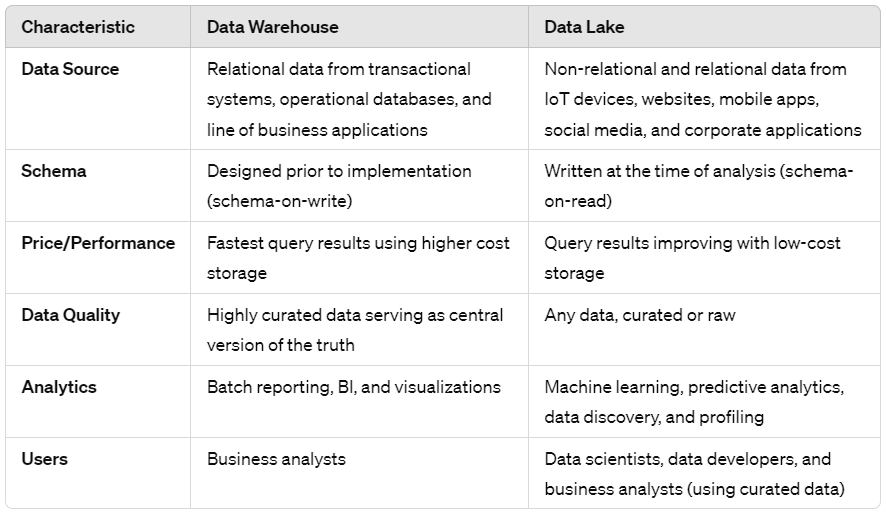


**A data lake** serves as a centralized repository designed to handle vast amounts of data in its original, unprocessed form, whether structured, unstructured, or semi-structured. Unlike traditional hierarchical DW, which organizes data in folders or files, a data lake adopts a flat architecture to store data.

**Key Characteristics:**

1. Data lakes store raw data from diverse sources like system records, sensors, and social media, supporting reporting, visualization, analytics, and machine learning.
2. Data lakes store data in its original format, without rigid schemas, enabling flexible analysis of various data types and formats.
3. Data lakes use a flat architecture, unlike hierarchical data warehouses, making data storage and access simpler and more scalable.



**Advantages of Data Lakes:**

1. *Data Variety:* Handles all data types, structured and unstructured, providing a single repository for diverse data sources.
2. *Flexibility:* Offer flexibility in storing and managing data, and adapts easily to changing business needs.
3. *Raw Data Storage:* Stores data as-is for refining later and analysis as understanding and insight improve over time.
4. *Flexible Querying:* Allows various ways to query and analyze data to extract insights using various tools and techniques.
5. *Tool Compatibility:* Works with a wide range of analytics tools, to gain deeper insights and understand the meaning behind the data.
6. *Value Extraction:* Helps derive insights from diverse data sources, leading to enhanced decision-making and business outcomes.
7. *Accessible Data:* Provides easy access to information for all stakeholders.

**Data Lake Architecture:**

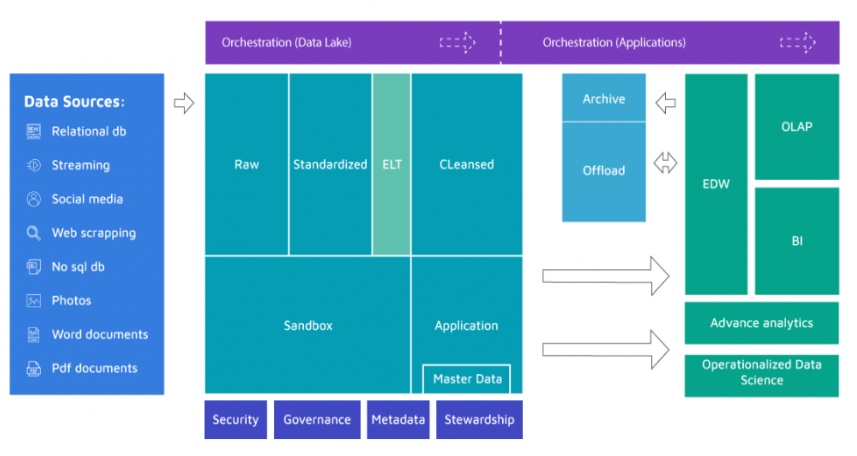
*Single Repository:* Acts as a centralized storage for all data, equipped with orchestration and job scheduling features. (Includes tools for managing workflows and executing tasks.)

*Diverse Data Sources:* Integrates various data types like Operational data: sales, finances, & inventory. Auto-generated data: IoT device data, logs. Human-generated data: social media posts, and emails. Native Format Storage: Stores data in its original format.

*Divided into layers for flexibility:* Raw, Standardized, Cleansed, Application, Sandbox

**Data Lake Architecture: Key Components**

1. *Security:* During initial phases and architectural planning, robust security measures are vital. Unlike RDB, Data Lakes lack extensive built-in security mechanisms.
2. *Governance:* Monitoring, logging, and lineage operations become crucial to track data operations for better Data Lake management and adjustments.
3. *Metadata:* Provides info about your data's structure and usage guidelines.
4. *Stewardship:* Assigns ownership and responsibilities to teams or individuals for data management based on scale and need.
5. *Master Data:* Stores essential, ready-to-use data for easy access and reference.
6. *Archive:* Stores older data, which might pose storage and speed challenges with large amounts.
7. *Offload:* Relieves Data Warehouses by moving resource-heavy tasks to the Data Lake for faster and cheaper processing.
8. *Monitoring/Orchestration and ELT Processes:* Tools help manage data flow and transformations efficiently.



**Data Lake Architecture: Layers**

1. *Raw Data Layer:*

Initial entry point for data, known as the Ingestion or Landing Area.

Data arrives untouched, organized in folders, without any changes made.

Primarily for swift data intake, with duplicates and versions managed.

1. *Standardized Data Layer:*

Optional layer improving data transfer from Raw to Curated.

Applies scheduled & on-demand transformations for cleansing, selecting suitable formats.

Structure similar to Raw but may be partitioned for detail.

1. *Cleansed Data Layer:*

Also called Curated or Conformed Layer, it transforms data into usable datasets, often stored in files or tables.

Complex transformations like denormalization occur, ready for end-user access.

1. *Application Data Layer:*

Applies business logic to Cleansed data, adding specific application requirements.

Maintains Cleansed data structure, adding features like security or machine learning.

1. *Sandbox Data Layer:*

Applies business logic to Cleansed data, adding specific application requirements.

Maintains Cleansed data structure, adding features like security or machine learning, before moving it to other layers.

