Al-Mustaqbal University

College of Sciences

Intelligent Medical System Department

جامـــــعة المــــستـقـبـل

AL MUSTAQBAL UNIVERSITY

كلية العلوم

قـســــــــــم الانظمة الطبية الذكية

Lecture: (2)

Data Warehouse Architecture

Subject: Clinical Data Mining

Level: Four

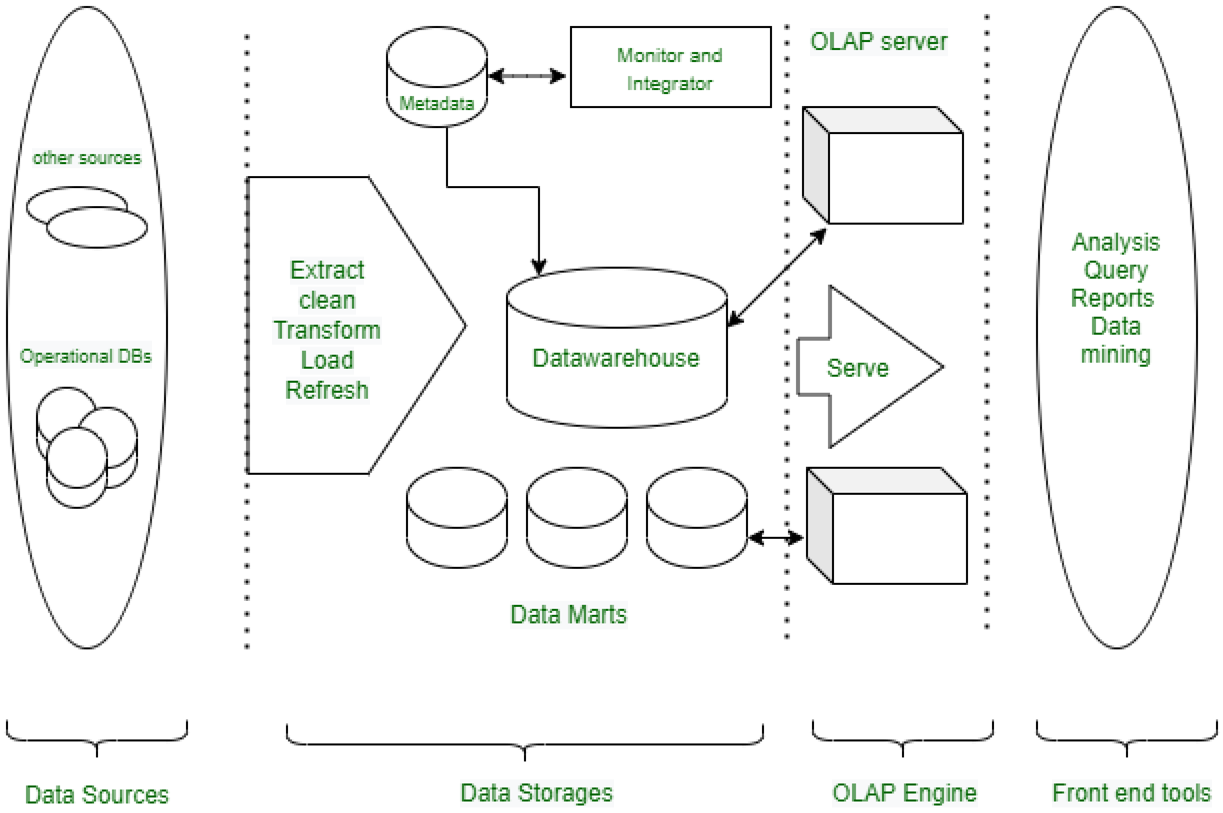
Lecturer: Dr. Maytham Nabeel Meqdad

Data Warehouse Architecture

Data warehousing helps businesses make informed decisions using large datasets.

The **Three-Tier Architecture** is widely used for its clear structure, dividing data processing into three layers for efficient access and management.

- Bottom Tier (Data Sources and Data Storage)
- Middle Tier (OLAP Engine)
- Top Tier (Front-End Tools)



Three/Multi-tier Architecture of Data Warehouse

Bottom Tier

Bottom Tier is the foundation of the data warehouse, responsible for collecting**,** processing**,** and storing data from multiple sources. It plays a critical role in preparing data for analysis.

Key Components:

- **Data Sources:**
  Includes **operational databases (OLTP systems)**, **flat files**, **spreadsheets**, **external APIs**, **CRM/ERP systems**, and **web logs**. These are the raw inputs feeding into the data warehouse.

- **Data Storage:**
  Processed data is stored in a Relational Database Management System **(RDBMS)** or a **multidimensional database** designed to support structured querying and analysis.

ETL Process (Extract, Transform, Load)

This is the core function of the bottom tier:

1. **Extract:**
   Gathers raw data from different, often incompatible sources.

2. **Transform:**
   Converts data into a consistent format, applying business rules, cleansing errors, handling missing values, and resolving duplicates.

3. **Load:**
   Loads the transformed data into the warehouse, organizing it for fast access and analysis.

This process ensures the warehouse contains clean, reliable, and business-ready data.

Common Challenges in Bottom Tier

Integrating data from diverse sources presents several challenges such as:

- **Data Quality:** Inconsistent data can lead to errors and unreliable analytics.

- **Data Compatibility:** Different data formats and structures can complicate integration.

- **Scalability:** Handling increasing volumes of data efficiently.

Solutions

- **Implement Robust ETL Tools:** Utilize powerful ETL tools like Informatica, Microsoft SSIS, or Confluent to streamline the data integration process.

- **Standardize Data Formats:** Standardizing data at the point of entry minimizes compatibility issues.

- **Continuous Data Quality Management:** Regularly check and clean data to maintain high quality.

- **Scalability Planning:** Design data storage solutions that can expand as data volume grows, ensuring that the architecture can handle future increases in data without performance degradation.

Middle Tier

The **Middle Tier** hosts the **OLAP server**, which processes complex analytical queries. It acts as a bridge between the **data storage layer (bottom tier)** and the **user interface (top tier)**, ensuring data is quickly retrieved, aggregated, and ready for reporting and analysis.

OLAP is a powerful technology for complex calculations, trend analysis, and data modeling. It is designed for high-speed analytical processing. OLAP server models come in three different categories, including:

- **ROLAP (Relational OLAP):** This model uses a relational database to store and manage warehouse data. It is ideal for handling large data volumes as it operates directly on relational databases.

- **MOLAP (Multidimensional OLAP):** This model stores data in a multidimensional cube. The storage and retrieval processes are highly efficient, making MOLAP suitable for complex analytical queries that require aggregation.

- **HOLAP (Hybrid OLAP):** It is combination of relational and multidimensional online analytical processing paradigms. HOLAP is the ideal option for a seamless functional flow across the database systems when the repository houses both the relational database management system and the multidimensional database management system.

Common Challenges in Middle Tier

- **Data Latency:** Delays in data availability can impact decision-making.
- **Query Performance:** Managing large volumes of data can slow down query performance.
- **Data Integration:** Combining data from different sources with varying formats can be challenging.

Solutions

- **Real-Time & Incremental Loading:**
  Update data frequently using real-time and incremental loading to reduce latency and support faster decision-making.
- **Query Optimization:**
  Improve performance with **indexing**, **partitioning**, and **optimized SQL** for faster data retrieval.
- **Standardization & Integration Tools:**
  Standardize data formats and use tools like **Talend** or **Informatica** for seamless integration and improved data quality.

Top Tier

The Top Tier in the Three-Tier Data Warehouse Architecture comprises the front-end client layer, which is essential for interacting with the data stored and processed in the lower tiers. This layer includes a variety of business intelligence (BI) tools and techniques designed to facilitate easy access and manipulation of data for reporting, analysis, and decision-making.

BI tools are critical components of the Top Tier, providing robust platforms through which users can query, report, and analyze data. Popular BI tools include:

- **IBM Cognos:** Offers comprehensive reporting capabilities.
- **Microsoft BI Platform:** Integrates well with existing Microsoft products, providing a familiar interface for users.
- **SAP BW:** Specializes in managing large datasets and integrating with other SAP products.

- **Crystal Reports:** Known for its powerful reporting features.
- **SAS Business Intelligence:** Provides advanced analytics.
- **Pentaho:** A versatile tool for data integration and visualization.

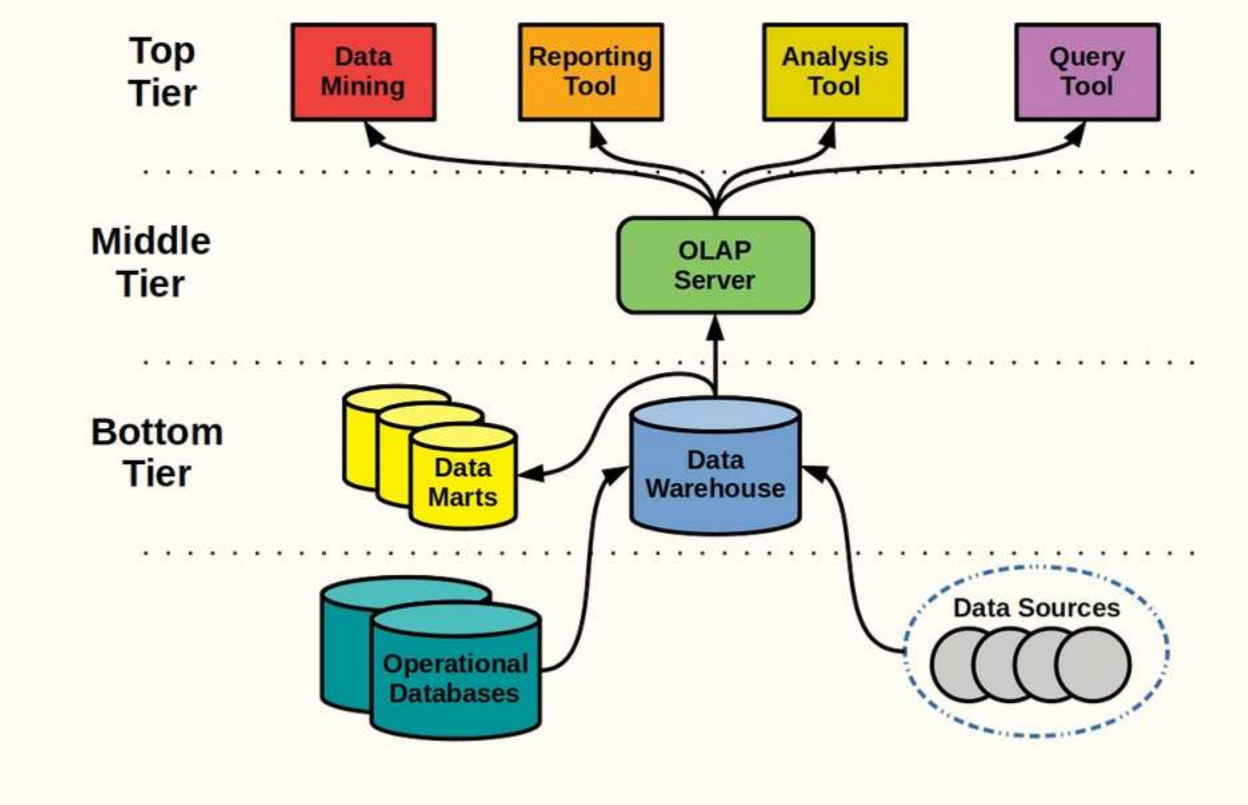
The Top Tier is crucial for decision-making as it provides the interface through which insights are accessed and explored. By presenting data in visual formats such as graphs, charts, and dashboards, these tools allow decision-makers to quickly grasp complex patterns, trends, and anomalies, leading to faster and more effective decision-making.

Common Challenges in Top Tier

- **Usability Issues:** Complex tools can hinder user adoption and effectiveness.
- **Integration Difficulties:** Ensuring seamless integration with other tiers can be challenging.

Solutions

- **User Training and Support:** Offering comprehensive training sessions to help users fully leverage the capabilities of BI tools.
- **Choosing Integrative Tools:** Select tools that easily integrate with existing systems in the data warehouse architecture, ensuring consistency and reliability in data handling.



What is 3-tier architecture in a data warehouse?

The 3-tier architecture in a data warehouse consists of three layers: the Bottom Tier (Data Sources and Data Storage), the Middle Tier (OLAP Engine), and the Top Tier (Front-End Tools).

What is the ETL process in a data warehouse?

ETL stands for Extract, Transform, and Load. It's a process where data is extracted from various sources, transformed for analysis, and loaded into a data warehouse.

What is the full form of OLAP?

OLAP stands for Online Analytical Processing, which is a category of software tools that enables analysis of data stored in a database.

What is the full form of KDD?

KDD stands for Knowledge Discovery in Databases, referring to the process of discovering useful knowledge from data.

References

[1] Digital Health and HealthcareQuality: A Primer on the Evolving4th Industrial RevolutionAhmed Umar Otokiti

[2] Oracle Help Center : https://docs.oracle.com › ... › Release 19

[3] Han and M. Kamber, “ Data Mining Tools and Technique s”, Morgan Kaufmann Publishers.

[4] .M.H. Dunham, “ Data Mining Introductory and Adv anced Topics”, Pear son Education

[5] Geeks for geeks https://www.geeksforgeeks.org/dbms/multi-tier-architecture-of-data-warehous

[6] Software sim https://softwaresim.com/blog/enterprise-data-warehouses-as-a-source-of-data-for-simulation