Comparing RDBMS, Data Warehouse, Data Lake, and Lakehouse

Data Platforms Comparison: RDBMS vs. Data Warehouse vs. Data Lake vs. Lakehouse

Abdel Dadouche
DJZ Consulting

adadouche@hotmail.com
@adadouche

Traditional RDBMS

- **Definition**:
  - Relational Database Management Systems store structured data in predefined schemas (tables, rows, columns).
  - Using a schema-on-write approach
- **Characteristics**:
  - ACID transactions
  - Optimized for OLTP (Online Transaction Processing)
  - Strong consistency and data integrity
- **Schema**:
  - Example: Customer table → CustomerID, Name, Email
- **Use Cases**: Banking systems, ERP, CRM

Data Warehouse

- **Definition**:
  - A centralized system for storing structured, historical data for analytics.
- **Characteristics**:
  - Optimized for OLAP (complex queries, reporting)
  - Schema-on-write (structured before loading)
  - Uses **star** or **snowflake schemas**
- **Schema Example**:
  - **Sales Fact Table** → Date, Customer, Product, Revenue (with dimension tables)
- **Use Cases**: Business Intelligence (BI), dashboards, trend analysis

Data Lake

- **Definition**:
  - Scalable storage system for structured, semi-structured, and unstructured data.
- **Characteristics**:
  - Schema-on-read (schema applied when queried)
  - Stores raw data in multiple formats (CSV, JSON, Parquet, images, videos)
  - High scalability and low cost
- **Schema Example**:
  - Raw files (csv, txt, json, parquet…Lakehouse) in object storage (e.g., S3 buckets, HDFS)
- **Use Cases**: Machine learning, advanced analytics, IoT data, raw log storage

Lakehouse

- **Definition**:
  - Hybrid architecture that unifies Data Lake and Data Warehouse capabilities.
- **Characteristics**:
  - ACID transactions with scalable data storage
  - Handles structured + unstructured data
  - Supports BI, ML, and AI workloads
  - Uses open file formats (Parquet, Delta Lake, Iceberg)
- **Schema Example** (Medallion Architecture):
  - **Bronze Layer**: Raw data
  - **Silver Layer**: Cleaned & conformed
  - **Gold Layer**: Curated, business-ready data
- **Use Cases**: Self-service BI, ML pipelines, unified enterprise analytics

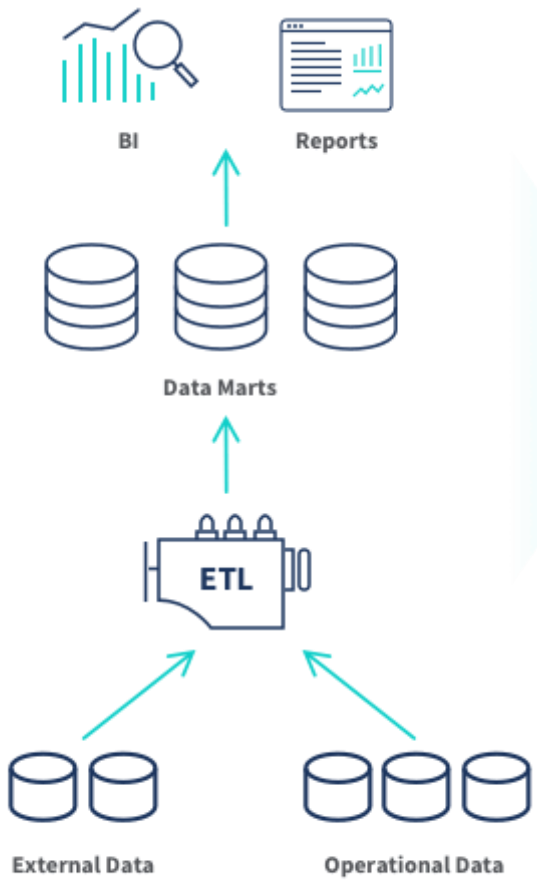
Comparison Table

Feature	RDBMS	Data Warehouse	Data Lake	Lakehouse
Data Types	Structured	Structured	All (structured + raw)	All (structured + raw)
Data Format	**Closed** proprietary format	**Closed** proprietary format	**Open** format	**Open** format
Schema	Schema-on-write	Schema-on-write	Schema-on-read	Both
Workload Optimized For	OLTP	OLAP	Big data & ML	BI + ML/AI
ACID Transactions	Yes	Limited	No	Yes
Scalability & Cost	Moderate	Expensive	Highly scalable, low-cost	Scalable, cost-effective

Source : https://www.qlik.com/us/data-lake/data-lakehouse

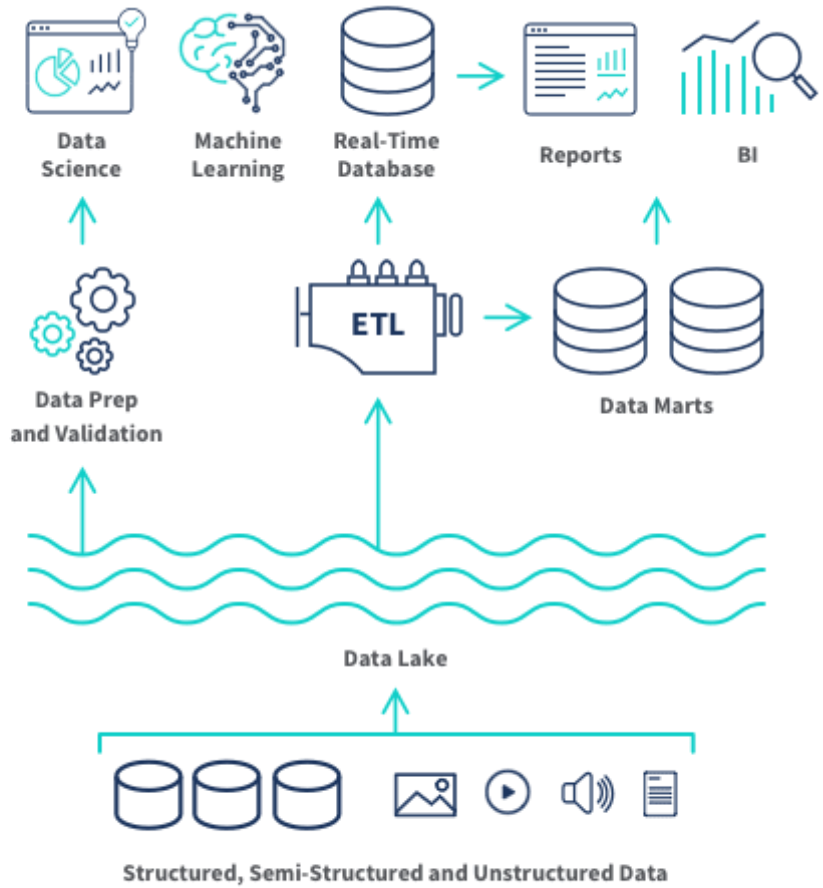
LATE 1980'S

Data Warehouse



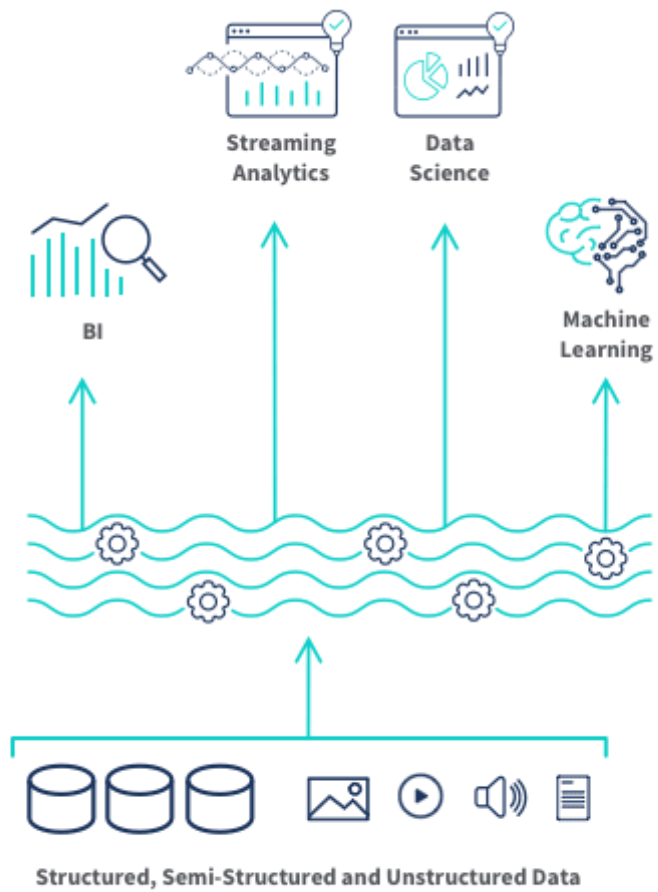
2011

Data Lake

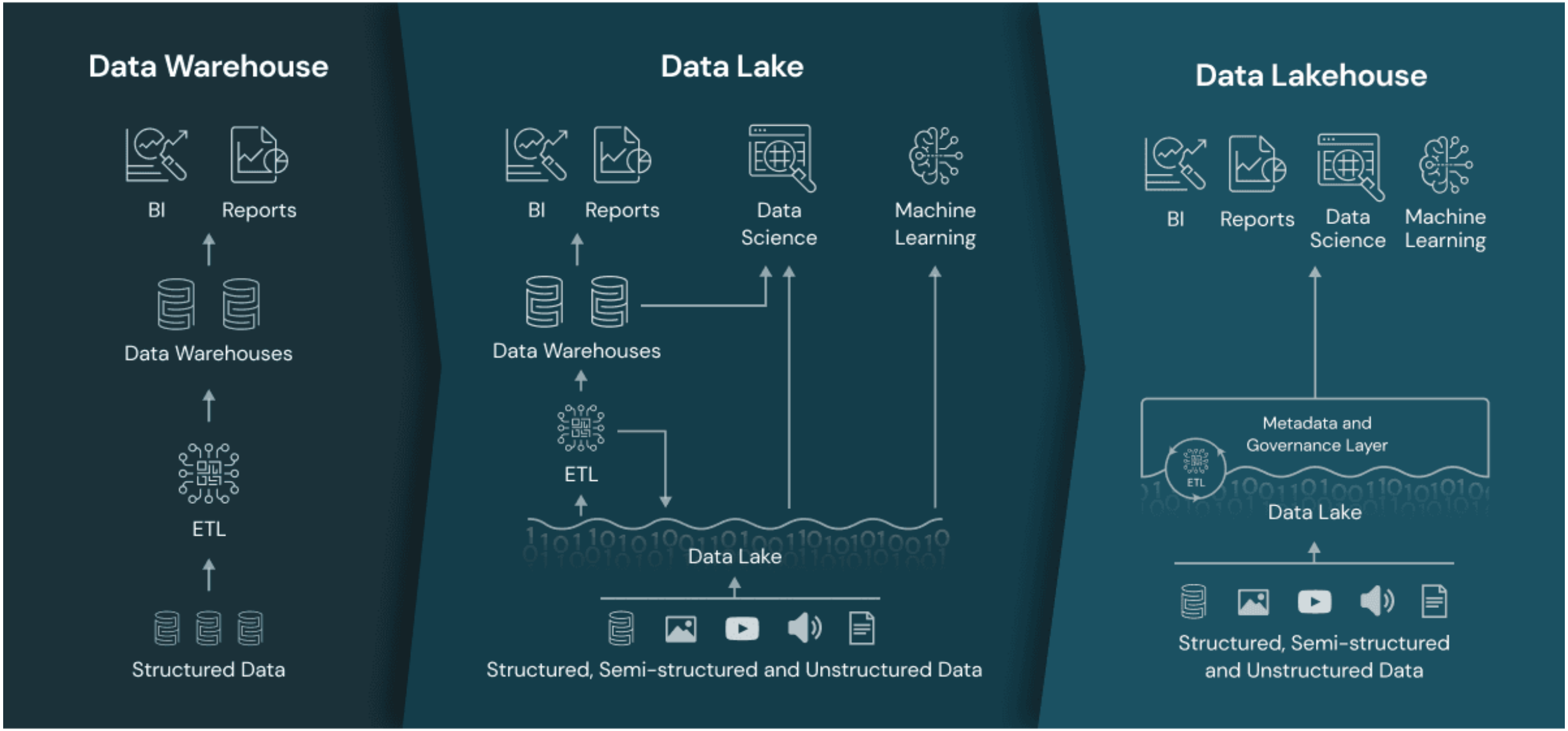


2020

Lakehouse



Source : https://www.qlik.com/us/data-lake/data-lakehouse



Source : https://www.databricks.com/glossary/data-lakehouse

Key Takeaway

- **RDBMS** → Operational systems
- **Data Warehouse** → Historical structured analytics
- **Data Lake** → Scalable raw data storage
- **Lakehouse** → Unified platform for BI + ML + AI

References

- Inmon, W.H. (2005) Building the Data Warehouse.
- Kimball, R. (2013) The Data Warehouse Toolkit.
- Stonebraker, M. (2024). What Goes Around Comes Around... And Around...?
- Azure Data Lake
- Databricks (2020). The Lakehouse Architecture.
- Google Cloud BigQuery