DMDW Unit 1

1. Define data warehouse . explain why a data wareshouse is seperated from operational database

Definition of Data Warehouse

- A **data warehouse** is a specialized database designed to store, integrate, and manage large volumes of historical data from multiple sources. The main goal is to enable complex queries, analytics, and reporting for decision-making.
- Data warehouses use a different structure (often optimized for read access and analysis) than the real-time transactional systems, supporting business intelligence tasks such as data mining, trend analysis, and forecasting.
- They maintain data over long time periods, transforming raw operational data into structured information for strategic analysis.

Reasons for Separating Data Warehouse from Operational Database

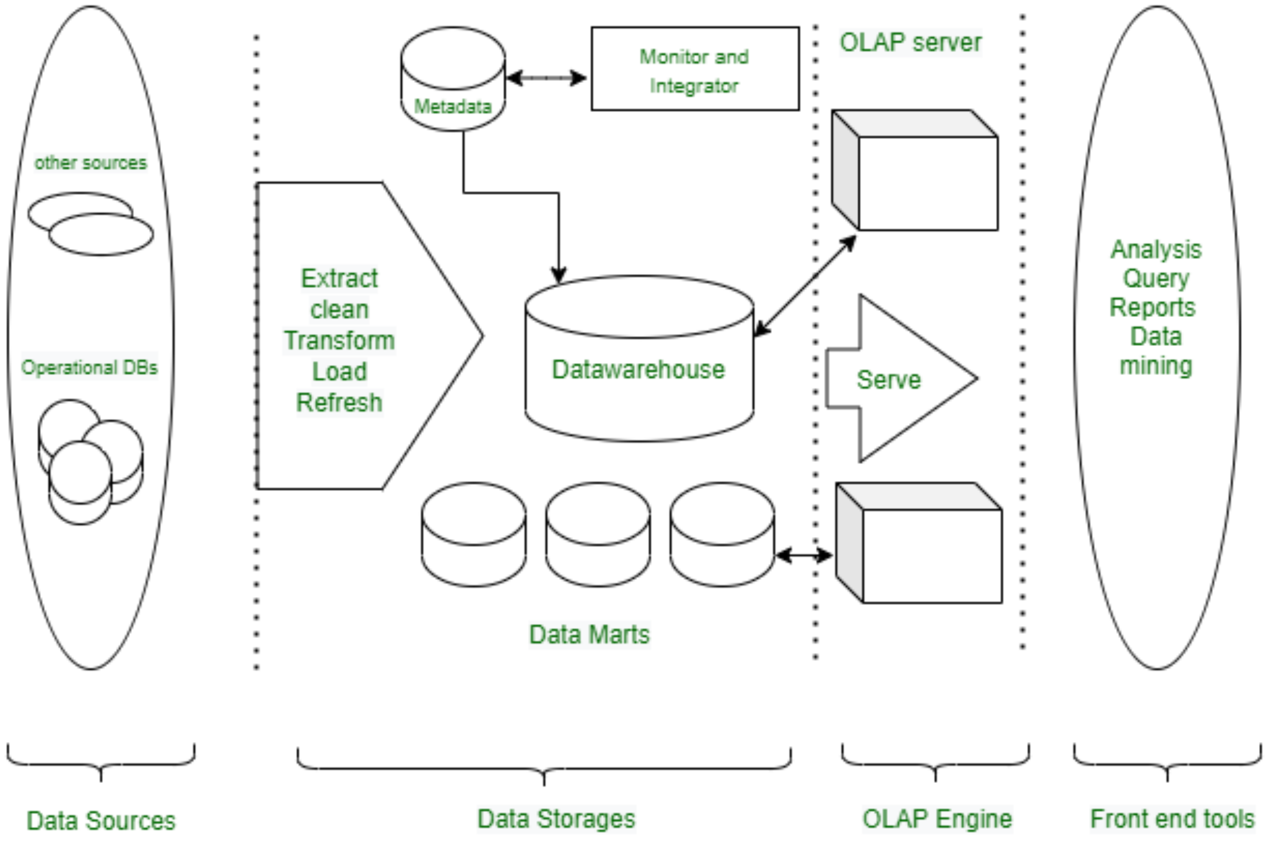
- **Performance Optimization:** Operational databases handle rapid transactions, while data warehouses manage heavy analytical queries; keeping them apart prevents slowdowns for both systems.
- **Data Consistency and Integrity:** Separation ensures smooth business operations, as day-to-day transactions remain unaffected by data restructuring and aggregation happening in the warehouse.
- **Historical Data Management:** Data warehouses store extensive historical records, unlike operational databases that only need current data, ensuring operational systems stay lean and efficient.
- **Security and Access Control:** Isolating the warehouse allows analysts wider access to information without putting live, critical operational data at risk.
- **Data Integration and Quality:** Combining and cleaning data from many sources is easier and safer in a dedicated data warehouse, instead of disrupting operational databases.

2. Differentiate between OLAP and OLTP

Feature	OLTP	OLAP
Primary Purpose	Transaction processing	Data analysis and reporting
Data Volume	Smaller, constantly changing	Larger, historical data
Data Model	Normalized relational	Multidimensional (cubes) or columnar
Read/Write Operations	Balanced	Read-intensive
Performance Focus	Speed of individual transactions	Speed of complex queries
Data Integrity	High priority	Less strict, focus on analysis
Typical Users	Frontline staff, customers	Data analysts, managers

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc-created or done for a particular purpose as necessary.
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

3. Explain and draw Architecture of Data Warehouse



Data Warehouse Architecture Explained

A **data warehouse architecture** is the structural design that organizes how data is collected, stored, managed, and accessed for analytical and business intelligence purposes. The most common and practical model is the **three-tier architecture**, described below:

1. Bottom Tier: Data Source Layer

- **Description:** This is where data is extracted from multiple sources including operational databases, legacy systems, cloud storage, and external resources.
- **Function:** Data is loaded into a staging area where it is cleansed, transformed, and integrated before entering the central data warehouse.
- **Tools:** ETL (Extract, Transform, Load) tools perform data integration and cleaning at this level[123].

2. Middle Tier: Data Storage and Processing Layer

- **Description:** Known as the data warehouse database or OLAP server layer.
- **Function:** Stores vast volumes of historical and current data, supports complex analytical queries, and structures data for fast, multidimensional analysis.
- **Technology:** Can use relational (ROLAP) or multidimensional (MOLAP) OLAP models, enabling users to perform fast analytical operations like slicing, dicing, and drilling into data[234].

3. Top Tier: Presentation and Access Layer

- **Description:** The front-end that users interact with.
- **Function:** Provides data access through dashboards, reporting tools, data mining, and analytics applications. This layer enables business users and analysts to generate reports and derive insights easily.
- **Tools:** Query, analysis, and data mining tools[34].

5. Explain Meta Data?

- **Definition:** Metadata is "data about data"—it describes the content, structure, source, format, and meaning of the actual data stored in the warehouse, acting as a roadmap or directory for users and systems135.
- **Categories:** Metadata is typically organized into **business metadata** (definitions, ownership, policies), **technical metadata** (table/column names, data types, constraints), and **operational metadata** (data currency, lineage, transformation history)356.
- **Functions:** It enables data cataloging (finding and understanding data), lineage tracking (showing data origin and transformations), governance (ensuring data quality and security), and discovery (helping users quickly locate relevant datasets)26.
- **Importance:** Metadata improves data usability, supports data integration and interoperability, ensures consistency, and is critical for data analysis, reporting, and decision-making in large, complex environments149.
- **Management:** Metadata is stored in repositories and managed using standards and tools, helping administrators maintain, secure, and govern the data warehouse effectively17.

Data Warehouse Model Types

A data warehouse can be structured and implemented with different architectural models, each serving distinct business analytics, reporting, and integration needs. The three widely recognized types are: **Enterprise Data Warehouse (EDW)**, **Operational Data Store (ODS)**, and **Data Mart**.

1. Enterprise Data Warehouse (EDW)

- **Definition:**
  An EDW is a centralized, comprehensive data repository that stores and manages historical business data from across an organization123.
- **Key Features:**
  - Integrates data from multiple departments, applications, and systems to provide a holistic organizational view.
  - Supports advanced analytics, business intelligence, and decision-making for the entire enterprise.
  - Employs ETL/ELT processes for data extraction, cleansing, transformation, and loading before making data available for analysis.
- **Typical Architecture:**
  - Three-tier structure: (1) Data sources and staging, (2) central data warehouse, (3) presentation/access layer with analytics tools.

    - Stores structured, semi-structured, and sometimes unstructured data.
- **Benefits:**
    - Ensures a “single source of truth” for all business data.
    - Enhances data consistency, accuracy, and supports regulatory compliance.
- **Example Use Case:**
    - An international retailer consolidating global sales, inventory, and customer data for executive analysis[1435].

2. Operational Data Store (ODS)

- **Definition:**
  An ODS is a centralized database that integrates current, real-time operational data from various transactional systems to support immediate reporting and monitoring needs[6789].
- **Key Features:**
    - Collects and stores data in its raw or minimally transformed state, focusing on up-to-date operational information.
    - Provides a near real-time, consolidated view for tactical, short-term decisions.
    - Data in an ODS is frequently refreshed and reflects the latest state of business operations, but typically does not contain extensive historical data.
- **Role in Data Architecture:**
    - Acts as a bridge between transactional systems (OLTP) and analytical systems (EDW).
    - Often serves as a staging area before data is moved to the EDW.
- **Benefits:**
    - Enables operational reporting and timely queries without impacting core transactional systems.
    - Supports fast, detailed data analysis for immediate business activities.
- **Example Use Case:**
    - A bank using ODS to aggregate real-time transactions from ATMs, online banking, and branch systems to monitor account activities[678].

3. Data Mart

- **Definition:**
  A data mart is a specialized, subject-oriented subset of a data warehouse, focused on a specific business unit, department, or function (e.g., sales, finance, marketing)[10111213].

- **Key Features:**
    - Contains data relevant only to its focus area, offering faster access and simplified analytics for specific teams.
    - Can be created independently (directly from source systems) or dependently (extracting from an EDW).
    - Supports tailored reports, dashboards, and analyses needed by business units without accessing the broader data warehouse.
- **Types:**
    - Dependent Data Mart: Sourced from an EDW for consistency.
    - Independent Data Mart: Sourced directly from operational or other systems.
- **Benefits:**
    - Quicker and more cost-effective to set up than an enterprise-wide warehouse.
    - Reduces complexity for end-users needing focused data.
- **Example Use Case:**
    - The marketing department maintaining a data mart with customer campaign response and sales data for targeted analytics101213.

6. what are various access tools used in data warehouse ? explain

Certainly! Here's a clear, detailed explanation of each step in the data warehousing process, expanded with purpose, methods, and significance:

Data Extraction

- **Definition:** The process of gathering data from various, often diverse, sources.
- **Sources:** Can include relational databases, flat files, APIs, cloud services, legacy systems, and external partners.
- **Methods:** Extraction can be **full** (all data at once), **incremental** (only new/changed data), or **real-time** (streaming). Data is moved to a staging area for further processing.
- **Purpose:** Ensures that all relevant data required for analysis is collected from across the organization and beyond, regardless of the source format or location.
- **Challenge:** Different sources may use different formats, schemas, or platforms, requiring flexible extraction mechanisms.

Data Cleaning

- **Definition:** The process of identifying and correcting errors, inconsistencies, and missing values in the extracted data.

- **Activities:** Includes deduplication, correcting typos, converting units, standardizing values (e.g., dates, currencies), handling missing data, and validating data ranges.

- **Purpose:** Improves data quality, making the data accurate, reliable, and suitable for analysis.

- **Significance:** Clean data is crucial for accurate reporting, analytics, and business decision-making. Poor data quality can lead to misleading insights and poor decisions.

Data Transformation

- **Definition:** Converting raw, often messy, data into a compatible and optimized format for the data warehouse.

- **Activities:** Involves restructuring, aggregating, sorting, applying business rules, and mapping source data fields to target warehouse fields.

- **Purpose:** Ensures that data from disparate sources is unified, consistent, and ready for meaningful analysis.

- **Example:** Converting customer birthdates (stored as strings in one system and integers in another) into a standardized date format in the warehouse.

Load

- **Definition:** Inserting the cleaned and transformed data into the data warehouse.

- **Activities:** Sorting, summarizing, consolidating, computing derived views, enforcing integrity constraints, and building indexes or partitions for faster query access.

- **Purpose:** Makes data available to end-users in an optimized and structured format, supporting efficient querying and reporting.

- **Challenge:** Large datasets may require careful scheduling and resource management to avoid overloading systems.

Refresh

- **Definition:** Regularly updating the warehouse with new or changed data from the source systems.

- **Frequency:** Can be scheduled (daily, weekly) or event-driven (real-time), depending on business needs.

- **Activities:** Extracting and propagating updates, additions, and deletions from operational systems to maintain the currency of the warehouse.

- **Purpose:** Keeps the warehouse up-to-date, ensuring that analytics and reports reflect the latest business reality.

- **Significance:** Critical for time-sensitive decision-making and operational reporting.