Microsoft Fabric - hidden gems

Adrian Chodkowski

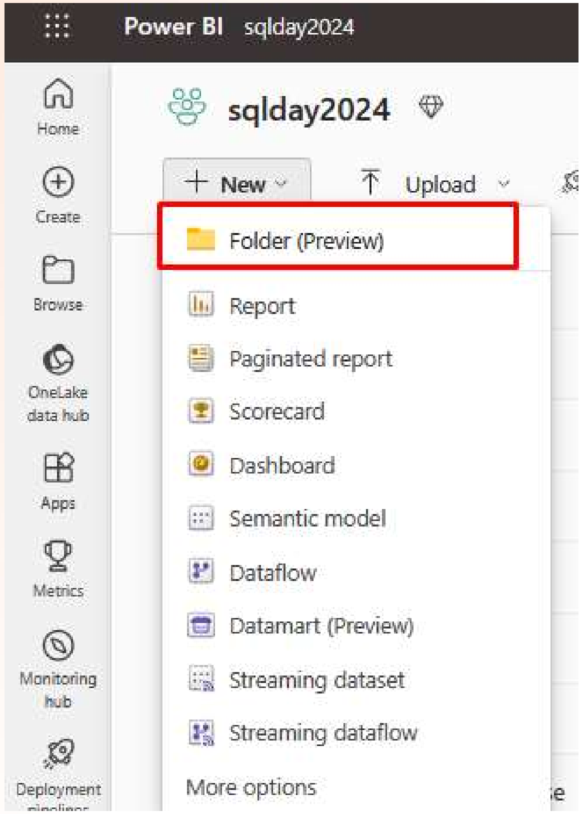
About

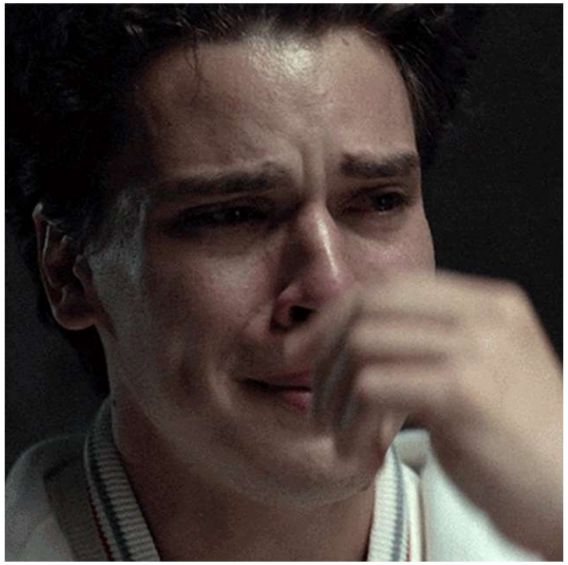
- Adrian Chodkowski
- Microsoft Data Platform MVP
- Architekt i konsultant
- Specjalizacja: Platforma danych Microsoft
- Data Community
- seequality.net
- Adrian.Chodkowski@outlook.com
- @Twitter: Adrian_SQL
- LinkedIn: http://tinyurl.com/adrian-sql

Agenda

- Introduction
- Orchestration
- Loading into delta
- Cloning and Restoring
- Report distribution
- Caching in warehouse
- FastCopy
- DirectLake Framing
- Other

Folders in workspace (Preview)

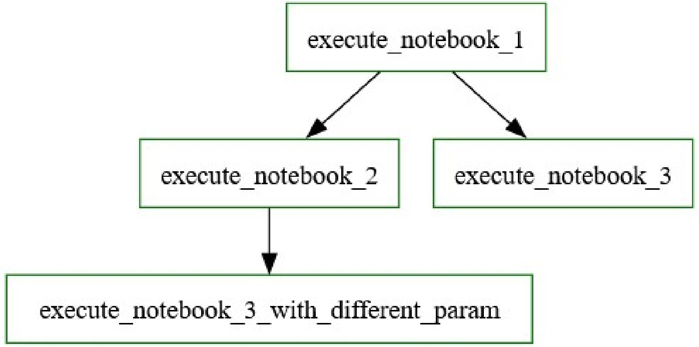




Orchestration

Orchestration via Notebooks

- Main choice for orchestration in Fabric is **Data Factory pipelines** or **Job Schedules**,
- Some people don’t like ADF but there is alternative to orchestrate notebooks via notebooks,
- You can use for that magic command **%run** (it can also execute py files!) , **mssparkutils.notebook.run** but if you have to execute multiple in parallel you must use python (ThreadPoolExecutor similar)
- Fortunately there is also **mssparkutils.notebooks.runMultiple**
- It give you possibility to run many notebooks in parallel and programmatically pass DAG definition to them.

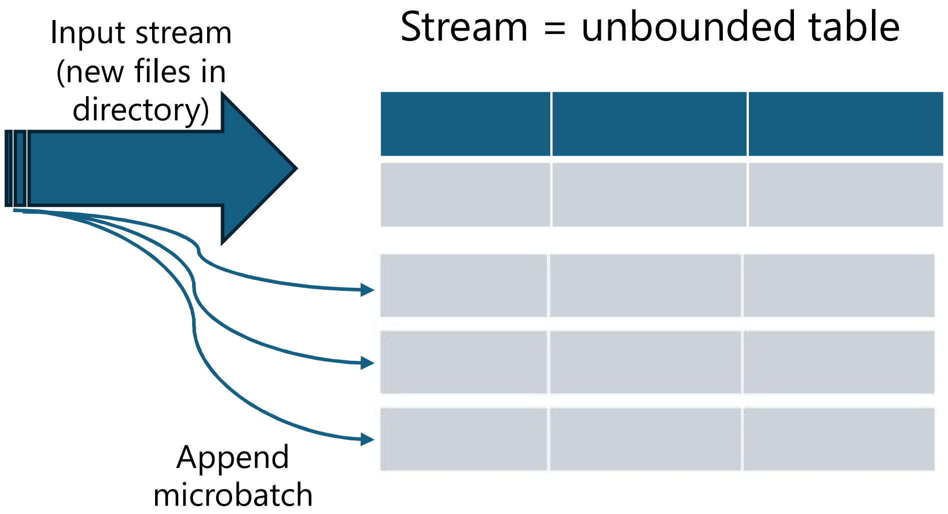


```
{'execute_notebook_1': {'exitVal': '999', 'exception': None},
 'execute_notebook_2': {'exitVal': '888', 'exception': None},
 'execute_notebook_3': {'exitVal': '777', 'exception': None},
 'execute_notebook_3_with_different_param': {'exitVal': '111',
  'exception': None}}
```

Loading into delta

Structured Streaming

- It is not always possible to save extracted data to delta and it must be staged in native format,
- Loading new files from staging can be challenging,
- **Structured Streaming** on the folder can list new files and load it with saving it's status in the checkpoint,
- Spark Structured Streaming is Spark component that is fault-tolerant and allows you to deal with streaming data in micro-batch manner,
- Both approaches are scalable and **idempotent**.



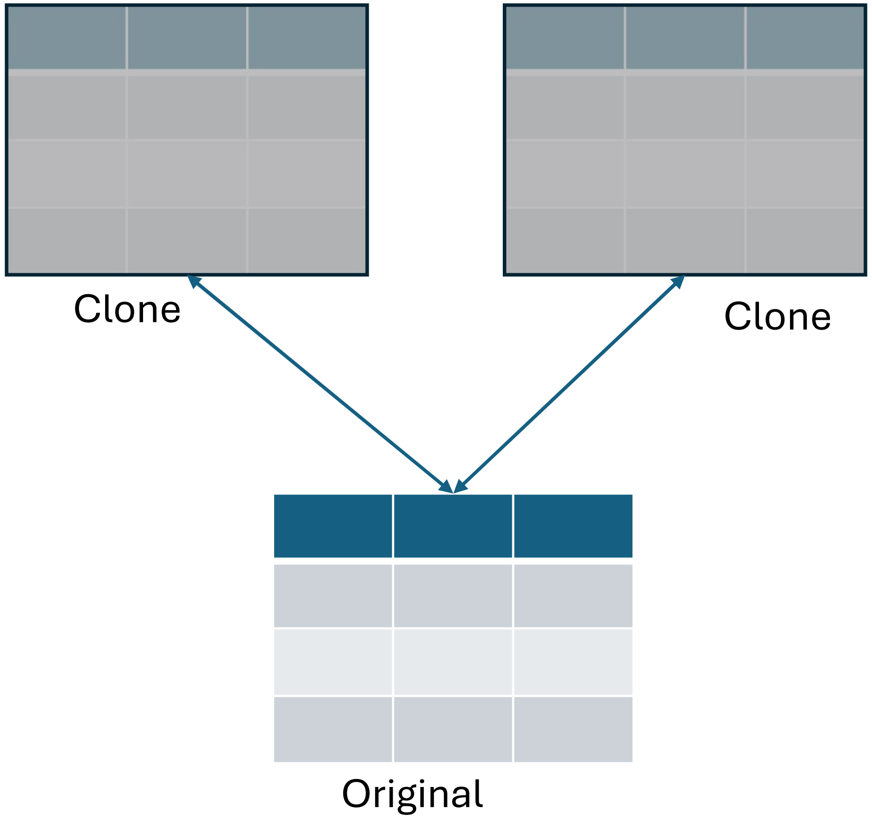
Cloning & restoring

Restore in-pleace of a warehouse

- Restore points are recovery points of a warehouse created by copying the metadata,
- Useful to recover warehouse to specific point in time,
- System-generated restore points are generated automatically every 8 hours and available for 7 days
- User-defined restore points are created manually by workspace administrators,
- User-defined restore points are available via REST API tools,
- Very useful for Disaster recovery scenarios,

Zero-copy clone in warehouse

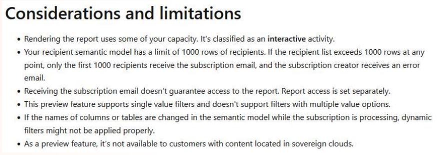
- Clone creates a replica of the table by copying metadata referencing same data files,
- Data history is stored for seven calendar days – clones can reference specific point in time,
- Any modifications to original table are not reflected in cloned table,
- Any modifications to clones are not reflected in original table,
- Useful for development and testing, reporting and data exploration.



Report distribution

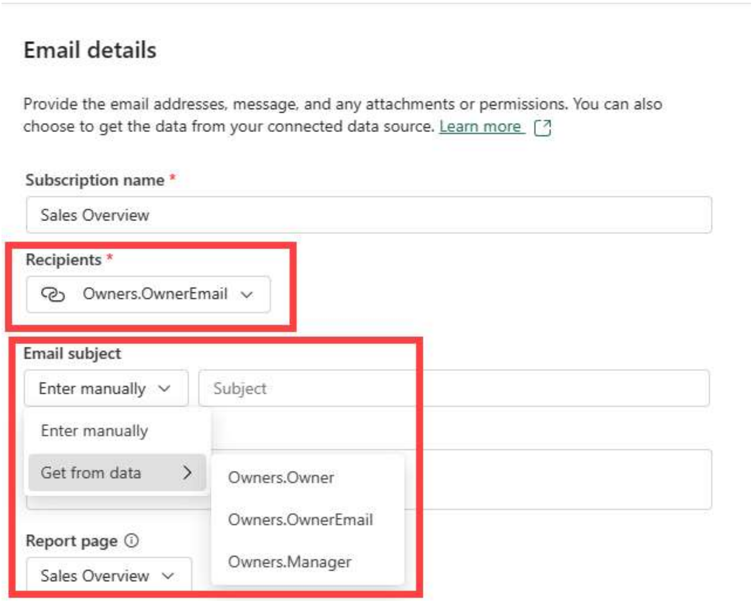
Dynamic per recipient subscriptions for reports (Preview)

- Reports should be distributed via browser and interactive exploration,
- If we have to send it, let’s do it automatically,
- From some time we have possibility to setup dynamic subscription!
- The only thing to do is to have table with emails in the model joined to the data,
- Works for both Power BI and Paginated reports,



Considerations and limitations

- Rendering the report uses some of your capacity. It's classified as an **interactive** activity.
- Your recipient semantic model has a limit of 1000 rows of recipients. If the recipient list exceeds 1000 rows at any point, only the first 1000 recipients receive the subscription email, and the subscription creator receives an error email.
- Receiving the subscription email doesn't guarantee access to the report. Report access is set separately.
- This preview feature supports single value filters and doesn't support filters with multiple value options.
- If the names of columns or tables are changed in the semantic model while the subscription is processing, dynamic filters might not be applied properly.
- As a preview feature, it’s not available to customers with content located in sovereign clouds.



Email details

Provide the email addresses, message, and any attachments or permissions. You can also choose to get the data from your connected data source. Learn more

Subscription name *: Sales Overview

Recipients *: Owners.OwnerEmail

Email subject: Enter manually | Subject

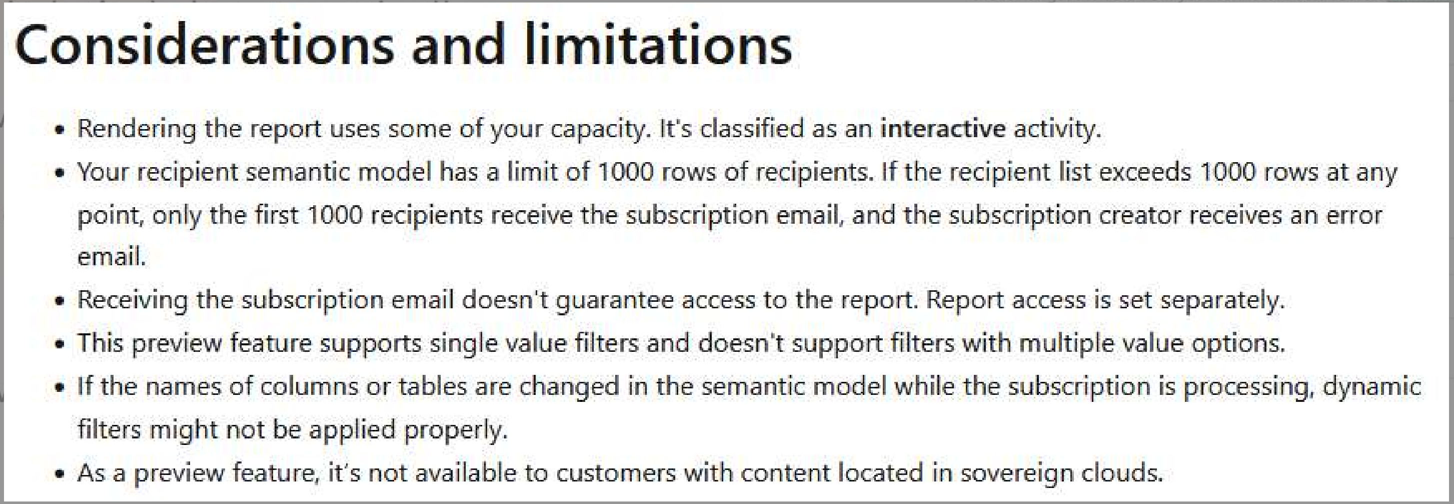
Report page: Sales Overview

Dynamic per recipient subscriptions for reports

Always start reading documentation from the end – from the limitations section!

- Reports should be distrib
  exploration,
- If we have to send
- From some time w
- The only thing to d
  to the data,
- Works for both Pov

Provide the email addresses, message, and any attachments or permissions. You can also choose to get the data from your connected data source. Learn more.



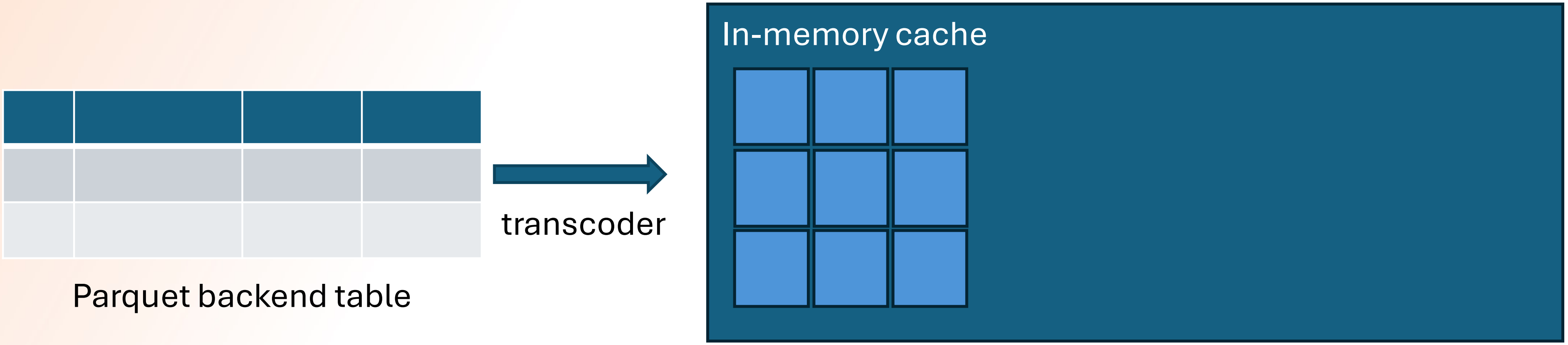
Considerations and limitations

- Rendering the report uses some of your capacity. It's classified as an **interactive** activity.
- Your recipient semantic model has a limit of 1000 rows of recipients. If the recipient list exceeds 1000 rows at any point, only the first 1000 recipients receive the subscription email, and the subscription creator receives an error email.
- Receiving the subscription email doesn't guarantee access to the report. Report access is set separately.
- This preview feature supports single value filters and doesn't support filters with multiple value options.
- If the names of columns or tables are changed in the semantic model while the subscription is processing, dynamic filters might not be applied properly.
- As a preview feature, it's not available to customers with content located in sovereign clouds.

Caching in warehouse

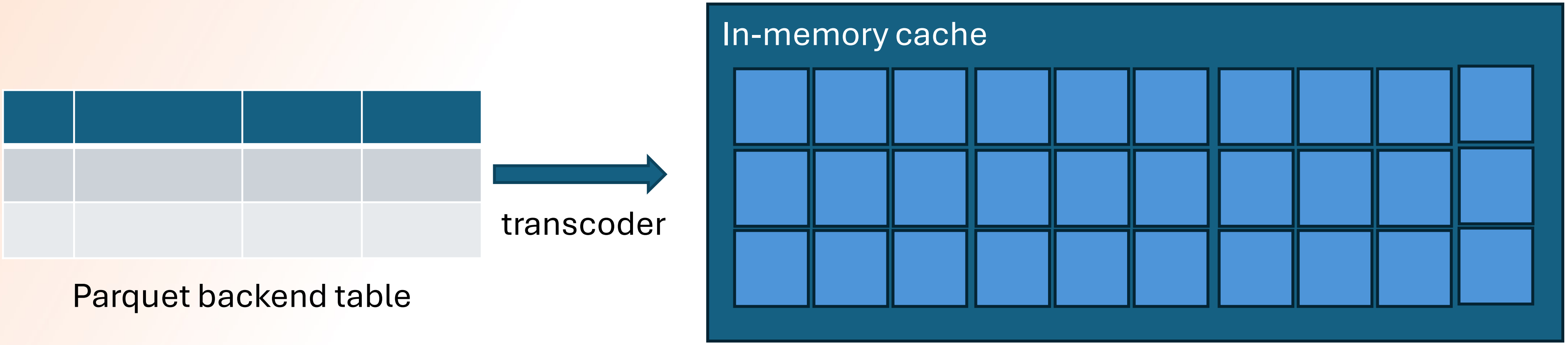
Caching in Warehouse

- Fabric stores data in parquet based delta format,
- When data is not in cache it must be transcoded from parquet into in-memory columnar format optimized for analytical queries,



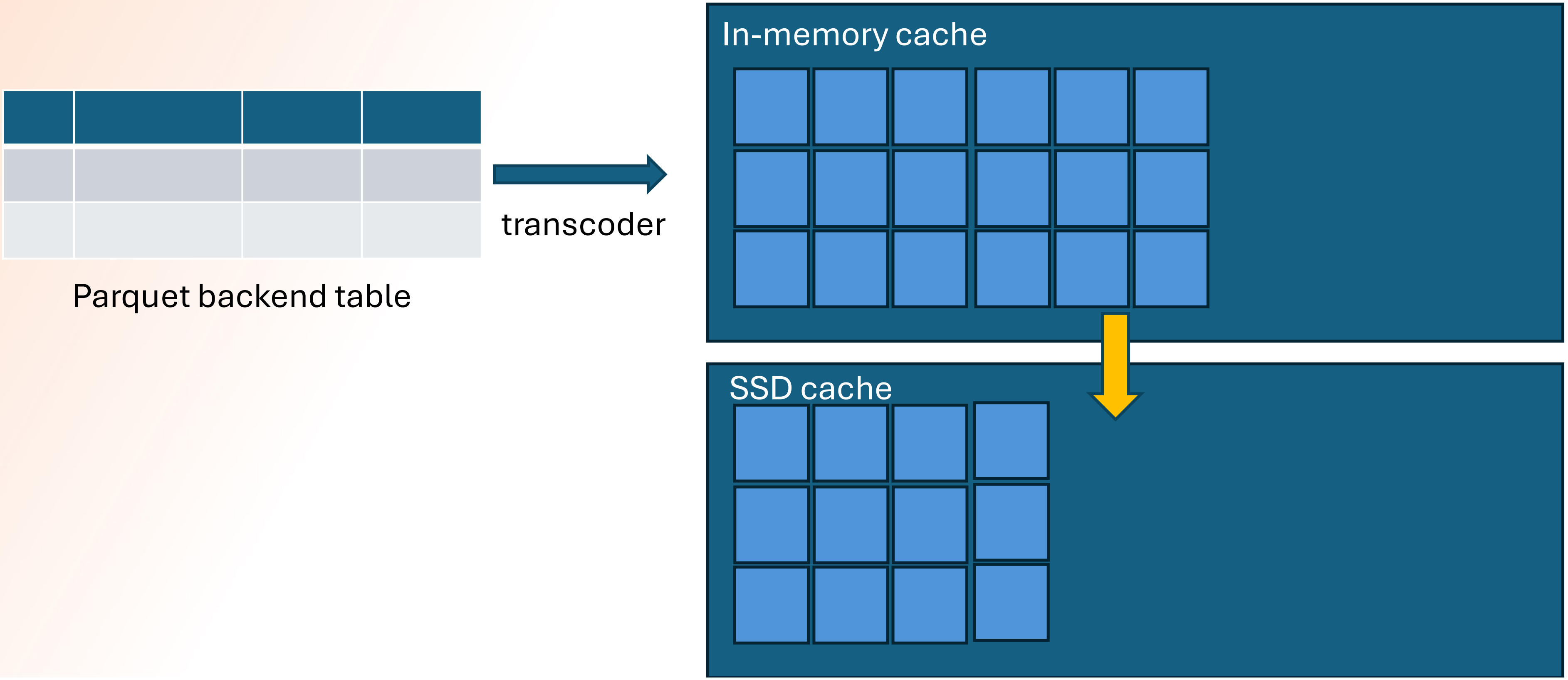
Caching in Warehouse

- Data cannot be fully loaded into cache so it must be evicted from there,
- Missing cache can lead to performance degradation,



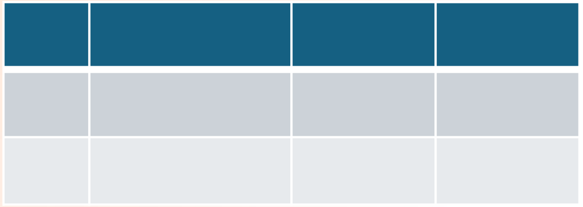
Caching in Warehouse

- Fabric has special SSD cache that is much larger than in-memory to minimize cache miss impact,
- Data is asynchronously saved in SSD in SQL native format + parquet metadata is also saved.

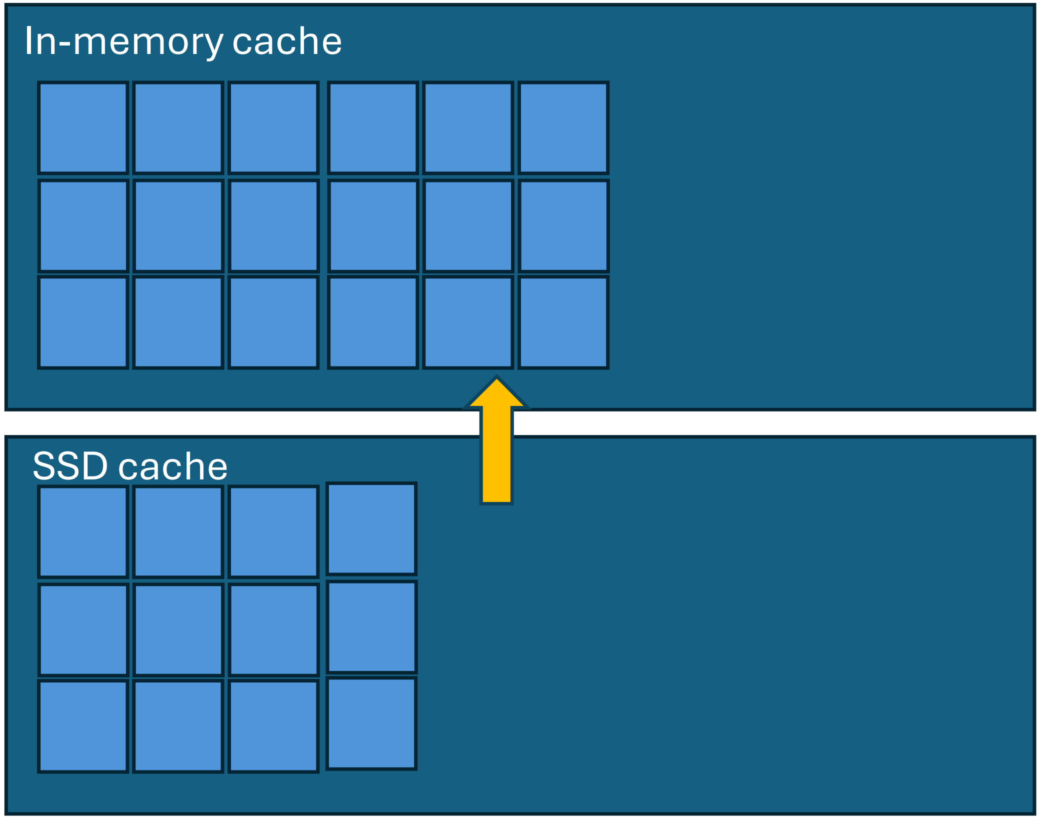


Caching in Warehouse

- Next time when query will read the data it will be read from SSD cache instead of remote storage
- Data is already in native format so transcoder is not needed.

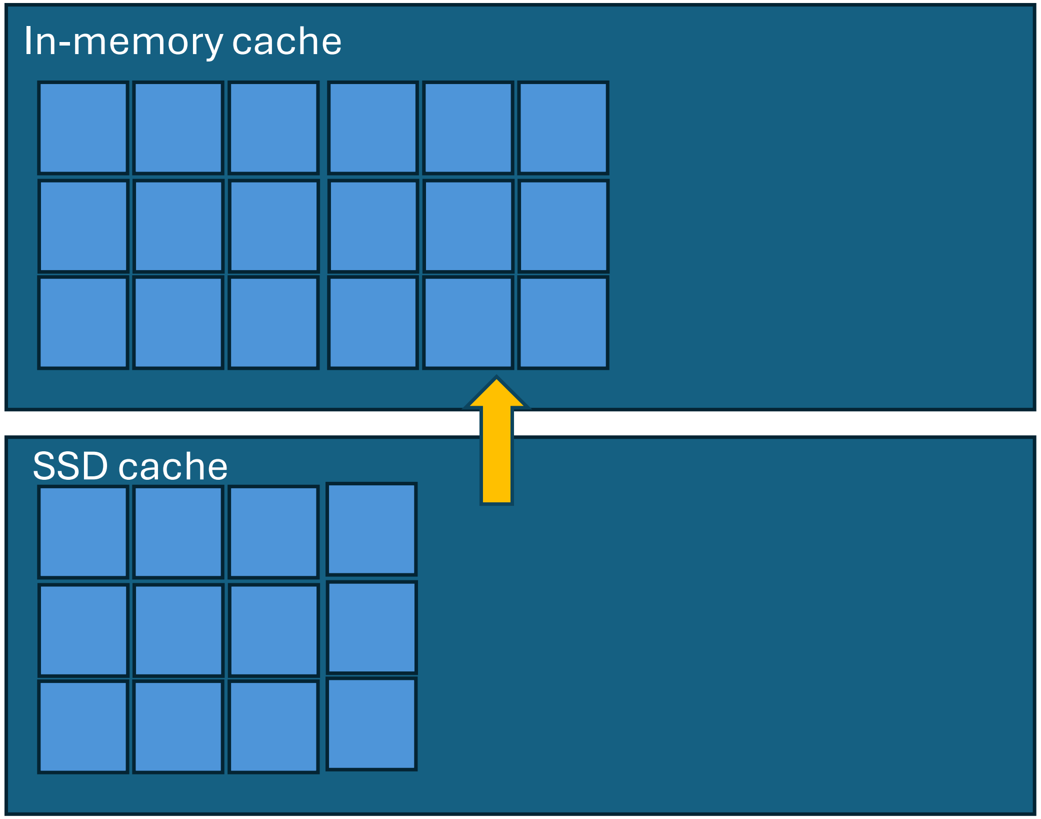


Parquet backend table



Caching in Warehouse

- Cache is constantly active in the background,
- No intervention is needed,
- Cache cannot be cleared,
- Cache is transactionally consistent.

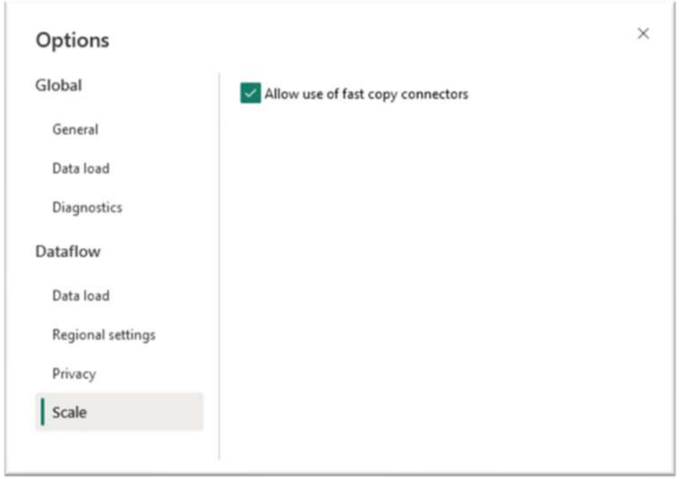


SQL Day

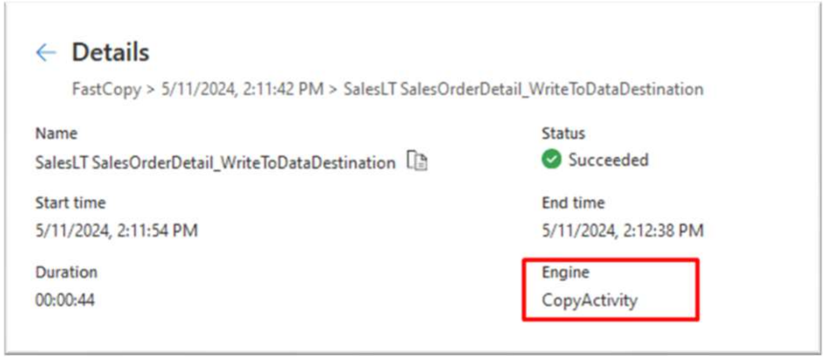
Fast copy in Dataflows gen2

Fast copy in Dataflows Gen2

- Improvement that will give possibility to ingest tons of data within Dataflows,
- If data size exceeds 100MB backend will switch to ADF's Copy Activity, for SQL it will be minimum 1 million rows,
- You can force fast copy,
- Supported sources: ADLS Gen2, Blob Storage, Azure SQL Database, PostgreSQL, Lakehouse
- Query Folding supported, For BLOB and ADLS only parquet and csv are supported,
- Data Gateway and VNET Gateway are not supported,



☑ Allow use of fast copy connectors



FastCopy > 5/11/2024, 2:11:42 PM > SalesLT SalesOrderDetail_WriteToDataDestination

Name: SalesLT SalesOrderDetail_WriteToDataDestination

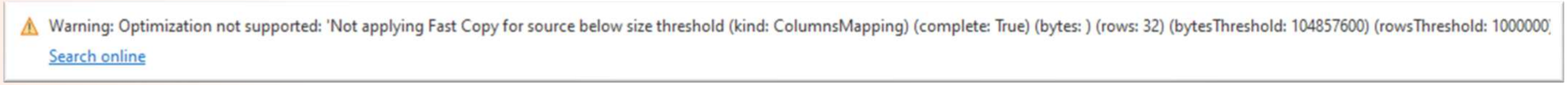
Status: Succeeded

Start time: 5/11/2024, 2:11:54 PM

End time: 5/11/2024, 2:12:38 PM

Duration: 00:00:44

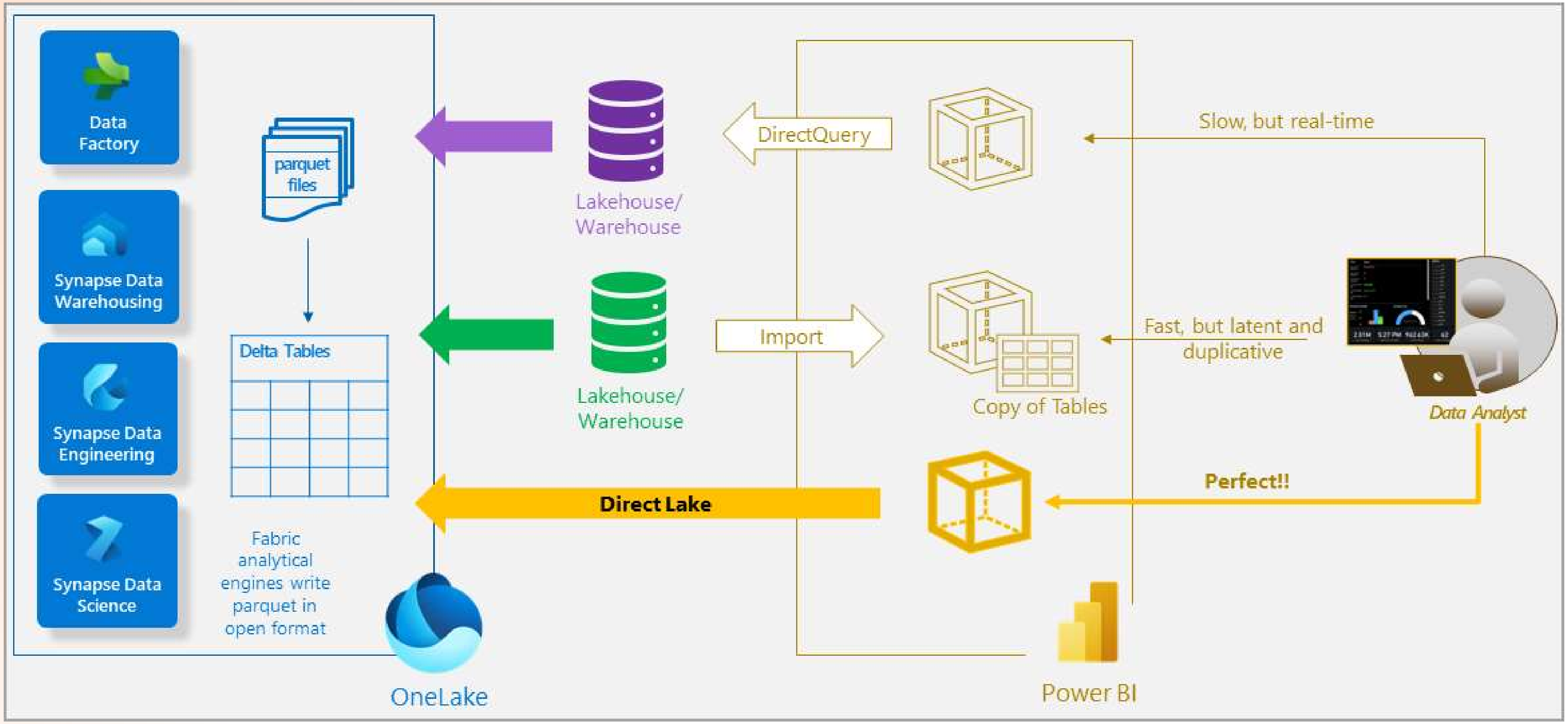
Engine: CopyActivity



Warning: Optimization not supported: 'Not applying Fast Copy for source below size threshold (kind: ColumnsMapping) (complete: True) (bytes:) (rows: 32) (bytesThreshold: 104857600) (rowsThreshold: 1000000)

Report refresh process

Power BI - data access methods



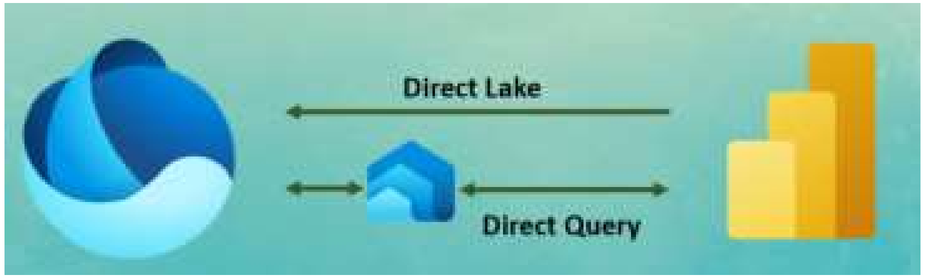
Direct Lake

- New mode that give us possibility to read delta files directly from Power BI report,
- If read cannot be done (i.e RLS is implemented) it can fallback to Direct Query to SQL Endpoint,
- Can be as good as Import or worse,
- If memory pressure occur then data can be spilled to disk
- It has many benefits over Import like:

  Direct Lake is loading into memory only used columns

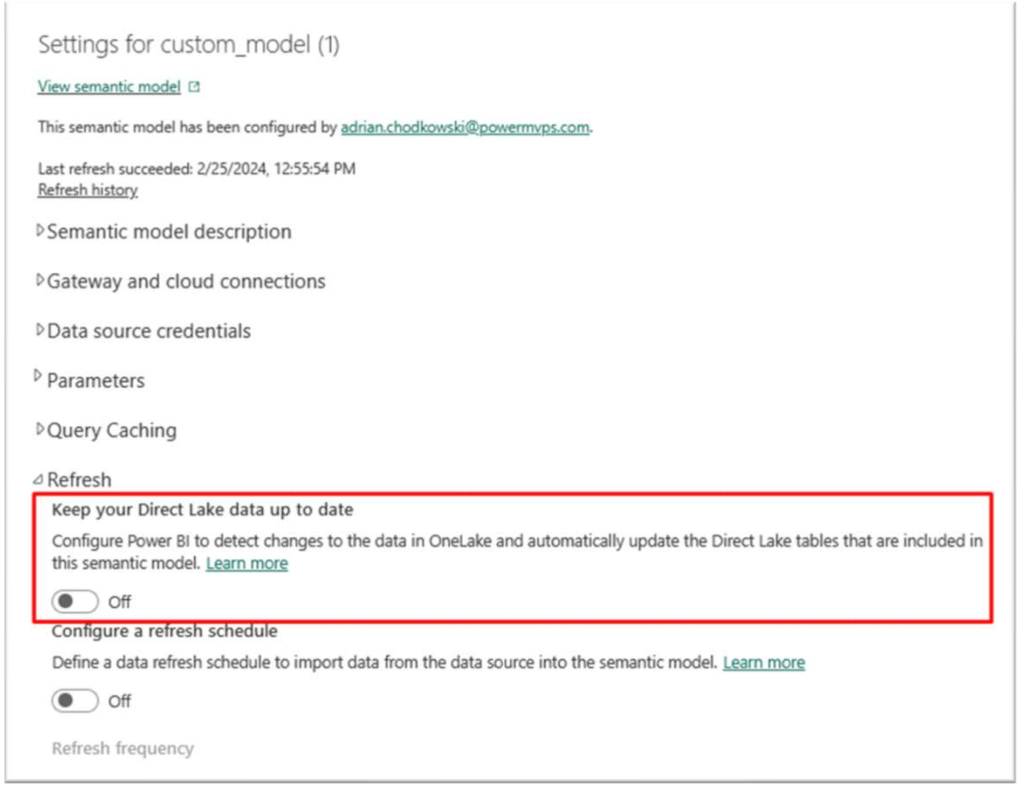
  It doesn't need typical refresh process

  Memory usage is much more efficient



Direct Lake refresh

- In Direct Lake data is not copied into the model but it must be refreshed,
- Refresh is turned on by default but it can be changed,
- Refresh means in this case that model will point to the newest delta version,
- You can start "refresh" on demand or based on schedule,
- It give you possibility to reflect changes in the model when "all the processes" finished and avoid "integrity" problem
- It is like automatic snapshoting!
- Direct Lake Refresh also called "Framing" removes data from cache!
- **Everything in Fabric consumes Capacity Units!**



Settings for custom_model (1)

View semantic model

This semantic model has been configured by adrian.chodkowski@powermvps.com.

Last refresh succeeded: 2/25/2024, 12:55:54 PM
Refresh history

▷Semantic model description

▷Gateway and cloud connections

▷Data source credentials

▷Parameters

▷Query Caching

◿Refresh

Keep your Direct Lake data up to date

Configure Power BI to detect changes to the data in OneLake and automatically update the Direct Lake tables that are included in this semantic model. Learn more

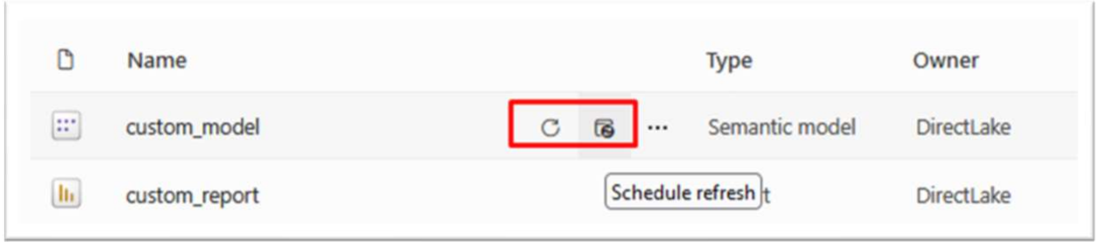
Off

Configure a refresh schedule

Define a data refresh schedule to import data from the data source into the semantic model. Learn more

Off

Refresh frequency



Name	Type	Owner
custom_model	Semantic model	DirectLake
custom_report	t	DirectLake

Other

- Throttling and smoothing
- Job Queueing
- Optimistic Job Admission
- Auto-tuned spark configurations
- Jupyter-black
- Pause-resume Event Stream
- External Sharing
- All available delta lake & Spark features!

Data
Community