MS Fabric

Data Flow:

Type of cloud based ETL 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.

The goal of dataflows Gen2 is to provide an easy, reusable way to perform ETL tasks using Power Query Online.

* Data pipeline-> copy data -> use preferred coding for ETL.
* Dataflow Gen2 -> extract and transform -> load into lakehouse and other destinations for easy consumption.

Adding data destination to dataflow -> optional;

Dataflow -> preserves all transformation steps.

To perform other tasks or load data to a different destination after transformation:

Create pipeline -> add dataflow Gen2 activity to orchestration

Other option:

* Use data pipeline + dataflow Gen2 for ELT
* Pipeline -> extract and load data to destination -> create dataflow gen2 to connect to LH to cleanse and transform data.

Dataflows -> can be horizontally partitioned as well.

* Also promote reusable ETL logic.
* Can be run manually/ on refresh schedule/ part of data pipeline orchestration.

Benefits:

* Extend data with consistent data.
* Allow self service users access to a subset of DW.
* Optimize performance (enable extracting data once for reuse, reducing data refresh time)
* Simplify data source complexity by exposing only dataflows.
* Ensure consistency and quality of data by enabling cleaning/ transformation before data load.
* Simplify data integration by providing low-code interface.

Limitations:

* Aren’t replacement for a data warehouse.
* Row level security isn’t supported.
* Capacity workspace is required.

**EXPLORING DATAFLOWS GEN2 IN MS FABRIC:**

You can create a dataflow Gen2 in ADF workload in PowerBI workspace/ directly in LH.

They use Power Query online to visualize transformations.

1. Power Query Ribbon:

Numerous data transformations are possible:

* + Filter & sort rows
  + Pivot and Unpivot
  + Merge and Append queries
  + Split and conditional split
  + Replace vals and remove duplicates
  + Add, rename, reorder, delete columns
  + Rand & 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.

1. Queries Pane:

Shows different data sources – now called queries.

These queries are called – tables when loaded to your data store.

You can duplicate/ reference a query if you need multiple copies of same data.

You can also disable load of query in case you only need one time import.

1. Diagram view:

Allows you to visually see how data sources are connected and diff applied transformations. You can turn this view on/off.

1. Data preview pane:

This only shows a subset od data to see which transformations (for selected query) you should make and how they affect the data.

1. Query Settings pane:

Includes applied steps.

Each transformation -> step (some are applied automatically when you connect the data source)

In this pane, you can see 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.

A screenshot of a computer

AI-generated content may be incorrect.

**INTEGRATE DATAFLOWS GEN2 AND PIPELINES IN MICROSOFT FABRIC:**

Combination of dataflows and pipelines is useful when you need to perform additional operations on the transformed data.

Some common activities:

* Copy data
* Incorporate dataflow
* Add notebook
* Get metadata
* Execute script or stored procedure

A screenshot of a computer

AI-generated content may be incorrect.

**Use dataflow for data ingestion, transformation and land into fabric data store -> then incorporate dataflow into pipeline to orchestrate extra activities, like execute scripts/ stored procedures after the dataflow is completed.**

Pipelines -> **scheduled / activated by trigger** to run dataflow.

By using 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 are dealing with enterprise of frequently changing data, automation allows you to focus on other responsibilities.

A screenshot of a computer

AI-generated content may be incorrect.