Microsoft

Microsoft Fabric
Lakehouse vs Warehouse
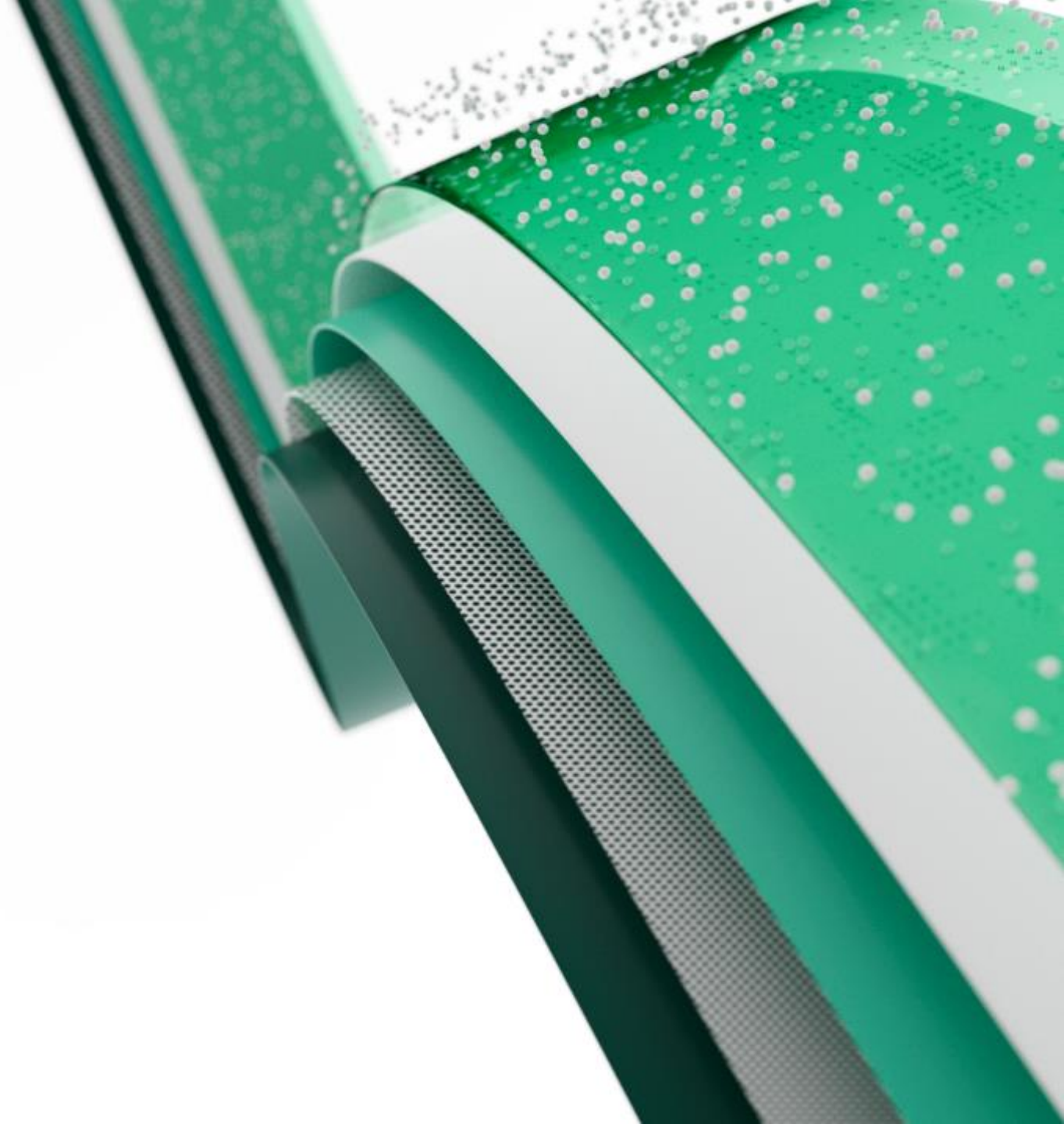
James Serra

Industry Advisor

Microsoft, Federal Civilian

jamesserra3@gmail.com

Blog: JamesSerra.com

7/17/23

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
- Former SQL Server MVP
- Author of book "Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh"
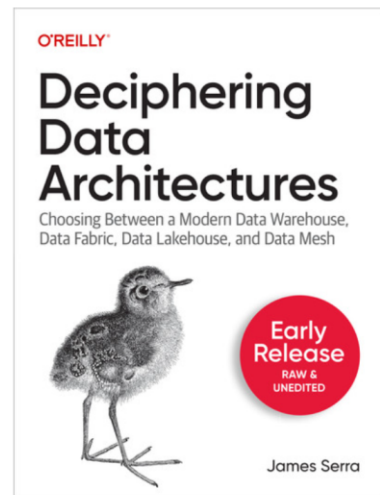
My upcoming book

Deciphering Data Architectures

Write the first review
By **James Serra**

O'REILLY®
Deciphering Data Architectures
Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh

Early Release RAW & UNEDITED

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

Five chapters available now:
Deciphering Data Architectures (oreilly.com)

- Foundation
    - 1. Big Data (available)
    - 2. Types of Data Architectures (available)
    - 3. The Architecture Design Session (available)
- Common data architecture concepts
    - 4. Relational Data Warehouse (available)
    - 5. Data Lake (available)
    - 6. Approaches to Data Stores
    - 7. Approaches to Design
    - 8. Approaches to Data Modeling
    - 9. Approaches to Data Ingestion
- Data Architectures
    - 10. Modern Data Warehouse (MDW)
    - 11. Data Fabric
    - 12. Data Lakehouse
    - 13. Data Mesh Foundation
    - 14. Data Mesh Adoption
- People, Process, and Technology
    - 15. People and process
    - 16. Technologies

What is the difference between lakehouse and warehouse?

Lakehouse

Lakehouse

Data Source   **Ingestion**   **Store**   **Expose**

Shortcut Enabled → Shortcuts → Lakehouse(s) → PBI

Structured / Unstructured → Pipelines & Dataflows → Lake Warehouse

Transform

Notebooks & Dataflows

OneLake -> Workspace -> Lakehouse

https://onelake.dfs.fabric.microsoft.com/SerraFabricDemo/DemoLakehouse.Lakehouse/Tables
https://onelake.dfs.fabric.microsoft.com/SerraFabricDemo/DemoLakehouse.Lakehouse/Files

Lakehouse – Lakehouse mode

Lakehouse explorer

- ∨ wwilakehouse
  - ∨ 📁 Tables Managed
    - > ⊞ TestTable ...
    - > 📁 Unrecognized ①
  - ∨ 📁 Files Unmanaged
    - > 📁 Folder1
    - ∨ 📁 wwi-data

Right click –> View table files

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
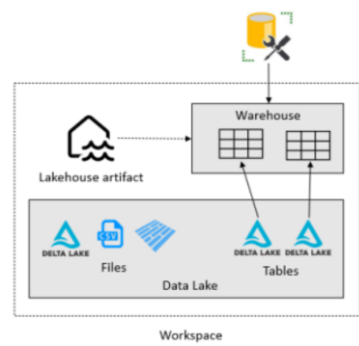
Lakehouse – SQL endpoint mode

Can query tables (not files).
Cannot modify data

SQL Query

Visual Query

Lakehouse – shortcuts (to lakehouse)

Lakehouse – shortcuts (to lakehouse)…cont

Select a data source type   ✕

⚠ When accessing this shortcut using a dataset or T-SQL, the identity of the calling item's owner is used to authorize access rather than the user's identity.

ⓘ **wwilakehouse** is located in the region **West Central US**. Any data sourced through this shortcut will be processed in the same region.

Find and connect to the data you want to use with your shortcut.

| | 🔲 All | 🗄 My data | 🏅 Endorsed in your org | | | 🔍 Filter by keyword | ☰ Filter ⌄ |

	Name	Type	Capacity region	Owner	Location	Endorsement	Sensitivity
🏠	wwilakehouse	Lakehouse	West Central US	James Serra	SerraTridentTraining	–	Confidential\Micro... ⓘ
🏠	TestWarehouse	Warehouse	West Central US	James Serra	SerraTridentTraining	–	General ⓘ
🗄	rta-tutorial-db ⓘ	KQL Database	West Central US	James Serra	SerraTridentTraining	–	Confidential\Micro... ⓘ
🏠	WorldWideImporters	Warehouse	West Central US	James Serra	SerraTridentTraining	–	Confidential\Micro... ⓘ
🏠	serratestlakehouse1	Lakehouse	West Central US	James Serra	TridentSerra	–	Confidential\Micro... ⓘ
🏠	SerraTridentLakehouse	Lakehouse	West Central US	James Serra	TridentSerra	–	Confidential\Micro... ⓘ
🏠	test1dw	Warehouse	West Central US	James Serra	TridentSerra	–	–

Explorer

| Previous | | | | Next | Cancel |

New shortcut

⚠ When accessing this shortcut using a dataset or T-SQL, the identity of th

ⓘ **wwilakehouse** is located in the region **West Central US**. Any data sourc

Find and connect to the data you want to use with your shortcut.

◥ **OneLake**

⌄   serratestlakehouse1

　　>   Tables

　　⌄   Files

　　　　> ⦿ 🗂 testfolder1

Lakehouse – shortcuts (to Data Warehouse)

You can create shortcuts to your Data Warehouse. This allows you to access tables in your Data Warehouse directly from your Lakehouse without having to orchestrate any data movement.

Home

Get data

New data pipeline

New Dataflow Gen2

New shortcut

Lakehous

wwilak

Tables
> australia
> TestTable
> Unidentified   2

Files
> Folder1

New shortcut

Use shortcuts to quickly pull data from internal and exter not affect the original data and its source.

Internal sources

Microsoft OneLake
Fabric

Select a data source type

⚠ When accessing this shortcut using a dataset or T-SQL, the identity of the

ⓘ **wwilakehouse** is located in the region **West Central US**. Any data sourced

Find and connect to the data you want to use with your shortcut.

All   My data   Endorsed in your org

Name	Type
TestWarehouse	Warehouse

Lakehouse explorer

wwilakehouse

Tables
> australia
> TestTable
> Unidentified
  > SerraTestTable

New shortcut

⚠ When accessing this shortcut using a dataset o

ⓘ **wwilakehouse** is located in the region **West C**

Find and connect to the data you want to use

OneLake

TestWarehouse
  Tables
    dbo
      > SerraTestTable

Note: Data Warehouse tables are automatically mirrored to OneLake using Delta Lake logs. This function is supported for tables that are Insert only. When a Delete or Update occurs, new data will not be available in OneLake or accessible via the shortcut.

Data Warehouse

Data warehouse

Synapse Data Warehouse

Infinitely scalable and open

Synapse Data Warehouse in Fabric

| Data Warehouse | Data Warehouse | Data Warehouse | Data Warehouse |

Relational Engine

Infinite serverless compute

1 **Open Storage Format**
in customer owned Data Lake
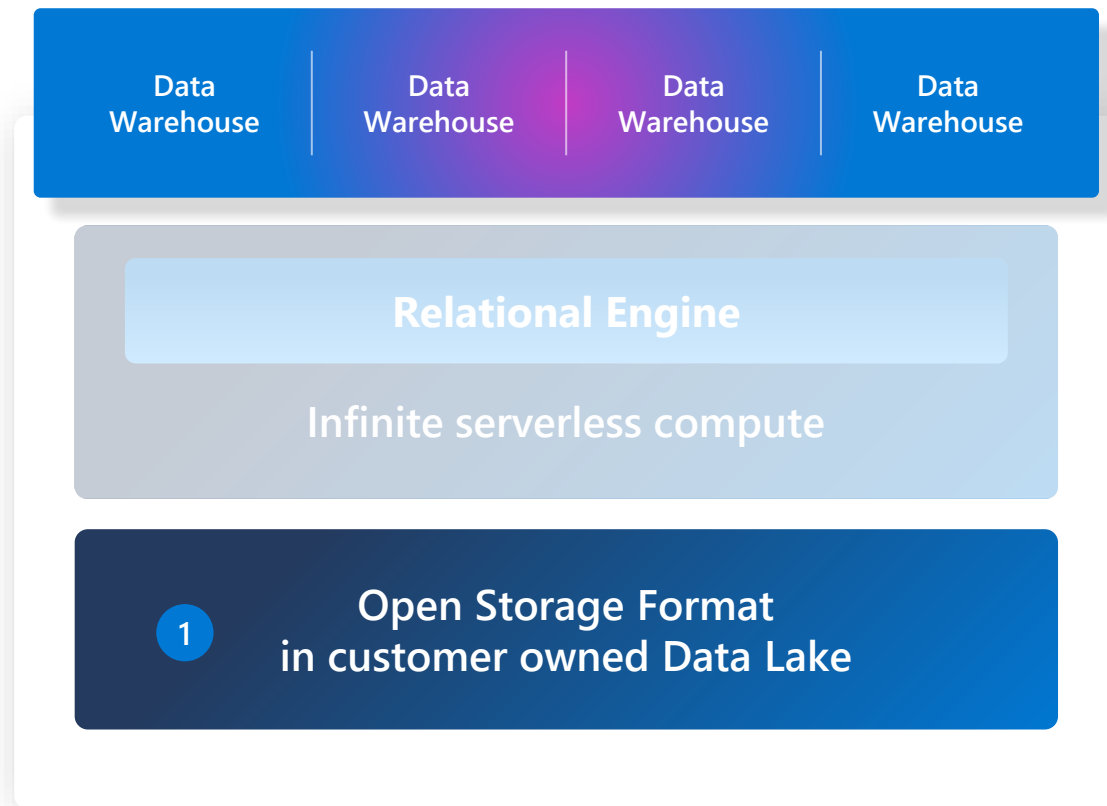
1 **Open standard format in an open data lake replaces proprietary formats as the native storage**

- 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

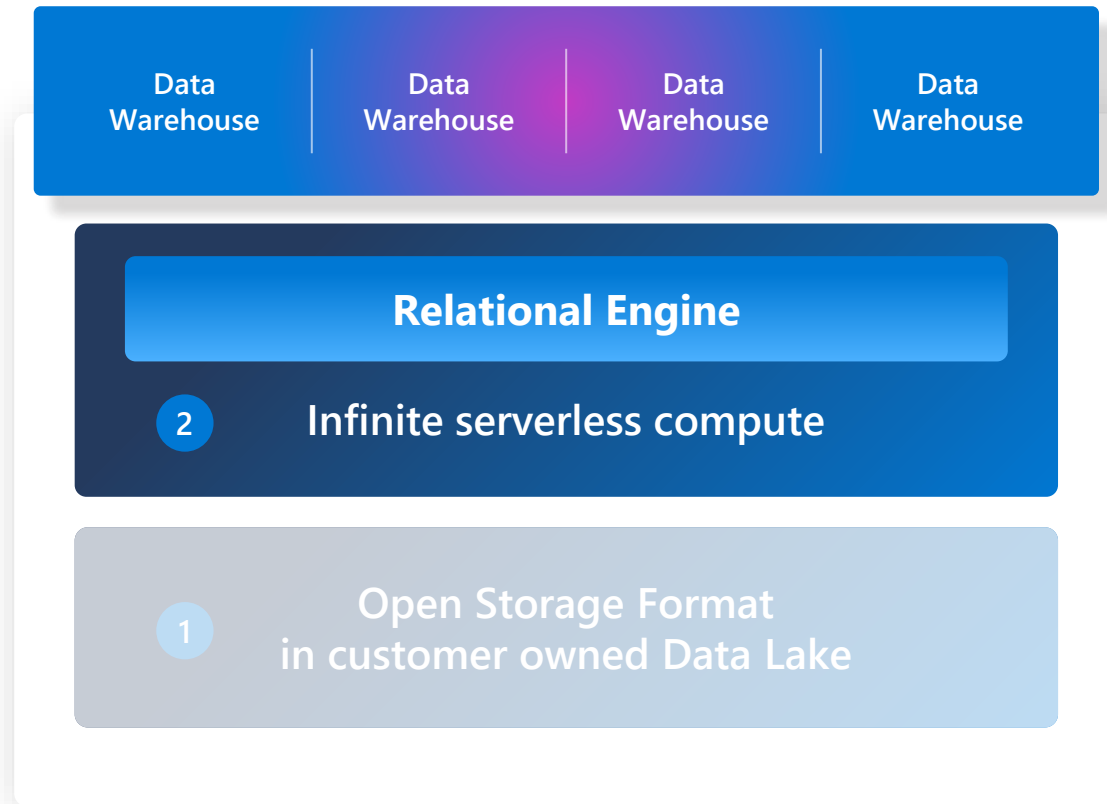- Extends full SQL ecosystem benefits

Synapse Data Warehouse

Infinitely scalable and open

Synapse Data Warehouse in Fabric

| Data Warehouse | Data Warehouse | Data Warehouse | Data Warehouse |

Relational Engine

2 Infinite serverless compute

1 Open Storage Format
in customer owned Data Lake

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

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

Coming June'23

Cloning

- Create zero-copy table clones as of current point-time-time
- Zero-copy Warehouse clones coming in CYQ3'23
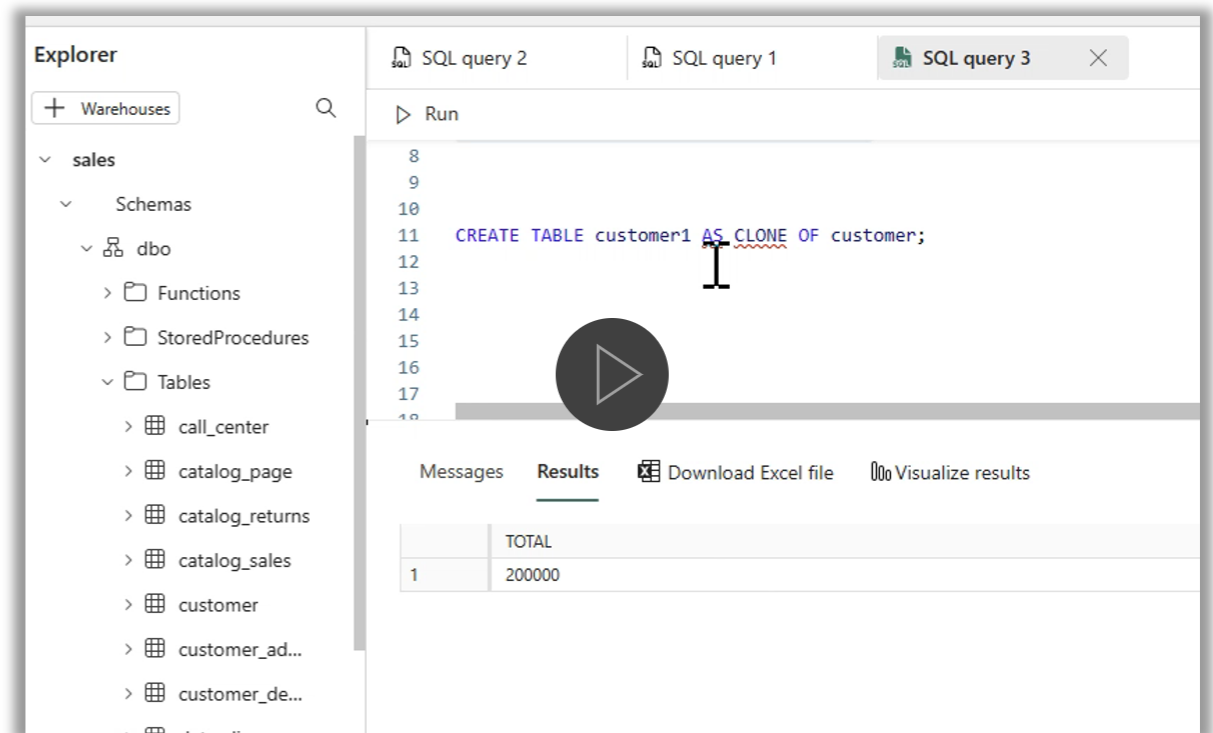
Data Warehouse – query Lakehouse tables

Anything in the tables section of the lakehouse (Delta today, potentially more formats will be available in the future) you can query this using a 3-part name from the data warehouse

Lakehouse vs Warehouse

Microsoft Fabric

Legend:
- = Lakehouse
- = Lakehouse table
- = Warehouse
- = Warehouse table
- = Shortcut
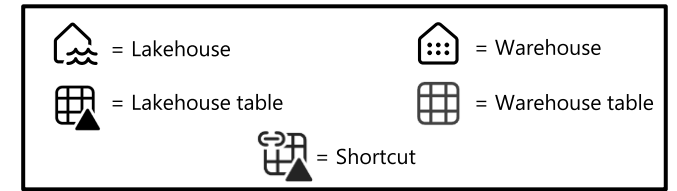
Spark Engine — Synapse Data Engineering

- Use Spark Notebooks
- Python
- R
- Scala
- SQL
- Write data into Lakehouse tables
- Data pipeline
- Dataflow Gen2

Synapse Data Warehousing — SQL Engine

- Use SQL Queries & Stored Procedures
- SSMS
- Azure Data Studio
- Full T-SQL support*
- Write data into Warehouse tables
- Data pipeline
- Dataflow Gen2
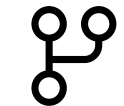
* See limitations at https://learn.microsoft.com/en-us/fabric/data-warehouse/tsql-surface-area

T-SQL limitations – SQL Endpoint or Warehouse

At this time, the following list of commands is NOT currently supported. Don't try to use these commands because even though they may appear to succeed, they could cause issues to your warehouse.

- ALTER TABLE ADD/ALTER/DROP COLUMN
- BULK LOAD
- CREATE ROLE
- CREATE SECURITY POLICY - Row Level Security (RLS)
- CREATE USER
- GRANT/DENY/REVOKE
- Hints
- Identity Columns
- Manually created multi-column stats
- MASK and UNMASK (Dynamic Data Masking)
- MATERIALIZED VIEWS
- MERGE
- OPENROWSET
- PREDICT
- Queries targeting system and user tables
- Recursive queries
- Result Set Caching
- Schema and Table names can't contain / or \
- SELECT - FOR (except JSON)
- SET ROWCOUNT
- SET TRANSACTION ISOLATION LEVEL
- sp_showmemo_xml
- sp_showspaceused
- sp_rename
- Temp Tables
- Triggers
- TRUNCATE

https://learn.microsoft.com/en-us/fabric/data-warehouse/tsql-surface-area

Data warehouse and lakehouse properties

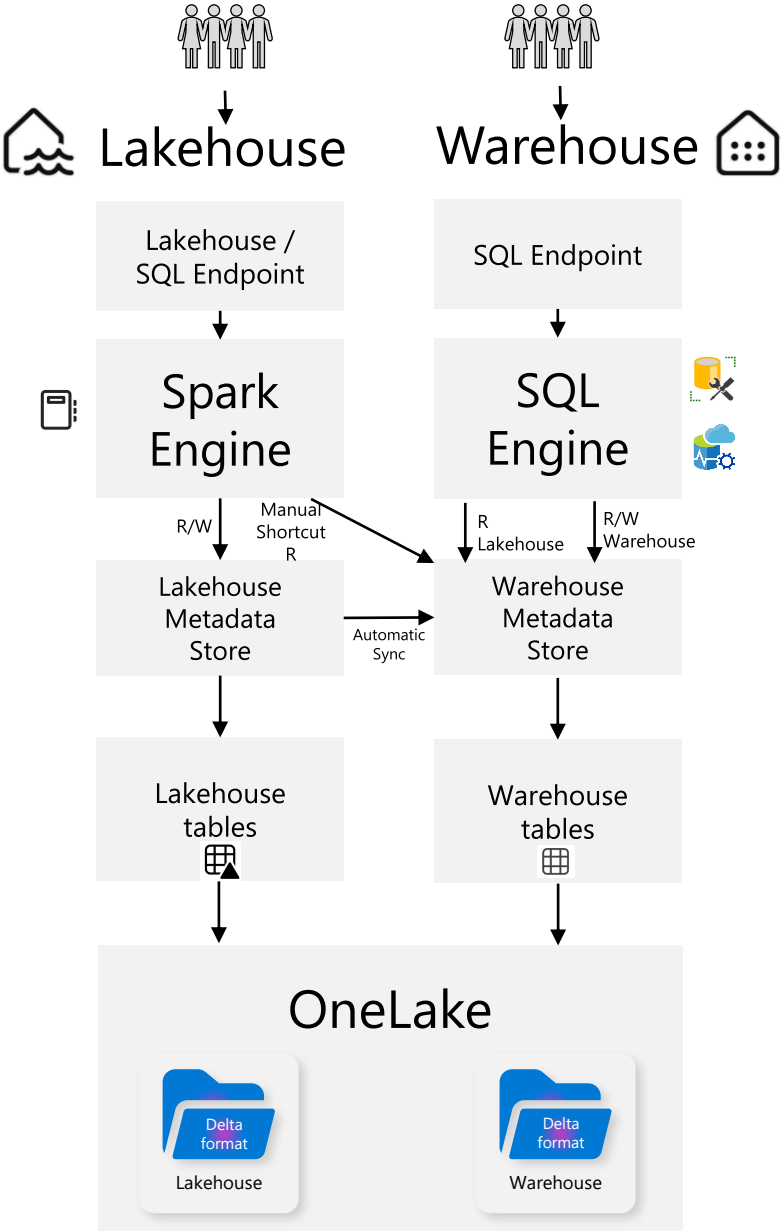|	Data warehouse	Lakehouse	Power BI Datamart
Data volume	Unlimited	Unlimited	Up to 100 GB
Type of data	Structured	Unstructured, semi-structured, structured	Structured
Primary developer persona	Data warehouse developer, SQL engineer	Data engineer, data scientist	Citizen developer
Primary developer skill 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 (can't update data with T-SQL, limited T-SQL support for reads)
- Performance problem for trickle transactions (updates to a few rows in Spark)
- Does not support schemas

https://learn.microsoft.com/en-us/fabric/get-started/decision-guide-warehouse-lakehouse

Warehouse vs Lakehouse SQL Endpoint

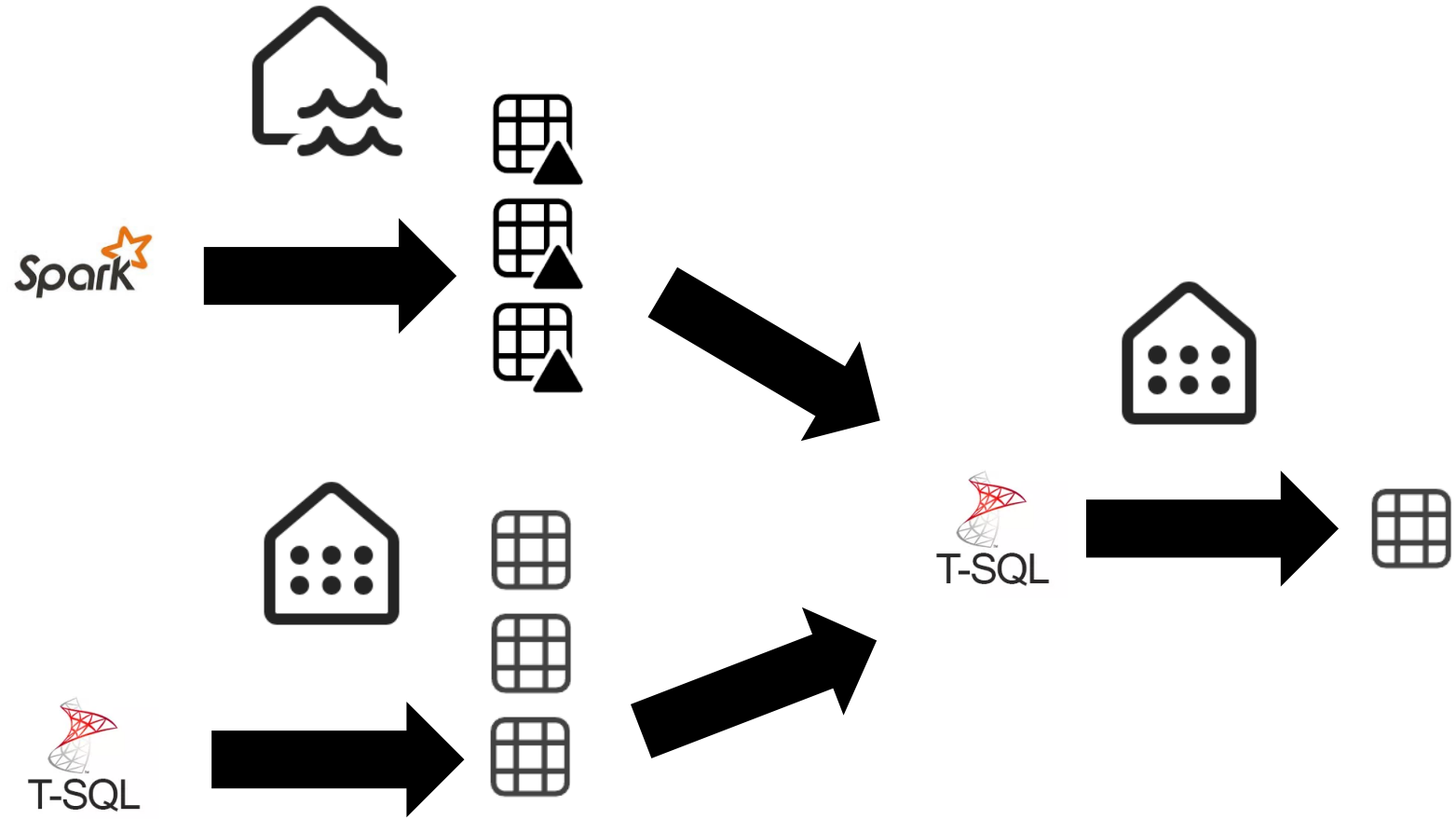|	Warehouse and Lakehouse SQL Endpoint (SQL Engine)
Warehouse tables in same workspace	R/W, no shortcuts needed
Warehouse tables in different workspace	R, within Lakehouse need shortcut to Warehouse in different workspace (only see inserts) Note: Warehouses don't support shortcuts
Lakehouse tables in same workspace	R, no shortcuts needed (use Spark engine for writes)
Lakehouse tables in different workspace	R, within Lakehouse need shortcut to Lakehouse in different workspace

- T-SQL endpoint: one per workspace. Each warehouse and each lakehouse shows up as a database. This is the way you would connect using SSMS.
- SQL Endpoint: a workspace item, which has a SQL designer, editor, and saved queries. One per lakehouse and warehouse.
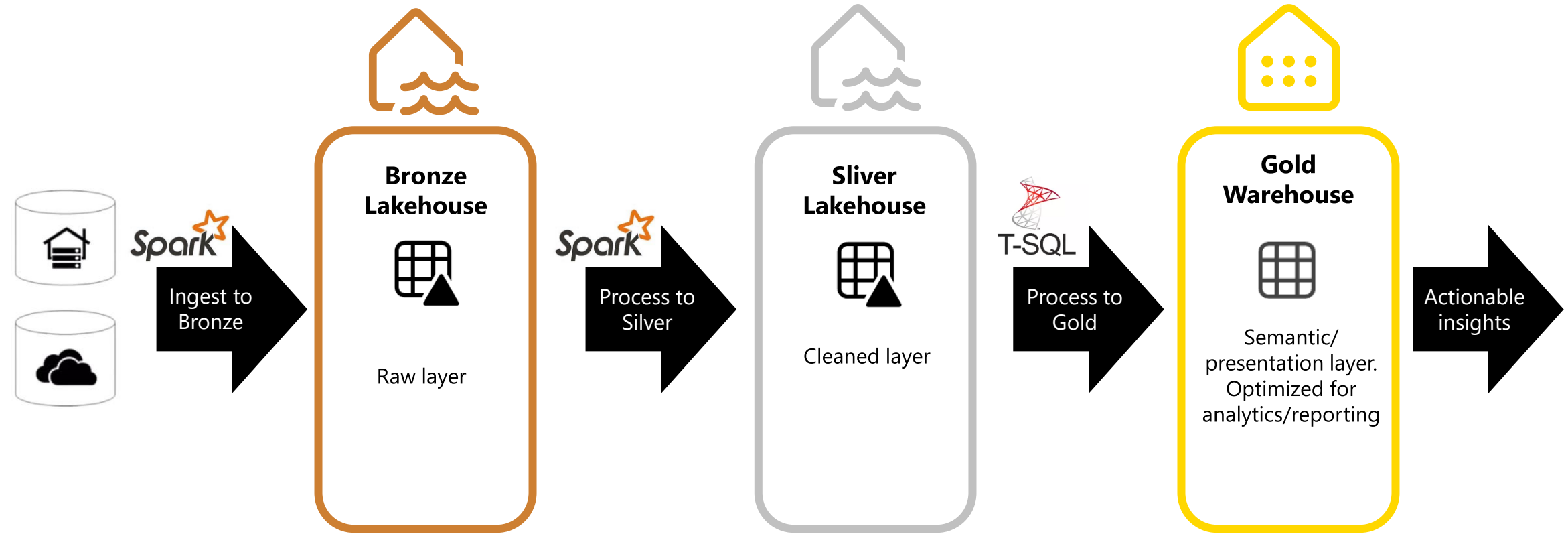
Spark/Notebook vs SQL Engine/SSMS

Microsoft Fabric

Microsoft Fabric

Bronze Lakehouse

Raw layer

Sliver Lakehouse

Cleaned layer

Gold Warehouse

Semantic/ presentation layer. Optimized for analytics/reporting

Spark — Ingest to Bronze

Spark — Process to Silver

T-SQL — Process to Gold

Actionable insights

Q & A

James Serra, Microsoft, Industry Advisor
Email me at: jamesserra3@gmail.com
Follow me at: @JamesSerra
Link to me at: www.linkedin.com/in/JamesSerra
Visit my blog at: JamesSerra.com