Chapter 2

Data Warehousing

What is a Data Warehouse?

- A **central storage system** holding organized, cleansed data in a standard format for the entire organization.
- **Key Features** (Inmon's Definition):
  - **Integrated**: Combines data from various sources (e.g., sales, inventory).
  - **Subject-Oriented**: Focused on specific areas (e.g., customer data).
  - **Non-Volatile**: Data doesn't change once stored, persistent (e.g., historical records).
  - **Time-Variant**: Tracks data over time (e.g., sales trends).
- **Purpose**: Supports decision-making (DSS) by providing reliable data for analysis.

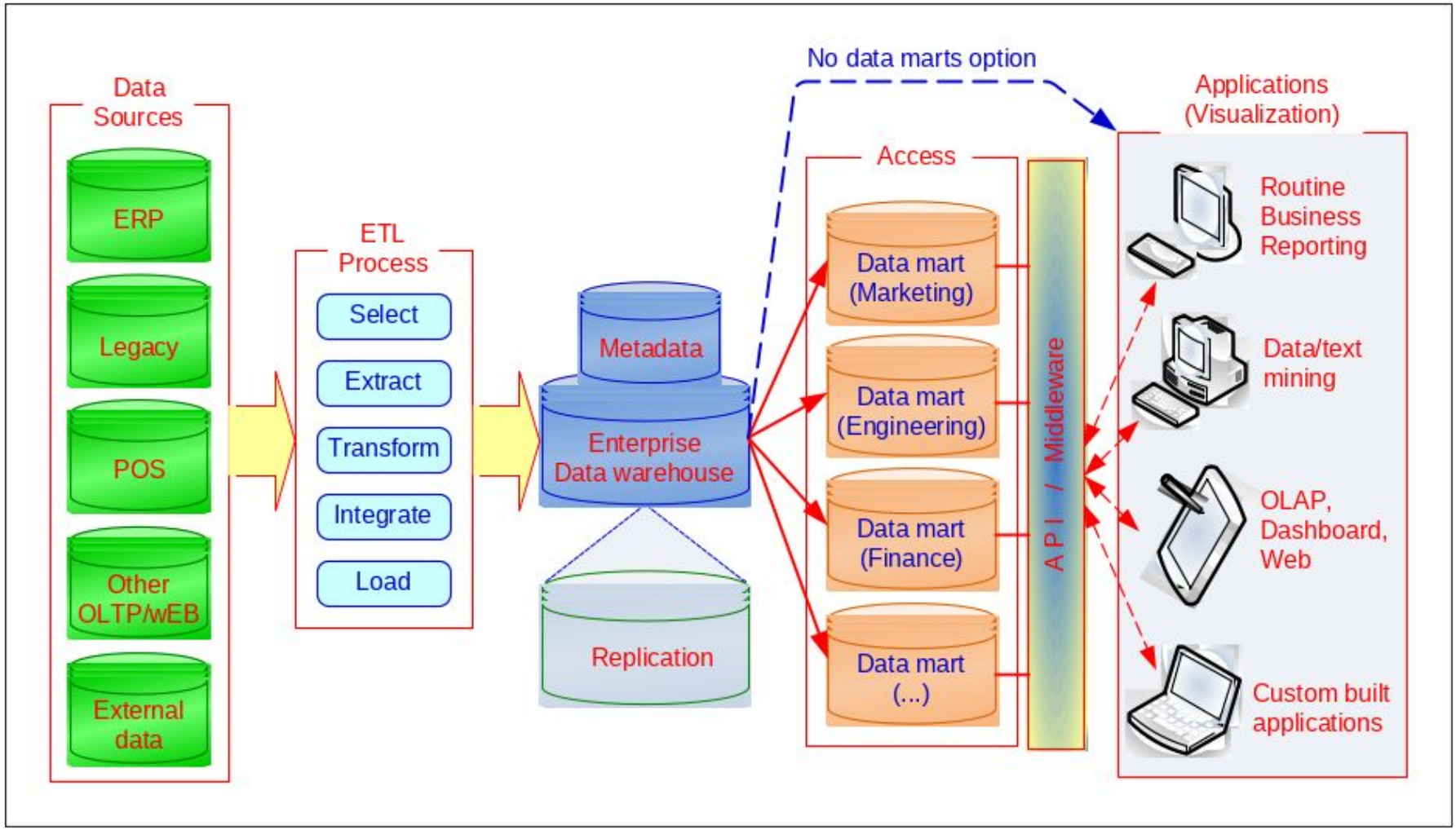
What is the place where unorganized data is stored? - Data

Characteristics of Data Warehouses (DWs)

- **Subject-Oriented**: Focused on business areas (e.g., sales, customers).
- **Integrated**: Combines data from multiple sources in a unified format.
- **Time-Variant**: Stores historical data as a time series (e.g., sales over years).
- **Nonvolatile**: Data doesn't change once stored (e.g., historical records stay fixed).
- **Summarized**: Provides aggregated data for analysis (e.g., total monthly sales).
- **Not Normalized**: Optimized for querying, not updates (unlike operational databases).
- **Metadata-Driven**: Includes data about the data (e.g., data source, format).
- **Technology Features**:
  - Web-based, relational, or multi-dimensional (e.g., OLAP cubes).
  - Client/server architecture, supports real-time or near real-time access.

Data Mart

- A **smaller-scale data warehouse** for a specific department, holding only relevant data.
- **Types**:
  - **Dependent Data Mart**: A subset pulled directly from the main data warehouse (e.g., sales data for the marketing team).
  - **Independent Data Mart**: A standalone mini data warehouse for a department or business unit (e.g., finance department's own data mart).
- **Purpose**: Provides focused, faster access to data for specific business needs.

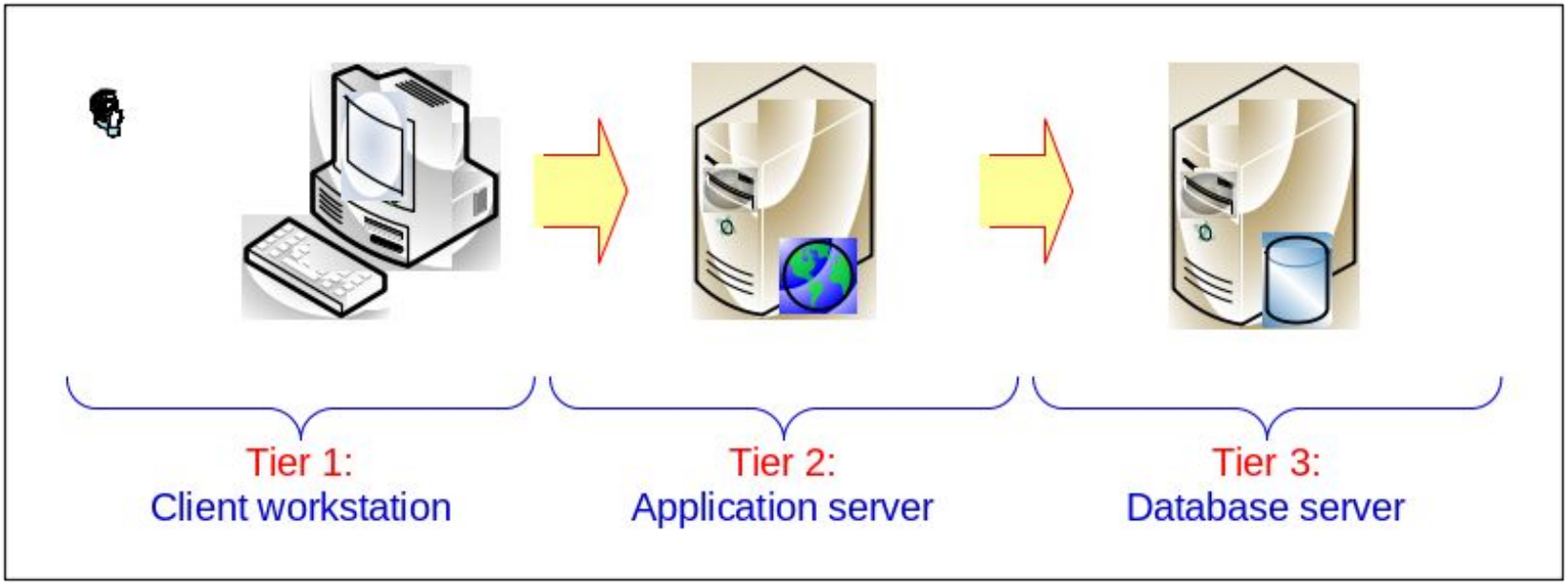

Data Sources
ERP
Legacy
POS
Other OLTP/wEB
External data
ETL Process
Select
Extract
Transform
Integrate
Load
Metadata
Enterprise Data warehouse
Replication
No data marts option
Access
Data mart (Marketing)
Data mart (Engineering)
Data mart (Finance)
Data mart (...)
A P I / Middleware
Applications (Visualization)
Routine Business Reporting
Data/text mining
OLAP, Dashboard, Web
Custom built applications

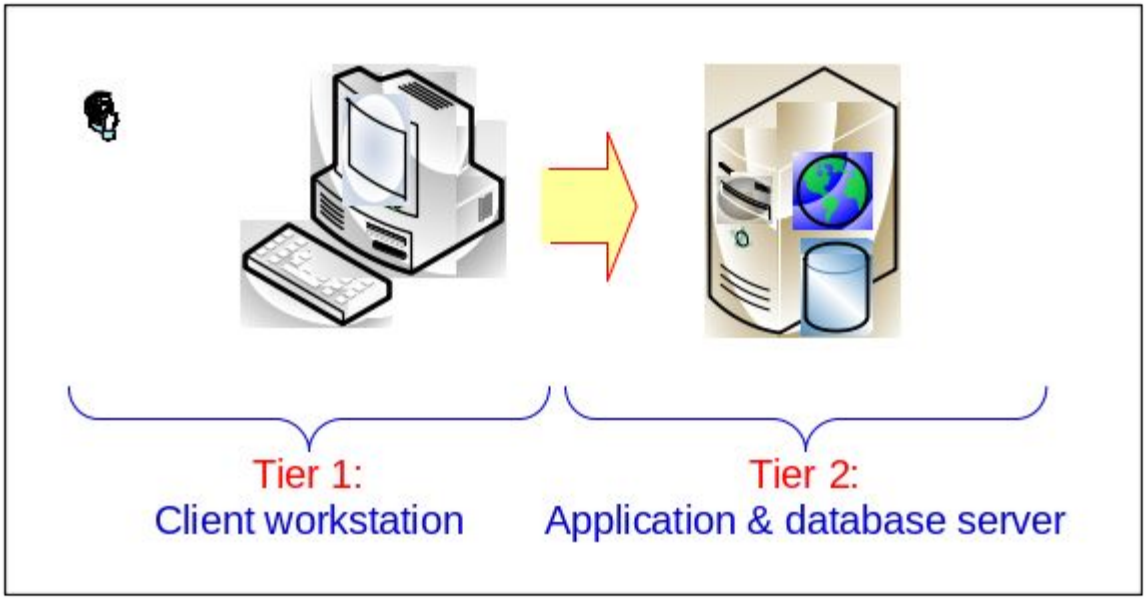
A Generic DW Framework

- **Data Sources**: Data comes from systems like ERP, POS, legacy systems, and external sources (e.g., web data).
- **ETL Process**: Extracts, transforms, and loads data:
  - **Extract**: Pulls data from sources.
  - **Transform**: Cleans and integrates data.
  - **Load**: Stores data into the warehouse.
- **Enterprise Data Warehouse**: Central storage with metadata and replication for reliability.
- **Data Marts (Optional)**: Smaller, department-specific stores (e.g., marketing, finance) accessed via APIs or direct access.
- **Applications**: Tools like dashboards, OLAP, data mining, and custom apps use the data for insights.

DW Architecture

- **Three-Tier Architecture**:
  - **Back-End (Data Acquisition)**: Software that extracts and loads data (e.g., ETL tools).
  - **Middle Tier (Data Warehouse)**: Stores the data and manages it with software.
  - **Front-End (Client)**: Tools for users to access and analyze data (e.g., dashboards, reports).
- **Two-Tier Architecture**:
  - Combines back-end and data warehouse into one tier.
  - Front-end remains separate for user access.
- **Single-Tier (Rare)**: Everything in one tier—data and tools together (less common).

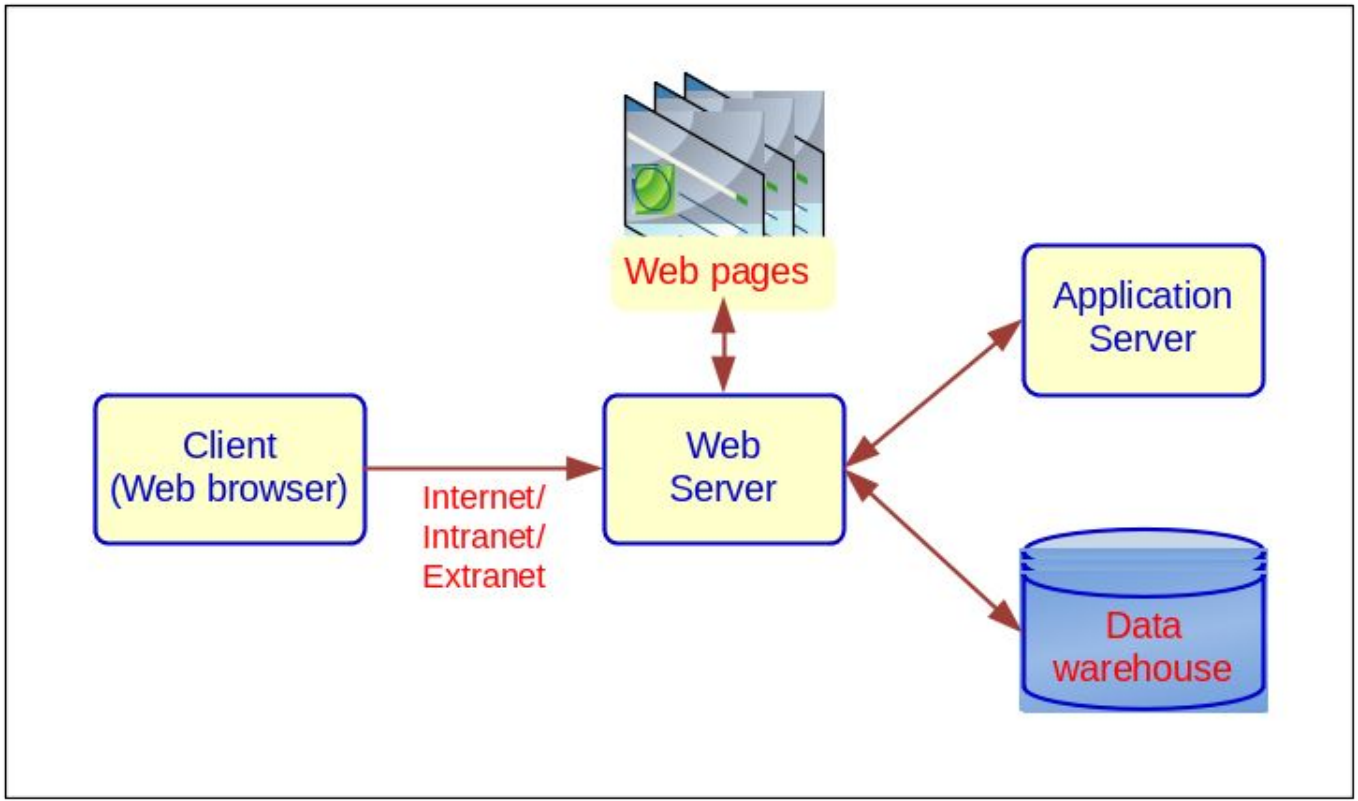

Tier 1:
Client workstation
Tier 2:
Application server
Tier 3:
Database server


Tier 1:
Client workstation
Tier 2:
Application & database server

Data Warehousing Architectures - Issues to Consider

- **Key Decisions**:
  - **Database Management System (DBMS)**: Which system to use for storing data (e.g., Oracle, SQL Server)?
  - **Parallel Processing/Partitioning**: Will data be split or processed simultaneously for speed?
  - **Data Migration Tools**: Will tools be used to move and load data into the warehouse?
  - **Analysis Tools**: Which tools will help users retrieve and analyze data (e.g., Tableau, Power BI)?
- **Why It Matters**: These choices affect performance, cost, and how well the warehouse supports

A Web-Based DW Architecture



A Web-Based DW Architecture

- **Structure**:
  - **Data Sources**: Collect data from systems (e.g., ERP, POS, web data).
  - **ETL Process**: Extracts, transforms, and loads data into the warehouse.
  - **Data Warehouse**: Central storage with data and management software.
  - **Web Layer**: Acts as the middleman, connecting users to the warehouse via the internet.
  - **Client (Web Browser)**: Users access and analyze data through a browser (e.g., dashboards on Chrome).
- **Benefits**:
  - Easy access from anywhere using a browser.
  - Supports real-time or near real-time data updates.

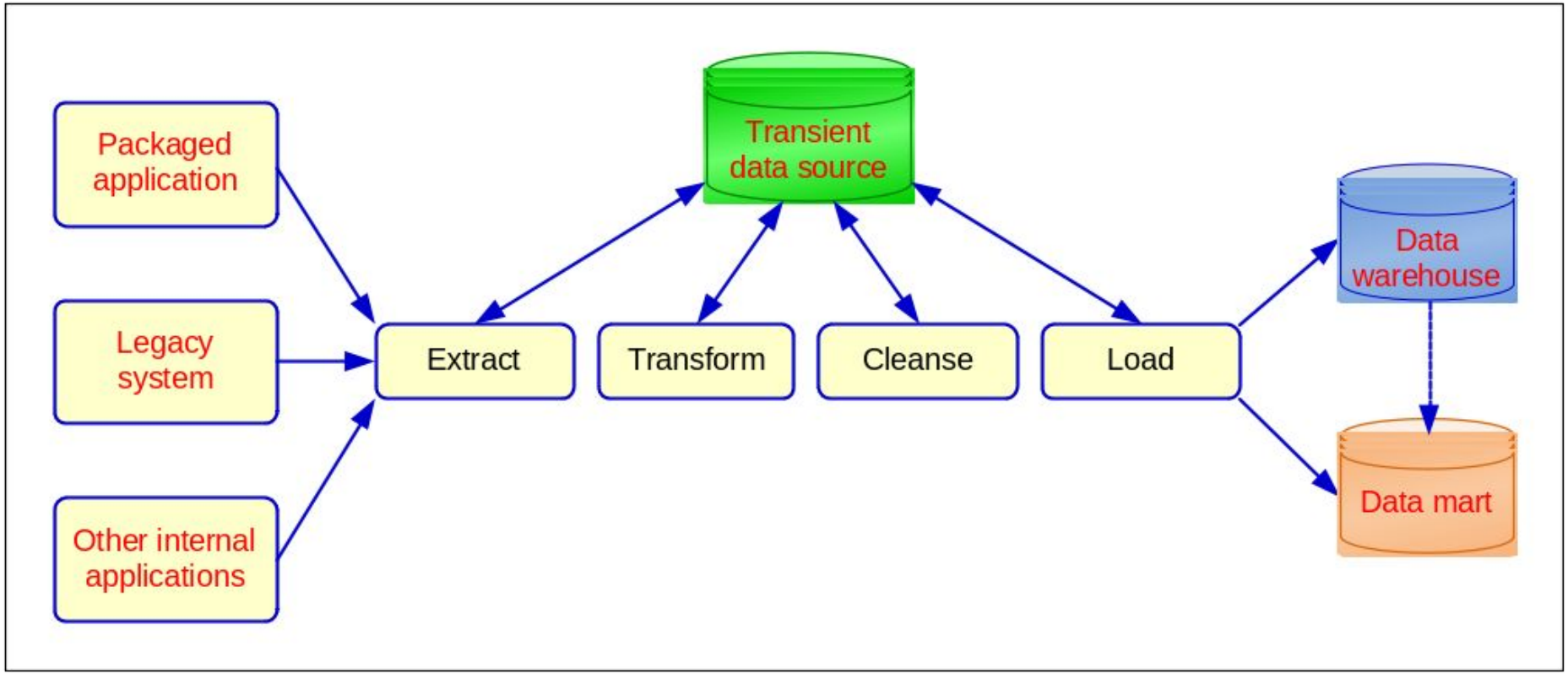
Ten Factors That Affect DW Architecture Selection

1. **Information Sharing Between Units**: How much do departments need to share data?
2. **Upper Management Needs**: What information do executives need (e.g., strategic reports)?
3. **Urgency**: How quickly is the data warehouse needed (e.g., tight deadlines)?
4. **End-User Tasks**: What will users do with the data (e.g., reports, analytics)?
5. **Resource Constraints**: What's the budget, time, or staff availability?
6. **Strategic Vision**: How does the warehouse fit into long-term goals?
7. **Compatibility**: Does it work with existing systems (e.g., ERP)?
8. **IT Staff Skills**: Can the in-house team handle the setup and maintenance?
9. **Technical Issues**: Are there hardware, software, or scalability concerns?
10. **Social/Political Factors**: Are there internal politics or stakeholder conflicts?

Data Integration and the ETL Process

- **ETL Defined**: Extract, Transform, Load—moves and prepares data for the warehouse.
- **Data Integration**:
  - Involves three steps:
    Access: Retrieves data from multiple systems.
  - **Federation**: Combines data into a unified view.
  - **Change Capture**: Tracks and updates data changes.
- **Enterprise Application Integration (EAI)**:
  - A technology to streamline data transfer from source systems (e.g., CRM, ERP) to the warehouse. (Customer relationship management)
- **Enterprise Information Integration (EII)**:
  - A toolset for real-time data integration from diverse sources (e.g., databases, web services).

Data Integration and the Extraction, Transformation, and Load Process



Data Integration and the ETL Process

- **Flow**:
  - **Data Sources**: Includes packaged apps, legacy systems, and other internal applications (transient data source).
  - **ETL Steps**:
    - **Extract**: Pulls data from sources.
    - **Transform**: Processes and organizes data.
    - **Cleanse**: Removes errors and inconsistencies.
    - **Load**: Stores data into the data warehouse and data marts.
- **Outcome**: Clean, integrated data in the data warehouse and department-specific data marts.

ETL (Extract, Transform, Load)

- **Challenges in Buying an ETL Tool**:
  - **Cost**: Data transformation tools can be expensive.
  - **Learning Curve**: Tools often take time to learn.
- **Key Criteria for Choosing an ETL Tool**:
  - **Versatility**: Must handle various data sources (e.g., ERP, web) and architectures (e.g., relational, cloud).
  - **Metadata Management**: Automatically captures metadata (e.g., data source, format) for transparency.
  - **Open Standards**: Ensures compatibility with other systems by following industry standards.
  - **User-Friendly**: Interface should be simple for developers (technical) and functional users (business).

Additional DW Considerations - Hosted Data Warehouses

- **Benefits**:
  - **Minimal Infrastructure Investment**: No need to buy or maintain servers; hosted by a provider.
  - **Frees Up In-House Systems**: Reduces the load on internal IT resources.
  - **Improves Cash Flow**: Avoids large upfront costs, paying as you go.
  - **Affordable Powerful Tools**: Access advanced features without high expenses.
  - **Supports Growth**: Scales easily as data or business needs grow.
  - **High-Quality Equipment/Software**: Providers offer top-tier tech and updates.
  - **Faster Connections**: Leverages provider's high-speed internet

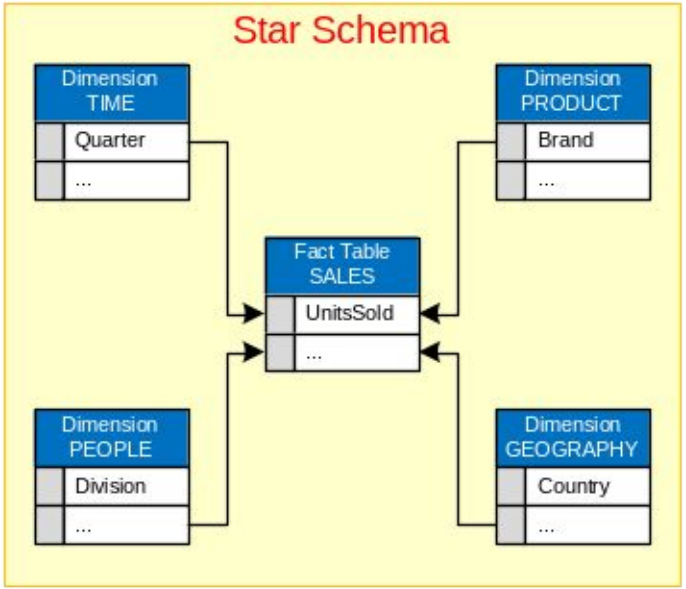
Representation of Data in DW

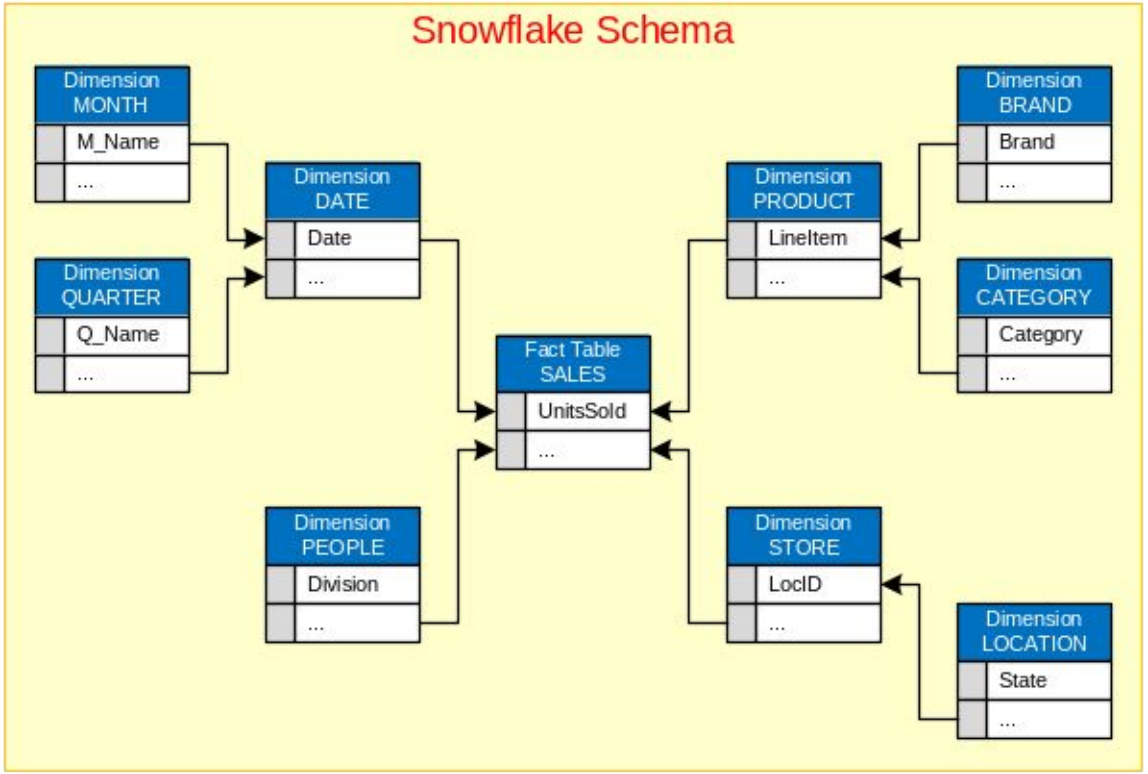
- **Dimensional Modeling**:
  - A system designed for fast, high-volume query access.
- **Star Schema**:
  - The simplest and most common dimensional model.
  - Features a central fact table (e.g., sales data) connected to dimension tables (e.g., time, product).
- **Snowflake Schema**:
  - An extension of the star schema, with more layers resembling a snowflake shape.

Multidimensionality

- **Definition**:
  - Ability to organize, present, and analyze data across multiple dimensions (e.g., sales by region, product, salesperson, and time).
- **Multidimensional Presentation**:
  - **Dimensions**: Products, salespeople, market segments, business units, locations, channels, countries, industries.
  - **Measures**: Money, sales volume, head count, inventory, profit, actual vs. forecast.
  - **Time**: Daily, weekly, monthly, quarterly, yearly.

Star versus Snowflake Schema





Star versus Snowflake Schema

- **Star Schema**:
  - Central fact table (e.g., sales) connected to single-layer dimension tables (e.g., product, time).
  - Simpler, faster for queries, less storage.
- **Snowflake Schema**:
  - Fact table connected to multi-layered dimension tables (e.g., time → year → month).
  - More complex, normalized, saves storage, slower queries.
- **Comparison**:
  - Star: Easier to use, faster for BI analytics.
  - Snowflake: More detailed, better for complex hierarchies.

Analysis of Data in DW - OLTP vs. OLAP

- **OLTP (Online Transaction Processing)**:
  - Captures and stores data from systems like ERP, CRM, POS.
  - Focus: Efficiency for routine tasks (e.g., recording sales).
- **OLAP (Online Analytical Processing)**:
  - Turns data into insights for decision-making.
  - Uses data cubes, drill-down/rollup, slice & dice.
  - Supports ad hoc reports, statistical analysis, multimedia apps.
- **Key Difference**:
  - OLTP: Day-to-day operations.
  - OLAP: Strategic analysis for BI.

OLAP vs. OLTP

TABLE 3.5 A Comparison Between OLTP and OLAP

Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or data mart (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational databases	Multiprocessor, large-capacity, specialized databases
Execution speed	Fast (recording of business transactions and routine reports)	Slow (resource intensive, complex, large-scale queries)

OLAP Operations

- **Slice**: Selects a subset of data (e.g., sales for one month).
- **Dice**: Selects data across multiple dimensions (e.g., sales for one month and region).
- **Drill Down/Up**: Moves between summarized (up) and detailed (down) data levels (e.g., yearly to daily sales).
- **Roll Up**: Summarizes data across dimensions (e.g., total sales by product).
- **Pivot**: Rotates the view to change data orientation (e.g., switch rows and columns in a report).

Variations of OLAP

- **Multidimensional OLAP (MOLAP)**:
  - Uses a specialized multidimensional database.
  - Pre-summarizes data into multidimensional views for fast queries.
- **Relational OLAP (ROLAP)**:
  - Builds OLAP on top of a relational database.
  - Queries data directly from relational tables.
- **Other Variations**:
  - **DOLAP**: Desktop OLAP (local analysis on user devices).
  - **WOLAP**: Web OLAP (browser-based analysis).

DW Implementation Issues

- Identifying data sources and setting data governance.
- Planning for data quality and designing the data model.
- Selecting the right ETL tool.
- Establishing service-level agreements (SLAs) for performance.
- Managing data transport and conversion.
- Reconciling data inconsistencies.
- Supporting end-users.
- Addressing political challenges (e.g., team conflicts).

Failure Factors in DW Projects

- Lack of executive support from top leaders.
- Unclear goals for the business.
- Ignoring cultural differences or resistance.
- Poor change management during the project.
- Setting unrealistic expectations for results.
- Choosing the wrong architecture design.
- Low data quality or missing key information.
- Loading unnecessary data just because it’s available.

Massive DW and Scalability

- **Scalability**: The ability to handle growing demands efficiently.
- **Key Issues**:
  - Volume of data in the warehouse.
  - Speed of warehouse growth.
  - Number of users accessing it at once.
  - Complexity of user queries.
- **Good Scalability**: Query performance grows linearly with warehouse size.

Real-Time/Active DW/BI

- **Overview**: Updates data in real-time for instant analysis and decisions—growing fast.
- **Push vs. Pull**:
  - Push: Data sent automatically to the warehouse.
  - Pull: Warehouse requests data when needed.
- **Concerns**:
  - Not all data needs constant updates.
  - Reports minutes apart may differ.
  - Can be expensive.
  - May not always be practical.

DW Administration and Security

- **Data Warehouse Administrator (DWA)**:
  - Needs expertise in high-performance software, hardware, and networking.
  - Requires strong business knowledge and insight.
  - Should understand decision-making to design and maintain the warehouse.
  - Must have great communication skills.
- **Security and Privacy**:
  - Protects valuable data assets.
  - Follows government rules (e.g., HIPAA).
  - Requires careful planning and execution.