Data Warehousing

SUNIL BIJARANIYA[1], NITESH KUMAR JANGIR[2]
[1,2] *MTECH (CSE), Shekhawati Institute of Engineering and Technology, Sikar*

Abstract -- The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. A data warehouses provides us generalized and consolidated data in multidimensional view. Along with generalized and consolidated view of data, a data warehouses also provides us Online Analytical Processing (OLAP) tools. These tools help us in interactive and effective analysis of data in a multidimensional space. This analysis results in data generalization and data mining.

Indexed Terms: Data warehouse, data mining, OLAP

I. INTRODUCTION

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

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively.

(i) 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.
(ii) Customer Analysis: Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
(iii) Operations Analysis: Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

II. DATA WAREHOUSE ARCHITECTURE

A data warehouses consists of a three-tier architecture. Following are the three tiers of the data warehouse architecture.

(i) Bottom Tier: The bottom tier of the architecture is the data warehouse database server. It is the relational database system.
(ii) Middle Tier: In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.

- By Relational OLAP (ROLAP), which is an extended relational database management system.
- By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.

(iii) Top-Tier: This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

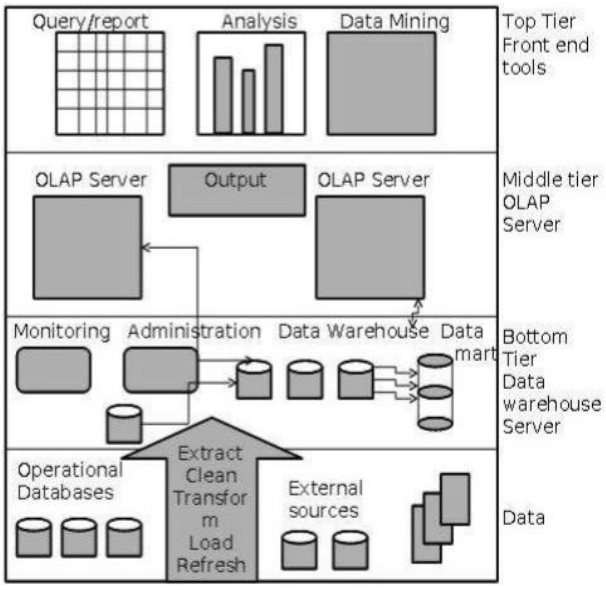

Fig 1: data warehouse three tier architecture

III. DATA WAREHOUSE MODEL

From the perspective of data warehouse architecture, we have the following data warehouse models −
(i) Virtual Warehouse

(ii) Data mart

(iii) Enterprise Warehouse

(i) Virtual Warehouse: The view over an operational data warehouse is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.

(ii) Data Mart: Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

Some important points to remember about data marts

- They are implemented on low-cost servers.
- The implementation data mart cycles is measured in short periods of time.
- Data marts are small in size.
- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse.
- Data mart are flexible.

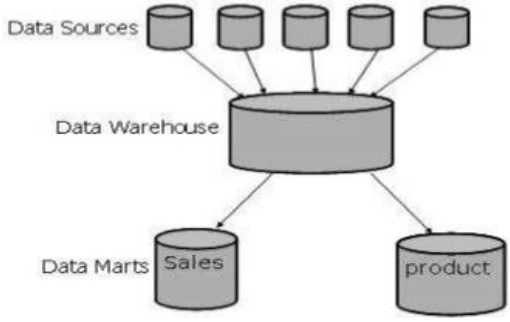

Fig 2: Data Mart Concept

(iii) Enterprise Warehouse:

- An enterprise warehouse collects all the information and the subjects spanning an entire organization
- It provides us enterprise-wide data integration.
- The data is integrated from operational systems and external information providers.
- This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

IV. ONLINE ANALYTICAL PROCESSING (OLAP)

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information. There are four types of OLAP servers −

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

(i) Relational OLAP: ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.
ROLAP includes the following −

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

(ii) Multidimensional OLAP: MOLAP uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the data set is sparse.

(iii) Hybrid OLAP: Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP.

(iv) Specialized SQL Servers: Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

V. CONCEPT OF MATA DATA

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In terms of data warehouse, we can define metadata as follows.

- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory.

Metadata can be broadly classified into three categories

(i) Business Metadata − It has the data ownership information, business definition, and changing policies.

(ii) Technical Metadata − It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.

(iii) Operational Metadata − It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.

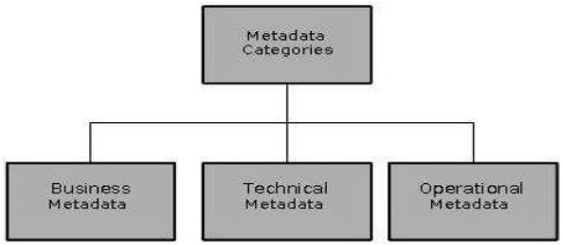

Fig 3: Types of Meta data

VI. DATA WAREHOUSE BACK UP

A data warehouse is a complex system and it contains a huge volume of data. Therefore it is important to back up all the data so that it becomes available for recovery in future as per requirement.

(i) Complete backup − It backs up the entire database at the same time. This backup includes all the database files, control files, and journal files.

(ii) Partial backup − As the name suggests, it does not create a complete backup of the database. Partial backup is very useful in large databases because they allow a strategy whereby various parts of the database are backed up in a round-robin fashion on a day-to-day basis, so that the whole database is backed up effectively once a week.

(iii) Cold backup − Cold backup is taken while the database is completely shut down. In multi-instance environment, all the instances should be shut down.

(iv) Hot backup − Hot backup is taken when the database engine is up and running. The requirements of hot backup varies from RDBMS to RDBMS.

(v) Online backup − It is quite similar to hot backup.

VII. DATA WAREHOUSE TESTING

Testing is very important for data warehouse systems to make them work correctly and efficiently. There are three basic levels of testing performed on a data warehouse −

(i) Unit testing

(ii) Integration testing

(iii) System testing

(i) Unit Testing

- In unit testing, each component is separately tested.
- Each module, i.e., procedure, program, SQL Script, Unix shell is tested.
- This test is performed by the developer.

(ii) Integration Testing

- In integration testing, the various modules of the application are brought together and then tested against the number of inputs.
- It is performed to test whether the various components do well after integration.

(iii) System Testing

- In system testing, the whole data warehouse application is tested together.
- The purpose of system testing is to check whether the entire system works correctly together or not.
- System testing is performed by the testing team.
- Since the size of the whole data warehouse is very large, it is usually possible to perform minimal system testing before the test plan can be enacted.

VIII. DATA WAREHOUSE FUTURE ASPECTS

Following are the future aspects of data warehousing.

- As we have seen that the size of the open database has grown approximately double its magnitude in the last few years, it shows the significant value that it contains.
- As the size of the databases grow, the estimates of what constitutes a very large database continues to grow.
- The hardware and software that are available today do not allow to keep a large amount of data online.
- The record contains textual information and some multimedia data. Multimedia data cannot be easily manipulated as text data.
- Apart from size planning, it is complex to build and run data warehouse systems that are ever increasing in size.
- With the growth of the Internet, there is a requirement of users to access data online.

REFERENCES

[1] "Data Mining: Concepts and Techniques" by Han
[2] "Data Warehousing" by Reema Thareja
[3] "Data Warehousing and Data Mining" by Singh M
[4] "Data Mining and Warehousing" by S Prabhu
[5] "Data Mining and Warehousing" by Khushboo and Sandeep
[6] "Introducing Data Science: Big Data, Machine Learning, and more, using Python tools" by Davy Cielen and Arno Meysman
[7] "Data Warehousing: OLAP and Data Mining" by Nagabhushana S
[8] "Data Mining and Warehousing" by M Sudheep Elayidom
[9] "Python for Data Science for Dummies" by John Paul Mueller and Luca Massaron
[10] "The Encyclopedia of Data Warehousing and Mining" by John Wang