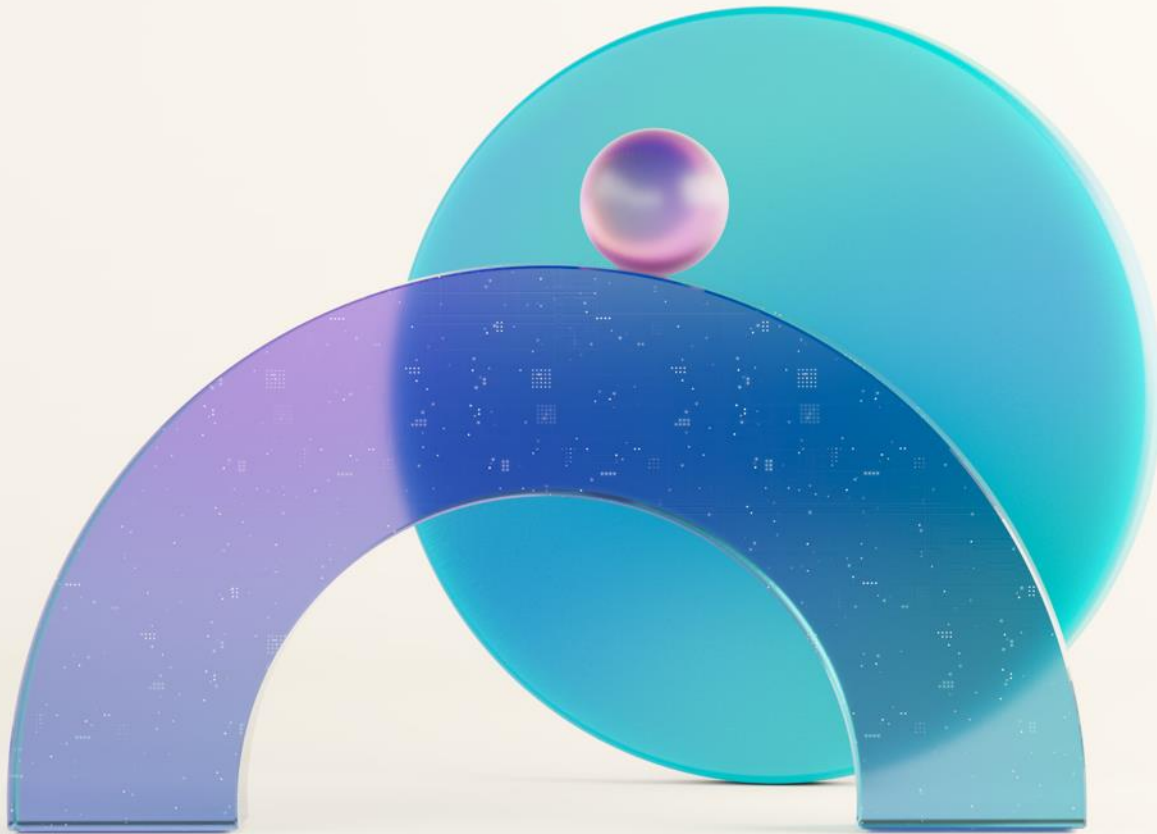


The background features abstract, overlapping geometric shapes in shades of pink, teal, and light blue. A small, reflective sphere sits on the edge of one of the teal blocks.

Microsoft Fabric

COMMUNITY CONFERENCE



Factory in Microsoft Fabric Technical Deep Dive

March 26-28, 2024

Mohan Sankaran

John Welch

Erwin de Kreuk

Mohan Sankaran

- Partner Director of Engineering
 - Data Integration - ADF/Fabric Pipeline
- Microsoft
- LinkedIn: [linkedin.com/in/mohansankaran](https://www.linkedin.com/in/mohansankaran)



John Welch

- Principal Architect
 - Data Integration – Citizen Data Integration
- Microsoft
- LinkedIn: [linkedin.com/in/johncwelch](https://www.linkedin.com/in/johncwelch)



Erwin de Kreuk

- Principal Consultant
- Lead Data & Analytics InSpark



Let's connect



 @erwindekrek
 [linkedin.com/in/erwindekrek](https://www.linkedin.com/in/erwindekrek)
 erwindekrek.com
 github.com/edkreuk
 <https://sessionize.com/erwin-de-kreuk/>



Objectives



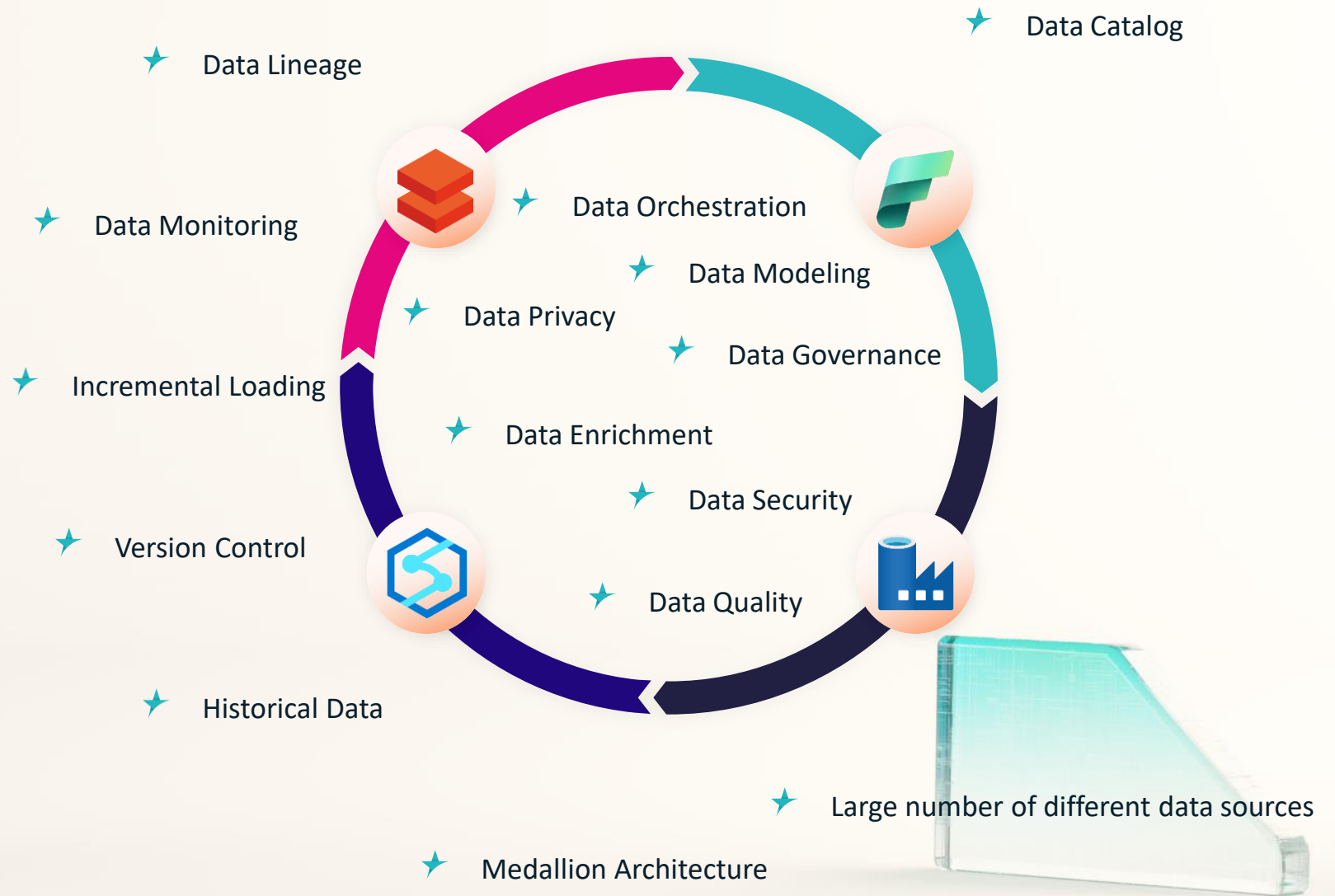
- Overview of approach to Medallion Architecture
- Pipeline Architecture
- Dataflows
- Dataflow Architecture
- Recap

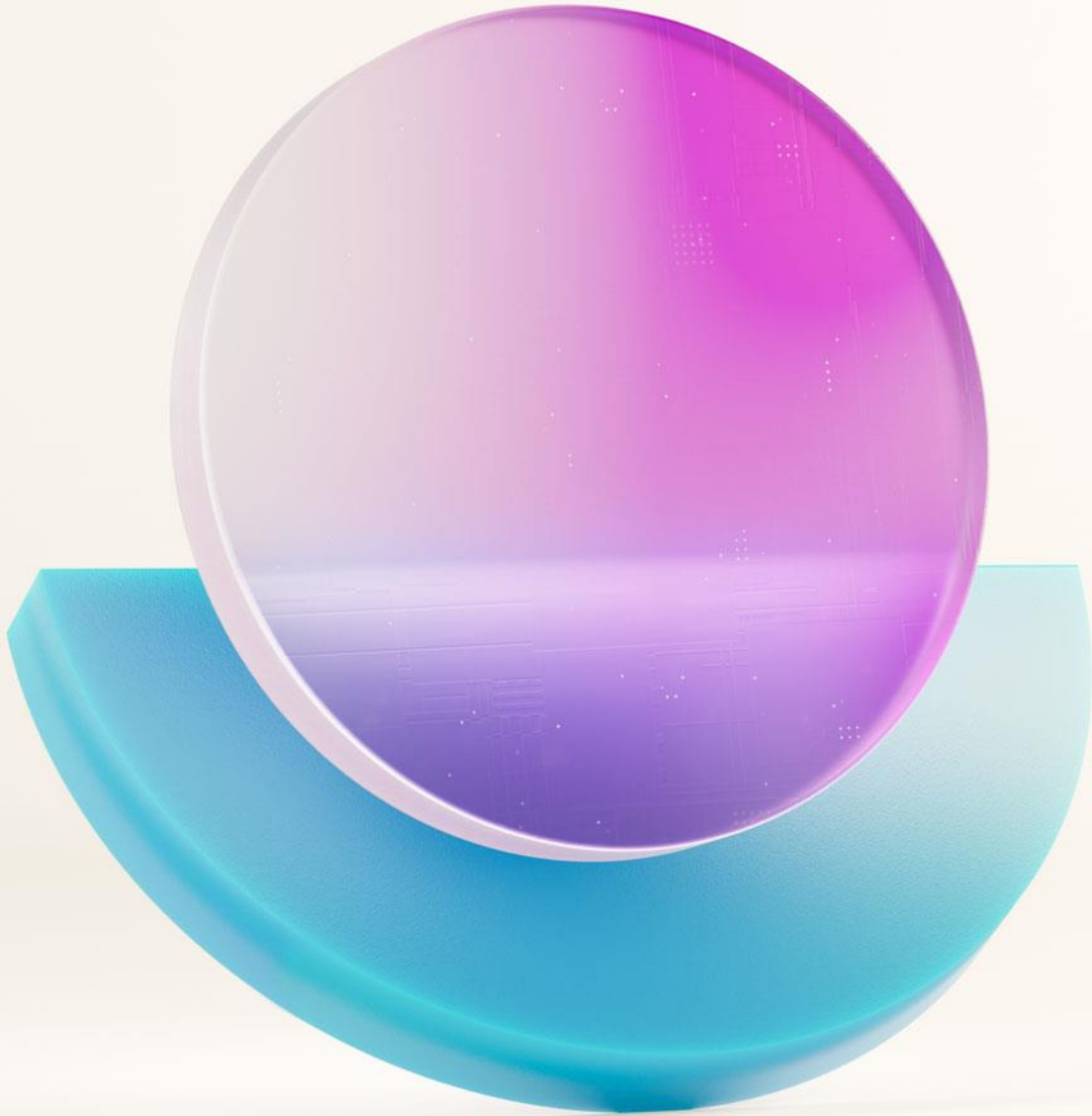




Data platform Challenges

'From data source to data model' to report



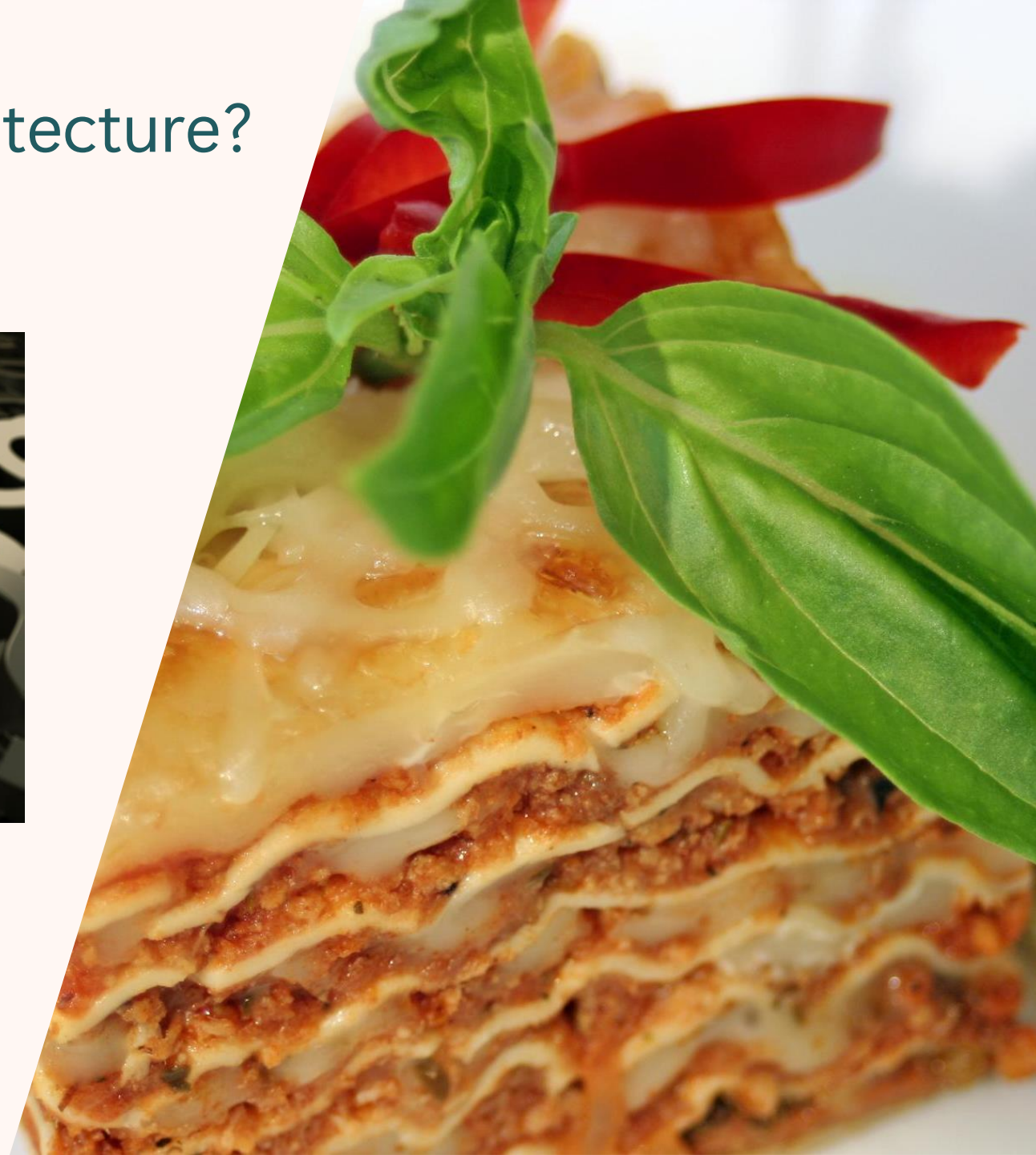


Lakehouse Medallion Architecture

Who is using a Medallion architecture?



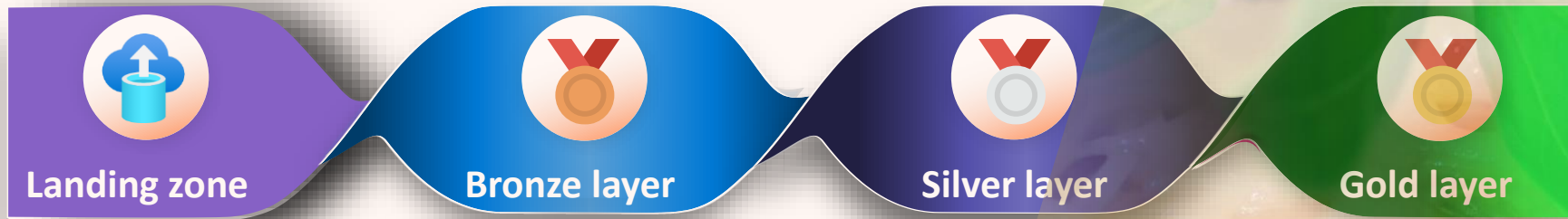
*‘Uniform data architecture’
From data “Spaghetti to Lasagna”*



Medallion Architecture

'Data processing in different stages'

Stages



Medallion Architecture

'Data processing in different stages'

Stage:

Definition:

Filetype:

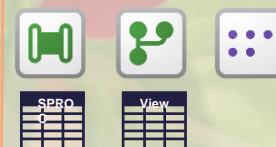
Files/Tables:

Fabric:



Gold layer

- Dimensions & Facts (Star Schema)
- Historical Analysis
- Business rules
- Documentation
- Aggregated data
- Logical table names



Silver layer

- Historical Data (Type 1 or 2)
- Data quality rules
- Data Cleansing
- Validated data
- No business model/data



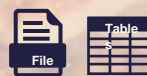
Bronze layer

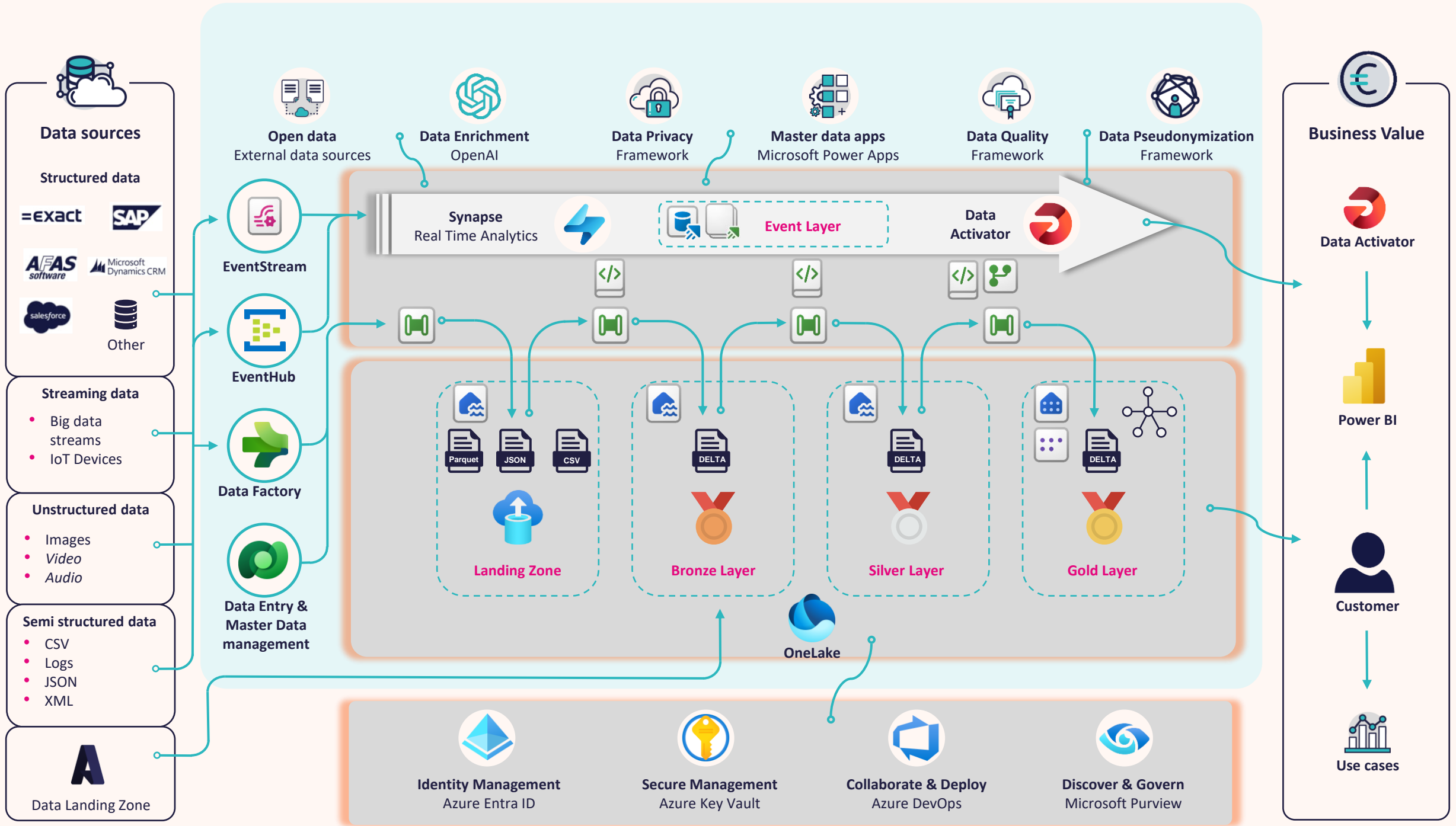
- Deduplicate data
- Add datatypes
- Data can be inconsistent
- Mostly a copy of the source
- Schema



Landing zone

- Structured data
- Unstructured data
- Incremental loads
- Data as is
- Stored in Datetime folder structure
- No Schema





Home

Create

Browse

OneLake data hub

Workspaces

FabCon

PL_FABCON_COPY

PL_FABCON_DEMO1

PL_FABCON_DEMO

PL_03_DEMO_LOAD_LA...

PL_03_DEMO_LOAD_AL...

...

Power BI

PL_FABCON_COPY

Internal

fabr

Trial: 59 days left

15

Home

Activities

Run

View

Validate

Run

Schedule

View run history

Copy data

Dataflow

Notebook

Lookup

Invoke pipeline

Start building your data pipeline

Add pipeline activity

Copy data


Choose a task to start

Copy Activity

- Queue 11 seconds
- Reading from Source 12 seconds
- Writing to Sink 35 seconds
- Transfer 60 seconds


Copy data details

Source

 Azure Data Lake Storage Gen2

→

Destination

 Lakehouse

Data read: ⓘ	3,65 GB	Data written: ⓘ	3,65 GB
Files read: ⓘ	21	Files written: ⓘ	21

Status

✔ Succeeded

Start time

2/29/2024, 2:42:15 PM

Activity run ID

cec360c0-ed69-49c1-a3ea-c2033ecba3e6

Throughput

60,839 MB/s

Total duration

00:01:13

▼ Duration breakdown

Start time

2/29/2024, 2:42:17 PM

Optimized throughput ⓘ

Standard

Used parallel copies ⓘ

6

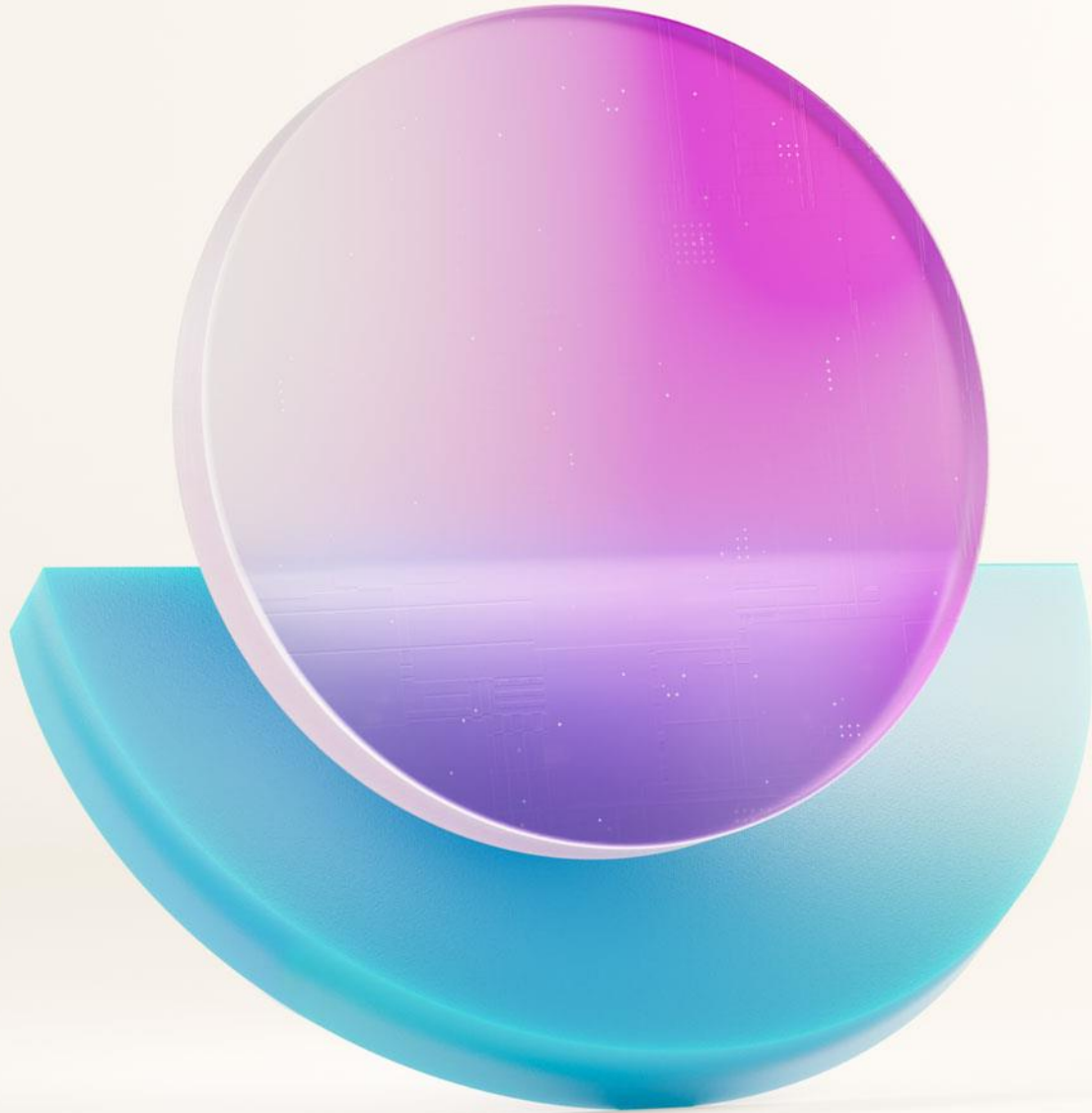
Queue

Transfer

Reading from...

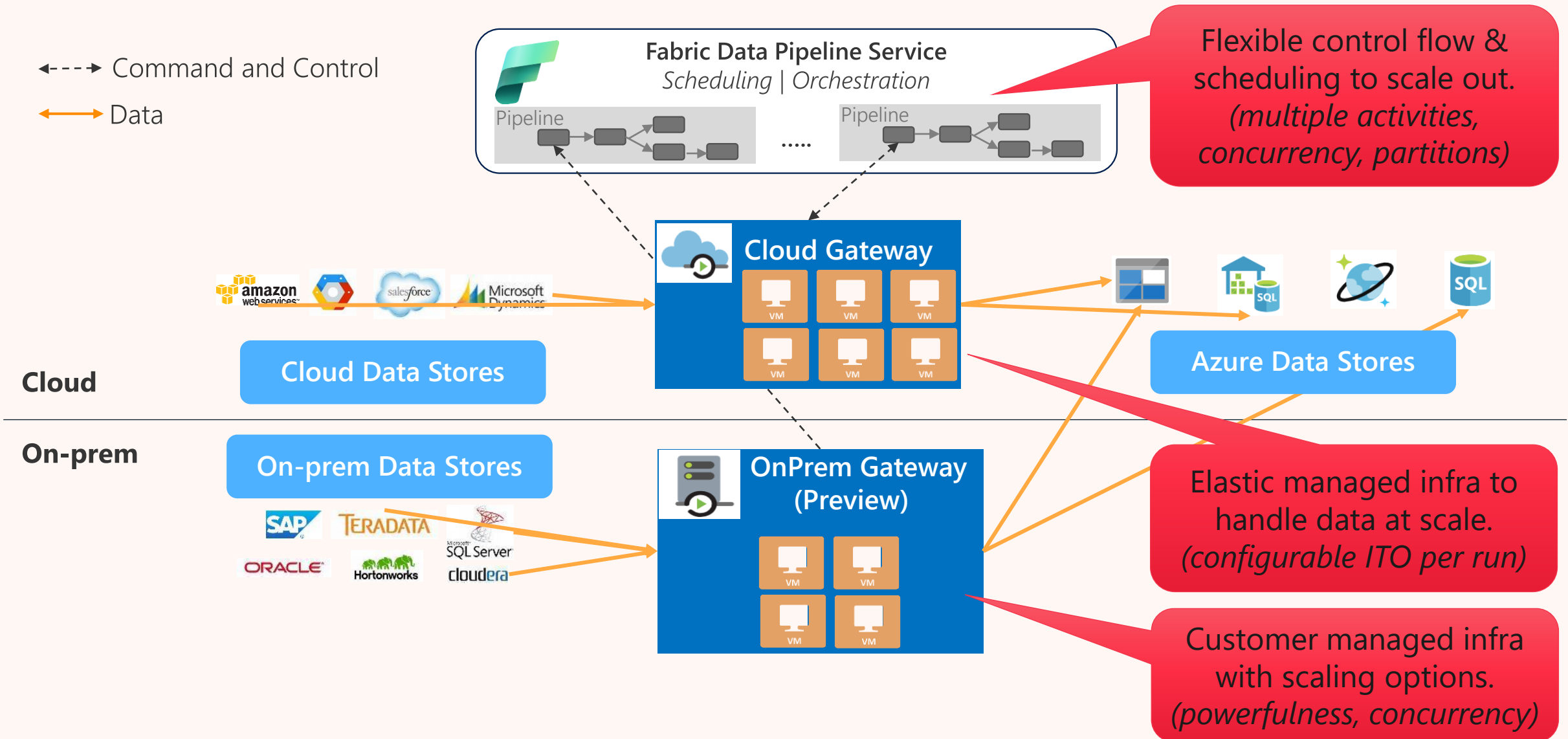
Writing to sink

Close



Pipeline Architecture

Understand How Pipeline Scales




Copy Performance Metrics

Parquet -> LH (binary)

Copy data details


loading 1t parquet to lh table binary

Source

 Azure Data Lake Storage Gen2

→

Destination

 Lakehouse

Data read: ⓘ

1.049 TB

Data written: ⓘ

1.049 TB

Files read: ⓘ

5,120

Files written: ⓘ

5,120

Status ✔ Succeeded

Start time 2/26/2024, 9:05:34 PM

Activity run ID b7fba58c-0504-462d-9eb9-fde707ad95a0

Throughput 5.141 GB/s

Total duration 00:03:31

▼ Duration breakdown


Start time 2/26/2024, 9:05:35 PM

CSV -> LH

Copy data details


Loading 1T csv to LH table

Source

 Azure Data Lake Storage Gen2

→

Destination

 Lakehouse

Data read: ⓘ

1.049 TB

Data written: ⓘ

353.654 GB

Files read: ⓘ

5,120

Files written: ⓘ

768

Rows read:

7,583,957,915

Rows written: ⓘ

7,583,957,915

Status ✔ Succeeded

Start time 2/26/2024, 3:21:25 PM

Activity run ID b4c34be5-f653-461b-b4f7-3237167b42e9

Throughput 1.284 GB/s

Total duration 00:13:44

▼ Duration breakdown


Start time 2/26/2024, 3:21:27 PM

SQL -> LH

Copy data details


sql perf

Source

 Azure SQL Database

→

Destination

 Lakehouse

Data read: ⓘ

127.292 GB

Data written: ⓘ

24.172 GB

Rows read:

512,000,000

Files written: ⓘ

256

Rows written: ⓘ

512,000,000

Status ✔ Succeeded

Start time 3/6/2024, 10:29:11 PM

Activity run ID 45f452c0-4172-40ea-9a09-f4a87d8cb0e6

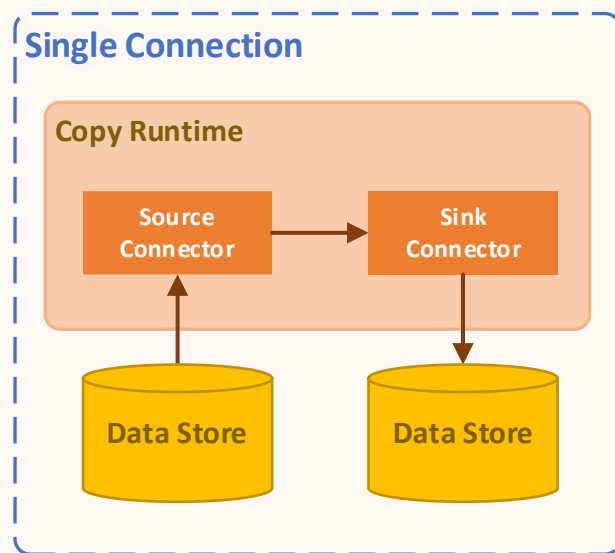
Throughput 1.354 GB/s

Total duration 00:01:41

▼ Duration breakdown

Start time 3/6/2024, 10:29:13 PM

How Copy Scales – Connection Level



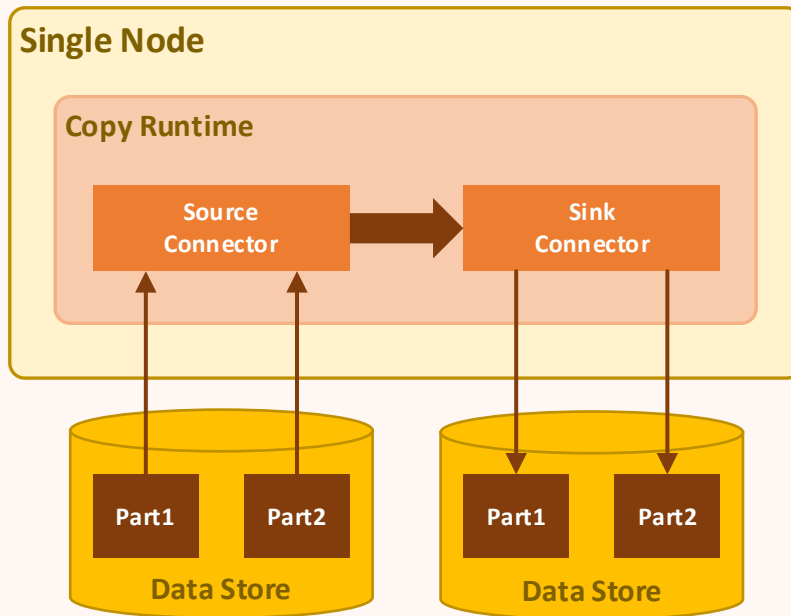
Pipeline processing

- **Less memory:** No need to load everything in memory and then write
- **Less total duration:** Read and Write are in parallel

	Clock					
	0	1	2	3	4	5
Four batches to read						
Source connector						
Destination connector						
Total Duration						

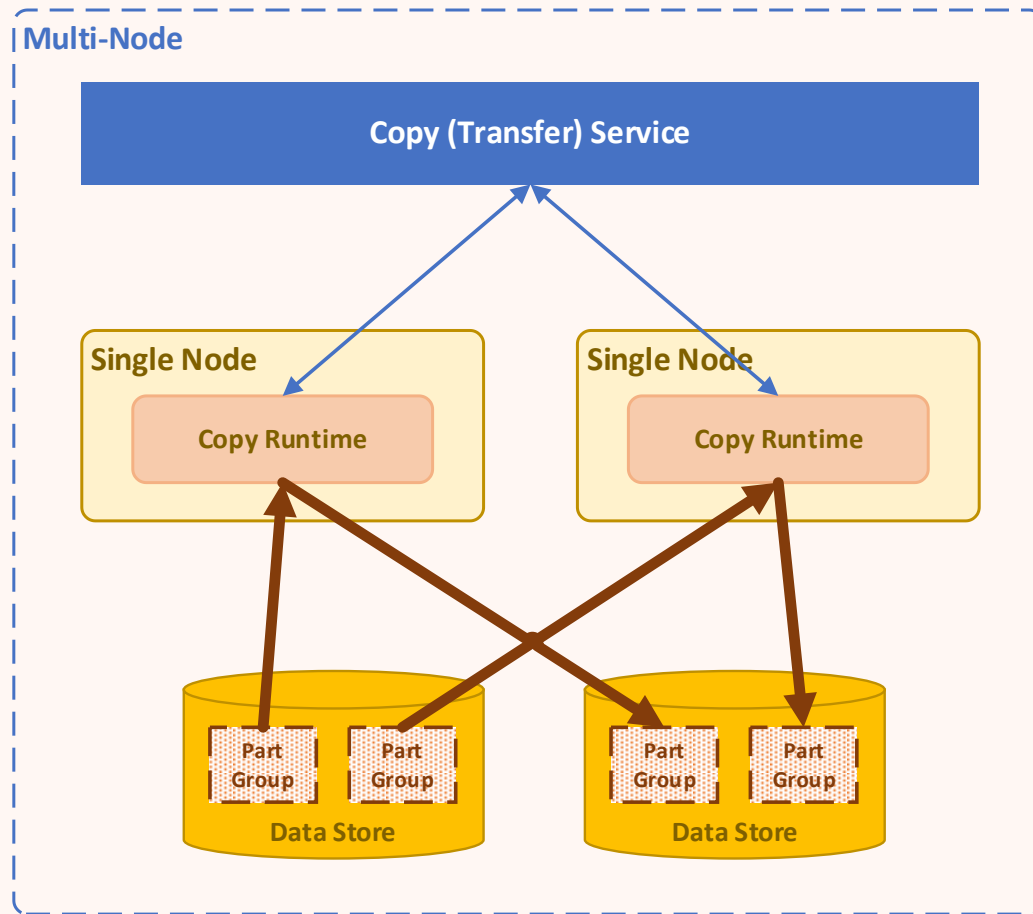
How Copy Scales – Node Level

Multi-Connection on Single Node



- Producer & Consumer design
 - Data is partitioned for multiple concurrent connections (even for single large file)
 - Full node utilization: Data can be partitioned differently between source and destination to avoid any starving / idle connection
- Partitions come from
 - Physical partitions setup on DB side
 - Dynamic partitions from different queries
 - Multiple files
 - Multiple parts from a single file

How Copy Scales – Multi-Node Level



- **Copy(Transfer) Service:** Manages Copy activities
 - Copy State Management / Monitoring
 - Cross Machine Parallelism
 - Billing
 - Manages Compute Clusters and tasks running on them
- Easily scale out to multiple stateless nodes when available partitions are more than what one node can afford.
- Theoretically no upper limit on the performance.

Copy Design Considerations

- Concepts
 - **Intelligent Throughput Optimization (ITO)**: A measure that represents the power (a combination of CPU, memory, and network resource allocation) used for a single Copy activity.
 - **Parallel Copy**: The maximum number of threads within the Copy activity that read from source and write to destination in parallel.
 - **Max Concurrent Connections**: The upper limit of concurrent connections established to the data store during the activity run. (Usually used to avoid throttling)
 - **Default perf settings**
 - Intelligent throughput optimization: **Auto** (Use maximum available VM resources)
- Implementation details
 - Files based connectors
 - 1 file = 1 partition
 - Tabular connectors that support Partition read
 - SQL family, Oracle, Teradata, Azure PostgreSQL, Netezza, SAP HANA, SAP Table, SAP Open Hub

Troubleshoot Performance Issues


- Common bottlenecks
 - Network (cross region/OnPrem)
 - Source/Destination data stores
 - Busy neighbors on the same data store
 - Low service/compute tier
 - Throttling mechanism on data store
 - Complex/unoptimized query (huge timeToFirstByte)

Troubleshoot Performance Issues – Example


Copy data details

Copy data1

Source

 Azure SQL Database

Destination

 Lakehouse

Data read: ⓘ

57.733 MB

Rows read:

232,590

Data written: ⓘ

47.61 MB

Files written: ⓘ

1

Rows written: ⓘ

232,590

Status

✓ Succeeded

Start time

3/6/2024, 7:45:31 PM

Activity run ID

19f15fea-2021-4be9-a5dd-3d14f81401f9

Throughput

4.441 MB/s

Total duration

00:00:27

▼ Duration breakdown

Start time

3/6/2024, 7:45:34 PM

Optimized throughput ⓘ

Standard

Used parallel copies ⓘ

1

Queue

Transfer

Reading from...

Writing to sink

> Advanced

Close

1

Bottleneck at Source

Parameters Variables Settings Output

Pipeline run ID: 7fdb813b-73c7-470c-bd04-cff8d959d107 ⓘ ↺ ⓘ

Pipeline status ✓ Succeeded

View run detail ⌵

Export to CSV | ⌵

⌵

⌵

Showing 1 - 1 items

Activity name ⬆️⬆️	Activity status ⬆️⬆️	Run start ⬆️⬆️	Duration ⬆️⬆️	Input	Output
Copy data1	✓ Succeeded	3/6/2024, 7:45:31 PM	33s	→	↗️

2

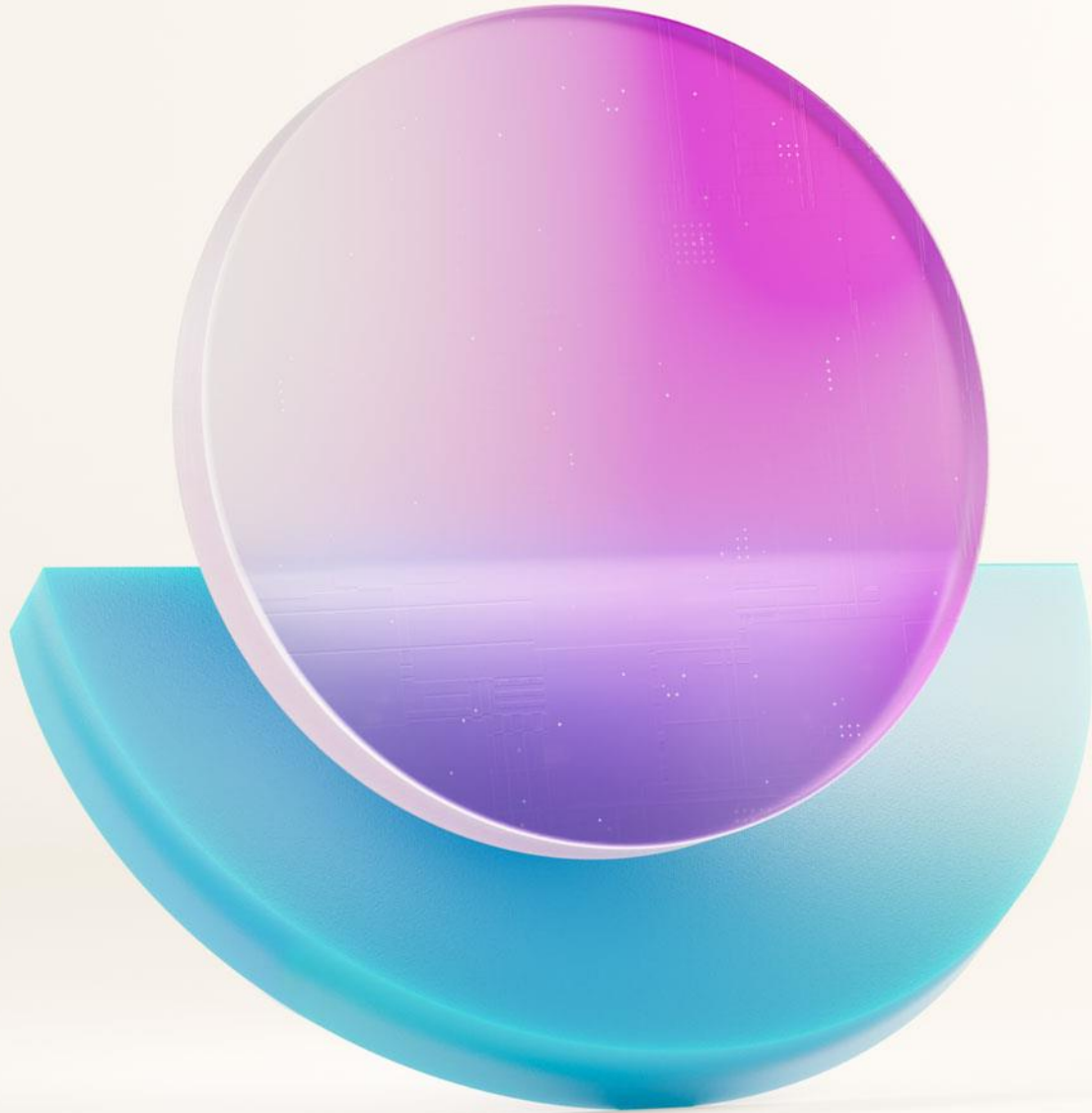
Indicating that most time was spent on preparing/running the query on DB side. Optimize the query or upgrade DB tier to eliminate this bottleneck.

3

```
usedParallelCopies: 1,
"profile": {
  "queue": {
    "status": "Completed",
    "duration": 1
  },
  "transfer": {
    "status": "Completed",
    "duration": 13,
    "details": {
      "readingFromSource": {
        "type": "AzureSqlDatabase",
        "workingDuration": 10,
        "timeToFirstByte": 9
      },
      "writingToSink": {
        "type": "Lakehouse",
        "workingDuration": 1
      }
    }
  }
},
"detailedDurations": {
```


Leveraging the Architecture

- Carefully consider number of parallel activities
- Understand Source and Destination capacity
- Set ITO to Auto and let Copy Engine tune resources



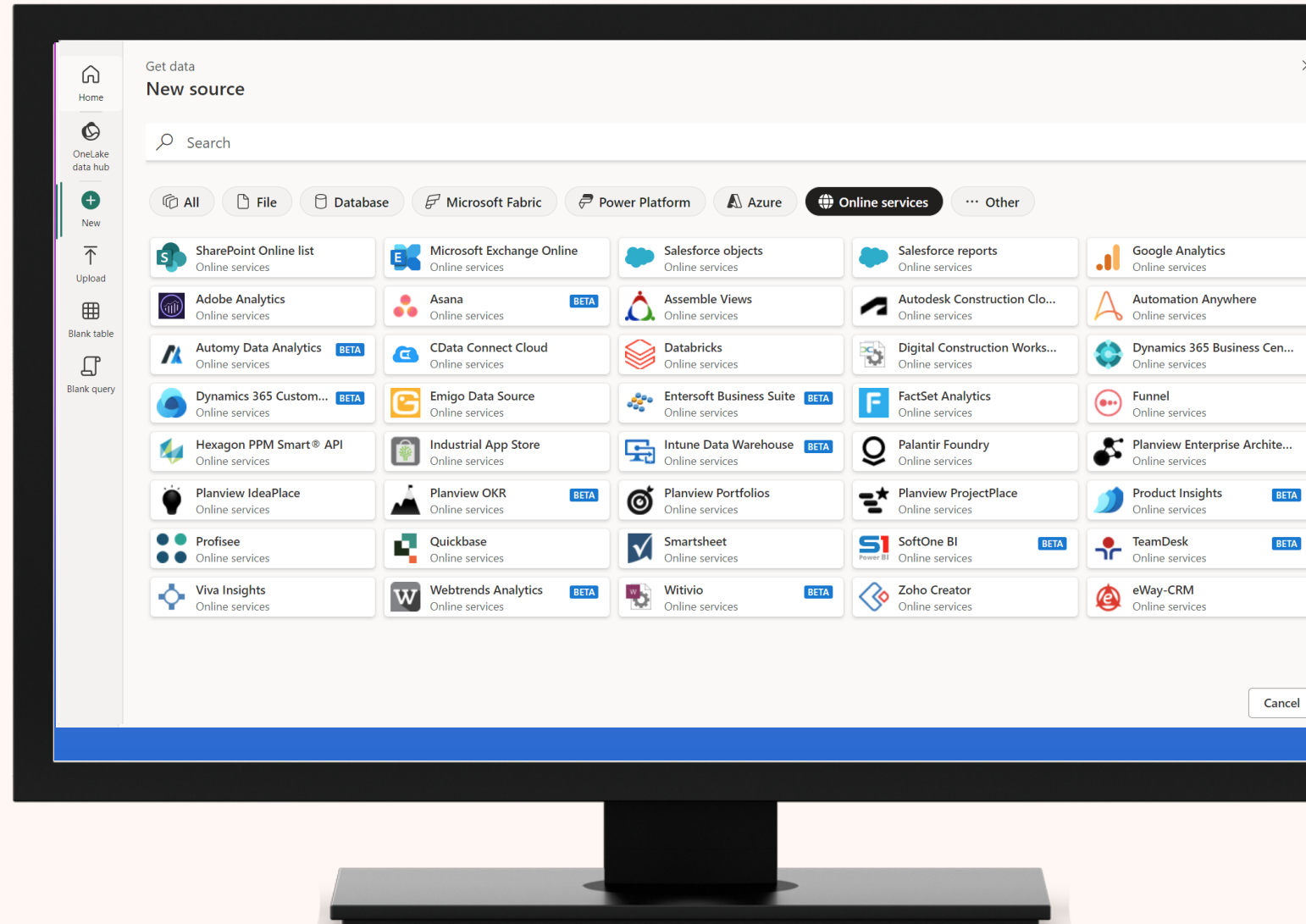
Dataflows

Dataflows Gen 2

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

Get Data

- Fabric
- Database
- Azure
- ~175 connectors



Apply Transformations

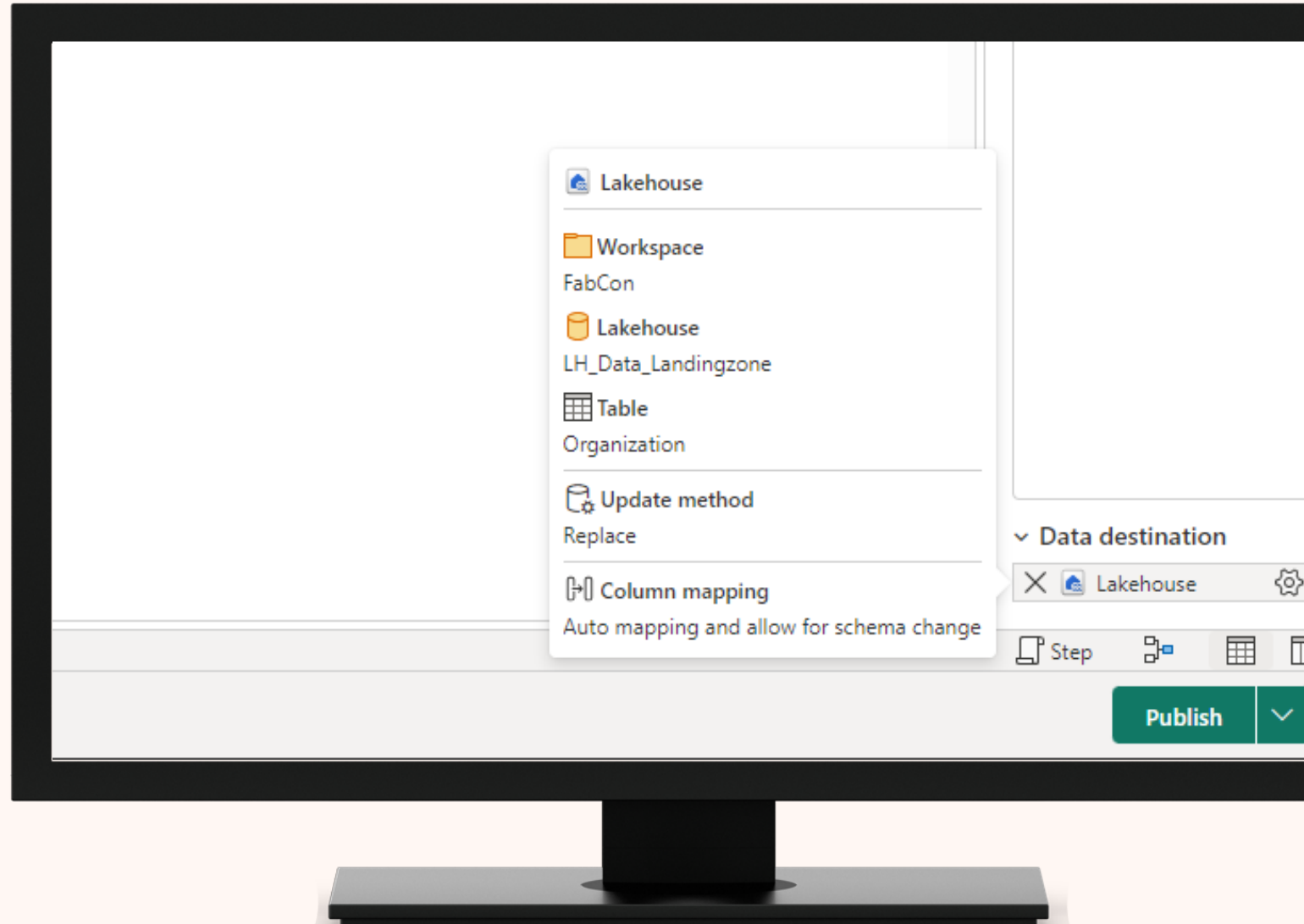
- Change data types
- Remove Columns
- Aggregations

The screenshot displays the Power Query Editor interface. The top ribbon shows the 'Transform' tab with various transformation options. The formula bar at the top indicates the current transformation: `Table.RemoveColumns(#"Changed column type", {"Index"})`. The main area shows a data table with the following columns: Organization Id, Name, Website, Country, Description, Founded, Industry, and Nu. The table contains 20 rows of data, including organizations like 'anner.com/', 'vine-marks.com/', 'www.ortiz.org/', etc.

Organization Id	Name	Website	Country	Description	Founded	Industry	Nu
74fc6fDa	anner.com/	anner.com/	Cape Verde	Horizontal bi-directional artificial intelligence	1971	Professional Training	
4C119be	vine-marks.com/	vine-marks.com/	Reunion	Progressive maximized instruction set	2008	Investment Management / Hedge Fund / Private Equ...	
347dbce	www.ortiz.org/	www.ortiz.org/	South Africa	Decentralized dynamic attitude	1993	Music	
4Dcd663	ine.info/	ine.info/	Congo	Intuitive actuating approach	2020	Information Technology / IT	
928B08e	www.barrera.com/	www.barrera.com/	Eritrea	Optional well-modulated budgetary management	1987	Recreational Facilities / Services	
22dCb8F	www.miranda.com/	www.miranda.com/	Burkina Faso	Cloned tertiary task-force	1997	Consumer Electronics	
FECE6fd7	www.anthony-braun.com/	www.anthony-braun.com/	Pitcairn Islands	Self-enabling attitude-oriented task-force	2007	Management Consulting	
2cdd825i	www.blackwell.com/	www.blackwell.com/	Nigeria	Customizable asymmetric initiative	1976	Photography	
706DC76	swrence-huffman.net/	swrence-huffman.net/	Greece	Cross-group bottom-line archive	2011	Government Relations	
Fdb4732i	www.mccormick-reed.com/	www.mccormick-reed.com/	Moldova	De-engineered maximized complexity	1988	Entertainment / Movie Production	
16FdD40	ievins.com/	ievins.com/	Netherlands Antilles	Synergized eco-centric process improvement	2005	Sports	
6c46Fd8i	wis.com/	wis.com/	Ghana	User-friendly motivating project	2003	Law Practice / Law Firms	
e0cd008i	anco-espinoza.biz/	anco-espinoza.biz/	Colombia	Versatile foreground collaboration	2009	Hospital / Health Care	
5d83EC7	naxwell.com/	naxwell.com/	Germany	Total 24/7 matrix	1989	Motion Pictures / Film	
31aDF88	nora-adkins.com/	nora-adkins.com/	Turks and Caicos Islands	Enhanced coherent functionalities	1976	Biotechnology / Greentech	
CFAE6F8i	www.morrison-ware.com/	www.morrison-ware.com/	Bulgaria	Public-key real-time groupware	2002	Architecture / Planning	
2D92015	www.velasquez-leonard.org/	www.velasquez-leonard.org/	Saint Helena	Upgradable demand-driven challenge	1987	Design	
0Fd92FBi	hite.net/	hite.net/	Faroe Islands	Assimilated demand-driven portal	1980	Animation	
48D8CdE	www.deleon.info/	www.deleon.info/	Uganda	Reduced client-server forecast	1971	Marketing / Advertising / Sales	
6bd1F8Ei	www.koch-phelps.com/	www.koch-phelps.com/	United Arab Emirates	Diverse next generation firmware	2001	Machinery	
e12C672i	www.morse.com/	www.morse.com/	Solomon Islands	Sharable even-keeled definition	2019	Mining / Metals	
6FCfbA4i	www.stewart.com/	www.stewart.com/	Tanzania	Reactive explicit task-force	1987	Motion Pictures / Film	
25D777e	hecker-lloyd.com/	hecker-lloyd.com/	Guernsey	Business-focused eco-centric firmware	1998	Religious Institutions	
231ec1di	illoway-soto.info/	illoway-soto.info/	Lao People's Democratic Republic	Secured secondary help-desk	1973	Health / Fitness	
614a968AA0bc86d	Little-Cross	https://barker.net/	Madagascar	Reactive non-volatile installation	2021	Tobacco	

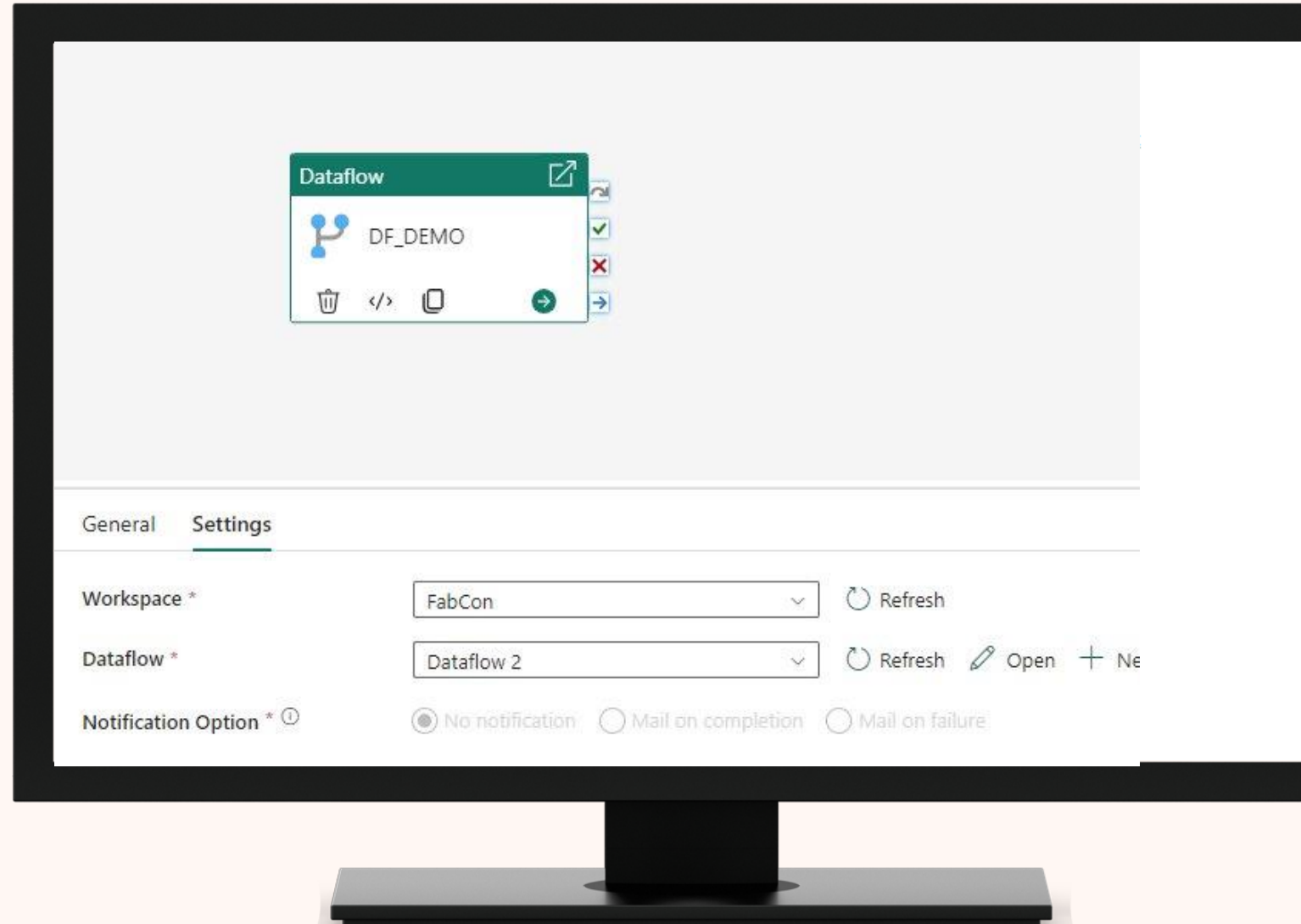
Publish Dataflows

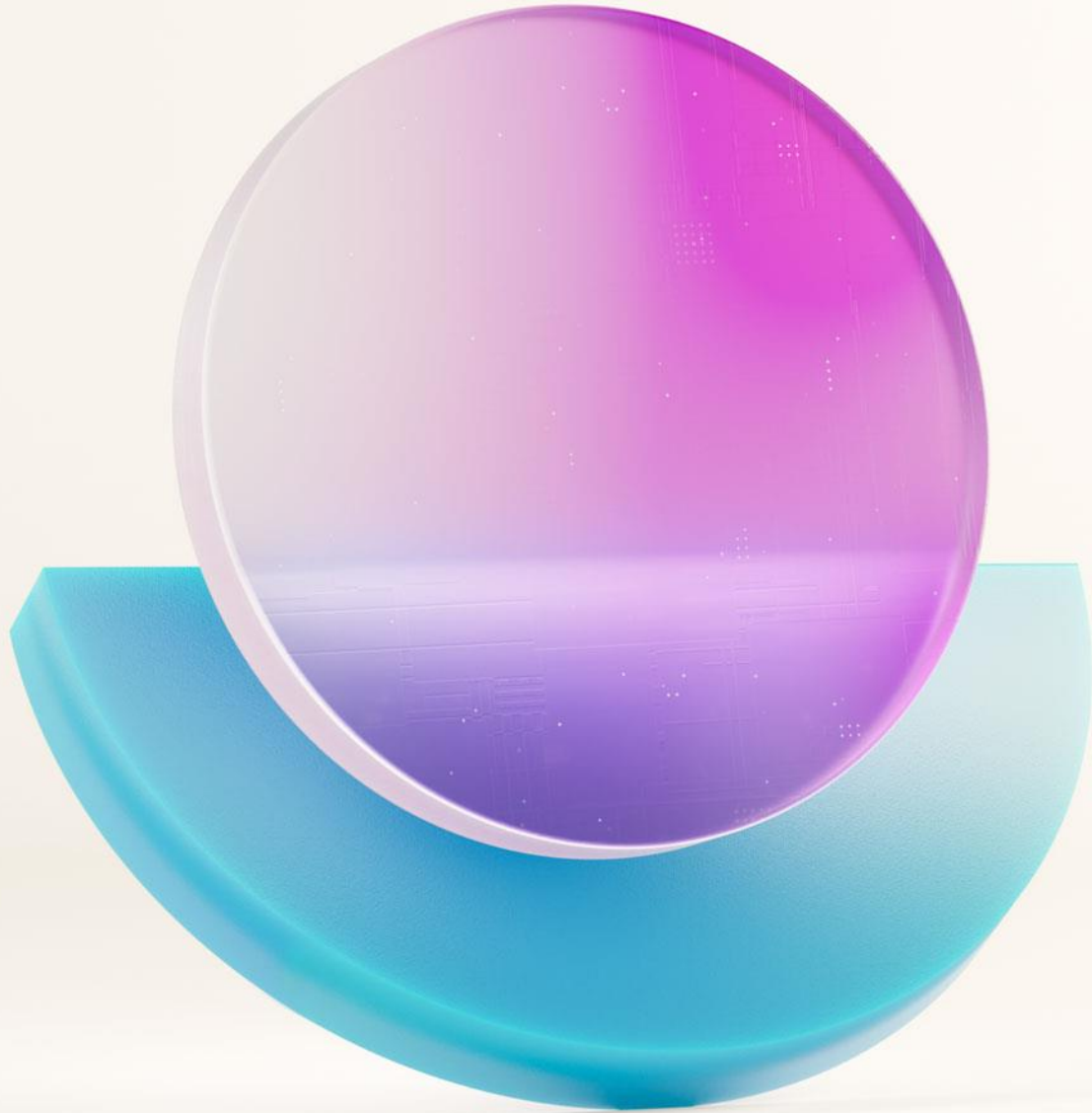
- Select Destination
 - Lakehouse
 - Warehouse
 - Azure Data Explorer
 - Azure SQL Database
 - More to come



Refreshing Dataflows

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



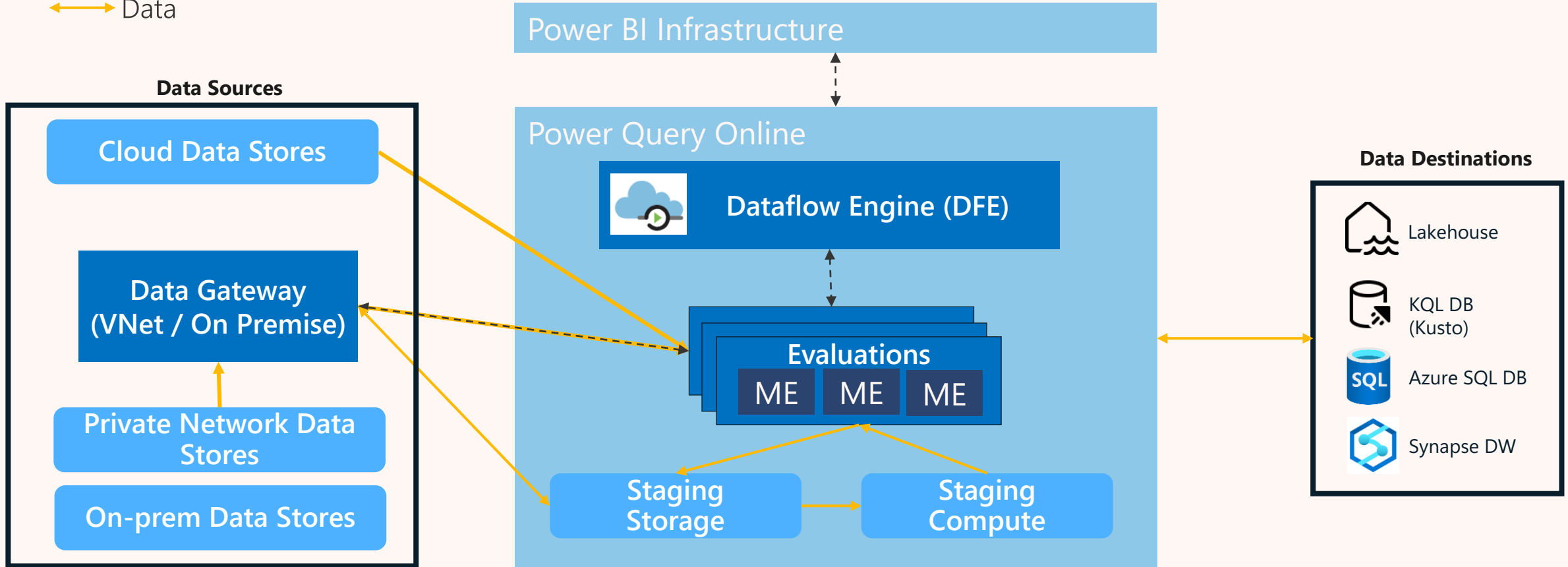


Dataflow Architecture

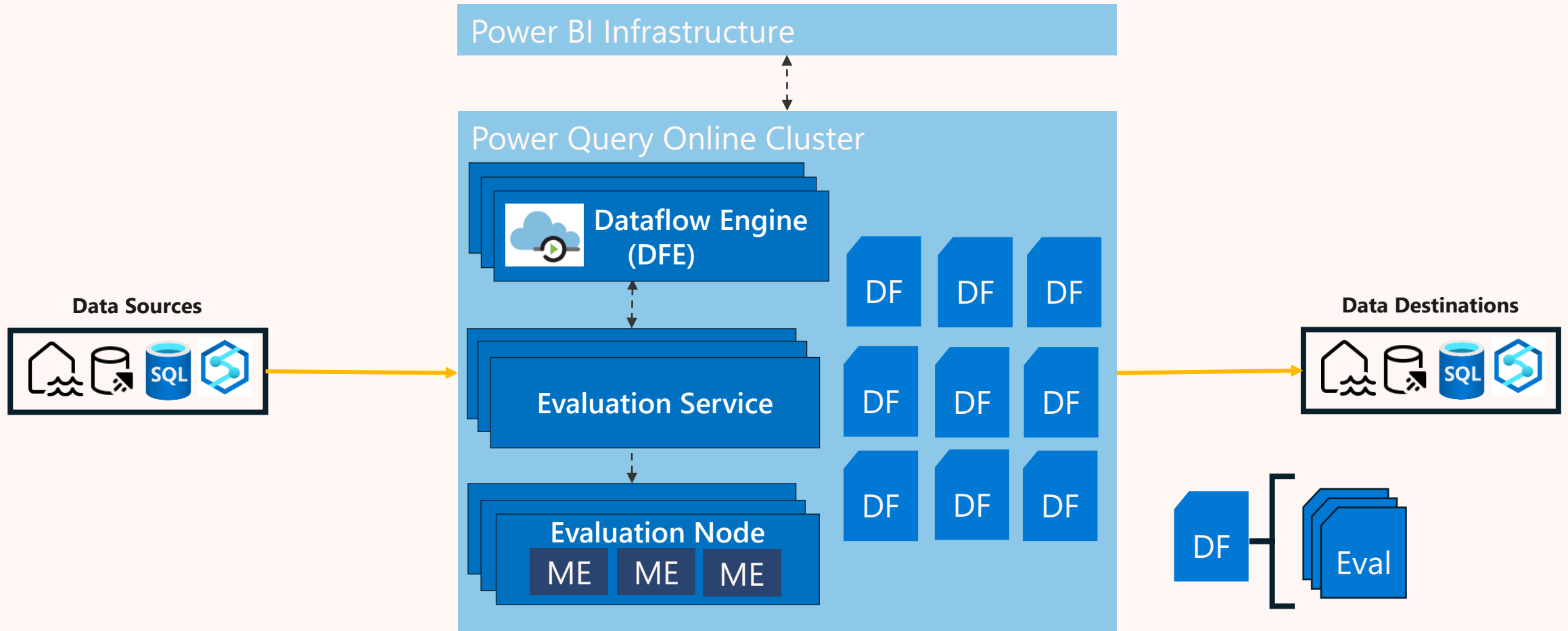
Fabric Dataflows (aka Gen2)

←---→ Command and Control

↔ Data

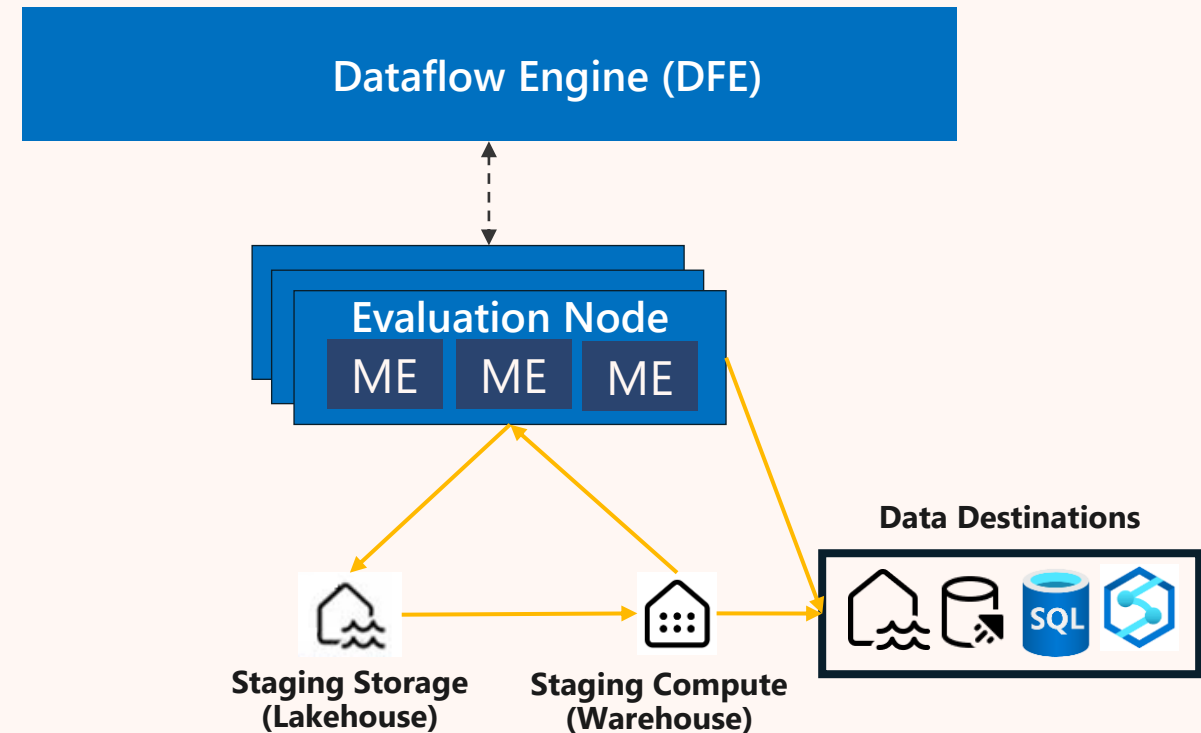


Fabric Dataflows Scale Out



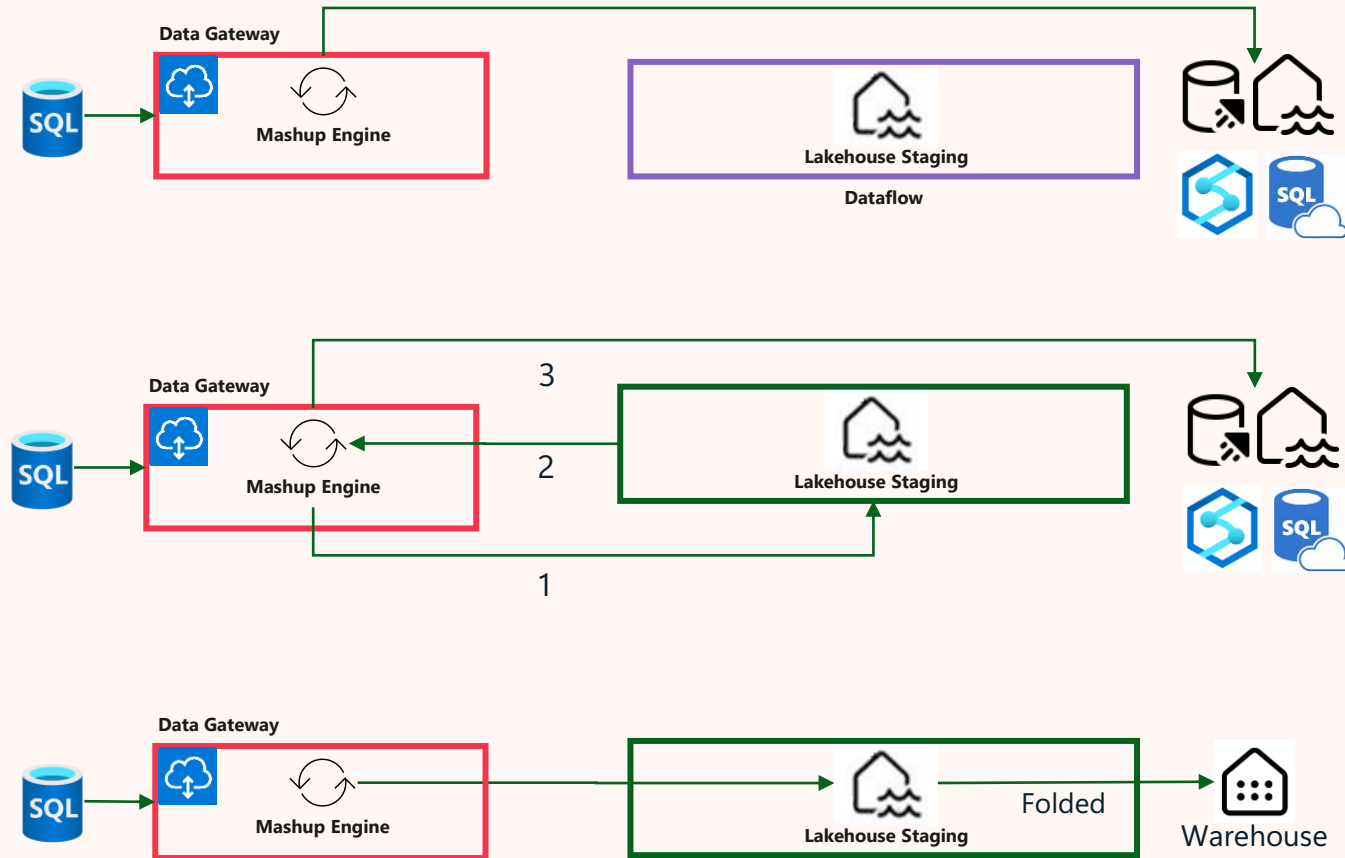
Staging

- Capabilities
 - Landing zone for incoming data
 - Provides a staging point tied to the dataflow / entity refresh
 - Provides WH Compute to speed up operations
- When to use it
 - Combining data from multiple sources
 - Stage data once for subsequent use
 - Writing to WH destination (required)
 - Merging, sorting, grouping, aggregating (uses WH compute)
- When to skip it
 - No value added for entities that write to non-WH destinations
 - See Gateway (next)



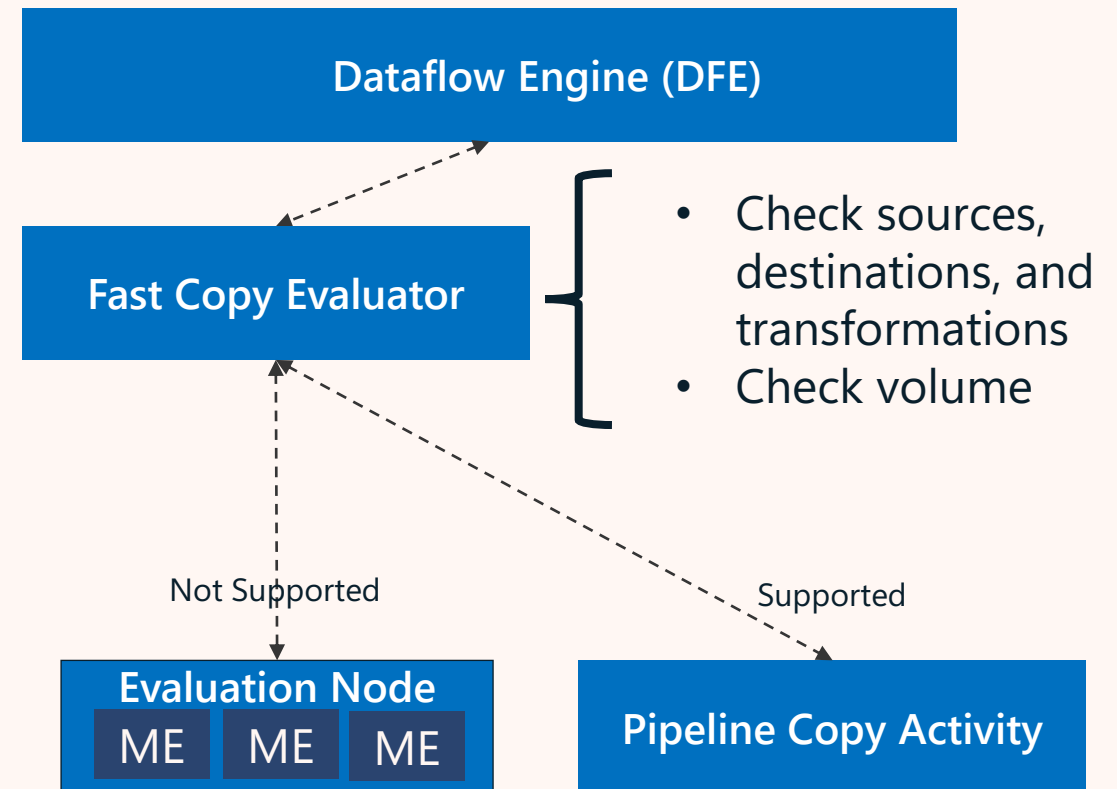
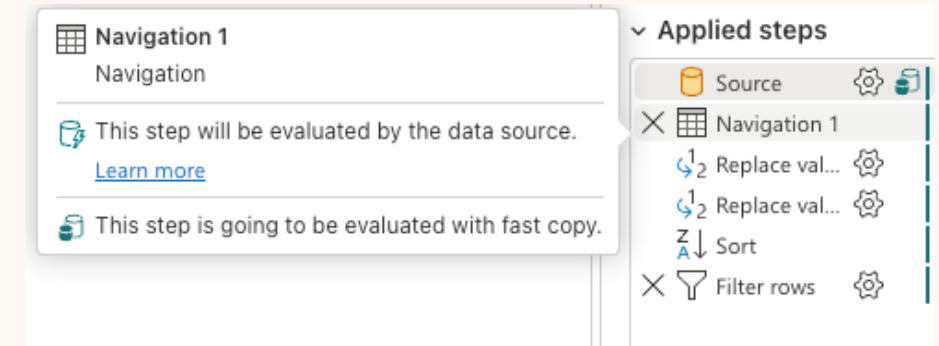
Data Gateway (VNET / On Premise)

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



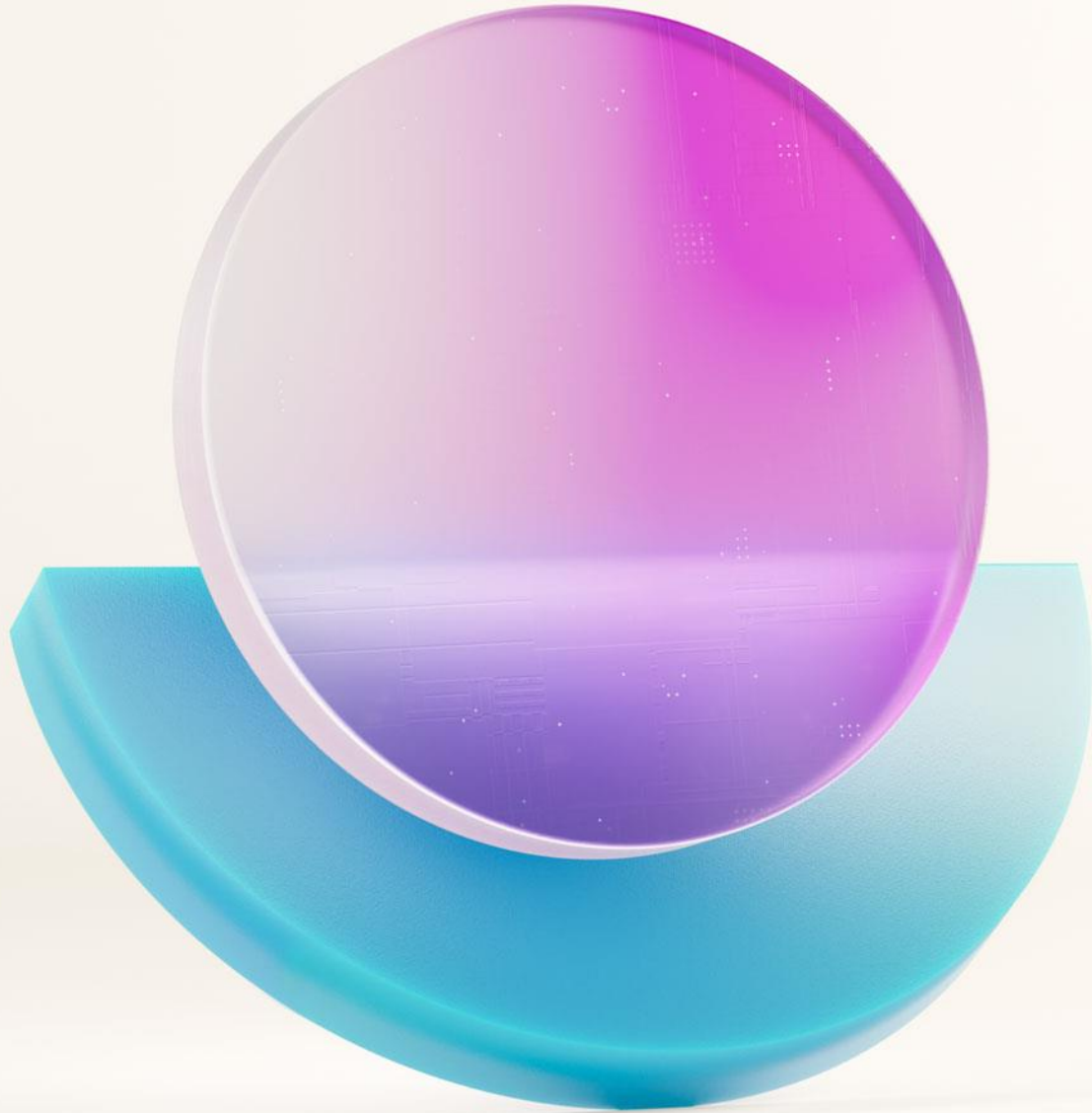
Fast Copy

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



Leveraging the Architecture

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



Q&A

Data Factory - Data Integration Sessions



TUESDAY

Getting Started with Data Factory

*Speakers: Shireen Bahadur,
Cathrine Wilhelmsen (MVP)*
TUESDAY @ 11:30am

Connecting to the World's data using Data Factory

Speakers: Matt Masson, Miguel Escobar, Jianlei Shen
TUESDAY @ 11:30am

Behind the Design: Crafting Data Factory Experiences in Fabric

Speakers: Cristin Ford, Arian Martinez, Vichita Jantlert
TUESDAY @ 2pm

From Data to Decisions: Leveraging Microsoft 365 with Data Factory

Speakers: Wilson Lee, Karan Shah, Rishi Girish
TUESDAY @ 3:15pm

WEDNESDAY

Modern Data Integration with Microsoft Fabric Data Factory

Speakers: Wee Hyong, Shabnam Watson (MVP), Penny Zhou
WEDNESDAY @ 8am

Customer Stories: Data Integration

Speakers: Andre Fomin, Tom Peplow
WEDNESDAY @ 8am

Data Factory in Microsoft Fabric Technical Deep Dive

Speakers: Mohan Sankaran, John Welch, Erwin de Kreuk (MVP)
WEDNESDAY @ 9:15am

Performance Tuning Secrets for Data Factory

Speakers: Sid Jayadevan, Mark Kromer, Matt Masson
WEDNESDAY @ 11:15am

Implement Enterprise Data Integration Patterns with Data Factory

Speakers: Abhishek Narain, Miquella de Boer, Noelle Li
WEDNESDAY @ 1:45pm

THURSDAY

Upgrade Pathways and Best Practices for Data Factory

Speakers: Mark Kromer, Miguel Escobar, Mike Carlo (MVP)
THURSDAY @ 11am

Using Azure AI Services with Data Factory

Speakers: Abhishek Narain, Joroen Luitwieler
THURSDAY @ 1:30pm

Empowering Self-service BI on SAP Data with Microsoft Fabric

Speakers: Abhishek Narain, Joroen Luitwieler
THURSDAY @ 2:45pm



New Features for DI at #FabCon

GENERAL AVAILABILITY

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

PREVIEW

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

Fast Copy for Dataflows

40 to 80 activity limit in Data Pipelines

Semantic Model Refresh

CI/CD in Data Pipelines

Cancel Dataflow Refresh

SPN support for VNET Data Gateway

Modern Get Data – browse Azure Connections

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

SNEAK PEAK

Incremental Refresh for Dataflows

Interested in Connecting with the Product Group?



Please email us at
Fabcon-DI-Speakers@microsoft.com
for any questions!



aka.ms/FabricCommunity

Ask and answer questions in
the Fabric Community forum



aka.ms/FabricUserGroups

Find a user group in your area
or to match your interests



Community Lounge Meet Ups

Check Whova for official meetups with
user group leaders, MVPs, Super Users
and more!



Meet Speakers & the Product Group

Check Whova for the full schedule of speaker Q&A and
PG meet & greets in the Community Lounge.

Thank you

