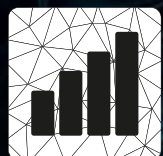




Partners

Data Point Prague 2024



**Data
Brothers**



OKškolení
Powered by OKsystem



Tabular Editor





Microsoft Fabric in a Day

Full-day hands-on workshop

Meet Your Instructors



Benni De Jagere

Senior Program Manager
Fabric CAT



Pawel Potasinski

Senior Program Manager
Fabric CAT

Kudos to:



Estera Kot, PhD
Principal Product Manager
Fabric DE&DS Product Group

Category	Description
Workshop Format	This is a hands-on workshop. You'll be actively involved in coding, implementing, and problem-solving.
What to Expect	Real-world scenarios and exercises. Direct application of concepts in live environments.
Participant Engagement	Please have your laptops and necessary tools ready. Prepare to code, collaborate, and share insights.
Hands-on	Please delve into the detailed descriptions provided for each exercise and step. Examine the attached screenshots carefully, and note the sequential visualization presented as 1), 2), 3). Review them thoroughly.
Support and Collaboration	Instructors and facilitators are here to guide you. Feel free to ask questions and help your fellow participants.
Outcome	By the end of this workshop, you'll have practical experience and new skills to apply directly to your projects.

Agenda

09:00 – 09:45 (45 min) Introduction, Set Up and Overview of Fabric Data Platform

09:45 – 11:00 (75 min) Exercise 1 - Ingest data with data pipelines and shortcuts

11:00 – 11:15 (15 min) Break

11:15 – 12:00 (45 min) Exercise 2 - Transform data using Notebooks and Spark clusters

12:00 – 13:00 (45 min) Lunch break

13:00 – 13:30 (30 min) Exercise 2 - Transform data using Notebooks and Spark clusters (c-d)

13:30 – 14:10 (40 min) Exercise 3 - Collaborate inside Notebooks and share Lakehouse. Use SQL Endpoint and SSMS

14:10 – 14:20 (10 min) Break

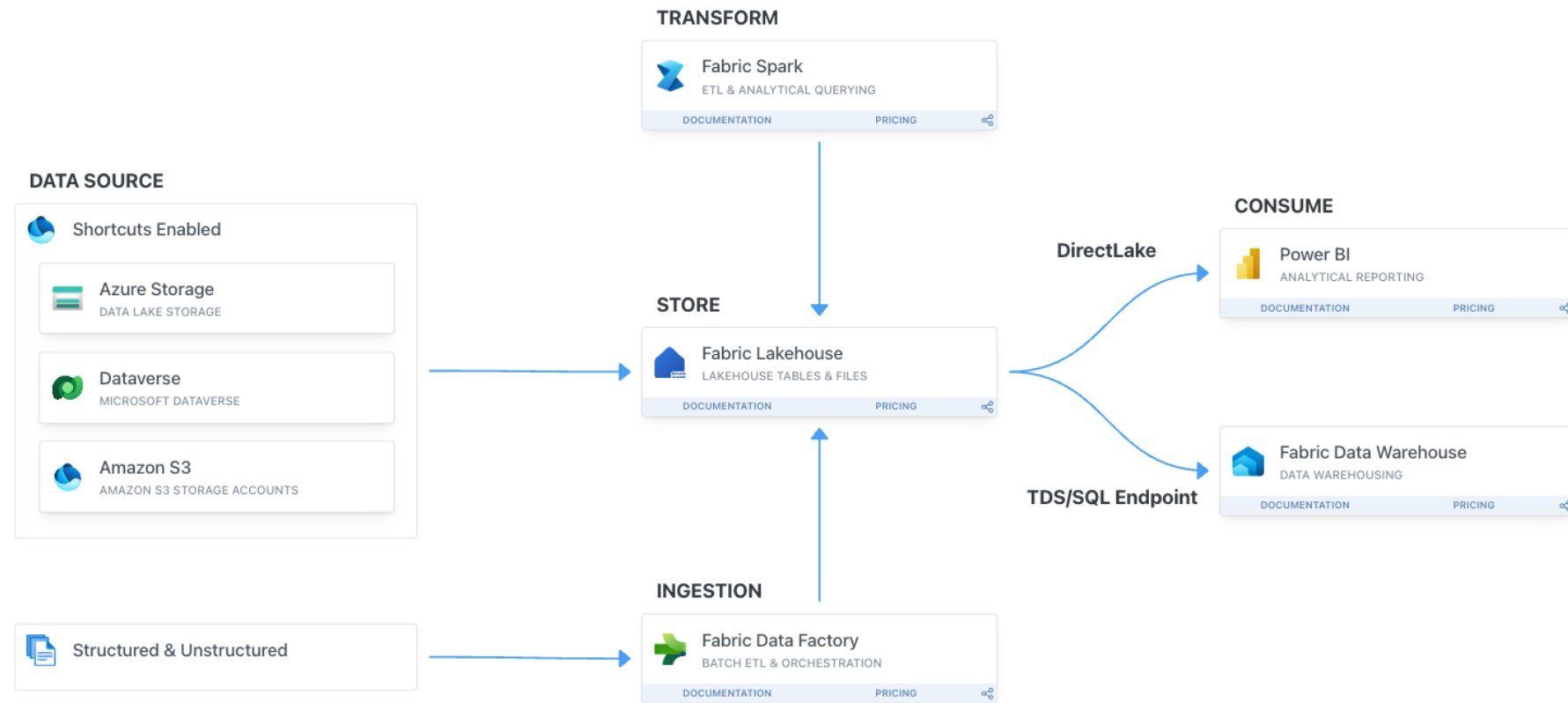
14:20 – 15:40 (80 min) Exercise 4 - Serve and consume data using Power BI and Data Science

15:40 – 15:50 (10 min) Break

15:50 – 16:40 (50 min) Exercise 5 - Latest Fabric Features

16:30 – 17:00 (30 min) Buffer, Recap and Extra exercises

High-level architecture



Enjoy, experience,
network and learn!

Let's get to know each other!

#1

Raise your hand if you have ever tried Microsoft Fabric before.

This could be in any context, just any prior exposure at all.

Let's get to know each other!

#2

Raise your hand if you have explored more than three workloads within Fabric, such as Power BI, Data Science, Data Factory, or Data Engineering.

Let's get to know each other!

#3

Raise your hand if you, or your company, have faced challenges with orchestrating and integrating different tools to ingest, process, and maximize the value from data.

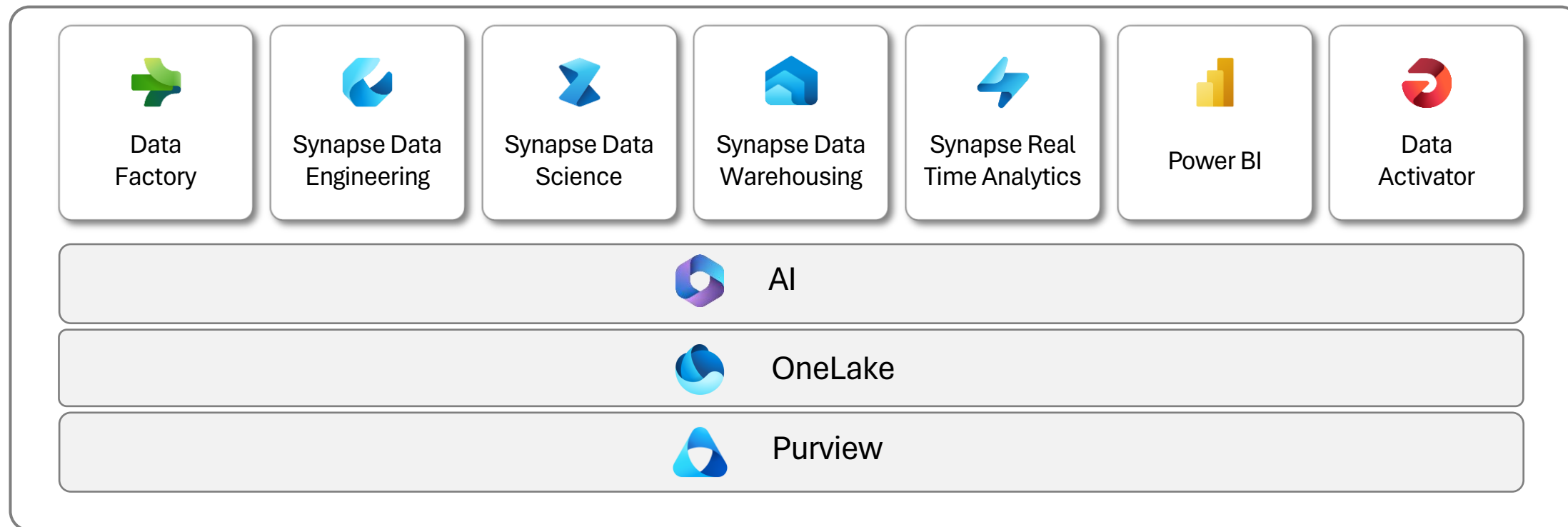
Introduction, Setup and Overview of Fabric Data Platform

Setup and Overview of Fabric Data Platform





The unified data platform for the era of AI



Unified
architecture

Unified
experience

Unified
governance

Unified
business model



The unified data platform for the era of AI

Complete Analytics Platform

Everything, unified

SaaS-ified

Secured 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

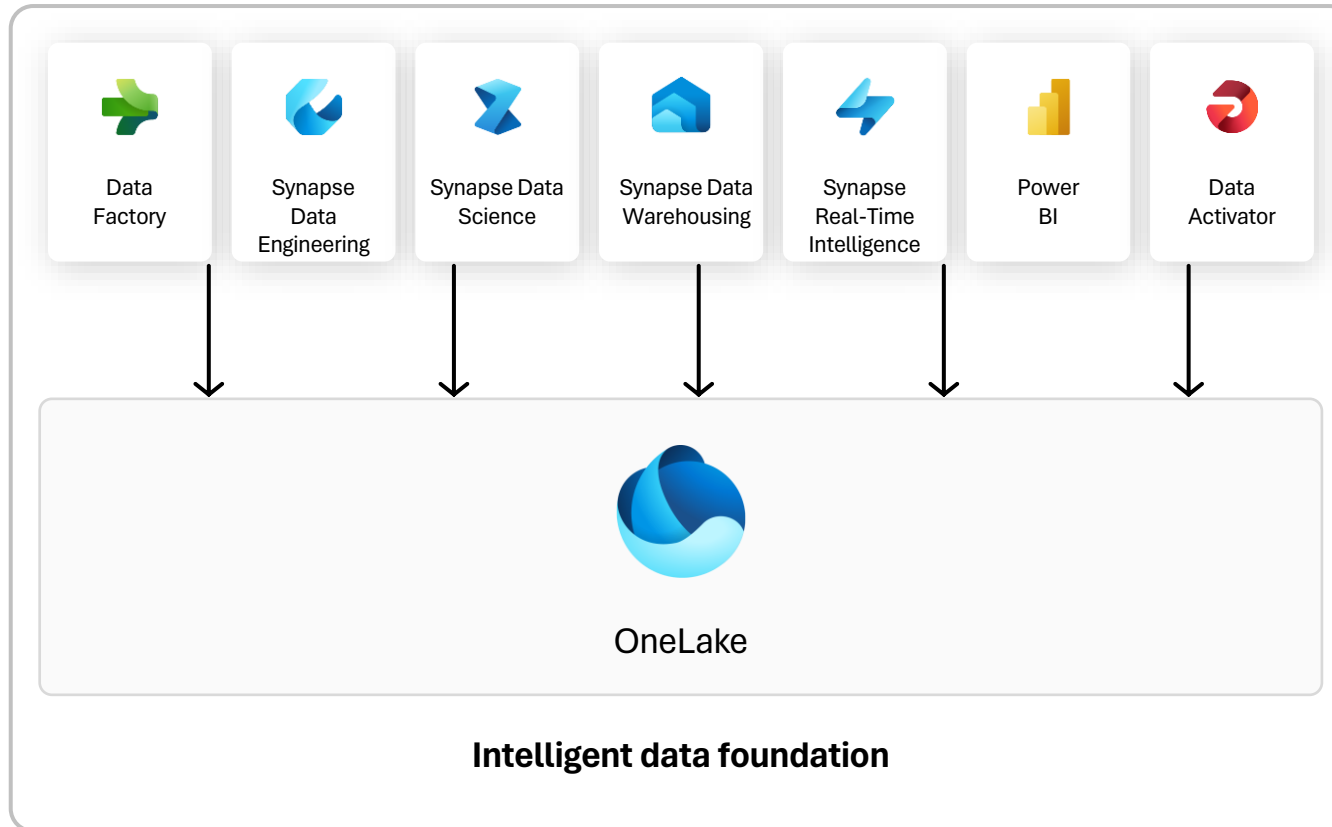
Copilot accelerated

ChatGPT on your data

AI driven insights

OneLake for all data

“The OneDrive for data”



A single SaaS lake for the whole organization

Provisioned automatically with the tenant

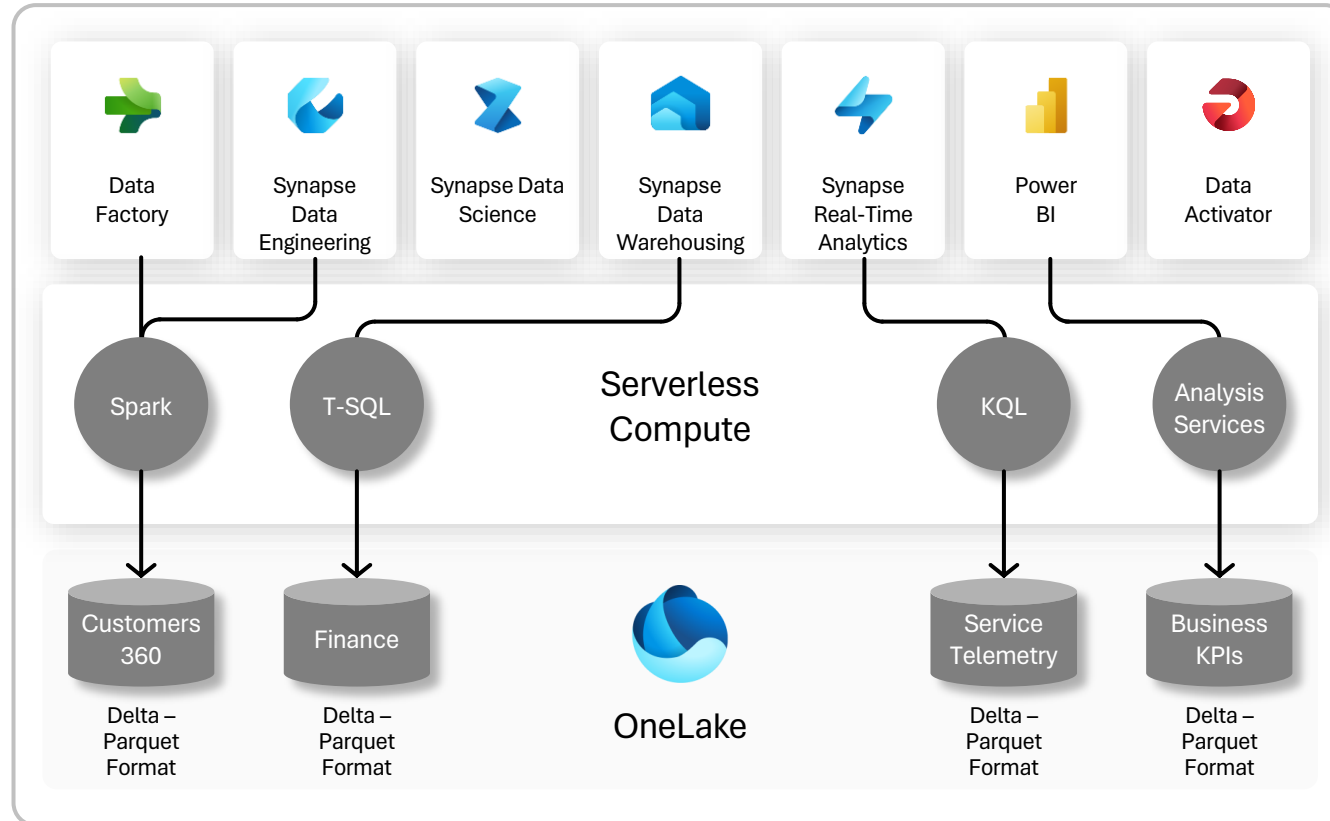
All workloads automatically store their data in the OneLake workspace folders

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

One Copy for all computes

Real separation of compute and storage



All the compute engines store their data automatically in OneLake

The data is stored in a single common format

[Delta - Parquet](#), an open standards format, is the storage format for all tabular data in Microsoft Fabric

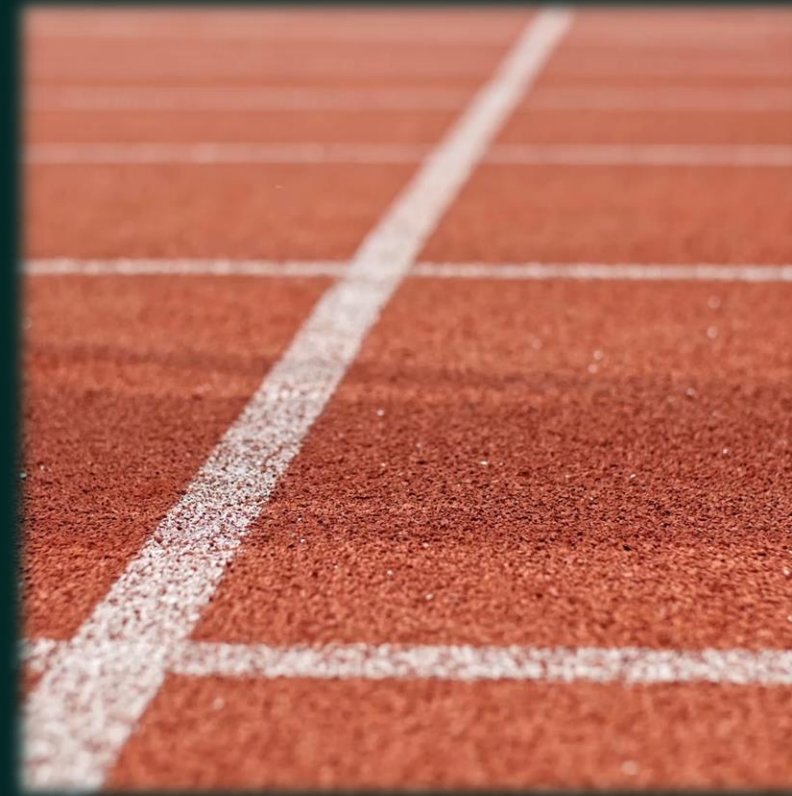
Once data is stored in the lake, it is directly accessible by all the engines without needing any import / export

All the compute engines have been fully optimized to work with Delta Parquet as their native format

Shared universal security model is enforced across all the engines

Exercise 1

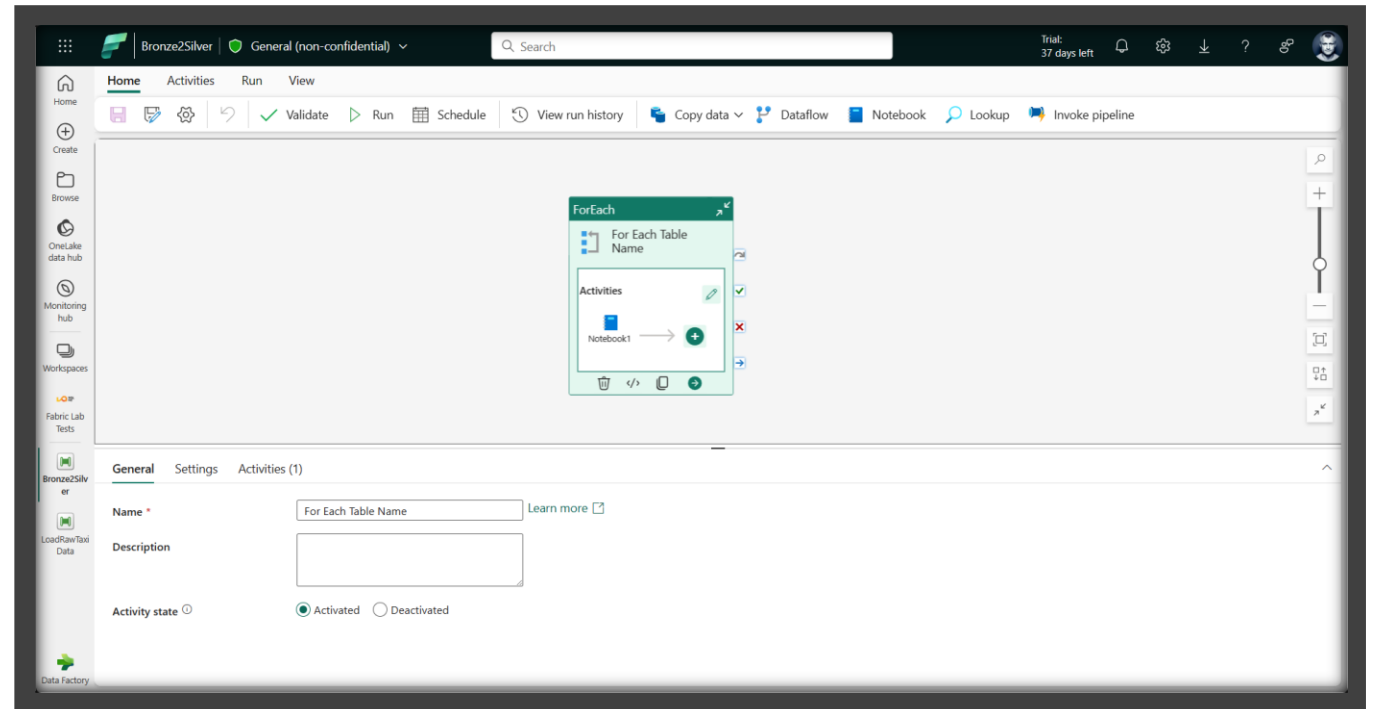
Ingest data with data pipelines and shortcuts



Data Pipelines

Data Pipelines enable powerful workflow capabilities at cloud-scale like building complex workflows, moving PB-size data, and defining sophisticated control flow pipelines.

Data pipelines can be used to build complex ETL and data factory workflows that can perform a number of different tasks at scale. Additionally, control flow capabilities are built into pipelines so you can build workflow logic which provide loops and conditional.

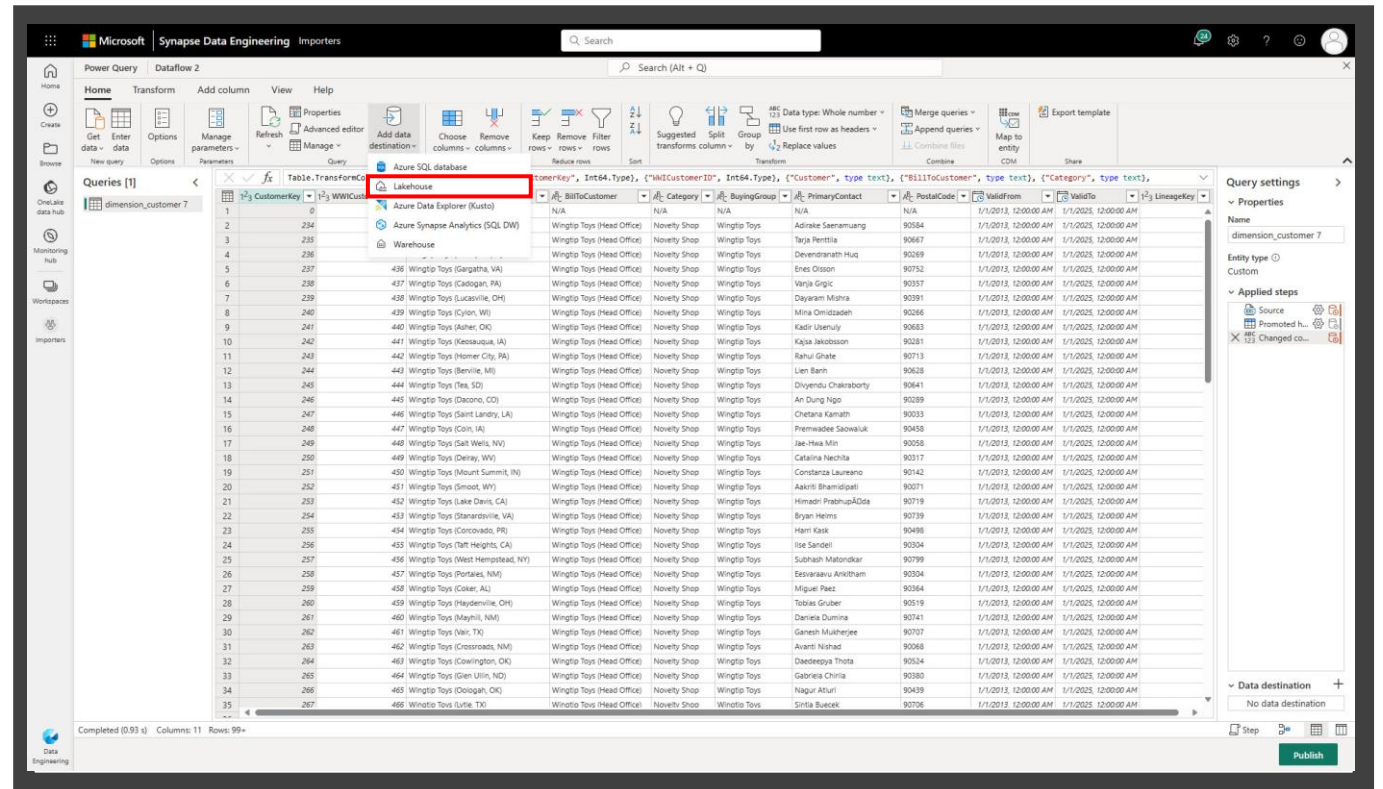


Dataflows Gen2

Dataflows provide a low-code interface for ingesting and transforming data from hundreds of data sources.

Key Capabilities:

- Accelerate data transformation with well-known Power Query user experience (no-code/low-code)
- Load results of data transformations into multiple destinations (Lakehouse, Warehouse, Azure SQL Database, etc.)
- Integrate with data pipelines for orchestration



Shortcuts



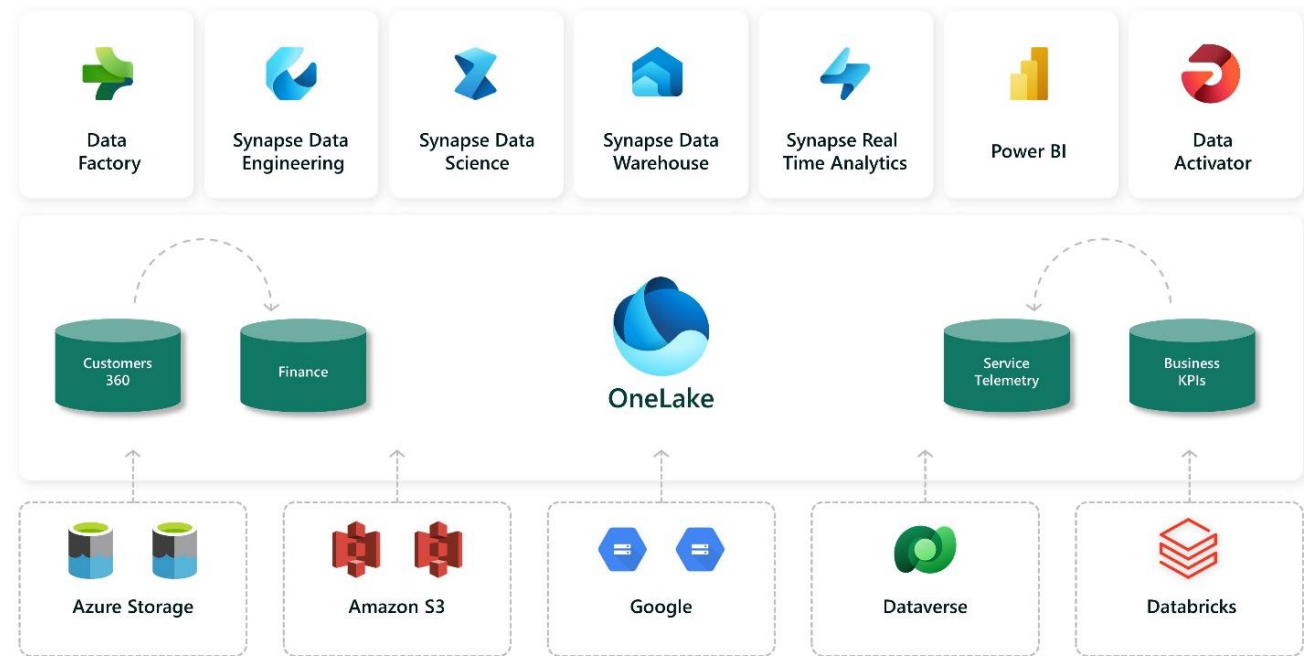
Shortcuts unify data without copying or moving existing data.

This means that data can be used multiple times without data duplication.



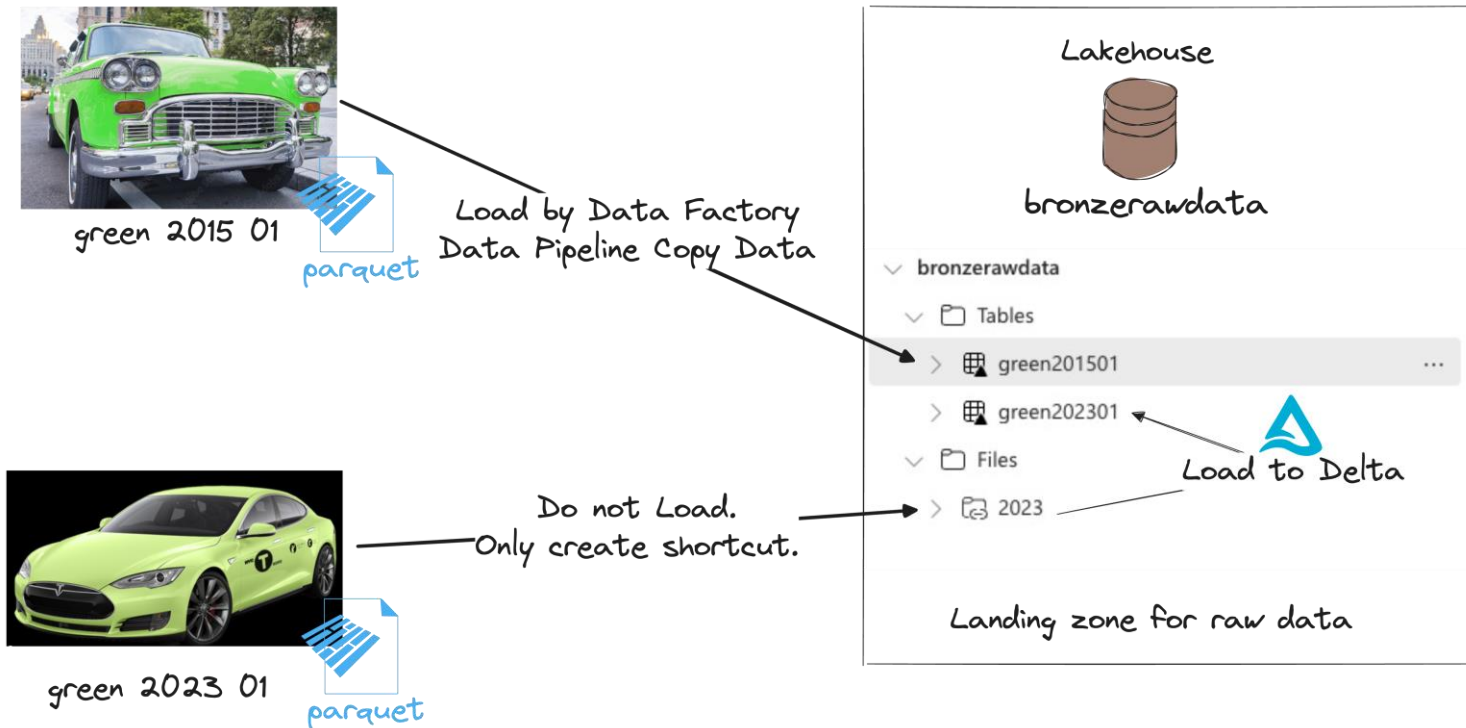
Key Capabilities:

- Create shortcuts within Microsoft Fabric to consolidate data across artifacts or workspaces, without changing ownership of the data
- With shortcuts, data throughout OneLake can be composed together without any data movement
- Shortcuts also allow instant linking of data already existing in Azure and in other clouds, without any data duplication and movement, making OneLake the first multi-cloud data lake
- With support for industry standard APIs, OneLake data can be directly accessed by any application or service



Exercise 1

Ingest data with data pipelines and shortcuts

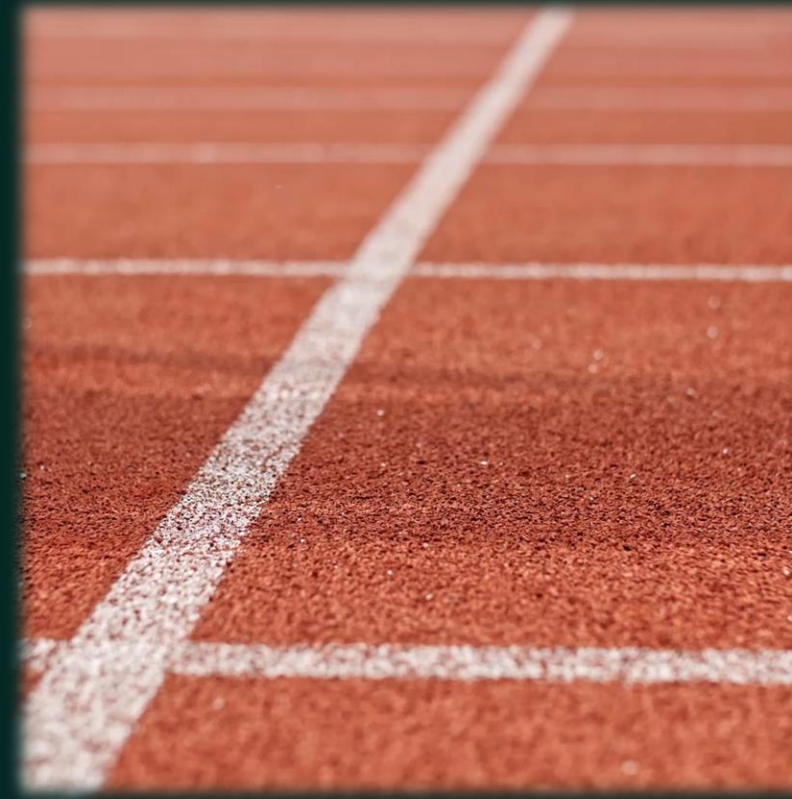


Let's get your hands dirty with Fabric!

aka.ms/fabriclab-datapoint

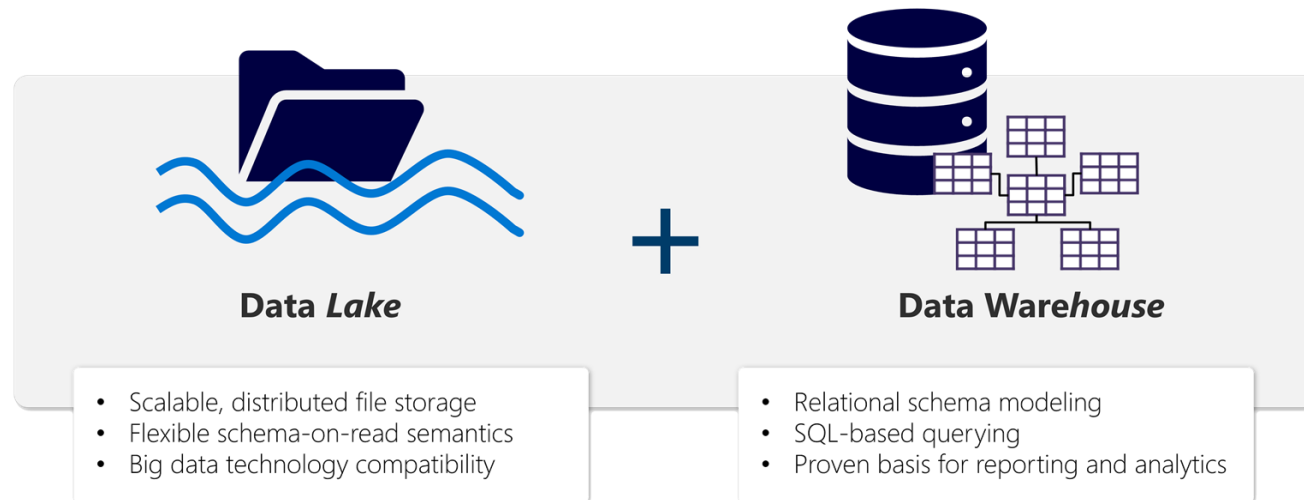
Exercise 2

Transform data using Notebooks and Spark clusters



Lakehouse concept

Explore the Microsoft Fabric Lakehouse



Lakehouses use Spark and SQL engines to process large-scale data and support machine learning or predictive modeling analytics.

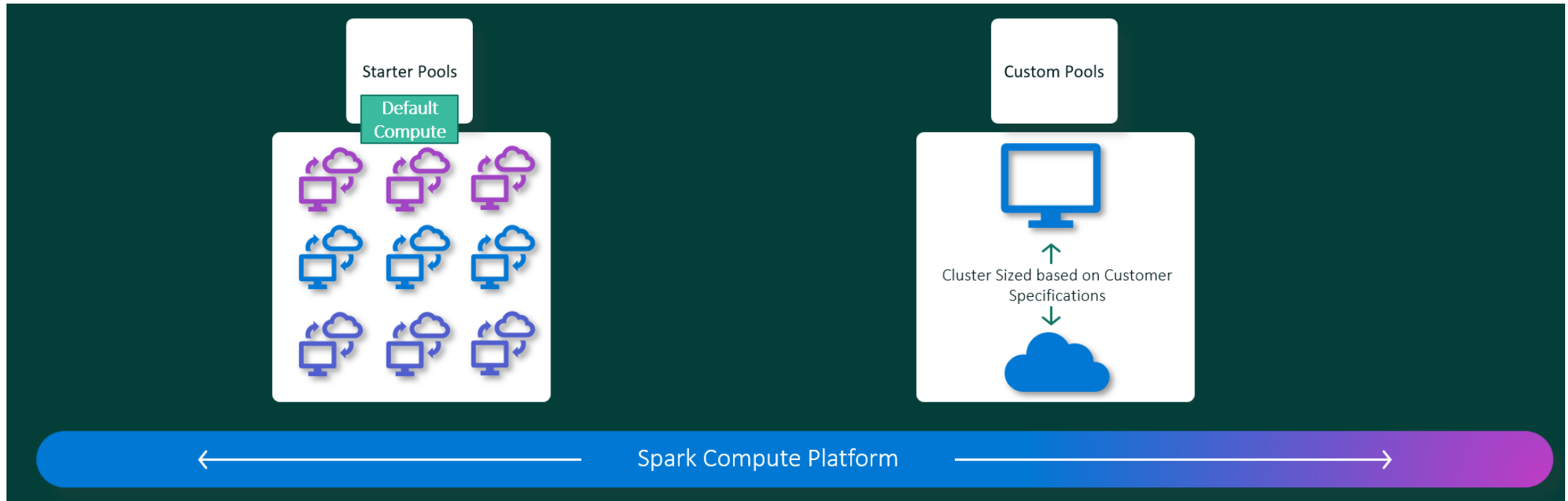
Lakehouse data is organized in a schema-on-read format, which means you define the schema as needed rather than having a predefined schema.

Lakehouses support **ACID** (Atomicity, Consistency, Isolation, Durability) transactions through Delta Lake formatted tables for data consistency and integrity.

Lakehouses are a single location for data engineers, data scientists, and data analysts to access and use data.

Spark compute for Data Engineering

Spark compute platform



Spark compute for Data Engineering

Starter pools

Starter Pool Configuration

Node family	Memory optimized
Node Size	Medium
Min and Max Nodes	[1, 10]
Autoscale	On
Dynamic Allocation	On

Data Engineering in Microsoft Fabric

Microsoft Fabric Data Engineering items



Lakehouse



Notebook



Environment
(Preview)



Spark Job
Definition



Data pipeline



Import notebook



Use a sample

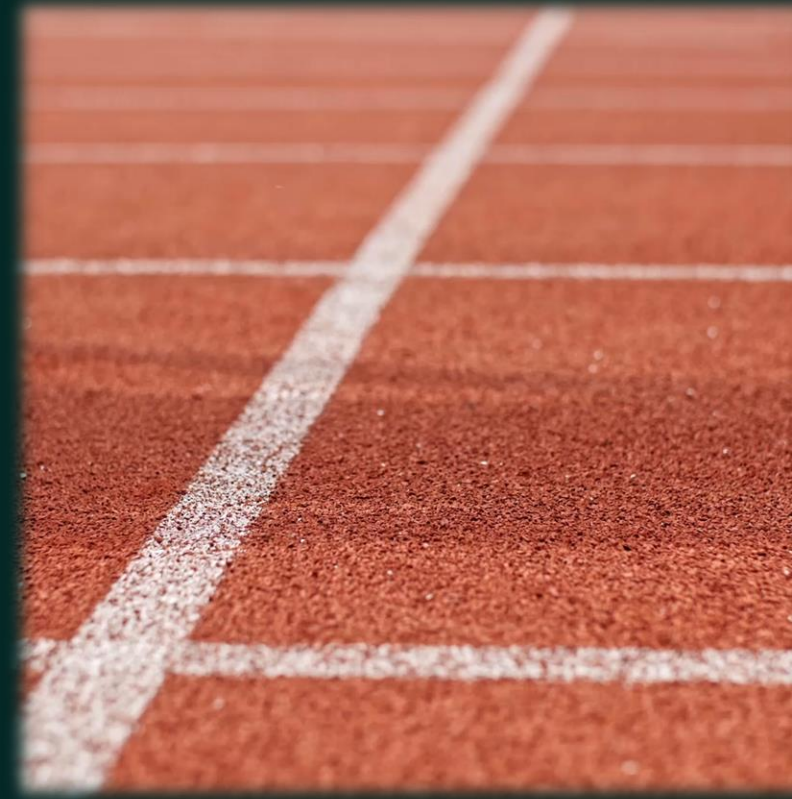
Exercise 2

Transform data using Notebooks and Spark clusters



Exercise 3

Collaborate inside Notebooks and share Lakehouse.
Use SQL Endpoint and SSMS



Exercise 3

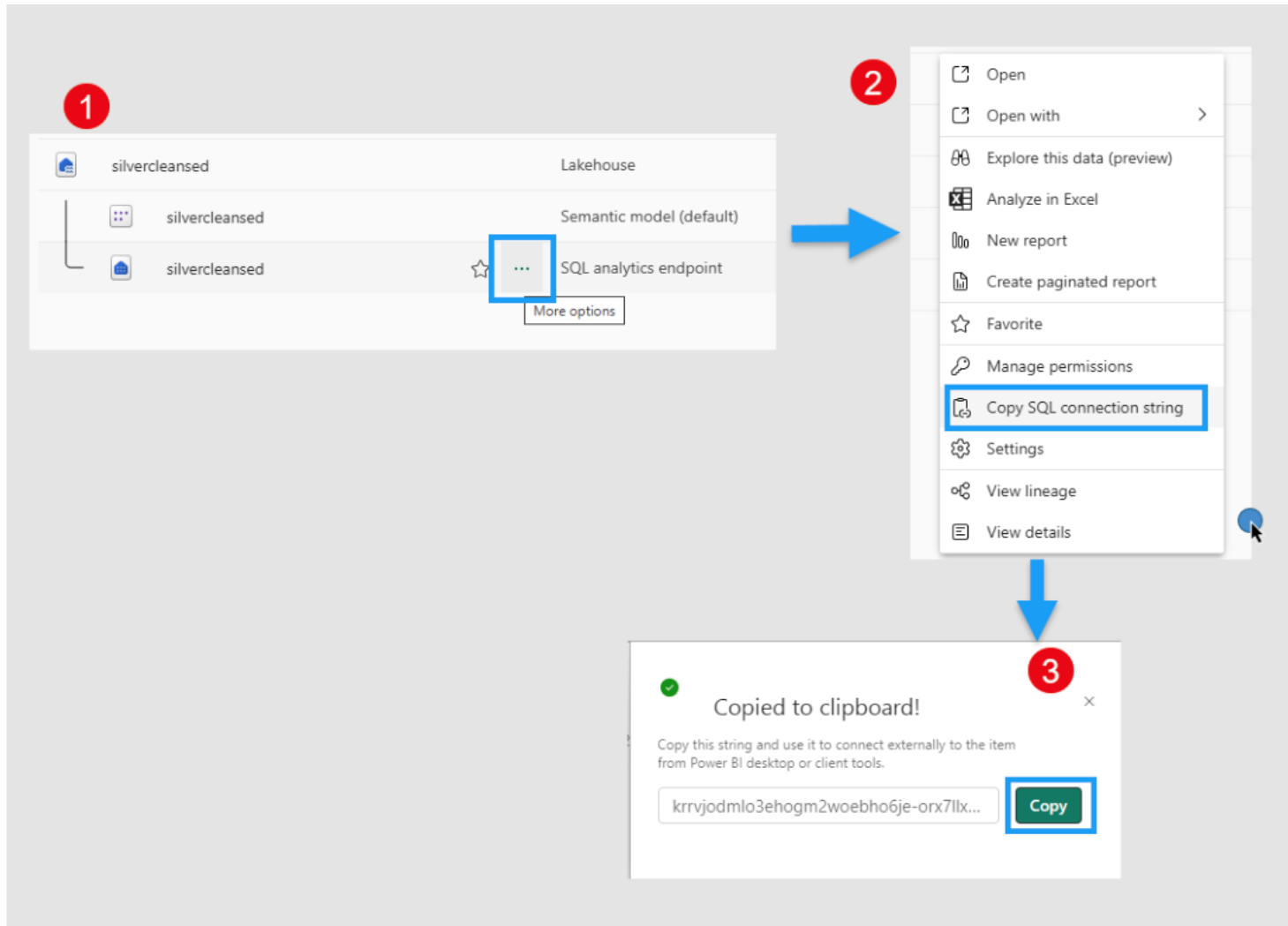
Collaborate inside Notebooks and share Lakehouse. Use SQL Endpoint and SSMS



Exercise 3

- Retrieve Lakehouse SQL analytics endpoint connection string
- Connect to Fabric SQL endpoint from SSMS
- Execute T-SQL queries on Lakehouse Delta Tables
- Share your Lakehouse
- Share your notebook for collaboration

Get lakehouse's SQL analytics endpoint connection string



Connect to Fabric SQL endpoint from SSMS

Connect to Server

SQL Server

Server type: Database Engine

Server name: 4d3e5fsgalhweq2eri.datawarehouse.fabric.microsoft.com

Authentication: Microsoft Entra Password

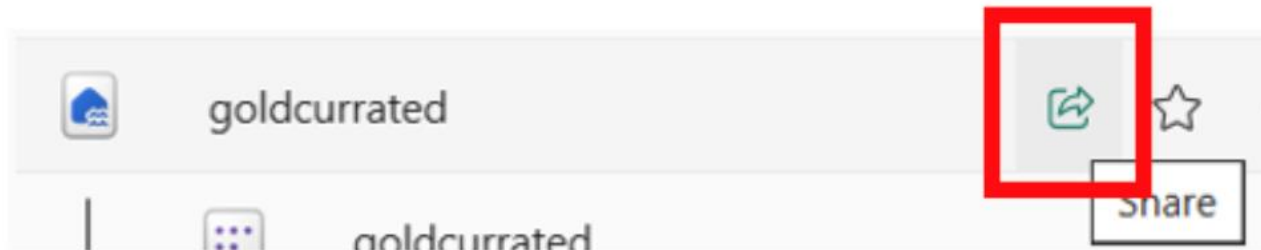
User name: BB001@fabricconf.onmicrosoft.com

Password:

☐ Remember password

Connect Cancel Help Options >>

Share your Lakehouse



Grant people access

goldcurrated

People you share this Lakehouse with can open it and its SQL endpoint and read the default dataset. To allow them to read directly in the Lakehouse, grant additional permissions.

1

DE001

×

Enter a name or email add

2

Additional permissions

☐ Read all SQL endpoint data ⓘ

☐ Read all Apache Spark ⓘ

☐ Build reports on the default dataset

3

Notification Options

☒ Notify recipients by email

Add a message (optional)

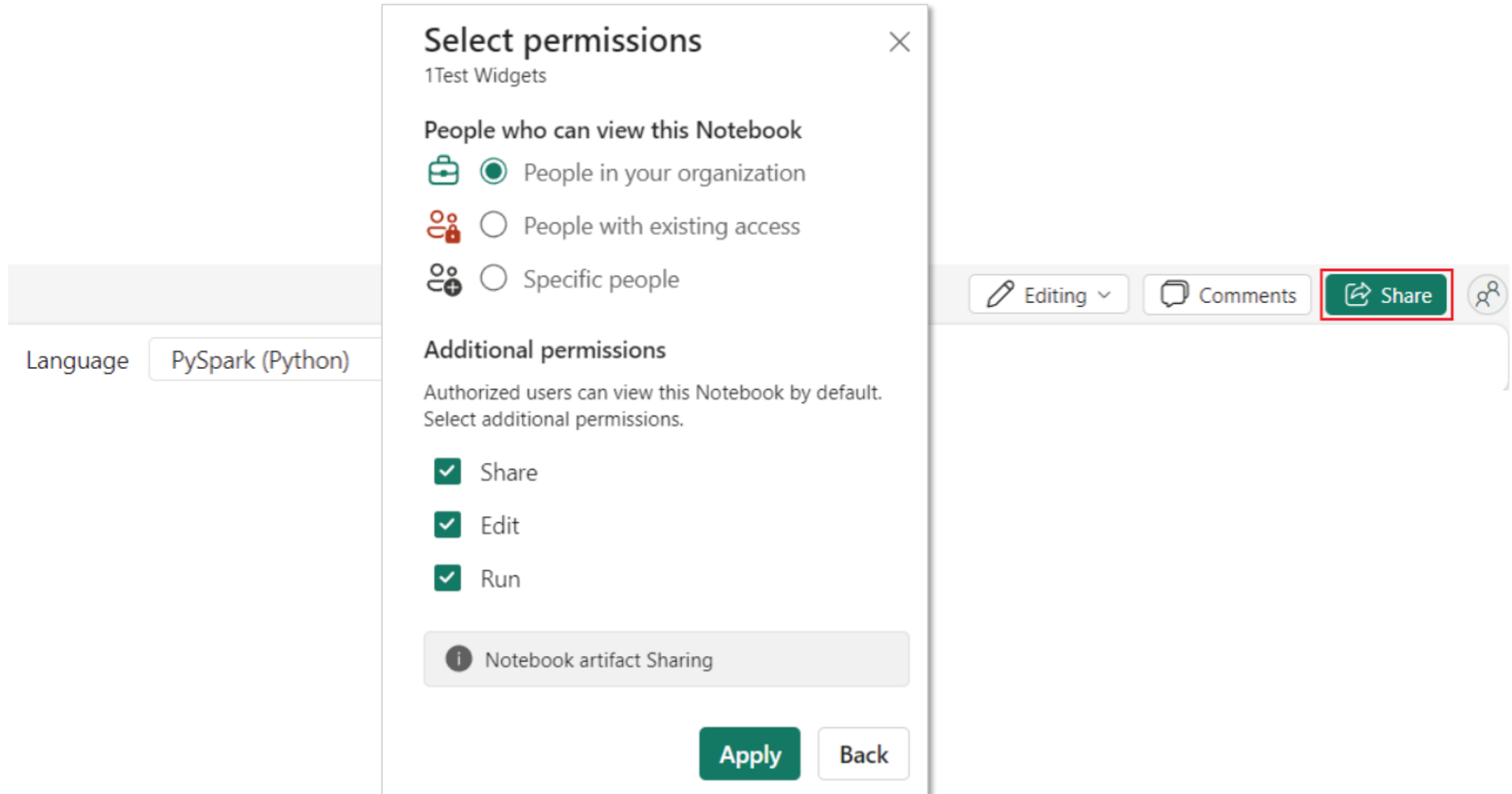
4

Grant

Back

ⓘ Depending on which additional permissions you select, recipients will have different access to the SQL endpoint, default dataset, and data in the lakehouse. For details, view lakehouse permissions documentation.

Share your notebook for collaboration



The image shows a notebook interface with a 'Select permissions' dialog box open. The dialog box is titled 'Select permissions' and has a close button (X) in the top right corner. Below the title, it says '1 Test Widgets'. The dialog is divided into two main sections: 'People who can view this Notebook' and 'Additional permissions'.

People who can view this Notebook

- ☒ People in your organization (indicated by a green briefcase icon)
- ☐ People with existing access (indicated by a red group of people icon)
- ☐ Specific people (indicated by a grey group of people icon with a plus sign)

Additional permissions

Authorized users can view this Notebook by default. Select additional permissions.

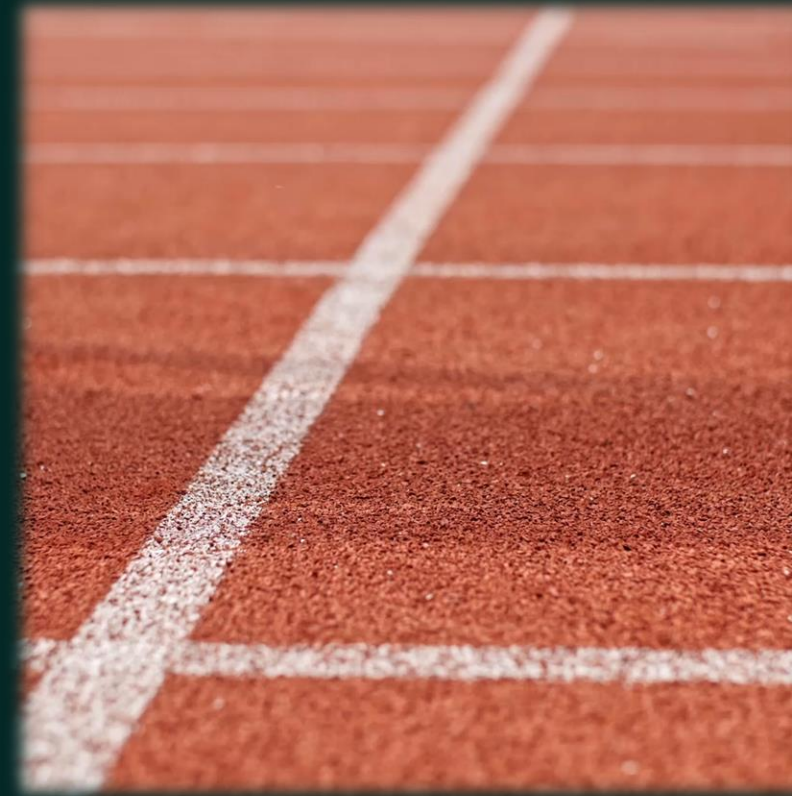
- ☒ Share
- ☒ Edit
- ☒ Run

At the bottom of the dialog, there is a grey bar with an information icon (i) and the text 'Notebook artifact Sharing'. Below this bar are two buttons: 'Apply' (green) and 'Back' (white).

In the background, the notebook interface is visible. It includes a 'Language' dropdown menu set to 'PySpark (Python)'. To the right of the dialog, there is a toolbar with three buttons: 'Editing' (with a pencil icon and a dropdown arrow), 'Comments' (with a speech bubble icon), and 'Share' (with a share icon and a red border). To the right of the 'Share' button is a user profile icon.

Exercise 4

Serve and consume data using
Power BI and Data Science



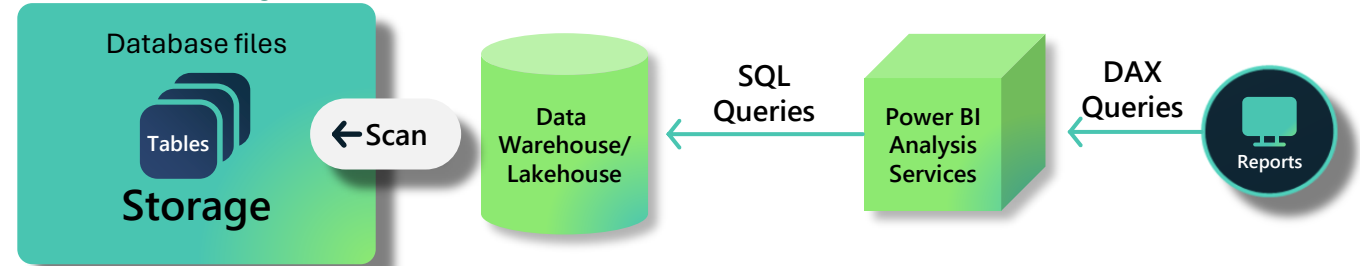
Power BI | Direct Lake mode



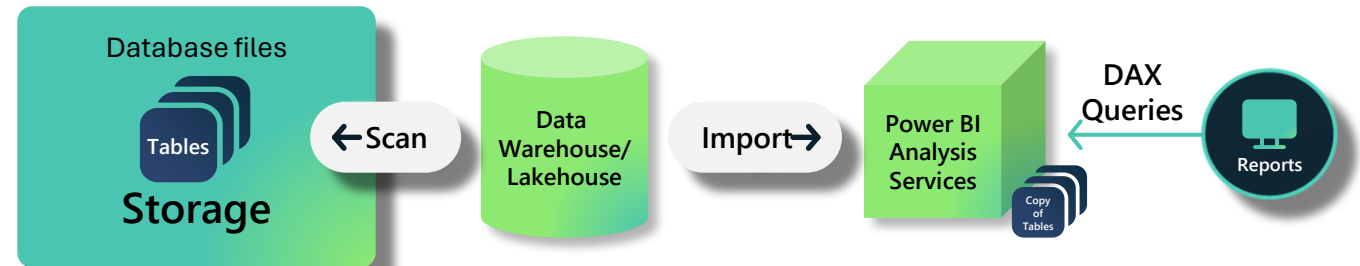
Direct Lake is a fast-path to load the data from the lake straight into the Power BI engine, ready for analysis.

Direct Lake is based on loading parquet-formatted files directly from a data lake without having to query a Lakehouse endpoint, and without having to import or duplicate data into a Power BI semantic model.

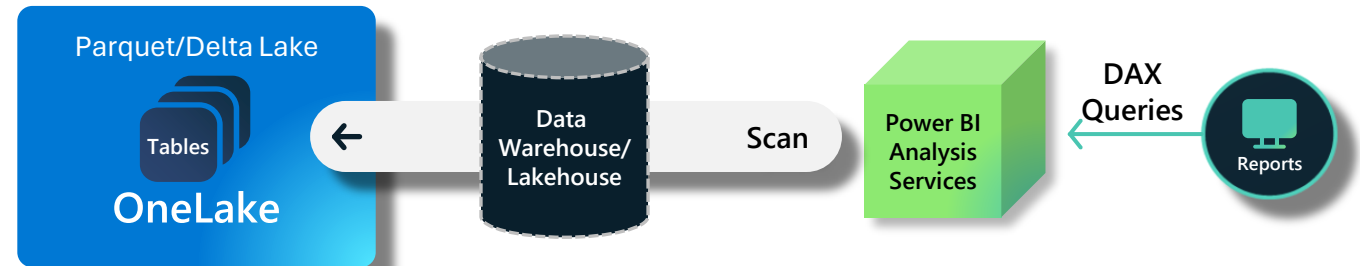
Direct Query mode. Slow, but real time



Import mode. Latent and duplicative, but fast



Direct Lake mode. Fast and real time



Introducing V-Ordering

Write time optimization to parquet files

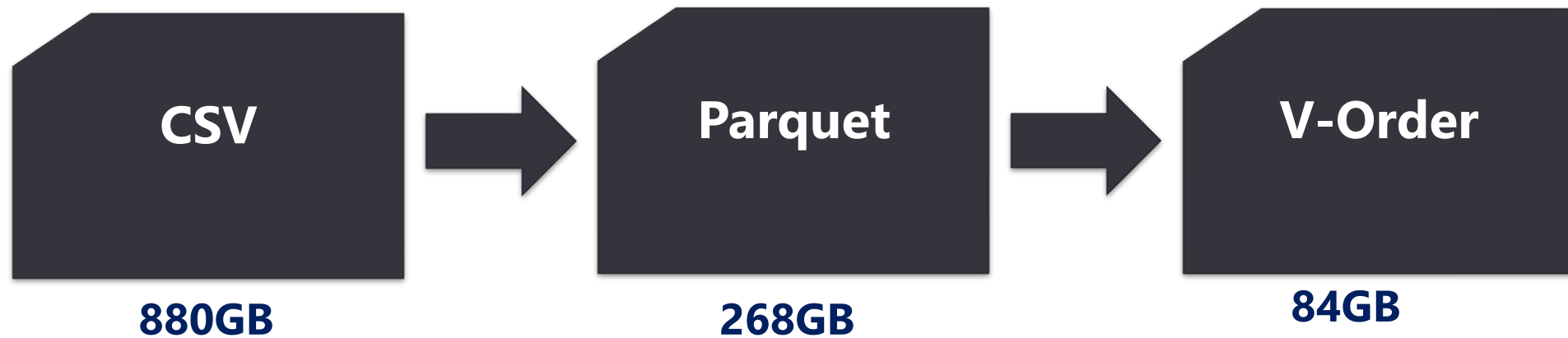
Sorting, row group distribution, dictionary encoding, and compression (Shuffling)

Complies to the open standard

Z-Order, compaction, vacuum, time travel, etc. are compatible with V-Order

V-ordering in action

Microsoft Internal DB (162 tables)



x3.2

Reduced IO for workloads

DirectLake Mode

- On start, no data is loaded in memory
- Column data is transcoded from Parquet files when queried
- Tables can have mix of resident and non-resident columns
- Column data can get evicted over time
- DirectLake fallback as an alternative
- "Framing" of dataset determines what gets loaded from Delta Lake

Optimizing Delta for Direct Lake mode

- V-Order makes a big difference, as it's tailored for Verti-Scan
- Direct Lake will work over Shortcuts to external data
 - Expect a performance impact, because reasons ..
- Direct Lake thrives on fewer, larger .parquet files
 - Physical structure will always be crucial
 - OPTIMIZE (bin-compaction) and VACUUM in the Data Engineering process will be key
 - Especially with streaming/small batch architectures, keep this in mind
- Principle of lean models will still apply
 - Only include what's needed for the reports and datasets

Framing

- What is framing
 - "point in time" way of tracking what data can be queried by DirectLake
- Framing is near instant and acts like a cursor
 - Determines the set of .parquet files to use/ignore for *transcoding* operations
- Why is this important
 - Delta-lake data is transient for many reasons
- Typical ETL Process
 - Ingest data to delta lake tables
 - Transform as needed using preferred tool
 - When ready, perform *Framing* operation on dataset

Limitations

- Relationships based on DateTime types
- Calculated Columns and Calculated Tables
- Complex delta table column types (i.e. Binary and GUID)
 - Some other
- T-SQL Based views will **always** fallback to DQ mode
- Composite models are not yet supported

Identifying Fallback

- You can tell when Fallback happens if ..
 - It's slower than usual 😊
 - Using the Performance Analyzer, you see a "DirectQuery" category
 - Performing a trace, you see DirectQueryBegin and DirectQueryEnd events
 - Depending on the behaviour, you get an error in the report(s)

Controlling Fallback Behaviour

- The FallbackBehaviour is set to 'Automatic' by default
- Alternative options are:
 - DirectLake only
 - DirectQuery only
- Be careful when making changes to this ..



Couldn't load the data for this visual

We cannot process the request because the table 'vw_Records' either does not exist or requires fallback to DirectQuery mode. Fallback to DirectQuery mode is disabled in this semantic model. Consider enabling fallback to DirectQuery mode and try again. See <https://go.microsoft.com/fwlink/?linkid=2248855> to learn more.

Close

Data science in Microsoft Fabric

End-to-end data science for predictive business insights



Data Centric

- Easy and secure access to lake-centric data
- Open Delta Lake support promotes reproducibility
- Native integration with data infrastructure



Developer friendly

- SaaS experiences with quick setup
- Code authoring experiences in Notebooks and IDE
- VS Code integration



Rich ML tools

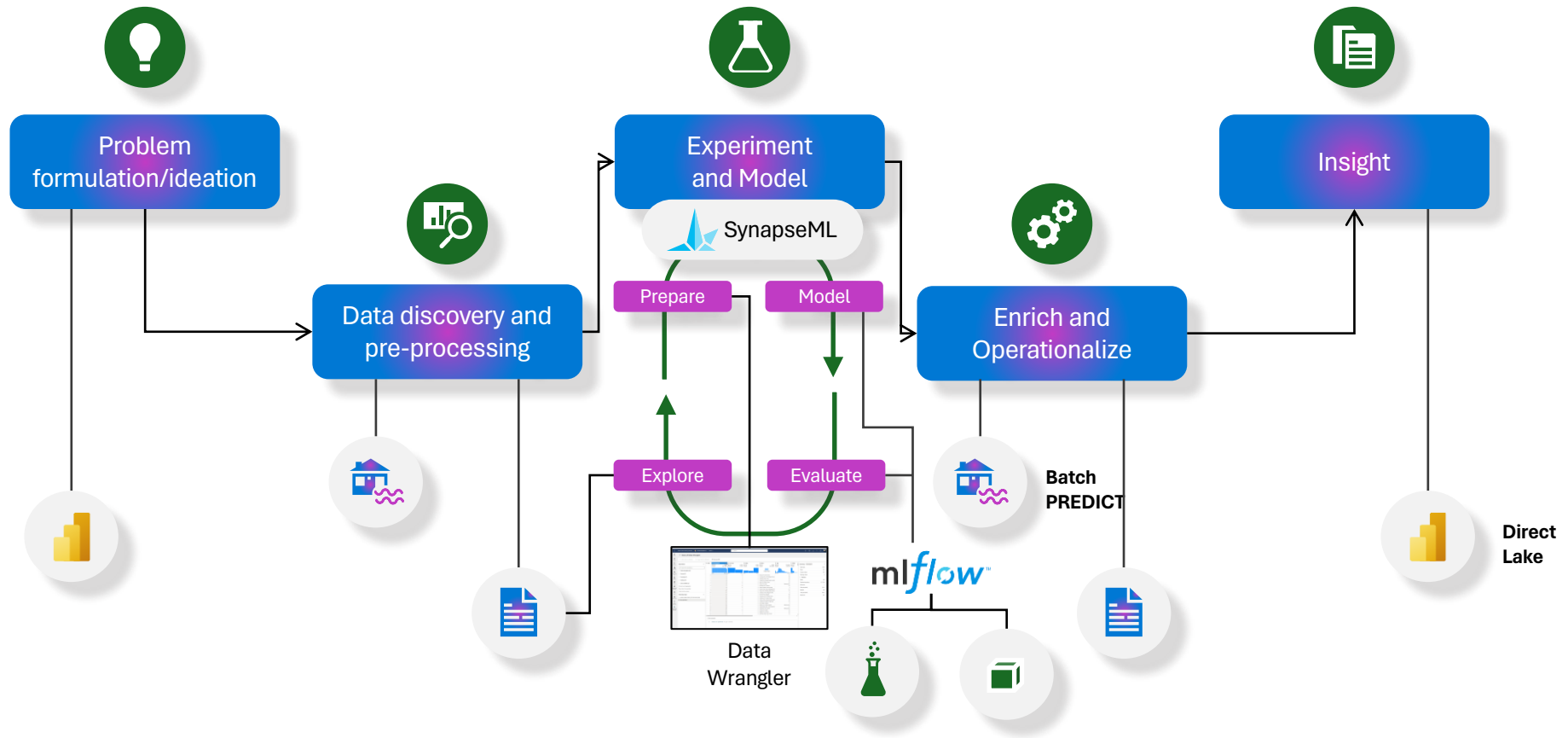
- Supports MLFlow model and experiment management
- Built-in, scalable ML tools with SynapseML
- Direct Lake mode for serving predictions to BI reports



Promotes collaboration

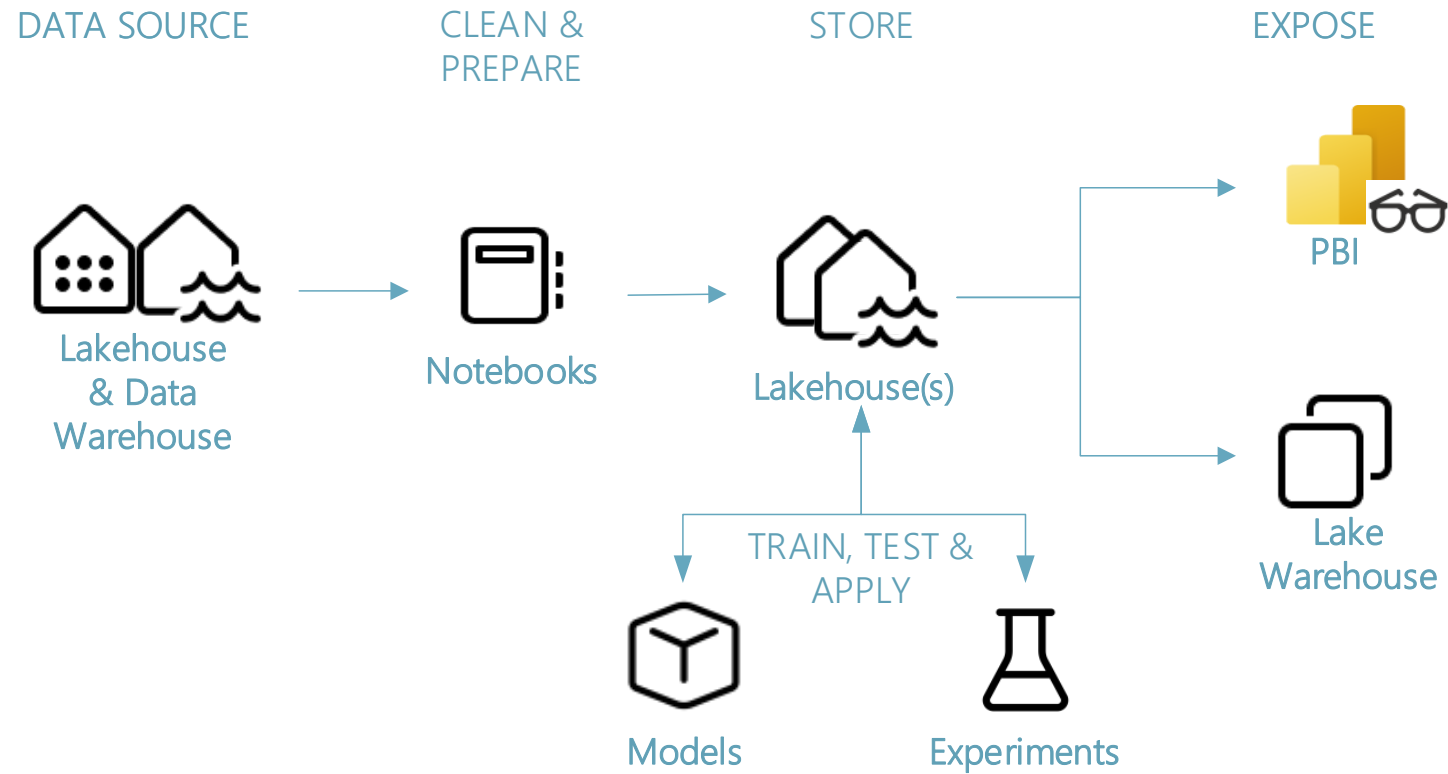
- Unified platform for all analytics roles incl. data scientists
- Secure and easy sharing of data, code, models and experiments

The Fabric Data Science experience



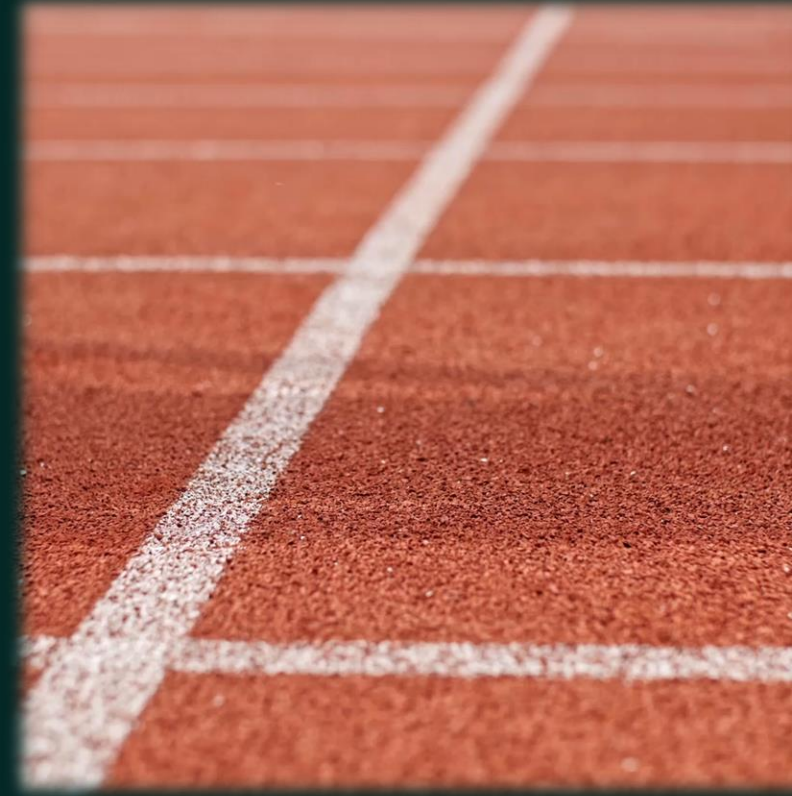
Exercise 4

Serve and consume data using Power BI and Data Science



Exercise 5

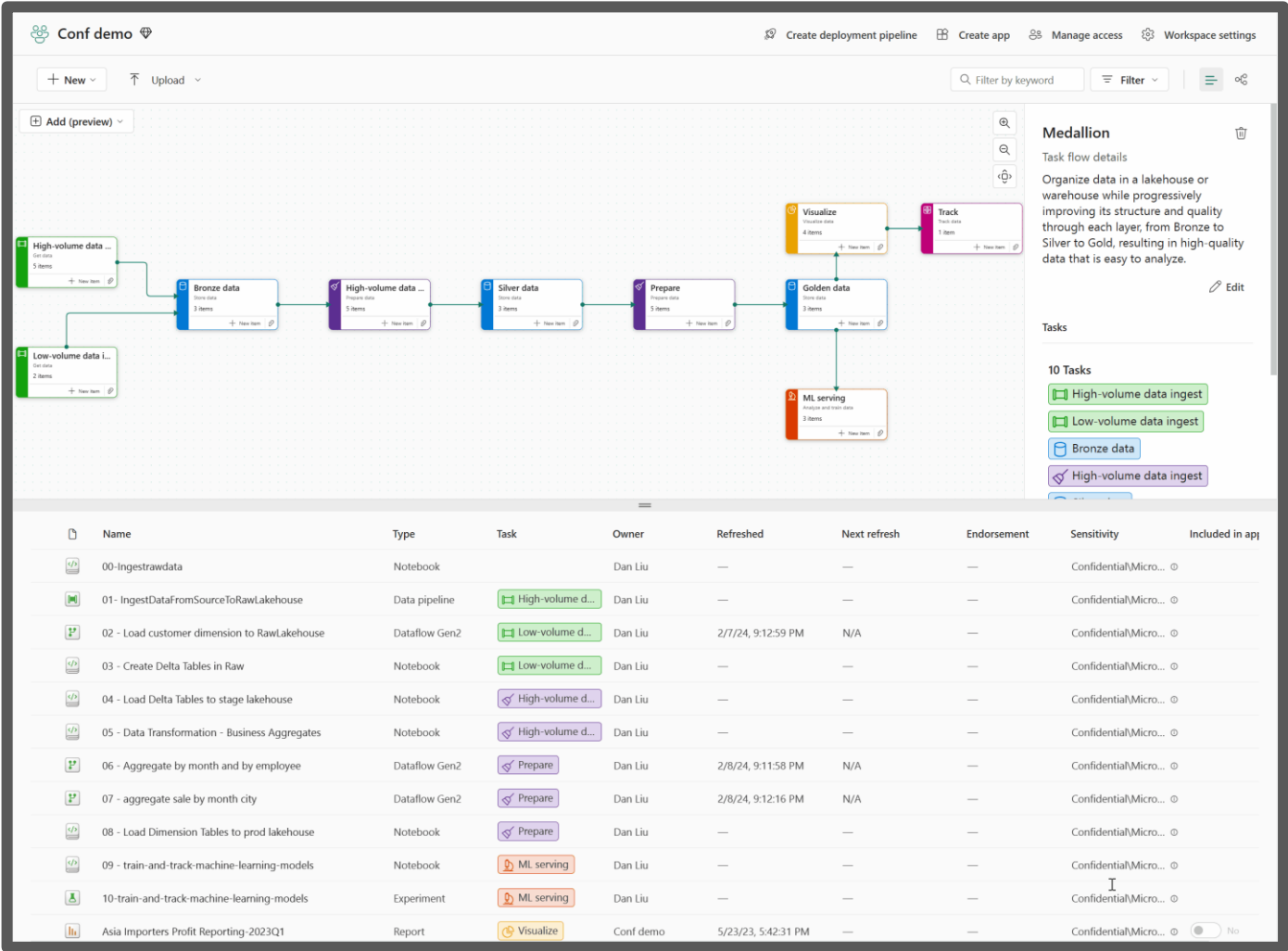
Latest Fabric Features



Task flows

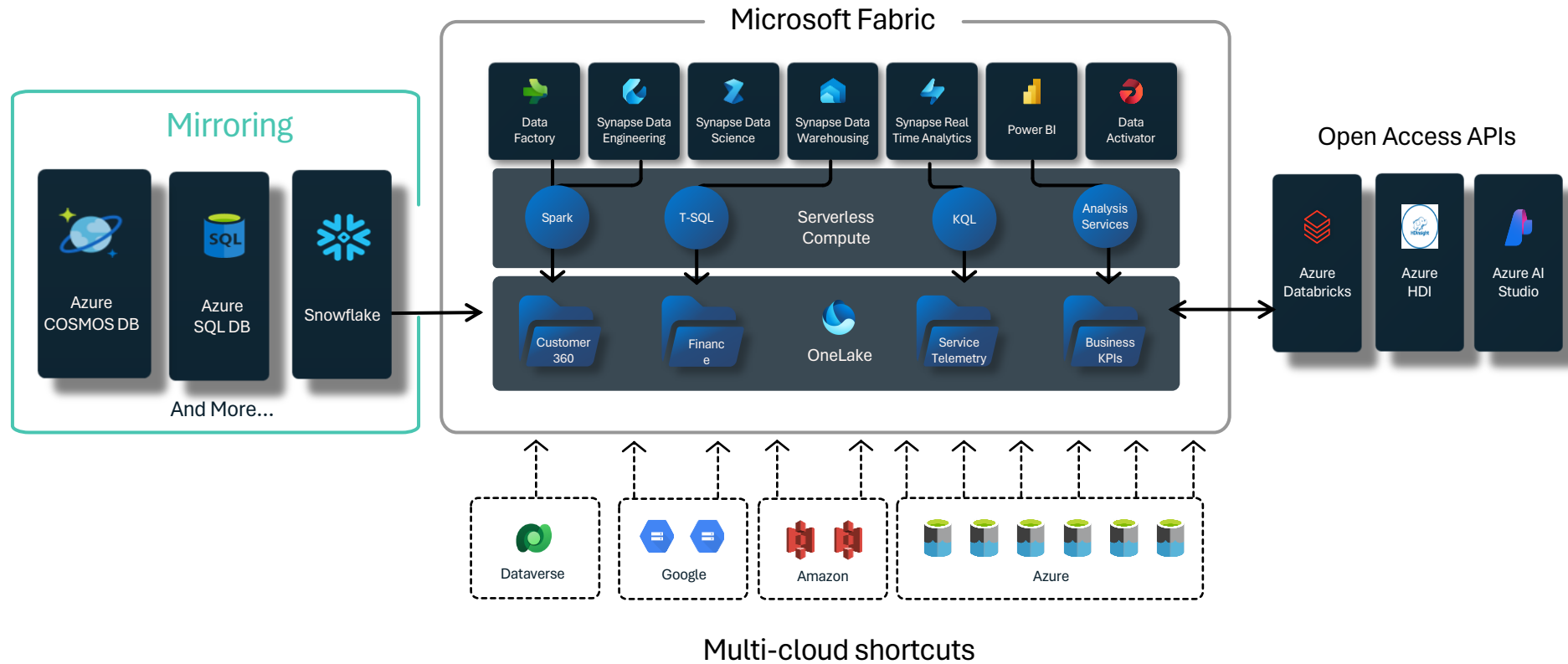
Fabric task flow is a workspace feature that enables you to build a visualization of the flow of work in the workspace.

The task flow helps you understand how items are related and work together in your workspace, and makes it easier for you to navigate your workspace, even as it becomes more complex over time.



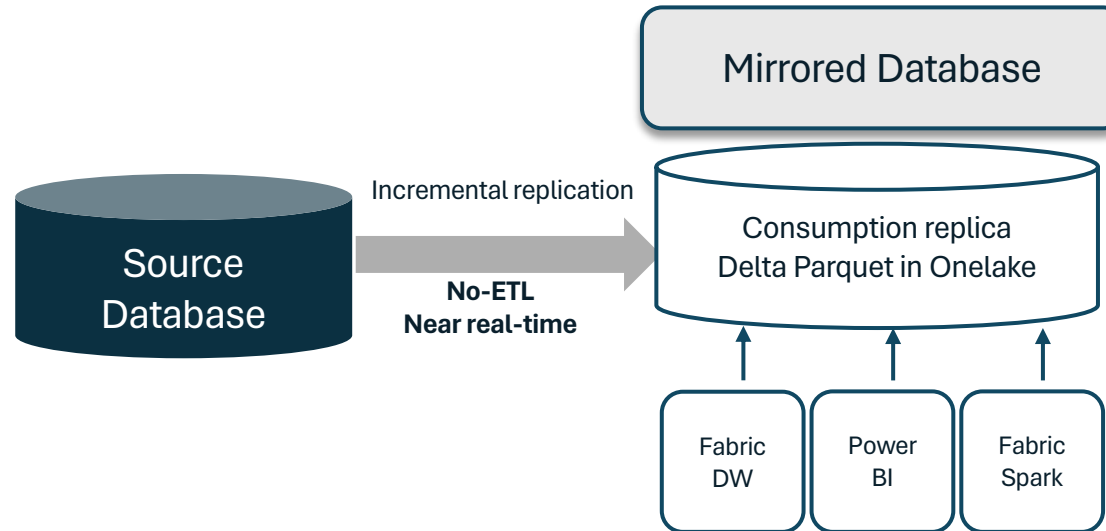
Creating Data Gravity in OneLake

Connect your entire data estate



Mirroring in Microsoft Fabric

Simplify near real-time analytics



Fabric Mirroring enables adding existing databases and data warehouses to Fabric without any ETL.

Data is replicated into OneLake in Delta format in near real-time

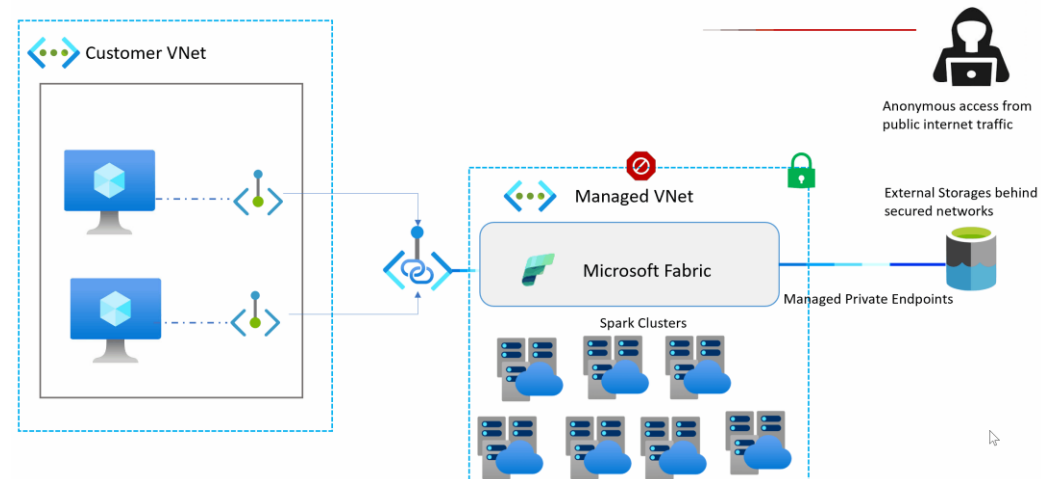
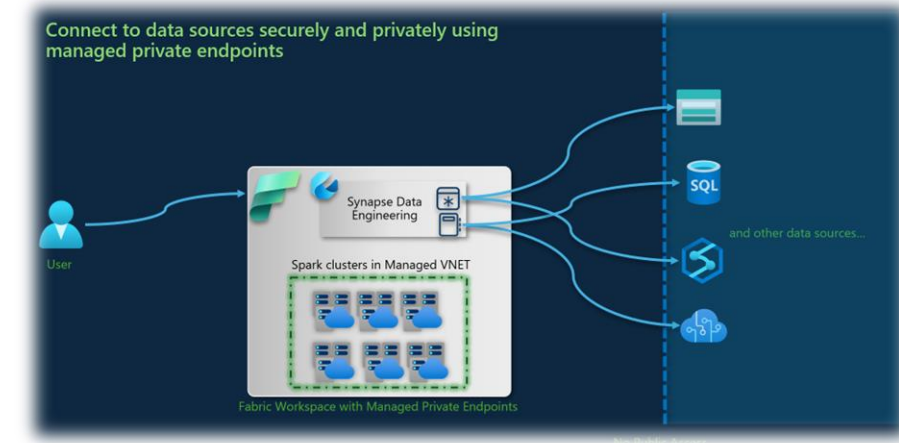
All of the Fabric analytics and AI experiences instantly work with the Mirrored Database

Your data is in an open format enabling limitless opportunities for data liberation

Managed private endpoints in Microsoft Fabric

Connect to data sources securely

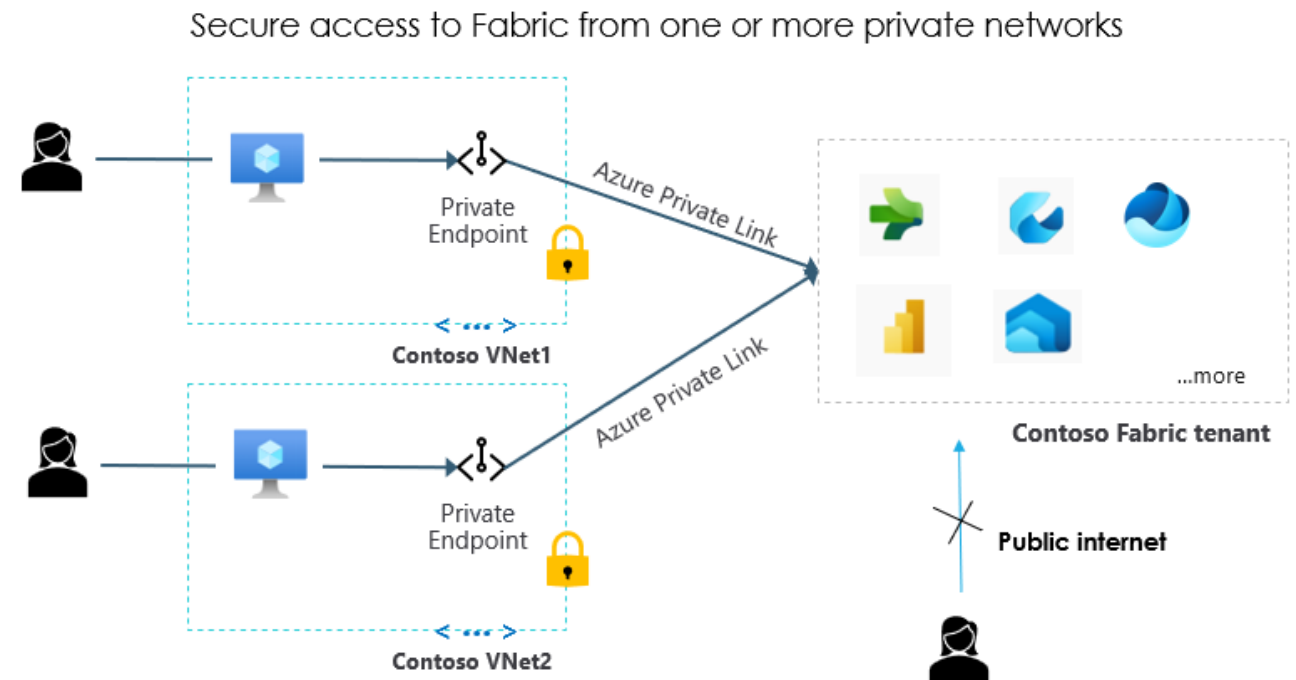
- With a managed virtual network you get complete network isolation for the Spark clusters running your Spark
- You don't need to create a subnet for the Spark clusters based on peak load, as this is managed for you by Microsoft Fabric.
- A managed virtual network for your workspace, along with managed private endpoints, allows you to access data sources that are behind firewalls or otherwise blocked from public access.



Private links in Microsoft Fabric

Connect to data sources securely

- Restrict traffic from the internet to Fabric and route it through the Microsoft backbone network
- Ensure only authorized client machines can access Fabric
- Comply with regulatory and compliance requirements that mandate private access to your data and analytics services.

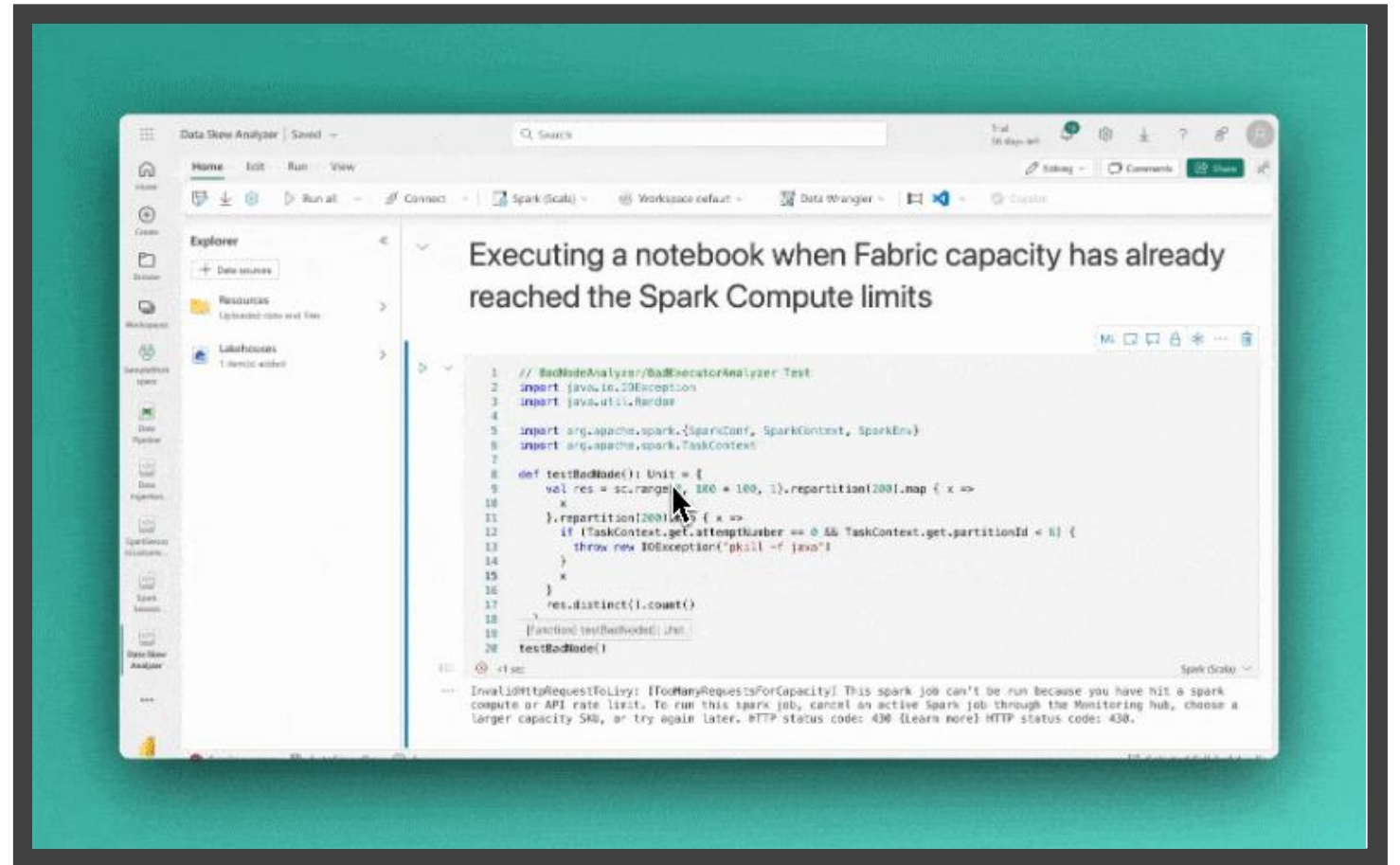


<https://aka.ms/fabricsecuritywhitepaper>

Job Queueing for Notebook

Job Queueing eliminates manual retries and improve the user experience for users who run notebook jobs on Microsoft Fabric.

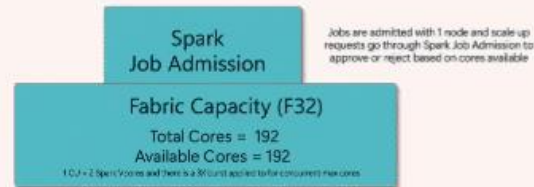
Notebook jobs that are triggered by pipelines or job scheduler will be added to a queue and will be retried automatically when the capacity frees up.



Optimistic Job Admission for Spark

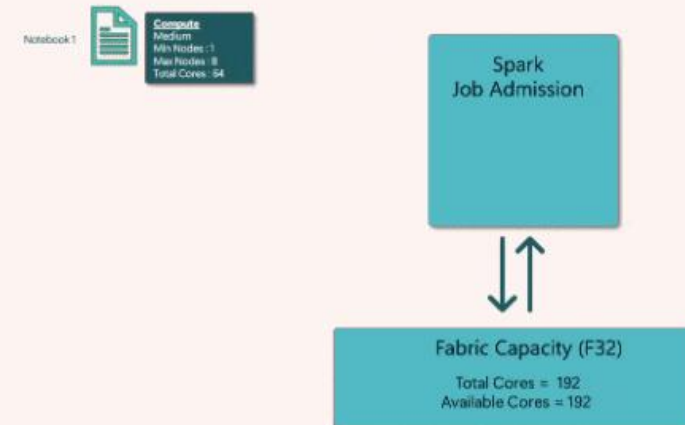
Job Concurrency (Without Optimistic Job Admission)

Job Admission (Optimistic Approach)



- By enabling autoscale, F32 capacity could support running 24 concurrent jobs (192 Total Cores/8 Core per job)
- Job scale up is approved/ rejected based on cores available in a fair manner
- Scale up or new job admission exceeding the available cores are queued or throttled

Job Scale Up Flow



Workshop recap



Microsoft Fabric Resources

Community Call to Action

- ✓ Try Microsoft Fabric for free: <https://aka.ms/try-fabric>
- ✓ Join the Fabric community: <https://aka.ms/fabriccommunity>
- ✓ Share and vote for ideas to improve Fabric: <https://aka.ms/fabricideas>
- ✓ Read and comment our blog: <https://aka.ms/fabricblog>

Learn More About Microsoft Fabric

- Product website: <https://aka.ms/microsoft-fabric>
- Documentation: <https://aka.ms/fabric-docs>
- Fabric e-book: <https://aka.ms/fabric-get-started-ebook>
- Microsoft Learn: <https://aka.ms/learn-fabric>
- End-to-end scenario tutorials: <https://aka.ms/fabric-tutorials>
- Fabric Notes: <https://aka.ms/fabric-notes>