

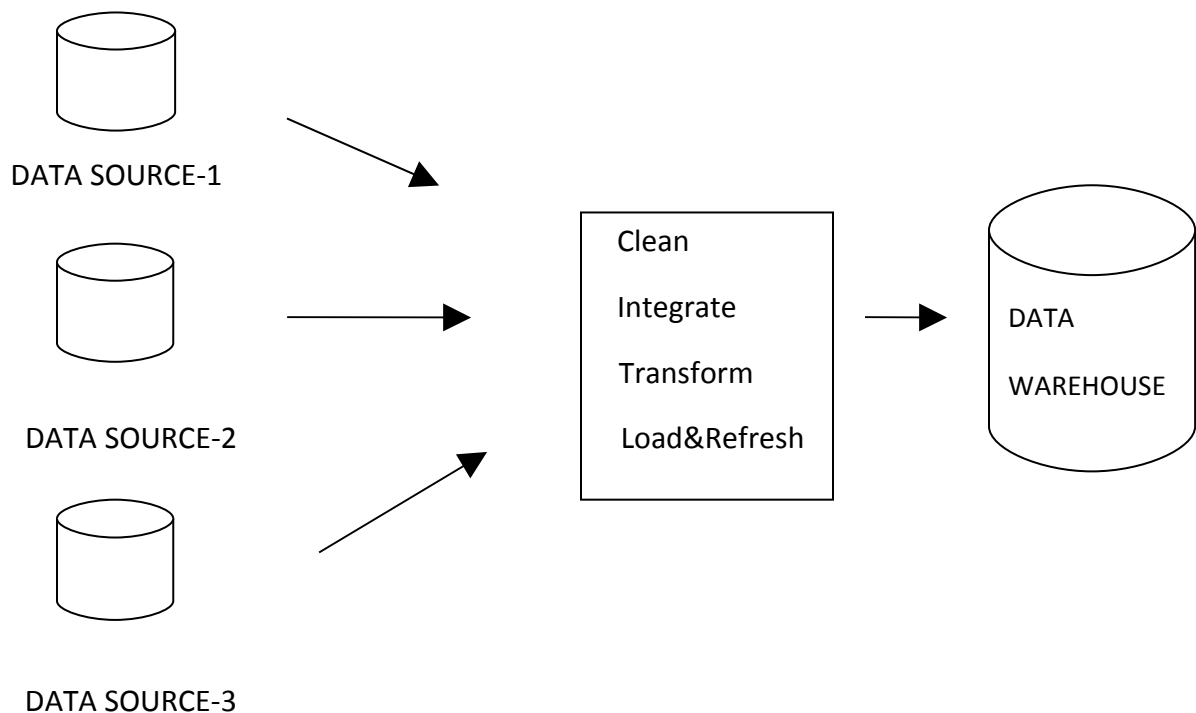
UNIT-1

(INTRODUCTION AND DATA WAREHOUSING)

1. Define data warehousing?

->Repository of information collected from multiple heterogenous datasources,stored under a unified schema ,and that usually resides at a single site.

->Data warehouse are constructed via a process of datacleaning,data integration,data transformation,data loading and periodic data refreshing.



2. Mention the uses of date warehouse?

->Provides an integrated and total view of the enterprise.

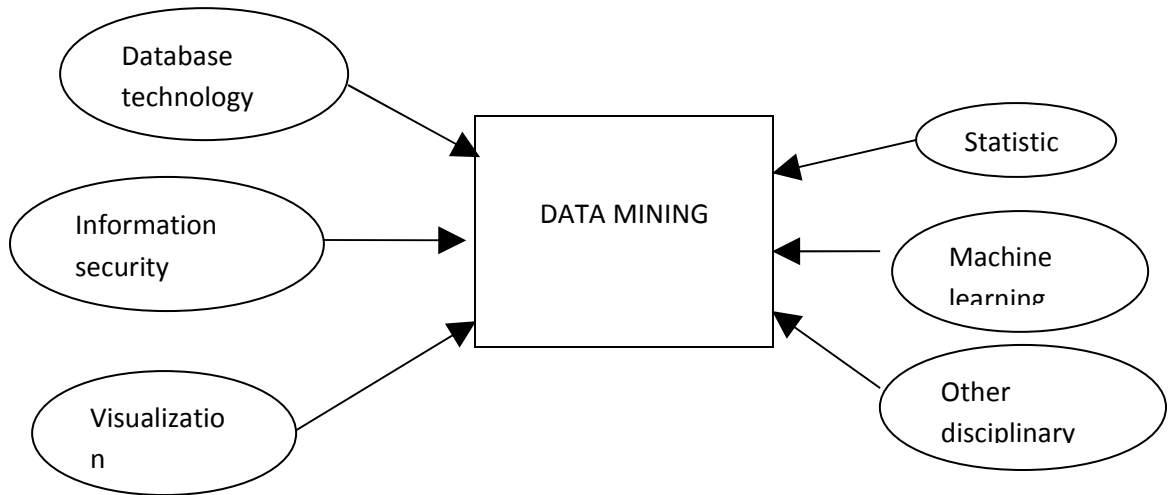
->provides architecture and tool for business executives to systematically,organize,understand&use the data to make decision management systems.

->Makes the enterprises current historical information easily available for decisionmaking.

3. Define data mining?

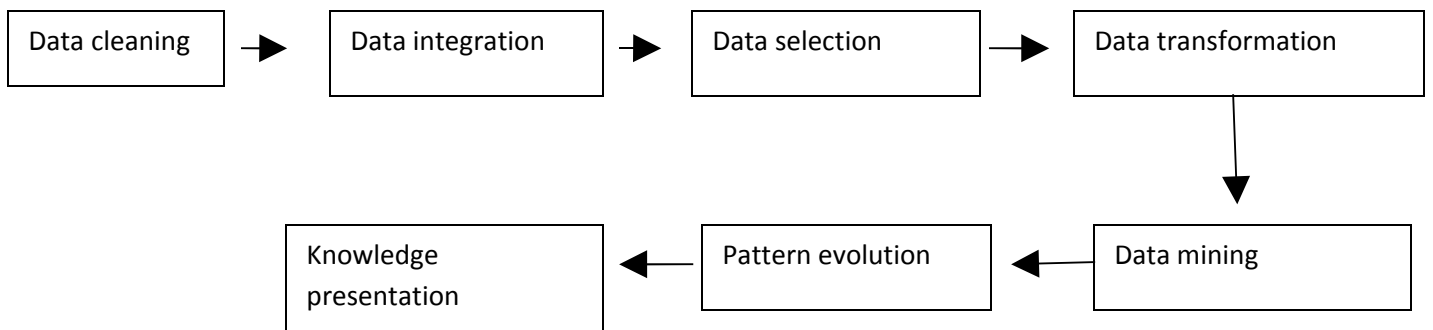
->Refers to extracting or " mining" knowledge from large amounts of data. In other terms, mining of knowledge from databases, data warehouses and any other information repository.

->Data mining popularly terms as "KDD"(KNOWLEDGE DISCOVERY IN DATABASES).



Write steps in data mining process?

- >Data cleaning
- >Data integration
- >Data selection
- >Data transformation
- >Data mining
- >Pattern Evolution
- >Knowledge presentation.



4. Explain the different types of data repositories on which mining can be performed?

The different types of data repositories on which mining can be performed are

- >Relational Databases
- >Data Warehouses
- >Transactional Databases
- >Advanced Databases
- >Flat files
- >World Wide Web

5. What are the characteristics of data warehouse?

- >Separate
- >Available
- >Integrated
- >Subject Oriented
- >Not Dynamic(Non-Volatile)
- >Consistency
- >Iterative Development
- >Aggregation Performance
- >Time-variant

6. Difference between operational database(OLTP)&Data warehouse(OLAP)

FEATURES	OLTP	OLAP
Orientation	Customer oriented	Market oriented
User	Clerk,IT professional	Knowledge workers,analyst,managers,executives
Function	Day-Day operation	Decision support
Database design	ER-model,application oriented model	Star&snowflake model,Subject oriented model
Focus	Focuses on current data within an enterprise or department	Multiple heterogenous data.
Access	Requires concurrency control&recovery mechanism	Read only operation.

7. Difference between datamarts and data warehouse?

Data warehouse	Data mart
Repository of information collected from multiple heterogeneous data sources, stored under a unified schema, and that usually resides at a single site.	Data mart is a pragmatic collection of related facts, but does not have to be exhaustive or exclusive. A datamart is both a kind of subject area and an application. Data mart is a collection of numeric facts.
Union of all data marts	A single business process
Co-operate or Enterprise wide	Departmental-wide
Collects information about subjects that span an entire organization.	Department subset of a data warehouse.
Structure for corporate view of data	Structure to suite the departmental view of data

8. What are the advantages of a data modelling tool?

- ~ Integrates the datawarehouse model with other corporate data models.
- ~ Helps assure consistency in naming.
- ~ Creates good documentation in a variety of useful formats.
- ~ Provides a reasonably intuitive user interface for entering comments about objects.

9.. What is datawarehouse performance issue?

The performance of a data warehouse is largely a function of the quantity and type of data stored within a database and the query/data loading work load placed upon the system.

10. What are the types of performance issue?

- >Capacity planning for the data warehouse
- >data placement techniques within a data warehouse
- >Application Performance Techniques.
- > Monitoring the Data Warehouse.

11. What is Data Inconsistency Cleaning?

This can be summarized as the process of cleaning up the small inconsistencies that introduce themselves into the data. Examples include duplicate keys and unreferenced foreign keys.

12. What is Column Level Cleaning?

This involved checking the contents of each column field and ensuring it conforms to a set of valid values. For example convert EBCDIC to ASCII.

13. What is Bottleneck Detection?

Bottleneck Detection is the process where by the database administrator will detect for what reason the database performance level has reached a plateau. To increase the performance from this plateau may require the addition of more hardware resources or reconfiguration of the system software(O/S, RDBMS or Application).

14. What is back room Meta data?

Back room Meta data is process related, and it guides the extraction, cleaning and loading process.

15. What is Front Room Meta data?

It is more descriptive and it helps query tools and report writers function smoothly.

16. What are the specifications of source system Meta data?

- >Repositories
- >Source Scheme
- > Copy Books
- > Spread Sheet Sources
- > Lotus notes database

17. What is active Meta data?

Active Meta data is Meta data that drives a process rather than documents.

18. What is Meat data catalogue?

Meat data catalogue is a single common storage point for information that drives the entire warehouse process.

19. Why do you need data warehouse life cycle process?

Data warehouse life cycle approach is essential because it ensures that the project pieces are brought together in the right order and at the right time.

20. What are the steps in the life cycle approach?

- >Project Planning
- >Business Requirements definition
- >Data track: Dimensional modeling, Physical Design, Data Staging Design & Development
- >Technology track: Technical Architecture design, Product Selection & Installation
- >Application track: End user Application Specification, End user Application Development
- >Deployment
- >Maintenance & Growth
- >Project Management

21. Name three-tier datawarehouse architecture.

Bottom tier-Warehouse database server

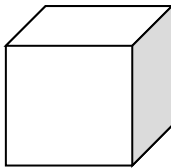
Middle tier-OLAP server

Top tier-Front end client layer

22. What is multidimensional data model?

A data warehouse is modeled by a multidimensional database structure called multidimensional data cube where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure.

It is used for the design of corporate data warehouse and departmental data marts. In this model, a star, snowflake schema or fact constellation schema is adopted.



23. What is a data cube?

-> Data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

-> It consists of a lattice of cuboids, each corresponding to a different degree of summarization of the given multi-dimensional data.

24. What is dimension?

Dimensions are the perspectives or entities with respect to which an organization wants to keep records. Each dimension may have a table associated with it, called a dimension table.

For example: All electronics may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch, and location.

25. What is fact?

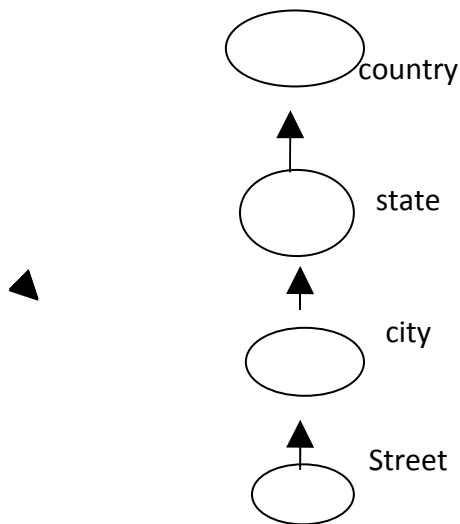
Facts are numerical measures. Examples of facts for a sales data warehouse include dollars_sold, units_sold, and amount_budgeted.

A fact table contains the names of the facts, or measures, as well as keys to each of the related dimensional levels of tables.

26. What is concept hierarchies?

A concept hierarchy defines a sequence of mappings from set of low-level concepts to higher-level, more general concepts.

-> it organizes the values of attributes or dimension into gradual levels of abstraction. They are useful in mining at multiple levels of abstraction.



27. In the context of data warehousing what is data transformation?

In data transformation, the data are transformed or consolidated into forms appropriate for mining.

Data transformation can involve the following:

- => smoothing
- => aggregation
- => generalization
- => normalization
- => attribute construction

28. Explain slice and dice operation, roll up, drill-down & pivot operation?

SLICE OPERATION:

The slice operation performs a selection on one dimension of the given cube, resulting in a sub cube.

DICE OPERATION:

It defines subcubes by performing a selection on two or more dimension.

ROLL UP :(drill up)

Roll up operation performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.

DRILL DOWN:

It is a reverse of roll up. It navigates from less detailed data to more detailed data and it can be realized by either stepping down by a concept hierarchy for a dimension or introducing additional dimensions.

PIVOT OPERATION:(Rotate)

It is a visualization operation that rotates the data axes in view in order to provide alternative presentation of the data.

29. Define data discrimination?

It is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

The output of data discrimination can be presented in various forms like:

Pie charts

Bar charts

Curves

Multi dimensional cubes

Multi dimensional tables

30. Define data characterization ?

It is summarization of the general characteristics or features of a target class of data. The output of data characterization can be presented in various forms like:

Pie charts

Bar charts

Curves

Multi dimensional cubes & tables

31. Define association?

It is the discovery of association rules showing attribute value conditions that occur frequently together in a given set of data. For example $X \Rightarrow Y$ where X & Y are the set of item.

For example: buys (X, "computer") \Rightarrow buys (X, "software")

32. Differentiate classification, prediction, clustering?

CLASSIFICATION	PREDICTION	CLUSTERING
It is the process of finding a model that describes and distinguishes data classes or concepts.	It refers to both data value prediction and class label prediction.	It is the method of grouping the data into different groups, so that data in each group share similar trends and patterns.
It analysis the training data set and constructs a model based on the class label and assigns a class label to the future unlabeled records.	It is used to predict missing or unavailable data values rather than class labels.	To uncover natural groupings To initiate hypothesis about the data. To find consistent and valid organization of the data.
Models like decision tree, neural networks genetic algorithm etc...	It can be modeled by statistical techniques of regression.	Based on the principles of maximizing intra-class similarity & minimizing interclass similarity.

UNIT-2

(DATA PREPROCESSING, LANGUAGE ARCHITECTURE, CONCEPT DESCRIPTION)

1. Why preprocessing or what is the need for preprocessing the data?

Incomplete, noisy and inconsistent data are common place properties of large real world database and data warehouses due to the typically large size and likely multiple heterogenous data sources, so need preprocessing of data.

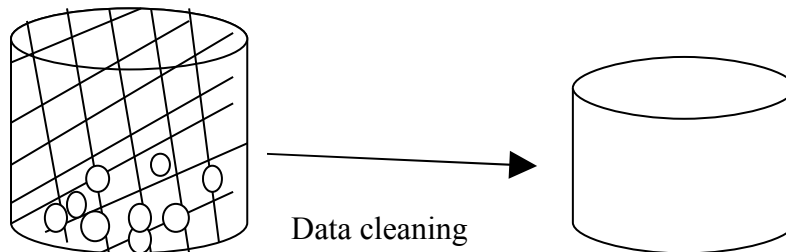
Preprocessing helps to improve the quality of the data, improve the efficiency and ease of mining process.

2. Mention the various tasks to be accomplished as a part of data preprocessing?

- >Data cleaning
- >Data integration
- >Data transformation
- >Data reduction

3. What is data cleaning?

Data cleaning routines attempt to fill in missing values, smooth out noises while identifying outliers and correct inconsistencies in the data



4. Write data mining primitives?

- >Task relevant data
- >Kin of knowledge
- >Background of knowledge
- >interestingness measures and threshold
- >Representation of visualizing

5. Define binning?

Binning is a top down splitting technique based on a specified number of bins. This method are also used as a discretization methods for numerosity reduction and concept hierarchy. Binning does not use class information and it is sensitive to the user specified number of bins as well as presence of outliers.

6. Define data discretization?

It can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute in to intervals.

Categories of data discretization

→Supervised

→Unsupervised

→Bottom up

→Top down

7. Define data reduction?

It can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintain the integrity of the original data.

Strategies for data reduction are

→data cube aggregation

→attribute subset selection

→dimensionality reduction

→numerosity reduction

→discretization and concept hierarchy

8. What is the need for data mining query language?

In an inductive database, ordinary data query can be used to access and manipulate data while inductive queries can be used to generate, manipulate and applied pattern.

KDD becomes an extended querying processes where the analyzed can control the whole process since he or she specified the data and pattern of interest.

DMQL consists of well accepted set of primitives.

9. What are the functional components of GUI in data mining?

- Data collection and mining query composition
- Presentation of discovered patterns
- Hierarchy specification and manipulation
- Manipulation of data primitives
- Intractive multilevel mining
- Other miscellaneous information

10. Define dimensionality reduction and numerosity reduction?

Dimensionality reduction:

Data encoding or transformations are applied so as to obtain a reduced or compressed representation of the original data.

Numerosity reduction:

It is used to reduce the data volume by choosing alternative, smaller forms of representation.

11. Define knowledge representation

Knowledge representation techniques are used to present the mined knowledge to the user.

12. What is Visualization?

Visualisation is for depiction of data and to gain intuition about data being observed. It assists the analysts in selecting display formats, viewer perspectives and data representation schema

13. Name some conventional visualization techniques

- Histogram
- Relationship tree
- Bar charts
- Pie charts
- Tables etc.

14. What is Descriptive and predictive data mining?

Descriptive datamining describes the data set in a concise and summarative manner and presents interesting general properties of the data. Predictive datamining analyzes the data in order to construct one or set of models and attempts to predict the behavior of new data sets.

15. What is Data Generalization?

It is process that abstracts a large set of task-relevant data in a database from a relatively low conceptual to higher conceptual levels

2 approaches for Generalization

- 1) Datacube approach
- 2) Attribute-oriented induction approach

16. Define Attribute Oriented Induction

These method collects the task-relevant data using a relational database query and then perform generalization based on the examination in the relevant set of data.

17. What do you meant by concept hierarchies?

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Concept hierarchies allow specialization, or drilling down where by concept values are replaced by lower-level concepts.

18. Define support.

Support is the ratio of the number of transactions that include all items in the antecedent and consequent parts of the rule to the total number of transactions. Support is an association rule interestingness measure.

19. Define Confidence.

Confidence is the ratio of the number of transactions that include all items in the consequent as well as antecedent to the number of transactions that include all items in antecedent. Confidence is an association rule interestingness measure.