Introduction to Data Warehouse & Data Mining

Tanzim Hossain
Lecturer
Department of Software Engineering
Daffodil International University

Why Data Warehouse?

Business Questions

- Imagine a corporate executive of a national electronics retailer asking the question, "**What retail stores were the top producers during the past 12 months in the Rocky Mountain region?**"

Business Questions

- Follow-up questions may include, "**What were the most profitable products in the top producing stores?"** and "**What were the most successful product promotions at the top producing stores?"**
- These are examples of typical decision support or business intelligence questions, asked every day by managers all over the world.

Shortcomings of SQL

- Answers to these questions often require complex SQL statements that may take hours to code and execute.
- Further more, formulating some of these queries may require data from a diverse set of internal legacy systems and external market sources, involving both relational and nonrelational databases.

Need for New Technology

- Decision-making questions such as those above pose new requirements on DBMSs.
- Data warehouse technology complements and extends relational database technology beyond online transaction processing and simple query capabilities such as the GROUP BY clause in SQL.

Transaction Processing Vs. Decision Support System

Transaction Processing

- Transaction processing allows organizations to conduct daily business in an efficient manner.
- Operational or production databases used in transaction processing assist with decisions such as tracking orders, resolving customer complaints, and staffing requirements.
- These decisions involve detailed data about business processes.,

Decision Support System

- Decision support processing helps management provide medium-term and long-term direction for an organization.
- Management needs support for decisions about capacity planning, product development, store location, product promotion, and other needs.

Different Question, Different Needs

- Historically, most organizations have assumed that operational databases can provide the data for decision support.
- As organizations have developed operational databases for various functions, an information gap has developed.
- Gradually, organizations have realized that the operational databases must be significantly transformed for decision support.

Different Question, Different Needs

- Operational databases must be transformed for decision support.
- Operational databases can contain inconsistencies in areas such as formats, entity identification, and units of measure that hamper usage in decision support.
- In addition, decision support needs a broad view that integrates business processes.

Different Question, Different Needs

- Because of the different requirements, operational databases are usually separate from databases for decision support.
- Using a common database for both kinds of pro cessing can significantly degrade performance and make it difficult to summarize activity across business processes.

Characteristics of Data Warehouse

What is a Data Warehouse?

- Data warehouse, a term coined by William Inmon in 1990, refers to a **central data repository** where data from operational databases and other sources are **integrated**, **cleaned**, and **standardized** to support decision making.
- The transformational activities (cleaning, integrating, and standardizing) are essential for achieving benefits.

What are the Benefits?

- Tangible benefits from a data warehouse can include increased revenue and reduced expenses enabled by business analysis that was not possible before the data warehouse was deployed.
- For example, a data warehouse may enable reduced losses due to improved fraud detection, improved customer retention through targeted marketing, and reduction in inventory carrying costs through improved demand forecasting.

Characteristics of Data Warehouse

- **Subject-Oriented:** A data warehouse is organized around the major business subjects or entities such as customers, orders, and products.
- This subject orientation contrasts to the more process orientation for transaction processing.

Characteristics of Data Warehouse

- **Integrated:** Operational data from multiple databases and external data sources are integrated in a data warehouse to provide a single, unified database for decision support.
- Consolidation of data requires consistent naming conventions, uniform data formats, and comparable measurement scales across databases and external data sources.

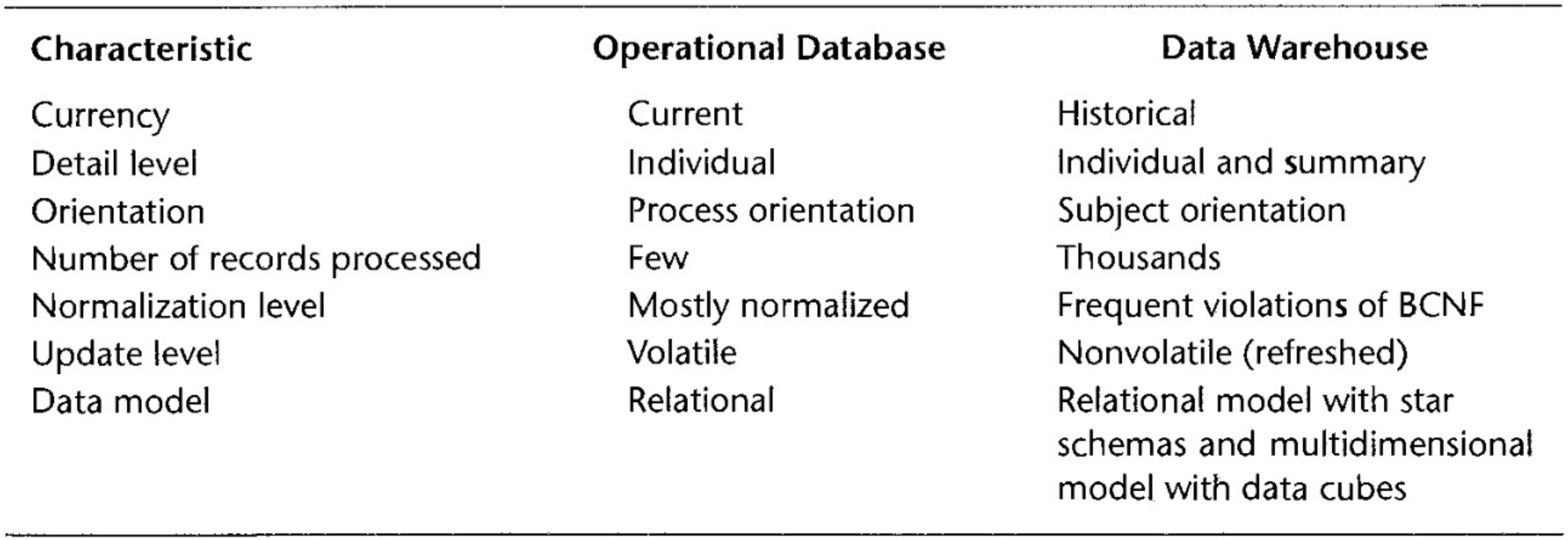
Characteristics of Data Warehouse

- **Time-Variant:** Data warehouses use time stamps to represent historical data.
- The time dimension is critical for identifying trends, predicting future operations, and setting operating targets.
- Data warehouses essentially consist of a long series of snapshots, each of which represents operational data captured at a point in time.

Characteristics of Data Warehouse

- **Non-volatile:** New data in a data warehouse are appended, rather than replaced, so that historical data are preserved.
- The act of appending new data is known as refreshing the data warehouse.
- Lack o f update and delete operations ensures that a data warehouse is free o f update or deletion anomalies.
- Transaction data are transferred to a data warehouse only when most updating activity has been completed.

Operational Database Vs. Data Warehouse



Characteristic	Operational Database	Data Warehouse
Currency	Current	Historical
Detail level	Individual	Individual and summary
Orientation	Process orientation	Subject orientation
Number of records processed	Few	Thousands
Normalization level	Mostly normalized	Frequent violations of BCNF
Update level	Volatile	Nonvolatile (refreshed)
Data model	Relational	Relational model with star schemas and multidimensional model with data cubes

Operational Database Vs. Data Warehouse

- Transaction processing relies on operational databases with current data at the individual level, while decision support processing utilizes data warehouses with historical data at both the individual and summarized levels.
- Individual-level data provides flexibility for responding to a wide range o f decision support needs while summarized data provides fast response to repetitive queries.

Operational Database Vs. Data Warehouse

- For example, an order-entry transaction requires data about individual customers, orders, and inventory items, while a decision support application may use monthly sales to customers over a period of several years.
- Operational databases therefore have a process orientation (e.g., all data relevant to a particular business process), compared to a subject orientation for data warehouses (e.g., all customer data or all order data).

Operational Database Vs. Data Warehouse

- A transaction typically updates only a few records, whereas a decision support application may query thousands to millions of records.
- Data integrity and performance of transaction processing require that operational data bases be highly normalized.
- In contrast, data warehouses are usually denormalized from Boyce-Codd Normal Form to reduce the effort to join large tables.

Architectures of Data Warehouse

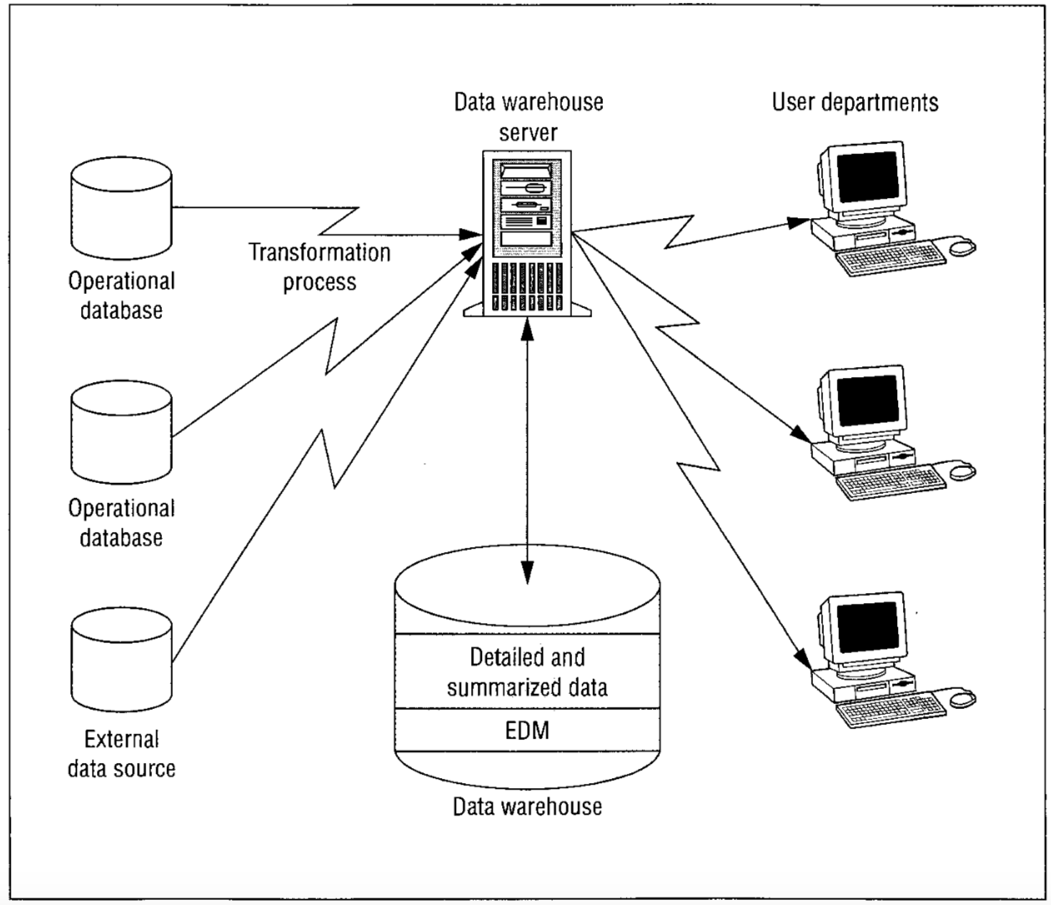
Two-Tier Architecture

- For most organizations, a two-tier or three-tier data warehouse architecture is appropriate.
- In a two-tier data warehouse architecture, operational data are transformed and then transferred to a data warehouse.
- A separate layer of servers may be used to support the complex activities of the transformation process.
- To assist with the transformation process, an enterprise data model (EDM) is created.

Two-Tier Architecture

- The EDM describes the structure of the data warehouse and contains the metadata required to access operational databases and external data sources.
- The EDM may also contain details about cleaning and integrating data sources.
- Management uses the data warehouse directly to retrieve data for decision support.

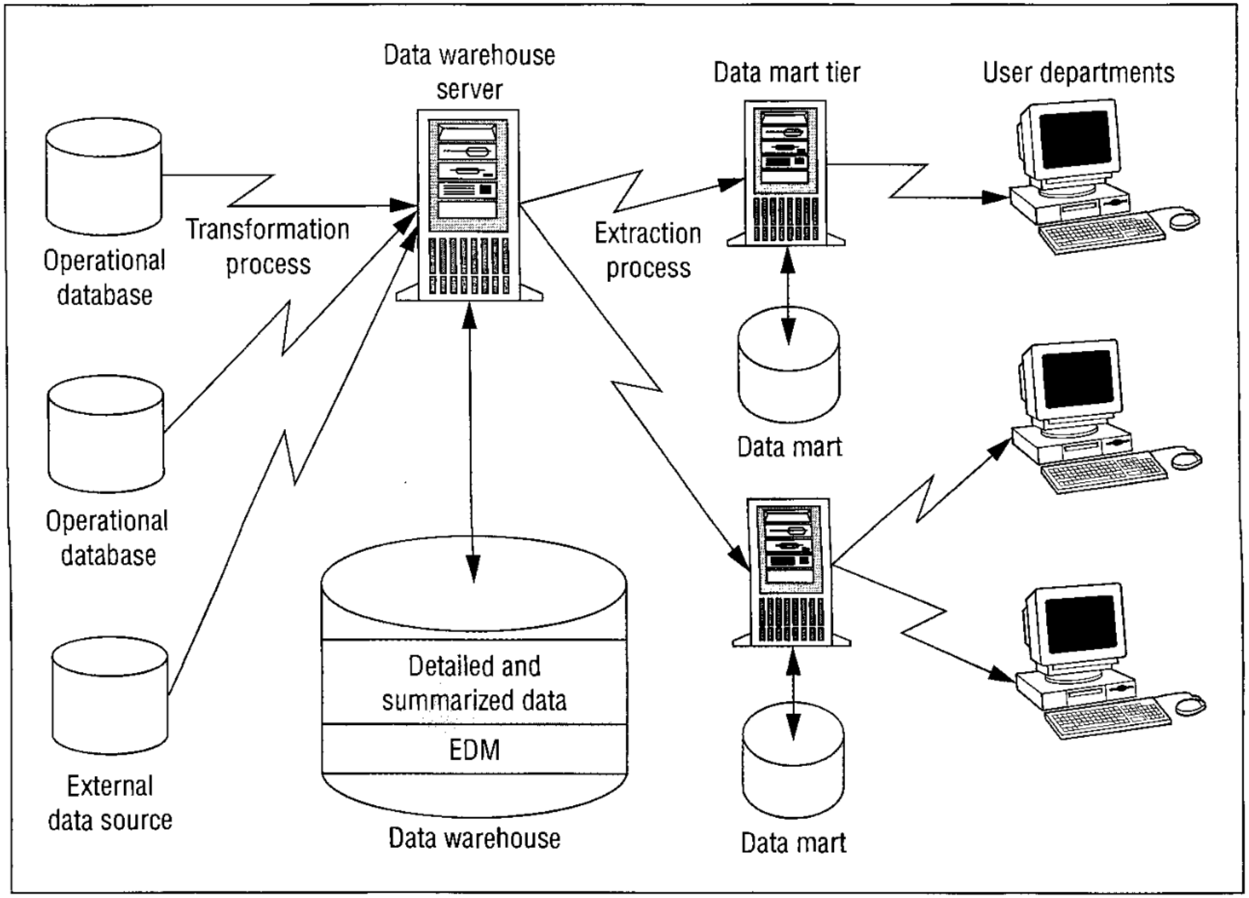
Two-Tier Architecture



Three-Tier Architecture

- The two-tier architecture can have performance problems for large data warehouses with data-intensive applications for decision support.
- To alleviate these difficulties, many large organizations use a three-tier data warehouse architecture.

Three-Tier Architecture



Data Marts

- Data marts act as the interface between end users and the corporate data warehouse, storing a subset of the warehouse data and refreshing those data on a periodic (e.g., daily or weekly) basis.
- Generally, the data warehouse and the data marts reside on different servers to improve performance and fault tolerance.
- Departmental users retain control over their own data marts, while the data warehouse remains under the control of the corporate information systems staff.

Data Marts

- Departmental users generally need access to small portions of the data warehouse, instead of the entire warehouse.
- To provide them with faster access while isolating them from data needed by other user groups, smaller data warehouses called **data marts** are often used.

Data Models

- Because of the different processing requirements, different data models have been developed for operational databases and data warehouses.
- The relational data model dominates for operational databases.
- In the early years of data warehouse deployment, the multidimensional data model dominated.

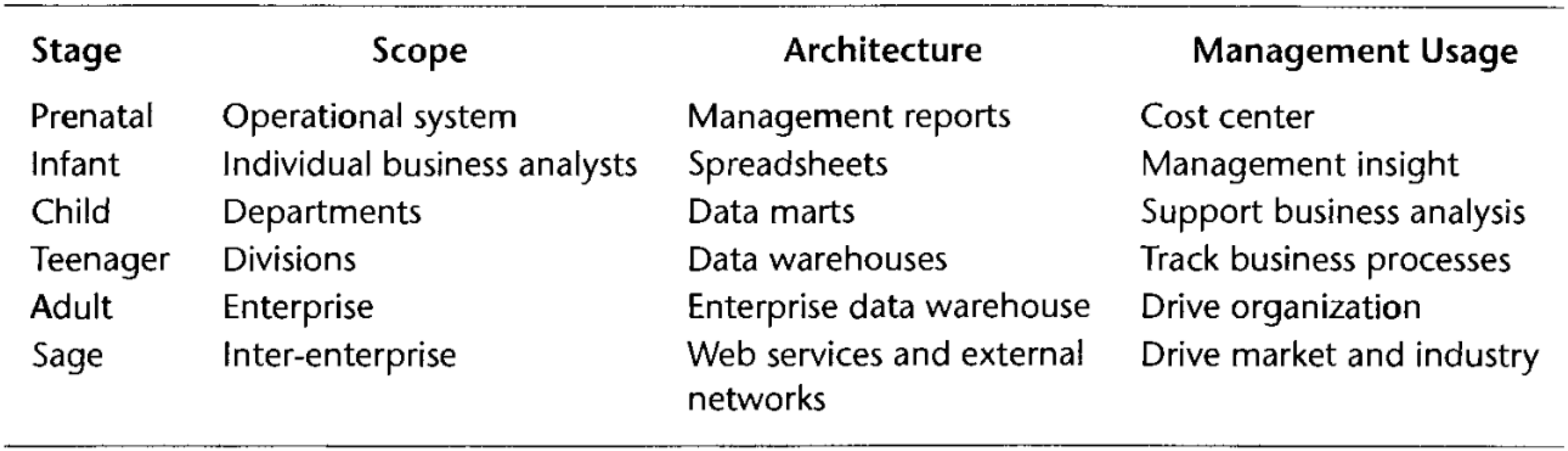
Data Models

- In recent years, relational databases have been increasingly used for data warehouses with a schema pattern known as a **star schema**.
- The multidimensional data model is now typically used as an end user representation of a view of a data warehouse.

Data Warehouse Maturity Model

- The data warehouse maturity model has been proposed to provide guidance for data warehouse investment decisions.
- The maturity model consists of six stages.
- The stages provide a framework to view an organization's progress, not an absolute metric as organizations may demonstrate aspects of multiple stages at the same time.

Data Warehouse Maturity Model



Stage	Scope	Architecture	Management Usage
Prenatal	Operational system	Management reports	Cost center
Infant	Individual business analysts	Spreadsheets	Management insight
Child	Departments	Data marts	Support business analysis
Teenager	Divisions	Data warehouses	Track business processes
Adult	Enterprise	Enterprise data warehouse	Drive organization
Sage	Inter-enterprise	Web services and external networks	Drive market and industry

Data Warehouse Maturity Model

- As organizations move from lower to more advanced stages, increased business value can occur.
- However, organizations may have difficulty justifying significant new data warehouse investments in the teenager and adult stages as benefits are sometimes difficult to quantify.

Data Mining

Data Mining

- Data warehouses improve the quality of decision making by consolidating and aggregating transactional data.
- The value of a data warehouse can be increased if hidden patterns in the data can be discovered.
- Data mining refers to the process of discovering implicit patterns in data and using these patterns for business advantage.
- Data mining facilitates the ability to detect, understand, and predict patterns.

Tools for Data Mining

- Data access tools to extract and sample transaction data according to complex criteria from large databases.
- Data visualization tools that enable a decision maker to gain a deeper, intuitive under standing of data.
- A rich collection of models to cluster, predict, and determine association rules from large amounts of data. The models involve neural networks, genetic algorithms, decision tree induction, rule discovery algorithms, probability networks, and other expert system technologies.
- An architecture that provides optimization, client-server processing, and parallel queries to scale to large amounts of data.

Questions

Questions

- Why are operational databases not particularly suited for decision support applications?
- How is a data warehouse different from a data mart?
- When is the three-tier data warehouse architecture more appropriate than the two-tier data warehouse architecture?
- What are the components of an enterprise data model?