Building a data lakehouse in Azure

Johan Ludvig Brattås

Microsoft

Evidi


blinQ


CloudWay


INNOFACTOR®


03C


POINT : TAKEN

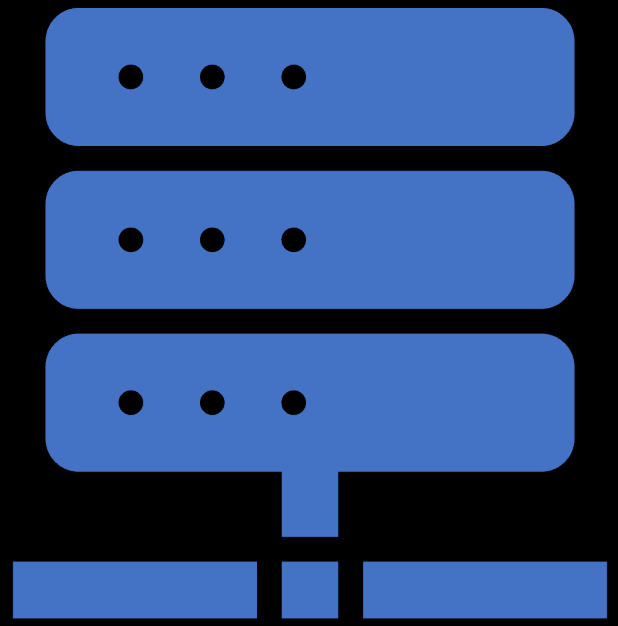
• A G E N D A

Data lakes and data warehouses

What is a data lakehouse?

What is a lakehouse file?

How to build your lakehouse on Azure



The Enterprise Data Warehouse

- Been around since 1988 (the Business Data Warehouse...)
- Evolved over time
- contains structured, schema on-write data
- relational data
- stores historical data that needs to be transformed
- typically transform before storing
- good for business decisions

The Enterprise Data Warehouse

- high data quality and focus on exact numbers
- «long» development life cycles

Data Lake

- gained foothold around 2011
- answer to pains in EDW on the 3 Vs of big data (volume, variety, velocity)
  or 4 or 5 Vs...
- typical varieties are:
- the raw data lake
- the business data lake or layered data lake

The principles of the Business Data Lake

1. Land all the information you can *as is with no modification*
2. Encourage LOB to create point solutions
3. Let LOB decide on the cost/performance for their problem
4. Concentrate governance on the critical points only
5. Consider the corporate view to be just another LOB view
6. Unstructured information is still information
7. Never assume the lake contains everything
8. Scale is driven by demands – scale down as well as up

Data Lake

- transformation only from raw layer to curated layers.
- various use cases and users on various layers of the data lake
- great for trends and ad-hoc analysis
- short time to market development cycles
- enables data science and self-service
- danger of data swamp
-

Data Lake

- EDW functionality on data lake traditionally a problem
- Data lakes lack ACID (Atomic, Consistent, Isolated, Durable) support
- Slow query response an issue for consumers

- The enterprise data warehouse is well established – why throw away the investment?
- Data lakes cannot fill the shoes of the EDW – without taking on parts of the form and function of one.

Data lakehouse

- delivers both regular EDW (data warehouse) and big data analytics (data lake)
- empowers self-service BI
- enables self-service analytics and data science

- still need proper services for performant EDW on large datasets
- as data size grows, still the same challenges of data swamp
- how can users find and make use of data?
- how to know what data shows trends and what shows exact truths?

Data Lakehouse

Features of a data lakehouse:

- Transaction support
- Schema enforcement and governance
- BI support
- Storage is decoupled from compute
- Openness
- Support for diverse data types
- Support for diverse workloads
- End-to-end streaming

What is a lakehouse file?

- Lakehouse files are at the heart of the lakehouse

- Open source
- Brings ACID principles to data lake (supports S3, ADLS, GCS, and HDFS)
- Unified batch and streaming source and sink
- Schema enforcement
- Schema evolution
- Transaction logs
- Time travel
- Updates and deletes
- Audit history
- Data stored in Parquet

APPEND DATA

pandas

memory-intensive

```python
# Read the data
existing_data = pd.read_parquet("data.parquet")

# Concat two dataframes
df3 = pd.concat([df1, df2])

# Save to a file
df3.to_parquet("data.parquet")
```

time-consuming

Delta Lake

one step

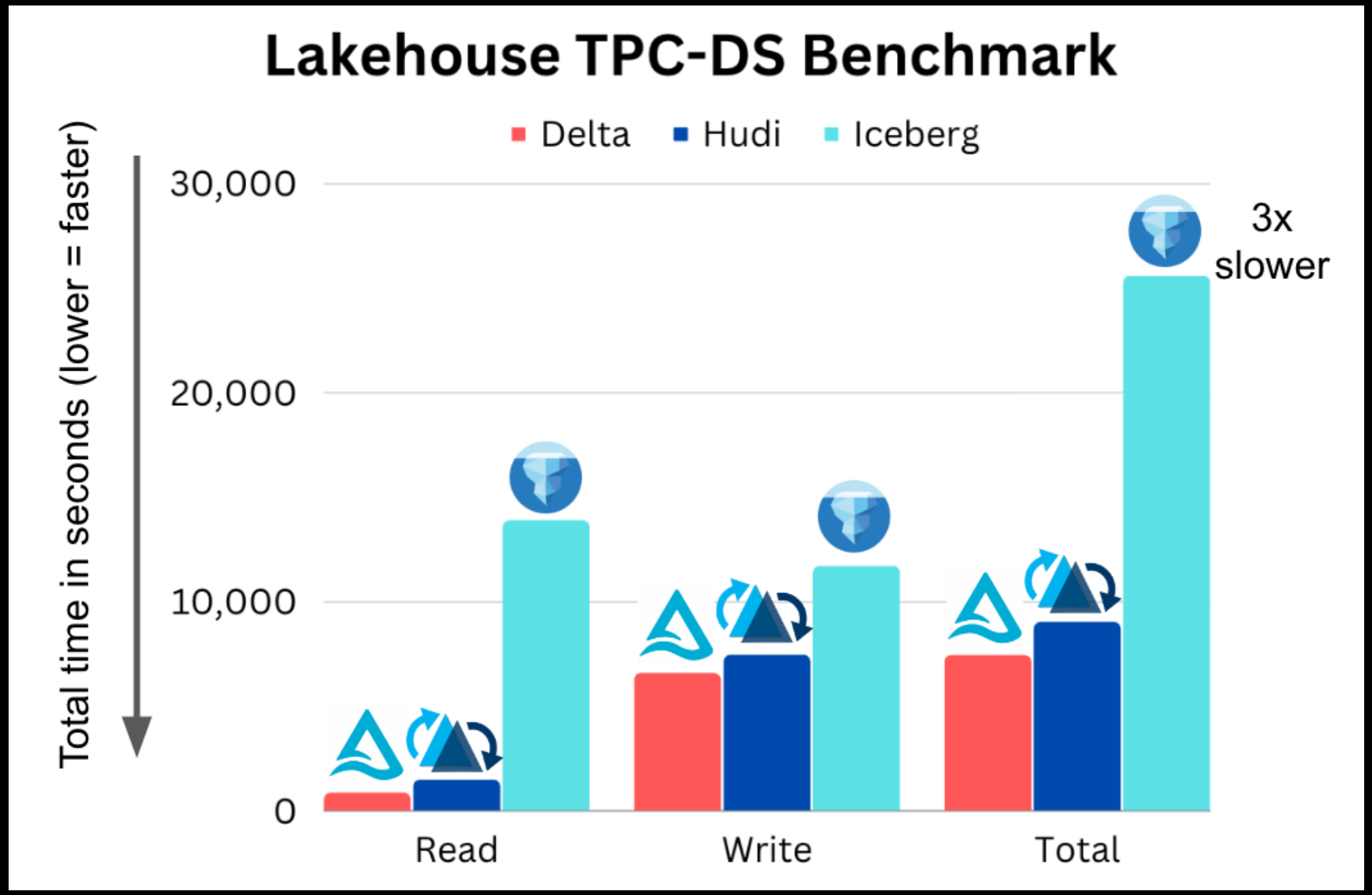
```python
from deltalake.writer import write_deltalake

# Append to Delta Lake
write_deltalake("data", df2, mode="append")
```

khuyentran1401

What is a lakehouse file

- 3 competing formats
- Apache Iceberg – from Netflix
- Apache Hudi – from Uber
- Linux Delta Lake – from Databricks
- All three can run in Spark, but only Delta Lake runs «natively» in Azure
- Delta Lake 3.0 has announced UniForm (Unified Format) – but only when underlying storage is parquet

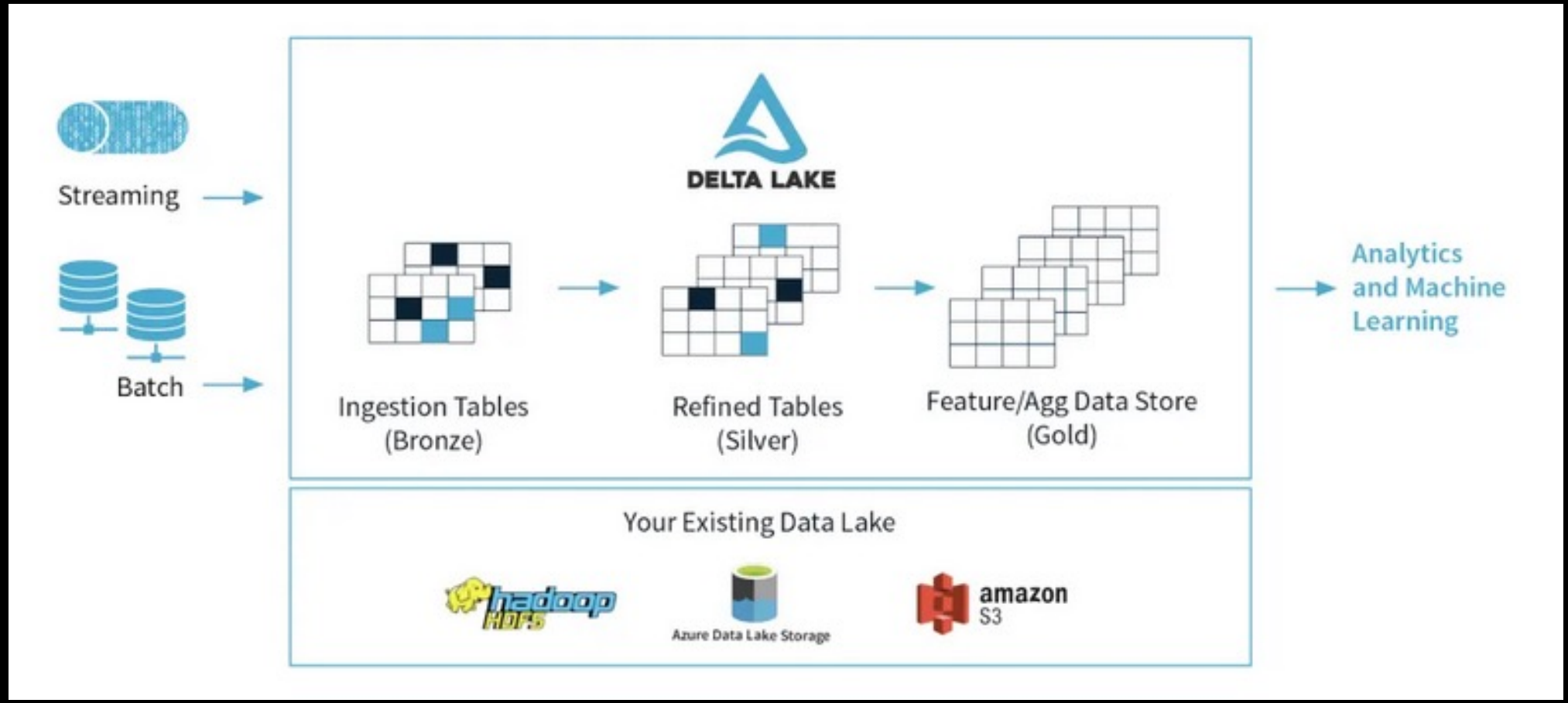
What is a lakehouse file



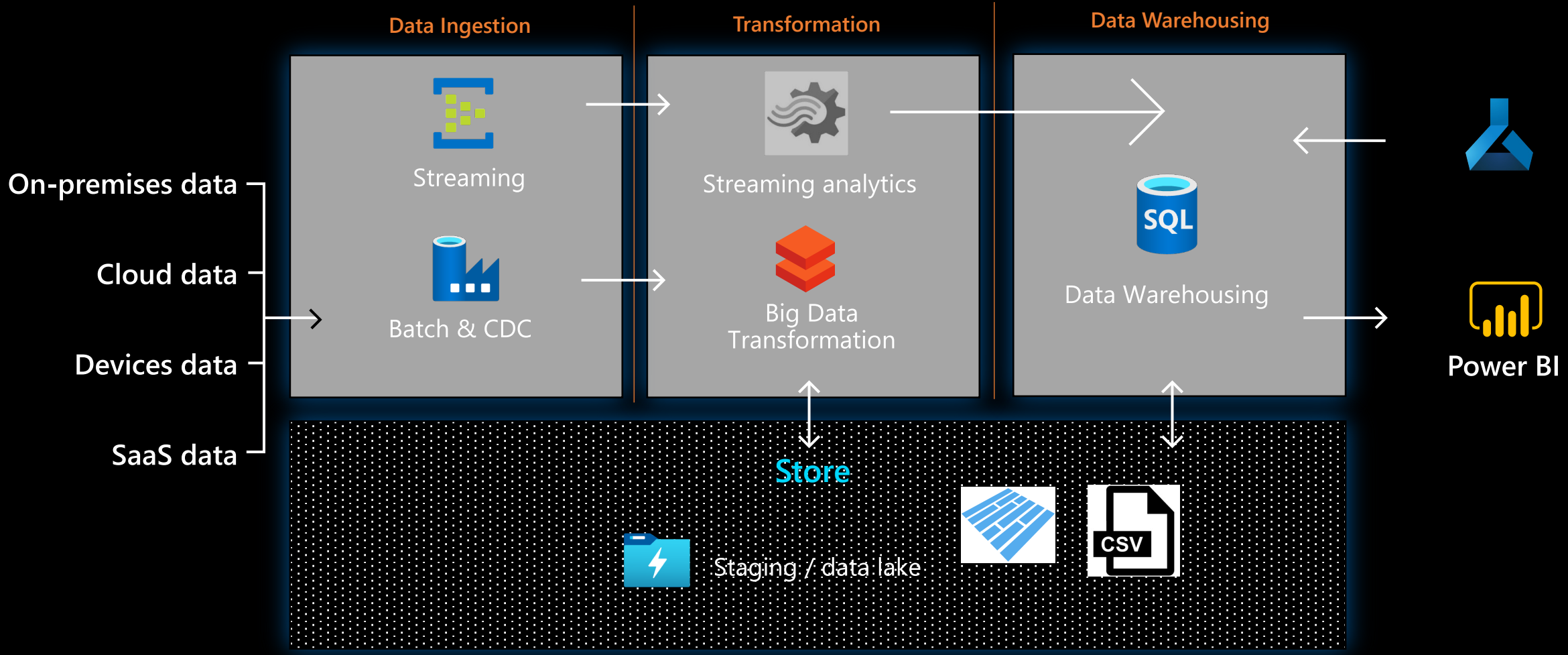
https://brooklyndata.co/blog/benchmarking-open-table-formats
https://www.onehouse.ai/blog/apache-hudi-vs-delta-lake-vs-apache-iceberg-lakehouse-feature-comparison

Data lakehouse

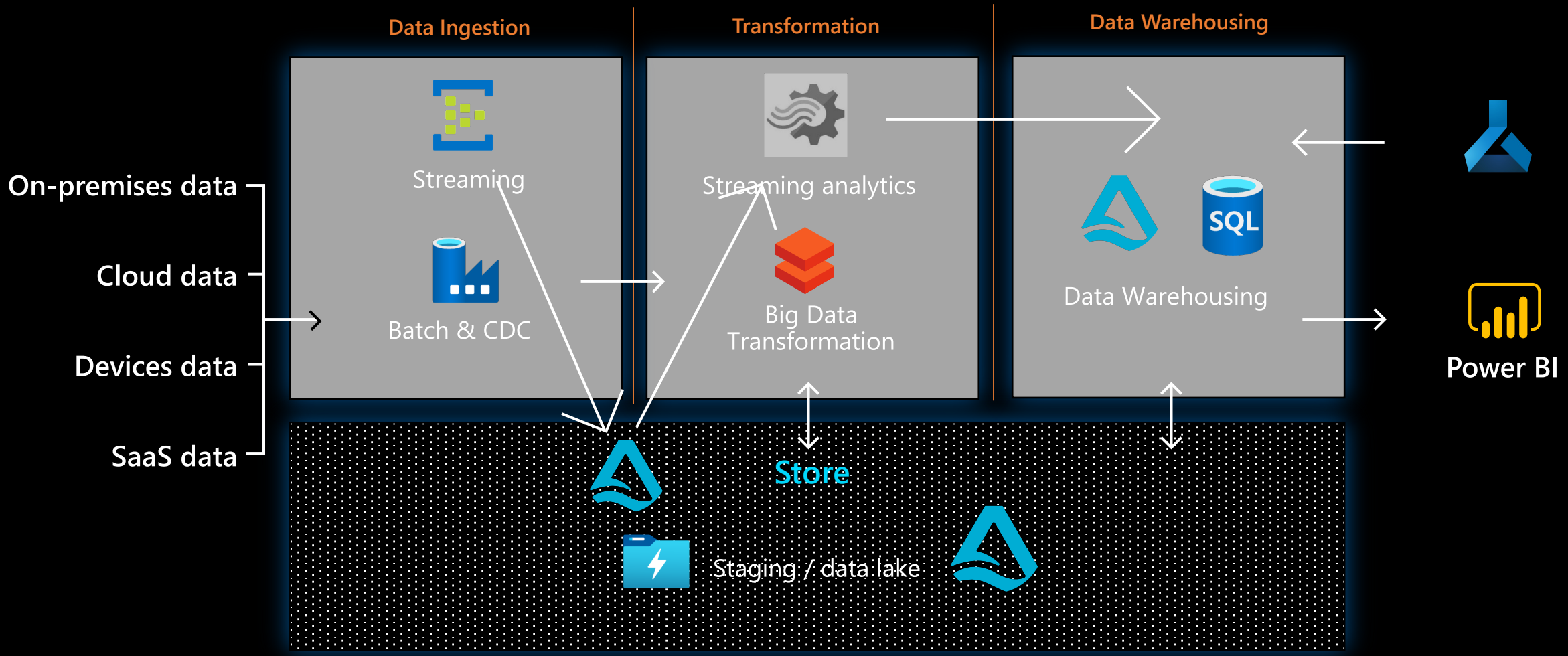
Delta lake at the heart of the Azure lakehouse



Typical data platform in Azure



Now with Delta Lake

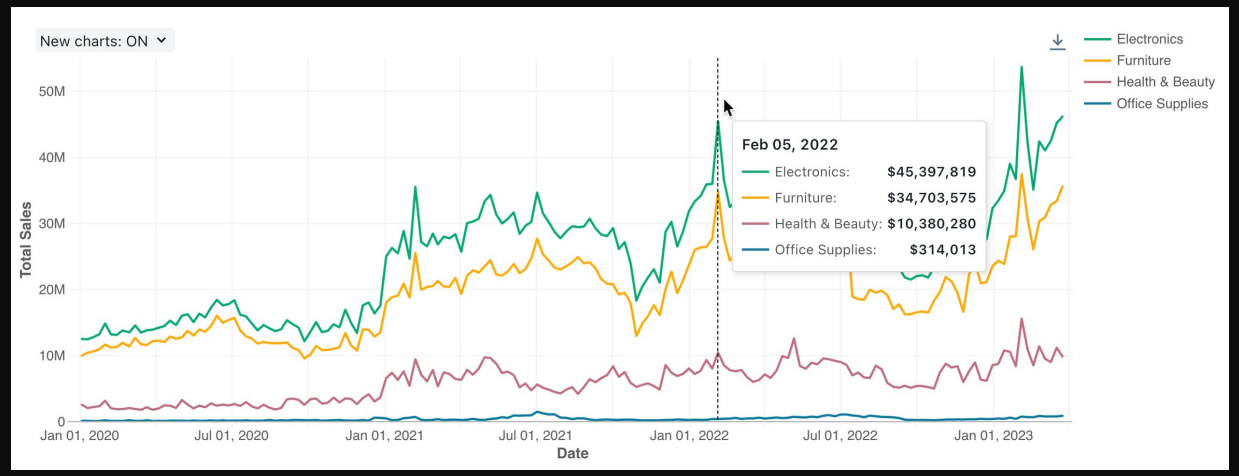


Data lakehouse

- Compressed files in data lake - cheap storage vs database
- Compute needed on read
- And which endpoints to use

- Databricks has announced lots of features in 20/21
- - Photon – analytical query engine
- - Serverless SQL
- - Delta Live Tables
- - Unity Data Catalog

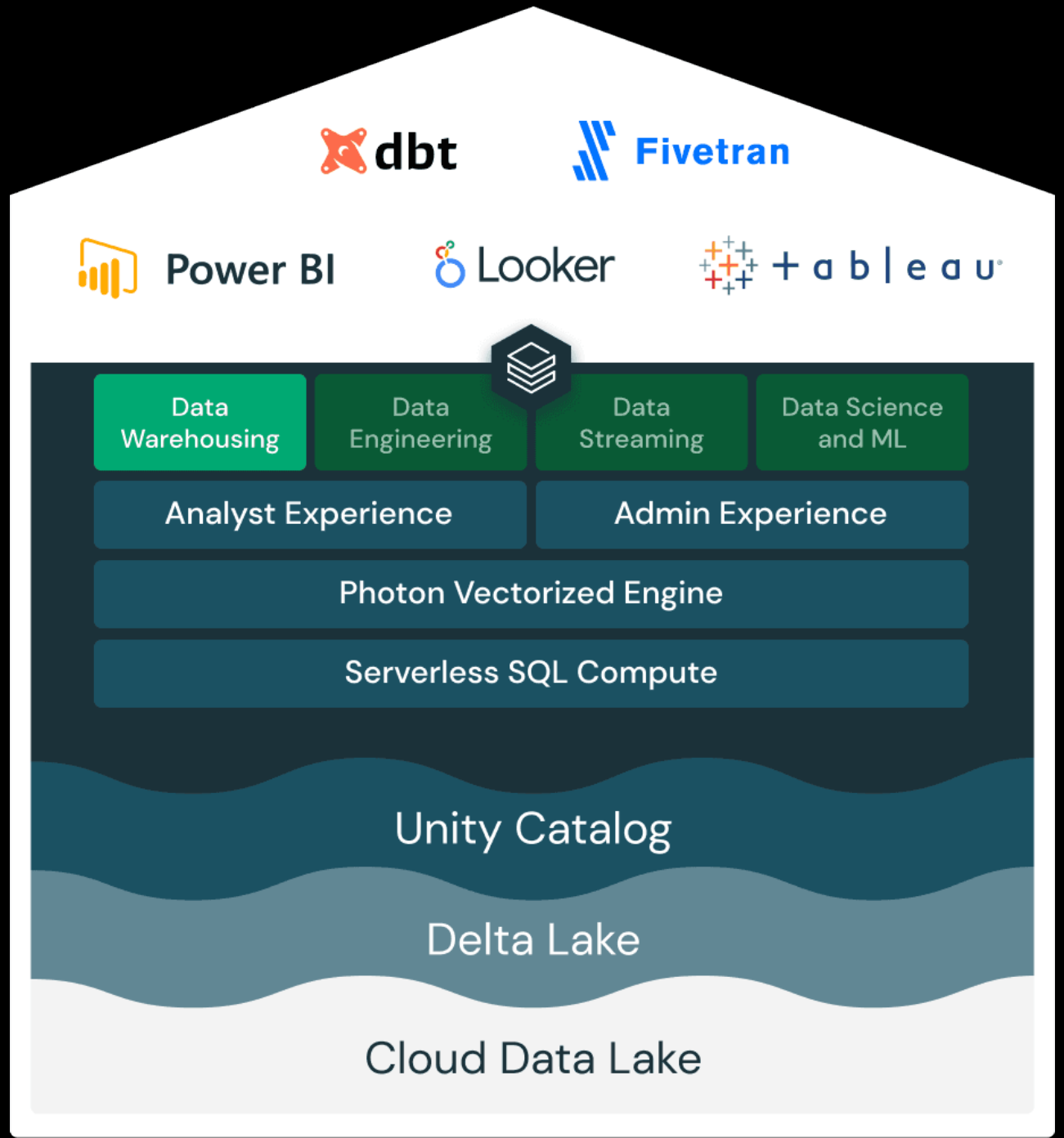
- All components in the data lakehouse architecture



Latest Databricks news...

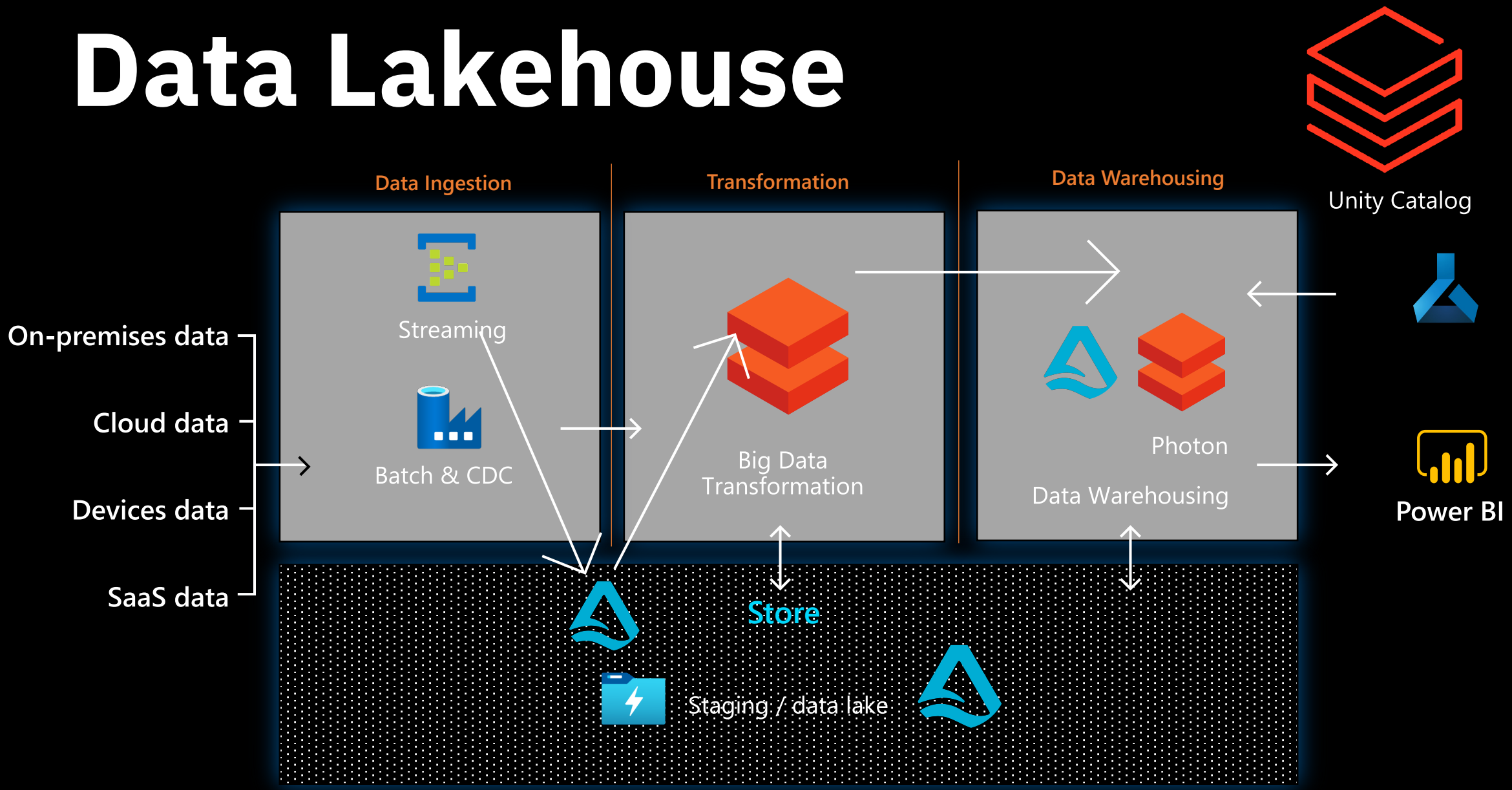
Databricks Agrees to Acquire Arcion, the Leading Provider for Real-Time Enterprise Data Replication Technology

October 23, 2023


dbt
Fivetran
Power BI
Looker
tableau
Data Warehousing
Data Engineering
Data Streaming
Data Science and ML
Analyst Experience
Admin Experience
Photon Vectorized Engine
Serverless SQL Compute
Unity Catalog
Delta Lake
Cloud Data Lake

Data lakehouse

Data Lakehouse



Azure Synapse analytics

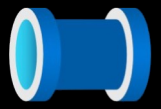
One name, but several different services...

- Azure Data Factory
- Azure SQL Serverless
- Azure SQL Data Warehouse
- Azure Spark
- Synapse Studio
- Azure Data Explorer (preview)
- Azure Stream Analytics
- Azure ML – partially... (ONNX)

Azure Data lakehouse

- Synapse Analytics – the second data lakehouse compatible system…

- Supports delta lake natively in Azure Spark

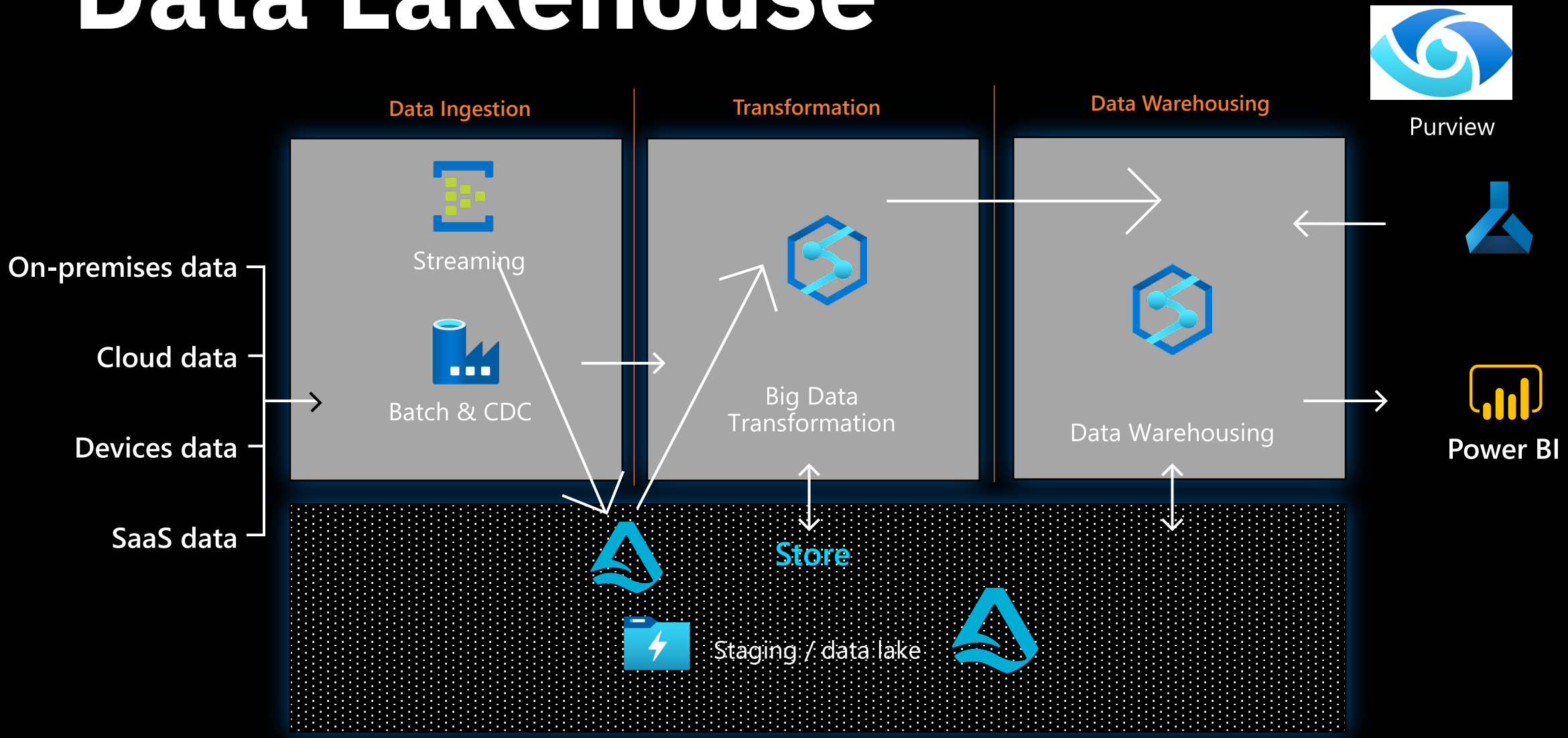


- Synapse pipelines (and Data Factory) supports delta lake source and sink

- Serverless SQL pools reads delta lake and can build views etc on top.

- Power BI supports delta lake as well

Data Lakehouse



Microsoft Fabric

Microsoft Fabric

One name, but several different services...

- Power BI
- Azure Data Factory
- Azure SQL Serverless
- Azure SQL Data Warehouse
- Azure Lakehouse
- Azure Spark
- Azure Data Explorer
- Azure Stream Analytics
- Azure ML

Microsoft Fabric

OneLake is the core of Fabric

- OneDrive for your data
- Fully managed data lake service
- Delta «Parquet» is default format – for most services

Microsoft Fabric

Azure Lakehouse is the native data lakehouse offering

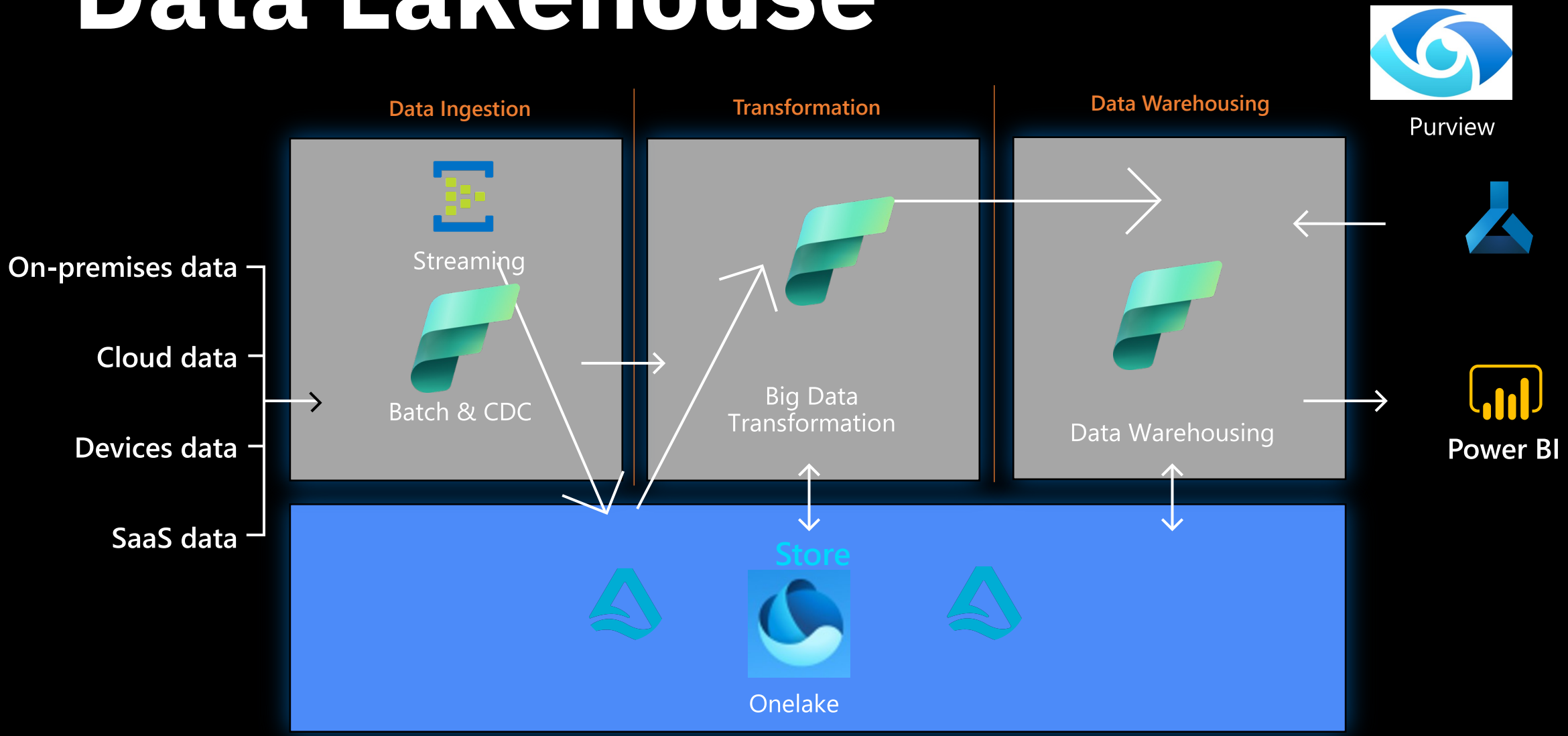
Data can be written with Spark or Dataflows

Data can be read with SQL, Spark or BI tools (Power BI)

Supports XMLA

Offers APIs

Data Lakehouse



Johan Ludvig Brattås

Director, Deloitte

/johanludvig

@intoleranse

jbrattas@deloitte.com

GitHub

Chronic volunteer

Co-organizer – DataSaturday Oslo
President – MDPUG Oslo
Frequent voulenteer in general

When not geeking out over new tech

Teaching coeliacs how to bake gluten free
Baking
Hiking
Gardening

MVP-Dagen 2023

Questions?