

Big Data Architectures

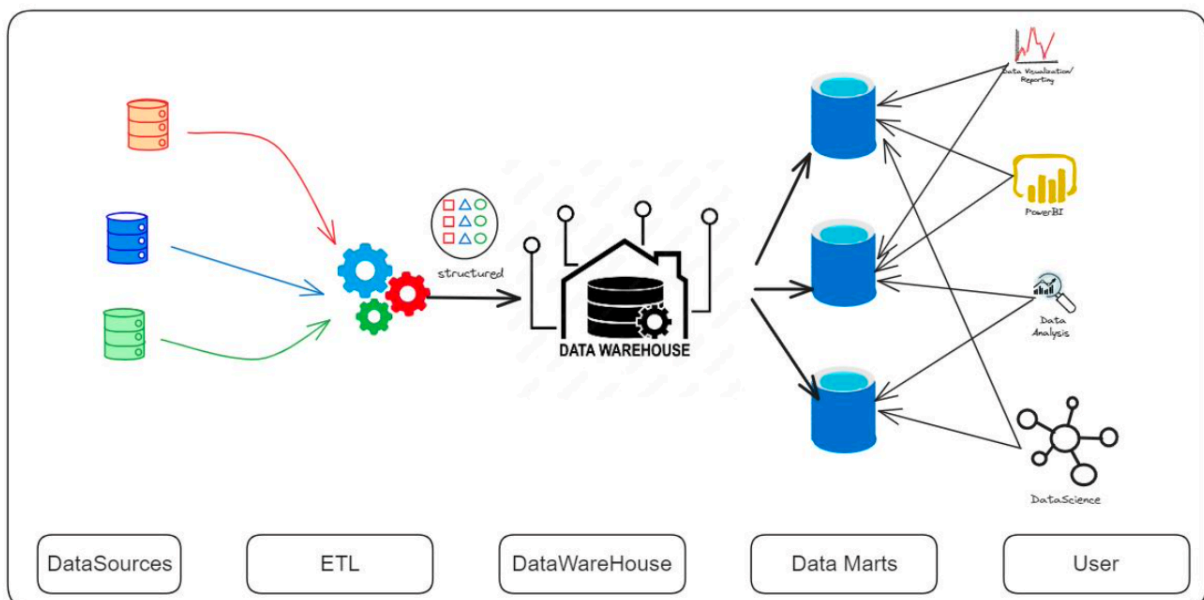
Source : Masters Report of Madiouni Mohsen (Supcom)

Data Warehouses : Centralised Data Management for Informed Decisions

A data warehouse holds a pivotal role in contemporary organizations, acting as a centralized reservoir for extensive data gathered from diverse sources. It serves as the bedrock for business analytics and reporting, offering a singular reference for data-driven activities. By focusing on storing historical data, a data warehouse aggregates relational data from various origins, including application and business data. This information is curated through transformation and cleansing steps before integration, ensuring its quality and unity.

The paramount value of data warehouses stems from their capacity to expedite the generation of vital business insights across the organisation. This ease of access to comprehensive data enables well-informed decision-making for various stakeholders. Business analysts, data engineers, and decision-makers can utilise different tools like BI platforms and SQL clients to navigate, query, and interpret data without extensive data science expertise.

Data warehouses offer strategic benefits by promoting collaboration and breaking down data silos. These benefits extend to retail, finance, healthcare, and manufacturing sectors, where data analytics supports personalised marketing, fraud detection, precision medicine, and supply chain optimisation.



However, data warehouses often function as black boxes, concealing their intricate workings. Data goes in, and insights come out, but the underlying processes remain hidden from users. This "black box" nature might limit users' understanding of the data transformation, integration, and query optimisation steps, potentially leading to a gap between the technicalities of data warehouse operations and user expectations. Moreover data

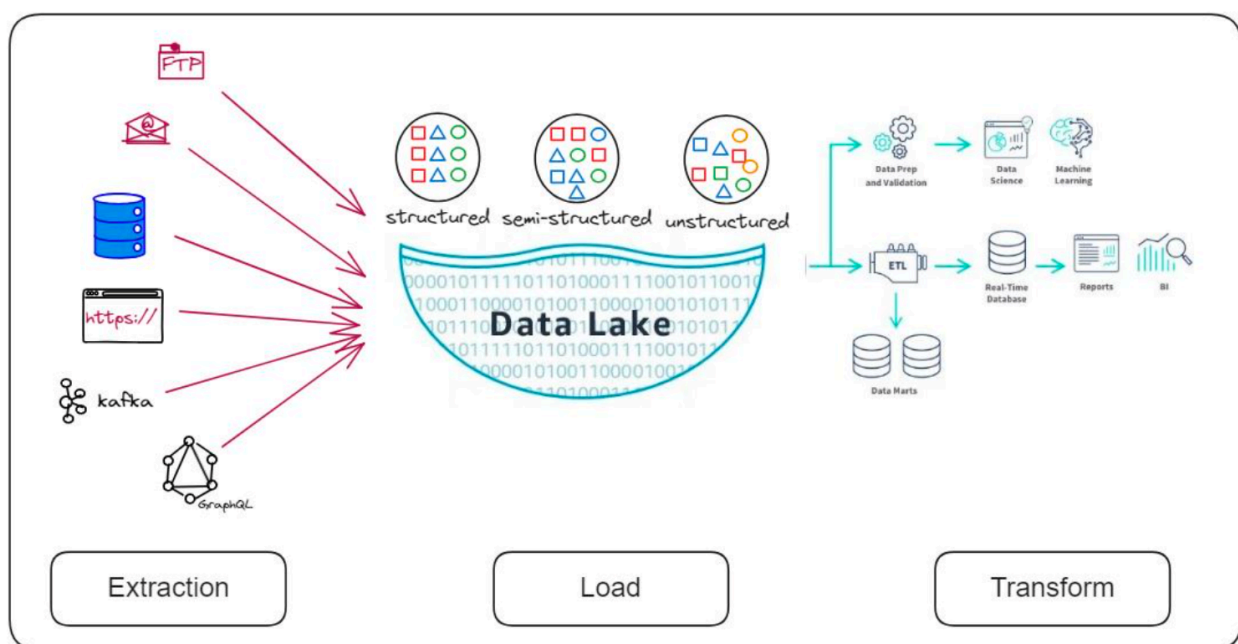
can't be ingested unless it has been cleaned and fully transformed which is not always be possible in big data era.

To harness data warehouses effectively, robust infrastructure, scalable storage, analytics tools, and skilled personnel are imperative. Ensuring data ethics, security, and privacy is vital. By embracing the potential of data warehouses, organisations unlock insights, competitive advantages, and a path to lasting success.

Data Lake : Flexible and Scalable Data Storage

A data lake acts as a centralised repository securely housing vast amounts of structured and unstructured data in its raw form. Unlike data warehouses that store organised relational data, data lakes employ a flat architecture and object storage, offering flexibility, durability, and cost-effectiveness. This enables insights from challenging unstructured data.

Crucially, data lakes capture data without a predefined schema, allowing Extraction, Loading, and Transformation (ELT) when needed. This adaptability suits various data types, such as IoT and social media data. Hence, data lakes facilitate machine learning and predictive analytics.



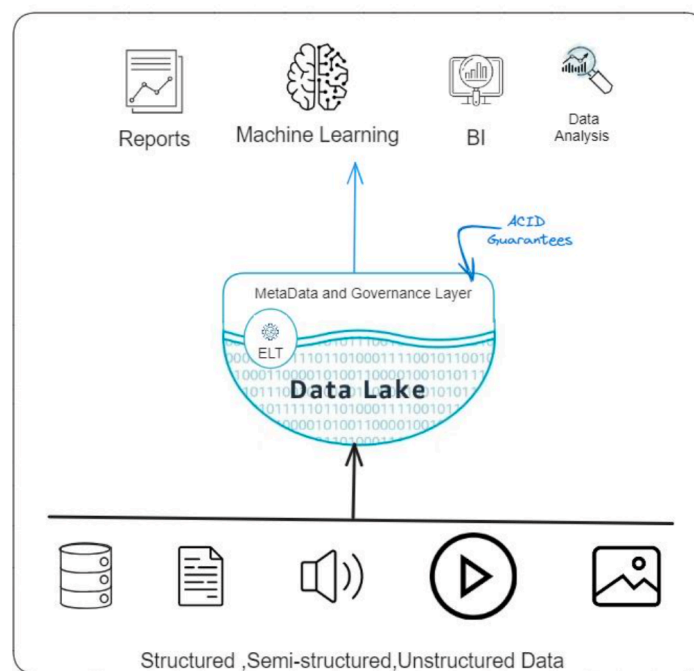
Leveraging data lakes, organisations uncover data potential, driving advanced analytics. This approach uncovers concealed patterns, enriching customer understanding and aiding data-driven decisions.

- Data lakes offer benefits:
 - Data centralization: Centralized storage for all data formats.
 - Data flexibility: Native format storage fosters adaptability.
 - Cost savings: Lower-cost storage via object storage.
- Challenges include:
 - Poor business intelligence performance: Disorganization can hinder analytics.
 - Data reliability and security: Lack of consistency complicates security

Data lakes present a transformative potential, yet their implementation isn't without challenges. A key concern is the lack of governance and metadata management, potentially leading to the "data swamp" problem where data becomes disorganised, impeding effective analytics; in other words there are no ACID guarantees. The very flexibility that makes data lakes appealing can become a challenge, as it allows for data of varying quality and relevance to be ingested, potentially leading to difficulties in data cleaning and accurate analysis. Additionally, the complexity of managing access controls and ensuring data security in a heterogeneous environment can be daunting. Despite these challenges, data lakes remain a compelling solution for organisations seeking to harness the full potential of their data.

Lake House: Merging Data Warehouses and Data Lakes

The emergence of the data lakehouse introduces an innovative storage paradigm that combines the strengths of both traditional data warehouses (ACID Guarantees) and modern data lakes (all data types). This fusion creates a dynamic solution for organisations dealing with vast volumes of data. The data lakehouse establishes a unified repository capable of accommodating various data formats, including structured, semi-structured, and unstructured data. This unification streamlines data storage and integration, eliminating the need for separate storage environments and facilitating seamless data consolidation.



One of the most significant advantages of the data lakehouse architecture as presented in the Figure is its capacity to support a wide spectrum of data processing needs. It bridges the gap between structured data management, which data warehouses excel at, and the flexibility and cost-effectiveness of data lakes. As a result, the architecture enables organisations to easily organise and query their data, while also allowing them to leverage the storage efficiency and adaptability of data lakes.

This hybrid approach to data storage and management holds substantial benefits:

- Reduced Data Redundancy:

- The unified storage model of data lakehouses minimizes data duplication, addressing various data storage and retrieval requirements.
- Cost Efficiency:
 - By tapping into the cost-effective storage capabilities of data lakes, organizations can store massive volumes of data without incurring exorbitant costs. Furthermore, the unified storage system eliminates the complexity of managing multiple storage systems, thus reducing operational expenses.
- Versatile Workloads:
 - The data lakehouse model provides direct access to widely used business intelligence (BI) tools, such as Tableau and PowerBI, facilitating advanced analytics and decision-making. Additionally, it employs open data formats and supports popular machine learning libraries, making data analysis accessible to data scientists and machine learning practitioners.
- Enhanced Governance and Security:
 - Through schema enforcement and data integrity measures, the data lakehouse architecture ensures a robust foundation for implementing data security and governance protocols.

However, it's important to note that the data lakehouse architecture is still relatively new and maturing. While it holds great promise, organisations need to navigate its evolving landscape and address potential challenges as the technology continues to develop.

Also Read: Data Lakehouse, by *databricks* : <https://www.databricks.com/glossary/data-lakehouse>

Data Design Patterns

In the realm of data design pattern, we encounter three prominent strategies: Lambda, Kappa, and Multihop, each presenting unique processing methodologies.

Lambda architecture segregates processing into batch and real-time streams, allowing for comprehensive data handling as shown in Figure 2.7. Kappa architecture, concentrates solely on real-time processing, streamlining operations for efficiency as shown in Figure 2.8.

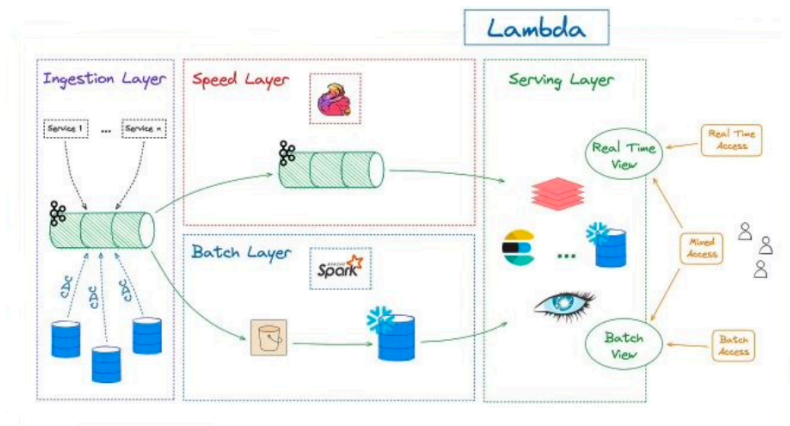
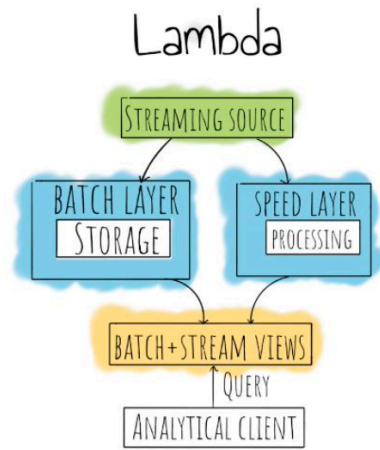


Figure 2.7: Lambda architecture

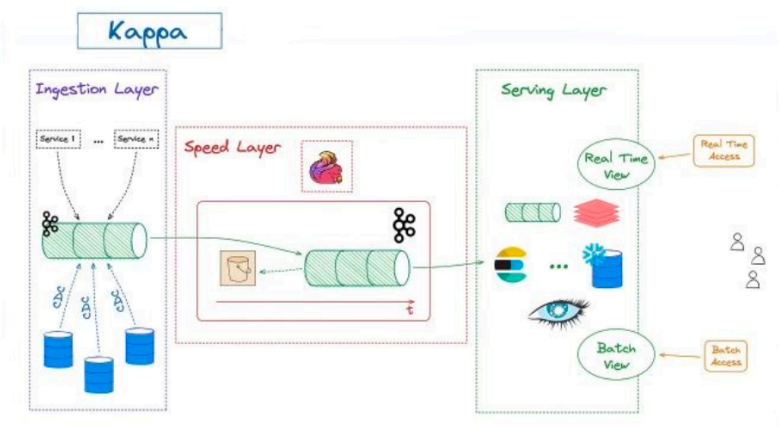
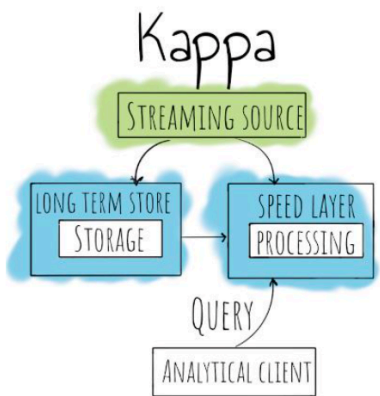


Figure 2.8: Kappa architecture

Multihop architecture Figure 2.9 emerges as a synthesis of Lambda and Kappa, effectively harnessing their strengths while mitigating their limitations. Excelling in both historical and real-time data scenarios, Multihop boasts automated conflict resolution for seamless integration. Opting for the Multihop approach signifies a deliberate choice for a versatile and efficient data architecture, placing it in the forefront of consideration. Transitioning to a more detailed exploration of the Multihop architecture, let's delve into its distinctive features and benefits that set it apart in the realm of data processing strategies.

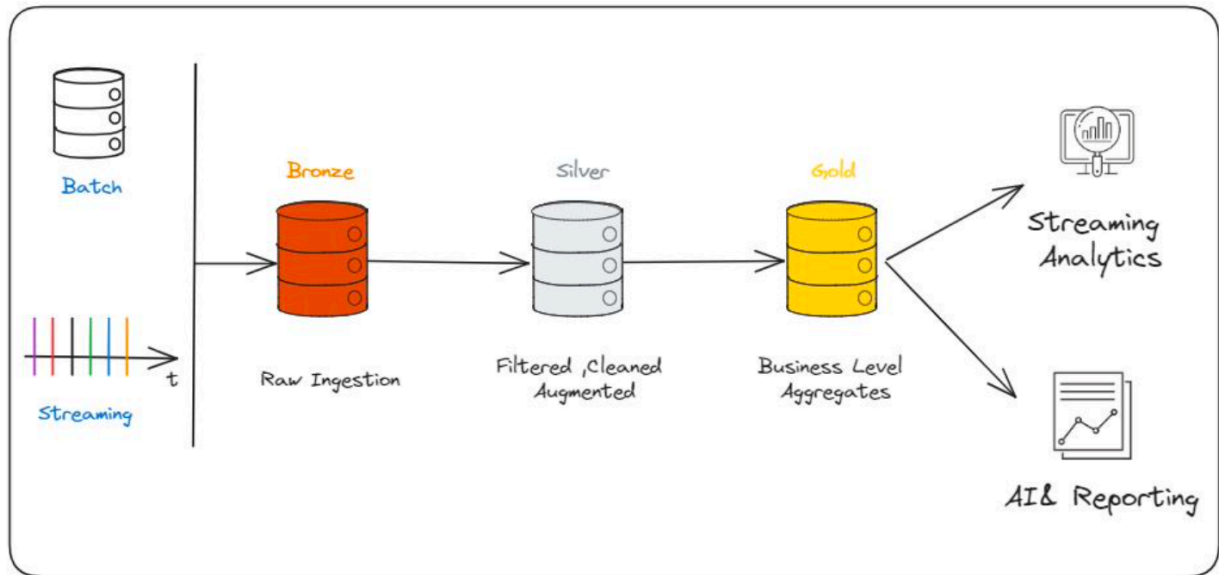


Figure 2.9: Medallion Architecture

Bronze layer

The Bronze phase acts as the primary entry point for data intake and retention. Information is preserved without alterations or conversions. This layer prioritises quick Change Data Capture, historical archiving of source data for reference (cold storage), data lineage tracing, auditing, and the provision for reprocessing without requiring re-reading of source data .

Silver layer

In the Silver stage, tables undergo cleansing, filtering, or adjustment to enhance their usability. It's important to emphasize that the modifications at this point should be minor, avoiding complex aggregations or enhancements. The Silver layer brings together data from diverse origins, enabling self-service analytics for rapid reporting, sophisticated analyses, and machine learning. It serves as a foundation for Departmental Analysts, Data Engineers, and Data Scientists to create solutions addressing business issues through both overall enterprise and specific departmental data endeavors, ultimately converging in the Gold Layer.

In the lakehouse data engineering concept, the Silver layer follows the ELT method, concentrating on minimal transformations and data purification during loading to ensure swift data lake integration and delivery. Regarding data modeling, the Silver Layer adopts data structures reminiscent of the 3rd-Normal Form or even employs Data Vault-like models optimized for efficient data writing.

Gold layer

In the Gold tier of the lakehouse architecture, information is organized into databases customized for specific projects, ensuring readiness. This stratum is optimized for reporting, employing streamlined and efficient data models to reduce the need for extensive connections. It serves as the stage for final data adjustments and quality enhancements. Within this layer, projects such as Customer Analytics, Product Quality Analytics, Inventory Analytics, and Customer Segmentation are housed. The Gold tier frequently integrates Kimball-style star schema-based data models or Inmon-style Data marts as part of the overall lakehouse framework.