RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-10

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

OUTLINE

- METADATA REPOSITORY
- DATA MARTING
- AGGREGATIONS
- QUERY FACILITY
- QUERY MANAGEMENT
- MCQ
- REFERENCES

Metadata Repository

Meta data is the data defining warehouse objects. It stores:

- **Description of the structure of the data warehouse**

  schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents

- **Operational meta-data**

  data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)

- **The algorithms used for summarization**

- **The mapping from operational environment to the data warehouse**

- **Data related to system performance**

  warehouse schema, view and derived data definitions

- **Business data**

  business terms and definitions, ownership of data, charging policies

Data Marting

- A data mart is a subset of an organizational data store, usually oriented to a specific purpose or major data subject, that may be distributed to support business needs. Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization. Data marts are often derived from subsets of data in a data warehouse, though in the bottom-up data warehouse design methodology the data warehouse is created from the union of organizational data marts.
- A data mart is a repository of data gathered from operational data and other sources that is designed to serve a particular community of knowledge workers
- Data Marts are created for the following reasons:
  - ❑ To speed up queries by reducing the volume of data to be scanned.
  - ❑ To structure data in a form suitable for a user access tool.
  - ❑ To partition data in order to impose access control strategies.
  - ❑ To segment data into different hardware platforms.

Data Marting

- **Situation for creation of Data Marts to identify whether**
  - ❑ There is a natural functional split within the organization.
  - ❑ There is a natural split of the data.
  - ❑ The proposed user access tool uses its own database structures.
  - ❑ Any infrastructure issues predicate the use of data marts.
  - ❑ There are any access control issues that require data marts to provide Chinese walls.
- **Costs of Data Marting :** data Marting strategies incur costs in the following areas:
  - ❑ Hardware and software
  - ❑ Network Access
  - ❑ Time-window constraints

Aggregations

- Data aggregation is an essential component of decision support DWH. It allows us to provide cost-effective query performance by avoiding the need for substantial.
- Aggregation strategies rely on the fact that most common queries will analyze either a subset or an aggregation of the detailed data.
- We can see that the appropriate aggregation will substantially reduce the processing time required to run a query, at the cost of processing and storing the intermediate results.
- The aggregations make trends more apparent, by providing an overview of the whole picture rather than small parts of the picture.

Aggregations

- **Designing Summary Tables :** The primary purpose of using summary tables is to cut down the time it takes to execute queries. The objective of using summary tables is to minimize the volume of data being scanned, by storing as many partial results as possible.
- **The summary tables are designed using following steps:**
  - ❑ Determine which dimensions are aggregated.
  - ❑ Determine the aggregation of multiple values.
  - ❑ Aggregate multiple facts into the summary table.
  - ❑ Determine the level of aggregation.
  - ❑ Determine the extent of embedding dimension data in the
  - ❑ summary.
  - ❑ Design time into the summary table.
  - ❑ Index the summary table.

QUERY FACILITY

- Generally, requests for information from a DW are usually complex and iterative queries of what happened in a business such as "finding the products' types, units sold and total cost that were sold week for all stores in west region". Most of the queries contain a lost of join operations involving a large number of records. Also aggregate function such as group – by are very common in these queries. Such complex queries could take several hours or days to process because the queries have to process through a large amount of data. A majority of requests for information from a data warehouse involve dynamic ad hoc queries; users can ask any question at any time for any reason against the base table in a data warehouse.
- The ability to answer these queries quickly is a critical issue in the data warehouse environment. There are many solutions to speed up query processing such as summary tables, indexes, parallel machines, etc the performance when using summary tables for predetermined queries is good. However when an unpredicted query arises, the system must scan, fetch and sort the actual data, resulting in performance degradation. Whenever the base table changes, the summary tables have to be recomputed. A lso building summary tables often supports only known frequent queries, and requires more time and more space than the original data . because we cannot build all possible summary tables, choosing which ones to be built is a difficult job. Moreover, summarized data hide valuable information.
- For example, we cannot know the effectiveness of the promotion on Monday by querying weekly summary. Indexing is the key to achieve this objective without adding additional hardware.

Query Management

- In information delivery process query management is very important because most of the information is delivered trough queries. The entire query process must be managed with maximum care. Following features of a managed query environment should be considered first.
  - ❑ Query initiation, formulation, and results presentation are provided on the client machine.
  - ❑ Metadata guides the query process.
  - ❑ Ability for the users to navigate easily through the data structure is absolutely essential.
  - ❑ Information environment must be flexible to accommodate different classes of users.
- There are three sections in the environment where queries are being processed. Essentially,. The first section deals with the user who needs the query management facility. The next section is about the types of queries themselves. Finally, we have the data that resides in the query processing environment with the three sections. Few important services to be made available in the managed query environment.
- **Query Definition :** Make it easy to translate the business need into the proper query syntax.
- **Query simplification :** Make the complexity of data and query formulation transport to the users. Provide simple views of the data structures showing tables and attributes. Make the rules for combining tables and statures easy to use.

Query Management

- **Query Recasting :** Even simple- looking queries can results in intensive data retrieval and manipulation. Therefore, provide for parsing incoming queries and recasting them to work more efficiently.
- **Ease of Navigation:** Use of metadata to browse through the data warehouse, easily navigation with business terminology and not technical phrases.
- **Query Execution:** Provide ability for the user to submit the query for execution without any intervention from IT
- **Results Presentation :** Present results of the query in a variety of ways.
- **Aggregate Awareness:** Query processing mechanism must be aware of aggregate fact tables and, whenever necessary, redirect the queries to the aggregate tables for faster retrieval.
- **Query Governance:** Monitor and intercept runaway queries before they bring down the data warehouse operations

Multiple Choice Question

1. The most common source of change data in refreshing a data warehouse is _____.
a) Query able change data.
b) cooperative change data.
c) logged change data.
d) snapshot change data.

2.. _____ are responsible for running queries and reports against data warehouse tables.
a) Hardware
b) Software
c) End users.
d) Middle ware.

3. Query tool is meant for _____.
a) data acquisition.
b) information delivery.
c) information exchange.
d) communication

4. Classification rules are extracted from _____.
a) root node.
b) decision tree.
c) siblings
d) branches

5. Dimensionality reduction reduces the data set size by removing _____.
a) relevant attributes.
b) irrelevant attributes.
c) derived attributes.
d) composite attributes.

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
- https://www.javatpoint.com/three-tier-data-warehouse-architecture
- M.H. Dunham, " Data Mining: Introductory & Advanced Topics" Pearson Education
- Jiawei Han, Micheline Kamber, " Data Mining Concepts & Techniques" Elsevier
- Sam Anahory, Denniss Murray," data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, " Pearson Education
- Mallach," Data Warehousing System", TMH