ORACLE Certified Professional

Oracle Certified Professional Java Programmer

Microsoft Certified Professional

sohailimran@yahoo.com

DATA LAKE

Sohail IMRAN سهيل عمران

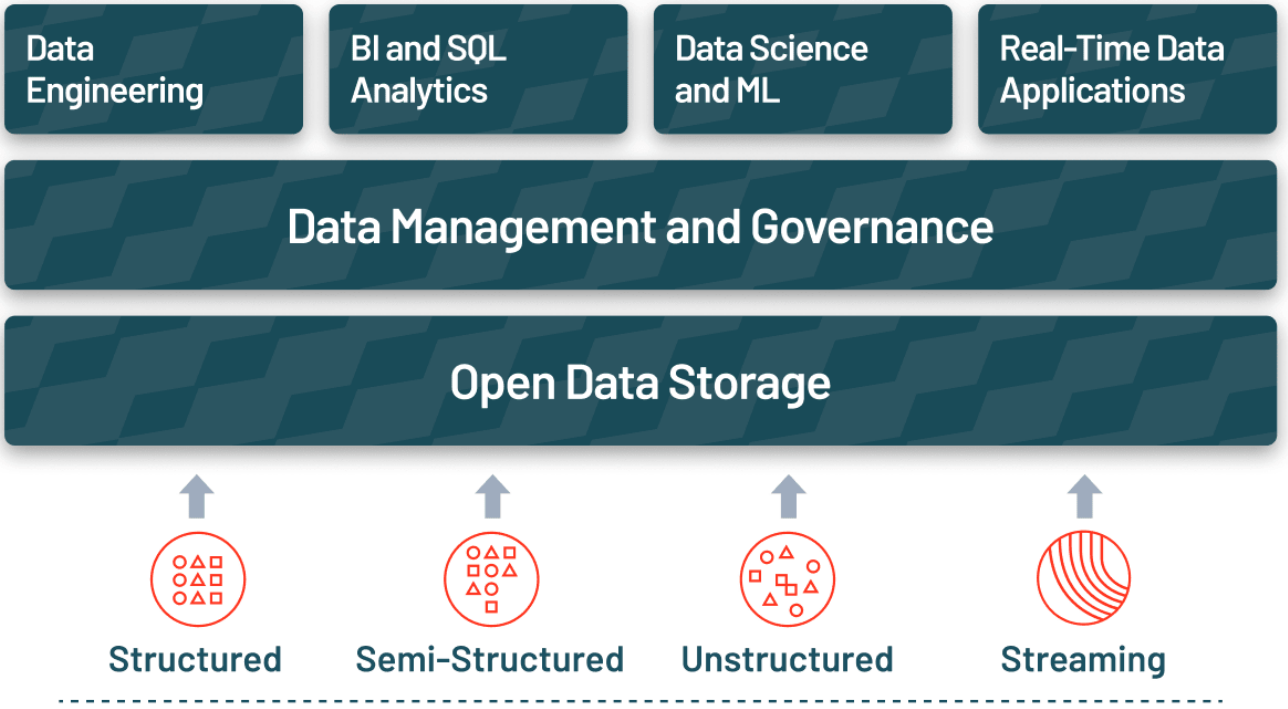
Introduction

Ref: Web

what?

A data lake is **a storage repository that holds a vast amount of raw data in its native format** until it is needed for analytics applications.

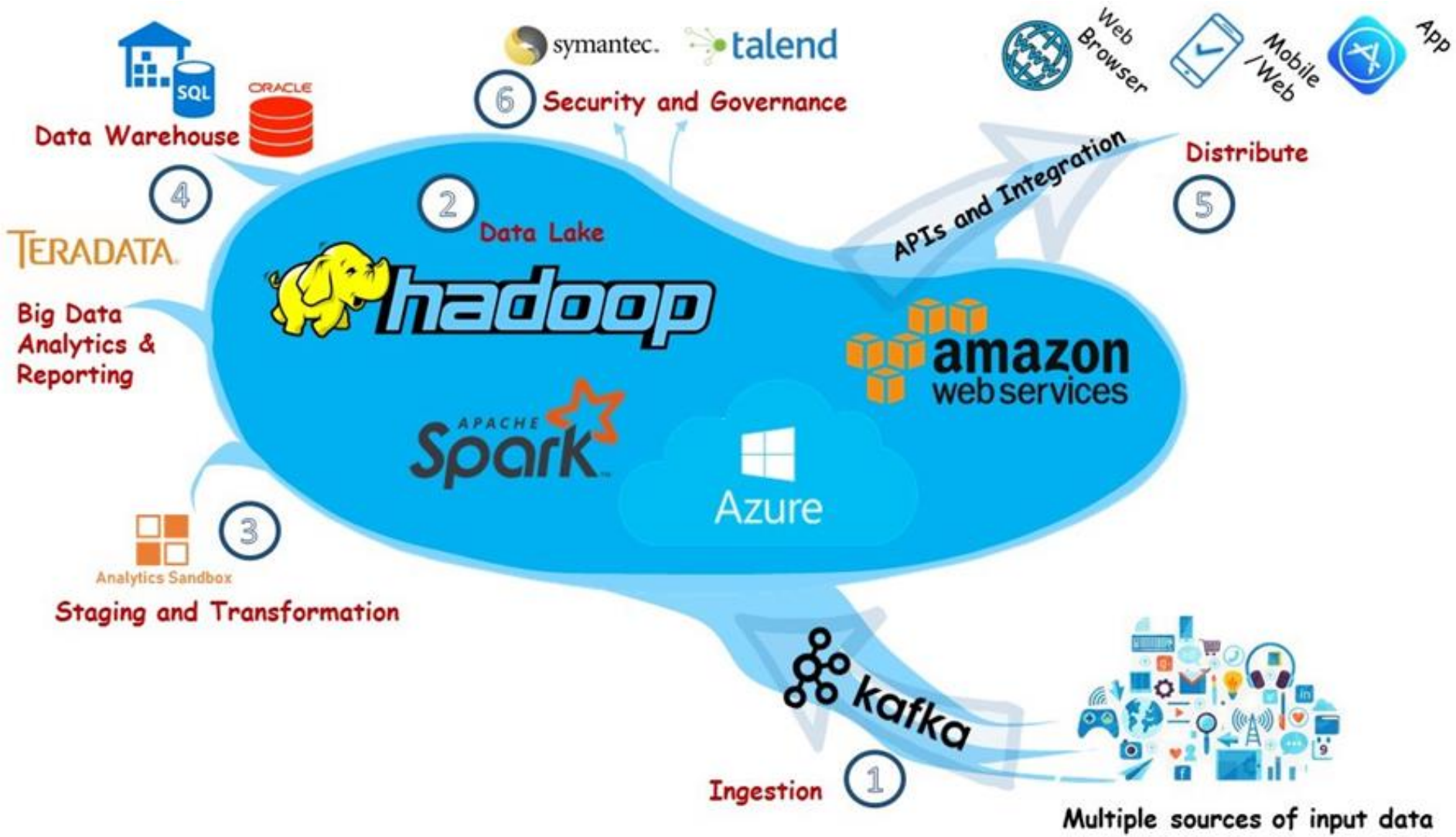
Data Lake is like a large container which is very similar to real lake and rivers. Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time.



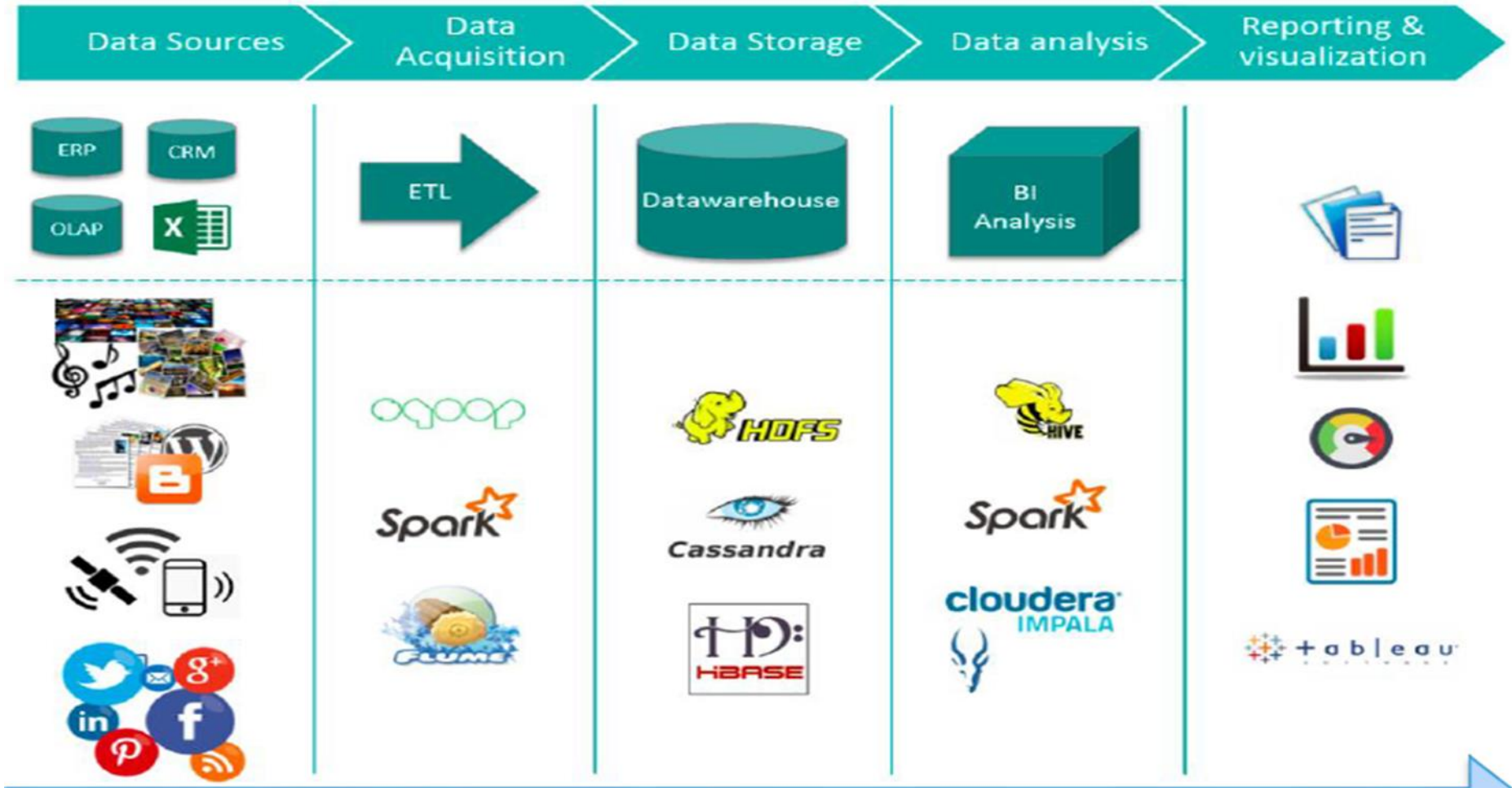
why?

- With the onset of storage engines like Hadoop storing disparate information has become easy. There is no need to model data into an enterprise-wide schema with a Data Lake.
- With the increase in data volume, data quality, and metadata, the quality of analyses also increases.
- Data Lake offers business Agility
- Machine Learning and Artificial Intelligence can be used to make profitable predictions.
- It offers a competitive advantage to the implementing organization.
- There is no data silo structure. Data Lake gives 360 degrees view of customers and makes analysis more robust.

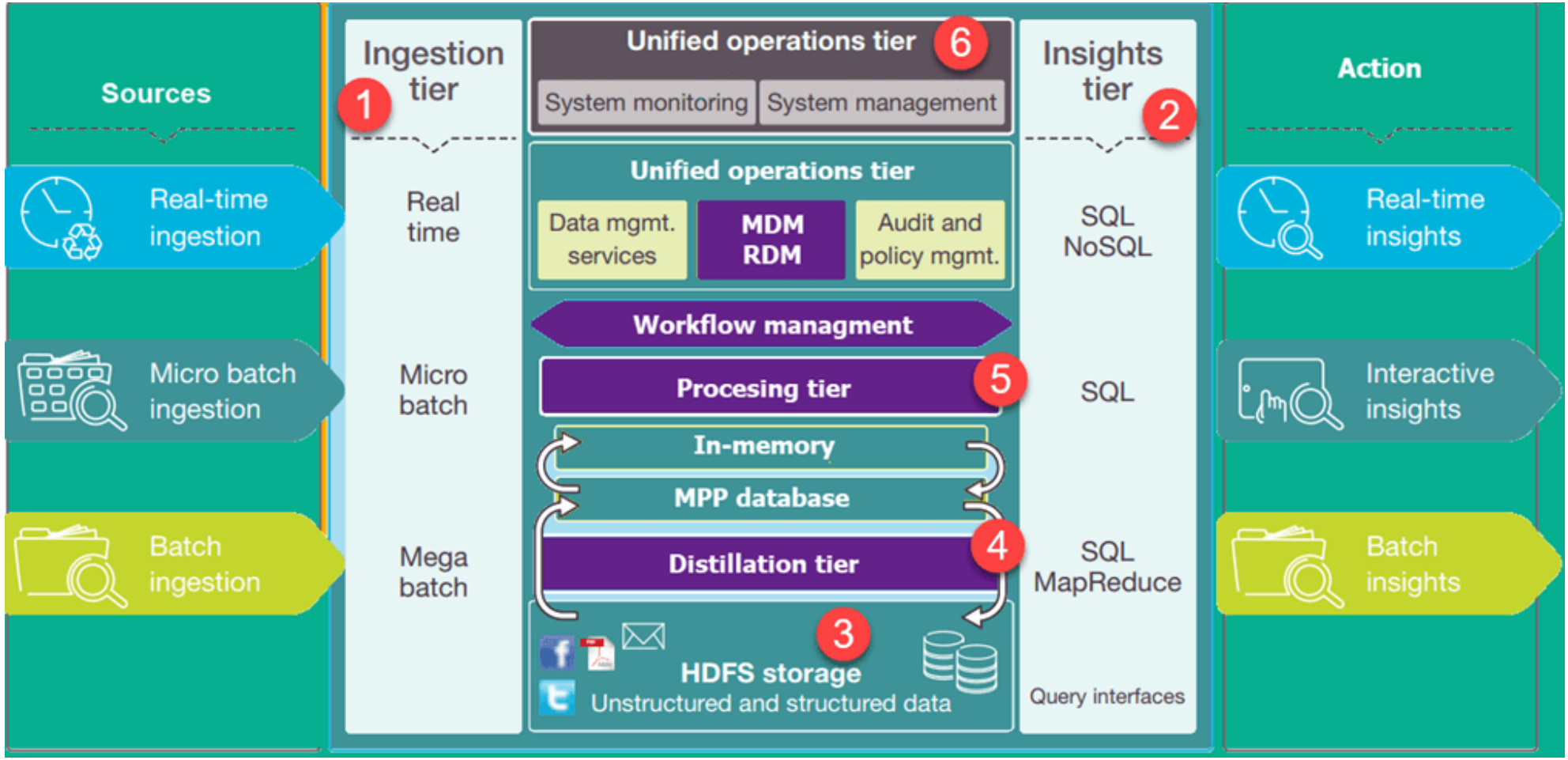
big data lake



how?


Data Sources
Data Acquisition
Data Storage
Data analysis
Reporting & visualization
ERP
CRM
OLAP
ETL
Datawarehouse
BI Analysis
Spark
HDFS
Cassandra
HBASE
HIVE
Spark
cloudera IMPALA
tableau
Traditional Enterprise Data warehouse

architecture



tiers

1. **Ingestion Tier:** Ingestion tiers depict the data sources. Data Ingestion allows connectors to get data from a different data sources and load into the Data lake. The data could be loaded into the data lake in batches or in real-time.

2. **Insights Tier:** The tiers on the right represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.

3. **Hadoop Distributed File System (HDFS):** is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.

4. **Distillation tier** takes data from the storage tire and converts it to structured data for easier analysis.

5. **Processing tier** run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.

6. **Unified operations tier** governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

metadata

Data lakes, manage data with variety for on-demand, ad-hoc analyses, are typically subdivided into three components:

1. a storage system: very often HDFS, though NoSQL DBMSs are also an option;
2. a **metadata** system, which we focus on below;
3. an access and analysis system generally relying on MapReduce or Spark.

Data lakes mostly bear a “flat” architecture, where each data element has a unique identifier and a set of characterizing tags.
Such tags are actually essential metadata to help comprehend data and access data, not to mention query effectiveness.
The advantage of data tagging is that new data and new sources can be inserted on the fly.
Once data are tagged, they just need to be connected to already stored data.
This feature allows users to formulate queries directly, without needing the help of a business intelligence expert.

Characteristics

1. **Data Fidelity:** A data lake stores data **as it is** in a business system. A data lake stores raw data, whose format, schema, and content cannot be modified.
2. **Data Flexibility:** A data lake adopts schema-on-read, meaning it sees business uncertainty as the norm and can adapt to unpredictable business changes. You can design a data schema in any phase as needed, so the entire infrastructure generates data that meets your business needs.
3. **Data Manageability:** A data lake provides comprehensive data management capabilities. Due to its fidelity and flexibility, a data lake stores at least two types of data: raw data and processed data. The stored data constantly accumulates and evolves. This requires robust data management capabilities, which cover data sources, data connections, data formats, and data schemas. This requires permission management capabilities.
4. **Data Traceability:** A robust data lake fully reproduces the data production process and data flow, ensuring that each data record is traceable through the processes of access, storage, processing, and consumption.
5. **Data Rich Computing Engines:** A data lake supports a diversity of computing engines, including batch processing, stream computing, interactive analytics, and machine learning engines.
6. **Multi-Modal Storage Engine:** In theory, a data lake should provide a built-in multi-modal storage engine to enable data access by different applications, while considering a series of factors, such as the response time (RT), concurrency, access frequency, and costs.

data lake Vs data warehouse

Parameters	Data Lakes	Data Warehouse
Data	Data lakes store everything.	Data Warehouse focuses only on Business Processes.
Processing	Data are mainly unprocessed	Highly processed data.
Type of Data	It can be Unstructured, semi-structured and structured.	It is mostly in tabular form & structure.
Task	Share data stewardship	Optimized for data retrieval
Agility	Highly agile, configure and reconfigure as needed.	Compare to Data lake it is less agile and has fixed configuration.
Users	Data Lake is mostly used by Data Scientist	Business professionals widely use data Warehouse
Storage	Data lakes design for low-cost storage.	Expensive storage that give fast response times are used
Security	Offers lesser control.	Allows better control of the data.
Replacement of EDW	Data lake can be source for EDW	Complementary to EDW (not replacement)
Schema	Schema on reading (no predefined schemas)	Schema on write (predefined schemas)
Data Processing	Helps for fast ingestion of new data.	Time-consuming to introduce new content.
Data Granularity	Data at a low level of detail or granularity.	Data at the summary or aggregated level of detail.
Tools	Can use open source/tools like Hadoop/ Map Reduce	Mostly commercial tools.

Delta lake

Delta Lake is an open format data management and governance layer that combines the best of both data lakes and data warehouses.
Across industries, enterprises are leveraging Delta Lake to power collaboration by providing a reliable, single source of truth. By delivering quality, reliability, security and performance on your data lake — for both streaming and batch operations — Delta Lake eliminates data silos and makes analytics accessible across the enterprise. With Delta Lake, customers can build a cost-efficient, highly scalable lakehouse that eliminates data silos and provides self-serving analytics to end-users.

Use the data lake as a landing zone for all of your data

Save all of your data into your data lake without transforming or aggregating it to preserve it for machine learning and data lineage purposes.

Mask data containing private information before it enters your data lake

Personally identifiable information (PII) must be pseudonymized in order to comply with GDPR and to ensure that it can be saved indefinitely

Secure your data lake with role- and view-based access controls

Adding view-based ACLs (access control levels) enables more precise tuning and control over the security of your data lake than role-based controls alone.

Build reliability and performance into your data lake by using Delta Lake

The nature of big data has made it difficult to offer the same level of reliability and performance available with databases until now. Delta Lake brings these important features to data lakes.

Catalog the data in your data lake

Use data catalog and metadata management tools at the point of ingestion to enable self-service data science and analytics

Challenges

Reliability issues

Without the proper tools in place, data lakes can suffer from data reliability issues that make it difficult for data scientists and analysts to reason about the data. These issues can stem from difficulty combining batch and streaming data, data corruption and other factors.

Slow performance

As the size of the data in a data lake increases, the performance of traditional query engines has traditionally gotten slower. Some of the bottlenecks include metadata management, improper data partitioning and others.

Lack of security features

Data lakes are hard to properly secure and govern due to the lack of visibility and ability to delete or update data. These limitations make it very difficult to meet the requirements of regulatory bodies.