



Data Virtualization: Hype Or Salvation?

Ben Kettner

Markus Raatz

SQL KONFERENZ 2024

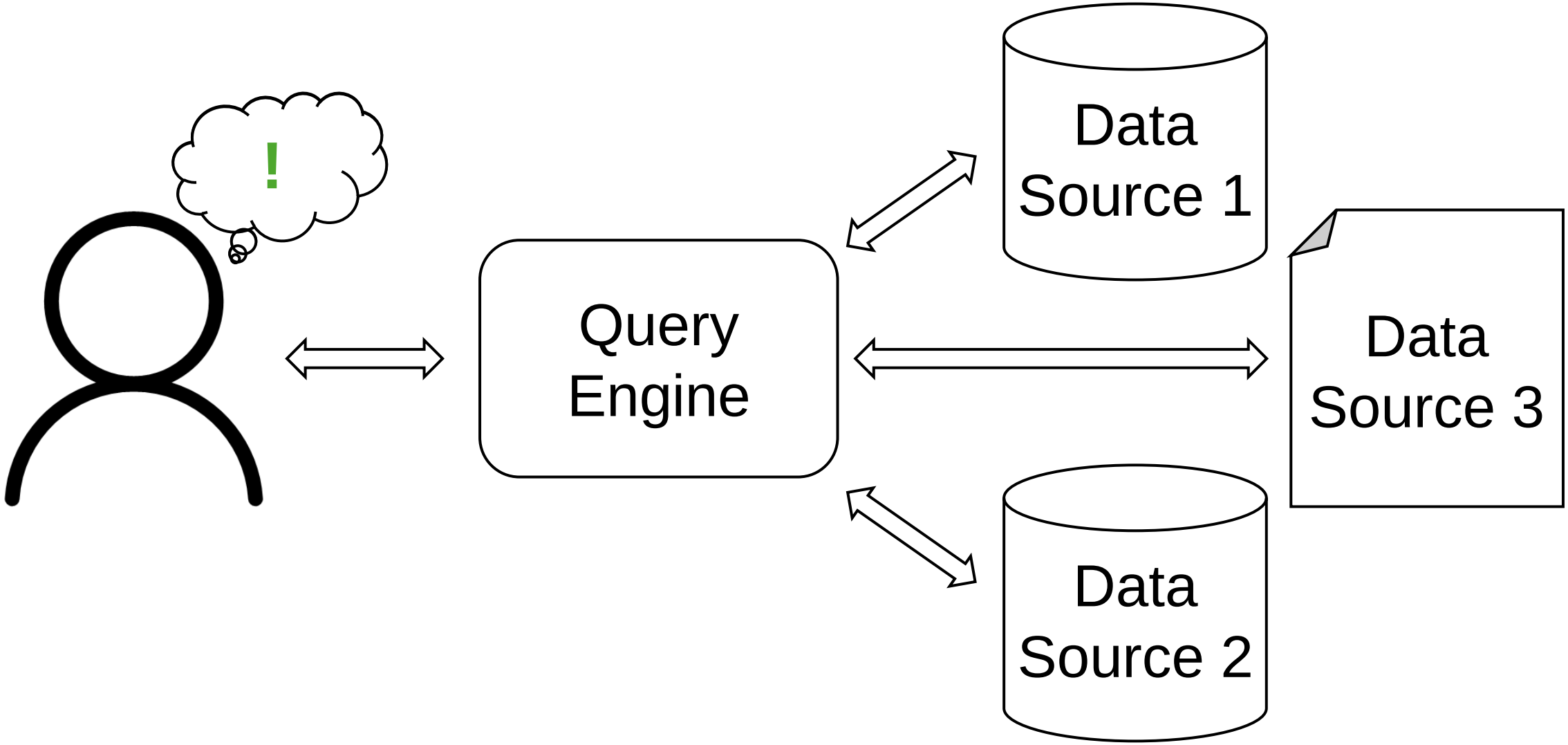
Data what?



Data Virtualization

Data virtualization is an approach to data management that allows an application to retrieve and manipulate data without requiring technical details about the data, such as how it is formatted at source, or where it is physically located, and can provide a single customer view (or single view of any other entity) of the overall data.

So... DWH without ETL?





Obstacles? Challenges!

Joining data? How can we be sure that data types match? How can we be sure that constraints are met?

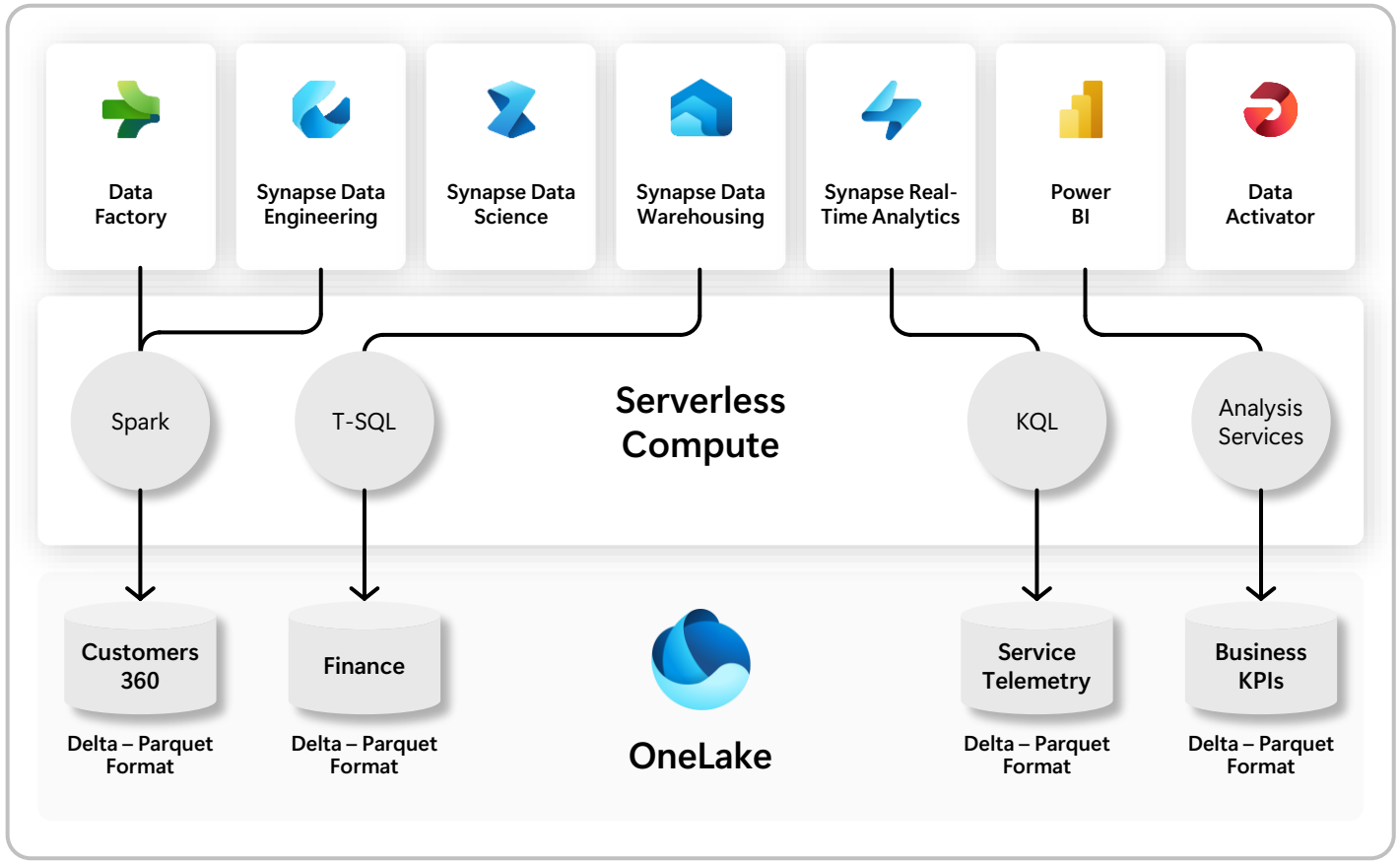
What was that thing about ACID?

What about performance? Can indexes in the source be used? Can filtering in the source be used?

I don't care, I still want it...

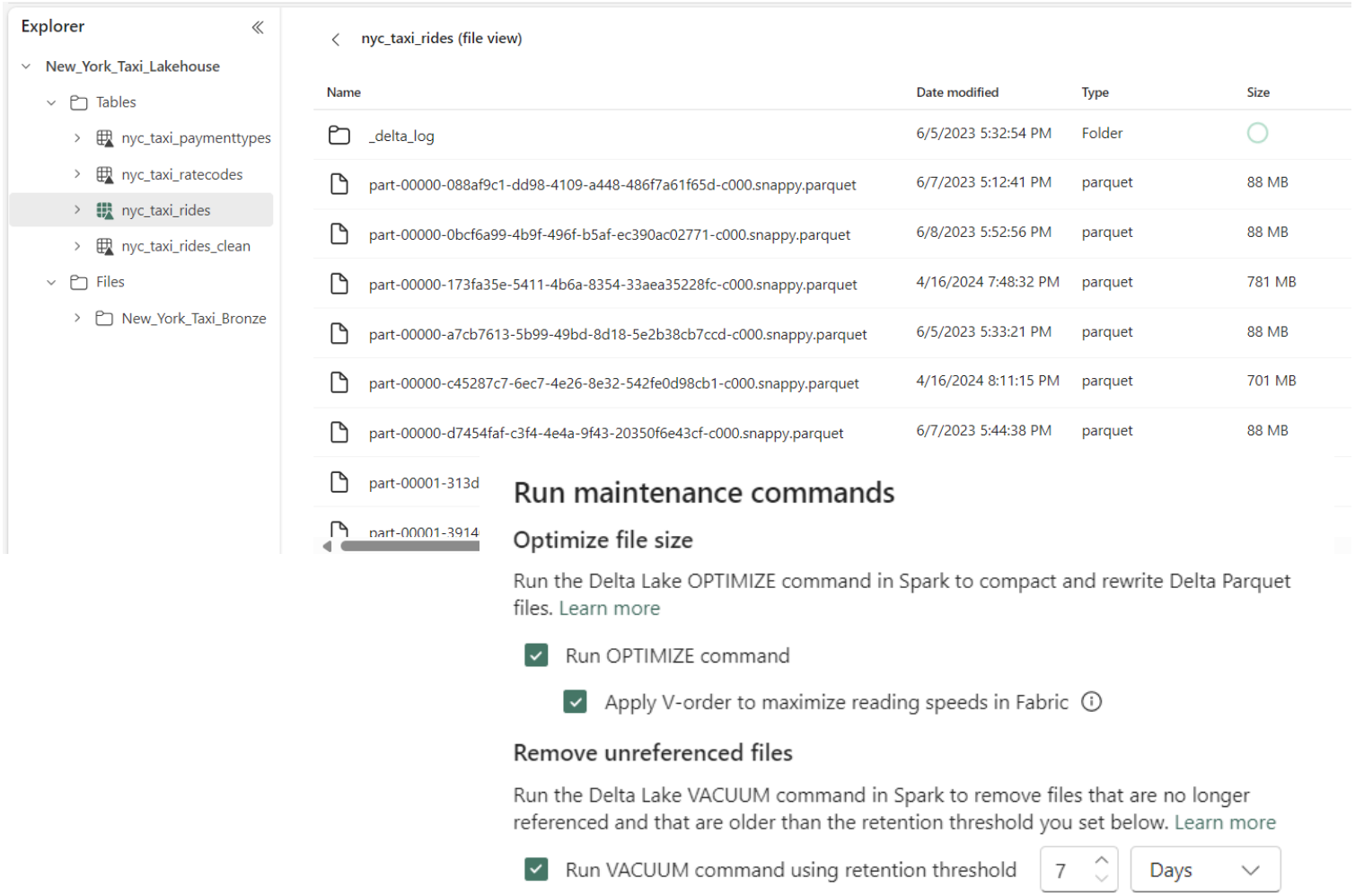
Best Data Virtualization Reviews 2024 | Gartner Peer Insights

OneLake is where Fabric stores data



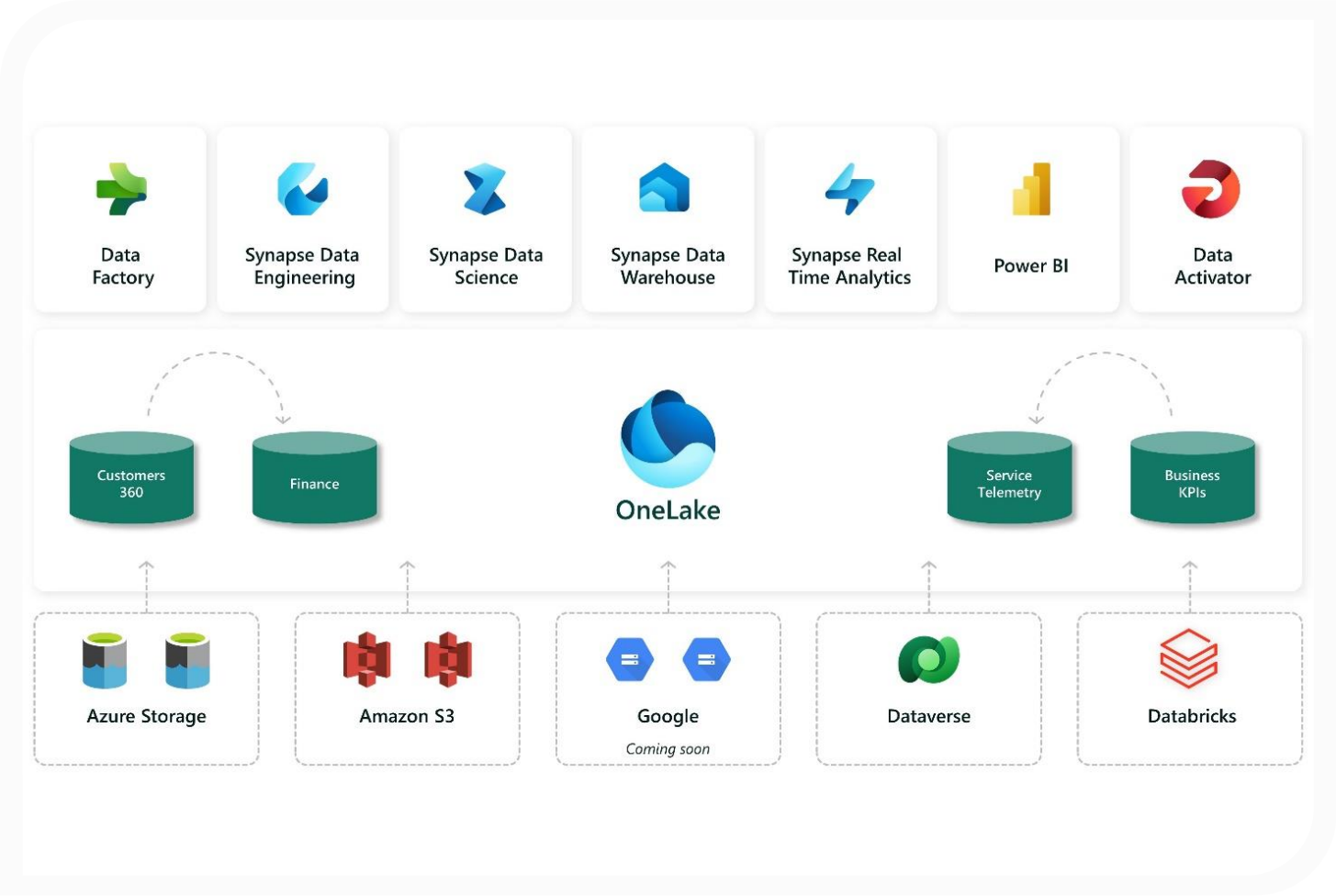
- No matter which workload, all data in Fabric ends in OneLake
- „One copy for all computes“
- All workloads store data in the same format: Delta Parquet!

DEMO: OneLake and Delta Parquet



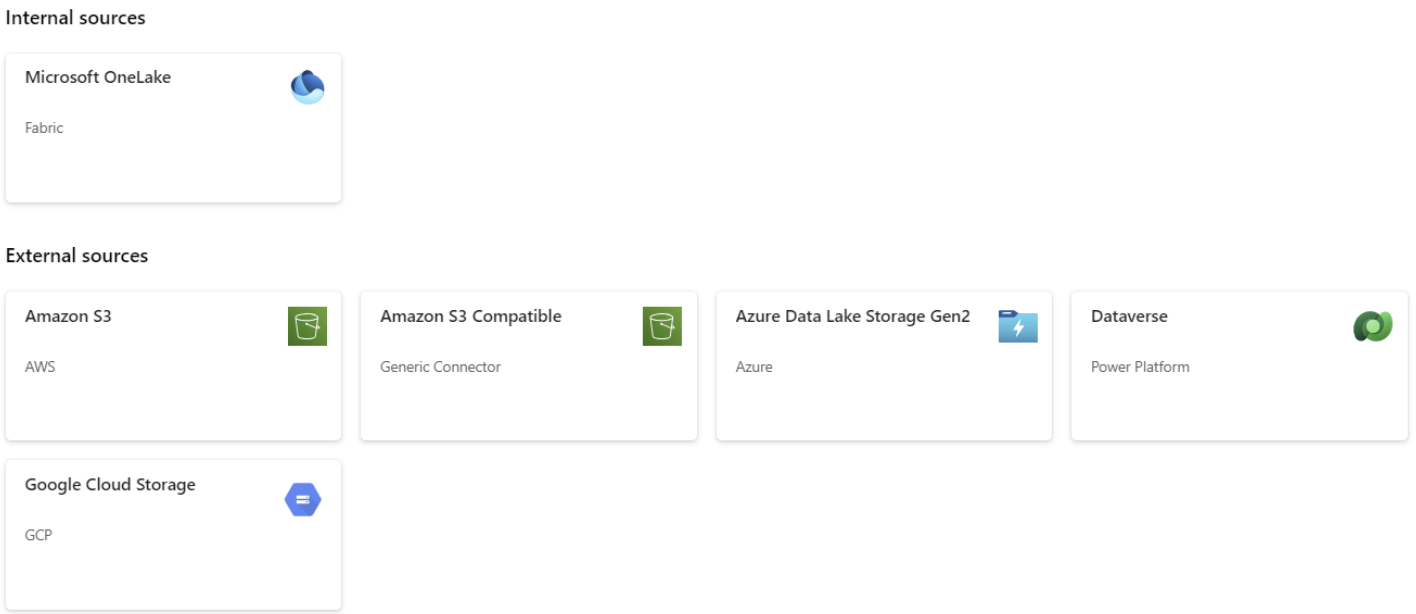
- OneLake supports Parquet in v-order and z-order (and both)
- Transactions are kept in JSON-format for ACID-properties.
- Don‘t forget to clean up from time to time!

Shortcuts bring in data from everywhere



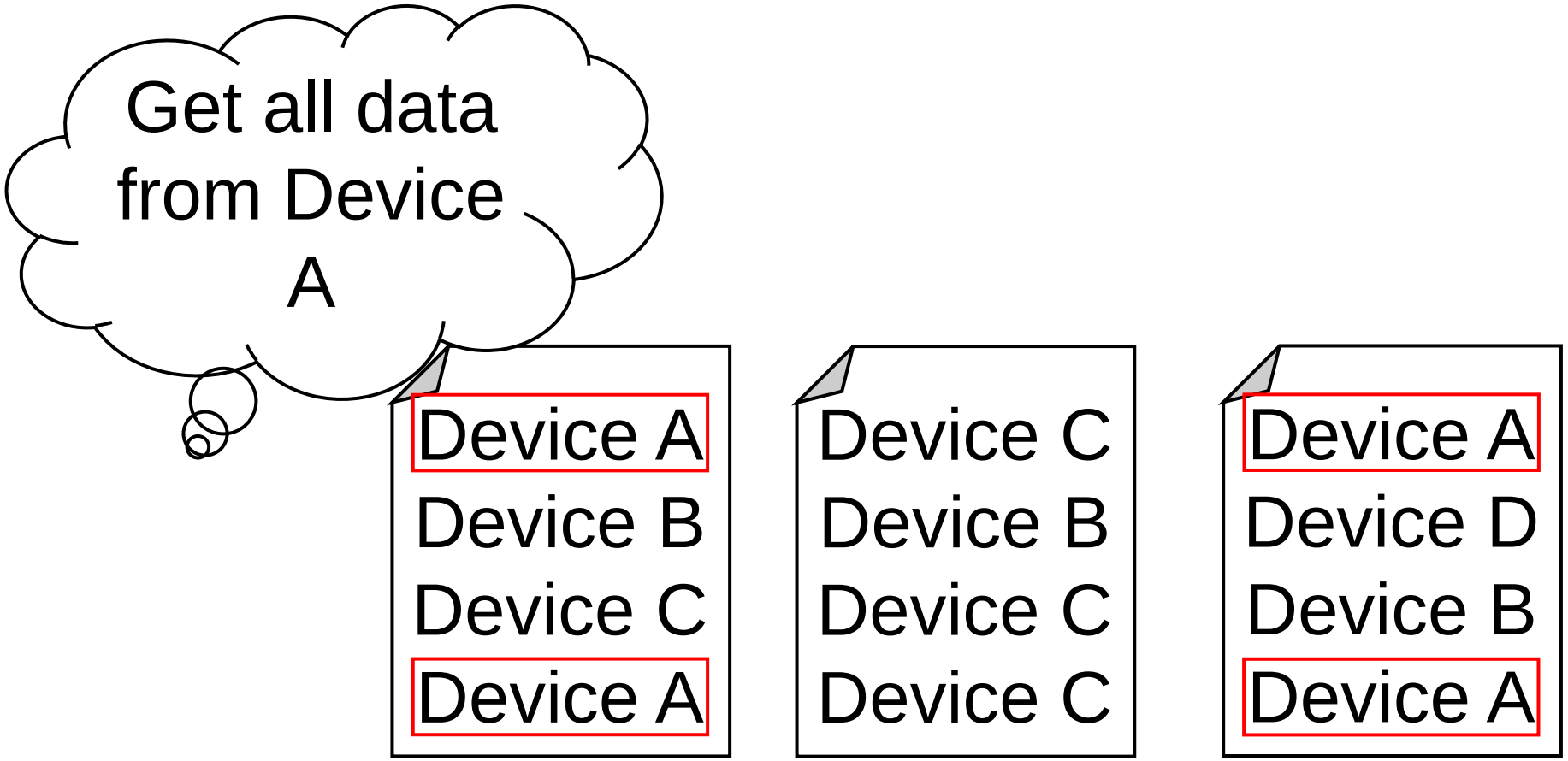
- If you have data in other storage locations like AWS or GCP, create a shortcut to it instead of copying it to your OneLake
- You can query it using SQL in your Fabric Lakehouse
- The first multi-cloud data lake!

DEMO: Shortcuts to ADLS Gen2 and AWS



- If data is stored in ADLS, you can even write to it using Spark SQL

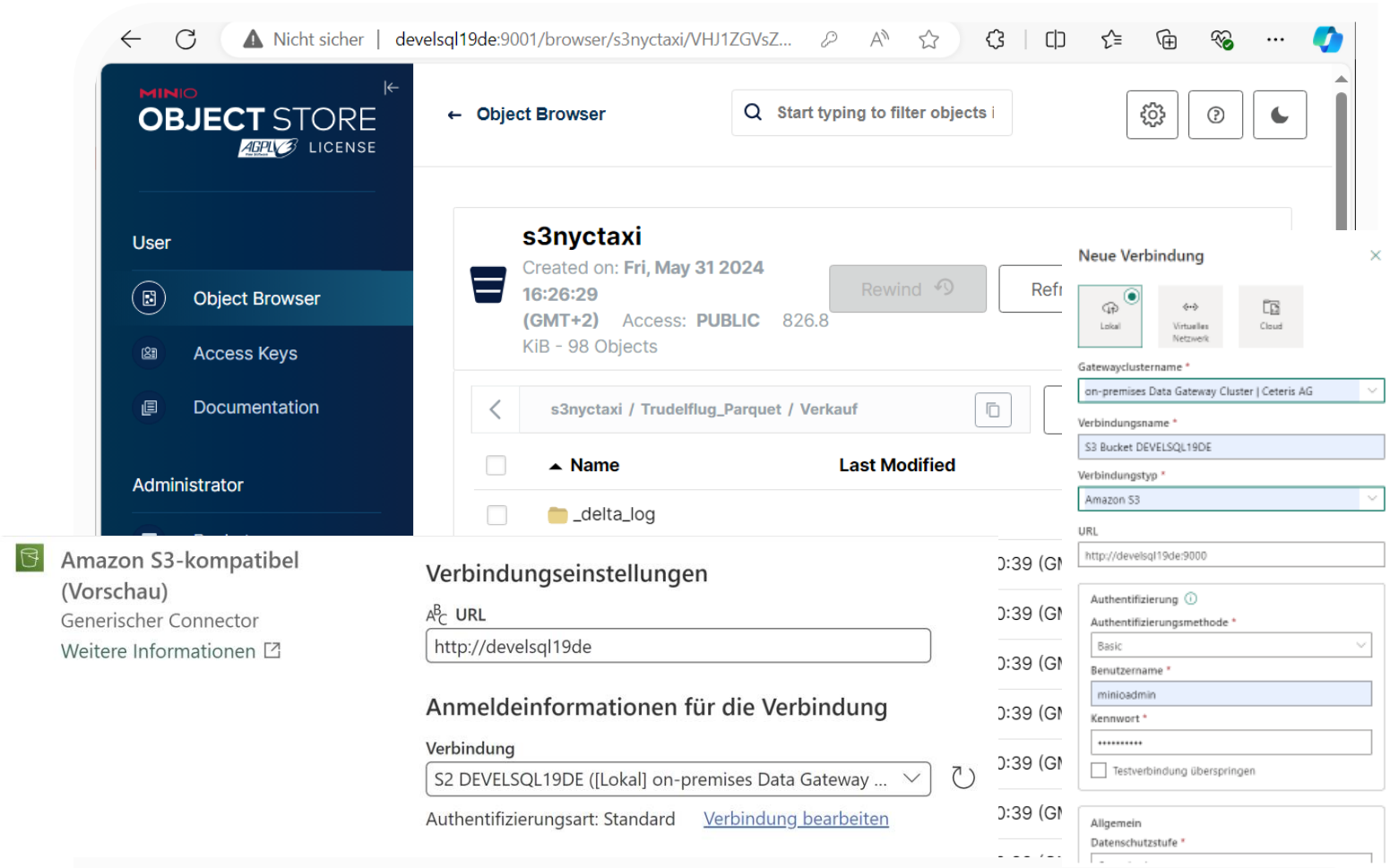
Careful! Data organization matters



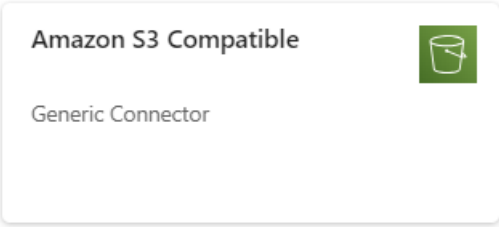
Shortcuts features

- Will automatically check for updates the moment you query it
- Views can be created using the SQL analytics endpoint
- Supports iceberg-format files as well now (see snowflake)!
- Also possible: Shortcuts to unmanaged files and folders

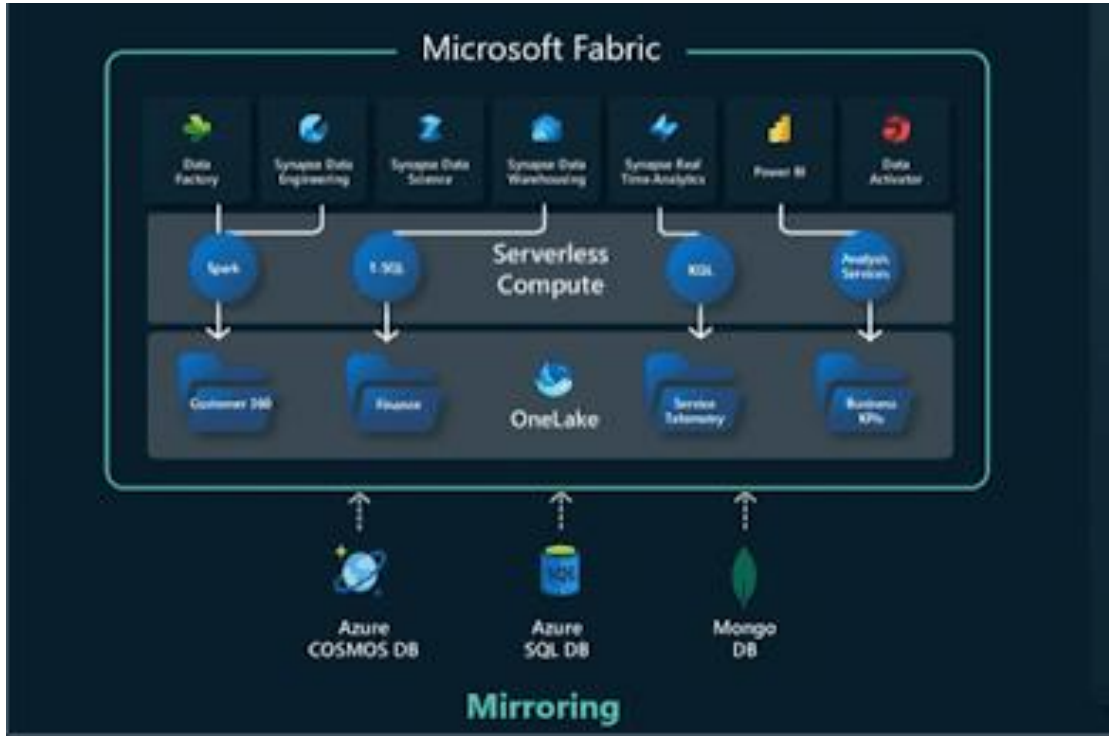
DEMO: Shortcuts to data on premise



- Use local software to act as an S3 bucket, like Minio
- You have to configure the local data gateway, of course



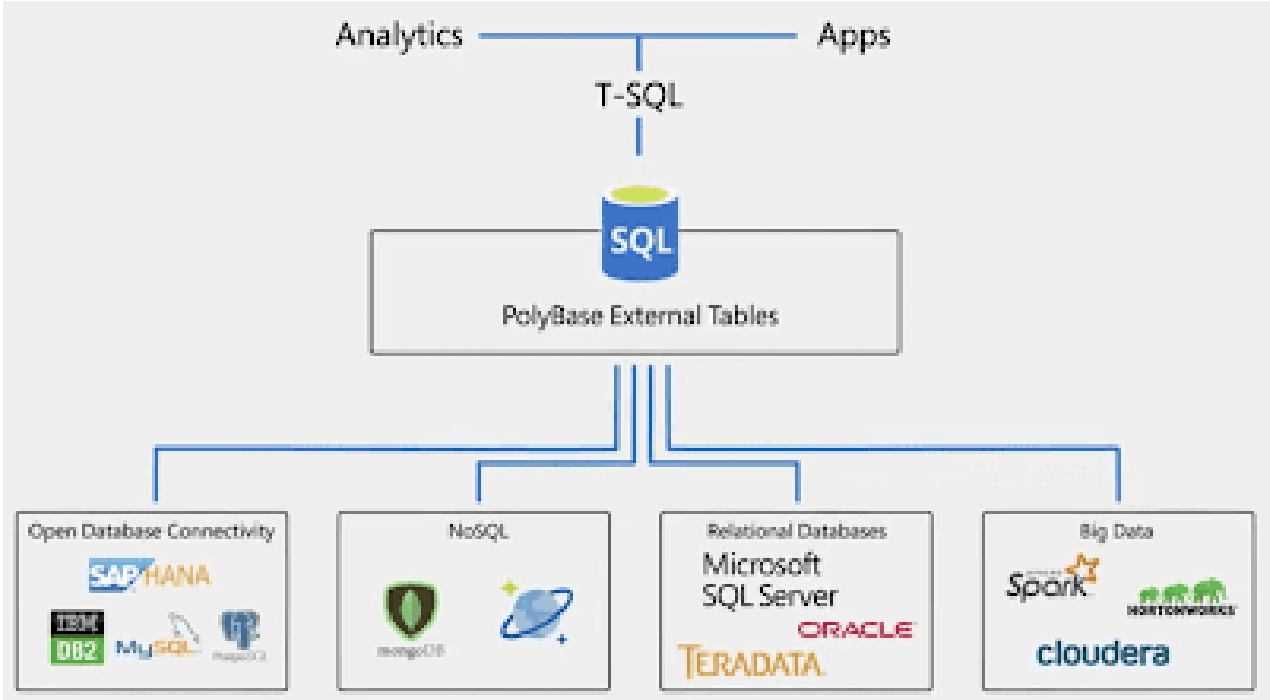
Fabric Mirroring: how virtual is it?



- Mirroring creates a local copy of your tables in OneLake and keeps them up to date, using CDC-technologies
- Entire database or individual tables
- Storage for the mirrored data is free!
- Currently only SQL Azure, Cosmos DB, Azure Databricks and Snowflake supported
- Sorry, no updates, just analytics

But what about Polybase?

- Available in SQL Server 2019 and later , SQL MI and Synapse dedicated server
- T-SQL queries against:
  - ODBC data sources,
  - Oracle,
  - Teradata,
  - MongoDB,
  - Hadoop clusters (deprecated),
  - Cosmos DB,
  - Amazon S3 with Delta Parquet
  - ADLS Gen2 with CSV



The demo was too fast for me…

- Go to https://www.youtube.com/@CeterisAG/





Conclusion

- Fabric Shortcuts virtualize over many data sources, but readonly
- The source has to be one format only: Delta Lake (Delta Parquet)

Feedback, please!

SQL KONFERENZ 2024