CM2606 Data Engineering

Modern Trends in Data Engineering

Week 11 | Piumi Nanayakkara

ROBERT GORDON UNIVERSITY ABERDEEN
TEF Gold
INFORMATICS INSTITUTE OF TECHNOLOGY

Learning Outcomes

- Covers 01 and 02 for Module

- On completion of this lecture, students are expected to be able to:
  - Understand thinking behind modern data concepts with respect to data eco systems
  - Understand and explain drawbacks of a data system implementation

Content

- Drawbacks of current setups
- Data Lakehouse
- Data Mesh
- Data Virtualization

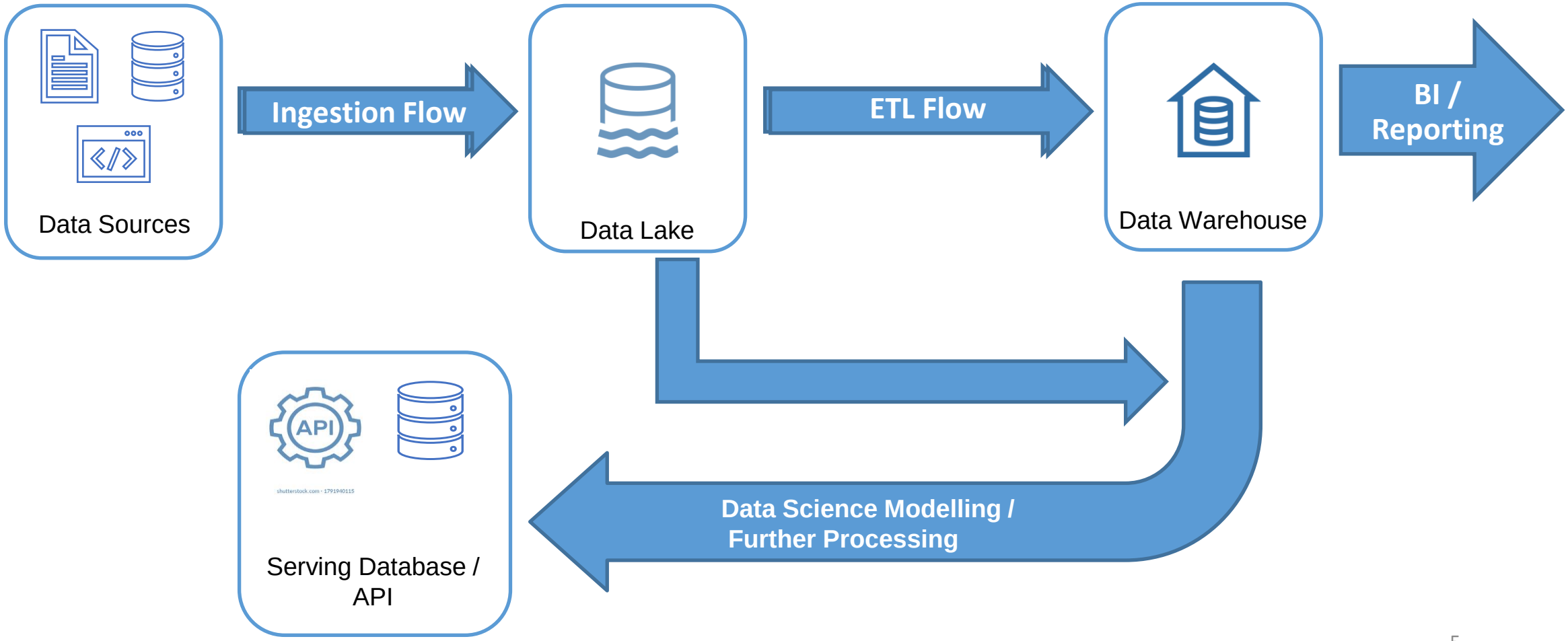
Big Data Storage

- In current data eco systems most organizations use both data warehouses and data lakes
  - leveraging data warehousing to derive valuable business insights and
  - using data lakes for storage and data science.

- Data warehouses were the first designed systems for analytics in 1980s.
  - Idea: ETL data directly from operational databases
  - Optimized for analytics / queries

- Data Lakes
  - Low-cost storage option introduced in 2010s
  - To support analytics, data is also loaded into a data warehouse

Two Tier Implementation



Drawbacks of Data Warehouse

- Lack of Flexibility
  - Less support for unstructured and semi structured data.
    - E.g., Time series, logs, images, documents
  - Less useful in machine learning and AI use cases
    - Data needs to be directly read from data lake
- Cost
  - Cost of holding historical data and running queries against them
  - Requires specialist knowledge in implementing data warehouse from scratch
  - Proprietary Data Warehouse software are expensive and struggle with integrating open-source tools e.g., Spark
  - Needs regular maintenance to prevent being outdated

Drawbacks of Data Lake

- Poor performance for BI and Reporting
  - Lack of consistent data structure can result in sub-optimal query performance
- Data Security and Reliability
  - As a data lake accommodate all data formats, it might be challenging to implement proper data security and governance policies

Issues with Two Tier Implementation

- Data Reliability
  - Additional Processing might lead to reduced data quality
  - Multiple storage systems with different semantics
- Timeliness
  - Add a delay in pipelines
- Cost
  - Duplicated storage
  - increased complexity and costs as data should be kept consistent between the two systems.

Data Lakehouse

- New data storage architecture with single tier
- Enables a single repository for all data (structured, semi-structured, and unstructured)
- Combines a data warehouse's data structure and management features with a data lake's low-cost storage and flexibility.

Data Lakehouse: Features

- Support for structured and semi-structured data types, including streaming data
- Standardized storage formats. E.g., Apache Parquet and ORC .
- Schema support with mechanisms for data governance
  - Control the schema of the tables thanks to the support of schema enforcement
  - Access control and auditing.
- Separation of storage and compute resources
  - Use of separate clusters for storage and compute. This ensures greater scalability:

Data Lakehouse - Implementation

- Implement DW management and performance features as a layer on top of directly accessible data in open formats

- Metadata handling:
  - a unified catalog that provides metadata for all objects in the lake storage
  - Tracks which files are part of which table versions

- Engine Optimization:
  - Add indexes, versioning, caches for performance – in RAM/SSD

Data Lakehouse - Implementation

- Declarative I/O interfaces for access:
  - SQL APIi or direct access to the data files (Ability to read open format files using TensorFlow and Spark MLlib)

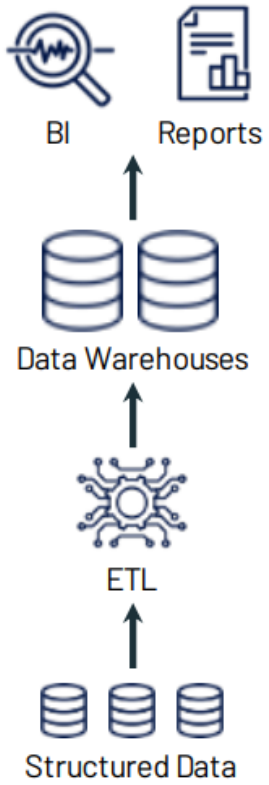
- Optionally can have data processing layers to reflect data warehouse
  - Bronze – row data
  - Silver – Cleaned data
  - Gold – Aggregated data which can be directly connected to dashboards

Data Lakehouse - Solutions

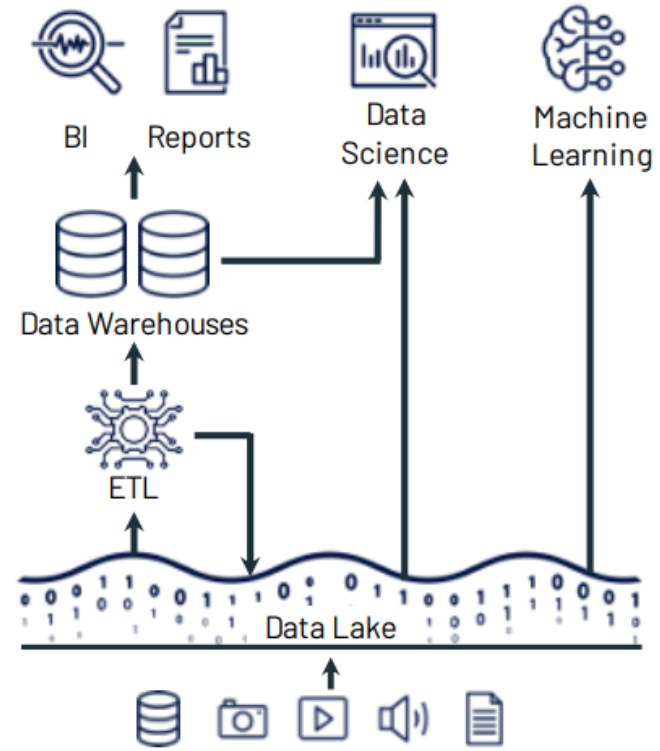
- Databricks is the industry leader and original creator of Lakehouse architecture
  - In the paper introduced by experts from Databricks, UC Berkeley, and Stanford University at the 11th Conference on Innovative Data Systems Research (CIDR) in 2021, L*akehouse officially came into picture*

- Amazon Web Services (AWS) is another pioneer with a Lake House architecture (i.e. Lake Formation + AWS Analytics).

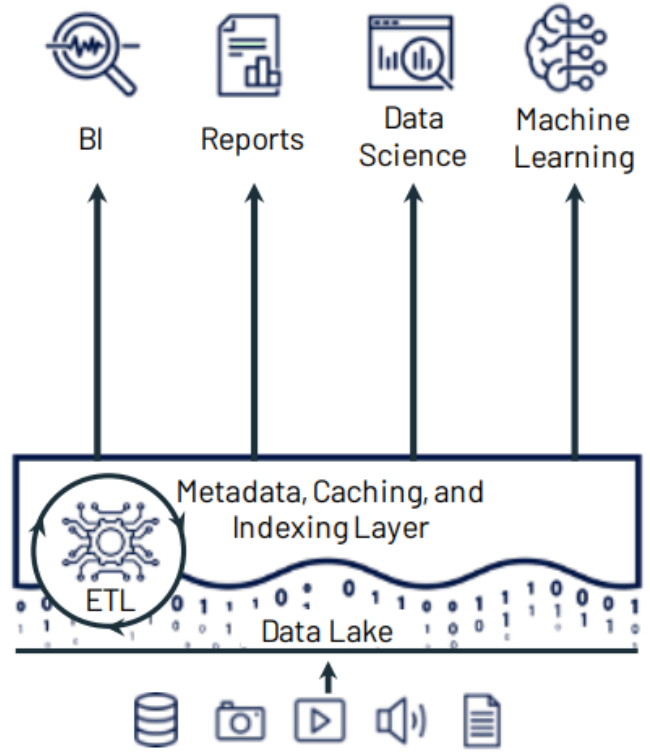
Changes in Architecture



(a) First-generation platforms.



(b) Current two-tier architectures.



(c) Lakehouse platforms.

Source

Benefits of a Data Lakehouse

- Unified data platform
  - Single Source of truth
- Less time and effort administrating
- Simplified schema and data governance
  - Fine Grained – Row/Column level rather than file level
- Increased usability
  - Advanced analytics, reporting, and machine learning.
  - Avoid vendor lock in of data warehouses

Benefits of a Data Lakehouse

- Direct access to data for analysis tools
- Cost-effective data storage
  - No redundancy
  - Reduced ETL cost
- Improved data reliability
  - Fewer cycles of ETL data transfers
- Reduce Data Latency and Staleness
  - Serve the need for agile analytics

Data Mesh

- A type of decentralized data platform architecture leveraging a domain-driven & self-serve design influenced by:
  - Data Marts, Domain Driven Design, Microservices etc.
- Centralized model
  - Suitable for organizations that have a simpler domain with few business use cases
  - Not appropriate for enterprises with rich domains, and large number of data sources / consumers.
- It helps solve the challenges that often come with quickly scaling a centralized data approach.

Data Mesh - Principles

- Data Ownership by Domain
  - Consumption, storage, transformation (ETL), access control and output of data are all decentralized and handled by specific business domain.
- Data as a Product
  - Data is considered a product by each team that publishes it
  - That team is responsible for the data, including quality, representation, and cohesiveness.
- Data available everywhere, self serve
  - To transfer data from one domain to another APIs can be used. E.g., FastAPI
- Data Governed wherever it is

Data Virtualization

- Without moving data from different platforms to a common place physically, use tools like denodo which virtualize the sources and provide the facilities to do analytics inside the tool
- Based on Self service data discovery
- Useful when new sources getting added rapidly and sources are in heterogenous platforms

Data Virtualization – Avoid Usage

- To replace a large-scale enterprise data warehouse
  - Instead, can be used to bridge data across data warehouses, data marts, data lakes and consumption layer
  - Existing data infrastructure can continue performing their core functions while the data virtualization layer just leverages the data from those sources via virtualization
- To access an Operational Data Systems

Data Virtualization - Benefits

- Reduces the risk of data errors
- Cost Saving in terms of storage/ processing resources, time and effort for ETL pipelines
- Provides instantaneous access to data to make real time business decisions
- Query processing push-down to data source

READING

- Matei Zaharia, Ali Ghodsi 0002, Reynold Xin, Michael Armbrust. **Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics**. In 11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings. www.cidrdb.org, 2021.