

Ingest Data with Dataflows Gen2 in Microsoft Fabric

1. Introduction

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/1-introduction>

Introduction

Completed

Microsoft Fabric offers a unified solution for data engineering, integration, and analytics. A crucial step in end-to-end analytics is data ingestion. Dataflows Gen2 are used to ingest and transform data from multiple sources, and then land the cleansed data to another destination. They can be incorporated into data pipelines for more complex activity orchestration, and also used as a data source in Power BI.

Imagine you work for a retail company with stores across the globe. As a data engineer, you need to prepare and transform data from various sources into a format that is suitable for data analysis and reporting. The business requests a semantic model that consolidates disparate data sources from the different stores. Dataflows Gen2 allow you to prepare the data to ensure consistency, and then stage the data in the preferred destination. They also enable reuse and make it easy to update the data. Without a dataflow, you'd have to manually extract and transform the data from every source, which is time-consuming and prone to errors.

In this module, we explain how to use Dataflows Gen2 in Microsoft Fabric to meet your data ingestion needs.

2. Understand Dataflows Gen2 in Microsoft Fabric

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/2-dataflows-gen-2>

Understand Dataflows Gen2 in Microsoft Fabric

Completed

In our scenario, you need to develop a semantic model that can standardize the data and provide access to the business. By using Dataflows Gen2, you can connect to the various data sources, and then prep and transform the data. You can land the data directly into your lakehouse or use a data pipeline for other destinations.

What is a dataflow?

Dataflows are a type of cloud-based ETL (*Extract, Transform, Load*) tool for building and executing scalable data transformation processes.

Dataflows Gen2 allow you to extract data from various sources, transform it using a wide range of transformation operations, and load it into a destination. Using Power Query Online also allows for a visual interface to perform these tasks.

Fundamentally, a dataflow includes all of the transformations to reduce data prep time and then can be loaded into a new table, included in a data pipeline, or used as a data source by data analysts.

How to use Dataflows Gen2

Traditionally, data engineers spend significant time extracting, transforming, and loading data into a consumable format for downstream analytics. The goal of Dataflows Gen2 is to provide an easy, reusable way to perform ETL tasks using Power Query Online.

If you only choose to use a data pipeline, you copy data, then use your preferred coding language to extract, transform, and load the data. Alternatively, you can create a Dataflow Gen2 first to extract and transform the data. You can also load the data into a lakehouse, and other destinations. Now the business can easily consume the curated semantic model.

Adding a data destination to your dataflow is optional, and the dataflow preserves all transformation steps. To perform other tasks or load data to a different destination after transformation, create a data pipeline and add the Dataflow Gen2 activity to your orchestration.

Another option might be to use a data pipeline and Dataflow Gen2 for ELT (*Extract, Load, Transform*) process. For this order, you'd use a Pipeline to extract and load the data into your preferred destination, such as the lakehouse. Then you'd create a Dataflow Gen2 to connect to Lakehouse data to cleanse and transform data. In this case, you'd offer the Dataflow as a curated semantic model for data analysts to develop reports.

Dataflows can be horizontally partitioned as well. Once you create a global dataflow, data analysts can use dataflows to create specialized semantic models for specific needs.

Dataflows allow you to promote reusable ETL logic that prevents the need to create more connections to your data source. Dataflows offer a wide variety of transformations, and can be run manually, on a refresh schedule, or as part of a data pipeline orchestration.

Tip

Make your dataflow discoverable so data analysts can also connect to the dataflow through Power BI Desktop. This reduces the data preparation for report development.

Benefits and limitations

There's more than one way to ETL or ELT data in Microsoft Fabric. Consider the benefits and limitations for using Dataflows Gen2.

Benefits:

- Extend data with consistent data, such as a standard date dimension table.
- Allow self-service users access to a subset of data warehouse separately.
- Optimize performance with dataflows, which enable extracting data once for reuse, reducing data refresh time for slower sources.
- Simplify data source complexity by only exposing dataflows to larger analyst groups.
- Ensure consistency and quality of data by enabling users to clean and transform data before loading it to a destination.
- Simplify data integration by providing a low-code interface that ingests data from various sources.

Limitations:

- Dataflows aren't a replacement for a data warehouse.
- Row-level security isn't supported.
- Fabric capacity workspace is required.

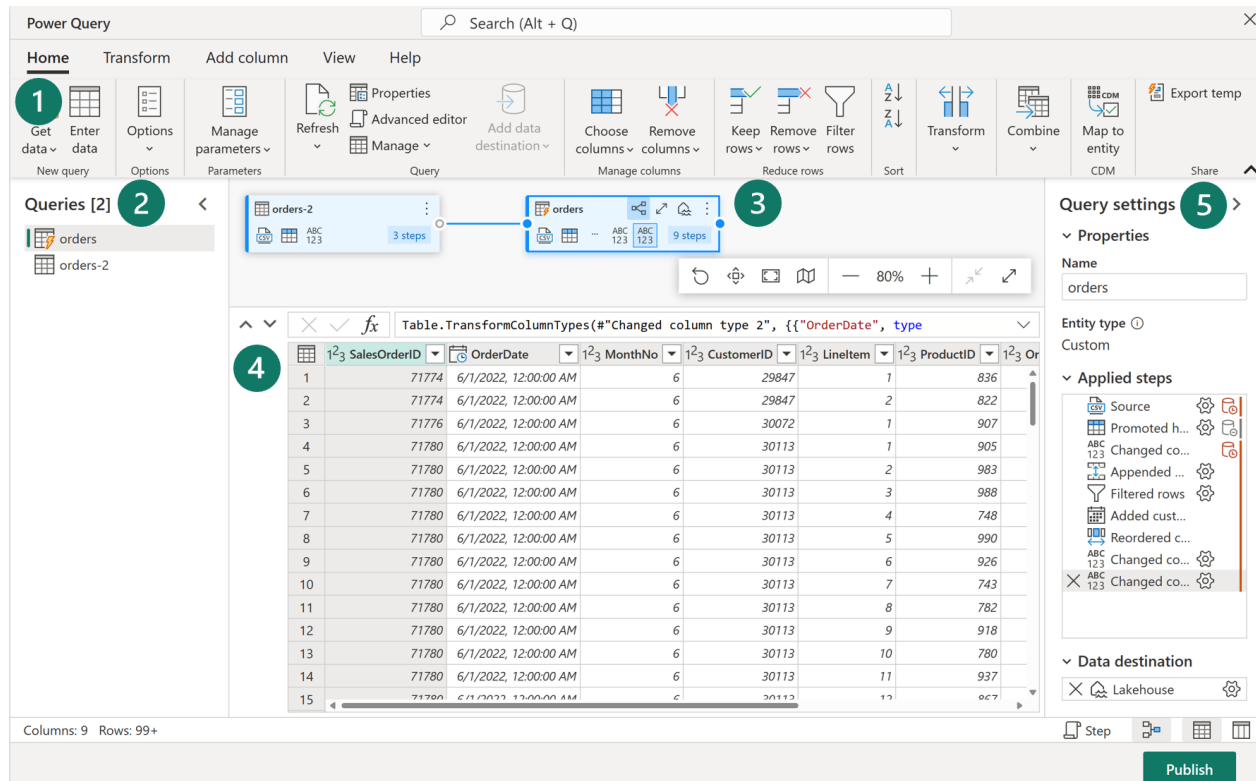
3. Explore Dataflows Gen2 in Microsoft Fabric

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/3-explore-dataflows-gen-2>

Explore Dataflows Gen2 in Microsoft Fabric

Completed

In Microsoft Fabric, you can create a Dataflow Gen2 in the Data Factory workload or Power BI workspace, or directly in the lakehouse. Since our scenario is focused on data ingestion, let's look at the **Data Factory** workload experience. Dataflows Gen2 use Power Query Online to visualize transformations. See an overview of the interface:



1. Power Query ribbon

Dataflows Gen2 support a wide variety of data source connectors. Common sources include cloud and on-premises relational databases, Excel or flat files, SharePoint, Salesforce, Spark, and Fabric lakehouses. Then there are numerous data transformations possible, such as:

- Filter and Sort rows
- Pivot and Unpivot
- Merge and Append queries
- Split and Conditional split
- Replace values and Remove duplicates
- Add, Rename, Reorder, or Delete columns
- Rank and Percentage calculator
- Choose Top N and Bottom N

You can also create and manage data source connections, manage parameters, and configure the default data destination in this ribbon.

2. Queries pane

The Queries pane shows you the different data sources - now called *queries*. These queries are called tables when loaded to your data store. You can duplicate or reference a query if you need multiple copies of the same data, such as creating a star schema and splitting data into separate, smaller tables. You can also disable the load of a query, in case you only need the one-time import.

3. Diagram view

The Diagram View allows you to visually see how the data sources are connected and the different applied transformations. For example, your dataflow connects to a data source, duplicates the query, removes columns from the source query, then unpivots the duplicate query. Each query is represented as a shape with all of the applied transformations and connected by a line for the duplicate query. You can turn this view on or off.

4. Data Preview pane

The Data Preview pane only shows a subset of data to allow you to see which transformations you should make and how they affect the data. You can also interact with the preview pane by dragging and dropping columns to change order or right-clicking on columns to filter or make changes. The data preview shows all of your transformations for the selected query.

5. Query Settings pane

The Query Settings pane includes the **Applied Steps**. Each transformation is represented as a step, some of which are automatically applied when you connect the data source. Depending on the complexity of the transformations, you might have several applied steps for each query. Most steps have a gear icon that allows you to modify the step, otherwise you must delete and repeat the transformation.

Each step also has a contextual menu when you right-click so you can rename, reorder, or delete the steps. You can also view the data source query when connecting to a data source that supports query folding.

While this visual interface is helpful, you can also view the M code through **Advanced editor**.

Advanced editor



```
1 let
2     Source = Lakehouse.Contents([]),
3     Navigation = Source{[workspaceId = "91dd9a9b-e5b1-4d41-88f1-3e315bdd9f7e"]}[Data],
4     #"Navigation 1" = Navigation{[lakehouseId = "5c04bd01-14eb-4393-8a2e-b15277568c26"]}[Data],
5     #"Navigation 2" = #"Navigation 1"{[Id = "COVID_data", ItemKind = "Table"]}[Data],
6     #"Added rank column" = Table.AddRankColumn(#"Navigation 2", "Rank", {{"positive", Order.Descending}}),
7     #"Added index" = Table.AddIndexColumn(#"Added rank column", "Index", 0, 1, Int64.Type),
8     #"Marked key columns" = Table.AddKey(#"Added index", {"Index"}, false),
9     #"Split column by delimiter" = Table.SplitColumn(#"Marked key columns", "load_time", Splitter.SplitTextByEachDelimiter
10     ({" "}, QuoteStyle.Csv), {"load_time.1", "load_time.2"}),
11     #"Changed column type" = Table.TransformColumnTypes(#"Split column by delimiter", {{"load_time.1", type date},
12     {"load_time.2", type time}}),
13     #"Filtered rows" = Table.SelectRows(#"Changed column type", each ([hospitalized] <> null)),
14     #"Changed column type 1" = Table.TransformColumnTypes(#"Filtered rows", {{"hospitalized", Int64.Type}}),
15     #"Duplicated Column" = Table.AddColumn(#"Changed column type 1", "iso_subdivision - Copy", each [iso_subdivision],
16     type text)
```

In the Query settings pane, you can see a **Data Destination** option to land your data in one of the following locations in your Fabric environment:

- Lakehouse
- Warehouse
- SQL database

You can also load your dataflow to Azure SQL database, Azure Data Explorer, or Azure Synapse Analytics.

Dataflows Gen2 provide a low-to-no-code solution to ingest, transform, and load data into your Fabric data stores. Power BI developers are familiar and can quickly begin to perform transformations upstream to improve performance for their reports.

Note

For more information, see the [Power Query](#) documentation to optimize your dataflows.

4. Integrate Dataflows Gen2 and Pipelines in Microsoft Fabric

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/4-dataflow-pipeline>

Integrate Dataflows Gen2 and Pipelines in Microsoft Fabric

Completed

Dataflows Gen2 provide an excellent option for data transformations in Microsoft Fabric. The combination of dataflows and pipelines is useful when you need to perform additional operations on the transformed data.

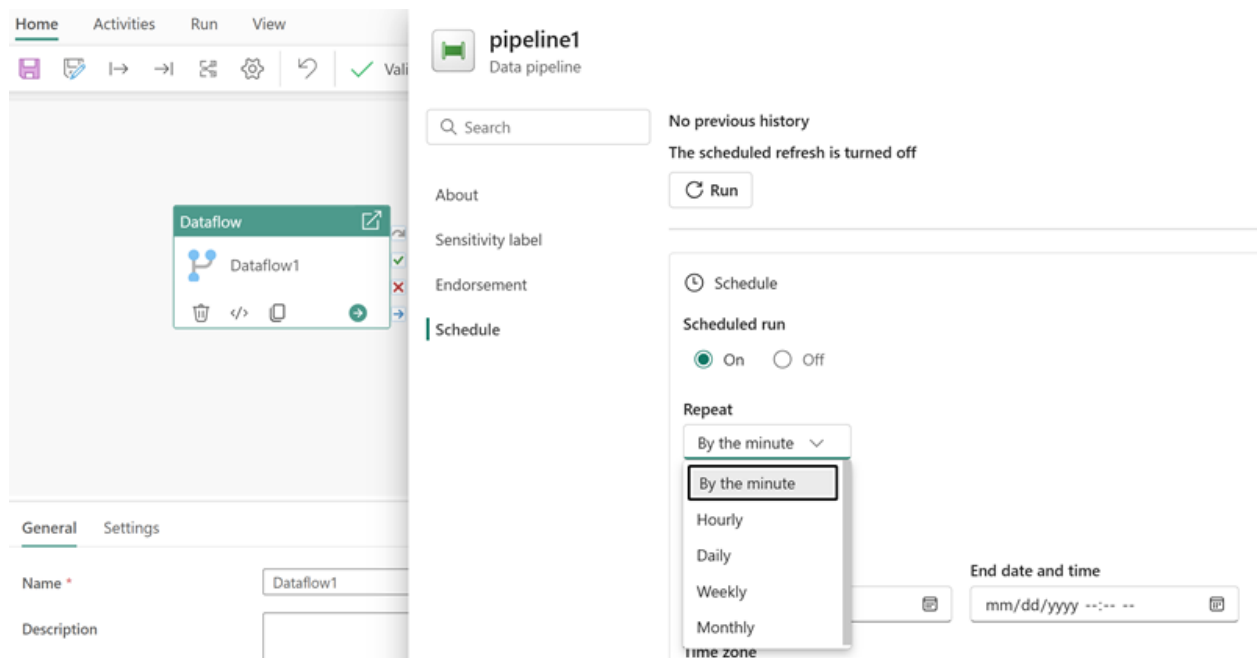
Data pipelines are a common concept in data engineering and offer a wide variety of activities to orchestrate. Some common activities include:

- Copy data
- Incorporate Dataflow
- Add Notebook
- Get metadata
- Execute a script or stored procedure

Pipelines provide a visual way to complete activities in a specific order. You can use a dataflow for data ingestion, transformation, and landing into a Fabric data store. Then incorporate the dataflow into a pipeline to orchestrate extra activities, like execute scripts or stored procedures after the dataflow has completed.

Pipelines can also be scheduled or activated by a trigger to run your dataflow. By using a pipeline to run your dataflow, you can have the data refreshed when you need it instead of having to manually

run the dataflow. When you're dealing with enterprise or frequently changing data, automation allows you to focus on other responsibilities.



5. Exercise - Create and use a Dataflow Gen2 in Microsoft Fabric

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/5-exercise>

Exercise - Create and use a Dataflow Gen2 in Microsoft Fabric

Completed

In this exercise, you'll use a Dataflow Gen2 to load transformed data into a lakehouse and add a dataflow to a pipeline.

Note

You need a Microsoft Fabric trial license with the Fabric preview enabled in your tenant. See [Getting started with Fabric](#) to enable your Fabric trial license.

Launch the exercise and follow the instructions.

[Launch Exercise](#)

6. Module assessment

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/6-knowledge-check>

Module assessment

Completed

7. Summary

<https://learn.microsoft.com/en-us/training/modules/use-dataflow-gen-2-fabric/7-summary>

Summary

Completed

In this module, we walked through a scenario where both data engineers need to ingest, transform, and load data into a Fabric data store such as a lakehouse. We also identified data analyst needs to perform transformations closer to the data source to support Power BI report development.

With Microsoft Fabric, you can create Dataflows Gen2 to perform data integration for your lakehouse, and optionally include the dataflow in a Data Pipeline as well. You learned about Dataflows Gen2 and how to use them as part of your data integration process. Power Query Online offers a visual interface to perform complex data transformations without writing any code.

To learn more about data integration, see the [Data Factory in Microsoft Fabric](#) documentation.