KARPAGAM ACADEMY OF HIGHER EDUCATION

Coimbatore – 641 021.

(For the candidates admitted from 2015 onwards)

DEPARTMENT OF COMPUTER SCIENCE, CA & IT

Subject Name: DATA MINING AND WAREHOUSING
Subject Code: 15CSU505A
Semester : V
Class : III B.Sc (CS)

SYLLABUS

		L	T	P	C
15CSU505A	**DATA MINING AND WAREHOUSING**	**5**	**0**	**0**	**5**

COURSE OBJECTIVE:

This course introduce students to the basic concepts and techniques of Data Mining, develop skills of using recent data mining software for solving practical problems and gain experience of doing independent study and research.

COURSE OUTCOME:

- To introduce students to the basic concepts and techniques of Data Mining.
- To develop skills of using recent data mining software for solving practical problems.
- To gain experience of doing independent study and research.
- Possess some knowledge of the concepts and terminology associated with database systems, statistics, and machine learning

UNIT I

Data Mining: Introduction: Basic Data Mining Tasks – Data Mining versus Knowledge Discovery in Database Mining Issues and Mechanisms. Data Mining Techniques: Statistical Perspective on Data Mining – Similarity Measures – Decision Trees – Neural Networks – Genetic Algorithms.

UNIT II

Classifications: Bayesian Classification – Distance Based Algorithms – K-Nearest Neighbor. Clustering: K-Means Clustering – Clustering with Genetic Algorithms – Clustering with Neural Networks. Association Rules – Basic Algorithms – Parallel and Distributed Algorithms – Comparing Approaches – Generalized and Multilevel Association Rules. Web Mining: Web Content Mining: Personalization.

UNIT III

Data Warehousing: Introduction – Architecture – System Process-Process Architecture. Design: Database Schema – Partitioning Strategy – Aggregations – Data Marting – Meta Data.

UNIT IV

Hardware and Operational Design : Hardware Architecture – Physical layout – Security –Backup and Recovery – Service Level Agreement – Operating and Data Warehousing.

UNIT V

Capacity planning – Tuning and Data Warehouse – Testing and Data Warehouse – Data Warehouse Futures. Application: Data warehousing and data mining in government: Introduction-national data warehouses-other areas for data warehousing and data mining.

TEXT BOOK

1. Sam Anahory and Dennis Murray. 2009. Data Warehousing in the Real World, Pearson Education, New Delhi.

REFERENCES

1. Margaret H. Dunham. 2004. Data Mining Introductory and Advanced Topics, Pearson Education, 2004.
2. Pieter Adriaans, Dolf Zantinge. 1998. Data Mining, Addison Wesley.
3. Jiawei Han and Micheline Kamber. 2006. Data Mining – Concepts and Techniques, 1st Edition, Morgan Kaufmann Publishers, Mumbai.

WEB SITES

1. Thedacs.Com
2. Dwreview.Com
3. Pcai.Com
4. Eruditionhome.Com

END SEMESTER MARK ALLOCATION

1.	**PART – A** **20*1=20** **ONLINE EXAMINATION**	**20**
2	**PART – B** **5*8=40** **EITHER OR TYPE**	**40**
3	**TOTAL**	**60**

KARPAGAM ACADEMY OF HIGHER EDUCATION

DEPARTMENT OF CS,CA & IT

III BSc CS BATCH 2015-2018

DATA MINING AND WAREHOUSING(15CSU505A)

SEMESTER V/ LECTURE PLAN

UNIT I

S.NO	Lecture Duration (Hours)	Topics To Be Covered	Support Materials/ Pg.No
1	1	Data Mining: Introduction	T1: 3-5, R2: 5-9
2	1	Basic Datamining Tasks	T1: 5-9, W3
3	1	Data Mining Versus Knowledge Discovery in Databases	T1: 9-14, W3
4	1	Data Mining Issues and Mechanisms	T1:14-16, W3
5	1	Data Mining Techniques: Introduction, A Statistical perspective on Datamining	T1: 46-51, W3
6	1	Models based om summarization, Bayes theorem	T1:51-54,W1
7	1	Hypothesis Testing, regression and Correlation	T1: 54-57,R1
8	1	Hypothesis Testing, regression and Correlation [Cont..]	T1: 54-58
9	1	Similarity Measures	T1: 57-58
10	1	Similarity Measures[Cont..]	T1: 57-59
11	1	Decision Trees	T1: 58-61, W3
12	1	Decision Trees[Cont..]	T1: 58-61, W3
13	1	Neural Networks, Genetic Algorithm	T1: 61-70,W1
14	1	Neural Networks, Genetic Algorithm[Cont..]	T1: 61-70
15	1	Recapitulation and Discussion of Important Questions	
Total No of Hours Planned for Unit I-15			
T1		Margaret H.Dunham, 2004 Datamining Introductory and Advanced Topics, Pearson Education, 2004.	
R1		Pieter Adriaans, Dolf Zantinge, 1998, datamining, Addison Wesley.	
W1		http://www.dwreview.com/DW_Overview.html	
W3		http://www.tutorialspoint.com/data_mining/	
		UNIT II	
S.NO	**Lecture Duration (Hours)**	**Topics To Be Covered**	**Support Materials/ Pg.No**
1	1	Classifications: Bayesian Classification	T1: 86-89, W3
2	1	Distance based algorithms, K-Nearest neighbor	T1:89-92
3	1	Distance based algorithms, K-Nearest neighbor [Cont..]	T1:89-92
4	1	Clustering:K-means Clustering	T1:140-141
5	1	Clustering with Genetic algorithms, Clustering with Neural Networks	T1: 146-149
6	1	Clustering with Genetic algorithms, Clustering with Neural Networks [Cont..]	T1: 146-149,R1
7	1	Association rules: Introduction, Basic Algorithms- Apriori Algorithm	T1: 164-173
8	1	Association rules: Introduction, Basic Algorithms- Apriori Algorithm [Cont..]	T1: 164-173

9	1	Sampling Algorithm, Partitioning	T1: 173-178,W2
10	1	Sampling Algorithm, Partitioning [Cont..]	T1: 173-178
11	1	Parallel and Distributed Algorithm, Comparing approaches	T1:178-184
12	1	Generalized and Multilevel association rules	T1: 184-185
13	1	Web Mining : Introduction, Web Content Mining- Personalization	T1: 202-204, W3
14	1	Web Mining : Introduction, Web Content Mining- Personalization [Cont..]	T1: 202-204, W3
15	1	Recapitulation and Discussion of Important Questions	
Total No of Hours Planned for Unit II-15			
T1		Margaret H.Dunham, 2004 Datamining Introductory and Advanced Topics, Pearson Education, 2004.	
R1		Pieter Adriaans, Dolf Zantinge, 1998, datamining, Addison Wesley.	
W2		http://www.pcai.com/web/ai_info/data_warehouse_mining.html	
W3		http://www.tutorialspoint.com/data_mining/	
		UNIT III	
S.NO	**Lecture Duration (Hours)**	**Topics To Be Covered**	**Support Materials/ Pg.No**
1	1	Data warehouse: Introduction, Architecture	W4
2	1	System Process	T2: 33-42,W4
3	1	Process Architecture	T2: 43-51
4	1	Cont.. Process Architecture	T2: 52-60
5	1	Cont.. Process Architecture	T2: 61-67, W4
6	1	Design: Database scheme	T2: 71-80,W3
7	1	Cont.. Database scheme	T2: 80-90
8	1	Cont.. Database scheme	T2: 90-98
9	1	Cont.. Database scheme	T2: 98-100
10	1	Partitionong Strategy: Introduction, Horizontal Partitioning	T2: 101-106
11	1	Cont.. Partitioning Strategy	T2: 106-108
12	1	Cont.. Partitioning Strategy	T2: 108-114
13	1	Aggregations: Introduction, Why Aggregate, What is an Aggregation?	T2: 115-118
14	1	Cont.. Aggregations	T2: 119-129
15	1	Data Marting	T2: 130-138
16	1	Meta data	T2: 139-148
17	1	Recapitulation and Discussion of Important Questions	
Total No of Hours Planned for Unit III-17			
T2		Sam Anahory and Dennis Murray,2009.Datawarehousing in the Real World, Pearson Education, New Delhi.	
W3		http://www.tutorialspoint.com/data_mining/	
W4		http://www.tutorialspoint.com/dwh/	
		UNIT IV	
S.NO	**Lecture Duration (Hours)**	**Topics To Be Covered**	**Support Materials/ Pg.No**

1	1	Hardware & Operational Design: Hardware Architecture- Introduction, process, Server Hardware	T2: 169-176
2	1	Cont.. Hardware & Operational Design	T2: 176-189
3	1	Physical Layout- Parallel Techniques, Disk Technology- RAID Technique, Size & Speed, Databse Layout	T2: 189-195
4	1	Cont.. Physical Layout	T2: 195-201
5	1	Security: Introduction, requirements, Audit Requirements	T2: 202-210, W4
6	1	Cont.. Security	T2:210-218
7	1	Cont.. Security	T2:210-218
8	1	Backup Recovery- Introduction, Definition, Hardware, Software	T2: 219-229
9	1	Backup Strategies, testing the Strategy, Disaster recovery	T2: 229-235
10	1	Service level agreement	T2: 236-243,R1
11	1	Service level agreement [Cont…]	T2: 236-243
12	1	Operating and Data warehousing	T2: 244-251,R2
13	1	Operating and Data warehousing [cont…]	T2: 244-251
14	1	Recapitulation and Discussion of Important Questions	
Total No of Hours Planned for Unit IV-14			
T2		Sam Anahory and Dennis Murray,2009.Datawarehousing in the Real World, Pearson Education, New Delhi.	
R1		Pieter Adriaans, Dolf Zantinge, 1998, datamining, Addison Wesley.	
R2		Pieter Adriaans, Dolf Zantinge, 1998, datamining, Addison Wesley.Publishers, Mumbai.	
W4		http://www.tutorialspoint.com/dwh/	
		UNIT V	
S.NO	**Lecture Duration (Hours)**	**Topics To Be Covered**	**Support Materials/ Pg.No**
1	1	Capacity Planning	T2: 255-266
2	1	Tuning the Data Warehouse	T2: 267-277,W4
3	1	Testing the Data Warehouse- Introduction, Developing the test plan, testing Backup recovery, testing the Operational environment	T2: 278-284,W4
4	1	Testing the Data Warehouse- Introduction, Developing the test plan, testing Backup recovery, testing the Operational environment	T2: 278-284,W4
5	1	Testing the Database, Testing the Application	T2: 284-288
6	1	Logistics of the test	T2: 291-293
7	1	Datawarehouse Futures, Application- Datawarehousing and datamining in government- introduction	T2:450-455
8	1	Datawarehouse Futures, Application- Datawarehousing and datamining in government- introduction [Cont…]	T2:450-455
9	1	National data warehouses, Other areas for datawarehousing and datamining	T2:502-510
10	1	National data warehouses, Other areas for datawarehousing and datamining [Cont..]	T2:502-510
11	1	Recapitulation and Discussion of Important Questions	
12	1	Revision-Previous year ESE question papers	
13	1	Revision-Previous year ESE question papers	
14	1	Revision-Previous year ESE question papers	
Total No of Hours Planned for Unit V-14			
T2		Sam Anahory and Dennis Murray,2009.Datawarehousing in the Real World, Pearson Education, New Delhi.	

W4	http://www.tutorialspoint.com/dwh/

UNIT I

SYLLABUS

Data Mining: Introduction: Basic Data Mining Tasks – Data Mining versus Knowledge Discovery in Database Mining Issues and Mechanisms. Data Mining Techniques: Statistical Perspective on Data Mining – Similarity Measures – Decision Trees – Neural Networks – Genetic Algorithms.

What is Data mining?

Data Mining is defined as the procedure of extracting information from huge sets of data. Data mining is mining knowledge from data.

Data mining is often defined as finding hidden information in a database. Alternatively it has been called exploratory data analysis, data driven discovery, and deductive learning.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

Database Processing vs. Data Mining Processing

Query	Query
-Well defined	- Poorly defined
- SQL	-No precise query language
Data	
- Operational Data	- Not operational data
Output	
- Precise	- Fuzzy
- Subset of database	- Not a subset of database

Examples

Data base:

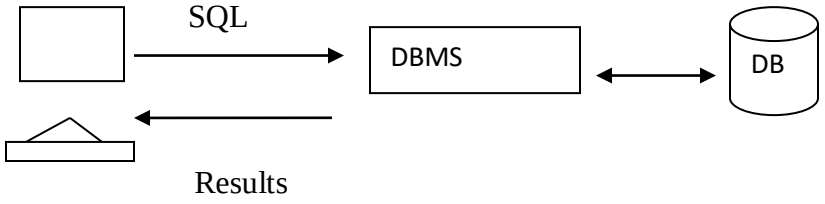
Find all credit applicants with last name of Smith. Identify customers who have purchased more than $10,000 in the last month. Find all customers who have purchased milk.

Data mining:

Find all credit applicants who are poor credit risks. (Classification)

Identify customers with similar buying habits.(Clustering)

Find all items which are frequently purchased with milk. (Association rules)

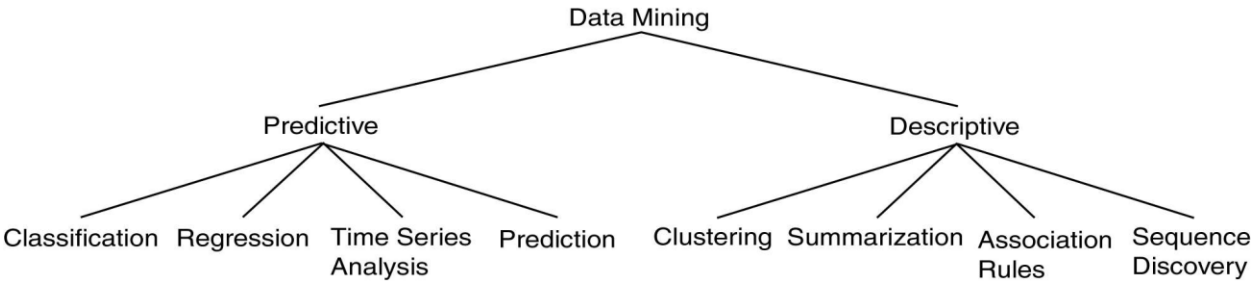


Data mining algorithms can be characterized as consisting of three parts:

Model: the purpose of the algorithm is to fit a model to the data.

Preference: Some criteria must be used to fit one model over another.

Search: All algorithms require some technique to search the data.



Predictive Model:

It makes a prediction about values of data using known results found from different data. Predictive model may be made based on the use of other historical data.

Descriptive Model:

It identifies patterns or relationship in data. Unlike predictive model, a descriptive model serves as a way to explore the properties of the data examined not to predict new properties.

Basic Data mining tasks:

Classification:

- Classification maps data into predefined groups or classes.
- It is often referred to as supervised learning because the classes are determined before examining the data.
- Pattern recognition is a type of classification where an input pattern is classified into one several classes based on its similarity to these predefined classes.

Regression:

- Regression is used to map a data item to a real valued prediction variable.
- Regression assumes that the target data fit into some known type of function and then determines the best function of this type that models in the given data.

Time series analysis:

- With time series analysis the value of an attribute is examines as it varies over time. The values are obtained as evenly space time points.
- There are three basic functions performed in time series analysis. In one case, distance measures are used to determine the similarity between different time series.
- In second case, the structure of the line is examined to determine its behavior.
- In third application would be to use the historical time series plot to predict future values.
- Example: Stock Market

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.

Clustering:

- Clustering is similar to classification except that the groups are not rather defined by the data alone.
- Clustering is alternatively referred to as unsupervised learning or segmentation.
- Clustering groups similar data together into clusters.
- A special type of clustering is called segmentation. With segmentation a database is partitioned into disjointed groupings of similar tuples called segments.

Summarization

It maps data into subsets with associated simple descriptions. It extracts or derives representative information about the database. It is also called as Characterization, Generalization

Association rules:

Link Analysis uncovers relationships among data. It is also referred to as Affinity Analysis or Association Rules, refers to the data mining task of uncovering relationship among data. An association rule is a model that identifies specific types of data associations.

Sequence discovery:

It is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data are found to be related, but the relationship is based on time. It determines sequential patterns.

Data mining Vs Knowledge Discovery in Databases:

The terms Knowledge discovery in databases and data mining are often used interchangeably. Many other names given to this process of discovering useful hidden patterns in data: knowledge extraction,

Information discovery,

Exploratory data analysis,

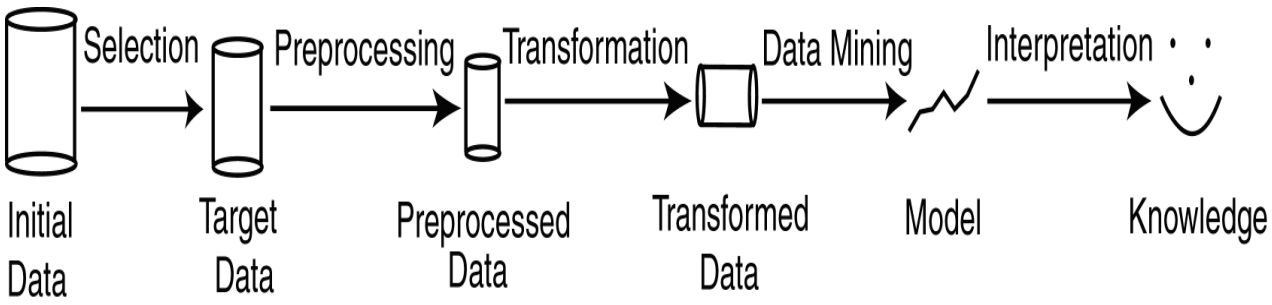
Information harvesting and

Unsupervised pattern recognition.

Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data.

Data Mining: Use of algorithms to extract the information and patterns derived by the KDD process

The KDD process consists of the following five steps:



Selection: The data needed for the data mining process may be Obtain data from various data sources.

Preprocessing: The data to be used by the process may have incorrect or missing data. Erroneous data may be corrected or removed, whereas missing data must be supplied or predicted.

Transformation: Data from different sources must be converted into a common format for processing.

Data Mining: This step applied algorithms to the transformed data to generate the desired results.

Interpretation/Evaluation: Present results to user in meaningful manner.

Visualization refers to the visual presentation of data. The use of visualization techniques allows use to summarize, extract and grasp more complex results than more mathematical or text type descriptions of the results.

Visualization techniques include:

Graphical: Bar charts, pie charts, histograms, and line graphs.

Geometric: Box Plot, Scatter diagram

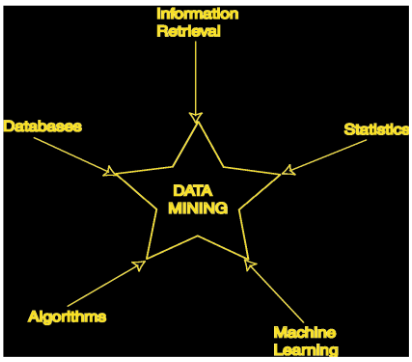
Icon based: Figures, colors, or other icons

Pixel based: uniquely colored pixel

Hierarchical: Divide display area into regions

Hybrid: Preceding approaches can be combined into one display.

Data mining development:



Different views of what datamining function actually are:

Induction: It used to proceed from very specific knowledge to more general information. This type of technique is often found in AI application.

Compression: The primary objective of data mining is to describe some characteristic of a set of data by a general model; this approach can be viewed as a type of compression. Here the detailed data within database are abstracted and compressed to a smaller description of the data characteristics that are found in the model.

Querying: The data mining process itself can be viewed as a type of querying the underline database.

Approximation: Describing a large dataset can be viewed as using approximation to help uncover hidden information about the data.

Search: When dealing with large database, the impact of size and efficiency of developing an abstract model can be thought of as a type of search problem.

Data mining Issues:

1. **Human Interaction**: data mining problems are often precisely stated interfaces may be needed with both domain and technical experts. Users are needed to identify training data and desired results.

2. **Over fitting**: When a model is generated that is associated with a given database state, it is desirable that the model also fit future database states. Over fitting occurs when the model does not fit future states.

3. **Outliers**: There are often many data entries that do not fit nicely into a derived model. This becomes even more of an issue with large databases.

4. **Interpretation of results**: data mining output may require experts to correctly interpret the results, which might otherwise be meaningless to the average database user.

5. **Visualization of results**: To easily view and understand the output of data mining algorithms, visualization of the results is helpful.

6. **Large Datasets**: The massive datasets associated with data mining create problems when applying algorithms designed for small datasets.

7. **High Dimensionality**: A conventional database schema may be composed of many different attributes. The problem here is that not all attributes may be needed to solve a given data mining problem.

8. **Multimedia Data**: Most previous data mining algorithms ate targeted to traditional data types. The use of multimedia data such as is found in GIS database complicates or invalidates many proposed algorithms.

9. **Missing data**: During the preprocessing phase of KDD, missing data may be replaced with estimates. This and other approaches to handling missing data can lead to invalid results in the data mining step.

10. **Irrelevamt data**: Some attributes in the database might not be of interest to the data mining task being developed.

11. **Noisy data**: Some attribute values might be invalid or incorrect.

12. **Changing data**: databases cannot be assumed to be static. This requires that the algorithm be completely rerun anytime the database changes.

13. **Integration**: KDD process is not currently integrated into normal data processing activities. KDD requests may be treated as special, unusual, one time needs. This makes them ineffective, inefficient and not general enough to be used on an ongoing basis.

14. **Application**: the intended use for the information obtained from the data mining function is a challenge. The data are of a type that has not previously been known, business practices may have to be modified to determine how to effectively use the information uncovered.

Data mining technique:

A statistical perspective on data mining:

1. Point Estimation:

Point Estimate: It refers to the process of estimate a population parameter.

May be made by calculating the parameter for a sample.

May be used to predict value for missing data.

The bias of an estimator is the difference between the expected value of the estimator and the actual value:

$$Bias = E(\hat{\Theta}) - \Theta$$

An unbiased estimator is one whose bias is 0.

Mean Squared Error:

Expected value of the squared difference between the estimate and the actual value:

$$MSE(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2$$

Root Mean Square:

RMS may also be used to estimate error or as another statistic to describe a distribution. The RMS can be used for this purpose. Given a set of n values $X=\{x_1 \ldots\ldots x_n\}$.

Jackknife estimate:

It estimate of parameter is obtained by omitting one value from the set of observed values.

Ex: estimate of mean for X={x1, … , xn}

$$\hat{\mu}_{(i)} = \frac{\sum_{j=1}^{i-1} x_j + \sum_{j=i+1}^{n} x_j}{n-1}$$

$$\hat{\theta}_{(.)} = \frac{\sum_{j=1}^{n} \hat{\theta}_{(j)}}{n}$$

Maximum Likelihood Estimate

Obtain parameter estimates that maximize the probability that the sample data occurs for the Specific model. Joint probability for observing the sample data by multiplying the individual probabilities. Likelihood function:

$$L(\Theta \mid x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i \mid \Theta)$$

Expectation Maximization:

Solves estimation with incomplete data. Obtain initial estimates for parameters. Iteratively use estimates for missing data and continue until convergence.

EM Algorithm:

Input:
$\Theta = \{\theta_1, ..., \theta_p\}$ //Parameters to be Estimated
$X_{obs} = \{x_1, ..., x_k\}$ //Input Database Values Observed
$X_{miss} = \{x_{k+1}, ..., x_n\}$ //Input Database Values Missing
Output:
$\hat{\Theta}$ //Estimates for Θ
EM Algorithm:
i := 0;
Obtain initial parameter MLE estimate,$\hat{\Theta}^i$;
repeat
Estimate missing data,$\hat{X}^i_{miss}$;
i++;
Obtain next parameter estimate,$\hat{\theta}^i$ to maximize data;
until estimate converges;

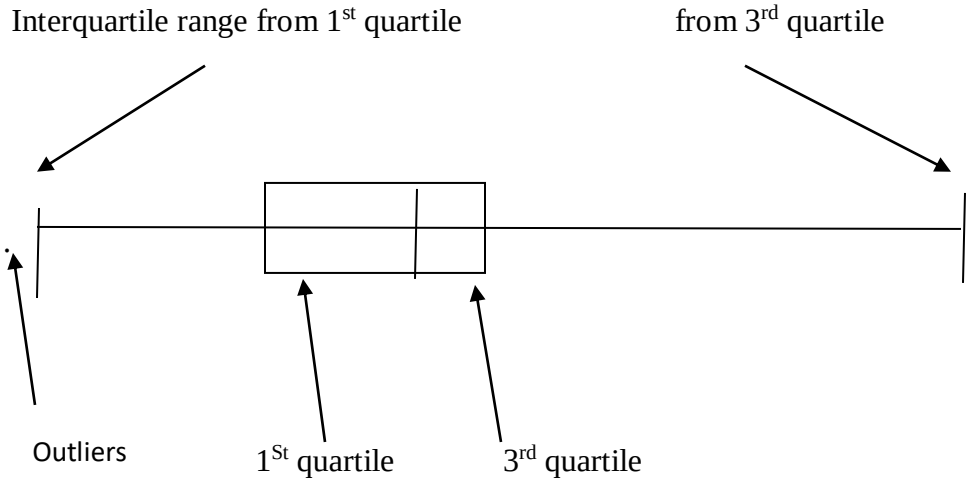
2. Models Based on Summarization:

There are many basic concepts that provide an abstraction and summarization of the data as a whole. Fitting a population to a specific frequency distribution provides an even better model of the data.

There are also many well known techniques to display the structure of the data graphically.

For ex,A histogram shows the distribution of the data. A box plot more sophisticated technique that illustrates several different features of the population at once.

Smallest value within 1.5 largest values within 1.5 interquartile ranges

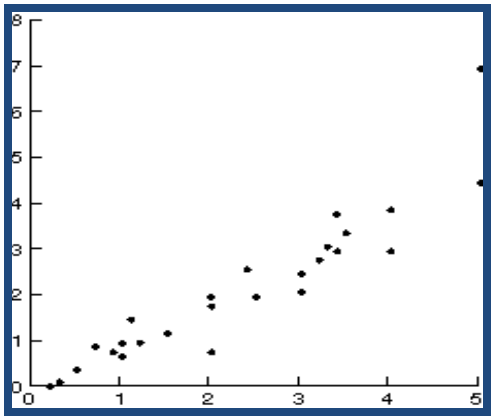


Box plot example

The total range of the data values is divided into four equal parts called quartiles. The box in center of the figure shows the range between first, second, third quartiles.

The lines in the box show the median. The lines exceeding from either end of the box are the values that are a distance of 1.5 of the interquartile range from the first and third quartiles.

Scatter Diagram:



3. Bayes theorem:

Given a set of data X={x_1…. X_n} a data mining problem is to uncover properties of the distribution from which the set comes.

Here P(h1|xi) is called the **Posterior Probability.** While P(h1) is the prior probability associated with hypothesis h_1.

$$P(x_i) = \sum_{j=1}^{m} P(x_i \mid h_j)\ P(h_j).$$

$$P(h_1 \mid x_i) = \frac{P(x_i \mid h_1)\ P(h_1)}{P(x_i)}.$$

Assign probabilities of hypotheses given a data value.

Credit authorizations (hypotheses): h1=authorize purchase, h2 = authorize after further identification, h3=do not authorize, h4= do not authorize but contact police

Assign twelve data values for all combinations of credit and income:

	1	2	3	4
Excellent	x1	x2	x3	x4

Good	x5	x6	x7	x8
Bad	x9	x10	x11	x12

From training data: P(h1) = 60%; P(h2)=20%; P(h3)=10%; P(h4)=10%.

ID	Income	Credit	Class	x_i
1	4	Excellent	h_1	x_4
2	3	Good	h_1	x_7
3	2	Excellent	h_1	x_2
4	3	Good	h_1	x_7
5	4	Good	h_1	x_8
6	2	Excellent	h_1	x_2
7	3	Bad	h_2	x_{11}
8	2	Bad	h_2	x_{10}
9	3	Bad	h_3	x_{11}
10	1	Bad	h_4	x_9

4. Hypothesis Testing

Hypothesis Testing attempts to find a model to explain behavior by creating and then testing a hypothesis about the data.

The hypothesis usually is verified by examining a data sample. Given a population , the initial hypothesis to be tested, H_0 is called the null hypothesis.

Rejection of the null hypothesis causes hypothesis, H_1 called alternative hypothesis, to be made.

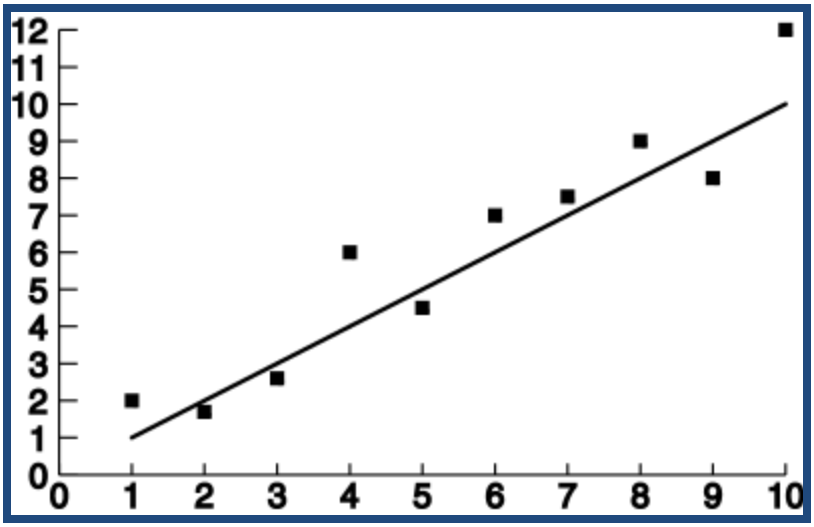
Chi Squared Statistic

There is set of procedures referred to as chi squared. Assuming that O represents the observed data and E is the expected values based on the hypothesis, the Chi-squared statistic,

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

5. Regression and Correlation:

It predicts future values based on past values. ***Linear Regression*** assumes linear relationship exists. $y = c_0 + c_1 x_1 + \ldots + c_n x_n$ Find values to best fit the data.



Examine the degree to which the values for two variables behave similarly.

Correlation coefficient r:

1 = perfect correlation

-1 = perfect but opposite correlation

0 = no correlation

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

Similarity measures:

The use of similarity measures is known to anyone who has performed Internet searches using a search engine.

The similarity between two Objects
.Similarity characteristics:

- $\forall t_i \in D, sim(t_i, t_i) = 1$
- $\forall t_i, t_j \in D, sim(t_i, t_j) = 0$ **if** t_i **and** t_j **are not alike at all.**
- $\forall t_i, t_j, t_k \in D, sim(t_i, t_j) < sim(t_i, t_k)$ **if** t_i **is more like** t_k **than it is like** t_j**.**

Alternatively, distance measure measure how unlike or dissimilar objects are.

$$\textbf{Dice:} \quad sim(t_i, t_j) = \frac{2\sum_{h=1}^{k} t_{ih} t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2}$$

$$\textbf{Jaccard:} \quad sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih} t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2 - \sum_{h=1}^{k} t_{ih} t_{jh}}$$

$$\textbf{Cosine:} \quad sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih} t_{jh}}{\sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2}}$$

$$\textbf{Overlap:} \quad sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih} t_{jh}}{min(\sum_{h=1}^{k} t_{ih}^2, \sum_{h=1}^{k} t_{jh}^2)}$$

In these formulas it is assumed that similarity is being evaluated between two vectors $t_i=\{t_{i1}…..t_{ik}\}$ and $t_j=\{t_{j1}…..t_{jk}\}$ and vector entries are assumed to be non negative numeric values.

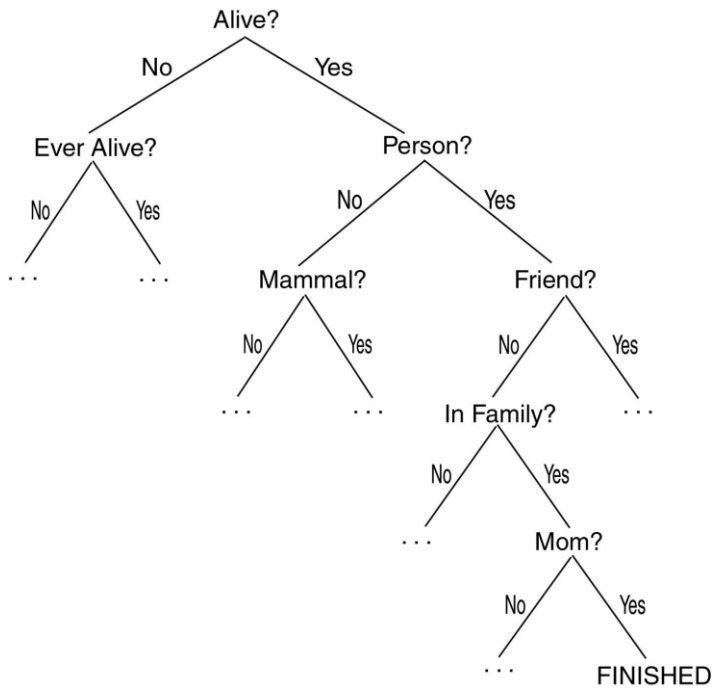
Be a count of the number of times an associated keyword appears in the document. If there is no overlap the resulting value is 0. If the two are identical, the resulting measure is 1.

Distance or dissimilarity measures are often used instead of similarity measures. Traditional distance measures may be used in a two-dimensional space. These include:

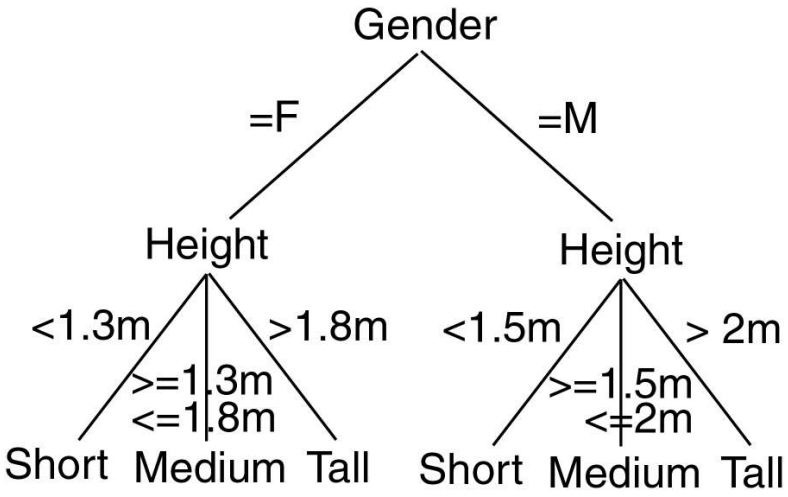
$$\textbf{Euclidean: } dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2}$$
$$\textbf{Manhattan: } dis(t_i, t_j) = \sum_{h=1}^{k} \mid (t_{ih} - t_{jh}) \mid$$

Decision Trees:

Decision tree is a predictive modeling technique used in classification, clustering and prediction tasks. Decision trees use a "divide and conquer" technique to split the problem search space into subsets.



Tree where the root and each internal node is labeled with a question. The arcs represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem. Popular technique for classification; Leaf node indicates class to which the corresponding tuple belongs.
Decision Tree Example:



A Decision Tree Model is a computational model consisting of three parts:

1. Decision Tree

2. Algorithm to create the tree

3. Algorithm that applies the tree to data

Creation of the tree is the most difficult part. Processing is basically a search similar to that in a

Binary search tree (although DT may not be binary).

Decision Tree Algorithm:

```
Input:
   T      //Decision Tree
   D      //Input Database
Output:
   M      //Model Prediction
DTProc Algorithm:
              //Illustrates Prediction Technique using DT
   for each t ∈ D do
      n = root node of T;
      while n not leaf node do
         Obtain answer to question on n applied t;
         Identify arc from t which contains correct answer;
         n = node at end of this arc;
      Make prediction for t based on labeling of n;
```

Advantages:

- Easy to understand.
- Easy to generate rules

Disadvantages:

- May suffer from over fitting.
- Classifies by rectangular partitioning.
- Does not easily handle nonnumeric data.
- Can be quite large – pruning is necessary.

Neural networks:

Based on observed functioning of human brain.(Artificial Neural Networks (ANN) Our view of neural networks is very simplistic. We view a neural network (NN) from a graphical viewpoint. Alternatively, a NN may be viewed from the perspective of matrices. Used in pattern recognition, speech recognition, computer vision, and classification.

Neural Network (NN) is a directed graph F=<V,A>

with vertices V={1,2,…,n} and arcs A={<i,j>|

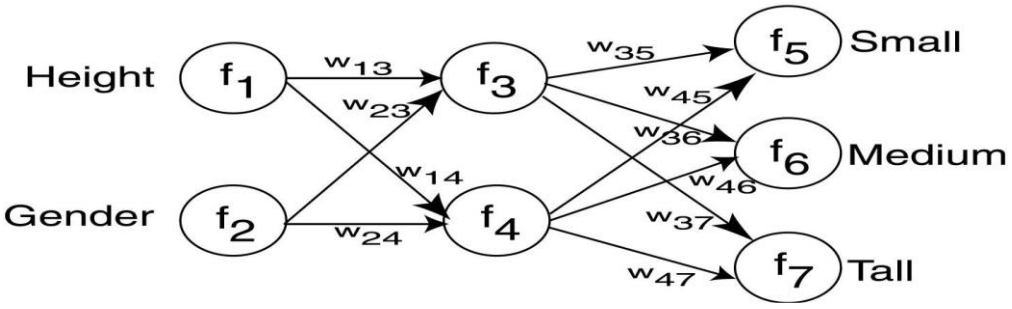
1<=i,j<=n}, with the following restrictions:

V is partitioned into a set of input nodes, VI, hidden nodes, VH, and output nodes, VO.

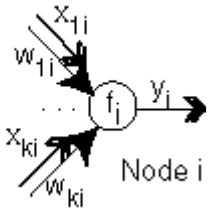
The vertices are also partitioned into layers Any arc <i,j> must have node i in layer h-1 and

Node j in layer h. Arc <i,j> is labeled with a numeric value wij. Node i am labeled with a function fi.

Neural Network Example:



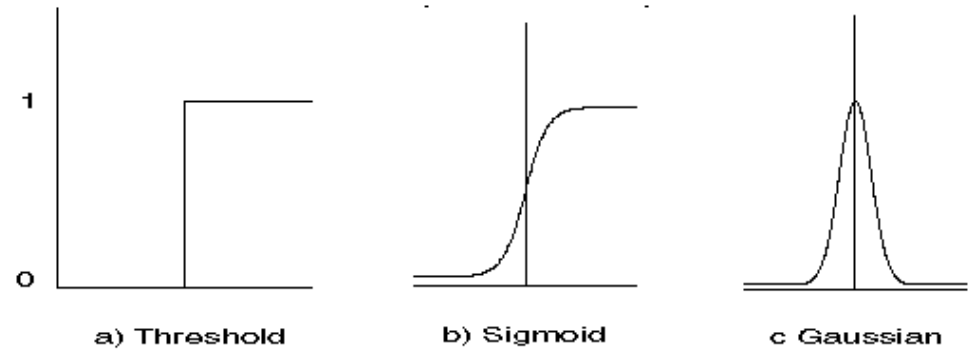
Neural Network node:



$$y_i = f_i(\sum_{j=1}^{k} w_{ji}\ x_{ji}) = f_i([w_{1i}...w_{ki}] \begin{bmatrix} x_{1i} \\ ... \\ x_{ki} \end{bmatrix})$$

NN Activation Functions

Functions associated with nodes in graph. Output may be in range [-1,1] or [0,1]



Neural Network Activation Functions:

Linear:

$$f_i(S) = c\,S$$

Threshold or Step:

$$f_i(S) = \left\{ \begin{array}{ll} 1 & if S > T \\ 0 & otherwise \end{array} \right\}$$

Ramp:

$$f_i(S) = \left\{ \begin{array}{ll} 1 & if S > T_2 \\ \frac{S-T_1}{T_2-T_1} & if T_1 \leq S \leq T_2 \\ 0 & if S < T_1 \end{array} \right\}$$

Sigmoid:

$$f_i(S) = \frac{1}{(1+e^{-c\,S})}$$

Hyperbolic Tangent:

$$f_i(S) = \frac{(1-e^{-S})}{(1+e^{-c\,S})}$$

Gaussian:

$$f_i(S) = e^{\frac{-S^2}{v}}$$

A Neural Network Model is a computational model consisting of three parts:

Neural Network graph learning algorithm that indicates how learning takes place. Recall techniques that determine how information is obtained from the network. We will look at propagation as the recall technique.

Advantages:

- Learning
- Can continue learning even after training set has been applied.
- Easy parallelization
- Solves many problems

Disadvantages:

- Difficult to understand
- May suffer from over fitting
- Structure of graph must be determined a priori.
- Input values must be numeric.
- Verification difficult.

Genetic Algorithms:

Optimization search type algorithms. Creates an initial feasible solution and iteratively

Creates new "better" solutions. Based on human evolution and survival of the fitness.

Must represent a solution as an individual. Individual: string I=I1,I2,…,In where Ij is in given

Alphabet A. Each character Ij is called a gene. Population: set of individuals.

A Genetic Algorithm (GA) is a computational model consisting of five parts:

A starting set of individuals, P.

Crossover: Technique to combine two parents to create offspring.

Mutation: Randomly change an individual.

Fitness: Determine the best individuals.

Algorithm which applies the crossover and mutation techniques to P iteratively using the fitness function to determine the best individuals in P to keep.

Genetic Algorithm:

Input:
P //Initial Population
Output:
P' //Improved Population
Genetic Algorithm:
//Illustrates Genetic Algorithm
repeat
$N = | P |;$
$P' = \emptyset;$
repeat
$i_1, i_2 = \textbf{select}(P);$
$o_1, o_2 = \textbf{cross}(i_1, i_2);$
$o_1 = \textbf{mutate}(o_1);$
$o_2 = \textbf{mutate}(o_2);$
$P' = P' \cup \{o_1, o_2\};$
until $| P' | = N;$
$P = P';$
until termination criteria satisfied;

Advantages

- Easily parallelized

Disadvantages

- Difficult to understand and explain to end users.
- Abstraction of the problem and method to represent individuals is quite difficult.
- Determining fitness function is difficult.
- Determining how to perform crossover and mutation is difficult.

POSSIBLE QUESTIONS

PART - B

8 MARKS

1. Describe about Basic Data mining tasks.
2. Write about Data mining Techniques.
    A. Point Estimation
    B. Models based on Summarization
    C. Bayes Theorem
3. Write a note on Data mining and Traditional Database
4. Explain decision trees with algorithm.
5. Describe Data mining Versus Knowledge Database Discovery
6. Explain similarity Measures in detail
7. Explain the issues in data mining
8. Write a note on Clustering with Genetic Algorithm
9. Describe the concept of Neural Networks with example
10. Write a note on Clustering with Neural Networks

UNIT II

SYLLABUS

Classifications: Bayesian Classification – Distance Based Algorithms – K-Nearest Neighbor. Clustering: K-Means Clustering –Clustering with Genetic Algorithms – Clustering with Neural Networks. Association Rules – Basic Algorithms – Parallel and Distributed Algorithms – Comparing Approaches – Generalized and Multilevel Association Rules. Web Mining: Web Content Mining: Personalization.

Classification:

Classification is perhaps the most popular data mining technique. Estimation and prediction may be viewed as types of classification.

Given a database $D=\{t_1,t_2,\ldots,t_n\}$ and a set of classes $C=\{C_1,\ldots,C_m\}$, the ***Classification Problem*** is to define a mapping f:DgC where each t_i is assigned to one class.

Actually divides D into ***equivalence*** *classes.*

Prediction is similar, but may be viewed as having infinite number of classes.

Examples:

Teachers classify students' grades as A, B, C, D, or F.

Identify mushrooms as poisonous or edible.

Predict when a river will flood.

Identify individuals with credit risks.

Speech recognition

Pattern recognition

Example: Grading

If x >= 90 then grade =A.

If 80<=x<90 then grade =B.

If 70<=x<80 then grade =C.

If 60<=x<70 then grade =D.

If x<50 then grade =F

Classification techniques:

Approach:

1. Create specific model by evaluating training data (or using domain experts' knowledge).

2Apply model developed to new data.

There are three basic methods used to solve the classification problem:

Specifying boundaries:

Here classification is performed by dividing the input space of potential database tuples into regions where each region is associated with one class.

Using probability distributions:

For any given class C_j , $P(t_i|C_j)$ is the PDF for the class evaluated at one point, t_i. If probability of occurrence for each class $P(C_j)$ is known, then $P(C_j)\ P(t_i|C_j)$ is used to estimate the probability that t_i is in class C_j .

Using posterior probabilities:

Given a data value t_i we would like to determine the probability that t_i is in a class C_j . This is denoted by $P(t_i|C_j)$ and is called posterior probabilities.

Classes must be predefined.

Most common techniques use DTs, NNs, or are based on distances or statistical methods.

Issues in Classification:

- Missing Data
  - Ignore
  - Replace with assumed value
- Measuring Performance

- Classification accuracy on test data
- Confusion matrix
- OC Curve

Height Example data

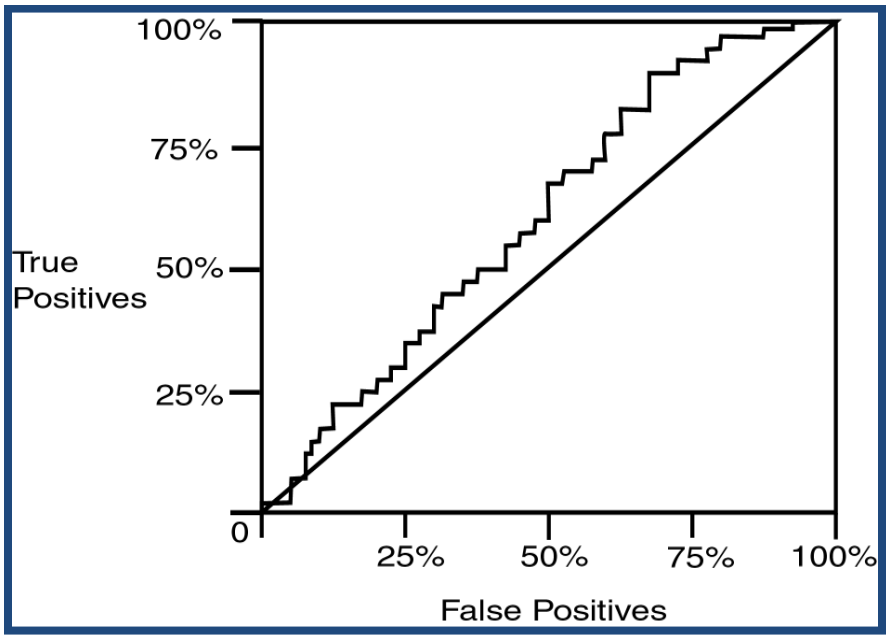
Name	Gender	Height	Output1	Output2
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

Confusion Matrix Example

Using height data example with Output1 correct and Output2 actual assignment

Actual Membership	Assignment		
	Short	Medium	Tall
Short	0	4	0
Medium	0	5	3
Tall	0	1	2

Operating Characteristic Curve



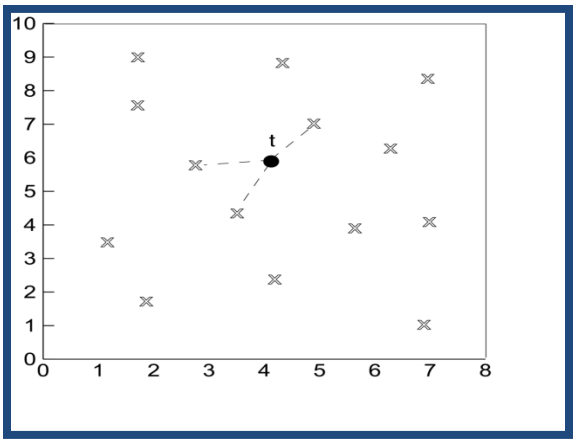
A confusion Matrix illustrates the accuracy of the solution to be classification problem. An Operating Characteristic curve or ROC curve shows the relationship between false positive and true positives. In the OC curve the horizontal axis has the percentage of false positives and the vertical axis has the percentage of true positives for a database sample. When evaluating the results for a specific sample, the curve looks like a jagged stair-step., as each new tuple is either a false positive or true positive.

Distance Based Algorithm:

K Nearest Neighbor:

One common classification scheme based on the use of distance measures is that of the K-Nearest neighbor (KNN), The KNN technique assumes that the entire training set includes not

only the set but also the desired classification of each item. New item placed in class that contains the most items from this set of K closest items.



Classification using KNN

Here the points in the training set are shown and K=3. The three closest items in the training set are shown; t will be placed in the class to which most of these are members.

```
Input:
    D       //Training data
    K       //Number of neighbors
    t       //Input tuple to classify
Output:
    c       //Class to which t is assigned
KNN Algorithm:
            //Algorithm to classify tuple using KNN
    N = ∅;
                //Find set of neighbors, N, for t
    foreach d ∈ D do
        if | N |≤ K then
            N = N ∪ d;
        else
            if ∃ u ∈ N such that sim(t, u) ≥ sim(t, d) then
                begin
                    N = N − u;
                    N = N ∪ d;
                end
                //Find class for classification
    c = class to which the most u ∈ N are classified;
```

T to represent the training data. Since each tuple to be classified must be compared to each element In the training data, if there are q elements in the training set, this is O(q). Given n elements to be classified, this becomes an O(nq) problem. Training data are of a constant size, this can then be viewed as an O(n) problem.

Clustering:

Clustering is similar to classification in that data are grouped. Unlike classification the groups are not predefined. Finding similarities between data according to characteristics found in actual data. The groups are called clusters.

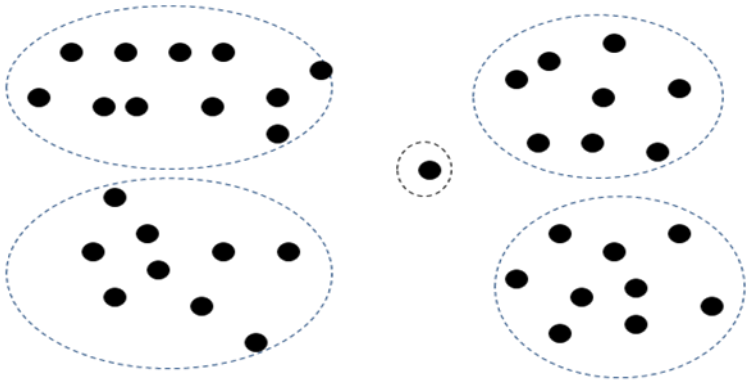
Many definitions for clusters:

Set of like elements. Elements from different clusters are not alike.

The distance between points in a cluster is less than the distance between a point in the cluster and any point outside it.

A term similar to clustering is database segmentation, where like tuples in a database are grouped together. Group houses in a town into neighborhoods based on similar features.

Income	Age	Children	Marital Status	Education
$25,000	35	3	Single	High School
$15,000	25	1	Married	High School
$20,000	40	0	Single	High School
$30,000	20	0	Divorced	High School
$20,000	25	3	Divorced	College
$70,000	60	0	Married	College
$90,000	30	0	Married	Graduate School
$200,000	45	5	Married	Graduate School
$100,000	50	2	Divorced	College



Geographic Distance Based

Clustering Houses

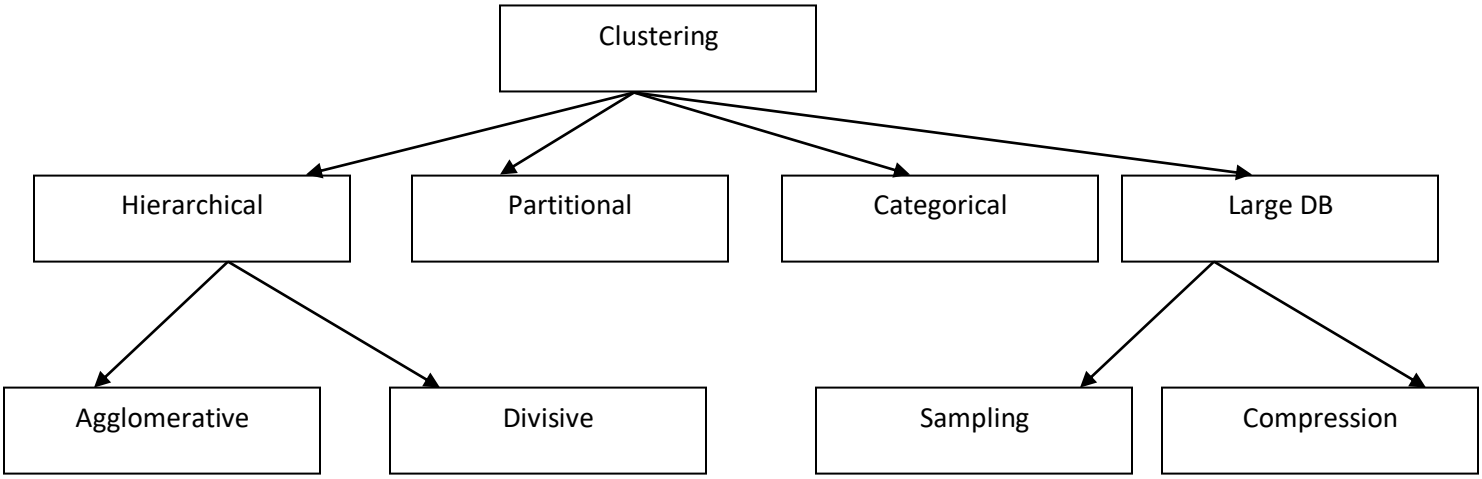
When Clustering is applied to a real world database, many interesting problems may occur

- Outlier handling- It is very difficult. Here the elements do not actually fall into any cluster. They can be viewed as solitary clusters. This process may result in the creation of poor clusters by combining two clusters and leaving the outlier in its own cluster.
- Dynamic data- Data in the database implies that cluster membership may change over time.
- Interpreting results: Interpreting the semantic meaning of each cluster may be difficult. Here is where a domain expert is needed to assign a label or interpretation for each cluster.
- Evaluating results: There is no one correct answer to a clustering problem. The exact number of clusters required is not easy to determine. Again a domain expert may be required.
- Number of clusters- Without any prior knowledge of plant classification, if we attempt to divide this set of data into similar groupings, it would not be clear how many groups should be created.
- Data to be used- What data should be used for clustering

Clustering Algorithm:

- Given a database $D=\{t_1,t_2,\ldots,t_n\}$ of tuples and an integer value k, the ***Clustering Problem*** is to define a mapping f:Dg{1,..,k} where each t_i is assigned to one cluster K_j, $1<=j<=k$.
- A ***Cluster***, K_j, contains precisely those tuples mapped to it.
- Unlike classification problem, clusters are not known a priori.

Classification of Clustering:



- ***Hierarchical*** –In Hierarchical Nested set of clusters created.
- ***Partitional*** – One set of clusters created.
- ***Incremental*** – Each element handled one at a time.
- ***Simultaneous*** – All elements handled together.
- ***Overlapping/Non-overlapping***

K-Means clustering:

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. High degree of similarity among elements in a cluster is obtained.

Given a cluster $K_i=\{t_{i1},t_{i2},\ldots,t_{im}\}$, the ***cluster mean*** is $m_i = (1/m)(t_{i1} + \ldots + t_{im})$

Algorithm:

Input:

$D = \{t_1, t_2, ..., t_n\}$ // Set of elements

A // Adjacency matrix showing distance between elements.

k // Number of desired clusters.

Output:

K // Set of clusters.

K-Means Algorithm:

assign initial values for means $m_1, m_2, ..., m_k$;

repeat

assign each item t_i to the cluster which has the closest mean ;

calculate new mean for each cluster;

until convergence criteria is met;

Example:

- Given: {2,4,10,12,3,20,30,11,25}, k=2
- Randomly assign means: m_1=3,m_2=4
- K_1={2,3}, K_2={4,10,12,20,30,11,25}, m_1=2.5,m_2=16
- K_1={2,3,4},K_2={10,12,20,30,11,25}, m_1=3,m_2=18
- K_1={2,3,4,10},K_2={12,20,30,11,25}, m_1=4.75,m_2=19.6
- K_1={2,3,4,10,11,12},K_2={20,30,25}, m_1=7,m_2=25
- Stop as the clusters with these means are the same.

The time complexity of K-means is O(tkn) where t is the number of iterations. K-means finds a local optimum and may actually miss the global optimum. K-means does not work on categorical data because the mean must be defined on the attribute type. One variation of K-means, k-modes does handle categorical data. Instead of using means, it uses modes.

Although the K-means algorithm often produces good results, it is not time efficient and does not scale well.

Clustering with Genetic Algorithm

There has been clustering technique based on the use of genetic algorithms. One simple approach would be to use a bit-map representation for each possible cluster.

So given a set of four items,{A,B,C,D}, we would represent one solution to creating two clusters as 1001 and 0110. This represents two clusters {A,D} and {B,C}.

Example:

- A database contain eight items {A,B,C,D,E,F,G,H}
- Randomly choose initial solution. Three clusters are:

  {A,C,E} {B,F} {D,G,H} which represented by

  10101000, 01000100, 00010011
- Suppose crossover at point four and choose 1st and 3rd individuals:

10100011, 01000100, 00011000

Algorithm:

```
Input:
   D = {t1, t2, ..., tn}    // Set of elements
   k      // Number of desired clusters.
Output:
   K      // Set of clusters.
GA Clustering Algorithm:
   randomly create an initial solution;
   repeat
      use crossover to create a new solution;
   until termination criteria is met;
```

The above algorithm shows one possible iterative refinement technique for clustering thar uses a genetic algorithm. A new solution is generated from the previous solution using cross over and mutation operation. Our algorithm shows only crossover.

Clustering with Neural network:

Clustering is a fundamental data analysis method. It is widely used for pattern recognition, feature extraction, vector quantization (VQ), image segmentation, function approximation, and data mining. As an unsupervised classification technique, clustering identifies some inherent structures present in a set of objects based on a similarity measure. Clustering methods can be based on statistical model identification or competitive learning. Importance is attached to a number of competitive learning based clustering neural networks such as the self-organizing map (SOM), the learning vector quantization (LVQ), the neural gas, and the ART model, and clustering algorithms such as the C-means, mountain/subtractive clustering, and fuzzy C-means (FCM) algorithms. Associated topics such as the under-utilization problem, fuzzy clustering, robust clustering, clustering based on non-Euclidean distance measures, supervised clustering, hierarchical clustering as well as cluster validity are also described. Two examples are given to demonstrate the use of the clustering methods.

Association Rule:

- ***Set of items:*** $I=\{I_1,I_2,\ldots,I_m\}$
- ***Transactions:*** $D=\{t_1,t_2,\ldots,t_n\}$, $t_j \subseteq I$
- ***Itemset:*** $\{I_{i1},I_{i2},\ldots,I_{ik}\} \subseteq I$
- ***Support of an itemset:*** Percentage of transactions which contain that itemset.
- ***Large (Frequent) itemset:*** Itemset whose number of occurrences is above a threshold.

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

I = { Beer, Bread, Jelly, Milk, PeanutButter}

Support of {Bread,PeanutButter} is 60%

- ***Association Rule (AR):*** implication $X \Rightarrow Y$ where $X,Y \subseteq I$ and $X \cap Y = ;$

- ***Support of AR (s) X ⇒ Y***: Percentage of transactions that contain X ∪Y
- ***Confidence of AR (a) X ⇒ Y:*** Ratio of number of transactions that contain X ∪ Y to the number that contain X

$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

- Given a set of items $I=\{I_1,I_2,\ldots,I_m\}$ and a database of transactions $D=\{t_1,t_2, \ldots, t_n\}$ where $t_i=\{I_{i1},I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, the ***Association Rule Problem*** is to identify all association rules X ⇒ Y with a minimum support and confidence.
- Link Analysis
- ***NOTE:*** Support of X ⇒ Y is same as support of X ∪ Y.
- Find Large Itemsets.

Generate rules from frequent itemsets

Input:
D //Database of transactions
I //Items
L //Large itemsets
s //Support
α //Confidence
Output:
R //Association Rules satisfying s and α
ARGen Algorithm:
$R = \emptyset$;
for each $l \in L$ **do**
for each $x \subset l$ **such that** $x \neq \emptyset$ **and** $x \neq l$ **do**
if $\frac{support(l)}{support(x)} \geq \alpha$ **then**
$R = R \cup \{x \Rightarrow (l - x)\}$;

Apriori

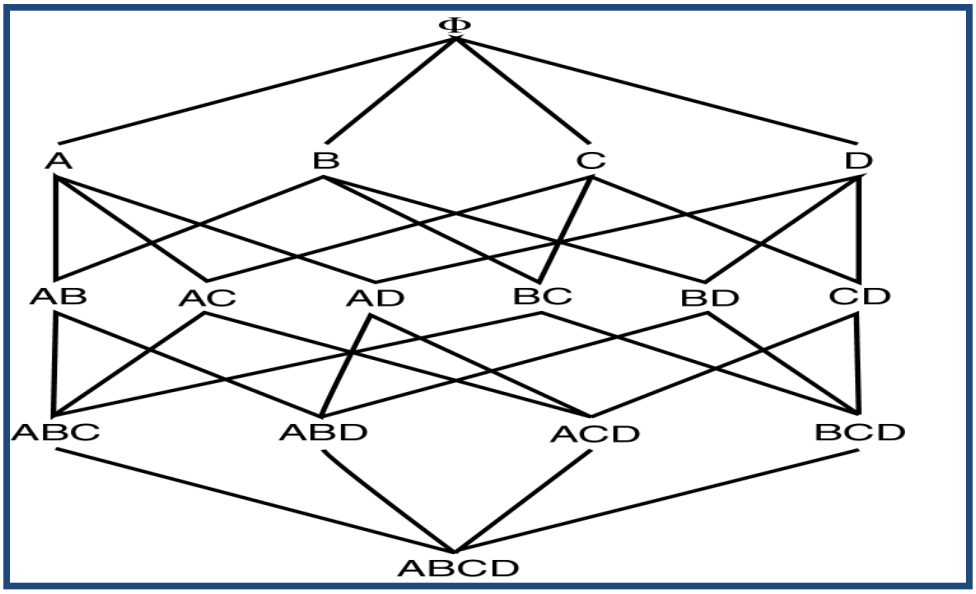
- ***Large Itemset Property:***

Any subset of a large itemset is large.

- Contrapositive:

If an itemset is not large,

none of its supersets are large



Pass	Candidates	Large Itemsets
1	{Beer},{Bread},{Jelly}, {Milk},{PeanutButter}	{Beer},{Bread}, {Milk},{PeanutButter}
2	{Beer,Bread},{Beer,Milk}, {Beer,PeanutButter},{Bread,Milk}, {Bread,PeanutButter},{Milk,PeanutButter}	{Bread,PeanutButter}

Apriori Algorithm

1. C_1 = Itemsets of size one in I;
2. Determine all large itemsets of size 1, $L_{1;}$
3. i = 1;
4. Repeat
5. i = i + 1;
6. C_i = Apriori-Gen(L_{i-1});
7. Count C_i to determine $L_{i;}$
8. until no more large itemsets found;
9. Generate candidates of size i+1 from large itemsets of size i.
10. Approach used: join large itemsets of size i if they agree on i-1
11. May also prune candidates who have subsets that are not large.

Transaction	Items
t_1	Blouse
t_2	Shoes,Skirt,TShirt
t_3	Jeans,TShirt
t_4	Jeans,Shoes,TShirt
t_5	Jeans,Shorts
t_6	Shoes,TShirt
t_7	Jeans,Skirt
t_8	Jeans,Shoes,Shorts,TShirt
t_9	Jeans
t_{10}	Jeans,Shoes,TShirt
t_{11}	TShirt
t_{12}	Blouse,Jeans,Shoes,Skirt,TShirt
t_{13}	Jeans,Shoes,Shorts,TShirt
t_{14}	Shoes,Skirt,TShirt
t_{15}	Jeans,TShirt
t_{16}	Skirt,TShirt
t_{17}	Blouse,Jeans,Skirt
t_{18}	Jeans,Shoes,Shorts,TShirt
t_{19}	Jeans
t_{20}	Jeans,Shoes,Shorts,TShirt

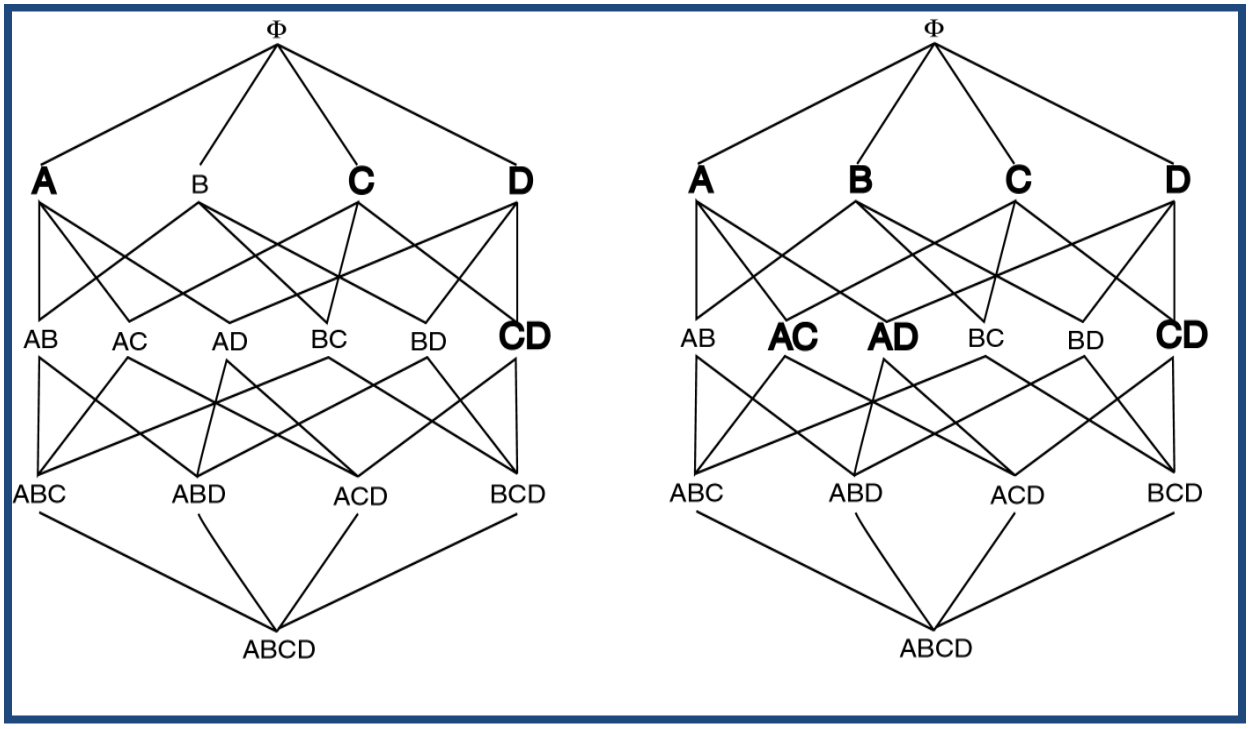
Scan	Candidates	Large Itemsets
1	{Blouse},{Jeans},{Shoes}, {Shorts},{Skirt},{TShirt}	{Jeans},{Shoes},{Shorts} {Skirt},{Tshirt}
2	{Jeans,Shoes},{Jeans,Shorts},{Jeans,Skirt}, {Jeans,TShirt},{Shoes,Shorts},{Shoes,Skirt}, {Shoes,TShirt},{Shorts,Skirt},{Shorts,TShirt}, {Skirt,TShirt}	{Jeans,Shoes},{Jeans,Shorts}, {Jeans,TShirt},{Shoes,Shorts}, {Shoes,TShirt},{Shorts,TShirt}, {Skirt,TShirt}
3	{Jeans,Shoes,Shorts},{Jeans,Shoes,TShirt}, {Jeans,Shorts,TShirt},{Jeans,Skirt,TShirt}, {Shoes,Shorts,TShirt},{Shoes,Skirt,TShirt}, {Shorts,Skirt,TShirt}	{Jeans,Shoes,Shorts}, {Jeans,Shoes,TShirt}, {Jeans,Shorts,TShirt}, {Shoes,Shorts,TShirt}
4	{Jeans,Shoes,Shorts,TShirt}	{Jeans,Shoes,Shorts,TShirt}
5	$\emptyset$	$\emptyset$

- ***Advantages:***
  - Uses large itemset property.
  - Easily parallelized
  - Easy to implement.
- ***Disadvantages:***
  - Assumes transaction database is memory resident.

– Requires up to m database scans.

Sampling

- Large databases
- Sample the database and apply Apriori to the sample.
- ***Potentially Large Itemsets (PL):*** Large itemsets from sample
- ***Negative Border (BD^-):***
  - Generalization of Apriori-Gen applied to itemsets of varying sizes.
  - Minimal set of itemsets which are not in PL, but whose subsets are all in PL.



1. D_s = sample of Database D;
2. PL = Large itemsets in D_s using smalls;
3. C = PL ∪ BD^-(PL);
4. Count C in Database using s;
5. ML = large itemsets in BD^-(PL);

6. If ML = ∅ then done

7. else C = repeated application of BD^{-};

8. Count C in Database;

Find AR assuming s = 20%

D_s = { t_1,t_2}

Smalls = 10%

PL = {{Bread}, {Jelly}, {PeanutButter}, {Bread,Jelly}, {Bread,PeanutButter}, {Jelly, PeanutButter}, {Bread,Jelly,PeanutButter}}

BD^{-}(PL)={{Beer},{Milk}}

ML = {{Beer}, {Milk}}

Repeated application of BD^{-} generates all remaining itemsets

Advantages:

Reduces number of database scans to one in the best case and two in worst.

Scales better.

Disadvantages:

Potentially large number of candidates in second passes

Partitioning

- Divide database into partitions D^1,D^2,…,D^p
- Apply Apriori to each partition
- Any large itemset must be large in at least one partition.

Algorithm

- Divide D into partitions D^1,D^2,…,D^p;
- For I = 1 to p do
- L^i = Apriori(D^i);

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- $C = L^1 \cup \ldots \cup L^p$;
- Count C on D to generate L;

L^1 ={{Bread}, {Jelly},
{PeanutButter},
{Bread,Jelly},
{Bread,PeanutButter},
{Jelly, PeanutButter},
{Bread,Jelly,PeanutButter}}

L^1 ={{Bread}, {Jelly},
{PeanutButter},
{Bread,Jelly},
{Bread,PeanutButter},
{Jelly, PeanutButter},
{Bread,Jelly,PeanutButter}}

- ***Advantages:***
  - Adapts to available main memory
  - Easily parallelized
  - Maximum number of database scans is two.
- ***Disadvantages:***
  - May have many candidates during second scan.

What is Web Mining?

Web mining - is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by mining the web. It makes utilization of automated apparatuses to reveal and extricate data from servers and web reports, and it permits organizations to get to both organized and unstructured information from browser activities, server logs, website and link structure, page content and different sources. Web mining can be divided into three different types – **Web usage mining**, **Web content mining** and **Web structure mining**.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

POSSIBLE QUESTIONS

PART - B

8 MARKS

1. Write a note on Bayesian Classification and K-Nearest Neighbor

2. Explain in detail about K-means Clustering.

3. Write about Classification and Issues in Classification

4. Explain Web mining in detail

5. Explain Distance based algorithm and K-Nearest neighbors in detail

6. Describe Basic Algorithms with algorithm.

7. Explain K-means Clustering and Clustering with Neural Networks.

8. Explain Parallel and Distributed Algorithm and Comparing Approaches.

UNIT – III

SYALLBUS

Data Warehousing: Introduction – Architecture – System Process-Process Architecture. Design: Database Schema – Partitioning Strategy – Aggregations – Data Marting – Meta Data.

Data warehouse:

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. This data helps analysts to take informed decisions in an organization.

An operational database undergoes frequent changes on a daily basis on account of the transactions that take place. Suppose a business executive wants to analyze previous feedback on any data such as a product, a supplier, or any consumer data, then the executive will have no data available to analyze because the previous data has been updated due to transactions.

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.

Data mining functions such as association, clustering, classification, prediction can be integrated with OLAP operations to enhance the interactive mining of knowledge at multiple level of abstraction. That's why data warehouse has now become an important platform for data analysis and online analytical processing.

Understanding a Data Warehouse

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.

- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

Why a Data Warehouse is separated from Operational Databases

A data warehouses is kept separate from operational databases due to the following reasons:

- An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data.
- Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database.
- An operational database query allows reading and modifying operations, while an OLAP query needs only **read only** access of stored data.
- An operational database maintains current data. On the other hand, a data warehouse maintains historical data.

Data Warehouse Features

The key features of a data warehouse are discussed below:

- **Subject Oriented** - A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations; rather it focuses on modeling and analysis of data for decision making.
- **Integrated** - A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.

- **Time Variant** - The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.

- **Non-volatile** - Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database are not reflected in the data warehouse.

Note: A data warehouse does not require transaction processing, recovery, and concurrency controls, because it is physically stored and separate from the operational database.

Data Warehouse Applications

As discussed before, a data warehouse helps business executives to organize, analyze, and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess "closed-loop" feedback system for the enterprise management. Data warehouses are widely used in the following fields:

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

Types of Data Warehouse

Information processing, analytical processing and data mining are the three types of data warehouse applications that are discussed below:

- **Information Processing** - A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.

- **Analytical Processing** - A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic OLAP operations, including slice-and-dice, drill down, drill up, and pivoting.

- **Data Mining** - Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using the visualization tools.

Data warehouse Architecture:

Business analysis framework for the data warehouse design and architecture of a data warehouse.

Business Analysis Framework

The business analyst get the information from the data warehouses to measure the performance and make critical adjustments in order to win over other business holders in the market. Having a data warehouse offers the following advantages:

- Since a data warehouse can gather information quickly and efficiently, it can enhance business productivity.
- A data warehouse provides us a consistent view of customers and items; hence, it helps us manage customer relationship.
- A data warehouse also helps in bringing down the costs by tracking trends, patterns over a long period in a consistent and reliable manner.

To design an effective and efficient data warehouse, we need to understand and analyze the business needs and construct a **business analysis framework**. Each person has different views regarding the design of a data warehouse. These views are as follows:

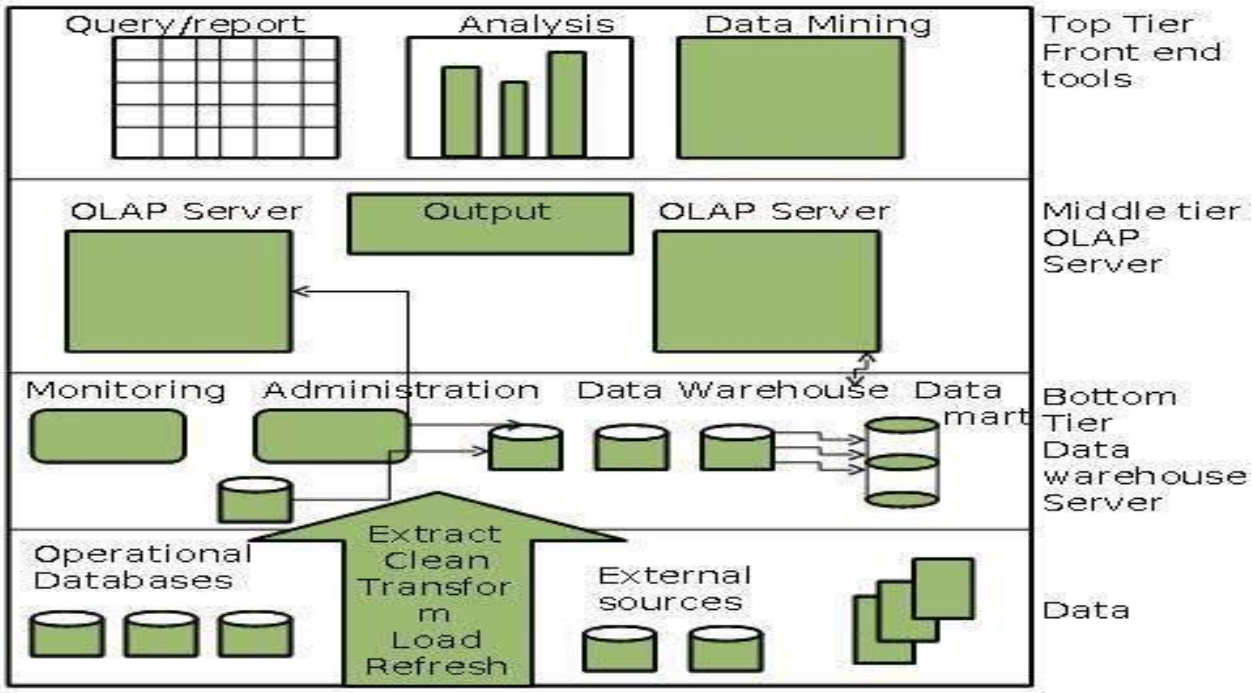
- **The top-down view** - This view allows the selection of relevant information needed for a data warehouse.
- **The data source view** - This view presents the information being captured, stored, and managed by the operational system.
- **The data warehouse view** - This view includes the fact tables and dimension tables. It represents the information stored inside the data warehouse.
- **The business query view** - It is the view of the data from the viewpoint of the end-user.

Three-Tier Data Warehouse Architecture

Generally a data warehouses adopts three-tier architecture. Following are the three tiers of the data warehouse architecture.

- **Bottom Tier** - The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.
- **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), this is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.
    - By Multidimensional OLAP (MOLAP) model, this directly implements the multidimensional data and operations.
- **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.

The following diagram depicts the three-tier architecture of data warehouse:



Data Warehouse Models

From the perspective of data warehouse architecture, we have the following data warehouse models:

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

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.

Data Mart

Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

In other words, we can claim that data marts contain data specific to a particular group. For example, the marketing data mart may contain data related to items, customers, and sales. Data marts are confined to subjects.

Points to remember about data marts:

- Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.
- 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 is flexible.

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.

Load Manager

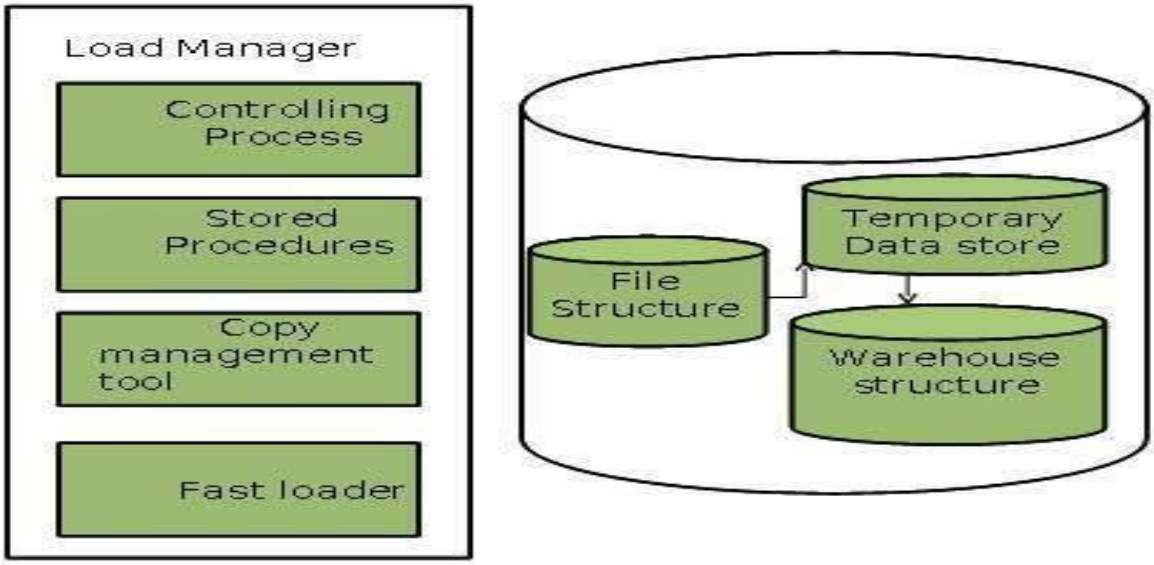
This component performs the operations required to extract and load process.

The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

Load Manager Architecture

The load manager performs the following functions:

- Extract the data from source system.
- Fast Load the extracted data into temporary data store.
- Perform simple transformations into structure similar to the one in the data warehouse.



Extract Data from Source

The data is extracted from the operational databases or the external information providers. Gateways are the application programs that are used to extract data. It is supported by underlying DBMS and allows client program to generate SQL to be executed at a server. Open Database Connection (ODBC), Java Database Connection (JDBC), is examples of gateway.

Fast Load

- In order to minimize the total load window the data need to be loaded into the warehouse in the fastest possible time.
- The transformations affect the speed of data processing.

- It is more effective to load the data into relational database prior to applying transformations and checks.
- Gateway technology proves to be not suitable; since they tend not be performing when large data volumes are involved.

Simple Transformations

While loading it may be required to perform simple transformations. After this has been completed we are in position to do the complex checks. Suppose we are loading the EPOS sales transaction we need to perform the following checks:

- Strip out all the columns that are not required within the warehouse.
- Convert all the values to required data types.

Warehouse Manager

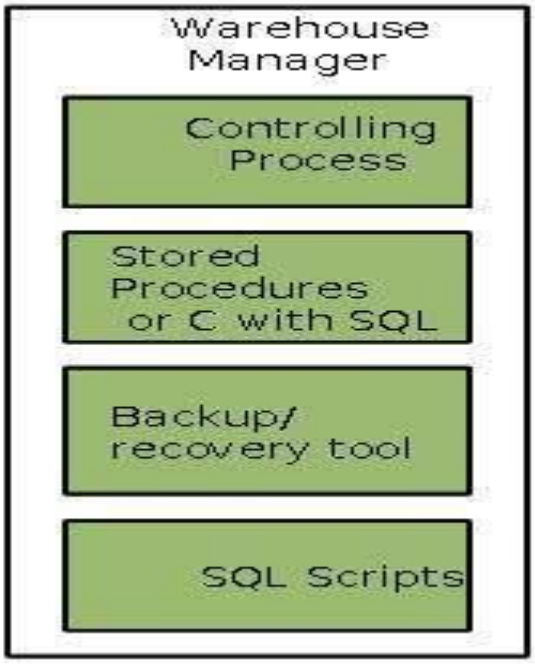
A warehouse manager is responsible for the warehouse management process. It consists of third-party system software, C programs, and shell scripts.

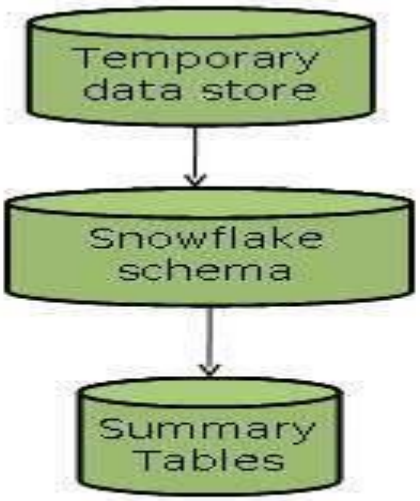
The size and complexity of warehouse managers varies between specific solutions.

Warehouse Manager Architecture

A warehouse manager includes the following:

- The controlling process
- Stored procedures or C with SQL
- Backup/Recovery tool
- SQL Scripts





Operations Performed by Warehouse Manager

- A warehouse manager analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.
- Generates new aggregations and updates existing aggregations. Generates normalizations.
- Transforms and merges the source data into the published data warehouse.
- Backup the data in the data warehouse.
- Archives the data that has reached the end of its captured life.

Note: A warehouse Manager also analyzes query profiles to determine index and aggregations are appropriate.

Query Manager

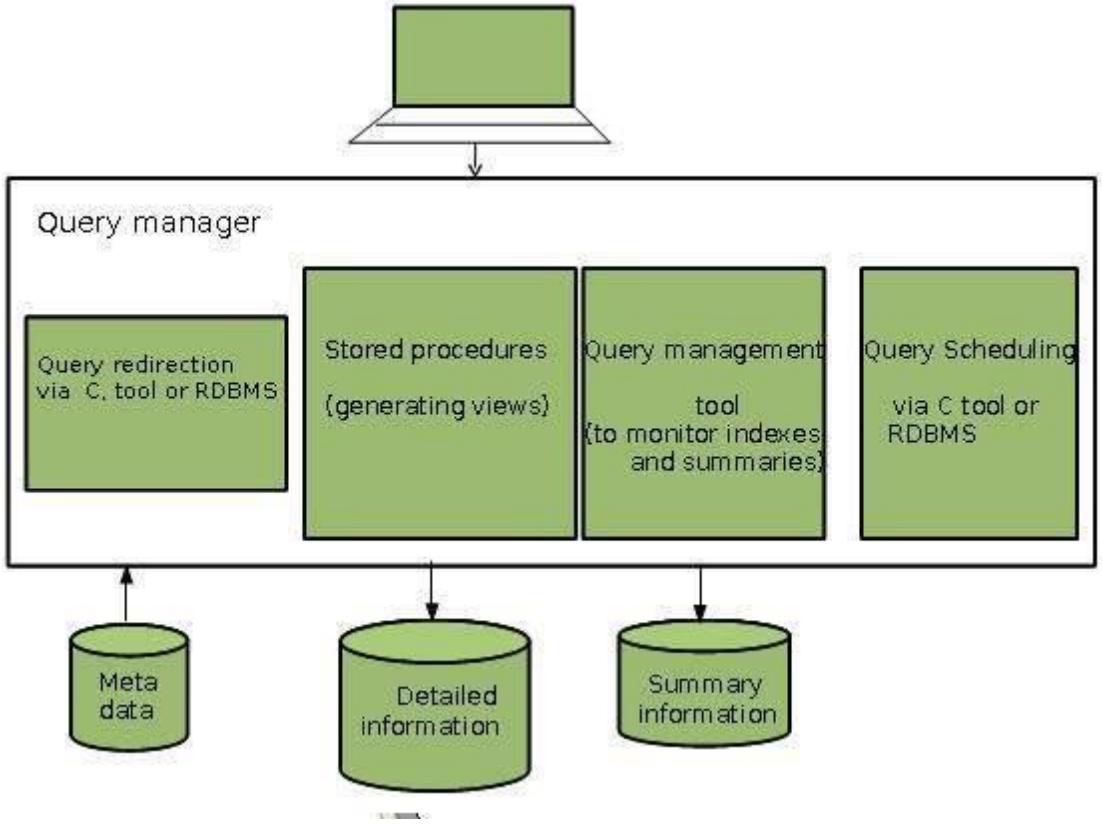
- Query manager is responsible for directing the queries to the suitable tables.

- By directing the queries to appropriate tables, the speed of querying and response generation can be increased.
- Query manager is responsible for scheduling the execution of the queries posed by the user.

Query Manager Architecture

The following screenshot shows the architecture of a query manager. It includes the following:

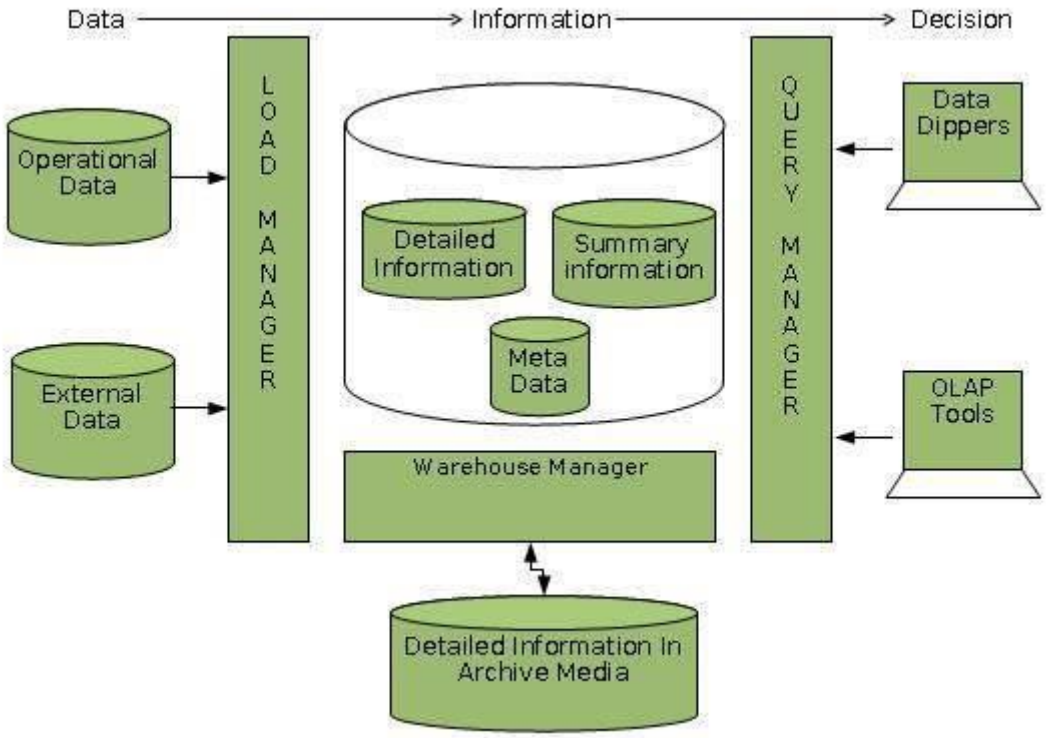
- Query redirection via C tool or RDBMS
- Stored procedures
- Query management tool
- Query scheduling via C tool or RDBMS
- Query scheduling via third-party software



Detailed Information

Detailed information is not kept online; rather it is aggregated to the next level of detail and then archived to tape. The detailed information part of data warehouse keeps the detailed information in the star flake schema. Detailed information is loaded into the data warehouse to supplement the aggregated data.

The following diagram shows a pictorial impression of where detailed information is stored and how it is used.



Note: If detailed information is held offline to minimize disk storage, we should make sure that the data has been extracted, cleaned up, and transformed into star flake schema before it is archived.

Summary Information

Summary Information is a part of data warehouse that stores predefined aggregations. These aggregations are generated by the warehouse manager. Summary Information must be treated as transient. It changes on-the-go in order to respond to the changing query profiles.

Points to remember about summary information.

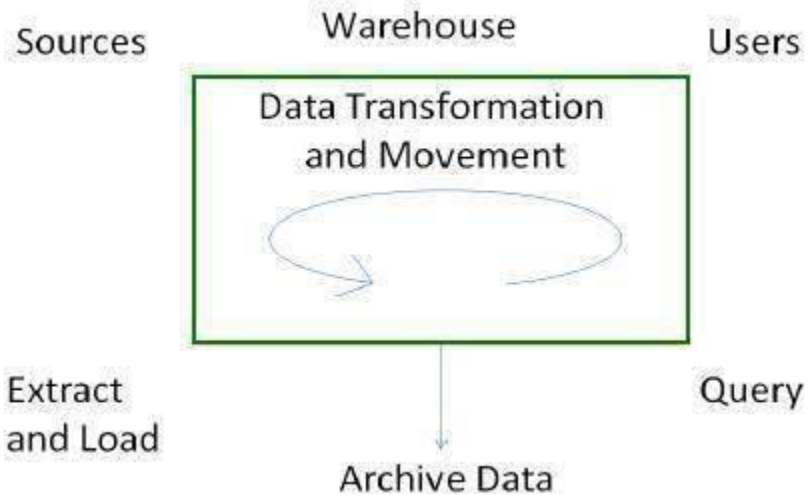
- Summary information speeds up the performance of common queries.
- It increases the operational cost.
- It needs to be updated whenever new data is loaded into the data warehouse.
- It may not have been backed up, since it can be generated fresh from the detailed information.

System Process:

Process Flow in Data Warehouse

There are four major processes that contribute to a data warehouse:

- Extract and load the data.
- Cleaning and transforming the data.
- Backup and archive the data.
- Managing queries and directing them to the appropriate data sources.



Extract and Load Process

Data extraction takes data from the source systems. Data load takes the extracted data and loads it into the data warehouse.

Note: Before loading the data into the data warehouse, the information extracted from the external sources must be reconstructed.

Controlling the Process

Controlling the process involves determining when to start data extraction and the consistency check on data. Controlling process ensures that the tools, the logic modules, and the programs are executed in correct sequence and at correct time.

When to Initiate Extract

Data needs to be in a consistent state when it is extracted, i.e., the data warehouse should represent a single, consistent version of the information to the user.

For example, in a customer profiling data warehouse in telecommunication sector, it is illogical to merge the list of customers at 8 pm on Wednesday from a customer database with the customer subscription events up to 8 pm on Tuesday. This would mean that we are finding the customers for whom there are no associated subscriptions.

Loading the Data

After extracting the data, it is loaded into a temporary data store where it is cleaned up and made consistent.

Note: Consistency checks are executed only when all the data sources have been loaded into the temporary data store.

Clean and Transform Process

Once the data is extracted and loaded into the temporary data store, it is time to perform Cleaning and Transforming. Here is the list of steps involved in Cleaning and Transforming:

- Clean and transform the loaded data into a structure
- Partition the data
- Aggregation

Clean and Transform the Loaded Data into a Structure

Cleaning and transforming the loaded data helps speed up the queries. It can be done by making the data consistent:

- Within itself.
- With other data within the same data source.

- With the data in other source systems.
- With the existing data present in the warehouse.

Transforming involves converting the source data into a structure. Structuring the data increases the query performance and decreases the operational cost. The data contained in a data warehouse must be transformed to support performance requirements and control the ongoing operational costs.

Partition the Data

It will optimize the hardware performance and simplify the management of data warehouse. Here we partition each fact table into multiple separate partitions.

Aggregation

Aggregation is required to speed up common queries. Aggregation relies on the fact that most common queries will analyze a subset or an aggregation of the detailed data.

Backup and Archive the Data

In order to recover the data in the event of data loss, software failure, or hardware failure, it is necessary to keep regular backups. Archiving involves removing the old data from the system in a format that allow it to be quickly restored whenever required.

For example, in a retail sales analysis data warehouse, it may be required to keep data for 3 years with the latest 6 months data being kept online. In such as scenario, there is often a requirement to be able to do month-on-month comparisons for this year and last year. In this case, we require some data to be restored from the archive.

Query Management Process

This process performs the following functions:

- Manages the queries.
- Helps speed up the execution time of queris.
- Directs the queries to their most effective data sources.
- Ensures that all the system sources are used in the most effective way.

- Monitors actual query profiles.

The information generated in this process is used by the warehouse management process to determine which aggregations to generate. This process does not generally operate during the regular load of information into data warehouse.

Process Architecture:

Process managers are responsible for maintaining the flow of data both into and out of the data warehouse. There are three different types of process managers:

- Load manager
- Warehouse manager
- Query manager

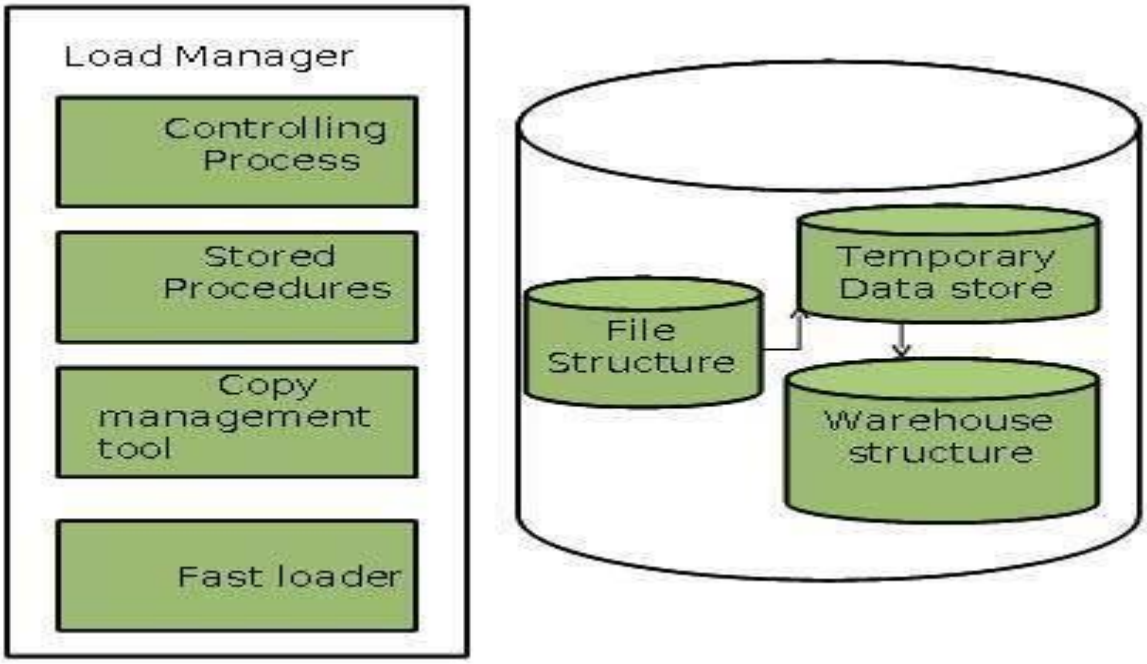
Data Warehouse Load Manager

Load manager performs the operations required to extract and load the data into the database. The size and complexity of a load manager varies between specific solutions from one data warehouse to another.

Load Manager Architecture

The load manager does perform the following functions:

- Extract data from the source system.
- Fast load the extracted data into temporary data store.
- Perform simple transformations into structure similar to the one in the data warehouse.



Extract Data from Source

The data is extracted from the operational databases or the external information providers. Gateways are the application programs that are used to extract data. It is supported by underlying DBMS and allows the client program to generate SQL to be executed at a server. Open Database Connection (ODBC) and Java Database Connection (JDBC) are examples of gateway.

Fast Load

- In order to minimize the total load window, the data needs to be loaded into the warehouse in the fastest possible time.
- Transformations affect the speed of data processing.
- It is more effective to load the data into a relational database prior to applying transformations and checks.
- Gateway technology is not suitable, since they are inefficient when large data volumes are involved.

Simple Transformations

While loading, it may be required to perform simple transformations. After completing simple transformations, we can do complex checks. Suppose we are loading the EPOS sales transaction, we need to perform the following checks:

- Strip out all the columns that are not required within the warehouse.
- Convert all the values to required data types.

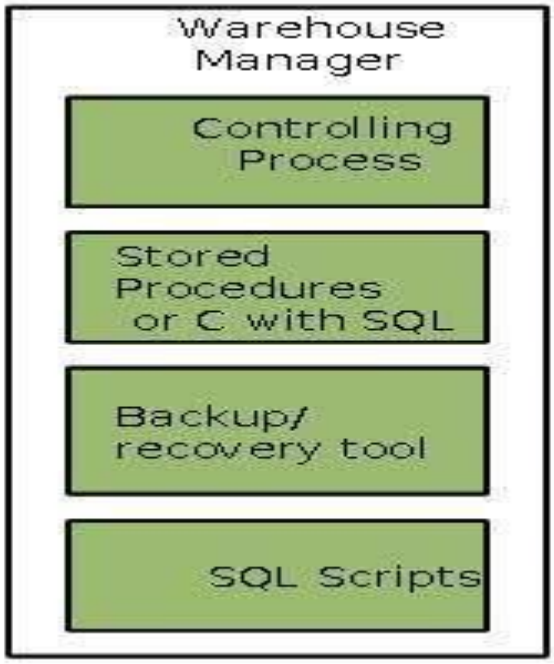
Warehouse Manager

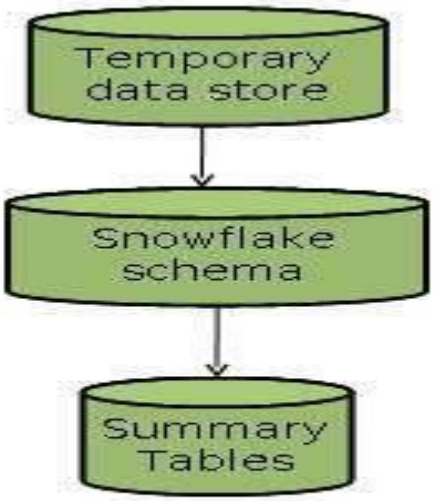
The warehouse manager is responsible for the warehouse management process. It consists of a third-party system software, C programs, and shell scripts. The size and complexity of a warehouse manager varies between specific solutions.

Warehouse Manager Architecture

A warehouse manager includes the following:

- The controlling process
- Stored procedures or C with SQL
- Backup/Recovery tool
- SQL scripts





Functions of Warehouse Manager

A warehouse manager performs the following functions:

- Analyzes the data to perform consistency and referential integrity checks.
- Creates indexes, business views, partition views against the base data.
- Generates new aggregations and updates the existing aggregations.
- Generates normalizations.
- Transforms and merges the source data of the temporary store into the published data warehouse.
- Backs up the data in the data warehouse.
- Archives the data that has reached the end of its captured life.

Note: A warehouse Manager analyzes query profiles to determine whether the index and aggregations are appropriate.

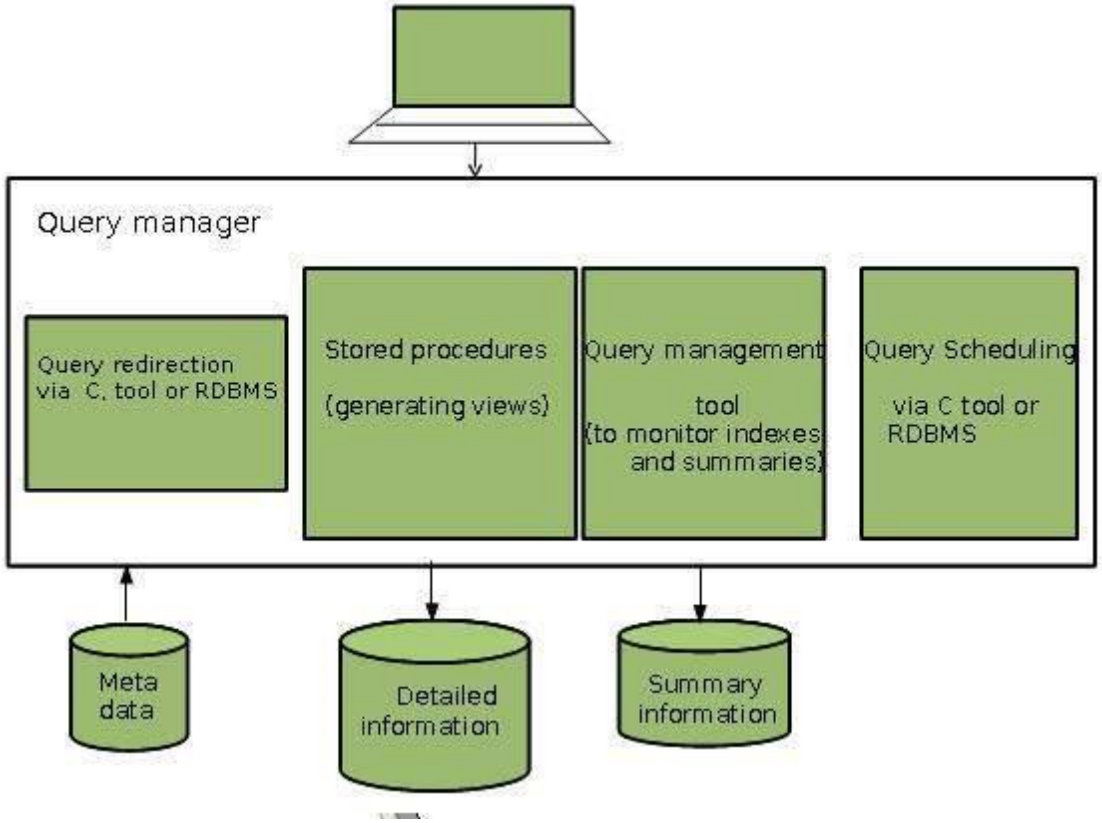
Query Manager

The query manager is responsible for directing the queries to suitable tables. By directing the queries to appropriate tables, it speeds up the query request and response process. In addition, the query manager is responsible for scheduling the execution of the queries posted by the user.

Query Manager Architecture

A query manager includes the following components:

- Query redirection via C tool or RDBMS
- Stored procedures
- Query management tool
- Query scheduling via C tool or RDBMS
- Query scheduling via third-party software



Functions of Query Manager

- It presents the data to the user in a form they understand.
- It schedules the execution of the queries posted by the end-user.

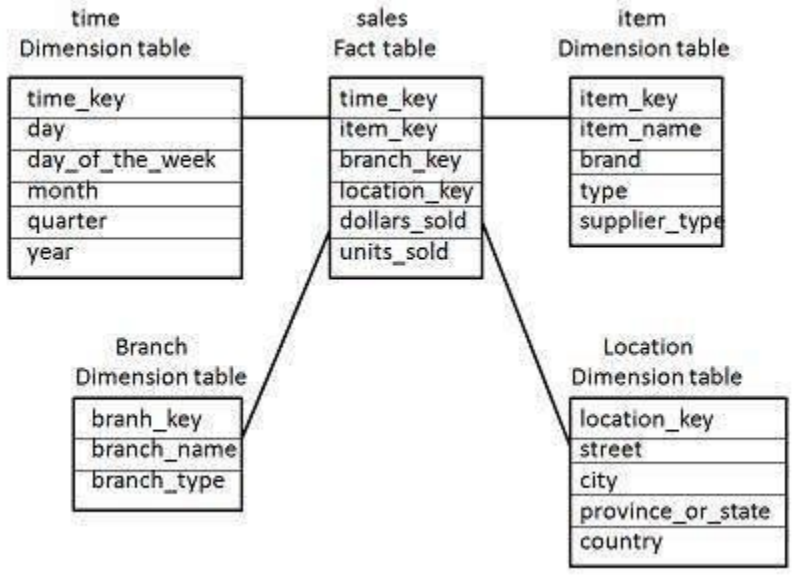
- It stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Schema:

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

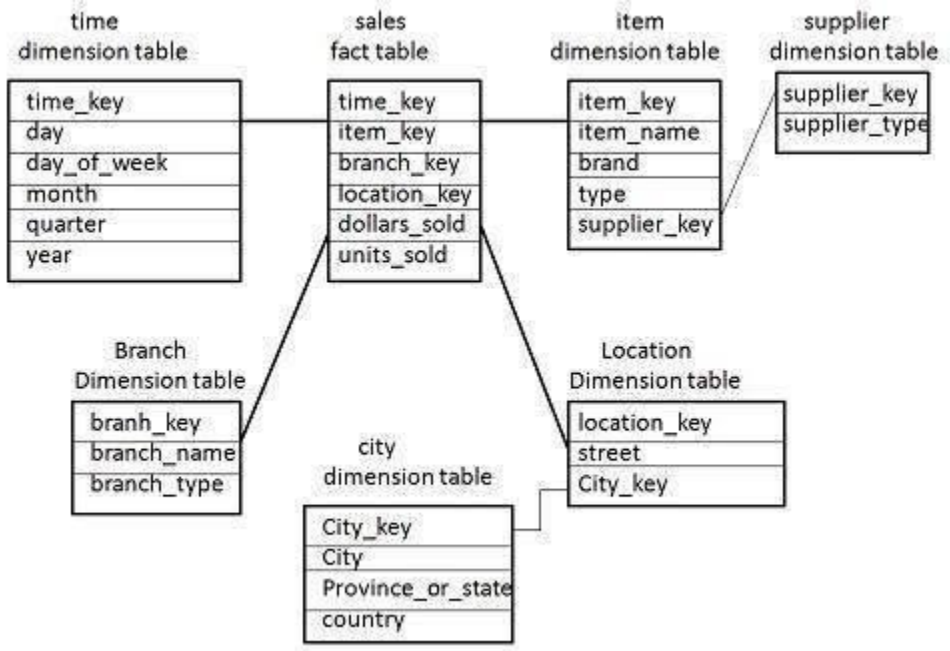


- There is a fact table at the center. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note: Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema is normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

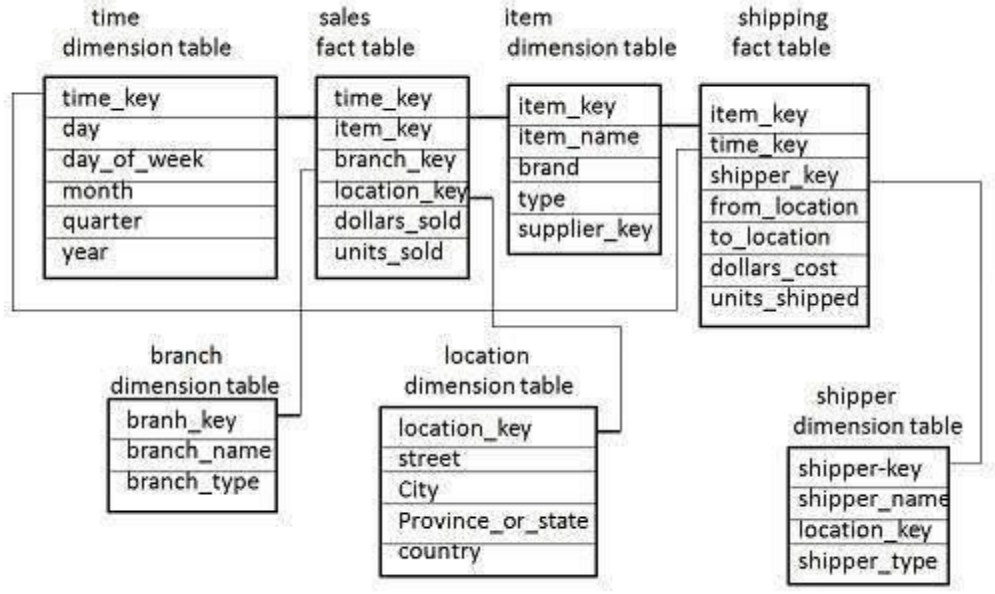


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

<b< style="box-sizing: border-box;">Note: Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.</b<>

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

Schema Definition

Multidimensional schema is defined using Data Mining Query Language (DMQL). The two primitives, cube definition and dimension definition, can be used for defining the data warehouses and data marts.

Syntax for Cube Definition

```
define cube < cube_name > [ < dimension-list > }: < measure_list >
```

Syntax for Dimension Definition

```
define dimension < dimension_name > as ( < attribute_or_dimension_list > )
```

Star Schema Definition

The star schema that we have discussed can be defined using Data Mining Query Language (DMQL) as follows:

```
define cube sales star [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier type)
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city, province or state, country)
```

Snowflake Schema Definition

Snowflake schema can be defined using DMQL as follows:

```
define cube sales snowflake [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier (supplier key, supplier type))
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city (city key, city, province or state, country))
```

Fact Constellation Schema Definition

Fact constellation schema can be defined using DMQL as follows:

```
define cube sales [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier type)
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city, province or state,country)
define cube shipping [time, item, shipper, from location, to location]:

dollars cost = sum(cost in dollars), units shipped = count(*)

define dimension time as time in cube sales
define dimension item as item in cube sales
define dimension shipper as (shipper key, shipper name, location as location in cube sales, shipper type)
define dimension from location as location in cube sales
define dimension to location as location in cube sales
```

Partitioning Strategy:

Partitioning is done to enhance performance and facilitate easy management of data. Partitioning also helps in balancing the various requirements of the system. It optimizes the hardware performance and simplifies the management of data warehouse by partitioning each fact table into multiple separate partitions. In this chapter, we will discuss different partitioning strategies.

Why is it Necessary to Partition?

Partitioning is important for the following reasons:

- For easy management,
- To assist backup/recovery,
- To enhance performance.

For Easy Management

The fact table in a data warehouse can grow up to hundreds of gigabytes in size. This huge size of fact table is very hard to manage as a single entity. Therefore it needs partitioning.

To Assist Backup/Recovery

If we do not partition the fact table, then we have to load the complete fact table with all the data. Partitioning allows us to load only as much data as is required on a regular basis. It reduces the time to load and also enhances the performance of the system.

Note: To cut down on the backup size, all partitions other than the current partition can be marked as read-only. We can then put these partitions into a state where they cannot be modified. Then they can be backed up. It means only the current partition is to be backed up.

To Enhance Performance

By partitioning the fact table into sets of data, the query procedures can be enhanced. Query performance is enhanced because now the query scans only those partitions that are relevant. It does not have to scan the whole data.

Horizontal Partitioning

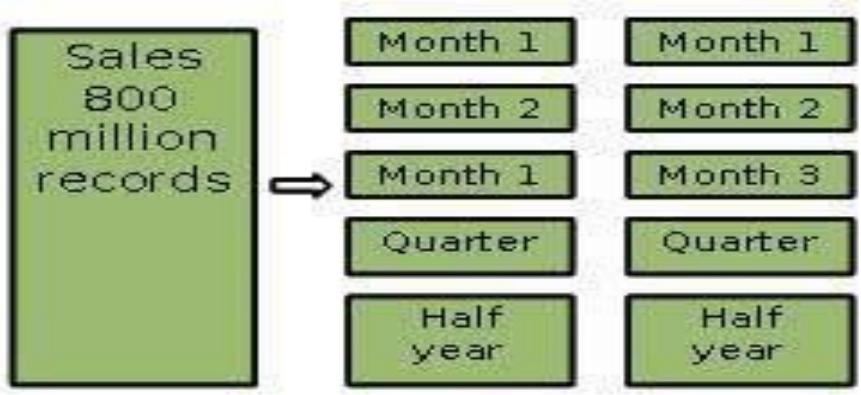
There are various ways in which a fact table can be partitioned. In horizontal partitioning, we have to keep in mind the requirements for manageability of the data warehouse.

Partitioning by Time into Equal Segments

In this partitioning strategy, the fact table is partitioned on the basis of time period. Here each time period represents a significant retention period within the business. For example, if the user queries for **month to date data** then it is appropriate to partition the data into monthly segments. We can reuse the partitioned tables by removing the data in them.

Partition by Time into Different-sized Segments

This kind of partition is done where the aged data is accessed infrequently. It is implemented as a set of small partitions for relatively current data, larger partition for inactive data.



Points to Note

- The detailed information remains available online.
- The number of physical tables is kept relatively small, which reduces the operating cost.
- This technique is suitable where a mix of data dipping recent history and data mining through entire history is required.
- This technique is not useful where the partitioning profile changes on a regular basis, because repartitioning will increase the operation cost of data warehouse.

Partition on a Different Dimension

The fact table can also be partitioned on the basis of dimensions other than time such as product group, region, supplier, or any other dimension. Let's have an example.

Suppose a market function has been structured into distinct regional departments like on a **state by state** basis. If each region wants to query on information captured within its region, it would prove to be more effective to partition the fact table into regional partitions. This will cause the queries to speed up because it does not require to scan information that is not relevant.

Points to Note

- The query does not have to scan irrelevant data which speeds up the query process.
- This technique is not appropriate where the dimensions are unlikely to change in future. So, it is worth determining that the dimension does not change in future.

- If the dimension changes, then the entire fact table would have to be repartitioned.

Note: We recommend performing the partition only on the basis of time dimension, unless you are certain that the suggested dimension grouping will not change within the life of the data warehouse.

Partition by Size of Table

When there is no clear basis for partitioning the fact table on any dimension, then we should **partition the fact table on the basis of their size.** We can set the predetermined size as a critical point. When the table exceeds the predetermined size, a new table partition is created.

Points to Note

- This partitioning is complex to manage.

It requires metadata to identify what data is stored in each partition.

Partitioning Dimensions

If a dimension contains large number of entries, then it is required to partition the dimensions. Here we have to check the size of a dimension.

Consider a large design that change over time. If we need to store all the variations in order to apply comparisons, that dimension may be very large. This would definitely affect the response time.

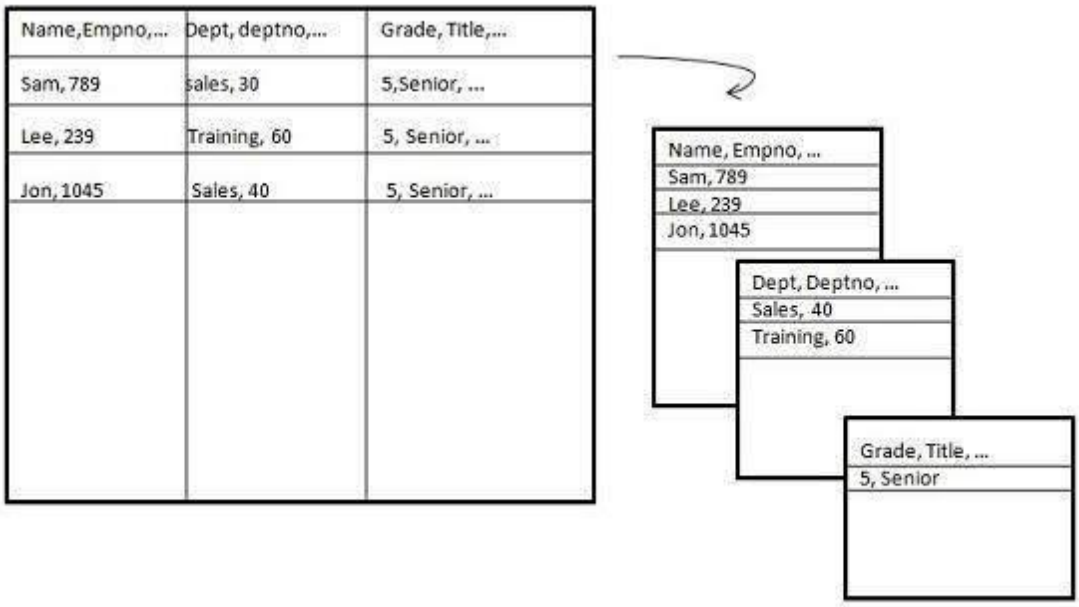
Round Robin Partitions

In the round robin technique, when a new partition is needed, the old one is archived. It uses metadata to allow user access tool to refer to the correct table partition.

This technique makes it easy to automate table management facilities within the data warehouse.

Vertical Partition

Vertical partitioning splits the data vertically. The following images depicts how vertical partitioning is done.



Vertical partitioning can be performed in the following two ways:

- Normalization
- Row Splitting

Normalization

Normalization is the standard relational method of database organization. In this method, the rows are collapsed into a single row, hence it reduce space. Take a look at the following tables that show how normalization is performed.

Table before Normalization

Product_id	Qty	Value	sales_date	Store_id	Store_name	Location	Region
30	5	3.67	3-Aug-13	16	sunny	Bangalore	S
35	4	5.33	3-Sep-13	16	sunny	Bangalore	S

40	5	2.50	3-Sep-13	64	san	Mumbai	W
45	7	5.66	3-Sep-13	16	sunny	Bangalore	S

Table after Normalization

Store_id	Store_name	Location	Region
16	sunny	Bangalore	W
64	san	Mumbai	S

Product_id	Quantity	Value	sales_date	Store_id
30	5	3.67	3-Aug-13	16
35	4	5.33	3-Sep-13	16
40	5	2.50	3-Sep-13	64
45	7	5.66	3-Sep-13	16

Row Splitting

Row splitting tends to leave a one-to-one map between partitions. The motive of row splitting is to speed up the access to large table by reducing its size.

Note: While using vertical partitioning, make sure that there is no requirement to perform a major join operation between two partitions.

Identify Key to Partition

It is very crucial to choose the right partition key. Choosing a wrong partition key will lead to reorganizing the fact table. Let's have an example. Suppose we want to partition the following table.

Account_Txn_Table
transaction_id
account_id
transaction_type
value
transaction_date
region
branch_name

We can choose to partition on any key. The two possible keys could be

- region
- transaction_date

Suppose the business is organized in 30 geographical regions and each region has different number of branches. That will give us 30 partitions, which is reasonable. This partitioning is good enough because our requirements capture has shown that a vast majority of queries are restricted to the user's own business region.

If we partition by transaction_date instead of region, then the latest transaction from every region will be in one partition. Now the user who wants to look at data within his own region has to query across multiple partitions.

Data Mart;

Why Do We Need a Data Mart?

Listed below are the reasons to create a data mart:

- To partition data in order to impose **access control strategies.**
- To speed up the queries by reducing the volume of data to be scanned.
- To segment data into different hardware platforms.

- To structure data in a form suitable for a user access tool.

Note: Do not data mart for any other reason since the operation cost of data marting could be very high. Before data marting, make sure that data marting strategy is appropriate for your particular solution.

Cost-effective Data Marting

Follow the steps given below to make data marting cost-effective:

- Identify the Functional Splits
- Identify User Access Tool Requirements
- Identify Access Control Issues

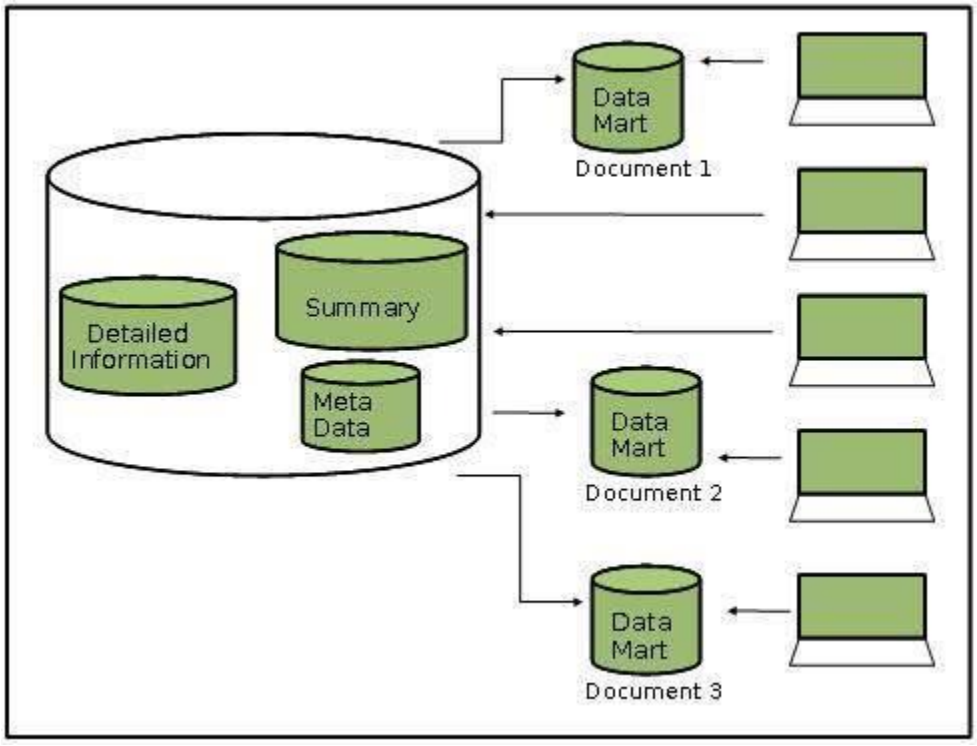
Identify the Functional Splits

In this step, we determine if the organization has natural functional splits. We look for departmental splits, and we determine whether the way in which departments use information tend to be in isolation from the rest of the organization. Let's have an example.

Consider a retail organization, where each merchant is accountable for maximizing the sales of a group of products. For this, the following are the valuable information:

- sales transaction on a daily basis
- sales forecast on a weekly basis
- stock position on a daily basis
- stock movements on a daily basis

As the merchant is not interested in the products they are not dealing with, the data marting is a subset of the data dealing which the product group of interest. The following diagram shows data marting for different users.



Given below are the issues to be taken into account while determining the functional split:

- The structure of the department may change.
- The products might switch from one department to other.
- The merchant could query the sales trend of other products to analyze what is happening to the sales.

Note: We need to determine the business benefits and technical feasibility of using a data mart.

Identify User Access Tool Requirements

We need data marts to support **user access tools** that require internal data structures. The data in such structures are outside the control of data warehouse but need to be populated and updated on a regular basis.

There are some tools that populate directly from the source system but some cannot. Therefore additional requirements outside the scope of the tool are needed to be identified for future.

Note: In order to ensure consistency of data across all access tools, the data should not be directly populated from the data warehouse, rather each tool must have its own data mart.

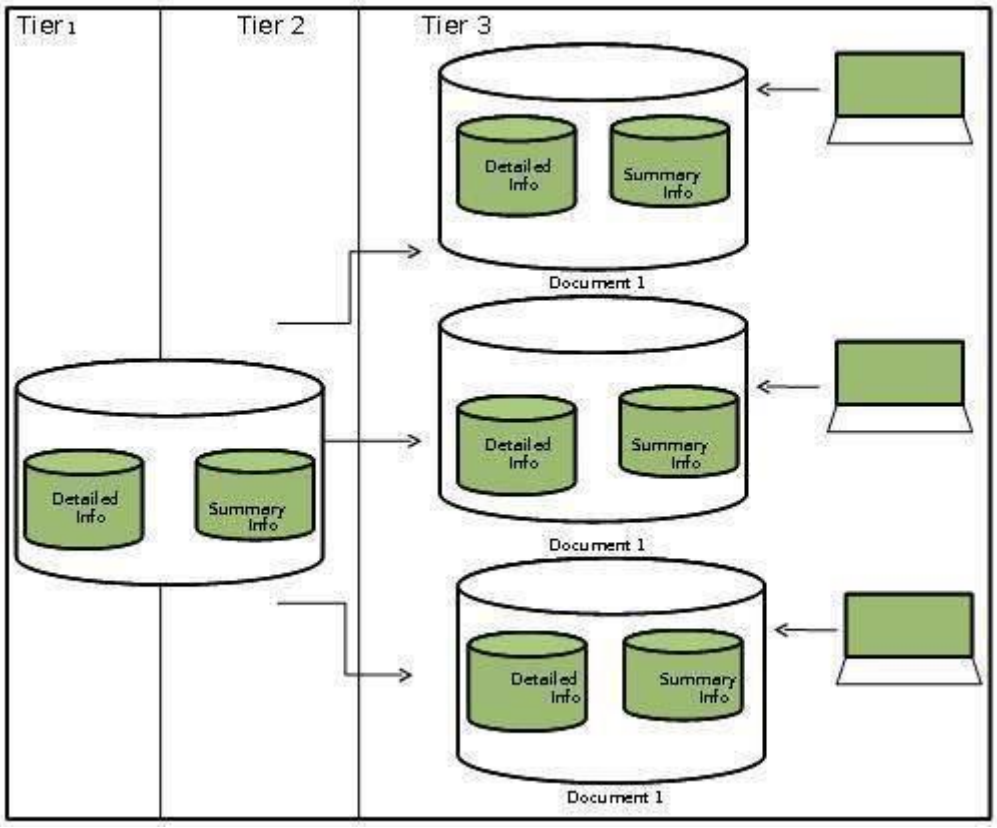
Identify Access Control Issues

There should to be privacy rules to ensure the data is accessed by authorized users only. For example a data warehouse for retail banking institution ensures that all the accounts belong to the same legal entity. Privacy laws can force you to totally prevent access to information that is not owned by the specific bank.

Data marts allow us to build a complete wall by physically separating data segments within the data warehouse. To avoid possible privacy problems, the detailed data can be removed from the data warehouse. We can create data mart for each legal entity and load it via data warehouse, with detailed account data.

Designing Data Marts

Data marts should be designed as a smaller version of starflake schema within the data warehouse and should match with the database design of the data warehouse. It helps in maintaining control over database instances.



The summaries are data marted in the same way as they would have been designed within the data warehouse. Summary tables help to utilize all dimension data in the starflake schema.

Cost of Data Marting

The cost measures for data marting are as follows:

- Hardware and Software Cost
- Network Access
- Time Window Constraints

Hardware and Software Cost

Although data marts are created on the same hardware, they require some additional hardware and software. To handle user queries, it requires additional processing power and disk storage. If detailed data and the data mart exist within the data warehouse, then we would face additional cost to store and manage replicated data.

Note: Data marting is more expensive than aggregations, therefore it should be used as an additional strategy and not as an alternative strategy.

Network Access

A data mart could be on a different location from the data warehouse, so we should ensure that the LAN or WAN has the capacity to handle the data volumes being transferred within the data mart load process.

Time Window Constraints

The extent to which a data mart loading process will eat into the available time window depends on the complexity of the transformations and the data volumes being shipped. The determination of how many data marts are possible depends on:

- Network capacity.
- Time window available
- Volume of data being transferred
- Mechanisms being used to insert data into a data mart

Meta Data:

What is Metadata?

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 other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.

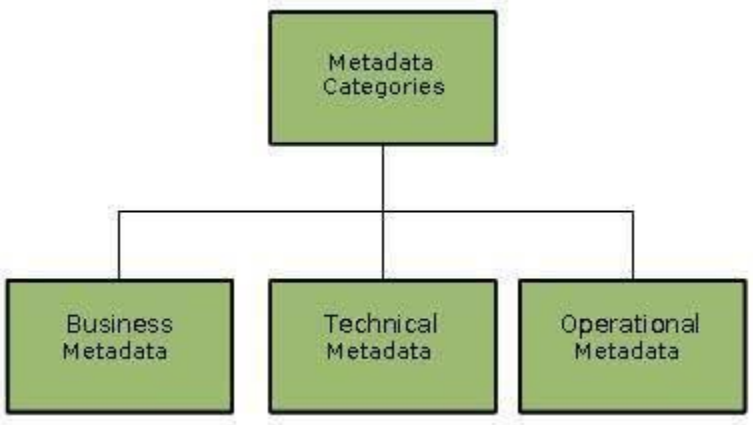
- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Note: In a data warehouse, we create metadata for the data names and definitions of a given data warehouse. Along with this metadata, additional metadata is also created for time-stamping any extracted data, the source of extracted data.

Categories of Metadata

Metadata can be broadly categorized into three categories:

- **Business Metadata** - It has the data ownership information, business definition, and changing policies.
- **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.
- **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.



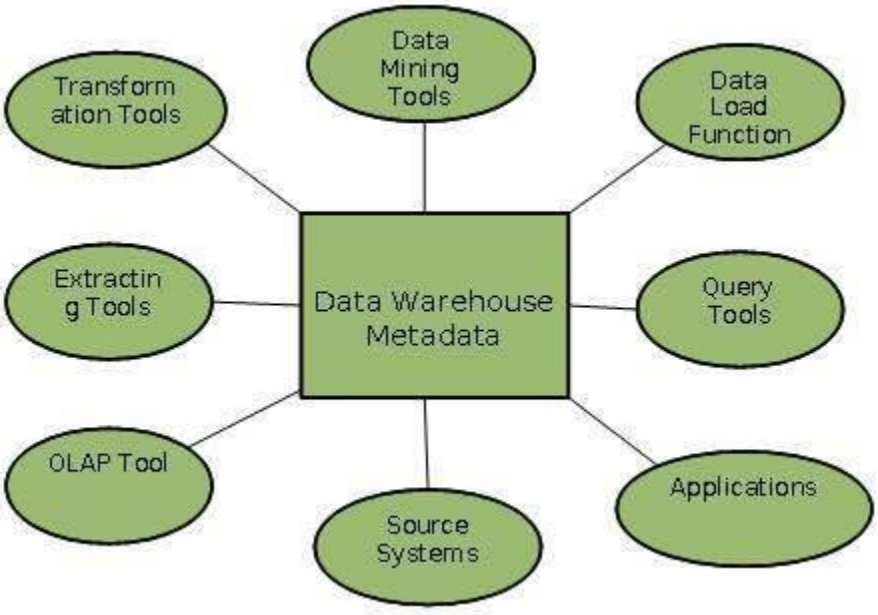
Role of Metadata

Metadata has a very important role in a data warehouse. The role of metadata in a warehouse is different from the warehouse data, yet it plays an important role. The various roles of metadata are explained below.

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.

- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

The following diagram shows the roles of metadata.



Metadata Repository

Metadata repository is an integral part of a data warehouse system. It has the following metadata:

- **Definition of data warehouse** - It includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.

- **Business metadata** - It contains has the data ownership information, business definition, and changing policies.

- **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.

- **Data for mapping from operational environment to data warehouse** - It includes the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.

- **Algorithms for summarization** - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Challenges for Metadata Management

The importance of metadata can not be overstated. Metadata helps in driving the accuracy of reports, validates data transformation, and ensures the accuracy of calculations. Metadata also enforces the definition of business terms to business end-users. With all these uses of metadata, it also has its challenges. Some of the challenges are discussed below.

- Metadata in a big organization is scattered across the organization. This metadata is spread in spreadsheets, databases, and applications.

- Metadata could be present in text files or multimedia files. To use this data for information management solutions, it has to be correctly defined.

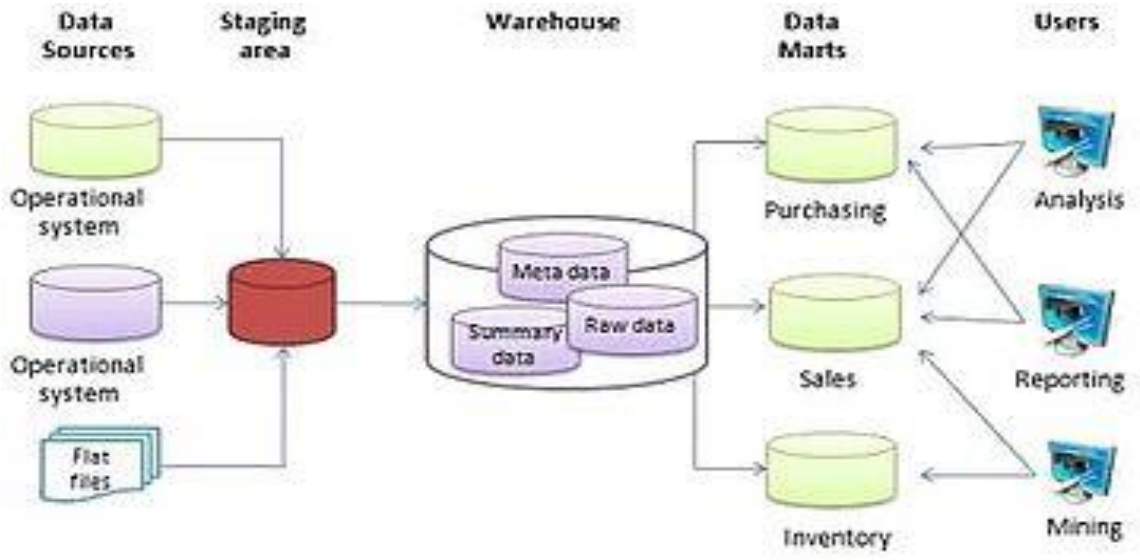
- There are no industry-wide accepted standards. Data management solution vendors have narrow focus.

- There are no easy and accepted methods of passing metadata.

Aggregation:

Data aggregation is any process in which 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.

Aggregates are used in dimensional models of the **data warehouse** to produce dramatic positive effects on the time it takes to query large sets of **data**. At the simplest form an **aggregate** is a simple summary table that can be derived by performing a Group by SQL query.



POSSIBLE QUESTIONS

PART – B

8 MARKS

1. Describe Data warehousing architecture.
2. Write a note on Database schema.
   A. Star flake Schemas
   B. Identifying facts and dimensions.
3. Discuss 1. Warehouse manager 2. Query manager with neat diagram.
4. Write about Horizontal Partitioning and Vertical Partitioning.
5. Explain 1. Extract and Load Process
   2. Clean and Transform the data
   3. Query Management Process with an example
6. Describe the concept of data marting in detail.
7. Explain process architecture with neat sketch.
8. Explain about aggregations with neat sketch.
9. Explain metadata in detail.

UNIT IV

SYLLABUS

Hardware and Operational Design : Hardware Architecture – Physical layout – Security – Backup and Recovery – Service Level Agreement – Operating and Data Warehousing.

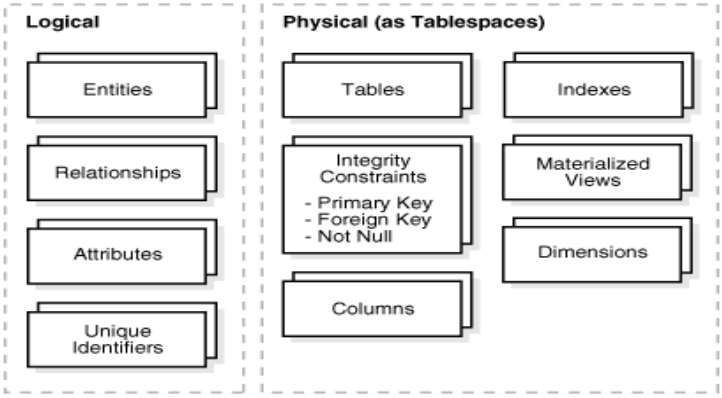
Physical Design in Data Warehouses

Physical Design

During the logical design phase, you defined a model for your data warehouse consisting of entities, attributes, and relationships. The entities are linked together using relationships. Attributes are used to describe the entities. The **unique identifier** (UID) distinguishes between one instance of an entity and another.

The following diagram illustrates a graphical way of distinguishing between logical and physical designs.

Logical Design Compared with Physical Design



Logical Design Compared with Physical Design

During the physical design process, you translate the expected schemas into actual database structures. At this time, you have to map:

- Entities to tables
- Relationships to foreign key constraints
- Attributes to columns
- Primary unique identifiers to primary key constraints
- Unique identifiers to unique key constraints

Physical Design Structures

Once you have converted your logical design to a physical one, you will need to create some or all of the following structures:

- Tablespaces
- Tables and Partitioned Tables
- Views
- Integrity Constraints
- Dimensions

Some of these structures require disk space. Others exist only in the data dictionary. Additionally, the following structures may be created for performance improvement:

- Indexes and Partitioned Indexes
- Materialized Views

Tablespaces

A tablespace consists of one or more datafiles, which are physical structures within the operating system you are using. A datafile is associated with only one tablespace. From a design perspective, tablespaces are containers for physical design structures.

Tablespaces need to be separated by differences. For example, tables should be separated from their indexes and small tables should be separated from large tables. Tablespaces should also represent logical business units if possible. Because a tablespace is the coarsest granularity for backup and recovery or the transportable tablespaces mechanism, the logical business design affects availability and maintenance operations.

You can now use ultralarge data files, a significant improvement in very large databases.

Tables and Partitioned Tables

Tables are the basic unit of data storage. They are the container for the expected amount of raw data in your data warehouse.

Using partitioned tables instead of nonpartitioned ones addresses the key problem of supporting very large data volumes by allowing you to divide them into smaller and more manageable pieces. The main design criterion for partitioning is manageability, though you will also see performance benefits in most cases because of partition pruning or intelligent parallel processing. For example, you might choose a partitioning strategy based on a sales transaction date and a monthly granularity. If you have four years' worth of data, you can delete a month's data as it becomes older than four years with a single, fast DDL statement and load new data while only affecting 1/48th of the complete table. Business questions regarding the last quarter will only affect three months, which is equivalent to three partitions, or 3/48ths of the total volume.

Partitioning large tables improves performance because each partitioned piece is more manageable. Typically, you partition based on transaction dates in a data warehouse. For example, each month, one month's worth of data can be assigned its own partition.

Table Compression

You can save disk space by compressing heap-organized tables. A typical type of heap-organized table you should consider for table compression is partitioned tables.

To reduce disk use and memory use (specifically, the buffer cache), you can store tables and partitioned tables in a compressed format inside the database. This often leads to a better scaleup for read-only operations. Table compression can also speed up query execution. There is, however, a cost in CPU overhead.

Table compression should be used with highly redundant data, such as tables with many foreign keys. You should avoid compressing tables with much update or other DML activity. Although compressed tables or partitions are updatable, there is some overhead in updating these tables, and high update activity may work against compression by causing some space to be wasted.

Views

A view is a tailored presentation of the data contained in one or more tables or other views. A view takes the output of a query and treats it as a table. Views do not require any space in the database.

Integrity Constraints

Integrity constraints are used to enforce business rules associated with your database and to prevent having invalid information in the tables. Integrity constraints in data warehousing differ from constraints in OLTP environments. In OLTP environments, they primarily prevent the insertion of invalid data into a record, which is not a big problem in data warehousing environments because accuracy has already been guaranteed. In data warehousing environments, constraints are only used for query rewrite. NOT NULL constraints are particularly common in data warehouses. Under some specific circumstances, constraints need space in the database. These constraints are in the form of the underlying unique index.

Indexes and Partitioned Indexes

Indexes are optional structures associated with tables or clusters. In addition to the classical B-tree indexes, bitmap indexes are very common in data warehousing environments. Bitmap indexes are optimized index structures for set-oriented operations. Additionally, they are necessary for some optimized data access methods such as star transformations.

Indexes are just like tables in that you can partition them, although the partitioning strategy is not dependent upon the table structure. Partitioning indexes makes it easier to manage the data warehouse during refresh and improves query performance.

Materialized Views

Materialized views are query results that have been stored in advance so long-running calculations are not necessary when you actually execute your SQL statements. From a physical design point of view, materialized views resemble tables or partitioned tables and behave like indexes in that they are used transparently and improve performance.

Dimensions

A dimension is a schema object that defines hierarchical relationships between columns or column sets. A hierarchical relationship is a functional dependency from one level of a hierarchy to the next one. A dimension is a container of logical relationships and does not require any space in the database. A typical dimension is city, state (or province), region, and country.

Data Warehousing - Security

The objective of a data warehouse is to make large amounts of data easily accessible to the users, hence allowing the users to extract information about the business as a whole. But we know that there could be some security restrictions applied on the data that can be an obstacle for accessing the information. If the analyst has a restricted view of data, then it is impossible to capture a complete picture of the trends within the business.

The data from each analyst can be summarized and passed on to management where the different summaries can be aggregated. As the aggregations of summaries cannot be the same as that of the aggregation as a whole, it is possible to miss some information trends in the data unless someone is analyzing the data as a whole.

Security Requirements

Adding security features affect the performance of the data warehouse, therefore it is important to determine the security requirements as early as possible. It is difficult to add security features after the data warehouse has gone live.

During the design phase of the data warehouse, we should keep in mind what data sources may be added later and what would be the impact of adding those data sources. We should consider the following possibilities during the design phase.

- Whether the new data sources will require new security and/or audit restrictions to be implemented?
- Whether the new users added who have restricted access to data that is already generally available?

This situation arises when the future users and the data sources are not well known. In such a situation, we need to use the knowledge of business and the objective of data warehouse to know likely requirements.

The following activities get affected by security measures:

- User access
- Data load
- Data movement
- Query generation

User Access

We need to first classify the data and then classify the users on the basis of the data they can access. In other words, the users are classified according to the data they can access.

Data Classification

The following two approaches can be used to classify the data:

- Data can be classified according to its sensitivity. Highly-sensitive data is classified as highly restricted and less-sensitive data is classified as less restrictive.
- Data can also be classified according to the job function. This restriction allows only specific users to view particular data. Here we restrict the users to view only that part of the data in which they are interested and are responsible for.

There are some issues in the second approach. To understand, let's have an example. Suppose you are building the data warehouse for a bank. Consider that the data being stored in the data warehouse is the transaction data for all the accounts. The question here is, who is allowed to see the transaction data. The solution lies in classifying the data according to the function.

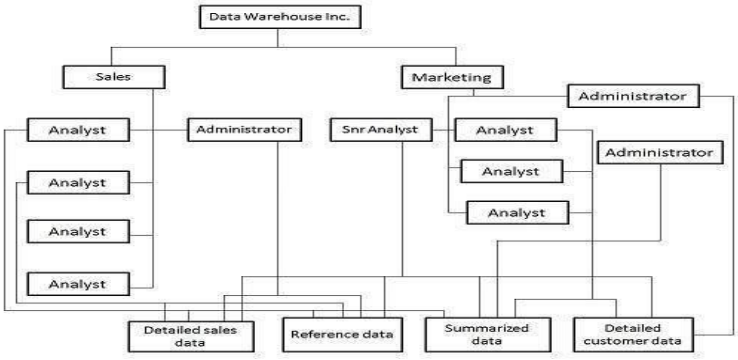
User classification

The following approaches can be used to classify the users:

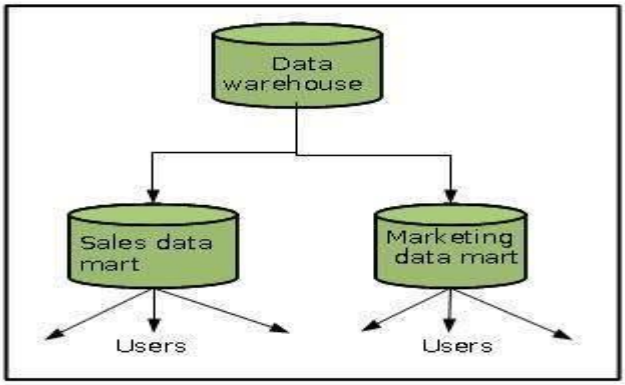
- Users can be classified as per the hierarchy of users in an organization, i.e., users can be classified by departments, sections, groups, and so on.
- Users can also be classified according to their role, with people grouped across departments based on their role.

Classification on basis of Department

Let's have an example of a data warehouse where the users are from sales and marketing department. We can have security by top-to-down company view, with access centered on the different departments. But there could be some restrictions on users at different levels. This structure is shown in the following diagram.

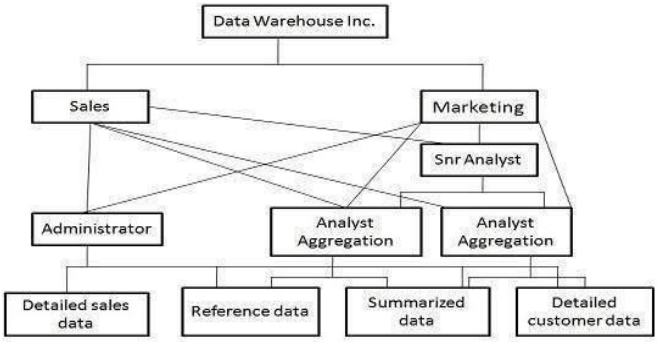


But if each department accesses different data, then we should design the security access for each department separately. This can be achieved by departmental data marts. Since these data marts are separated from the data warehouse, we can enforce separate security restrictions on each data mart. This approach is shown in the following figure.



Classification on basis of Role

If the data is generally available to all the departments, then it is useful to follow the role access hierarchy. In other words, if the data is generally accessed by all If the data is generally available to all the departments, then it is useful to follow the role access hierarchy. In other words, if the data is generally accessed by all



Audit Requirements

Auditing is a subset of security, a costly activity. Auditing can cause heavy overheads on the system. To complete an audit in time, we require more hardware and therefore, it is recommended that wherever possible, auditing should be switched off. Audit requirements can be categorized as follows:

- Connections
- Disconnections
- Data access
- Data change

Note : For each of the above-mentioned categories, it is necessary to audit success, failure, or both. From the perspective of security reasons, the auditing of failures are very important. Auditing of failure is important because they can highlight unauthorized or fraudulent access.

Network Requirements

Network security is as important as other securities. We cannot ignore the network security requirement. We need to consider the following issues:

- Is it necessary to encrypt data before transferring it to the data warehouse?
- Are there restrictions on which network routes the data can take?

These restrictions need to be considered carefully. Following are the points to remember:

- The process of encryption and decryption will increase overheads. It would require more processing power and processing time.
- The cost of encryption can be high if the system is already a loaded system because the encryption is borne by the source system.

Data Movement

There exist potential security implications while moving the data. Suppose we need to transfer some restricted data as a flat file to be loaded. When the data is loaded into the data warehouse, the following questions are raised:

- Where is the flat file stored?
- Who has access to that disk space?

If we talk about the backup of these flat files, the following questions are raised:

- Do you backup encrypted or decrypted versions?
- Do these backups need to be made to special tapes that are stored separately?
- Who has access to these tapes?

Some other forms of data movement like query result sets also need to be considered. The questions raised while creating the temporary table are as follows:

- Where is that temporary table to be held?
- How do you make such table visible?

We should avoid the accidental flouting of security restrictions. If a user with access to the restricted data can generate accessible temporary tables, data can be visible to non-authorized users. We can overcome this problem by having a separate temporary area for users with access to restricted data.

Documentation

The audit and security requirements need to be properly documented. This will be treated as a part of justification. This document can contain all the information gathered from:

- Data classification
- User classification
- Network requirements
- Data movement and storage requirements
- All auditable actions

Impact of Security on Design

Security affects the application code and the development timescales. Security affects the following area.

- Application development
- Database design

- Testing

Application Development

Security affects the overall application development and it also affects the design of the important components of the data warehouse such as load manager, warehouse manager, and query manager. The load manager may require checking code to filter record and place them in different locations. More transformation rules may also be required to hide certain data. Also there may be requirements of extra metadata to handle any extra objects.

To create and maintain extra views, the warehouse manager may require extra codes to enforce security. Extra checks may have to be coded into the data warehouse to prevent it from being fooled into moving data into a location where it should not be available. The query manager requires the changes to handle any access restrictions. The query manager will need to be aware of all extra views and aggregations.

Database design

The database layout is also affected because when security measures are implemented, there is an increase in the number of views and tables. Adding security increases the size of the database and hence increases the complexity of the database design and management. It will also add complexity to the backup management and recovery plan.

Testing

Testing the data warehouse is a complex and lengthy process. Adding security to the data warehouse also affects the testing time complexity. It affects the testing in the following two ways:

- It will increase the time required for integration and system testing.
- There is added functionality to be tested which will increase the size of the testing suite.

Data Warehousing – Backup and Recovery

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. In this chapter, we will discuss the issues in designing the backup strategy.

Backup Terminologies

Before proceeding further, you should know some of the backup terminologies discussed below.

- **Complete backup** - It backs up the entire database at the same time. This backup includes all the database files, control files, and journal files.
- **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.
- **Cold backup** - Cold backup is taken while the database is completely shut down. In multi-instance environment, all the instances should be shut down.
- **Hot backup** - Hot backup is taken when the database engine is up and running. The requirement of hot backup varies from RDBMS to RDBMS.
- **Online backup** - It is quite similar to hot backup.

Hardware Backup

It is important to decide which hardware to use for the backup. The speed of processing the backup and restore depends on the hardware being used, how the hardware is connected, bandwidth of the network, backup software, and the speed of server's I/O system. Here we will discuss some of the hardware choices that are available and their pros and cons. These choices are as follows:

- Tape Technology
- Disk Backups

Tape Technology

The tape choice can be categorized as follows:

- Tape media
- Standalone tape drives
- Tape stackers
- Tape silos

Tape Media

There exists several varieties of tape media. Some tape media standards are listed in the table below:

Tape Media	Capacity	I/O rates
DLT	40 GB	3 MB/s
3490e	1.6 GB	3 MB/s
8 mm	14 GB	1 MB/s

Other factors that need to be considered are as follows:

- Reliability of the tape medium
- Cost of tape medium per unit
- Scalability
- Cost of upgrades to tape system
- Cost of tape medium per unit
- Shelf life of tape medium

Standalone tape drives

The tape drives can be connected in the following ways:

- Direct to the server
- As network available devices
- Remotely to other machine

There could be issues in connecting the tape drives to a data warehouse.

- Consider the server is a 48node MPP machine. We do not know the node to connect the tape drive and we do not know how to spread them over the server nodes to get the optimal performance with least disruption of the server and least internal I/O latency.
- Connecting the tape drive as a network available device requires the network to be up to the job of the huge data transfer rates. Make sure that sufficient bandwidth is available during the time you require it.
- Connecting the tape drives remotely also require high bandwidth.

Tape Stackers

The method of loading multiple tapes into a single tape drive is known as tape stackers. The stacker dismounts the current tape when it has finished with it and loads the next tape, hence only one tape is available at a time to be accessed. The price and the capabilities may vary, but the common ability is that they can perform unattended backups.

Tape Silos

Tape silos provide large store capacities. Tape silos can store and manage thousands of tapes. They can integrate multiple tape drives. They have the software and hardware to label and store the tapes they store. It is very common for the silo to be connected remotely over a network or a dedicated link. We should ensure that the bandwidth of the connection is up to the job.

Disk Backups

Methods of disk backups are:

- Disk-to-disk backups
- Mirror breaking

These methods are used in the OLTP system. These methods minimize the database downtime and maximize the availability.

Disk-to-disk backups

Here backup is taken on the disk rather on the tape. Disk-to-disk backups are done for the following reasons:

- Speed of initial backups
- Speed of restore

Backing up the data from disk to disk is much faster than to the tape. However it is the intermediate step of backup. Later the data is backed up on the tape. The other advantage of disk-to-disk backups is that it gives you an online copy of the latest backup.

Mirror Breaking

The idea is to have disks mirrored for resilience during the working day. When backup is required, one of the mirror sets can be broken out. This technique is a variant of disk-to-disk backups.

Note: The database may need to be shutdown to guarantee consistency of the backup.

Optical Jukeboxes

Optical jukeboxes allow the data to be stored near line. This technique allows a large number of optical disks to be managed in the same way as a tape stacker or a tape silo. The drawback of this technique is that it has slow write speed than disks. But the optical media provides long-life and reliability that makes them a good choice of medium for archiving.

Software Backups

There are software tools available that help in the backup process. These software tools come as a package. These tools not only take backup, they can effectively manage and control the backup strategies. There are many software packages available in the market. Some of them are listed in the following table:

Package Name	Vendor
Networker	Legato
ADSM	IBM
Epoch	Epoch Systems
Omniback II	HP
Alexandria	Sequent

Data Warehousing Service-Level Agreements

A service-level agreement (SLA) is a contract that spells out in measurable terms what services a provider will deliver to a customer. Though conventional wisdom suggests that SLAs accompany data warehouses, little has been written about data warehouse SLAs. Let's take a closer look at data warehouse SLAs—their benefits, the areas of quality they should cover, and some of the implications and solution components for successful implementation.

Benefits

Among others, three key SLA benefits include:

- **User satisfaction**—When service levels are discussed, set, and measured, users will know what to expect. As a result, they will develop confidence in the data warehouse and the people who build it.
- **Data-driven decisions**—Data warehouses are built so that users can make data-driven decisions. Measuring data warehousing service levels is a good way to practice what we preach, thus promoting good analytic behavior. Also, service-level measures facilitate problem diagnosis, capacity planning, ROI, and cost justification.
- **Operational excellence**—Service-level agreements help us measure our progress toward efficiency and operational excellence.

In the days of operational transaction-processing systems, a service level agreement (SLA) was the agreement between the users and the IT department governing the expectations of the online environment. SLAs were created as a means of measuring and managing the service levels delivered to the end users for that environment. This agreement was a formal statement regarding proper service levels. Although it was somewhat like a contract, it was never a legally enforceable instrument.

SLAs covered two aspects of the online environment: response time and availability. Some SLAs were elaborate; some SLAs were very simple. A typical SLA would be:

- Response time – 95% of all transactions will be executed in less than 4 seconds.
- Availability – the system will be up and available 99% of the time between the hours of 8:00 a.m. and 6:00 p.m.

Now that the world has embraced business intelligence and data warehousing, these environments also need SLAs. However, the SLAs that govern these environments are fundamentally different from the SLAs that govern the operational environment.

In order to understand the difference between the SLAs, consider what can be termed *data warehouse velocity*.

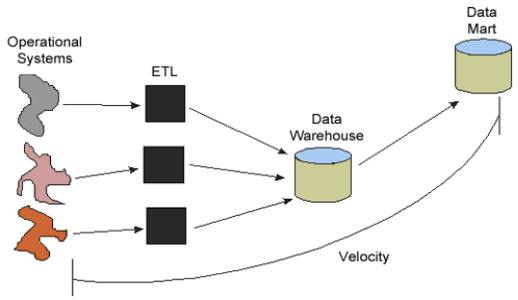


Figure 1: Data Warehouse Velocity

Data warehouse velocity refers to the speed at which data moves through the business intelligence/data warehouse environment, from the initial entry into the operational environment, through ETL (extract, transform and load) and into the data warehouse, and finally to the data mart environment. Data warehouse velocity measures how quickly data becomes available throughout the data warehouse environment.

Note that velocity is a measure of availability. Some data marts are "pull" processes and only pull the data infrequently. Data may not flow for three months, for example. However, if needed, the data could be pulled on a daily basis. Therefore, actual velocity is different from potential velocity. Potential velocity refers to the speed that could be attained if the data warehouse operated on push processes, not pull processes.

The data warehouse SLAs rely on potential velocity. A typical business intelligence/data warehouse SLA would be:

Potential velocity – 48 hours from entry to data mart usage

One aspect of SLAs that remains fairly constant over the online and the business intelligence/data warehouse environment is that of availability. Both environments need a definition of when the system is available. It is response time and velocity that are very different from one environment to another.

Note that the flow of data may be for a data element that changes. The ETL process may alter the data, and the entry into the data mart environment also may alter the data.

Therefore, the exact unit of data may not flow through the business intelligence/data warehouse environment at all. Instead, the mutated form of the data flows through the business intelligence/data warehouse environment.

SLA Categories

Before we delve into data warehousing–specific SLA categories, let's review some universal categories:

1. Uptime (data and system availability)—Users and other systems depend on the data warehouse to be open and ready for business. Time-of-day data availability targets are normally considered to be one of the key components of a data warehouse SLA. Uptime agreements include the times the system will be available for use; communication methods for planned outages; and consequences and penalties for unplanned outages.

2. Performance—Performance agreements with users are often written in terms of average and worst-case response times, and average and peak concurrent users. Performance agreements with other systems involve delivery/consumption latencies for event-driven interactions and batch windows for bulk interactions.

3. Problem resolution—Problem resolution agreements define problem classes. For each class, they define responsible parties, maximum resolution times, and communication processes.

4. Business continuity—Continuity measures and recovery times from catastrophic system failures should be established. Recovery times should not only address data loss and the time required to restore the database, but should also take into account data collection, data staging, and "catch-up" processing times.

POSSIBLE QUESTIONS

PART – B

8 MARKS

1. Explain Physical layout in detail.
2. Write about operating and Data warehousing.
3. Explain Server hardware and Network Hardware with neat sketch.
4. Describe User Requirement and System Requirement in detail.
5. Describe Security in Hardware and Operational Design.
6. Write a note on Hardware, Software in Backup and Recovery.
7. Discuss Hardware architecture in detail.
8. Explain service level agreement.
9. Explain the concept of backup and recovery in detail.
10. Write about Disk technology and database layout.

UNIT – V

SYLLABUS

Capacity planning – Tuning and Data Warehouse – Testing and Data Warehouse – Data Warehouse Futures. Application: Data warehousing and data mining in government: Introduction-national data warehouses-other areas for data warehousing and data mining.

Data Warehousing - Tuning

A data warehouse keeps evolving and it is unpredictable what query the user is going to post in the future. Therefore it becomes more difficult to tune a data warehouse system. In this chapter, we will discuss how to tune the different aspects of a data warehouse such as performance, data load, queries, etc.

Difficulties in Data Warehouse Tuning

Tuning a data warehouse is a difficult procedure due to following reasons:

- Data warehouse is dynamic; it never remains constant.
- It is very difficult to predict what query the user is going to post in the future.
- Business requirements change with time.
- Users and their profiles keep changing.
- The user can switch from one group to another.
- The data load on the warehouse also changes with time.

Note: It is very important to have a complete knowledge of data warehouse.

Performance Assessment

Here is a list of objective measures of performance:

- Average query response time
- Scan rates
- Time used per day query
- Memory usage per process
- I/O throughput rates

Following are the points to remember.

- It is necessary to specify the measures in service level agreement (SLA).
- It is of no use trying to tune response time, if they are already better than those required.
- It is essential to have realistic expectations while making performance assessment.
- It is also essential that the users have feasible expectations.
- To hide the complexity of the system from the user, aggregations and views should be used.
- It is also possible that the user can write a query you had not tuned for.

Data Load Tuning

Data load is a critical part of overnight processing. Nothing else can run until data load is complete. This is the entry point into the system.

Note: If there is a delay in transferring the data, or in arrival of data then the entire system is affected badly. Therefore it is very important to tune the data load first.

There are various approaches of tuning data load that are discussed below:

- The very common approach is to insert data using the**SQL Layer**. In this approach, normal checks and constraints need to be performed. When the data is inserted into the table, the code will run to check for enough space to insert the data. If sufficient space is not available, then more space may have to be allocated to these tables. These checks take time to perform and are costly to CPU.
- The second approach is to bypass all these checks and constraints and place the data directly into the preformatted blocks. These blocks are later written to the database. It is faster than the first approach, but it can work only with whole blocks of data. This can lead to some space wastage.

- The third approach is that while loading the data into the table that already contains the table, we can maintain indexes.
- The fourth approach says that to load the data in tables that already contain data, **drop the indexes & recreate them** when the data load is complete. The choice between the third and the fourth approach depends on how much data is already loaded and how many indexes need to be rebuilt.

Integrity Checks

Integrity checking highly affects the performance of the load. Following are the points to remember.

- Integrity checks need to be limited because they require heavy processing power.
- Integrity checks should be applied on the source system to avoid performance degrade of data load.

Tuning Queries

We have two kinds of queries in data warehouse:

- Fixed queries
- Ad hoc queries

Fixed Queries

Fixed queries are well defined. Following are the examples of fixed queries:

- regular reports
- Canned queries
- Common aggregations

Tuning the fixed queries in a data warehouse is same as in a relational database system. The only difference is that the amount of data to be queried may be different. It is good to store the most successful execution plan while testing fixed queries. Storing these executing plan will allow us to spot changing data size and data skew, as it will cause the execution plan to change.

Note: We cannot do more on fact table but while dealing with dimension tables or the aggregations, the usual collection of SQL tweaking, storage mechanism, and access methods can be used to tune these queries.

Ad hoc Queries

To understand ad hoc queries, it is important to know the ad hoc users of the data warehouse. For each user or group of users, you need to know the following:

- The number of users in the group
- Whether they use ad hoc queries at regular intervals of time
- Whether they use ad hoc queries frequently
- Whether they use ad hoc queries occasionally at unknown intervals.
- The maximum size of query they tend to run
- The average size of query they tend to run
- Whether they require drill-down access to the base data
- The elapsed login time per day
- The peak time of daily usage
- The number of queries they run per peak hour

Points to Note

- It is important to track the user's profiles and identify the queries that are run on a regular basis.
- It is also important that the tuning performed does not affect the performance.
- Identify similar and ad hoc queries that are frequently run.
- If these queries are identified, then the database will change and new indexes can be added for those queries.
- If these queries are identified, then new aggregations can be created specifically for those queries that would result in their efficient execution.

Data Warehousing - 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:

- Unit testing
- Integration testing
- System testing

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.

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.

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.

Test Schedule

First of all, the test schedule is created in the process of developing the test plan. In this schedule, we predict the estimated time required for the testing of the entire data warehouse system.

There are different methodologies available to create a test schedule, but none of them are perfect because the data warehouse is very complex and large. Also the data

warehouse system is evolving in nature. One may face the following issues while creating a test schedule:

- A simple problem may have a large size of query that can take a day or more to complete, i.e., the query does not complete in a desired time scale.
- There may be hardware failures such as losing a disk or human errors such as accidentally deleting a table or overwriting a large table.

Note: Due to the above-mentioned difficulties, it is recommended to always double the amount of time you would normally allow for testing.

Testing Backup Recovery

Testing the backup recovery strategy is extremely important. Here is the list of scenarios for which this testing is needed:

- Media failure
- Loss or damage of table space or data file
- Loss or damage of redo log file
- Loss or damage of control file
- Instance failure
- Loss or damage of archive file
- Loss or damage of table
- Failure during data failure

Testing Operational Environment

There are a number of aspects that need to be tested. These aspects are listed below.

- **Security** - A separate security document is required for security testing. This document contains a list of disallowed operations and devising tests for each.
- **Scheduler** - Scheduling software is required to control the daily operations of a data warehouse. It needs to be tested during system testing. The scheduling software requires an interface with the data warehouse, which will need the scheduler to control overnight processing and the management of aggregations.

- **Disk Configuration.** - Disk configuration also needs to be tested to identify I/O bottlenecks. The test should be performed with multiple times with different settings.

- **Management Tools.** - It is required to test all the management tools during system testing. Here is the list of tools that need to be tested.

    - Event manager
    - System manager
    - Database manager
    - Configuration manager
    - Backup recovery manager

Testing the Database

The database is tested in the following three ways:

- **Testing the database manager and monitoring tools** - To test the database manager and the monitoring tools, they should be used in the creation, running, and management of test database.

- **Testing database features** - Here is the list of features that we have to test:

    - Querying in parallel
    - Create index in parallel
    - Data load in parallel

- **Testing database performance** - Query execution plays a very important role in data warehouse performance measures. There are sets of fixed queries that need to be run regularly and they should be tested. To test ad hoc queries, one should go through the user requirement document and understand the business completely. Take time to test the most awkward queries that the business is likely to ask against different index and aggregation strategies.

Testing the Application

- All the managers should be integrated correctly and work in order to ensure that the end-to-end load, index, aggregate and queries work as per the expectations.

- Each function of each manager should work correctly
- It is also necessary to test the application over a period of time.
- Week end and month-end tasks should also be tested.

Logistic of the Test

The aim of system test is to test all of the following areas.

- Scheduling software
- Day-to-day operational procedures
- Backup recovery strategy
- Management and scheduling tools
- Overnight processing
- Query performance

Note: The most important point is to test the scalability. Failure to do so will leave us a system design that does not work when the system grows.

Data Warehousing - 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 sizes of the databases grow, the estimate 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. For example, a Telco call record requires 10TB of data to be kept online, which is just a size of one month's record. If it requires keeping records of sales, marketing customer, employees, etc., then the size will be more than 100 TB.
- The record contains textual information and some multimedia data. Multimedia data cannot be easily manipulated as text data. Searching the multimedia data is

not an easy task, whereas textual information can be retrieved by the relational software available today.

- Apart from size planning, it is complex to build and run data warehouse systems that are ever increasing in size. As the number of users increases, the size of the data warehouse also increases. These users will also require accessing the system.

- With the growth of the Internet, there is a requirement of users to access data online.

Hence the future shape of data warehouse will be very different from what is being created today.

ROLE OF DATA WAREHOUSING & DATA MINING IN E-GOVERNANCE

E-governance is the application of information & communication technologies to transform the efficiency, effectiveness, transparency and accountability of informational & transactional exchanges with in government, between govt. & govt. agencies of National, State, Municipal & Local levels, citizen & businesses, and to empower citizens through access & use of information. Governments deal with large amount of data. To ensure that such data is put to an effective use in facilitating decision-making, a data warehouse is constructed over the historical data. It permits several types of queries requiring complex analysis on data to be addressed by decision-makers.
Here we will deals with scope and use of data warehousing & Data mining in all the dimensions of e-governance like **Government to Citizen (G2C) Citizen to Government (C2G) Government to government (G2G) Government to Business, Government to NGO (G2N).** There are many methodology used to increase the efficiency of E-governance .Three complimentary trends are **Data warehousing, OLAP , Data mining.** By using these technique we find that data warehousing is very helpful in analyzing Current & Historical data finding useful pattern & support decision strategies .OLAP is useful in solving complex queries & views , interactive online analysis of data .Using Data mining technique & algorithm, automatic discovery of pattern & other interesting trends are find out

INTRODUCTION

E-governance involves the application of Information and Communication Technologies by government agencies for information and service delivery to citizens, business and government employees. It is an emerging field, faced with various implementation problems related to technology, employees, flexibility and change related issues, to mention a few. Global shifts towards increased deployment of IT infrastructure

by governments emerged with the advent of the World Wide Web. With the increase in Internet and mobile connections, the citizens are learning to exploit their new mode of access in wide ranging ways. They have started expecting more and more information and services online from governments and corporate organizations to enhance their civic, professional and personal lives .The concept of e-governance come into existence in India during the seventies with a focus on development of government applications in the areas of defense, economic monitoring, planning and the inclusion of Information Technology to manage data intensive functions related to elections, census, tax administration etc. The role & efforts made by National Informatics Center (NIC) to connect all the district headquarters during the eighties was a very new innovative approach.

From the early nineties, IT technologies were supplemented by ICT technologies to extend its use for wider applications with policy implementation & emphasis on reaching out to rural areas and taking in greater inputs from NGOs and private sector. There has been an increase involvement of international agencies under the framework of e-governance for development to catalyze the development of e-governance laws and technologies in developing countries. For governments, the motivation to shift from manual processes to IT-enabled processes may be increased efficiency in administration and service delivery, but this shift can be conceived as a worthwhile investment with potential for returns.

E-governance is the process of service delivery and information dissemination to citizens using electronic means providing the following benefits over the conventional system

- Increased efficiency in various Governmental processes
- Transparency and anticorruption in all transactions
- Empowerment of citizens and encouragement of their participation in governance.

The main objective of E-Governance is to change organization into e-organization. An e-organization needs to focus on the following things:-

- ✓ develop customer orientation
- ✓ manage customer relationships
- ✓ streamline business processes
- ✓ communicate better
- ✓ organize information
- ✓ work more flexibly
- ✓ make better decisions.
- ✓ coordinate activities better

NEED FOR E-GOVERNANCE

Pre conditions of E-Governance

Some of the pre conditions for an effective e-governance that could be listed as

- Formulation of new set of cyber laws to replace traditional set of rule & regulations for effective replacement of e-governance
- Simplification of procedure, rationalization of various administrative processes restructuring of government and mindset of bureaucrats to adopt according to e-governance.
- De-layering or re-layering of decision-making of levels
- Security and privacy are the two major concerns

Factor necessary for successful e-governance

- Political commitment
- Effective administrative leadership
- Effective handling of HR issues
- Involvement of staff at design stage
- Innovative funding strategy and revenue model
- Appropriate administrative structure
- Common infrastructure and database creation
- Training & Motivation

IMPLEMENTATION ISSUE IN E-GOVERNANCE

The government of India, like all over the world, has began investing large amounts in Information and Communication Technology(ICT).The object behind these investment is to improve the efficiency of government function by, especially enabling citizen centric services. There are some technical issue which need to be discussed apart from above mentioned issue. The Above mentioned issue can be resolved by the government but as far as technical issue are concerned they need more focus to resolve the issue.
Some of technical issue related to e-governance are

- ✓ Technical Infrastructure support by the government
- ✓ Collection of Large amount of data
- ✓ Analysis of the data So that accurate Decision can be made
- ✓ Online Support to all department of Government organization
- ✓ Retrieval of meaningful Data
- ✓ Presentation of meaningful data so fast decision can be made

E-governance ,meaning the electronic-governance ,has evolved as an information age model of governance that seeks to realize process and structure for harshening the potentialities of information & communication technologies at various level of government and public sector .E-governance is the commitment to utilize appropriate technologies to enhance governmental relationships in order to encourage the fair & efficient delivery of services .The ICT model uses the new technologies to maintain the data in government organization .Some of these are discussed in this paper which is very popular technologies now a days .

Increasingly, government organization, are analyzing current and historic data to identify useful patterns from the large database so that they can support their business strategy Their main emphasis is on complex, interactive, exploratory analysis of very large dataset created by the integration of data from across all the part of the organization and that data is fairly static Three complementary trends are their

1) Data warehouse
2) OLAP
3) Data Mining

ROLE OF DATA WARE HOUSE IN E-GOVERNANCE

Need for data warehouse

Governments deal with enormous amount of data. In order that such data is put to an effective use in facilitating decision-making, a data warehouse is constructed over the historical data. It permits several types of queries requiring complex analysis on data to be addressed by decision-makers.

When used properly, it can help planners and decision makers in making informed decisions leading to positive impact on targeted group of citizens. However to use information to it's fullest potential, the planners and decision makers need instant access to relevant data in a properly summarized form. In spite of taking lots of initiative for computerization, the Government decision makers are currently having difficulty in obtaining meaningful information in a timely manner because they have to request and depend on IT staff for making special reports which often takes long time to generate. An Information Warehouse can deliver strategic intelligence to the decision makers and provide an insight into the overall situation. This greatly facilitates decision-makers in taking micro level decisions in a timely manner without the need to depend on their IT staff. By organizing person and land-related data into a meaningful Information Warehouse, the Government decision makers can be empowered with a flexible tool that enables them to make informed policy decisions for citizen facilitation and accessing their impact over the intended section of the population.

Benefit of a data warehouse for e-governance

Citizen facilitation is the core objective of any Government body. For facilitating the citizens of a state or a country, it is important to have the right information about the people and the places of the concerned territory. Hence a data warehouse built for eGovernance can typically have data related to person and land. Such a data warehouse can be beneficial to both the Government decision makers and citizens as well in the following manner:

Benefit for the decision makers

- They do not have to deal with the heterogeneous and sporadic information generated by various state-level computerization projects as they can

access current data with a high granularity from the information warehouse.

- They can take micro-level decisions in a timely manner without the need to depend on their IT staff.
- They can obtain easily decipherable and comprehensive information without the need to use sophisticated tools.
- They can perform extensive analysis of stored data to provide answers to the exhaustive queries to the administrative cadre. This helps them to formulate more effective strategies and policies for citizen facilitation

Benefit for the citizens

- They are the ultimate beneficiaries of the new policies formulated by the decision makers and policy planner's extensive analysis on person and land-related data.
- They can view frequently asked queries whose results will already be there in the database and will be immediately shown to the user saving the time required for processing.
- They can have easy access to the Government policies of the state.
- The web access to Information Warehouse enables them to access the public domain data from anywhere.

The data warehouse has enough potential to access the impact of various welfare schemes across the population of the state. The planners can design schemes focused on specific target groups and achieve high impact. The decision-makers can carry out analysis of population profile across the state in areas of economy, education, family units, shelter, etc. The warehouse can also be used for rural and urban development planning, agricultural yield and cropping pattern analysis and much more. These analyses will help in making decisions that are focused and the benefit of the government policies can reach the intended group. The various types and number of queries that can be handled by the data warehouse are limited only by the intelligence of the person using the data warehouse and the data fed to it.

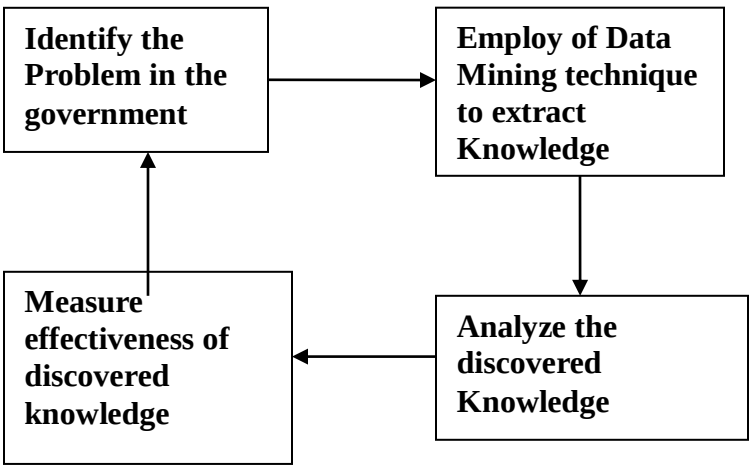
ROLE OF DATA MINING IN E-GOVERNANCE

It is well known that in Information Technology (IT) driven society, knowledge is one of the most significant assets of any organization. The role of IT in E-governance is well established. Knowledge Pragmatic use of Database systems, Data Warehousing and Knowledge Management technologies can contribute a lot to decision support systems in E-governance. Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as :"Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data". Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the E-governance administrators to improve

the quality of service. Traditionally, decision making in E-governance is based on the ground information, lessons learnt in the past resources and funds constraints. However, data mining techniques and knowledge management technology can be applied to create knowledge rich environment. An organization may implement Knowledge Discovery in databases (KDD) with the help of a skilled employee who has good understanding of organization. KDD can be effective at working with large volume of data to determine meaningful pattern and to develop strategic solutions. Analyst and policy makers can learn lessons from the use of KDD in other industries E-governance data is massive. It includes centric data, resource management data and transformed data. E-governance organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to organization .

Knowledge Discovery in E-governance

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven. Data mining is a young interdisciplinary field closely connected to data warehousing, statistics, machine learning, neural networks and inductive logic programming. Data mining provides automatic pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Without data mining it is difficult to realize the full potential of data collected within healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain.



Data Mining Cycle

For Government organization to succeed they must have the ability to capture, store and analyze data Online analytical processing (OLAP) provides one way for data to be analyzed in a multi-dimensional capacity. With the adoption of data warehousing and data analysis/OLAP tools, an organization can make strides in leveraging data for better decision making. Many organizations struggle with the utilization of data collected through an organization online transaction processing (OLTP) system that is not integrated for decision making and pattern analysis. For successful E-governance organization it is important to empower the management and staff with data warehousing based on critical thinking and knowledge management tools for strategic decision making. Data warehousing can be supported by decision support tools such as data mart, OLAP and data mining tools. A data mart is a subset of data warehouse. It focuses on selected subjects. Online analytical processing (OLAP) solution provides a multi-dimensional view of the data found in relational databases. With stored data in two dimensional format OLAP makes it possible to analyze potentially large amount of data with very fast response times and provides the ability for users to go through the data and drill down or roll up through various dimensions as defined by the data structure. The traditional manual data analysis has become insufficient and methods for efficient computer assisted analysis indispensable. A Data Warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. A data warehouse is also often viewed as architecture constructed by integrating data from multiple heterogeneous sources to support structured and/or ad-hoc queries, analytical reporting and decision making

Data mining technique in E-governance

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and measured before they are applied to data mining

CLASSIFICATION OF DATA MINING TECHNIQUE

Some of the data mining technique we discuss as they are very useful in e-Governance organization

Rule Induction

Rule induction: is the process of extracting useful ‘if then’ rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form IF conditions THEN conclusion This kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction.
Rule Induction Method has the potential to use retrieved cases for predictions.

Decision tree

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree based models which include classification and regression trees, are the common implementation of induction modeling. Decision tree models are best suited for data mining. They are in expensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5

POSSIBLE QUESTIONS

PART – B

8 MARKS

1. Explain Tuning the Data warehouse in detail.

2. Describe other areas for Data warehousing and Data mining.

3. Explain the concept of backup and recovery in detail.

4. Write about Disk technology and database layout.

5. Describe capacity planning and estimating the load in detail.

6. Explain national data warehouse in detail.

7. Explain a) Developing the test plan.

   b) Testing backup recovery

   c) Testing the operational environment

8. Discuss about data warehouse futures.

9. Explain a) testing the database

   b) Testing the application

   c) Logistics of the test

10. Describe Capacity planning in detail.

11. Describe Tuning the data warehouse in detail with example

12. Explain National Data warehouse in detail.

KARPAGAM

Enable | Enlighten | Enrich
(Deemed to be University)
(Under Section 3 of UGC Act 1956)

DEPAR

PAR

ONLINE EXAM

S.no	Questions
1	_____ is finding information from the database
2	The _____ is usually a subset of the database in SQL
3	The purpose of the algorithm is fit a model to the data is _____
4	Which model makes a prediction about values of data using known results found from different data _____
5	which modelling may be made based on the use of other historical data _____
6	_____ model identifies pattern or relationship in data.
7	_____ maps data into predefined groups or classes
8	____ is a type of clssification where an input pattern is classified into one of several classes.
9	_____model serves as a way to explore the preoperties of tye data examined not to predict new properties.
10	___ is used to map a data item to a real valued predictive variable
11	_____ is used to determine which function is best.
12	_____ the value of an attribute is examined as it varies over time.
13	_____ are used to determine the similarity between different time series
14	How many functions performed in time analysis.?
15	_____ can be viewed as a type of classification.
16	Which task is used to predicting a future state rather than the current state.?
17	Which application include flooding, speech recognition, and pattern recognition?
18	Future values may be predicted using _____
19	_____ is similar to classification except that the groups are not predefined
20	_____ is alternatively referred to as unsupervised learning or segmentation
21	The most similar data are gropued into ____
22	A special type of clustering is called ____
23	A database is partitioned into disjointed groupings of similar tuples called ____

24	_____ maps data into subsets with associated simple descriptions.
25	Summarization is also called _____
26	_____ alternatively referred to as affinity analysis
27	_____ is a model that identifes specific type of data associations.
28	____ is used to determine sequential patterns in data.
29	Which is used in predicting the failure of telecommunication switvches
30	KDD stands for
31	Which is the process of finding useful information and patterns in data?
32	____ is the use of algorithm to extract the information and patterns derives by the KDD process
33	KDD process consists of ___ steps
34	____ is the first step obtains the data from various databases
35	supplies or predicted in _____
36	processing is _____
37	provide more meaningful results
38	_____ refers to the visual presentation of data
39	IR stands for _____
40	Which describe the relationship between input and output through the use of algeberic equation _____ models
41	It refers to the process of estimating a population parameter_____
42	estimator and the actual value
43	MSE stands for _____
44	A popular estimating technique technique is the _____
45	Which can be defined as a value proportional to the actual probability_____
46	Which algorithm is an approach that solves the estimation problem with incomplete data _____
47	A ____ is a more sophisticated technique that illustrates several features of the population at once.
48	Another visual technique to display data is called ____
49	__ can be used to evaluate the strength of a relationship between two variables
50	_____ is generally used to predict future values based on past values by fitting a set of points to acurve
51	If there is more than one predictor it is called as an _____

52	___ is used to measure the overlap of two sets as related to the whole set caused by their union.
53	_____ use a divide and conquer technique to split the problem search space into subsets.
54	An activation function is sometimes called a _____ function
55	Linear threshold function also called a _____
56	One technique to perform hypothesis testing is based on the use of the _____
57	The total range of the data values is divided into four equal parts called _____
58	RMS stands for _____
59	Determine a range of values within which the true parameter value should fall is_____
60	A definition of a concept is-----if it recognizes all the instances of that concept

M ACADEMY OF HIGHER EDUCATION

Coimbatore – 641 021.

TMENT OF COMPUTER SCIENCE, CA & IT

III B.Sc(CS) BATCH 2015-2018

T - A OBJECTIVE TYPE / MULTIPLE CHOICE QUESTIONS

INATION **ONE MARK QUESTIONS**

UNIT - I

Choice 1	Choice 2	Choice 3	Choice 4
Datawarehouse	Datamining	Pattern recognition	Classification
Input	Query	Output	Process
Model	Preference	Serach	Output
Descriptive	Predictive	Association	Sequence
Association	Descriptive	Predictive	Sequence
Predictive	Descriptive	reference	Clustering
Classification	Clustering	regression	Summarization
regression	Summarization	Pattern recognition	Clustering
Association	Descriptive	Predictive	Sequence
Clustering	regression	Time series analysis	Summarixzation
Error analysis	Pattern recognition	Time series analysis	Regression
Time series analysis	Prediction	Regression	Clustering
regression	clustering	Pattern recognition	Distance measures
Three	Two	Five	Four
Clustering	Prediction	Sequence discovery	Association rules
Regression	Clustering	Prediction	Association
Clustering	Prediction	Sequence discovery	Association rules
Time series analysis	Prediction	Clustering	Descriptive
Clustering	Prediction	Sequence discovery	Association rules
Regression	Clustering	Prediction	Association
Input	Query	Output	Clusters
Segmentation	Characterization	simulation	Sequence
Model	Preference	segments	Process

Clustering	regression	Time series analysis	Summarization
Model	characterization	Sequence discovery	Association rules
Clustering	Prediction	Sequence discovery	Link analysis
Clustering	Prediction	Sequence discovery	Association rules
Association rules	Sequential analysis	Clustering	Link analysis
Model	characterization	Sequence discovery	Association
in database	in Database	description	Discovery
KDD	MDD	DDK	DKD
Datawarehouse	Datamining	Pattern recognition	Classification
Five	Four	Three	Ten
Selection	Preprocessing	Tranformation	Interpretation
Tranformation	Selection	Preprocessing	Interpretation
Selection	Interpretation	Sequence	Tranformation
Transformation	Selection	Preprocessing	Interpretation
Visualization	Tranformation	Selection	Interpretation
Instant Retrieval	Information Redundant	Information Retrieval	Information Replace
Predictive	Descriptive	parametric	Segment
Point estimation	Co-relation	Baytes theorem	Hypotheis testing
KDD	MDD	BIAS	BISA
Mean squared Error	Main sequence error	Estimation	Main squared Arror
Jackknife estimate	Likelihood	regression	Visualization
regression	Visualization	Likelihood	Jackknife estimate
expectation-maximization	expectation-minimization	maximization	minimization
Quartiles	box-plot	IQR	outlier
box-plot	scatter diagram	outlier	BIAS
Co-relation	regression	Jackknife estimate	Descriptive
regression	Summarization	Pattern recognition	Clustering
single linear regression	multiple linear regression	Visualization	Co-relation

outlier	IQR	Jaccard's co-efficient	Decision tree
Jackknife estimate	Co-relation	Decision tree	Neural network
processing element	processing event	Pattern recognition	Tranformation
ramp function	Tranformation	Preprocessing	Pattern recognition
Jackknife estimate	ramp function	chi-squared statistic	Jaccard's co-efficient
Quartiles	box-plot	IQR	outlier
Root mean square	Random mean square	Rapid mean square	Real medium square
Confidence interval	Quartiles	box-plot	IQR
Complete	Consistent	Constant	Quartiles

Answer
Datamining
Output
Model
Predictive
Predictive
Descriptive
Classification
Pattern recognition
Descriptive
regressiohn
Error analysis
Time series analysis
Distance measures
Three
Prediction
Prediction
Prediction
Time series analysis
Clustering
Clustering
Clusters
Segmentation
segments

Summarization
characterization
Link analysis
Association rules
Sequential analysis
Association
KDD
Datamining
Five
Selection
Preprocessing
Tranformation
Tranformation
Visualization
Information Retrieval
parametric
Point estimation
BIAS
Mean squared Error
Jackknife estimate
Likelihood
expectation-maximization
box-plot
scatter diagram
Co-relation
regression
multiple linear regression

Jaccard's co-efficient
Decision tree
processing element
ramp function
chi-squared statistic
Quartiles
Root mean square
Confidence interval
Complete

K

Enable | Enlighten | Enrich
(Deemed to be University)
(Under Section 3 of UGC Act 1956)

ONLINE EX

S.No	Questions
1	_ is mining of data related the world wide web
2	_ structure is the actual linkage structure between web pages
3	_ examines the content of webpages as well as result of web searching
4	With _ information is obtained from the actual organization of pages on the web
5	_ may be applied to web pages to identify similar pages
6	_ is any technique that is used to direct business marketing or advertising to the most beneficial subset of the total population
7	_____ goes beyond the basic IR technology
8	Web mining divided web content mining into ___
9	Agent based approaches have software systems that performs the_____
10	_____ use information about user preferences to direct their search
11	The database approaches view the ___ as belonging to a database
12	Basic content mining is a type of _____
13	_ retrieve relevant documents usually using a keyword base retrieval technique
14	Web pages are defined using ___
15	Webpages created using HTML are only _____
16	Example for web content mining is in the area of _____
17	_____ technique predcts future needs based on past needs
18	_____ uses the interestingness of a document to determine if a user is interested in it
19	Another approach to automatic personalization is that used by _____
20	_ is based on the concept that humans often base decisions on what they hear from others
21	Another collaborative approach is called _____
22	____ are used to show the relationship between data items
23	_____ is the most well known association rule algorithm and is used in commercial products
24	The large itemsets are also said to be ____

25	An algorithm called _____ is used to generate the candiadate itemsets for each pass after the first
26	PL stands for _
27	The ____ is an generalization of the apriori gen algorithm
28	ML stands for ____
29	_____ improve the performance of finding large itemsets
30	_ would be easy to parallize using the task parallelism appraoch
31	_ algorithm have reduced communication cost over the task
32	One data parallelism algorithm is the ___
33	CDA stands for ___
34	The _____ demonstrates tak parallelism
35	DDA stands for _____
36	The most common strategy to generate association rules is that of _____
37	_____ provide an effective technique to store, access amd count itemsets
38	A _____ is a multiway search tree
39	OC stands for _____
40	._____ curve is used in information retrieval to examine fall act versus recall.
41	ROC stands for _____
42	A major issues associated with classification is that of _____
43	_____ is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached
44	The most familia data mining technique is _____
45	Which of the following is a clustering algorithm?
46	Strategic value of data mining is _____.
47	The power of self-learning system lies in _____.
48	Incorrect or invalid data is known as _____.
49	A priori algorithm is otherwise called as _____.
50	The A Priori algorithm is a _____.
51	The first phase of A Priori algorithm is _____.
52	The second phaase of A Priori algorithm is _____.
53	Box plot and scatter diagram techniques are _____
54	The left hand side of an association rule is called _____.
55	The right hand side of an association rule is called _____.
56	The a priori frequent itemset discovery algorithm moves _____ in the lattice.
57	After the pruning of a priori algorithm, _____ will remain.
58	A definition or a concept is---------- if it classifies any examples as coming within the concept
59	A goal of data mining includes which of the following?
60	Which is the technique used for classification in data mining?

ARPAGAM ACADEMY OF HIGHER EDUCATION

Coimbatore – 641 021.

DEPARTMENT OF COMPUTER SCIENCE, CA & IT

III B.Sc(CS) BATCH 2015-2018

PART - A OBJECTIVE TYPE / MULTIPLE CHOICE QUESTIONS

XAMINATION **ONE MARK QUESTIONS**

UNIT 2

Choice 1	Choice 2	Choice 3
Datamining	Web content mining	Webmining
Middle page	Interpage	Data
Web content mining	Content mining	Web content mining
Web structure mining	Content mining	Web content mining
Clustering	Classification	Prediction
Targetting	Automated personalization	Personalized web agents
Web mining	Content mining	Web content mining
Agnet based	Personalized web agents	News dude
Web mining	Content mining	Web content mining
Personalized web agents	News dude	Firefly
Webdata	content of text	images only
Text mining	Web mining	Web content mining
Search engine	Monitor	Chip
HTML	C	Java
Semi structured	Full structured	unstructured
Characterization	Summarization	Personalization
Automated personalization		Data Distribution Algorithm
Simple Approach	News dude	Regression
Firefly	K-Means	Outliers
Firefly	K-Means	Outliers
Web watcher	Data parellisim	Partitioning
Hash tree	Association rules	Hypothesis
K nearest	Apriori algorithm	Data Distribution Algorithm
downward closed	openward closed	upward closed

K nearest	Apriori algorithm	Data Distribution Algorithm
Potentially Large	Positive Large	Parallelism Large
Activation function	Negative border function	task parallelism
Middle Large itemsets	Maximum large itemsets	Missing Large Itemsets
Partitioning	Apriori algorithm	Data Distribution Algorithm
Partition algorithm	DT technique	Distance based
Data parellisim	Partition algorithm	Apriori algorithm
Count Distribution Algorithm	Combining Distribution Alg	Character Distribution Algori
Character Data Algorithm	Count Distribution Algorithr	Combining Distribution Algo
Data Derived Algorithm	Data Distribution Algorithm	Disk Distribution Algorithm
Data Derived Algorithm	Data Distribution Algorithm	Disk Distribution Algorithm
Finding small items	Finding similar items	Finding large itemsets
Hypothesis	DT technique	Combining technique
Hash tree	Decision Tree	Hypothesis
Over Customized	Operating Characteristic	Open Class
OC curve	ROC curve	OCR curve
Receiver Operating Characteri	Read Operating Characterist	Real Operating Characteristic
Time Consumption	Security	Overfitting
K-Means	Squared Error	Partitional algorithm
Statistical	clustering	Estimation
A priori.	CLARA.	Pincer-Search.
cost-sensitive.	work-sensitive.	time-sensitive.
cost.	speed.	accuracy.
changing data.	noisy data.	outliers.
width-wise algorithm.	level-wise algorithm.	pincer-search algorithm.
top-down search	breadth first search.	depth first search.
Candidate generation.	Itemset generation.	Pruning.
Candidate generation.	Itemset generation.	Pruning.
Graphical.	Geometric.	Icon-based.
consequent.	onset.	antecedent.
consequent.	onset.	antecedent.
upward.	downward.	breadthwise.
Only candidate set.	No candidate set.	Only border set.
Complete	Consistent	Constant
To explain some observed eve	To confirm that data exists	To analyze data for expected
Descriptive pattern	Associations	Decision tree classifiers

S

Choice 4	Answer
Spatial mining	Webmining
Personalization	Interpage
Personalization	Web content mining
Personalization	Web structure mining
regression	Clustering
Data Distribution Algorithm	Targetting
Personalization	Web content mining
Firefly	Agnet based
Personalization	Content mining
Hash tree	Personalized web agents
None of the above	Webdata
Data mining	Text mining
Browser	Search engine
C++	HTML
None of the above	Semi structured
Content mining	Personalization
Hash tree	Automated personalization
Multiple level association	News dude
Web watcher	Firefly
Web watcher	Firefly
Apriori algorithm	Web watcher
Neural	Association rules
Character Distribution Algorithm	Apriori algorithm
backward closed	downward closed

Character Distribution Algorithm	Apriori algorithm
Positive Learning	Potentially Large
Assemble function	Negative border function
Minimum large itemsets	Missing Large Itemsets
Hash tree	Partitioning
Data parellisim	Partition algorithm
K nearest	Data parellisim
Character Data Algorithm	Count Distribution Algorithm
Character Distribution Algorithm	Count Distribution Algorithm
Data Distribution Activation	Data Distribution Algorithm
Data Distribution Activation	Data Distribution Algorithm
Distance based	Finding large itemsets
Hash tree	Hash tree
Neural	Hash tree
Overfitting Characteristic	Operating Characteristic
CO curve	OC curve
Random Operating Characteristic	Receiver Operating Characteristic
Human Intervention	Overfitting
Divisive	K-Means
Distance Measure	clustering
FP-growth.	CLARA.
technical-sensitive.	time-sensitive.
simplicity.	accuracy.
missing data.	noisy data.
FP growth algorithm.	level-wise algorithm.
bottom-up search.	bottom-up search.
Partitioning.	Candidate generation.
Partitioning.	Pruning.
Pixel-based.	Geometric.
precedent.	antecedent.
precedent.	consequent.
both upward and downward.	upward.
No border set.	No candidate set.
Geometric.	Consistent
To create a new data warehouse	To explain some observed event or condition
Regression	Decision tree classifiers

KARPAGAM ACA

DEPARTMEN

III

PART - A C

ONLINE EXAMINATION

S.No	Questions
1	_____ process is used to remove noise and inconsistent data
2	Which process multiple data sources may be combined
3	_____ process is used to retrieve the relevant data from the database
4	_____ process is used to identify the truly interesting patterns
5	A set of _____ that describe the objects. These corresponds to attributes in
6	A set of _____ helps the objects communicate with other objects
7	_____ is a repository for long term storage of data from multiple sources,
8	The _____ of the set is the difference between the largest and smallest values
9	_____ routines attempt to fill in missing values, smooth out noise
10	_____ techniques such as data cube aggregation attribute subset
11	_____ function is used to extract information from other data sources
12	Data mart is _____
13	Data warehouse is _____
14	_____ in a data warehouse is similar to the data dictionary
15	The _____ schema is used for data design is a relational model consisting of
16	Analytical capabilities of Data Warehouse system is _____
17	Data for a single session in OLTP system is _____
18	The _____ schema has three business dimensions, namely product ,
19	AGNES stands for
20	OLAP stands for
21	In which process pollution is detected
22	In which process data are collected from the operation database
23	_____ is a creative process
24	Data warehouse must be architected to support three major driving factors
25	_____ takes data from source systems and makes it available to the data
26	clean and _____ the loaded data into a structure that speeds up
27	The _____ process is the system process that manages the queries and
28	The _____ is the system component that performs all the operations

29	In bottom up approach _____ are used by end-users
30	__ are built to support large data volume cost effective
31	The size and complexity of datawarehouse systems make them very different
32	The architecture of datawarehouse is defined within the __ stage of thye
33	_ stage shouls have identified the initial user requirement and have developed
34	_____ evolution tends to be the most complex aspect of a datwarehouse
35	_____ takes data from source system and makes it available to the data
36	_____ takes extracted data and loads it into the data warehouse
37	data should be in a _____ state when it is estimated from the source
38	The information in a datawarehouse represents a _ of corporate information, so
39	Once the data is extracted from the source systems it is then typically loaded
40	_____ process will take a reasonable amount of alapsed time to
41	_ is the system process that takes the loaded data structures it for query
42	_____ is the system process that manages the queries and speeds them
43	_____ manager extracts and load the data
44	_ manager transform and manage the data
45	_____ manager backup and archieves the datawarehouse
46	_____ manager directs and manages queries
47	_____ is the system component that performs all the operation necessary
48	_____ can be collected by the query manager as it intercepts any query hitting
49	_____ is the system component that performs all the operations necessary to
50	Starflake schemas are _____ that structure the data to exploit a typical
51	The central factual transaction table is called the _____
52	The surrounding reference tables are called _____
53	_____ is an essential component of decision support datawarehouse
54	What is ETL Stand for?
55	A data warehouse is which of the following?
56	Fact tables are which of the following?
57	A data warehouse is which of the following?
58	snowflake schema is which of the following types of tables?
59	Fact tables are which of the following?
60	_____ is the heart of the warehouse.

ADEMY OF HIGHER EDUCATION

Coimbatore – 641 021.

T OF COMPUTER SCIENCE, CA & IT

B.Sc(CS) BATCH 2015-2018

)BJECTIVE TYPE / MULTIPLE CHOICE QUESTIONS

ONE MARK QUESTIONS

UNIT - III

Choice 1	Choice 2	Choice 3	Choice 4
data cleaning	data	data selection	data transformation
data cleaning	data	data selection	data transformation
data cleaning	data	data selection	data transformation
data cleaning	data	data selection	pattern evaluation
Messages	methods	variables	None of the above
Messages	methods	variables	None of the above
database	data	data mining	None of the above
range	median	mode	None of the above
data cleaning	data	data selection	data transformation
data integration	data selection	data	data reduction
Data Extraction	Data Transformatio	Data loading	None
departmental	corporate	organized	none
departmental	corporate	organized	none
internal	production	metadata	external
STAR	Cube	Dimension	none
low	moderate	very low	small
Large	moderate	very limited	small
star	cube	diamond	none
AGlomerative	Against	All Network	None
Online arithmetic	Online	Operation	none
data cleaning	data selection	data	data coding
data cleaning	data selection	data	data coding
data cleaning	data selection	data	data coding
3	4	2	5
data cleaning	data	data extraction	none
transform	extract	partition	aggregate
process	load	system	query management
process manager	load manager	system	query manager

data mart	large data	central	operational database
Datawarehouse	central	system	process management
OLTP	OLAP	SMP	OLPT
loading	selection	technical blue	process
business	technical blue	process	loading
requirement	analysis	process	evaluate
Data Extraction	data load	snapshot	evaluate
process	requirement	data load	business requirement
selection	consistent	process	data mart
data load	snapshot	data mart	Data Extraction
temporary	permanent	volatile	evaluate
selection	load	evaluate	transform
clean and	Data	requirement	system management
Query	system	load	process management
Load	temporary	permanent	Data Extraction
Query	Load	Warehouse	Load manager
Data Extraction	Warehouse	Query	Load
temporary	Load	Data	Query
Load manager	Data	Load	temporary
Query	Data aggregatic	SMP	Query statistics
SMP	Query Manage	central	Data Extraction
Detailed informati	Query	Query	process management
fact table	Query Manage	fact design	fact tuple
fact tuple	fact table	dimension table	fact design
Data aggregation	system	Data	load management
Execute tramit and	Extract transfo	Excute Transfo	Evaluate Transform and
Can be updated by	Contains nume	Contains only c	Contains only current da
Completely denor	Partially denor	Completely nor	Partially normalized
Can be updated by	Contains nume	Organized arou	Contains only current da
Fact	Dimension	Helper	All of the above
Completely denor	Partially denor	Completely nor	Partially normalized
Data mining datab	Data warehous	Data mart datab	Relational data base serv

Answer
data cleaning
data integration
data selection
pattern evaluation
variables
Messages
data warehouse
range
data cleaning
data reduction
Data Extraction
departmental
corporate
metadata
STAR
moderate
very limited
star
AGlomerative NESting
Online Analytical Processing
data cleaning
data selection
data coding
3
data extraction
transform
query management
load manager

data mart
Datawarehouse
OLTP
technical blue print
business requirement
requirement
Data Extraction
data load
consistent
snapshot
temporary
load
clean and transform data
Query Management process
Load
Warehouse
Warehouse
Query
Load manager
Query statistics
Query Manager
Detailed information
fact table
dimension table
Data aggregation
Extract transform and load
Contains only current data.
Completely normalized
Organized around important subject areas.
All of the above
Completely normalized
Data warehouse database servers.

KARPAGAM ACA

DEPARTMENT

III B

PART - A O

ONLINE EXAMINATION

S.No	Questions	Choice 1
1	The hardware architecture of a datawarehouse is defined with in _____	technical blueprint
2	The backup and security strategies are also determined during the _____	technical blueprint
3	The ____ is a crucial part of the datawarehouse environment	Hardware
4	Warehouse application generally require ____ hardware configuration	Small
5	SMP stands for _____	Symmetric multi processing
6	MPP stands for _____	Massively Parallel Processing
7	_____ Machine is a set of tightly coupled CPU that share memory and disk	SMP
8	___ machine is a set of loosely couples CPU's each of which has its own memory	SMP
9	NUMA Stands for	Non uniform Memory Architecture
10	SMP machine are the natural progression of the traditional ____ cpu midrange servers	Single
11	_____ gives the CPU bandwidth needed to support the large adhoc queries of decision support systems	SMP
12	SMP machine is also called _____	Shared multi environment
13	____ is a set of loosely couples SMP machine connected by a high speed interconnect	Cluster
14	each machine has its own CPU and memory but they share access to disk. It is called _____	S\Distributed lock Manager
15	Each machine in the cluster is called ____	head
16	Most MPP systems allow a disk to be dual connected between two nodes	Five
17	An extra layer of software called __	Virtual system drive
18	VSD stands for ____	Virtual system drive
19	A ____ machione is basically a tightly coupled cluster of SMP nodes with an extremely high speed interconnect	SMP

20	____ are any requests that cause data to be transferred out over the network	Data Cleaning
21	_____ is black art	System Hardware
22	Heart of any computer is _____	Monitor
23	_____ is where a process requests for the data to be shipped to the location where the process is running	Datamart
24	____ is where the function to be performed is moved to locale of the data	Datamart
25	_____ are separate operations that occur within the conbines of the SQL statement	Intra Statement Operations
26	_____ is where operations are carried out sequentially on a given piece of data	Infra Structure
27	_____ uses stored statistics about the tables and their indexes to calculate the best strategy for executing the SQL statement	Infra Structure
28	___ is an extension of the parallel query functionality	Parallel load
29	RAID stands for _____	rapid Array of interactive Disks
30	____ technology is to provide resilence against disk failure	Pipelining
31	_____ is vital part of a daily operation	Query Management
32	The warehouse application managers are likely to have a no of service or background process running. This process is called	Query Management
33	_____ are usually avoided during the user day to avoid connections with the users queries	Query Management
34	_____- are subset of data that are offloaded from the server machine onto other machines	Data Cleaning
35	_____ is an area that needs to be slearly defined and documented	Problem management
36	_____ processing required will again vary from data warehouse to datawarehouse	data Integration
37	SLA stands for ___	Service level agreement
38	___ is a system that has responsibilities to the operations of the business	Operational system
39	_____ is a system that the business absolutely depends on to function	Operational system
40	_ is system that needs to be available all day everyday, except for small periods of planned downtime	Operational system
41	_____ is one of the most important regular operations carriedout on any system	Backup
42	A _____ is a backup that is taken while the database is completel shut down	Partial backup
43	tape Stackers are a method of loading multiple tapes in a ___ tape drive	Single

44	_____databases are owned by particular departments or business groups.	Informational.
45	The key used in operational environment may not have an element of_____.	time
46	Data can be updated in _____environment.	data warehouse.
47	Record cannot be updated in _____	OLTP
48	The source of all data warehouse data is the_____.	operational environment.
49	Data warehouse contains_____data that is never found in the operational environment.	normalized
51	_____ is a data transformation process.	Comparison
52	MDDB stands for _____.	multiple data doubling.
53	EIS stands for _____.	Extended interface system.
54	contains the measurement of business processes, and it contains foreign keys for the dimension tables.	Dimension table
55	The data Warehouse is_____.	read only.
56	Expansion for DSS in DW is_____.	Decision Support system.
57	The time horizon in Data warehouse is usually _____	1-2 years.
58	The data is stored, retrieved & updated in _____.	OLAP
59	_____ is the specialized data warehouse database.	Oracle
60	Which storage has the drawback that storing or fetching a single tuple requires multiple I/O operations?	Row oriented

ADEMY OF HIGHER EDUCATION

Coimbatore – 641 021.

' OF COMPUTER SCIENCE, CA & IT

B.Sc(CS) BATCH 2015-2018

BJECTIVE TYPE / MULTIPLE CHOICE QUESTIONS

ONE MARK QUESTIONS

UNIT IV

Choice 2	Choice 3	Choice 4
operational	automation	strategy
operational	automation	strategy
Server	Software	Architecture
Medium	Large	tiny
Symbol Multi Processing	Sign Multi Processing	Simultaneous Multi Process
mart Passive Processing	Multi Pre Processing	Machinery Process Port
MPP	SPM	MMS
MPP	SPM	MMS
Numeric Uniform Memory Architecture	Negative Uniform Memory Architecture	Null Uniform Memory Architecture
Multi	Two	three
MPP	SPM	MMS
Shared Everything environment	Shared single environment	Shared Similar environment
regression	prediction	Scalability
Shared disk systems	Shared data systems	Simultaneous data system
arc	node	leaf
Four	Two	Three
Virtual Suppoted Drive	Various shred disk	virtual shared disk
Virtual Suppoted Drive	Various shred disk	virtual shared disk
MPP	NUMA	NMUA

Data Integration	Data Extraction	data selection
Network Management	System Software	Hardware Management
Mouse	Speaker	CPU
Function Shipping	data shipping	Meta data
Function Shipping	data shipping	Meta data
Infra Statement Operations	Item Sequence Operation	item Sequence operation
pipelining	parallel Query	Cost based optimizer
pipelining	parallel Query	Cost based optimizer
parallel technique	parallel index build	parallel Query
redundant array of inexpensive disks	redundant array of intensive disks	redundant array of interactive disks
RAID	Cost optimizer	RAPID
Backups	datafeeds	Daemons
Backups	datafeeds	Daemons
Backups	datafeeds	Daemons
Data Integration	Data Extraction	data selection
Process management	hardware management	software management
data Cleanup	data transformation	None of the above
Service low argument	Several level agreement	Server level argument
Mission critical system	7 x 24 system	7 x 24 x 52 system
Mission critical system	7 x 24 system	7 x 24 x 52 system
Mission critical system	7 x 24 system	7 x 24 x 52 system
Security	Testing	None of the above
Cold backup	Hot backup	Online baclup
Double	Multi	Standalone

Operational.	Both informational and operational.	Flat
cost	frequency	quality
data mining	operational.	informational
files	RDBMS	data warehouse
informal environment.	formal environment.	technology environment.
informational	summary	denormalized
Projection	Selection	Filtering
multidimensional databases.	multiple double dimension.	multi-dimension doubling.
Executive interface system.	Executive information system.	Extendable information system.
Fact table	Multi table	data table
write only.	read write only.	none
Decision Single System.	Data Storable System.	Data Support System.
3-4years.	5-6 years.	5-10 years.
OLTP	SMTP	FTP
DBZ	Informix	Redbrick
Column oriented	Operational.	Informational.

Answer
technical blueprint
technical blueprint
Server
Large
Symmetric multi processing
Massively Parallel Processing
SMP
MPP
Non uniform Memory Architecture
Single
SMP
Shared Everything environment
Cluster
Shared disk systems
node
Two
virtual shared disk
virtual shared disk
NUMA

Data Extraction
Network Management
CPU
data shipping
Function Shipping
Infra Statement Operations
pipelining
Cost based optimizer
parallel index build
redundant array of inexpensive disks
RAPID
Query Management
Daemons
Backups
Data Extraction
Problem management
data Cleanup
Service level agreement
Operational system
Mission critical system
7 x 24 system
Backup
Cold backup
Single

Operational.
time
operational.
data warehouse
operational environment.
summary
Filtering
multidimensional databases.
Executive information system.
Fact table
read only.
Decision Support system.
5-10 years.
OLTP
Redbrick
Column oriented

KARPAGAM ACADEMY

Coin

DEPARTMENT OF COM

III B.Sc(CS) I

PART - A OBJECTIV

ONLINE EXAMINATION

UN

S.No	Questions	Choice 1	Choice 2
1	The hardware architecture of a datawarehouse is defined with in _____	technical blueprint	operational
2	The backup and security strategies are also determined during the _____	technical blueprint	operational
3	The ____ is a crucial part of the datawarehouse environment	Hardware	Server
4	Warehouse application generally require ____ hardware configuration	Small	Medium
5	SMP stands for _____	Symmetric multi processing	Symbol Multi Processing
6	MPP stands for _____	Massively Parallel Processing	mart Passive Processing
7	_____ Machine is a set of tightly coupled CPU that share memory and disk	SMP	MPP
8	___ machine is a set of loosely couples CPU's each of which has its own memory	SMP	MPP
9	NUMA Stands for	Non uniform Memory Architecture	Numeric Uniform Memory Architecture
10	SMP machine are the natural progression of the traditional ____ cpu midrange servers	Single	Multi
11	_____ gives the CPU bandwidth needed to support the large adhoc queries of decision support systems	SMP	MPP
12	SMP machine is also called _____	Shared multi environment	Shared Everything environment
13	____ is a set of loosely couples SMP machine connected by a high speed interconnect	Cluster	regression

14	each machine has its own CPU and memory but they share access to disk. It is called _____	S\Distributed lock Manager	Shared disk systems
15	Each machine in the cluster is called ____	head	arc
16	Most MPP systems allow a disk to be dual connected between two nodes	Five	Four
17	An extra layer of software called __	Virtual system drive	Virtual Suppoted Drive
18	VSD stands for ____	Virtual system drive	Virtual Suppoted Drive
19	A ____ machione is basically a tightly coupled cluster of SMP nodes with an extremely high speed interconnect	SMP	MPP
20	____ are any requests that cause data to be transferred out over the network	Data Cleaning	Data Integration
21	_____ is black art	System Hardware	Network Management
22	Heart of any computer is _____	Monitor	Mouse
23	_____ is where a process requests for the data to be shipped to the location where the process is running	Datamart	Function Shipping
24	____ is where the function to be performed is moved to locale of the data	Datamart	Function Shipping
25	_____ are separate operations that occur within the conbines of the SQL statement	Intra Statement Operations	Infra Statement Operations
26	_____ is where operations are carried out sequentially on a given piece of data	Infra Structure	pipelining
27	_____ uses stored statistics about the tables and their indexes to calculate the best strategy for executing the SQL statement	Infra Structure	pipelining
28	___ is an extension of the parallel query functionality	Parallel load	parallel technique
29	RAID stands for _____	rapid Array of interactive Disks	redundant array of inexpensive disks
30	____ technology is to provide resilence against disk failure	Pipelining	RAID

31	_____ is vital part of a daily operation	Query Management	Backups
32	The warehouse application managers are likely to have a no of service or background process running. This process is called	Query Management	Backups
33	_____ are usually avoided during the user day to avoid connections with the users queries	Query Management	Backups
34	_____- are subset of data that are offloaded from the server machine onto other machines	Data Cleaning	Data Integration
35	_____ is an area that needs to be slearly defined and documented	Problem management	Process management
36	_____ processing required will again vary from data warehouse to datawarehouse	data Integration	data Cleanup
37	SLA stands for ___	Service level agreement	Service low argument
38	___ is a system that has responsibilities to the operations of the business	Operational system	Mission critical system
39	_____ is a system that the business absolutely depends on to function	Operational system	Mission critical system
40	_ is system that needs to be available all day everyday, except for small periods of planned downtime	Operational system	Mission critical system
41	_____ is one of the most important regular operations carriedout on any system	Backup	Security
42	A _____ is a backup that is taken while the database is completel shut down	Partial backup	Cold backup
43	tape Stackers are a method of loading multiple tapes in a ___ tape drive	Single	Double
44	_____databases are owned by particular departments or business groups.	Informational.	Operational.
45	The key used in operational environment may not have an element of_____.	time	cost
46	Data can be updated in _____environment.	data warehouse.	data mining
47	Record cannot be updated in _____	OLTP	files

48	The source of all data warehouse data is the_____.	operational environment.	informal environment.
49	Data warehouse contains_____data that is never found in the operational environment.	normalized	informational
51	_____ is a data transformation process.	Comparison	Projection
52	MDDB stands for _____.	multiple data doubling.	multidimensional databases.
53	EIS stands for _____.	Extended interface system.	Executive interface system.
54	contains the measurement of business processes, and it contains foreign keys for the dimension tables.	Dimension table	Fact table
55	The data Warehouse is_____.	read only.	write only.
56	Expansion for DSS in DW is_____.	Decision Support system.	Decision Single System.
57	The time horizon in Data warehouse is usually _____	1-2 years.	3-4years.
58	The data is stored, retrieved & updated in _____.	OLAP	OLTP
59	_____ is the specialized data warehouse database.	Oracle	DBZ
60	Which storage has the drawback that storing or fetching a single tuple requires multiple I/O operations?	Row oriented	Column oriented

Y OF HIGHER EDUCATION

nbatore – 641 021.

MPUTER SCIENCE, CA & IT

BATCH 2015-2018

VE TYPE / MULTIPLE CHOICE QUESTIONS

ONE MARK QUESTIONS

NIT - VI

Choice 3	Choice 4	Answer
automation	strategy	technical blueprint
automation	strategy	technical blueprint
Software	Architecture	Server
Large	tiny	Large
Sign Multi Processing	Simultaneous Multi Process	Symmetric multi processing
Multi Pre Processing	Machinery Process Port	Massively Parallel Processing
SPM	MMS	SMP
SPM	MMS	MPP
Negative Uniform Memory Architecture	Null Uniform Memory Architecture	Non uniform Memory Architecture
Two	three	Single
SPM	MMS	SMP
Shared single environment	Shared Similar environment	Shared Everything environment
prediction	Scalability	Cluster

Shared data systems	Simultaneous data system	Shared disk systems
node	leaf	node
Two	Three	Two
Various shred disk	virtual shared disk	virtual shared disk
Various shred disk	virtual shared disk	virtual shared disk
NUMA	NMUA	NUMA
Data Extraction	data selection	Data Extraction
System Software	Hardware Management	Network Management
Speaker	CPU	CPU
data shipping	Meta data	data shipping
data shipping	Meta data	Function Shipping
Item Sequence Operation	item Sequence operation	Infra Statement Operations
parallel Query	Cost based optimizer	pipelining
parallel Query	Cost based optimizer	Cost based optimizer
parallel index build	parallel Query	parallel index build
redundant array of intensive disks	redundant array of interactive disks	redundant array of inexpensive disks
Cost optimizer	RAPID	RAPID

datafeeds	Daemons	Query Management
datafeeds	Daemons	Daemons
datafeeds	Daemons	Backups
Data Extraction	data selection	Data Extraction
hardware management	software management	Problem management
data transformation	None of the above	data Cleanup
Several level agreement	Server level argument	Service level agreement
7 x 24 system	7 x 24 x 52 system	Operational system
7 x 24 system	7 x 24 x 52 system	Mission critical system
7 x 24 system	7 x 24 x 52 system	7 x 24 system
Testing	None of the above	Backup
Hot backup	Online baclup	Cold backup
Multi	Standalone	Single
Both informational and operational.	Flat	Operational.
frequency	quality	time
operational.	informational	operational.
RDBMS	data warehouse	data warehouse

formal environment.	technology environment.	operational environment.
summary	denormalized	summary
Selection	Filtering	Filtering
multiple double dimension.	multi-dimension doubling.	multidimensional databases.
Executive information system.	Extendable information system.	Executive information system.
Multi table	data table	Fact table
read write only.	none	read only.
Data Storable System.	Data Support System.	Decision Support system.
5-6 years.	5-10 years.	5-10 years.
SMTP	FTP	OLTP
Informix	Redbrick	Redbrick
Operational.	Informational.	Column oriented

Register Number_____

[15CSU505A]

KARPAGAM ACADEMY OF HIGHER EDUCATION

KARPAGAM UNIVERSITY

Coimbatore-641021.

B.Sc COMPUTER SCIENCE

FIRST INTERNAL EXAMINATION - JULY 2017

Fifth Semester

DATA MINING AND DATAWAREHOUSING

Class :III B.Sc(CS) A & B Duration : 2 Hours

Date & Session : 21.7.2017 & N Maximum : 50 Marks

SECTION A – (20 X 1 = 20 Marks)
ANSWER ALL THE QUESTIONS

1. Finding hidden information from the database is known as _____
   a) Data warehouse b) Data mining c) Pattern Recognition d) Classification
2. The _____ is usually a subset of database in SQL.
   a) Input b) Query c) Output d) Process
3. The purpose of the algorithm is fit a model to the data is _____
   a) Model b) Preference c) Search d) Output
4. _____ Maps data into pre-defined groups or classes.
   a) Classification b) Clustering c) Regression d) Summarization
5. _____ is used to map a data item to a real valued predictive variable.
   a) Clustering b) Regression c) Time series Analysis d) Summarization
6. MSE stands for _____
   a) Mean Squared Error b) Main Sequence Error
   c) Multiple Sequence Estimation d) Main Squared Error
7. KDD stands for _____
   a)Knowledge Discovery in Database b)Knowledge Distribution in Database
   c) Knowledge Data Description d)Knowledge Data mining Discovery
8. KDD process consists of _____ steps.
   a) Five b) Four c) Three d) Ten
9. _____is the first step obtain the data from various databases.
   a) Selection b) Preprocessing c) Transformation d) Interpretation
10. _____ refers to the process of estimating a population parameter
    a) Point Estimation b) Co-relation c) Bayes Theorem d) Hypothesis testing
11. Each tuple in the database is assigns to exactly _____
    a) one class b) two class c) multiple class d) five class
12. OC stands for _____
    a) Over Customized b) Operating Characteristic
    c) Open Class d) Overfitting Characteristic
13. _____ curve is used in information retrieval to examine fall act versus recall
    a) OC curve b) ROC curve c) OCR curve d) CO curve
14. ROC stands for _____
    a) Receiver Operating Characteristic a) Read Operating Characteristic
    c) Real Operating Characteristic d) Random Operating Characteristic
15. A major issues associated with classification is that of _____
    a) Time Consumption b) Security c) Overfitting d)Human Intervention
16. _____ is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached.
    a) K-Means b) Squared Error c) Partitional algorithm d) Divisive
17. _____ algorithm is the most well-known association rule algorithm and is used in most commercial products.
    a) K-means b) Apriori c) Sampling d) Partitioning
18. The most familiar data mining technique is _____
    a) Statistical b) Clustering c) Estimation d) Distance Measure
19. The large item sets are also said to be _____
    a) Openward b) Upward c) Downward d) Closed
20. CDA stands for _____
    a) Count Distribution Algorithm b) Character Distribution Algorithm
    c) Count Database Algorithm d)Casual Descriptive Algorithm

SECTION –B (10 X 3 =30 Marks)
ANSWER ALL THE QUESTIONS

21. a) Explicate Data mining tasks.

    [OR]

    b) Write about Data mining Techniques.

    (i) Point Estimation (ii) Models based on Summarization

22. a) Explain in detail about K-means Clustering.

    [OR]

    b) Write about Clustering with Genetic Algorithm.

23. a) Write a note on Bayesian Classification and K-Nearest Neighbor.

    [OR]

    b) Explain Decision Trees and Neural Networks.

Register Number_____

[15CSU505A]

KARPAGAM ACADEMY OF HIGHER EDUCATION

Coimbatore-641021.

B.Sc COMPUTER SCIENCE

FIRST INTERNAL EXAMINATION - JULY 2017

Fifth Semester

DATA MINING AND DATAWAREHOUSING

Class :III B.Sc(CS) A & B **Duration : 2 Hours**

Date & Session : 21.7.2017 & N **Maximum : 50 Marks**

SECTION A – (20 X 1 = 20 Marks)
ANSWER ALL THE QUESTIONS

1. Finding hidden information from the database is known as _____
   a) Data warehouse **b) Data mining** c) Pattern Recognition d) Classification
2. The _____ is usually a subset of database in SQL.
   a) Input b) Query **c) Output** d) Process
3. The purpose of the algorithm is fit a model to the data is _____
   a) Model b) Preference c) Search d) Output
4. _____ Maps data into pre-defined groups or classes.
   a) Classification b) Clustering c) Regression d) Summarization
5. _____ is used to map a data item to a real valued predictive variable.
   a) Clustering **b) Regression** c) Time series Analysis d) Summarization
6. MSE stands for _____
   a) Mean Squared Error b) Main Sequence Error
   c) Multiple Sequence Estimation d) Main Squared Error
7. KDD stands for _____
   a)Knowledge Discovery in Database b)Knowledge Distribution in Database c) Knowledge Data Description d)Knowledge Data mining Discovery
8. KDD process consists of _____ steps.
   a) Five b) Four c) Three d) Ten
9. _____is the first step obtain the data from various databases.
   a) Selection b) Preprocessing c) Transformation d) Interpretation
10. _____ refers to the process of estimating a population parameter

a) Point Estimation b) Co-relation c) Bayes Theorem d) Hypothesis testing

11. Each tuple in the database is assigns to exactly _____

a) one class b) two class c) multiple class d) five class

12. OC stands for _____

a) Over Customized **b) Operating Characteristic**

c) Open Class d) over fitting Characteristic

13. _____ curve is used in information retrieval to examine fall act versus recall.

a) OC curve b) ROC curve c) OCR curve **d) CO curve**

14. ROC stands for _____

a) Receiver Operating Characteristic a) Read Operating Characteristic

c) Real Operating Characteristic d) Random Operating Characteristic

15. A major issues associated with classification is that of _____

a) Time Consumption b) Security **c) Over fitting** d)Human Intervention

16. _____ is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached.

a) K-Means b) Squared Error c) Partitional algorithm d) Divisive

17. _____ algorithm is the most well-known association rule algorithm and is used in most commercial products.

a) K-means **b) Apriori** c) Sampling d) Partitioning

18. The most familiar data mining technique is _____

a) Statistical **b) Clustering** c) Estimation d) Distance Measure

19. The large item sets are also said to be _____

a) Open ward b) Upward **c) Downward** d) Closed

20. CDA stands for _____

a) Count Distribution Algorithm b) Character Distribution Algorithm

c) Count Database Algorithm d)Casual Descriptive Algorithm

SECTION –B (10 X 3 =30 Marks)
ANSWER ALL THE QUESTIONS

21. a) Explicate Data mining tasks.

Basic Data mining tasks:

Classification:

- Classification maps data into predefined groups or classes.

- It is often referred to as supervised learning because the classes are determined before examining the data.
- Pattern recognition is a type of classification where an input pattern is classified into one several classes based on its similarity to these predefined classes.

Regression:

- Regression is used to map a data item to a real valued prediction variable.
- Regression assumes that the target data fit into some known type of function and then determines the best function of this type that models in the given data.

Time series analysis:

- With time series analysis the value of an attribute is examines as it varies over time. The values are obtained as evenly space time points.
- There are three basic functions performed in time series analysis. In one case, distance measures are used to determine the similarity between different time series.
- In second case, the structure of the line is examined to determine its behavior.
- In third application would be to use the historial time series plot to predict future values.
- Example: Stock Market

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.

Clustering:

- Clustering is similar to classification except that the groups are not rather defined by the data alone.
- Clustering is alternatively referred to as unsupervised learning or segmentation.
- Clustering groups similar data together into clusters.
- A special type of clustering is called segmentation. With segmentation a database is partitioned into disjointed groupings of similar tuples called segments.

Summarization

It maps data into subsets with associated simple descriptions. It extracts or derives representative information about the database. It is also called as Characterization, Generalization

Association rules:

Link Analysis uncovers relationships among data. It is also referred to as Affinity Analysis or Association Rules, refers to the data mining task of uncovering relationship among data. An association rule is a model that identifies specific types of data associations.

Sequence discovery:

Is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data are found to be related, but the relationship is based on time. It determines sequential patterns.

[OR]

b) Write about Data mining Techniques.

Data mining technique:

A statistical perspective on data mining:

a) Point Estimation:

Point Estimate: It refers to the process of estimate a population parameter. May be made by calculating the parameter for a sample. May be used to predict value for missing data.

The bias of an estimator is the difference between the expected value of the estimator and the actual value:

$$Bias = E(\hat{\Theta}) - \Theta$$

An unbiased estimator is one whose bias is 0.

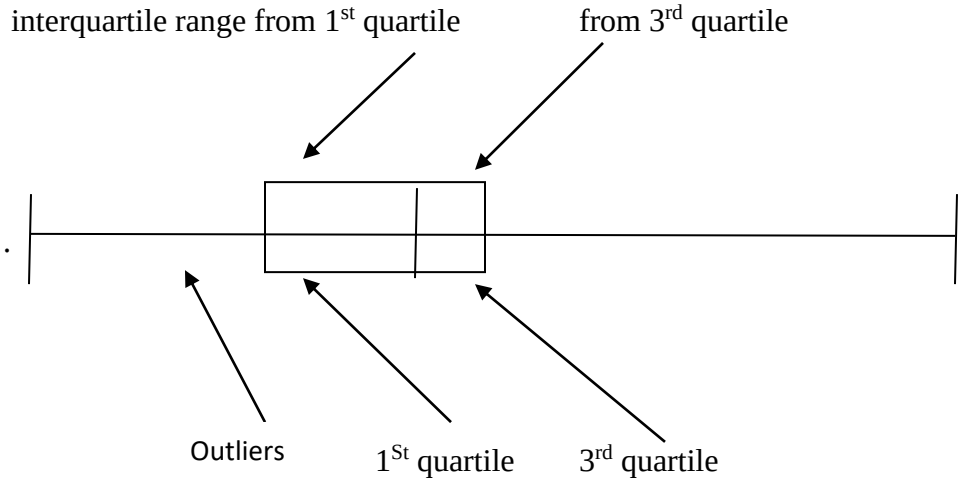
b) Models based on Summarization

There are many basic concepts that provide an abstraction and summarization of the data as a whole. Fitting a population to a specific frequency distribution provides an even better model of the data.

There are also many well known techniques to display the structure of the data graphically.

For ex,A histogram shows the distribution of the data. A box plot more sophisticated technique that illustrates several different features of the population at once.

Smallest value within 1.5 largest values within 1.5 interquartile ranges

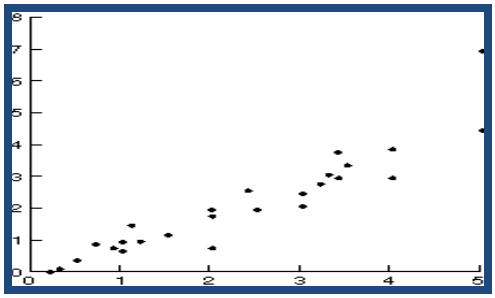


Box plot example

The total range of the data values is divided into four equal parts called quartiles. The box in center of the figure shows the range between first, second, third quartiles.

The lines in the box show the median. The lines exceeding from either end of the box are the values that are a distance of 1.5 of the inter quartile range from the first and third quartiles.

Scatter Diagram:



22. a) Explain in detail about K-means Clustering.

K-Means clustering:

K-means is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. High degree of similarity among elements in a cluster is obtained.

Given a cluster $K_i=\{t_{i1},t_{i2},\ldots,t_{im}\}$, the ***cluster mean*** is $m_i = (1/m)(t_{i1} + \ldots + t_{im})$

Algorithm:

Input:
  $D = \{t_1, t_2, ..., t_n\}$ // Set of elements
  A // Adjacency matrix showing distance between elements.
  k // Number of desired clusters.
Output:
  K // Set of clusters.
K-Means Algorithm:
  assign initial values for means $m_1, m_2, ..., m_k$;
  repeat
    assign each item t_i to the cluster which has the closest mean ;
    calculate new mean for each cluster;
  until convergence criteria is met;

Example:

- Given: {2,4,10,12,3,20,30,11,25}, k=2
- Randomly assign means: $m_1=3, m_2=4$
- $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$, $m_1=2.5, m_2=16$
- $K_1=\{2,3,4\}$, $K_2=\{10,12,20,30,11,25\}$, $m_1=3, m_2=18$
- $K_1=\{2,3,4,10\}$, $K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$
- $K_1=\{2,3,4,10,11,12\}$, $K_2=\{20,30,25\}$, $m_1=7, m_2=25$
- Stop as the clusters with these means are the same.

The time complexity of K-means is O(tkn) where t is the number of iterations. K-means finds a local optimum and may actually miss the global optimum. K-means does not work on categorical data because the mean must be defined on the attribute type. One variation of K-means, k-modes does handle categorical data. Instead of using means, it uses modes.

Although the K-means algorithm often produces good results, it is not time efficient and does not scale well.

[OR]

b) Write about Clustering with Genetic Algorithm.

Clustering with Genetic Algorithm

There has been clustering technique based on the use of genetic algorithms. One simple approach would be to use a bit-map representation for each possible cluster.

So given a set of four items,{A,B,C,D}, we would represent one solution to creating two clusters as 1001 and 0110. This represents two clusters {A,D} and {B,C}.

Example:

- A database contain eight items {A,B,C,D,E,F,G,H}
- Randomly choose initial solution. Three clusters are:

  {A,C,E} {B,F} {D,G,H} which represented by

  10101000, 01000100, 00010011

- Suppose crossover at point four and choose 1st and 3rd individuals:

10100011, 01000100, 00011000

Algorithm:

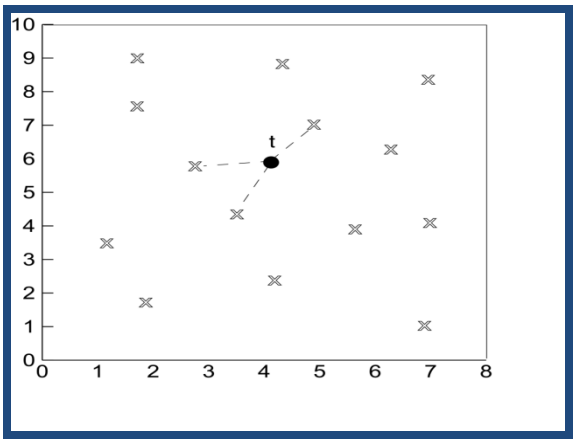
```
Input:
   D = {t1, t2, ..., tn}    // Set of elements
   k      // Number of desired clusters.
Output:
   K      // Set of clusters.
GA Clustering Algorithm:
   randomly create an initial solution;
   repeat
      use crossover to create a new solution;
   until termination criteria is met;
```

The above algorithm shows one possible iterative refinement technique for clustering thar uses a genetic algorithm. A new solution is generated from the previous solution using cross over and mutation operation. Our algorithm shows only crossover.

23. a) Write a note on Bayesian Classification and K-Nearest Neighbor.

K Nearest Neighbour:

One common classification scheme based on the use of distance measures is that of the K-Nearest neighbor (KNN), The KNN technique assumes that the entire training set includes not only the set but also the desired classification of each item. New item placed in class that contains the most items from this set of K closest items.



Classification using KNN

Here the points in the training set are shown and K=3. The three closest items in the training set are shown; t will be placed in the class to which most of these are members.

```
Input:
   D       //Training data
   K       //Number of neighbors
   t       //Input tuple to classify
Output:
   c       //Class to which t is assigned
KNN Algorithm:
           //Algorithm to classify tuple using KNN
   N = ∅;
                //Find set of neighbors, N, for t
   foreach d ∈ D do
      if | N |≤ K then
         N = N ∪ d;
      else
         if ∃ u ∈ N such that sim(t, u) ≥ sim(t, d) then
            begin
               N = N − u;
               N = N ∪ d;
            end
            //Find class for classification
   c = class to which the most u ∈ N are classified;
```

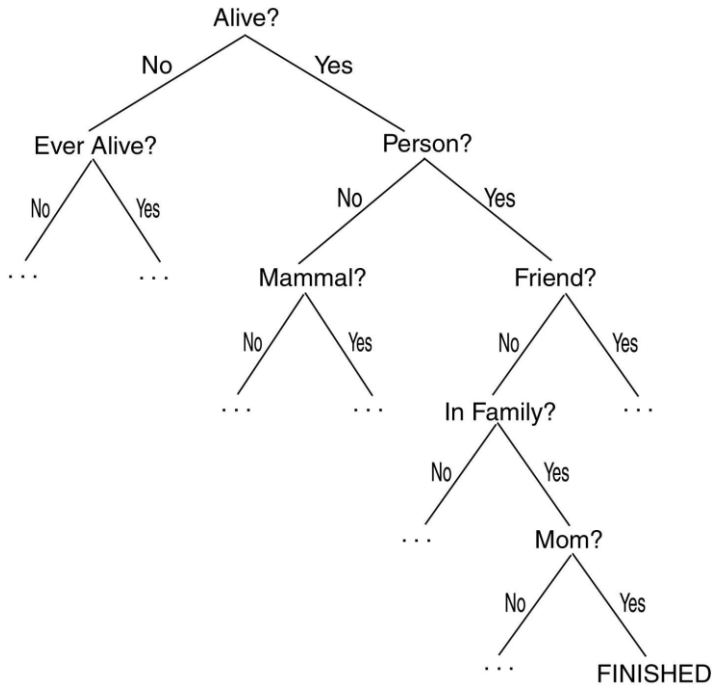
T to represent the training data. Since each tuple to be classified must be compared to each element In the training data, if there are q elements in the training set, this is O(q). Given n elements to be classified, this becomes an O(nq) problem. Training data are of a constant size, this can then be viewed as an O(n) problem.

[OR]

b) Explain Decision Trees and Neural Networks.

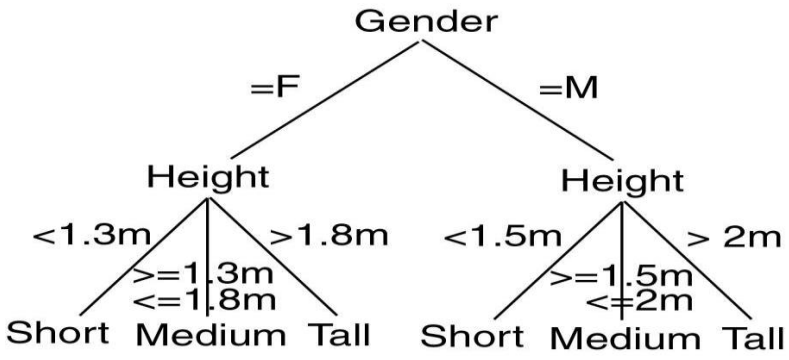
Decision Trees:

Decision tree is a predictive modeling technique used in classification, clustering and prediction tasks. Decision trees use a "divide and conquer" technique to split the problem search space into subsets.



Tree where the root and each internal node is labeled with a question. The arcs represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem. Popular technique for classification; Leaf node indicates class to which the corresponding tuple belongs.

Decision Tree Example:



A Decision Tree Model is a computational model consisting of three parts:

1. Decision Tree

2. Algorithm to create the tree

3. Algorithm that applies the tree to data

Creation of the tree is the most difficult part. Processing is basically a search similar to that in a binary search tree (although DT may not be binary).

Decision Tree Algorithm:

```
Input:
   T      //Decision Tree
   D      //Input Database
Output:
   M      //Model Prediction
DTProc Algorithm:
            //Illustrates Prediction Technique using DT
   for each t ∈ D do
      n = root node of T;
      while n not leaf node do
         Obtain answer to question on n applied t;
         Identify arc from t which contains correct answer;
         n = node at end of this arc;
      Make prediction for t based on labeling of n;
```

Advantages:

- Easy to understand.
- Easy to generate rules

Disadvantages:

- May suffer from over fitting.
- Classifies by rectangular partitioning.
- Does not easily handle nonnumeric data.
- Can be quite large – pruning is necessary.

Neural networks:

Based on observed functioning of human brain.(Artificial Neural Networks (ANN) Our view of neural networks is very simplistic. We view a neural network (NN) from a graphical viewpoint. Alternatively, a NN may be viewed from the perspective of matrices. Used in pattern recognition, speech recognition, computer vision, and classification.

Neural Network (NN) is a directed graph F=<V,A> with vertices V={1,2,…,n} and arcs A={<i,j>|

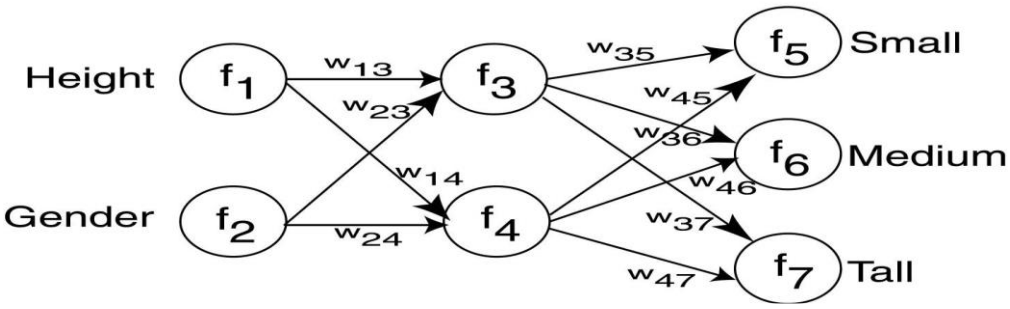
1<=i,j<=n}, with the following restrictions:

V is partitioned into a set of input nodes, VI, hidden nodes, VH, and output nodes, VO.

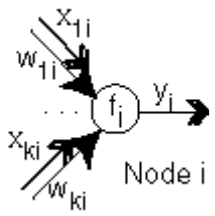
The vertices are also partitioned into layers Any arc <i,j> must have node i in layer h-1 and

node j in layer h. Arc <i,j> is labeled with a numeric value wij. Node i is labeled with a function fi.

Neural Network Example:



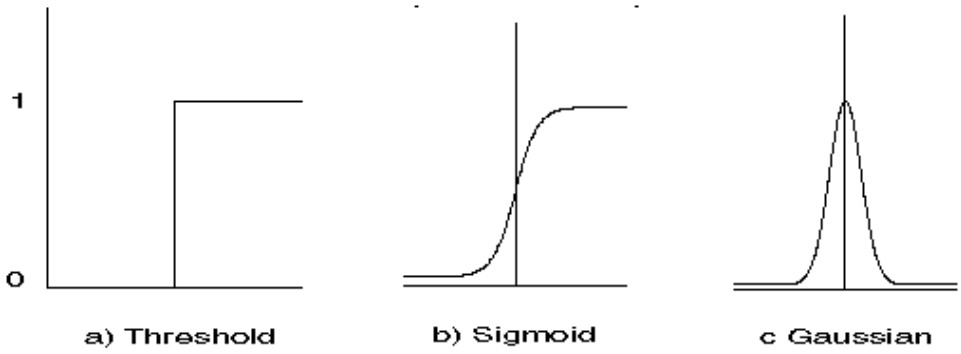
Neural Network node:



$$y_i = f_i(\sum_{j=1}^{k} w_{ji}\ x_{ji}) = f_i([w_{1i}...w_{ki}] \begin{bmatrix} x_{1i} \\ ... \\ x_{ki} \end{bmatrix})$$

NN Activation Functions

Functions associated with nodes in graph. Output may be in range [-1,1] or [0,1]



Neural Network Activation Functions:

Linear:

$$f_i(S) = c\,S$$

Threshold or Step:

$$f_i(S) = \left\{ \begin{array}{ll} 1 & if S > T \\ 0 & otherwise \end{array} \right\}$$

Ramp:

$$f_i(S) = \left\{ \begin{array}{ll} 1 & if S > T_2 \\ \frac{S-T_1}{T_2-T_1} & if T_1 \leq S \leq T_2 \\ 0 & if S < T_1 \end{array} \right\}$$

Sigmoid:

$$f_i(S) = \frac{1}{(1+e^{-c\,S})}$$

Hyperbolic Tangent:

$$f_i(S) = \frac{(1-e^{-S})}{(1+e^{-c\,S})}$$

Gaussian:

$$f_i(S) = e^{\frac{-S^2}{v}}$$

A Neural Network Model is a computational model consisting of three parts:

Neural Network graph learning algorithm that indicates how learning takes place. Recall techniques that determine how information is obtained from the network. We will look at propagation as the recall technique.

Advantages:

- Learning
- Can continue learning even after training set has been applied.
- Easy parallelization
- Solves many problems

Disadvantages:

- Difficult to understand
- May suffer from over fitting
- Structure of graph must be determined a priori.
- Input values must be numeric.
- Verification difficult.