

**WiFi:** FabConEurope  
**Password:** FabCon24



# EUROPEAN MiCROSOFT FABRiC

## Community Conference

STOCKHOLM 24-27 SEPTEMBER 2024

► JOIN THE CONVERSATION

#FABCONEUROPE





# Data Factory in a Day

with Microsoft Fabric

# Speakers



**Sunil Sabat**

Principal Program Manager,  
Microsoft



**Alex Powers**

Senior Program Manager,  
Microsoft



**Jeroen Luitwieler**

Senior Program Manager,  
Microsoft

# Agenda (times are approximate and will be fluid with the event)

## Morning

9:00 AM – 9:15 AM		Introduction & setup
9:15 AM – 10:30 AM	Alex, Sunil and Jeroen	Microsoft Fabric and Data Factory overview
10:30 AM – 11:00 AM		Break & Q&A
11:00 AM – 12:45 PM	Sunil, Alex	Data pipeline
12:45 PM – 2:00 PM		Break for lunch

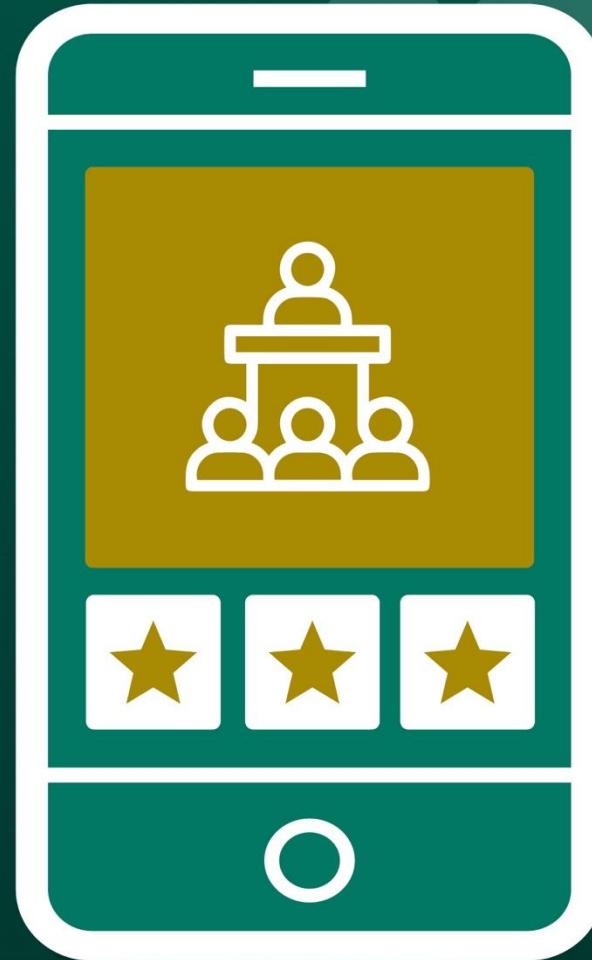
## Afternoon

2:00 PM – 3:15 PM	Jeroen, Alex	Dataflow Gen2
3:15 PM – 3:45 PM		Break & Q&A
3:45 PM – 5:00 PM	Alex, Sunil and Jeroen	Production patterns, Q&A and closing thoughts
5:00 PM		Conclusion

This workshop assumes that you have a working knowledge of accessing and creating content via the Microsoft Fabric



Please rate  
this session  
on the app



cvent



A collection of translucent, multi-colored geometric shapes (cubes, spheres, and hexagons) in shades of blue, green, yellow, and pink, arranged in a scattered, overlapping pattern against a white background.

# Session material and setup

# Attendee learning styles



## Visual

- Studies notes on presentations.
- Reads diagrams.
- Great sense of direction.
- Takes detailed notes.



## Auditory

- Enjoys story-telling.
- Understands changes in tone.
- Participates vocally.
- Engages in open discussions.



## Kinesthetic

- Learns by doing.
- Likes to explore.
- Gets satisfaction from building.
- Enjoys the “clicks” and “keys”.



## (Optional) Tutorial and trial

All materials available at:

<https://aka.ms/pbiworkshops>

Folder:

Data Factory in a Day

Fabric trial:

<https://aka.ms/try-fabric>



## (Optional) Demo user profile

Open a new incognito browser session:

**Ctrl+Shift+N**

Login with demo user profile.

Don't use Multi-factor authentication.



A collection of abstract, translucent 3D geometric shapes, including cubes, hexagons, and spheres, in shades of blue, cyan, and yellow, floating against a white background.

# Microsoft Fabric overview



# Microsoft Fabric

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



Data  
Factory



Data  
Engineering



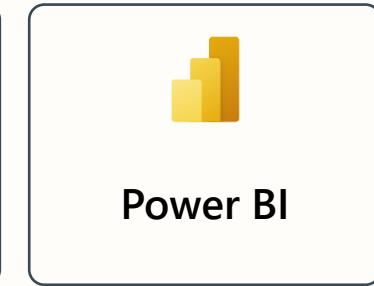
Data  
Warehouse



Data  
Science



Real-Time  
Intelligence



Power BI



AI-powered



OneLake



Purview

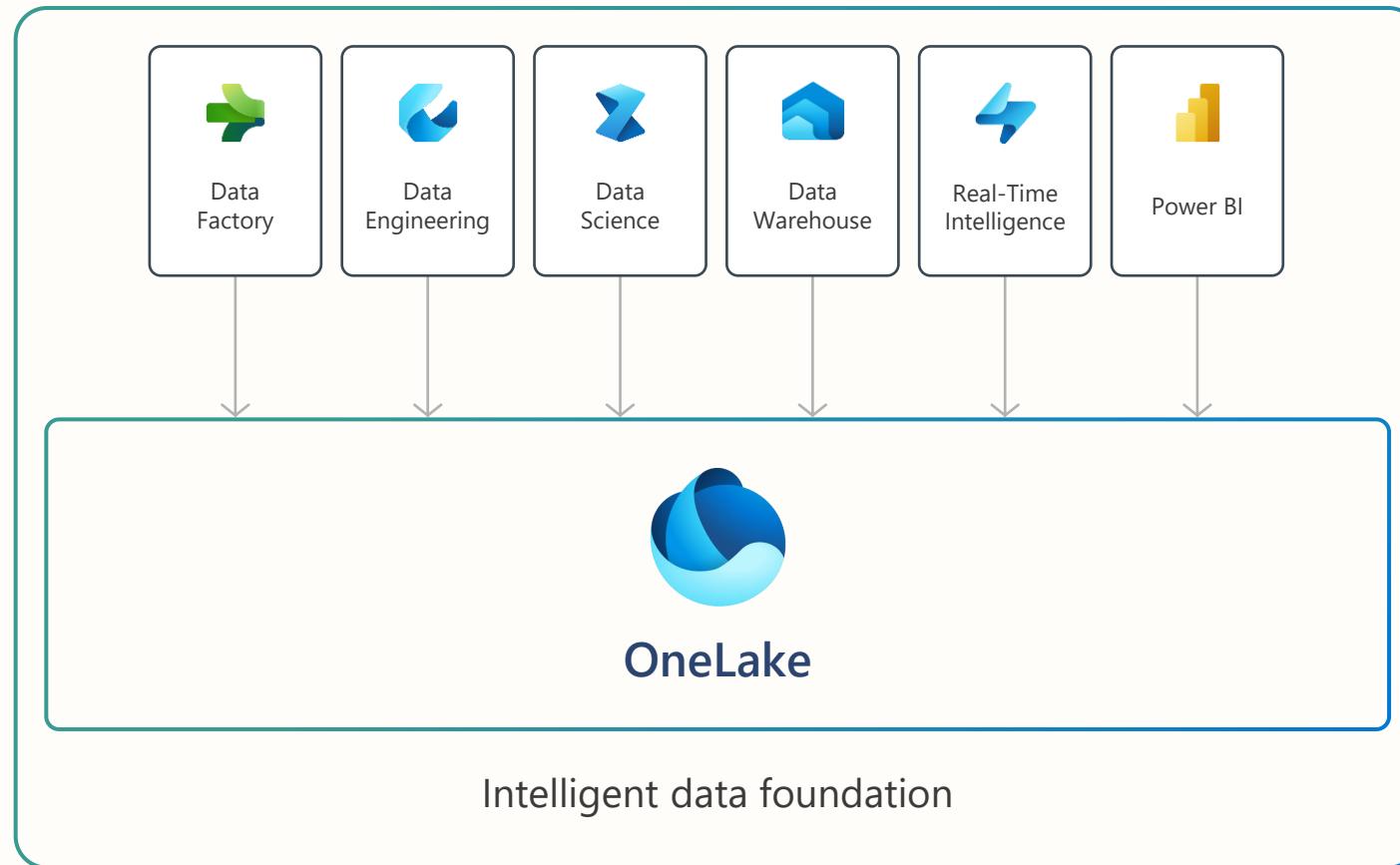
**“The OneDrive  
for Data”**



**OneLake**

# OneLake for All Data

“The OneDrive for Data”



A single SaaS lake for the whole organization

Provisioned automatically with the tenant

All workloads automatically store their data in the OneLake workspace folders

All the data is organized in an intuitive hierarchical namespace

The data in OneLake is automatically indexed for discovery, MIP labels, lineage, PII scans, sharing, governance and compliance



# Microsoft Fabric

The data platform for the era of AI

## Complete analytics platform

Everything, unified

SaaS-ified

Secure and governed

## Lake centric and open

OneLake

One Copy

Open at every tier

## Empower every business user

Familiar and intuitive

Built into Microsoft 365

Insight to action

## AI-powered experiences

Copilot accelerated

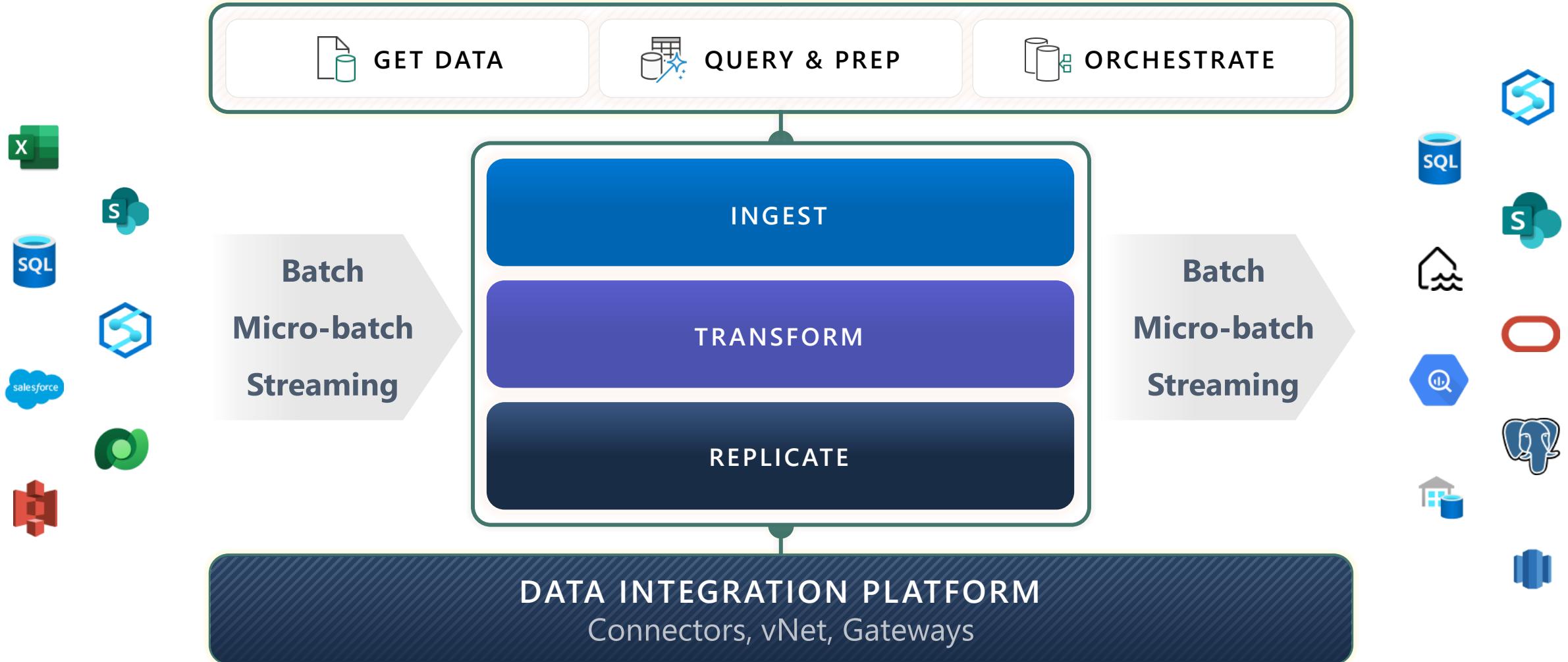
ChatGPT on your data

AI driven insights

A collection of translucent, multi-colored 3D geometric shapes, including cubes, spheres, and hexagons, arranged in a dynamic, overlapping composition against a white background. The colors range from blue and purple to green, yellow, and pink. Some shapes have internal circuit board patterns visible through their translucent surfaces.

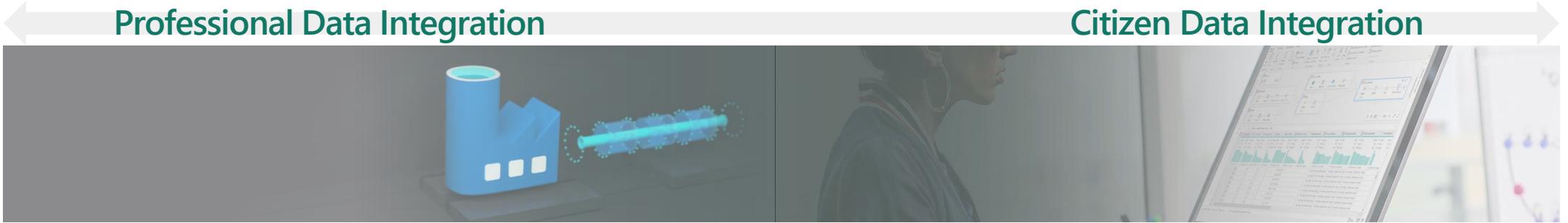
# Data Factory overview

# Microsoft Data Integration



# Microsoft Data Integration

## Products



### Azure Data Factory, Azure Synapse Analytics, SQL Server Integration Services

- Fully managed, with serverless data integration services
- Visually integrate data sources with more than 100 built-in connectors
- Easily construct ETL and ELT processes code-free

### Power Query

- Seamlessly integrated into many popular Microsoft products
- An easy to use, engaging, no-code experience
- Includes powerful and smart AI-based data preparation

# Microsoft Data Integration

## Unified Product Portfolio

### Professional & Citizen Data Integration

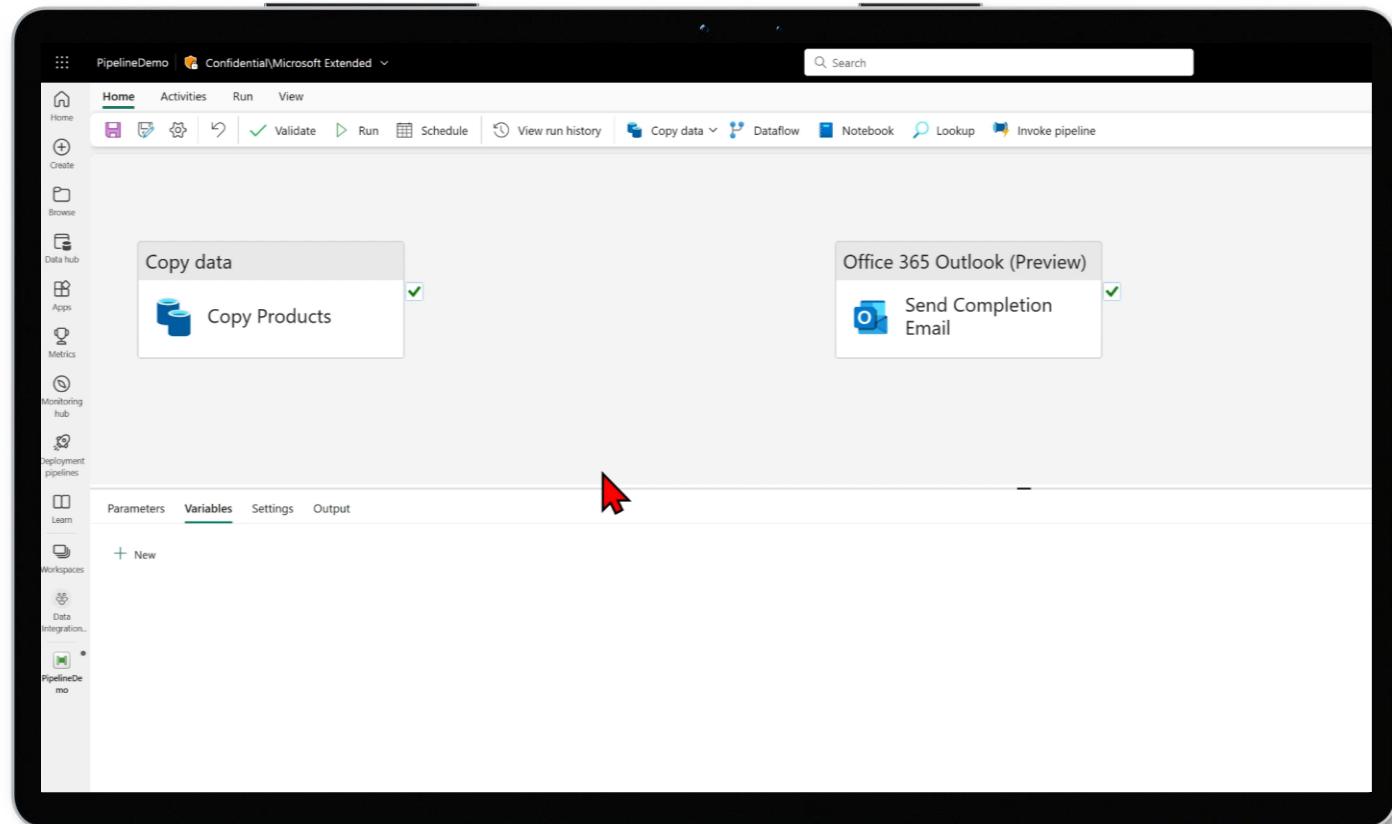


### Data Factory in Microsoft Fabric

- Brings together the best of Power Query and Azure Data Factory, into a modern data integration experience
- Empowers both professional and citizen developers
- Ingest and transform data as well as orchestrate data workflows
- Data Factory enables everyone to connect to diverse data sources and to bring that data to where it can best help derive insights for better business decisions.

# Data Factory in Microsoft Fabric

Data Factory converges our data capabilities into a single SaaS interface to provide the world's most complete data integration experience.





# Data pipeline

## Ingest and orchestrate activities at scale

Familiar authoring canvas experience

myPipeline | Confidential\Microsoft Extended

Search

Trial: 59 days left

Home Activities Run View

Copy data Dataflow Notebook Get metadata Lookup Script Stored procedure Set variable If condition

Start building your data pipeline

Add pipeline activity Copy data Choose a task to start



OneLake



# Data pipeline

## Ingest and orchestrate activities at scale

Empower every person to integrate data

The screenshot shows the Microsoft Fabric Data Pipeline interface. The top navigation bar includes Home, Activities, Run, View, and several action buttons like Validate, Run, Schedule, Copy data, Dataflow, Notebook, Lookup, and Invoke pipeline. On the left, a sidebar lists Home, Create, Browse, Data hub, Apps, Metrics, Monitoring hub, Deployment pipelines, Learn, Workspaces, Data Integration..., and PipelineDe... (with a red dot). The main area displays two selected activities: 'Copy data' (Copy Products) and 'Office 365 Outlook (Preview)' (Send Completion Email). Below these, tabs for Parameters, Variables, Settings, and Output are visible, along with a '+ New' button. A red arrow points from the pipeline interface towards the right side of the slide.



OneLake



Public Preview

# Copy job

## Easily ingest and move data at scale

Supports both batch and near real-time incremental copy (CDC)

The screenshot shows the Microsoft Fabric Data factory interface. In the top navigation bar, there are links for Home, Create, Browse, OneLake data hub, Monitoring hub, Workspaces, and My workspace. The main area has a "New" section with icons for Dataflow Gen2 (Preview), Data pipeline, Copy job, Apache Airflow project, and Data Factory mount. Below this is a "Recommended" section with cards for "Not sure where to start? Launch Dataflow Copilot", "Not sure where to start? Launch Pipelines Copilot", "Introduction to data integration Watch video", "Getting started with dataflow Watch the tutorial video", and "Getting started with data pipelines Watch the tutorial video". At the bottom, there is a "Quick access" table with columns for Name, Type, Opened, Owner, Endorsement, Sensitivity, and Workspace. The table lists four items: Cool data mart (Datamart, 7m ago, Tim Deboar, —, General, Contoso workspace), Data flow for triggers (Data flow, 7m ago, Tim Deboar, —, —, Contoso workspace), User data (Datamart, 7m ago, Tim Deboar, Certified, —, Contoso workspace), and Copy data pipeline (Pipeline, 7m ago, Tim Deboar, —, —, Contoso workspace).

Name	Type	Opened	Owner	Endorsement	Sensitivity	Workspace
Cool data mart	Datamart	7m ago	Tim Deboar	—	General	Contoso workspace
Data flow for triggers	Data flow	7m ago	Tim Deboar	—	—	Contoso workspace
User data	Datamart	7m ago	Tim Deboar	Certified	—	Contoso workspace
Copy data pipeline	Pipeline	7m ago	Tim Deboar	—	—	Contoso workspace





# Dataflow Gen 2

## Ingest and transform data at scale

Enterprise-scale data ingestion and transformation

The screenshot shows the Microsoft Power Query Editor interface. On the left, the navigation pane includes sections for Home, Queries (13), Data staging, Data load, and Data transformation. A cursor points to the 'DimCustomer' step under 'Data transformation'. The main area displays a table with columns such as CustomerKey, GeographyKey, FirstName, MiddleName, LastName, BirthDate, MaritalStatus, Suffix, Title, EmailAddress, and Education. Each column has a corresponding histogram showing data distribution. The 'Query settings' pane on the right shows the step name 'DimCustomer' and its properties, including 'Source' set to 'Merged queries' and 'Applied steps' showing the 'Expanded DimGeography.raw' step.



OneLake



# Unifying data in OneLake

## Seamlessly connect to more than 170+ data sources

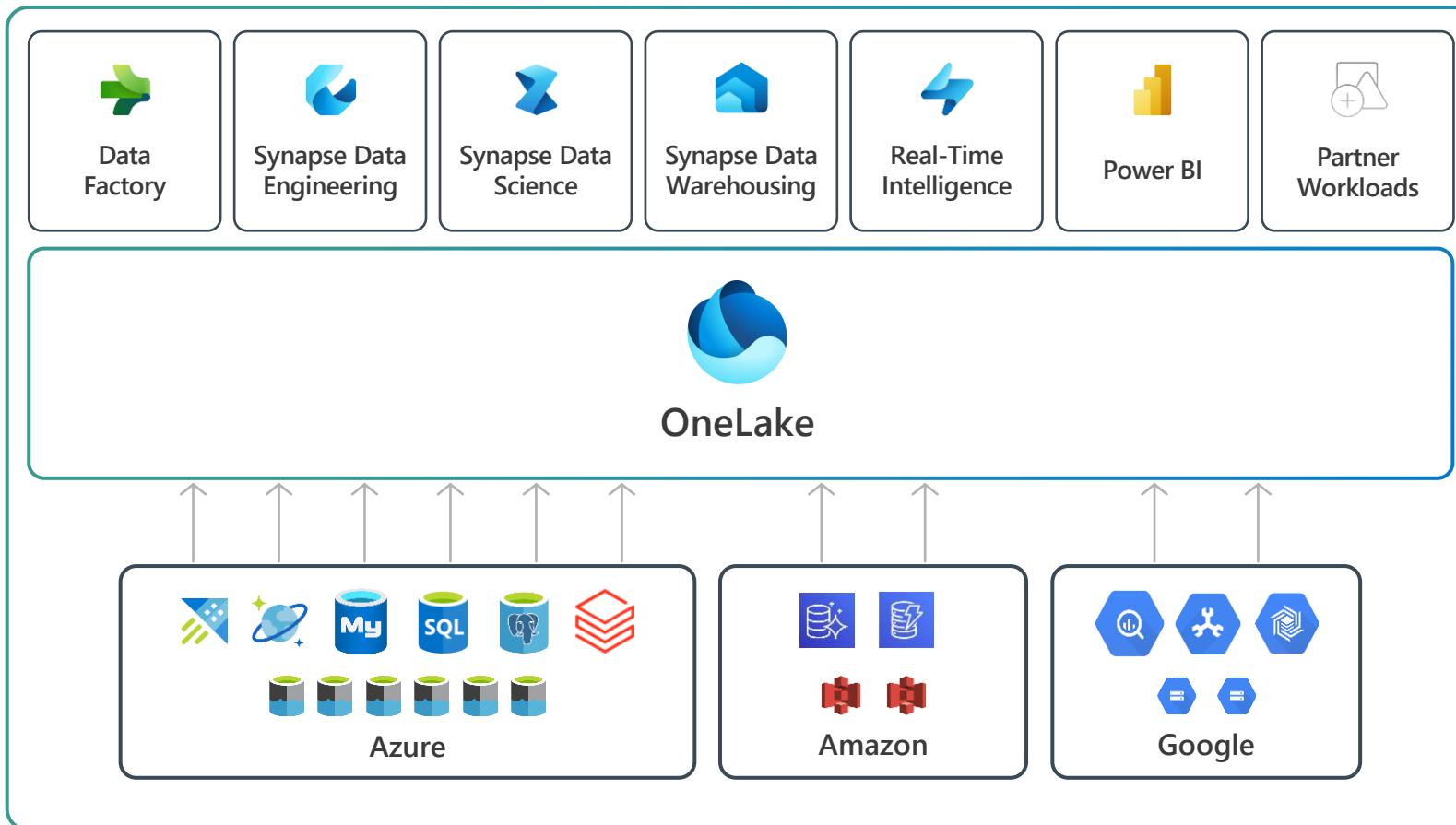


Azure Database for PostgreSQL	Azure Databricks Delta Lake	Amazon RDS for Oracle	Amazon RDS for SQL Server	Amazon Redshift	Phoenix	PostgreSQL	Presto	Magento (Preview)
Azure SQL Database	Azure SQL Database Managed Instance	Apache Impala	Azure SQL Database Managed Instance	DB2	SAP BW	SAP BW	SAP HANA	Oracle Eloqua (Preview)
Azure Table Storage	MongoDB Atlas	Drill	Google AdWords	Google BigQuery	SAP TABLE	SQL server	Spark	PayPal (Preview)
Azure Cosmos DB (MongoDB API)	Azure Cosmos DB (SQL API)	Greenplum	HBase	Hive	Amazon S3	Amazon S3 Compatible	FTP	SAP Cloud For Customer
Azure Data Lake Storage Gen1	Azure Data Lake Storage Gen1 for Cosmos Structured Stream	Informix	MariaDB	Microsoft Access	File system	Google Cloud Storage (S3 API)	HDFS	Salesforce Marketing Cloud
Azure Data Lake Storage Gen2 for Cosmos Structured Stream	Azure Database for MariaDB	MySQL	Netezza	Oracle	HTTP	Oracle Cloud Storage (S3 API)	SFTP	Shopify (Preview)
Teradata	Vertica	ODBC	OData	REST	Amazon Marketplace Web Service	Concur (Preview)	Dataverse (Common Data Service for Apps)	Web Table
Jira	Kusto	SharePoint Online List	Dynamics 365	Dynamics AX	Dynamics CRM	cassandra	Couchbase (Preview)	...



# Unifying data in OneLake

## Mirroring of External Databases



Frictionless linking of external databases, with full replicas created with a couple of clicks

Available for both multi-cloud and on-premises databases

Real time updates of the replicas using the CDC feeds of the database

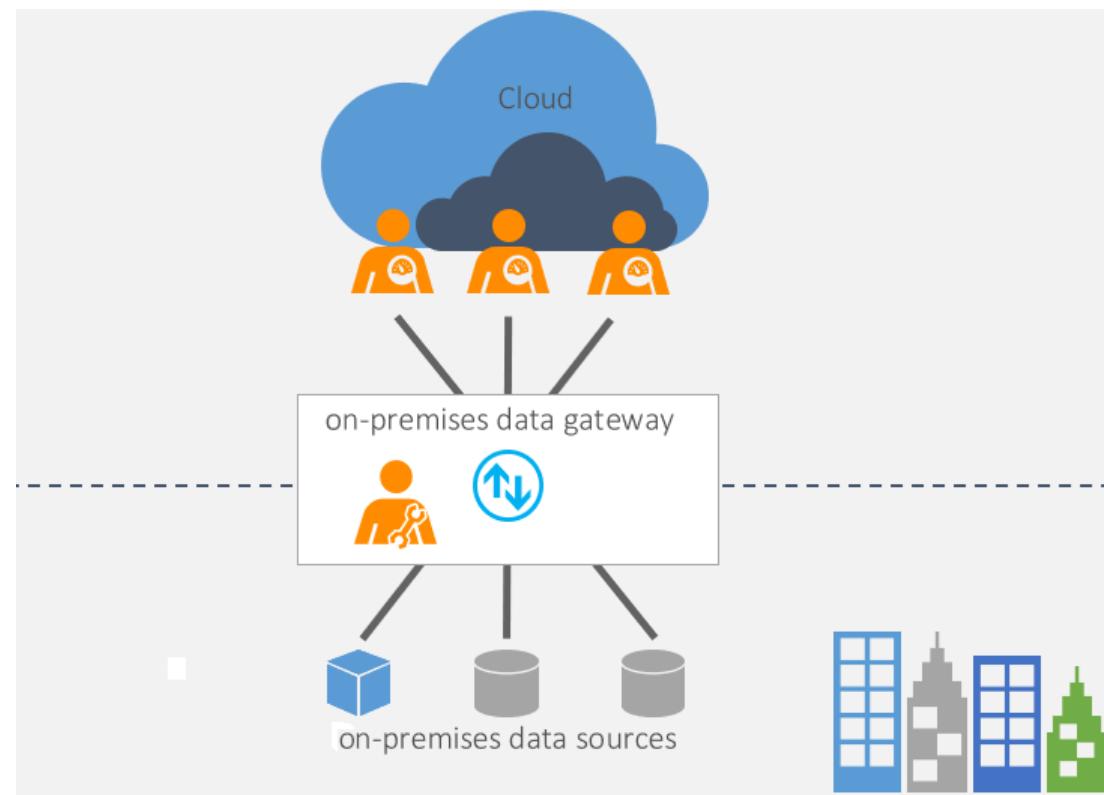
Data is stored in Delta Parquet tables, with all Fabric services instantly available



# On-premises data gateway

Secure connectivity to on-premises and private endpoints

Bridge between your secure data sources and the Microsoft cloud

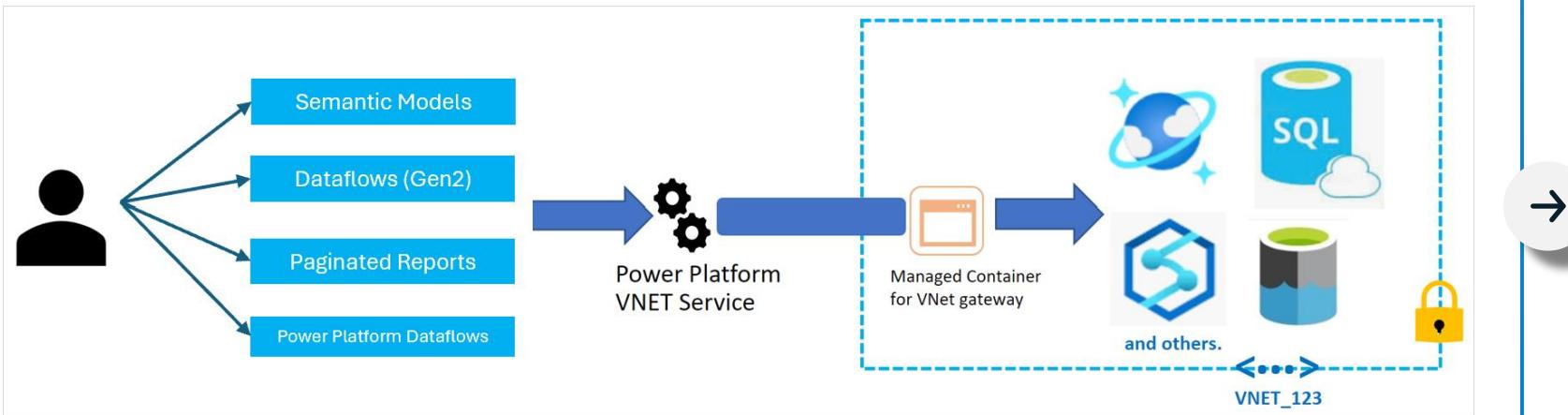




# VNet data gateway

Secure connectivity to your Azure and private data services

Bridge between your secure data sources and the Microsoft cloud





# Data integration + AI

## Copilot in Data Factory

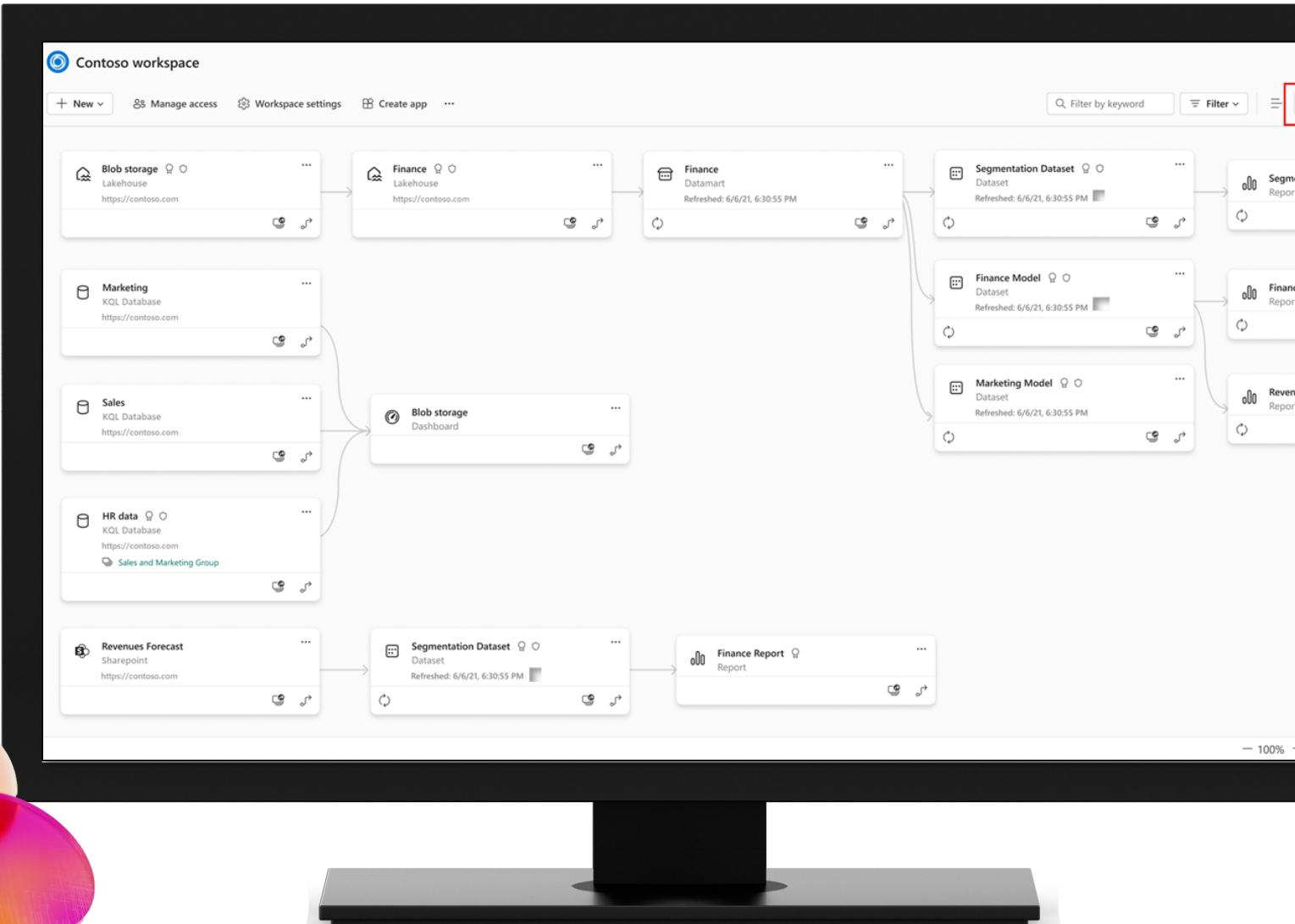
Easily integrate generative AI into your dataflows and \*pipelines using Copilot

- Chat with **Copilot** to transform data using natural language
- Tap into generative AI capabilities from **Azure Open AI** as data transformation steps
- Use **Copilot** to describe your data transformations

The screenshot shows the Microsoft Data Factory interface. On the left, there's a sidebar with navigation links: Home, Create, Browse, Data hub, Monitoring hub, Workspaces, My workspace, and Orders. The main area displays a dataflow named 'Orders'. It consists of two stages: 'Orders - Staging' (Source) and 'Orders' (Source). Between these stages is a 'Navigation' step. The 'Orders' stage includes 'Split column' and 'Open AI' options. To the right, the 'Copilot' pane is open, showing a preview of the dataflow. The 'Applied steps' section lists 'Reference' and 'Split column by...'. The 'Copilot' pane also contains several cards: 'Create dataflow transforms with Copilot' (describing the transformation), 'Bring orders and split location column by comma', 'Your dataflow has been updated with two queries: Orders - Shipping and Orders.', '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.', 'Ask a question or type / for suggestions', and 'AI-generated content can have mistakes. Make sure it's accurate and appropriate before using it. Read preview terms'. At the bottom, there are buttons for 'Step', 'Publish', and 'Data destination'.

# Lineage view

- Understand the relationships between analytical solutions
- Available automatically with every workspace
- Assess the potential impact of change on downstream items using impact analysis



# Monitoring hub

- Monitor activities from a central location
- Gain deep insights into how your Fabric items are performing – including dataflows and pipelines

The screenshot shows the Microsoft Fabric Monitoring hub interface. On the left, there is a vertical navigation bar with icons for Home, Create, Browse, Data hub, Analytics, Metrics, Monitoring hub (which is selected and highlighted in blue), and Deployment pipelines. The main area is titled "Monitoring hub" and contains a message: "Monitoring hub is a station to view and track active activities across different products." Below this is a button labeled "Last 7 days X". A table lists 12 completed activities from the past 7 days:

Activity name	Status	Item type	Start time	Submitter	Location
SMC StoryBoard	Completed	Dataset	10:29 AM, 9/4/23	Submitter	Workspace name
MSX1 - Sell-With - IP Co-Sell - BizApps Performa...	Completed	Dataset	10:29 AM, 9/4/23	Submitter	Workspace name
Data Refresh	Completed	Dataset	10:29 AM, 9/4/23	Submitter	Workspace name
Support Services Customer Report	Completed	Dataset	10:29 AM, 9/4/23	Submitter	Workspace name
MSX Insights - Azure Close Rate	Completed	Dataset	10:29 AM, 9/4/23	Submitter	Workspace name
Pipeline Management_PrepProdRefresh	Completed	Dataset	10:27 AM, 9/4/23	Submitter	Workspace name
All Training Consumption Aggregates	In progress	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name
SOXReports	Completed	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name
PendingRequest	Completed	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name
GPS Insights Hub - Partner Mastering Search	Completed	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name
Partner Parenting User History Page	Completed	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name
Microsoft_Github	Completed	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name
SOXReports_UAT_00202018	Completed	Dataset	10:26 AM, 9/4/23	Submitter	Workspace name

A cluster of abstract, translucent geometric shapes in shades of blue, green, and yellow, resembling floating crystals or data blocks, are positioned on the left side of the slide.

# Lakehouse overview

# Lakehouse overview

- Store, manage and analyze all your data in a single location and easily share across the entire enterprise
- Flexible and scalable solution that enables organizations to handle large data volumes of all types and sizes
- Built-in SQL endpoint unlocks Data Warehouse capabilities on top of your Lakehouse with no data movement
- Address the challenges of traditional Data Lakes by adding a **Delta Lake storage** layer directly on top of the cloud Data Lake



The screenshot shows the Microsoft Fabric Explorer interface. On the left, there's a sidebar with icons for Home, Create, Browse, OneLake data hub, Monitoring hub, Workspaces, My workspace, and TFLakehouse. The TFLakehouse icon is highlighted with a green bar at the bottom. The main area has a header with 'Home' and various navigation links like 'Get data', 'New Power BI dataset', and 'Open notebook'. A message at the top says: 'A SQL endpoint for SQL querying and a default dataset for reporting were created and will be updated with any tables added to the lakehouse.' Below this is the 'Explorer' section, which shows a tree view under 'TFLakehouse' with 'Tables' expanded, showing 'taxi\_zone\_lookup' selected, and 'Files' expanded, showing 'TaxiData'. To the right is a data grid titled 'taxi\_zone\_lookup' with the following schema and data:

	LocationID	Borough	Zone	service_zone
1	1	EWR	Newark Air...	EWR
2	132	Queens	JFK Airport	Airports
3	138	Queens	LaGuardia ...	Airports
4	264	Unknown	NV	N/A
5	265	Unknown	NA	N/A
6	4	Manhattan	Alphabet City	Yellow Zone
7	12	Manhattan	Battery Park	Yellow Zone
8	13	Manhattan	Battery Par...	Yellow Zone

# Automatic table discovery and registration

The screenshot shows the Microsoft Power BI Data Explorer interface. On the left, the sidebar includes icons for Home, Create, Browse, OneLake data hub, Monitoring hub, Workspaces, and Data Engineering. The main area has tabs for Home, Get data, New Power BI dataset, and Open notebook. The Explorer pane shows a tree view with 'NYCCabData' expanded, revealing 'Tables' and 'Files'. A folder named 'TaxiData' is selected. The main workspace displays a file list under 'Files > TaxiData'. The table has the following columns:

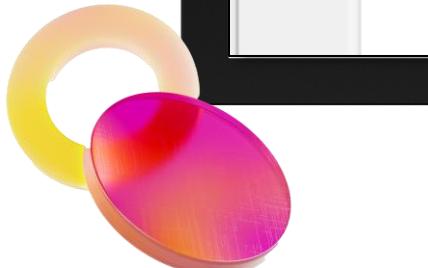
Name	Date modified	Type	Size
taxi_zone_lookup.csv	5/3/2023 11:30:02 AM	CSV	12 KB

# Automatic discovery of Delta Lake tables

The Lakehouse explorer provides a tree-like view of the objects in the Microsoft Fabric Lakehouse item.

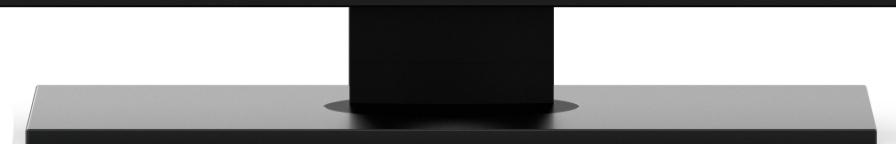
It has a key capability of discovering and displaying tables that are described in the metadata repository and in OneLake storage.

**Delta Lake** - Achieve seamless data access across all compute engines in Microsoft Fabric



A screenshot of the Microsoft Fabric Lakehouse Explorer interface. On the left, a sidebar shows navigation options: Create, Browse, OneLake data hub, Monitoring hub, Workspaces, My workspace, and TFLakehouse (which is selected). The main area is titled 'Explorer' and shows a tree view under 'TFLakehouse'. The 'Tables' node is expanded, showing a table named 'taxi\_zone\_lookup'. This table has four columns: LocationID, Borough, Zone, and service\_zone. Below the table, there are four rows of data. To the right of the table, there is a note: 'A SQL endpoint for SQL querying and a default dataset for reporting were created and will be updated with any tables added to the lakehouse.'

	LocationID	Borough	Zone	service_zone
1	1	EWR	Newark Air...	EWR
2	132	Queens	JFK Airport	Airports
3	138	Queens	LaGuardia ...	Airports
4	264	Unknown	NV	N/A
5	265	Unknown	NA	N/A
6	4	Manhattan	Alphabet City	Yellow Zone
7	12	Manhattan	Battery Park	Yellow Zone
8	13	Manhattan	Battery Par...	Yellow Zone
9	24	Manhattan	Bloomingd...	Yellow Zone



# Navigating the Fabric Lakehouse explorer

The screenshot shows the Microsoft Fabric Lakehouse explorer interface. On the left is a navigation sidebar with icons for Home, Create, Browse, Data hub, Monitoring hub, Workspaces, LHRedesign\_AviWS, ContosoDailySales, and Power BI. The main area has a header with 'Home' selected, 'Lakehouse' dropdown, and buttons for 'Get data', 'New Power BI dataset', and 'Open notebook'. A message bar at the top says: 'A SQL endpoint for SQL querying and a default dataset for reporting were created and will be updated with any tables added to the lakehouse. You can access the SQL endpoint using the dropdown.' Below this is the 'Lakehouse explorer' section, which lists 'ContosoDailySales' and 'Unidentified' datasets. Under 'ContosoDailySales', there are 'Tables' (Customer, inventory, product, sales, Transactions) and 'Files' (CustomerFeedbackAudio). The 'Customer' table is currently selected, showing a preview of 1000 rows. The columns are: Index, UserId, FirstName, LastName, Sex, Email, Phone, DateOfBirth, and JobTitle. The data includes rows for Jo Rivers, Sheryl Lowery, Lindsey Rice, Eddie Barnes, Ralph Lowe, Carly Abbott, Natasha Macias, Courtney Jenkins, Perry McMahon, Norman Walton, Roaer Sweenev, and leblanciohn... (partial data shown).

Index	UserId	FirstName	LastName	Sex	Email	Phone	DateOfBirth	JobTitle
1	3d5AD30A...	Jo	Rivers	Female	fergusonkat...	-10395	7/26/1931	Dancer
2	810Ce0F27...	Sheryl	Lowery	Female	fhoward@e...	(599)782-0...	11/25/2013	Copy
3	9afFEafAe1...	Lindsey	Rice	Female	elin@exam...	(390)417-1...	4/15/1923	Biomedical ...
4	CDA21B6e8...	Eddie	Barnes	Female	brandy23@...	801.809.91...	2/27/1975	Dramathera...
5	1CC30c5F2...	Ralph	Lowe	Female	dleon@exa...	+1-511-127...	4/10/1938	Presenter, b...
6	bFCFDdE54...	Carly	Abbott	Female	stricklando...	(416)979-0...	10/27/2007	Therapeutic...
7	aCeff56E59...	Natasha	Macias	Female	dorothyme...	(929)366-8...	10/31/1971	Recruitmen...
8	CF091D6b9...	Courtney	Jenkins	Female	estesana@...	(973)243-9...	1/20/1948	Accounting...
9	462EF46dca...	Perry	Mcmahon	Female	allison66@...	060-611-93...	11/24/2006	Education o...
10	3Cb9Fe3aB...	Norman	Walton	Female	samanthas...	(590)187-8...	6/19/1973	Personnel o...
11	be6BBa9EB...	Roaer	Sweenev	Female	leblanciohn...	-8153	9/9/2008	Race relatio...

# Lakehouse SQL endpoint

The Lakehouse creates a serving layer by automatically generating a SQL endpoint and a default dataset during creation

You can only read data from delta tables using SQL endpoint

Save functions, views, and set SQL object-level security



A screenshot of the Microsoft Fabric TFLakehouse interface. The top navigation bar includes 'TFLakehouse', a search bar, and links for 'Get data', 'New Power BI dataset', and 'Open notebook'. A red box highlights a message: 'A SQL endpoint for SQL querying and a default dataset for reporting were created and will be updated with any tables added to the lakehouse.' The left sidebar shows 'Home', 'Create', 'Browse', 'OneLake data hub', 'Monitoring hub', 'Workspaces', 'My workspace', and 'TFLakehouse'. The main area is titled 'Home' and contains an 'Explorer' section with 'TFLakehouse' expanded to show 'Tables' and 'Files'. To the right, there's a 'Get data in your lakehouse' section with three buttons: 'New Dataflow Gen2', 'New data pipeline', and 'Open notebook'. At the bottom, there's a decorative graphic of a monitor on a stand.

# Share your Lakehouse with consumers

- Provide users access to a Lakehouse without adding them to your workspace
- Use SQL Security or customer permission to grant access through SQL Endpoint
- Discover Lakehouses you have access to in the OneLake data hub
- Grant access to Lakehouse data through Spark, SQL Endpoint, and Power BI semantic models



Name	Type	Owner	Location	Refreshed	Endorsement	Sensitivity
ContosoDailySales	Lakehouse	Avinanda Chattapad...	LHRedesign_AvIWS	-	-	Confidential\Microsoft Ext...
Test	Lakehouse	Avinanda Chattapad...	Customer360WS	-	-	Confidential\Microsoft Ext...
Test2	Lakehouse	Avinanda Chattapad...	Customer360WS	-	-	Confidential\Microsoft Ext...
Customer360	Lakehouse	Avinanda Chattapad...	Customer360WS	-	-	Confidential\Microsoft Ext...
Test2	Warehouse (default)	Avinanda Chattapad...	Customer360WS	12/31/52, 4:07:02 PM	-	Confidential\Microsoft Ext...
Test2	Dataset (default)	Avinanda Chattapad...	Customer360WS	4/10/23, 1:52:44 PM	-	Confidential\Microsoft Ext...

# Open and accessible endpoints

Item	Connectivity options
Lakehouse	SQL analytics endpoint / Azure Blob File System (ABFS) / URL endpoint
Warehouse	SQL endpoint / OneLake availability
Event house	Cluster address (Kusto) / OneLake availability
Semantic model	Extensible Markup Language for Analysis (XMLA) endpoint address / OneLake availability
GraphQL	URL endpoint



# Table and column name validation and rules

Table names can only contain alphanumeric characters and underscores

Text files without column headers are replaced with standard col# notation as the table column names

Column names are validated during the load action and the action fails if column names are invalid.



# Interacting with the Lakehouse

## OneLake file explorer

Explore data in the Lakehouse using Windows File Explorer

## Notebooks

Use the Notebook to write code to read, transform and write directly to Lakehouse as tables and/or folders

## Data pipelines

Use pipeline copy tool to pull data from other sources and land into the Lakehouse

## Copy job

Develop robust applications and orchestrate the execution of compiled Spark jobs in Java, Scala, and Python

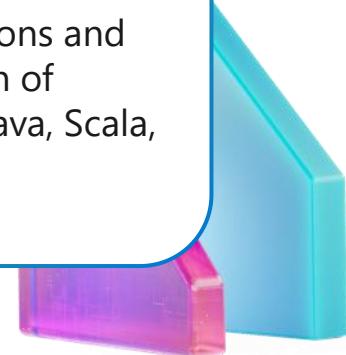
## Dataflows Gen 2

Use Dataflows Gen 2 to ingest and prepare the data

## Apache Spark job definitions

Develop robust applications and orchestrate the execution of compiled Spark jobs in Java, Scala, and Python

and more...



# Medallion Lakehouse Architecture

Design and implementation guidance

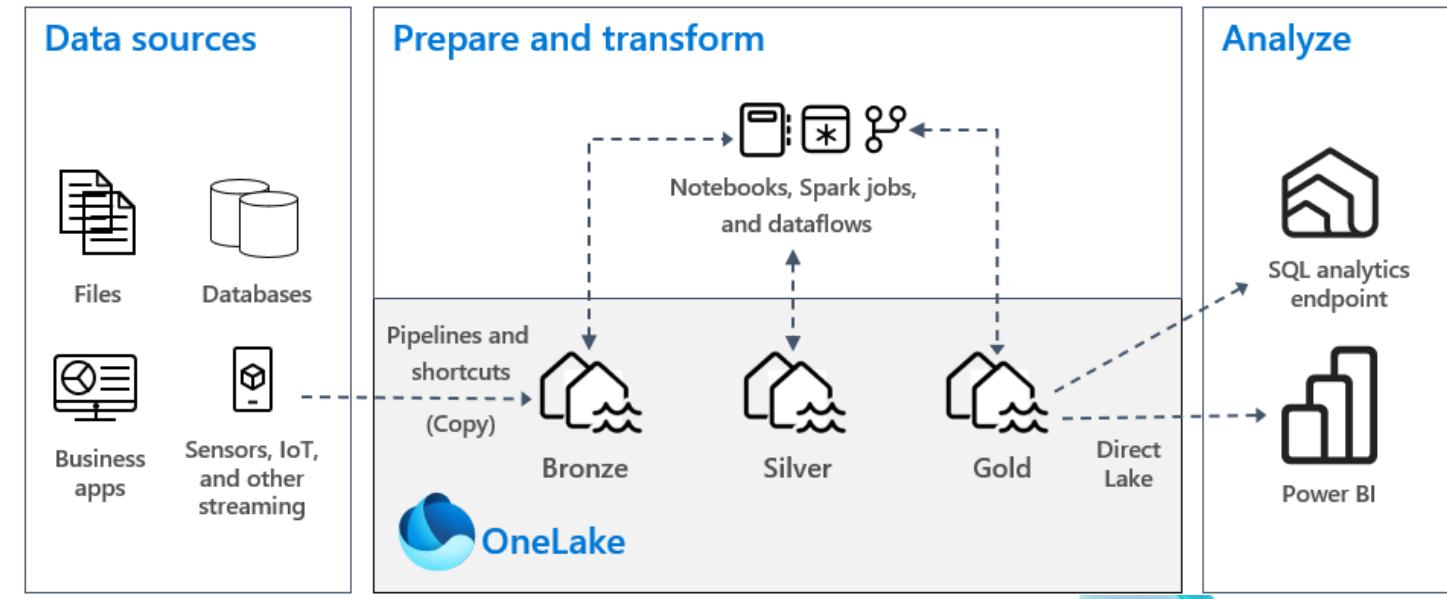


# Medallion Lakehouse architecture

Build a single source of truth for enterprise data products with a multi-layered approach.

Medallion architecture has three layers that represent different data quality levels: **bronze** (raw data), **silver** (validated data), and **gold** (enriched data).

As data moves through these layers, it undergoes validations and transformations to ensure it meets the ACID properties (Atomicity, Consistency, Isolation, and Durability) and is optimized for analytics.



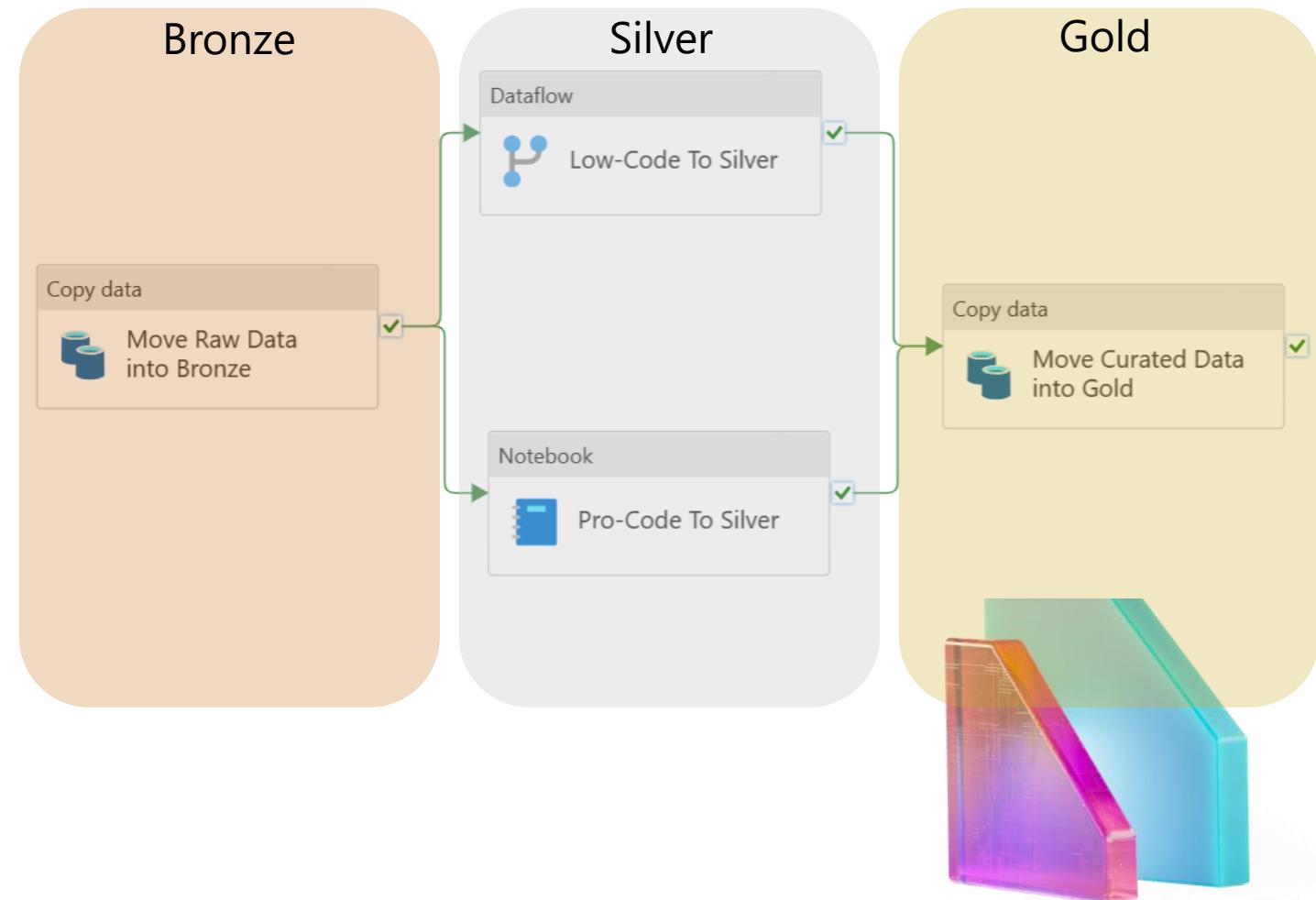
In a typical medallion architecture implementation in Fabric, the bronze zone stores the data in the same format as the data source. When the data source is a relational database, Delta tables are a good choice. The silver and gold zones contain Delta tables.



# Medallion Lakehouse architecture

Distinct layers (or zones).

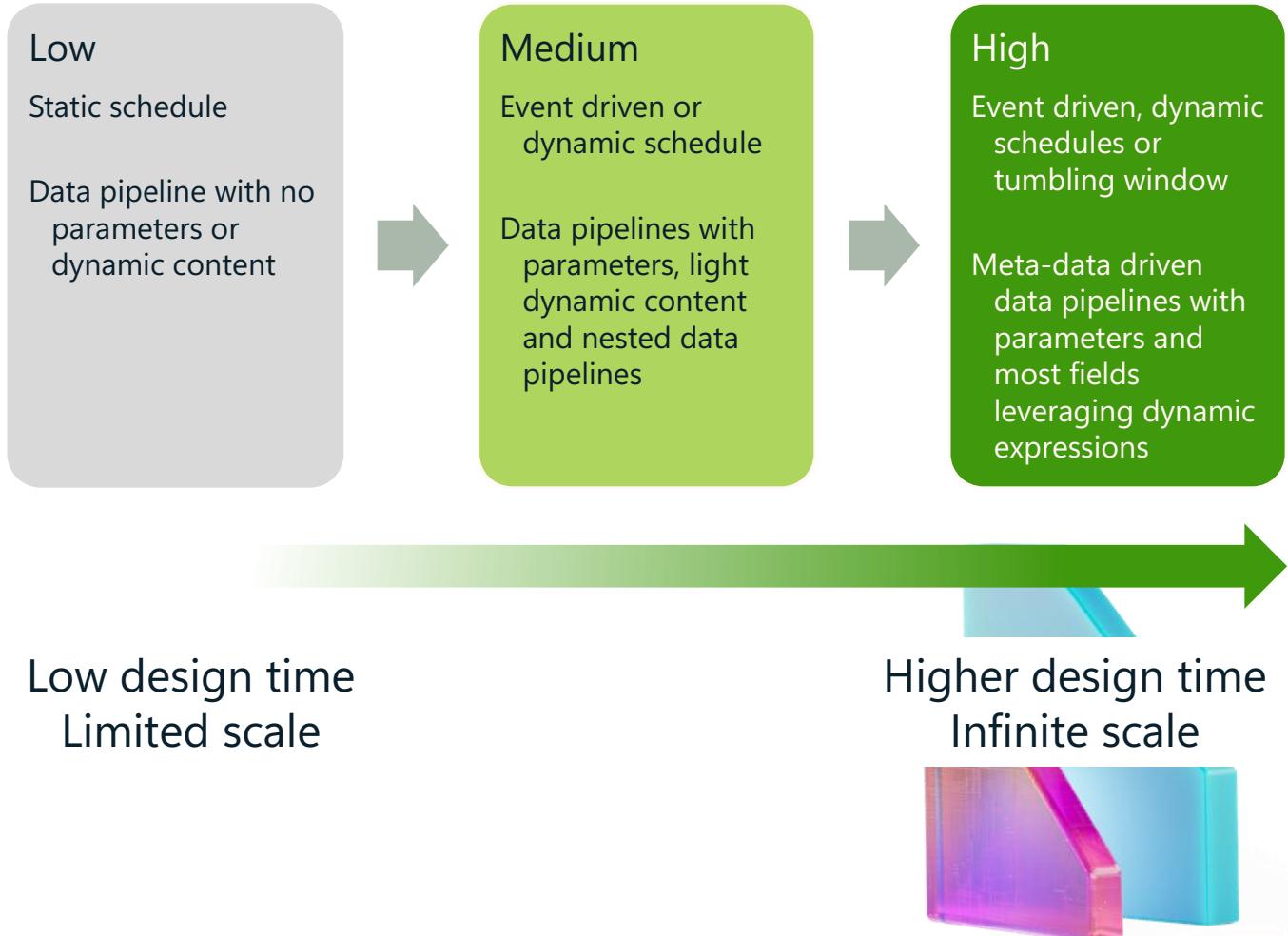
- **Bronze (raw zone):** This layer stores source data in its original format, which is typically append-only and immutable.
- **Silver (enriched zone):** Data from the bronze layer is cleansed, standardized, and structured as tables. It may also be integrated with other data to provide a comprehensive view of business entities like customers and products.
- **Gold (curated zone):** Data from the silver layer is refined to meet specific business and analytics requirements. Tables in this layer usually follow a star schema design, optimizing them for performance and usability.



# Design and scale options

Distinct layers (or zones).

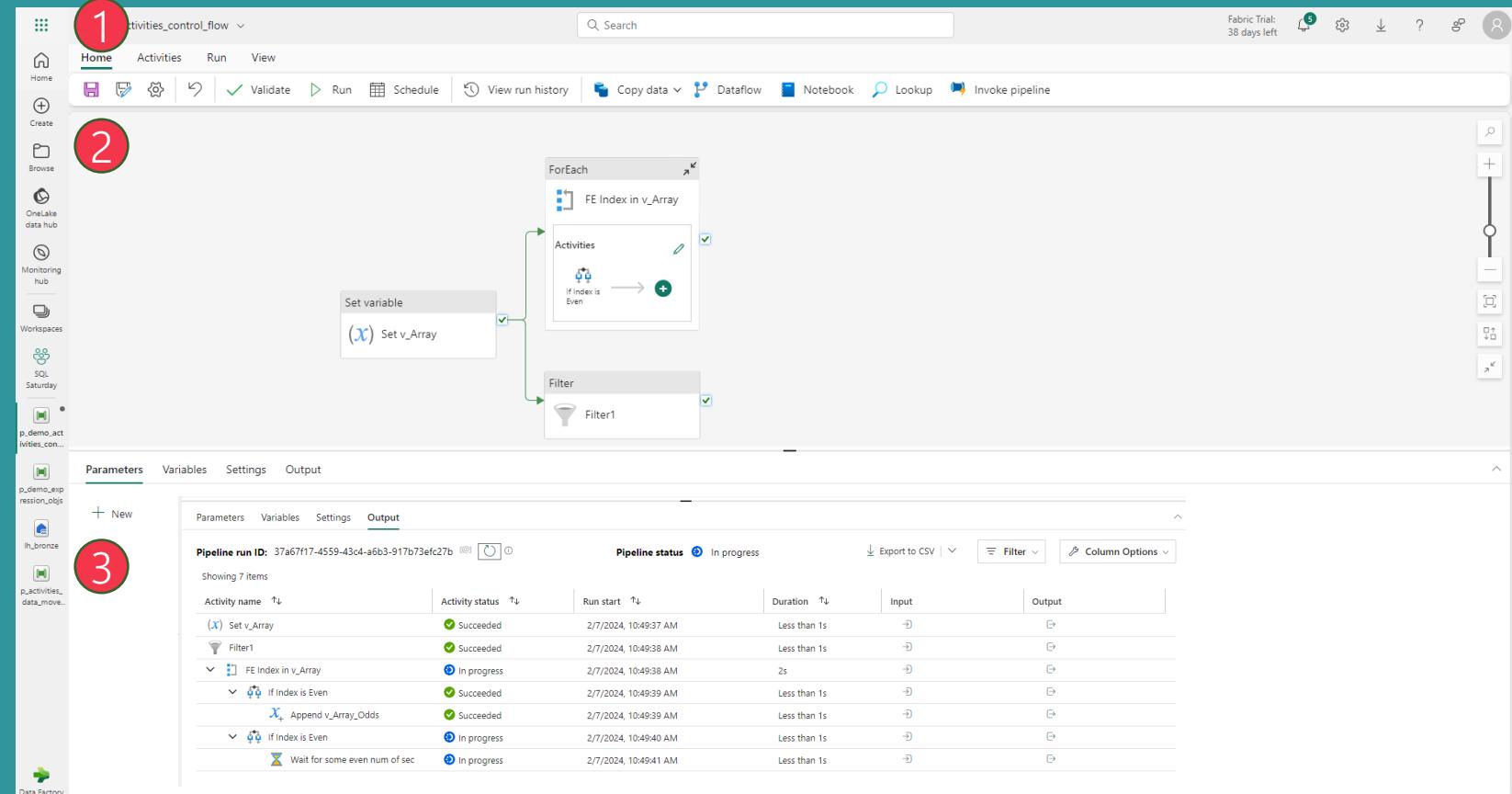
- **Varied Complexity:** Architecture designs of data pipelines can range widely in complexity due to the vast array of design options.
- **Reusability:** Design with reusability in mind, incorporating parameters and components that can be reused, such as invoking other pipelines, running notebooks or data functions.
- **Dynamic Designs:** Aims for designs that are fully dynamic and driven by expressions, utilizing metadata parameterization.



# Exploring Data pipelines

## Navigation

1. Command ribbon
2. Authoring canvas
3. Properties/output



# DEMO

Creating a...

- Task flow
- Lakehouse
- Sample data pipeline



A collection of translucent, multi-colored 3D geometric shapes, including cubes, hexagons, and spheres, arranged in a dynamic, overlapping composition against a white background. The colors range from blue and purple to green, yellow, and pink.

# Getting data into the Lakehouse

# Getting data into the Fabric Lakehouse

- File upload from local computer or OneLake file explorer
- Run a copy job, data pipeline, or mirroring database to ingest data
- Ingest and transform data with a dataflow gen2 or eventstream
- Apache Spark libraries in Notebook code

The screenshot shows the Microsoft Fabric Home interface. At the top, there's a navigation bar with icons for Home, Get data (which is highlighted with a red box), New semantic model, Open notebook, and Manage OneLake data access. Below the navigation bar is the Explorer sidebar, which lists a workspace named 'testLake' containing several items: 'DimCustomer' (selected and highlighted with a gray background), 'DimCustomer\_raw', 'DimGeography\_raw', and 'DimProduct'. To the right of the Explorer is a large data grid titled 'DimCustomer' displaying 8 rows of data. The columns are CustomerKey (12L), GeographyKey (12L), and FirstName. The data includes Clifford, Franklin, Brandon, Hector, Glenn, Adrian, Drew, and Jon.

	12L CustomerKey	12L GeographyK...	ABC FirstName
1	3704	586	Clifford
2	4260	443	Franklin
3	4279	546	Brandon
4	4668	674	Hector
5	7900	628	Glenn
6	8284	626	Adrian
7	9832	515	Drew
8	10097	480	Jon

# Shortcuts in Lakehouse

Shortcuts in a Lakehouse allow users to reference data without copying it from **Tables** or **Files**

Shortcuts unify data from different Lakehouses, workspaces, or external storage, such as

- ADLS Gen2
- AWS S3
- Dataverse



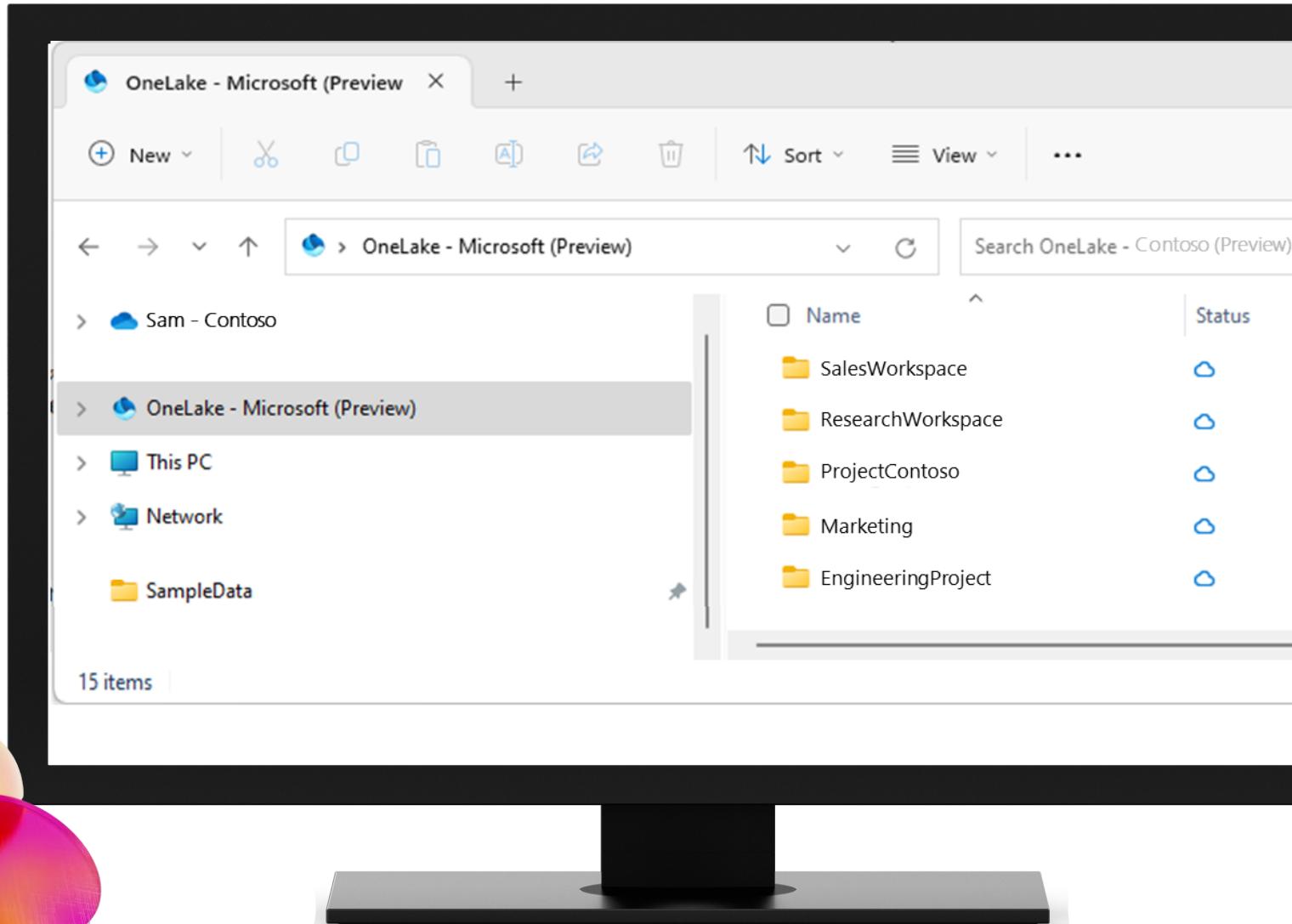
The screenshot shows the Microsoft Fabric Data Explorer interface. On the left, there's a sidebar with icons for Home, Create, Browse, OneLake data hub, Monitoring hub, Workspaces, My workspace, and TFLakehouse. The main area shows an 'Explorer' view for 'TFLakehouse'. Under 'Tables', there's a table named 'taxi\_zone\_lookup' with columns: ID, Borough, Zone, and service\_zone. A context menu is open over this table, with 'New shortcut' highlighted by a red box. Other options in the menu include 'New subfolder', 'Upload', 'Rename', 'Delete', 'Properties', and 'Refresh'. At the bottom of the screen, there are some decorative 3D spheres.

ID	Borough	Zone	service_zone
EWR	Newark Air...	EWR	
Queens	JFK Airport	Airports	
Queens	LaGuardia ...	Airports	
Unknown	NV	N/A	
Unknown	NA	N/A	
Manhattan	Alphabet City	Yellow Zone	
Manhattan	Battery Park	Yellow Zone	

# OneLake file explorer

Seamlessly integrates OneLake with Windows File Explorer, **automatically syncing** all OneLake items that you have access to in file explorer.

Syncing doesn't download the data, it creates placeholders. You must double-click on a file to download the data locally.



# Data ingestion scenarios | Check your knowledge

Connecting to existing SQL Server and copying data into Delta table on the Lakehouse

Uploading files from your computer

Copying and merging multiple tables from other Lakehouses into a new Delta table

Referencing data without copying it from other internal Lakehouses or external sources



# Considerations when loading data

## Use case

Small file upload from local machine

Data ingestion only

Orchestrated data ingestion and transformation

Complex data transformations and orchestration

## Recommendation

Use OneLake file explorer / Upload file

Use Copy job or Data pipeline

Use Data pipeline or Dataflow Gen2

Use Notebook code and Data pipeline



# Copy activity, dataflow, and Spark decision guide

	Data pipeline copy activity	Dataflow Gen 2	Spark
Use case	Data lake and data warehouse migration, data ingestion, lightweight transformation	Data ingestion, data transformation, data wrangling, data profiling	Data ingestion, data transformation, data processing, data profiling
Primary developer persona	Data engineer, data integrator	Data engineer, data integrator, business analyst	Data engineer, data scientist, data developer
Primary developer skill set	ETL, SQL, JSON	ETL, M, SQL	Spark (Scala, Python, Spark SQL, R)
Code written	No code, low code	No code, low code	Code
Data volume	Low to high	Low to high	Low to high
Development interface	Assistant, 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

[Microsoft Learn: Data ingestion decision guide](#)



A collection of translucent, multi-colored 3D geometric shapes, including cubes, spheres, and cylinders, arranged in a dynamic, overlapping composition. The colors range from blue and purple to green, yellow, and pink, with some shapes showing internal circuit board patterns. They are set against a white background with faint, light gray wireframe hexagonal grids.

# Data engineering overview

# Data engineering in Fabric

Empowers data engineers to transform data at scale and build a Lakehouse architecture



Build a Lakehouse  
for all your  
organizational data



Spark runtime with great  
out of the box  
performance and robust  
admin controls



Delightful authoring  
experience in your  
tools of choice



Completely  
integrated into the  
Fabric foundation



# Data engineering in Fabric

Empowers data engineers to transform data at scale and build a Lakehouse architecture



## Lakehouse

Store and manage structured and unstructured data in a single location



## Apache Spark job definition

Submit batch/streaming job to Spark cluster, apply different transformation logic to the data



## Notebook

Create and share documents that contain live code, equations, visualizations, and narrative text



## Data pipeline

Collect, process, and transform data from its raw form to a format that you can use for analysis and decision-making



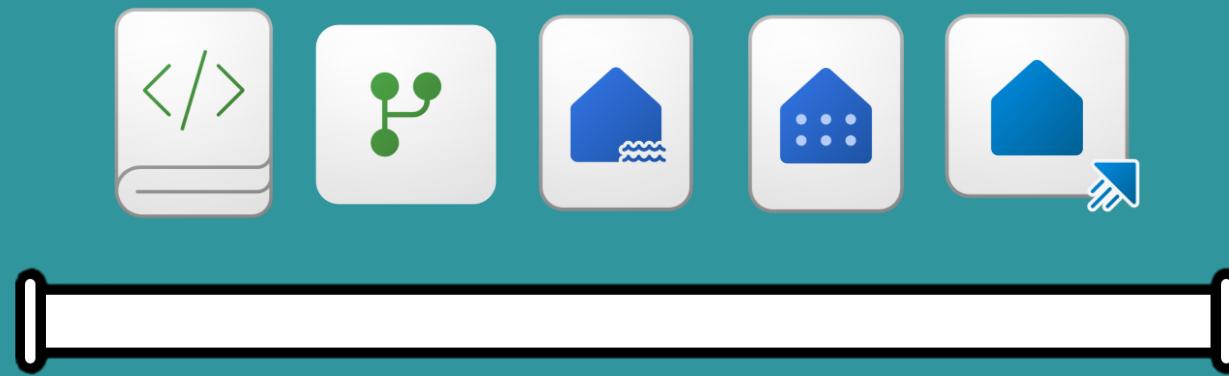
A collection of translucent, multi-colored 3D geometric shapes (cubes, spheres, and hexagons) in shades of blue, green, yellow, and pink, scattered across a white background. A faint, large hexagonal grid pattern is visible in the center.

# Ingesting and orchestrating data with Data pipelines

# Data pipelines

Seamlessly connect and ingest data into Fabric using a **no-code** interface.

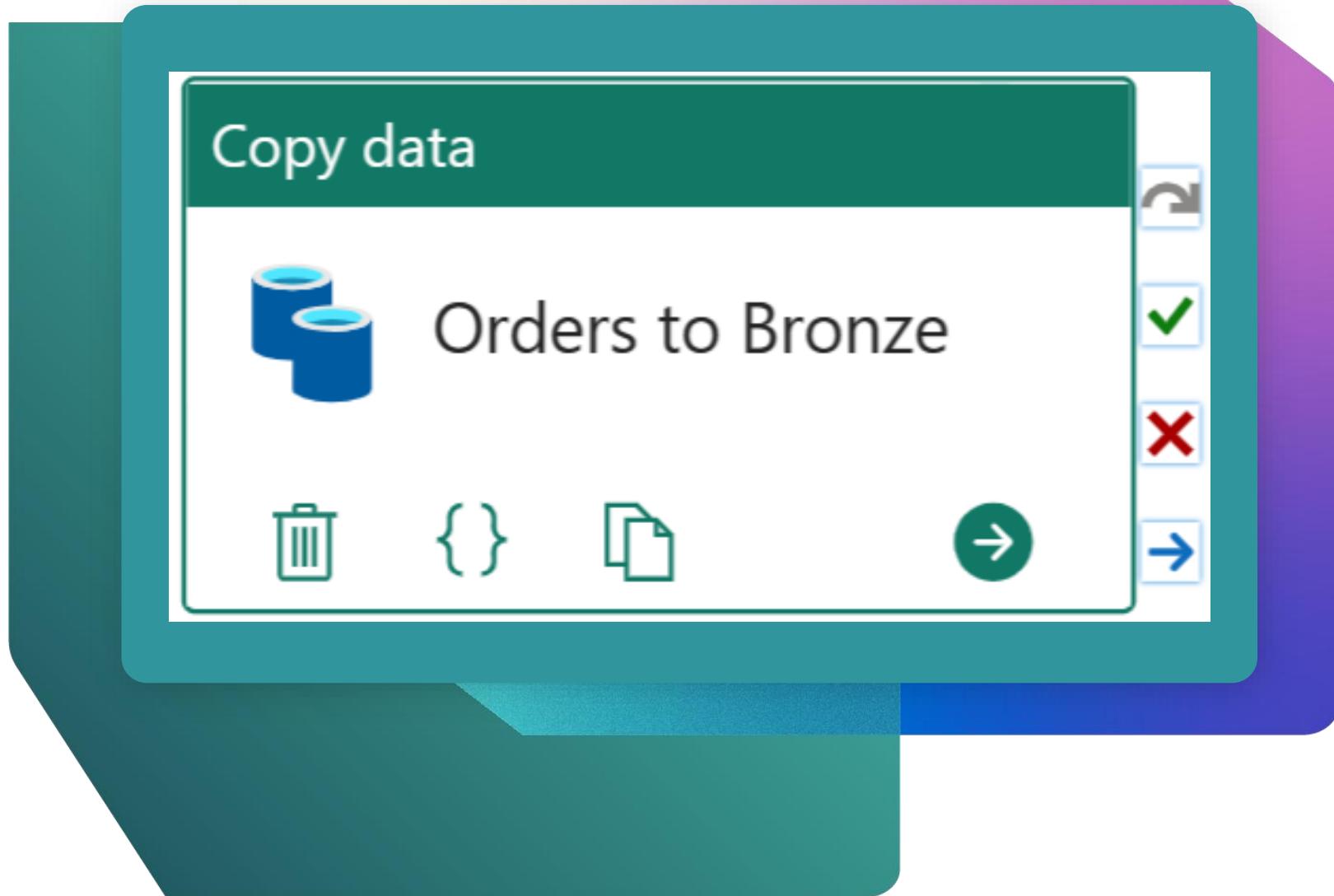
- Evolution of **Azure Data Factory** pipelines
- Rich library of **activities**
- Fast copy of binary files using Az-Copy
- Jumpstart with **Copy Assistant**



# Copy activity

The Copy Activity is the **core** of data pipelines

- Jumpstart with **Copy job** or **Copy assistant**
- **Fine grain control** of file type conversion, column mapping, merging/splitting of files, and more..
- Workspace Source and Destinations include Lakehouse (Tables & Files), Warehouse, and KQL Database
- Supports external sources and destinations



# And so much more...

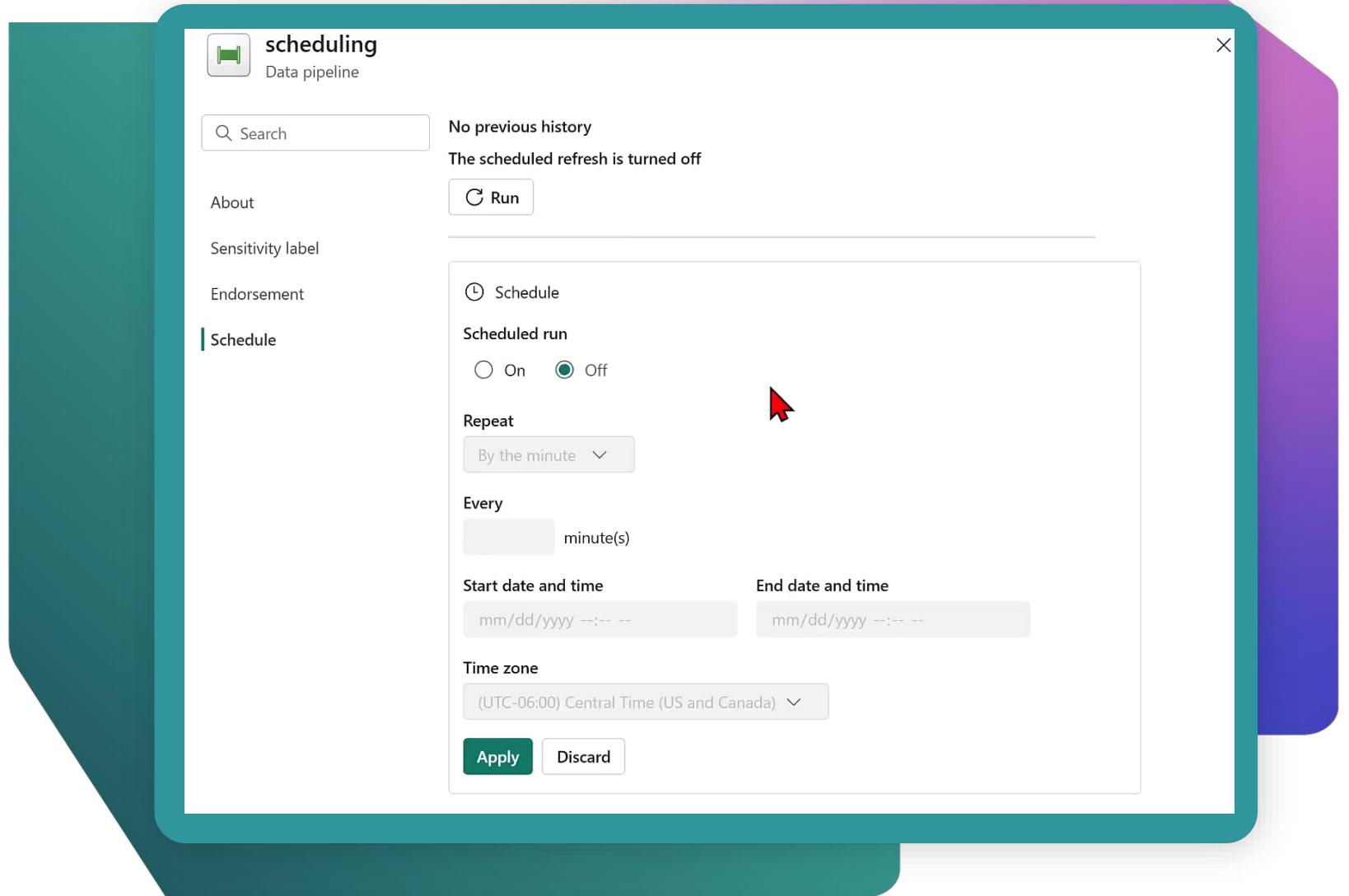
Capability	Description
Control flow	<ul style="list-style-type: none"><li>• Conditional paths</li><li>• Loops</li></ul>
Nesting	<ul style="list-style-type: none"><li>• Invoking data pipelines from other data pipelines</li><li>• Bi-directional communication via output variables</li></ul>
Parameterization	<ul style="list-style-type: none"><li>• Parameters of varying data types</li></ul>
Expression language	<ul style="list-style-type: none"><li>• Almost every field can be made dynamic!</li><li>• Extensive built-in functions allow for massive customizations (data driven designs)</li></ul>
Repeatable	<ul style="list-style-type: none"><li>• Schedulable</li></ul>



# Platform scheduler

## Schedule events at desired intervals

- Deeply integrated with monitoring hub
- Scheduled:
  - By the minute
  - Hourly
  - Daily
  - Weekly
  - \*Monthly

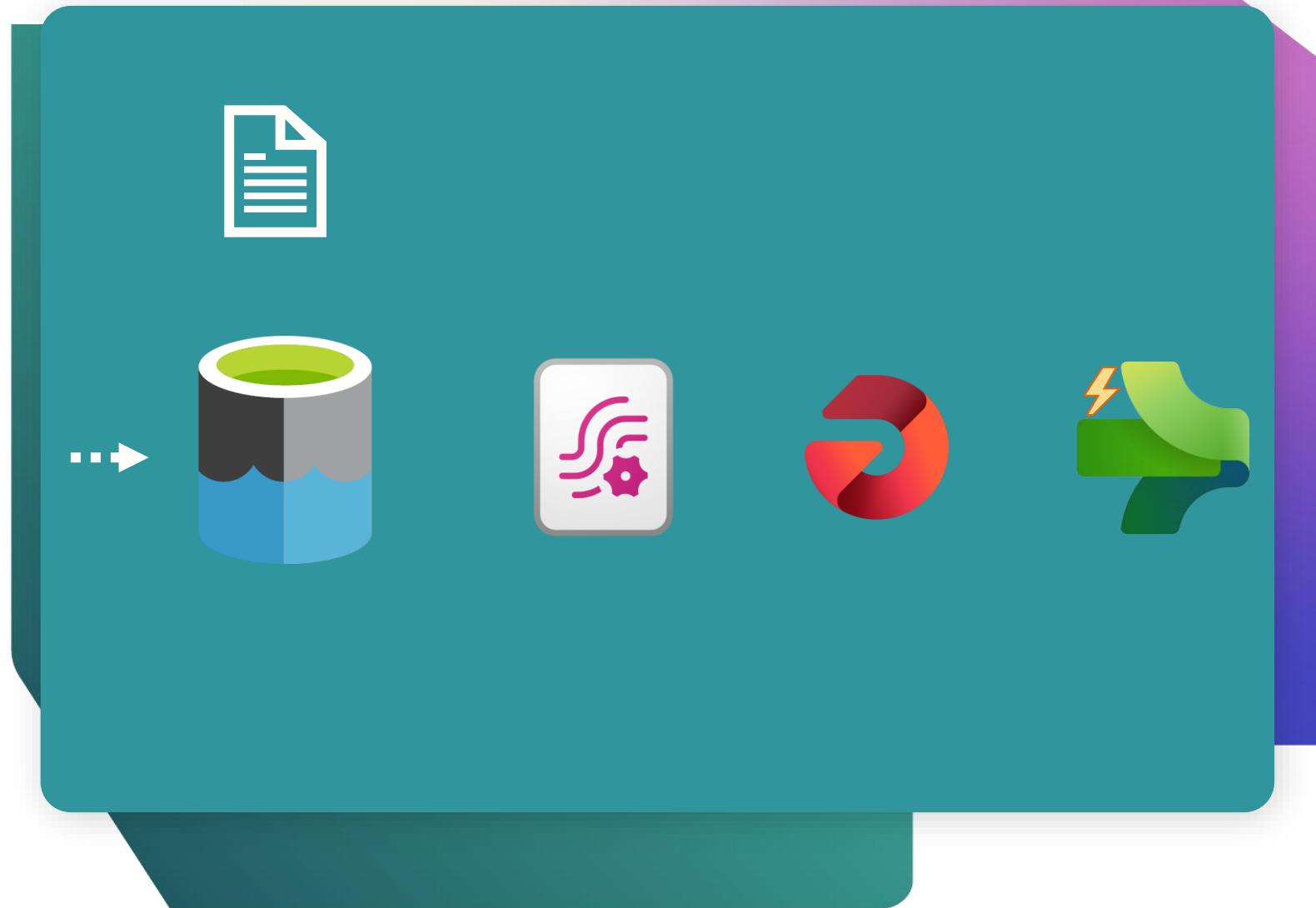


# Event-driven triggers with data pipelines

Invoke a Data Pipeline upon a file event using event streams and Reflex triggers

Extensive event types and filtering capabilities

Ability to link multiple pipelines to a single event

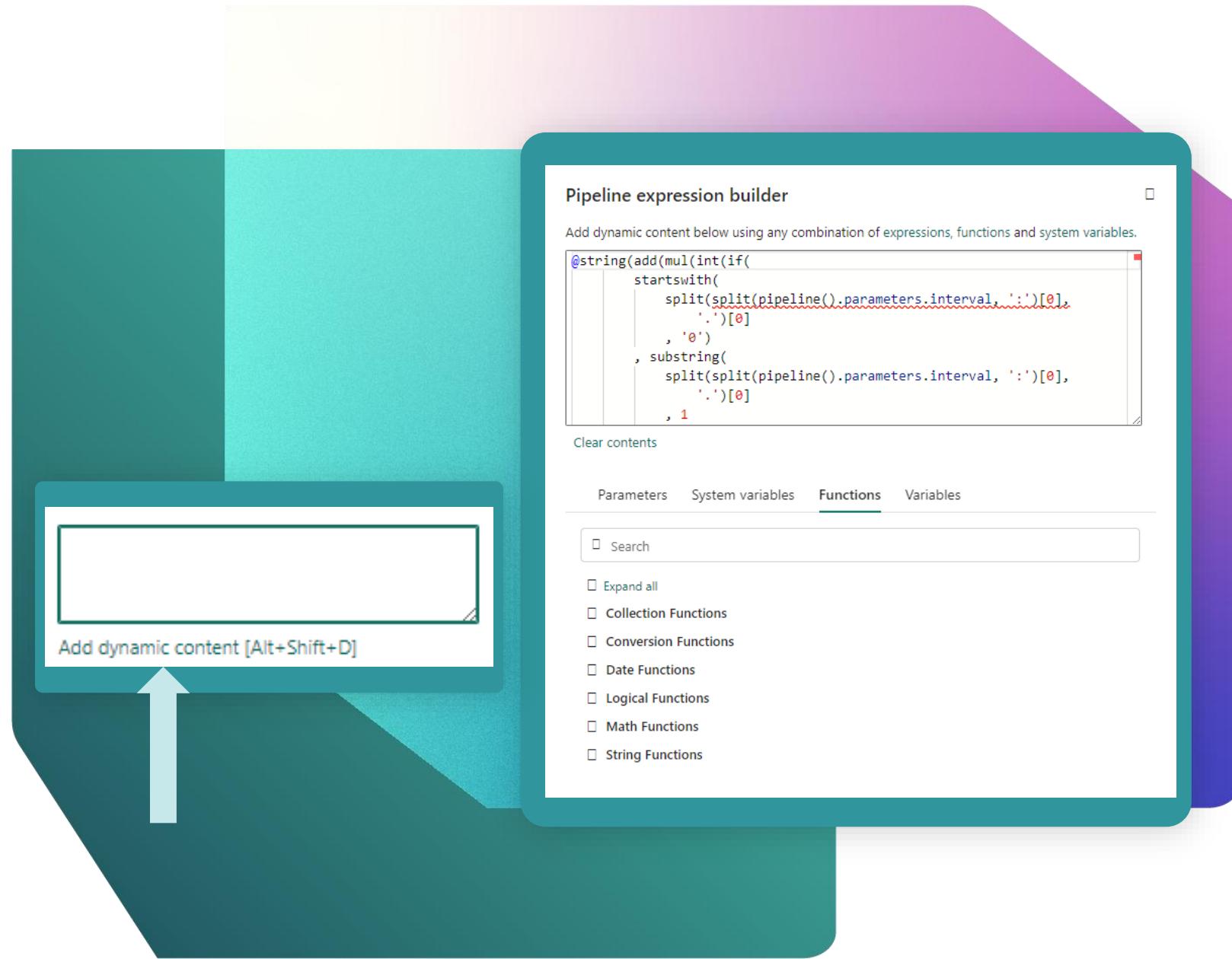


# Expression builder

Allows you to create and manage dynamic data-driven content

Wide range of functions and operators that can be used to manipulate and transform data

Greater flexibility and reusability, and to perform advanced data transformations without the need for custom code

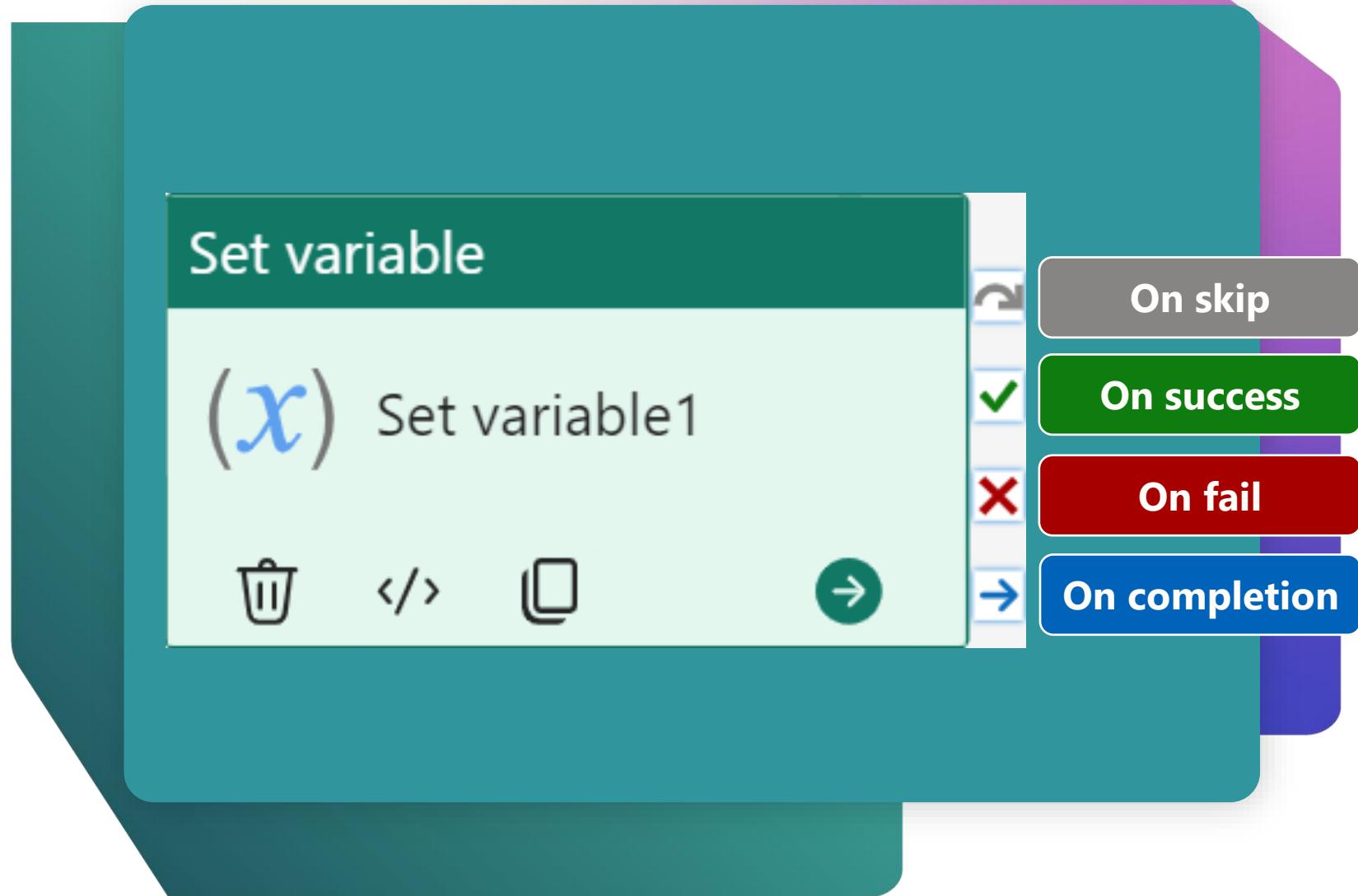


# Conditional paths

Enable you to build **robust pipelines** with error handling and branching logic

There are four types:

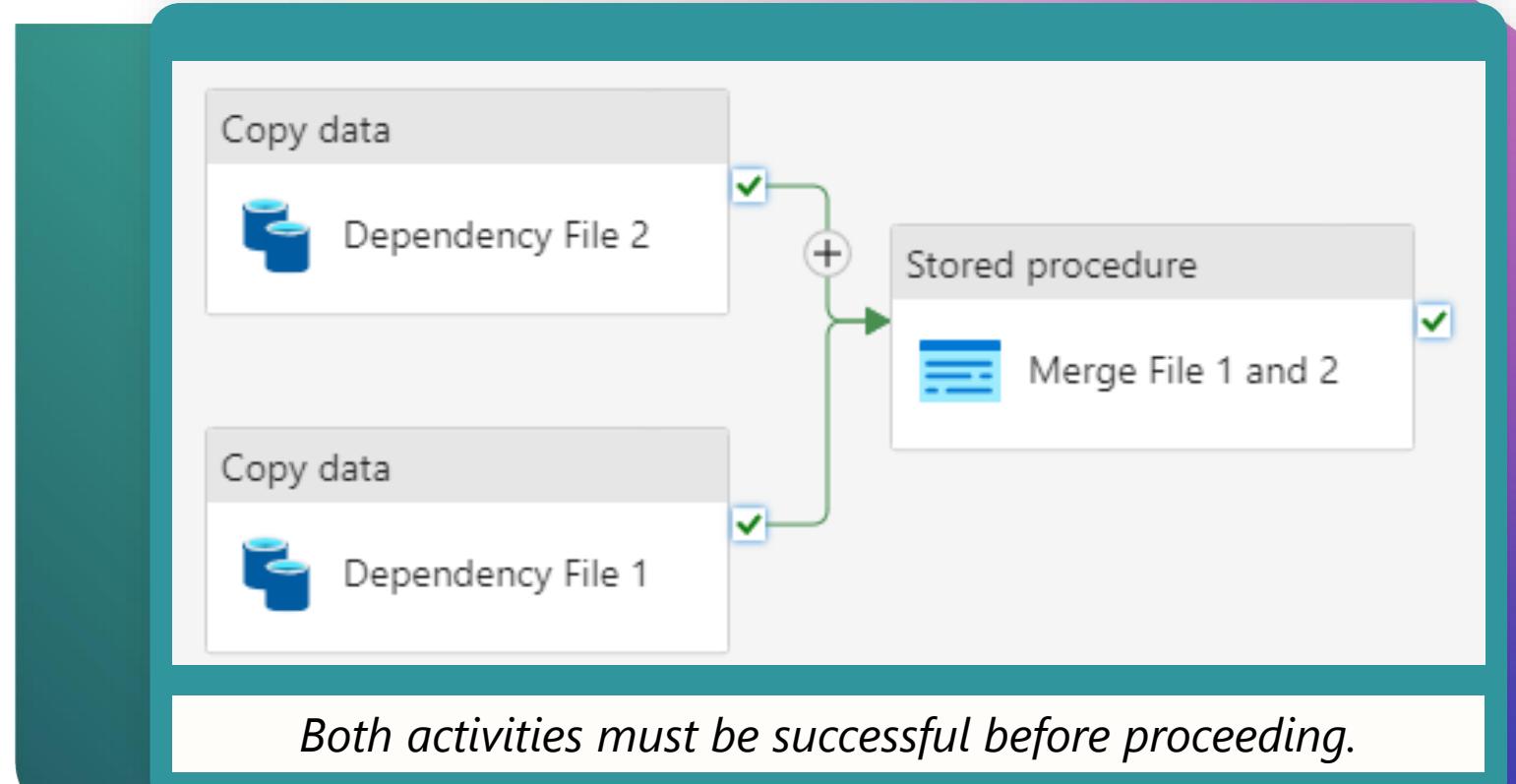
- On Skip
- On success
- On fail
- On completion



# Conditional paths

Define different execution paths based on the outcome of a previous activity

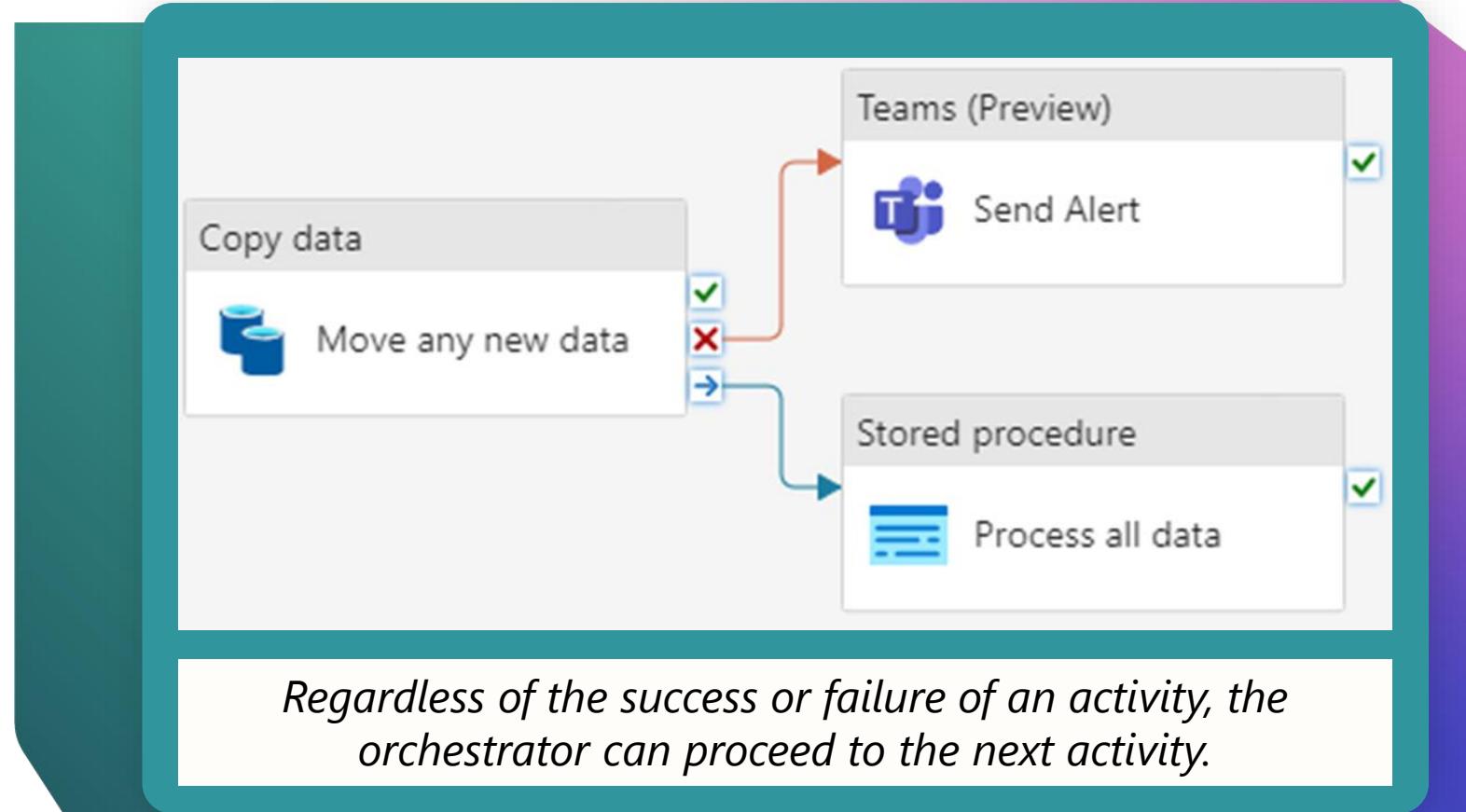
- **Blocking dependencies**
- Non-blocking dependencies
- Error handling



# Conditional paths

Define different execution paths based on the outcome of a previous activity

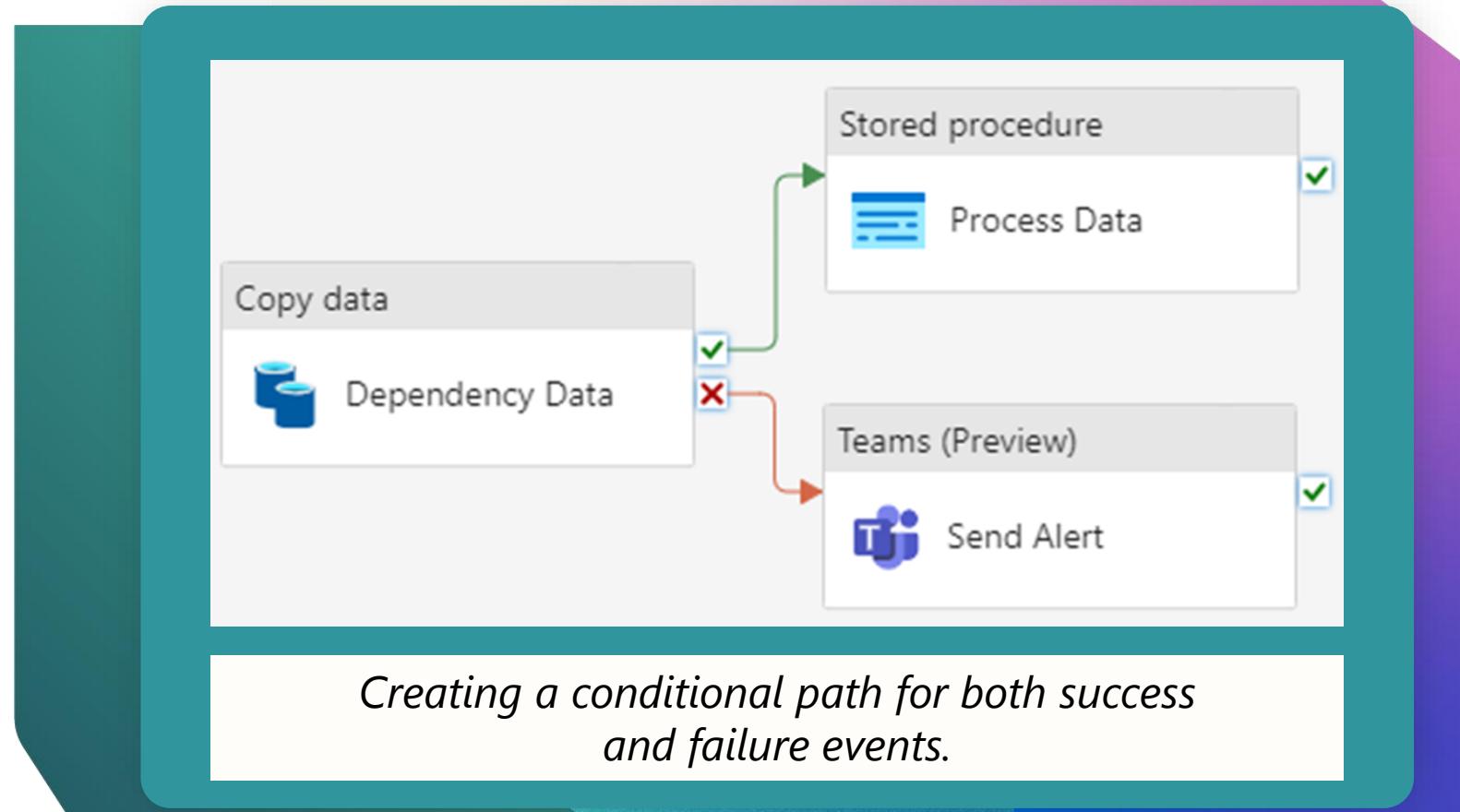
- Blocking dependencies
- **Non-blocking dependencies**
- Error handling



# Conditional paths

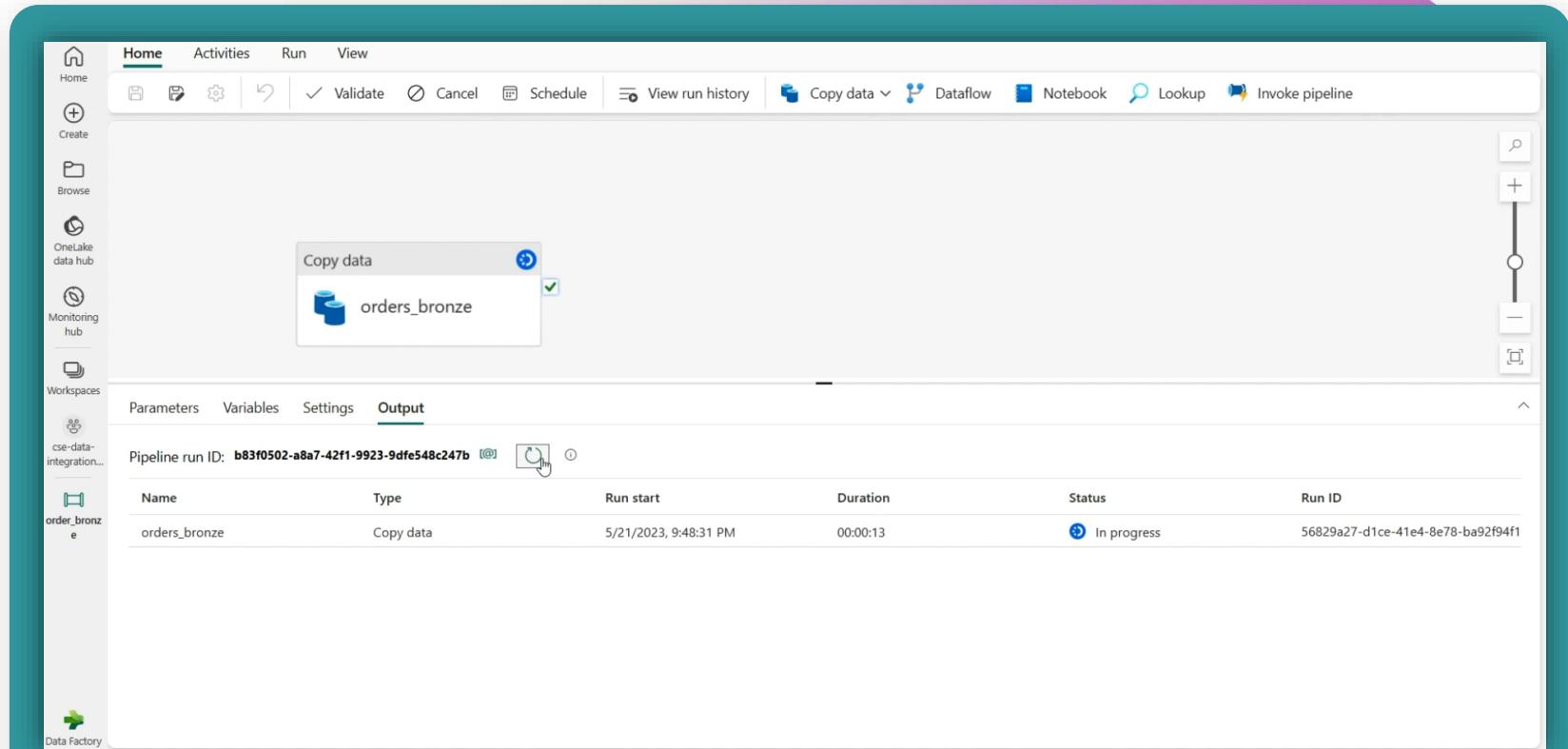
Define different execution paths based on the outcome of a previous activity.

- Blocking dependencies
- Non-blocking dependencies
- **Error handling**



# Troubleshooting tips

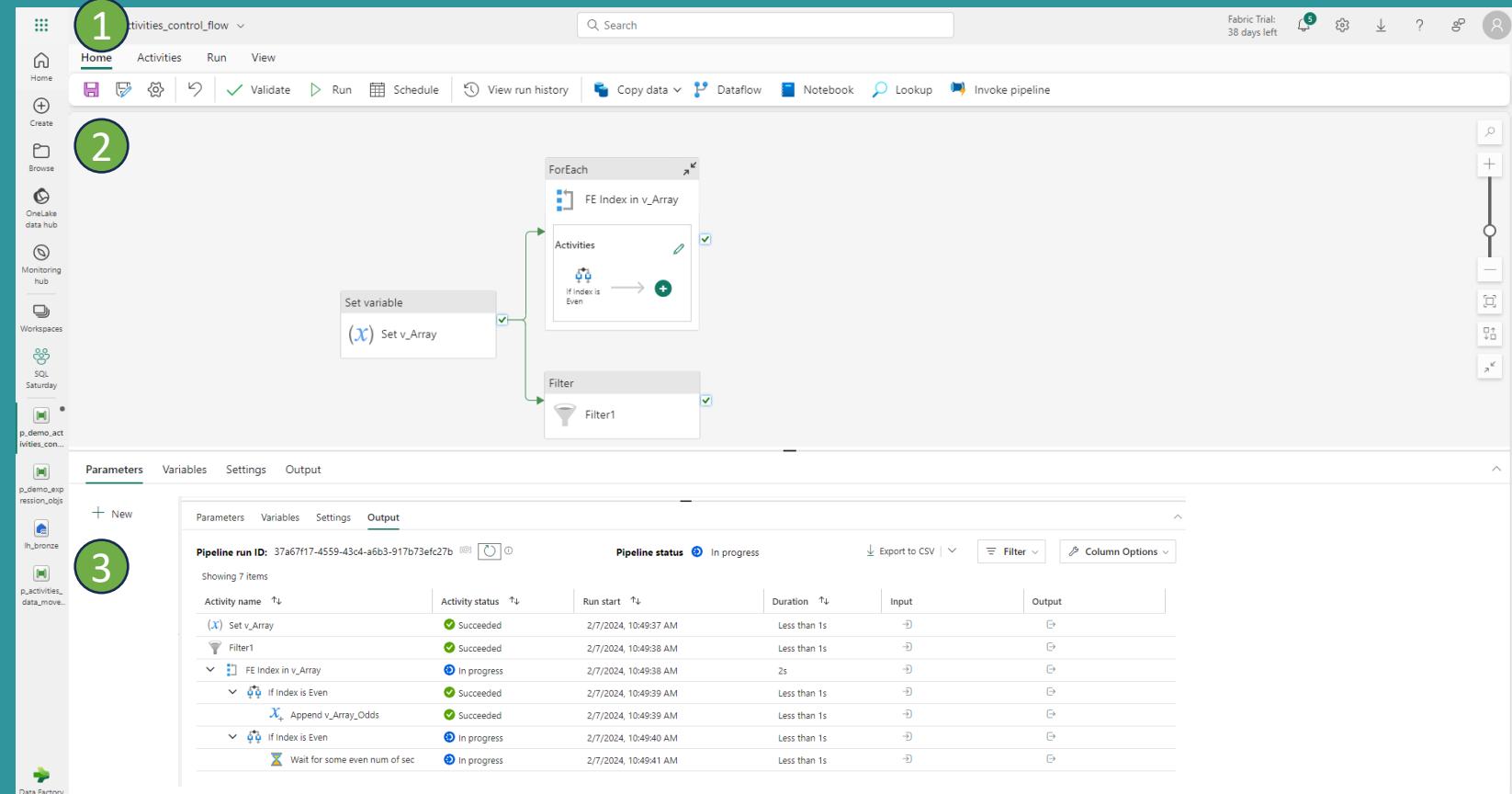
- Leverage the Output window to find and review errors
- Determine if it is an activity error or an error from the source/destination
- Set a retry attempt for activities



# Exploring Data pipelines

## Navigation

1. Command ribbon
2. Authoring canvas
3. Properties/output



# DEMO

Creating a...

- Lakehouse
- Data pipeline



# Migrate from ADF/Synapse Pipelines to Fabric Data pipelines

Feature	Fabric Data Factory	Azure Data Factory	Azure Synapse Pipelines
Office 365 Outlook Activity (Email activity)	Yes	*No	*No
MS Teams Activity	Yes	*No	*No
Refresh a Dataflow Gen2	Yes	*No	*No
Refresh a Power BI semantic model	Yes	*No	*No
Change Data Capture	Yes	Yes	No
Disable Activity	Yes	Yes	Yes
Managed Airflow	Yes	Yes (Preview)	No
Validation activity	No	Yes	Yes

\* Supported by external services / REST APIs



A collection of translucent, multi-colored 3D geometric shapes, including cubes, spheres, and hexagons, arranged in a dynamic, overlapping composition against a white background. The colors range from blue and purple to red and yellow, with some shapes showing internal circuit board patterns.

# Ingesting and transforming data with Dataflow Gen2

# Dataflow Gen2

## Next generation of data preparation

- **Easy to use**, no-code ETL & ELT
- Includes **smart AI-based** data prep
- More than **300+** transformations
- **Output data destinations**  
Write output of dataflows to Azure SQL database, Data warehouse, Lakehouse and more

The screenshot shows the Microsoft Power Query interface with the following details:

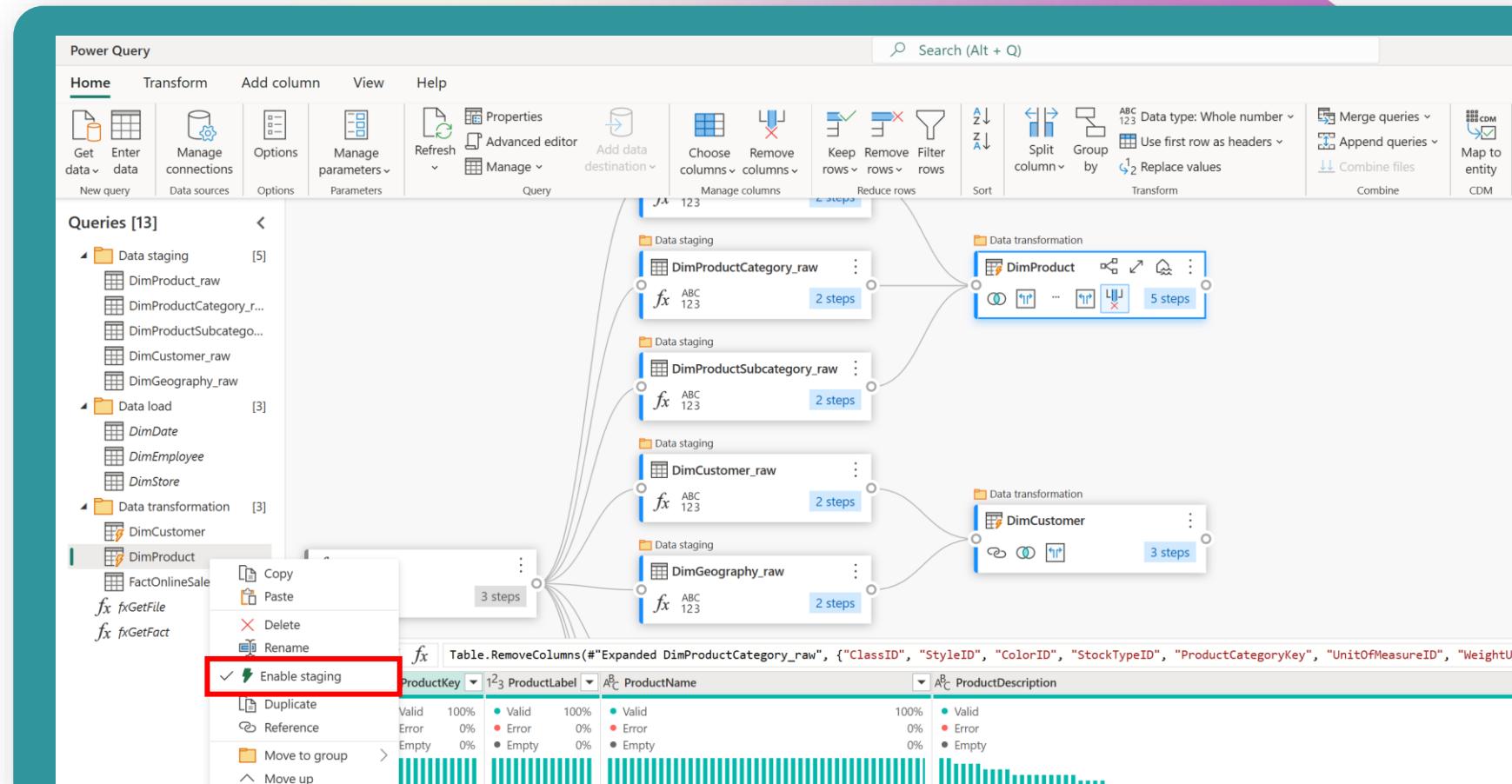
- Queries [13]:** A list of queries including "Data staging", "Data load", and "Data transformation". The "DimCustomer" query is selected and highlighted with a red arrow.
- Table View:** A preview of the "DimCustomer" data with columns: CustomerKey, GeographyKey, FirstName, MiddleName, LastName, BirthDate, MaritalStatus, Suffix, Title, EmailAddress, YearlyIncome, and Education.
- Column Profiling:** A grid at the top of the table view shows the count of distinct values and the percentage of valid data for each column.
- Power BI Integration:** The bottom left corner shows the "Power BI" logo.
- Query Settings:** On the right, it shows "Properties" for "DimCustomer" and "Entity type: Custom".
- Applied Steps:** Shows the history of steps taken, including "Source" (Merged queries) and "Expanded DimGeography.raw".
- Data Destination:** Shows "No data destination".

# Dataflow Gen2 staging

Highly scalable using  
**Fabric compute**

A **seamless experience** -  
yielding fast, easy and powerful  
results

**Abstracts away** the  
complexities of traditional ETL  
and ELT

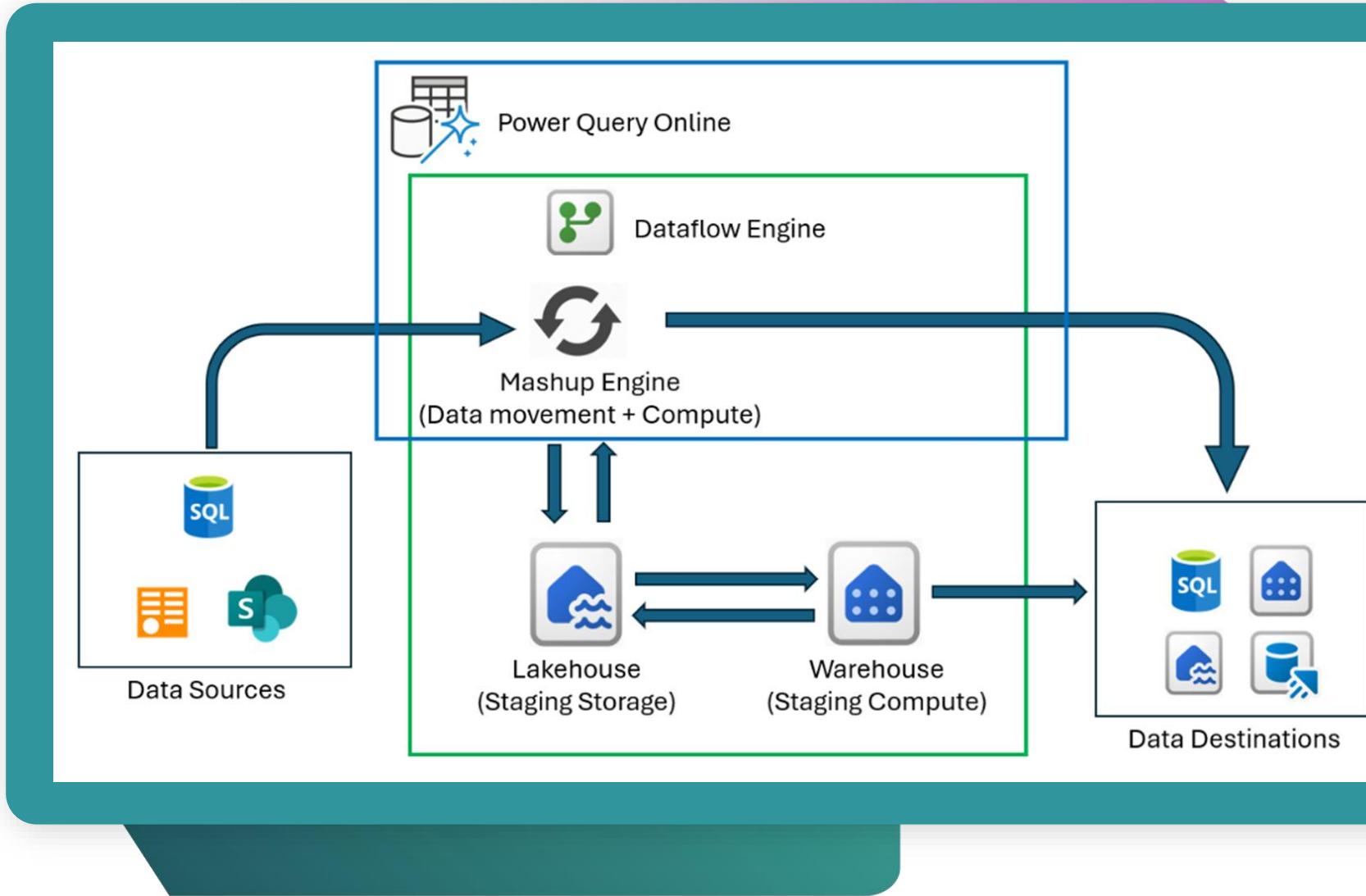


\*Previously titled "Enable load"

# Dataflow Gen2 staging

**Optimize** the use of dataflows with Fabric compute

1. Connect to your data and **copy** it into Fabric using the **\*Enable staging** option (**\*On by default**)
2. Create a **reference** query in a new query.
3. Apply transformation steps to the **computed** table for complex ETL operations such as join, distinct, filter and group by – leveraging Fabric compute.



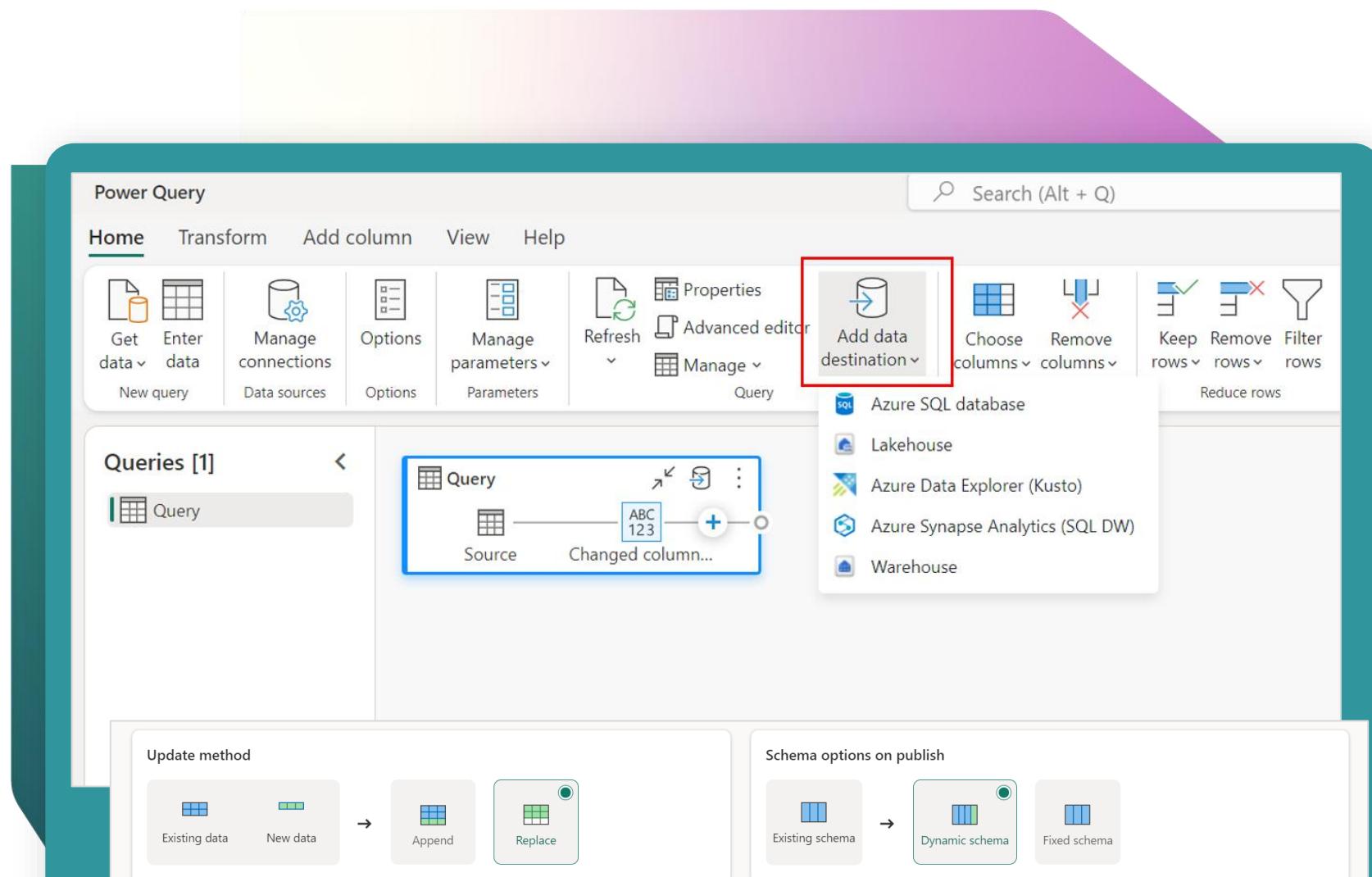
# Data destinations

Supported destinations:

- Lakehouse
- Warehouse
- Eventhouse
- Azure SQL database
- Azure Synapse Analytics

Update methods:

- Replace
- Append



# Fast Copy

Ingest terabytes of data **effortlessly** with dataflows, powered by the scalable backend of the Copy activity

A limited set of transformations are supported:

- Combine files
- Select or remove columns
- Change data types
- Rename a column

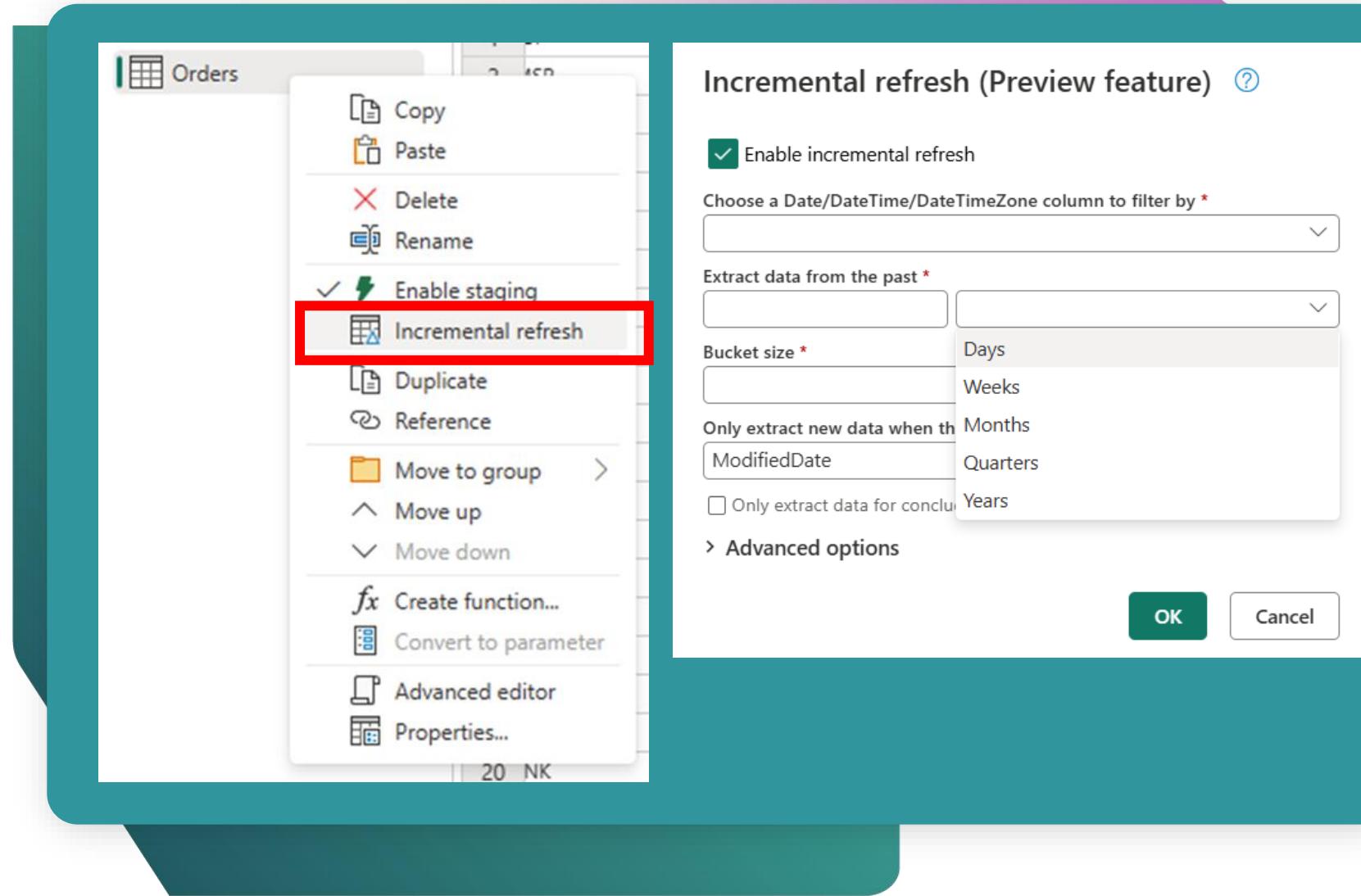
The screenshot shows the Microsoft Dataflow interface. On the left, there's a preview pane displaying a table with five columns labeled r\_1 through r\_5. The first row contains values 27450000, 3240000, 26640000, 22770000, and 2016. The second row contains 1080000, 28800000, 29070000, 15210000, and 2547. The third row contains 16020000, 24210000, 26460000, 15750000, and 846. The fourth row contains 2070000, 24840000, 12150000, 3150000, and 1926. The fifth row contains 30870000, 9720000, 18270000, 8010000, and 594. The sixth row contains 4320000, 17640000, 21420000, 18990000, and 1458. A red box highlights a message in the center of the screen: "This step is going to be evaluated with fast copy." To the right, there's a sidebar titled "Applied steps" which lists various transformation steps: "Filtered hid...", "Invoke cust...", "Renamed c...", "Removed o...", "Expanded t...", and "Changed c...".

# Incremental refresh

Public Preview

Process only changed data since the last refresh to **save time and resources**

- Evaluate changes by comparing the maximum **DateTime** value with the previous refresh.
- Retrieve and load data for changed buckets in parallel, loaded to staging
- Replaces the destination data with the new data, affecting only updated buckets.



# Dataflow Gen2 | Check your knowledge

A single dataflow has a limit of how many queries/tables?

25

50

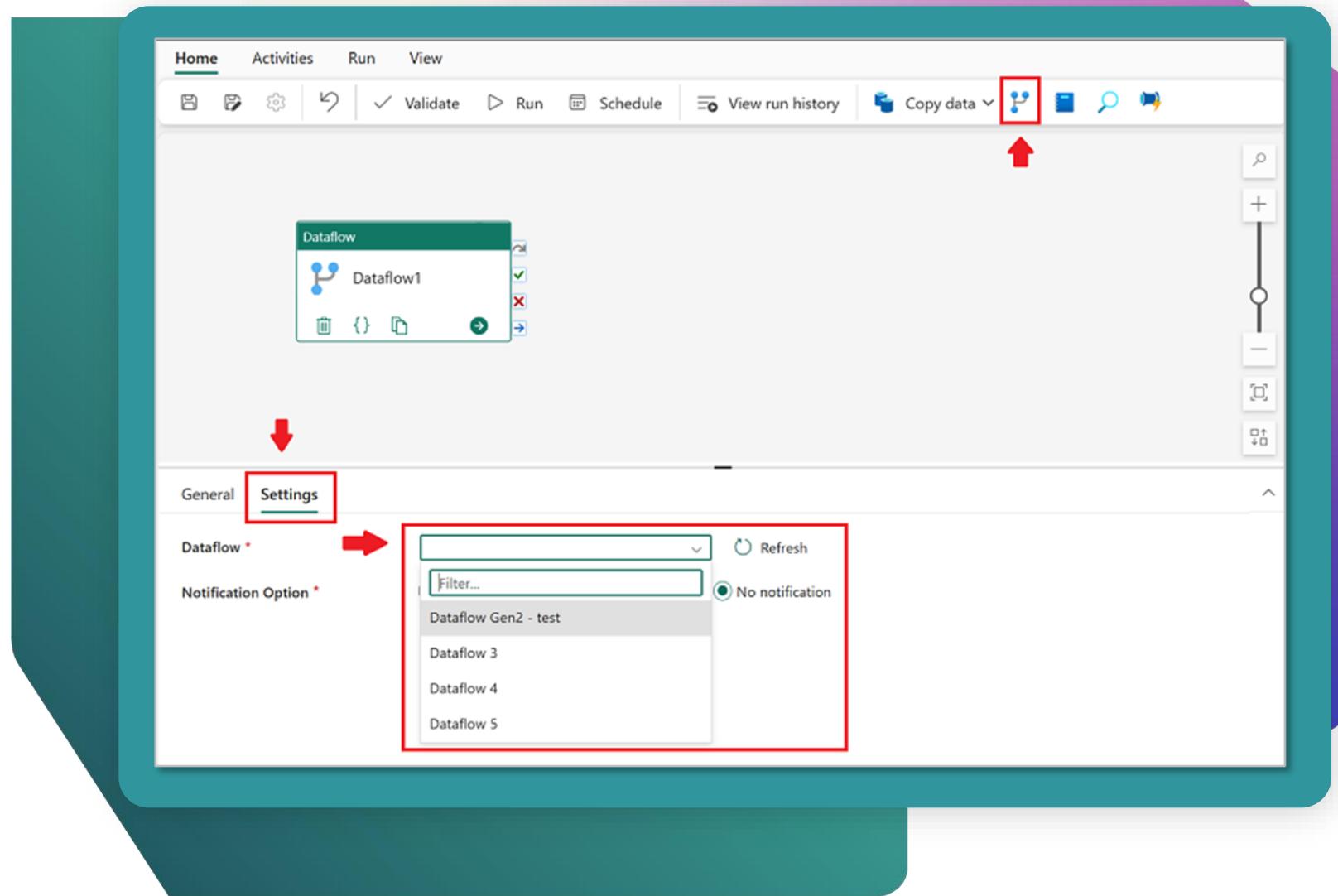
100



# Orchestrating dataflows gen2 with data pipelines

A dataflow for data ingestion and transformation, and landing into a lakehouse using dataflows

Then incorporate the dataflow into a pipeline to orchestrate additional activities



# Dataflow Gen2 scale recommendations

Separate **dimension** tables and **fact** tables into separate dataflows based on update method

Separate tables **with staging** and **without staging** into separate dataflows

Separate tables **with fast copy** support and **without fast copy** support into separate dataflows

Separate tables based on the data source and if they **support query folding** or do **not support query folding**

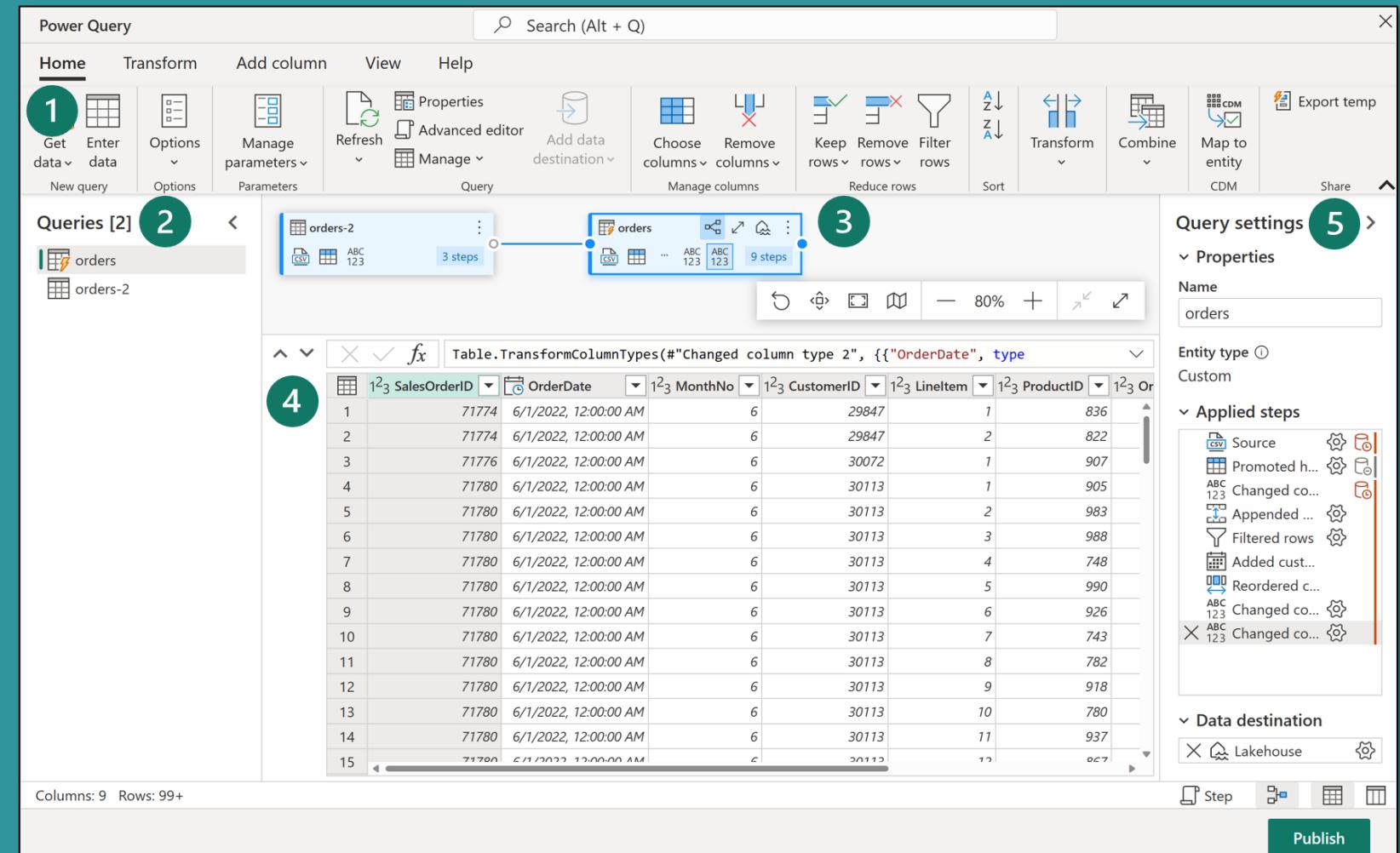
Use **copy job, mirroring** or **data pipelines** to ingest data and dataflows to transform



# Exploring Dataflows Gen2

## Navigation

1. Command ribbon
2. Queries pane
3. Visual diagram
4. Data preview grid
5. Query settings



# DEMO

Creating a...

- Lakehouse
- Dataflow Gen2



# Migrate from Power BI dataflow to Fabric dataflow gen2

Feature	Power BI Dataflow	Dataflow Gen2
Incremental refresh	Yes	<b>YES!</b>
Accessible file outputs	*No	Yes
Fast copy	No	Yes
Data destination output	No	Yes
Premium capacity required	No	Yes
AI Insights	Yes	*No
AutoML	Deprecated	Deprecated
Attach Common Data Model (CDM) folder	Yes	No
Linked Tables	Yes	No (use Shortcuts)
On-premises data gateway	Yes	Yes
Vnet data gateway	No	Yes

\* Supported by external services / REST APIs



# Metadata-Driven design benefits

- Configuration Driven
- Scalable
- Adaptable
- Independent Components



# Metadata-Driven design concepts

**Metadata Management** – Empower users to configure new data models

**Logging** – Leveraging native and custom logging for detailed insights and performance metrics

**Alerting** – Point in Time notifications directly to MS Teams or Outlook

**Lineage** - End to end visibility of processes and operations

**Extendibility** – Easily integrate additional features



# Metadata-Driven design components

- **Metadata Storage**
  - History, Audit
- **Custom Logging Framework**
  - Status, Rollback, Performance
- **Custom Alerting Framework**
  - MS Teams & Outlook
- **Isolation and compartmentalization of tasks**
  - Independent components that can operate together
- **Medallion Architecture**



**One more thing...**





Thank you!



# Microsoft Fabric

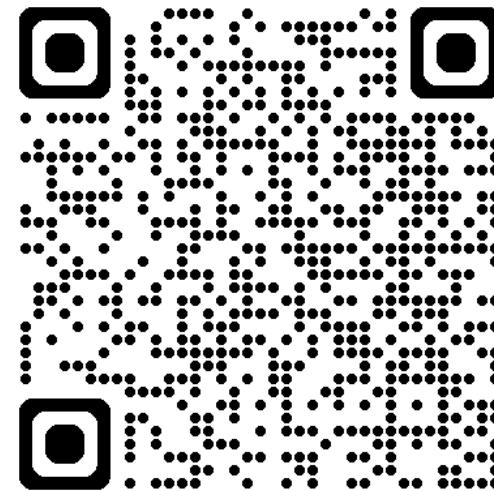
## 60-Day Free Trial

No Credit Card

No Azure Subscription

# F64 SKU

\$17,000 value



[aka.ms/try-fabric](http://aka.ms/try-fabric)



# Get Involved in the Fabric Community



## [aka.ms/FabricCommunity](https://aka.ms/FabricCommunity)

Connect with community members, ask questions, and learn more about Fabric



## [aka.ms/FabricUserGroups](https://aka.ms/FabricUserGroups)

Find a user group that matches your interests in your area or online



## [aka.ms/SuperUsers](https://aka.ms/SuperUsers)

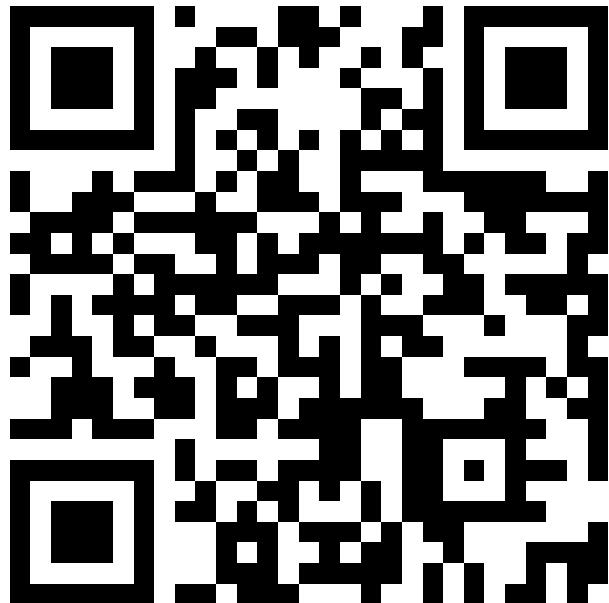
Spread your Fabric knowledge, insights, and best practices with others



## [aka.ms/MVP](https://aka.ms/MVP)

Technology experts that share their knowledge and passion with the community

# FREE Microsoft Fabric Exam



## Certification Exam

Are you ready to get Fabric certified by the end of October?

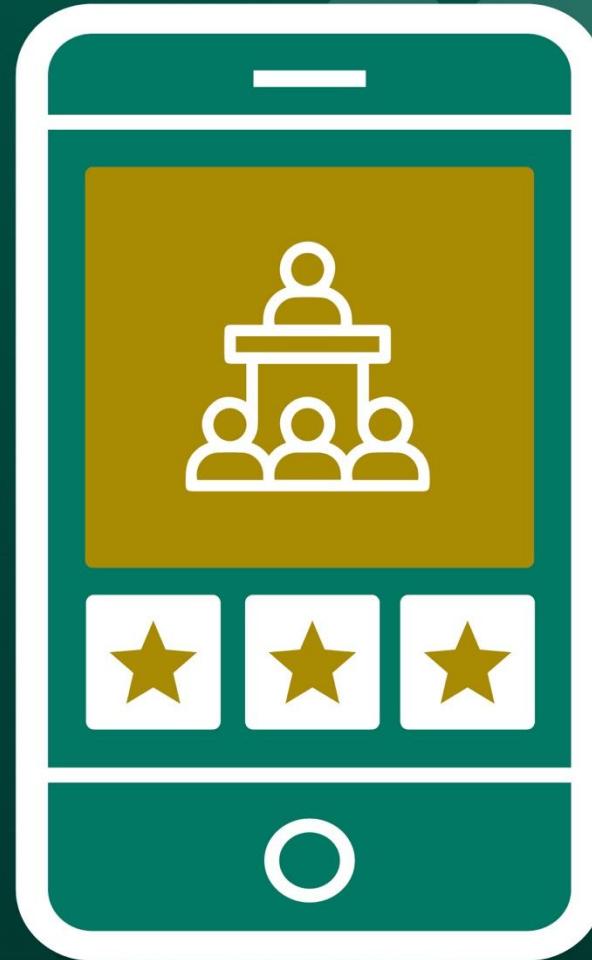
Claim your **100% discount voucher** for Exam DP-600: Fabric Analytics Engineer.

Come find us in the Community Lounge under the Get Certified banner to learn more!

[aka.ms/FabCon24/IAmReady](https://aka.ms/FabCon24/IAmReady)



Please rate  
this session  
on the app



cvent





**Learn more about Microsoft Fabric**

Unify your data to unlock AI innovation