Capgemini

Building a data lakehouse in Azure

Johan Ludvig Brattås

AGENDA

- Data lakes and data warehouses
- What is a data lakehouse?
- 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

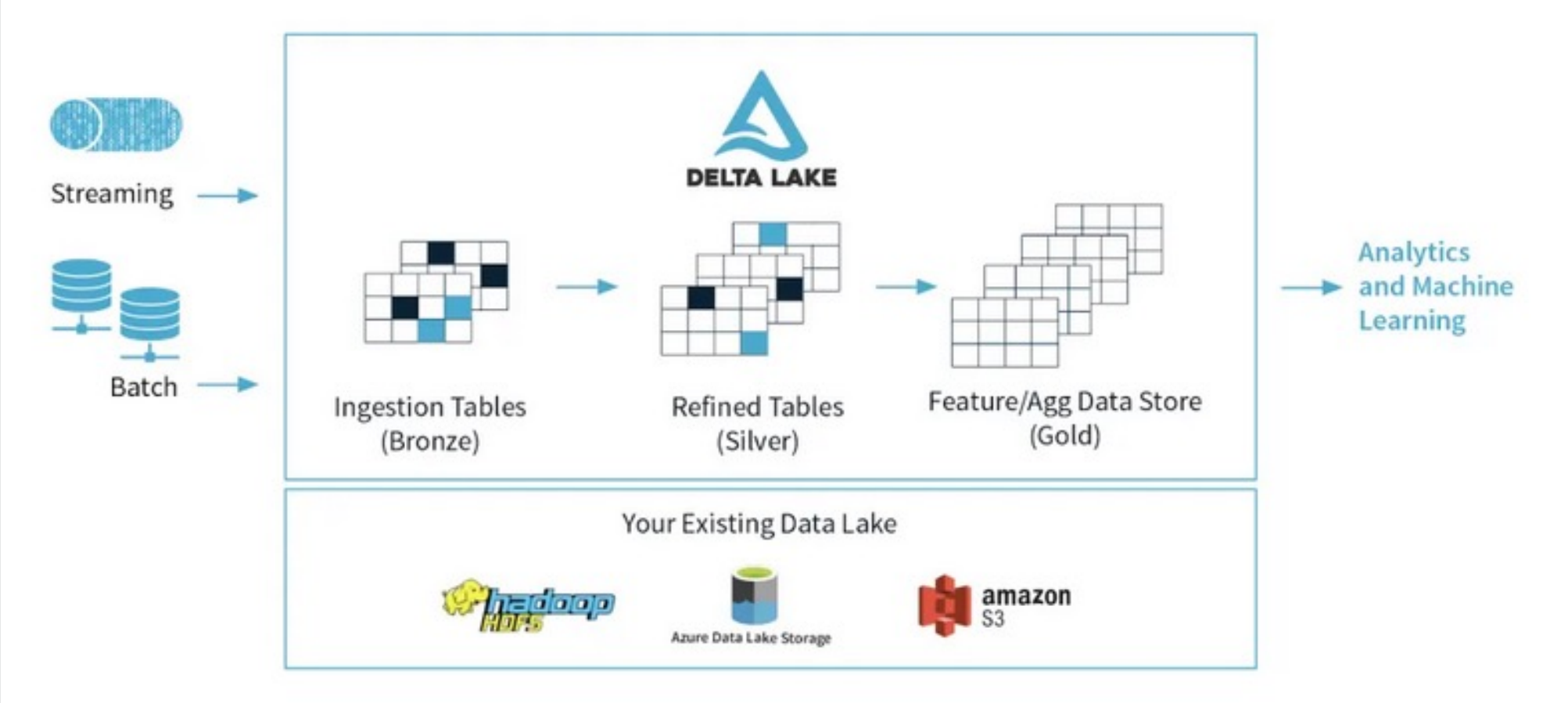
What is a delta lake?

Delta lake at the heart of the lakehouse

- Delta lake is 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

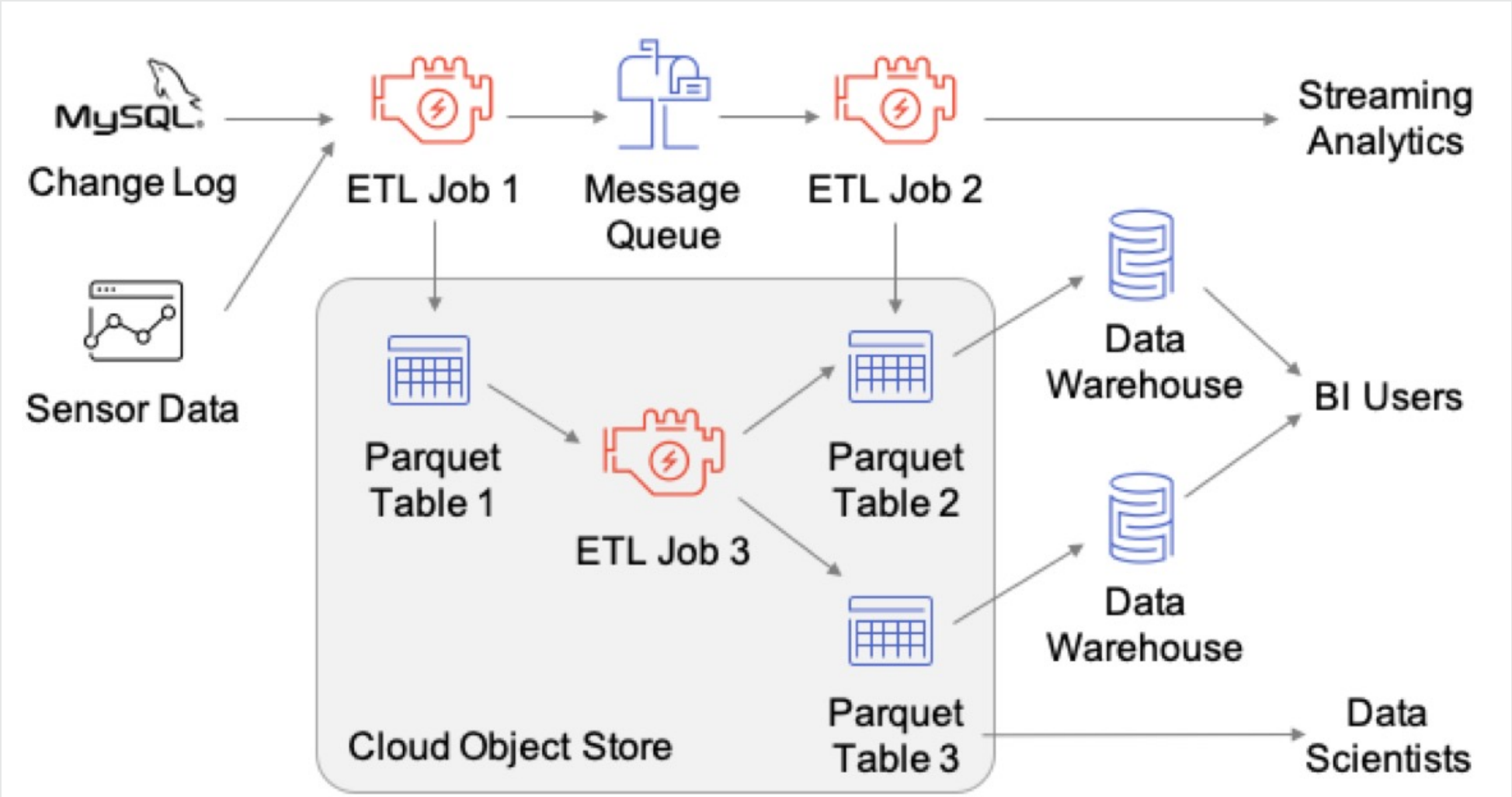
Data lakehouse

Delta lake at the heart of the lakehouse



www.delta.io

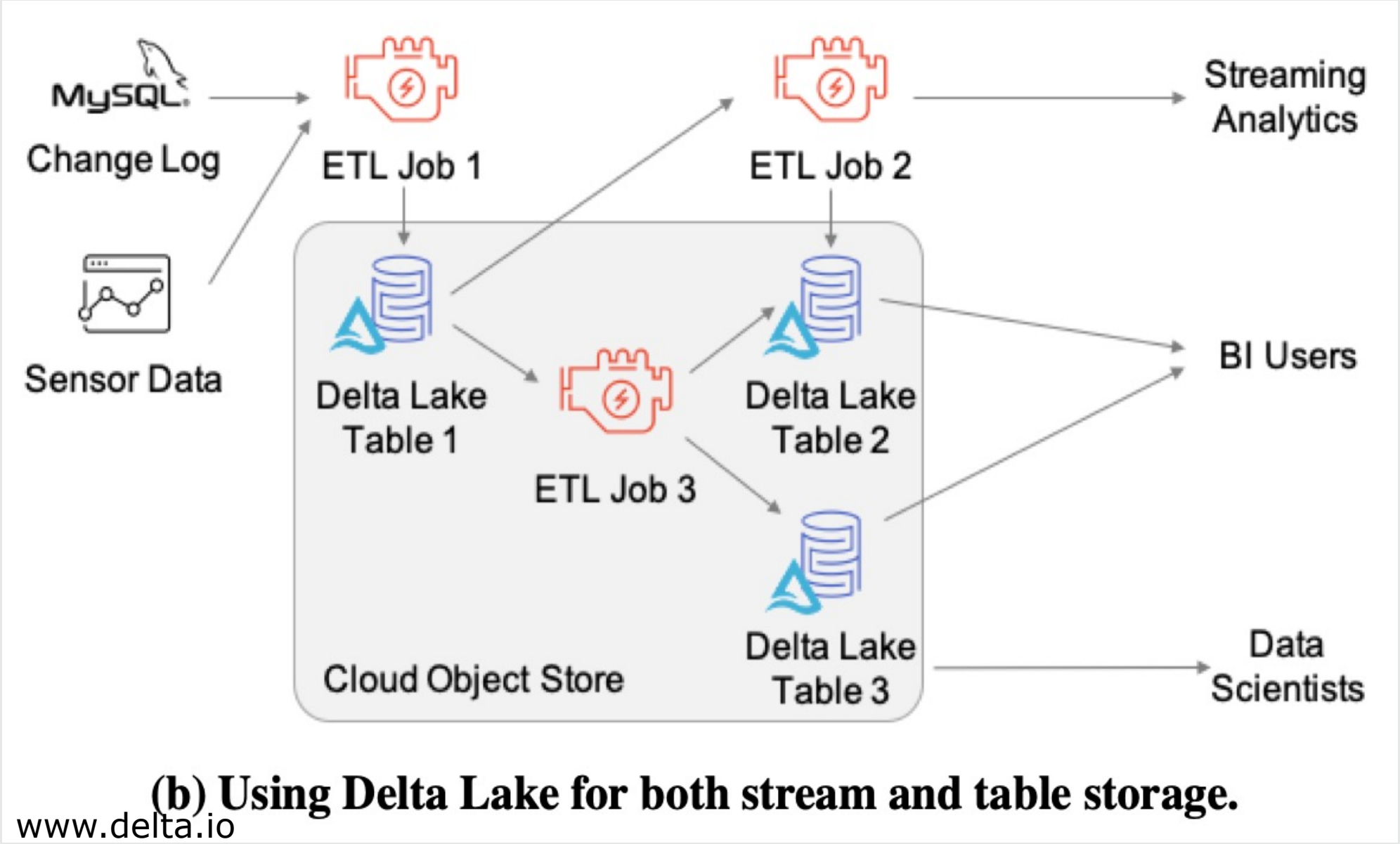
Data lakehouse



(a) Pipeline using separate storage systems.

www.delta.io

Data lakehouse



(b) Using Delta Lake for both stream and table storage.

www.delta.io

Data lakehouse

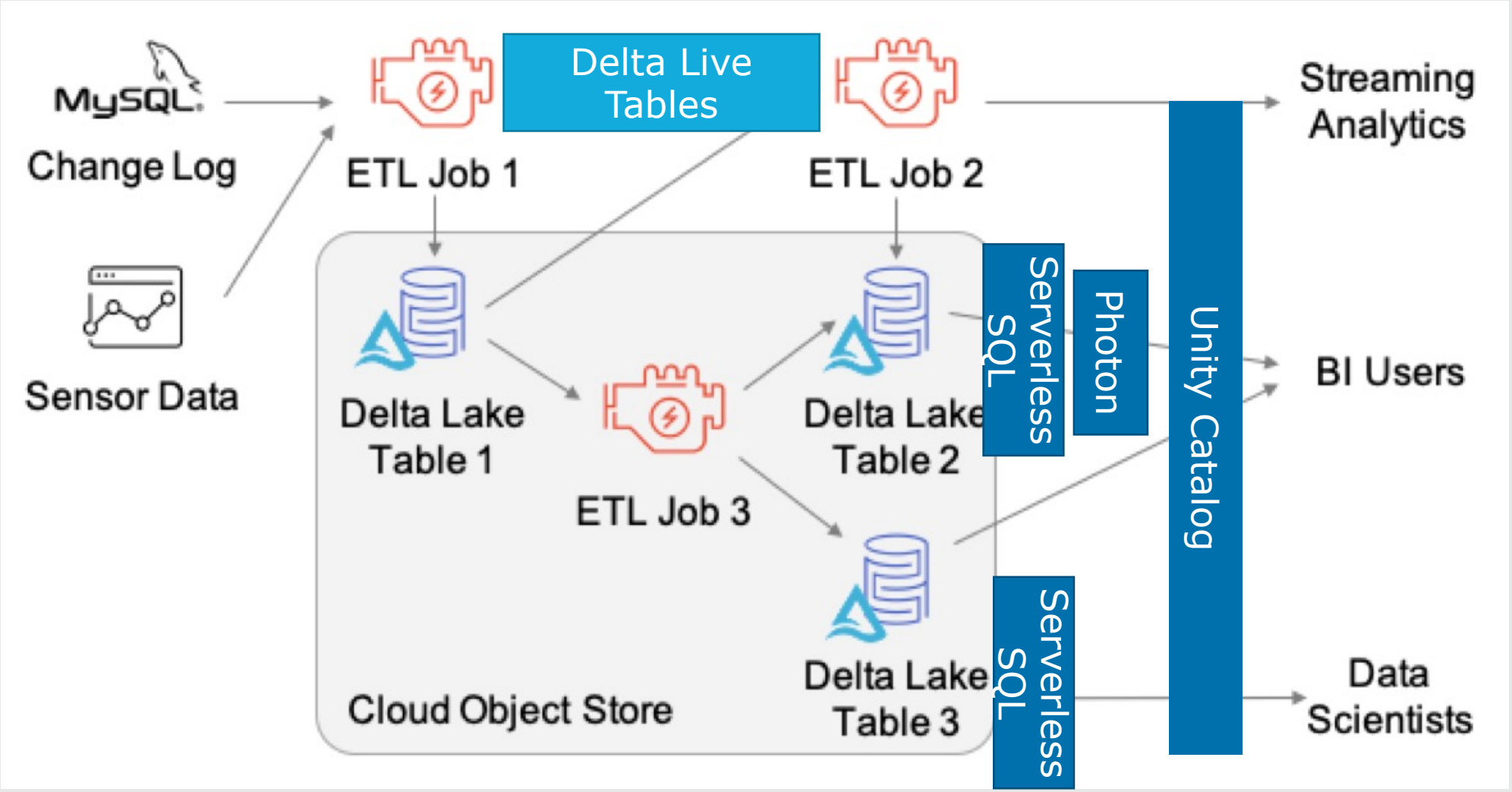
Data lakehouse

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

Data lakehouse



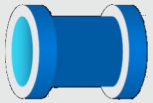
Delta Lake used for both stream and table storage

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)

Synapse Analytics – the second data lakehouse compatible system…

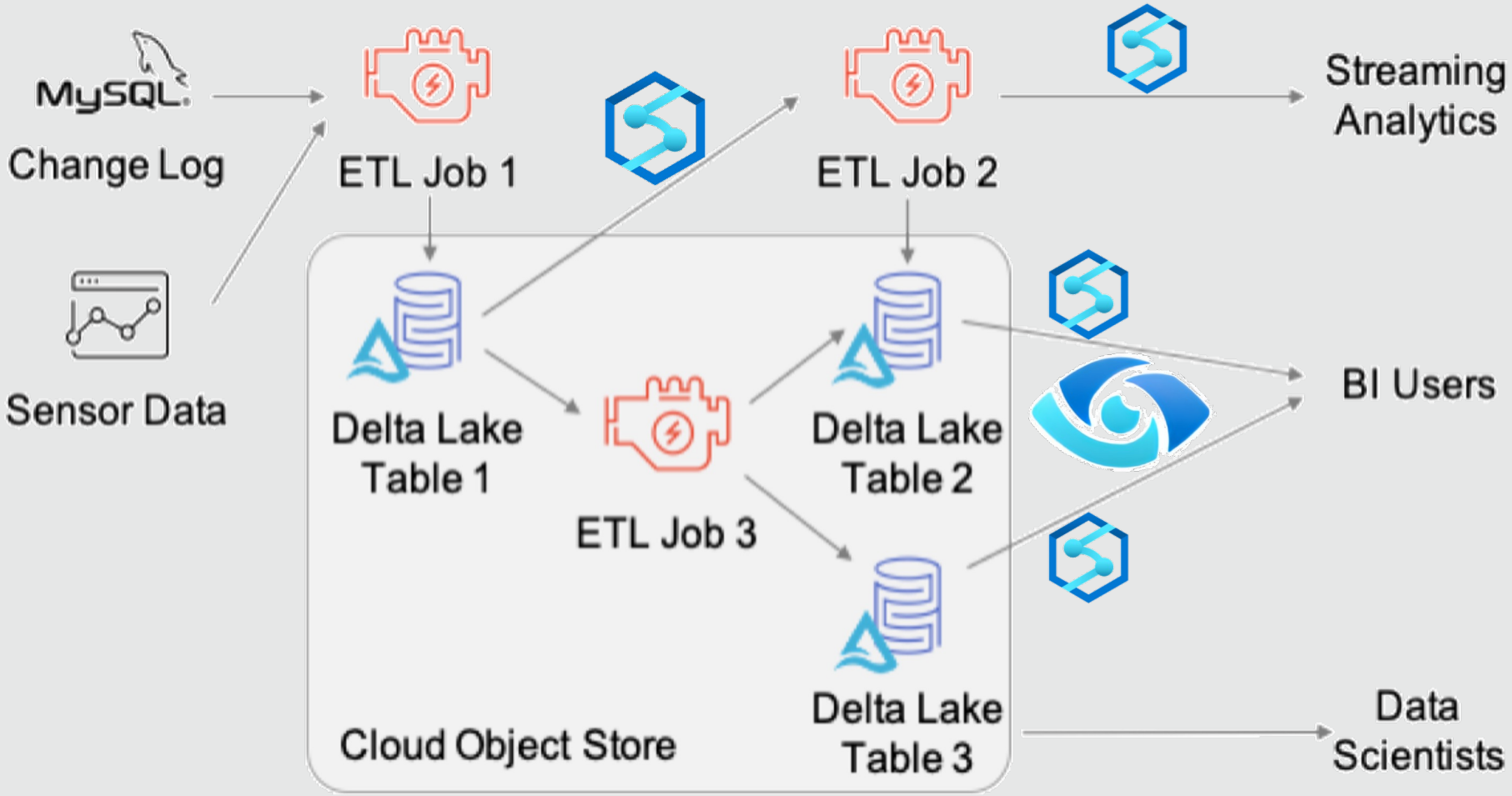
Supports delta lake natively in Azure Spark



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



Delta Lake used for both stream and table storage

Johan Ludvig Brattås

Lead Solutions Architect, Capgemini

/johanludvig

@intoleranse

johan-ludvig.brattas@capgemini.com

Chronic volunteer

Co-organizer – SQLSatOslo
Ex Virtual Group lead – Excel BI
Board member – MDPUG Oslo
Frequent voulenteer in general for PASS

When not geeking out over new tech

Teaching coeliacs how to bake gluten free
Baking
Hiking
Gardening

Questions?

Thank you!