Datalake y DataWarehouse

BASES DE DATOS PARA IA

QUE ES UN DATAWAREHOUSE?

- CONTIENE INFORMACIÓN DE BASES DE DATOS TRANSACCIONALES
- SUMA INFORMACIÓN EXTERNA PARA VISTAS 360
- ORIENTADO A BIG DATA
- DATA ESTRUCTURADA DEL TIPO RELACIONAL
- ACID COMPLIANCE ES MENOS ESTRICTO
- AYUDA A CONSEGUIR INSIGHTS DE NEGOCIO EN INTELIGENCIA DE NEGOCIOS O BUSINESS INTELLIGENCE (BI)

Google BigQuery

CARACTERÍSTICAS

- SUBJECT ORIENTED:

  SE ORIENTA A SOLUCIONAR PROBLEMAS DE NEGOCIO CON DATOS

- INTEGRADO:

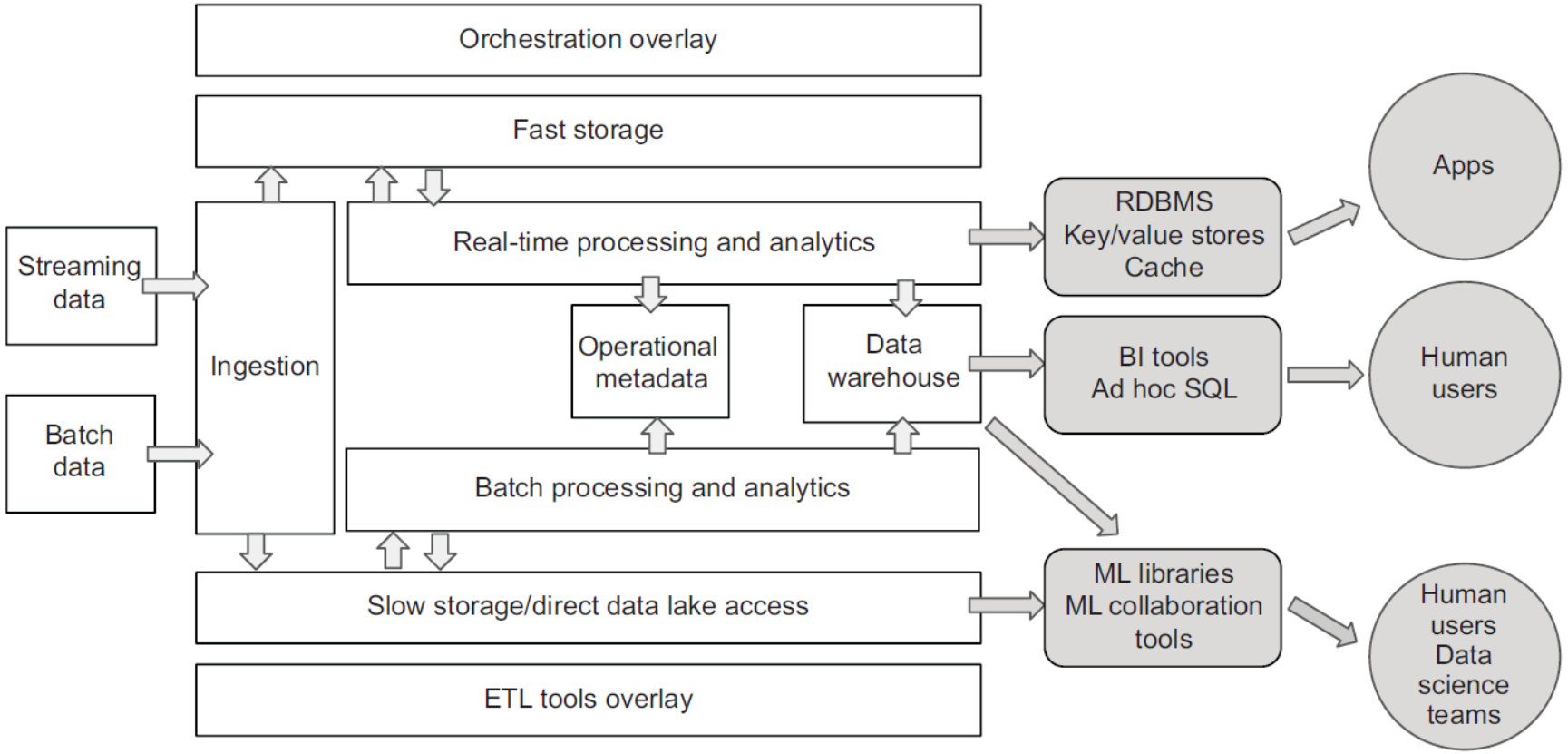
  RECIBE INFORMACIÓN DE DIFERENTES LUGARES CON EL FIN DE INTEGRARLOS

- VARIANTE EN EL TIEMPO

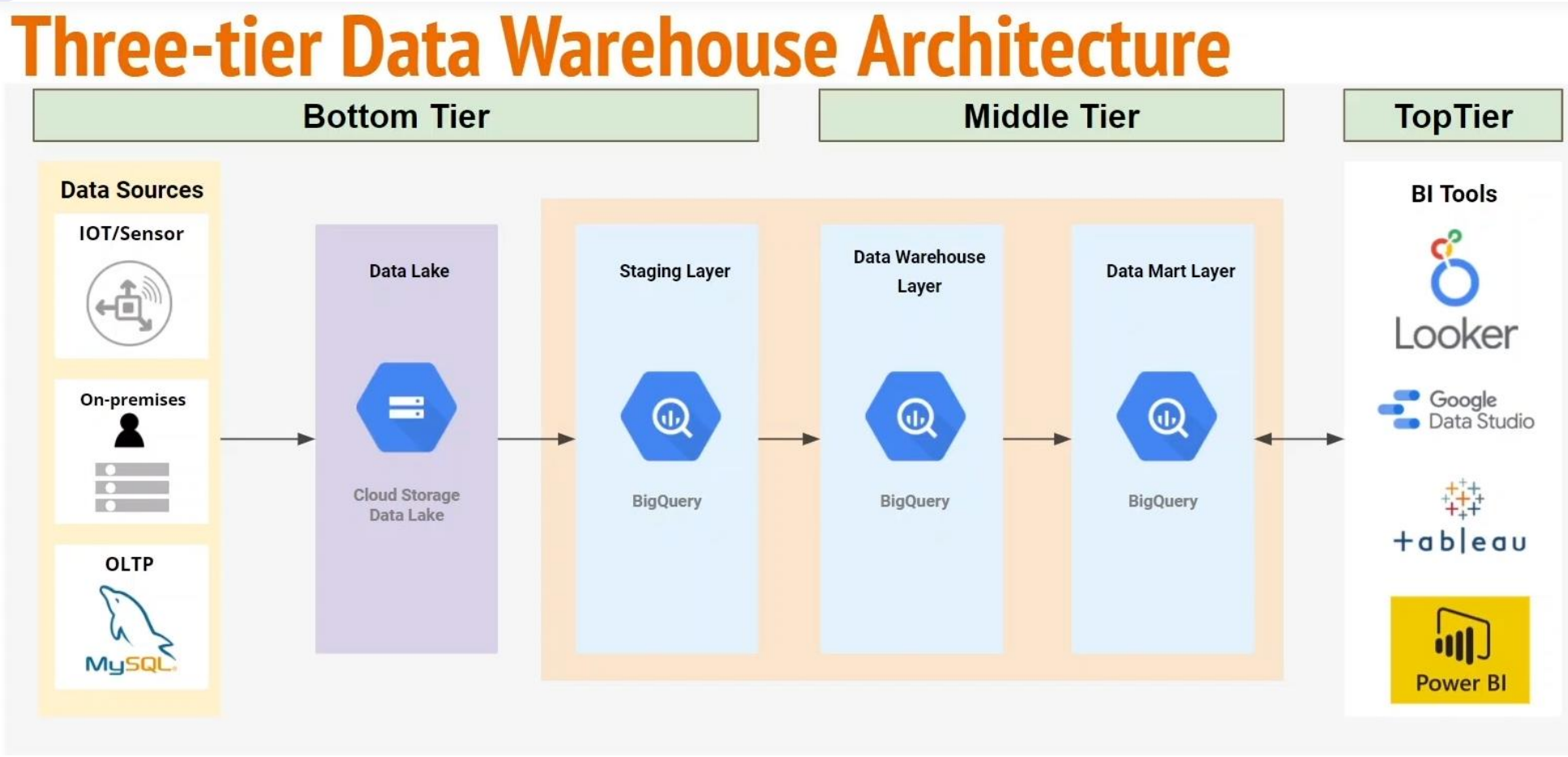
  SE GUARDA INFORMACIÓN CON CIERTA FRECUENCIA (DIARIO, SEMANAL) Y SE GUARDA LA INFORMACIÓN HISTÓRICA

- NO ES VOLATIL:
  - NUEVOS DATOS ENTRAN SIN QUE VIEJOS DATOS SE BORREN

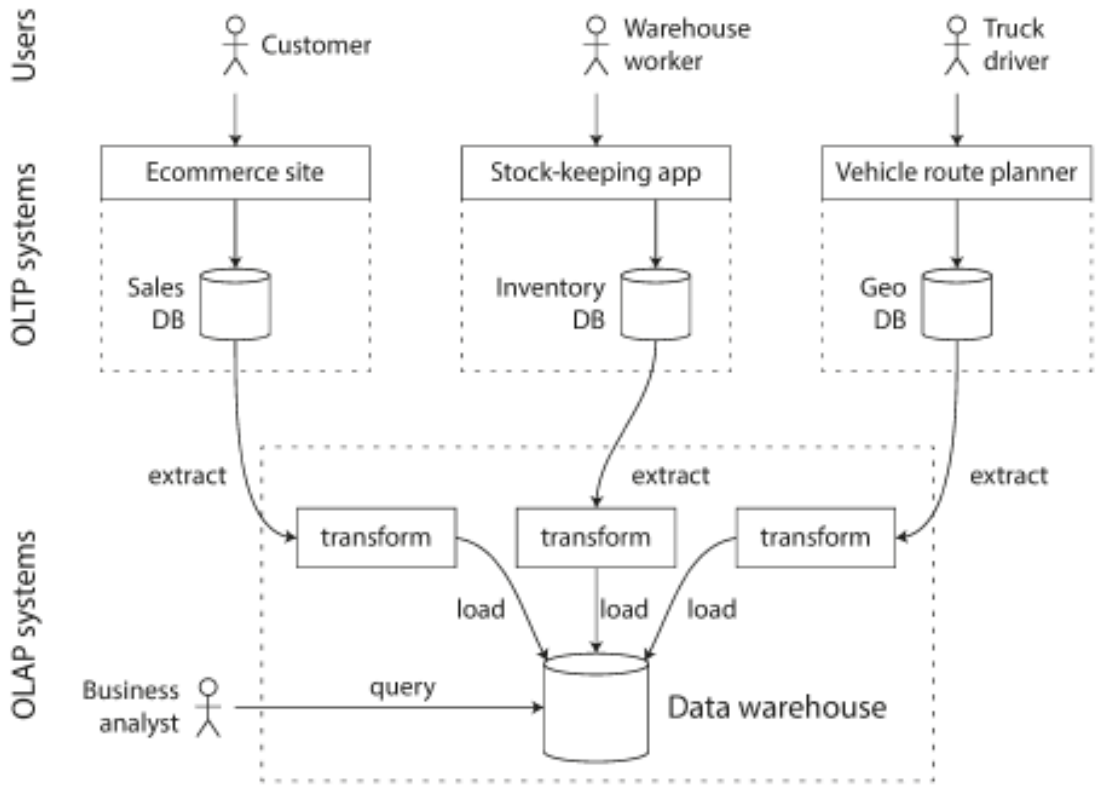
Arquitectura general de un DataWarehouse



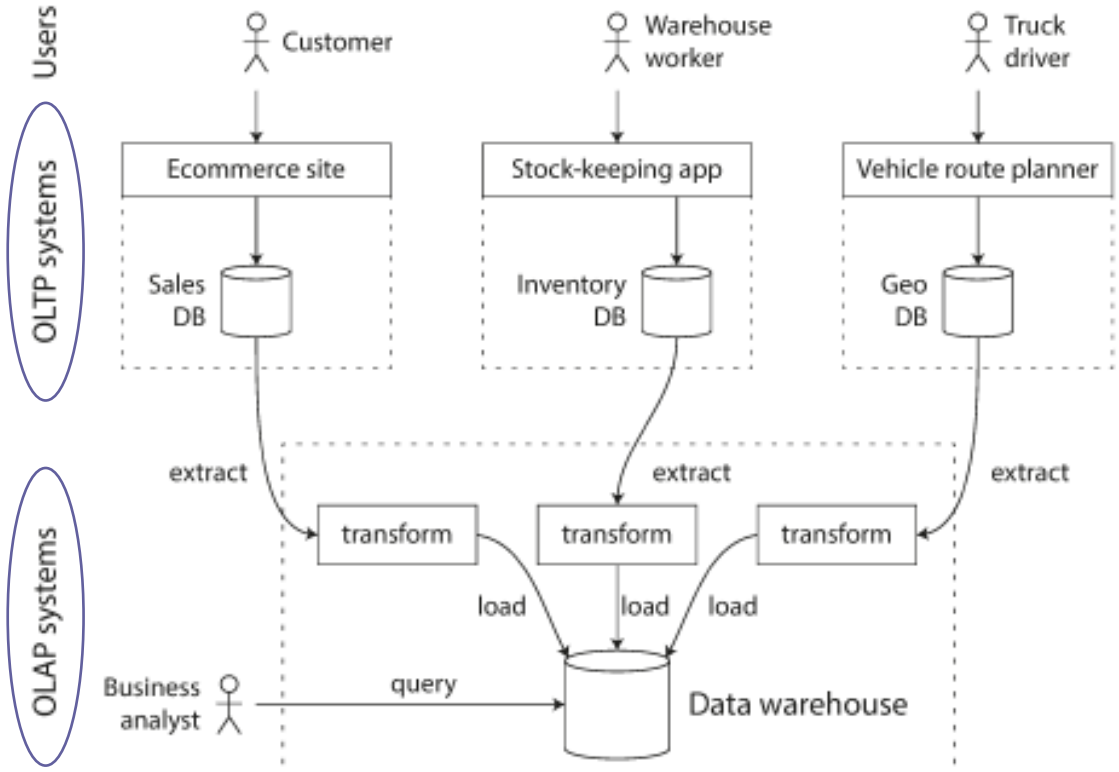
COMO SE VE UN DW?



Proceso de ETL



Proceso de ETL



Row Oriented Storage vs Column Oriented Storage

OLTP
ONLINE TRANSACTION PROCESSING

Data in logical table

X	Y	Z
x1	y1	z1
x2	y2	z2
x3	y3	z3

Row-oriented storage on disk

All columns of a given row are stored together

OLAP
ONLINE ANALYTICAL PROCESSING

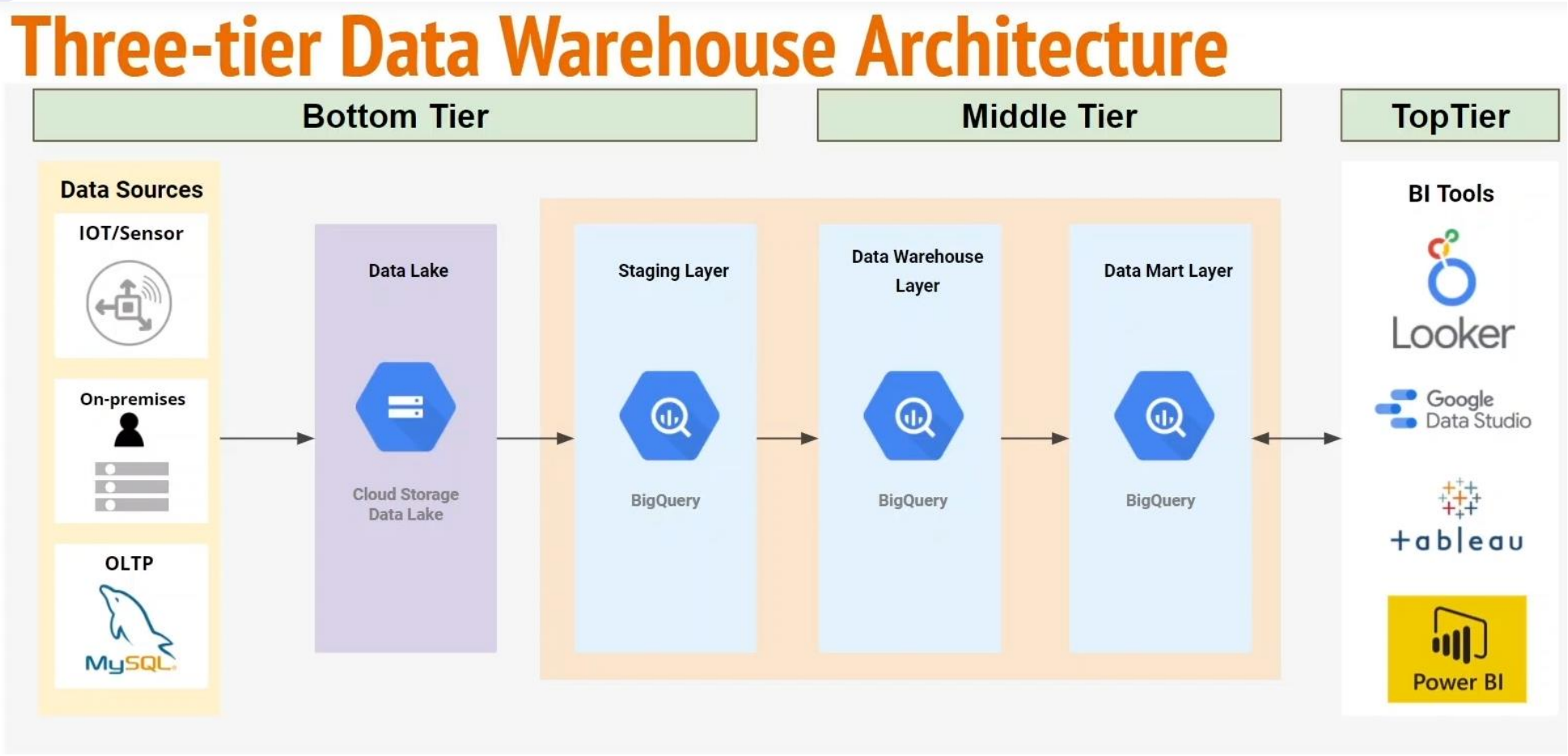
Data in logical table

X	Y	Z
x1	y1	z1
x2	y2	z2
x3	y3	z3

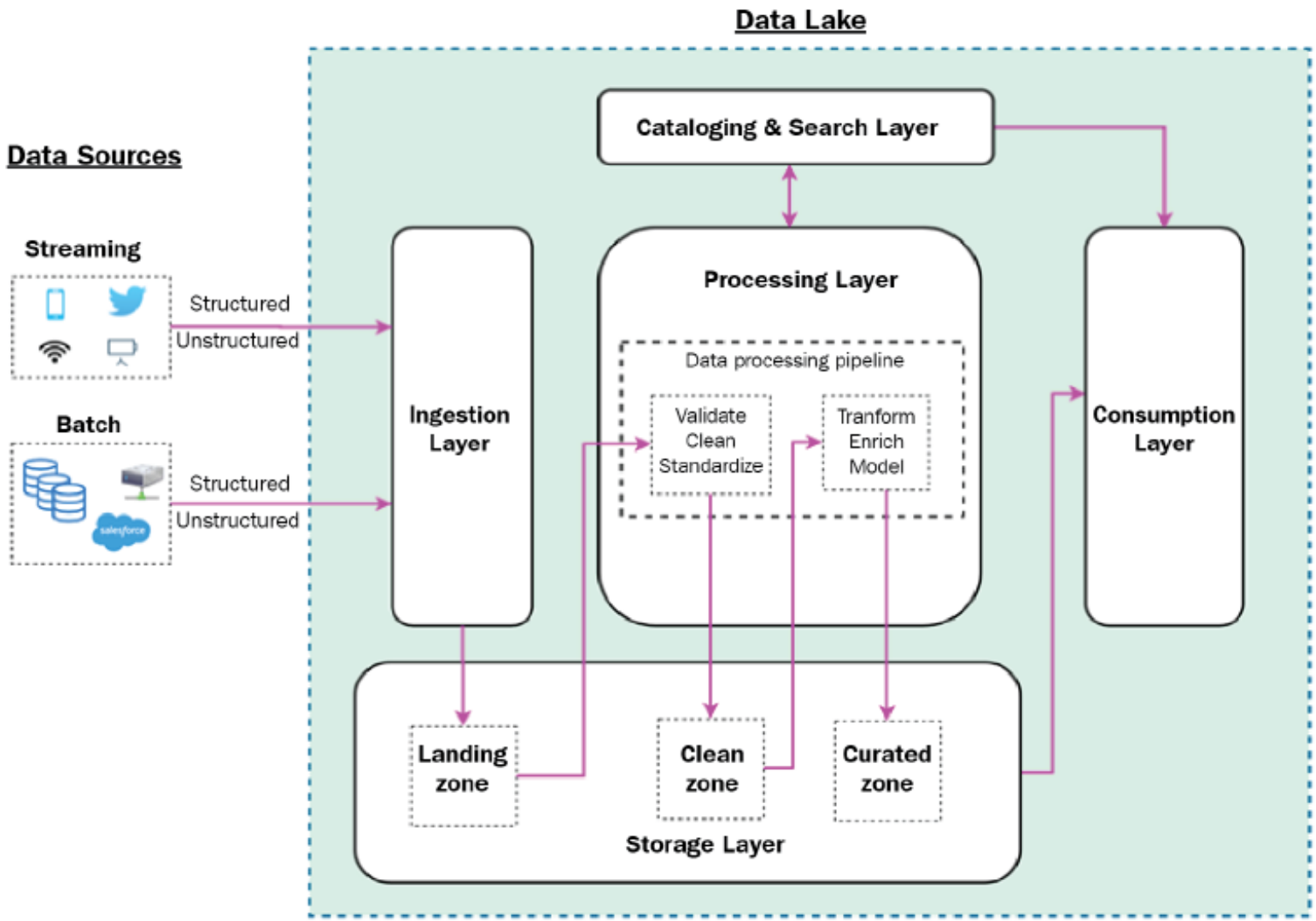
Column-oriented storage on disk

For a set of rows (a.k.a. "chunk"), all values of per column are stored together

COMO SE VE UN DW?

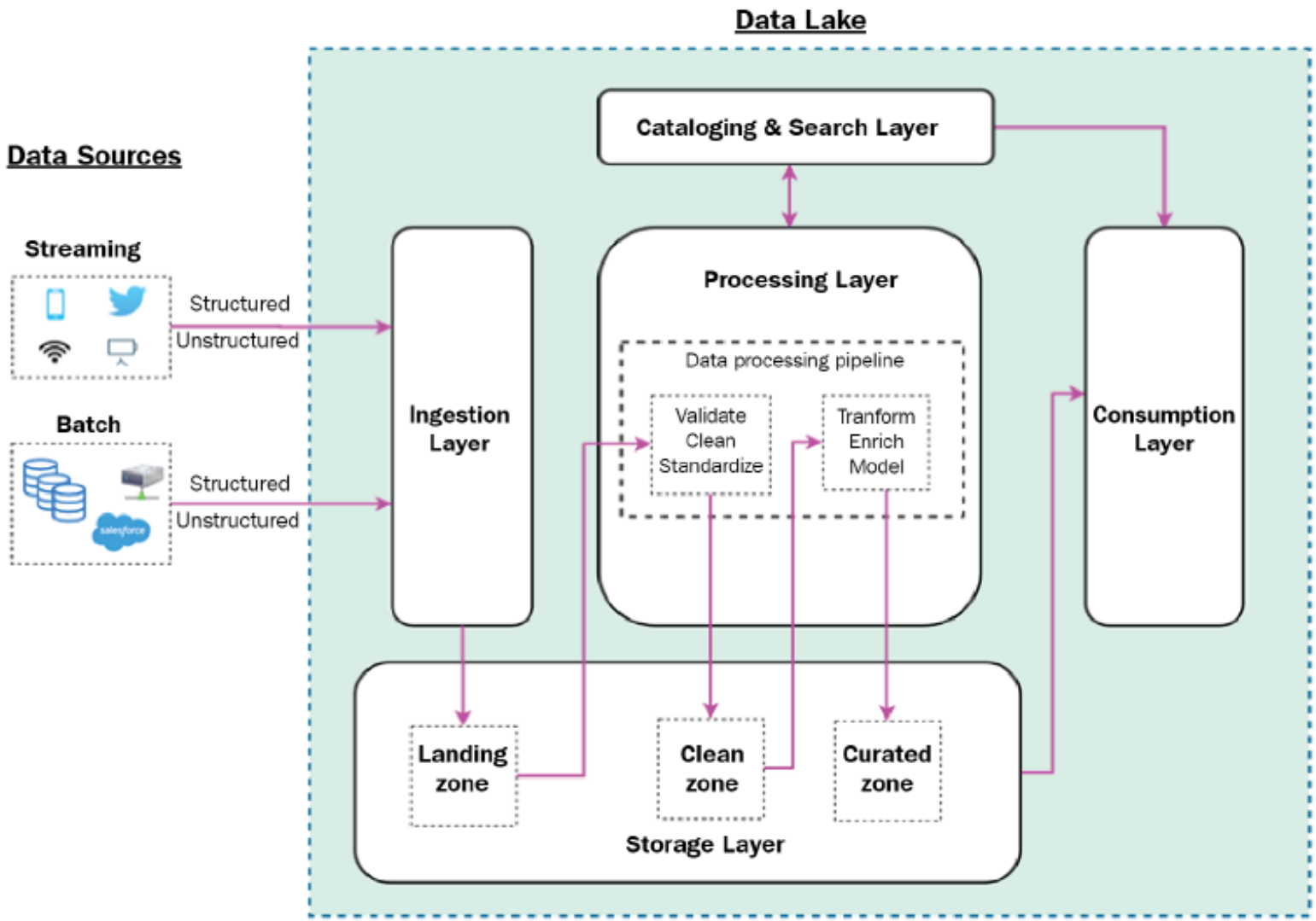


Data Lake Architecture



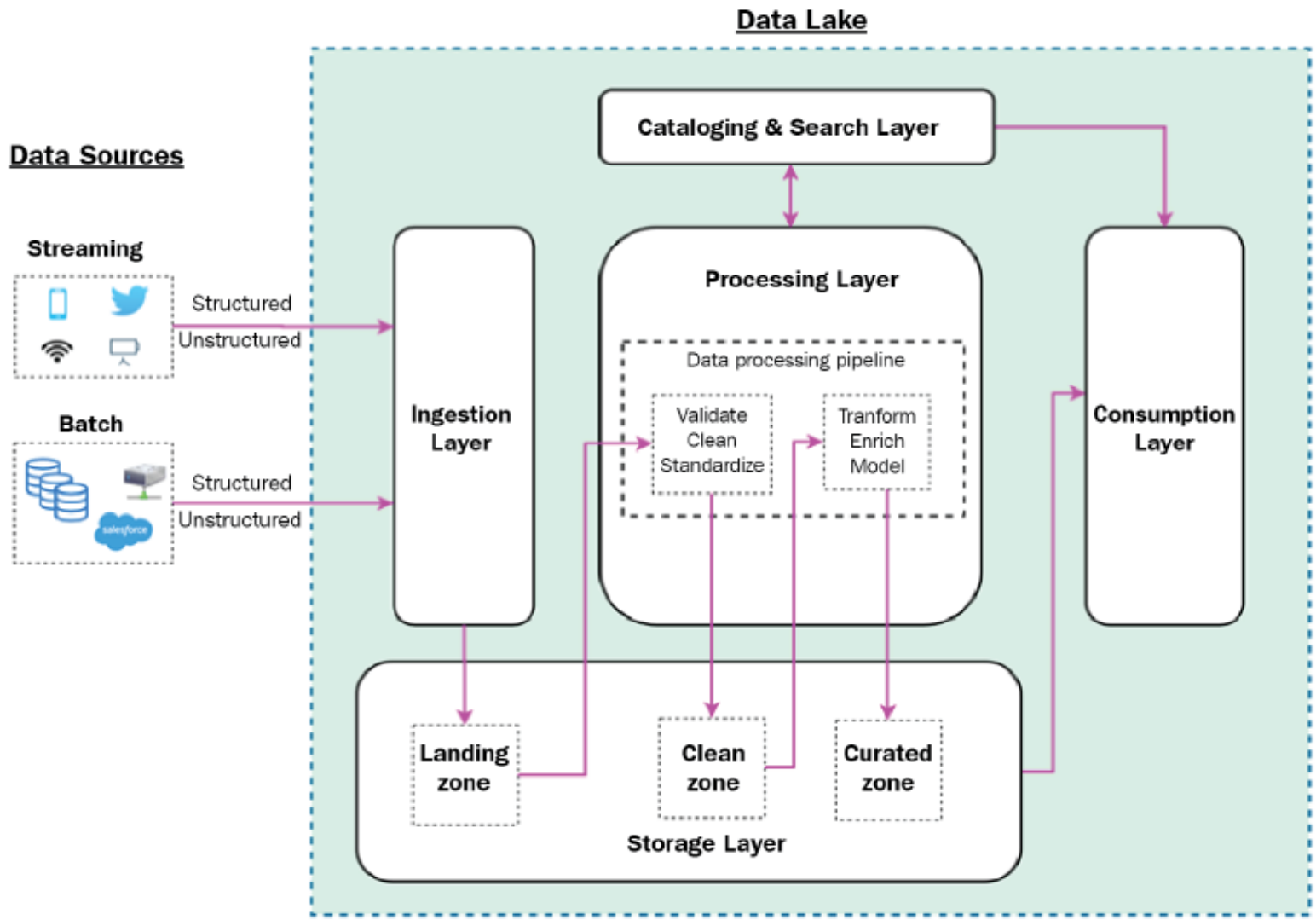
Data Lake Architecture

- LANDING ZONE:
  - Ingesta de datos
  - Se guarda “como vino”
  - Política de mantenimiento subjetiva



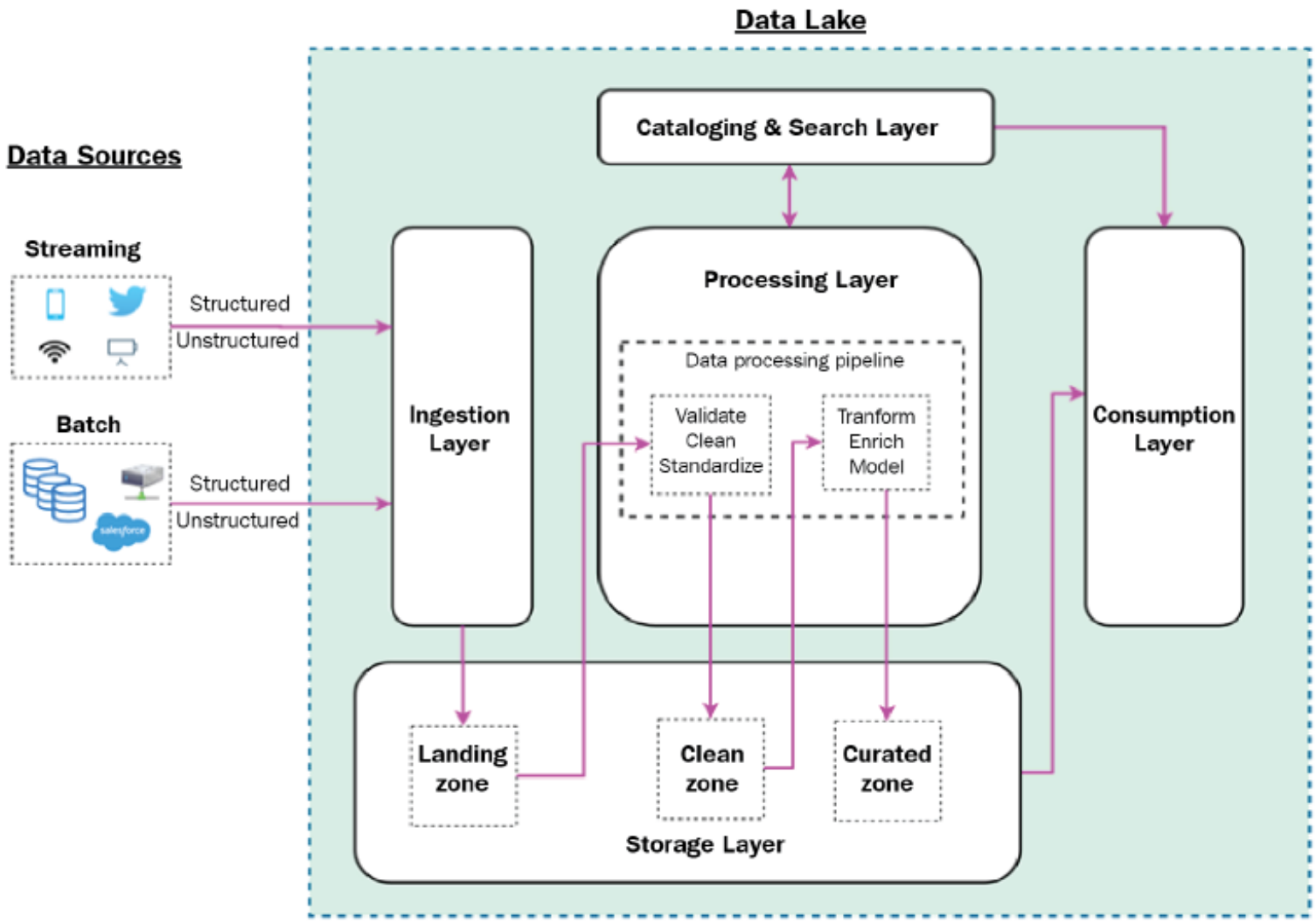
Data Lake Architecture

- CLEAN ZONE:
  - Se procesa de la landing zone
  - Se valida
  - Se limpia
  - Se optimiza el guardado (parquet)
  - Se enmascaran datos PII

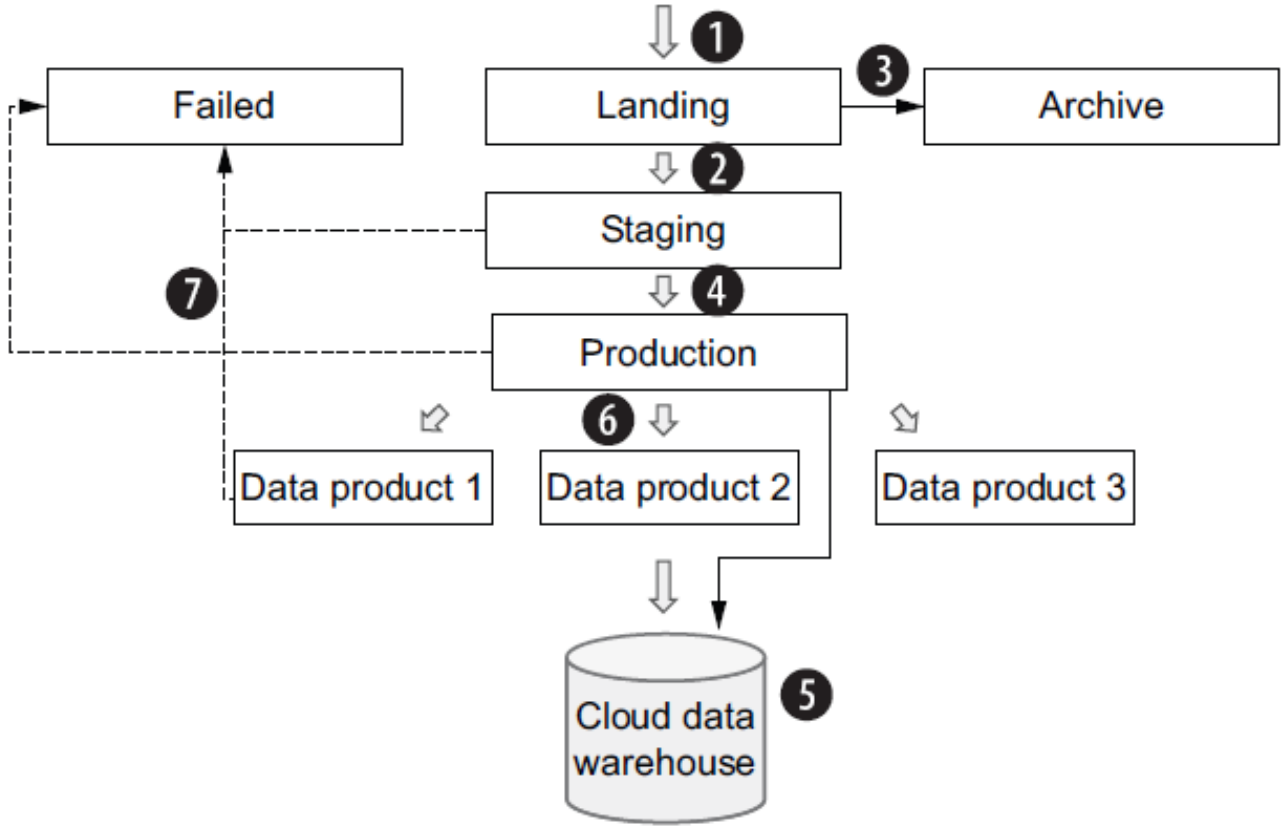


Data Lake Architecture

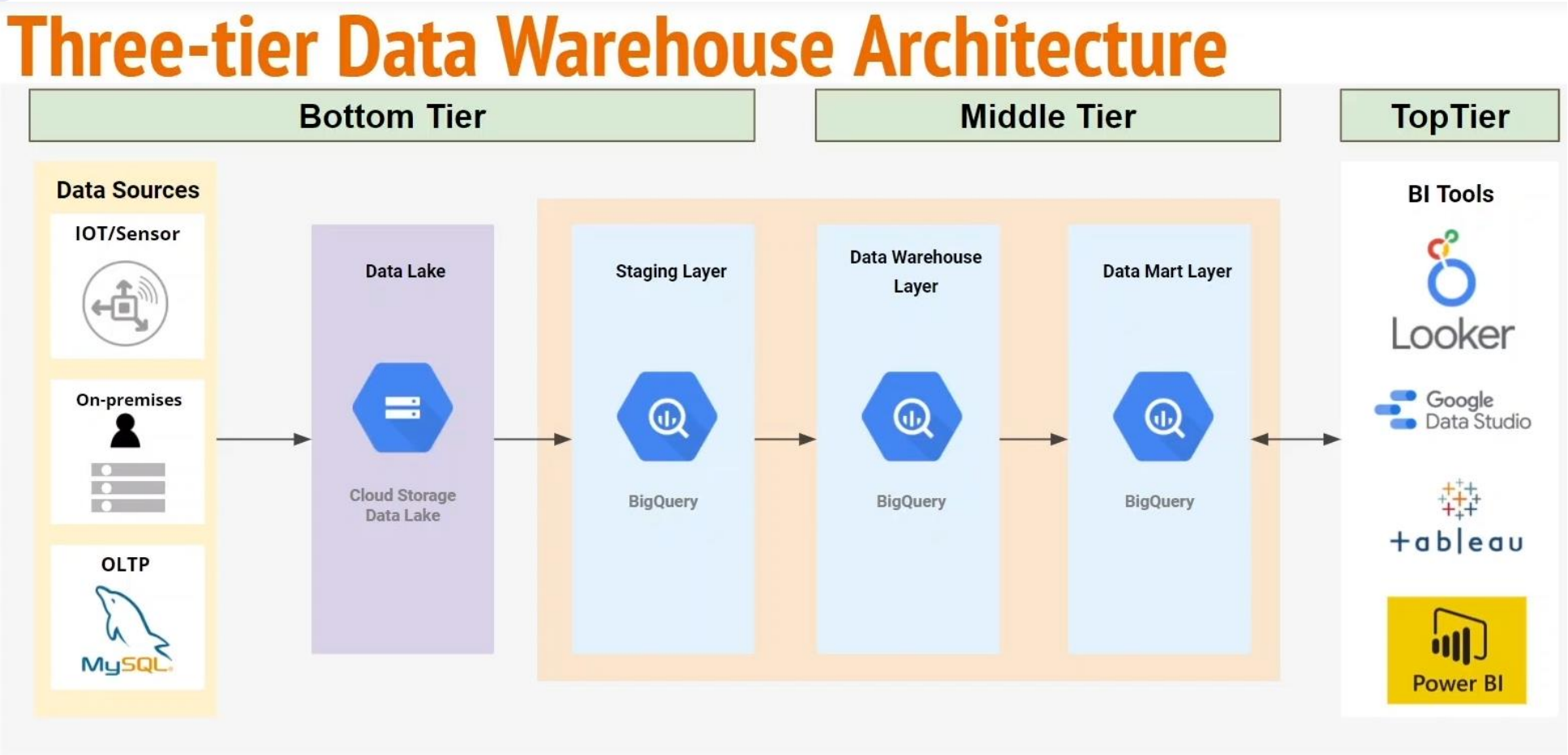
- CURATED ZONE:
  - Se refina con información externa/extra
  - Se agrega validación y lógica de negocio
  - Se cataloga para hacer gobierno



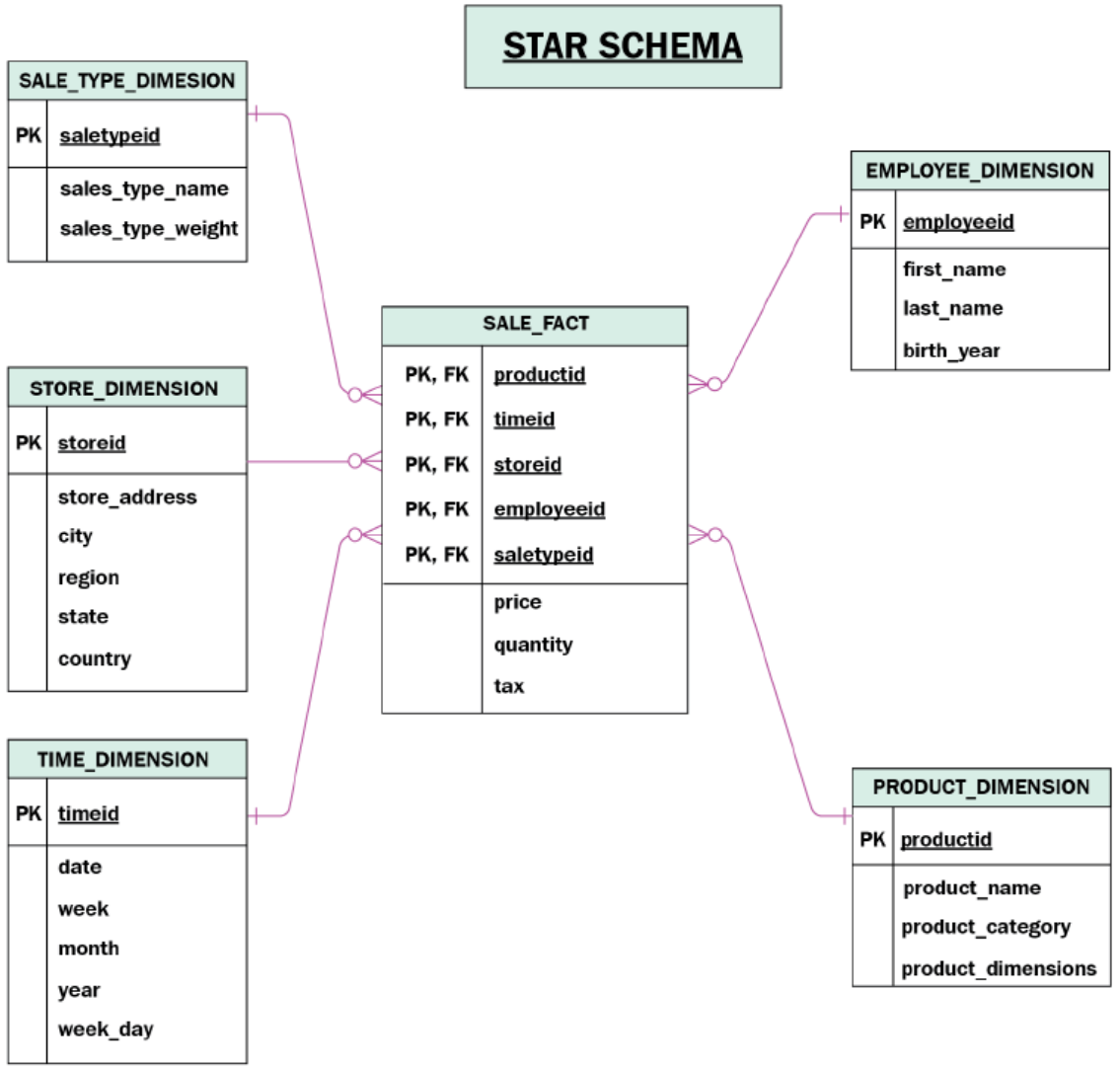
Pasos de ELT al DataLake



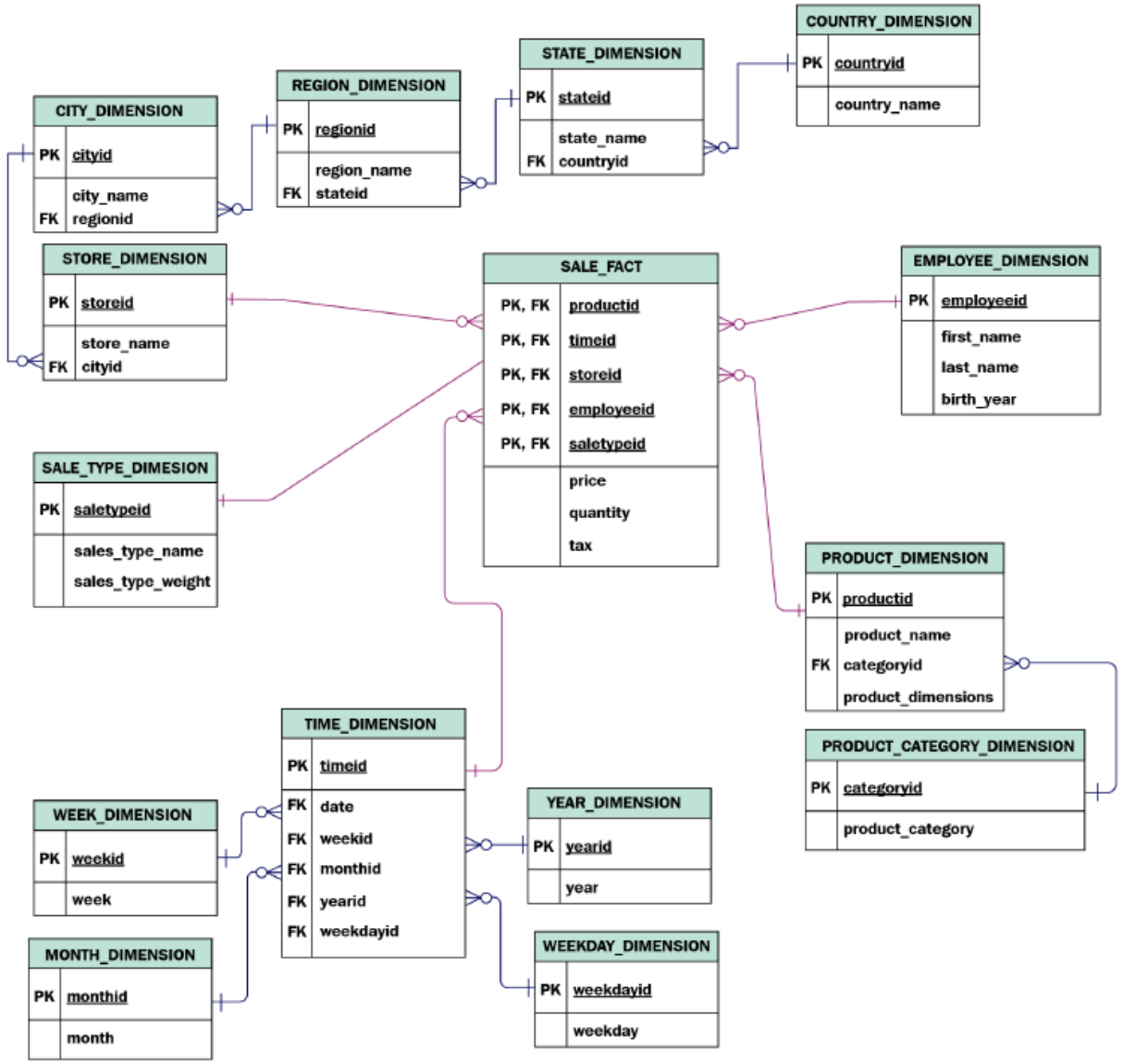
COMO SE VE UN DW?



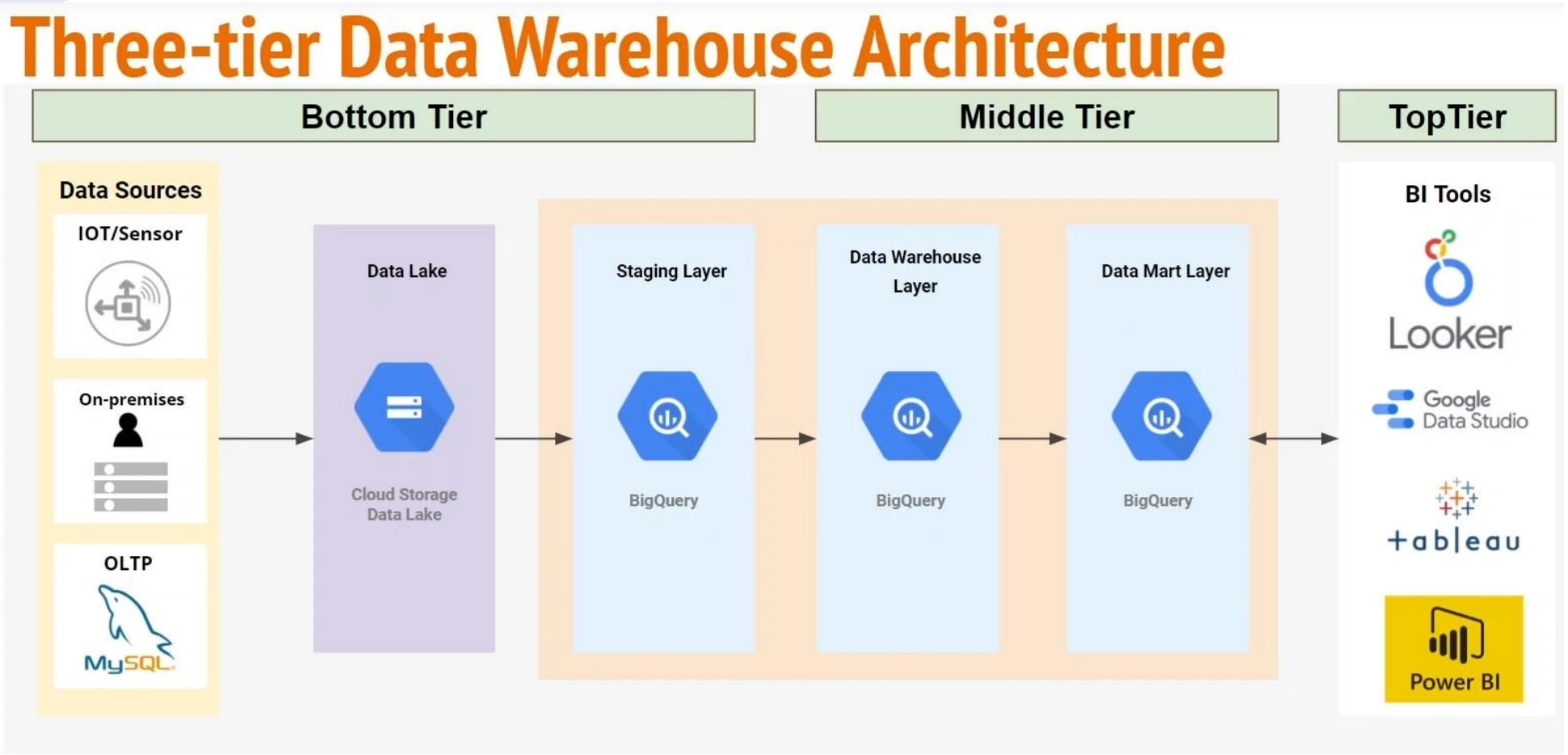
Dimensional Modelling
Star Schema



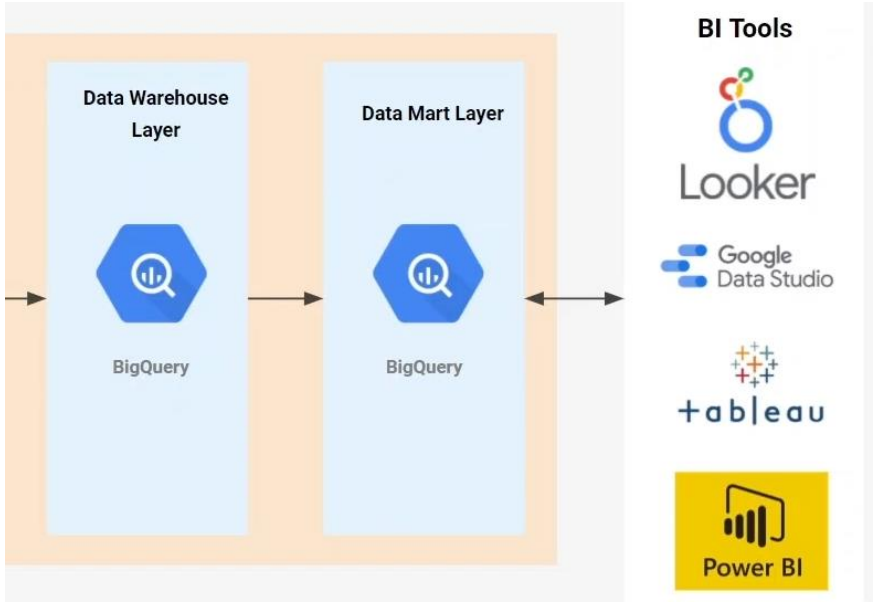
Dimensional Modelling - Snowflake Schema

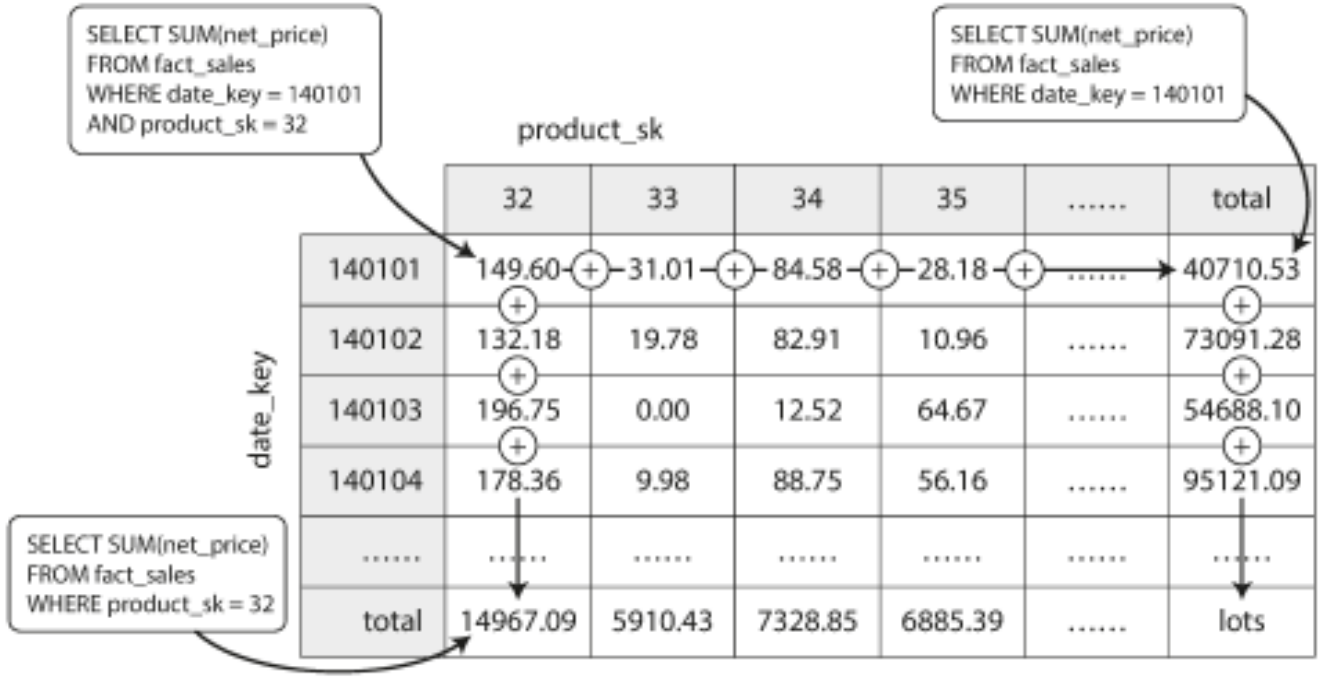


DATA MART

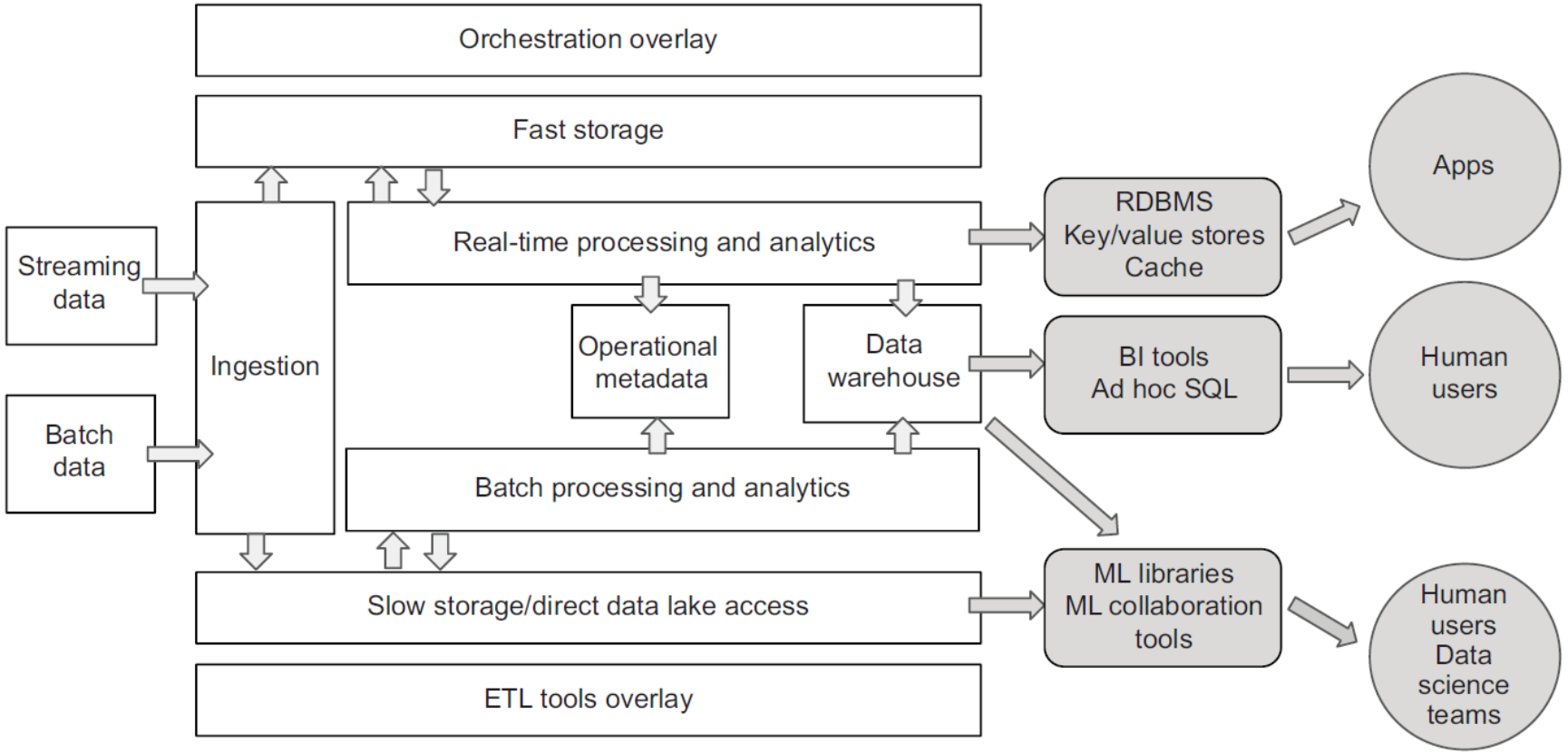

Three-tier Data Warehouse Architecture
Bottom Tier
Middle Tier
TopTier
Data Sources
IOT/Sensor
On-premises
OLTP
MySQL
Data Lake
Cloud Storage
Data Lake
Staging Layer
BigQuery
Data Warehouse
Layer
BigQuery
Data Mart Layer
BigQuery
BI Tools
Looker
Google
Data Studio
+ableau
Power BI

Data Mart Cubes

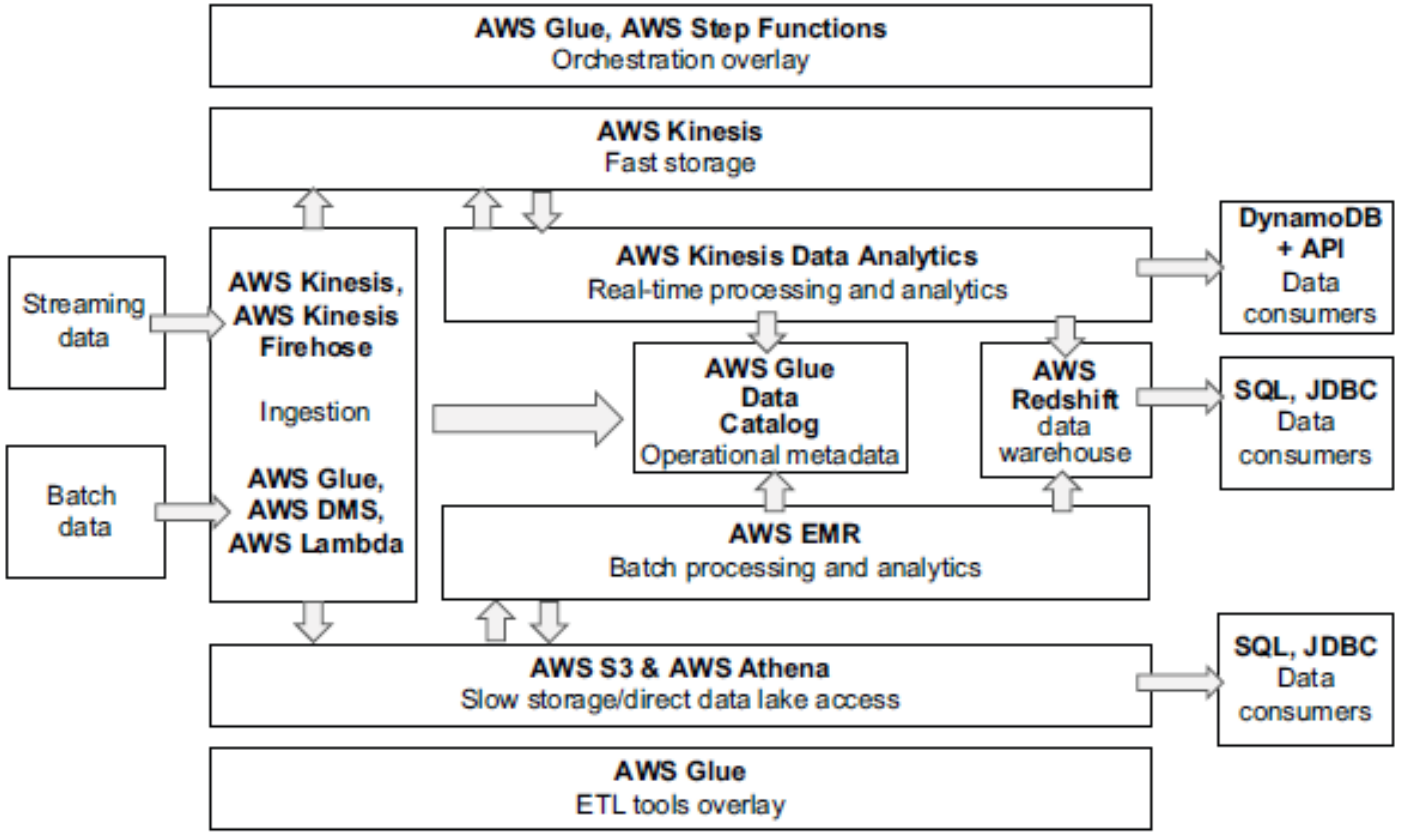




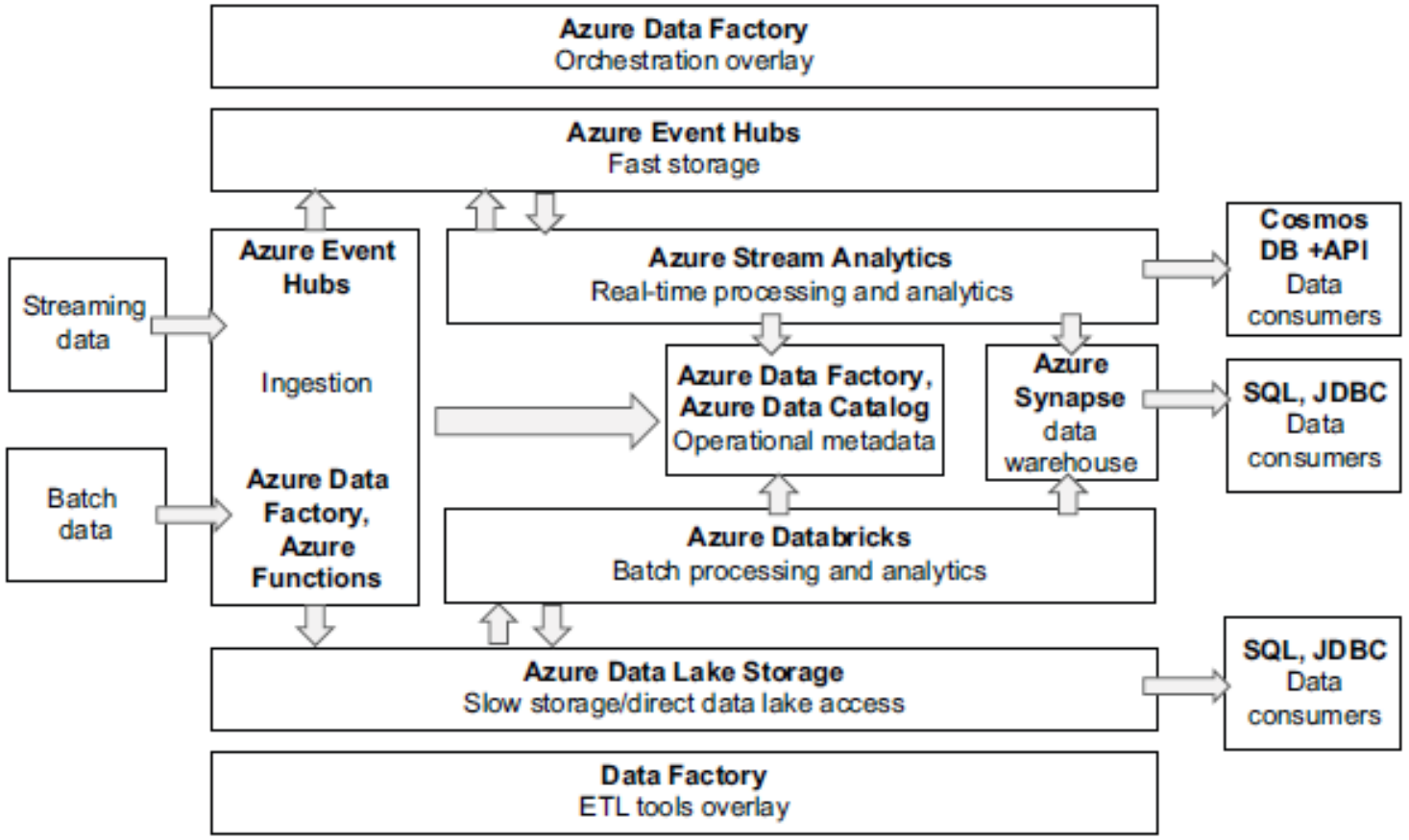
Arquitectura general de un DataWarehouse



Arquitectura general de un DataWarehouse - AWS



Arquitectura general de un DataWarehouse - AZURE



Arquitectura general de un DataWarehouse - GCP