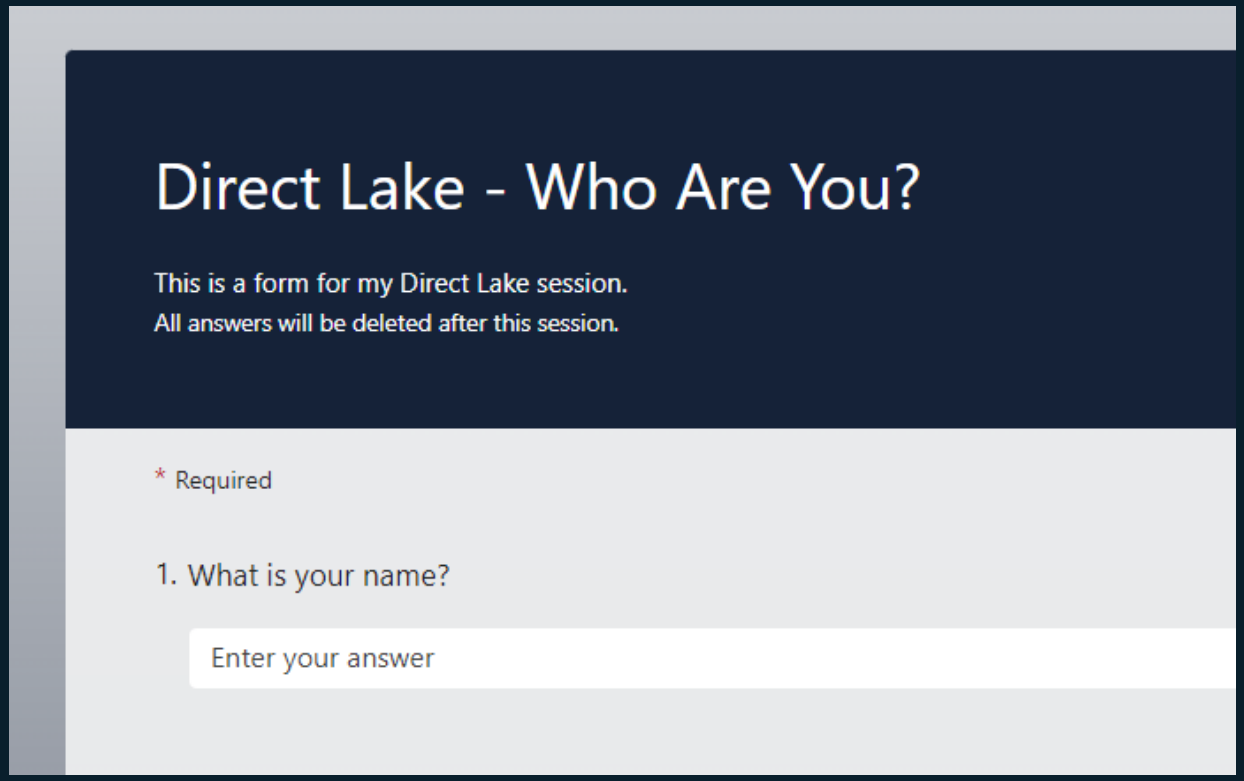
A Deep Dive into Power BI Direct Lake

https://bit.ly/dl-pbi

A Deep Dive into Power BI Direct Lake

Nicky van Vroenhoven

Who Are You?



https://bit.ly/dl-pbi

Check out my blog for a how-to:
nickyvv.com/2019/12

THANK YOU

Data Community Austria

Platinum

smart casual datadesign

redgate

TIMEXTENDER

Gold

b.telligent
smart data. smart decisions.

Lucient

Measure Killer

Spot by NetApp

Silver

veeam

Bronze

ACP CUBIDO Digital Solutions

Power BI Camp
powered by linearis.at

Savory DATA

SOLISYON

Nicky van Vroenhoven

powerdobs

Unit Lead Fabric & Power BI

sessionize

/nicky-van-vroenhoven

/NickyvV

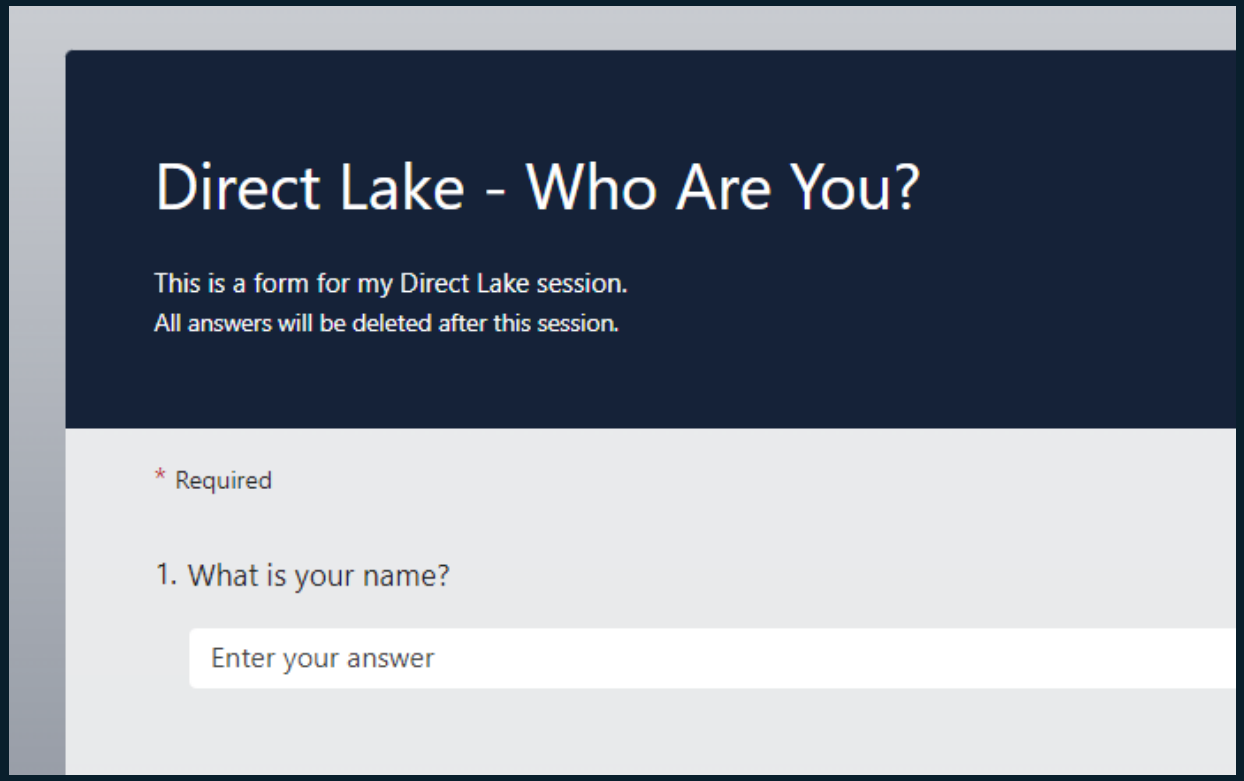
nickyvv.com

/in/nickyvanvroenhoven

Microsoft® Most Valuable Professional

Data Platform MVP

Who Are You?



https://bit.ly/dl-pbi

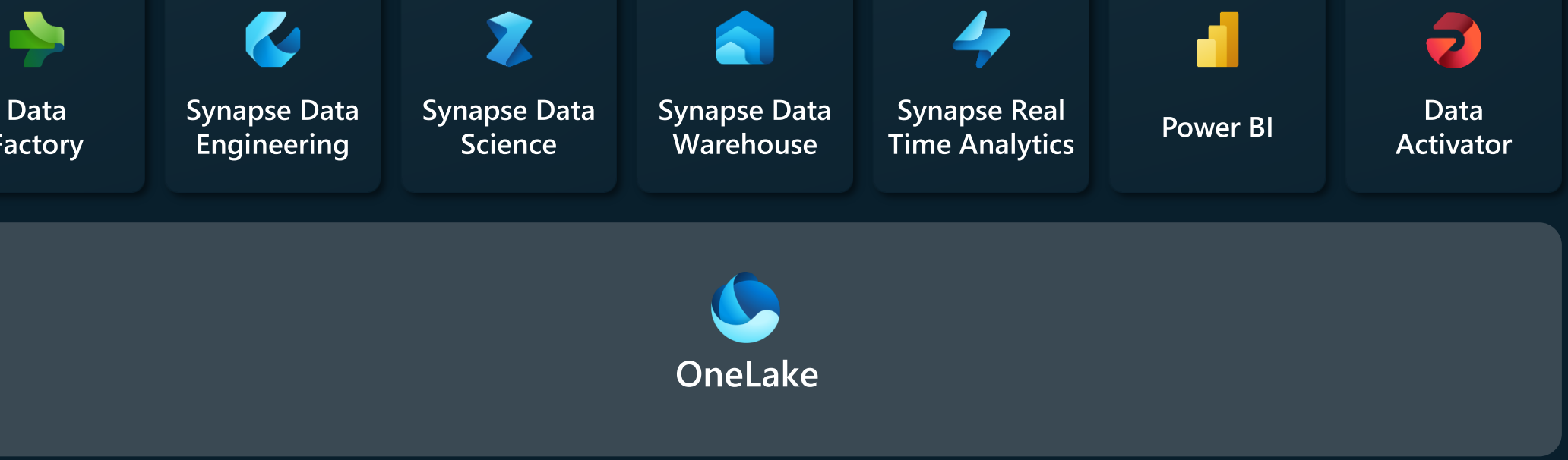
Check out my blog for a how-to:
www.nickyvv.com/2019/12

Today's Objectives

- Introduction to Fabric
- Introduction to Direct Lake
- Parquet, Delta, Z Order
- V-Order
- D-Emo's

Microsoft Fabric

The data platform for the era of AI



Microsoft Fabric

Data analytics 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

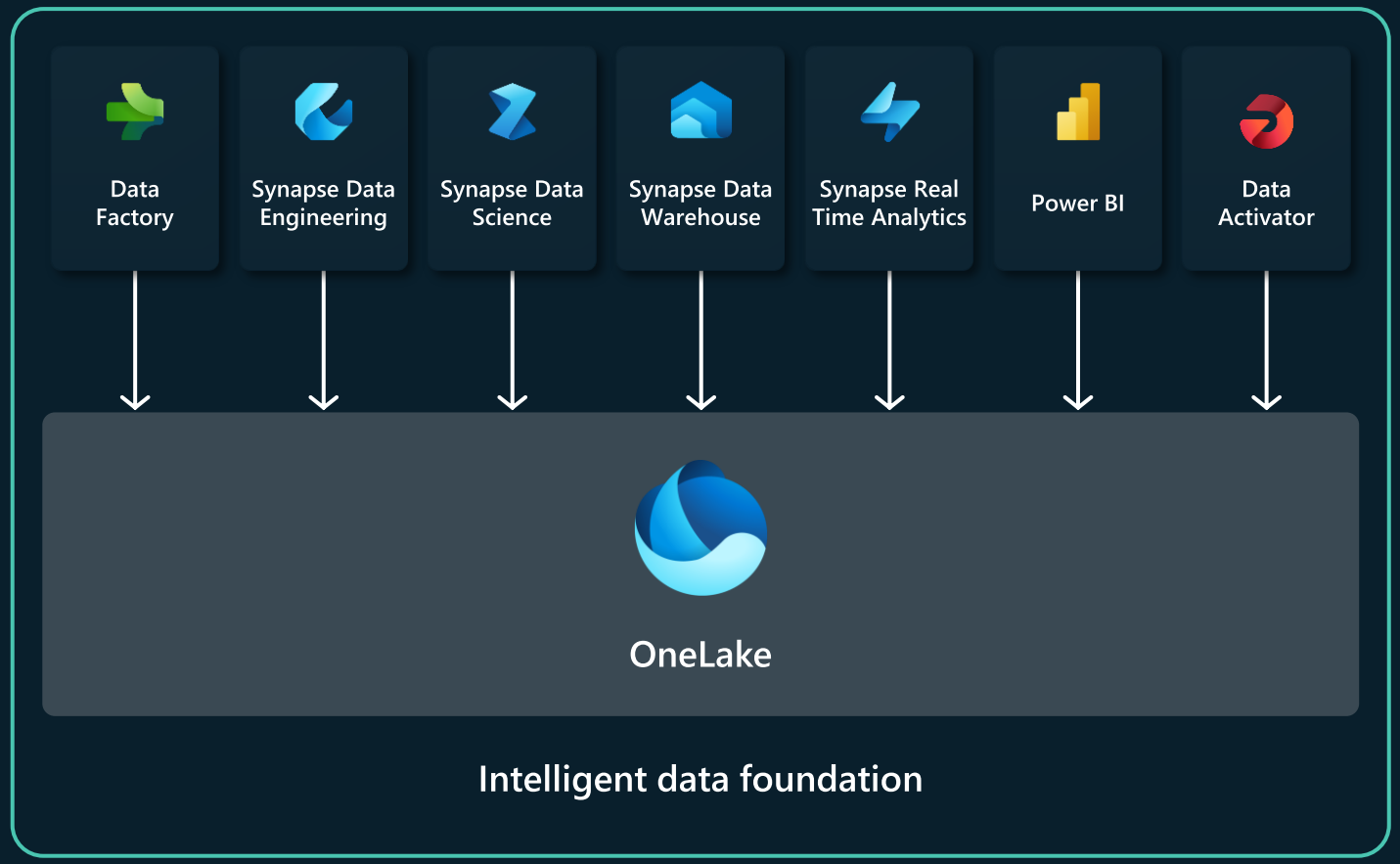
- Familiar and intuitive
- Built into Microsoft 365
- Insight to action

AI Powered

- Copilot accelerated
- Gen AI on your data
- AI-driven insights

OneLake for all Data

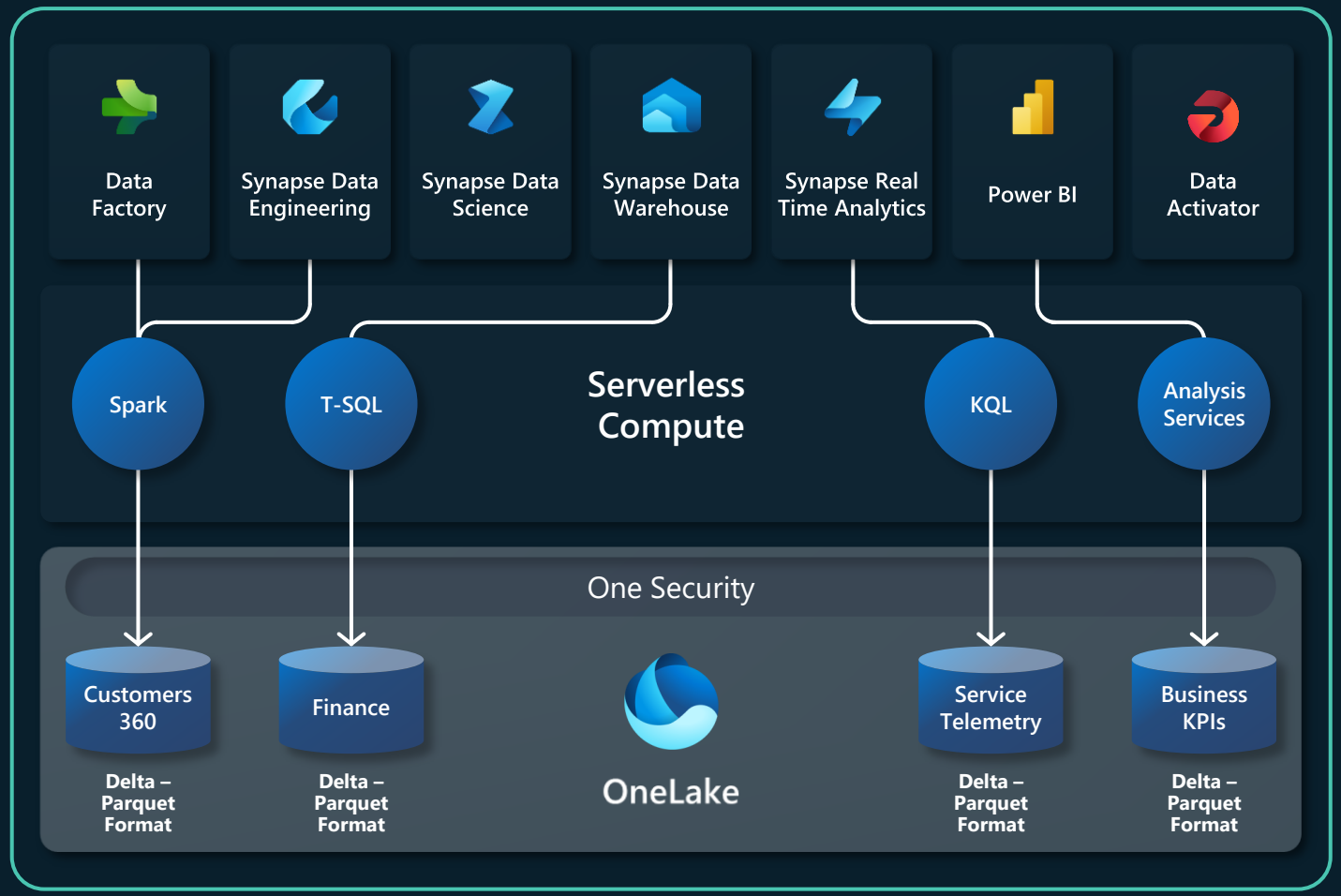
"The OneDrive for Data"



- 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

Universal security makes it real



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 Analytics vNext

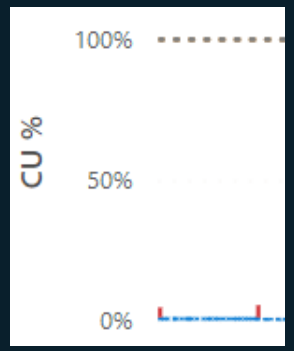
Once data is stored in the lake, it is directly accessible by all the engines without needing any import/export

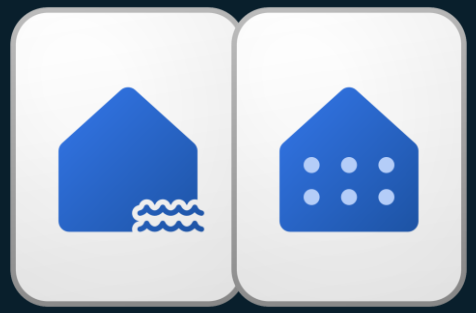
All the compute engines have been fully optimized to work with Delta Parquet as their native format

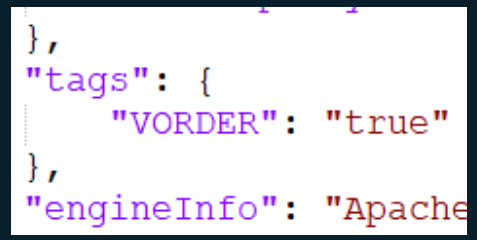
Shared universal security model is enforced across all the engines

Direct Lake prerequisites

1 + 2 + 3 + 4







Active!
Fabric (or P) capacity

Available Consumption Units (CUs)

Storage artefact

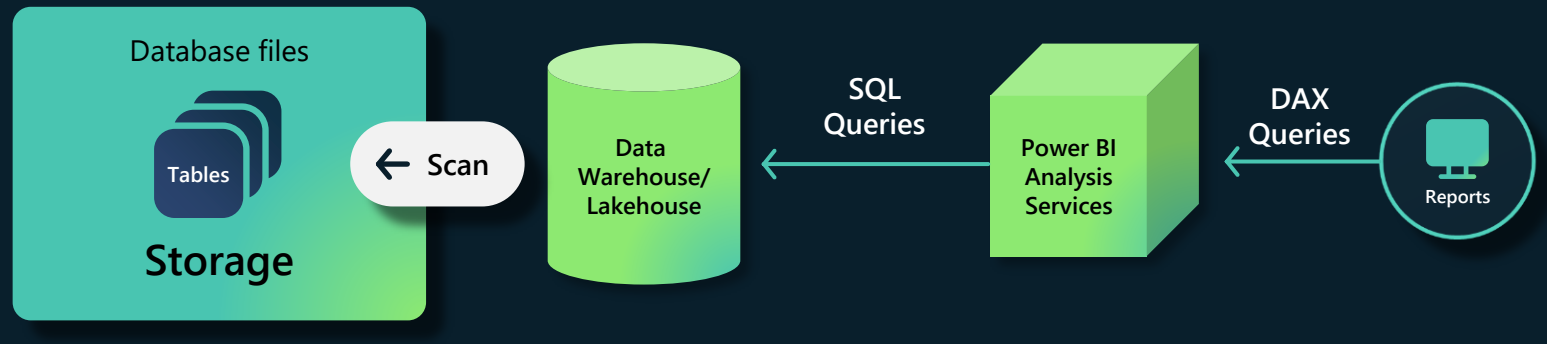
V-Ordered delta tables

Microsoft Fabric

Introducing Direct Lake Mode

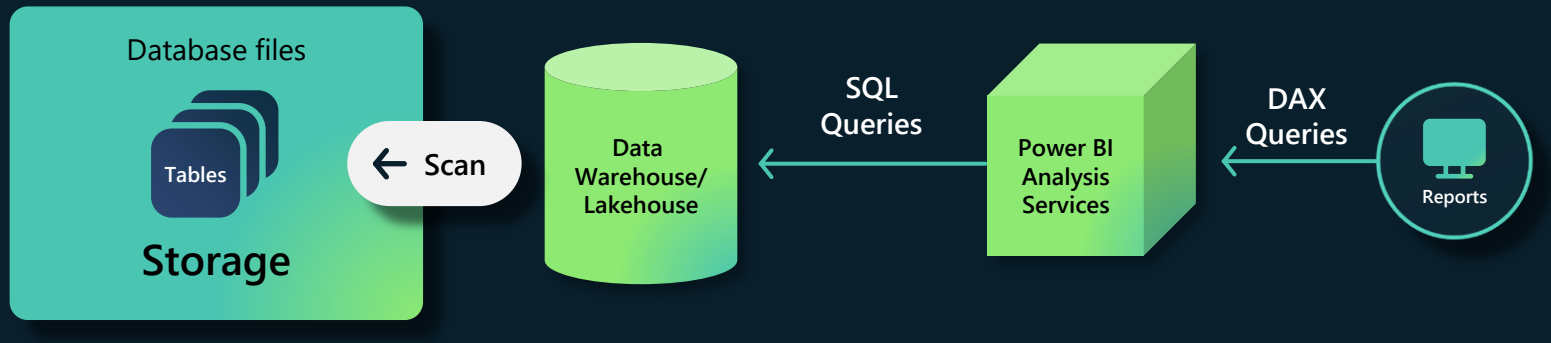
“Direct Query Mode”

Slow, but real time



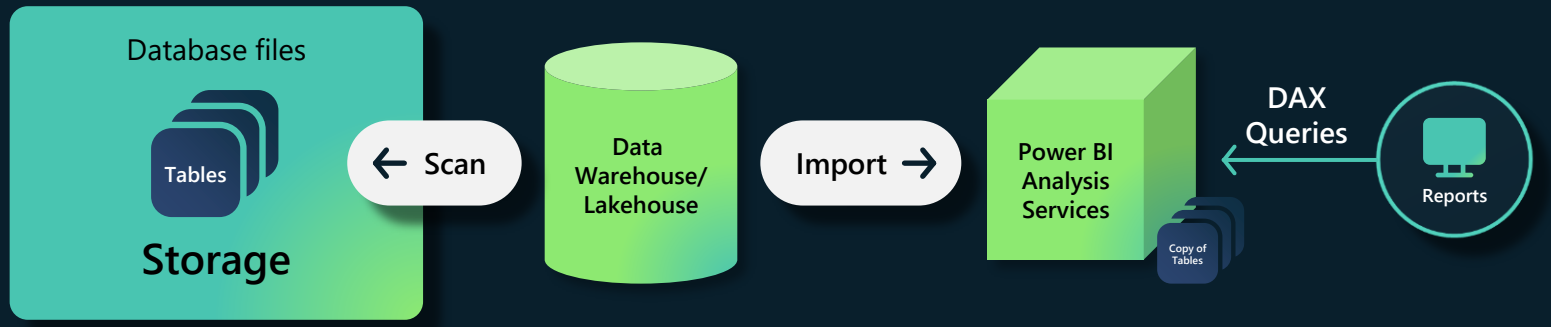
“Direct Query Mode”

Slow, but real time



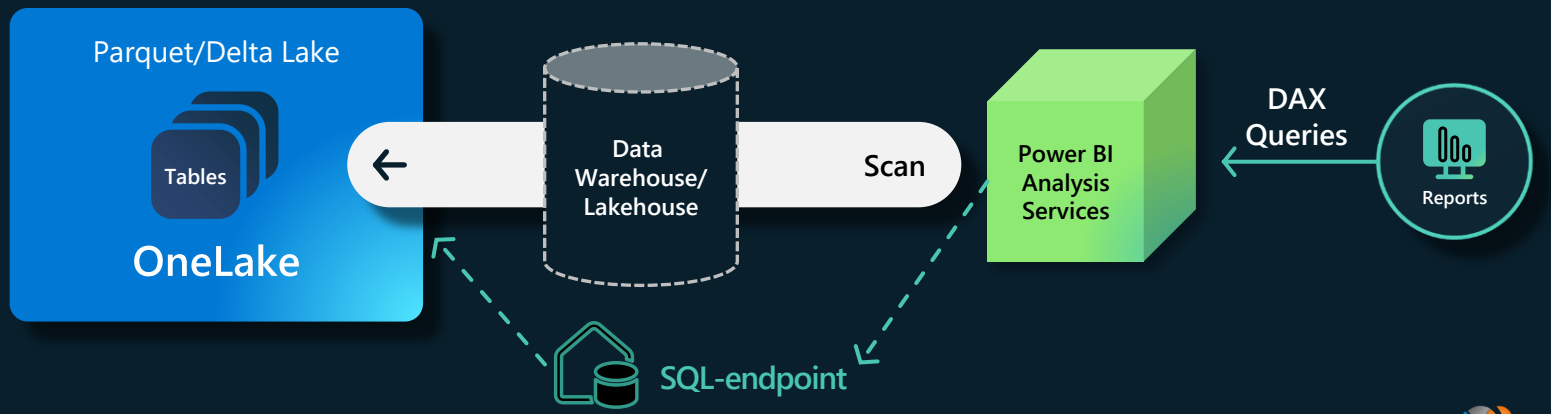
“Import Mode”

Latent & duplicative, but fast



“Direct Lake Mode”

Perfect!



Direct Lake mode

Isn’t that what a

Live connection

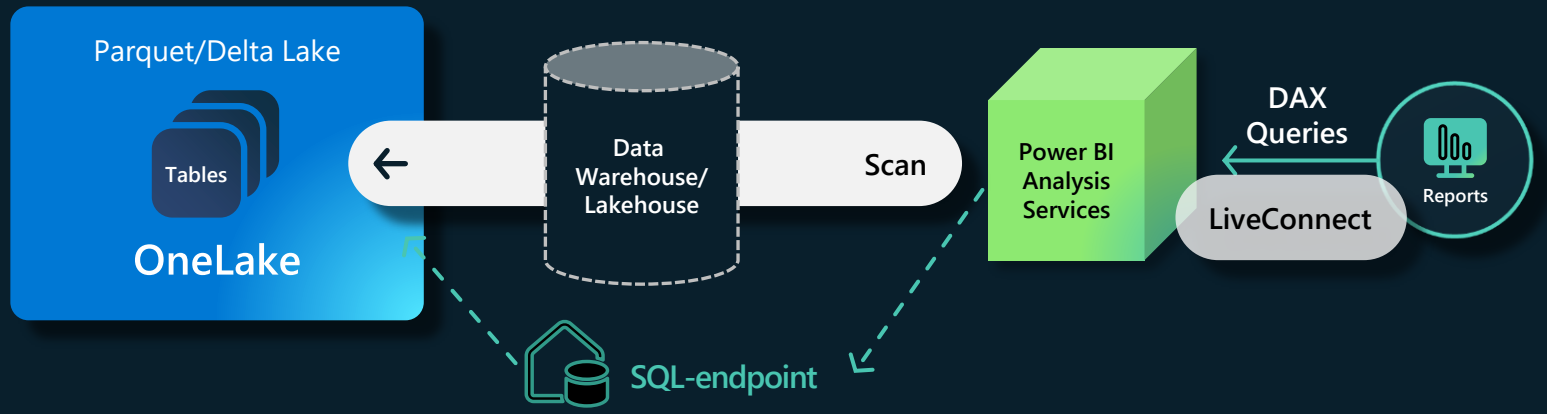
also does

?

Live Connection

- A semantic model that already exists in the Power BI service
- An Azure Analysis Services (AAS) database
- An on-premises instance of SQL Server Analysis Services (SSAS)

So it connects a **report** to a **semantic model**

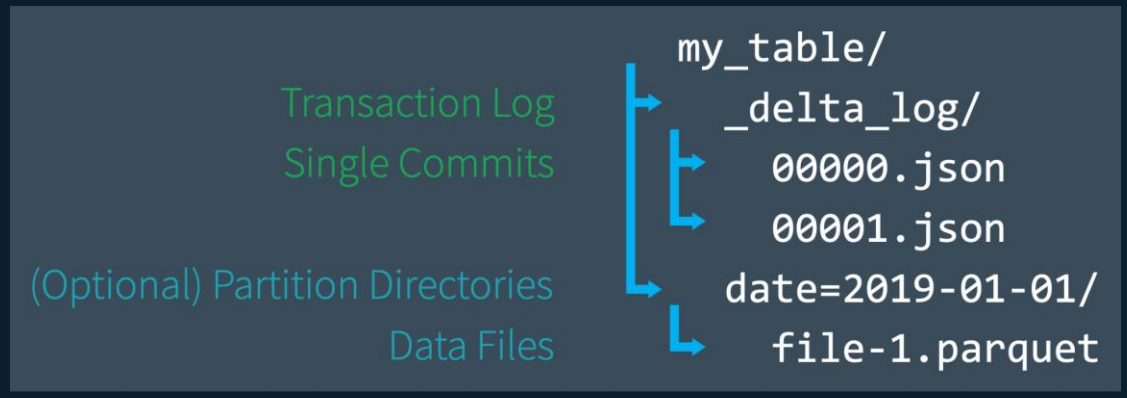


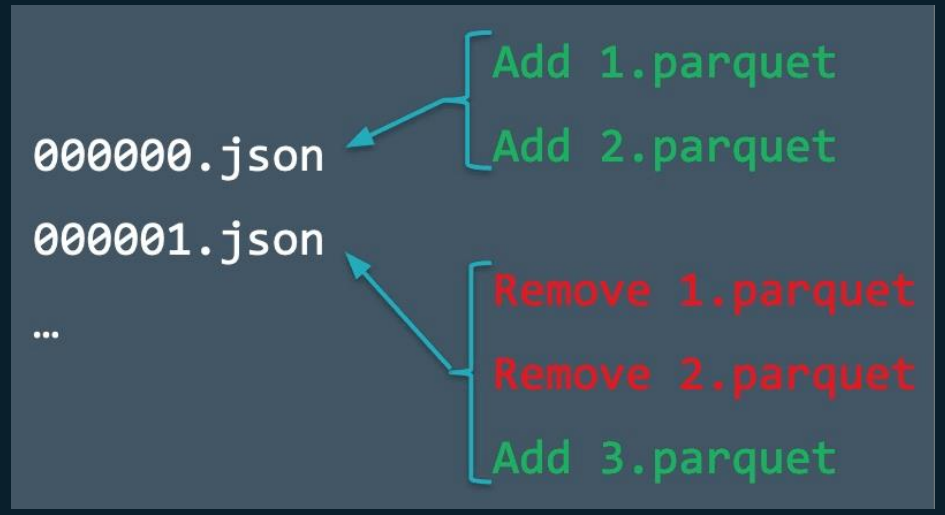
Parquet?

- Binary, columnar file format
- Open industry standard
- Metadata (in file footer) contains schema information
- Efficient data storage and retrieval
- Efficient data compression and encoding
- Thrives on bulk operations
- Well suited for pruning (column, rowgroup elimination)

Delta?

- Also, open industry standard
- Metadata statistics in transaction log
- Supports ACID transactions & Time-travel
- Optimized for querying and pruning
- OPTIMIZE & VACUUM
- Z-Order indexing





Direct Lake mode

On start, no data is loaded in-memory

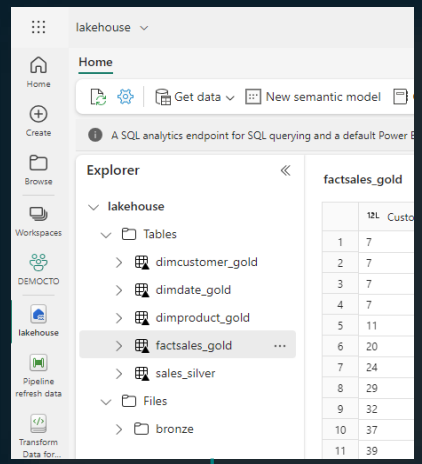
Column data is transcoded from Parquet files when queried

Tables can have mix of transcoded (resident) and non-resident

Column data can get evicted over time

Direct Lake fallback to Direct Query for *reasons*

“Framing” of dataset determines what gets loaded from DeltaLake



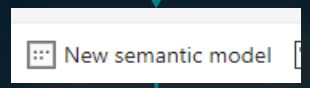
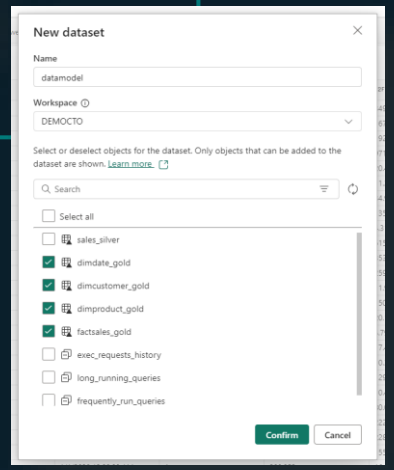
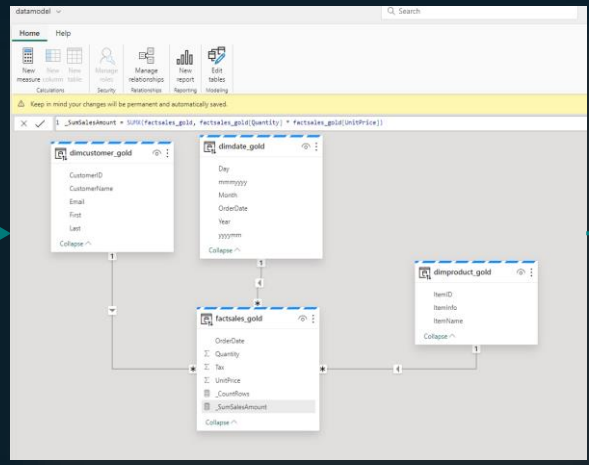
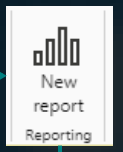


Table selection

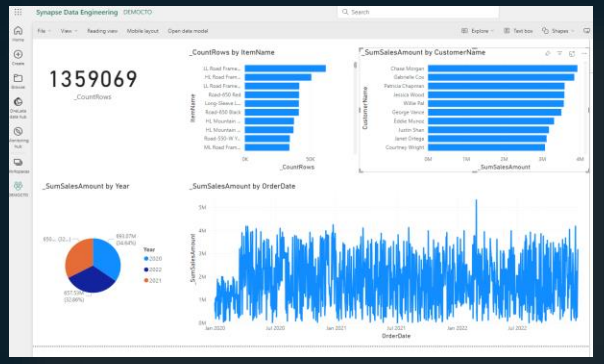


Edit model : create relationships and measures





Report creation



V-Ordering

- Write time optimization to parquet
- 100% open-source format compliant
- Special sorting, row group distribution, dictionary encoding, and compression
- Other Delta options are compatible with V-Order
- 15% slower write times, but up to 50% more compression and 10-50% faster read times

Delta-parquet transaction log

V-Ordering in action

Microsoft Internal DB (162 tables)

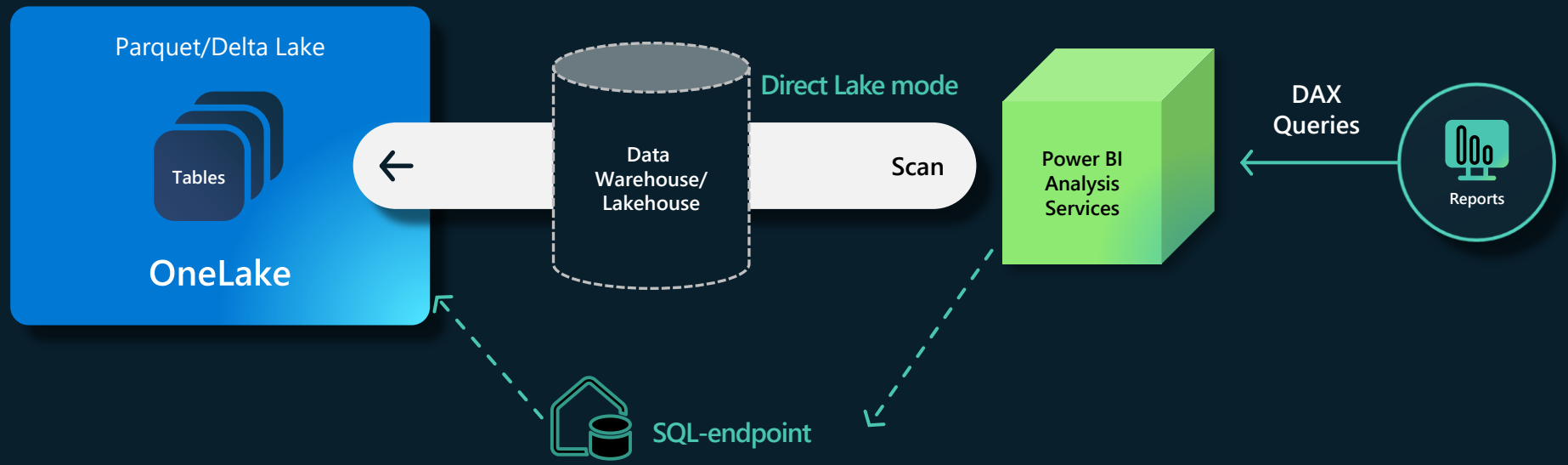
x3.2

Reduced IO for workloads

Direct Lake Fallback

Uses DirectQuery via SQL-Endpoint

- Unsupported features
- Resource limits
  - # Parquet files
  - # Rowgroups per table
  - # Rows per table



Direct Lake Fallback

The following table lists both resource guardrails and MaxMemory:

Expand table

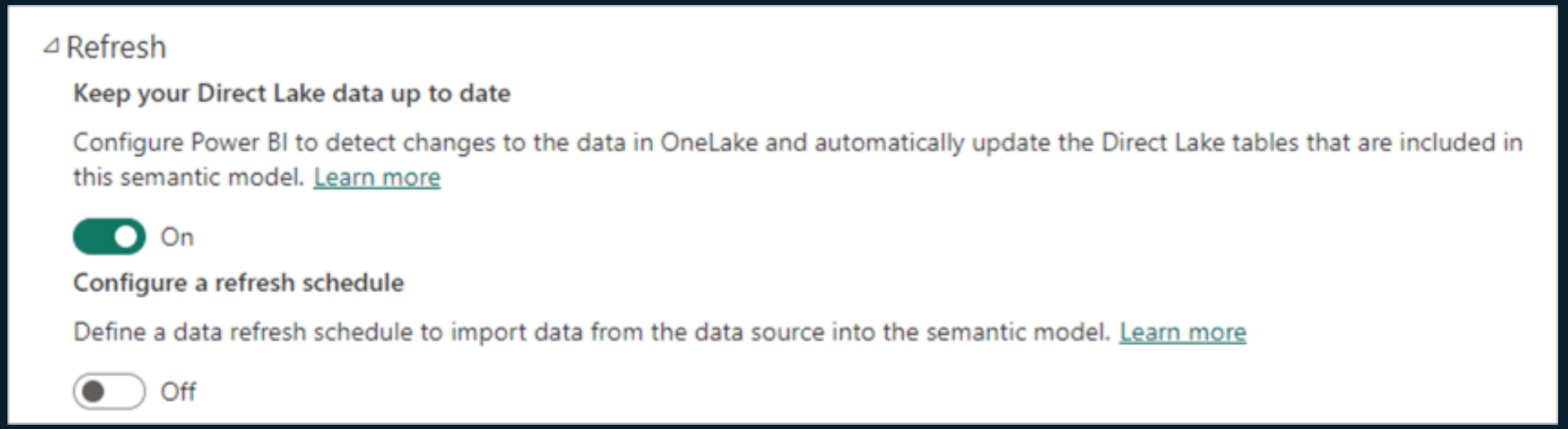
Fabric/Power BI SKUs	Parquet files per table	Row groups per table	Rows per table (millions)	Max model size on disk/OneLake[1] (GB)	Max memory (GB)
F2	1,000	1,000	300	10	3
F4	1,000	1,000	300	10	3
F8	1,000	1,000	300	10	3
F16	1,000	1,000	300	20	5
F32	1,000	1,000	300	40	10
F64/FT1/P1	5,000	5,000	1,500	Unlimited	25
F128/P2	5,000	5,000	3,000	Unlimited	50
F256/P3	5,000	5,000	6,000	Unlimited	100
F512/P4	10,000	10,000	12,000	Unlimited	200
F1024/P5	10,000	10,000	24,000	Unlimited	400
F2048	10,000	10,000	24,000	Unlimited	400

1 - If exceeded, Max model size on disk/Onelake will cause all queries to the model to fallback to DirectQuery, unlike other guardrails that are evaluated per query.

Depending on your Fabric or Power BI SKU, additional **Capacity unit** and **Max memory per query** limits also apply to Direct Lake models. To learn more, see Capacities and SKUs.

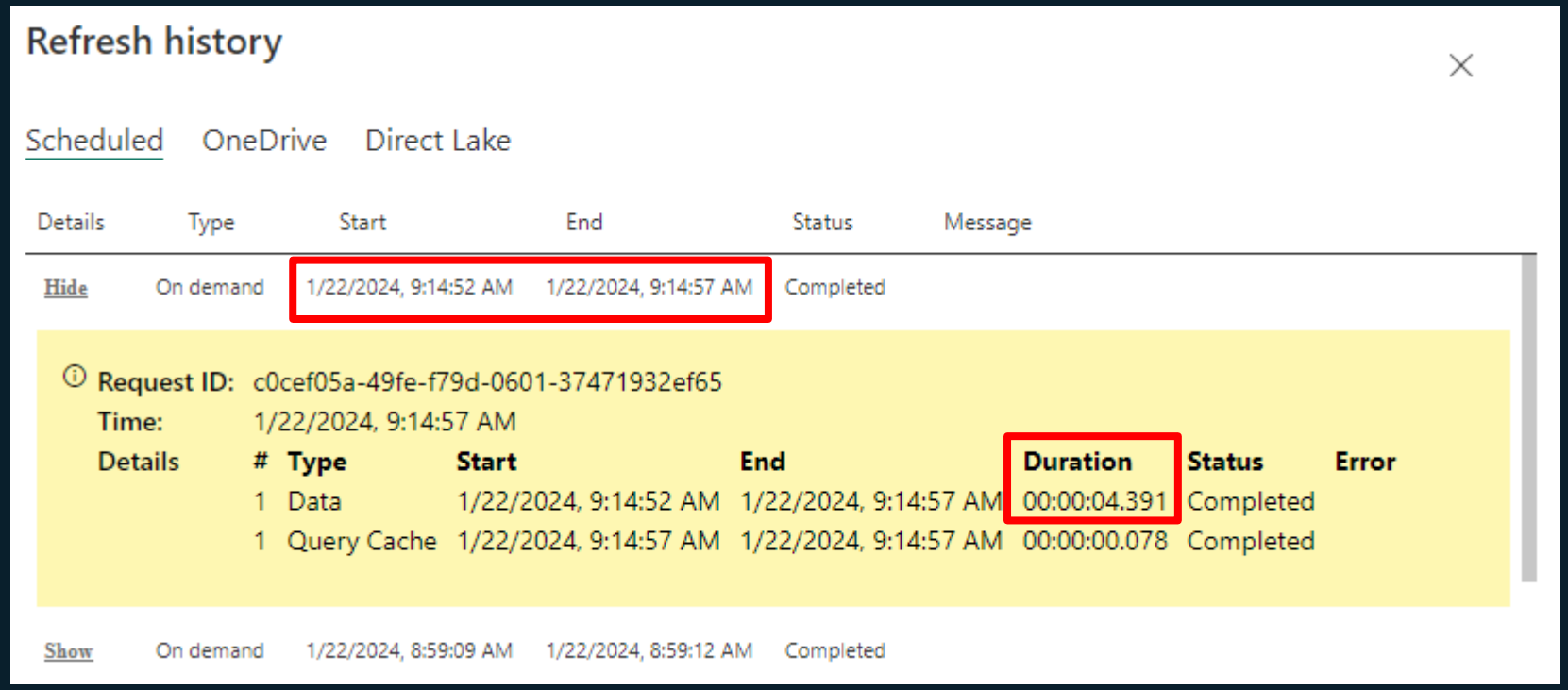
https://learn.microsoft.com/power-bi/enterprise/directlake-overview#fallback

Direct Lake Refresh



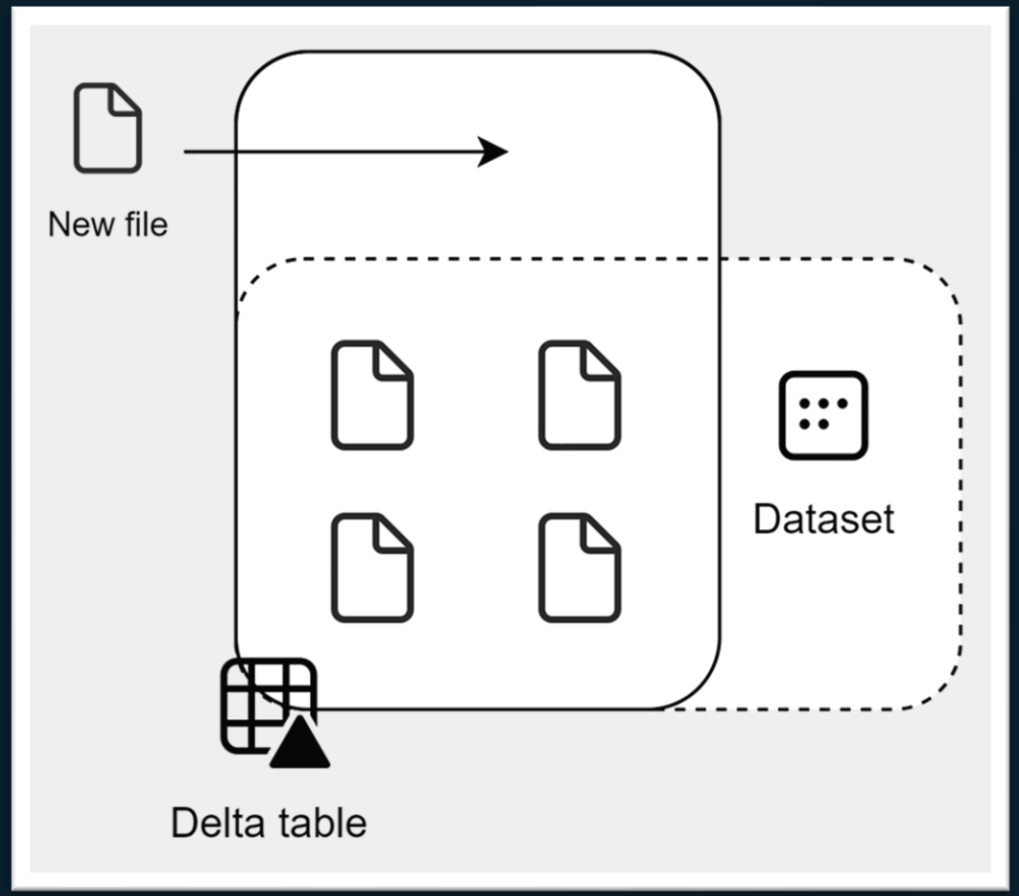
Direct Lake Framing

- Acts like a “Refresh” operation, but only metadata
- Determines the set of .parquet files to use
- **Clears everything currently in memory!**

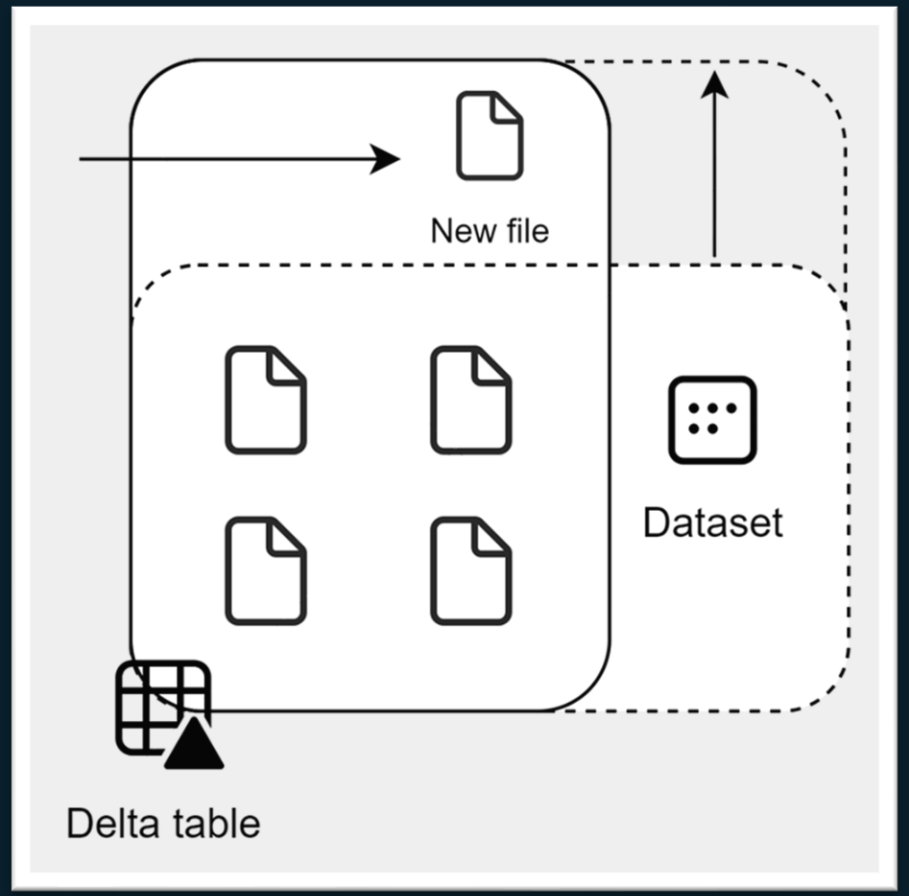


Framing – how it works

Initial state



"Refreshed" state



Credits of pictures: https://data-marc.com/

Dictionary Temperature

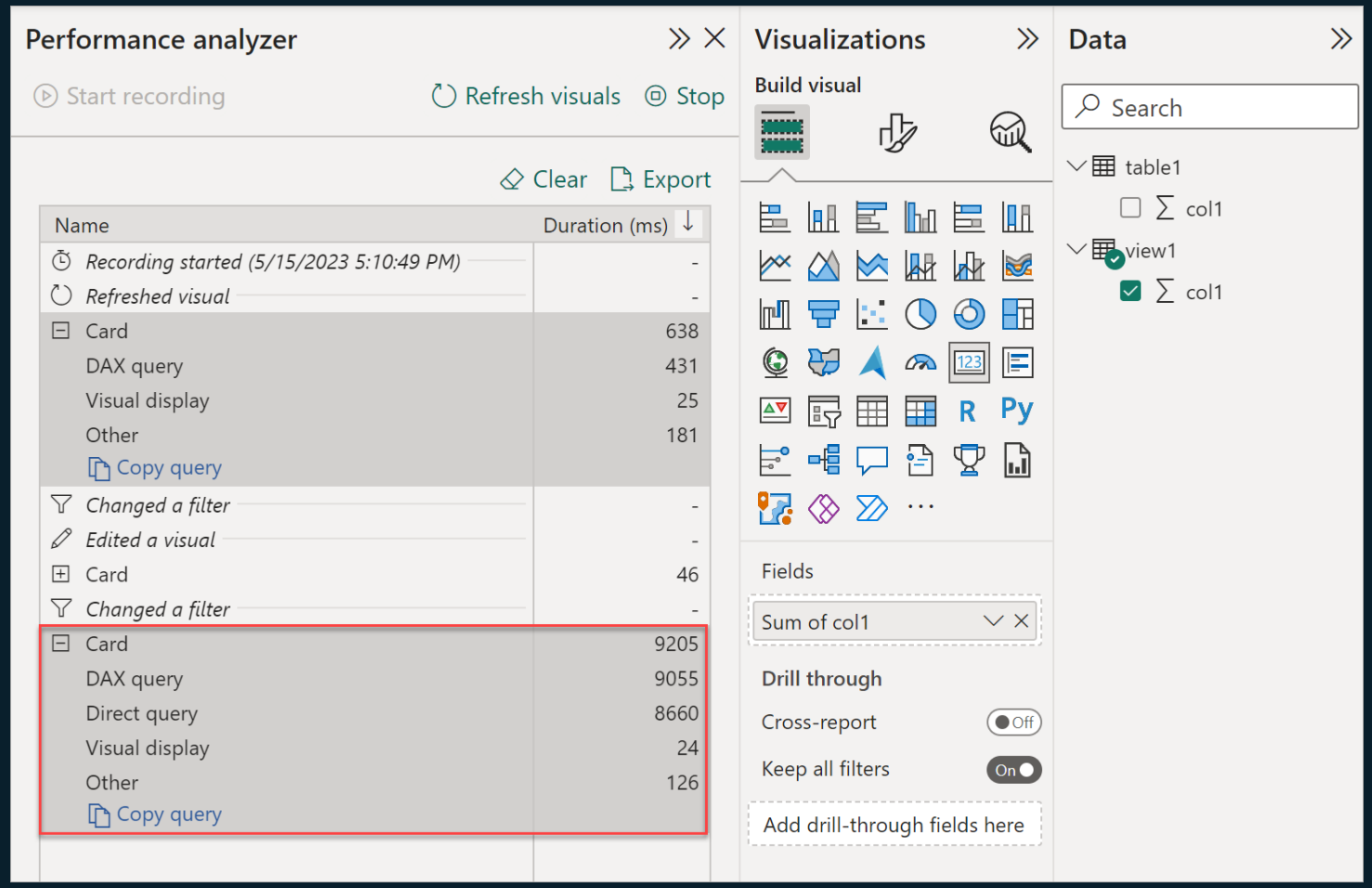
- Can be checked with DAX Studio
- Shows the “activity” /column (/segment)

Log **Results** History

DIMENSION_NAME	COLUMN_ID	DICTIONARY_SIZE	DICTIONARY_TEMPERATURE	DICTIONARY_LAST_ACCESSED
InternetSales	SalesAmount (67)	2640	14,2296194711652	28-9-2023 19:54:45
Product	Color (35)	17308	3,35256194575053	28-9-2023 19:48:26
InternetSales	ProductKey (55)	5032	0,60062313079834	28-9-2023 19:48:26
Product	ProductKey (29)	19112	0,60062313079834	28-9-2023 19:48:26
Product	EnglishProductName (33)	0		
Product	StandardCost (34)	0		
Product	ProductAlternateKey (30)	0		

More info: https://data-marc.com/2023/09/28/

Analyze query processing for Direct Lake models



https://learn.microsoft.com/power-bi/enterprise/directlake-analyze-qp

(Common) Best practices still apply!

@KoVer

(Common) Best practices still apply!

Roche's Maxim

"Transform data
as far upstream as possible,
and **as far downstream as necessary**"

- Matthew Roche (Fabric CAT Microsoft)

Best practices still apply!

- There are no DAX limitations in Direct Lake
  - *But DQ on the other hand…*
- Only include necessary tables/columns
- Fewer, larger parquet files is better
  - *Small file-problem* > OPTIMIZE
  - Table Maintenance options in a Lakehouse

Be aware of current limitations

- T-SQL-based views will fallback to DQ
- No calculated columns/tables
- No composite models with Direct Lake yet

https://learn.microsoft.com/en-us/power-bi/enterprise/directlake-overview

External tools support for Direct Lake

- Enable XMLA read-write on the capacity
- DAX Studio (DMV's)
- Creating a Direct Lake dataset with Tabular Editor 3.0 (paid)
  - Unable to edit in the service afterwards!

    Direct Lake models created or modified by using XMLA-based tools cannot be opened in the Web modelling feature
  - Possible actions in TE3
    - Create a Direct Lake model
    - Create tables
    - Preview data
    - Add tables (to an existing DL model)

Fabric Capacities - How it works

COMPUTE

Capacity units (CUs) = Compute power

Capacity units (CUs) are units of measure that represent a pool of compute power needed. Compute power is required to run queries, jobs, or tasks.

CU Consumption

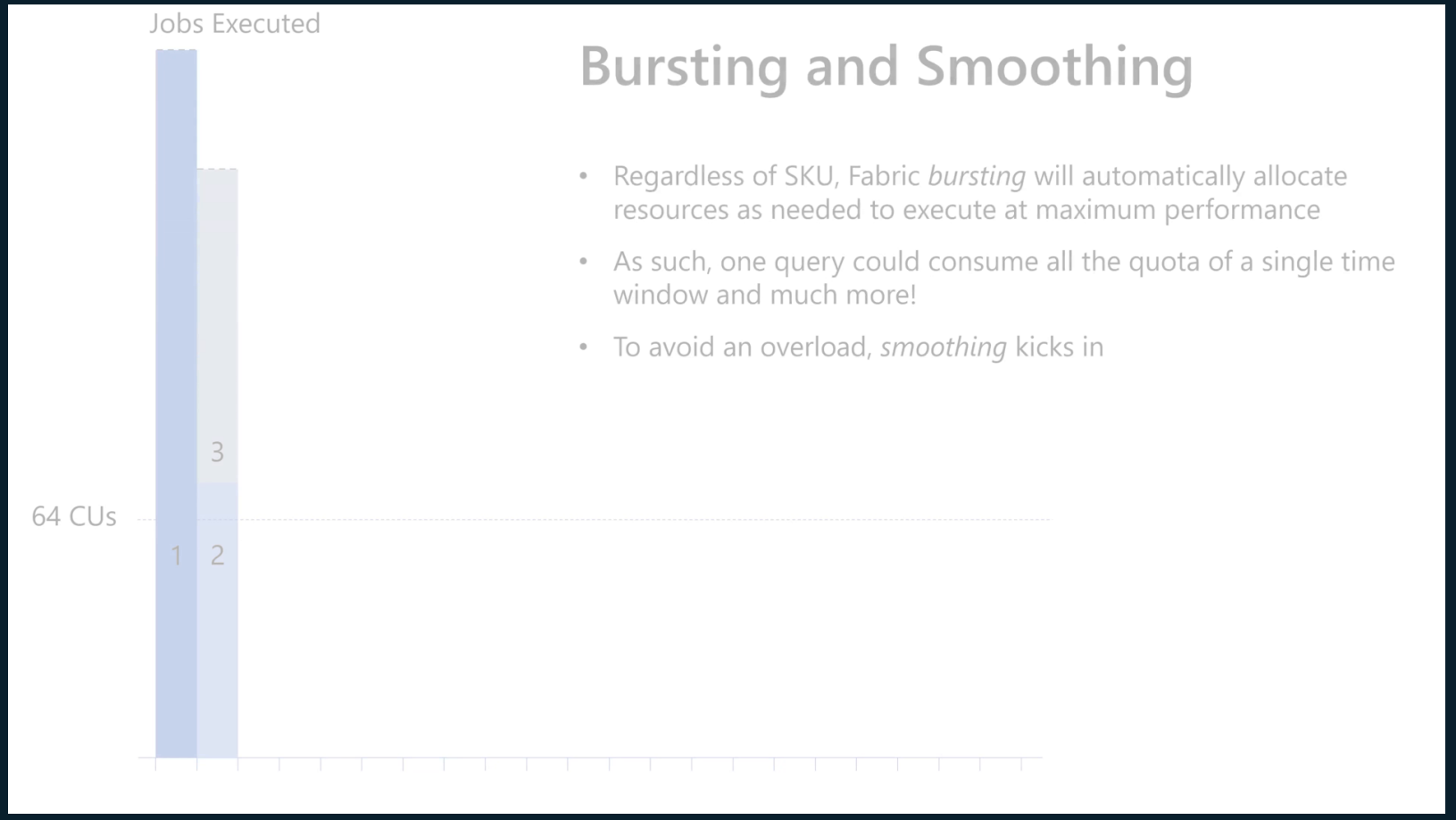
The CU consumption is highly correlated to the underlying compute effort needed for the tasks performed by the capability during the processing time.

Each capability, such as Power BI, Spark, Data Warehouse, with the associated queries, jobs, or tasks has a unique consumption rate.

Note: For additional documentation on purchasing experience and capacity models, please see here: aka.ms/fabriclicensing

Unified business model

Pay for the average, not the peak



Pause a capacity

Usable artifacts

Created by nickyvv.com With a paused capacity	Access	Refresh	Download/ Save copy
Direct Lake in Fabric workspace	No	No	Yes
Direct Lake in other workspace	No[1]	No	Yes
DirectQuery in Fabric workspace	No	No	Yes
DirectQuery in other workspace	No	No	Yes
Import in Fabric workspace	**No**	No	**Yes**
Import in other workspace	Yes	No	Yes
Capacity Metrics App	Yes	No	NA
(New) Usage Metrics report	No	No	No

https://www.nickyvv.com/2023/09/

OneLake pricing

OneLake is a data lake built into Microsoft Fabric and provides a single place to store all organizational data. Data storage is charged at a rate of $ per GB per month and priced uniquely across regions.

Data Storage

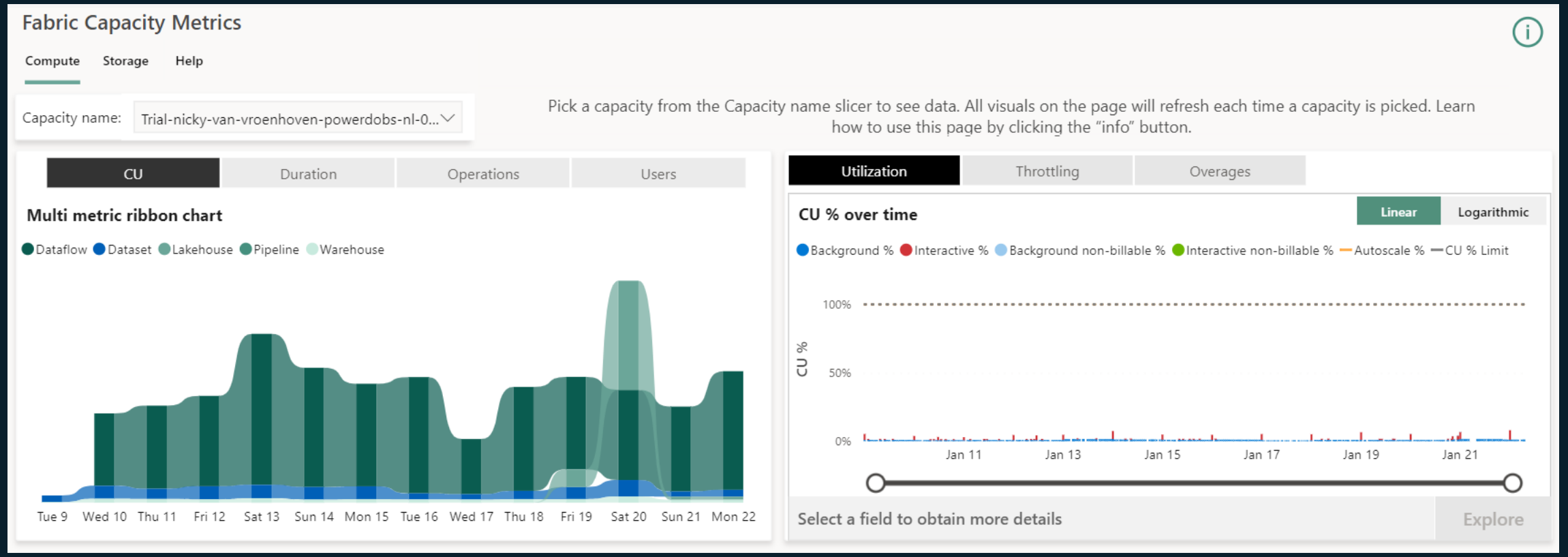
Type	Pay-as-you-go price at West US 2
OneLake storage	$0.023 per GB / month
OneLake BCDR storage	$0.0414 per GB / month
OneLake cache	$0.2 per GB / month

Data Transfer & Internet Egress

Cross-region data transfer network charges may apply based on source/destination of each storage access. Learn more at the Bandwidth Pricing.

Note: OneLake cache is billed for KQL cache storage and Data Activator data retained.

Fabric Capacity Metrics App



Direct Lake recap

- Auto-managed by Fabric
- Delta, Parquet, V-Order
- Fallback > F-SKU
- Use best practices
- Check limitations

New Microsoft Credentials

Microsoft Certifications

Microsoft Certified: Fabric Analytics Engineer Associate

Exam DP-600: Implementing Analytics Solutions Using Microsoft Fabric

In BETA *now*

Microsoft Applied Skills

Fabric credentials

Covering scenarios like real-time analytics, data lakehouses, and data warehouses using Microsoft Fabric

NEW in early 2024

Next steps

Try Microsoft Fabric
aka.ms/try-Fabric

Explore the Fabric Roadmap
aka.ms/FabricRoadmap

Evaluations, evaluations...

https://evals.datagrillen.com/evals_vienna.aspx

Resources

- https://learn.microsoft.com/power-bi/enterprise/directlake-overview
- https://learn.microsoft.com/power-bi/enterprise/directlake-fixed-identity
- https://learn.microsoft.com/en-us/power-bi/enterprise/directlake-analyze-qp
- https://learn.microsoft.com/power-bi/enterprise/directlake-overview#fallback
- https://blog.crossjoin.co.uk/2023/10/08/what-does-it-mean-to-refresh-a-direct-lake-power-bi-dataset-in-fabric/
- https://data-marc.com/2023/10/09/exploring-direct-lake-framing-and-warm-up-data-using-semantic-link-in-fabric-notebooks/
- https://fabric.guru/power-bi-direct-lake-mode-frequently-asked-questions
- https://blog.tabulareditor.com/2023/09/26/fabric-direct-lake-with-tabular-editor-part-2-creation/
- https://www.mssqltips.com/sqlservertip/7894/power-bi-direct-lake-mode-in-microsoft-fabric/

Nicky van Vroenhoven

powerdobs

sessionize

Unit Lead Fabric & Power BI

/nicky-van-vroenhoven

/NickyvV

nickyvv.com

/in/nickyvanvroenhoven

Microsoft® Most Valuable Professional

Data Platform MVP

Evaluations, evaluations...

https://evals.datagrillen.com/evals_vienna.aspx