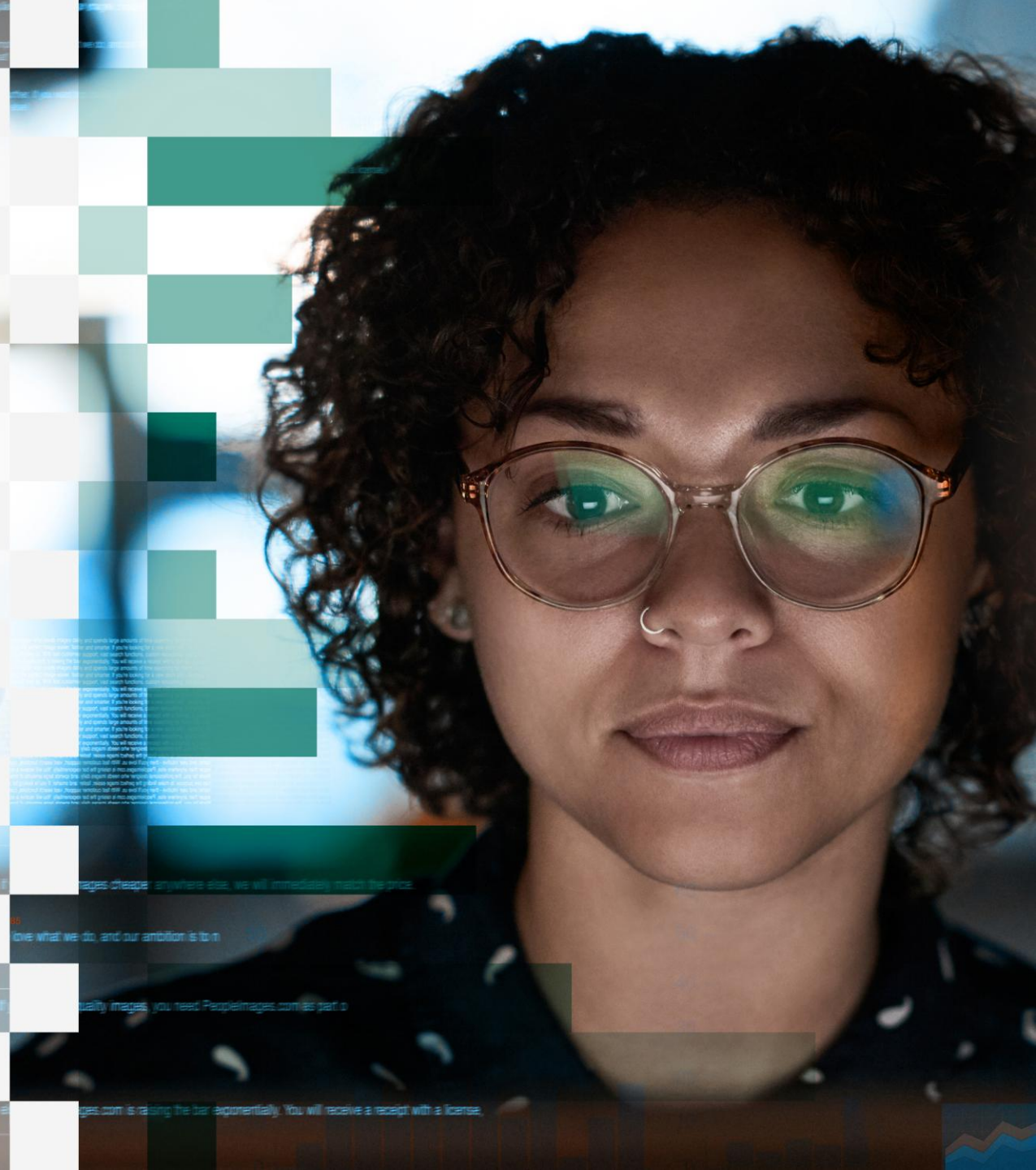
WiFi: FabConEurope
Password: FabCon24

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

GET IT ON Google Play

Download on the App Store

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

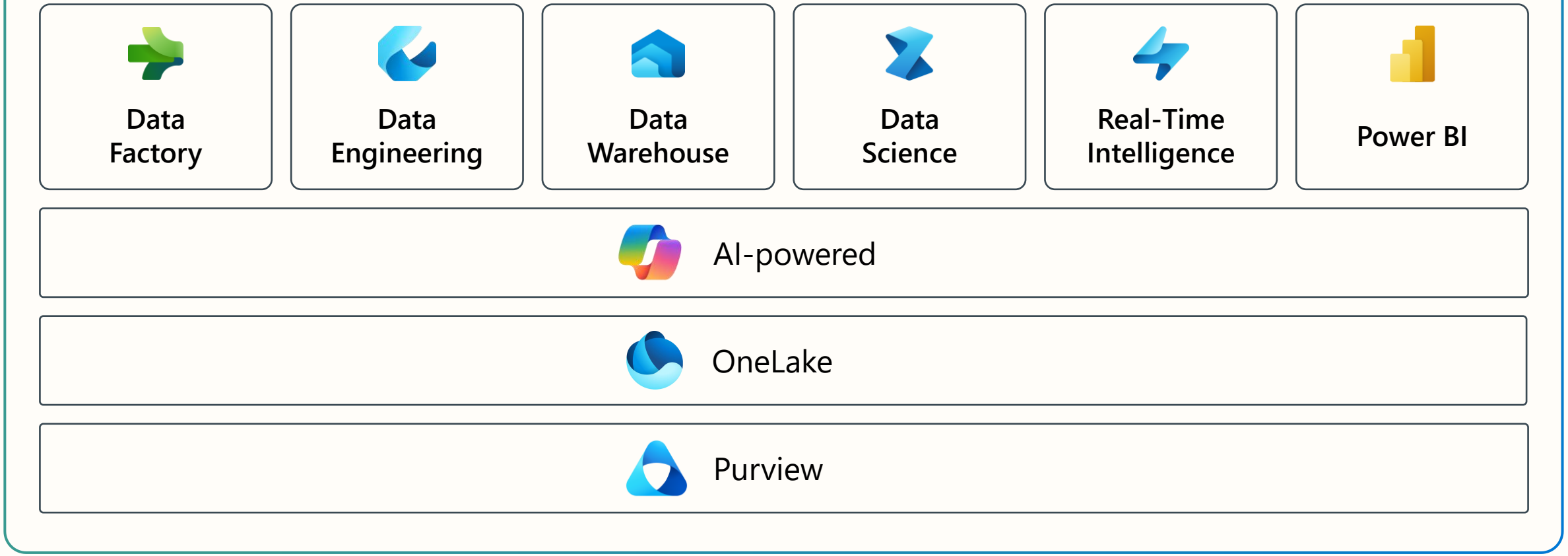
Login with demo user profile.

Don’t use Multi-factor authentication.

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

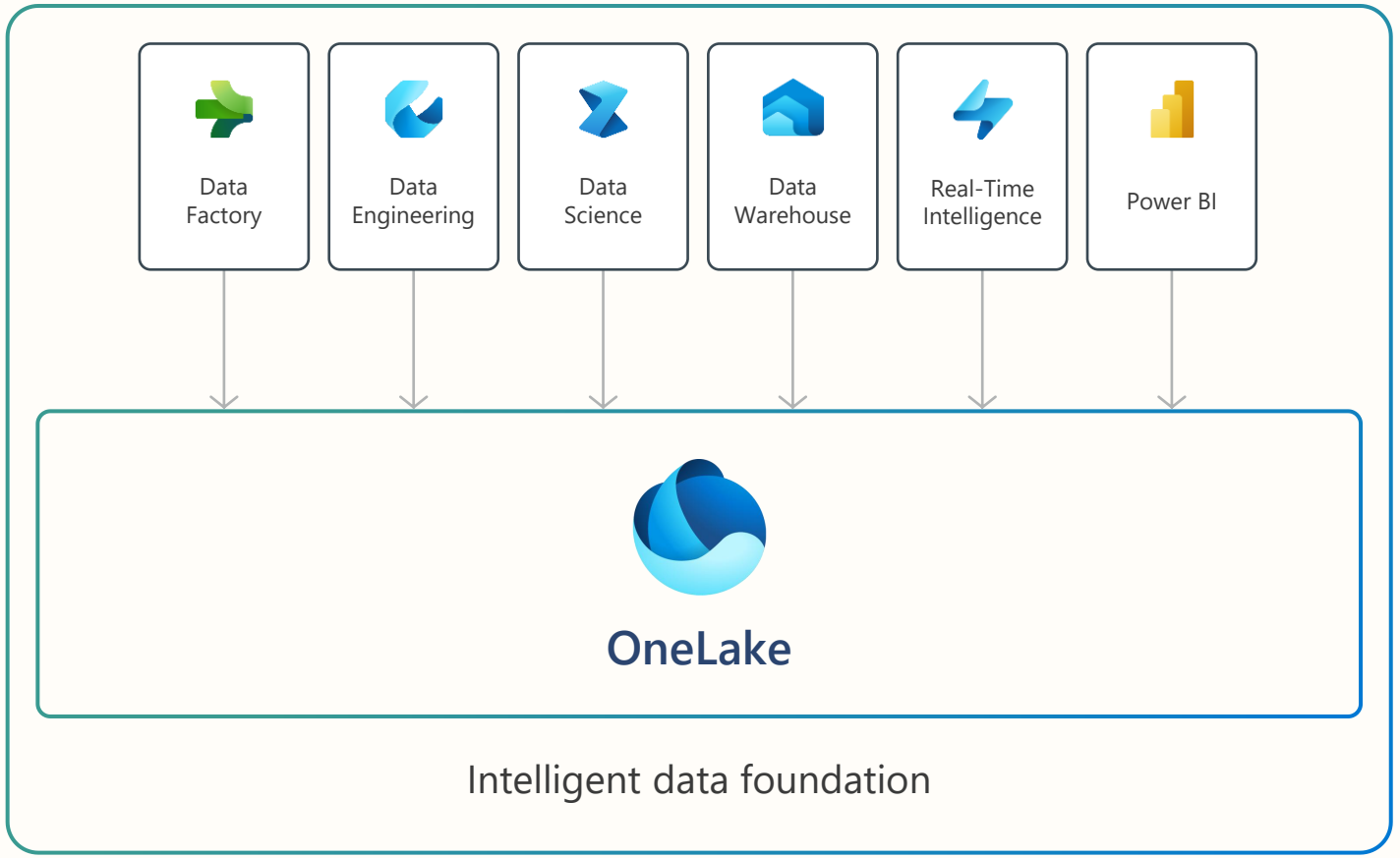


“The OneDrive for Data”

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

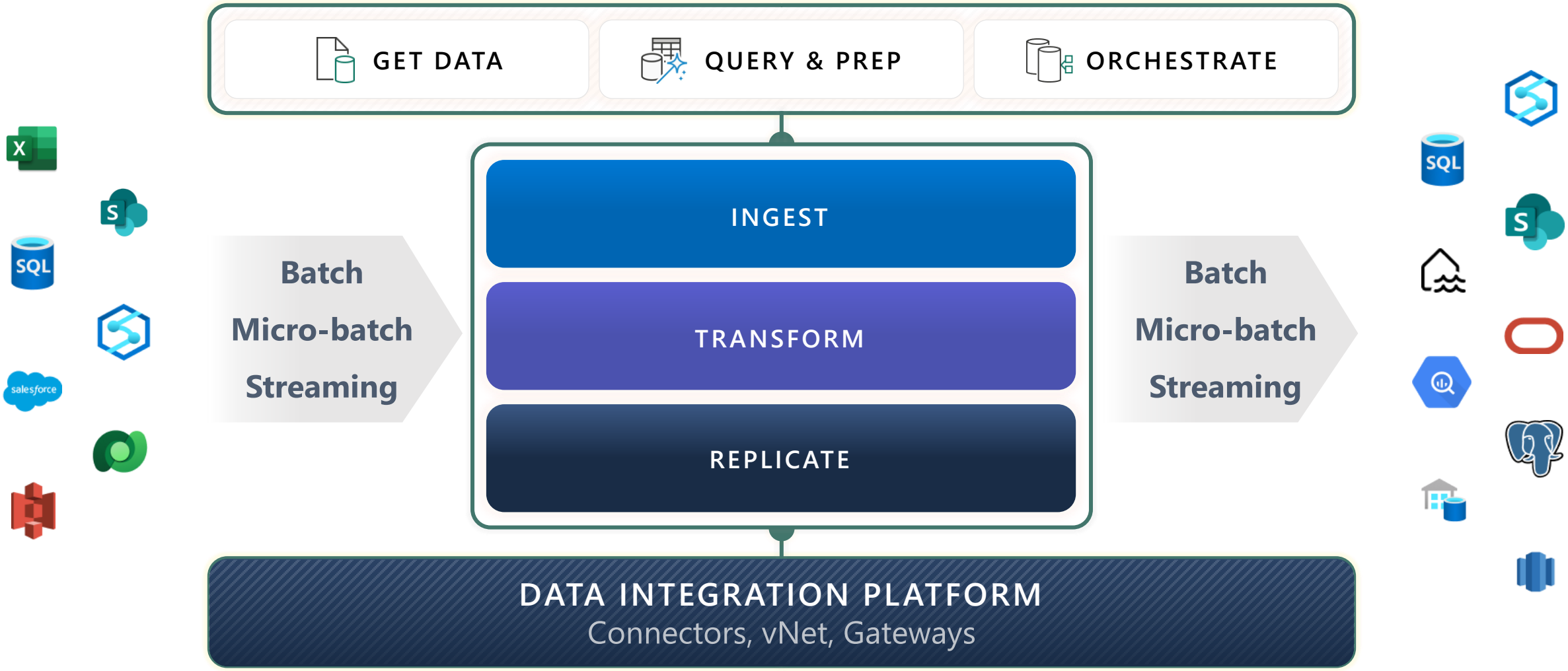
Copilot accelerated

ChatGPT on your data

AI driven insights

Data Factory overview

Microsoft Data Integration

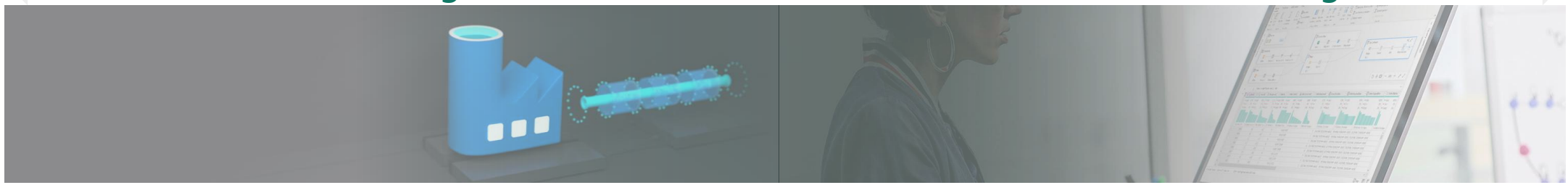


Microsoft Data Integration

Products

Professional Data Integration

Citizen Data Integration



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

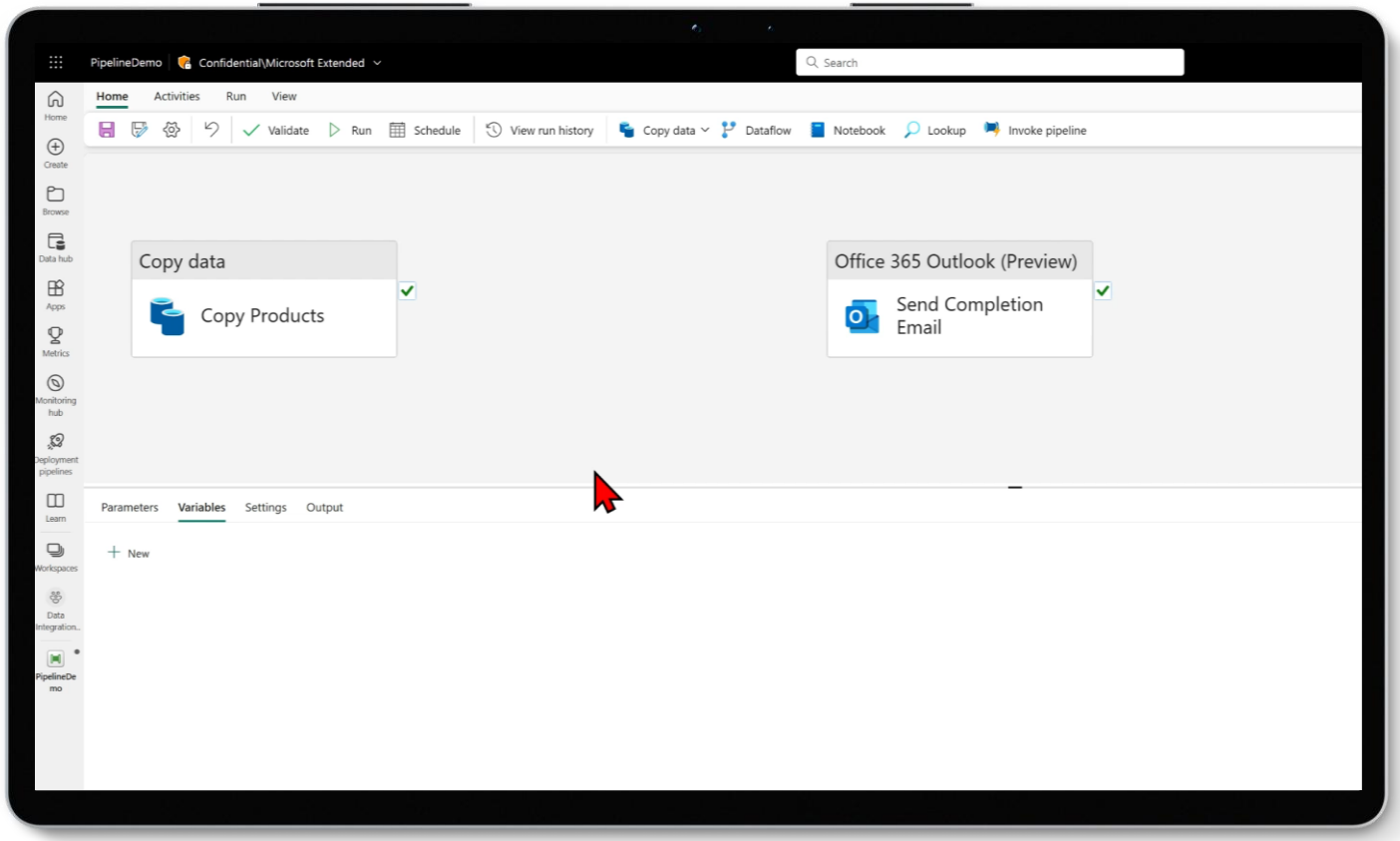
Professional & Citizen Data Integration

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

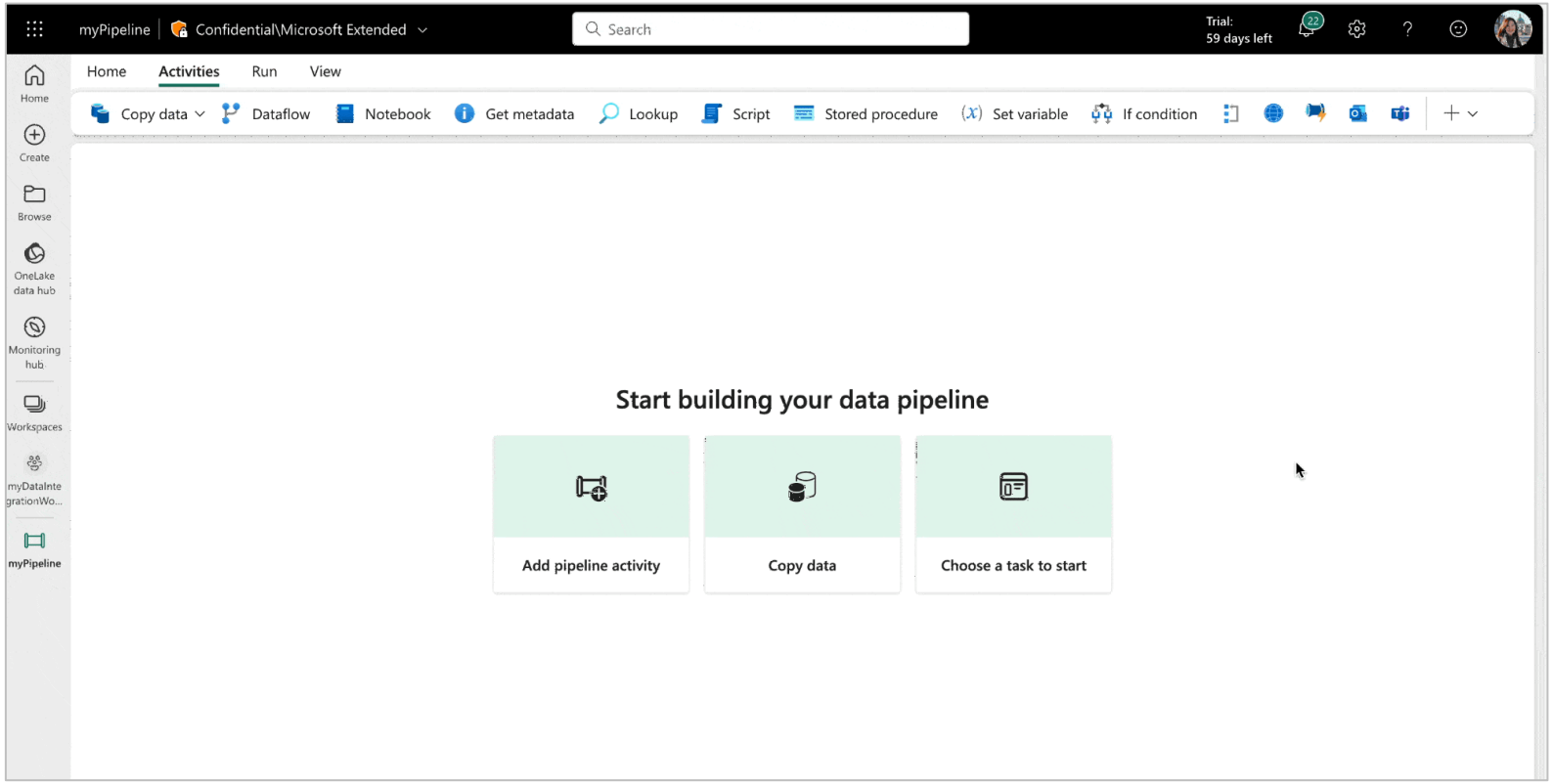
Data Factory converges our data capabilities into a single SaaS interface to provide **the world's most complete data integration experience.**



Data pipeline

Ingest and orchestrate activities at scale

Familiar authoring canvas experience

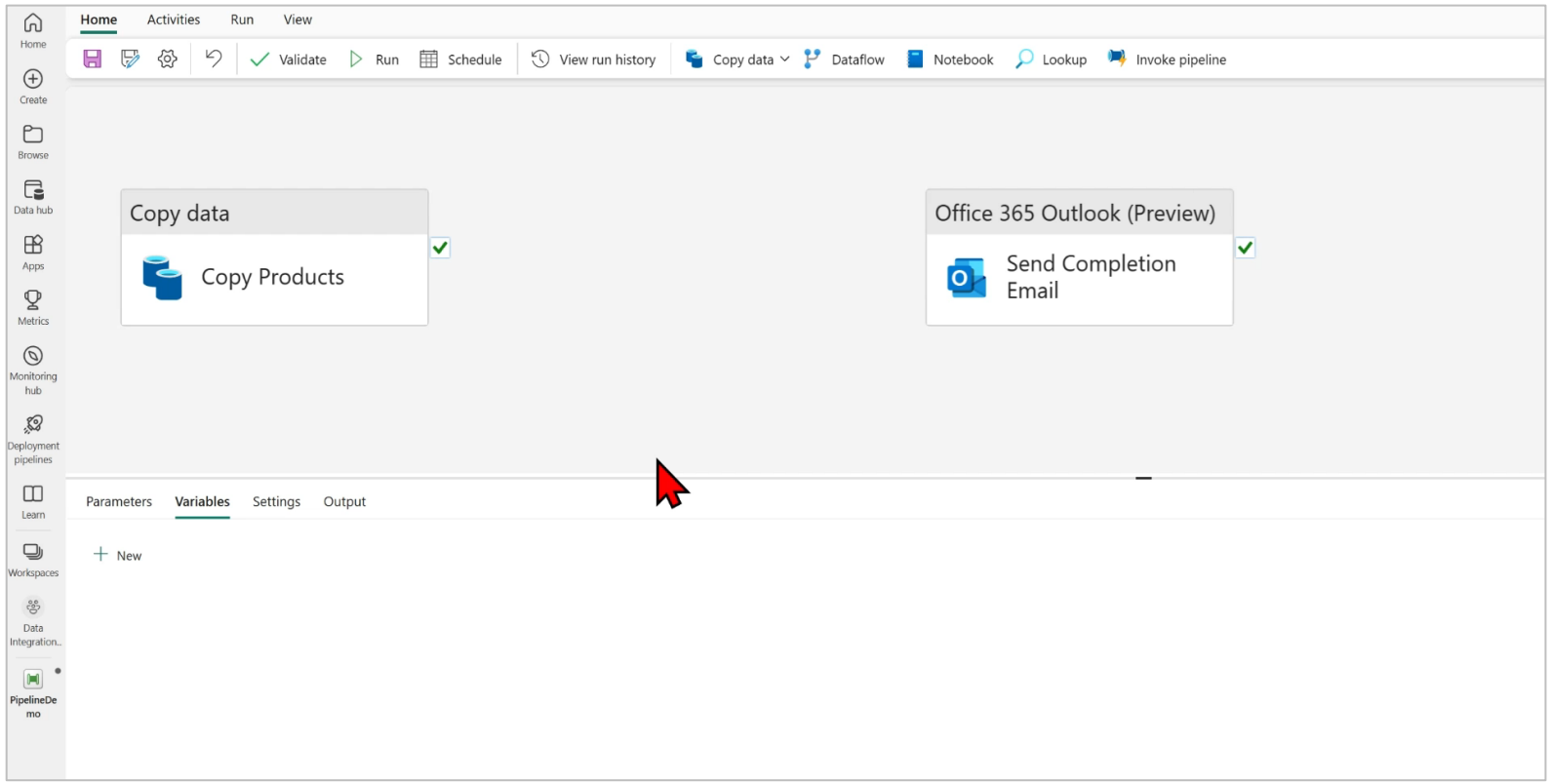


OneLake

Data pipeline

Ingest and orchestrate activities at scale

Empower every person to integrate data



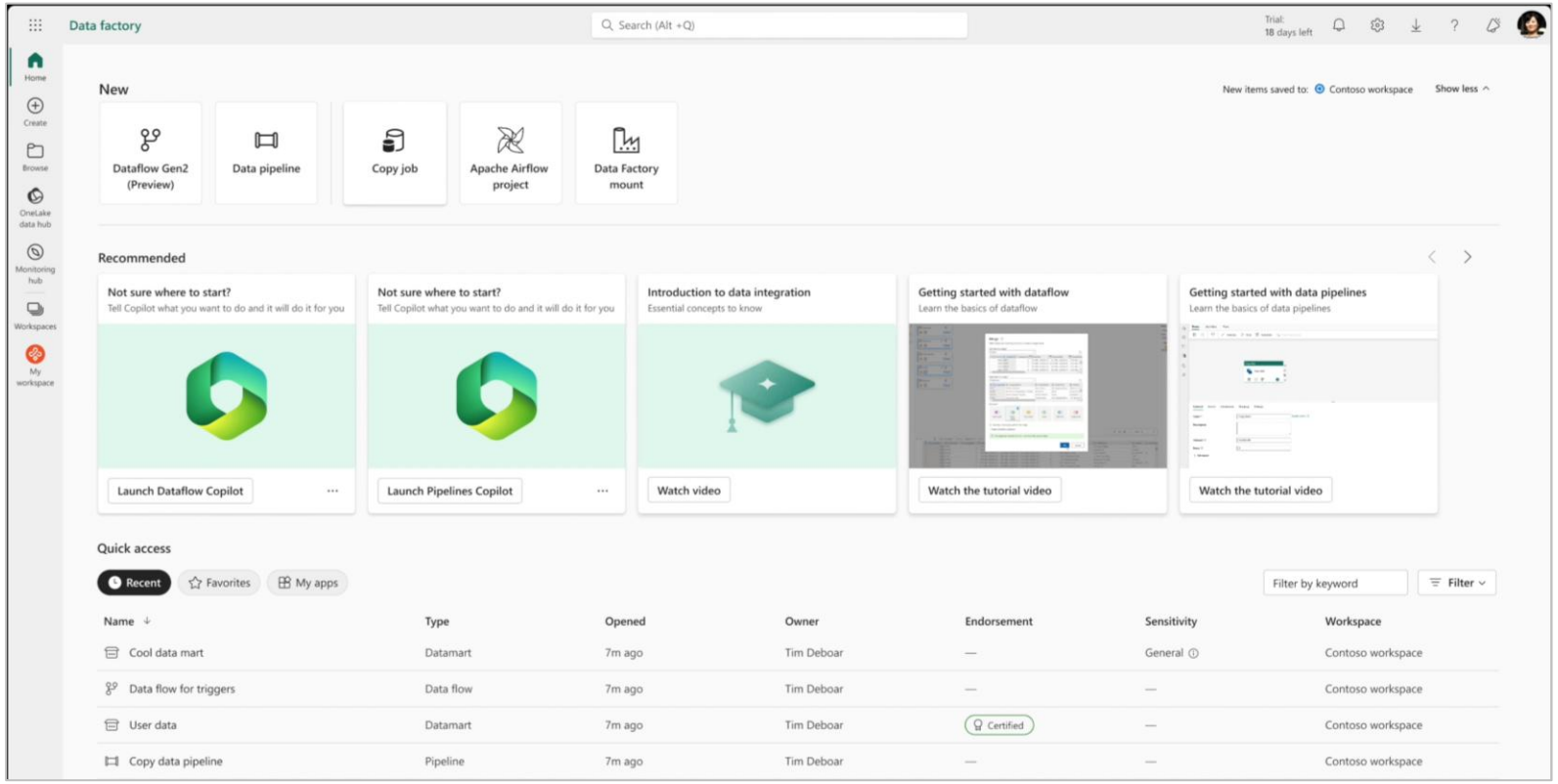
OneLake

Copy job

Easily ingest and move data at scale

Public Preview

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

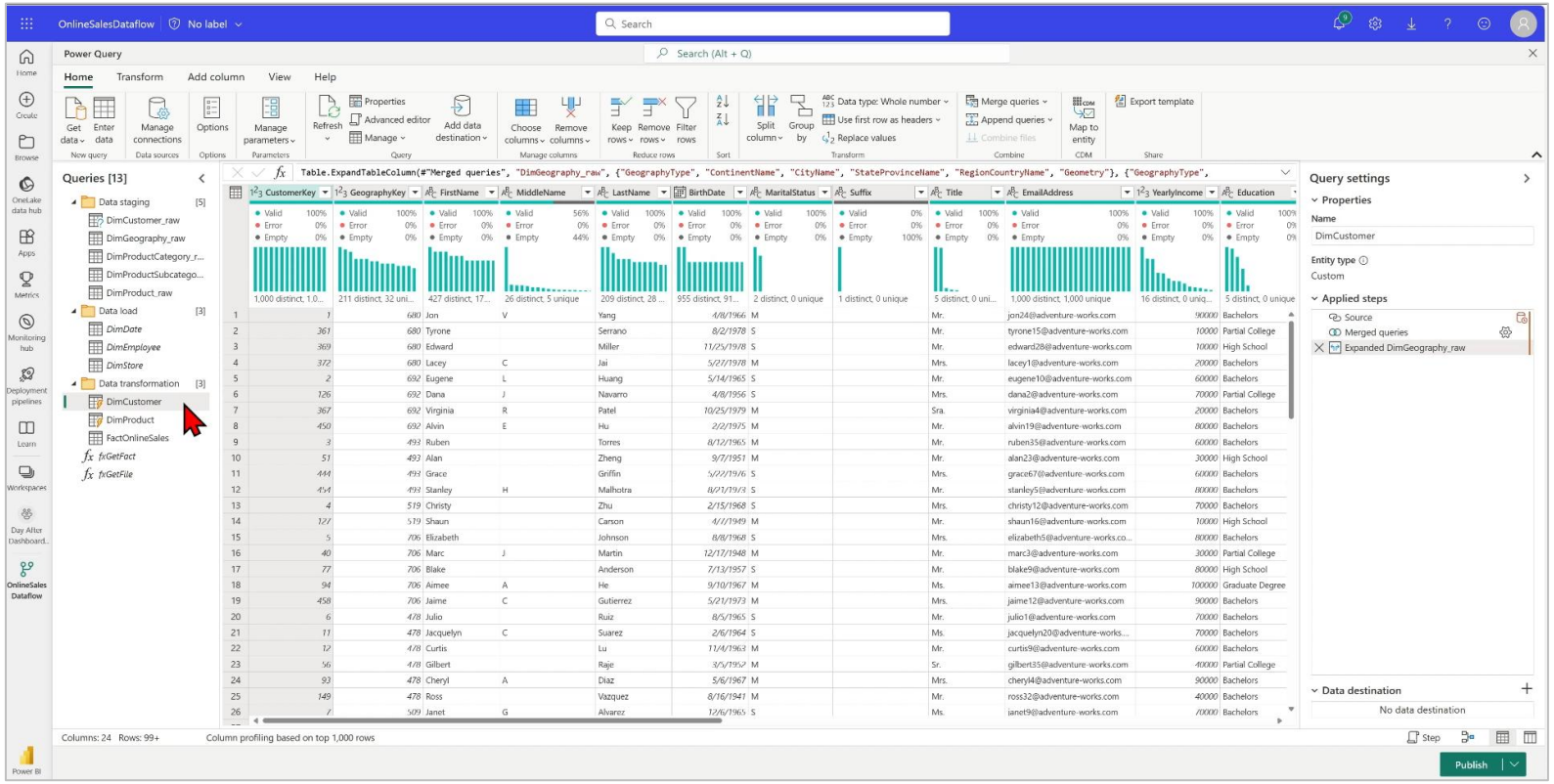


OneLake

Dataflow Gen 2

Ingest and transform data at scale

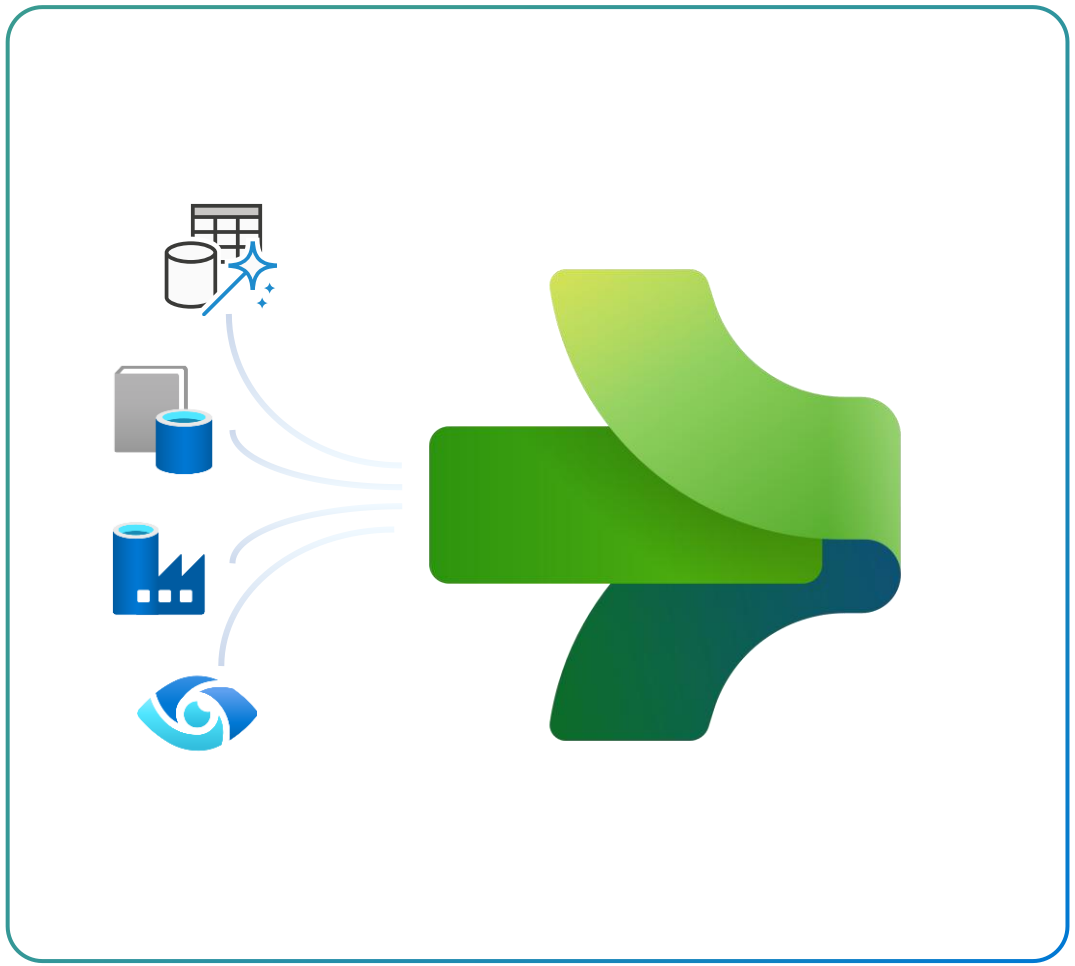
Enterprise-scale data ingestion and transformation

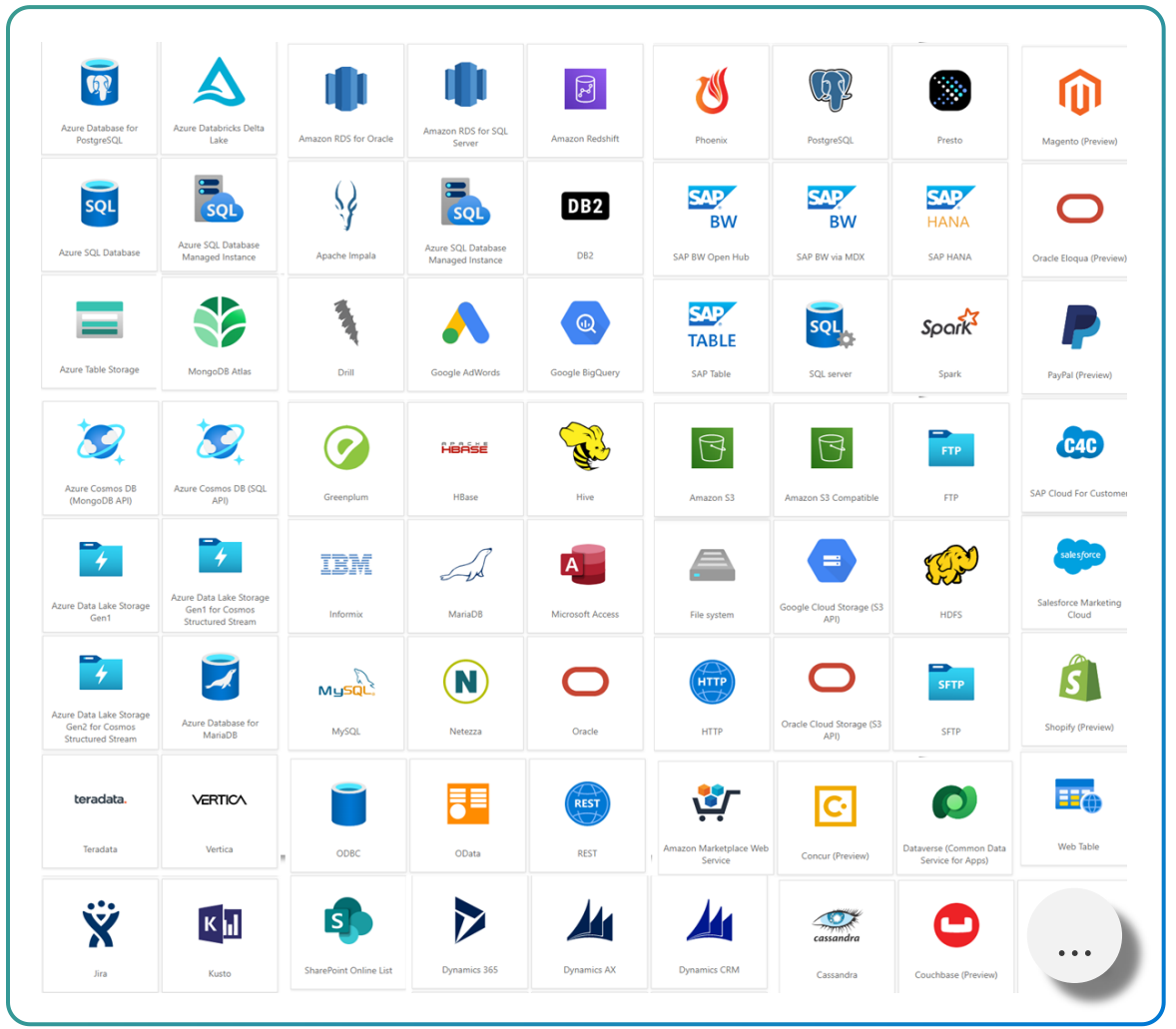


OneLake

Unifying data in OneLake

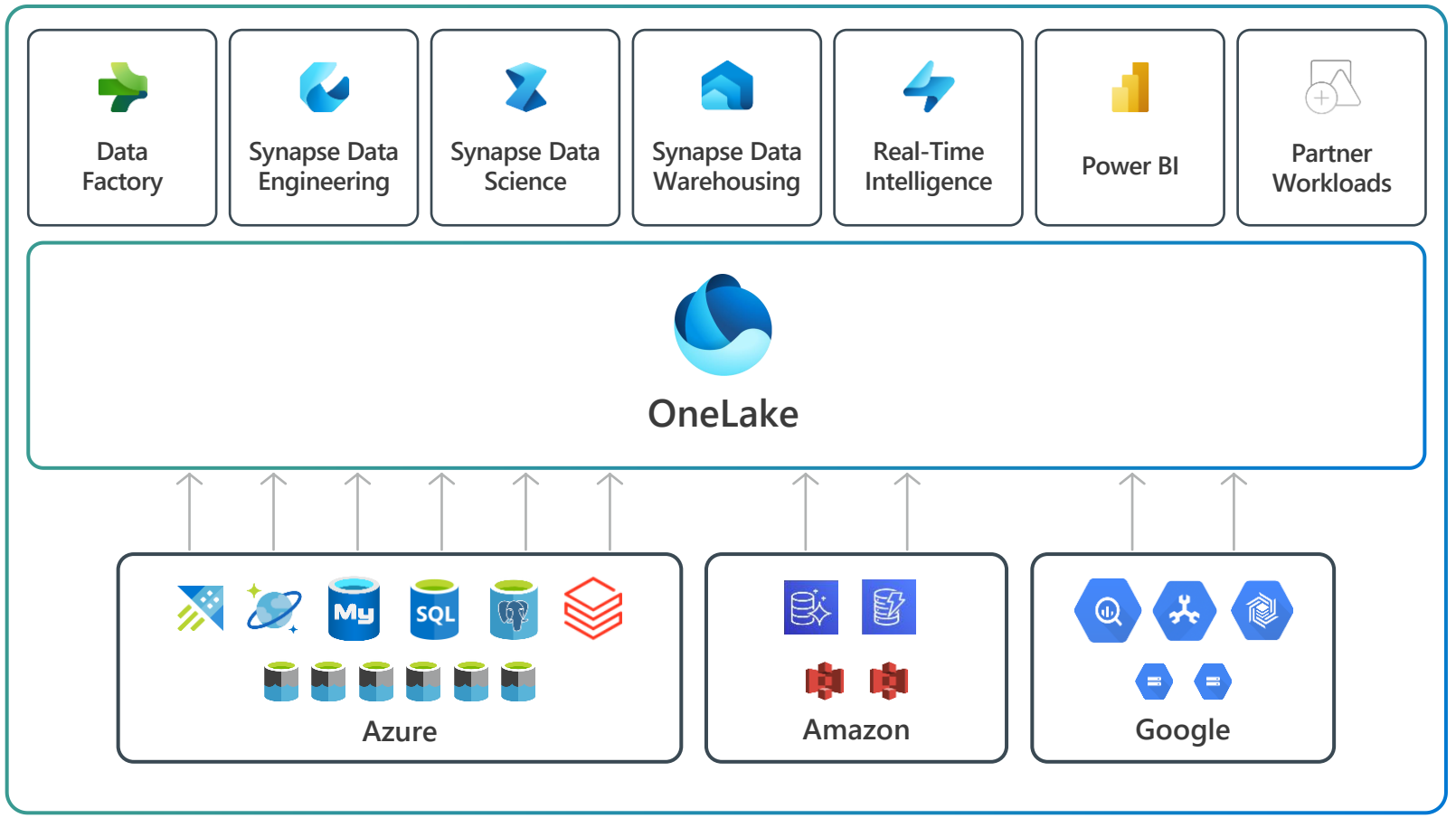
Seamlessly connect to more than 170+ data sources





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

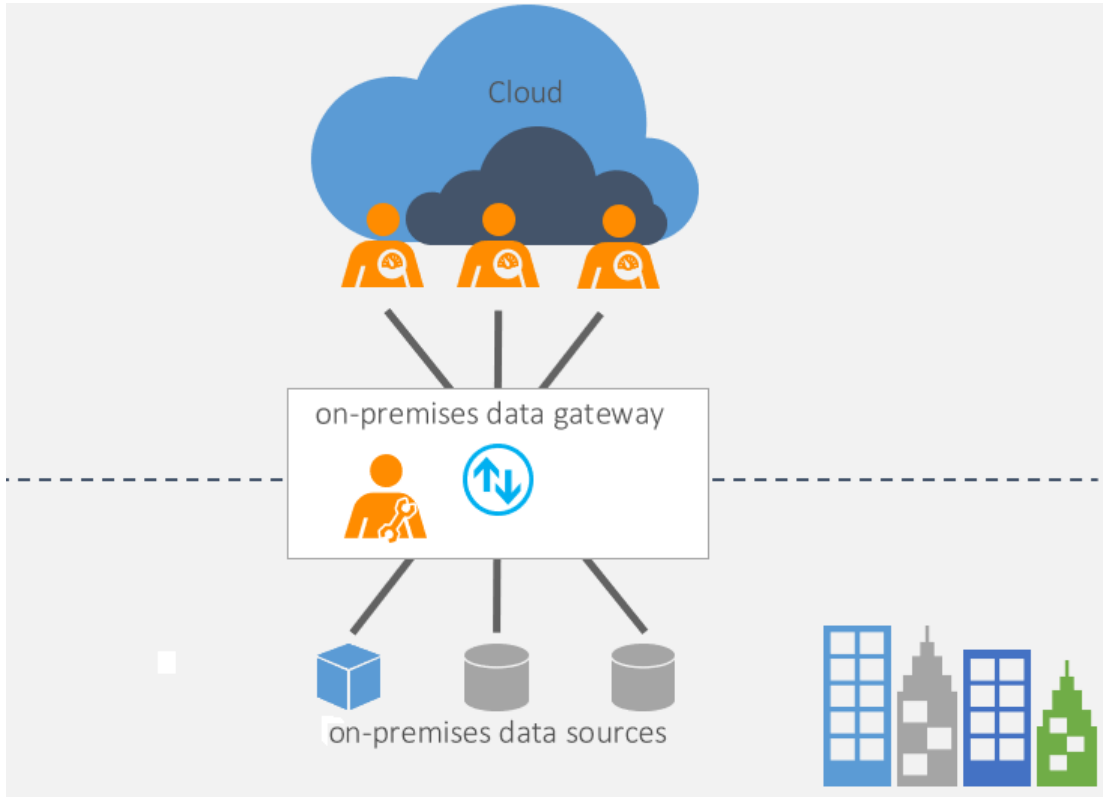
Real time updates of the replicas using the CDC feeds of the database

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

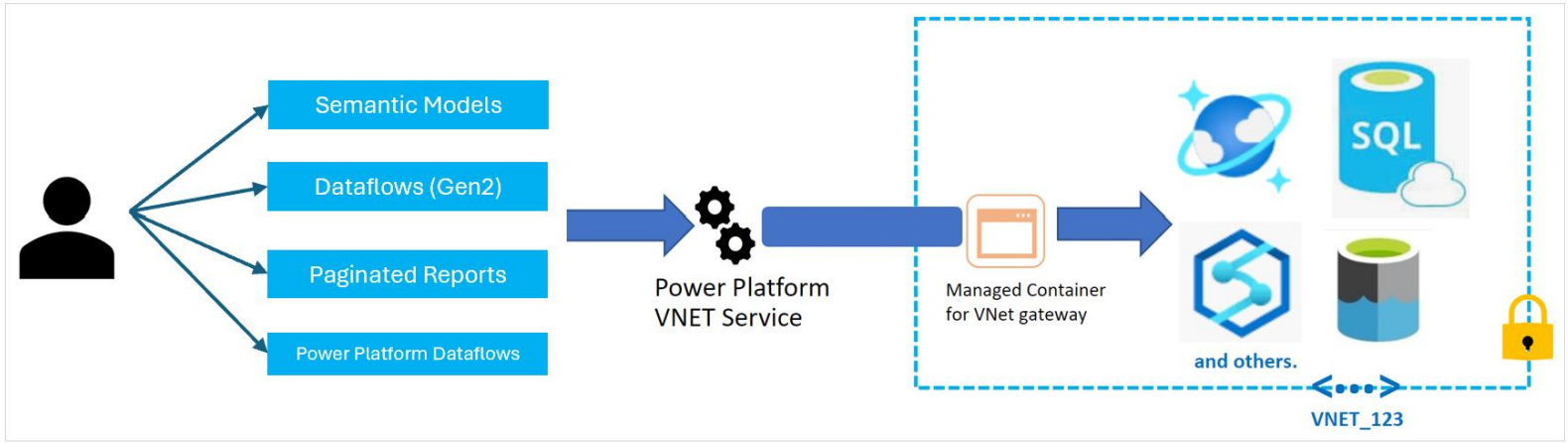


OneLake

VNet data gateway

Secure connectivity to your Azure and private data services

Bridge between your secure data sources and the Microsoft cloud



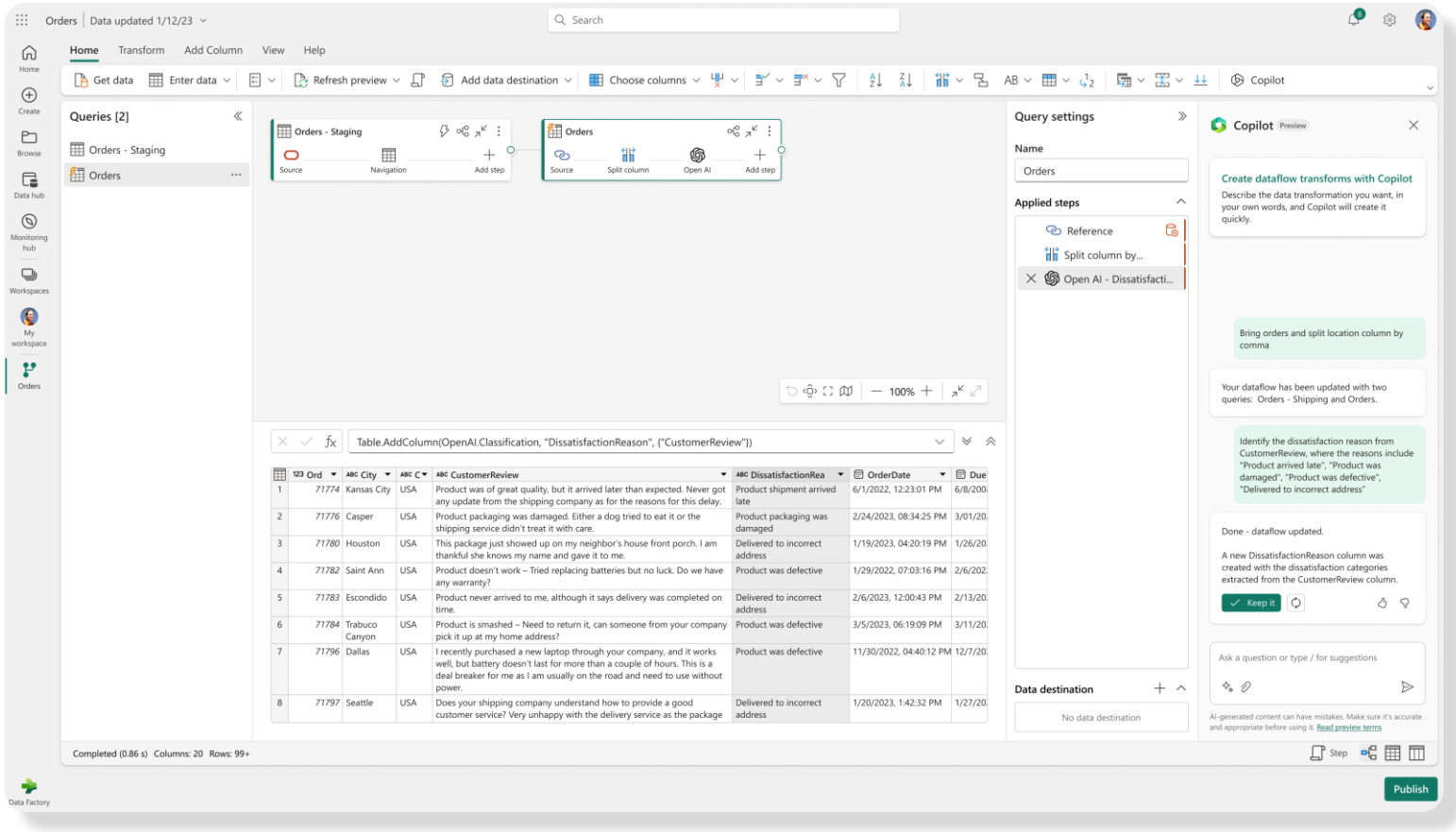
OneLake

Data integration + AI

Copilot in Data Factory

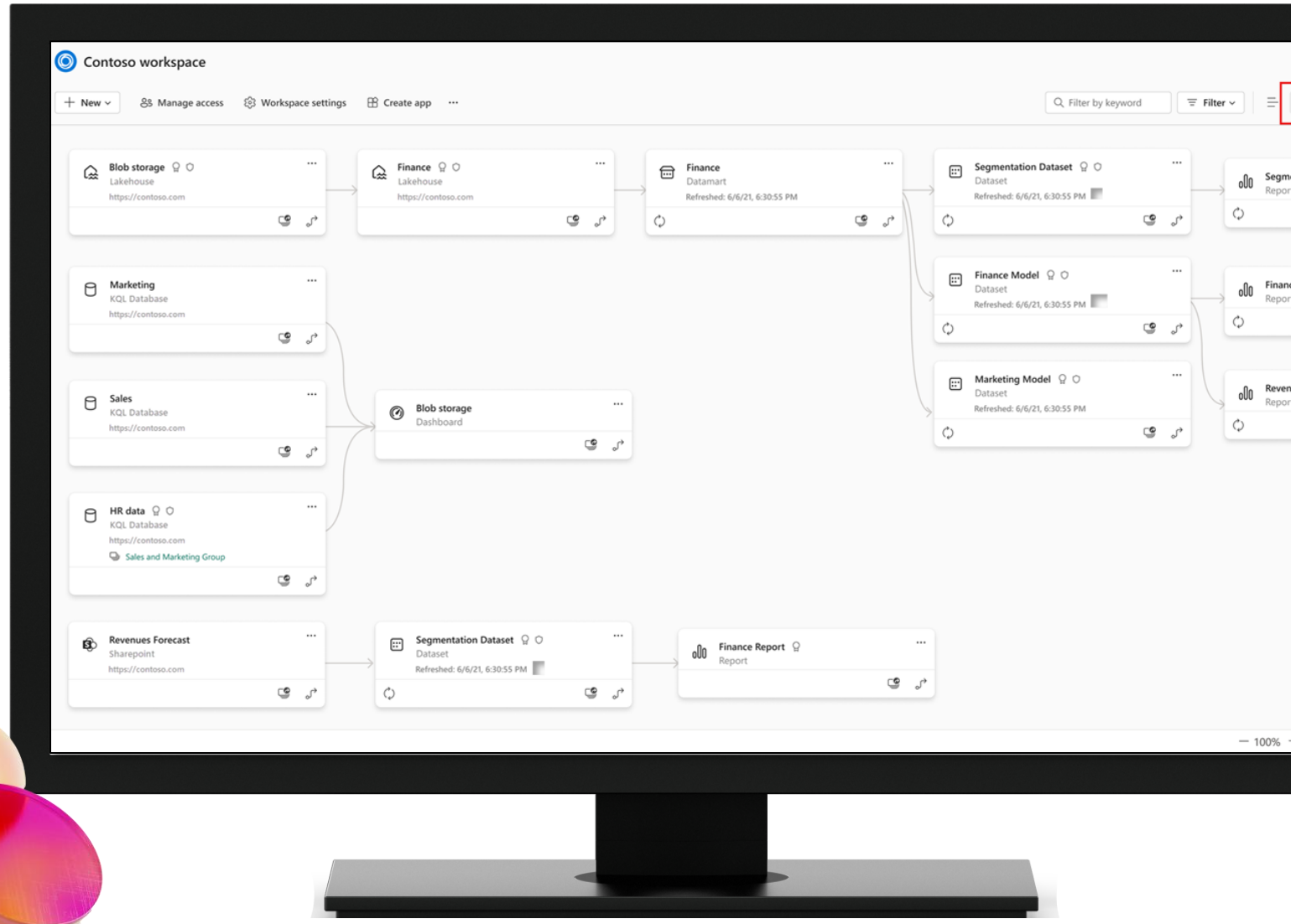
Easily integrate generative AI into your dataflows and *pipelines using Copilot

- 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



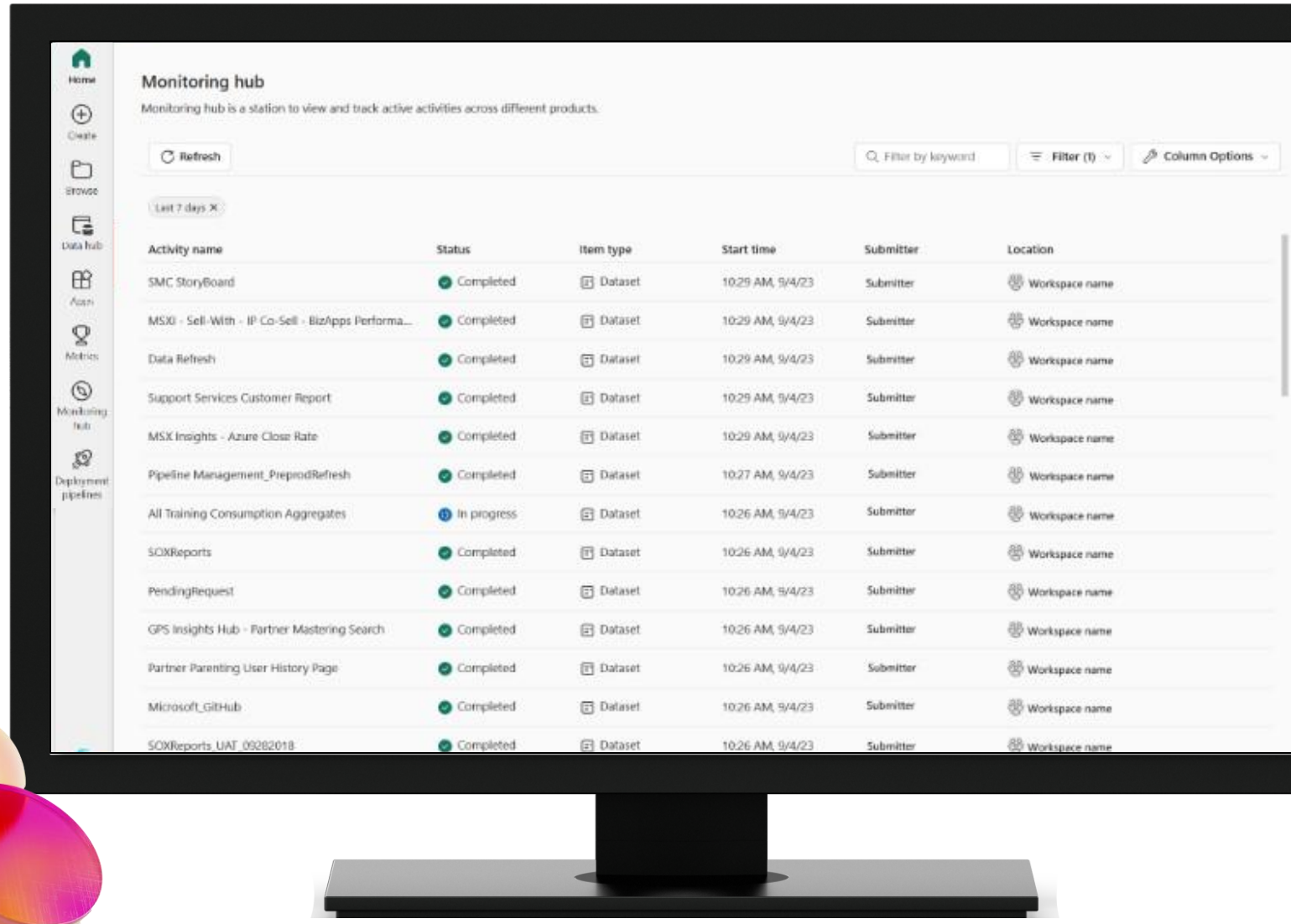
Lineage view

- 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

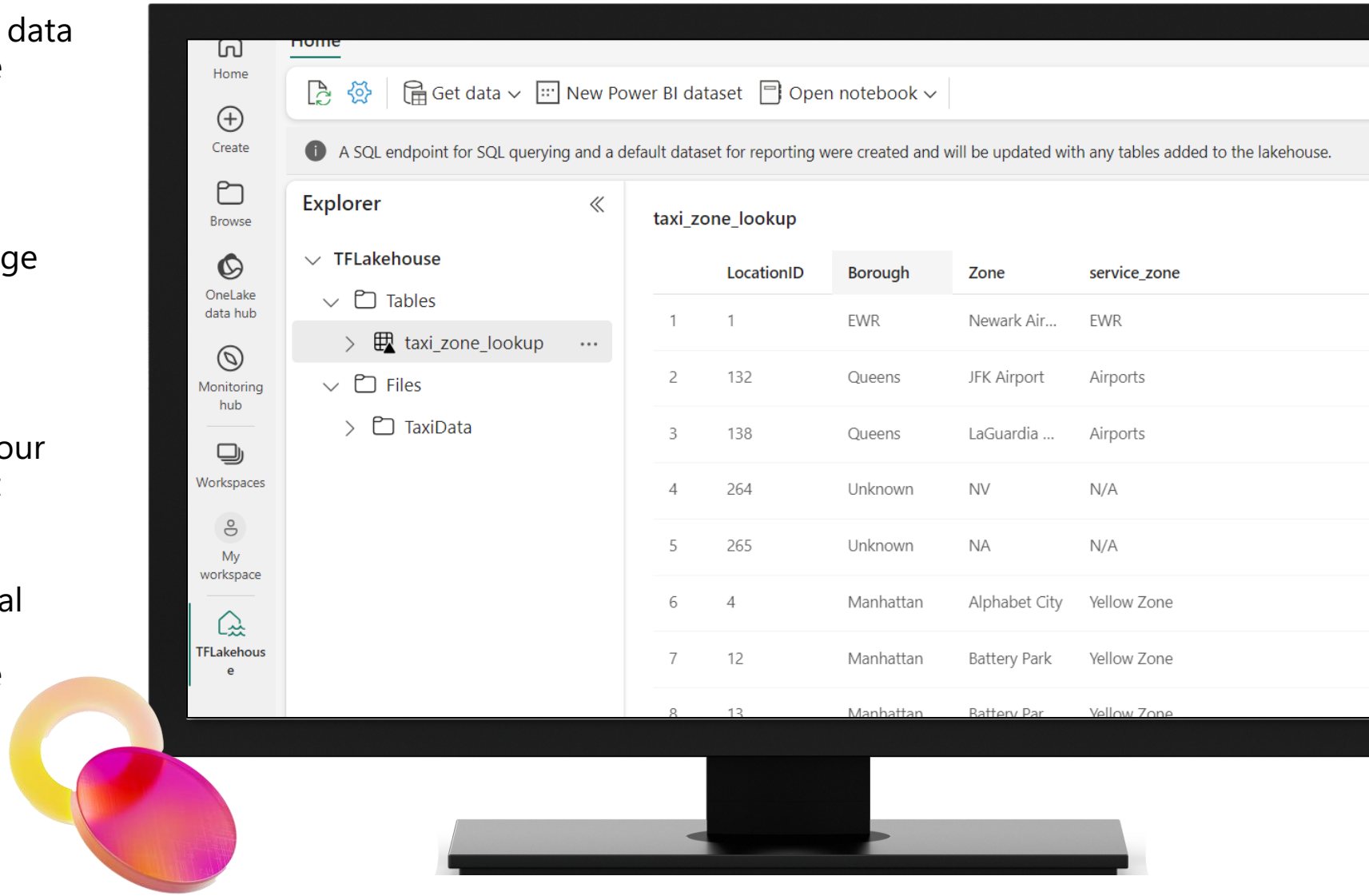
- Monitor activities from a central location
- Gain deep insights into how your Fabric items are performing – including dataflows and pipelines



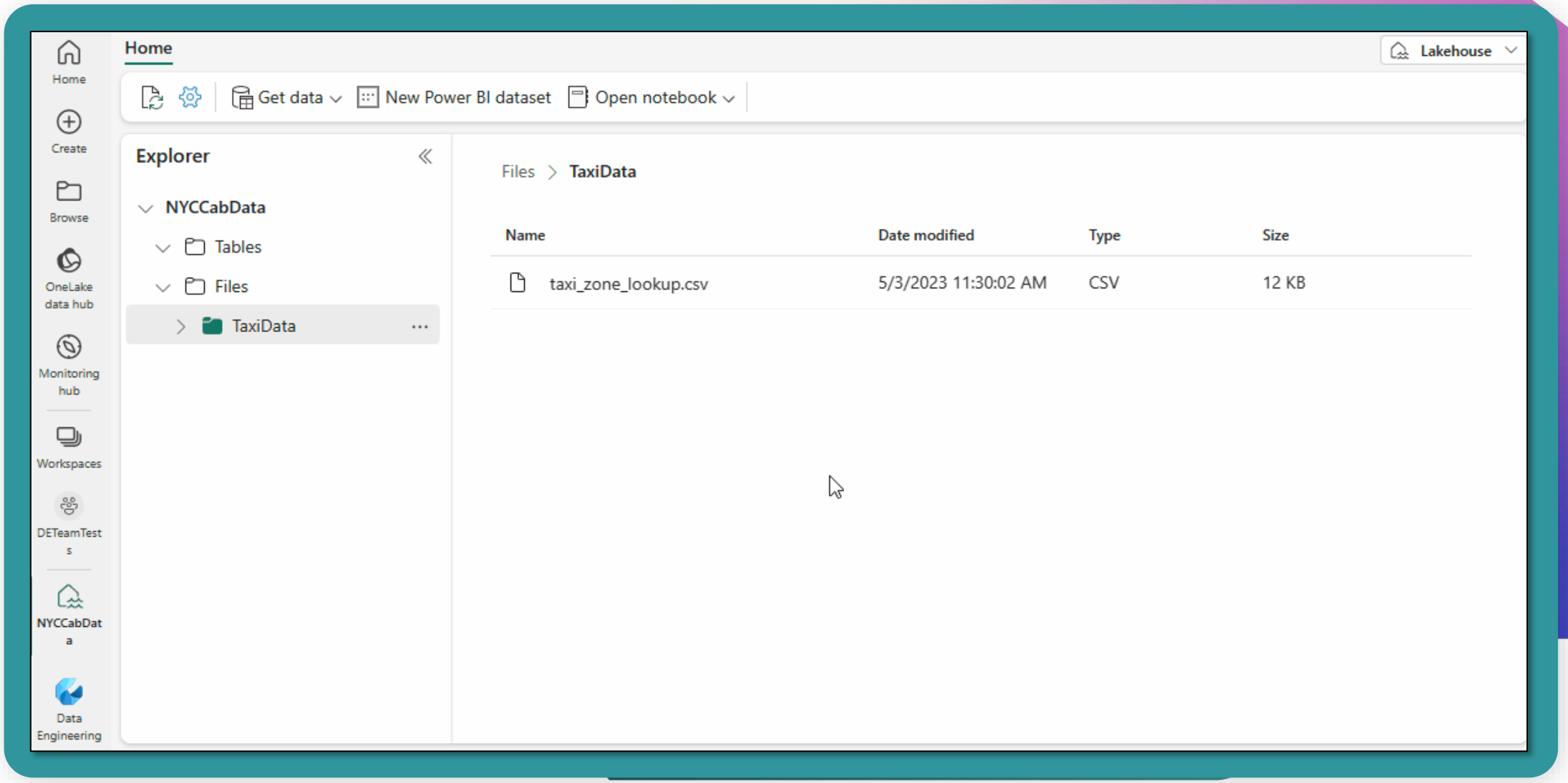
Lakehouse overview

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



Automatic table discovery and registration

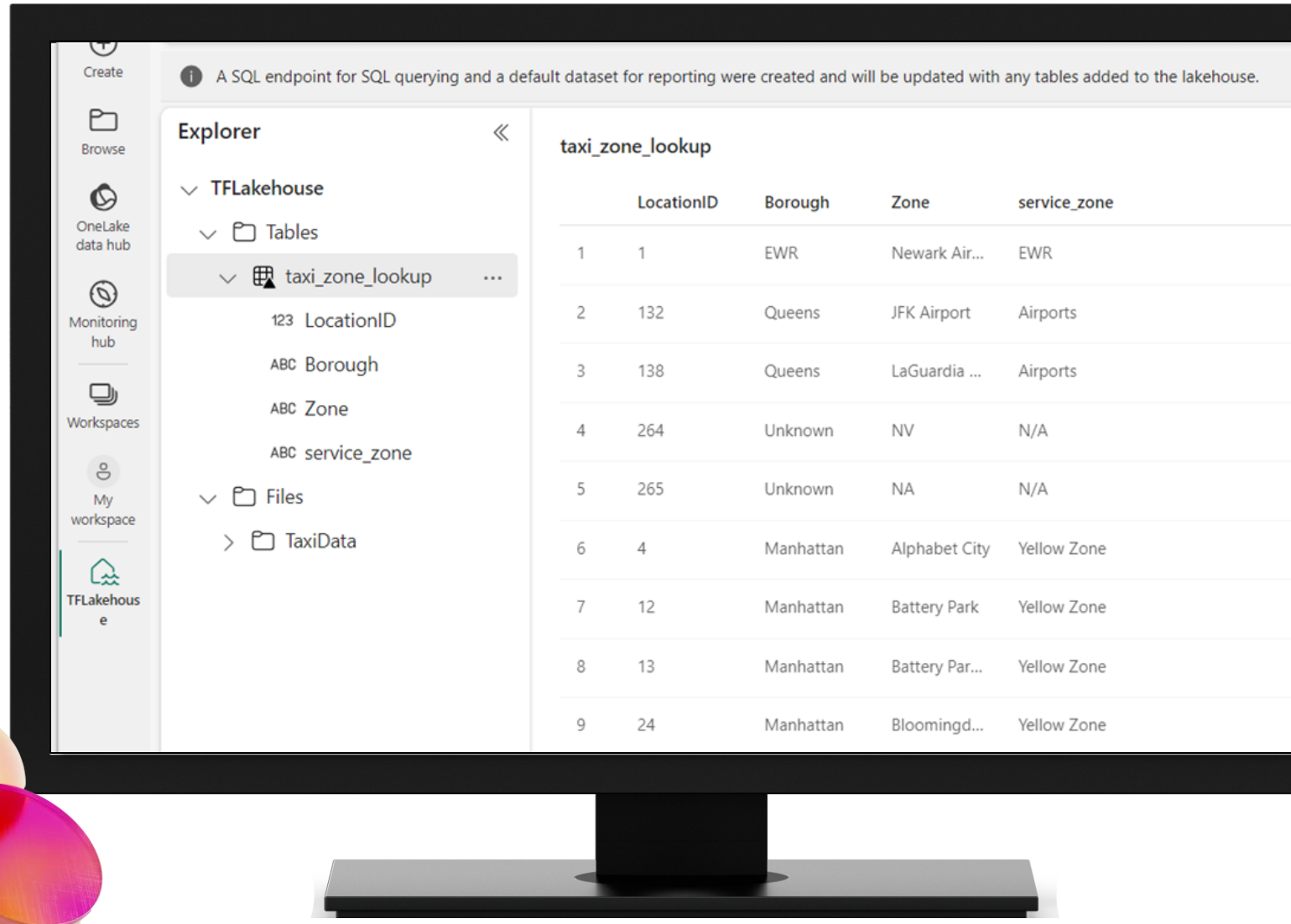


Automatic discovery of Delta Lake tables

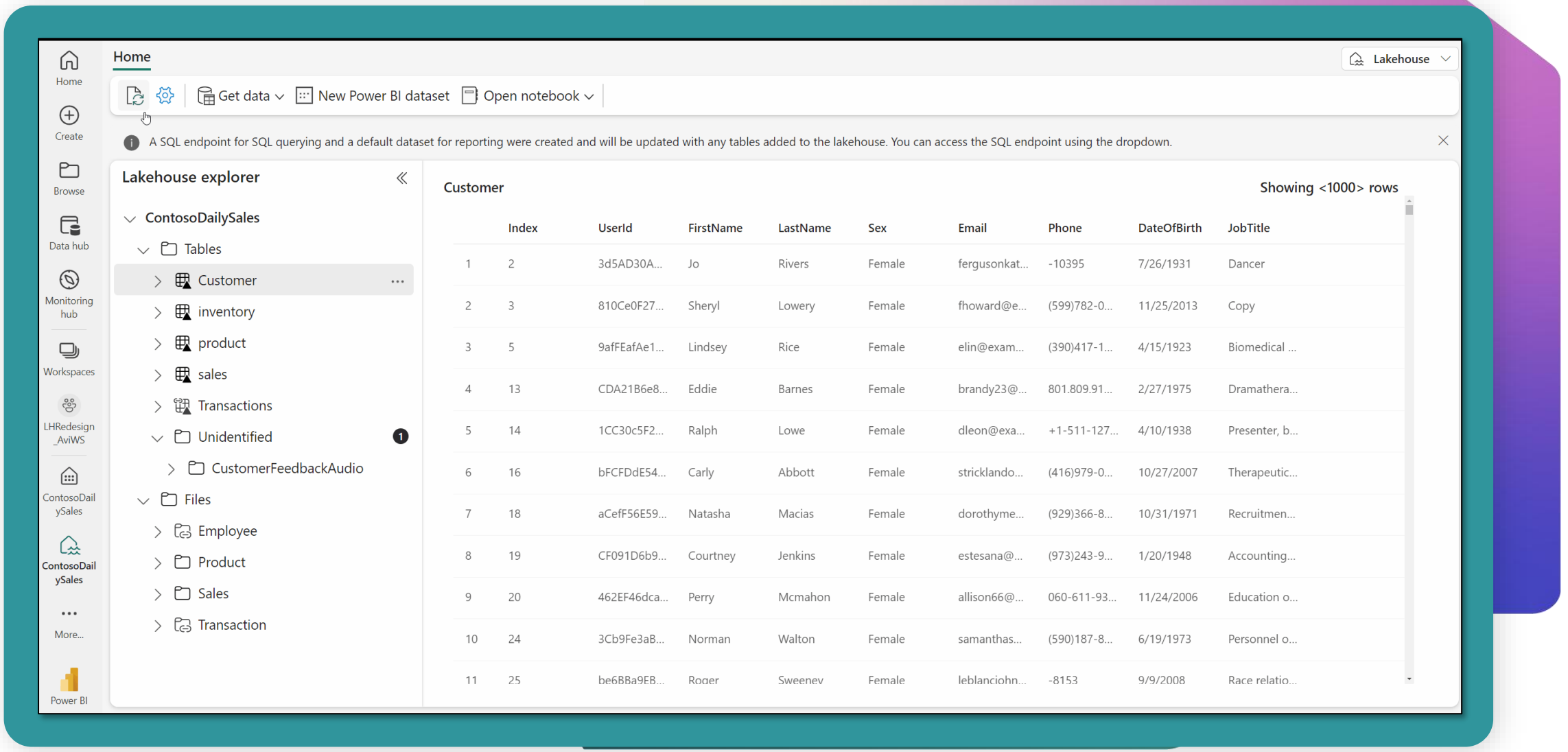
The Lakehouse explorer provides a tree-like view of the objects in the Microsoft Fabric Lakehouse item.

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



Navigating the Fabric Lakehouse explorer

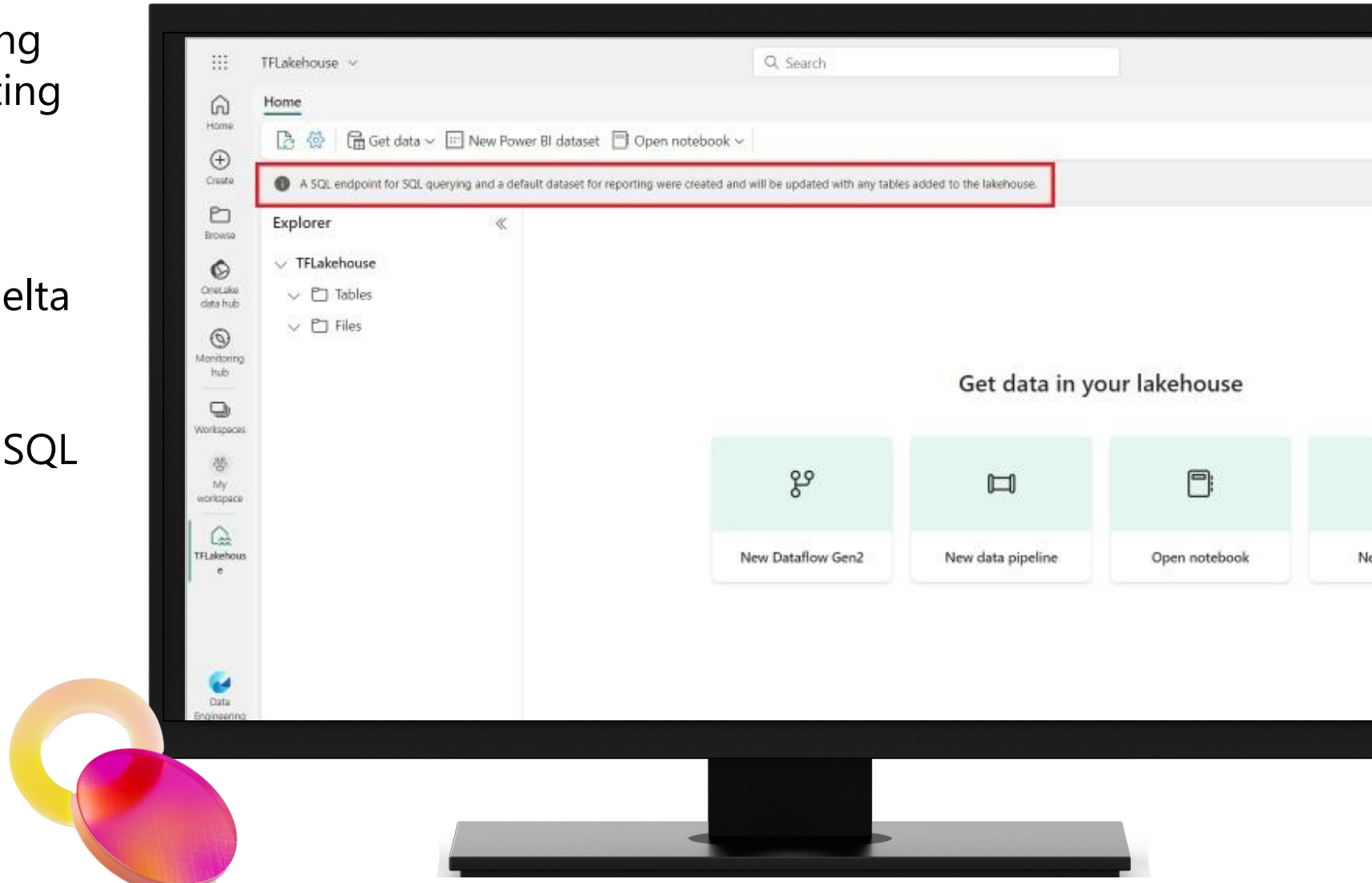


Lakehouse SQL endpoint

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

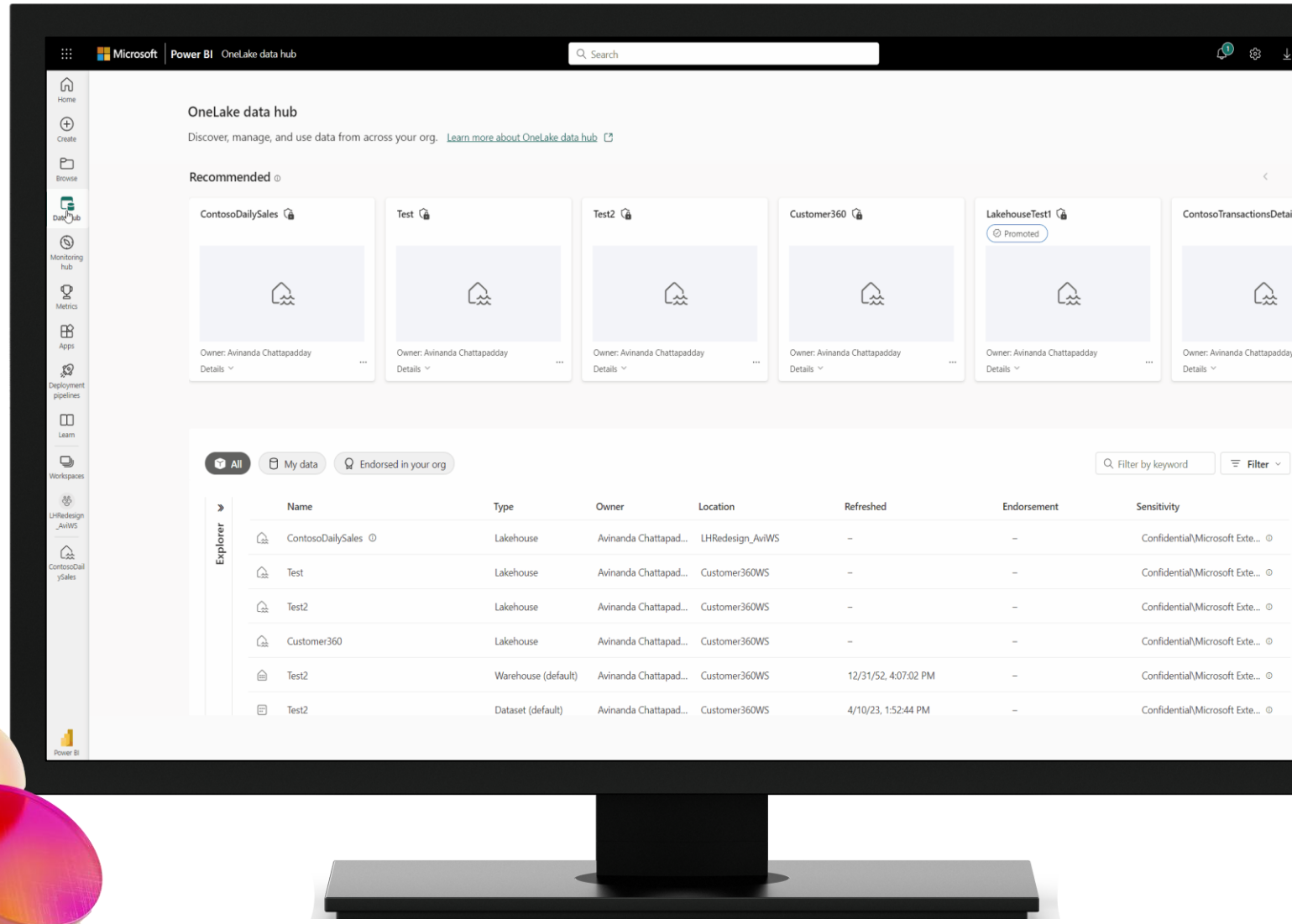
You can only read data from delta tables using SQL endpoint

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



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



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

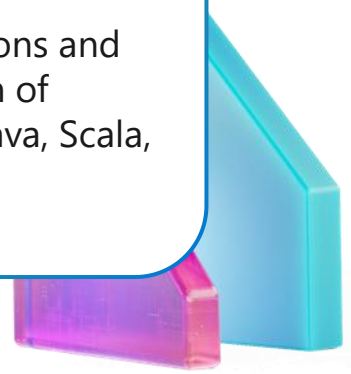
Develop robust applications and orchestrate the execution of compiled Spark jobs in Java, Scala, and Python

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

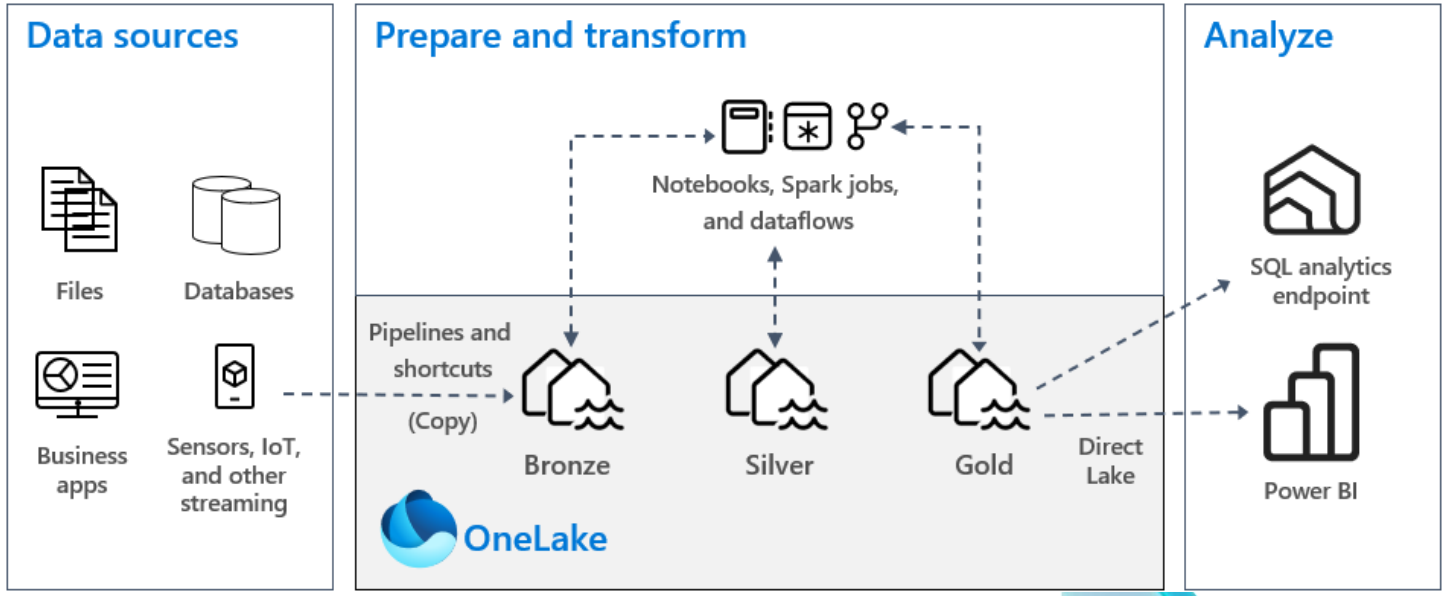
Design and implementation guidance

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.

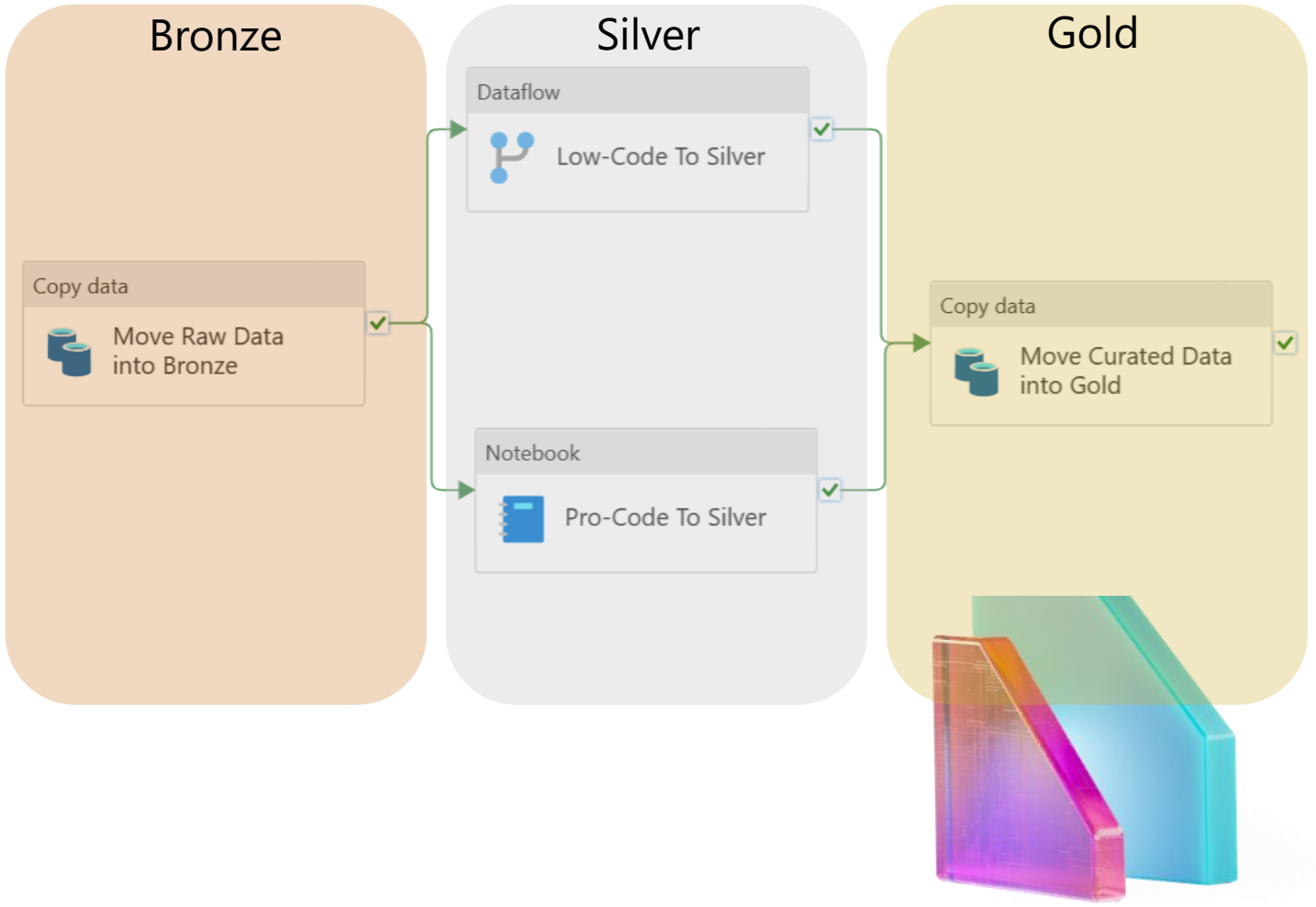


In a typical medallion architecture implementation in Fabric, the bronze zone stores the data in the same format as the data source. When the data source is a relational database, Delta tables are a good choice. The silver and gold zones contain Delta tables.

Medallion Lakehouse architecture

Distinct layers (or zones).

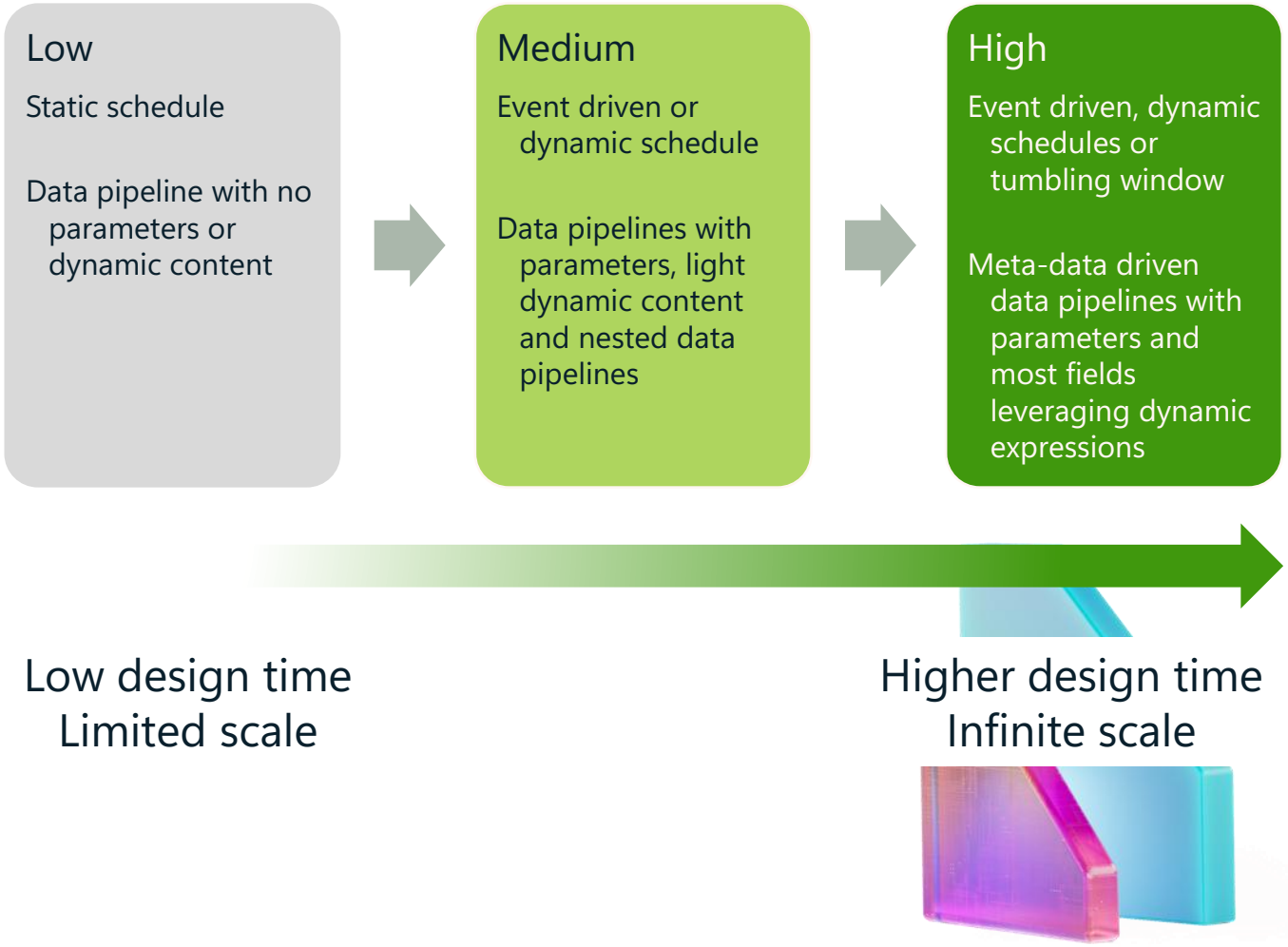
- **Bronze (raw zone):** This layer stores source data in its original format, which is typically append-only and immutable.
- **Silver (enriched zone):** Data from the bronze layer is cleansed, standardized, and structured as tables. It may also be integrated with other data to provide a comprehensive 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).

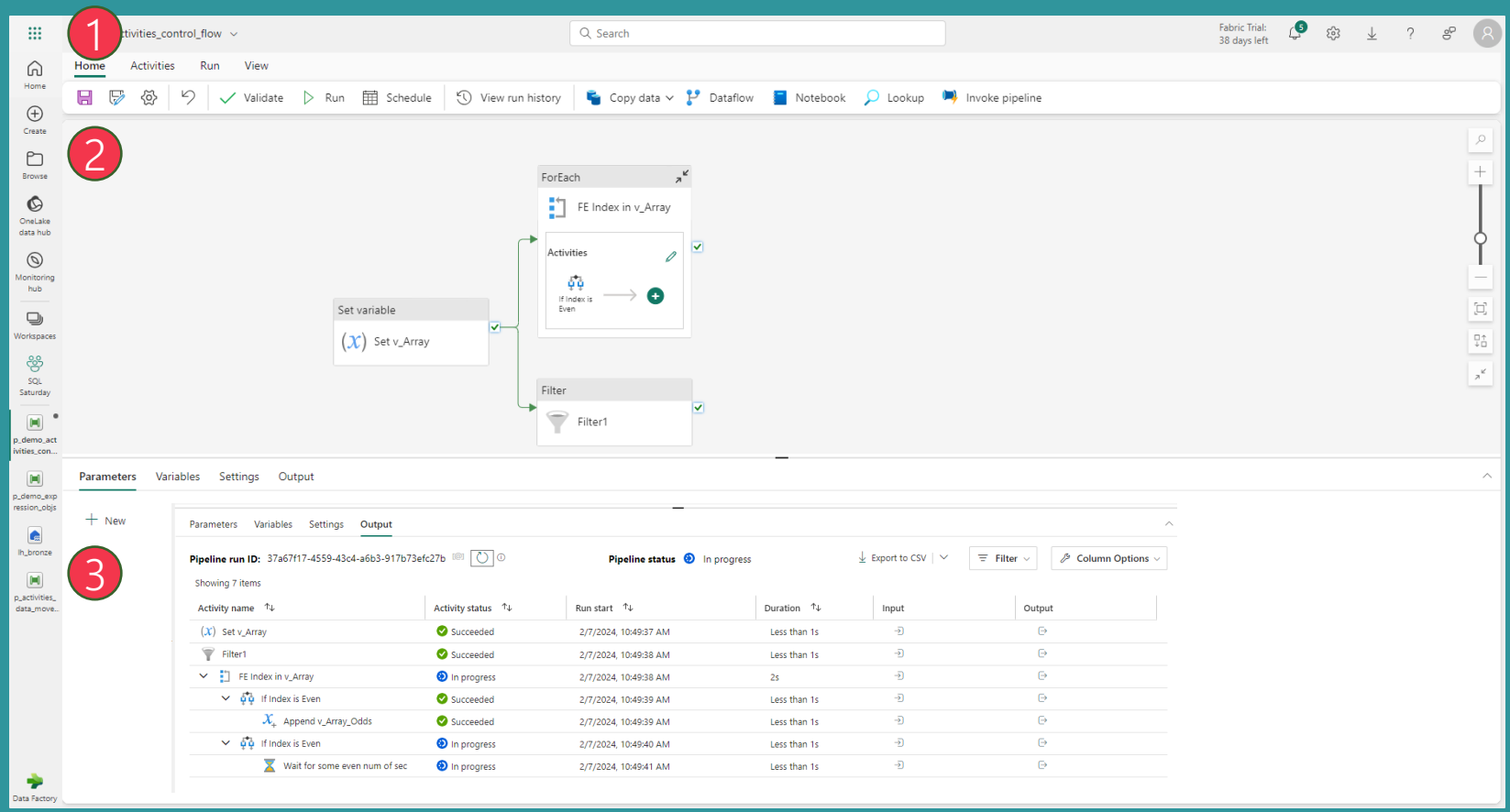
- **Varied Complexity**: Architecture designs of data pipelines can range widely in complexity due to the vast array of design options.
- **Reusability**: Design with reusability in mind, incorporating parameters and components that can be reused, such as invoking other pipelines, running notebooks or 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

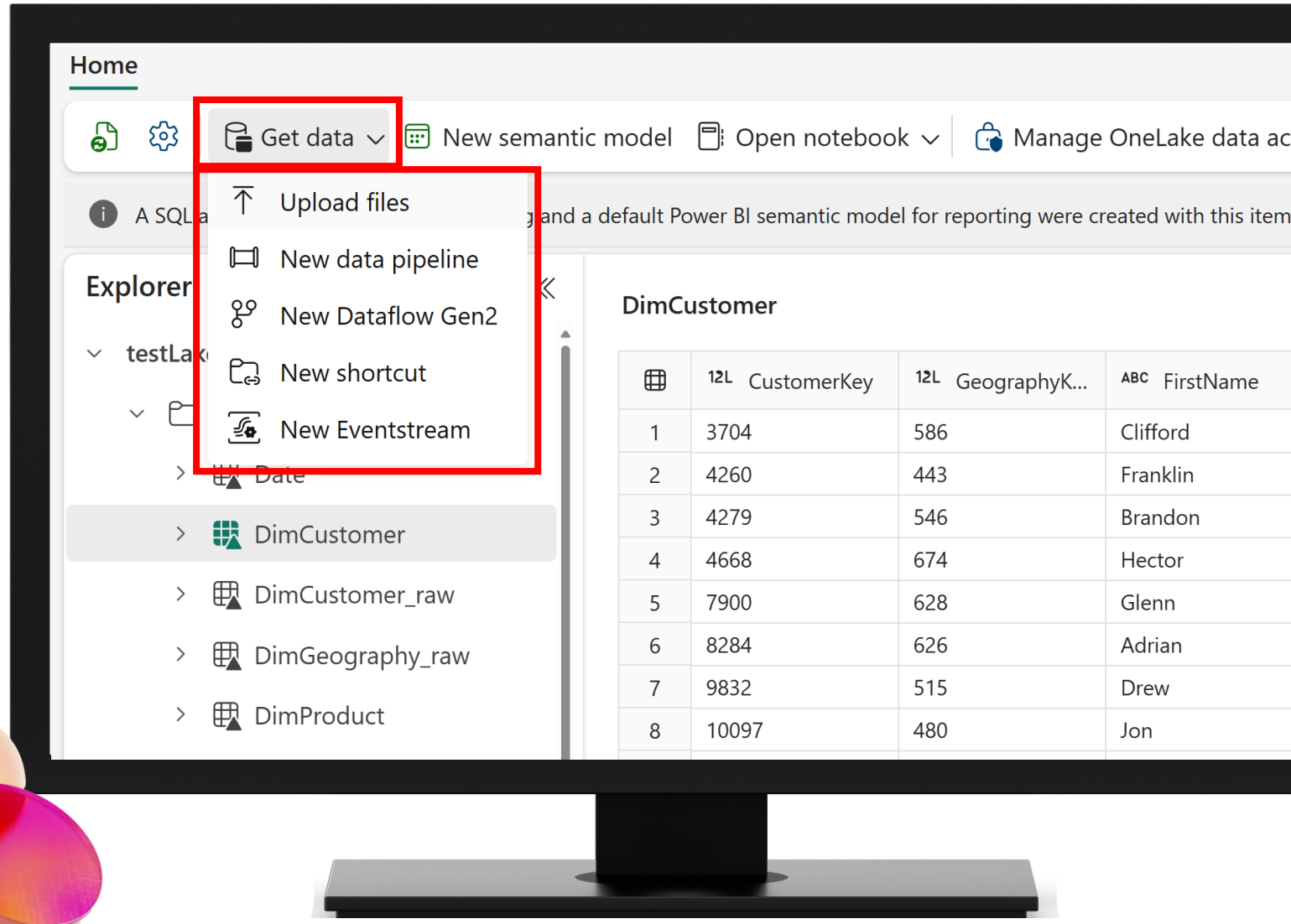
Creating a…

- Task flow
- Lakehouse
- Sample data pipeline

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

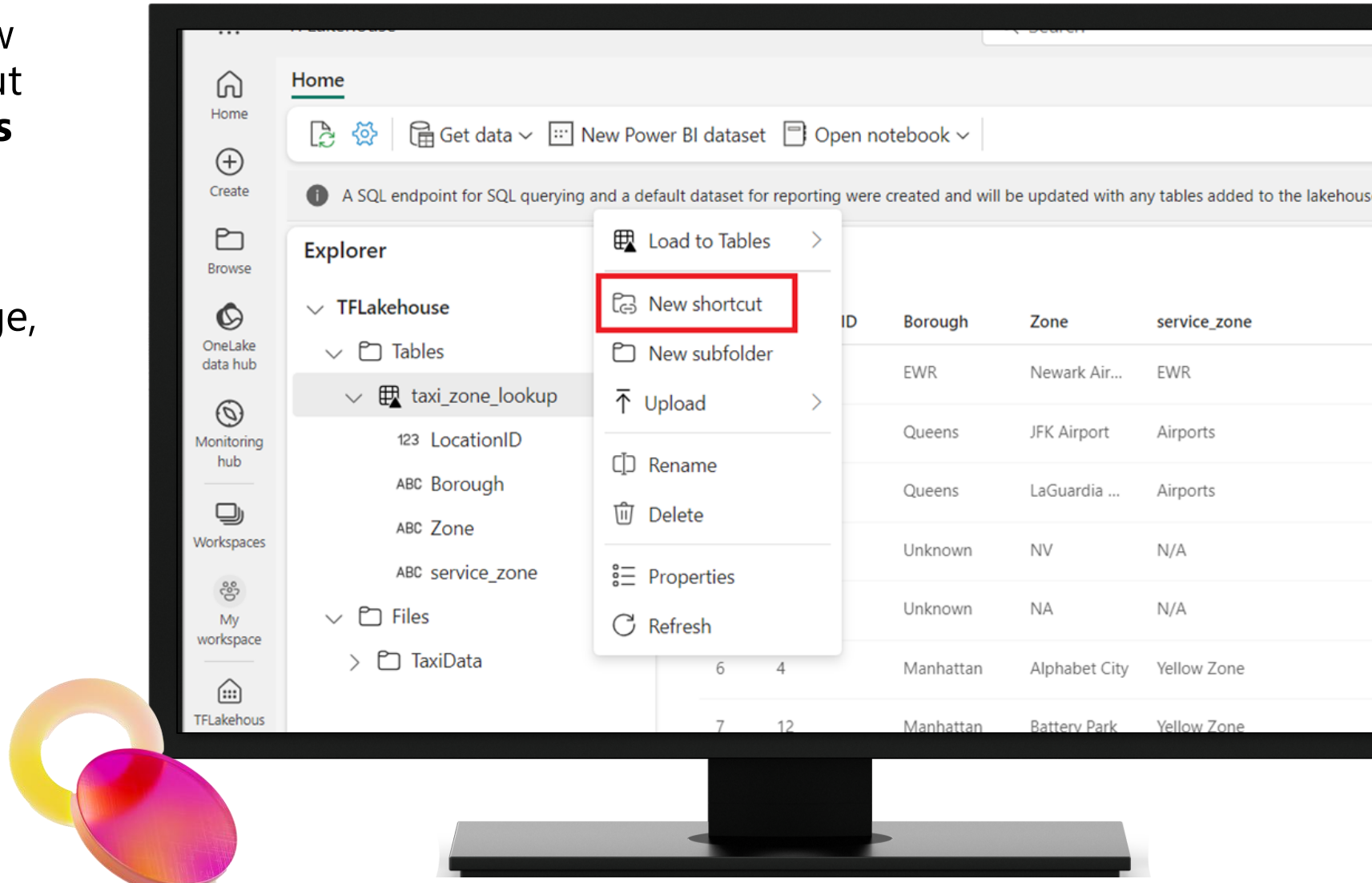


Shortcuts in Lakehouse

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

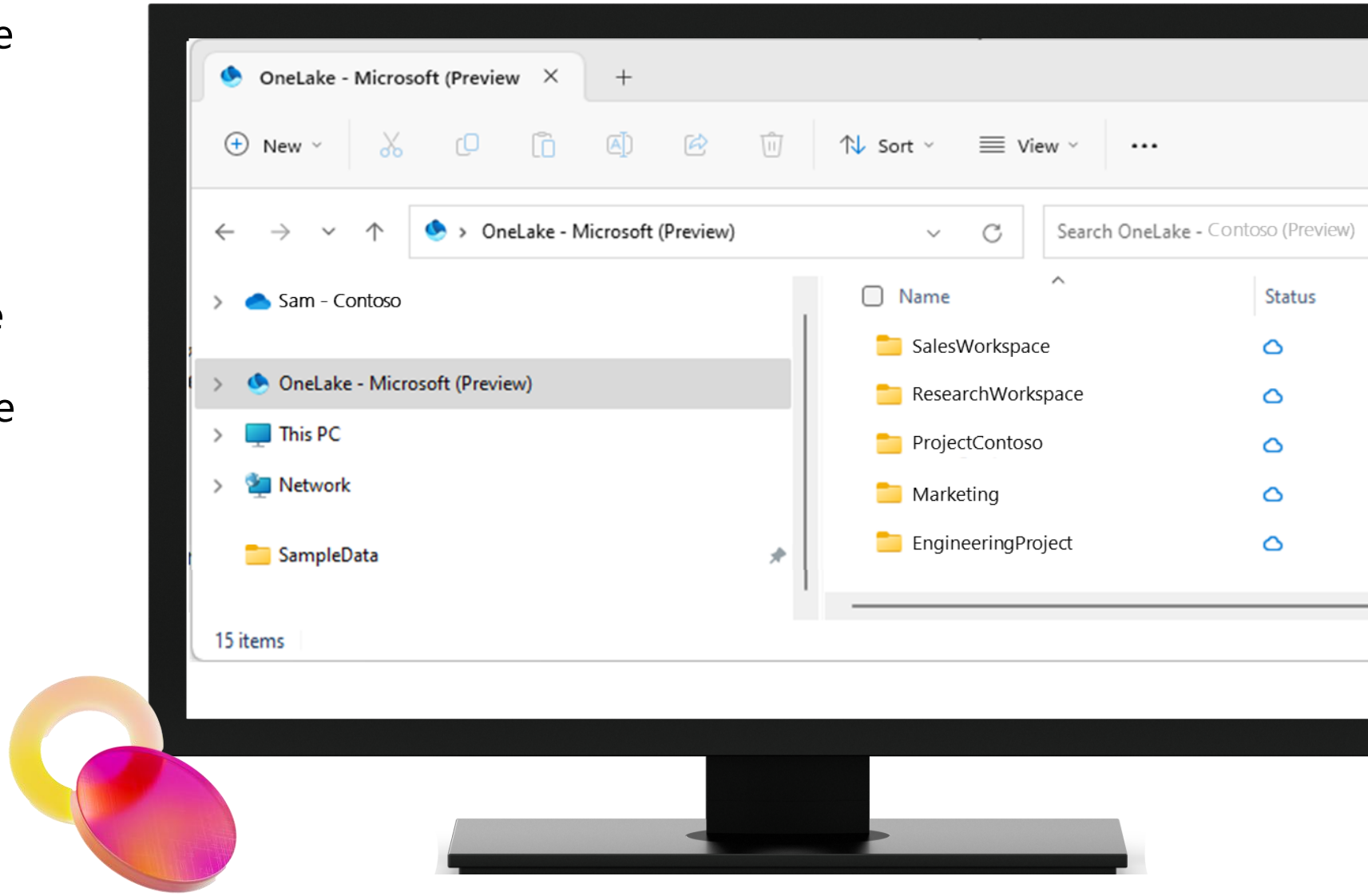
- ADLS Gen2
- AWS S3
- Dataverse



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

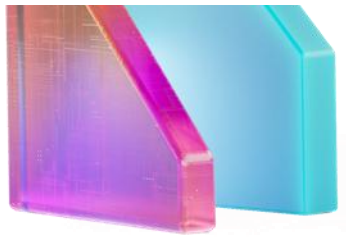
Uploading files from your computer

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	Recommendation
Small file upload from local machine	Use OneLake file explorer / Upload file
Data ingestion only	Use Copy job or Data pipeline
Orchestrated data ingestion and transformation	Use Data pipeline or Dataflow Gen2
Complex data transformations and orchestration	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]

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

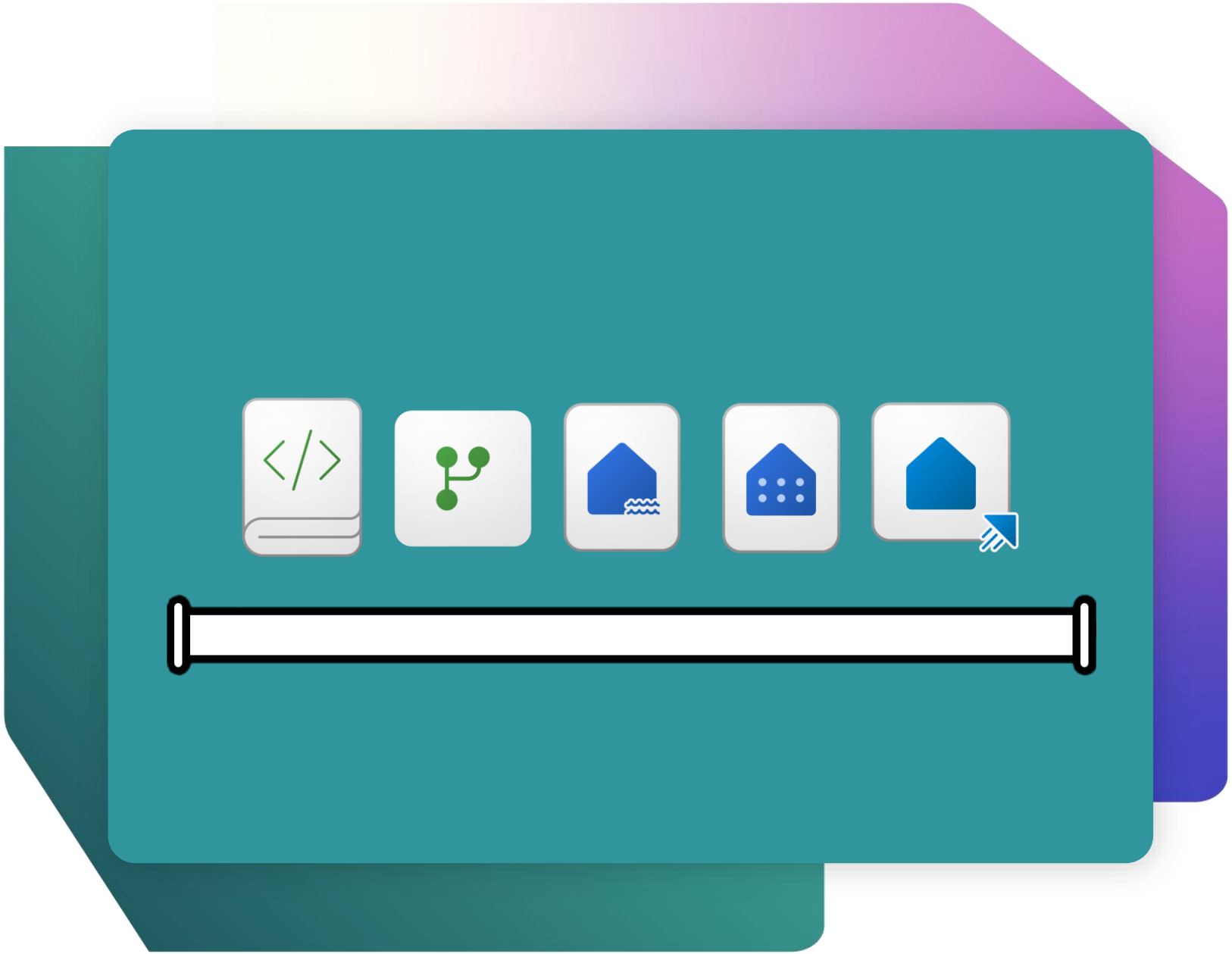
Collect, process, and transform data from its raw form to a format that you can use for analysis and decision-making

Ingesting and orchestrating data with Data pipelines

Data pipelines

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

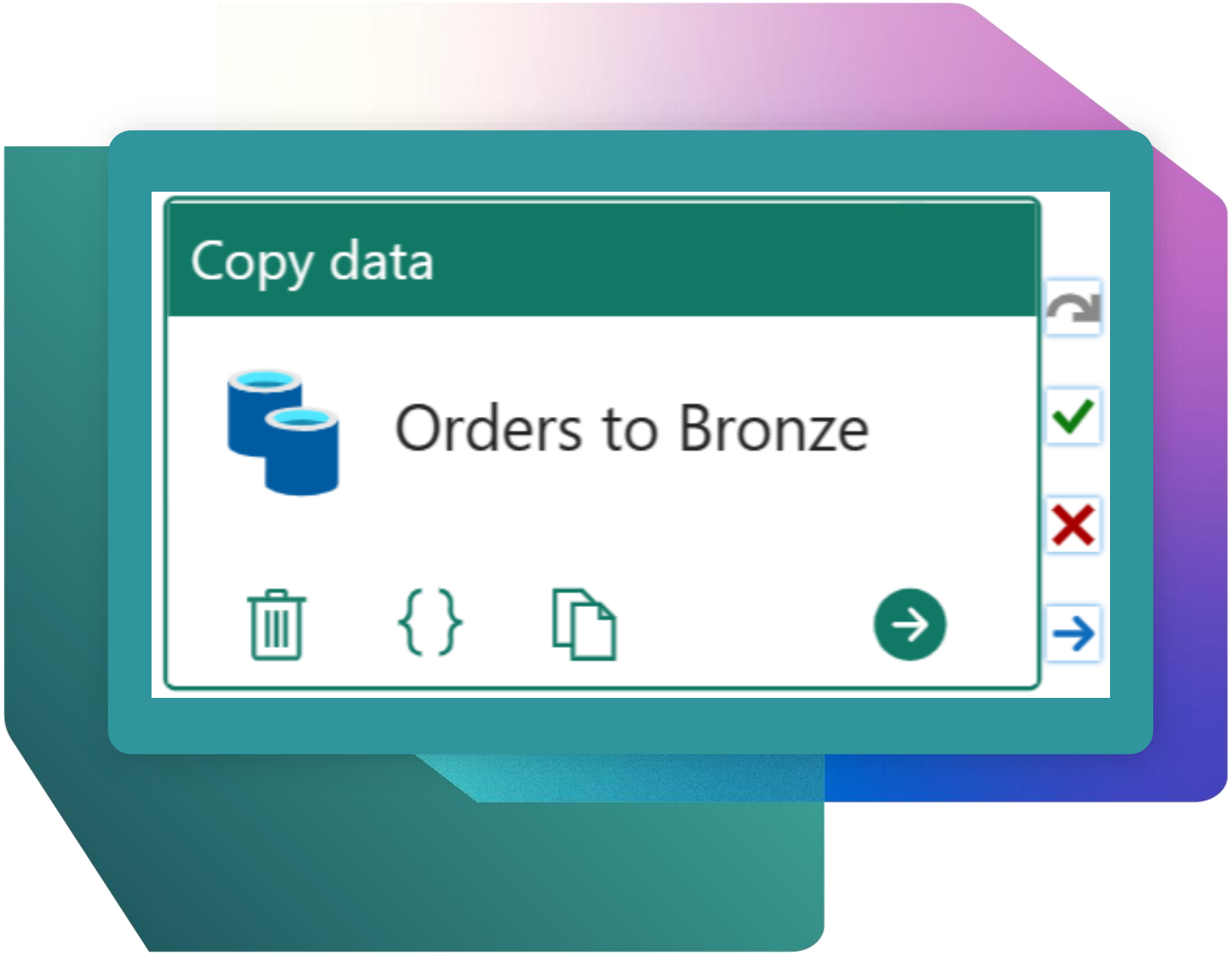
- Evolution of **Azure Data Factory** pipelines
- Rich library of **activities**
- Fast copy of binary files using Az-Copy
- Jumpstart with **Copy Assistant**



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



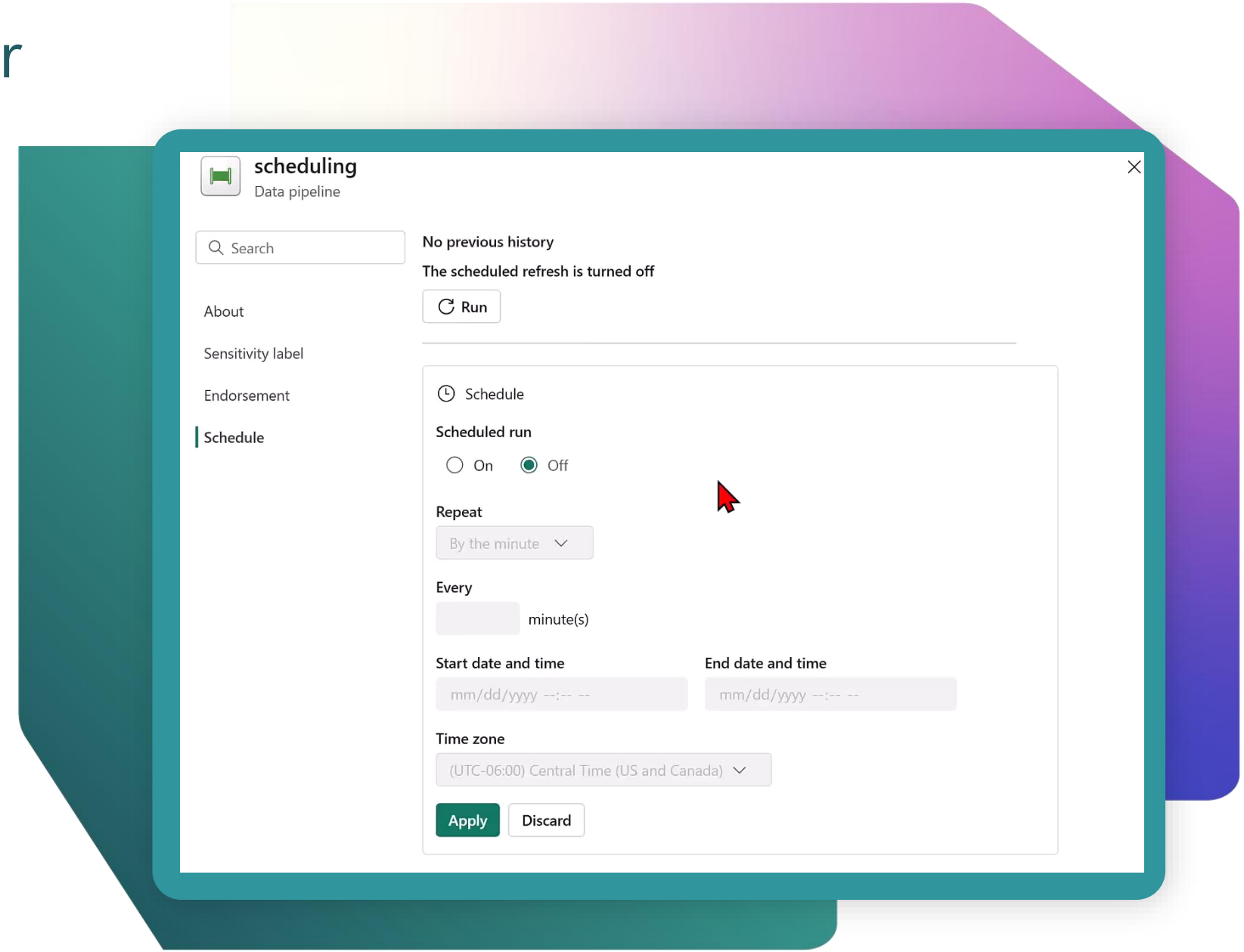
And so much more…

Capability	Description
Control flow	• Conditional paths • Loops
Nesting	• Invoking data pipelines from other data pipelines • Bi-directional communication via output variables
Parameterization	• Parameters of varying data types
Expression language	• Almost every field can be made dynamic! • Extensive built-in functions allow for massive customizations (data driven designs)
Repeatable	• Schedulable

Platform scheduler

Schedule events at desired intervals

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

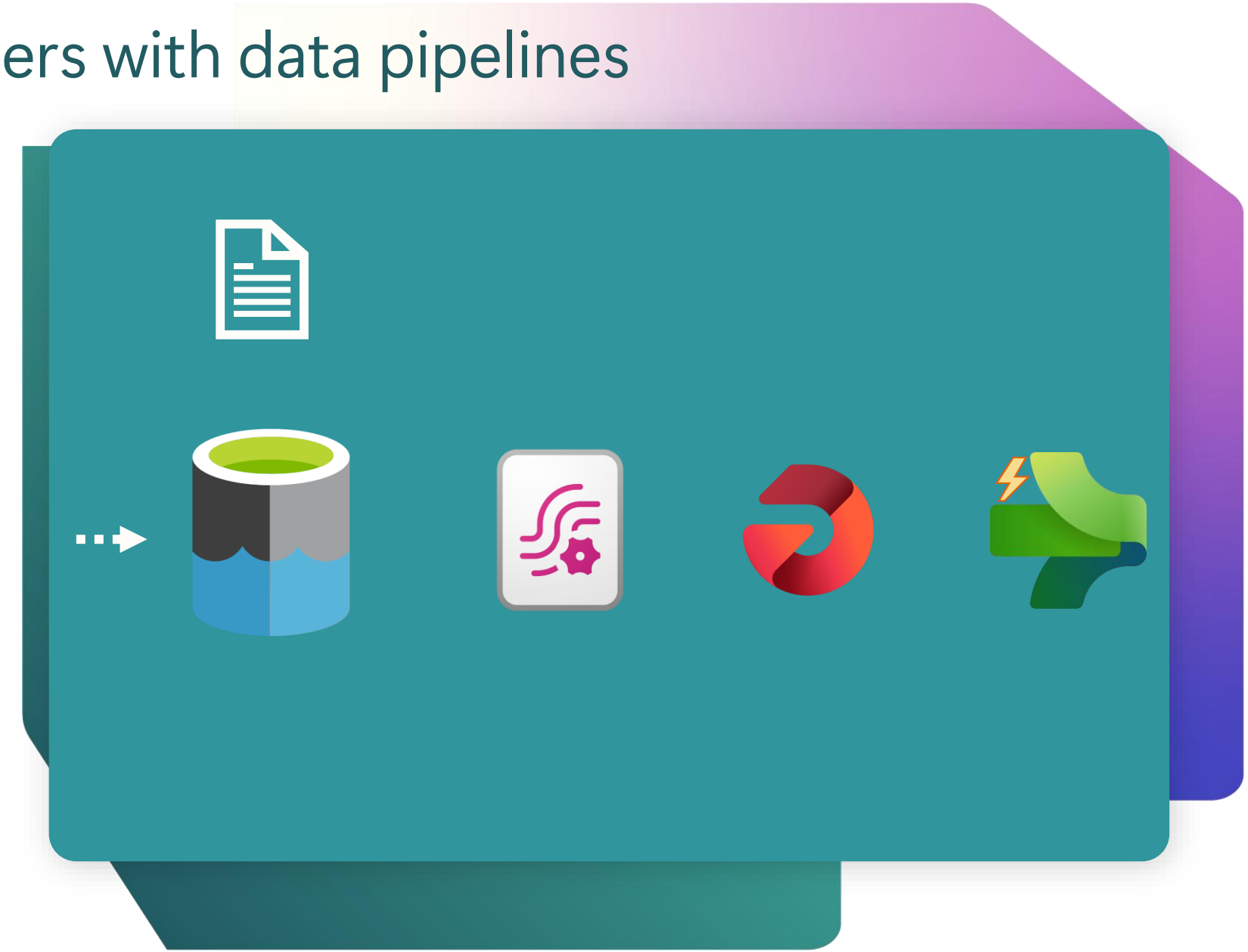


Event-driven triggers with data pipelines

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

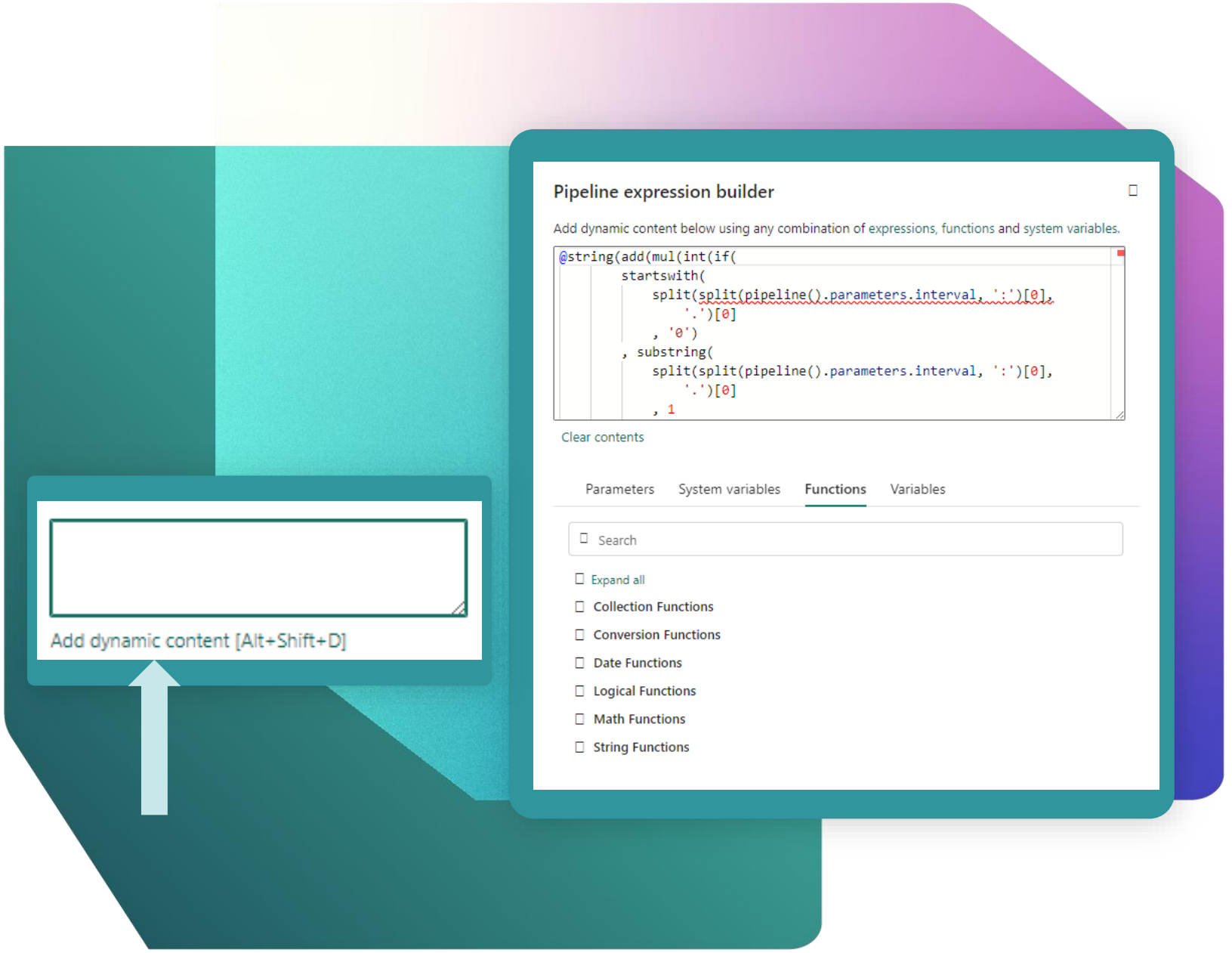


Expression builder

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

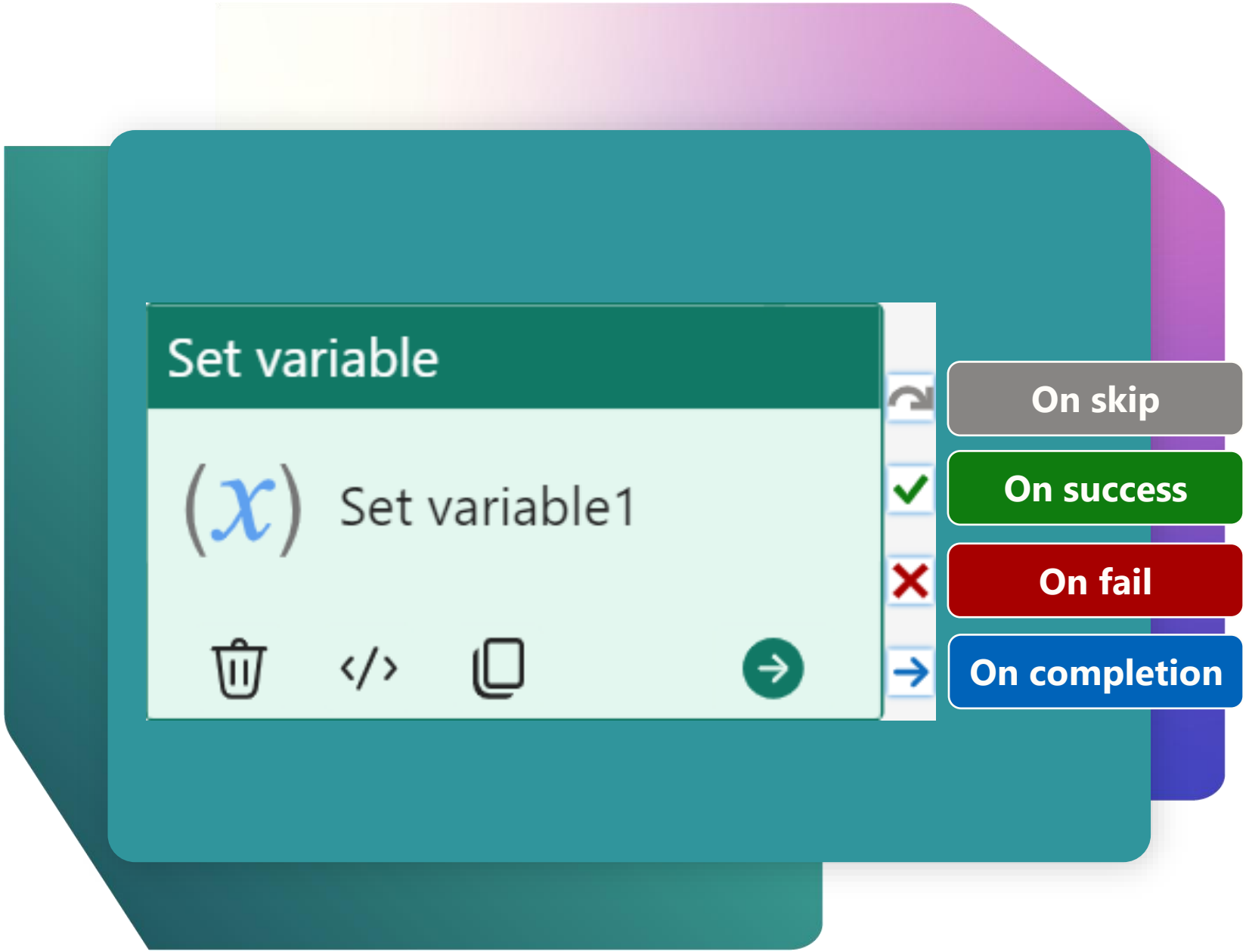
Greater flexibility and reusability, and to perform advanced data transformations without the need for custom code



Conditional paths

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

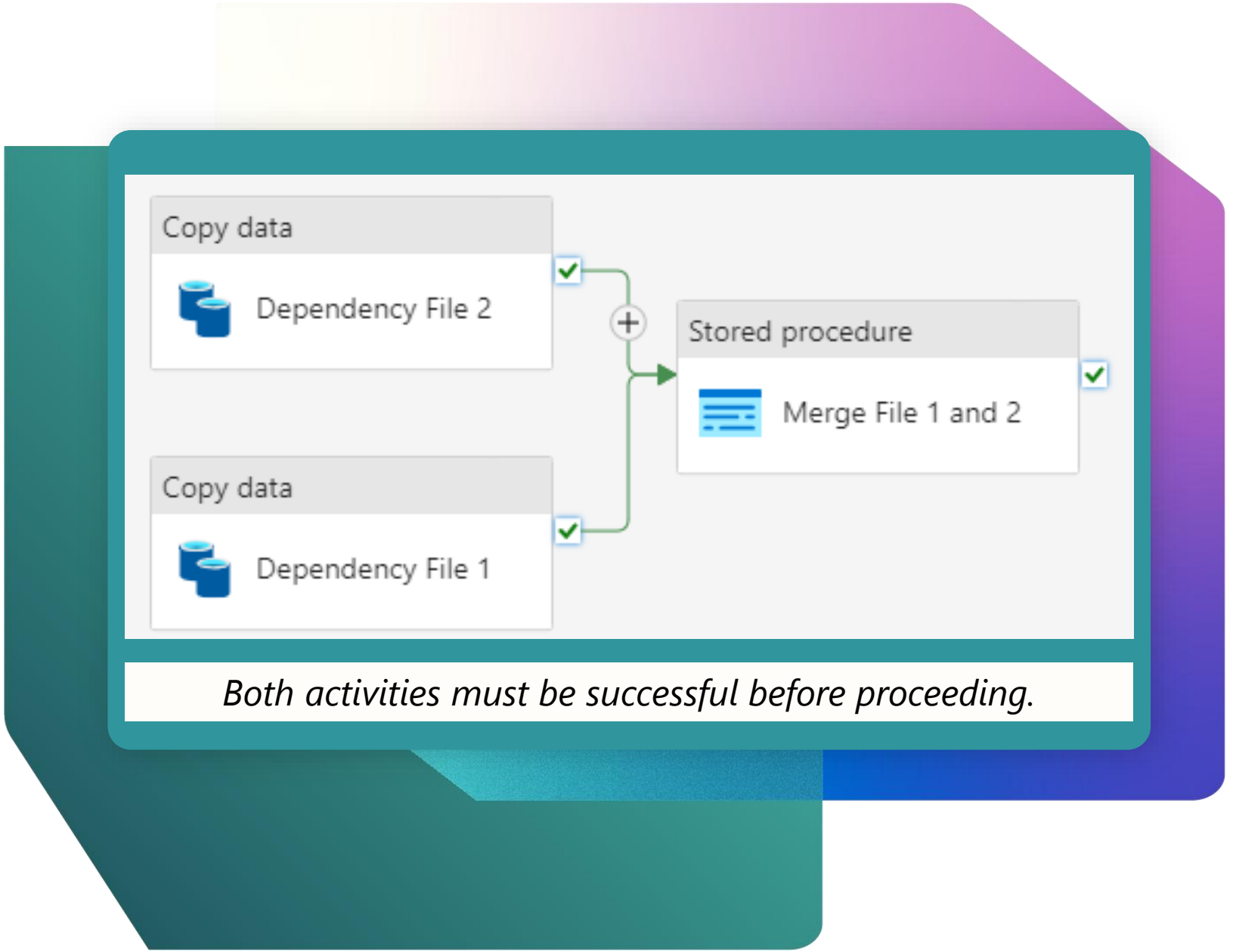
There are four types:
- On Skip
- On success
- On fail
- On completion



Conditional paths

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

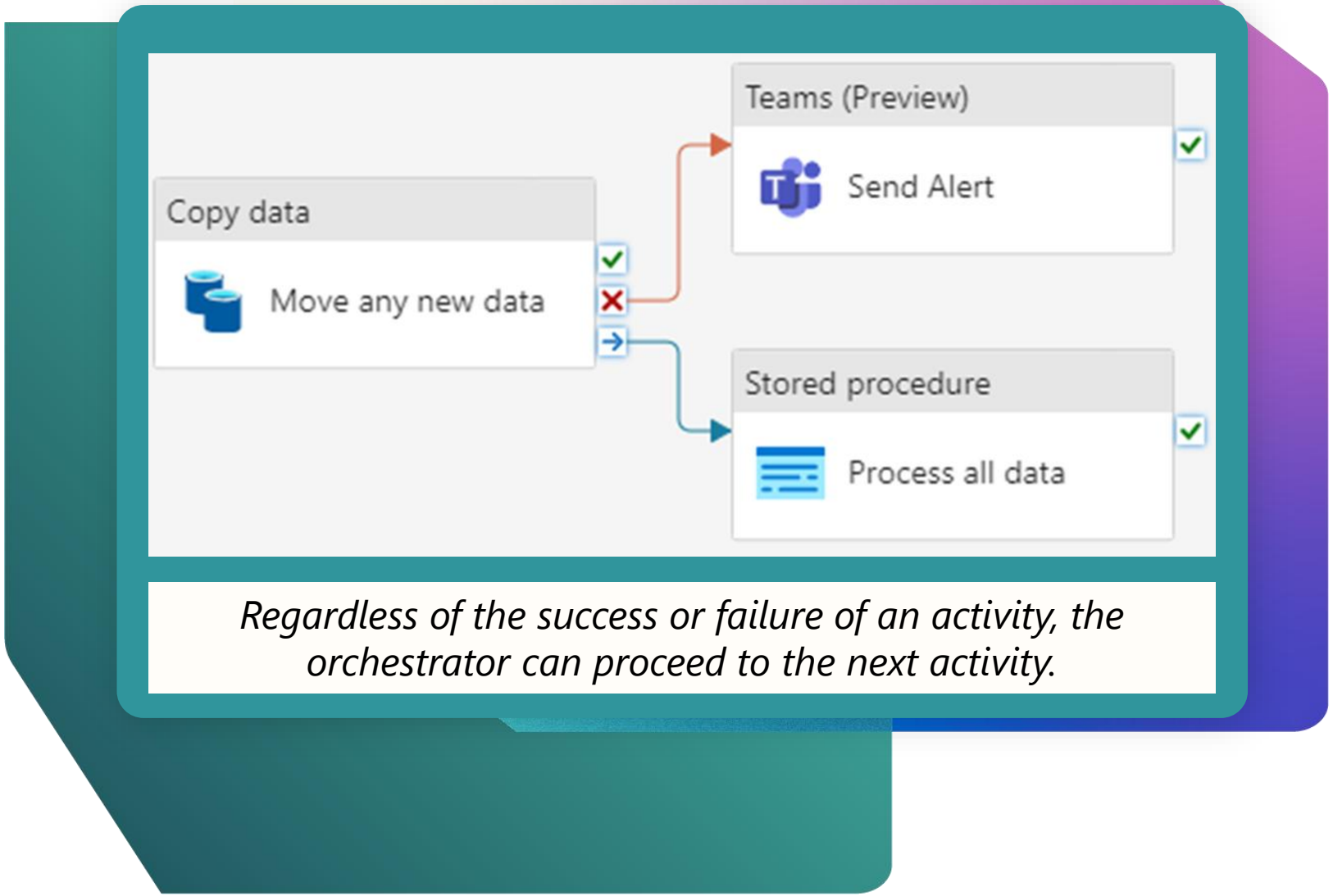
- **Blocking dependencies**
- Non-blocking dependencies
- Error handling



Conditional paths

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

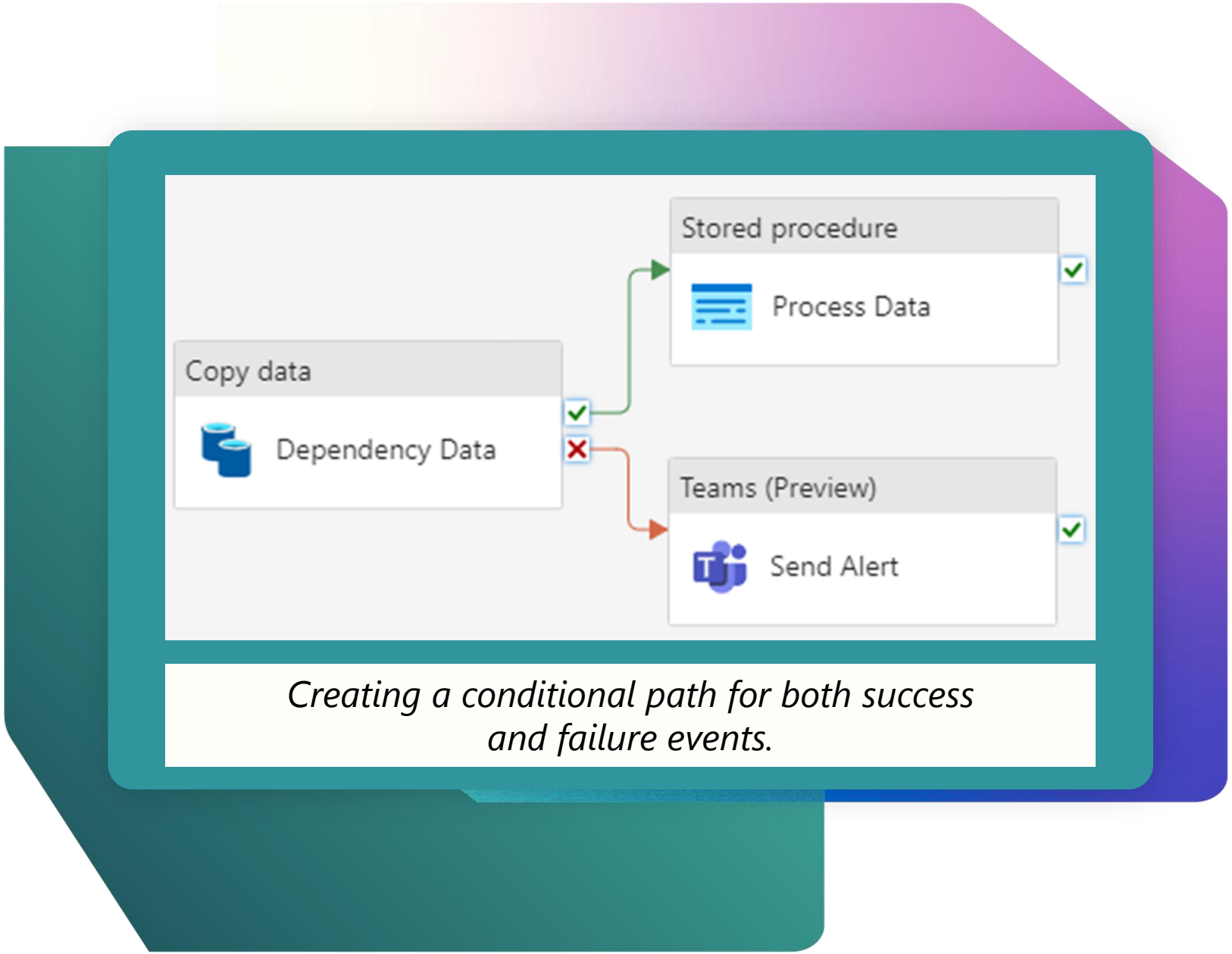
- Blocking dependencies
- **Non-blocking dependencies**
- Error handling



Conditional paths

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

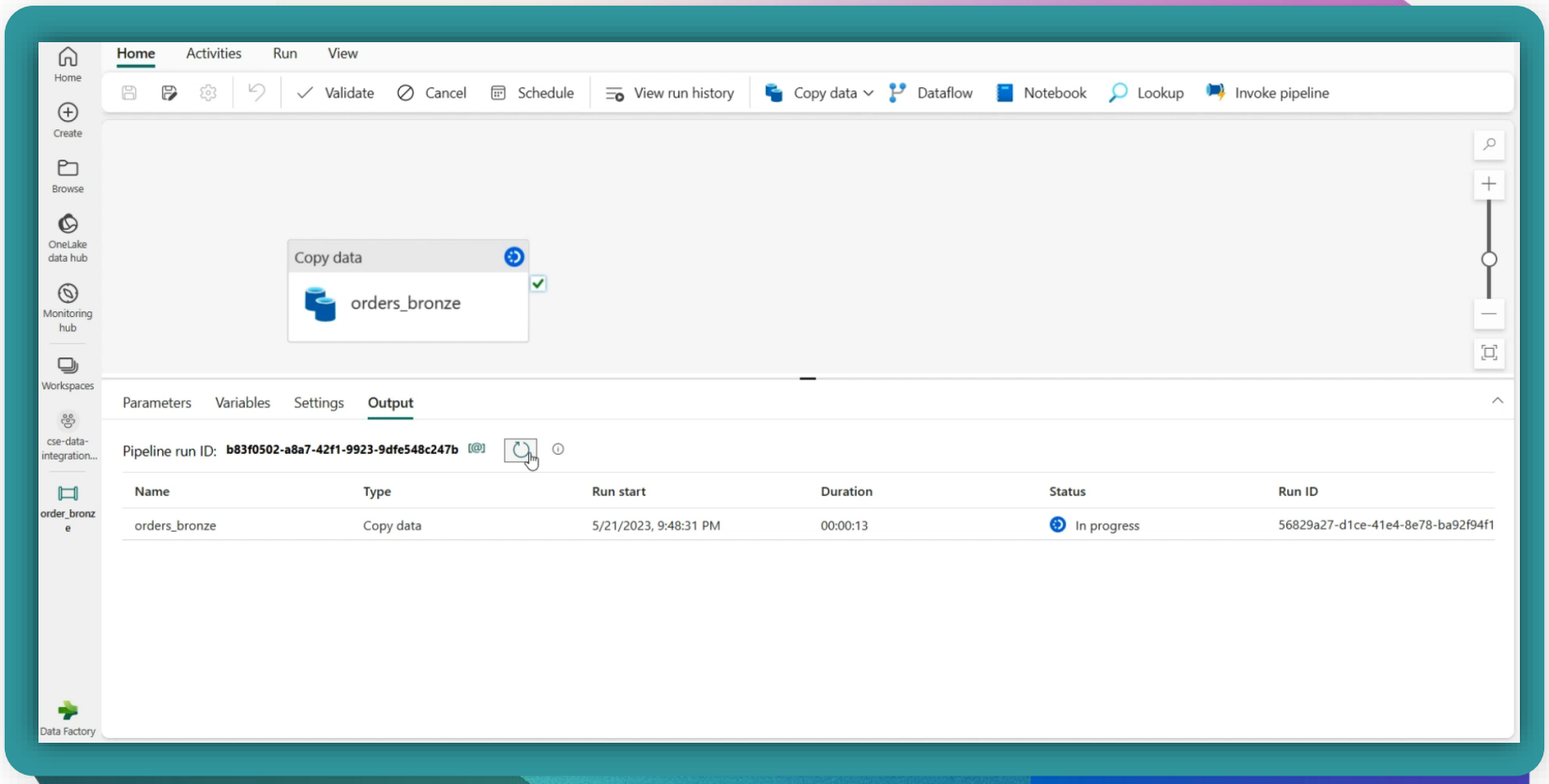
- Blocking dependencies
- Non-blocking dependencies
- **Error handling**



Creating a conditional path for both success and failure events.

Troubleshooting tips

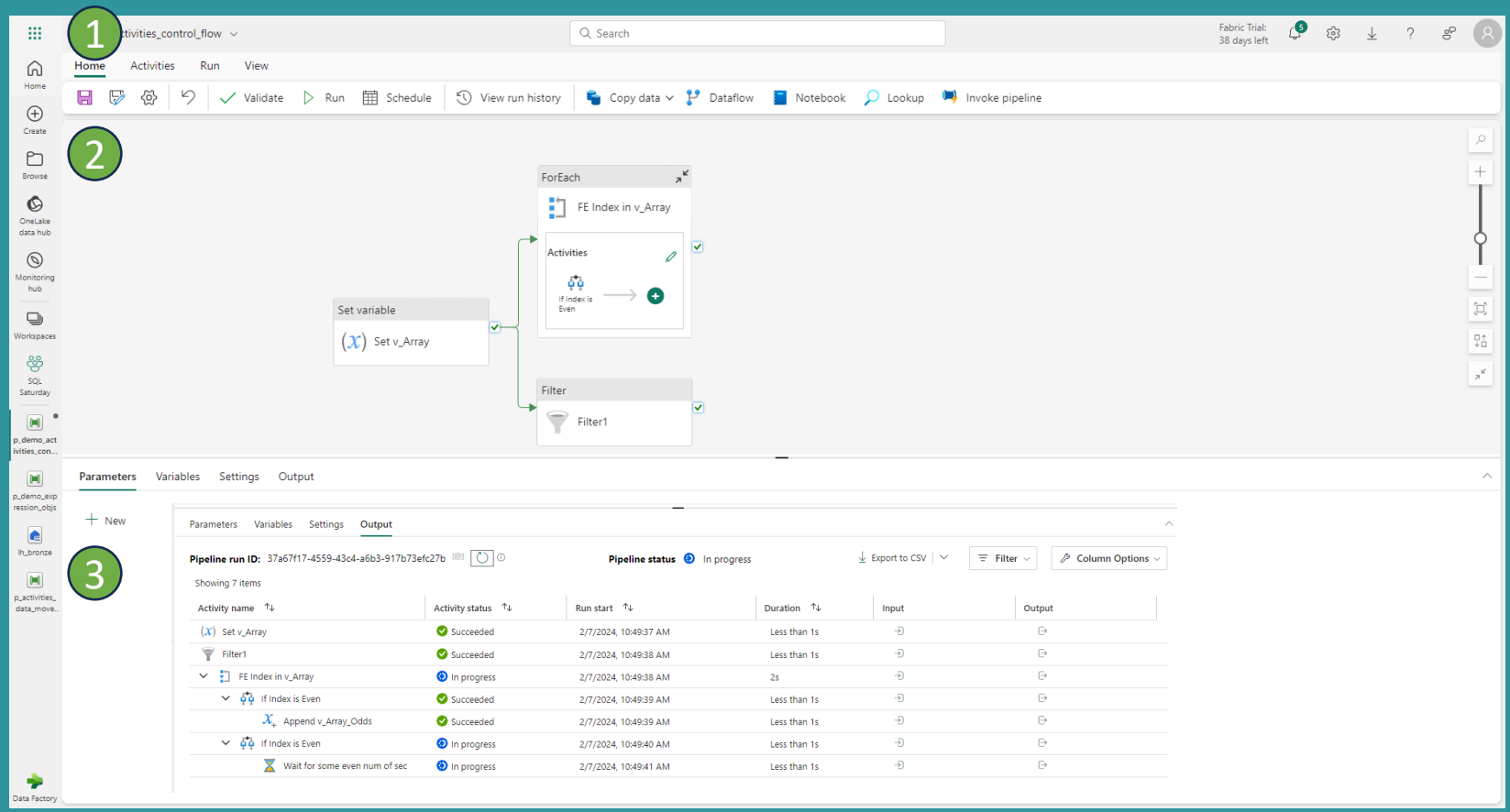
- Leverage the Output window to find and review errors
- Determine if it is an activity error or an error from the source/destination
- Set a retry attempt for activities



Exploring Data pipelines

Navigation

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



DEMO

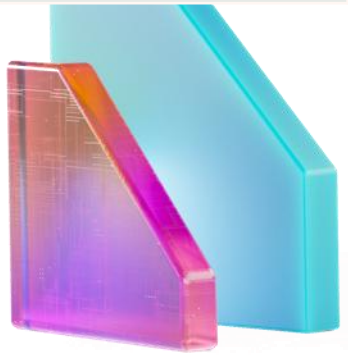
Creating a…

- 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

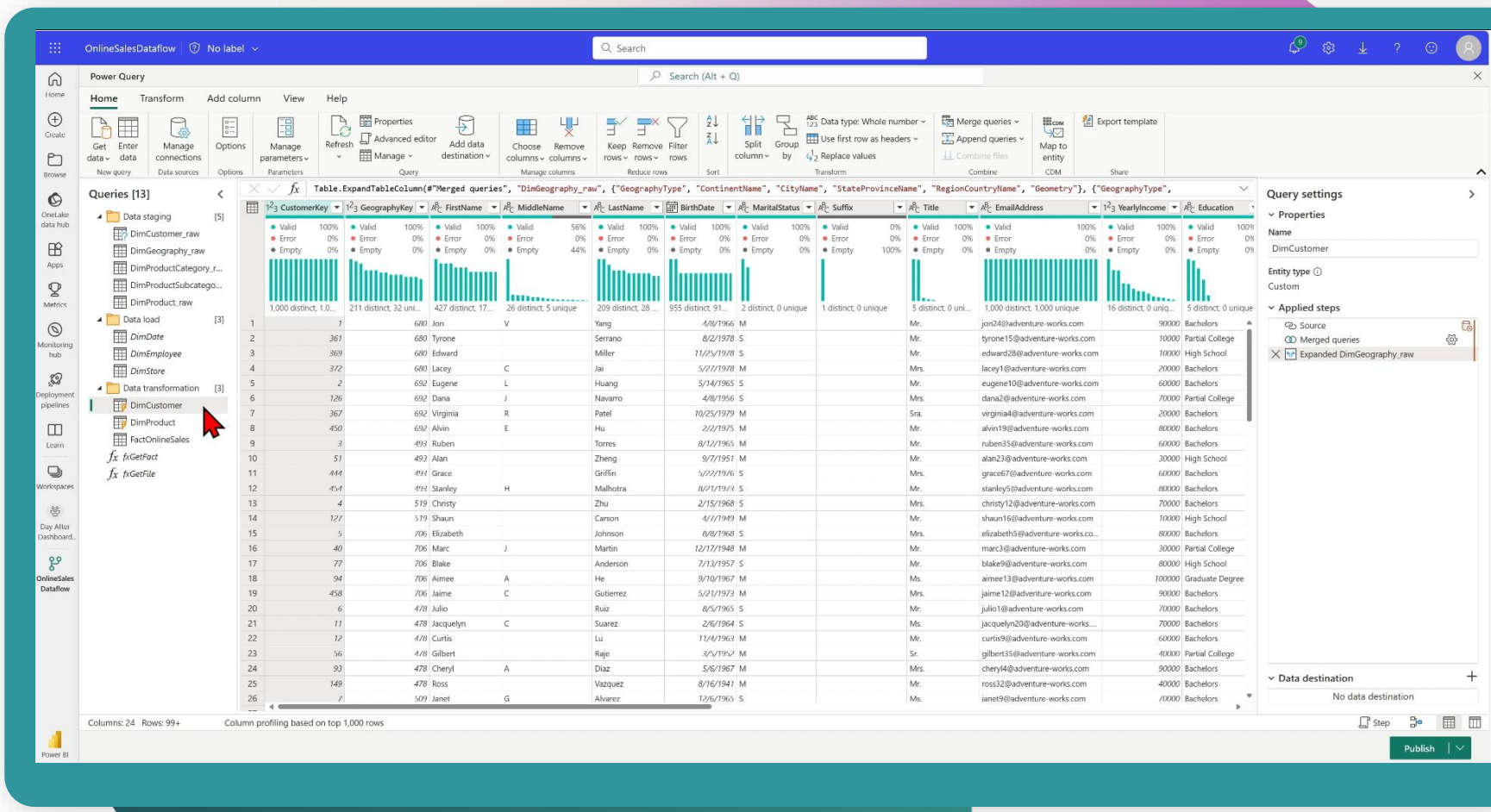


Ingesting and transforming data with Dataflow Gen2

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

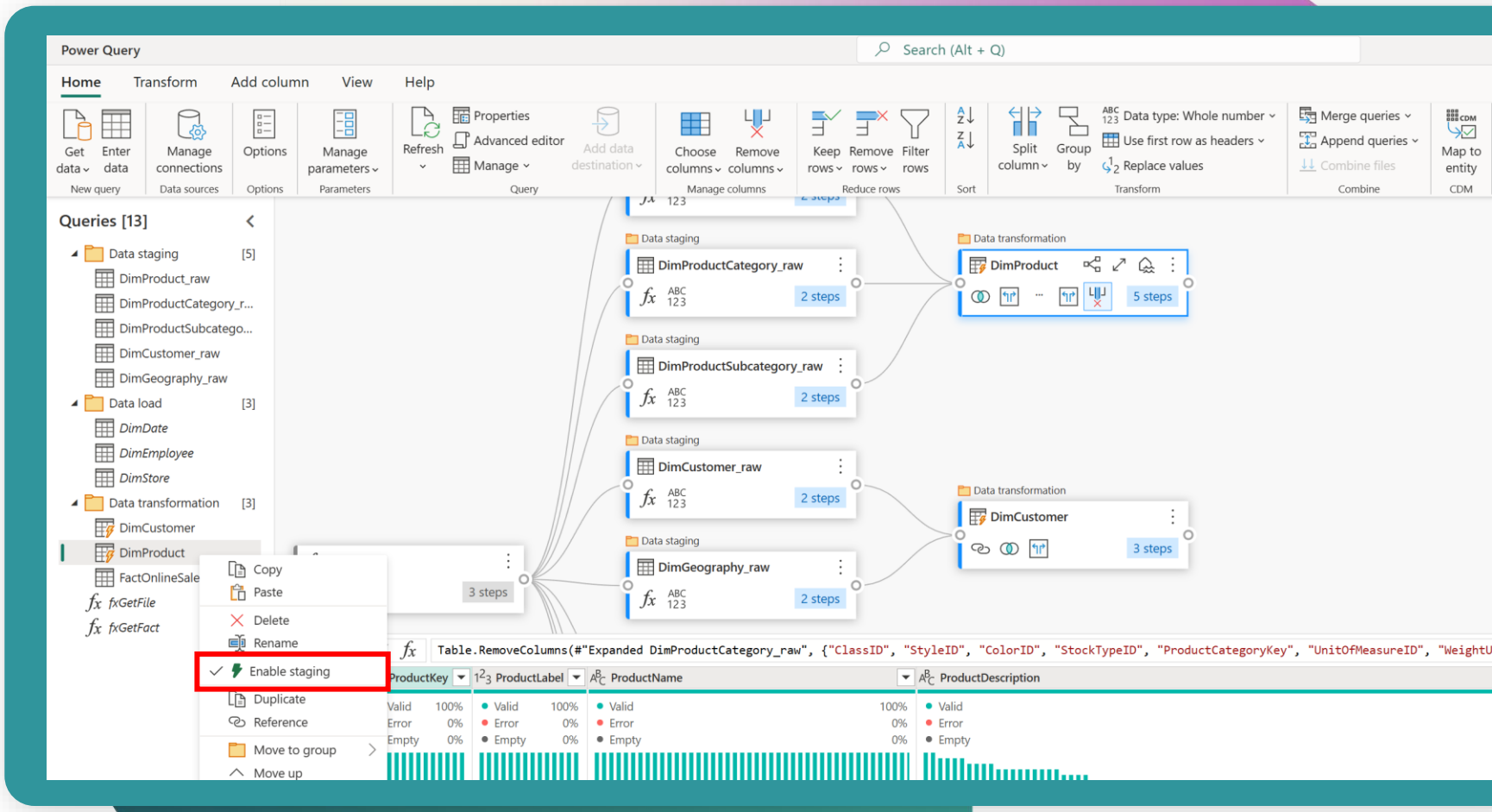


Dataflow Gen2 staging

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”

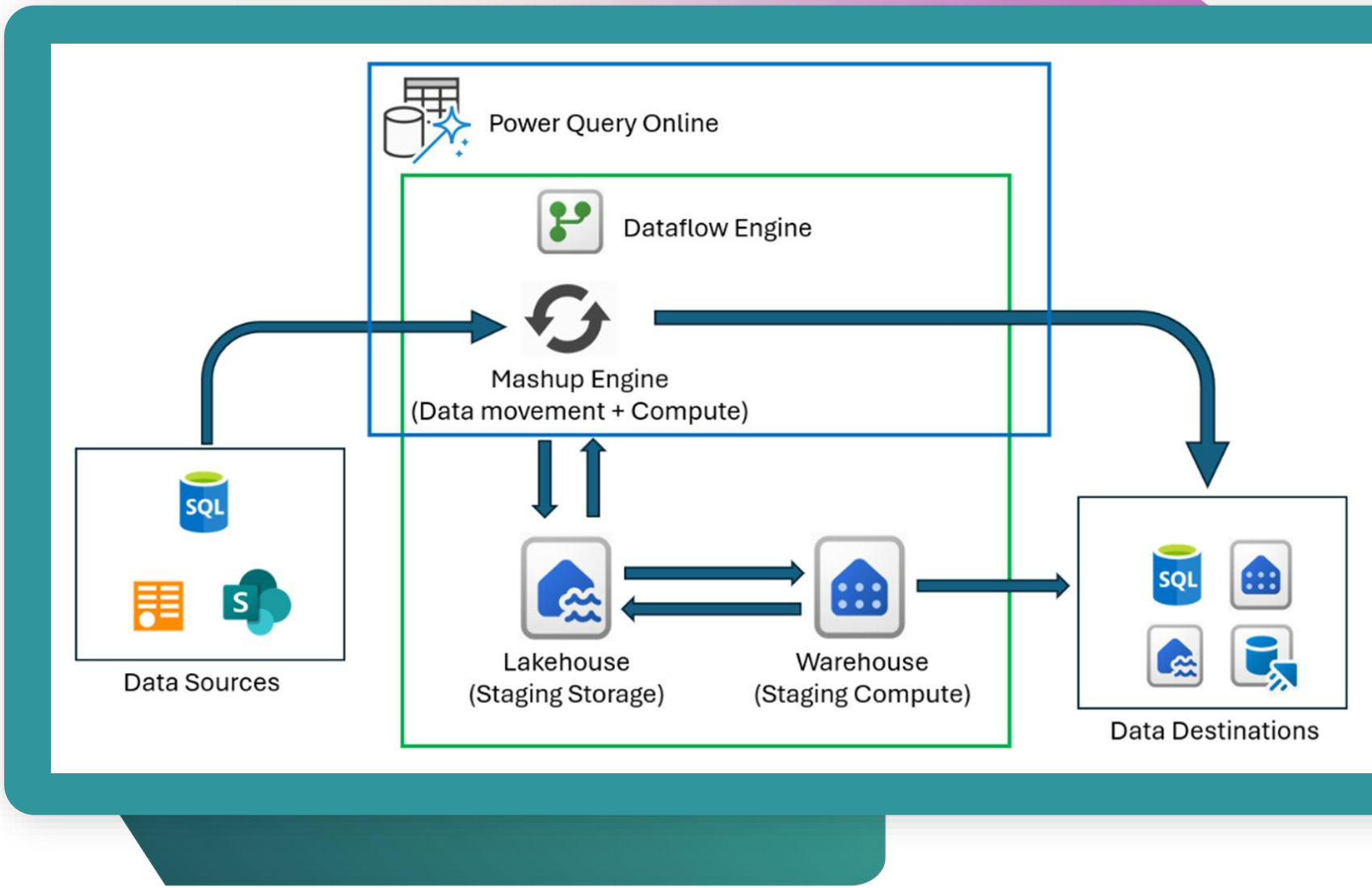
Dataflow Gen2 staging

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.

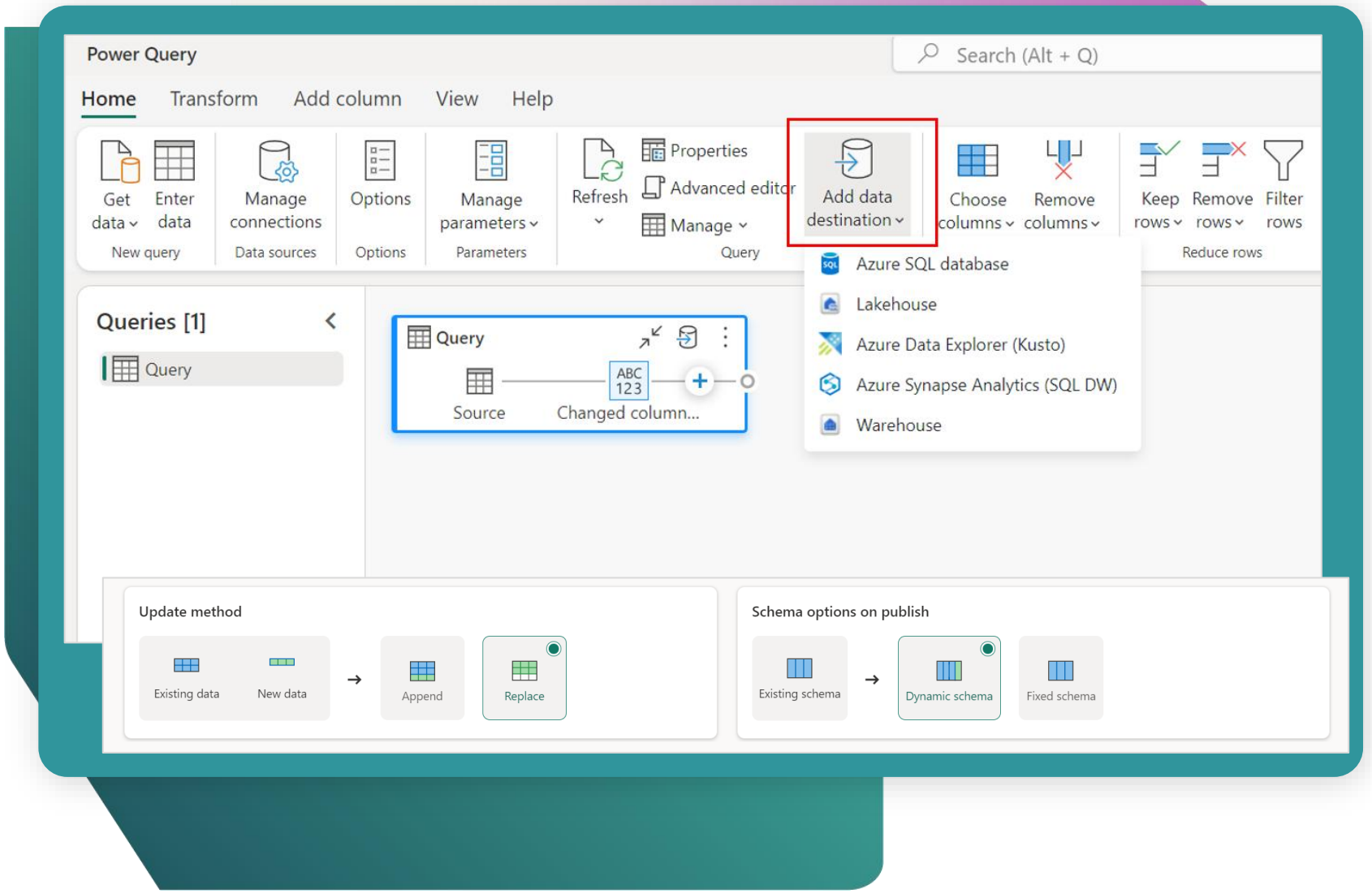
3. Apply transformation steps to the **computed** table for complex ETL operations such as join, distinct, filter and group by – leveraging Fabric compute.



Data destinations

Supported destinations:
- Lakehouse
- Warehouse
- Eventhouse
- Azure SQL database
- Azure Synapse Analytics

Update methods:
- Replace
- Append



MS Learn | Dataflow Gen2 data destinations and managed settings

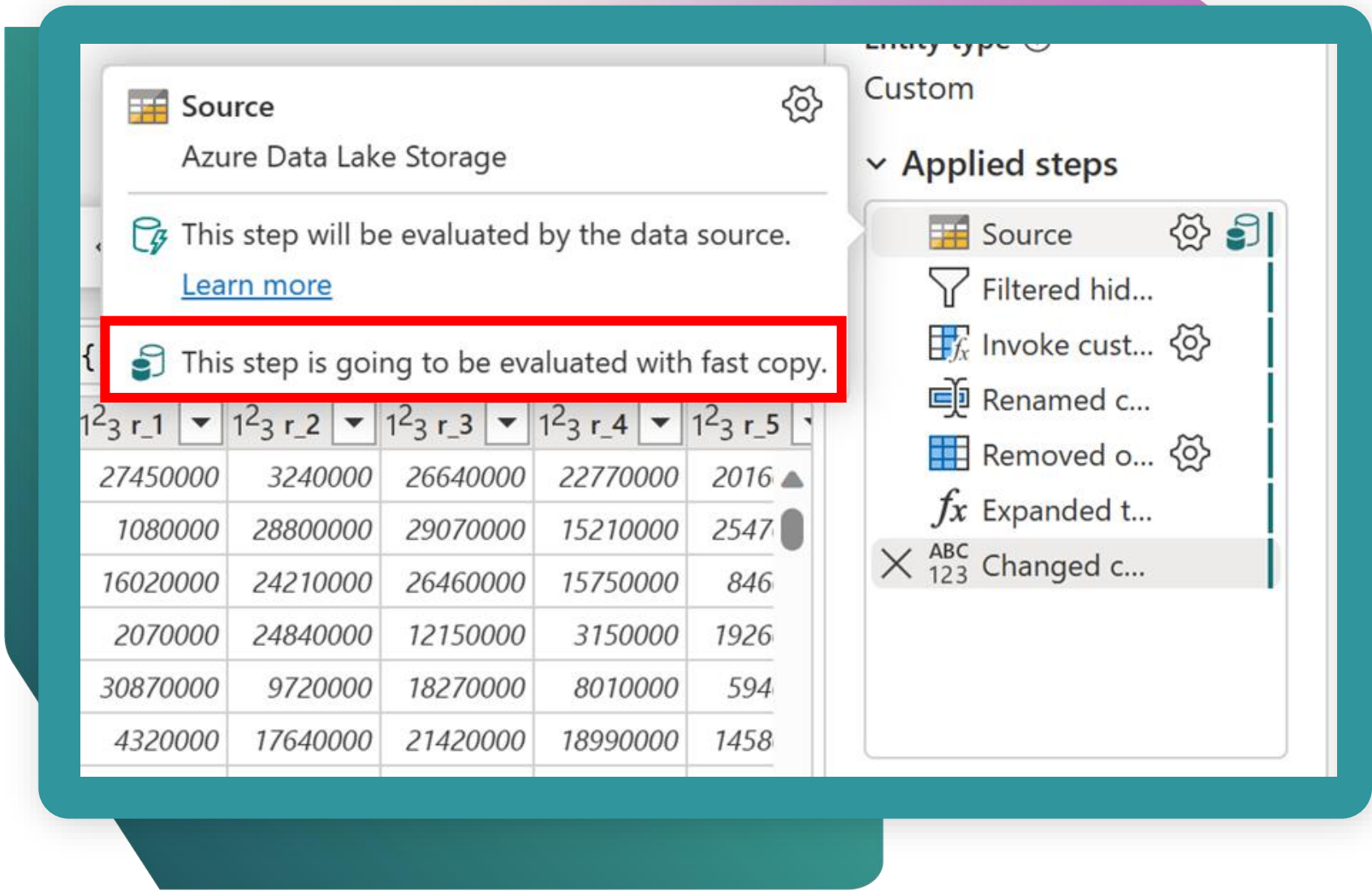
Fast Copy

General Availability

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



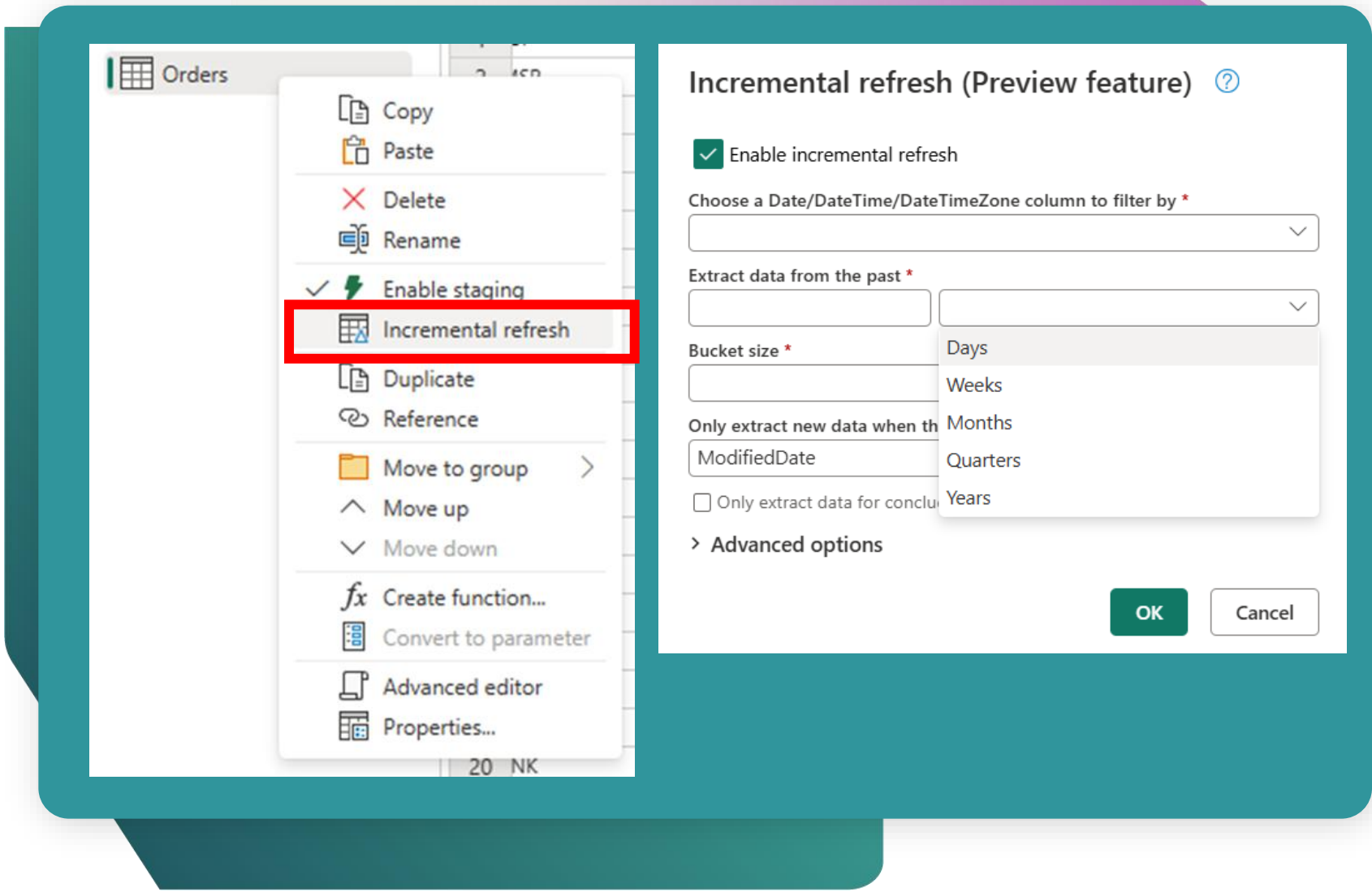
MS Learn | Fast Copy in Dataflows Gen2

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

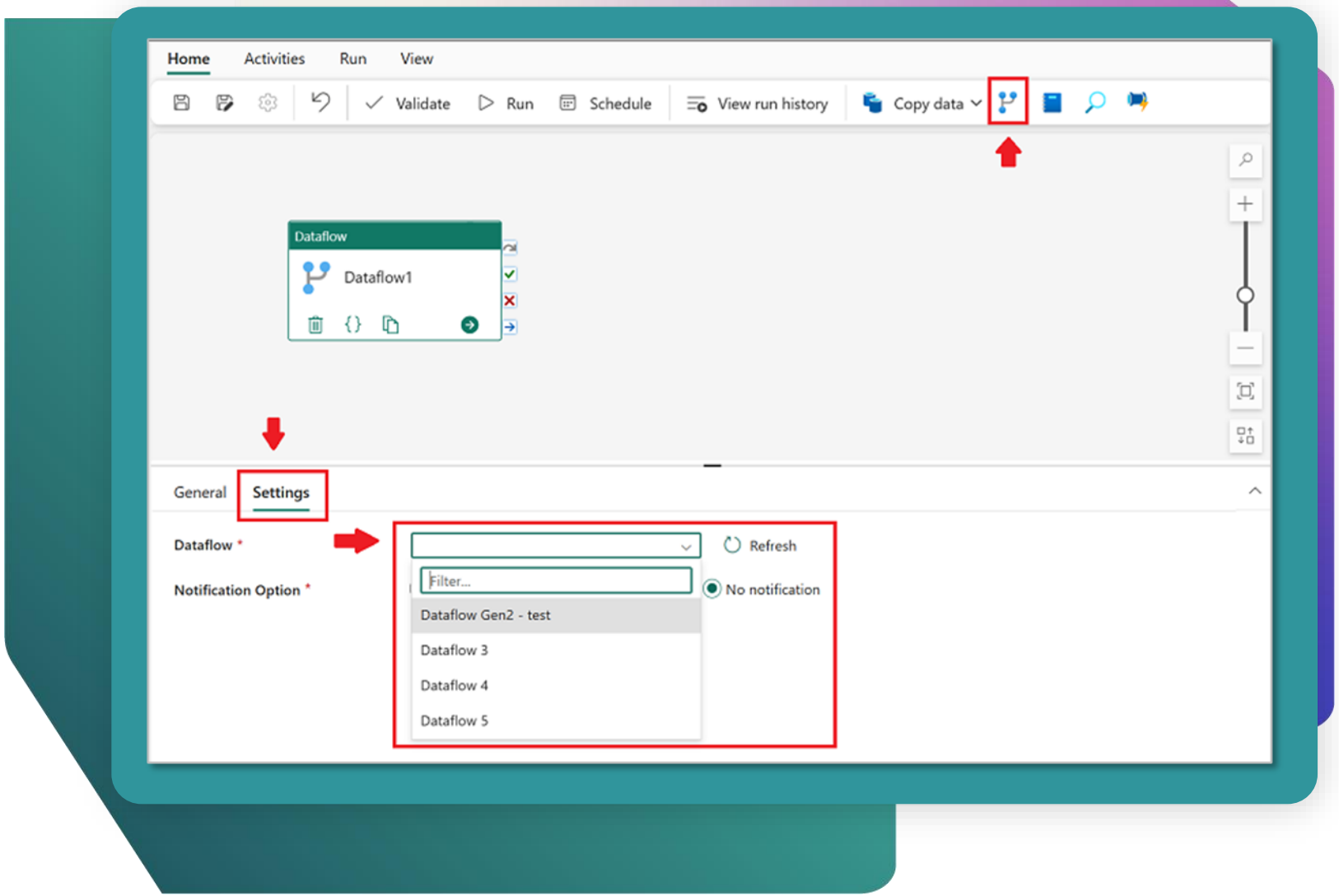
50

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

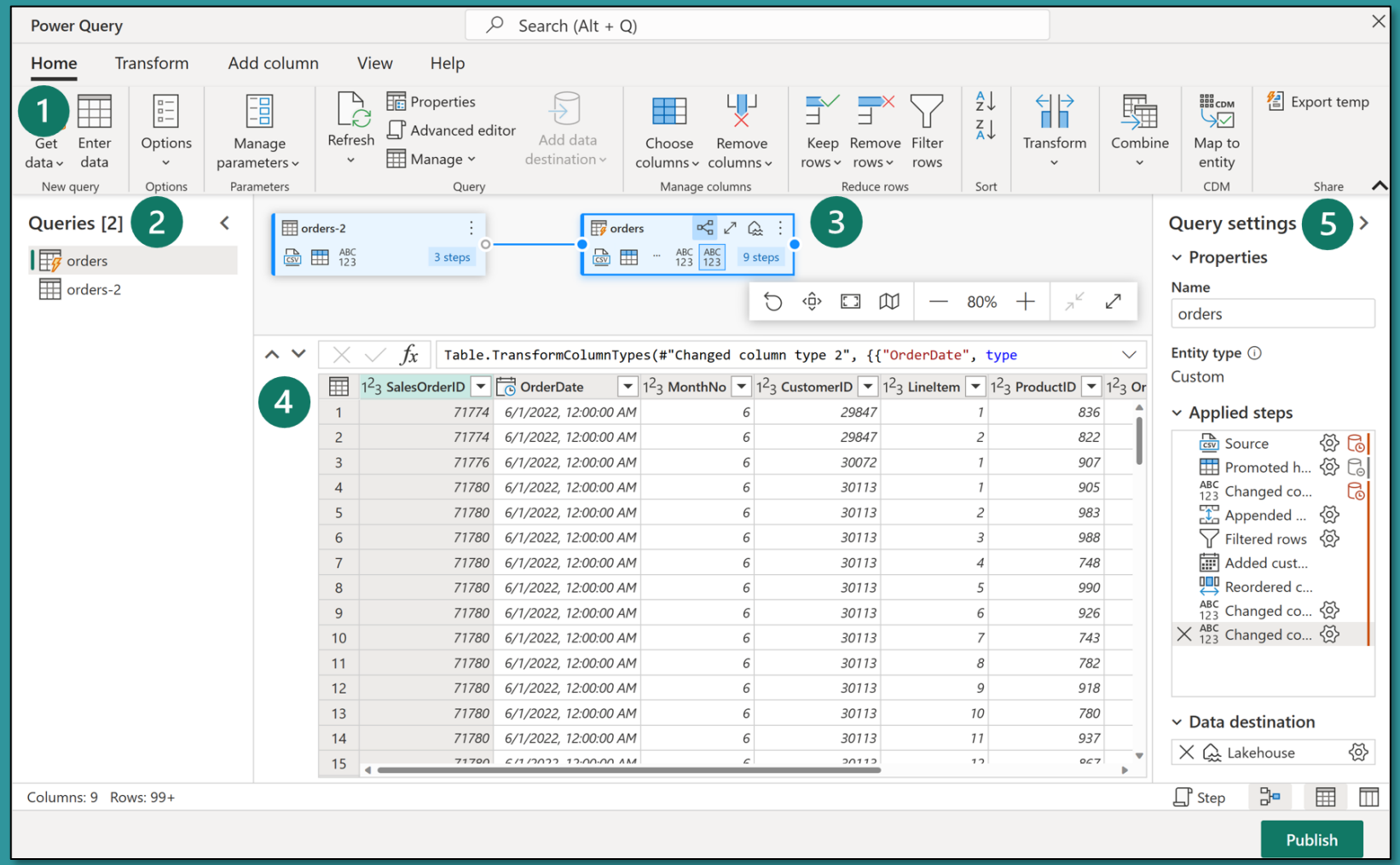
Separate tables based on the data source and if they **support query folding** or do **not support query folding**

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

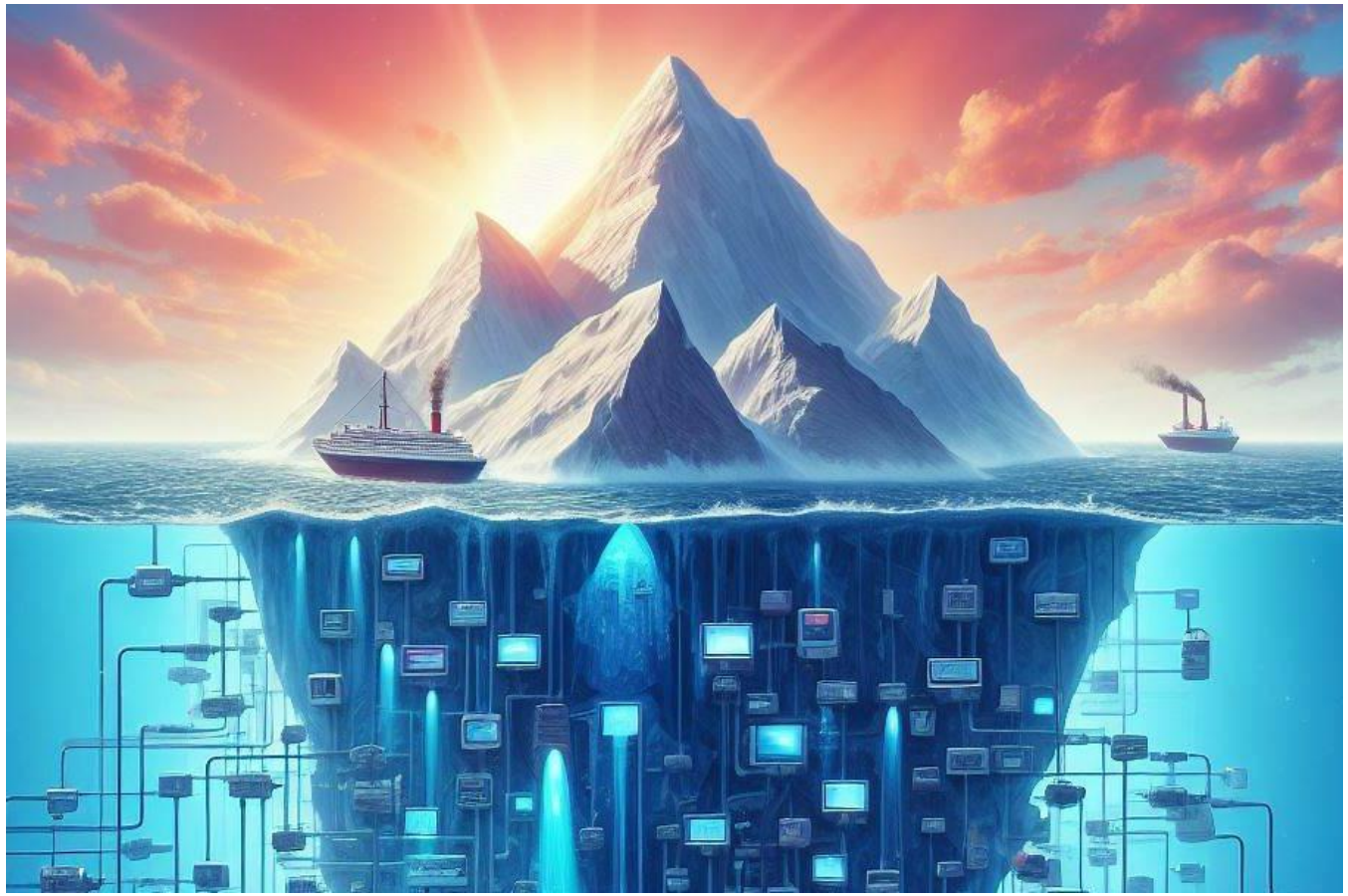
Migrate from Power BI dataflow to Fabric dataflow gen2

Feature	Power BI Dataflow	Dataflow Gen2
Incremental refresh	Yes	**YES!**
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...

Microsoft

Thank you!

Microsoft Fabric

60-Day Free Trial

No Credit Card

No Azure Subscription

F64 SKU

$17,000 value

aka.ms/try-fabric

Microsoft

Get Involved in the Fabric Community

aka.ms/FabricCommunity
Connect with community members, ask questions, and learn more about Fabric

aka.ms/FabricUserGroups
Find a user group that matches your interests in your area or online

aka.ms/SuperUsers
Spread your Fabric knowledge, insights, and best practices with others

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

Please rate this session on the app

cvent

GET IT ON Google Play

Download on the App Store

EUROPEAN MICROSOFT FABRIC Community Conference

Microsoft | Microsoft Fabric

Learn more about Microsoft Fabric

Unify your data to unlock AI innovation