Data Lakes

Brief by Melita Saldanha, and Neha Shetty

Why should you care? - A **Data Lake** is a **centralized repository** that allows storing of structured and unstructured data in its **raw format**, usually object blobs or files. It offers high data volume and quality to increase real-time analytic performance required in Big Data and Machine Learning and boost native integration for dashboards. Data Lakes offer business agility without the need of modeling an enterprise-wide schema or data silo structure. It can be built on multiple technologies such as Hadoop, NoSQL, Amazon S3 and RDBMS.

Approach – Data Lake has a flat architecture ie. there is no hierarchy in the data, but it can be divided into separate data layers: Raw, Standardized, Cleansed, Application and Sandbox. Every data element is marked with a set of metadata information and is given a unique identifier.

Data flows through the system with little or no latency. The Data Lake architecture consists of these tiers: **Ingestion** (Data sources), **Insights** (Data Analysis), **HDFS** (Data at rest), **Distillation** (Converts to structured data), **Processing** (Realtime algorithms and queries), **Unified Operations** (System Management and Monitoring). Organization of Data Lakes can be guided by various factors such as Time Partitioning, Data Load Patterns, Security Boundaries, Stewardship, Retention Policies, and Confidential Data Classification.

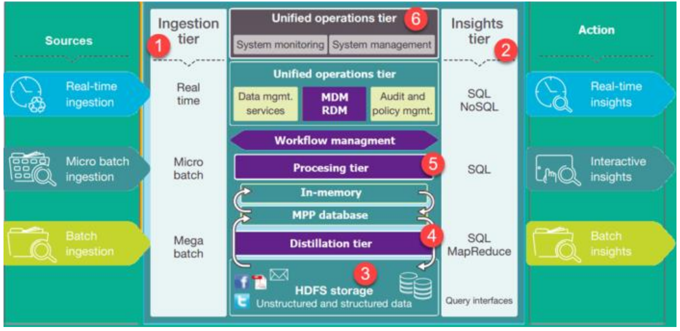

Figure 1: Data Lake Architecture (Figure from Guru99)

Results - Data lakes and data warehouses are being widely used for big data depending on the requirements of the organization. They serve different needs and use cases. A data lake is a vast pool of raw data, the purpose for which is defined later. A data warehouse is a repository for processed, filtered and structured data. Organizations often need both. The ones that are using data warehouses are now evolving their warehouse to include data lakes as these unstructured raw data are used for machine learning. But for analytics use we still need data warehouses.

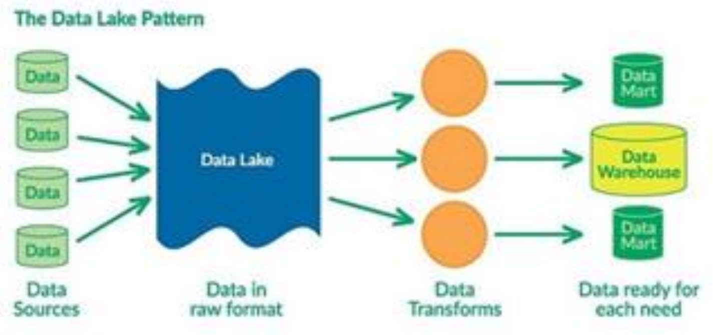

Figure 2: Data Lake Pattern(Figure from Medium)

	Data Lakes	Data Warehouse
Data	Retains all data	Focuses on Business Processes and profiling data.
Accessibility	Highly accessible and flexible	More complicated and costly to make changes.
Data Type	Support all data types.	Consist of data extracted from transactional systems
Type of Users	Data Scientists	Business professionals
Data Processing	Raw data	Highly processed data.

Table 1: Data Lakes v/s Data Warehouse

Pros/Cons

- Pro: Retains all types of data from all data sources and applies schema only when it's ready to use.
- Pro: Customizes management to extract maximum value, aids productionization, and offers cost effective scalability and flexibility.
- Con: Raw data stored with no oversight of the content requires good data governance, semantic consistency, security and access controls, and also increases storage and computational costs.
- Con: Unruly storage of data may lead to ungoverned chaos, unusable data, disparate and complex tools, resulting in a data swamp.

Conclusion - A Data Lake is a storage repository that can store large amount of structured and unstructured data. In the modern era, for companies looking to collect and store the raw data needed for next-generation data analysis, artificial intelligence and machine learning the data lake has emerged as an attractive data architecture.