



# Microsoft Fabric

A unified analytics solution for the era of AI

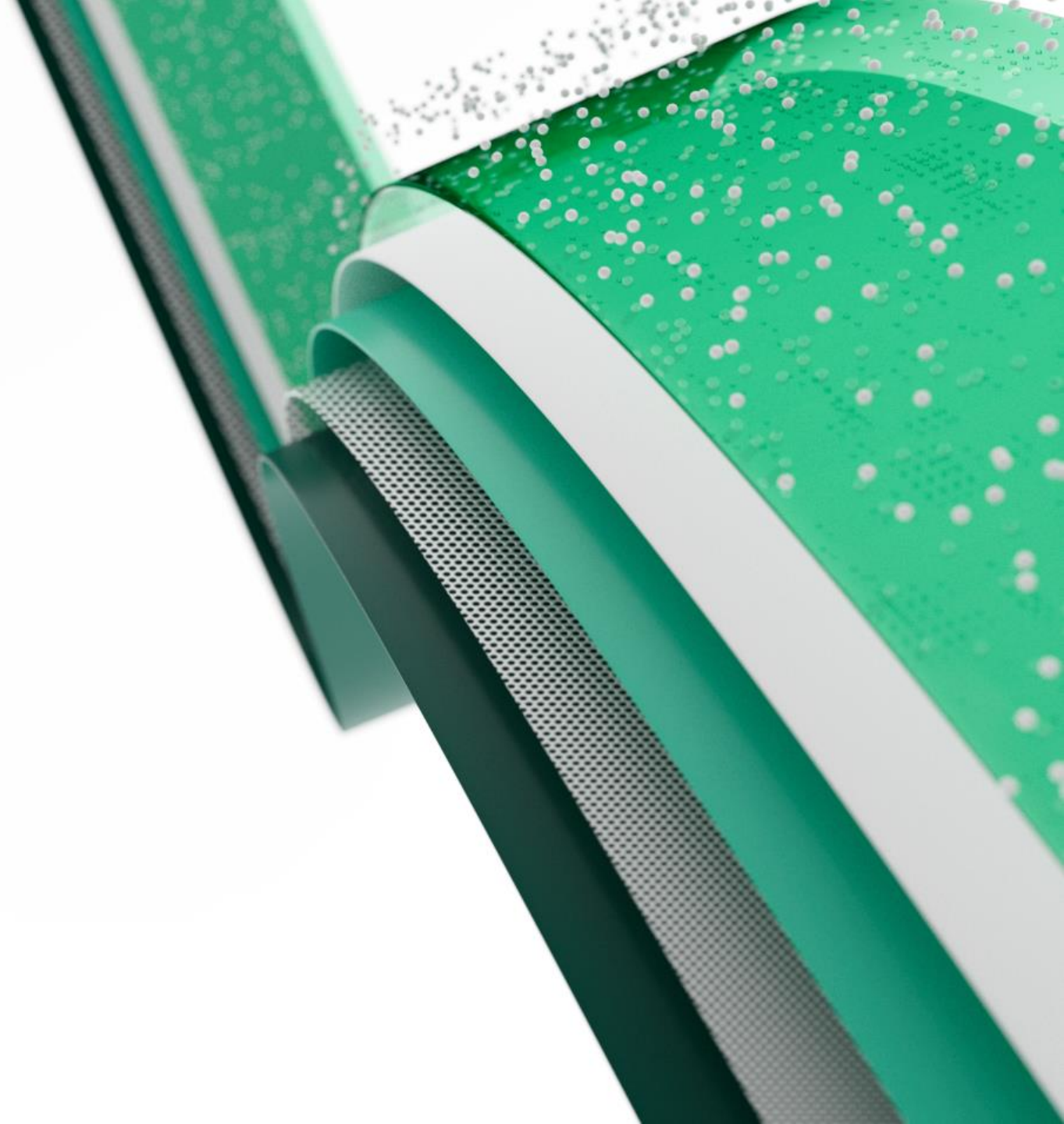
James Serra

Industry Advisor

Microsoft, Federal Civilian

[jamesserra3@gmail.com](mailto:jamesserra3@gmail.com)

6/16/23



# About Me

- Microsoft, Data & AI Solution Architect in Microsoft Federal Civilian
- At Microsoft for most of the last nine years as a Data & AI Architect , with a brief stop at EY
- In IT for 35 years, worked on many BI and DW projects
- Worked as desktop/web/database developer, DBA, BI and DW architect and developer, MDM architect, PDW/APS developer
- Been perm employee, contractor, consultant, business owner
- Presenter at PASS Summit, SQLBits, Enterprise Data World conference, Big Data Conference Europe, SQL Saturdays, Informatica World
- Blog at [JamesSerra.com](https://JamesSerra.com)
- Former SQL Server MVP
- Author of book "Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh"

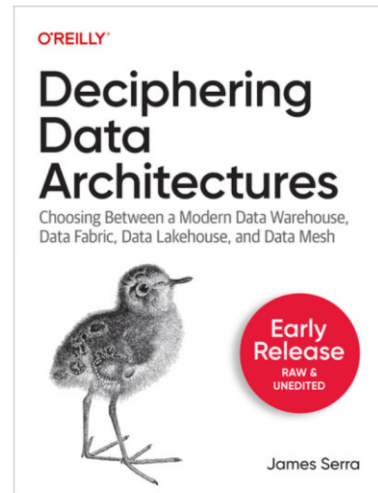


# My upcoming book

## Deciphering Data Architectures

Write the [first review](#)

By [James Serra](#)



TIME TO COMPLETE:  
49m

TOPICS:  
[Data Lake](#)

PUBLISHED BY:  
[O'Reilly Media, Inc.](#)

PUBLICATION DATE:  
September 2024

PRINT LENGTH:  
32 pages

[Table of contents](#)

Continue

Data fabric, data lakehouse, and data mesh have recently appeared as viable alternatives to the modern data warehouse. These new architectures have solid benefits, but they're also surrounded by a lot of hyperbole and confusion. This practical book provides a guided tour of each architecture to help data professionals understand its pros and cons.

In the process, James Serra, big data and data warehousing solution architect at Microsoft, examines common data architecture concepts, including how data warehouses have had to evolve to work with data lake features. You'll learn what data lakehouses can help you achieve, and how to distinguish data mesh hype from reality. Best of all, you'll be able to determine the most appropriate data architecture for your needs. By reading this book, you'll:

- Gain a working understanding of several data architectures
- Know the pros and cons of each approach
- Distinguish data architecture theory from the reality
- Learn to pick the best architecture for your use case
- Understand the differences between data warehouses and data lakes
- Learn common data architecture concepts to help you build better solutions
- Alleviate confusion by clearly defining each data architecture
- Know what architectures to use for each cloud provider

First two chapters available now:

[Deciphering Data Architectures \(oreilly.com\)](#)

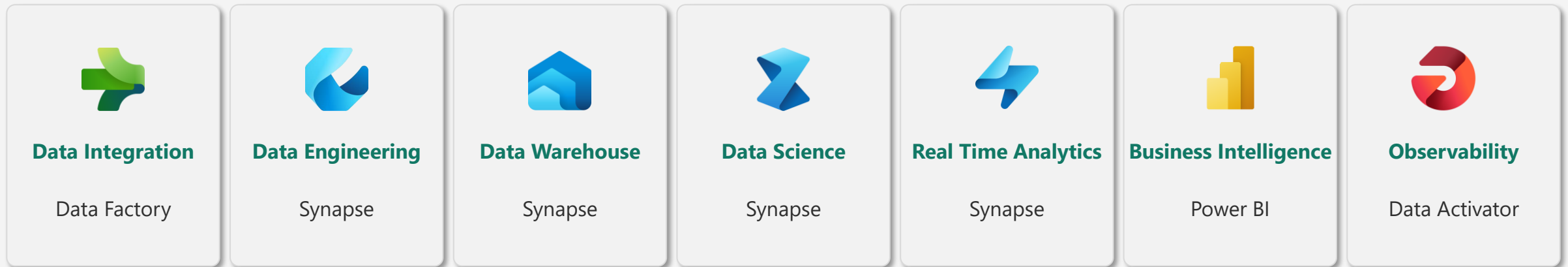
- Foundation
  - Big data
  - Types of data architectures
  - Architecture Design Session
- Common data architecture concepts
  - Relational Data Warehouse
  - Data Lake
  - Approaches to Data Stores
  - Approaches to Design
  - Approaches to Data Modeling
  - Approaches to Data Ingestion
- Data Architectures
  - Modern Data Warehouse (MDW)
  - Data Fabric
  - Data Lakehouse
  - Data Mesh Foundation
  - Data Mesh Adoption
- People, Process, and Technology
  - People and process
  - Technologies
  - Data architectures on Microsoft Azure

# Agenda

- [What is Microsoft Fabric?](#)
- [Workspaces and capacities](#)
- [OneLake](#)
- [Lakehouse](#)
- [Data Warehouse](#)
- [ADF](#)
- [Power BI / DirectLake](#)
- [Resources](#)
  
- Not covered:
  - Real-time analytics
  - Spark
  - Data science
  - Fabric capacities
  - Billing / Pricing
  - Reflex / Data Activator
  - Git integration
  - Admin monitoring
  - Purview integration
  - Data mesh
  - Copilot

# Microsoft Fabric does it all—in a unified solution

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



**Unified data foundation**  
OneLake

## UNIFIED

SaaS product experience

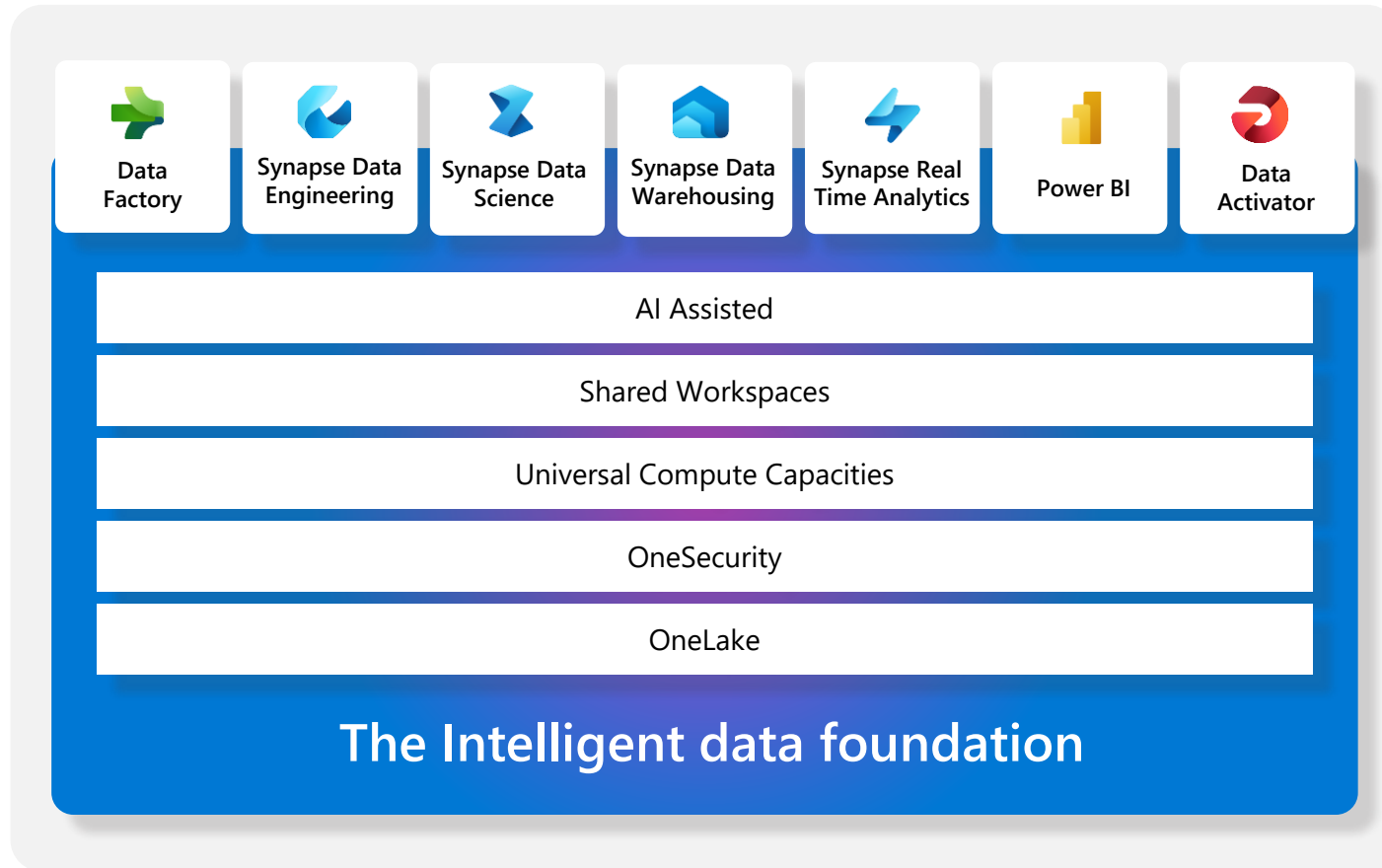
Security and governance

Compute and storage

Business model

# Microsoft Fabric

## The data platform for the era of AI



### Single...

Onboarding and trials  
Sign-on  
Navigation model  
UX model  
Workspace organization  
Collaboration experience  
Data Lake  
Storage format  
Data copy for all engines  
Security model  
CI/CD  
Monitoring hub  
Data Hub  
Governance & compliance

# SaaS

"it just works"

*5 seconds to signup, 5 minutes to wow*

5x5

Success  
by Default

Centralized  
administration

Frictionless onboarding

Minimal knobs

Tenant-wide governance

Instant Provisioning

Auto optimized

Centralized security  
management

Quick results w/ Intuitive UX

Auto Integrated

Compliance built-in

# Old vs New

## Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review

Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

### Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription \* MCAPS-Hybrid-REQ-41740-2022-jamesserra  
Resource group \* AzureDataExplorerDemo  
[Create new](#)

### Instance details

If you need to create a legacy storage account type, please click [here](#).

Storage account name ⓘ \*  
Region ⓘ \* (US) East US 2  
[Deploy to an edge zone](#)

Performance ⓘ \*  
☒ Standard: Recommended for most scenarios (general-purpose v2 account)  
☐ Premium: Recommended for scenarios that require low latency.

Redundancy ⓘ \*  
Geo-redundant storage (GRS)  
☒ Make read access to data available in the event of regional unavailability.

Review

< Previous

Next : Advanced >

## New lakehouse

Name \*

Create

Cancel

## Create Synapse workspace

\* Basics \* Security Networking Tags Review + create

Create a Synapse workspace to develop an enterprise analytics solution in just a few clicks.

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.

Subscription \* ⓘ MCAPS-Hybrid-REQ-41740-2022-jamesserra  
Resource group \* ⓘ  
[Create new](#)  
Managed resource group ⓘ Enter managed resource group name

### Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name \* Enter workspace name  
Region \* South Central US  
Select Data Lake Storage Gen2 \* ⓘ  
☒ From subscription ☐ Manually via URL  
Account name \* ⓘ  
[Create new](#)  
File system name \*  
[Create new](#)

**i** We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account, using the [Storage Blob Data Contributor](#) role. To enable other users to use this storage account after you create your workspace, perform these tasks:

- Assign other users to the **Contributor** role on workspace
- Assign other users the appropriate [Synapse RBAC roles](#) using Synapse Studio
- Assign yourself and other users to the **Storage Blob Data**

Review + create

< Previous

Next: Security >

## New warehouse

Name \*

Create

Cancel

# Understanding Microsoft Fabric / FAQ

- Think of it as taking the PBI workspace and adding a SaaS version of Synapse to it
- You will wake up one day and PBI workspaces will be automatically migrated to Fabric workspaces: PBI capacities will become fabric capacities. Your PBI tenant will have the Fabric workloads automatically built-in
- Aligned to backend fabric capacity. Similar to Power BI capacity – specific amount of compute assigned to it. A universal bucket of compute. No more Synapse DWU's, Spark clusters, etc
- Serverless Pool and Dedicated Pool combined into one – no more relational storage or dedicated resources. Everything is serverless. All about data lakehouse
- No Azure portal, subscriptions, creating storage. User won't even realize they are using Azure
- Fabric has strong separation between person who buys and pays the bill, with person who builds stuff. In Azure, the person building the solution has to also have the power to buy
- This is not just for departmental use. It's not PaaS services (i.e., Synapse) vs Fabric. Fabric is the future. Fabric is going to run your entire data estate: departmental projects as well as the largest data warehouse, data lakehouses and data science projects
- One platform for enterprise data professional and citizen developer (next slide)



## Data Engineers

- **Execute faster** with the ability to spin up a Spark VM cluster in seconds, or configure with familiar experiences like Git DevOps pipelines for data engineering artifacts
- **Streamline your work** with a single platform to build and operate real-time analytics pipelines, data lakes, lake houses, warehouses, marts, and cubes using your preferred IDE, plug-ins, and tools.
- **Reduce costly data replication** and movement with the ability to produce base datasets that can serve data analysts and data scientists without needing to build pipelines

Serve data via  
warehouse or  
lakehouse

### Supporting experiences:



Data Factory



Data Warehouse



Data Engineering



Real-time analytics



## Data Scientists

- **Quickly tune a custom model** by integrating a model built and trained in Azure ML in a Spark notebook
- **Work faster** with the ability to use your preferred data science frameworks, languages, and tools
- **Bypass engineering dependencies** with the ability to use your preferred no-code ML Ops to deploy and operate models in production
- **Tap into proven-at-scale models and services** to accelerate your AI differentiation (AOAI, Cognitive Services, ONNX integration, etc).

Serve  
transformed  
data

### Supporting experiences



Data Science



Azure ML



## Data Analysts

- **Avoid slow, progress-stagnating data wrangling** by seamlessly triggering a workflow that can unlock data engineering tools and capabilities quickly.
- **Accelerate your work** with visual and SQL based tools for self-serve data transformations and modeling as well as self-serve tools for reporting, dashboards, and data visualizations
- **Turn data into impact** with industry-leading BI tools and integration with the apps your people use everyday like Microsoft 365

Serve insights  
via  
embedding

### Supporting experiences



Data  
Warehouse



Real-time  
analytics



Power BI



## Data Citizens

- **Make more data-driven decisions** with actionable insights and intelligence in your preferred applications
- **Maintain access to all the data you need**, without being overwhelmed by data ancillary to your role thanks to fine grain data access management controls

### Supporting experiences



Power BI



Microsoft 365

Serve data via warehouse or lakehouse

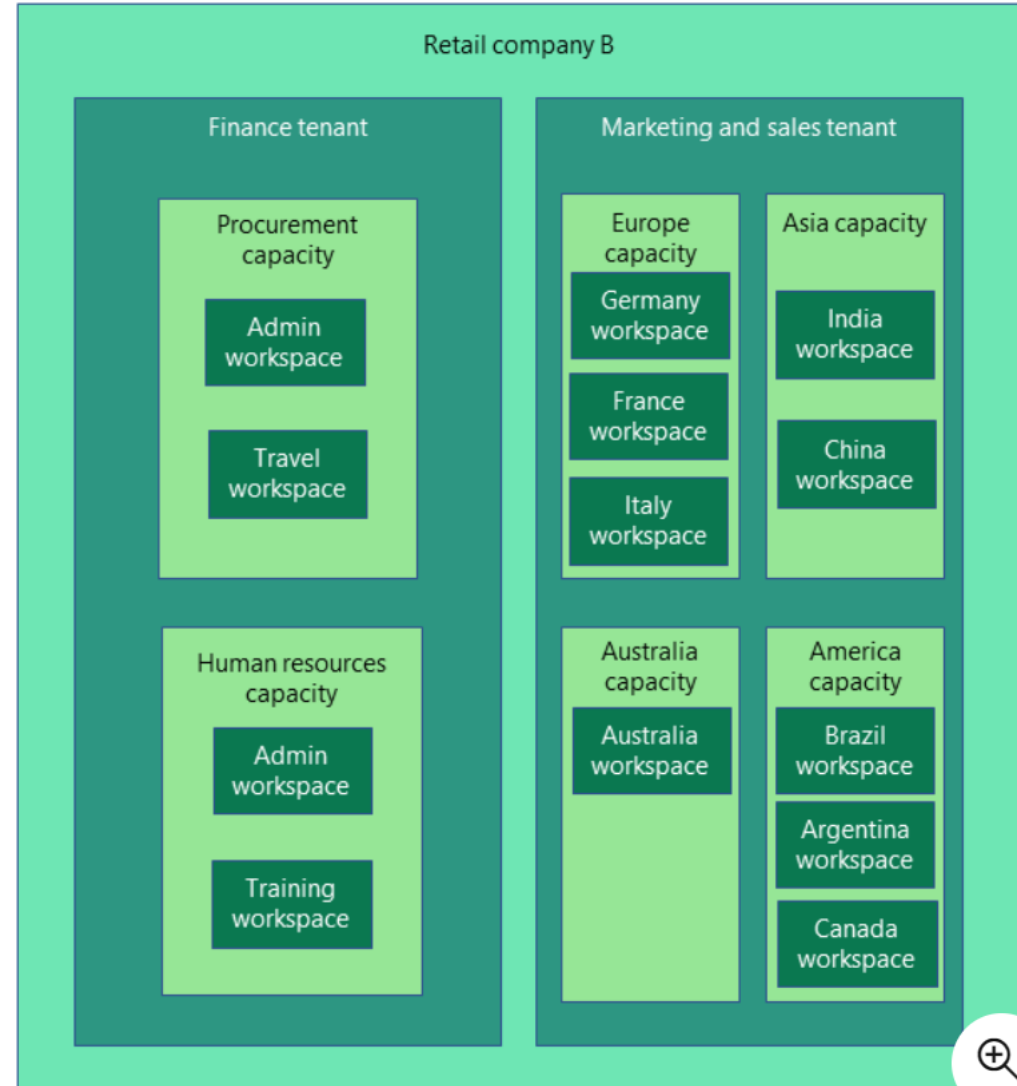


## Data Stewards


- **Maintain visibility and control of costs** with a unified consumption and cost model that provides evergreen spend optics on your end-to-end data estate
- **Gain full visibility and governance** over your entire analytics estate from data sources and connections to your data lake, to users and their insights

Workspaces and capacities

# Company examples




# Create fabric capacity



### Welcome to Microsoft Fabric

Fabric delivers an end-to-end analytics platform from the data lake to the business user.

Find out more



\* Basics   Tags   Review + create

Create Fabric capacity that you can use with your Fabric workspaces.

### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all of your resources.

Subscription \* ⓘ

Azure subscription 1

⌵

ⓘ

The Fabric resource provider is now registered with this subscription.

Resource group \* ⓘ

(New) Fabric

⌵

Create new

### Capacity details

Name your Capacity and select a location.

Capacity name \* ⓘ

chktestfabriccap

✓

Region \*

East US

⌵

Size ⓘ

F8

8 Capacity units

Change size

Fabric capacity administrator \* ⓘ

james@serracs.com

✓

Select

## Select the resource size

SKU	Capacity Units	COST (ESTIMATED/MONTH)
F2	2	\$262.80
F4	4	\$525.60
F8	8	\$1,051.20
F16	16	\$2,102.40
F32	32	\$4,204.80
F64	64	\$8,409.60
F128	128	\$16,819.20
F256	256	\$33,638.40
F512	512	\$67,276.80
F1024	1024	\$134,553.60
F2048	2048	\$269,107.20

Capacity is a dedicated set of resources reserved for exclusive use. It offers dependable, consistent performance for your content. Each capacity offers a selection of SKUs, and each SKU provides different resource tiers for memory and computing power. You pay for the provisioned capacity whether you use it or not.

A capacity is a quota-based system, and scaling up or down a capacity doesn't involve provisioning compute or moving data, so it's instant.

# Create fabric capacity

Admin portal

- Tenant settings New
- Usage metrics
- Users
- Premium Per User
- Audit logs
- Domains (preview) New
- Capacity settings
- Refresh summary
- Embed Codes
- Organizational visuals
- Azure connections
- Workspaces
- Custom branding
- Protection metrics
- Featured content

Power BI Premium   Power BI Embedded   Trial   Fabric Capacity

PREMIUM CAPACITIES

CAPACITY NAME	CAPACITY ADMINS	ACTIONS	CAPACITY SKU	CAPACITY UNITS	REGION	STATUS
chktestfabriccap	James Serra	⚙️	F2	2	East US	Active

[Set up a new capacity in Azure](#)

Once the capacity is created, we can see the capacity on the Admin portal- Capacity Settings pane under the "Fabric Capacity" tab

# Create fabric capacity

## Create a workspace

Name \*

MynewWorkspaceusingFabricCapacity

Available

Description

Describe this workspace

Domain (preview) ⓘ

Assign to a domain (optional) ▾

[Learn more about workspace settings](#) ↗

Workspace image




↑ Upload

↶ Reset

Advanced ^

Contact list \* ⓘ

 james (Owner) ×

Enter users and groups

License mode ⓘ

☐ Pro

Select Pro to use basic Power BI features and collaborate on reports, dashboards, and scorecards. To access a Pro workspace, users need Pro per-user licenses. [Learn more](#) ↗

☐ Trial

Select the free trial per-user license to try all the new features and experiences in Microsoft Fabric for 60 days. A Microsoft Fabric trial license allows users to create Microsoft Fabric items and collaborate with others in a Microsoft Fabric trial capacity. Explore new capabilities in Power BI, Data Factory, Data Engineering, and Real-Time Analytics, among others. [Learn more](#) ↗

☐ Premium per-user

Select Premium per-user to collaborate using Power BI Premium features, including paginated reports, dataflows, and datamarts. To collaborate and share content in a Premium per-user workspace, users need Premium per-user licenses. [Learn more](#) ↗

☐ Premium capacity

Select premium capacity if the workspace will be hosted in a premium capacity. When you share, collaborate on, and distribute Power BI and Microsoft Fabric content, users in the viewer role can access this content without needing a Pro or Premium per-user license. [Learn more](#) ↗

☐ Embedded ⓘ

Select embedded if the workspace will be hosted in an Azure embedded capacity. ISVs and developers use Power BI Embedded to embed visuals and analytics in their applications. [Learn more](#) ↗

☒ Fabric capacity

Select Fabric capacity if the workspace will be hosted in a Microsoft Fabric capacity. With Fabric capacities, users can create Microsoft Fabric items and collaborate with others using Fabric features and experiences. Explore new capabilities in Power BI, Data Factory, Data Engineering, and Real-Time Analytics, among others. [Learn more](#) ↗

Default storage format

☒

Small dataset storage format

☐

Large dataset storage format

[Learn more about dataset storage formats](#) ↗

Capacity \*

chktestfabriccap - East US ▾

# Turning on Microsoft Fabric

## Admin portal

Tenant settings New

Usage metrics

Users

Premium Per User

Audit logs

Domains (preview) New

Capacity settings

Refresh summary

Embed Codes

Organizational visuals

Azure connections

Workspaces

Custom branding

Protection metrics

Featured content

There are new or updated tenant settings. Expand to review the changes.

### Microsoft Fabric (Preview)

Users can create Fabric items (public preview) New

*Unapplied changes*

Users can create Fabric items with new capabilities in Microsoft Fabric. This setting can be managed at both the tenant and the capacity levels. [Learn More](#)

Microsoft has released the Fabric Public Preview in an Off state by default, if you don't make changes to this setting, Microsoft will change the state to On after July 1. If you decide to opt out before July 1, Fabric Preview will remain off until you turn it on.

☐ Accept Microsoft's default selection (Off for the entire organization)

☒ Enabled

Apply to:

☒ The entire organization

☐ Specific security groups

☐ Except specific security groups

Delegate setting to other admins

☒ Capacity admins can enable/disable

Apply

Cancel

## Admin portal

Usage metrics

Users

Premium Per User

Audit logs

Domains (preview) New

Tenant settings

Capacity settings

Refresh summary

Embed Codes

Organizational visuals

Azure connections

Workspaces

Custom branding

Protection metrics

Featured content

Microsoft Purview setting

## Power BI Premium > Admin Monitoring Upgra...

Details 

Delegated tenant settings

### Microsoft Fabric (Preview)

Users can create Fabric items (public preview)

*Enabled for all users in capacity*

Users can create Fabric items with new capabilities in Microsoft Fabric. This setting can be managed at both the tenant and the capacity levels. [Learn More](#)

☐ Override tenant admin selection

☒ Enabled

Apply to:

☒ All the users in capacity

☐ Specific security groups

☐ Except specific security groups

Apply

Cancel

Demo

OneLake

# OneLake for all data

## "The OneDrive for data"

### OneLake

- › A single unified logical SaaS data lake for the whole organization (no silos)
- › Organize data into domains
- › Foundation for all Fabric data items
- › Provides full and open access through industry standard APIs and formats to any application (no lock-in)

---

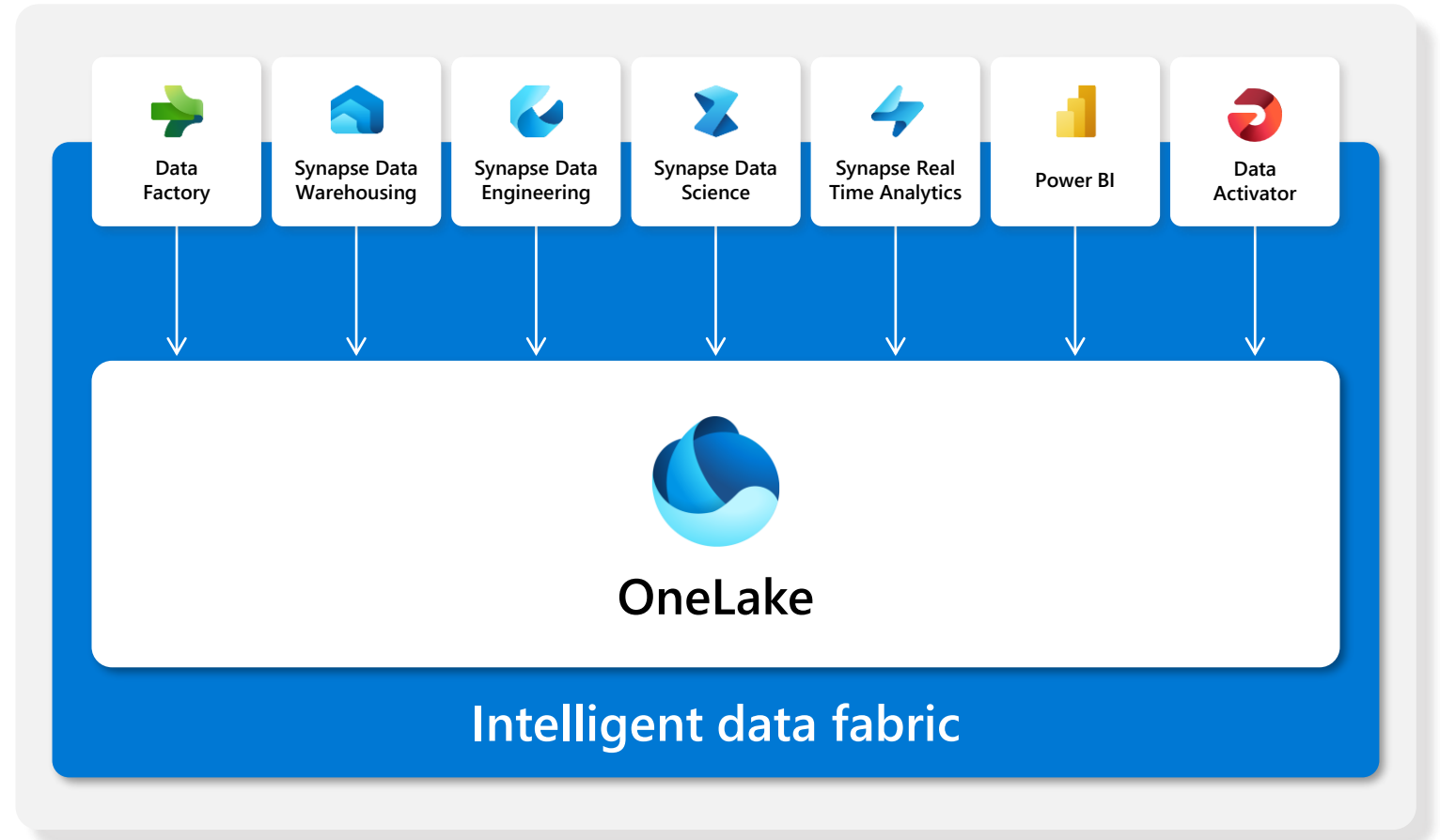
One Copy

---

One Security

---

OneLake Data Hub



# One Copy for all computes

## Real separation of compute and storage

Compute powers the applications and experiences in Fabric. The compute is separate from the storage.

Multiple compute engines are available, and all engines can access the same data without needing to import or export it. You are able to choose the right engine for the right job.

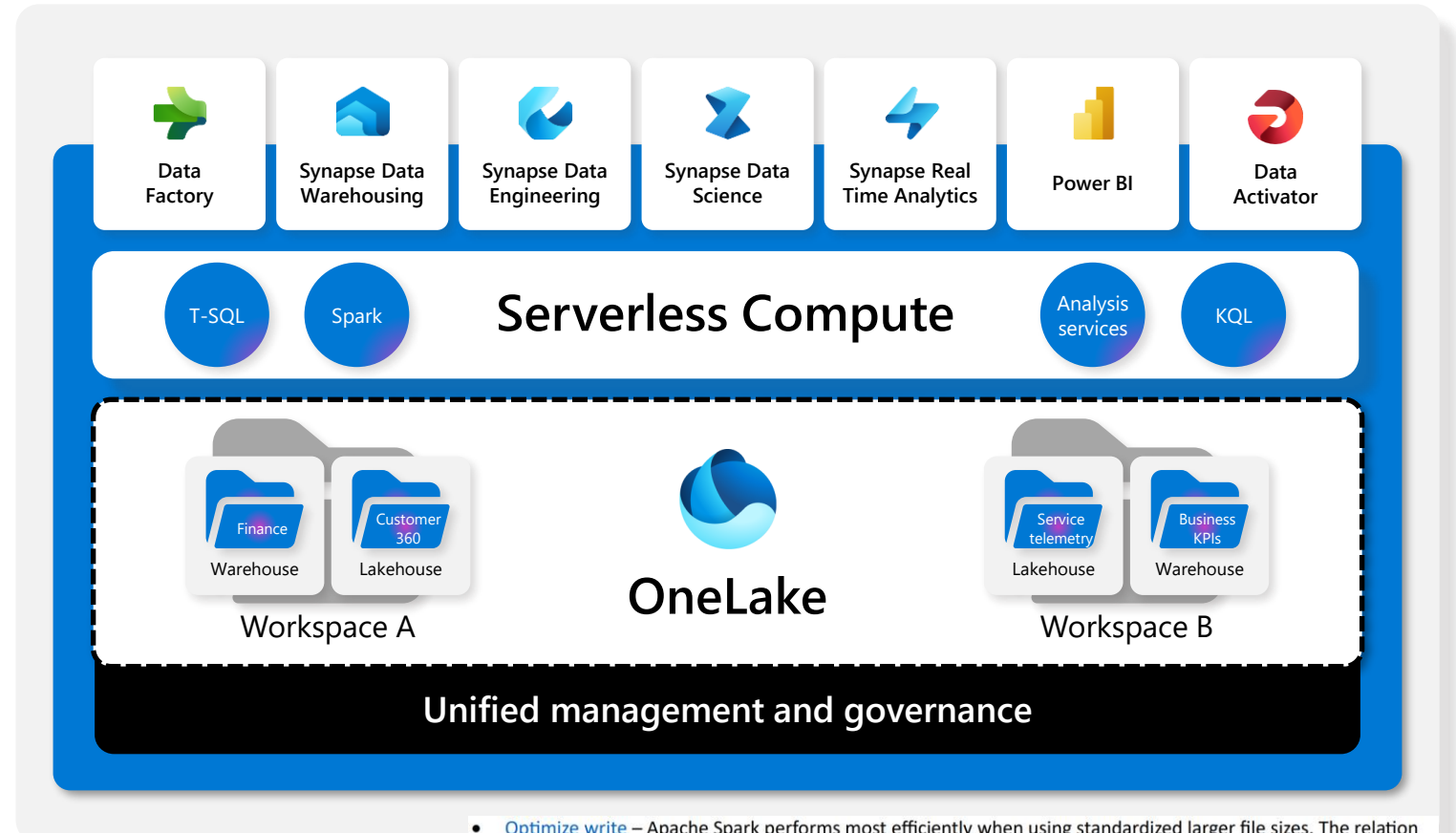
Non-Fabric engines can also read/write to the same copy of data using the ADLS APIs or added through shortcuts

No matter which engine or item you use, everyone contributes to building the same lake.

### ***Engines are being optimized to work with Delta Parquet as their native format***

Trident provides these unique capabilities for writing optimized delta lake files:

- **Verti-Parquet** – Trident includes Microsoft's unique VertiParquet IP. VertiParquet transparently optimizes the Delta Lake files in a way that is highly optimized by Trident compute engines, often resulting in 3x-4x compression improvement and up to 10x performance acceleration over Delta Lake files not optimized using VertiParquet while still maintaining full Delta Lake format compliance.



- **Optimize write** – Apache Spark performs most efficiently when using standardized larger file sizes. The relation between the file size, the number of files, the number of Spark workers and Spark's configurations play a critical role in performance. Ingestion workloads into Delta Lake tables may have the inherited characteristic of constantly writing lots of small files; this scenario is commonly known as the "small files problem". To overcome this problem, Spark in Trident includes an Optimize Write feature that reduces the number of files written and aims to increase individual file size of the written data. It dynamically optimizes partitions while generating files with a default 128 MB size. The target file size may be changed per workload requirements using configurations.

# Shortcuts virtualize data across domains and clouds

## No data movements or duplication

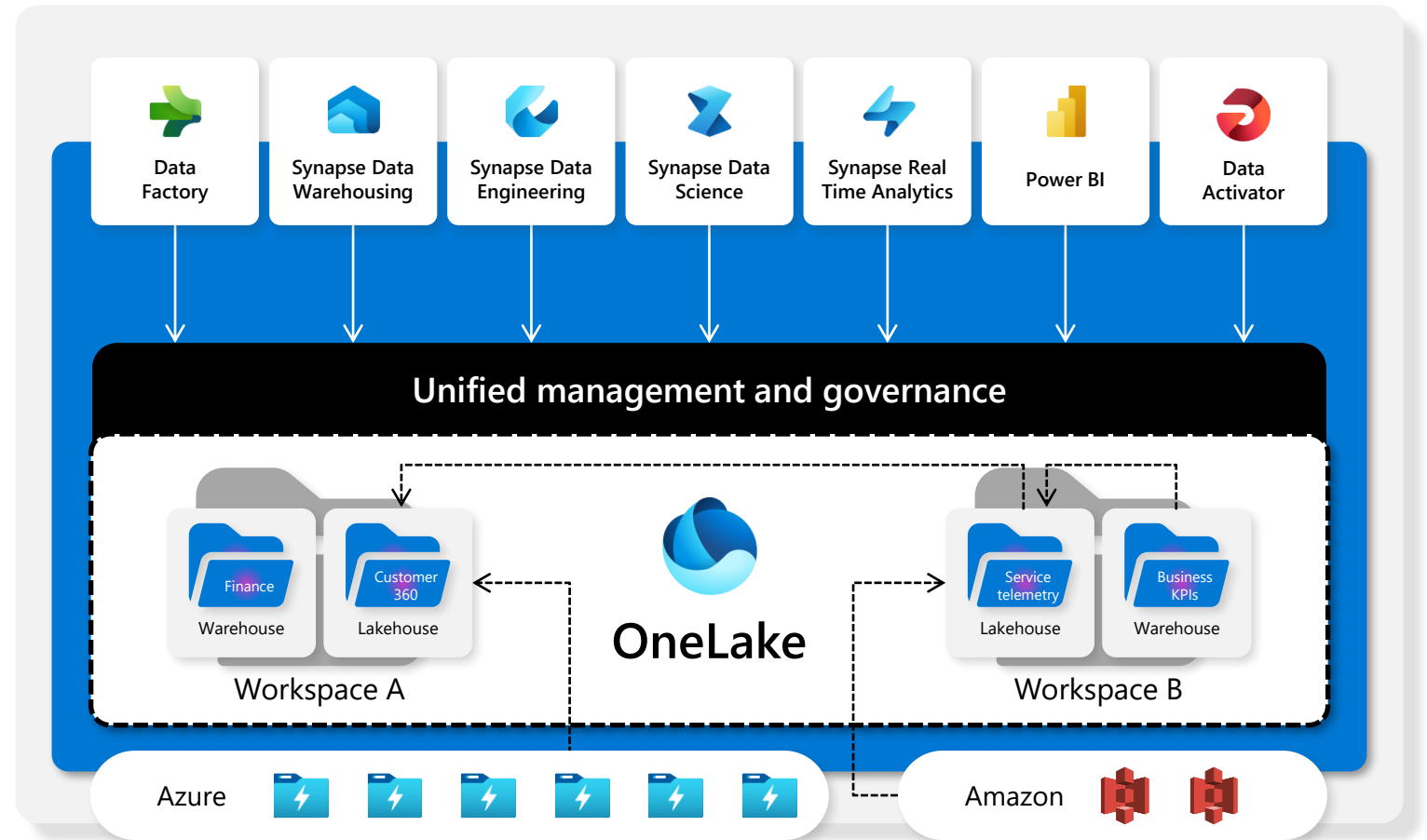
A shortcut is a symbolic link which points from one data location to another

Create a shortcut to make data from a warehouse part of your lakehouse

Create a shortcut within Fabric to consolidate data across items or workspaces without changing the ownership of the data. Data can be reused multiple times without data duplication.

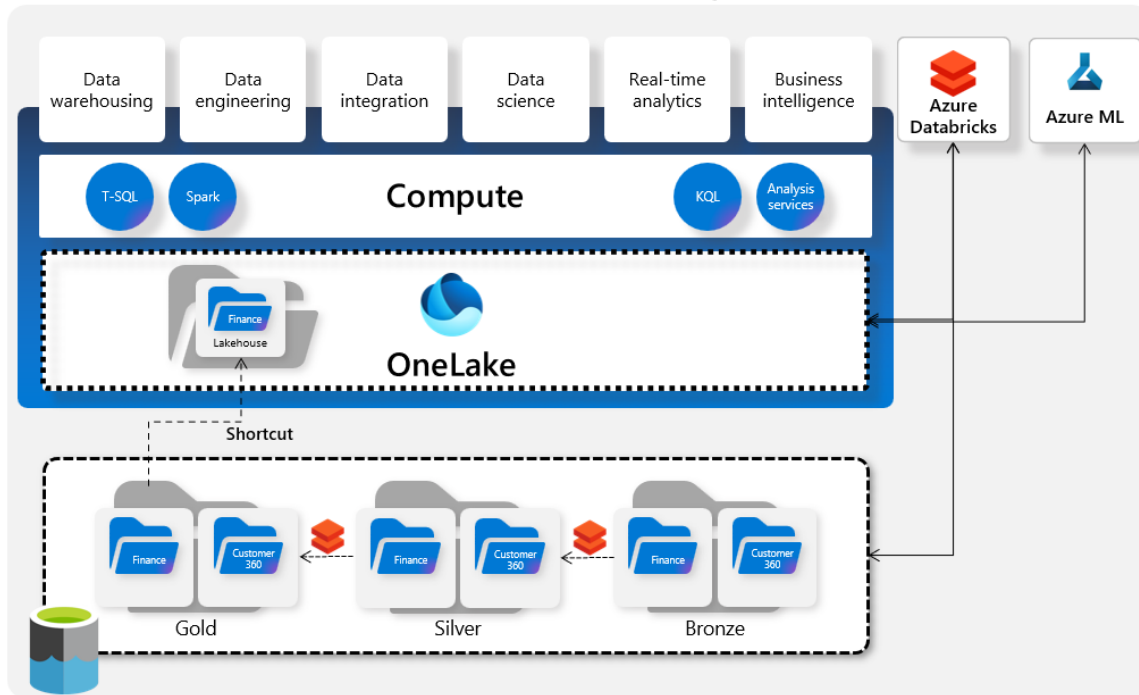
Existing ADLS gen2 storage accounts and Amazon S3 buckets can be managed externally to Fabric and Microsoft while still being virtualized into OneLake with shortcuts

All data is mapped to a unified namespace and can be accessed using the same APIs including the ADLS Gen2 DFS APIs

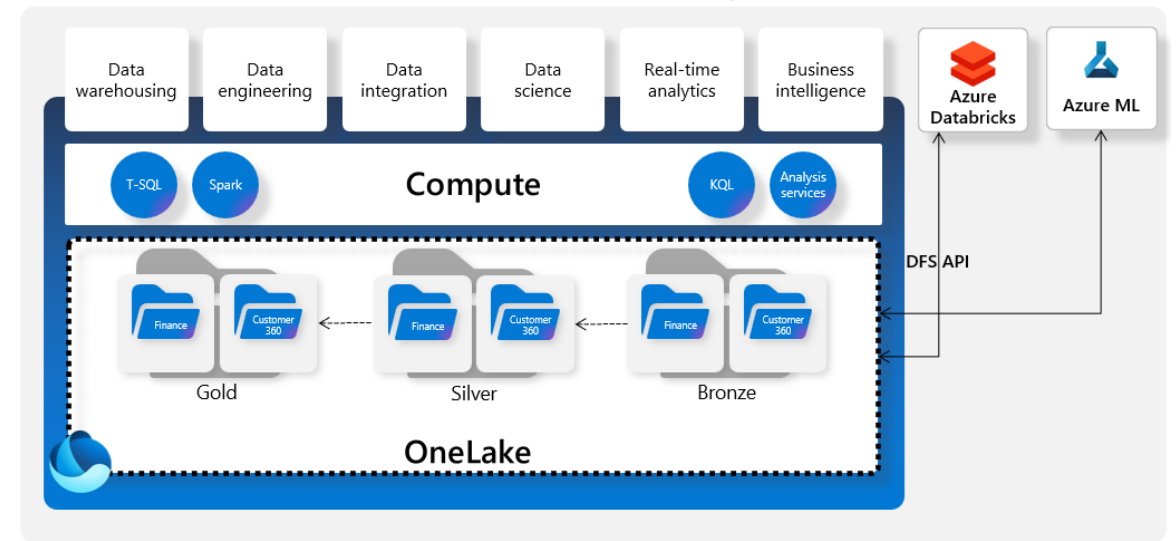


# OneLake Scenarios

## Use OneLake with existing data lakes



## Use and land data directly in OneLake



# OneLake Data Hub

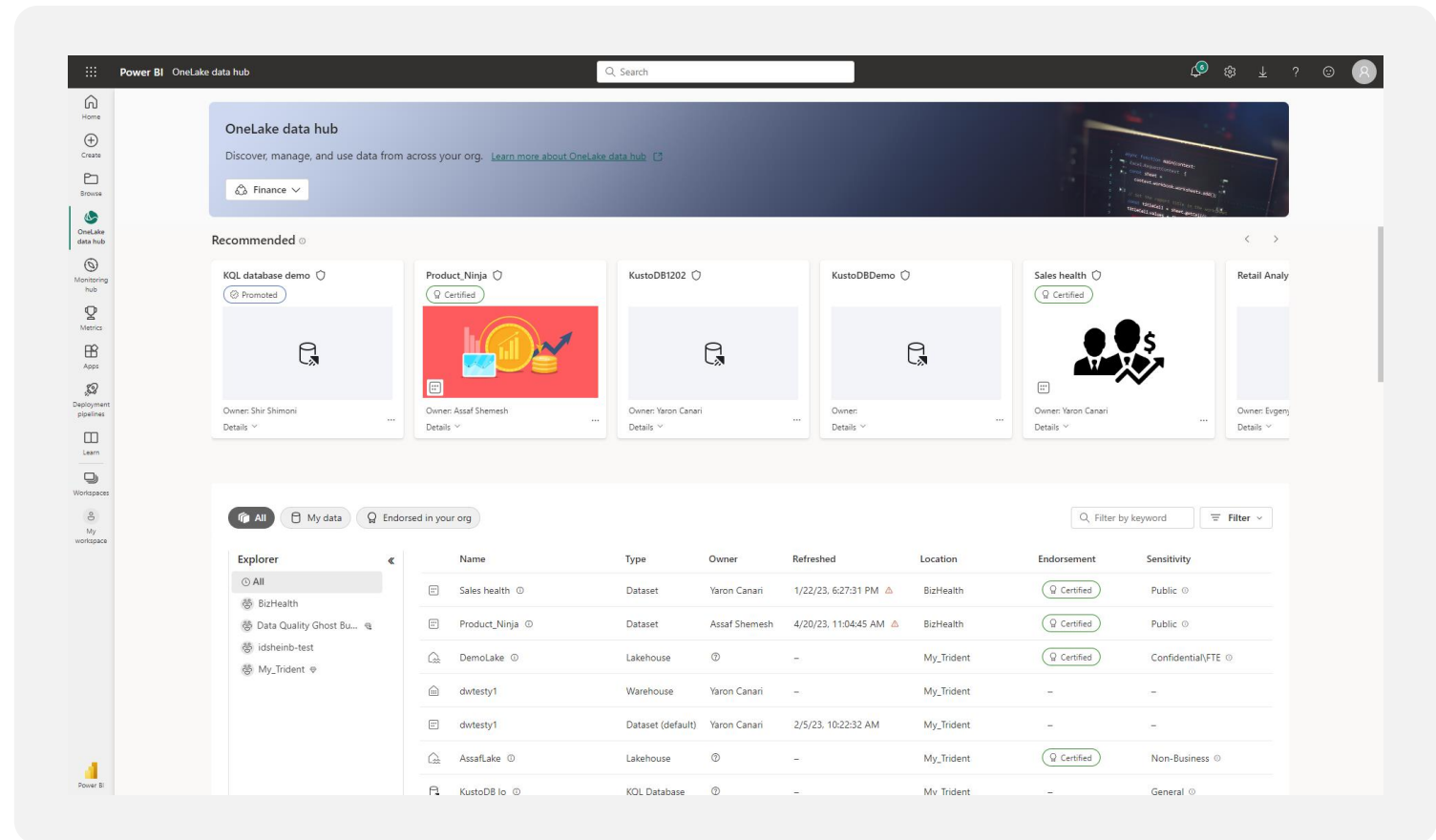
Discover, manage and use data in one place

**Central location** within Fabric to discover, manage, and reuse data

Data can be easily discovered by its **domain** (e.g. Finance) so users can see what matters for them

Efficient data discovery using **search, filter and sort**

**Explorer** capability to easily browse and find data by its folder (workspace) hierarchy



Lakehouse

# Lakehouse

## Data Source



Shortcut Enabled



Structured /  
Unstructured

## Ingestion



Shortcuts



Pipelines &  
Dataflows

## Store



Lakehouse(s)



## Transform

Notebooks &  
Dataflows

## Expose



PBI



Lake Warehouse

# Lakehouse – Lakehouse mode

Lakehouse explorer

wwilakehouse

Tables **Managed**

TestTable ...

Unrecognized 1

Files **Unmanaged**

Folder1

wwi-data

TestTable

	Column1	Column2
1	Team	Subject
2	test3	test4
3	test5	test99

Lakehouse

Explore your data files and folders

SQL endpoint

Query data using SQL

Lakehouse explorer

wwilakehouse

Tables

TestTable

Unrecognized 1

\_mashup\_temporary

Files

Folder1

wwi-data ...

Files > wwi-data

Name
Australia.csv
Canada.csv

Double-click file to view it

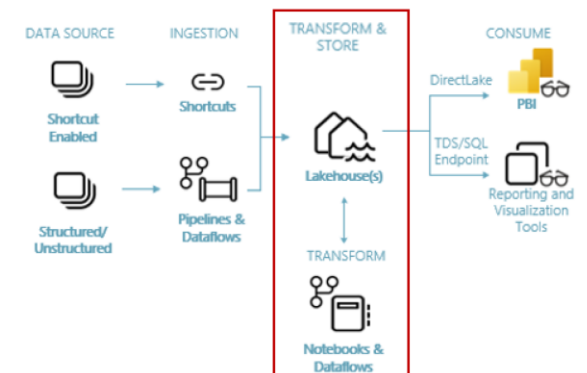
Right-click → Load to Delta table

**Table** - This is a virtual view of the managed area in your lake. This is the main container to host tables of all types (CSV, Parquet, Delta, Managed tables and External tables). All tables, whether automatically or explicitly created, will show up as a table under the managed area of the Lakehouse. This area can also include any types of files or folder/subfolder organizations.



**Files** - This is a virtual view of the unmanaged area in your lake. It can contain any files and folders/subfolder's structure. The main distinction between the managed area and the unmanaged area is the automatic delta table detection process which runs over any folders created in the managed area. Any delta format files (parquet + transaction log) will be automatically registered as a table and will also be available from the serving layer (TSQL)


## Automatic Table Discovery and Registration


Lakehouse Table Automatic discovery and registration is a feature of the lakehouse that provides a fully managed file to table experience for data engineers and data scientists. Users can drop a file into the managed area of the lakehouse and the file will be automatically validated for supported structured formats, *which is currently only Delta tables*, and registered into the metastore with the necessary metadata such as column names, formats, compression and more. Users can then reference the file as a table and use SparkSQL syntax to interact with the data. So don't need to explicitly call CREATE TABLE statement to create tables to use with SQL



# Lakehouse – SQL endpoint mode

	wwilakehouse	Warehouse (default)
	wwilakehouse	Lakehouse

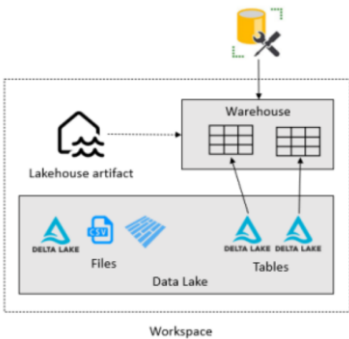
 **Lakehouse**  
Explore your data files and folders

 **SQL endpoint**  
Query data using SQL

NOTE: “Warehouse mode” was renamed “SQL endpoint”

Can query tables (not files).  
Cannot modify data

## SQL Query



**Explorer**

- Warehouses
- WorldWideImporters
  - Schemas
    - dbo
      - Functions
      - StoredProcedures
      - Tables
        - ag...
        - ag...
        - din...
        - din...
        - din...
        - din...
        - dimension\_sto...

**Data preview**

	Date	Calenx
1	2000-11-21T00:00:00.0000000	CY200
2	2000-07-15T00:00:00.0000000	CY200
3	2000-01-25T00:00:00.0000000	CY200
4	2000-11-30T00:00:00.0000000	CY200
5	2000-03-17T00:00:00.0000000	CY200
6	2000-06-20T00:00:00.0000000	CY200
7	2000-11-21T00:00:00.0000000	CY200
8	2000-08-03T00:00:00.0000000	CY200
9	2000-05-08T00:00:00.0000000	CY200
10	2000-03-28T00:00:00.0000000	CY200
11	2000-01-18T00:00:00.0000000	CY200
12	2000-04-22T00:00:00.0000000	CY200
13	2000-02-17T00:00:00.0000000	CY200
14	2000-10-07T00:00:00.0000000	CY200
15	2000-05-17T00:00:00.0000000	CY200
16		

**Context Menu:**

- New SQL query
- Select TOP 100 rows
- Hide in report view
- New measure
- Unhide all
- Remove from BI model

## Visual Query

**Visual Query Tools:**

- Choose columns
- Remove columns
- Keep rows
- Filter rows
- Sort
- Group by
- Replace values
- Merge queries
- Append queries

**Visual Query Diagram:**

The diagram shows a flow from a 'Source' (gold\_aggregate\_sale\_by\_date\_city) to a 'table' (Merged queries), then to 'Expanded silver...', and finally to 'Grouped rows'.

**Home**

**SQL endpoint**

**Explorer**

- Warehouses
- DemoLakehouse
  - Schemas
    - dbo
      - Functions
      - StoredProcedures
      - Tables
        - canada
        - customer
        - customer2
  - Views
  - guest
  - INFORMATION\_SCHEMA
  - sys
  - Security
- Queries
  - My Queries

**SQL query 2**

```
SELECT TOP (100) [Country]
, [Date]
, [ProductID]
, [Revenue]
, [Units]
, [Zip]
FROM [DemoLakehouse].[dbo].[canada]
```

**Results**

	Country	Date	ProductID	Revenue
1	Canada	2017-02-26T00:00:00.0000000	1982	1763.68!
2	Canada	2016-07-10T00:00:00.0000000	2054	2729.73!
3	Canada	2018-12-29T00:00:00.0000000	1935	787.185
4	Canada	2018-01-21T00:00:00.0000000	862	1249.13!
5	Canada	2017-03-27T00:00:00.0000000	1006	634.882!
6	Canada	2018-12-29T00:00:00.0000000	1934	787.185
7	Canada	2017-02-27T00:00:00.0000000	764	1826.68!
8	Canada	2017-02-27T00:00:00.0000000	763	1826.68!
9	Canada	2019-07-20T00:00:00.0000000	862	1264.88!

# Lakehouse – shortcuts (to lakehouse)

Lakehouse explorer

wwilakehouse

Tables

australia

TestTable

Unidentified

\_mashup\_temporary

testfolder1

Files

Folder1

book-recommendation

fraud-detection

wwi-data

Test

Files > wwi-data

Name

Australia.csv

Canada.csv

Refresh

New shortcut

New subfolder

Upload

Rename

Delete

Properties

## New shortcut

Use shortcuts to quickly pull data from internal and external locations into your lakehouses, warehouses, or datasets. Shortcuts can be updated or removed from your item, but these changes will not affect the original data and its source.

### Internal sources

Microsoft OneLake

Fabric

### External sources

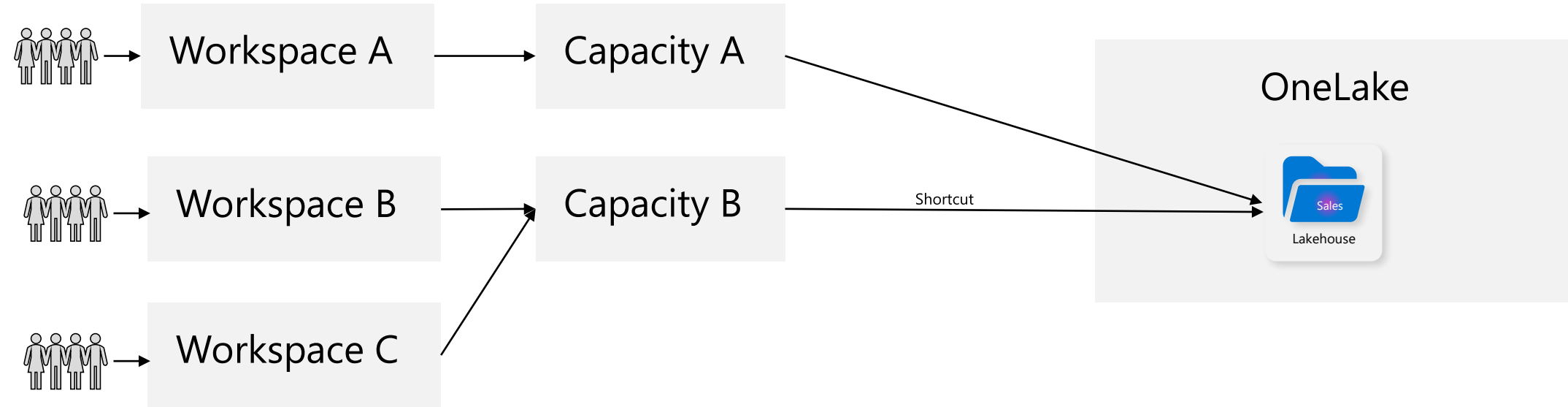
Azure Data Lake Storage Gen2

Azure

Amazon S3

File

# Workspaces and capacities accessing OneLake

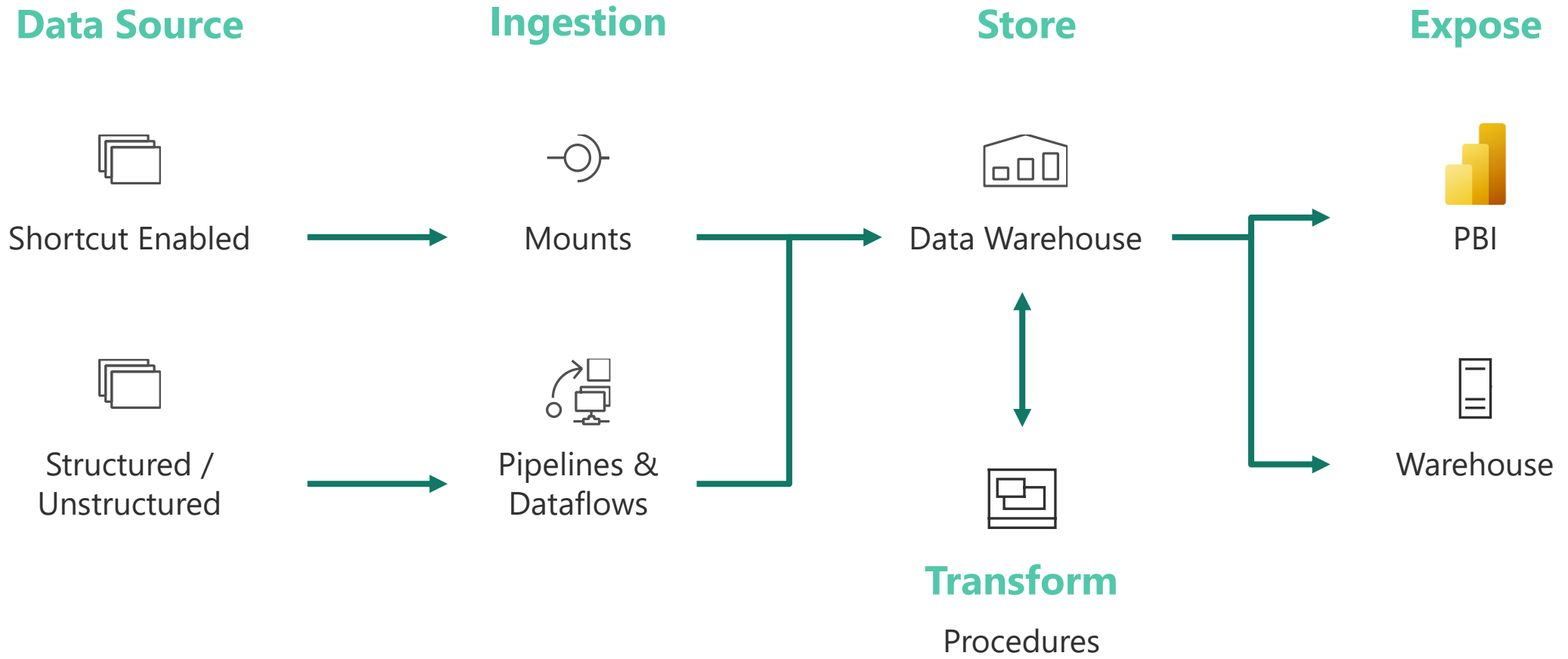


Each tenant will have only one OneLake, and any tenant can access files in a OneLake from other tenants via shortcuts

Demo

Data Warehouse

# Data warehouse

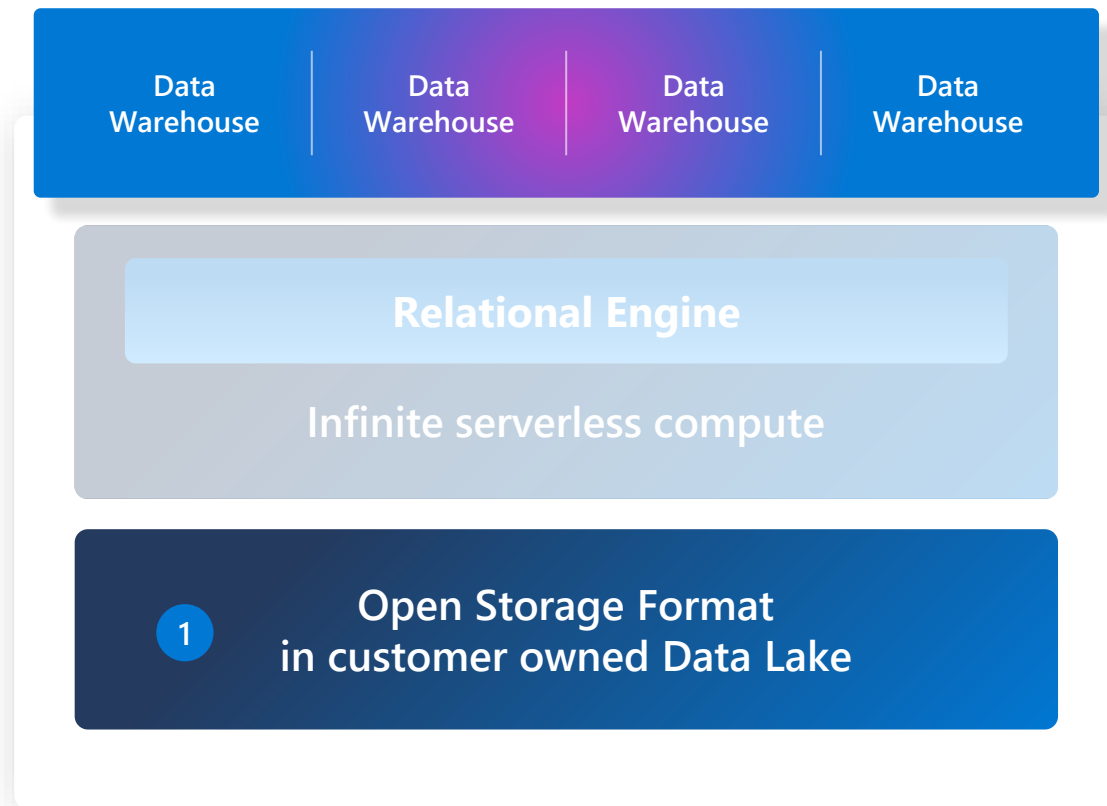


# Synapse Data Warehouse

Infinitely scalable and open



## Synapse Data Warehouse in Fabric



### 1 Open standard format in an open data lake replaces proprietary formats as the native storage

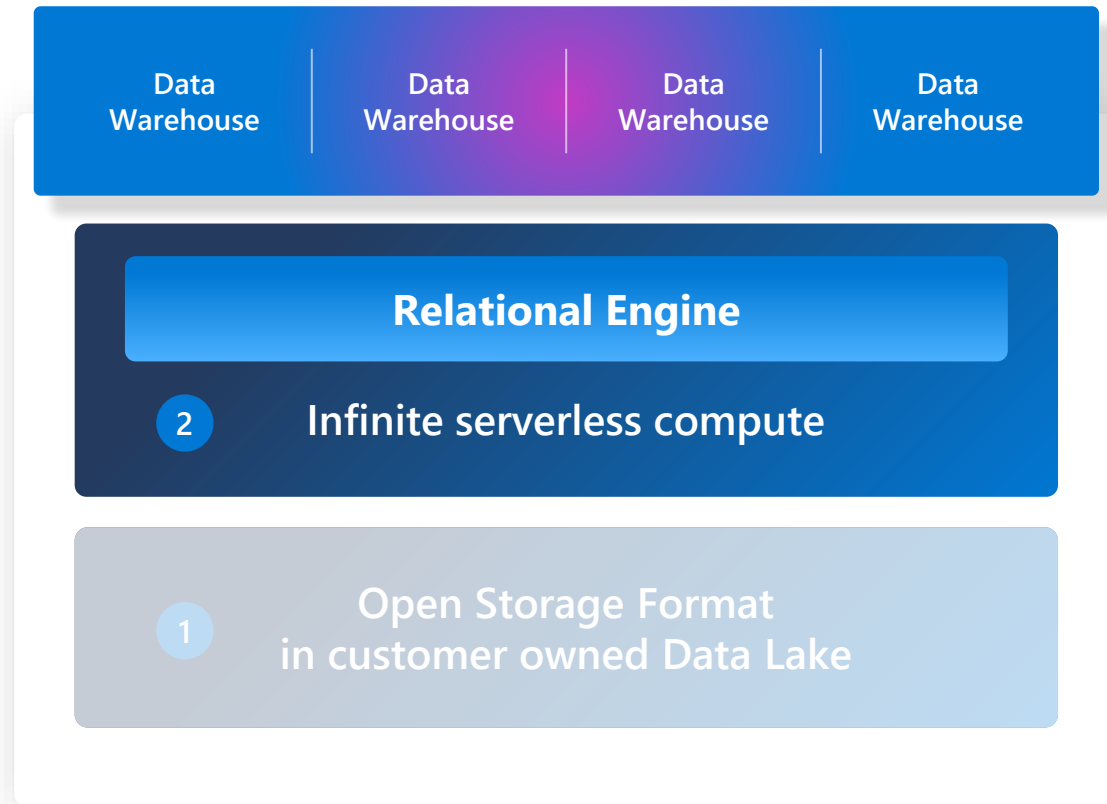
- First transactional data warehouse natively embracing an open standard format
- Data is stored in Delta – Parquet with no vendor lock-in
- Is auto-integrated and auto-optimized with minimal knobs
- Extends full SQL ecosystem benefits

# Synapse Data Warehouse

Infinitely scalable and open



## Synapse Data Warehouse in Fabric



## 2 Dedicated clusters are replaced by serverless compute infrastructure

- Physical compute resources assigned within milliseconds to jobs
- Infinite scaling with dynamic resource allocation tailored to data volume and query complexity
- Instant scaling up/down with no physical provisioning involved
- Resource pooling providing significant efficiencies and pricing

Available now

## Virtual warehouses and cross database querying

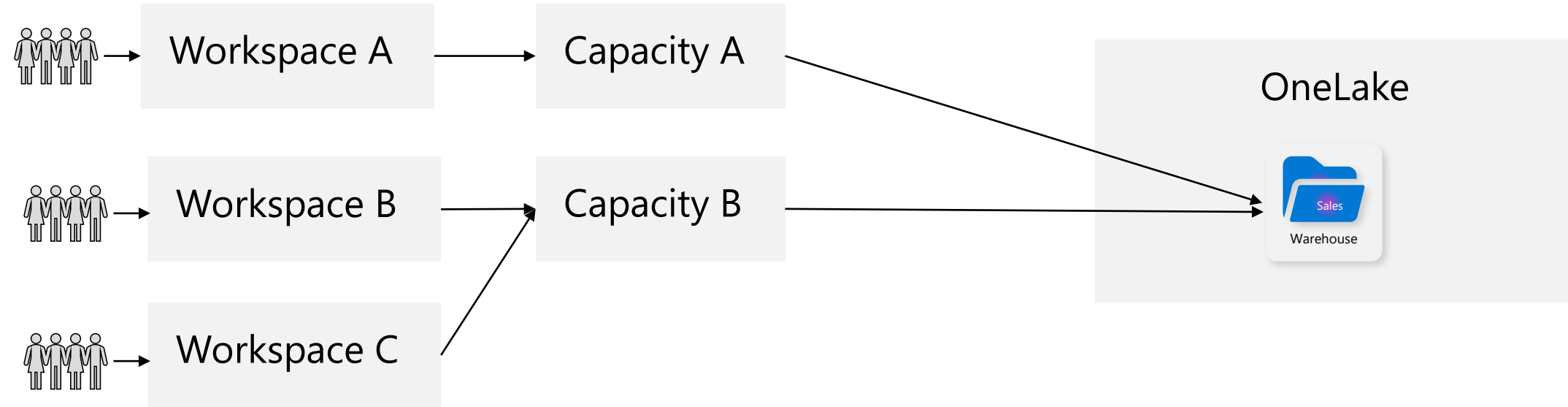
- Build virtual warehouses by creating Lakehouses with shortcuts to data in the Lake
- Query across Warehouse and Lakehouse SQL Endpoint with zero data movement
- Write T-SQL query with 3-part-naming for cross-joining

The screenshot displays the Databricks workspace interface. On the left, the 'Explorer' sidebar is visible, with a red box highlighting the 'Warehouses' section. The main area shows a SQL query in the 'SQL query DSQL...' editor. Below the query, the 'Results' tab displays a table with 4 columns: d\_year, brand\_id, brand, and sum\_agg. The table contains 13 rows of data. The status bar at the bottom indicates 'Succeeded (2 sec 451 ms)' and 'Columns: 4 Rows: 100'.

```
1 SELECT TOP 100 dt.d_year, item.i_brand_id brand_id, item.i_brand brand, SUM(ss_ext_sales_price) sum_agg
2 FROM date_dim dt, store_sales, item
3 WHERE dt.d_date_sk = store_sales.ss_sold_date_sk
4 AND store_sales.ss_item_sk = item.i_item_sk
5 AND item.i_manufact_id = 931
6 AND dt.d_moy=11
7 GROUP BY dt.d_year, item.i_brand, item.i_brand_id
8 ORDER BY dt.d_year, sum_agg desc, brand_id;
```

	d_year	brand_id	brand	sum_agg
1	1998	3002002	importoexporti #2	57072.23
2	1998	2002001	importoimporto #1	55121.24
3	1998	5003002	exportischolar #2	43818.97
4	1998	5003001	exportischolar #1	43427.27
5	1998	5004001	edu packscholar #1	35833.07
6	1998	7005004	scholarbrand #4	35621.42
7	1998	8012006	importomaxi #6	33982.90
8	1998	1004002	edu packamalig #2	32570.71
9	1998	8004005	edu packnameless #5	31966.71
10	1998	3004002	edu packexporti #2	30722.59
11	1998	9012003	importounivamalig #3	28539.16
12	1998	7015007	scholarnameless #7	28369.62
13	1998	7016008	corpnameless #8	27960.94

# Workspaces and capacities accessing OneLake



Each tenant will have only one OneLake, and any tenant can access files in a OneLake from other tenants via shortcuts

# Data Warehouse

WorldWideImporters	Dataset (default)
WorldWideImporters	Warehouse

Use this to build a relational layer on top of the physical data in the Lakehouse and expose it to analysis and reporting tools using T-SQL/TDS end-point.

This offers a transactional data warehouse with T-SQL DML support, stored procedures, tables, and views

How can I control “bad actor” queries?


Fabric compute is designed to automatically classify queries to allocate resources and ensure high priority queries (i.e. ETL, data preparation, and reporting) are not impacted by potentially poorly written ad hoc queries.

How is the classification for an incoming query determined? Queries are intelligently classified by a combination of the source (i.e., pipeline vs. Power BI) and the query type (i.e., INSERT vs. SELECT)

Where is the physical storage for the Data Warehouse? All data for Fabric is stored in OneLake in the open Delta format. A single COPY of the data is therefore exposed to all the compute engines of Fabric without needing to move or duplicate data

The screenshot displays the Microsoft Fabric user interface. On the left is a vertical sidebar with icons for Home, Create, Browse, Data hub, Metrics, Apps, Deployment pipelines, Learn, Workspaces, and SerraTrident Training. The main area is divided into two panes. The 'Explorer' pane on the left shows a tree structure under 'Schemas' > 'dbo'. It includes 'Functions', 'StoredProcedures' (with 'populate\_aggregate\_...' entries), 'Tables' (with 'aggregate\_sale\_b...' and other dimension tables like 'dimension\_city', 'dimension\_custo...', 'dimension\_date', 'dimension\_empl...', 'dimension\_stock...', and 'fact\_sale'), and 'Views'. The 'Queries' section shows 'Sales by Country StatePro...'. The 'Data preview' pane on the right shows a table with 29 rows and 4 columns: 'Date', 'CalendarMonthLabel', 'Day', and 'ShortMonth'. The table contains date and month data for various years. At the bottom, there's a 'Power BI' logo and a navigation bar with 'Data', 'Query', and 'Model' tabs.

# Access via other tools

 DemoWarehouse  
Warehouse

About

Sensitivity Label

Name

DemoWarehouse

Description

Describe this item


Details

Owned by James Serra  
Last modified by James Serra, 10:20 AM, 6/13/23

SQL connection string

Copy this string and use it to connect externally to the item from Power BI desktop or client tools.

grtn7k2lwpce7o44gm5hvhip4a-2drnpq3o6zeulgqiebkfq7gu3e.datawarehouse.pbidedicated.window

 DemoLakehouse  
Lakehouse

About

Endorsement

SQL endpoint

SQL connection string

Copy this string and use it to connect externally to the item from Power BI desktop or client tools.

grtn7k2lwpce7o44gm5hvhip4a-2drnpq3o6zeulgqiebkfq7gu3e.datawarehouse.pbidedic...

New Server Registration

General Connection Properties Always Encrypted Additional Connection Parameters

Login

Type the server name, or choose it from the drop-down list.

Server type: Database Engine

Server name: grtn7k2lwpce7o44gm5hvhip4a-2drnpq3o6zeulgqiebkfq7gu3e.datawarehouse.pbidedicated.window

Authentication: Azure Active Directory - Universal with MFA

User name:

Microsoft SQL Server Management Studio

File Edit View Tools Window Help

New Query

Execute

Object Explorer

Connect

grtn7k2lwpce7o44gm5hvhip4a-2drnpq3o6zeulgqiebkfq7gu3e.datawarehouse.pbidedicated.window

Databases

System Databases

Bronze

DataflowsStagingLakehouse

DemoLakehouse

DemoWarehouse

DemoWarehouse2

Security

Demo

# Microsoft Fabric



Synapse Data  
Engineering

Use Spark Notebooks



Python



R



Scala



Write data into  
Lakehouse tables



Synapse Data  
Warehousing



Use SQL Queries &  
Stored Procedures



Full T-SQL support



Write data into  
Warehouse tables

# Data warehouse and lakehouse properties

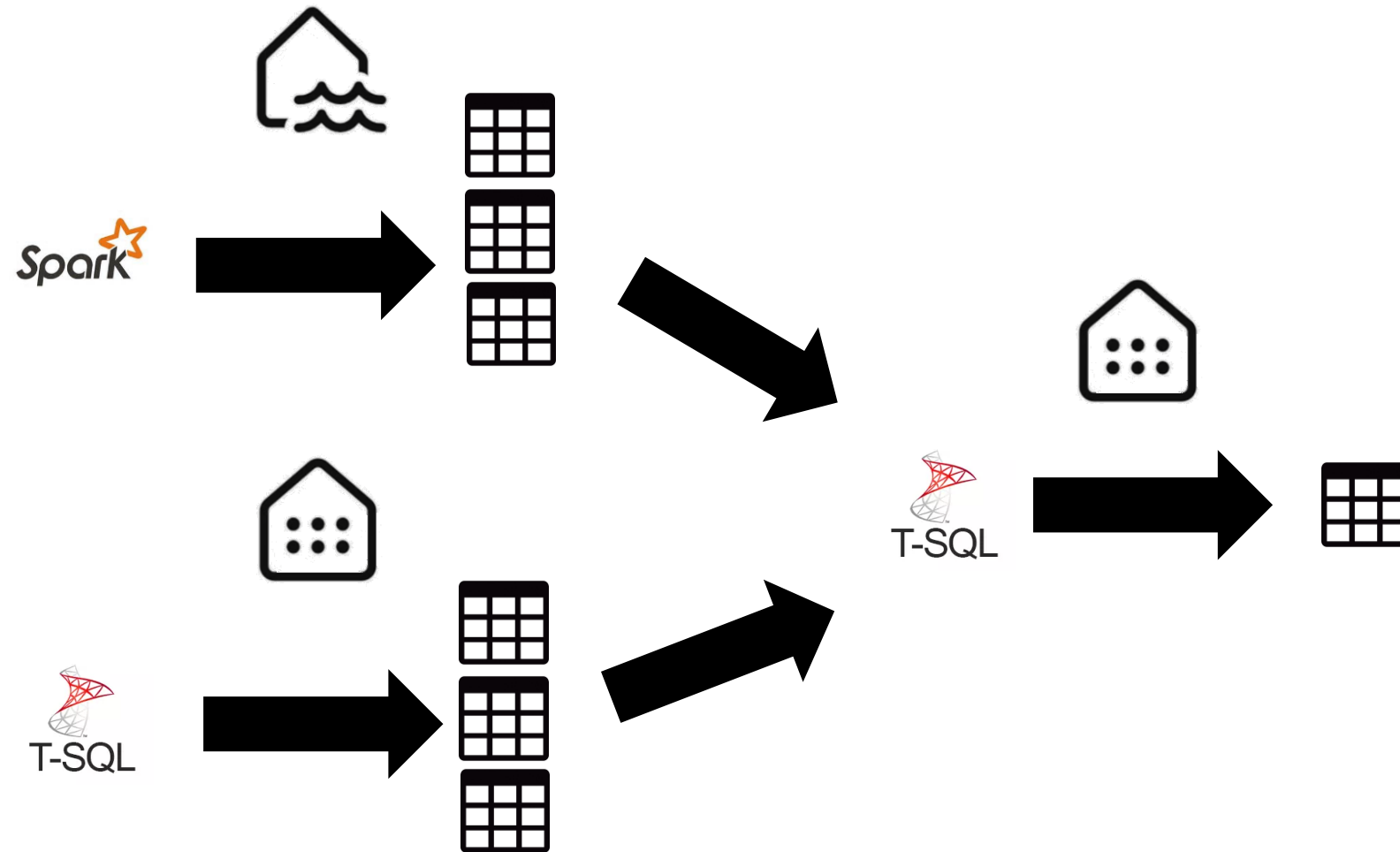
	Data warehouse	Lakehouse	Power BI Datamart
Data volume	Unlimited	Unlimited	Up to 100 GB
Type of data	Structured	Unstructured, semi-structured, structured	Structured
Primary developer persona	Data warehouse developer, SQL engineer	Data engineer, data scientist	Citizen developer
Primary developer skill set	SQL	Spark (Scala, PySpark, Spark SQL, R)	No code, SQL
Data organized by	Databases, schemas, and tables	Folders and files, databases and tables	Database, tables, queries
Read operations	Spark, T-SQL	Spark, T-SQL	Spark, T-SQL, Power BI
Write operations	T-SQL	Spark (Scala, PySpark, Spark SQL, R)	Dataflows, T-SQL
Multi-table transactions	Yes	No	No
Primary development interface	SQL scripts	Spark notebooks, Spark job definitions	Power BI
Security	Object level (table, view, function, stored procedure, etc.), column level, row level, DDL/DML	Row level, table level (when using T-SQL), none for Spark	Built-in RLS editor
Access data via shortcuts	Yes (indirectly through the lakehouse)	Yes	No
Can be a source for shortcuts	Yes (tables)	Yes (files and tables)	No
Query across items	Yes, query across lakehouse and warehouse tables	Yes, query across lakehouse and warehouse tables; query across lakehouses (including shortcuts using Spark)	No

Why two options?

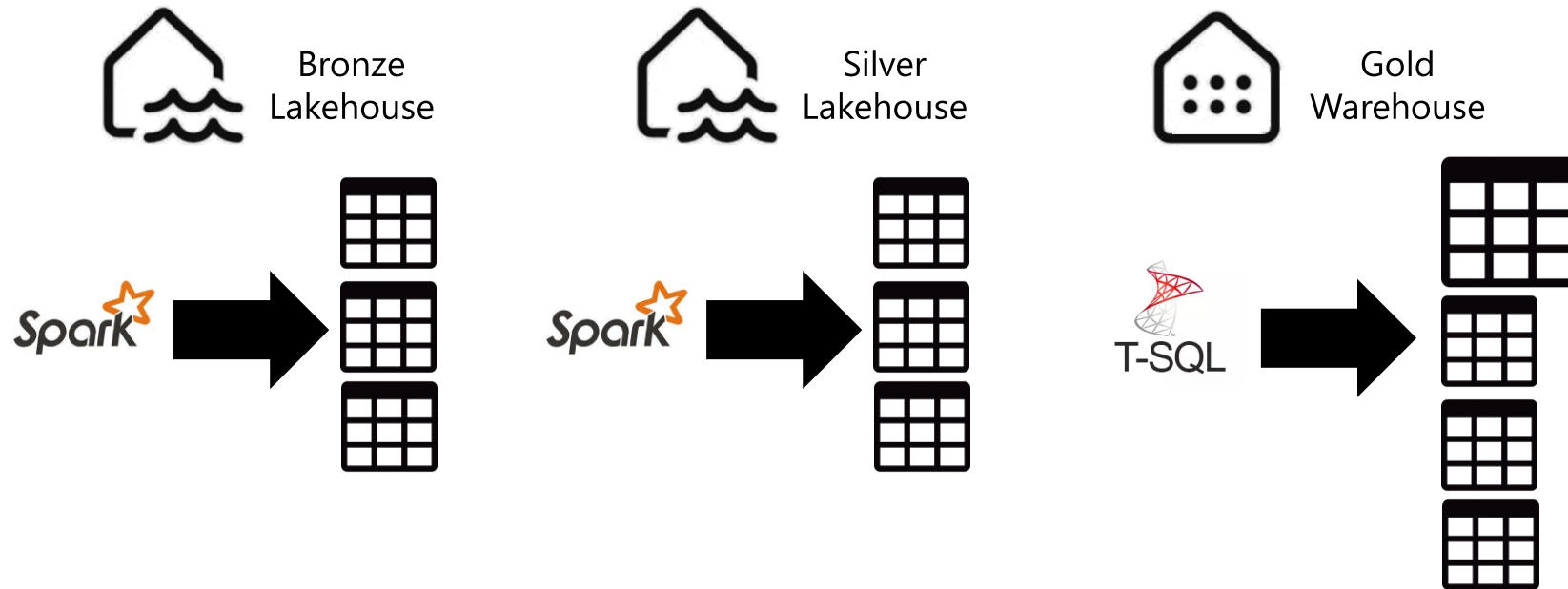
- Delta lake shortcomings:
- No multi-table transactions
  - Lack of full T-SQL support (no updates, limited reads)
  - Performance problem for trickle transactions



# Microsoft Fabric



# Microsoft Fabric



ADF

# ADF Review

Pipelines

● pipeline1

Activities

Search activities

Move & transform

Copy data

Data flow

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

Power Query

✓ Validate

✓ Validate copy runtime

▶ Debug

⚡ Add trigger

Copy data

Copy data1

Data flow

Data flow1

Power Query

Power Query1

General

Source

Sink

Mapping

Settings

User properties

Name

Copy data1

Description

Timeout

0.12:00:00

Retry

0

Retry interval (sec)

30

Secure output

Secure input

Synapse: Pipelines

## Mapping data flows

Data flows

● dataflow1

dataflow1

✓ Validate

⏸ Data flow debug

source1

Add source dataset

join1

Add second stream to the join from settings

compose1

Columns: 0 total

Add Source

Flowlet settings

Parameters

Inspect

Data preview

Output stream name

compose1

Description

Choose a flowlet to execute.

Flowlet

Select...

Synapse: Data flows

## Wrangling data flows

Power Query

● powerquery1

Standard View

Diagram View

Power BI: Dataflows

Synapse: No

Power BI: Dataflows  
Synapse: No

# ADF Review

Pipelines

● pipeline1

Activities

Search activities

Move & transform

Copy data

Data flow

Synapse

Azure Data Explorer

Azure Function

Batch Service

Databricks

Data Lake Analytics

General

HDInsight

Iteration & conditionals

Machine Learning

Power Query

Power Query

Copy data

Copy data1

Data flow

Data flow1

Power Query

General

Source

Sink

Mapping

Settings

User properties

Name

Copy data1

Description

Timeout

0.12:00:00

Retry

0

Retry interval (sec)

30

Secure output

Secure input

Don't Exist

Synapse: Pipelines

## Mapping data flows

Data flows

● dataflow1

dataflow1

Validate

Data flow debug

source1

Add source dataset

join1

Add second stream to the join from settings

compose1

Columns: 0 total

Add Source

Flowlet settings

Parameters

Inspect

Data preview

Output stream name

compose1

Description

Choose a flowlet to execute.

Flowlet

Select...

Synapse: Data flows

## Wrangling data flows

Power Query

● powerquery1

Standard View

Diagram View

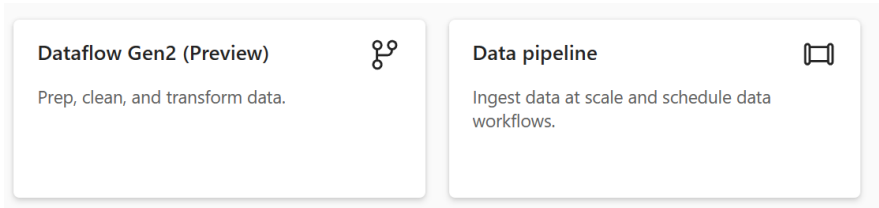
Power BI: Dataflows

Synapse: No

Don't Exist

Dataflow Gen1

# Data Factory in Fabric



## ADF: Power Query

PQ UI with the power of ADF (think of it as the next version of ADF PQ). Scale is still Excel/PBI scale, not yet ADF cloud scale

## ADF Pipelines

New interface, but basically same as ADF

- ADF Data flows do not exist in Fabric
- Power Query is now called Dataflow Gen2 (which helps in that Power Query does more than just query). Scalable
- Power BI Dataflows are now called Dataflows Gen1
- Mounting option available to use ADF mapping data flows in Fabric (no option for Synapse yet). Can then do changes in Fabric (but not in ADF)

## What is Dataflows Gen2?

This is the new generation of Dataflows Gen1. Dataflows provide a low-code interface for ingesting data from 100s of data sources, transforming your data using 300+ data transformations and loading the resulting data into multiple destinations such as Azure SQL Databases, Lakehouse, and more

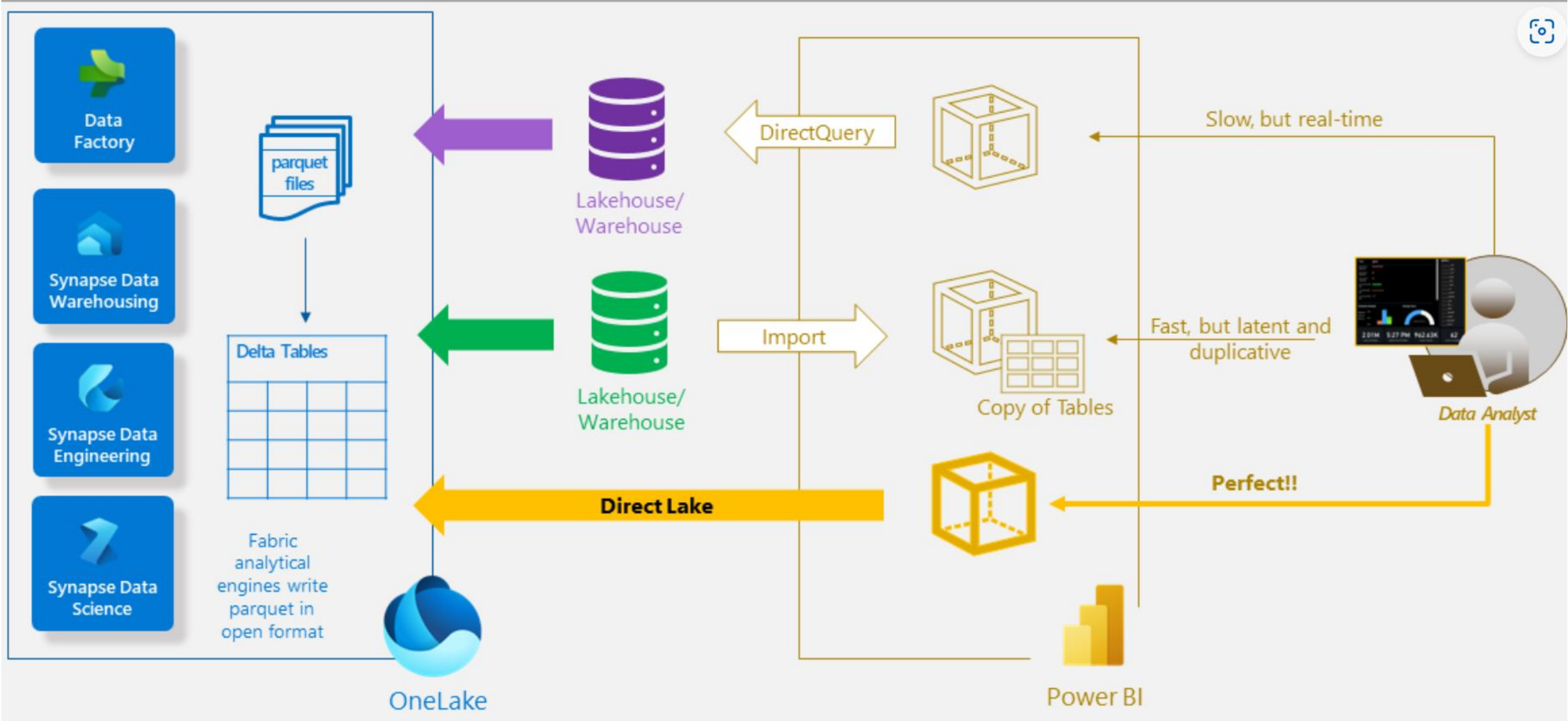
## We currently have multiple Dataflows experiences with Power BI Dataflows Gen1, Power Query Dataflows and ADF Data flows. What is the strategy with Fabric with these various experiences?

Our goal is to evolve over time with a single Dataflow that combines the ease of use of PBI, Power Query and the scale of ADF

## What is Fabric Pipelines?

Fabric pipelines enable powerful workflow capabilities at cloud-scale. With data pipelines, you can build complex workflows that can refresh your dataflow, move PB-size data, and define sophisticated control flow pipelines. Use data pipelines to build complex ETL and Data factory workflows that can perform a number of different tasks at scale. Control flow capabilities are built into pipelines that will allow you to build workflow logic which provides loops and conditional.

Power BI / DirectLake



# SeeThrough/DirectLake

- It's an “on-demand” import model – loading into memory the columns in tables which are needed for the query
- There is an initial slower cold cache performance
- There is a fall-back mechanism to Direct Query mode in certain scenarios
  - When the data cannot be loaded into memory
  - query is too complex
  - PBI Desktop is used for authoring
  - RLS is used
- We can also look at it as a Direct Query mode to the lake with some import cache – combining best of both worlds
- Dataset is tied to the version of the delta file when the dataset was created. Refresh is needed to pick up the new version. At GA this will be configurable behavior – show the latest or need to refresh

**For best performance you should compress the data using the VORDER compression method (50%-70% more compression). Stored this way by ADF by default**

Benefits:

- No more scheduled imports

# Should I use Fabric now?

- Yes, for prototyping
- Yes, if you won't be in production for several months
- You have to be OK with bugs, missing features, and possible performance issues
- Don't use if have hundreds of terabytes

# If building in Synapse, how to make transition to Fabric smooth?

- Do not use dedicated pools, unless needed for serving and performance
- Don't use any stored procedures to modify data in dedicated pools
- Use ADF for pipelines and for PowerQuery, and don't use ADF mapping data flows. Don't use Synapse pipelines or mapping data flows
- Embrace the data lakehouse architecture

# Resources

Microsoft Fabric webinar series: <https://aka.ms/fabric-webinar-series>

New documentation: <https://aka.ms/fabric-docs>. Check out the [tutorials](#).

[Data Mesh](#), [Data Fabric](#), [Data Lakehouse](#) – (video from Toronto Data Professional Community on 2/15/23)

Build videos:

[Build 2-day demos](#)

[Microsoft Fabric Synapse data warehouse, Q&A](#)

My [intro blog](#) on Microsoft Fabric (with helpful links at the bottom)

[Fabric notes](#)

[Advancing Analytics videos](#)

[Ask me Anything \(AMA\) about Microsoft Fabric!](#)

Q & A



James Serra, Microsoft, Industry Advisor

Email me at: [jamesserra3@gmail.com](mailto:jamesserra3@gmail.com)

Follow me at: @JamesSerra

Link to me at: [www.linkedin.com/in/JamesSerra](https://www.linkedin.com/in/JamesSerra)

Visit my blog at: [JamesSerra.com](http://JamesSerra.com)