AI FOR FINANCE LEADERS

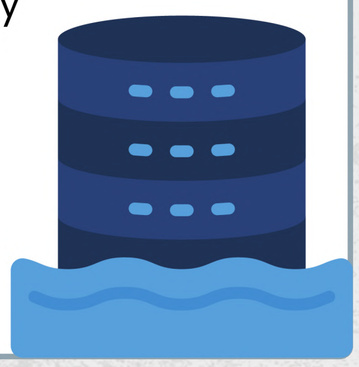
DATA LAKE FOR EVERYONE

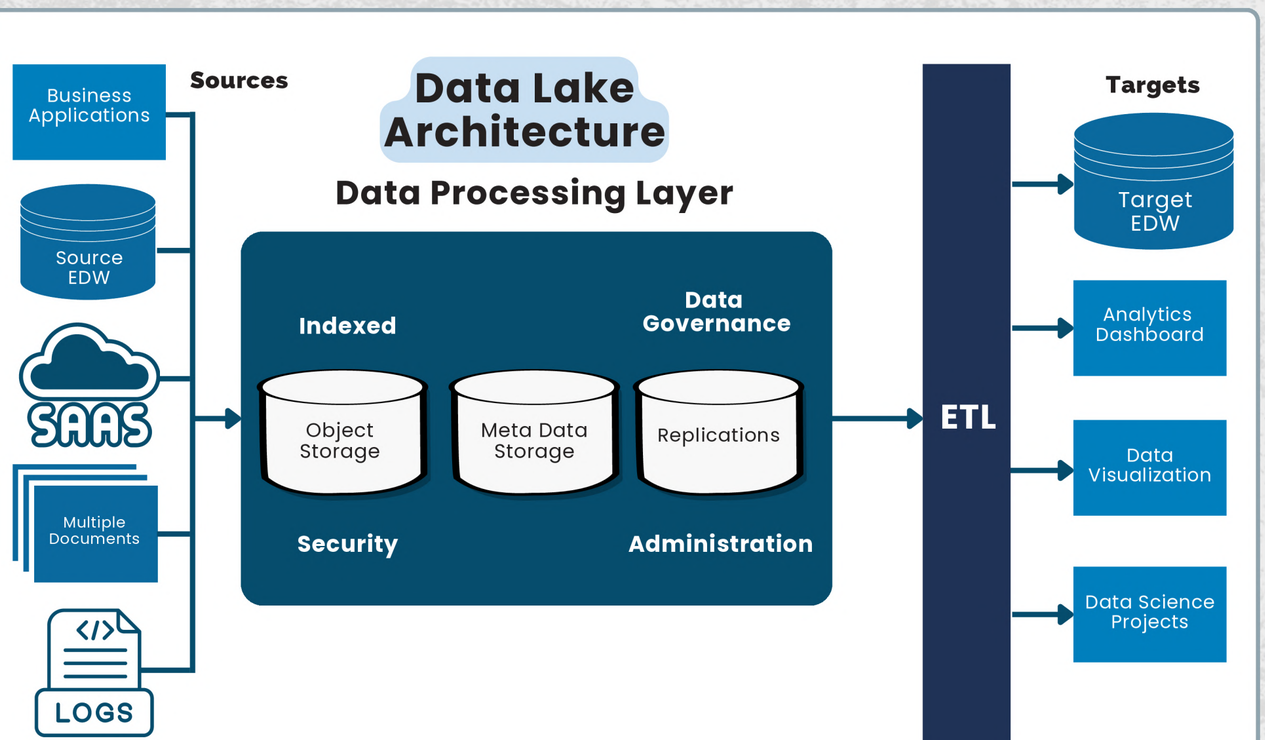
By Tariq Munir

What is a Data Lake?

A **data lake** is a centralized data repository that stores structured, semi-structured, or unstructured data in its raw format.

It Allows organizations to **easily access** and **analyze** diverse data types. To maintain efficiency, it's crucial to prioritize storing only the most important data.

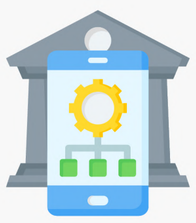




Key Best Practices

To maximize the effectiveness of a Data Lake, consider these best practices:

Incremental Scaling: Start with a small, focused implementation. As your understanding and needs evolve, scale your Data Lake incrementally.



Governance Framework: Implement strong data governance from the outset. This includes data lineage tracking, access controls, and ensuring compliance with regulatory requirements.

Data Cataloging: Develop a comprehensive data knowledge catalog that helps users discover, understand, and access the data. This improves data transparency and reduces redundancy.

Performance Optimization: Regularly monitor and optimize the performance of your Data Lake, particularly in how data is ingested, stored, and retrieved. This ensures it remains efficient and effective over time.

Data Warehouse Vs Data Lake

Data Warehouse	Data Lake
Contains **structured data** pre-processed data	Contains data in a raw, **unstructured form** in addition to structured one.
Difficult and expensive to scale as data needs to be in transformed and query-ready form.	Easy to scale at a low cost as it mostly uses distributed cloud storage.
ETL (Extract, Transform, Load). Data is extracted from its sources and then structured to make it ready for business-end analysis	ELT (Extract, Load, Transform). In this process, the data is extracted from its source for storage in the data lake, and transformed when needed.
Primarily used by business analysts for core reporting and visualization.	Utilized by data scientists, developers, and business analysts for Machine Learning, AI applications, and Advanced Analytics.

Key Considerations for CFOs

1 Understand the Data Landscape

Align data sources to be ingested in Data Lake, with strategic goals and business use cases. Collaborate with CDOs/CIOs to integrate non-financial data for better forecasting and decision-making.

2 Data Capture and Transformation

Establish processes for capturing and harmonizing data, starting with relevant data and expanding gradually.

3 Single Source vs. Multiple Versions of Truth

Balance having a Single Source of Truth (SSoT) and Multiple Versions of Truth (MVoT) to enabled tailored usage across the organization.

4 Privacy and Security

Prioritize security and privacy, implementing strong governance and ensuring compliance with industry regulations.

Follow me to learn more about Digital Transformation, Data, and AI. /in/tariq-munir/