A Brief Introduction to Data Management Architectures

PRESENTED BY: SARIM ASRAR & ZOYA SHAFIQUE

BIG DATA AND SCALABLE COMPUTATION

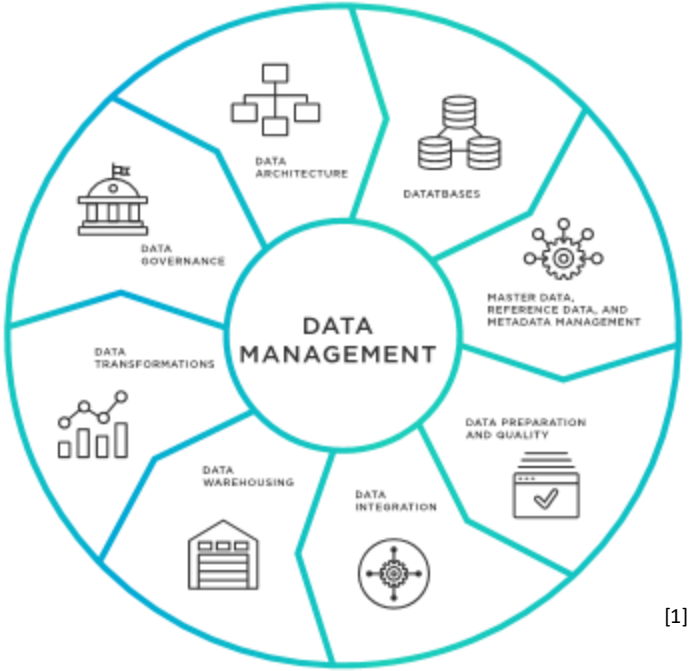
SPRING 2024

Agenda

- Introduction to Data Management
- History of Data Management Architecture
  - Data Warehouses
  - Data Lakes
  - The need for Lakehouses
- Modern Data Management: Data Lakehouses
- Demo

Introduction

- **Data management (DM):** The methods, architectural techniques, and tools for accessing and managing data
  - Needed to fill all data requirements for use cases, applications, and business processes



[1]

[1] https://www.tibco.com/glossary/what-is-data-management

History of Data Management Infrastructure

DATA WAREHOUSES

Data Warehouses

Businesses use relational databases for data management (1970s – 1980s)

With the rise of the internet, companies have more data than they ever had before

Relational databases were sufficient for business needs as

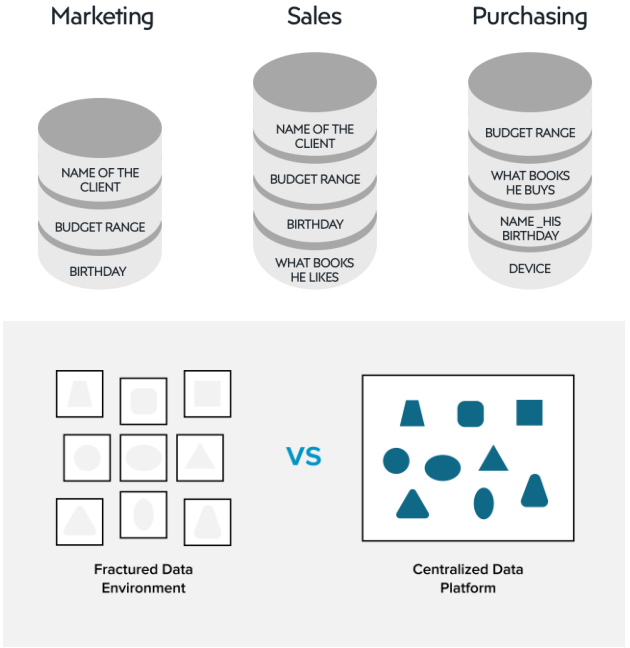
- Data was relatively small
- Relational databases were simple and reliable

With increases in data volume, a single database was no longer sufficient.

- Companies built multiple databases
- Led to the formation of data silos and inefficient data management

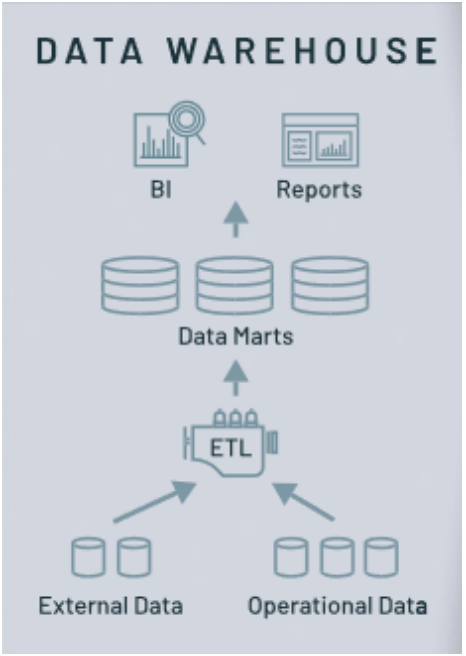
Data Warehouses

- **Data silo**: decentralized, fragmented stores of data
  - Inefficient and costly
  - Multiple copies of data in different places
- Data warehouses were created to centralize and efficiently manage data as data volume increased
  - "Single Source of Truth" [2]
- **Data warehouses**: a system for aggregating data from different sources into a single, central, consistent data store to support data analysis.



[2] https://www.oracle.com/database/what-is-a-data-warehouse/

Data Warehouses



- Main components:
  - Extraction, Transformation, Load (ETL) tools
  - Central database
  - Metadata management
  - Access tools

- **Data mart**: a partitioned segment of a data warehouse that is oriented to a specific business area or team, such as finance or marketing.
  - Data mart = store of bottled water

[3] The Data Lakehouse Platform for dummies

Characteristics of Data Warehouses

"Data warehouse architecture refers to a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process." -- Bill Inmon

Subject-oriented	Integrated	Time-variant	Non-volatile
Data is organized around subjects rather than the applications/sources that generate the data	Data from each source is consolidated and made consistent	Data is maintained over time / historical data is tracked	Data written does not get overwritten or deleted

[4] https://www.thoughtspot.com/data-trends/data-modeling/data-warehouse-architecture

Data Warehouses

Advantages	Disadvantages
Most sensible choice for data platforms whose primary use case is for data analysis and reporting	Only supports business intelligence and reporting use cases
Centralized repository for storing and managing data	Lack scalability and flexibility
Provide features for data security and governance	High initial setup and maintenance costs
Retain historical data over time	Vendor lock-in
Improve data quality and consistency with built-in ETL processes	

An Overview of Data Lakes

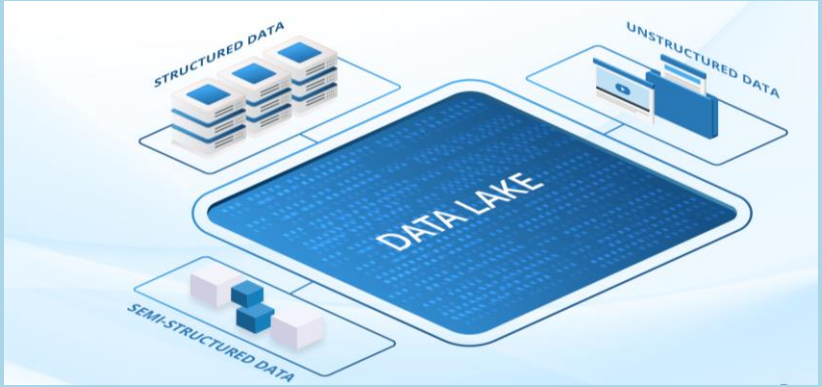
In the presentation, the evolution of data lakes from traditional data warehouses is discussed, highlighting their unique characteristics and significance in the realm of big data and analytics. It emphasizes the role of data lakes in enabling data-driven decision-making by providing an overview of their key components, including storage, data management, processing, analytics, and security measures. The presentation also sheds light on the architecture of a data lake, the importance of metadata, and how it integrates with existing IT infrastructures to support effective data handling.

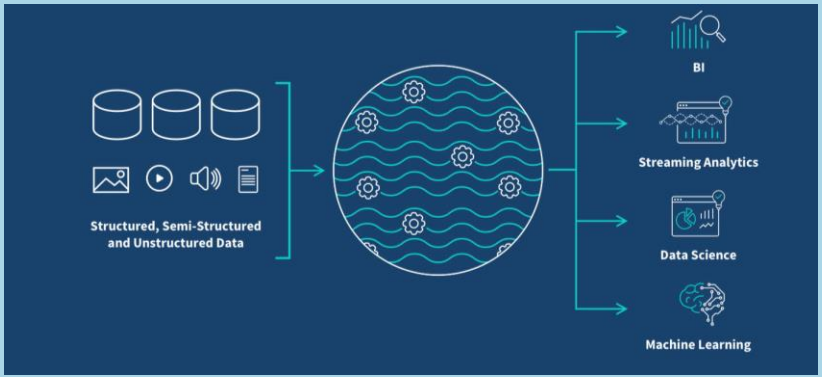
Following are the presentation contents:

- Evolution of data storage: From Data Warehouse to Data Lakes
- Characteristics of Data Lakes
- Importance of Data Lakes
- Components of Data Lakes
- Best practices for implementing a data lake
- Real world applications of Data Lakes
- Challenges and Consideration
- Future of Data Lake

What is Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed for insights like Machine Learning, Data Science, Streaming Analytics and Business Intelligence.

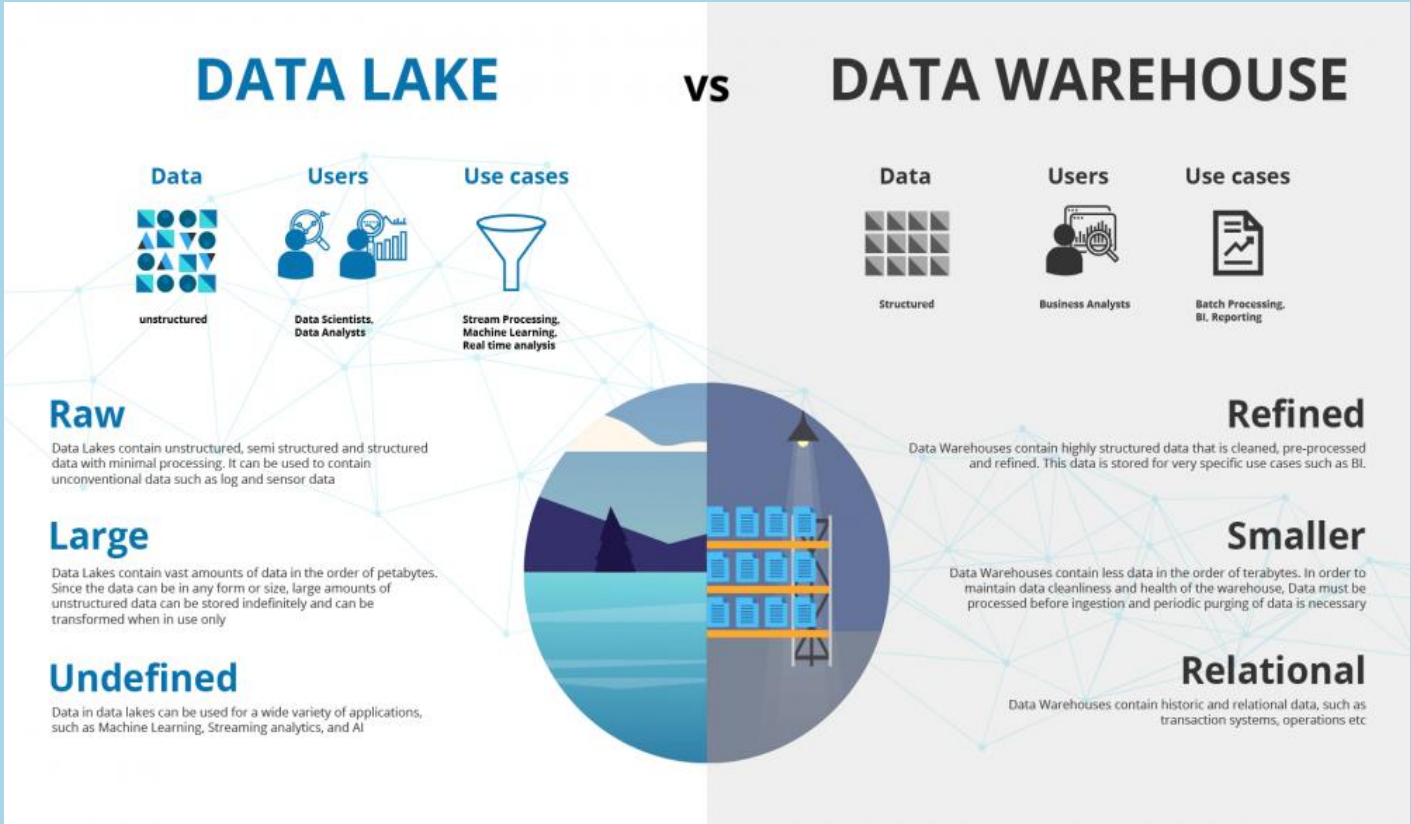




Evolution from Data Warehouses to Data Lakes

Traditional data warehouses were designed for structured data, while data lakes are designed to handle structured, semi-structured, and unstructured data. This makes data lakes more flexible and scalable.

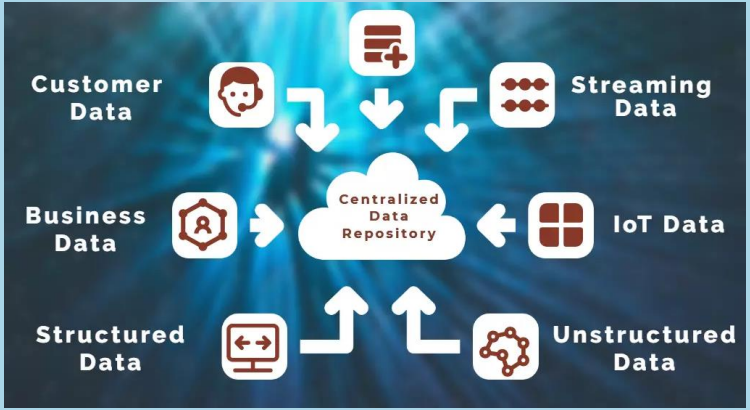
- Data lakes accommodate diverse data types
- Single place to look for data sources
- Scalability to meet evolving data needs



Characteristics of Data Lakes

Data lakes are characterized by their ability to store large volumes of data, support diverse data types, and provide flexibility in data usage and analytics.

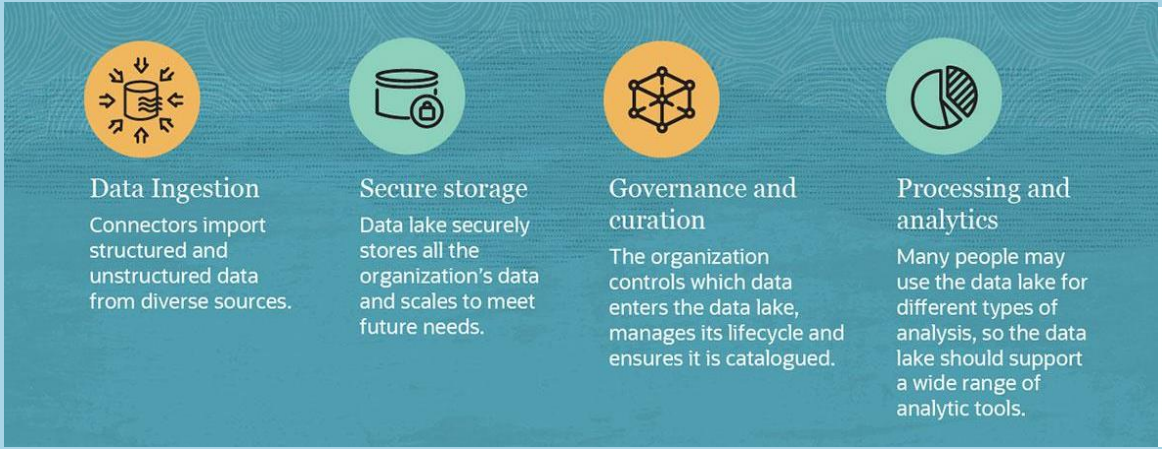
- On-read Schema: Unlike traditional databases that use a schema-on-write, data lakes typically use a schema-on-read approach, where data structure and schema are applied at the time of analysis or query, not when data is ingested.
- Perform Advanced Analytics
- Real-Time Data Processing: Capable of processing streaming data in real-time for timely insights and actions.



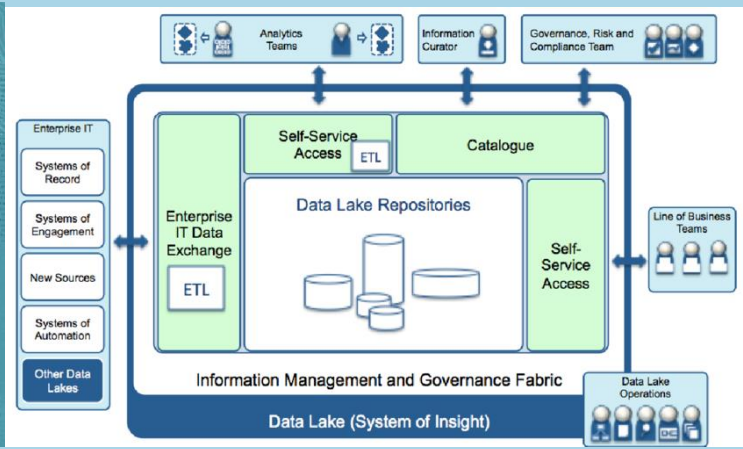
Components of a Data Lake

- Data Ingestion: Involves bringing data into the data lake, cataloging involves indexing and organizing the data, and quality control ensures the data is accurate and reliable.

- Storage: Data lakes typically use object storage for its scalability and costeffectiveness. Technologies like Amazon S3, Azure Data Lake Storage, and HDFS are commonly used.
- Security & Compliance: Security measures include encryption, access controls, and monitoring. Compliance involves adhering to regulations like GDPR and HIPAA.
- Processing & Analytics: Tools like Apache Spark, and Hadoop, are used for processing and analyzing data in a data lake.

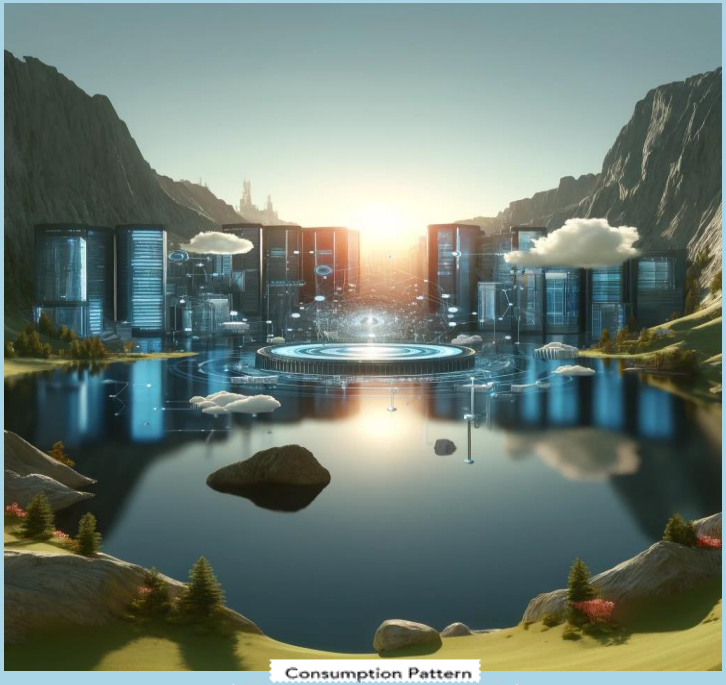


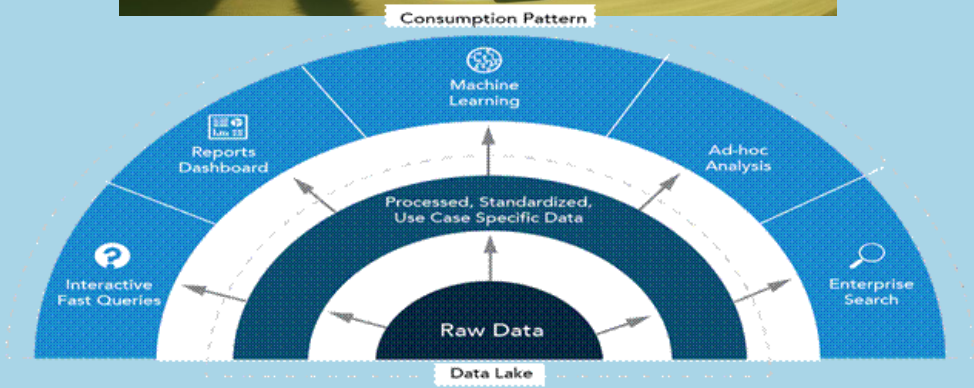
Data Lake Architecture

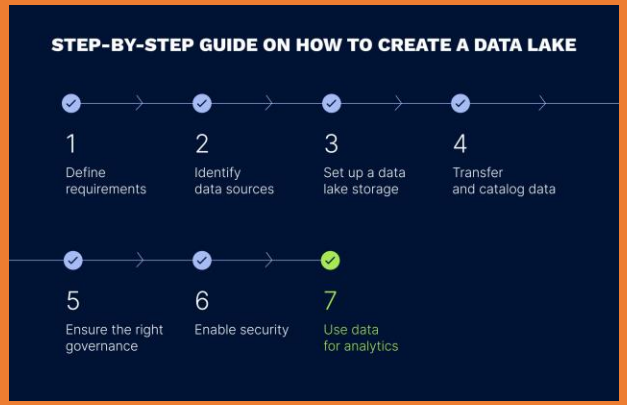


Importance of Data Lakes

- Unified Data Access: Consolidates disparate data sources, providing a single point of truth for the organization.
- Supports Big Data Initiatives: Enables storage and analysis of large volumes of data, essential for big data analytics.
- Reduces Costs: Offers cost-effective storage solutions compared to traditional data warehousing.
- Increases Operational Efficiency: Improves the speed and efficiency of data processing and analysis.
- Encourages Innovation: Provides the raw data necessary for experimentation and new analytics applications.







Best Practices for Implementing a Data Lake

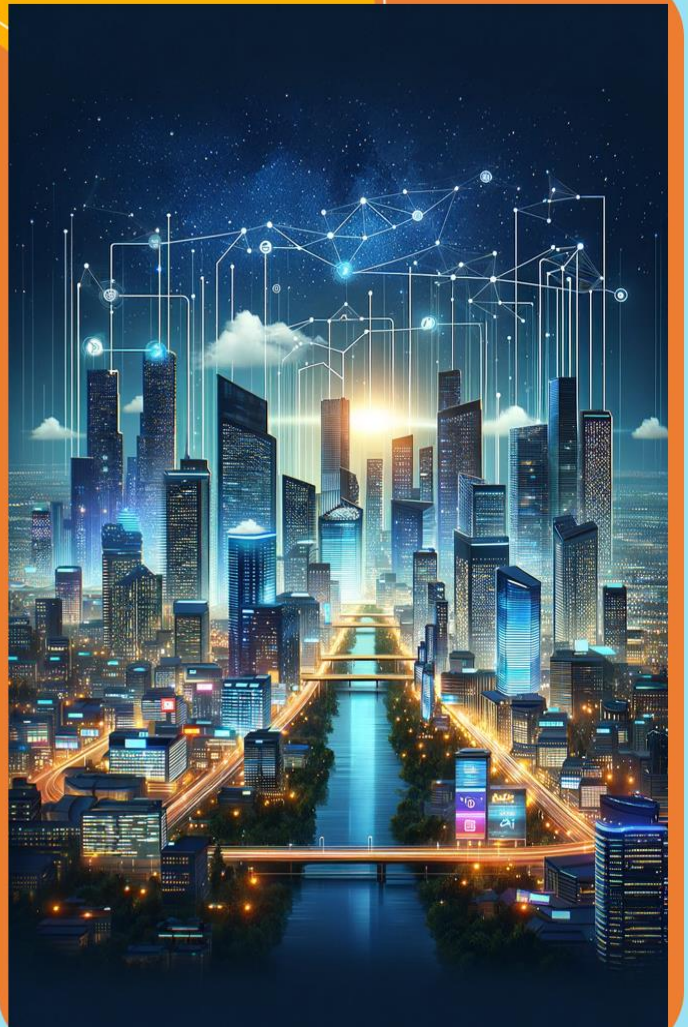
- **Strategic Planning and Business Alignment:** It's important to align the data lake strategy with business objectives to ensure it supports the organization's goals.
- **Data Governance and Quality Management:** Implementing data governance policies and quality management processes is crucial to ensure the reliability and security of the data.
- **Scalability and Performance Optimization**: Designing the data lake for scalability and performance ensures it can handle growing data volumes and complex analytics.
- **Security Considerations:** Robust security measures, including encryption, access controls, and monitoring, are essential to protect sensitive data.

Real-World Applications of Data Lakes

01 **Benefits Realized and Lessons Learned**: Benefits include improved data access, enhanced analytics, and cost savings. Lessons learned include the importance of data governance and the need for skilled personnel.

02 **Successful Implementations:** Companies like Netflix, Amazon, and Shell have successfully implemented data lakes to support their big data and analytics needs.

03 **Industry-Specific Uses:** In healthcare, data lakes are used to store patient data for analysis. In finance, they are used for fraud detection and risk analysis. In retail, they are used for customer behavior analysis and inventory management.



Challenges and Considerations

Despite their benefits, many of the promises of the data lakes have not been realized due to the lack of some critical features. Since data is in multiple formats, data can be huge, dirty, repeated and difficult to transform. Hence, data lakes can turn into data swamps.

01

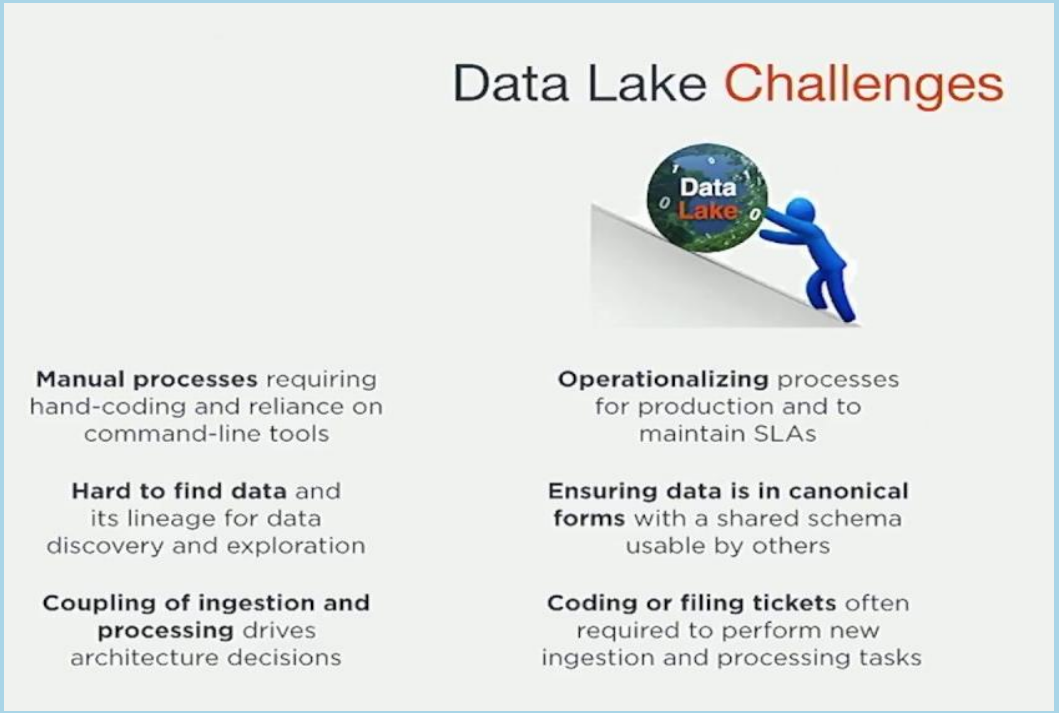
Addressing the Skills Gap and Training Needs: Investing in training and hiring skilled personnel is essential for the successful implementation and management of a data lake.

02

Management Difficulty
Even for skilled engineers, data lakes are hard to manage. Data lakes can store large amounts of unstructured data; therefore, businesses need to have good data management practices otherwise their data lake may turn into an unusable data swamp.

03

Managing Data Lake Evolution Over Time: Regularly reviewing and updating the data lake architecture and processes is crucial to keep up with evolving business and technology needs.

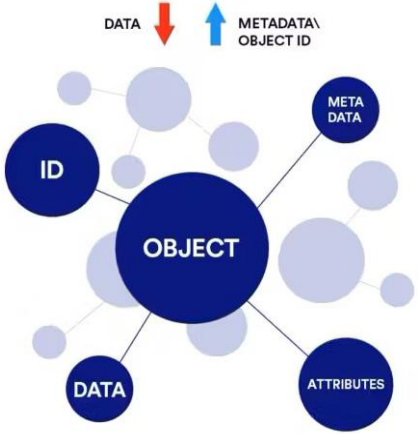


Modern Data Infrastructure

DATA LAKEHOUSES

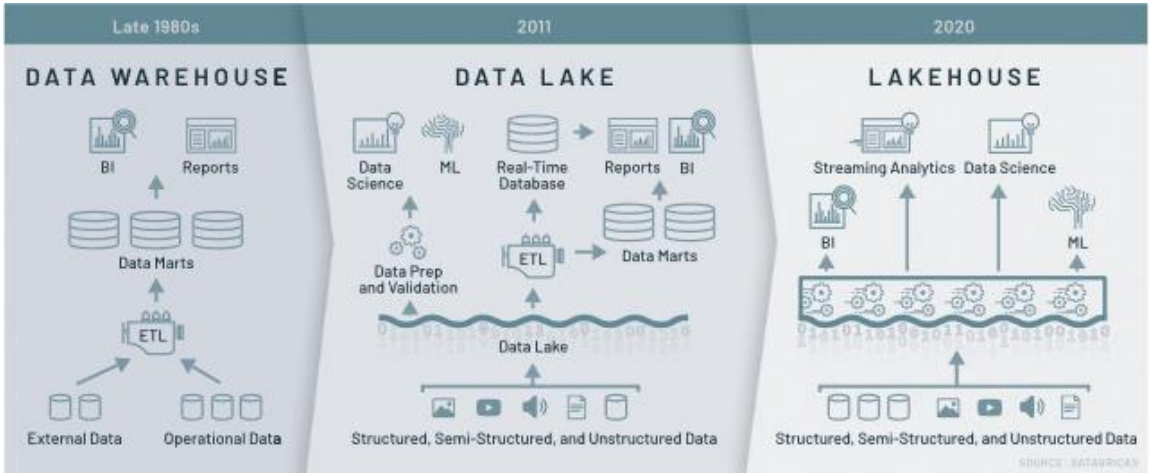
Object Storage

- A data storage architecture that manages data as distinct units. All objects are at the same "level".
- Each unit, or "object", contains
  - Data (all data may not be stored together)
  - Metadata
  - Globally unique identifier
- Abstracts some of the lower layers of storage away
- Key benefits of object storage:
  - "Unlimited" scale



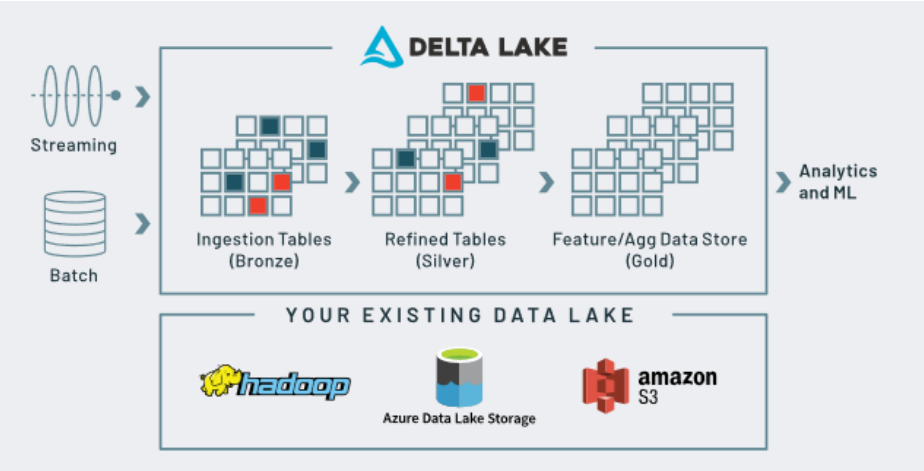
Modern Data Management:
Data Lakehouses

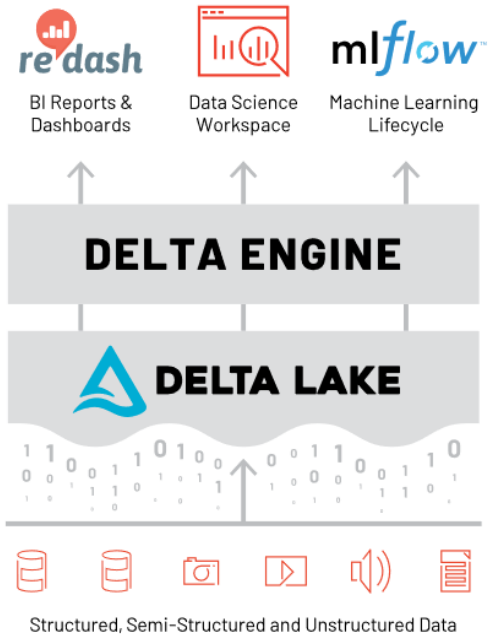
- Data Lakehouses = Modernized data warehouse with data lake type storage
  - Implement low-cost object storage used for data lakes with data structures and management processes from data warehouses



[3] The Data Lakehouse Platform for dummies

Modern Data Management:
What Does a Lakehouse look like?





[5] https://docs.databricks.com/en/delta/index.html

Modern Data Management:
What Does a Lakehouse look like?

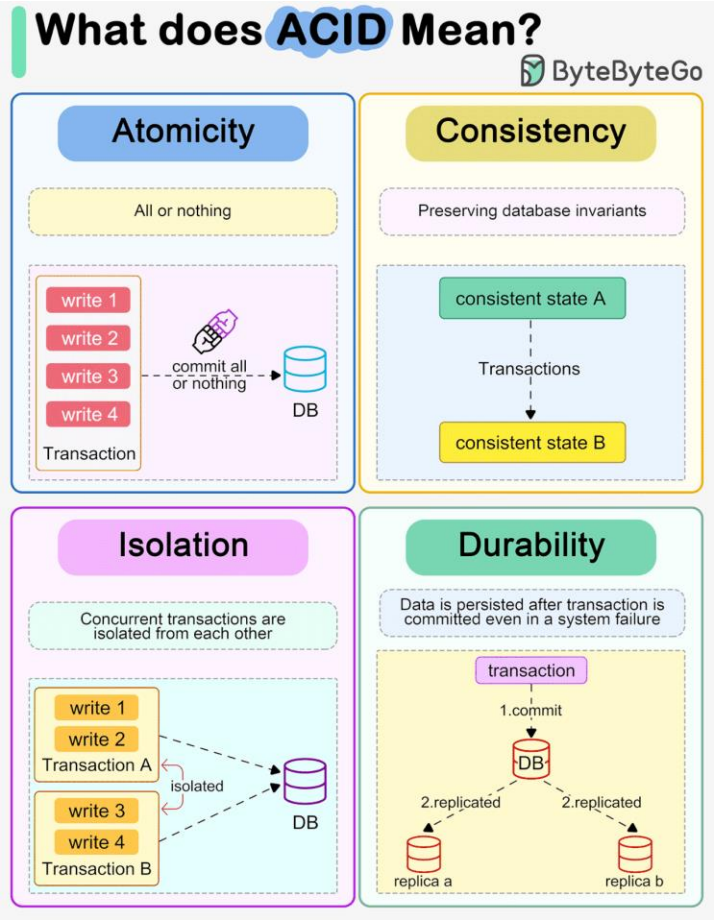
[6] https://www.datastackhub.com/top-tools/open-source-data-lake-tools/
[7] https://www.sprinkledata.com/blogs/10-best-data-lake-tools-2023

Modern Data Management:
Comparing Warehouses, Lakes, and Lakehouses

Data Warehouses	Data Lakes	Data Lakehouses
Works well with structured data	Useful for semi-structured and unstructured data	Can handle structured, semi-structured and unstructured data
Optimal for data analytics and business intelligence use-cases	Suitable for machine learning and artificial intelligence use-cases	Suitable for both data analytics and machine learning cases
Storage is costly and time-consuming	Storage is cost-effective, fast, and flexible	Storage is cost-effective, fast, and flexible
ACID-complaint	Non-ACID compliant	ACID-compliant

Modern Data Management: ACID Transactions

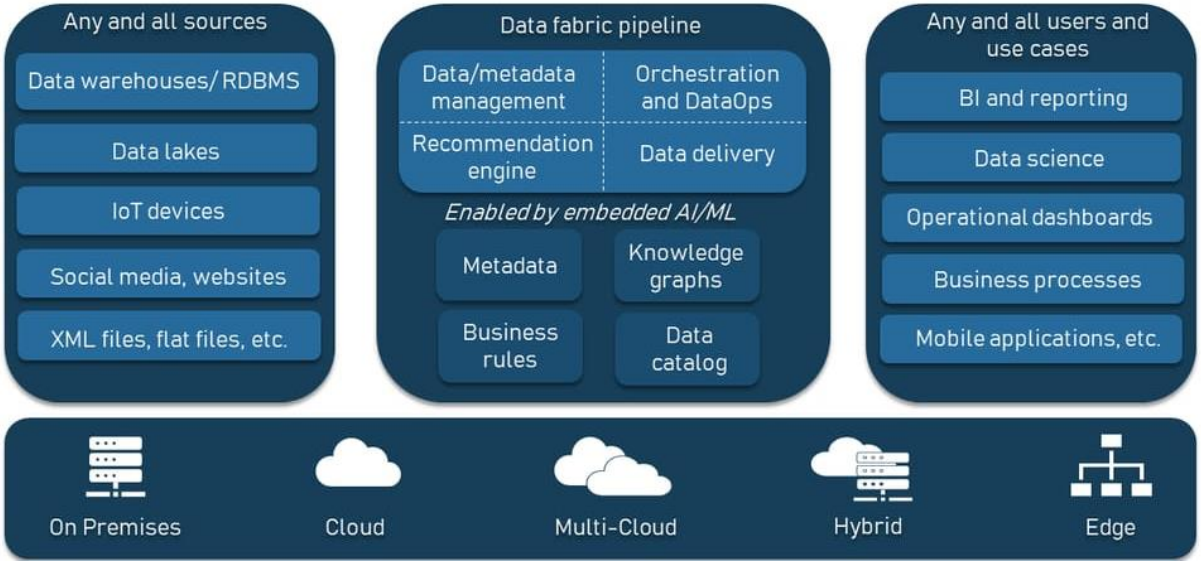
Data warehouses	ACID complaint
Data lakes	Non-ACID complaint
Data lakehouses	ACID complaint



[8] https://blog.bytebytego.com/p/ep89-what-does-acid-mean

What's Next?

- Data Fabric



[9] https://www.ibm.com/data-fabric

Demo

A BRIEF INTRODUCTION TO SNOWFLAKE

Thank you!

Supplementary Material

Modern Data Management:
Aren't Data Lakes Enough?

- Traditional data lakes
  - Don't support ACID transactions
  - Don't enforce data quality or consistency
  - Failed jobs go undetected and aren't discovered until you try to access the data
  - Cannot efficiently handle metadata
  - Initially stored data in files

  - One approach:
  - Use multiple systems
    - A data lake, multiple data warehouses, other specialized systems such as streaming and image databases
    - Introduces complexity and delay

Modern Data Management:
Which Data Management Structure Should I Choose?

Data Warehouse

- Good choice for companies who need a structured data solution that focuses on BI and data analytics with structured data

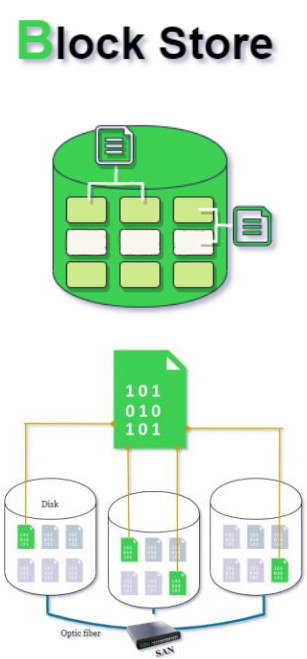
Data Lake

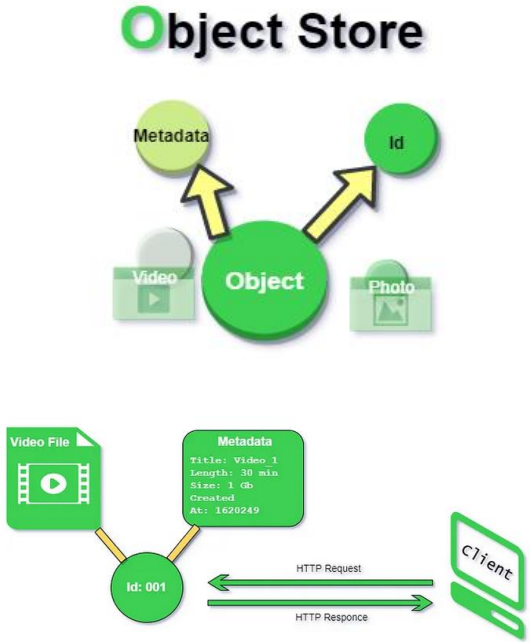
- Ideal for organizations who need a flexible, low-cost, big-data solution for machine learning and data science workloads
- Good choice for companies working with large amounts of unstructured data

Data Lakehouse

- Optimal for companies dealing with BI, data analytics, and machine learning workflows.

Object Storage

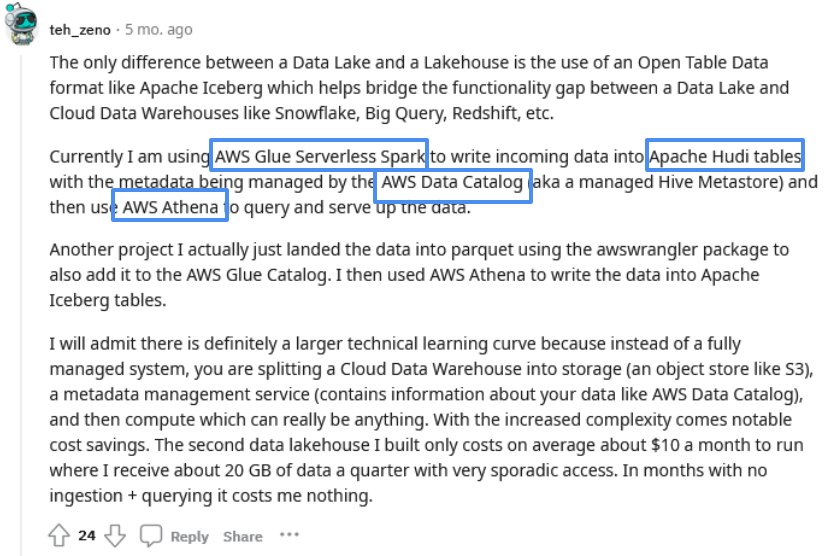




[5] https://rajkumaraug20.medium.com/file-storage-vs-block-storage-vs-object-storage-2519031a2646

Current Challenges

- Relatively new technology --> still an evolving technology
- Increased overhead in tool version compatibility and upgrades
- Difficult learning curve to set up and use



teh_zeno · 5 mo. ago

The only difference between a Data Lake and a Lakehouse is the use of an Open Table Data format like Apache Iceberg which helps bridge the functionality gap between a Data Lake and Cloud Data Warehouses like Snowflake, Big Query, Redshift, etc.

Currently I am using AWS Glue Serverless Spark to write incoming data into Apache Hudi tables with the metadata being managed by the AWS Data Catalog (aka a managed Hive Metastore) and then use AWS Athena to query and serve up the data.

Another project I actually just landed the data into parquet using the awswrangler package to also add it to the AWS Glue Catalog. I then used AWS Athena to write the data into Apache Iceberg tables.

I will admit there is definitely a larger technical learning curve because instead of a fully managed system, you are splitting a Cloud Data Warehouse into storage (an object store like S3), a metadata management service (contains information about your data like AWS Data Catalog), and then compute which can really be anything. With the increased complexity comes notable cost savings. The second data lakehouse I built only costs on average about $10 a month to run where I receive about 20 GB of data a quarter with very sporadic access. In months with no ingestion + querying it costs me nothing.

24 Reply Share ...

Understanding Data Lakes

This presentation offers an in-depth look at data lakes, covering essential aspects such as their characteristics, key components, and best practices for implementation. It delves into the practical applications of data lakes across various industries, addresses common challenges and considerations, and looks ahead to future developments in this area.

The aim is to equip data practitioners with the knowledge needed to make informed decisions when handling large datasets.

The Future of Data Lakes (with Data Warehouse)

The future of data lakes is tied to evolving data management technologies and the increasing importance of data in driving business decisions.

- ✓ **Improved Query Performance:** Technologies will improve to allow SQL and NoSQL queries to run more efficiently on data lakes, enabling them to handle a wider range of workloads with better performance.

- ✓ **Edge Computing:** As IoT devices proliferate, data lakes will extend to the edge, with edge devices performing some data processing and only sending necessary data back to the central data lake, reducing latency and bandwidth usage.

- ✓ **Predictions for the Future Impact on Businesses and Technology**: Data lakes are expected to become even more central to data strategies, driving innovation and enabling more advanced analytics and decision making. They are combined with Data Warehouses to form Data Lakehouses.