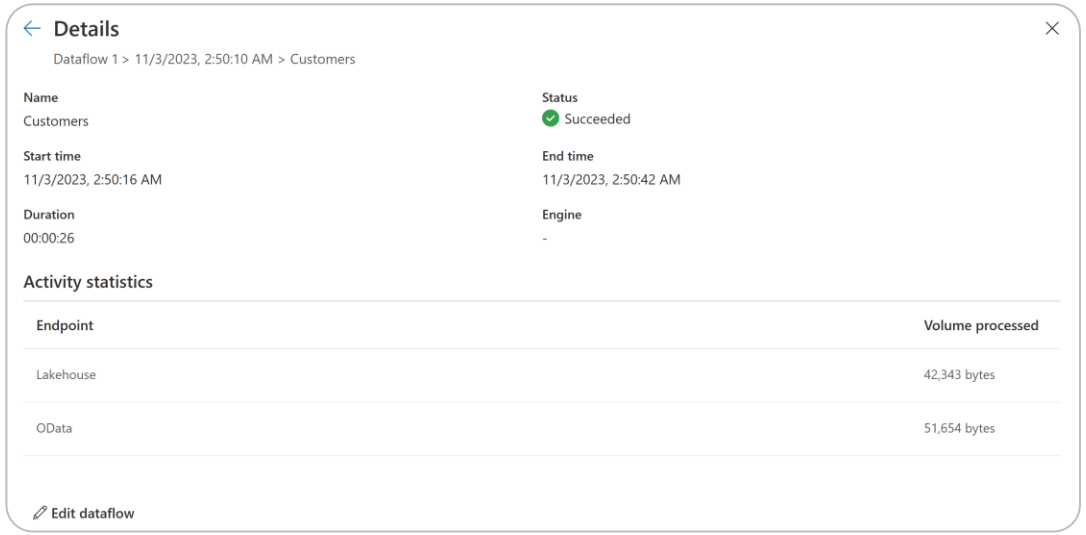
- Provides list of all your data refreshes
- Shows you up to 50 refresh histories or up to 6 months back
- Stores up to 250 refresh histories or 6 months back in the onelake



Reviewing your dataflow refresh from the UI

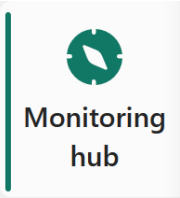
Drill down into one of the refreshes

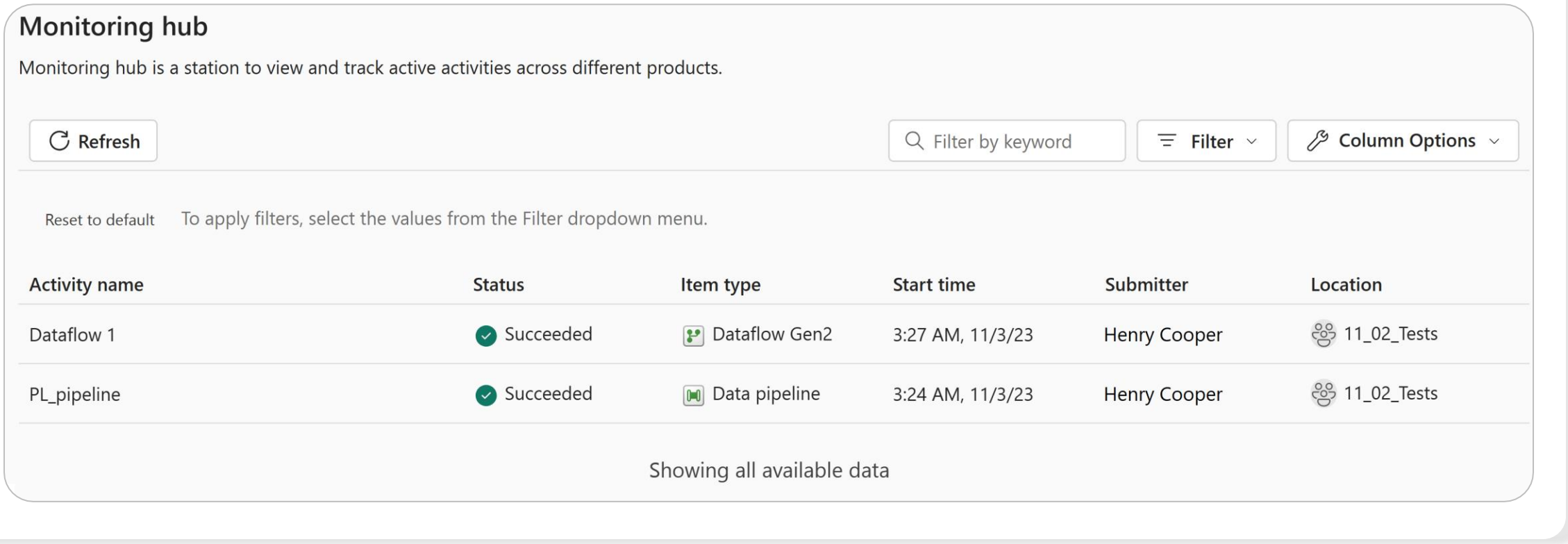




Monitoring hub

Provides you with a dashboard that gives you an overview of the status of your dataflows





Copilot for Data Factory in Microsoft Fabric

Data integration + AI

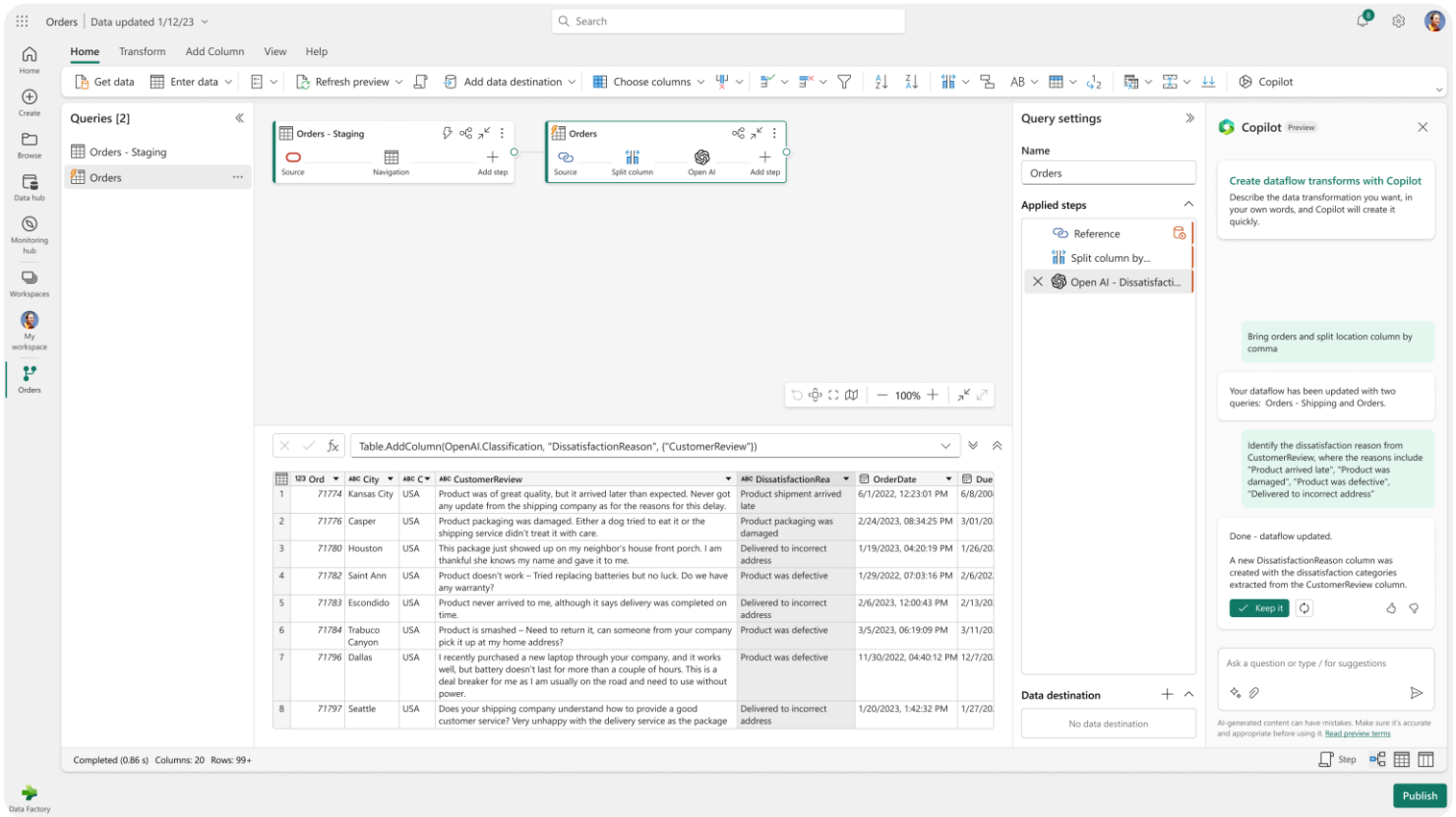
Copilot in Data Factory

- **Easily integrate generative AI into your dataflows and pipelines using Copilot**

Chat with **Copilot** to describe data transformations in natural language

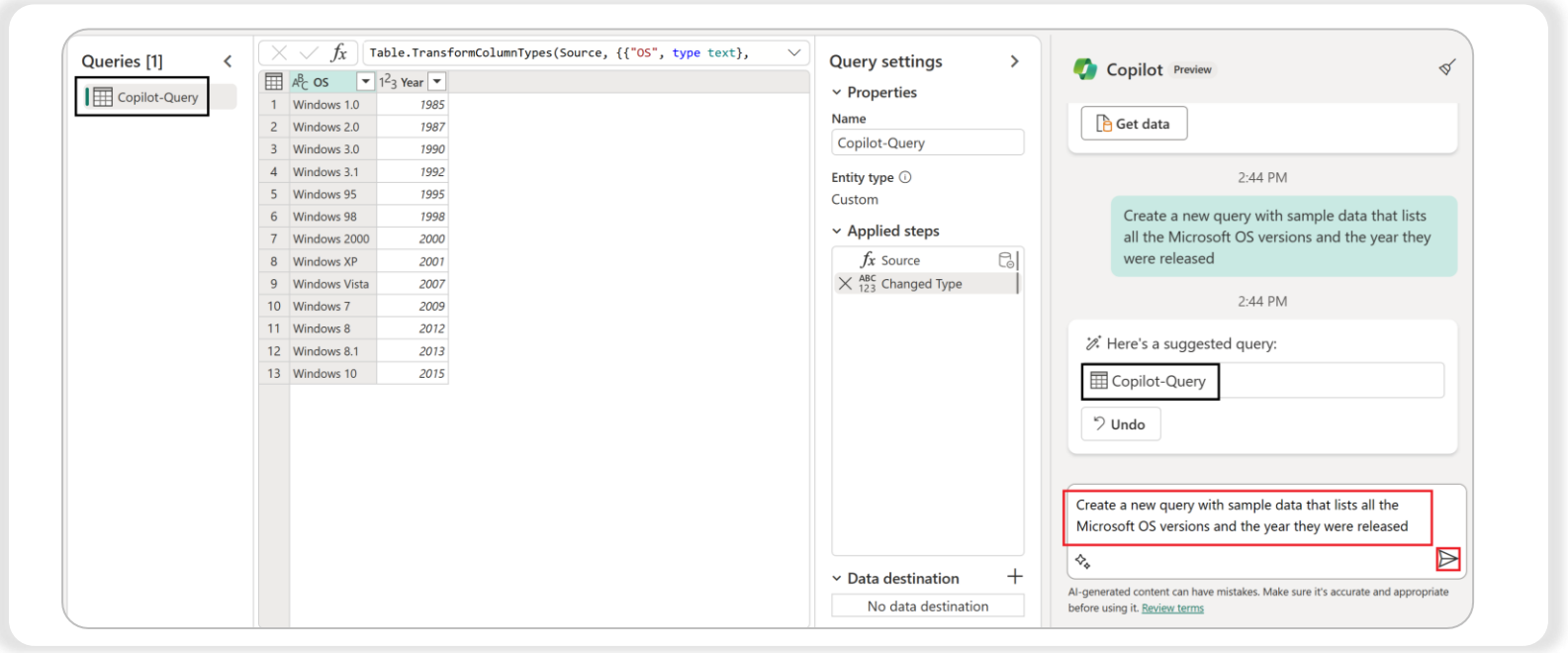
Tap into generative AI capabilities from **Azure Open AI** as data transformation steps

Use **Copilot** to schedule and run and manage dataflows



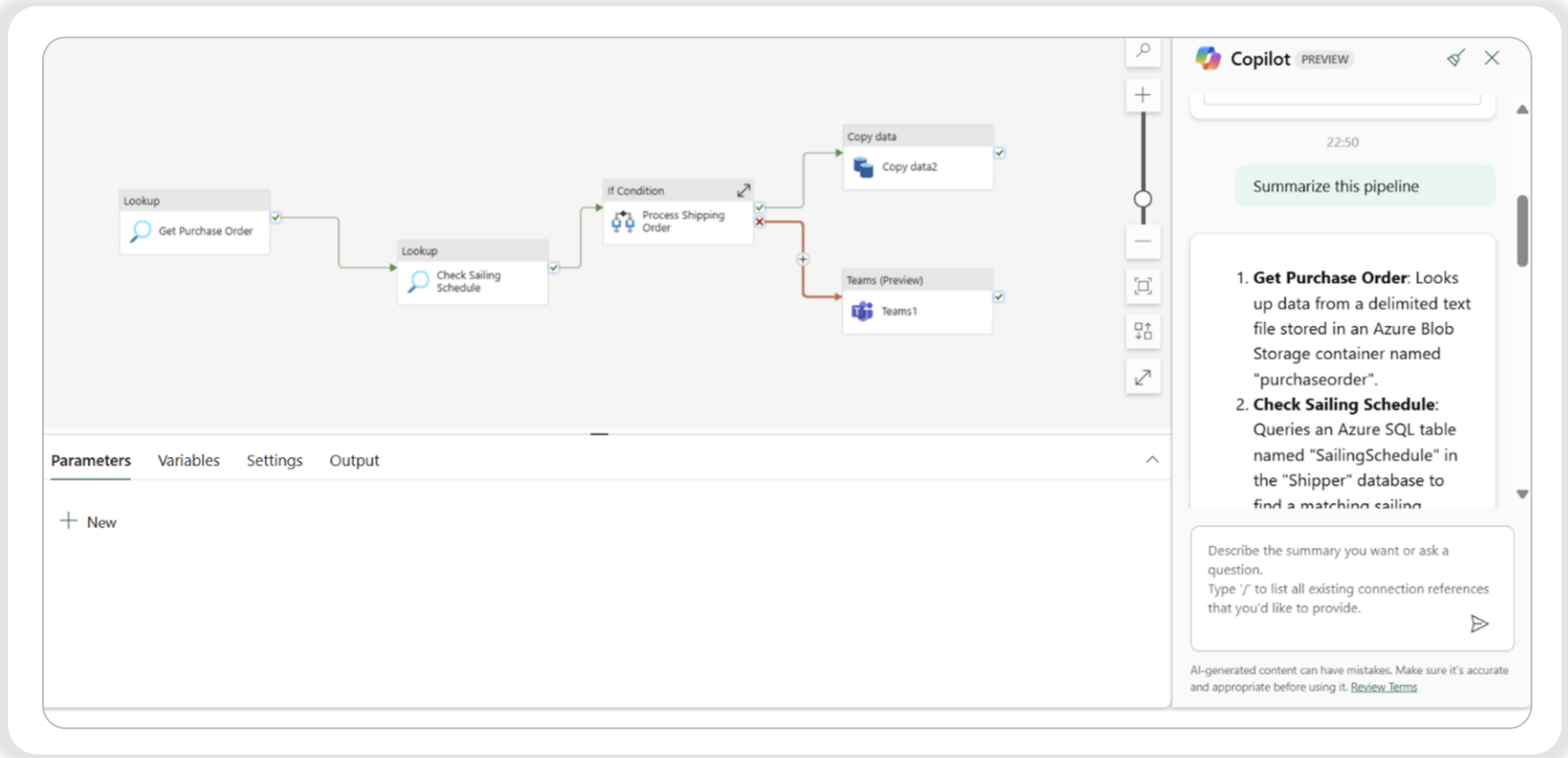
Supported capabilities in Dataflow Gen2

- Generate new transformation steps for an existing query
- Provide a summary of the query and the applied steps
- Generate a new query



Supported capabilities in Data pipelines

- Pipeline Generation
- Error message assistant
- Summarize Pipeline



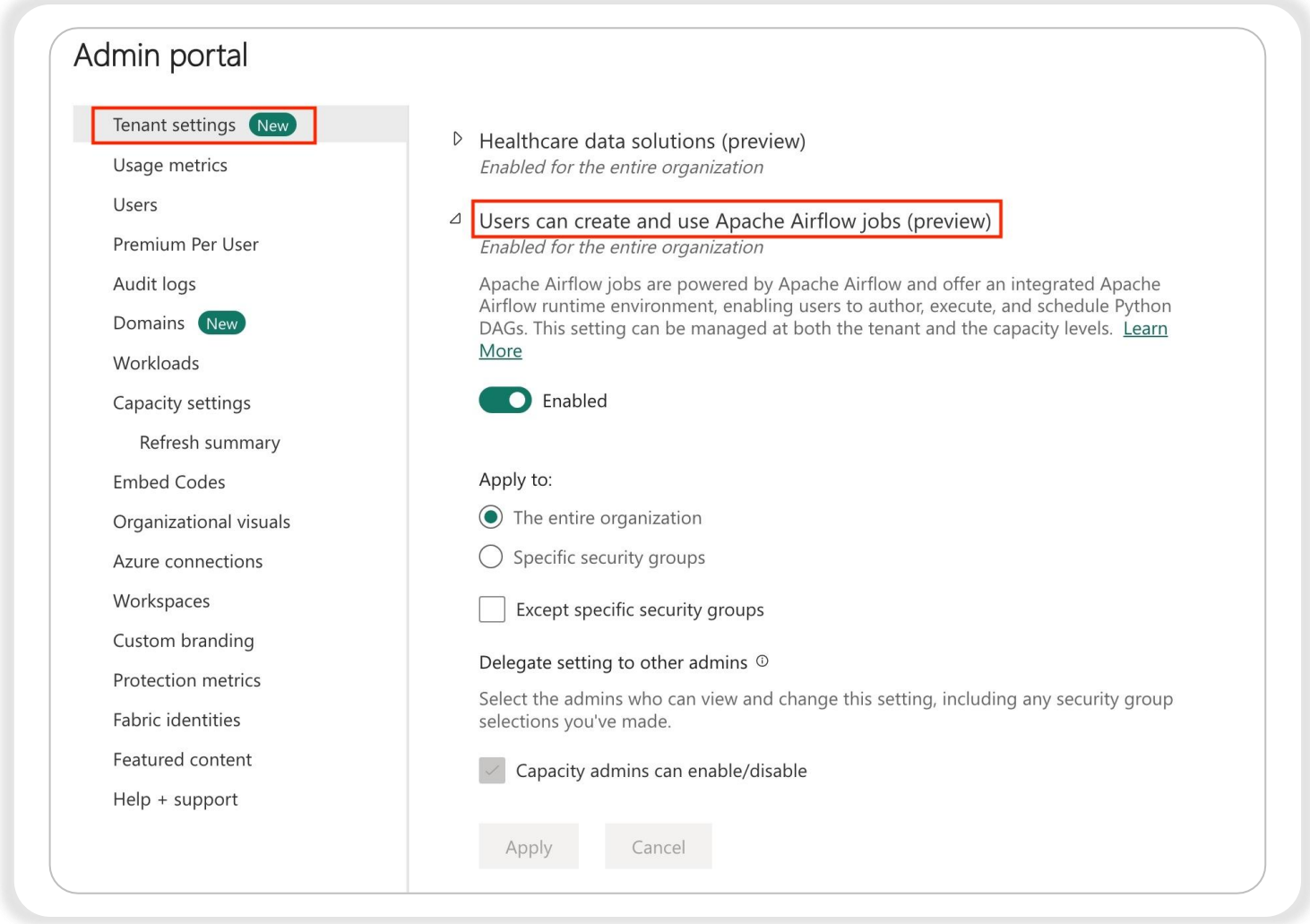
Limitations of Copilot for Data Factory

- Copilot can't perform transformations or explanations across multiple queries in a single input
- Copilot doesn't understand previous inputs and can't undo changes
- Copilot can't make layout changes to queries in your session
- Copilot doesn't produce a message for the skills that it doesn't support
- Copilot may produce inaccurate results

Workflow Orchestration with Apache Airflow job

Introduction to Apache Airflow job

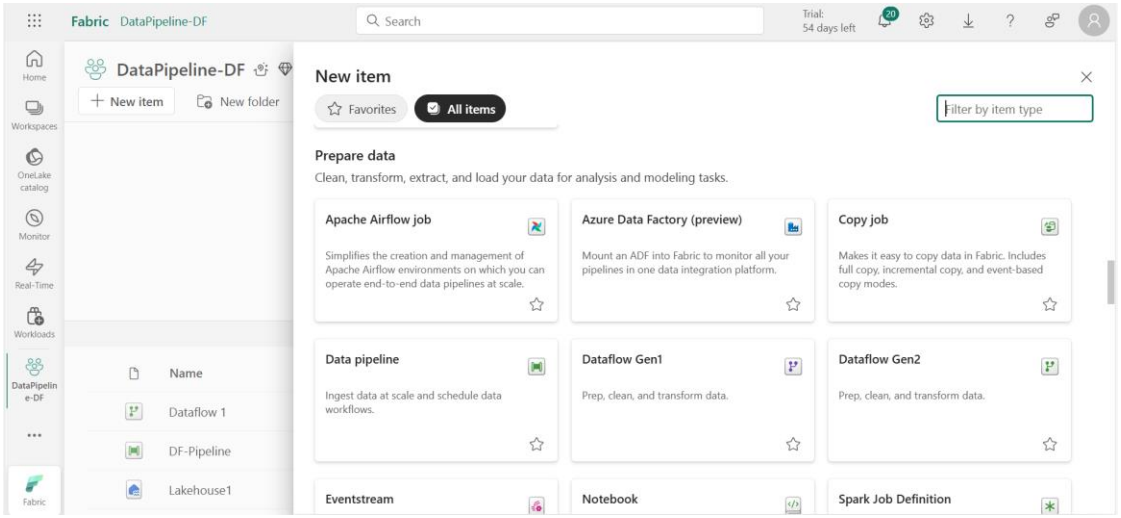
- Next generation of Azure Data Factory's Workflow Orchestration Manager
- Brings a SaaS-like experience to running DAGs in a fully managed Apache Airflow environment



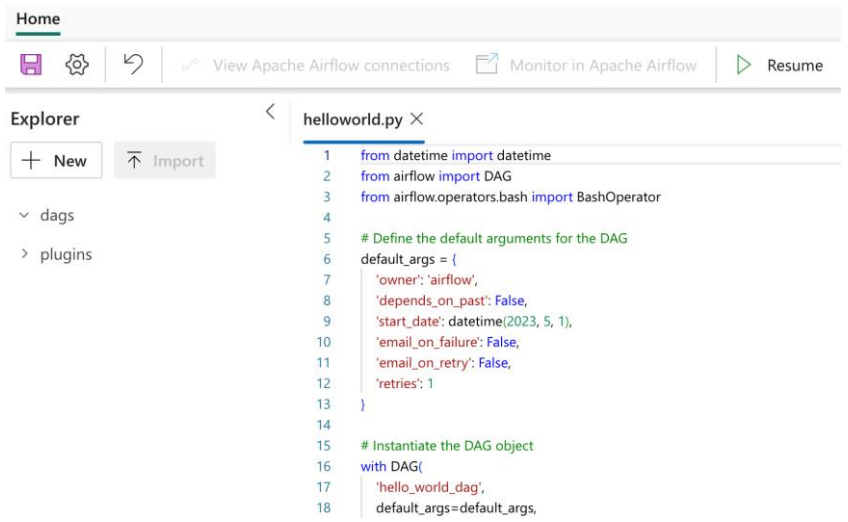
Key Features

Key Features	Apache Airflow Job in Fabric	Workflow Orchestration Manager in Azure Data Factory
Git sync	Yes	Yes
Enable AKV (Azure Key Vault) as backend	Yes	Yes
Install private package as requirement	Yes	Yes
Diagnostic logs and metrics	No	Yes
Blob Storage	No	Yes
Apache Airflow cluster IP address	Yes	Yes
Autoscale for managing production workload execution spikes	Yes	Partial
High Availability for mitigating outage/downtime	Yes	No
Deferrable Operators for suspending idle operators and free up workers	Yes	No
Pause and Resume TTL (Time to live)	Yes	No
SaaSified Experience - 10 secs to get started - Authoring DAGs - Fabric Free Trial	Yes	No

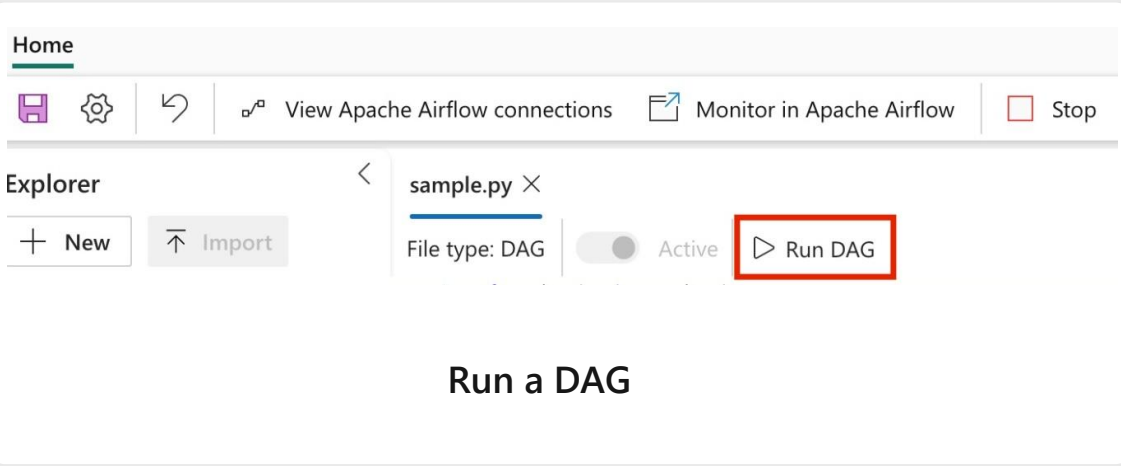
Create an Apache Airflow Job



Create an Apache Airflow Job in your workspace



Create a DAG File



Run a DAG

Product comparison

Dataflows vs. Dataflows Gen2

Feature	Dataflows	Dataflows Gen2
Output file format	CSV	Parquet (V-order)
Fast copy	No	Yes
Output data destinations	No	Yes
Premium only	No	*Yes
AI Insights	Yes	*No
AutoML	Yes	**No (???)
Attach Common Data Model (CDM) folder	Yes	**No (???)
Linked Tables	Yes	**No (? / use Shortcuts)

* To be determined
** Unknown

Important note, AI Insights and AutoML are not owned by the data integration team but by the product host (Power BI AI PMs)

Coming up tomorrow.....

Day 1

240 mins

Module 1 - Introduction to Data Factory in Microsoft Fabric

- Microsoft Fabric - The unified data platform for AI transformation
- Data Factory in Microsoft Fabric

Module 2 - Ingest data with Data Factory in Microsoft Fabric

- Data ingestion with Pipelines
- Copy activity with Pipelines
- Use parameters and expressions in Pipelines
- Data ingestion with Copy Job
- Mirroring databases in Microsoft Fabric
- Execute, monitor and troubleshoot Pipelines
- Data Pipeline storage event triggers
- REST API capabilities and CI/CD for Pipelines
- Fabric Pipelines vs. ADF/Synapse Pipelines

Hands-on labs

- Use case 01: Implementing Medallion Architecture with Data Factory in Microsoft Fabric for scalable data processing

Day 2

240 mins

Module 3 - Data Transformation with Dataflows Gen2

- Dataflows Gen2 in Microsoft Fabric
- Fast Copy in Dataflows Gen2
- Dataflow Gen2 with CI/CD and Git integration support
- Monitor your Dataflows
- Copilot for Data Factory in Microsoft Fabric
- Workflow Orchestration with Apache Airflow job

Module 4 - Migrate to Data Factory in Microsoft Fabric

- Plan your migration from Azure Data Factory to Data Factory in Microsoft Fabric
- Migrate from Dataflow Gen1 to Dataflow Gen2
- Move queries from Dataflow Gen1 to Dataflow Gen2
- Ingest data into Microsoft Fabric using the Azure Data Factory Copy Activity

Hands-on labs

- Use Case 02: Data Factory solution for moving and transforming data with dataflows and data pipelines

Microsoft

Technofocus

Thank you