Program : **B.E**

Subject Name: **Data Mining**

Subject Code: **CS-8003**

Semester: **8th**

XII

CS-8003 Elective-V (2) Data Mining

Unit-I

Introduction to Data warehousing, needs for developing data Warehouse, Data warehouse systems and its Components, Design of Data Warehouse, Dimension and Measures, Data Marts:-Dependent Data Marts, Independents Data Marts & Distributed Data Marts, Conceptual Modeling of Data Warehouses:-Star Schema, Snowflake Schema, Fact Constellations. Multidimensional Data Model & Aggregates.

Data Mining:

Introduction: Data mining is the process of analyzing large amount of data sets to identify patterns and establish relationships to solve problems through data analysis. Data Mining is processing data to identify patterns and establish relationships.

Data mining techniques are used in many research areas, major industry areas like Banking, Retail, Medicine, cybernetics, genetics and marketing. While data mining techniques are a means to drive efficiencies and predict customer behavior, if used correctly, a business can set itself apart from its competition through the use of predictive analysis.

Data mining can be applied on any kind of data or information stored. Data mining is also known as data discovery and knowledge discovery.

Data Warehousing

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. DW is combining data from multiple and usually varied sources in to one comprehensive and easily manipulated database. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. DW is commonly used by companies to analyze trends over time.

As compare to the relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing

(OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users.

"A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process"

Needs for developing data Warehouse:

- Provides an integrated and total view of the enterprise.
- Make the organizations current and historical information easily available for decision making.
- Make decision support transactions possible without hampering operational system.
- Provide consistent organizations information
- Provide a flexible and interactive sources of strategic information.
- End user creation of reports: The creation of reports directly by end users is much easier to accomplish in a BI environment.
- Dynamic presentation through dashboards: Managers want access to an interactive display of up-to-date critical management data.
- Drill-down capability
- Metadata creation: This will make report creation much simpler for the end-user
- Data mining
- Security

Data warehouse systems and its Components:

Data warehousing is typically used by larger companies analyzing larger sets of data for enterprise purposes. The data warehouse architecture is based on a relational database system server that functions as the central warehouse for informational data. Operational data and processing is purely based on data warehouse processing. This central information system is used some key components designed to make the entire environment for operational systems. Its mainly created to support different analysis, queries that need extensive searching on a larger scale.

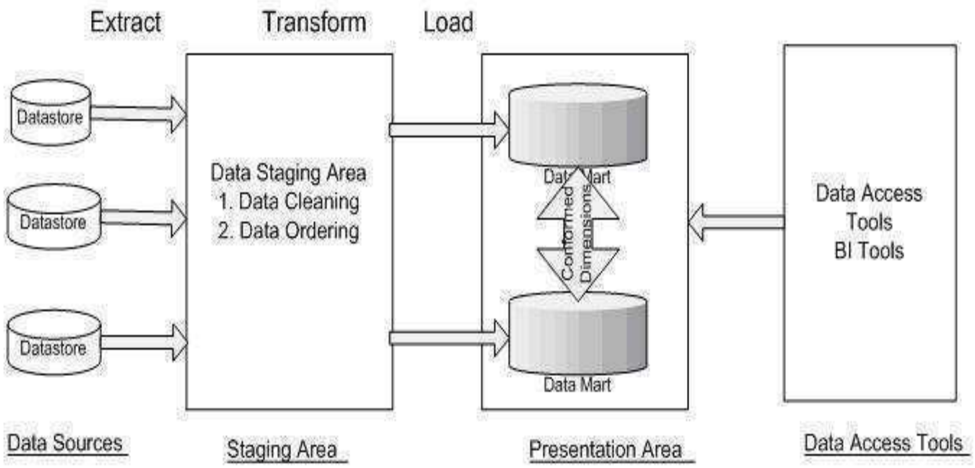


Figure 1.1: Data Warehouse Components

Operational Source System

Operational systems are tuned for known transactions and workloads, while workload is not known a priori in a data warehouse. Traditionally data base system used for transaction processing systems which stores transaction data of the organizations business. Its generally used one record at any time not stores history of the information's.

Data Staging Area

As soon as the data arrives into the Data staging area it is set of ETL process that extract data from source system. It is converted into an integrated structure and format.

Data is extracted from source system and stored, cleaned, transform functions that may be applied to load into data warehouse.

Removing unwanted data from operational databases.

Converting to common data names and definitions.

Establishing defaults for missing data accommodating source data definition changes.

Data Presentation Area

Data presentation area are the target physical machines on which the data warehouse data is organized and stored for direct querying by end users, report writers and other applications. It's the place where cleaned, transformed data is stored in a dimensionally structured warehouse and made available for analysis purpose.

Data Access Tools

End user data access tools are any clients of the data warehouse. An end user access tool can be a complex as a sophisticated data mining or modeling applications.

Design of Data Warehouse

Design Methods

Bottom-up design

This architecture makes the data warehouse more of a virtual reality than a physical reality. In the bottom-up approach, starts with extraction of data from operational database into the staging area where it is processed and consolidated for specific business processes. The bottom-up approaches reverse the positions of the data warehouse and the data marts. These data marts can then be integrated to create a comprehensive data warehouse.

Top-down design

The data flow in the top down OLAP environment begins with data extraction from the operational data sources. The top-down approach is designed using a normalized enterprise data model. The results are obtained quickly if it is implemented with iterations. It is time consuming process with an iterative method and the failure risk is very high.

Hybrid design

The hybrid approach aims to harness the speed and user orientation of the bottom up approach to the integration of the top-down approach. To consolidate these various data models, and facilitate the extract transform load process, data warehouses often make use of an operational data store, the information from which is parsed into the actual DW. The hybrid approach begins with an ER diagram of the data mart and a gradual extension of the data marts to extend the enterprise model in a consistent linear fashion. It will provide rapid development within an enterprise architecture framework.

Dimension and Measures

Data warehouse consists of dimensions and measures. It is a logical design technique used for data warehouses. Dimensional model allow data analysis from many of the commercial OLAP products available today in the market. For example, time dimension could show you the breakdown of sales by year, quarter, month, day and hour.

Measures are numeric representations of a set of facts that have occurred. The most common measures of data dispersion are range, the five number summery (based on quartiles), the inter-quartile range, and the standard deviation. Examples of measures include amount of sales, number of credit hours, store profit percentage, sum of operating expenses, number of past-due accounts and so forth.

Types

Conformed dimension

Junk dimension

Degenerate dimension

Role-playing dimension

Data Marts

A data mart is a specialized system that brings together the data needed for a department or related applications. A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), such as educational, sales, operations, collections, finance or marketing data. The sources may contain internal operational systems, central data warehouse, or external data. It is a small warehouse which is designed for the department level.

Dependent, Independent or stand-alone and Hybrid Data Marts

Three basic types of data marts are dependent, independent or stand-alone, and hybrid. The categorization is based primarily on the data source that feeds the data mart.

Dependent data marts : Data comes from warehouse. It is actually created a separate physical data-store.

Independent data marts: A standalone systems built by drawing data directly from operational or external sources of data or both. Independent data mart are independent and focuses exclusively on one subject area. It has a separate physical data store.

Hybrid data marts : Can draw data from operational systems or data warehouses.

Dependent Data Marts

A dependent data mart allows you to unite your organization's data in one data warehouse. This gives you the usual advantages of centralization. Figure 1.2 shows a dependent data mart.

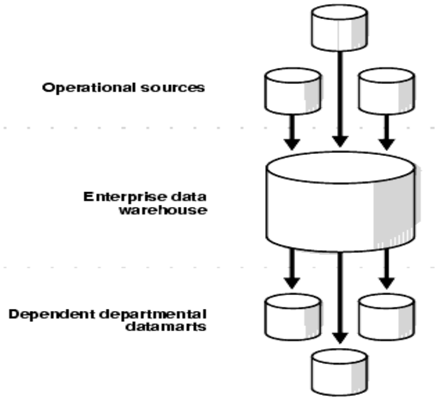


Figure 1.2: Dependent Data Mart

Independent or Stand-alone Data Marts

An independent data mart is created without the use of a central data warehouse. This could be desirable for smaller groups within an organization. It is not, however, the focus of this Guide. Figure 1.3 shows an independent data mart.

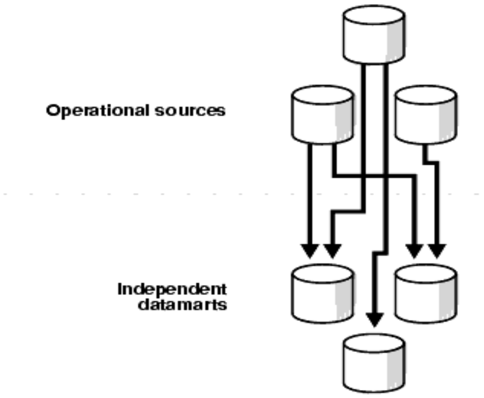


Figure 1.3: Independent Data Marts

Hybrid Data Marts

A hybrid data mart allows you to combine input from sources other than a data warehouse. This could be useful for many situations, especially when you need ad hoc integration, such as after a new group or product is added to the organization. Provides rapid development within an enterprise architecture framework. Figure 1.4 shows hybrid data mart.

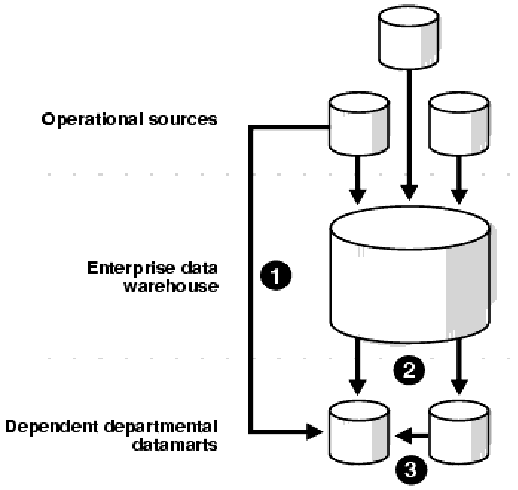


Figure 1.4: Hybrid Data Mart

Conceptual Modeling of Data Warehouses

It may also be defined by discretizing or grouping values for a given dimension or attribute, resulting in a set-grouping model. A conceptual data model identifies the highest-level relationships between the different entities. Features of conceptual data model include:

- Includes the important entities and the relationships among them.
- No attribute is specified.
- No primary key is specified.

Figure 1.5 below is an example of a conceptual data model.

Conceptual Data Model

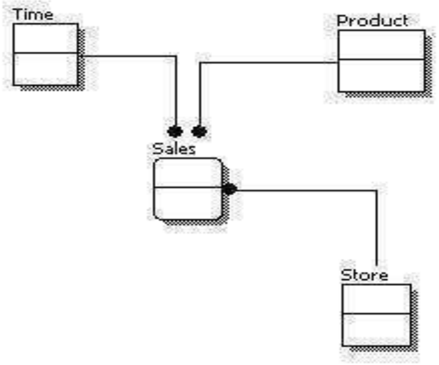


Figure 1.5: Conceptual data model

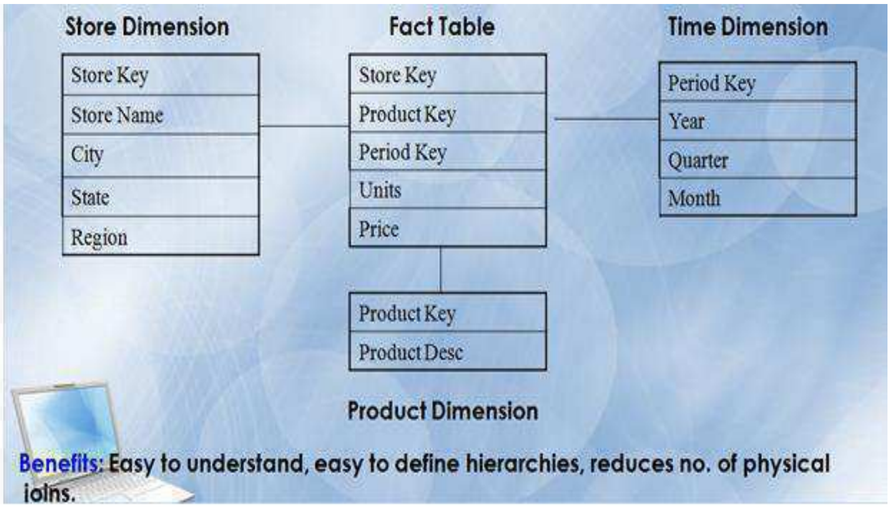
Data Warehousing Schemas

From the figure above, we can see that the only information shown via the conceptual data model is the entities that describe the data and the relationships between those entities. There may be more than one concept hierarchy for a given attribute or dimension, based on different users view points. No other information is shown through the conceptual data model.

- **Star Schema**
- **Snowflake Schema**
- **Fact Constellation**

Star Schema

- Consists of set of relations known as Dimension Table (DT) and Fact Table (FT)
- A single large central fact table and one table for each dimension.
- A fact table primary key is composition of set of foreign keys referencing dimension tables.
- Every dimension table is related to one or more fact tables.
- Every fact points to one tuple in each of the dimensions and has additional attributes
- Does not capture hierarchies directly.



Store Dimension

Store Key
Store Name
City
State
Region

Fact Table

Store Key
Product Key
Period Key
Units
Price

Time Dimension

Period Key
Year
Quarter
Month

Product Key
Product Desc

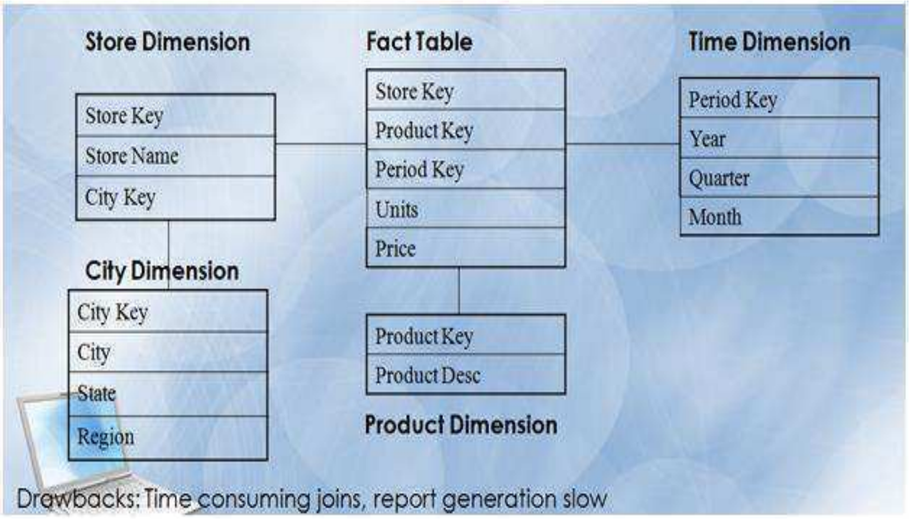
Product Dimension

Benefits: Easy to understand, easy to define hierarchies, reduces no. of physical joins.

Snowflake Schema

- Variant of star schema model.
- Used to remove the low cardinality.
- A single, large and central fact table and one or more tables for each dimension.

- Dimension tables are normalized split dimension table data into additional tables. But this may affect its performance as joins needs to be performed.
- Query performance would be degraded because of additional joins. (delay in processing)



Store Dimension
Store Key
Store Name
City Key

Fact Table
Store Key
Product Key
Period Key
Units
Price

Time Dimension
Period Key
Year
Quarter
Month

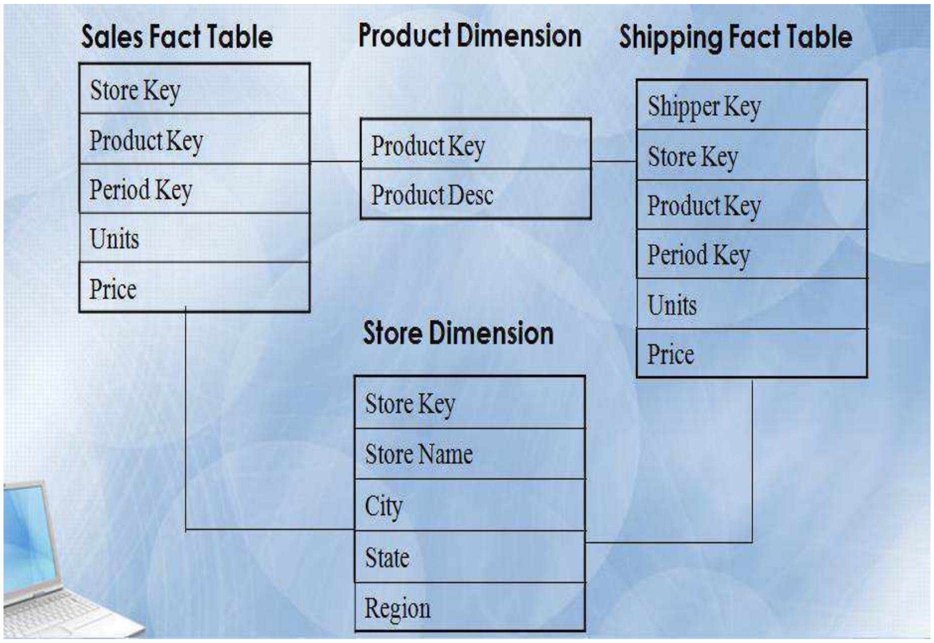
City Dimension
City Key
City
State
Region

Product Dimension
Product Key
Product Desc

Drawbacks: Time consuming joins, report generation slow

Fact Constellation:

- As its name implies, it is shaped like a constellation of stars (i.e. star schemas).
- Allow to share multiple fact tables with dimension tables.
- This schema is viewed as collection of stars hence called galaxy schema or fact constellation.
- Solution is very flexible, however it may be hard to manage and support.
- Sophisticated application requires such schema.



Sales Fact Table
Store Key
Product Key
Period Key
Units
Price

Product Dimension
Product Key
Product Desc

Shipping Fact Table
Shipper Key
Store Key
Product Key
Period Key
Units
Price

Store Dimension
Store Key
Store Name
City
State
Region

Multidimensional Data Model

Data warehouses are generally based on ‘multi-dimensional” data model. The multidimensional data model provides a framework that is intuitive and efficient, that allow data to be viewed and analyzed at the desired level of details with a good performance. The multidimensional model start with the examination of factors affecting decision-making processes is generally organization specific facts, for example sales, shipments, hospital admissions, surgeries, and so on. One instances of a fact correspond with an event that occurred. For example, every single sale or shipment carried out is an event. Each fact is described by the values of a set of relevant measures that provide a quantitative description of events. For example, receipts of sales, amount of shipment, product cost are measures.

The multidimensional data model is an integral part of On-Line Analytical Processing, or OLAP. Because OLAP is on-line, it must provide answers quickly; analysts pose iterative queries during interactive sessions, not in batch jobs that run overnight. And because OLAP is also analytic, the queries are complex. Dimension tables support changing the attributes of the dimension without changing the underlying fact table. The multidimensional data model is designed to solve complex queries in real time. The multidimensional data model is important because it enforces simplicity.

The multidimensional data model is composed of logical cubes, measures, dimensions, hierarchies, levels, and attributes. The simplicity of the model is inherent because it defines objects that represent real-world business entities. Analysts know which business measures they are interested in examining, which dimensions and attributes make the data meaningful, and how the dimensions of their business are organized into levels and hierarchies. Figure shows the relationships among the logical objects. Figure 1.6 shows the Logical Multidimensional Model

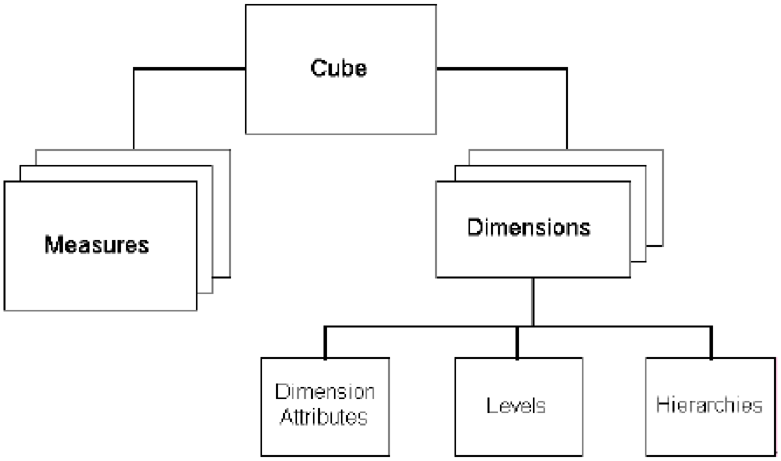


Figure 1.6: Logical Multidimensional Model

Aggregates

In data warehouse huge amount of data is stored that makes analyses of data very difficult. This is the basic reason why selection and aggregation is required to examine specific part of data. Aggregations are the way by which information can be divided so queries can be run on the aggregated part and not the whole set of data. These are pre-calculated summaries derived from the most granular fact table. It is a process for information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, profession, or income. The information about such groups can then be used for web site personalization. Tables are always changing along with the needs of the users so it is important to define the aggregations according to what summary tables might be of use.

*****