Fabric Direct Lake Deep Dive

Chris Webb

Fabric Customer Advisory Team

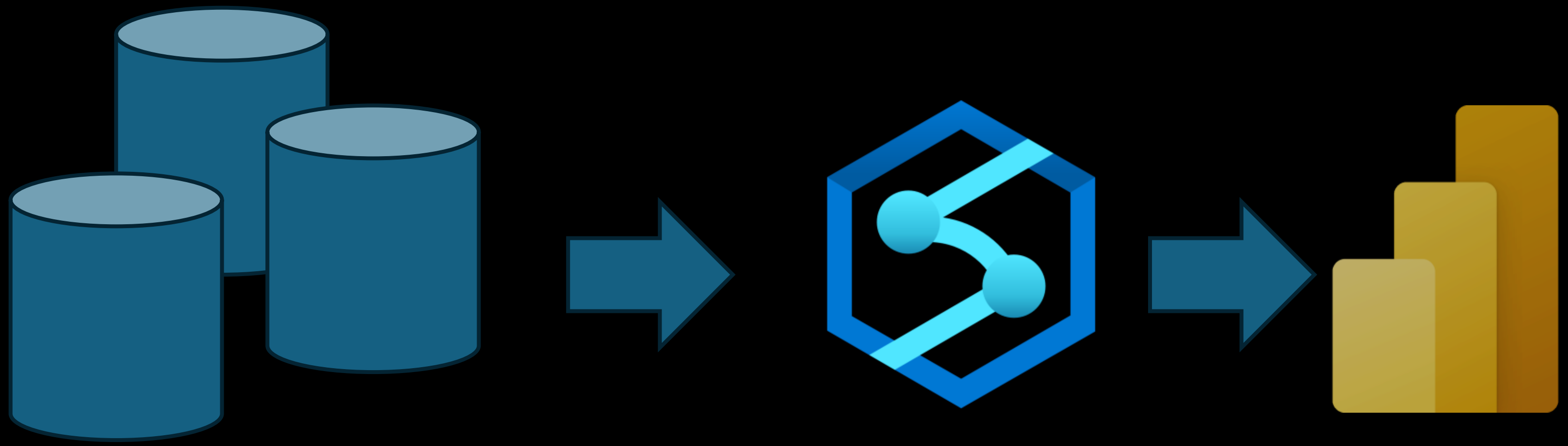
Microsoft

https://blog.crossjoin.co.uk/

https://twitter.com/cwebb_bi

What Is Direct Lake?

Traditional Microsoft BI Architecture



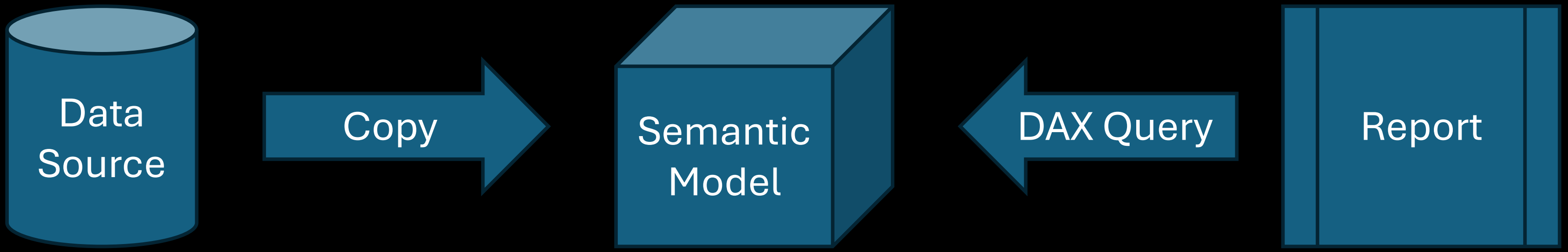
Source Data

Relational Serving Layer (eg Synapse, Azure SQL DB, Snowflake, BigQuery etc)

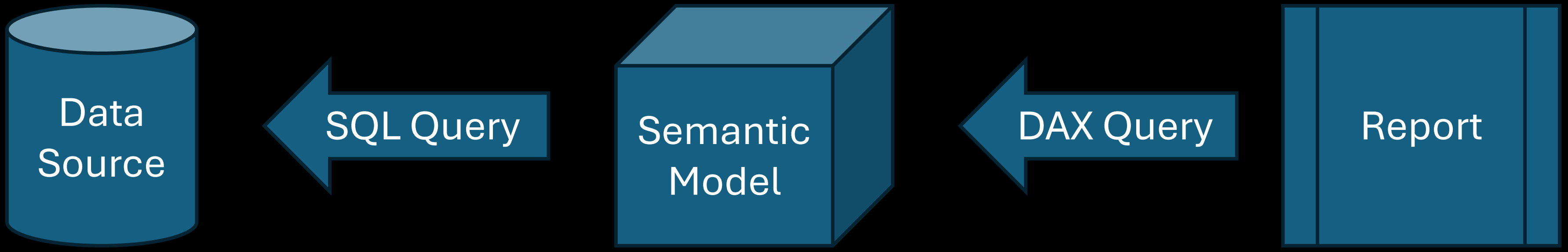
Power BI

Import mode and DirectQuery mode

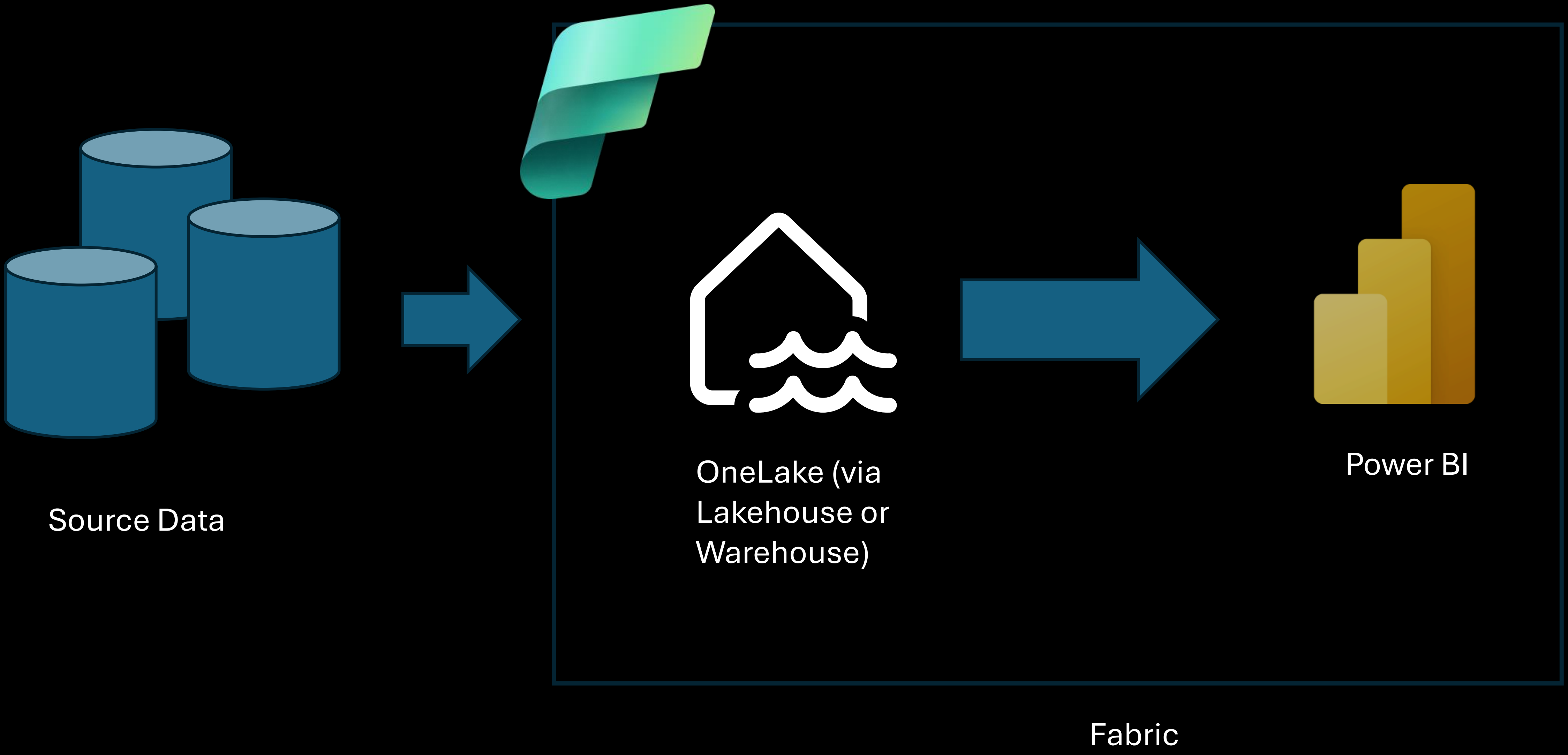
Import Mode:
Fast query performance
Refresh can be slow



DirectQuery Mode:
No refresh
Query performance
can be slow

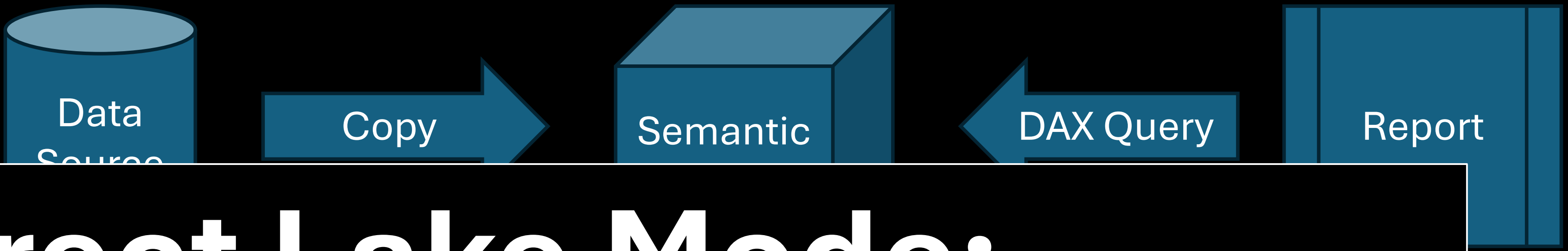


Fabric Data Lake Architecture



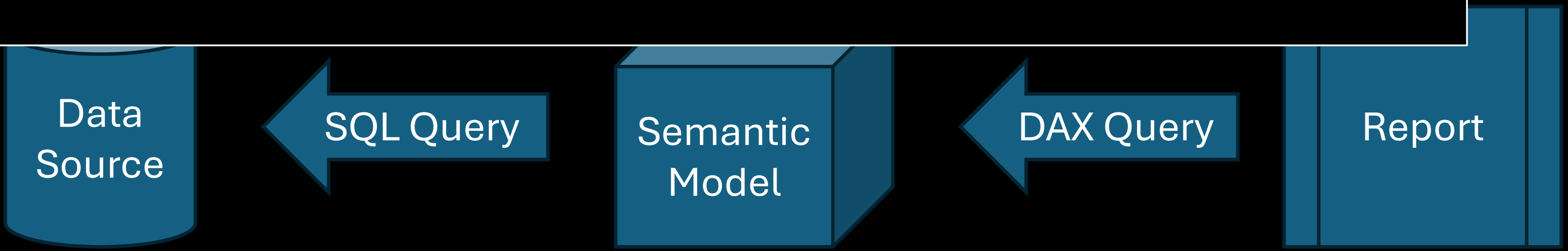
Direct Lake mode

Import Mode:
Fast query performance
Refresh can be slow



Direct Lake Mode:
Fast query performance!
No refresh*!

DirectQuery Mode:
No refresh
Query performance
can be slow



Direct Lake mode

- Direct Lake is a new third data storage mode for Power BI semantic models in Fabric
- It gives you the performance of Import mode and the latency of DirectQuery mode
- Queries run on data stored in memory just like Import mode – which is why it’s fast
- The data required by a query is loaded into memory on-demand from OneLake

Direct Lake requirements

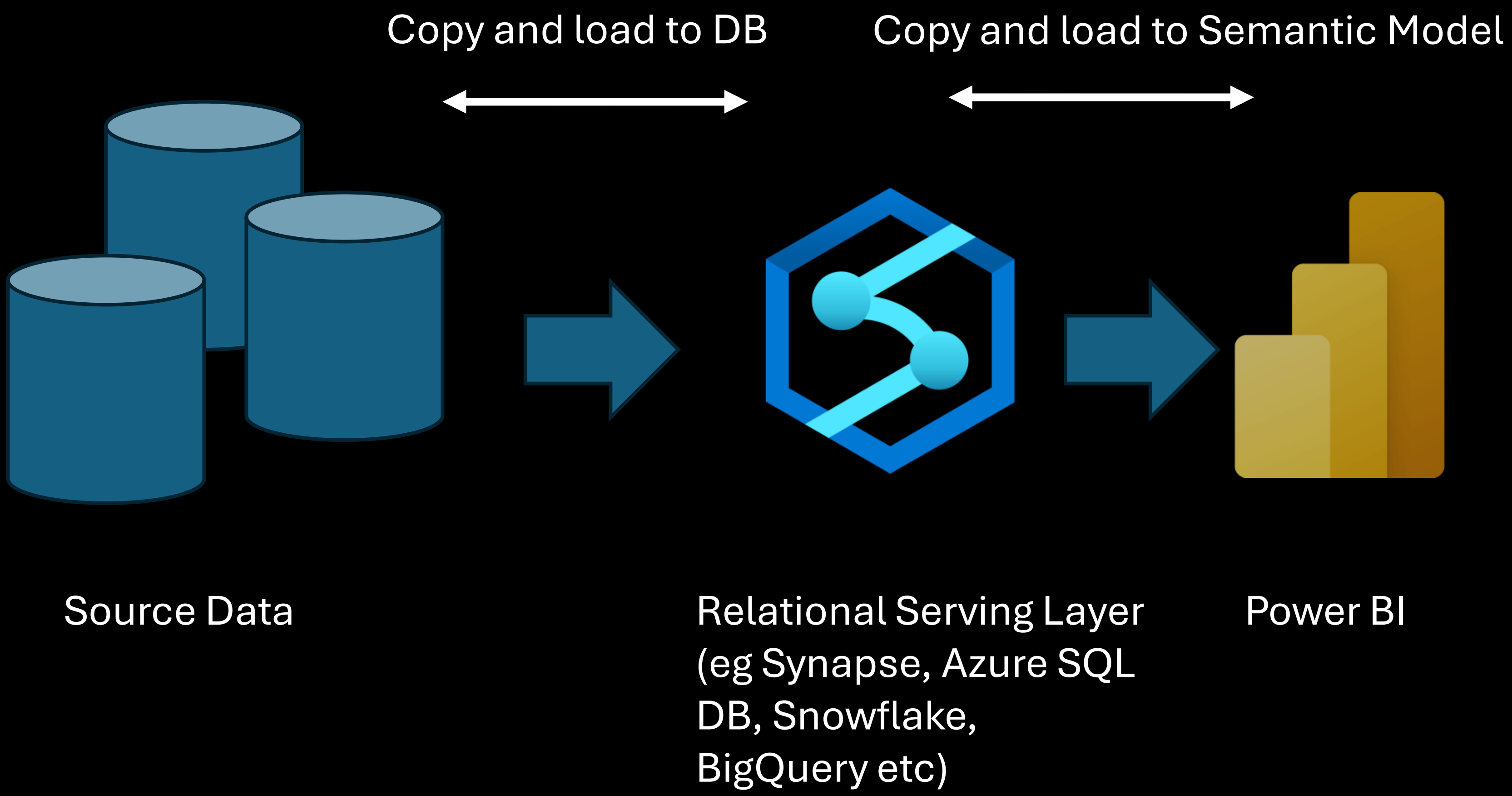
- You need a Fabric/Premium capacity – F or P SKU
  - Free trials are available for most customers
  - The capacity with the semantic model in must not be paused
- Fabric must be enabled for
  - The entire tenant
  - An individual capacity
- You can also limit Fabric to specific security groups
- Other limits and requirements may also prevent its use

Creating Direct Lake semantic models

- A Direct Lake semantic model is automatically created for you but you may want to create custom models too
- You cannot (yet) build or edit Direct Lake semantic models in Power BI Desktop
- Instead you must use either
  - The web editor
  - Tabular Editor 2 or 3
  - TMSL scripts run on the XMLA Endpoint
- Some features like RLS are supported but can’t be created (yet) in the web editor

Direct Lake benefits

How long does Import mode take end to end?



Copy and load to DB

Copy and load to Semantic Model

Source Data

Relational Serving Layer
(eg Synapse, Azure SQL
DB, Snowflake,
BigQuery etc)

Power BI

How long does Direct Lake take end to end?

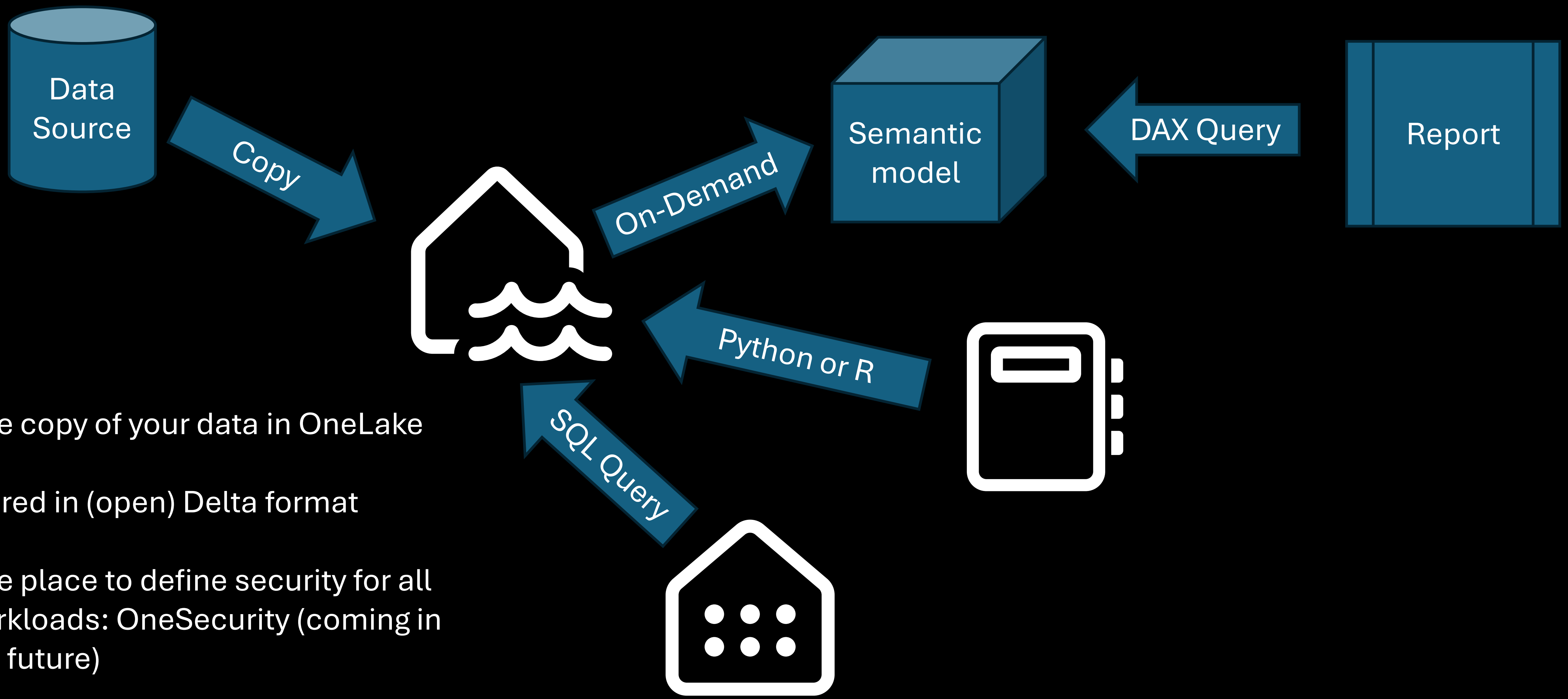
Direct Lake Mode: Faster time to report (maybe!)

Cost, complexity and maintenance

Direct Lake Mode: May remove a layer from your architecture

DB, Snowflake,
BigQuery etc)

OneLake: a single copy of data for everyone



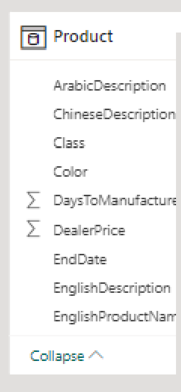
One copy of your data in OneLake

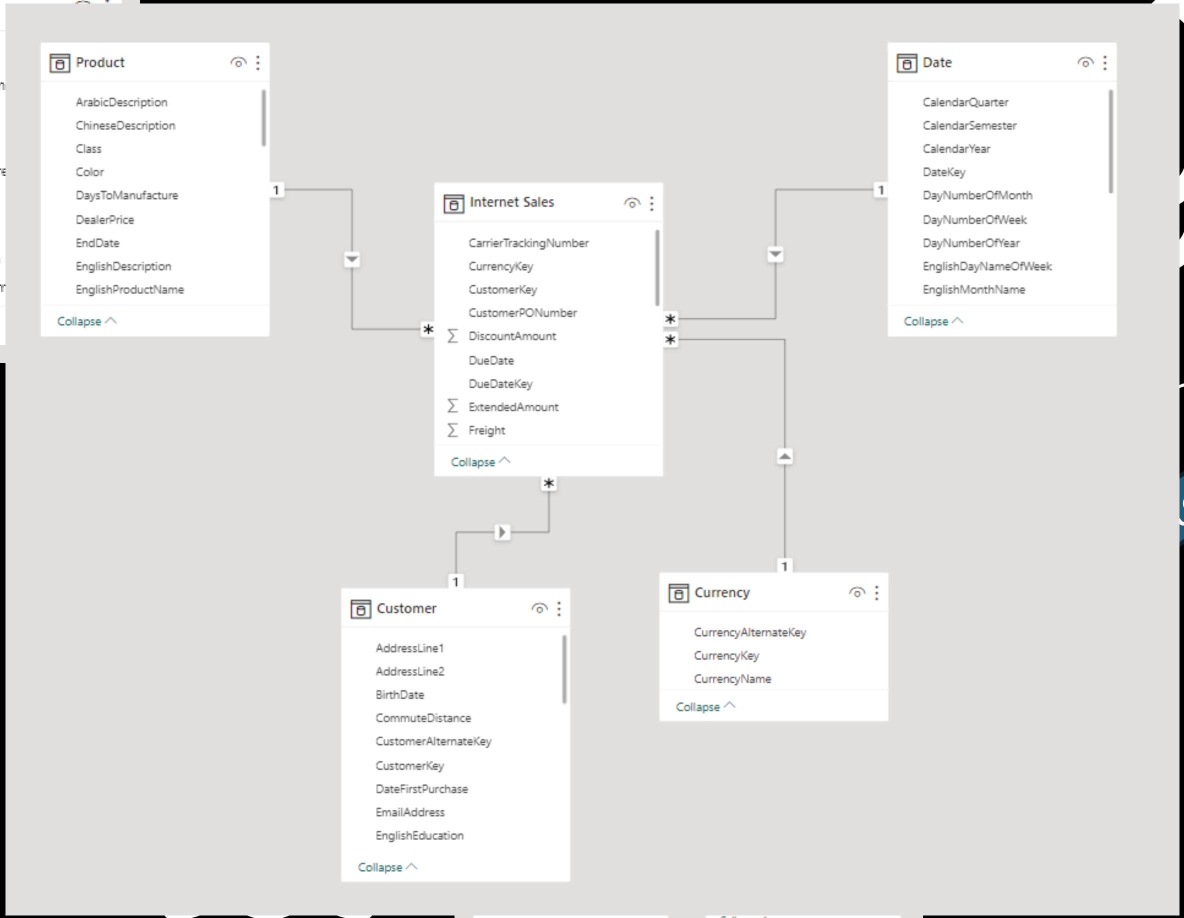
Stored in (open) Delta format

One place to define security for all workloads: OneSecurity (coming in the future)

Direct Lake and table shortcuts

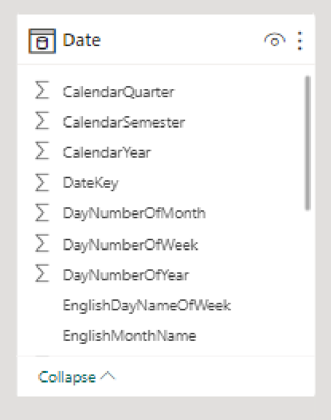
Product dimension





nsion

cut



Direct Lake limitations

Direct Lake limitations (for now)

- All data must come from a single Lakehouse or Warehouse
  - You can use shortcuts to bring data in from other places
- No calculated columns or calculated tables
- No composite models
  - Although calculation groups and field parameters are now allowed
- Can only be used with tables, not views, in a Warehouse
- Can only be used with security defined in the semantic model
- Not all data types supported
  - No structured data types, binary or GUID columns
  - DateTime relationships not supported
  - String length limited to 4000 characters
- No support for hierarchies or Excel drillthrough

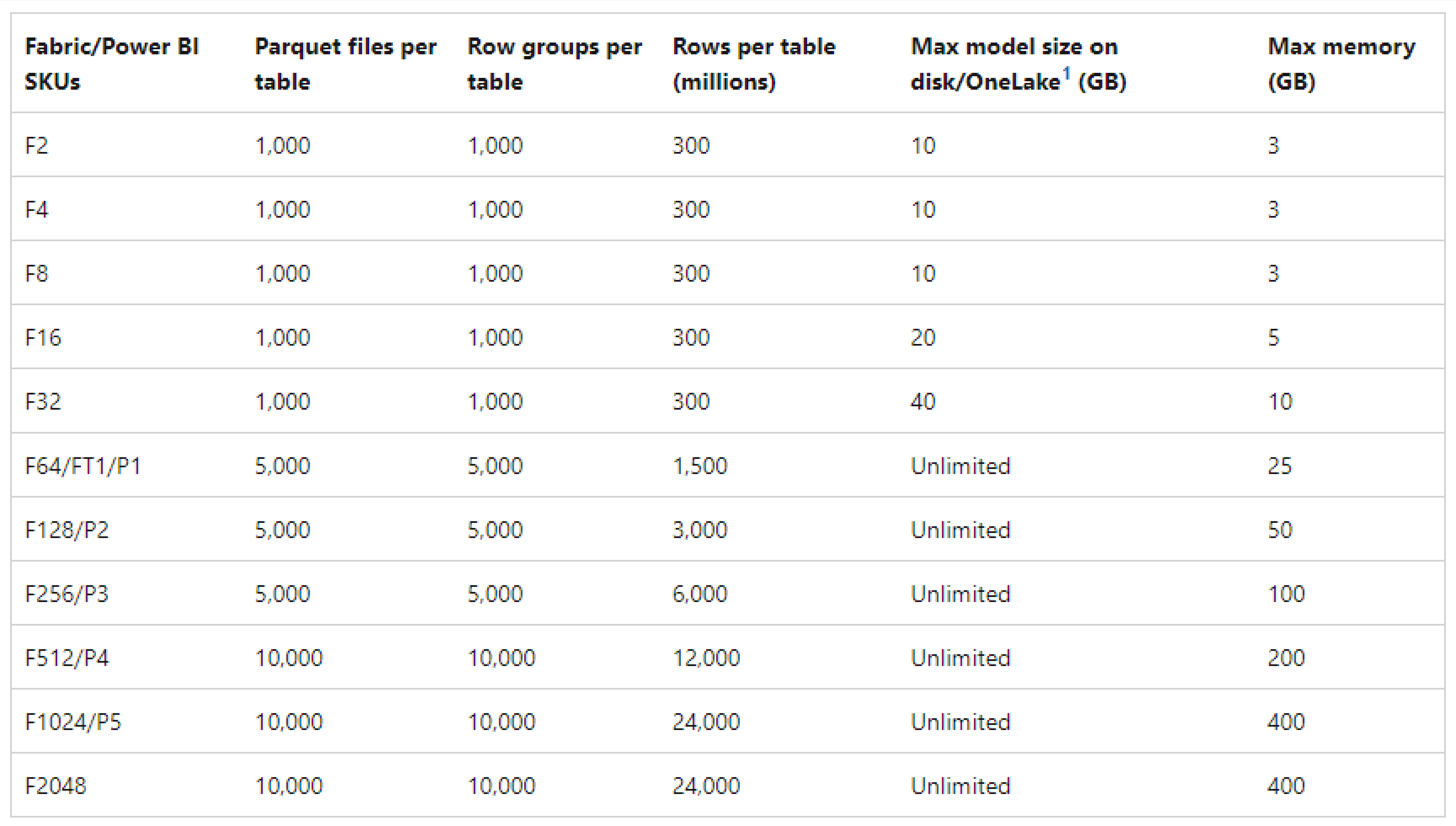
Fallback to DirectQuery – unsupported features

- A semantic model may fall back to DirectQuery mode because you’re using features that prevent Direct Lake
- Views are not allowed because they don’t have corresponding tables stored in a Lakehouse
- If RLS or OLS is defined in a Warehouse, the semantic model has to use DirectQuery to ensure that security is respected
  - In the future OneSecurity will allow you to define security once and apply it to all workloads including Warehouse and Semantic Models

Fallback to DirectQuery – data volumes

- There are limits on how much data can be used with Direct Lake mode
- These limits vary by capacity SKU size
- If you exceed these limits, your semantic model will fall back to DirectQuery mode
  - Query performance will be noticeably worse
- Fabric checks these limits when the semantic model is loaded into memory

Fallback to DirectQuery



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

Detecting fallback to DirectQuery

- Performance Analyzer, Profiler traces and/or Log Analytics will show what happens for individual queries
- The TMSCHEMA_DELTA_TABLE_METADATA_STORAGES DMV shows whether you have used a feature that prevents Direct Lake being used
- Limits on data volumes can be checked with Python notebooks (see Delta Analyzer from Phil Seamark) and in some cases DMVs

Controlling fallback to DirectQuery

- The semantic model DirectLakeBehavior property controls fallback behaviour
    - Automatic (default): allows fallback to DirectQuery if data can’t be loaded into memory
    - DirectLakeOnly: allows use of DirectLake but prevents fallback and returns an error instead of using DirectQuery
    - DirectQueryOnly: forces all queries to use DirectQuery mode
- This can also be set from the Web Editor in the Model view

Direct Lake internals

V-Order

- V-Order is a Microsoft-proprietary optimisation for writing data in parquet files (as used in Delta tables)
- V-Order = the same algorithms used by Power BI Import mode semantic models to compress data
- V-Ordered Delta tables are accessible by any application that can read Delta
- Direct Lake will perform better on V-Ordered Delta tables
- Direct Lake will work on all Delta tables even without V-Order

Refreshing Direct Lake semantic models

- Direct Lake semantic models still need to be refreshed
  - Refresh typically takes a few seconds
- Refresh does **not** involve copying data into the semantic model!
- It means the semantic model points to the latest version of the data held in each table
  - Called “framing”
  - If a model is not framed correctly this can also cause fallback
- Semantic models can be set to refresh automatically or be refreshed manually

Direct Lake refresh

Delta table contents

Product	Year	Sales
Apples	2023	10
Oranges	2023	20
Apples	2024	30
Oranges	2024	40
Pears	2024	50
Grapes	2024	60

DAX query results

Product	Year	Sales
Apples	2023	10
Oranges	2023	20
Apples	2024	30
Oranges	2024	40
Pears	2024	50
Grapes	2024	60

Monitoring paging of data

- Only data that is queried needs to be loaded into memory
- That means columns required:
  - By your query output
  - By any measures used in your query
  - For relationships used to join tables
- Paging data takes up to a few seconds depending on the volume
- DMVs can tell you what data has been paged into memory
  - DISCOVER_STORAGE_TABLE_SEGMENTS tells you if a column segment has been paged into memory and how recently it was used
  - DISCOVER_STORAGE_TABLE_COLUMNS tells you the same thing about column dictionaries

Testing Direct Lake performance

- Test Direct Lake query performance in DAX Studio by:
  - Refreshing the semantic model to ensure all data is paged out
  - Use the Clear Cache button to clear the cache
  - Run query and capture duration – to get worst-case performance (including paging data into memory)
  - Run query again and capture duration – to get performance when data is already paged into memory