RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-01

MR. DHIRENDRA

ASSISTANT PROFESSOR

RAMA UNIVERSITY

OUTLINE

- DATA WAREHOUSING
- DATA WAREHOUSE INFORMATION
- INTEGRATING HETEROGENEOUS DATABASES
- QUERY-DRIVEN APPROACH
- UPDATE-DRIVEN APPROACH
- FUNCTIONS OF DATA WAREHOUSE TOOLS AND UTILITIES
- MCQ
- REFERENCES

DATA WAREHOUSING

- Process of constructing and using a data warehouse.
- constructed by integrating data from multiple heterogeneous sources
- that support analytical reporting, structured and/or ad hoc queries, and decision making.
- involves data cleaning, data integration, and data consolidations.

DATA WAREHOUSE INFORMATION

- decision support technologies that help utilize the data available in a data warehouse.
- help executives to use the warehouse quickly and effectively.
- They can gather data, analyze it, and take decisions based on the information present in the warehouse.
- The information gathered in a warehouse can be used in any of the following domains

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.
- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

INTEGRATING HETEROGENEOUS DATABASES

- **Query-driven Approach**
- **Update-driven Approach**

QUERY-DRIVEN APPROACH

- **Query-Driven Approach**
  - traditional approach to integrate heterogeneous databases.
  - used to build wrappers and integrators on top of multiple heterogeneous databases
  - integrators are also known as mediators.
- **Process of Query-Driven Approach**
  - When a query is issued to a client side, a metadata dictionary translates the query into an appropriate form for individual heterogeneous sites involved.
  - Now these queries are mapped and sent to the local query processor.
  - The results from heterogeneous sites are integrated into a global answer set.
- **Disadvantages**
  - Query-driven approach needs complex integration and filtering processes.
  - This approach is very inefficient.
  - It is very expensive for frequent queries.
  - This approach is also very expensive for queries that require aggregations.

UPDATE-DRIVEN APPROACH

Update-Driven Approach

- Alternative to the traditional approach.
- Information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse.
- This information is available for direct querying and analysis.

Advantages

- This approach has the following advantages −
- This approach provide high performance.
- The data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.
- Query processing does not require an interface to process data at local sources.

FUNCTIONS OF DATA WAREHOUSE TOOLS AND UTILITIES

- **Data Extraction** –

  Involves gathering data from multiple heterogeneous sources.

- **Data Cleaning** –

  Involves finding and correcting the errors in data.

- **Data Transformation** –

  Involves converting the data from legacy format to warehouse format.

- **Data Loading** –

  Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.

- **Refreshing** –

  Involves updating from data sources to warehouse.

Multiple Choice Question

1. The time horizon in Data warehouse is usually _____.
a) 1-2 years.
b) 3-4years.
c) 5-6 years.
d) 5-10 years

2.. The data is stored, retrieved & updated in _____.
a) OLAP
b) OLTP
c) SMTP
d) FTP

3. _____describes the data contained in the data warehouse.
a) Relational data.
b) Operational data.
c) Metadata
d) Informational data.

4. _____ is the heart of the warehouse.
a) Data mining database servers.
b) Data warehouse database servers.
c) Data mart database servers.
d) Relational data base servers.

5. _____ is the specialized data warehouse database.
a) Oracle
b) DBZ
c) Informix
d) Redbrick

REFERENCES

- https://www.tutorialspoint.com/dwh/dwh_overview.htm

- http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
  DATA MINING BOOK WRITTEN BY Micheline Kamber