A DATA LAKEHOUSE STORY

09 April 2024

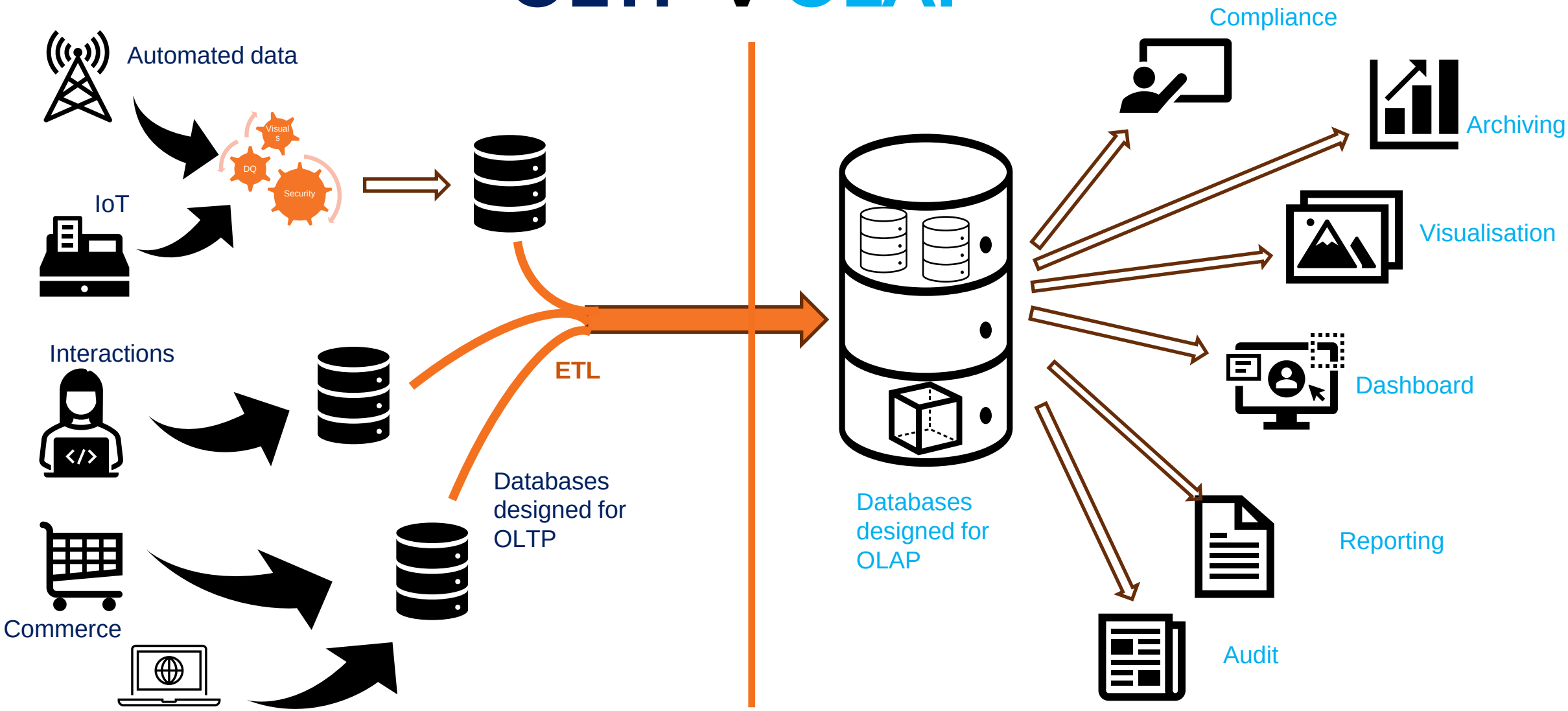
Data Architecture Event, Corinium

Michel Gehin, Data Architect

Case Study - how new technology has enabled:

- the deployment of a Data Lakehouse
- Its multipurpose
- Business benefits and savings

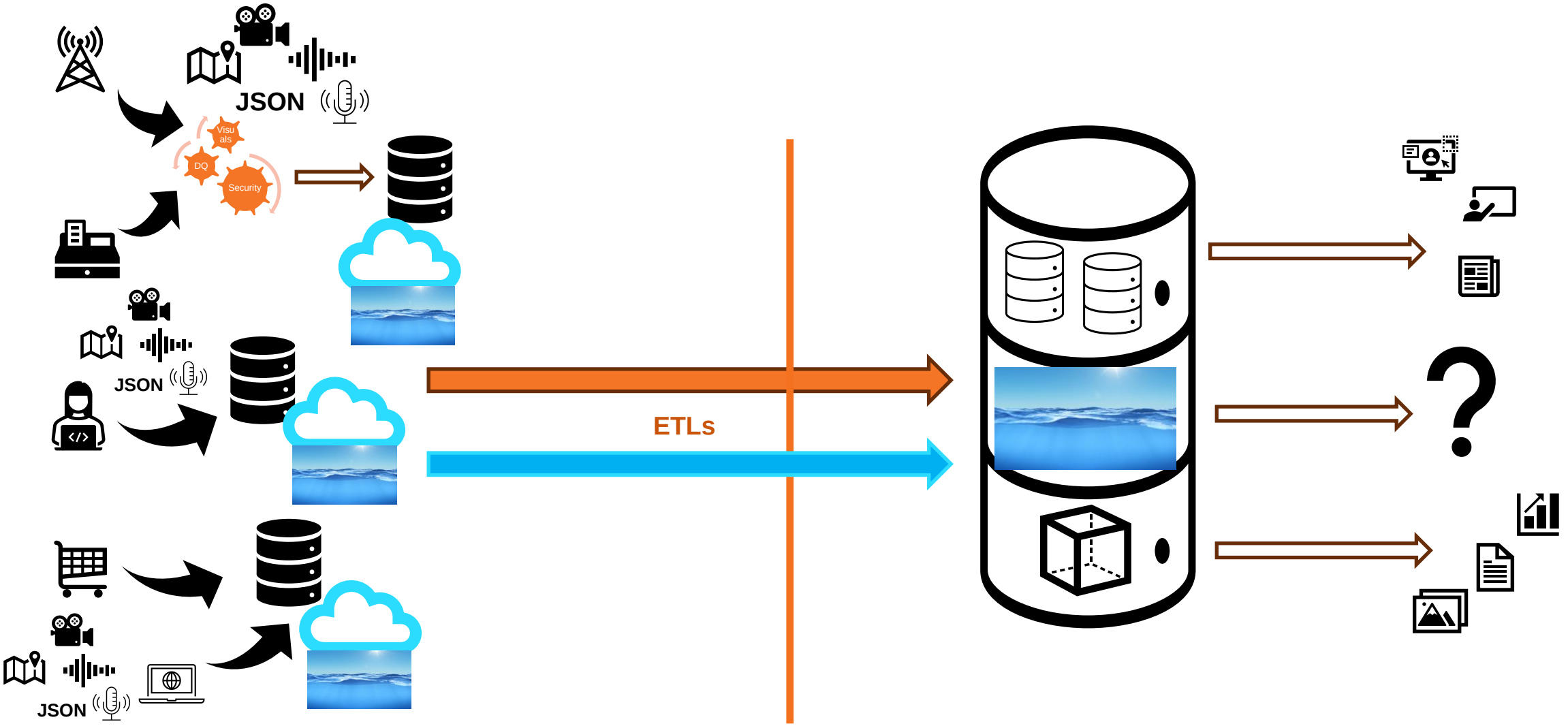
OLTP v OLAP



Fast transactional capabilities
Creation and updates of individual records
Lots of narrow transactions on a small amount of data

Fast querying of large datasets
Smaller number of attributes over aggregated stores and cubes.
Paired with a BI tool, semantic layers

New data sources and formats



+ Unstructured data
+ Semi-structured data

Traditional DW
+ Data Lake

OLAP-specific Spend

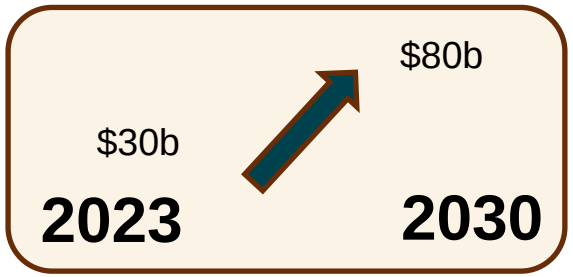
Warehouse Builders

Specialised skills: DW developers, DW architects

ETL Tools

STORAGE

Cost points



Sources

https://www.scnsoft.com/analytics/data-warehouse/pricing

https://www.integrate.io/blog/the-true-cost-of-a-data-warehouse/

Implementation Costs

A typical NZ company, with 5 internal sources of data (ERP, CRM etc), with just structured data batch-processed, hybrid of cloud and on-premise DW stores, manual DQ tools and limited ML/AI capabilities will **spend between $50,000 and $200,000 *building* a data warehouse**, not including BI tools

Oracle Warehouse Builder

Amazon REDSHIFT

Snowflake

ETL Costs

will depend on organization, between $800 and $8,000 per month

WhereScape RED

IBM Infosphere

Talend

Running Costs

Current estimations:
ON-PREMISE: $1,000 per terabyte per month

CLOUD: up to $84 per terabyte per month in the cloud (hot storage + basic compute – cold is cheaper).

The average size of a DW worldwide is around **100 terabytes**.

SAS Data Management

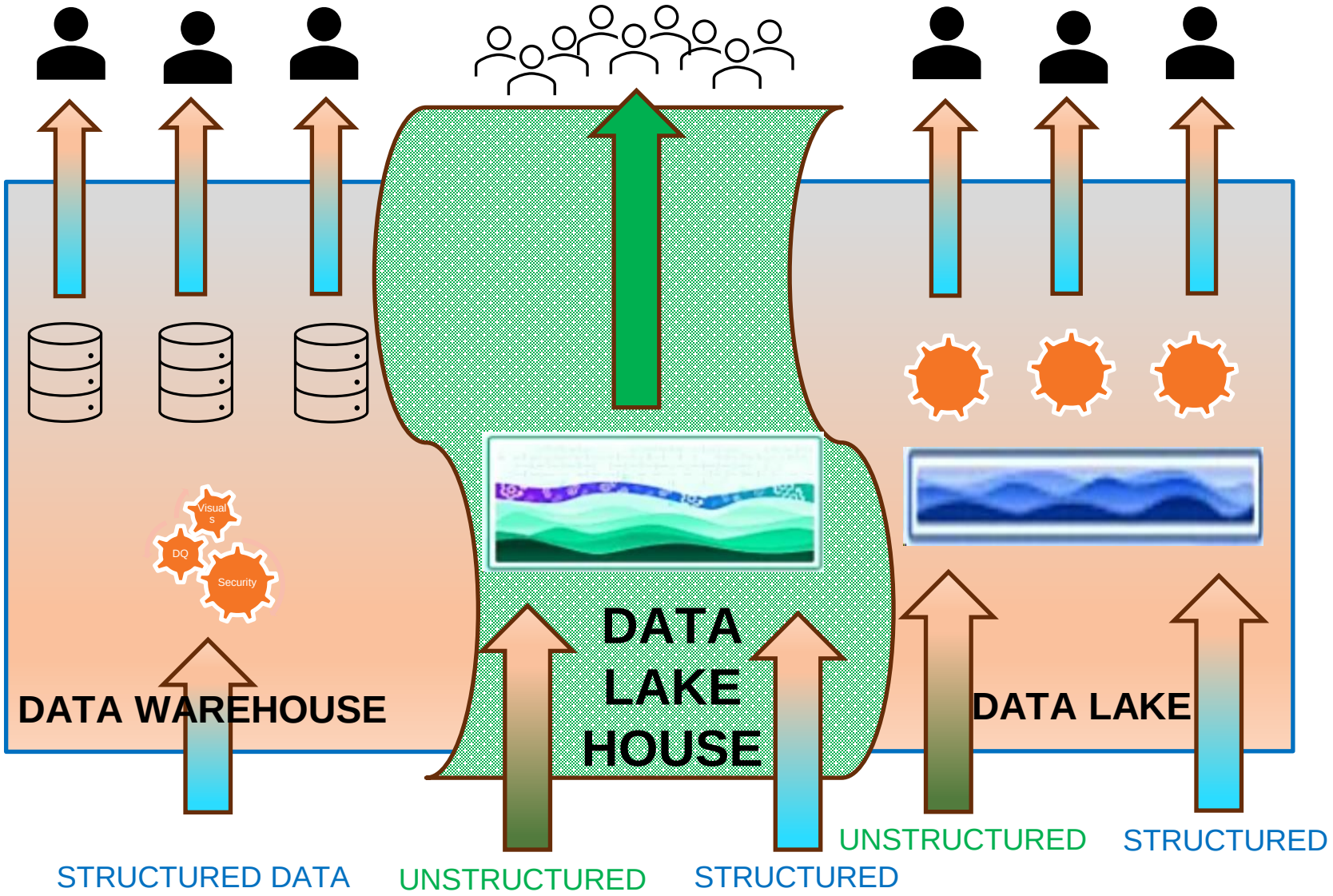
Google Big Query

Microsoft SSAS

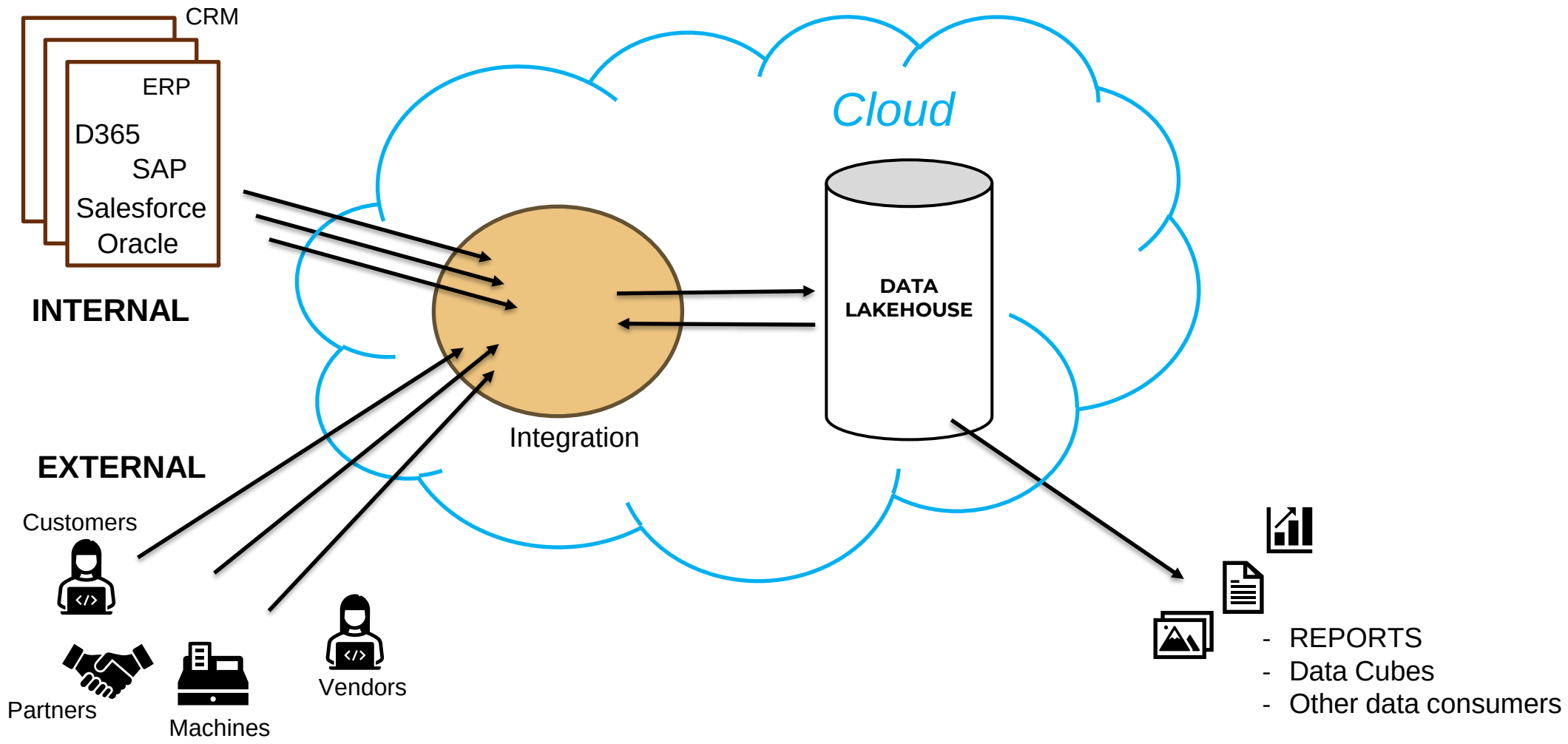
Specialised Staff Costs

$???k per month

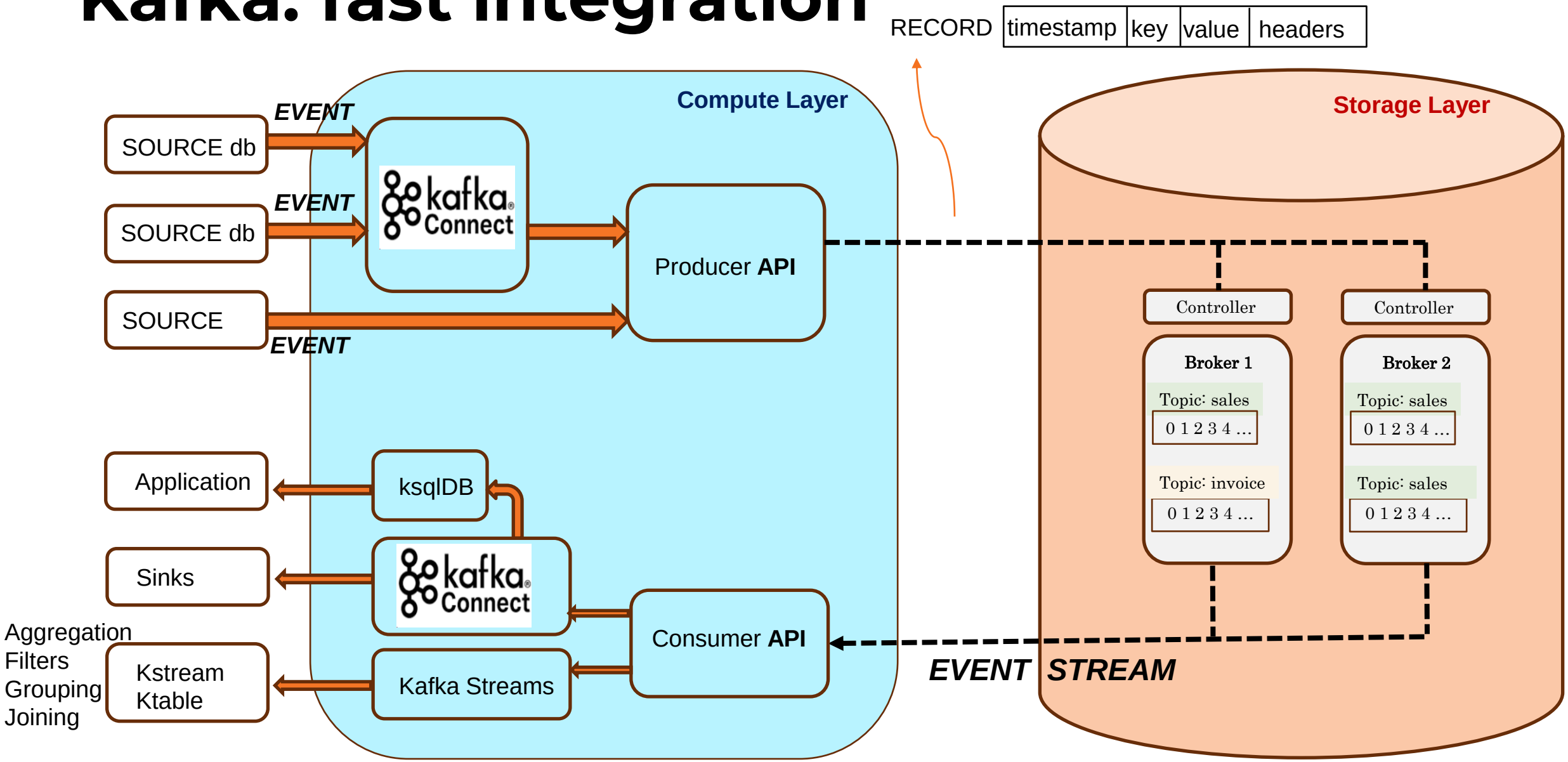
The Data Lakehouse



Integration and Lakehouse

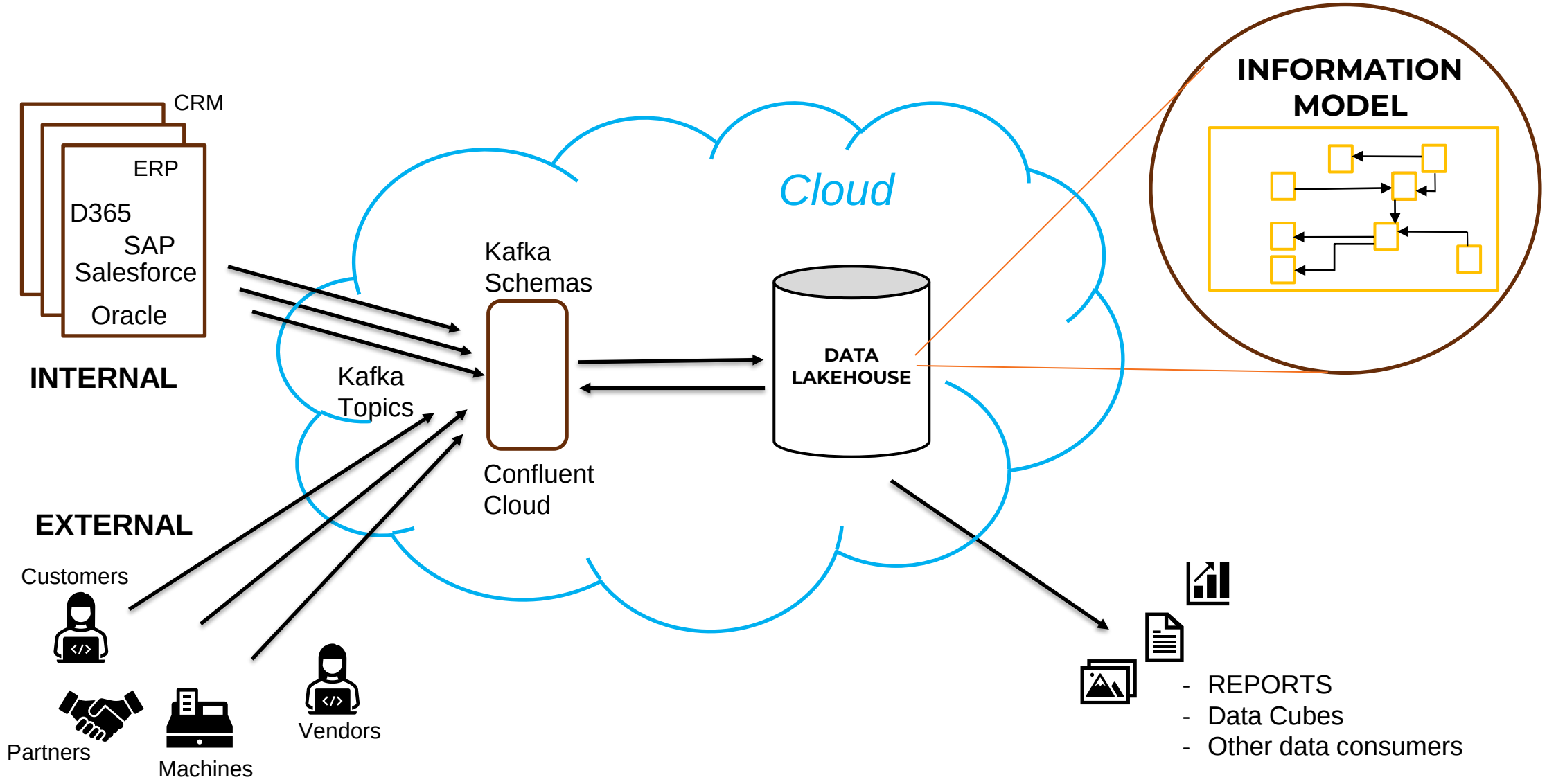


Kafka: fast integration



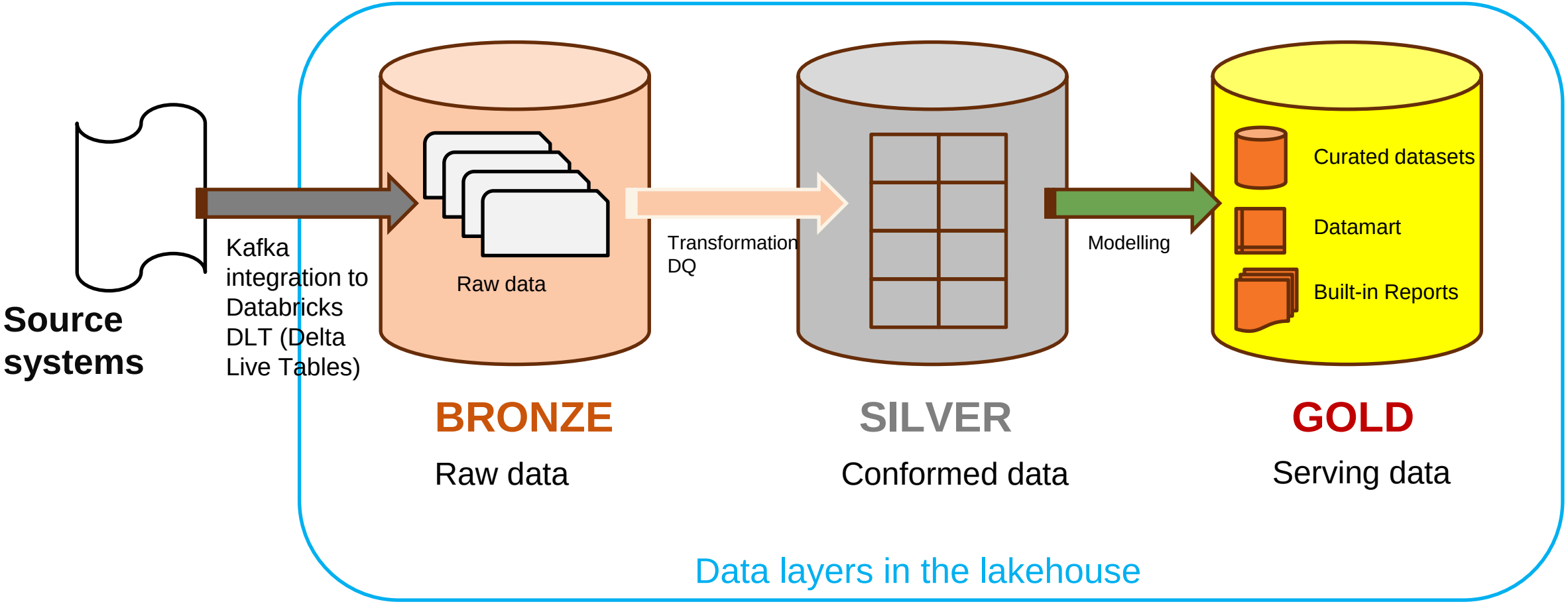
Benchmark: https://engineering.linkedin.com/kafka/benchmarking-apache-kafka-2-million-writes-second-three-cheap-machines

Data Architecture Overview

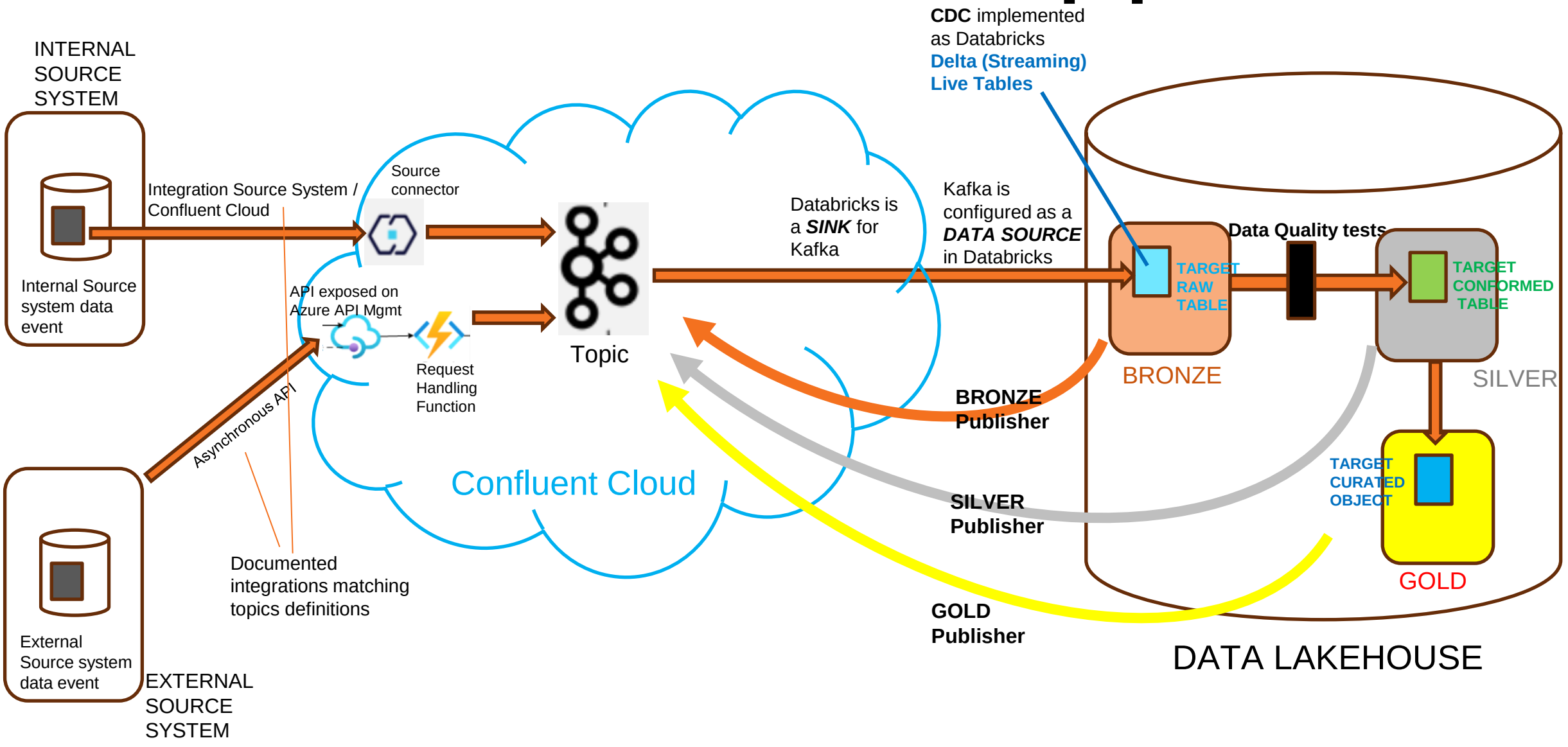


The Data Lakehouse – the physical view

The Data Lakehouse is composed of several physical asset layers, as follows:



The Data Lakehouse – data pipelines



Data event example

SAP updates the status of an order in its SalesOrder table

1

Integration: the Event Notification integration will ensure the producer system publish the data change in the salesorder.upd.db.prd Kafka Topic

Most of those will be **near real-time** although there are some exceptions depending on the source system:
- near real-time will be milliseconds
- Other sources might be batched

2

The Kafka topic is published by a Kafka Broker in the SalesOrder raw table in the lakehouse's **BRONZE** instance

CDC using Databricks Delta Streaming Live Tables:
"APPLY CHANGES INTO … FROM STREAM(.. "

SAP is **master** of that data so **the update is valid**

3a

The raw entry is transformed and data put in the conformed table [SALES ORDER] in the lakehouse's **SILVER** instance

The transformation includes **DQ checks**

4

The conformed entry is published onto curated data sets, datamarts in the **GOLD** instance; any report launched from that point on will contain the new data.

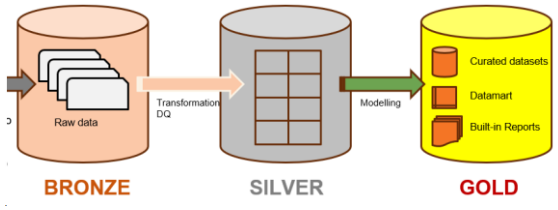
3b

5

D365 needs to know the change in status as soon as possible – before anything else

The Finance portal needs to know the status change but only once validated, and needs augmented data – even if it means a slight delay

How to consume data from the GOLD layer of the Lakehouse



CURATED DATA SETS

The GOLD layer will include curated data sets designed and built according to core business needs – for example, total sales with stores physical locations together.

DATAMART ACCESS

The data in the conformed information model will also be available, with data consumers able to join tables together with knowledge of the documented Information Model.

BUILT-IN REPORTS

The GOLD layer will also contain ready-to-use reports

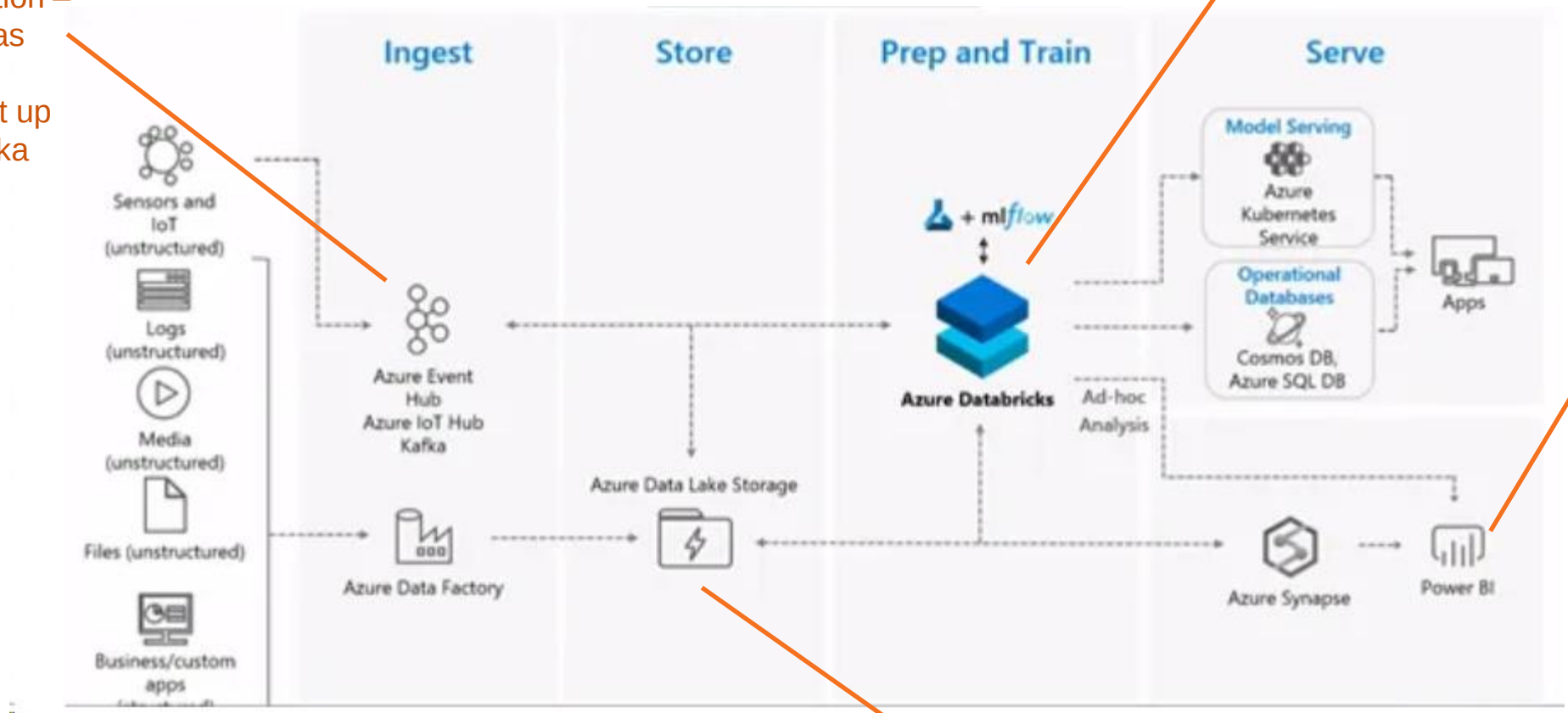
The data is DQd

The data has been classified and tagged, with security classification, privacy, sensitivity, domain, version, etc.

Databricks

Includes tools for analytics, ML, AI, ..

Kafka integration – Kafka set up as source, while Databricks set up as sink in Kafka



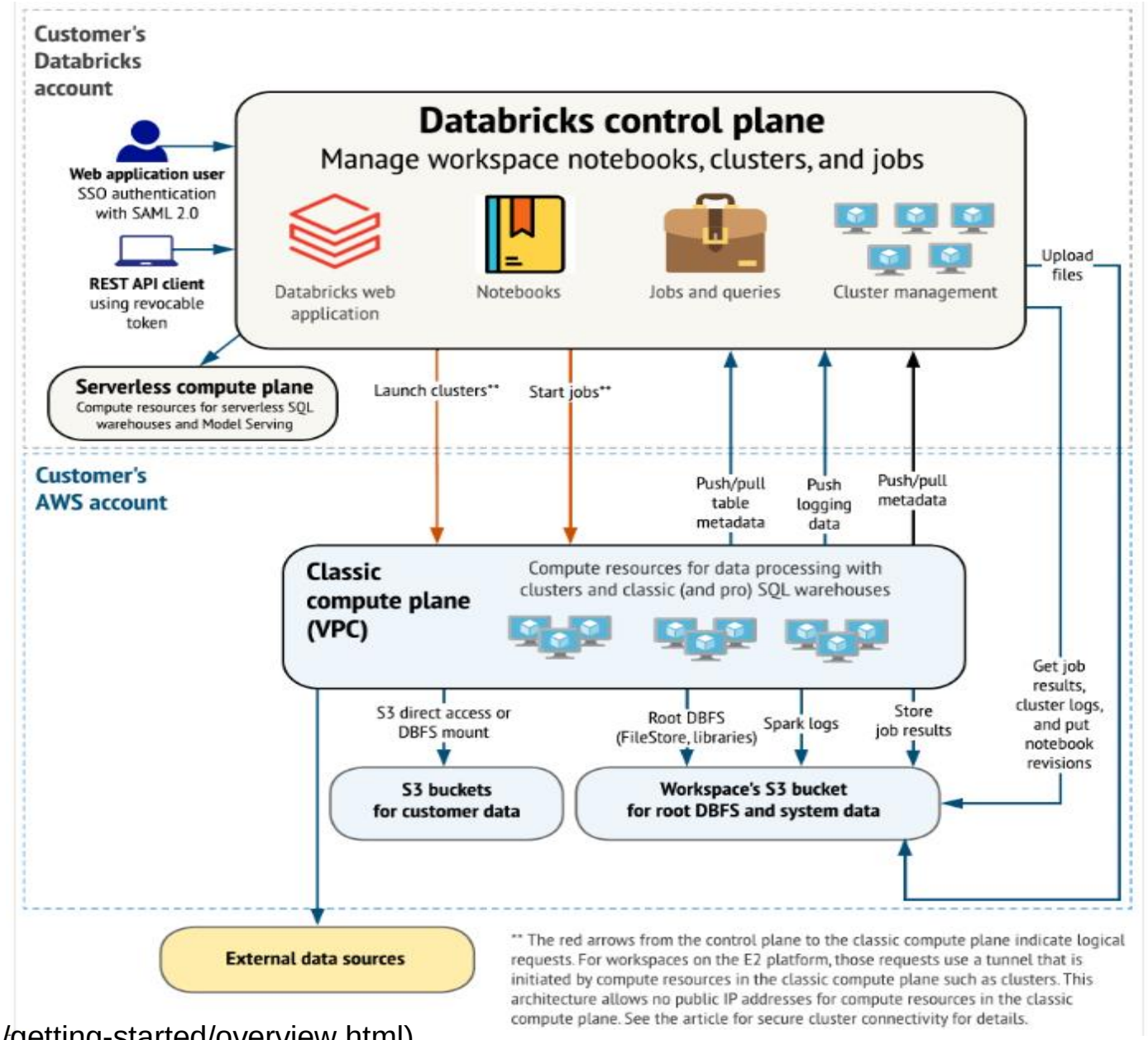
Connectivity for reporting tools such as Power BI

Physical Data Lake: Azure Gen 2 Data Storage

Includes backup, recovery, DR

(SOURCE: https://anuragsingh6701.medium.com/some-of-azure-databricks-best-practices-ba8317d0509e)

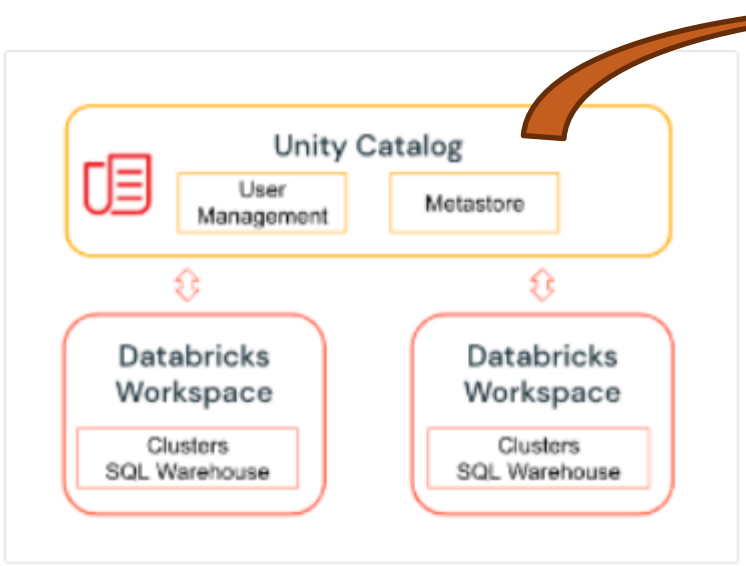
Databricks Architecture

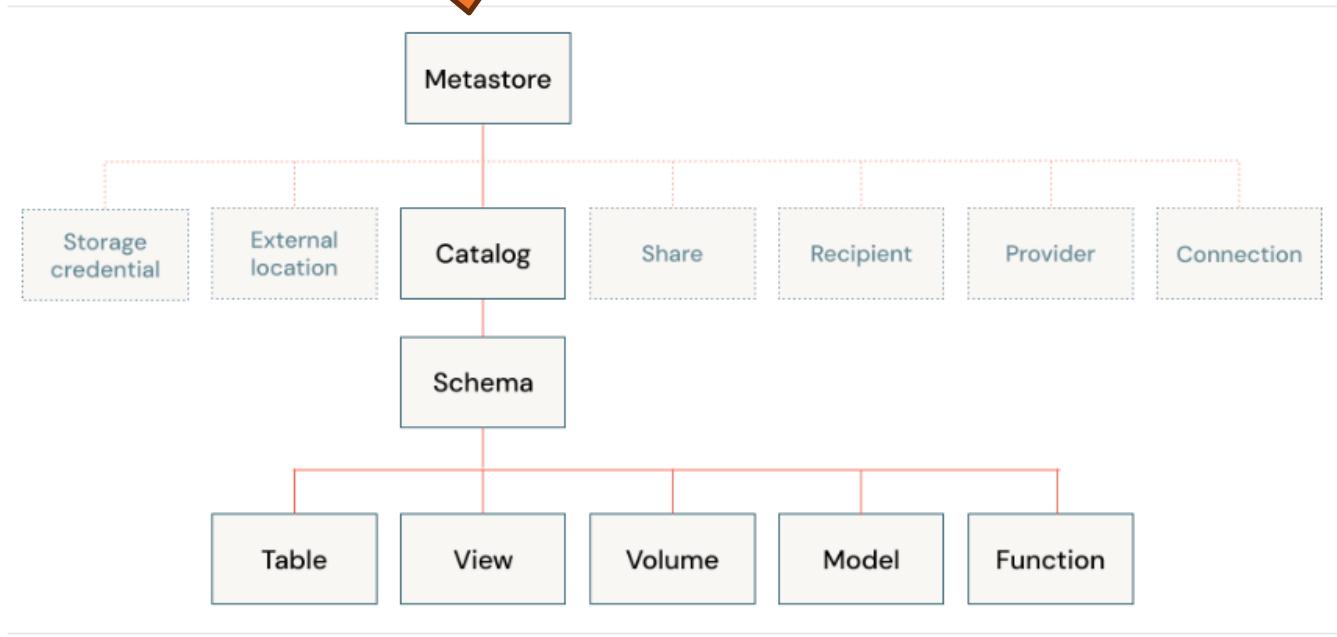


(SOURCE: https://docs.databricks.com/en/getting-started/overview.html)

Databricks Unity Catalog

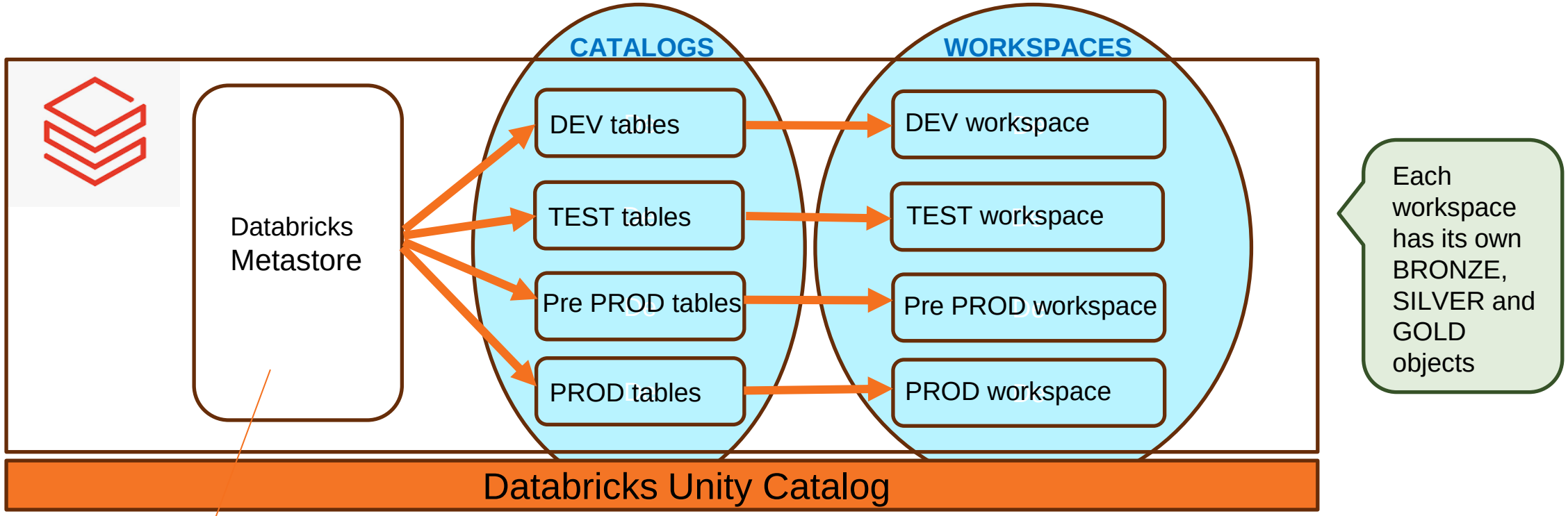
A one-stop, single-view management tool for all workspaces, for all layers, across all environments





Environments on Databricks

Leverage Databricks Unity Catalog to manage all 4 environments, each with 3 layers, with one Data Lakehouse / one Catalog
Unity Catalog enables the separation of data objects across catalogs and workspaces, isolating catalogs from one another.

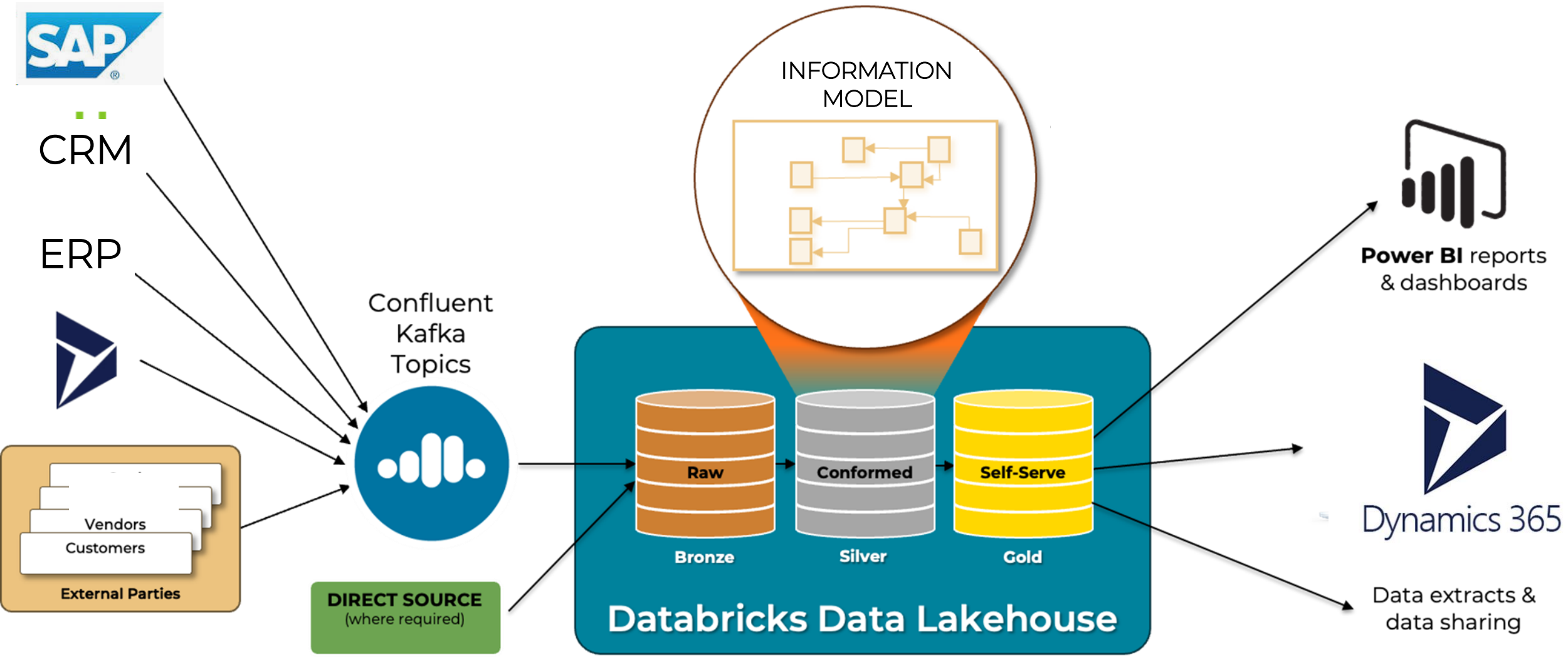


Users management
Security, access policies
Volumes management

Data Lake

THE LAKEHOUSE

FOR BOTH OLTP AND OLAP



Summary

Smart use of tech has enabled a more simple, coherent, data landscape:

- Each function carried by a **component that fits** the business needs
- **Integration** between each component is key
- Track **un-necessary duplication** of function and eliminate
- Pay for the **speed** and **volume** you need now, ensure scalability for future needs
- You don’t necessarily need best of breed, just **best fit** to your business
- Today’s offerings in **cloud** and **big data** enable you to pick each component separately, as long as the integration is at least “almost native”

Thank you