

Low friction data transformation and data movement using Fabric Dataflows

Benni De Jagere





Benni De Jagere

Senior Program Manager | Fabric Customer Advisory Team (FabricCAT)



Fabric CAT

.be Member

@BenniDeJagere

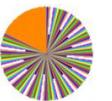
/bennidejagere



/bennidejagere

/bennidejagere

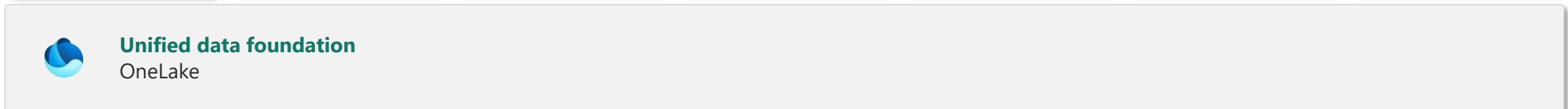
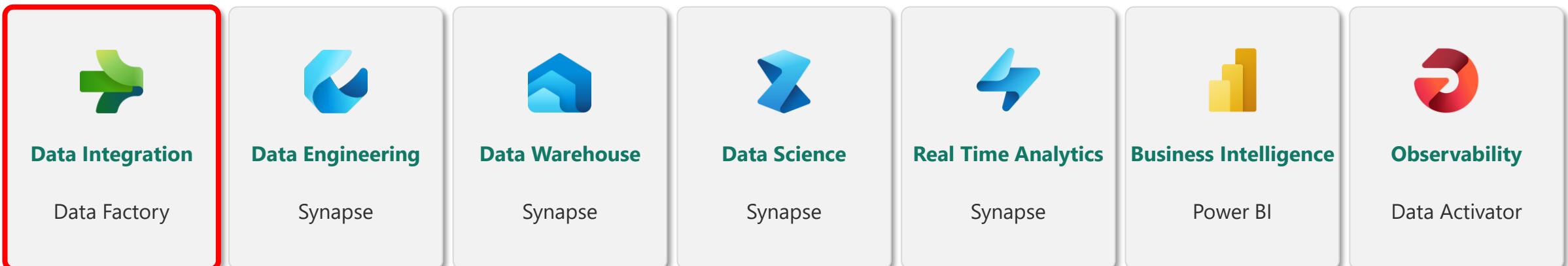
#SayNoToPieCharts



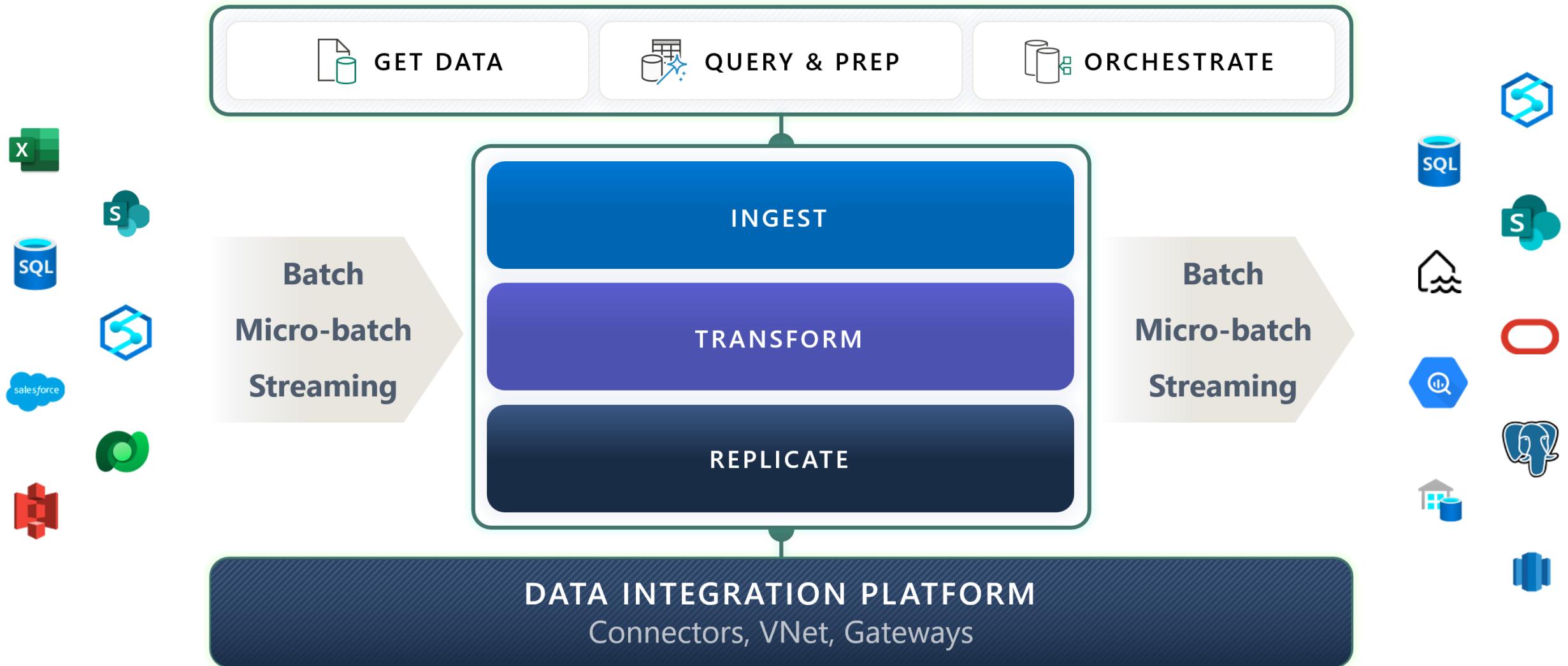


Microsoft Fabric does it all—in a unified solution

An end-to-end analytics platform that brings together all the data and analytics tools that organizations need to go from the data lake to the business user



Fabric Data Factory



What is Data Factory in Microsoft Fabric?

No-code / low-code data integration experience with:

- Scale and power of Azure Data Factory
- Ease-of-use of Power Query
- Intelligence of Copilot



What are Dataflows Gen 2?

Self-serve Data preparation

Familiar experience with **Power Query + M Language**

File, relational, multi-dimensional, SaaS data experiences

Embedded into Microsoft citizen experiences – Excel, Power BI, Power Apps & Automate, Dynamics 365



A screenshot of the Microsoft Power Query interface. It shows a data flow with several steps: 'Orders' and 'Order_Details' tables being combined via 'Merge', then 'Table.FirstN' being used to sort the rows. The resulting table is then grouped by 'CustomerID'. The preview pane shows a list of 10 customers with their details and total sales. The interface includes various transformation tools like 'Choose columns', 'Remove columns', 'Group by', and 'Sort'.

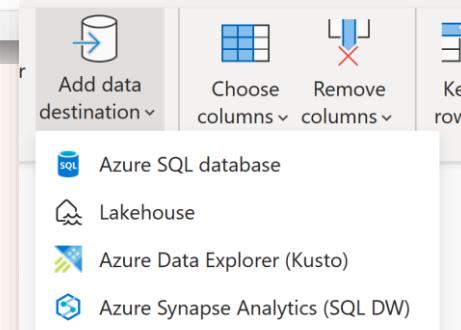
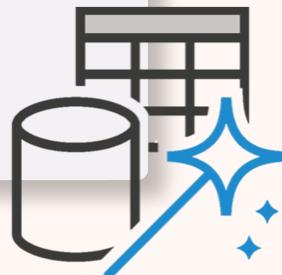
Next Generation of Power BI Dataflows

Shorter authoring experience

New output **Data Destinations**

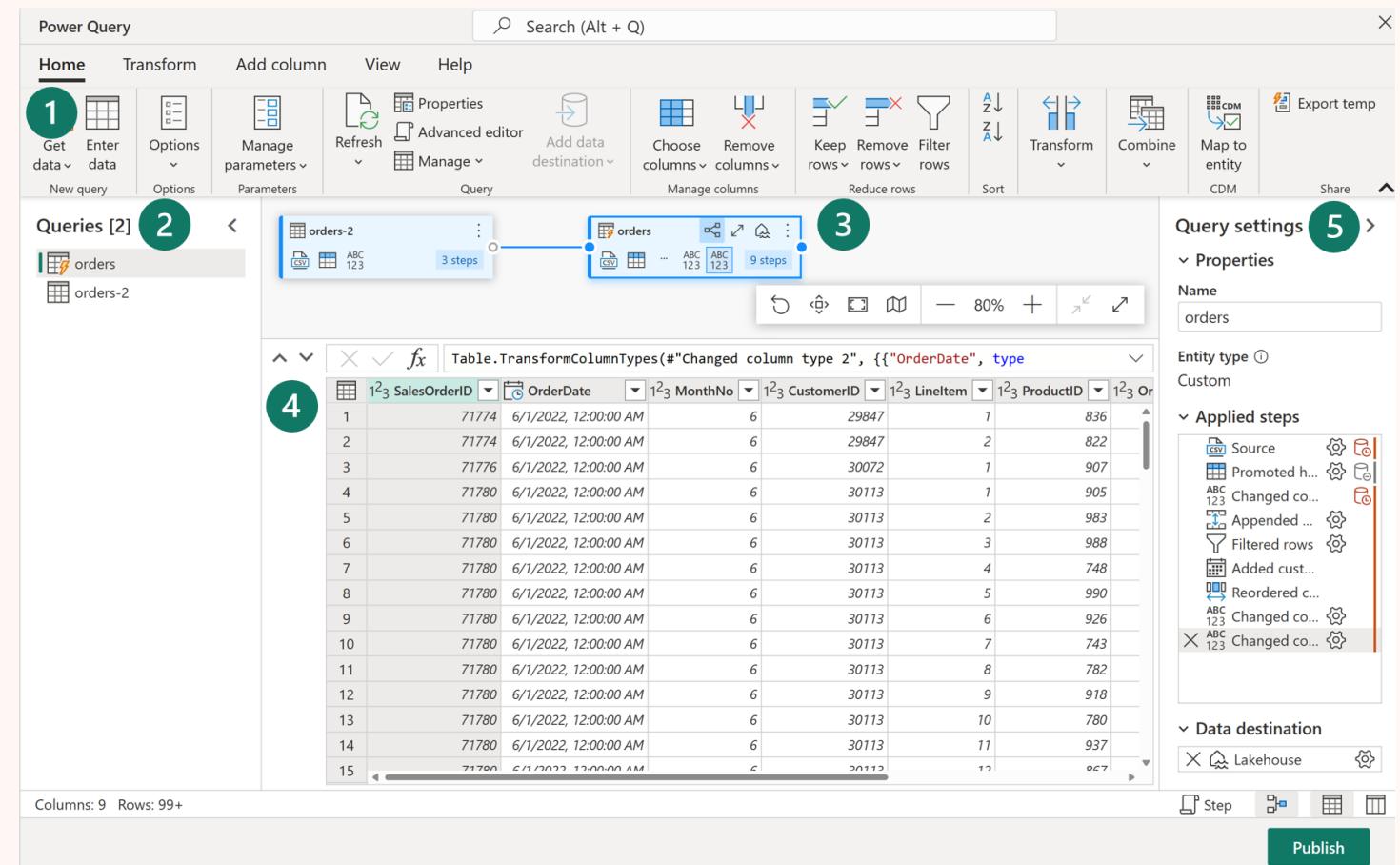
New **refresh history** and **monitoring**

Integration with Data pipelines



Components of a Dataflow

1. Get Data (connectors)
2. Queries Pane
3. Diagram View
4. Data Preview pane
5. Query settings pane

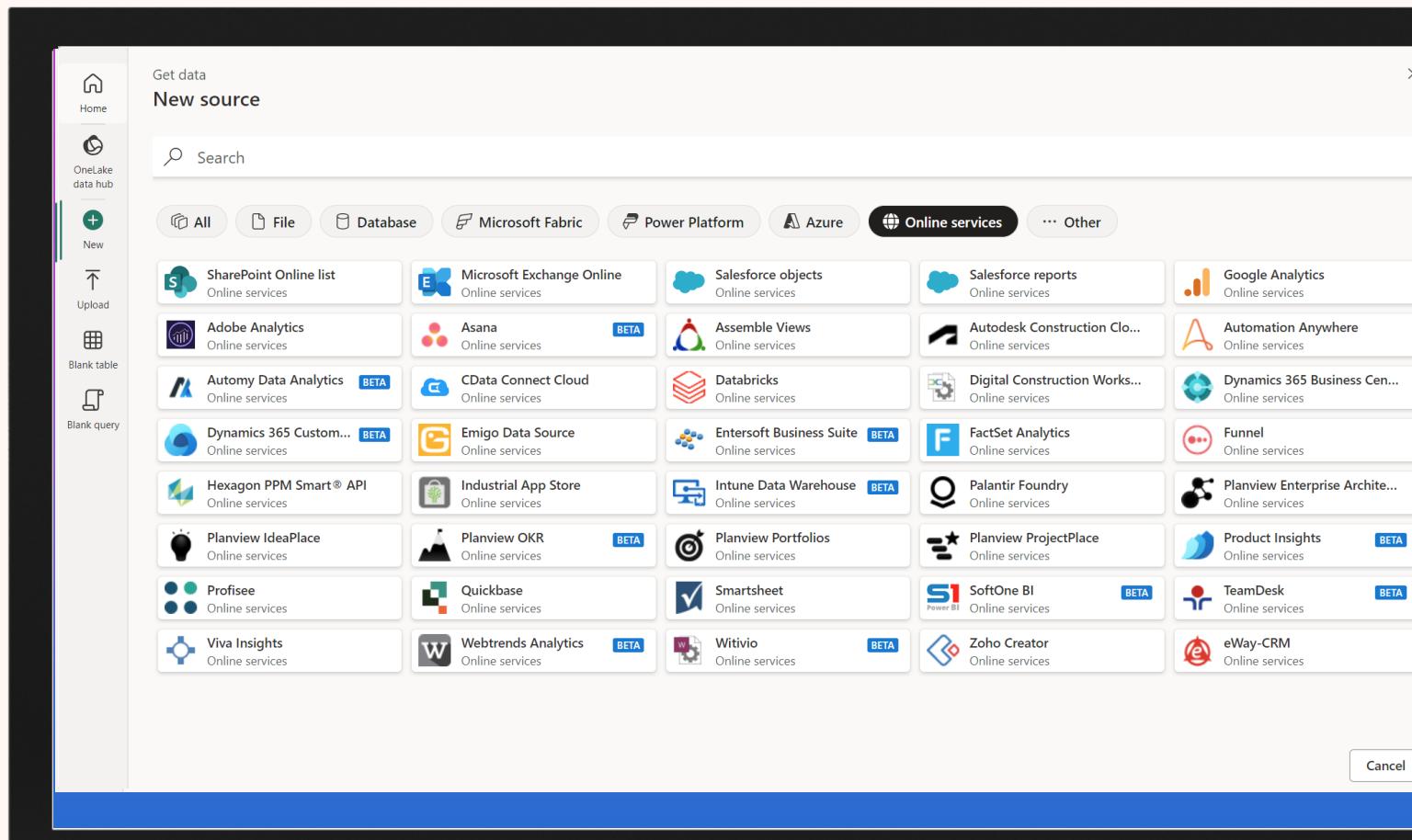


Dataflows Gen 2

- Dataflow Gen 2 in **Microsoft Fabric** is a powerful data preparation technology that allows you to create, transform, and load data into Fabric and Azure destinations. Here's what you can do with it:
- **Create Dataflows:**
 - Dataflows are self-service, cloud-based tools for data preparation.
- **Get Data:**
 - Dataflows enable you to retrieve data from 100s of on-premise and cloud data sources.
- **Apply Transformations:**
 - Once you've connected to your data source, it's time to shape it according to your needs.
 - Use the Power Query editor to apply transformations. For instance:
 - Calculate the total number of orders per customer using the Group By feature, combine, remove columns, etc.
- **Configure Destinations:**
 - Specify a destination to store the results of the query and transformations.
- **Publish Dataflows:**
 - After applying transformations, you can publish your dataflow so that it can start processing data.

Get Data

- Fabric
- Database
- Azure
- 175+ connectors



Apply Transformations

- Change data types
- Remove Columns
- Aggregations

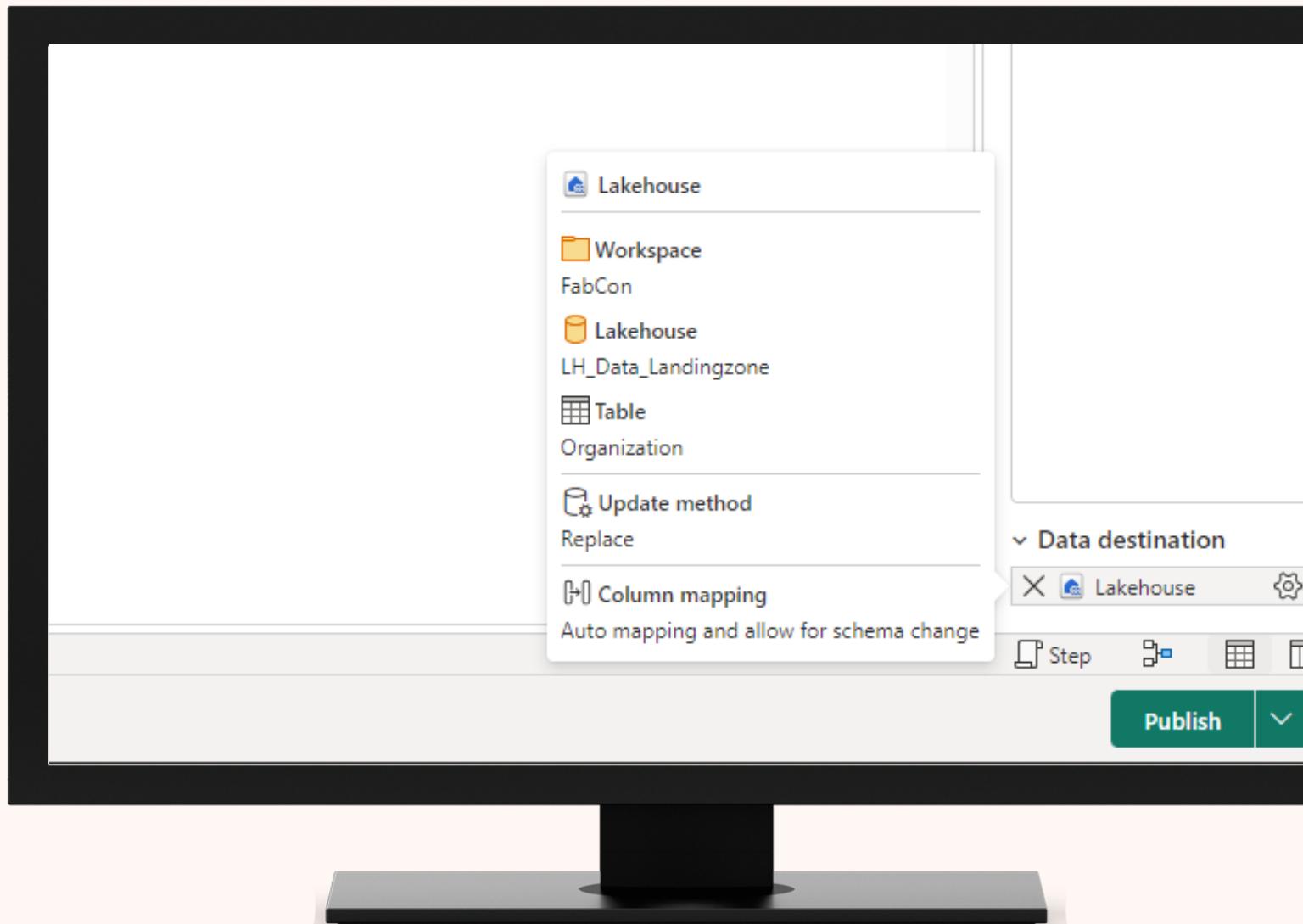
• ..

The screenshot shows the Microsoft Power BI desktop application's ribbon at the top. The 'Transform' tab is selected. Below the ribbon is a table with several columns: Organization Id, Name, Website, Country, Description, Founded, and Industry. A context menu is open over the first row of the table, specifically over the 'Name' column. The menu items visible are: Copy preview data, Remove columns (which is highlighted with a blue border), Remove other columns, Duplicate column, Add column from examples..., Remove duplicates, Remove errors, Split column, Replace values..., Replace errors..., Change type, Transform column, Group by..., Fill, Unpivot columns, Unpivot other columns, Unpivot only selected columns, Rename..., Move, Drill down, and Add as new query.

Organization Id	Name	Website	Country	Description	Founded	Industry
74fc6fD8	anner.com/	Cape Verde	Horizontal bi-directional artificial intelligence	1971	Professional Training	
4C119be1	evine-marks.com/	Reunion	Progressive maximized instruction set	2008	Investment Management / Hedge Fund / Private Equity	
347d1dcE	www.ortiz.org/	South Africa	Decentralized dynamic attitude	1993	Music	
4Dcd6629	ine.info/	Congo	Intuitive actuating approach	2020	Information Technology / IT	
92BB0B8e	www.barrera.com/	Eritrea	Optional well-modulated budgetary management	1987	Recreational Facilities / Services	
22dC0BF	www.miranda.com/	Burkina Faso	Cloned tertiary task-force	1997	Consumer Electronics	
FEC6f6d7	www.anthony-braun.com/	Pitcairn Islands	Self-enabling attitude-oriented task-force	2007	Management Consulting	
2cd8252	www.blackwell.com/	Nigeria	Customizable asymmetric initiative	1976	Photography	
706DC76	www.lawrence-huffman.net/	Greece	Cross-group bottom-line archive	2011	Government Relations	
Fdb4732	www.mccormick-reed.com/	Moldova	De-engineered maximized complexity	1988	Entertainment / Movie Production	
16FdD40	kevins.com/	Netherlands Antilles	Synergized eco-centric process improvement	2005	Sports	
bc46f8d8	wis.com/	Ghana	User-friendly motivating project	2003	Law Practice / Law Firms	
e0cd00B	anco-espinoza.biz/	Colombia	Versatile foreground collaboration	2009	Hospital / Health Care	
5d83EC7	nawxwell.com/	Germany	Total 24/7 matrix	1989	Motion Pictures / Film	
31aDF88	nora-adkins.com/	Turks and Caicos Islands	Enhanced coherent functionalities	1976	Biotechnology / Greentech	
CFAE6FB	www.morrison-ware.com/	Bulgaria	Public-key real-time groupware	2002	Architecture / Planning	
2D92015	www.velasquez-leonard.org/	Saint Helena	Upgradable demand-driven challenge	1987	Design	
0Fd92FB	hite.net/	Faroe Islands	Assimilated demand-driven portal	1980	Animation	
4BD8C4d	www.deleon.info/	Uganda	Reduced client-server forecast	1971	Marketing / Advertising / Sales	
5bd1F8E	www.koch-phelps.com/	United Arab Emirates	Diverse next generation firmware	2001	Machinery	
e12C672	www.morse.com/	Solomon Islands	Sharable even-keeled definition	2019	Mining / Metals	
aFCfbA42	www.stewart.com/	Tanzania	Reactive explicit task-force	1987	Motion Pictures / Film	
25D7F7e	iecker-lloyd.com/	Guernsey	Business-focused eco-centric firmware	1998	Religious Institutions	
C31ec1d	alloway-soto.info/	Lao People's Democratic Republic	Secured secondary help-desk	1973	Health / Fitness	
514aB6AA0bc66d	Little-Cross	https://barker.net/	Reactive non-volatile installation	2021	Tobacco	

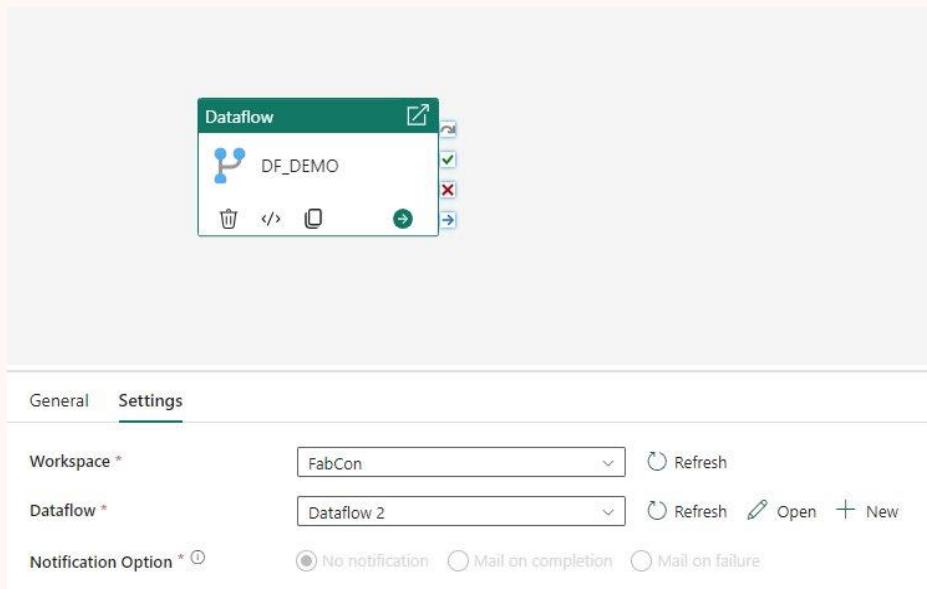
Publish Dataflows

- Select Destination
 - Lakehouse
 - Warehouse
 - Azure Data Explorer
 - Azure SQL Database
 - More to come

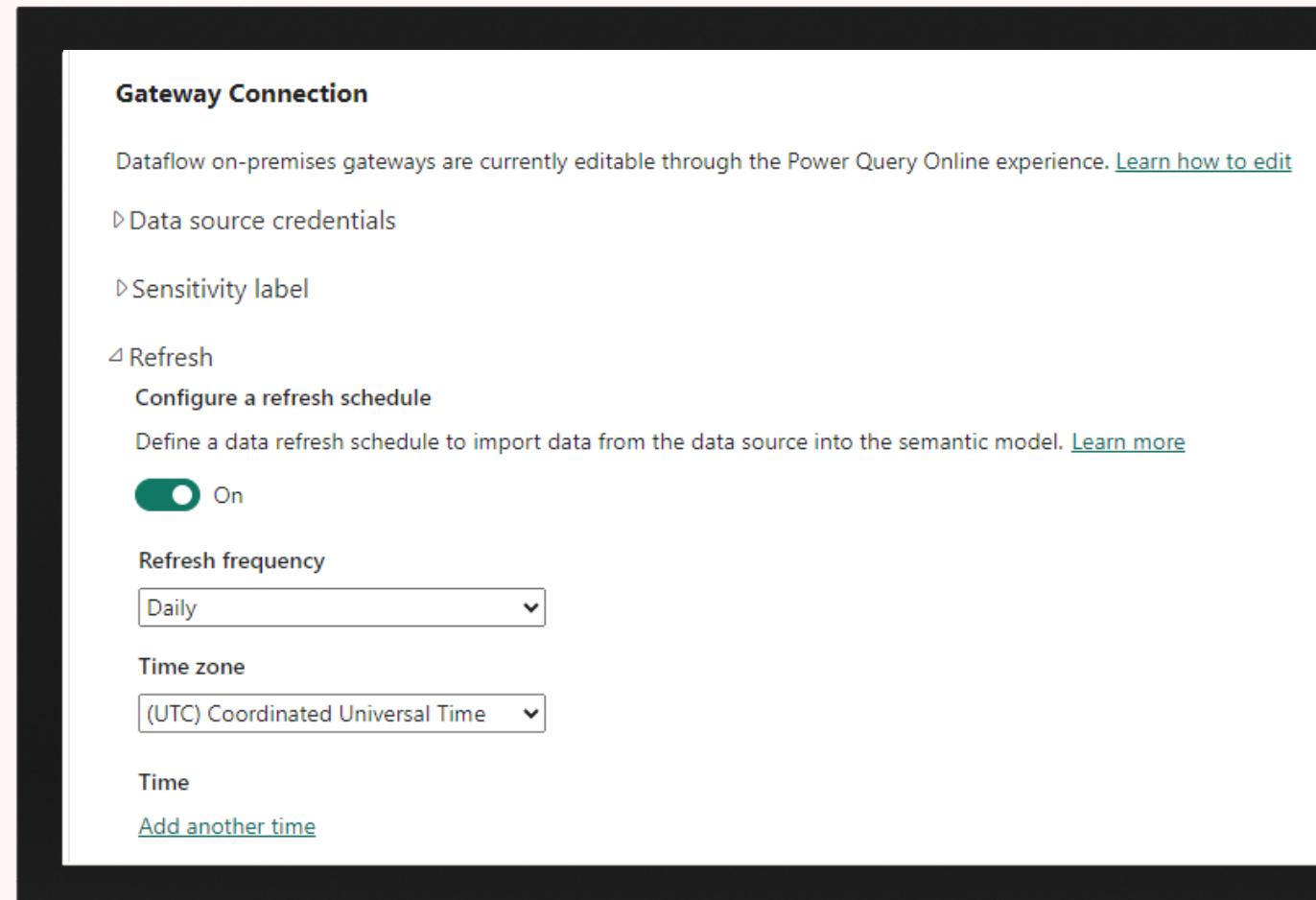


Refreshing Dataflows

- Schedule dataflow refreshes for automatic operation
- Execute from a pipeline



The screenshot shows the 'Settings' tab of the Power BI Dataflows interface. At the top, there's a list of dataflows: 'DF_DEMO' (selected) and 'Dataflow 2'. Below this, there are fields for 'Workspace' (FabCon), 'Dataflow' (Dataflow 2), and 'Notification Option' (No notification selected). A 'Refresh' button is also present.



The screenshot shows the 'Gateway Connection' settings in Power Query Online. It includes sections for 'Data source credentials', 'Sensitivity label', and 'Refresh'. The 'Refresh' section allows configuring a schedule to import data from the data source into the semantic model. The 'On' toggle is turned on, and the 'Refresh frequency' is set to 'Daily'. The 'Time zone' is '(UTC) Coordinated Universal Time'. There is also an option to 'Add another time'.



Smart data preparation
and deeply integrated
experiences

Add column from examples

Transform, combine, extract or enrich your data, in one or more columns within the Power Query Editor tables.

Simply specify a few output values for your new columns and Power Query generates the right column generation logic for you.

The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table with two columns: "ContactName" and "ContactTitle". The table contains 22 rows of data. The "ContactName" column lists names like Maria Anders, Ana Trujillo, Antonio Moreno, etc., and the "ContactTitle" column lists titles like Sales Representative, Owner, Marketing Manager, etc. The Power Query ribbon is visible at the top, and the "Applied steps" pane on the right shows the "Removed other columns" step.

ContactName	ContactTitle
Maria Anders	Sales Representative
Ana Trujillo	Owner
Antonio Moreno	Owner
Thomas Hardy	Sales Representative
Christina Berglund	Order Administrator
Hanna Moos	Sales Representative
Frédérique Citeaux	Marketing Manager
Martín Sommer	Owner
Laurence Lebihan	Owner
Elizabeth Lincoln	Accounting Manager
Victoria Ashworth	Sales Representative
Patricia Simpson	Sales Agent
Francisco Chang	Marketing Manager
Yang Wang	Owner
Pedro Afonso	Sales Associate
Elizabeth Brown	Sales Representative
Sven Ottlieb	Order Administrator
Janine Labrune	Owner
Ann Devon	Sales Agent
Roland Mendel	Sales Manager
Aria Cruz	Marketing Assistant
Diego Roel	Accounting Manager

Table by example (Web)

Extract any data from any HTML page in the world. Just preview the source page and specify output values.

Power Query will do the time-consuming work and extract all the appropriate data for you.

Out-of-the-box connectivity to hundreds of sources

We're building the connectors data analysts want most into a lot of Microsoft products you use every day. And if you can't find what you're looking for, we've made it easy to build your own. Even [get them certified by Microsoft](#).

Choose a connector category:

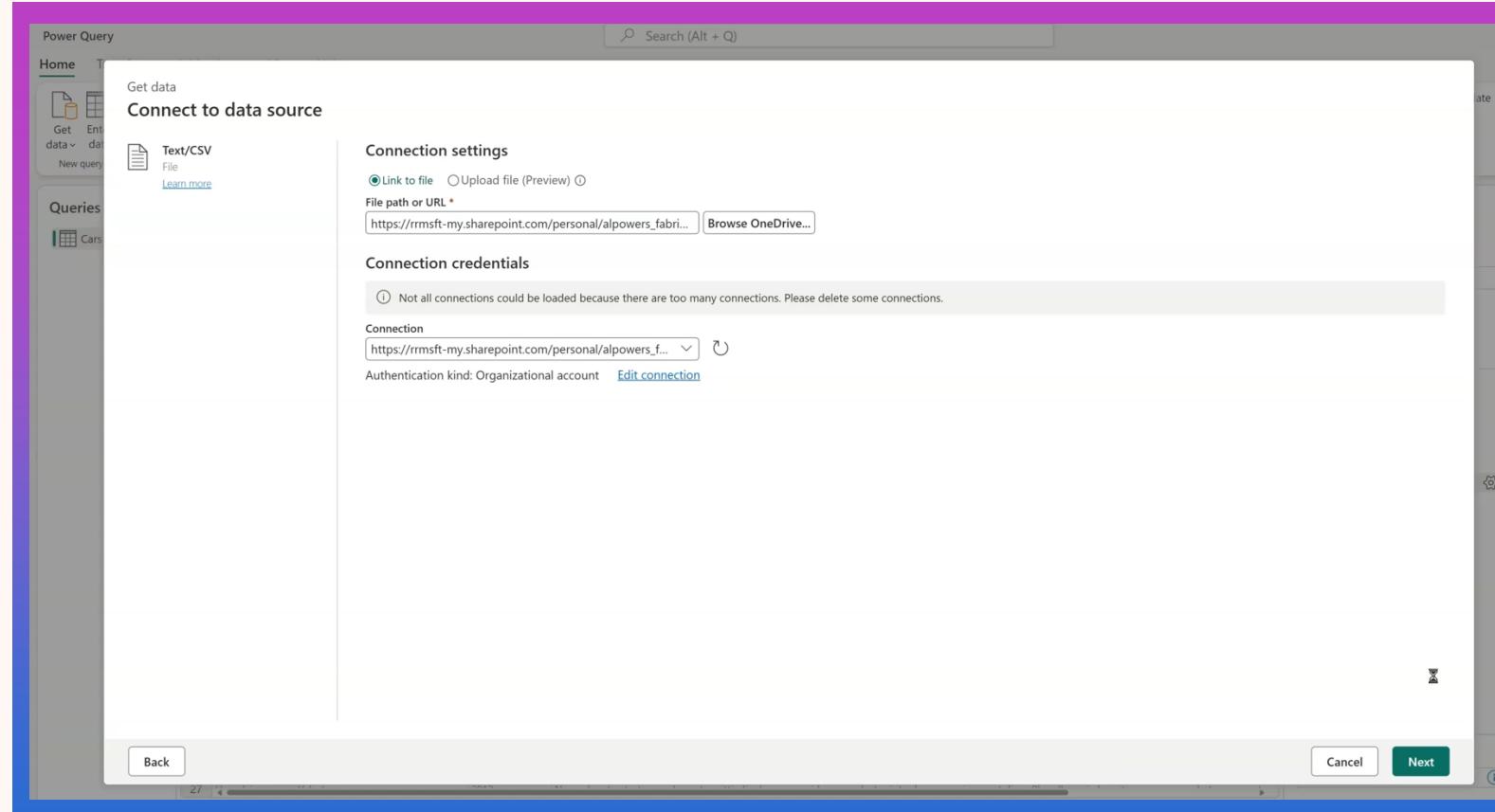
All connectors

Category	Connector	Provider	Products
Database	Access database	By Microsoft	Power BI, Excel, Analysis Services
Other	Acterys (Beta)	By Acterys	Power BI
Database	Action (Beta)	By Action	Power BI
Other	Active Directory	By Microsoft	Power BI, Excel, Analysis Services
Online Services	Adobe Analytics	By Microsoft	Power BI
Database	Amazon Athena (Beta)	By Amazon	Power BI
Other	Amazon OpenSearch Service (Beta)	By Amazon	Power BI
Database	Amazon Redshift	By Microsoft	Power BI
Other	Anaplan	By Anaplan	Power BI
Online Services	appFigures (Beta)	By Microsoft	Power BI
Online Services	Asana (Beta)	By Asana	Power BI
Online Services	Assemble Views	By Autodesk	Power BI

Table by example (Text/CSV)

Extract data from a Text/CSV file. Just specify sample output values from the source.

Power Query will do the time-consuming work and extract all the appropriate data for you.



Fuzzy merge

Power Query's built-in Fuzzy Matching algorithm lets you to merge multiple tables using an approximate match to correlate things like slightly different versions of product names, customer names or address information—to name a few examples.

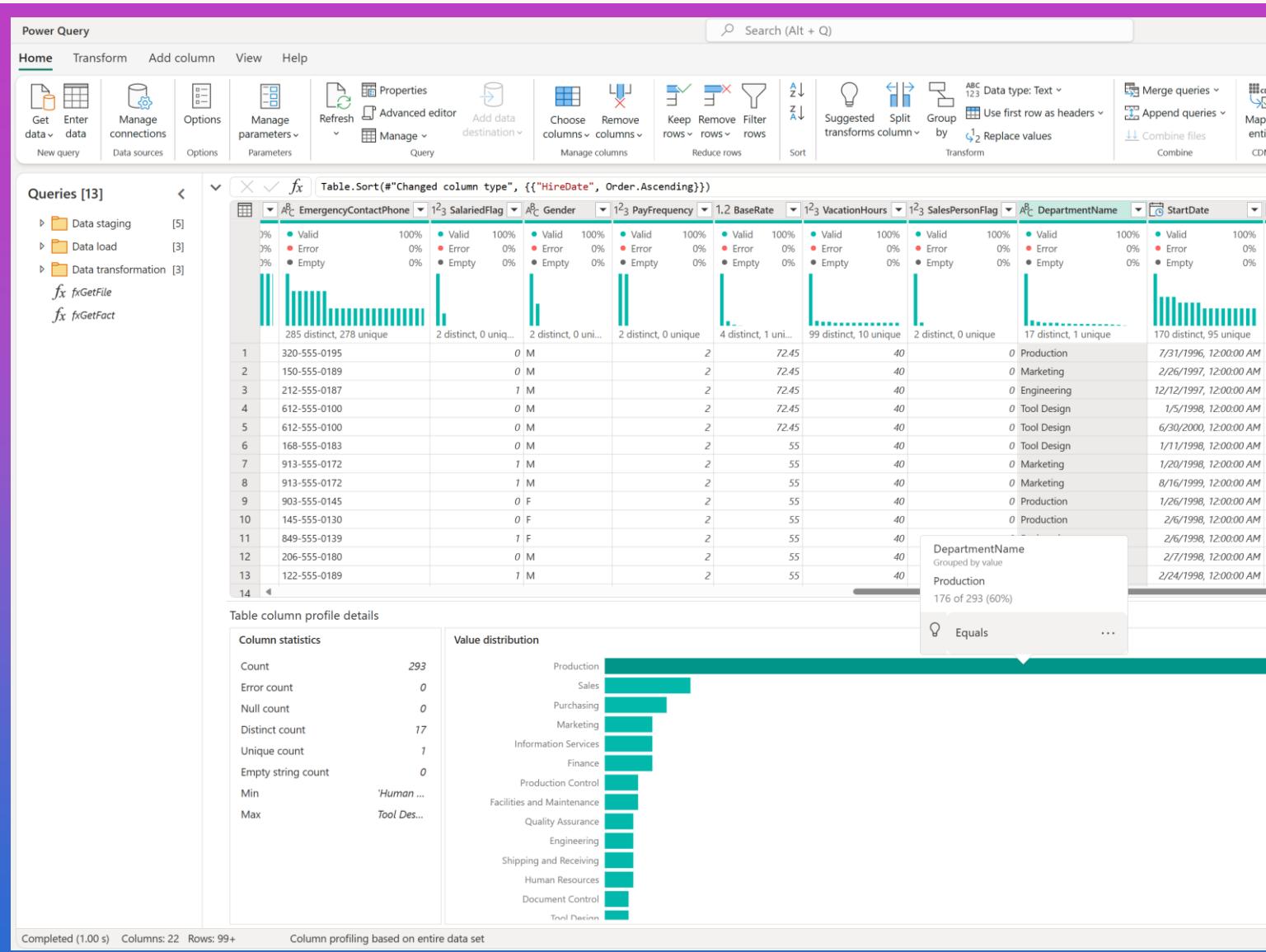
The screenshot shows the Microsoft Power Query interface. The ribbon at the top includes Home, Transform, Add column, View, and Help tabs. The Home tab is selected. Below the ribbon are various icons for data management, such as Get data, Manage connections, Options, and Refresh. The main area displays a table titled "Question" with the following data:

	Question
1	Apple
2	Aple
3	Pineapple
4	Water melon
5	watermln
6	watermleon
7	banana
8	Bananans
9	apls

The "Transform" tab is visible at the bottom of the ribbon. The status bar at the bottom right indicates "Data type: Text".

Data profiling

- **Inline Column Quality bars**
Quickly spot erroneous or empty values across all your columns.
- **Inline Value Distribution histograms**
Understand the number of different values, which are unique, and most/least common.
- **Detailed Column Profiles pane**
Select a specific column and dig deeper into a profile to fully understand what's in the data.



Column-pair suggestions

- The Merge dialog now intelligently detects matching columns from both the left and right tables, displayed in a convenient lightbulb icon in the top right corner.
 - Precise or approximate column title matches trigger these helpful suggestions.

Merge [?](#)

Select a table and matching columns to create a merged table.

Left table for merge *

1 ² ProductKey	1 ² ProductLabel	A ^B ProductName	A ^B ProductDescription
1	101001	Contoso 512MB MP3 Player E51 Silver	512MB USB driver p
2	101002	Contoso 512MB MP3 Player E51 Blue	512MB USB driver p
3	101003	Contoso 1G MP3 Player E100 White	1GB flash memory a
4	101004	Contoso 2G MP3 Player E200 Silver	2GB flash memory, l

Right table for merge *

1 ² ProductSubcategoryKey	1 ² ProductSubcategoryLabel	A ^B ProductSubcategoryName	A ^B ProductSubcategoryDescription
1	101	MP4&MP3	MP4&MP3
2	102	Recorder	Recorder
3	103	Radio	Radio
4	104	Recording Pen	Recording I

Join kind

Left outer

Right outer

Full outer

Inner

Left anti

Right anti

Use fuzzy matching to perform the merge

› Fuzzy matching options

The selection matches 2,517 rows from both the tables

OK Cancel

Suggestions [Learn more](#)

Select any of the suggested column-pair mappings for the selected tables.

DimProduct_raw	DimProductSubcategory_raw
1 ² ProductSubcategoryKey	→ 1 ² ProductSubcategoryKey

DimProduct_raw	DimProductSubcategory_raw
1 ² ProductKey	→ 1 ² ProductCategoryKey

Applied steps

- Source
- Expanded
- Merged
- Expanded
- Removed
- Get column
- Select none
- Select column

32 distinct, 0 unique



Copilot for Data Factory

Copilot for Data Factory

Easily integrate generative AI into your **dataflows** 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

The screenshot shows the Azure Data Factory interface. On the left, the sidebar includes Home, Create, Browse, Monitoring hub, Workspaces, and My workspace. The main area displays a dataflow named 'Orders'. The dataflow consists of two stages: 'Orders - Staging' and 'Orders'. The 'Orders' stage has a 'Source' step, a 'Navigation' step, and an 'Add step' button. The 'Add step' button is highlighted, indicating it's being used to add an AI-powered transformation. A tooltip for this step says 'Create dataflow transforms with Copilot. Describe the data transformation you want, in your own words, and Copilot will create it quickly.' To the right of the dataflow is a 'Copilot Preview' pane. It shows the 'Query settings' section with 'Name: Orders' and the 'Applied steps' section which contains a 'Split column by...' step and an 'Open AI - DissatisfactionReason...' step. The 'Copilot Preview' pane also contains a text input field with the placeholder 'Bring orders and split location column by comma' and a note stating 'Your dataflow has been updated with two queries: Orders - Shipping and Orders.' Below this, there are several green callout boxes with text such as 'Identify the dissatisfaction reason from CustomerReview, where the reasons include "Product arrived late", "Product was damaged", "Product was defective", "Delivered to incorrect address"', 'Done - dataflow updated.', 'A new DissatisfactionReason column was created with the dissatisfaction categories extracted from the CustomerReview column.', 'Keep it', and 'Ask a question or type / for suggestions'. At the bottom of the pane, it says 'AI-generated content can have mistakes. Make sure it's accurate and appropriate before using it. [Read preview terms](#)' and includes 'Step', 'Publish', and 'Data destination' buttons.

Column	Header	Content					
1	123 Ord	ABC City	ABC C	ABC CustomerReview	ABC DissatisfactionRea	OrderDate	Due
1	71774	Kansas City	USA	Product was great quality, but it arrived later than expected. Never got any update from the shipping company as for the reasons for this delay.	Product shipment arrived late	6/1/2022, 12:23:01 PM	6/8/2020
2	71776	Casper	USA	Product packaging was damaged. Either a dog tried to eat it or the shipping service didn't treat it with care.	Product packaging was damaged	2/24/2023, 08:34:25 PM	3/01/20
3	71780	Houston	USA	This package just showed up on my neighbor's house front porch. I am thankful she knows my name and gave it to me.	Delivered to incorrect address	1/19/2023, 04:20:19 PM	1/26/20
4	71782	Saint Ann	USA	Product doesn't work – Tried replacing batteries but no luck. Do we have any warranty?	Product was defective	1/29/2022, 07:03:16 PM	2/6/2022
5	71783	Escondido	USA	Product never arrived to me, although it says delivery was completed on time.	Delivered to incorrect address	2/6/2023, 12:00:43 PM	2/13/20
6	71784	Trabuco Canyon	USA	Product is smashed – Need to return it, can someone from your company pick it up at my home address?	Product was defective	3/5/2023, 06:19:09 PM	3/11/20
7	71796	Dallas	USA	I recently purchased a new laptop through your company, and it works well, but battery doesn't last for more than a couple of hours. This is a deal breaker for me as I am usually on the road and need to use without power.	Product was defective	11/30/2022, 04:40:12 PM	12/7/20
8	71797	Seattle	USA	Does your shipping company understand how to provide a good customer service? Very unhappy with the delivery service as the package	Delivered to incorrect address	1/20/2023, 1:42:32 PM	1/27/20

Completed (0.86 s) Columns: 20 Rows: 99+

Copilot skills

Create new transformations

- Add a single transformation step to an existing query
- Chain multiple transformation steps to an existing query

Create a new query

- Create a new query from scratch or by referencing existing data

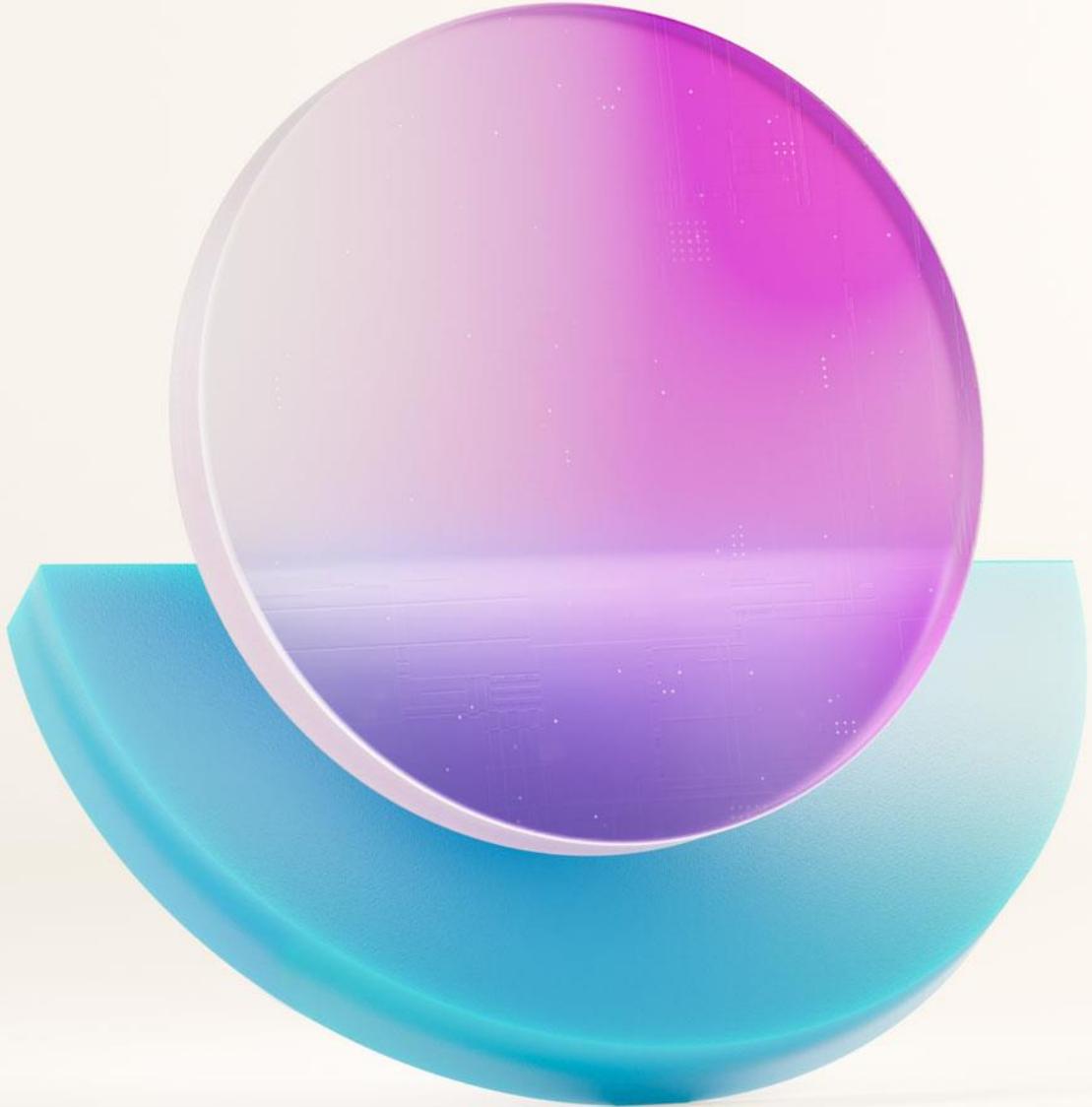
Explain a single step

- Generates a description of a transformation step

Explain the full query

- Generates a description of a complete query and its steps



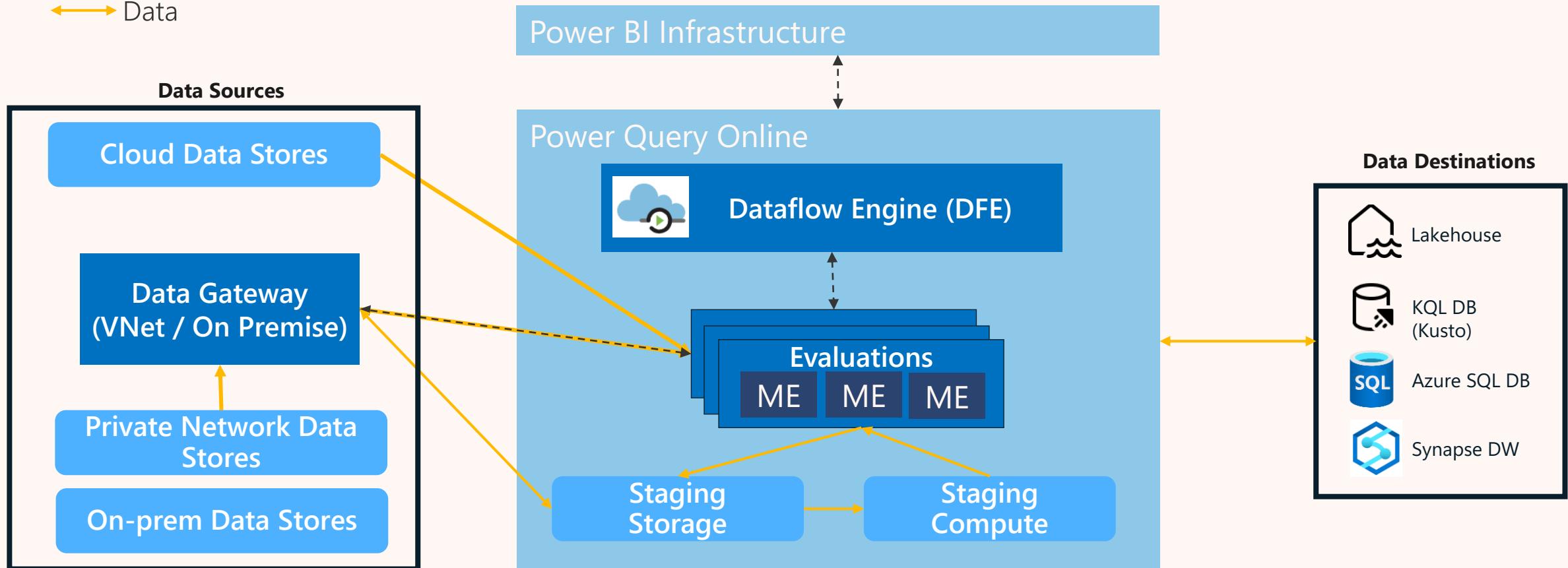


Dataflows In-Depth

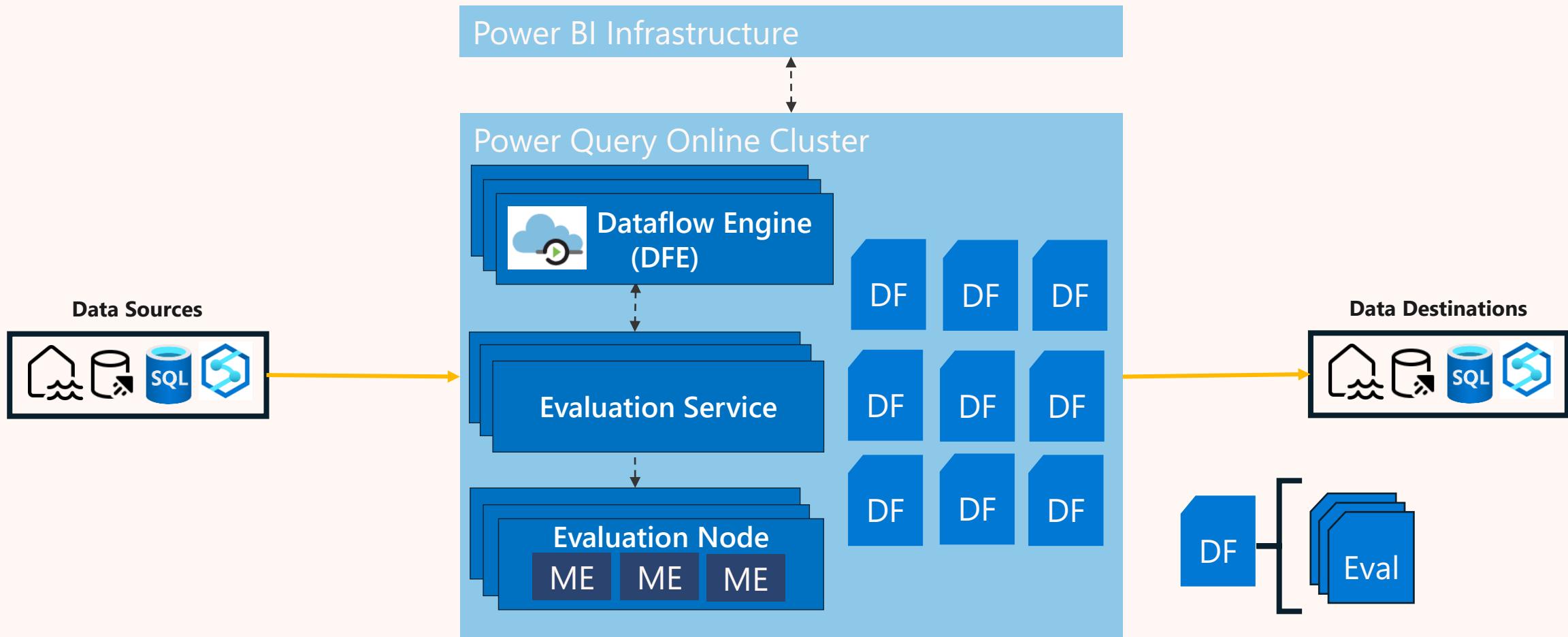
Fabric Dataflows (aka Gen2)

↔ Command and Control

↔ Data

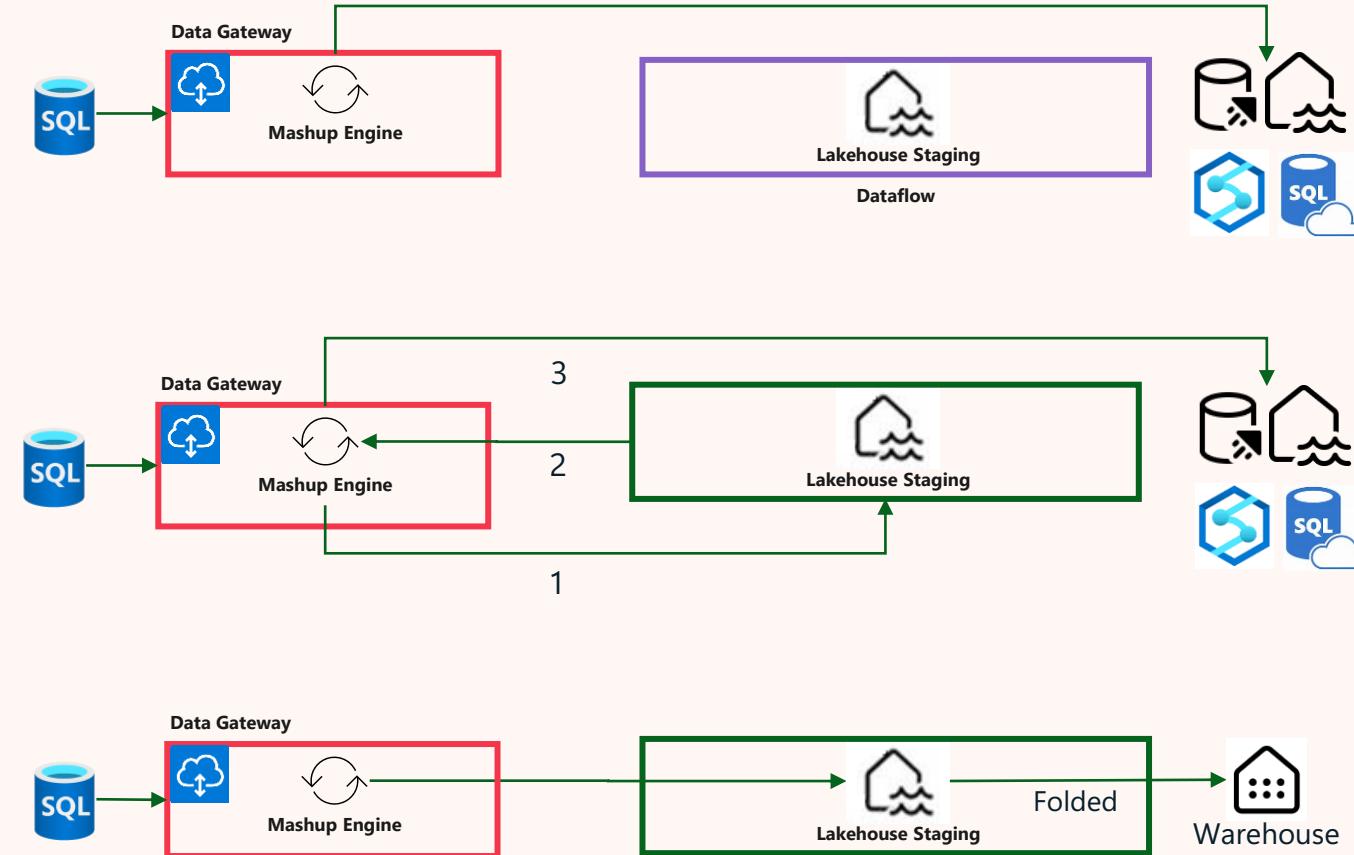


Fabric Dataflows Scale Out



Data Gateway (VNET / On Premise)

- Capabilities
 - Enables access to data sources without exposing them to the internet
 - Data is processed locally
- When to use it
 - Data source is inside your private network
 - Move compute closer to source
 - Control where evaluations happen
- When to skip it
 - Entities that stage data and write to destination (excluding WH)
 - Split Ingest and Destination into separate dataflows (consider where to Transform)
 - Don't stage



Dataflows Performance Principles

Four Principles

- Delegate to the most capable resource
- Do the expensive work first
- Divide and conquer
- Do as little work as possible



Dataflows Performance Capabilities

Principle	Capability	High-Level Approach
Delegate	Query Folding	Send as much of the query as possible to the underlying data source
Expensive work first	Staging	For large volumes of data reused by multiple query, first land the shared data in staging
Divide & conquer	Fast Copy	Leverage the Copy Activity for parallelized, high-scale data movement prior to transformation (ELT)
	Partitioning	Divide expensive queries into sub-queries that can be executed in parallel
Be lazy	Incremental Refresh	During refreshes, process only the data that has changed
	Lazy Evaluation	Leverage the feature of the "M" language that minimizes the amount of data traversed

Query Folding – What it is

- Terms that are synonymous to Query Folding
 - Query Delegation
 - Query Push Down
 - Remote/Distributed Query Evaluation
- Wherever possible, the script in the Power Query Editor ("M") is translated to a "native query"
- The native query is then executed by the underlying data source.

<https://learn.microsoft.com/en-us/power-bi/guidance/power-query-folding>

Query Folding - Example

The screenshot shows the Power Query ribbon with several tabs visible. The 'Transform' tab is active, showing options like 'Copy preview data', 'Remove columns', 'Remove other columns', 'Duplicate column', and 'Add column from examples...'. Below the ribbon, there's a context menu for a column named 'Event' with options like 'Copy preview data', 'Number filters', 'Replace values...', 'Drill down', and 'Add as new query'. The 'Number filters' option is expanded, showing filter conditions: Equals, Does not equal, Greater than, Greater than or equal to, Less than, and Less than or equal to. At the bottom, there are more transformation options: 'Transform column', 'Group by...', 'Fill', and 'Unpivot columns'. A table below shows data for 'Event' types: Thunderstorm Wind, Hail, Flash Flood, and Winter Weather.

```
let
    Source = Sql.Databases("server"),
    Navigation1 = Source{[Name = "Samples"]}[Data],
    Navigation2 = Navigation1{[Name = "StormEvents"]}[Data],
    SelectColumns = Table.SelectColumns(Navigation2, {"State", "EventType", "DamageProperty"}),
    Filter = Table.SelectRows(SelectColumns, each [DamageProperty] > 0),
    Group = Table.Group(Filter, {"State", "EventType"}, {"AvgDmg": each List.Average([DamageProperty])}),
    Rename = Table.RenameColumns(Group, {"EventType", "Event"})
in
    Rename
```

Query Folding – SQL native queries

The Dataflow runtime translates the M query to the data source's native query language, pushing down as much work as possible to the backend. In this example, all transformations steps can be folded to SQL.

```
let
  Source = Sql.Databases("server"),
  Navigation1 = Source{[Name = "Samples"]}[Data],
  Navigation2 = Navigation1{[Name = "StormEvents"]}[Data],
  SelectColumns = Table.SelectColumns(Navigation2, {"State", "EventType", "DamageProperty"}),
  Filter = Table.SelectRows(SelectColumns, each [DamageProperty] > 0),
  Group = Table.Group(Filter, {"State", "EventType"}, {{"AvgDmg", each List.Average([DamageProperty])}}),
  Rename = Table.RenameColumns(Group, {"EventType", "Event"})
in
  Rename

SELECT      [State], [EventType] as "Event", AVG([DamageProperty]) as AvgDmg
FROM        StormEvents
WHERE       [DamageProperty] > 0
GROUP BY    [State], [Event]
```

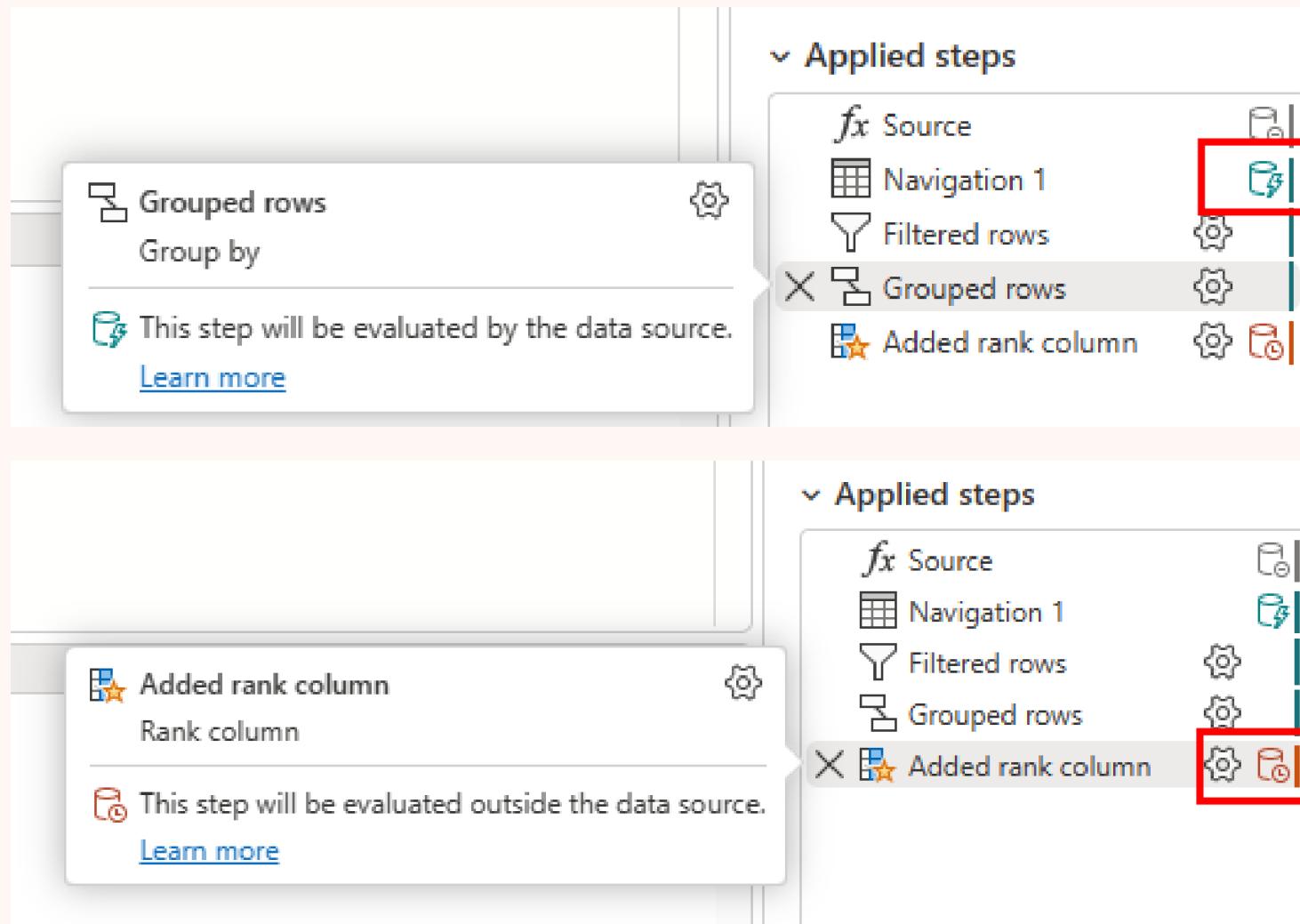
Query Folding – Indicators

Green “lightning bolt” indicator for folded steps

(typically, faster)

Red “clock” indicator for inmemory steps evaluated outside the data source

(typically, slower)



Query Folding – Why it's Faster

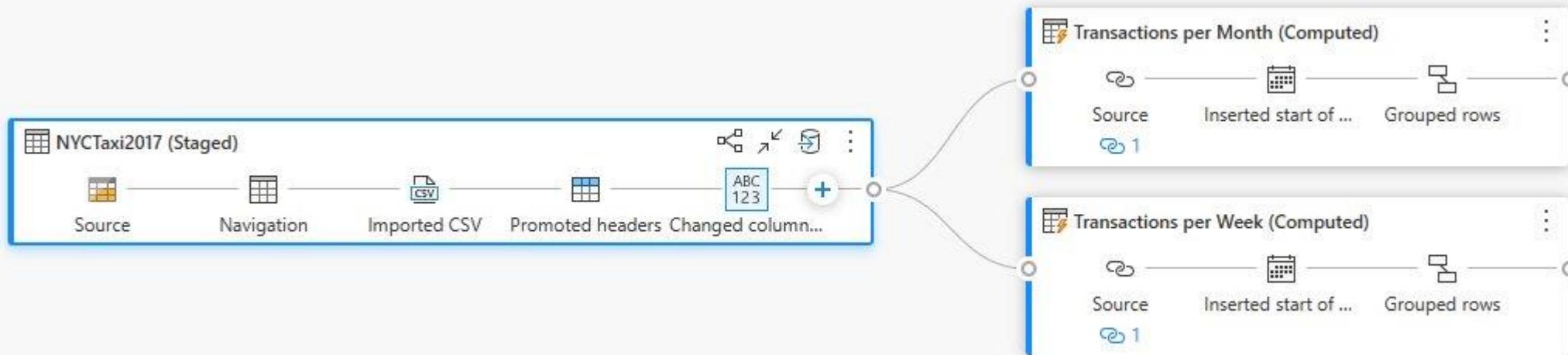
- Reduced data movement
 - In-memory evaluation engine would have to first transfer the data from the source
 - A filter that selects 10 rows from a 10M row table, can avoid transferring 10M rows
- Many data sources have highly efficient “query processors”
 - Optimized query plans
 - Ability to leverage indices, keys, etc.

Tips

1. Perform foldable operations early on
2. Do data type conversions last as they frequently break folding
3. Use the Table.StopFolding function to force local evaluation of subsequent steps

Staging – What it is

- Loading data into Fabric storage (Staging Lakehouse) as a first step
- The staged data can then be referenced by downstream queries that benefit from SQL Compute over the staged data



Staging - Tips

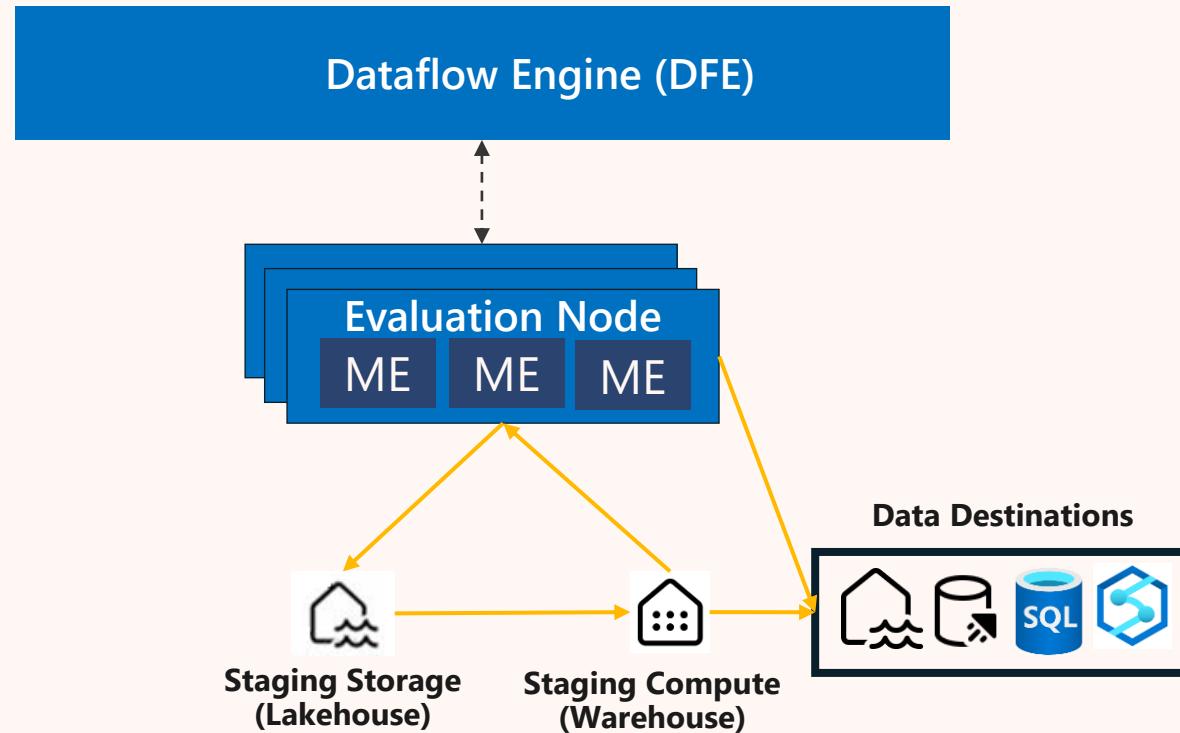
- **Data sources without rich query folding (e.g. files) are good candidates for staging**, especially when the files are large
- Data sources with rich query folding (e.g. databases) often do not require staging
- For a faster authoring experience create **separate dataflows for staging and transformation**
 - The transformation dataflow references the staging dataflow's data via the Dataflows connector

Staging – Why it's Faster

- In Fabric (Gen 2) Dataflows, staged data is stored in a Staging Lakehouse
- Queries against the Staging Lakehouse benefit from a SQL analytics endpoint
- SQL analytics endpoint queries can be orders of magnitude faster than directly querying “slow sources” like files (or SharePoint..)

File	Size	Direct Grouping Query	Staged Grouping Query
2020 Yellow Taxi Trip Data (CSV)	2.2 GB	4 mins 12 secs	2.8 secs
2017 Yellow Taxi Trip Data (CSV)	9.8 GB	Timeout (> 10 mins)	4.6 secs

Staging - Architecture



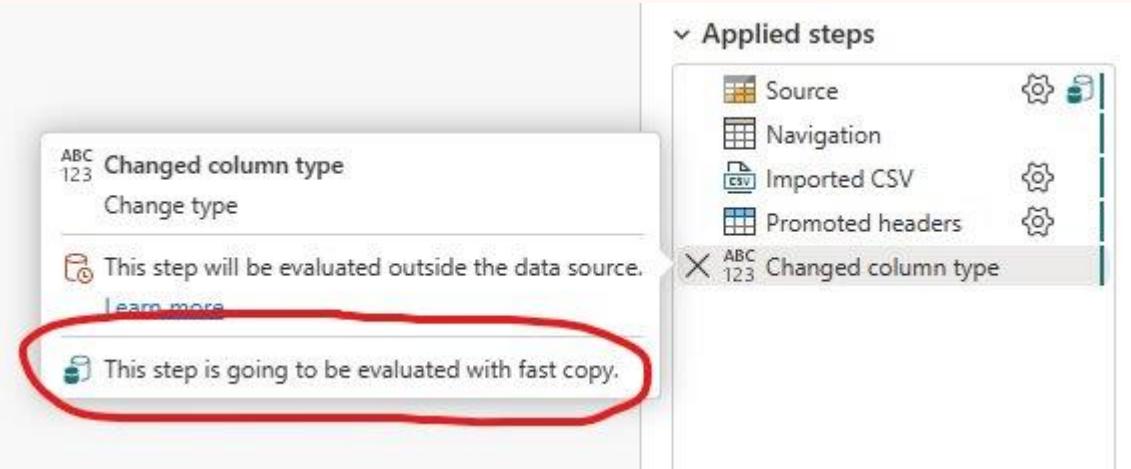
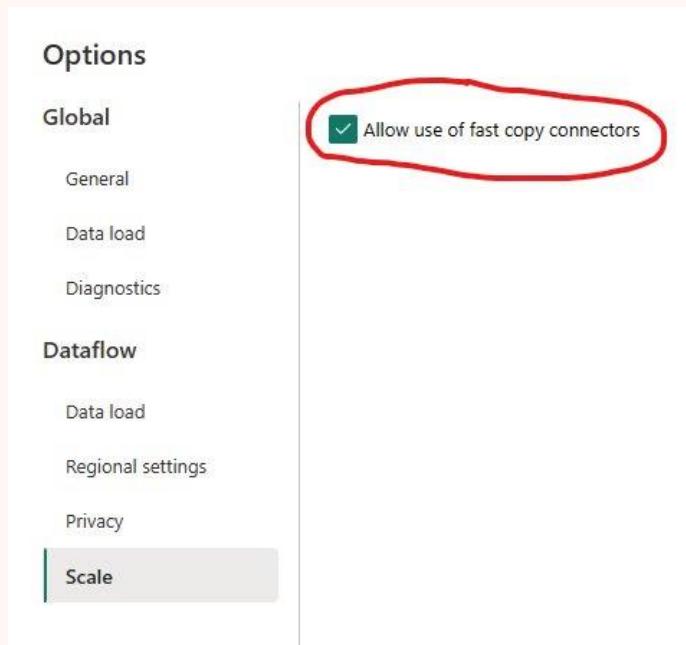
Fast Copy – What it is

Highly parallelized data movement

- Powered by Azure Data Factory Copy Tasks
- Standard Power Query experience

Step 2: Use Indicators in Steps pane to verify use of Fast Copy

Step 1: Enable Fast Copy



Step 3: Win

Details	
Postgres Test > 3/7/2024, 3:38:17 PM > public Address	
Name	Status
public Address	Succeeded
Start time	End time
3/7/2024, 3:38:27 PM	3/7/2024, 3:40:51 PM
Duration	Engine
00:02:24	CopyActivity

Fast Copy – Why it's Faster

Default is Serialized ($T_1 + T_2 + T_3 + T_4$)

Partition 1

Partition 2

Partition 3

Partition 4

Fast Copy is Parallel (T_4)

Partition 1

Partition 2

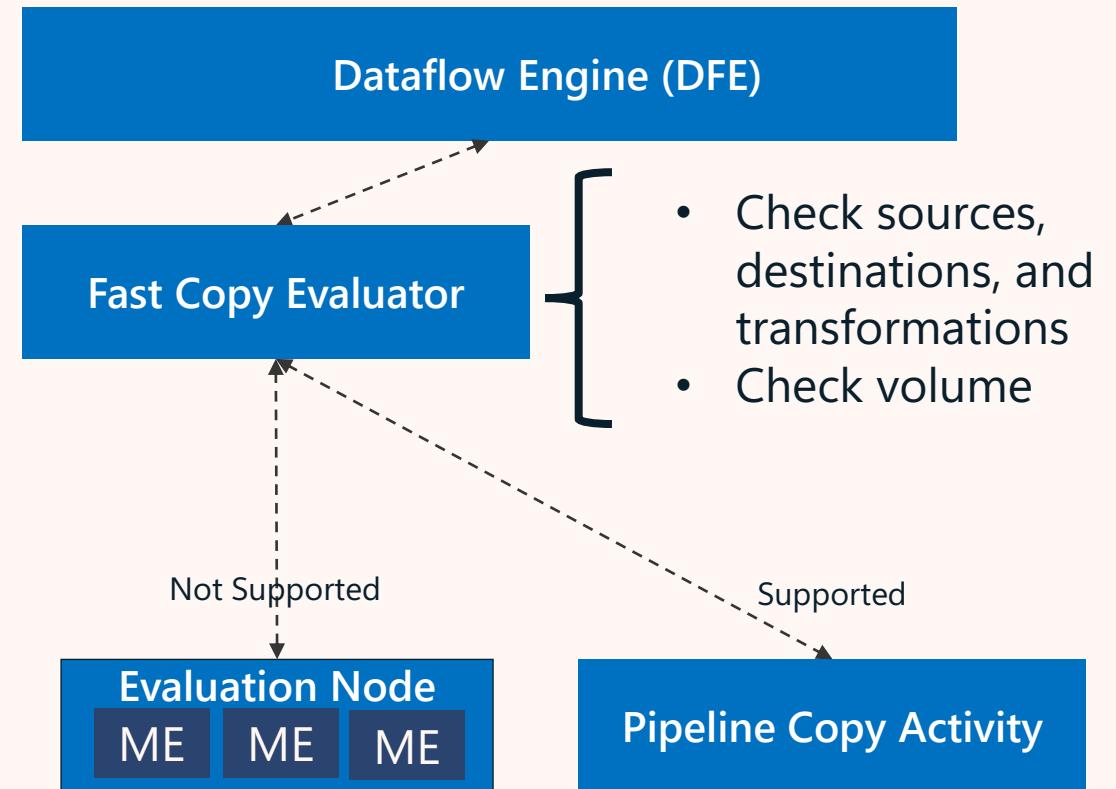
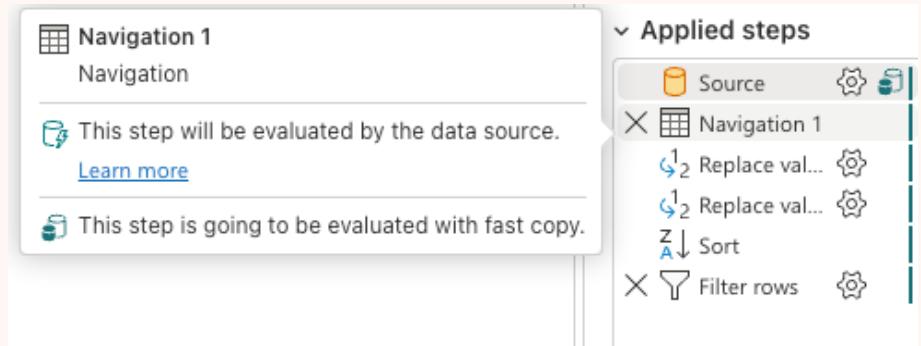
Partition 3

Partition 4

High degree of parallelism (up to 100)

Fast Copy - Architecture

- Capabilities
 - Leverages Pipeline Copy Activity for large performance boost in ingest
 - Automatically used based on pattern matching and volume
 - Transparent (no pipeline to manage)
- When to use it
 - Whenever possible
 - Defer transformations to post ingest if they affect Fast Copy use
 - Enable in Options..Scale..Allow use of fast copy connectors
- When to skip it
 - Don't – if it's an option, use it
 - Mark as "Require fast copy" to enforce



Partitioning – What it is

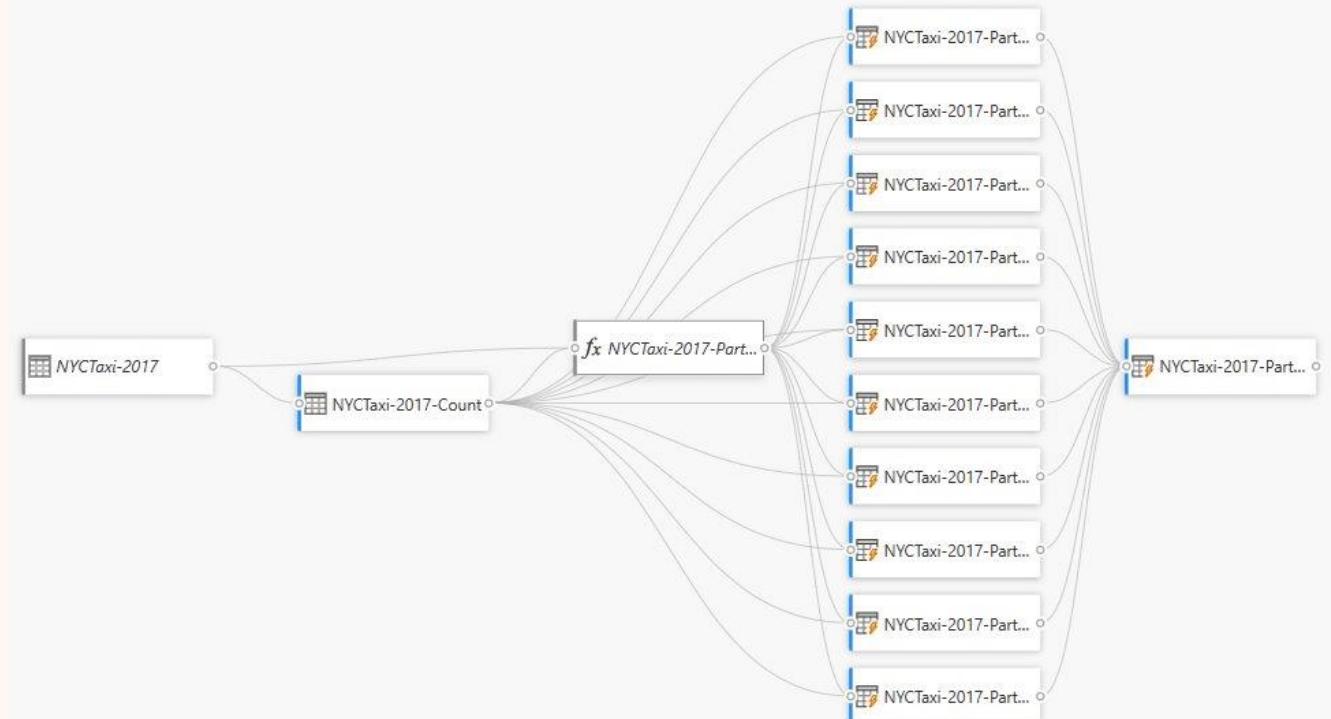
- Breaking long-running queries into smaller queries that can be run in parallel
- Explicit specification of what Fast Copy does implicitly
- Faster for the same reason as Fast Copy – parallelism

Partitioning – Sample

2017 Yellow Taxi Trip Data (9.8GB)

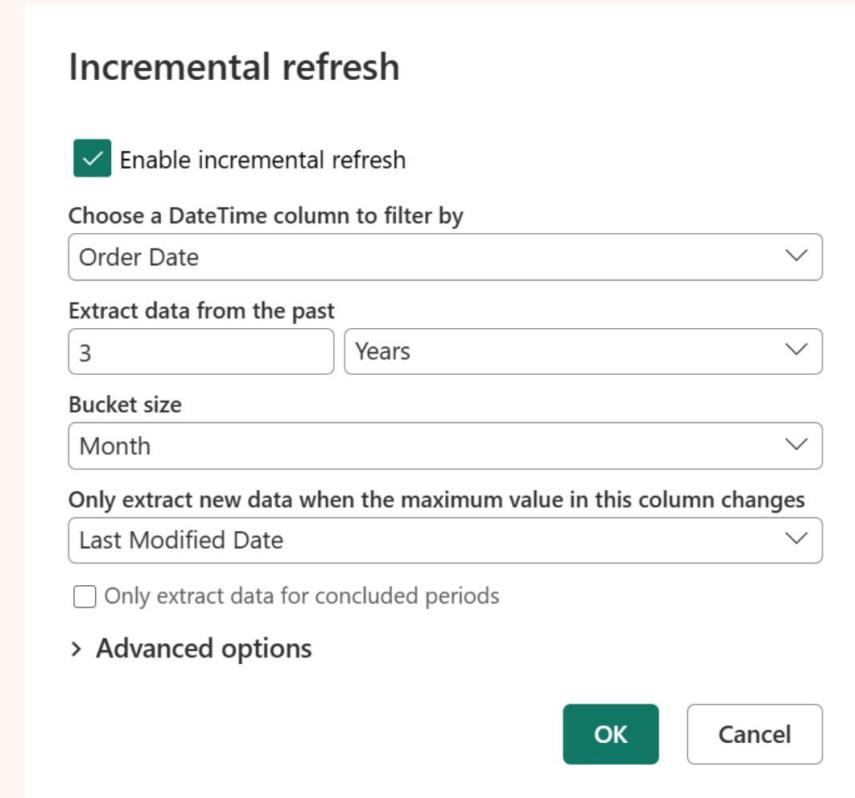
1. Calculate count (~113M rows)
2. Divide into 10 dynamic buckets using Table.Range (~11.3M rows per bucket)
3. Union the buckets

7x faster load time!



Incremental Refresh – What it is

- Specifying a timeframe
 - Example: Past 3 years
- Breaking a query into time-based buckets
 - Example: Month-based buckets based on “Order Date”
 - Bucket types: Days, Weeks, Months, Quarters, Years
- Specifying a “watermark” column
 - Example: Refresh a bucket only if “Last Modified Date” has increased since the last refresh
- Public Preview in first half of 2024!



Incremental Refresh – Why it's Faster

- Primarily about “doing less work”, but also embodies many of the other principles:
- Dividing and Conquering (partitions)
 - Queries are broken into sub-buckets (partitions) that are processed in parallel
 - Even though the initial load processes all data, it benefits from parallelization
- Doing Less Work
 - If a bucket has not changed (based on the “high watermark check”), it is not processed
- Delegating (query-folding)
 - The check to see whether a bucket has changed is a filter that is folded to the underlying data source

Leveraging the Architecture

- Carefully consider ETL vs ELT when laying out dataflows
- Know why you are using staging
 - Accelerates some operations, but adds no value to others
- Avoid “double-hops” with Data Gateways
- Take advantage of Fast Copy wherever possible
 - As patterns are added, dataflows can automatically benefit
 - Enable “use fast copy connectors”
 - Ensure that queries fully fold to maximize fast copy usage
- Build for parallel processing
 - Evaluations, dataflows
 - Incremental refresh when available

New Features for DI at #FabCon



GENERAL AVAILABILITY

VNET Data Gateway support with Private Links for Dataflows Gen 2 in Fabric

PREVIEW

Data Pipelines access on-premises data using “On Premises Data Gateway” (OPDG)

Fast Copy for Dataflows

40 to 80 activity limit in Data Pipelines

Semantic Model Refresh

CI/CD in Data Pipelines

Cancel Dataflow Refresh

SPN support for VNET Data Gateway

Modern Get Data – browse Azure Connections

Dataflow output destinations – Support for schema changes for Lakehouse & Azure SQL DB

SNEAK PEAK

Incremental Refresh for Dataflows

Resources

- <https://blog.fabric.microsoft.com/en-us/blog/data-factory-spotlight-dataflows-gen2>
- <https://blog.fabric.microsoft.com/en-us/blog/modern-get-data-in-dataflows>
- <https://blog.fabric.microsoft.com/en-us/blog/service-principal-support-to-connect-to-data-in-dataflow-datamart-dataset-and-dataflow-gen-2>
- <https://blog.fabric.microsoft.com/en-us/blog/dataflows-gen-2-data-destinations-and-managed-settings>
- <https://blog.fabric.microsoft.com/en-us/blog/updates-to-default-data-destination-behavior-dataflow-gen-2>
- <https://blog.fabric.microsoft.com/en-us/blog/fast-copy-in-dataflows-gen-2>



Slides



https://github.com/BenniDeJagere/Presentations/{Year}/{YYYYMMDD}_{Event}

