

**B.E.**

Seventh Semester Examination, Dec.-2009

**Data Warehousing and Data Mining (IT-401-E)**

Note : Attempt any *five* questions. All questions carry equal marks.

Q. 1. (a) Differentiate between the following :

Database, Data Warehouse, Data Mining, KDD.

**Ans. Database :** Database is an organized body of related information. A database is a collection of data for one or more multiple uses. One way of classifying databases involves the type of content, for example bibliographic, full-text, numeric, image.

**Data Warehouse :** A data warehouse is a repository of an organization's electronically stored data. Datawarehouses are designed to facilitate reporting and analysis.

This definition of data warehouse focuses on data storage. However, the means to retrieve and analyze data, to extract, transform and load data and to manage data dictionary are also considered essential component of a data warehousing system.

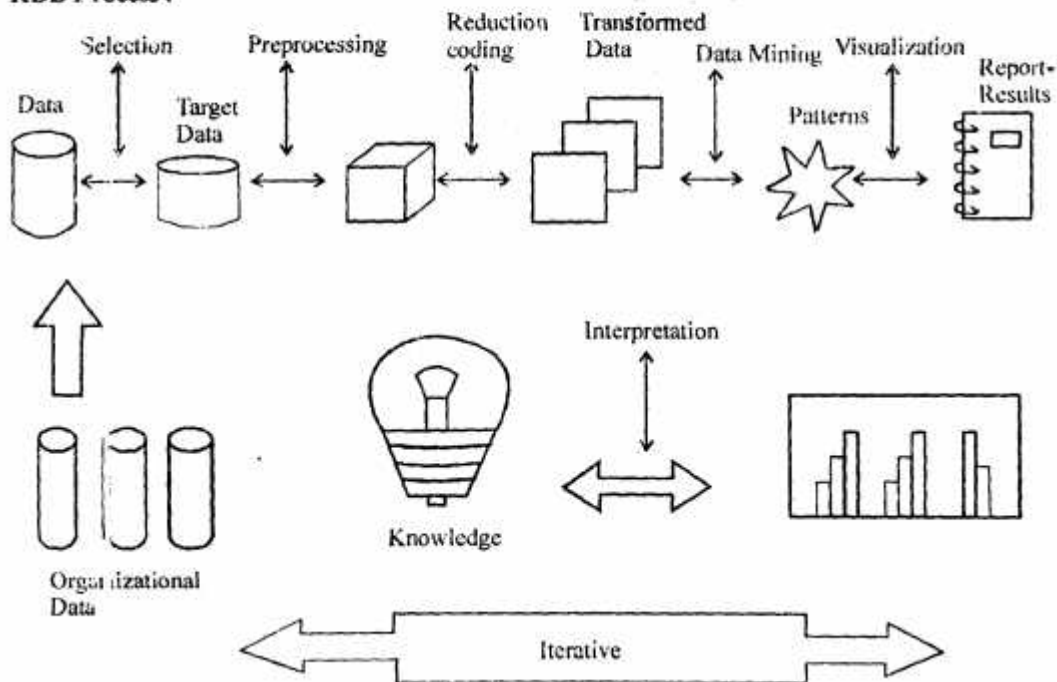
**Data Mining :** Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform the data into information. It is commonly used in a wide range of profiling practiced, such as marketing, surveillance, fraud detection and scientific discovery.

**KDD :** KDD stands for knowledge discovery in databases.

KDD is synonymous with large databases and automated discovery of patterns and relationships.

KDD is "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data."

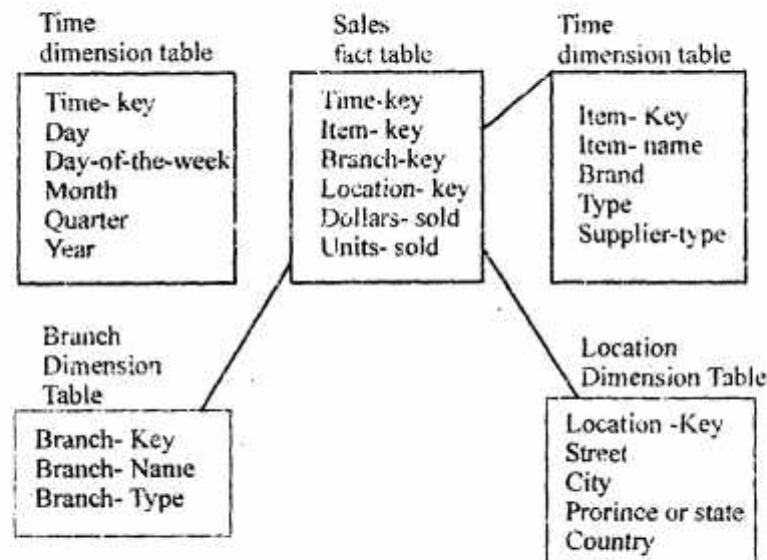
**KDD Process :**



**Q. 1. (b) Explain the concept of star, snowflake and galaxy schema with the help of suitable example.**

**Ans. Star Schema :** Single fact table with n dimension table linked to it :

- (i) There is a central large fact table with no redundancy.
- (ii) Each type in the fact table has a foreign key to a dimensional table which describes the details of that dimensions.



*Fig. Data Mining Concepts and Tech.*

**Q. 2. (a) Explain in detail the three-tier Data Warehouse architecture. How a query is mapped between three tiers, explain.**

**Ans.**

Data warehouse adopt a three tier architecture, these are :

- (i) Bottom tier (datawarehouse server).
- (ii) Middle tier (OLAP server)
- (iii) Top tier (front end tools).

**Bottom Tier :**

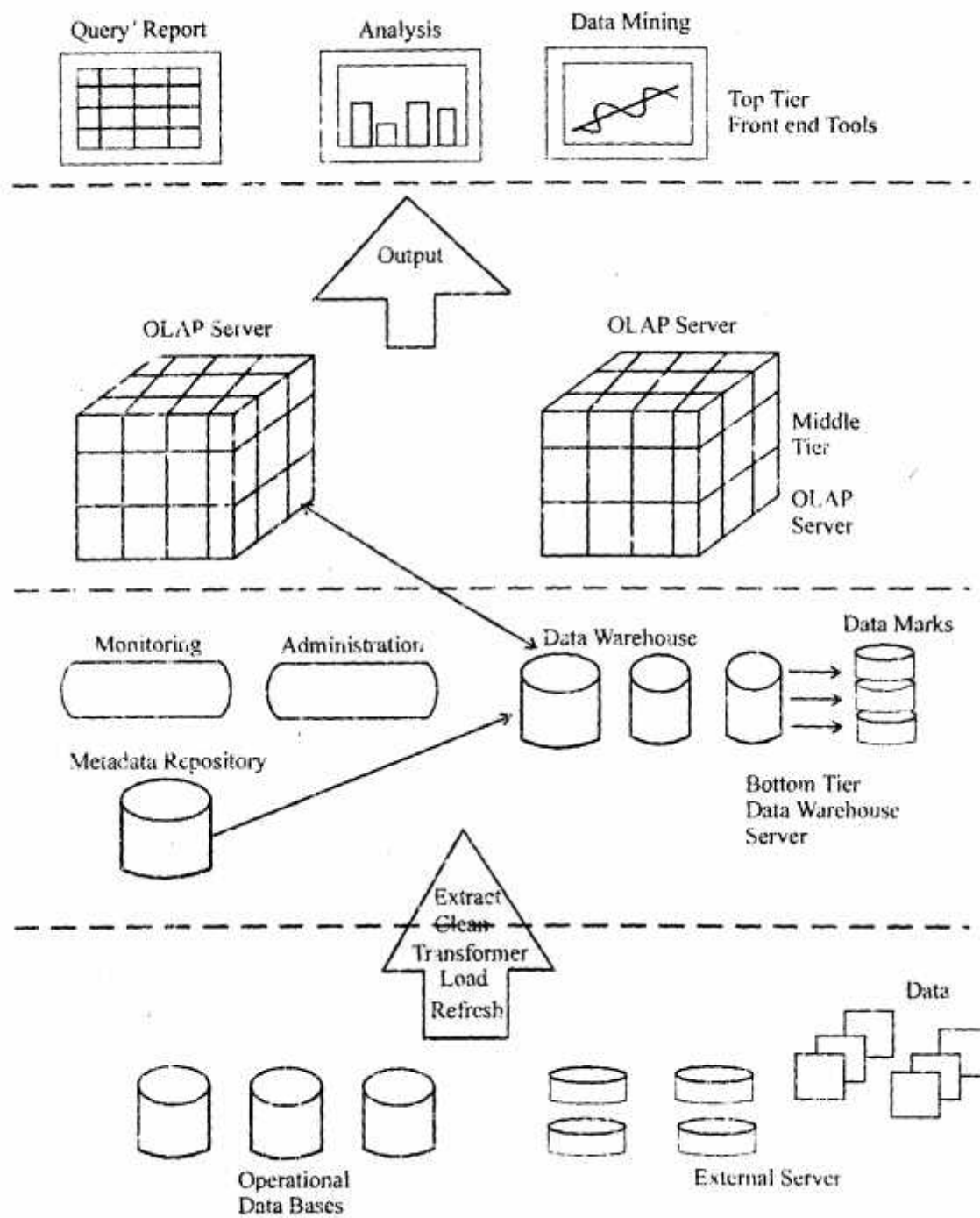
- (i) It is a warehouse database, server.
- (ii) Data is fed using back end tools and utilities.
- (iii) Data extracted using programs called gateways.
- (iv) It also contains meta data repository.

**Middle Tier :** Middle tier is an OLAP server that is typically implemented using either :

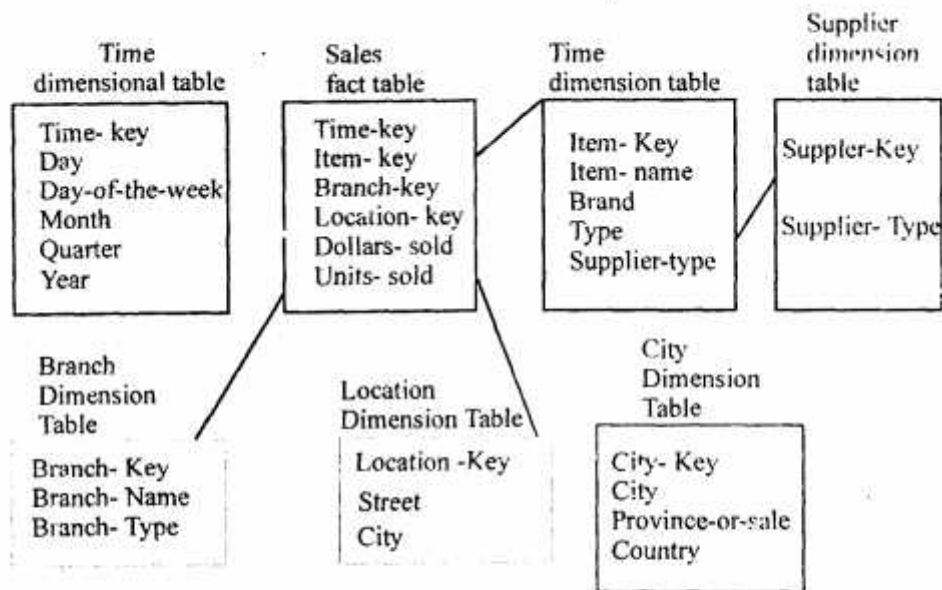
- (i) Relational OLAP model that is, extended relational DBMS that maps operations on multidimensional data standard relational operations; or
- (ii) A multidimensional OLAP model, that is a special purpose server that directly implements multidimensional data end operations.

**Top Tier :** The top tier is a front end client layer, which contains query and reporting tools, analysis tools

and or data mining tools.



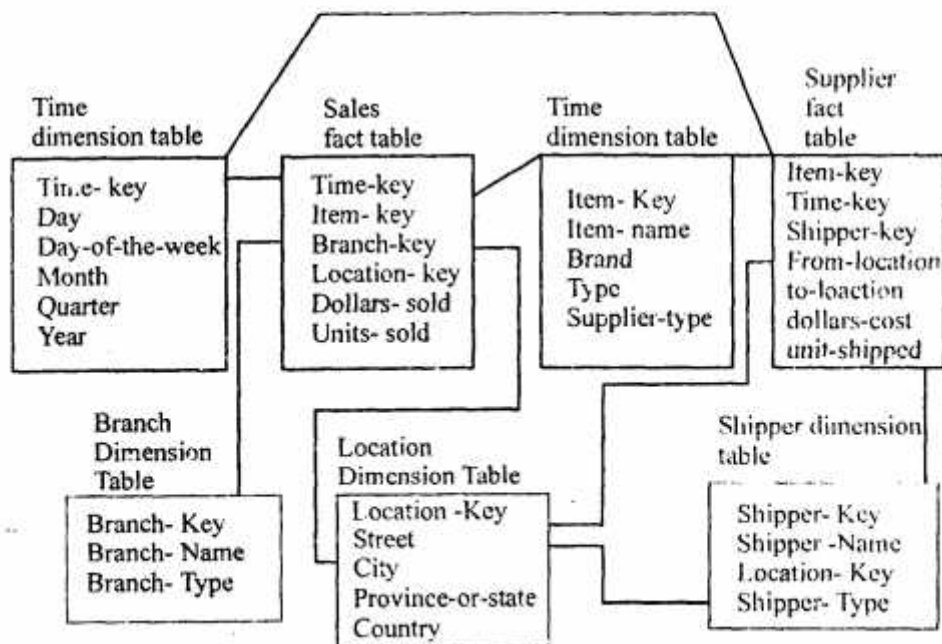
**Snowflake Schema :** Single fact table with n-dimining tables organized as a hierarchy. Some of the dimension tables are normalized thus splitting data into additional tables.



**Galaxy Schema :** Also known as fact constellation schema.

\* Multiple facts table sharing dimension tables.

In the fig. given below the 'sales' fact table and 'shipping' fact table share the dimension tables.



**Q. 2. (b) Discuss various OLAP operations which can be performed on a multidimensional data cube.**

**Ans. OLAP Operations :** The analyst can understand the meaning contained in the databases using multi-dimensional analysis. By aligning the data content with the analyst's mental model, the chances of confusion and erroneous interpretations are reduced. The analyst can navigate through the database and screen for a particular subset of the data, changing the data's orientations and defining analytical relations. 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 die". Common operations include slice and die, drill down, roll up and pivot.

**Slice :** A slice is a subset of a multidimensional array corresponding to a single value for one or more member of the dimensions not in the subset.

**Die :** The die operation is a slice on more than two dimensions of a data cube.

**Drill/Down/Up :** Drilling down or up is a specific analytical technique whereby the user navigate 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. A computational relationship or formula might be defined.

**Pivot :** This operation is also called rotate operation that rotates the data in order to provide an alternative presentation of data. To change the dimensional orientation of a report or page display.

**Q. 3. Suppose a database has four transactions. Let min-support = 60%, min-confidence = 80%**

TID	Date	Items-bought
T100	15/10/08	{A, B, D, K}
T200	15/10/08	{D, A, C, E, B}
T300	19/10/08	{C, A, B, E}
T400	22/10/08	{B, A, D}

(i) Find all frequent itemsets using a priori algorithm.

(ii) List all strong association rules matching the following meta-rule, where X is a variable representing customers and items i denotes variables representing items (e.g., A, B, etc.)  $\forall i \in \text{transaction}$ ,

$$\text{buys}(X, \text{item}_1) \wedge \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3).$$

**Ans. (i)** A priori algorithm employs BFS and uses a hash tree structure

$$FI(\text{Frequent Itemset}) = \{A, B, D\}$$

**(ii) Association Rules :**

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{Sup}(X)}$$

$$X \Rightarrow Y \text{ where } X, Y, \subseteq I \text{ and } X \cap Y = \phi$$

**Meta Rules :**

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(Y) \times \text{Supp}(X)}$$

$$\text{Conv}(X \Rightarrow Y) = \frac{1 - \text{Supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$$

**Q. 4. Explain the concept of Query Language employed in data mining and standardization of data mining. How pattern presentation and visualization specification can be carried out in data Mining Query Language?**

**Ans.** Data mining query languages are based on mine rule. MINE rule has been designed at the university of Torsion and the politechniq di Miland. It is an extension of sql which is coupled with a relation DBMS. Data can be selected using the full power of SQL. Mine association rules are materialized into relational tables as well. MINE RULE extracts association rules between values of attributes in a relational table. However, it is up to the user to specify the form of the rules to be extracted. The user can specify the cardinality of body and head of the desired rules and the attributes on which rule components can be built.

An interesting aspect of mine rule is that it is possible to work on different levels on grouping during the extraction. If there is one level of grouping, rule support will be completed w.r.t. the number of groups in the table. Defining a second level of grouping leads to the definition of clusters.

Rules components can be taken in two different clusters, eventually ordered, inside the same group. It is thus possible to extract some elementary sequential patterns (by clustering on a time related attributes). For instance, grouping purchases by customers who buy first. Butter and milk tend to by oil after. Concerning interestingness measures, MINE RULE enables to specify minimal frequency and confidence thresholds. The general syntax of a mine rule query for extracting rules is :

```
MINE RULE <Table Name> As
SELECT DISTINCT [<Cardinality>] <Attributes>
AS BODY
    [<Cardinality>] <Attributes>
AS HEAD
    [, SUPPORT] [, CONFIDENCE]
FROM <Table> [WHERE <whereclause>]
GROUP BY <Attributes> <HAVING <HAVING CLAUSE>]
    [CLUSTER BY <Attributes>
    [HAVING <Having clause>]]
EXTRACTING RULES WITH
SUPPORT : <real>, CONFIDENCE : <real>.
```

**Q. 5. (a) How decision trees assist in the process of data mining, explain?**

**Ans.** The decision tree is one of the most popular classification algorithm in current use in data mining and machine learning.

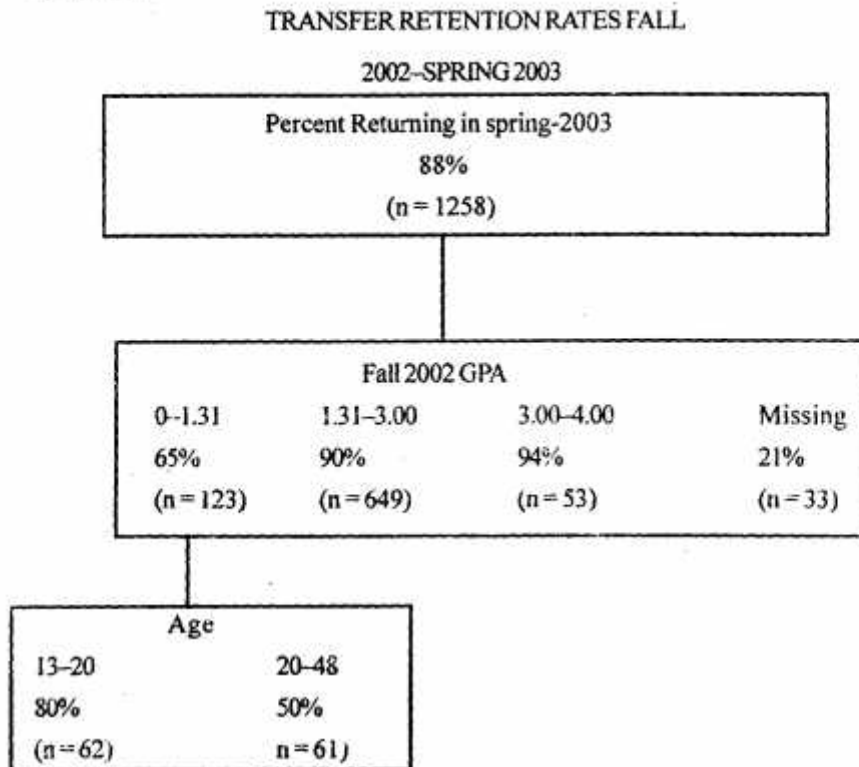
Data mining is all about automating the process of searching of patterns in the data.

A decision tree is a tree structured plan for a set of test in order to predict the output. To decide which attribute should be tested first, simply find the one with the highest information gain.

Implementation of decision trees in data mining :

1. Select significant independent variables.
2. Identify category grouping or interval breaks to create groups most different with respect to the different variables.
3. Select as the primary independent variables the one identifying groups with the most different values of the dependent variables.
4. Select additional variables to extend each branch if there are further significant differences.

Example :



Data mining basically is used now a days for development of large, integrated enterprises databases, data mining techniques and software.

One major significance is to develop a simplified user interface.

**Q. 5. (b) How data Warehouses are implemented? Discuss computation of data cubes in data Warehousing.**

**Ans. Data Warehouse Implementation :** Data warehousing was once devoted to business intelligence and reporting, usually at the departmental or business unit level. However, today the data warehouse is becoming a strategic corporate initiative supporting an entire enterprise across a multitude of business applications. The brisk pace of change, coupled with industry consolidation and mounting regulatory requirements, demands and data warehouses step into a mission critical, operational role.

It plays a arrival role in delivering the data foundation for key performance indicators such as revenue growth, margin improvement and asset efficiency at the corporate, business unit and departmental lends. The most effective approach is an enterprise wide, integrated hub to track and improve fundamental business measures.

**Value :**

- (i) Experienced inforamtical consultants guide and deliver a successful interprise data warehouse solution.
- (ii) Reduced development time implementing proven solutions.
- (iii) Experience and best practice methods based project execution.

- (iv) Project definition provides clear road map and realistic estimation up-front.
- (v) Knowledge transfer throughout to enable development of internal experts.
- (vi) Optimal implementation methods and functionality.
- (vii) Benefit of a balanced development team committed to delivering proven solutions.

**Q. 6. (a) Discuss various usage and trends in data Warehousing?**

**Ans. Usage in Data Warehousing :** DW appliances provide solutions for many analytic application uses, including :

- (i) Enterprise data warehousing.
- (ii) Super sized sand boxes which isolate power users with resource intensive queries.
- (iii) Off loading projects from the enterprise datawarehouse.
- (iv) Application with specific performance or loading requirements.
- (v) Data marks that have out grown their present environment.
- (vi) Turnkey data warehouses or data marks.
- (vii) Application requiring data warehouse encryption.

**Trends :** The DW appliance market has started to shift trends in many areas as it evolves :

- (i) Vendors have started moving toward using commodity technologies rather than proprietary assemble of commodity proprietary assemble of commodity components.
- (ii) Implemented applications show usage expansion from technical & data mark solutions to strategic & enterprise datawarehouse use.
- (iii) Mainstream vendor participation has become apparent as of 2009.
- (iv) With a lower total cost of ownership, reduced maintenance & high performance to address business analytics on growing data volumes.

**Q. 6. (b) What is meant by distributed and virtual data Warehouse, explain.**

**Ans. Distributed and Virtual Datawarehouse :** In today's world of global business, world wide partnership and corporate merger, decision making plays a major role in the steady growth of a business providing in a competitive edge. Decision making is the key to smooth day to day operations as well as for effective future planning in this ever competitive world. Several sources of data exist in the business from which valuable information can be extracted to help make a wide range of decisions. In order to facilitate querying and analysis, the data from these sources need to be integrated.

Applications are notorious for getting rid of historical data as quickly as possible to perform efficiently. What does this do for virtual data warehouse that operates on top of the application data? The answer is that a virtual data warehouse that operates on top of applications is limited to the historical data that resides in the applications, which isn't much.

The problem with existing datawarehouse is that when the business problem is addressed by multiple development organizations, each developer reduces the size and complexity of his/her problem. The result that one application has only a faint resemblance or connection to another application. There is no integration between applications. Therefore merely accessing data from multiple sources does not solve the much more profound problem of integrating data before the data is accessed.

A virtual data warehouse that provides aggregated views of the complete inventory. The virtual data warehouse contains metadata, which is used to form a logical enterprise data model that is part of the database record (DBOR) infrastructure. Each legacy back end database system is published on the infrastructure with its metadata extracted and used. The infrastructure software uses standard J2EE, JMS and reusable EJB's for

transactional unit requests and ETL tools for real time bulk loading of data.

**Q. 7. (a) What are Spatial Databases? Discuss the mining process of such databases.**

**Ans. Spatial Database :** A spatial database is a database that is optimized to store and query data related to objects in space, including points, lines and polygons. While typical databases can understand various numeric and character types of data, additional functionality needs to be added for databases to process spatial datatypes. These are typically called geometry or feature :

**Spatial Datamining Techniques :**

**Data Primitives for Spatial Data Mining :** We have developed a set of database primitives for mining in spatial databases which are sufficient to express most of the algorithms for spatial data mining and which can be efficiently supported by a DBMS. We believe that the use of these database primitives will enable the integration of spatial data mining with existing DBMS's & will speed up the development of new spatial data mining algorithms.

**Efficient DBMS Support :** Effective filters allow to restrict the search to such neighbourhood paths "leading away" from a starting object. Neighbourhood indices materialize certain neighborhood graphs to support efficient processing of the database primitives by a DBMS. The database primitives has been implemented on top of the DBMS illustra & are being ported to informs universal server.

**Algorithm for Spatial Datamining :** New algorithms for spatial characterization & spatial trend analysis were developed. For spatial characterization it is important that class membership of a database object is not only determined by its non-spatial attributes but also by the attributes of objects in its neighborhood. In spatial trend analysis, patterns of change of some non-spatial attributes in the neighborhood of a database object are determined.

**Q. 7. (b) How mining can be employed in Multimedia and time series databases, explain?**

**Ans. Data Mining in Multimedia and Time Series Databases :** Time series and multimedia data are ubiquitous : large volumes of such data are routinely created in scientific, industrial, entertainment, medical and biological 'domains.' Example includes gene expression data, X-rays, electroparadigms, electrocephlograms, gait analysis, stock market quotes, space telemetry etc. A decade ago, a seminal paper by fallouts or rangathan, Mandopoulos appeared in SIGMOD.

Time series databases consist of sequences of values or events changing with time. Data is recorded as regular intervals. Basic characteristics are :

- (i) Financial : Stock price, inflation
- (ii) Biomedical : Blood pressure
- (iii) Trend
- (iv) Cycle
- (v) Seasonal
- (vi) Irregular.

Categories of time series movements :

- (i) Long term or trend movement.
- (ii) Cycle or cycle variations.
- (iii) Irregular or random movements.

**Q. 8. Write short note on any four :**

- (i) Association Rules
- (ii) Back end tools & utilities in Data Warehousing

(iii) Genetic Algorithms

(iv) Data Warehouse Manager

(v) Complex aggregation at multiple granularities.

**Ans. (i) Association Rules :** Association rule mining is a technique for discovering unsuspected data dependencies & is one of the best known datamining techniques. The basic idea is to identify from a given database, consisting of item sets. Whether the occurrence of specific items, implies also the occurrence of other items with a relatively high probability. Association rule mining solves the problem of how to search efficiently for those dependencies.

**(ii) Back End Tools & Utilities in Data Warehousing :** Data warehousing systems use a variety of data extraction and cleaning tools & load & refresh utilities for populating warehouses. Data extraction from "foreign" sources is usually implemented via gateways & standard interfaces.

**Data Cleaning :** However, since large volumes of data from multiple sources are involved, there is a high probability of errors and anomalies in the data. Therefore, tools that help to detect data anomalies and correct them can have a high payoff.

Data migration tools

Data scrubbing tools

Data auditing tools.

**Load :** After extracting, cleaning & transforming, data must be loaded into the warehouse load utilities are used for this purpose. The load utilities for data warehouses have to deal with much larger data volumes than for operational databases.

**Refresh :** Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse.

**(iii) Genetic Algorithms :** Genetic algorithm is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms, which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover.

**(iv) Data Warehouse Manager :** The data warehouse manager's job is potentially huge; offering many opportunities & just as many risks. The DW manager has been given control of one of the most valuable assets of any organization the data.

Further more, the data warehouse manager is expected to interpret and deliver that asset to the rest of the organization in a way that makes it most useful. All eyes are on the datawarehouse manager.

**(v) Complex Aggregation at Multiple Granularities :** Decision support systems aim to provide answers to complex queries posed over very large databases. The databases may represent business information (such as transaction data), medical information (such as patient treatments and outcomes) or scientific data (such as large sets of experimental measurements). The vast quality of data contain enough information to answer questions of importance to the application users.