UNIT 3

TYPES OF DATA WAREHOUSES

1. Host Based Data warehouses

- *Host Based (MVS) Data Warehouses:* The data warehouses that reside on high-volume databases on MVS are the host based type of data warehouses. Such data warehouses

    1. usually have very high volumes of data storage
    2. Require support for both MVS and client-based report and query facilities.
    3. have very complex source systems
    4. Require continuous maintenance since these must be used for mission-critical purposes.

Steps to build such a data warehouse-

- *Unload Phase* involves selecting and scrubbing the operational data.
- *Transform Phase* for translating it into an appropriate form and defining the rules for accessing and storing it.
- *Load phase* for moving the data directly into DB2 tables or a special file for moving it to another database or non-MVS warehouse.

Host Based (Unix) Data Warehouses: Oracle and Informix RDBMSs provide the facilities for such data warehouses. Both of these databases can extract data from MVS-based databases as well as a larger number of other UNIX-based databases.

2. Host Based single-stage (LAN) Data warehouses

With a LAN-based warehouse, data delivery can be managed either centrally or from the workgroup environment so that business groups can meet and manage their own information needs without burdening centralized IT resources.

Limitations/challenges:

- LAN-based warehousing solutions are normally limited by both DBMS and hardware scalability factors.
- Many LAN based enterprises have not implemented adequate job scheduling, recovery management, organized maintenance, and performance monitoring procedures to support robust warehousing solutions.

- Often these warehouses are dependent on other platforms for source data. Building an environment that has data integrity, recoverability, and security needs careful design, planning and implementation. Otherwise, synchronization of changes and loads from sources to server could cause innumerable problems.

3. LAN Based workgroup Data warehouses

In this warehouse, you extract data from a variety of sources (like Oracle, IMS, DB2) and provide multiple LAN-based warehouses.

Designed for workgroup environment, it is ideal for any business organization that wishes to build a data warehouse, often called a data mart. Usually requires minimal initial investment and technical training. Its low startup cost and ease of use allow a workgroup to quickly build and easily manage its own custom data mart.

Common Issues:

- Lack of understanding how to distribute data and supporting intentional data redundancy for performance reasons.
- Many organizations may not have adequate job scheduling, recovery management, and performance monitoring to support robust warehousing solutions.
- Although providing +ve cost benefits, LAN-based warehousing solutions can be limited by both hardware and DBMS limitations.
- For many large enterprises, similar skills in database design, maintenance, and recovery are not present in every workgroup environment.

4. Multistage Data warehouses

This configuration is well suited to environments where end-users in different capacities require access to both summarized data for up-to-the-minute tactical decisions as well as summarized cumulative data for long-term strategic decisions. Both ODS (Operation Data Store) and the data warehouse may reside on host-based on LAN-based databases, depending on volume and usage requirements. Typically the ODS stores only the most recent records. The data warehouse stores the historical evolution of the records.

5. Stationary Data warehouses

In this type of a data warehouse, user is given direct access to the data, instead of moving from the sources. For many organizations, infrequent access, volume issues or corporate necessities dictate such an approach.

This is likely to impact performance since users will be competing with the production data stores.
Such a warehouse will require sophisticated middleware, possible with a single interface to the user. An integrated metadata repository becomes an absolute necessity under this environment.

6. Distributed Data warehouses

There are at least two types of distributed data warehouses and their variations for the enterprise: local warehouses distributed throughout the enterprises and a global warehouse. Useful when there are diverse businesses under the same enterprise umbrella. This approach may be necessary if a local warehouse already existed, prior to joining the enterprise. Local data warehouses have the following common characteristics:

1. Activity occurs at local level
2. Majority of the operational processing is done at the local site.
3. Local site is autonomous
4. Each local data warehouse has its own unique structure and content of data.
5. The data is unique and of prime importance to that locality only.
6. Majority of the data is local and not replicated.
7. Any intersection of data between local data warehouses is coincidental.
8. Local site serves different geographic regions.
9. Local site serves different technical communities.

The primary motivation in implementing distributed data warehouses is that integration of the entire enterprise data does not make sense. It is reasonable to assume that an enterprise will have at least some natural intersections of data from one local site to another. If there is any intersection, then it is usually contained in a global data warehouse.

7. Virtual Data warehouse

The data warehouse is a great idea, but it is complex to build and requires investment. Why not use a cheap and fast approach by eliminating the transformation steps of repositories for metadata and another database. This approach is termed the 'virtual data warehouse'. To accomplish this there is need to define 4 kinds of information:

i. A data dictionary containing the definitions of the various databases.

ii. A description of the relationship among the data elements.

iii. The description of the way user will interface with the system.

iv. The algorithms and business rules that define what to do and how to do it.

Disadvantages:

1. Since queries compete with production data transactions, performance can be degraded.
2. There is no metadata, no summary data or no individual DSS (Decision Support System) integration or history. All queries must be repeated, causing additional burden on the system.
3. There is no refreshing process, causing the queries to be very complex.

DESIGNING DATA WAREHOUSE DATABASE

Main Phases of Database Design

1. **Requirements specification**: The requirements and the collection analysis phase produce both data requirements and functional requirements. The data requirements are used as a source of database design. The data requirements should be specified in as detailed and complete form as possible.

2. **Conceptual Design**: Once all the requirements have been collected and analyzed, the next step is to create a conceptual schema for the database, using a high level conceptual data model. This phase is called conceptual design.

   The result of this phase is an Entity-Relationship (ER) diagram or UML class diagram. It is a high-level data model of the specific application area. It describes how different entities (objects, items) are related to each other. It also describes what attributes (features) each entity has. It includes the definitions of all the concepts (entities, attributes) of the application area.

3. **Logical design:** Translates the conceptual schema from the previous phase into an implementation model common to several DBMSs, e.g., relational or object-relational.

4. **Normalization:** Normalization is the last part of the logical design. The goal of normalization is to eliminate redundancy and potential update anomalies.

   Redundancy means that the same data is saved more than once in a database. Update anomaly is a consequence of redundancy. If a piece of data is saved in more than one place, the same data must be updated in more than one place.

Normalization is a technique by which one can modify the relation schema to reduce the redundancy. Each normalization phase adds more relations (tables) into the database.

5. **Physical Design**: The goal of the last phase of database design, physical design, is to implement the database. At this phase one must know which database management system (DBMS) is used. For example, different DBMS's have different names for data-types and have different data-types.

DATABASE DESIGN METHODOLOGY FOR DATA WAREHOUSES

Nine steps methodology-

'Nine-Step Methodology' includes following steps: „

a. Choosing the process „
b. Choosing the grain „
c. Identifying and conforming the dimensions „
d. Choosing the facts „
e. Storing pre-calculations in the fact table „
f. Rounding out the dimension tables „
g. Choosing the duration of the database „
h. Tracking slowly changing dimensions „
i. Deciding the query priorities and the query modes

a. Choosing the process: The process (function) refers to the subject matter of a particular data mart. „ E.g. Lease, Property Sale. First data mart built should be the one that is most likely to be delivered on time, within budget, and to answer the most commercially important business questions.

b. Choosing the grain: Decide what a record of the fact table is to represents. Identify dimensions of the fact table. The grain decision for the fact table also determines the grain of each dimension table. Also include time as a core dimension, which is always present in star schemas.

c. Identifying and conforming the dimensions: Dimensions set the context for asking questions about the facts in the fact table. If any dimension occurs in two data marts, they must be exactly the same dimension, or one must be a mathematical subset of the other. A dimension used in more than one data mart is referred to as being conformed.

d. Choosing the facts: The grain of the fact table determines which facts can be used in the data mart. Facts should be numeric and additive. Unusable facts include: non-numeric facts; non-additive facts; fact at different granularity from other facts in table.

e. **Storing pre-calculations in the fact table**: Once the facts have been selected each should be re-examined to determine whether there are opportunities to use pre-calculations.

f. **Rounding out the dimension tables:** Text descriptions are added to the dimension tables. Text descriptions should be as intuitive and understandable to the users as possible. Usefulness of a data mart is determined by the scope and nature of the attributes of the dimension tables.

g. **Choosing the duration of the database:** Duration measures how far back in time the fact table goes. Very large fact tables raise at least two very significant data warehouse design issues.

- Often difficult to source increasing old data. „
- It is mandatory that the old versions of the important dimensions be used, not the most current versions. Known as the 'Slowly Changing Dimension' problem.

h. **Tracking slowly changing dimensions:** Slowly changing dimension problem means that the proper description of the old dimension data must be used with the old fact data.

- Often, a generalized key must be assigned to important dimensions in order to distinguish multiple snapshots of dimensions over a period of time.
- There are three basic types of slowly changing dimensions: „
- Type 1, where a changed dimension attribute is overwritten „
- Type 2, where a changed dimension attribute causes a new dimension record to be created „
- Type 3, where a changed dimension attribute causes an alternate attribute to be created so that both the old and new values of the attribute are simultaneously accessible in the same dimension record

i. **Deciding the query priorities and the query modes**: Most critical physical design issues affecting the end-user 's perception includes: „
- physical sort order of the fact table on disk „
- presence of pre-stored summaries or aggregations „
- Additional physical design issues include administration, backup, indexing performance, and security.

OLAP AND DATA MINING

OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view.

OLAP allows users to analyze database information from multiple database systems at one time.

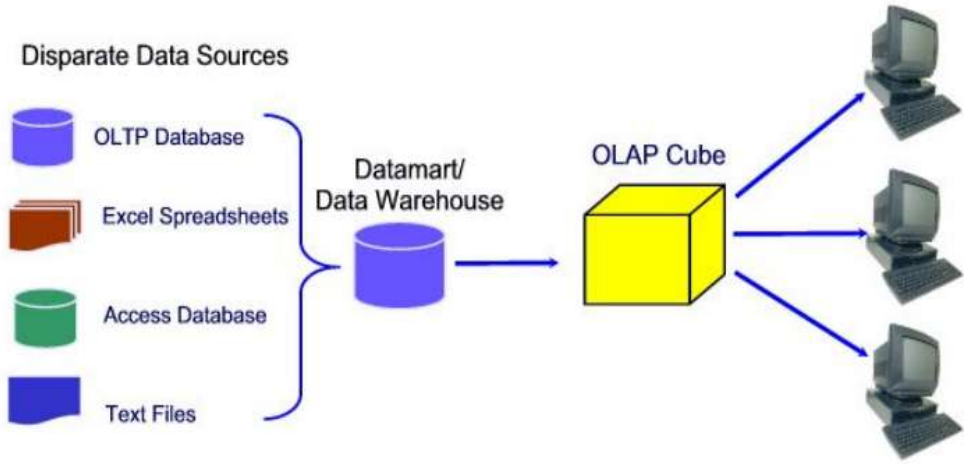
OLAP data is stored in multidimensional databases.

OLAP processing is often used for data mining. *f*

OLAP products are typically designed for multiple-user environments, with the cost of the software based on the number of users.

THE OLAP CUBE

- ✓ An OLAP Cube is a data structure that allows fast analysis of data.
- ✓ The arrangement of data into cubes overcomes a limitation of relational databases.
- ✓ It consists of numeric facts called measures which are categorized by dimensions.
- ✓ The OLAP cube consists of numeric facts called measures which are categorized by dimensions.
- ✓ A multidimensional cube can combine data from disparate data sources and store the information in a fashion that is logical for business users



OLAP OPERATIONS

The user-initiated process of navigating by calling for page displays interactively, through the specification of slices via rotations and drill down/up is sometimes called "slice and dice".

Slice: A slice is a subset of a multi-dimensional array corresponding to a single value for one or more members of the dimensions not in the subset.

Dice: The dice operation is a slice on more than two dimensions of a data cube (or more than two consecutive slices).

Drill Down/Up: Drilling down or up is a specific analytical technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).

Roll-up: A roll-up involves computing all of the data relationships for one or more dimensions. To do this, a computational relationship or formula might be defined.

Pivot: To change the dimensional orientation of a report or page display.

The output of an OLAP query is typically displayed in a matrix (or pivot) format. The dimensions form the row and column of the matrix; the measures, the values.

TYPES OF OLAP

Relational OLAP (ROLAP): Extended RDBMS with multidimensional data mapping to standard relational operation.

Multidimensional OLAP (MOLAP): Implemented operation in multidimensional data.

Hybrid Online Analytical Processing (HOLAP) is a hybrid approach to the solution where the aggregated totals are stored in a multidimensional database while the detail data is stored in the relational database. This is the balance between the data efficiency of the ROLAP model and the performance of the MOLAP model.

Relational OLAP

- Provides functionality by using relational databases and relational query tools to store and analyze multidimensional data.
- Build on existing relational technologies and represent extension to all those companies who already used RDBMS.
- Multidimensional data schema support within the RDBMS.
- Data access language and query performance are optimized for multidimensional data.
- Support for very large databases.

Multidimensional OLAP

- MOLAP extends OLAP functionality to MDBMS.
- Best suited to manage, store and analyze multidimensional data.
- Proprietary techniques used in MDBMS.
- MDBMS and users visualize the stored data as a 3-Dimensional Cube i.e Data Cube.
- MOLAP Databases are known to be much faster than the ROLAP counter parts.
- Data cubes are held in memory called “Cube Cache”

APPLICATIONS OF OLAP ™

- OLE DB for OLAP

OLE DB for OLAP (abbreviated ODBO) is a Microsoft published specification and an industry standard for multi-dimensional data processing.

ODBO is the standard application programming interface (API) for exchanging metadata and data between an OLAP server and a client on a Windows platform.

ODBO was specifically designed for Online Analytical Processing (OLAP) systems by Microsoft as an extension to Object Linking and Embedding Database (OLE DB).

- Marketing and sales analysis ™
- Consumer goods industries ™
- Financial services industry (insurance, banks etc) ™
- Database Marketing

Benefits of OLAP-

- ✓ One main benefit of OLAP is consistency of information and calculations.
- ✓ "What if" scenarios are some of the most popular uses of OLAP software and are made eminently more possible by multidimensional processing.
- ✓ It allows a manager to pull down data from an OLAP database in broad or specific terms.
- ✓ OLAP creates a single platform for all the information and business needs, planning, budgeting, forecasting, reporting and analysis.

DATA MINING

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications −

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Data Mining Applications

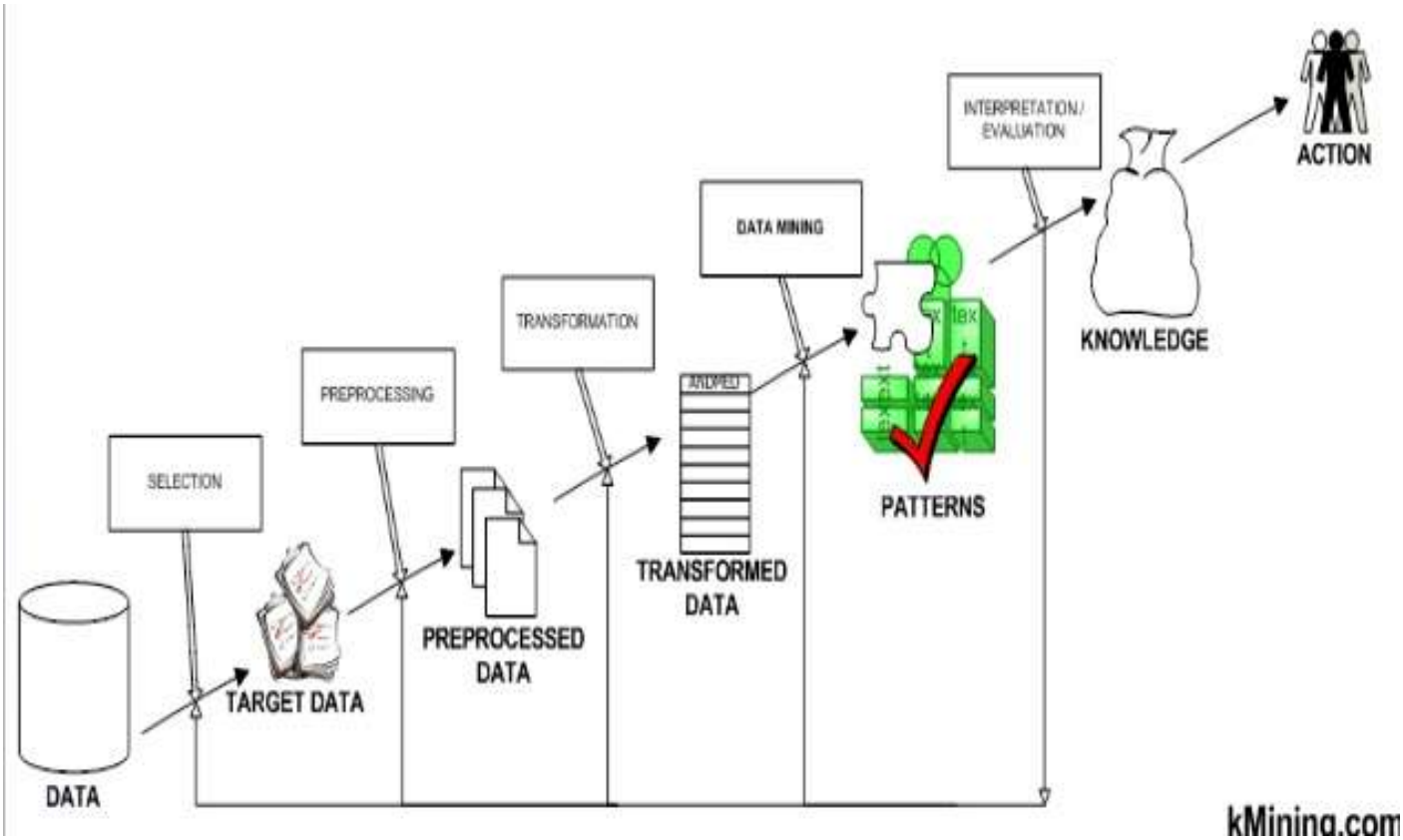
Data mining is highly useful in the following domains −

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid.

Steps of Data Mining

- Data integration
- Data selection
- Data cleaning
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation



Data Mining Techniques

- **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.

- **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.

- **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.

- **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.